ABSTRACT

ROY, ARKAPRAVA. Bayesian Methods for High Dimensional Models in Brain Imaging. (Under the direction of Subhashis Ghosal and Ana-Maria Staicu.)

In this thesis, we have worked on several methodology developments with the focus on brain imaging data. There is a general introduction to the materials covered in this thesis in Chapter 1.

In Chapter 2, we study possible relations between the structure of the connectome, white matter connecting different regions of the brain and Alzheimer disease. Regression models in covariates including age, gender, and disease status for the extent of white matter connecting each pair of regions of the brain are proposed. Subject inhomogeneity is also incorporated in the model through random effects with an unknown distribution. Due to a large number of pairs of regions, we also adopt a dimension reduction technique through graphon (Lovasz and Szegedy (2006)) functions reducing functions of pairs of regions to functions of regions. The connecting graphon functions are considered unknown but assumed smoothness allows putting priors of low complexity on them. We pursue a nonparametric Bayesian approach by assigning a Dirichlet process scale mixture of zero mean normal prior on the distributions of the random effects and finite random series of tensor products of B-splines priors on the underlying graphon functions. Presence of interaction between region effects with covariates with sparsity restrictions is allowed in the model. Posterior computation techniques are developed. Posterior consistency of the proposed Bayesian approach is established under the asymptotic regime of increasing number of subjects and a fixed number of brain regions. Performance of the proposed Bayesian method is compared with ANCOVA models through simulations under both well-specified and misspecified cases. The proposed Bayesian approach is applied to a dataset obtained from ADNI consisting of 100 subjects and 83 brain regions. Bayes estimates of connectome strengths for a person with a given profile are obtained which can be used as a benchmark for the actual connectome data of a future patient with the same profile.

In Chapter 3, we examine the effects of age, genetic variation and Alzheimer disease on atrophy of the brain regions. In the real data analysis section, we add subject-specific random effect to capture subject inhomogeneity. A nonparametric single index Bayesian model is proposed to study the data with B-spline series prior on the unknown functions and Dirichlet process scale mixture of zero mean normal prior on the distributions of the random effects. The posterior rate of contraction of the proposed model without the random effect is established for a fixed number of regions and time points with increasing sample size. A new Bayesian estimation procedure for high dimensional single index model is introduced in this paper. Performance of the proposed Bayesian method is compared with the corresponding least square estimator in the
linear model with lasso penalization on the high dimensional covariate. The proposed Bayesian method is applied to a dataset of 748 individuals with 620,901 SNP and 8 other covariates for each individual.

In Chapter 4, we study the problem of sparse signal detection on a spatial domain. We propose a novel approach to model continuous signals that are sparse and piecewise smooth as the product of independent Gaussian processes (PING) with a smooth covariance kernel. The smoothness of the PING process is ensured by the smoothness of the covariance kernels of Gaussian components in the product, and sparsity is controlled by the number of components. The bivariate kurtosis of the PING process shows more components in the product results in a thicker tail and sharper peak at zero. The simulation results demonstrate the improvement in estimation using the PING prior to Gaussian process (GP) prior for different image regressions. We apply our method to a longitudinal MRI dataset to detect the regions that are affected by multiple sclerosis (MS) in the greatest magnitude through an image-on-scalar regression model. Due to the huge dimensionality of these images, we transform the data into the spectral domain and develop methods to conduct computation in this domain. In our MS imaging study, the estimates from the PING model are more informative than those from the GP model.
Bayesian Methods for High Dimensional Models in Brain Imaging

by
Arkaprava Roy

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics
Raleigh, North Carolina
2018

APPROVED BY:

Soumendra Nath Lahiri
Brian J. Reich
Subhashis Ghoshal
Co-chair of Advisory Committee
Ana-Maria Staicu
Co-chair of Advisory Committee
DEDICATION

To my parents and all well wishers.
BIOGRAPHY

Arkaprava Roy was born on 10th of October 1991 in Kolkata, India. After finishing his schooling at Behala Arya Vidyamandir (till 10th standard) and Jodhpur Park Boys’ (for 11 and 12), he went to Indian Statistical Institute, Kolkata to pursue Bachelors degree and Masters degree in Statistics. Then he joined the Department of Statistics at North Carolina State University as a Ph.D. student.
I would like to extend my gratitude towards my advisors Dr. Subhashis Ghoshal and Dr. Ana Maria Staicu for giving exposure and guidance to lay the foundation for my research career. Their vast experience has helped me to finish my research works smoothly. I would like to thank Dr. Brian Reich for being part of the advisory committee as well as mentoring one of the research works of my thesis and also Dr. Soumendra Nath Lahiri and Dr. Ralph Smith to be part of the advisory committee and giving helpful comments on my work.

I met great people in this statistics department at North Carolina State University. All the support staffs have been extremely helpful so that everything went smoothly. I would like to thank Dr. Howard Bondel, Lanakila Alexander and Dr. Wenbin Lu for their help in any official issues as well as Terry Byron and Chris Waddell for any computing issues that I came across during my time here in the department. The computing facility in our department is not very best. But computing liaisons helped us at every stage. Also, I would like to thank the office of international students (OIS) for their help in any bureaucratic issues.

These four years have definitely been one of the best four years of my life. My stay at the USA has been very smooth. I am very fortunate to find a family with a wonderful group of friends including Sayan Banerjee, Sapna Rao, Suman Chakraborty, Sujatro Chakladar, Mitul Biswas, Priyam Das, Debraj Das, Arnab Hazra, Indranil Sahoo, Mounita Chakraborty, Sohini Raha, Arkopal Choudhury, Arnab Chakraborty, Suman Majumdar, Rahul Ghoshal, Dhrobojyoti Ghosh, Salil Konar, Aniket Bera, Abhishek Banerjee, Sayak Mukherjee, Abhra Sarkar, Rimli Sengupta, Jyotiska Dutta and Shalini Choudhury.

Finally, I owe all of my achievements to my parents. Their constant guidance and support have made it possible for me to stay focussed and reach my target.
# TABLE OF CONTENTS

LIST OF TABLES .............................................................. vii

LIST OF FIGURES ........................................................... viii

LIST OF SYMBOLS ........................................................... x

Chapter 1 Introduction ...................................................... 1

Chapter 2 Bayesian Modeling of the Structural Connectome for Studying Alzheimer Disease ................................................. 4
  2.1 Introduction ................................................................... 4
  2.2 Data description and modeling ........................................ 6
  2.3 Prior specification and posterior computation ................. 10
    2.3.1 Prior specification ..................................................... 10
    2.3.2 Posterior updating .................................................... 12
    2.3.3 Tuning ................................................................. 16
  2.4 Simulation ................................................................. 17
  2.5 Real data analysis ......................................................... 20
  2.6 Large-sample Properties ............................................. 23
    2.6.1 Proof ...................................................................... 30
  2.7 Conclusions and discussion ........................................... 36

Chapter 3 High dimensional Single Index Bayesian Modeling of the Brain Atrophy over time ......................................................... 38
  3.1 Introduction ............................................................. 38
  3.2 Data description and modeling ....................................... 40
  3.3 Prior specification ....................................................... 42
  3.4 Computation ............................................................. 45
  3.5 Simulation ............................................................... 47
  3.6 Large-sample Properties ............................................ 50
  3.7 Real data analysis ....................................................... 58
    3.7.1 Modification of the model for real data application ....... 58
  3.8 Conclusions and discussion .......................................... 61
  3.9 Acknowledgments ....................................................... 64
  3.10 Proofs ......................................................................... 66

Chapter 4 The Product Independent Gaussian Process Prior for Continuous, Sparse and Piecewise Smooth Signal Estimation ......................... 71
  4.1 Introduction ............................................................ 71
  4.2 The modeling framework .............................................. 73
    4.2.1 PING process ...................................................... 74
    4.2.2 Error distribution and Matérn ................................... 76
  4.3 Extension to other image regression models .................... 79
### List of Tables

Table 2.1 Comparison of estimation accuracy for $n = 500$ in well-specified case ............... 18
Table 2.2 Comparison of estimation accuracy for $n = 500$ in misspecified case ................. 23
Table 2.3 Demographic table ................................................................................. 23

Table 3.1 For non-linear case ............................................................................. 48
Table 3.2 For linear case ....................................................................................... 48
Table 3.3 Demographic table ................................................................................. 58
Table 3.4 Estimates of covariates for Total Brain for slope from non-linear model ....... 60
Table 3.5 Estimates for Ventricles for slope .......................................................... 60
Table 3.6 Estimates for Left Hippocampus for slope .............................................. 61
Table 3.7 Estimates for Right Hippocampus for slope ........................................... 61
Table 3.8 Estimates for Left inferior lateral ventricle for slope ......................... 62
Table 3.9 Estimates for Right inferior lateral ventricle for slope ....................... 62
Table 3.10 Estimates for Left Medial Temporal for slope ....................................... 63
Table 3.11 Estimates for Right Medial Temporal for slope ..................................... 63
Table 3.12 Estimates for Left Inferior Temporal for slope ...................................... 64
Table 3.13 Estimates for Right Inferior Temporal for slope ..................................... 64
Table 3.14 Estimates for Left Fusiform for slope ................................................... 65
Table 3.15 Estimates for Right Fusiform for slope ............................................... 65
Table 3.16 Estimates for Left Entorhin for slope .................................................... 66
Table 3.17 Estimates for Right Entorhin for slope ................................................. 66
Table 3.18 Estimates of covariates for Total Brain for slope from linear model ........ 67
Table 3.19 SNPs selected after variable selection from non-linear model along with gene names and main effect on slope ................................................................. 67

Table 4.1 Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type 1 error for the slope of the image-on-scalar simulation with different SNRs for Gaussian, PING 3 and PING 5 as choices of prior .................................................. 85
Table 4.2 Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type 1 error for the slope of the image-on-image simulation with different SNRs for Gaussian, PING 3 and PING 5 as choices of prior .................................................. 88
Table 4.3 Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type 1 error for the slope of the scalar-on-image regression model with different true variances for Gaussian, PING 3 and PING 5 as choices of prior .................................................. 91
Table 4.4 Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power and Type1 error for the slope of the scalar-on-image simulation with different true variances for soft-thresholded Gaussian process (STGP) and PING 8 as choices of prior .................................................. 92
LIST OF FIGURES

Figure 2.1  Heatmap of number in log-scale and average width of white matter fibers where each pixel represents each pair of region. .................................................. 7
Figure 2.2  Heatmap of truth along with Bayesian methods and ANCOVA estimates for sample size 500 in the well specified case with 20 nodes. ................................. 19
Figure 2.3  Plot of squared bias and variance for different parameters against sample sizes for Bayesian method and ANCOVA estimates in the well-specified case. 20
Figure 2.4  Heatmap of truth along with Bayesian methods and ANCOVA estimates for sample size 500 in the misspecified case with 20 nodes. ................................. 21
Figure 2.5  Plot of squared bias and variance for different parameters against sample sizes for Bayesian method and ANCOVA estimates in the misspecified case. 22
Figure 2.6  Effect of different covariates on connection width, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge. .................................................. 24
Figure 2.7  Effect of covariates on connection-probability, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge. .................................................. 25
Figure 2.8  Effect of covariates on number of connections, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge. .................................................. 26

Figure 3.1  Anatomic parcellation of cortical surface from different angles showing brain regions used for analysis. .................................................. 38
Figure 3.2  Spike distribution with $M_1 = 0.01$ and $M_2 = 10$. ........................................ 43
Figure 4.1  Comparison of Gaussian density with PING-3 and PING-5 densities. ............... 75
Figure 4.2  Conditional density plot of $\beta_2$ given different values of $\beta_1$ for different correlations and components for bivariate realization $(\beta_1, \beta_2)$ at two locations. .. 76
Figure 4.3  3-D image of the true slope $\beta(s)$ for the image-on-scalar regression simulation study. ................................. 84
Figure 4.4  Plot of the estimates for the slope of the image-on-scalar regression simulation with different signal to noise ratios (SNRs) for Gaussian, Product of independent Gaussian with 3 components (PING 3) and PING 5 as choices of prior. .................................................. 86
Figure 4.5  Plots of the truth for last five regression coefficients (first five were zero) from second simulation corresponding to image-on-image regression. ................................. 87
Figure 4.6  Comparison plot of the estimates for the slope of the scalar-on-image simulation with different true variances for Gaussian, PING 3 and PING 5 as choices of prior. .................................................. 90
Figure 4.7  Variation in MSE of estimated slope for varying number of components $q$ for scalar-on-image simulation. .................................................. 92
Figure 4.8  Data for each visit of the middle slice of the MRI image for 12 visits of the selected subject. .................................................. 95
Figure 4.9  Estimated slope $\beta(s)$ of the middle slice using different priors along with the color scale.
LIST OF SYMBOLS

\((A)_{i,j}\) : a matrix \(A\) with \((i,j)\)th element being \(A_{i,j}\).

\(A^T\) : transpose of the matrix \(A\).

\(A_{i,:}\) : the \(i\)th row of \(A\).

\(A_{:,j}\) : the \(j\)th column of \(A\).

\(\text{vec}(A)\) : the vector obtained by stacking the columns of the matrix \(A\) one over another.

\(A \otimes B\) : the Kronecker product between \(A\) and \(B\).

\(I_p\) : the identity matrix of order \(p\).

\(\text{maxeig}(A)\) : the maximum eigenvalue of the matrix \(A\).

\(\text{mineig}(A)\) : the minimum eigenvalue of the matrix \(A\).

\(\|x\| = \left( \sum_{i=1}^{p} x_i^2 \right)^{1/2}\), the \(L_2\) norm of the vector \(x\).

\(a \cdot b\) : element wise product of two vectors \(a\) and \(b\).

\(\mathbb{1}_A(\cdot)\) : indicator function of the set \(A\).

\(a_n = o(b_n)\) : \(a_n/b_n \to 0\) as \(n \to \infty\) for numerical sequences \(a_n\) and \(b_n\).

\(a_n = O(b_n)\) : \(a_n/b_n\) is bounded.

\(a_n \asymp b_n : a_n = O(b_n)\) and \(b_n = O(a_n)\).

\(a_n \lesssim b_n : a_n = O(b_n)\).

\(a_n \gtrsim b_n : b_n = O(a_n)\).

\(a_n \ll b_n : a_n = o(b_n)\).

\(o_P(1)\) : a sequence of random variables which converges in probability to zero.

\(O_P(1)\) : a sequence of random variables bounded in probability.

\(\mathbb{E}(\cdot)\) : the mean vector of a random vector.

\(\text{Var}(\cdot)\) : the dispersion matrix of a random vector.

\(d_H(P,Q) = \sqrt{\int \left( \sqrt{p} - \sqrt{q} \right)^2}\), the Hellinger distance between \(P\) and \(Q\).

\(K(p,q) = p \log \frac{p}{q}\), the Kullback-Leibler distance between \(P\) and \(Q\).

\(\mathcal{L}\) : class of real-valued smooth continuous functions on \([0,1]^2\).

\(a \land b\) : the minimum of two real numbers \(a\) and \(b\).

\(|r| = \sum_{j=1}^{s} r_j\) for a vector \(r = (r_1,\ldots,r_s)^T\).
Chapter 1

Introduction

Developing methodologies for high dimensional datasets are new challenges in the field of statistics. There are many sources of occurrences of such datasets like genomics, neuroscience, finance, earth sciences etc. We apply our methods to brain image data. Thus our methods are mostly motivated by application in brain imaging data but can be useful for other kinds of high dimensional datasets as well.

There are different kinds of neurological datasets available currently. There are data of structural brain, human brain connectome, magnetic resonance imaging (MRI) data, functional magnetic resonance imaging (fMRI) data etc. Some diseases are known to affect the human brain. The motivation to study these datasets is to understand the direct or indirect effect of these diseases on the human brain. Potentially there are three kinds of brain image datasets, namely brain connectome, brain anatomy and neuronal signal images.

The brain connectome is a comprehensive map of all the neural connection within the brain. It is very difficult to construct the connectome from the MRI images. Several researchers are using these connectome datasets to answer various kinds of questions. There are many scientific questions that the researchers are trying to answer as well several novel statistical methodologies are being developed for these complex network datasets. This kind of data is used to study diseases with the neurodegenerative disorder like Parkinson disease, Alzheimer disease etc. In this thesis, we focus on the studying brain connectome in connection with Alzheimer disease.

Alzheimer disease (AD) is a neurodegenerative disorder that affects approximately 5 million people in the US and 30 million people worldwide, with incidence increasing with age. There is no available treatment which modifies the disease once symptoms of cognitive impairment or dementia are clinically apparent (Ferri et al. (2005)). Current thought is that detecting pathologic changes in the brain before the development of clinical symptoms will allow for successful treatment. It has been observed that the white matter connections between regions of the brain that are vital for brain’s functioning are affected by Alzheimer disease, in particular in relation
to amyloid plaque burden, which is one of the pathological hallmarks of AD and has been shown to become elevated years before the onset of clinical symptoms (Prescott et al. (2014)). It may be that these structural connections are part of the anatomic substrate underlying cognition. In Chapter 2 of this thesis, we model changes in the brain connectome caused by the onset of the disease using a graphical structure.

Studying brain anatomy and its changes over time in a healthy or a disease patients are also some interesting works that researcher from various domains is interested in. One such issue is with brain atrophy. The shape and size of the brain change with time and these variations change across individuals. In Chapter 3 we study this problem. It is believed to have a prolonged preclinical phase initially characterized by the development of silent pathologic changes when patients appear to be clinically normal, followed by mild cognitive impairment (MCI) and then dementia (AD) (Petrella (2013)). Apart from its manifestation in the impairment of cognitive abilities, disease progression also produces a number of structural changes in the human brain, which includes the deposition of amyloid protein and the shrinkage or atrophy for certain regions of the brain over time (Thompson et al. (2003)). Previous studies have shown that the rate of brain atrophy is significantly modulated by a number of factors, such as gender, age, baseline cognitive status and most markedly, allelic variants in the APOE gene (Hostage et al. (2014)). In this paper, we examine if any other genes are also implicated in modulating the rate of brain atrophy. This analysis represents a technical challenge because the genomic data is high dimensional and needs to be incorporated into a model for longitudinal progression of brain volumes measured in multiple parts of the brain. We examine the effects of age, gender, APOE genes, genetic variation and Alzheimer disease on atrophy of the brain regions. To summarize the effects of the said covariates a nonparametric single model is proposed with two separate inputs for high dimensional and low dimensional covariates separately. We develop a Bayesian estimation procedure to get the estimates of the model parameters. The proposed Bayesian method is applied to a dataset of 748 individuals with 620,901 SNP and 8 other covariates for each individual.

In Chapter 4, we discuss linear regression models for two or three dimensional image responses, image covariates, or both, in which the signal is assumed to be continuous, sparse and piecewise smooth. The methodological development is motivated by a study of multiple sclerosis using magnetic resonance imaging (Sweeney et al., 2016; Pomann et al., 2016; Mejia et al., 2016), where subjects with multiple sclerosis (MS) are imaged repeatedly over multiple hospital visits, and the objective is to identify the brain regions that are damaged over time. Although a healthy brain would not change much during the study period, a diseased brain is expected to exhibit changes in a small number of regions of interest that are associated with the disease. This is an example of image-on-scalar regression, in which the signal is desired to
be continuous, sparse and piecewise smooth.

Apart from the applications discussed in this thesis, the proposed Bayesian methods are applicable in a variety of datasets.
Chapter 2

Bayesian Modeling of the Structural Connectome for Studying Alzheimer Disease

This chapter’s material is a joint work with Subhashis Ghosal, Jeffrey Prescott and Kingshuk Roy Choudhury.

2.1 Introduction

Our study is performed using data obtained by Alzheimer Disease Neuroimaging Initiative (ADNI – adni.loni.usc.edu). Using T1-weighted magnetic resonance (MR) images from ADNI, the cortex of the brain is divided into several regions using a standard anatomic atlas. These regions are connected by white matter fibers, identified using diffusion tensor imaging (DTI) MR. The graphical representation of these white matter connections between cortical regions is referred to as the connectome. It is thought that some of these connections between brain regions become weakened over time due to the AD. Some other factors like age or sex might also affect this connectome, as well as subject-specific random effects. We model the connectome using graph theoretic metrics, accounting for patients specific effects and implement a Bayesian analysis.

We consider a connectome with 83 cortical regions in the brain. Mathematically, the connectome can be viewed as a graph \((V, E)\), where \(V\) denotes the set of nodes or vertices standing for brain regions and \(E\) for the edges between pairs of vertices whenever present. As in a graphical model, edges are marked with certain measurements. In our context, the measurements consist of observing the presence or absence of white matter fibers connecting two regions, the number
of white matter fiber and the mean width of white matter fibers between them when at least a connecting white matter fiber exists. Our aim is to identify the pairs of regions between which the connections are affected by the disease and other covariates differently than in other pairs in the connectome by the covariates. The aspects of the presence of a connection, the number of connections and the mean width are modeled respectively by a binary regression model with a probit link, a Poisson regression model with an exponential link and a normal regression model. Interactions of covariates with pairs of regions are considered, but a sparsity condition is imposed in the analysis through the prior, which imply that for most pairs, these effects are the same and possibly non-zero while for a few of pairs, there could be different effects. In fact, detecting such region-pairs is the main objective of our analysis. Because the number of region-pairs is prohibitively high, it leads to a very high dimensional regression model if a naive approach is taken. Hence we use a dimension reduction technique that introduces a latent variable for each region and expresses functions of region-pairs in terms of a single smooth unknown function of each pair of latent variables as in a graphon model (Lovasz and Szegedy (2006)). If \((a_{jk} : j, k = 1, \ldots, J)\) is an array of parameters, then we model \(a_{jk} = g(u_j, u_k)\), where \(u_1, \ldots, u_J\) are latent variables and \(g\) is a function of two arguments. Symmetry of the matrix \(\{(a_{jk})\}\) is respected if \(g\) is symmetric in its arguments. The original motivation for this representation is that if \(a_{jk}, j, k = 1, \ldots, J\) are random variables such that the matrix \(\{(a_{jk})\}\) is distributionally invariant under permutations of rows and columns, then \(a_{jk} = g(u_j, u_k)\) for some latent variables \(u_1, \ldots, u_J\) that can be assumed to be uniformly distributed and a function \(g\) called a graphon function. In the present context, if the parameters \(a_{jk}, j, k = 1, \ldots, J\) are treated as random, then such row and column wise exchangeability conditions are natural non-informativeness conditions. Given the graphon function, thus the strength of connections is determined by only 83 latent variables linked with each region instead of being \(\binom{83}{2} = 3486\) making huge computational savings. The graphon function is also treated as the unknown without any specific parametric form and is nonparametrically estimated from the data. More specifically, we put a finite random series prior based on tensor products of B-splines, where the coefficients are given appropriate prior distributions. Finite random series priors are widely used in the literature to construct priors on various functions and are systematically studied, for instance, in Shen and Ghosal (2015), but it seems to have not been used before for putting prior distributions on graphon functions. The assumed smoothness of the graphon function helps keep the number of basis function required for the basis expansion relatively small. This is because to approximate a function in \([0, 1]^d\) of smoothness index \(f\) by a finite series within accuracy \(\epsilon\), one needs to use only \(O(\epsilon^{-d/f})\) many elements of standard bases like polynomials, B-splines or wavelets, that is fewer functions are needed for smoother functions. It is sensible to think that the covariates affect most region-pairs in the same way except for a few ones which
typically show more pronounced effects. This feature can be captured by a prior which introduces sparsity in the coefficients by making most of the latent variables cluster at a point with high probability. To reduce additional computational burden arising out of selection of active latent variables which do not cluster at the benchmark point, we use a continuous shrinkage prior instead of more traditional point mass spike and slab prior such as the horseshoe prior (Carvalho et al. (2010)), normal-gamma prior (Griffin and Brown (2010)), double-Pareto prior (Armagan et al. (2013)) or Dirichlet-Laplace prior (Bhattacharya et al. (2015)).

The rest of the chapter is organized as follows. In the next section, we describe the details of modeling the connectome dataset from ADNI. In Section 2.3, we describe the prior construction and develop posterior computing techniques. A simulation study comparing the proposed Bayesian procedure with ANCOVA-based ones is conducted in Section 2.4. The real data on connectome from ADNI is analyzed in Section 2.5. We study the posterior consistency of the proposed Bayesian procedure in Section 2.6 under the asymptotic regime that the number of subjects is going to infinity but the number of regions is remaining fixed. The result justifies the use of the proposed Bayesian procedure also from a frequentist perspective. Proof of the posterior consistency result is given in the Section 2.6.1.

### 2.2 Data description and modeling

In the ADNI dataset on connectome, the brain is divided into $J = 83$ regions. For each pair of brain regions, the number of white matter fibers between them and their average widths are obtained.

The data is obtained for $n = 100$ subjects, for whom information regarding disease prognosis, sex and age are obtained. For many edges in Figure 2.1, the mean widths are not defined where there are no white matter fibers. There are three disease prognosis states, Alzheimer (AD), mild cognitive impairment (MCI) and no cognitive impairment (NC). There may not be any white matter fiber connecting two brain regions for some subjects.

Except for age, the other two covariates, sex and disease prognosis, are categorical. Since disease prognosis has three possible states, we introduce dummy variables $Z_{\text{MCI}}$ and $Z_{\text{AD}}$ respectively standing for the onset of MCI and AD, setting NC at the baseline. Similarly, the dummy variable $Z_{\text{M}}$ indicating male gender is introduced setting females at the baseline. Let $Z = (Z_{\text{MCI}}, Z_{\text{AD}}, Z_{\text{M}}, \text{Age})'$ stand for the whole vector of covariates and $Z_i$ stand for its value for the $i$th subject. Let $N_{ijk}$ stand for the number of white matter fiber connecting brain regions $j$ and $k$ in the $i$th subject, and $W_{ijk}$ the mean width of such fibers, provided that $N_{ijk} \geq 1$. In ADNI dataset, there were no missing values in $N_{ijk}$ or $W_{ijk}$. Some of the covariates were missing. Any categorical missing covariate (i.e. disease state or sex) value is replaced by its mode and for the continuous covariate age, missing values are replaced by the mean age over
It seems natural to consider a Poisson model for the counts of fibers connecting two regions in a subject. However, as shown in Figure 2.1, the abundance of zero connections makes the Poisson model somewhat inappropriate. We overcome the problem by considering a zero-inflated Poisson model, by boosting the probability of zero through a binary latent variable $\Xi_{ijk}$ with parameter $\Phi(\pi_{ijk})$, where $\Phi$ stands for the standard normal distribution function and $\pi_{ijk}$ is a real-valued parameter. If $\Xi_{ijk} = 0$, then $N_{ijk}$ is set at zero, while if $\Xi_{ijk} = 1$, the number of connections $N_{ijk}$ is assumed to be Poisson distributed with some positive mean $e^{\lambda_{ijk}}$. Note that in our formulation $\Xi_{ijk}$ is not completely identifiable since the value $N_{ijk} = 0$ is compatible with both possible values of $\Xi_{ijk}$. If $N_{ijk} \geq 1$, we assume that the mean fiber width, in the logarithmic scale, is normally distributed with some mean $\mu_{ijk}$, and variance $\sigma^2/N_{ijk}$ for some unknown $\sigma > 0$. The heuristic justification of the choice of the variance $\sigma^2/N_{ijk}$ stems from the fact that $W_{ijk}$ is an average of $N_{ijk}$ (independent) variables and should be approximately normal with variance inversely proportional to the number $N_{ijk}$ of averaging variables. Since

Figure 2.1: Heatmap of number in log-scale and average width of white matter fibers where each pixel represents each pair of region.
fiber widths are positive, the model seems to fit the data better in the logarithmic scale, and the heuristics for the choice of the variance extends to the logarithmic scale by the delta method, at least when \( N_{ijk} \) is large. Thus we can represent the data generating process as

\[
\begin{align*}
\Xi_{ijk} & \sim \text{Bin}(1, \Phi(\pi_{ijk})) \\
N_{ijk} | \{\Xi_{ijk} = 1\} & \sim \text{Poisson}(e^{\lambda_{ijk}}) \\
\log W_{ijk} & = \mu_{ijk} + \epsilon_{ijk}, \\
\epsilon_{ijk} | \{N_{ijk} \geq 1\} & \sim N(0, \sigma^2 / N_{ijk}).
\end{align*}
\]

A simple analysis of covariance (ANCOVA)-type model can be formulated to describe linear effects of the covarites \( Z_i \) on each unrestricted parameter \( \pi_{ijk} \), \( \lambda_{ijk} \) and \( \mu_{ijk} \) for each pair of brain regions \((j,k)\):

\[
\begin{align*}
\mu_{ijk} & = (\mu_0)_{j,k} + Z_i'\chi_{jk} + \eta_1i, \\
\pi_{ijk} & = (\pi_0)_{j,k} + Z_i'\beta_{jk} + \eta_2i, \\
\lambda_{ijk} & = (\lambda_0)_{j,k} + Z_i'\nu_{jk} + \eta_3i,
\end{align*}
\]

where \((\mu_0)_{j,k}\), \((\pi_0)_{j,k}\) and \((\lambda_0)_{j,k}\) are baseline values of \( \mu_{ijk} \), \( \pi_{ijk} \) and \( \lambda_{ijk} \) respectively for covariate value \( Z_i = 0 \).

\[
\begin{align*}
\chi_{jk} & = ((\chi_{MCI})_{j,k}, (\chi_{AD})_{j,k}, (\chi_{M})_{j,k}, (\chi_{Age})_{j,k})', \\
\beta_{jk} & = ((\beta_{MCI})_{j,k}, (\beta_{AD})_{j,k}, (\beta_{M})_{j,k}, (\beta_{Age})_{j,k})', \\
\nu_{jk} & = ((\nu_{MCI})_{j,k}, (\nu_{AD})_{j,k}, (\nu_{M})_{j,k}, (\nu_{Age})_{j,k})'
\end{align*}
\]

are regression coefficients for the average width, connection probability and number of connections respectively and \( (.)_{j,k} \) denotes \((j,k)\)th element of a matrix, and \( \eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})' \), \( i = 1, \ldots, n \), are independent random effects of the ith subject distributed according to an unknown common distribution. It may be noted that the normal distribution function \( \Phi \) and the exponential function are used respectively as link functions for binary and Poisson regression. For the latter, the exponential link is almost a universal choice, while for binary regression both logistic and probit (i.e. \( \Phi \)) links are commonly used and usually give similar results. Our preference for the probit link is due to its computational advantage in a Gibbs sampling scheme for Bayesian computation, through a data-augmentation technique (see Albert and Chib (1993)).

For a preliminary analysis, we fit the model using a generalized heteroscedastic ANCOVA, ignoring the zero-inflation aspect and the random effects in the model.

The model thus has \( 3 \times (\binom{83}{2}) = 10458 \) parameters and 34860 observations of mean width and number of white matter fibers corresponding to 100 subjects and 3486 potential edges between different brain regions. We observed that for several edges \((j,k)\), the maximum likelihood
method failed to give estimates of either $\mu_{0,jk}$ or $\chi_{jk}$. For the Poisson regression, the `glm` function in R did not converge for several pairs $(j,k)$. This suggests using a dimension reduction of the parameter space through further modeling if we want to conduct an edge-wise analysis. The dimension reduction also helps with computation and gives easy interpretability of the results.

Since the parameters are indexed by edges, a substantial reduction of dimension will be possible if these can be viewed as arising from some latent characteristics of nodes through some fixed but unknown function. This can be motivated from exchangeability considerations. In the absence of initial information about connections between regions, exchangeability seems to be an appealing assumption. By a well known representation theorem of exchangeable random graphs (c.f. Aldous (1981), Hoover (1979)), a function of edge $(j,k)$ can then be represented as $f(\xi_i, \xi_j)$ where $\xi_i$, for each node $i$ is a latent variable independently and identically distributed and $f$ is a fixed function, called a graphon, irrespective of the size of the network. Assuming that the function $f$ is sufficiently smooth, a basis expansion can approximate it using only fewer terms. Thus the graphon technique in our context will be able to reduce a parameter array of size $83^2 = 3486$ to only a parameter vector of size $83 + K$, where $K$ is the number of parameters used to approximate the unknown smooth graphon function. Typically a modest number of terms suffices for well-behaved functions using standard bases such as B-splines or polynomials. As a result, a substantial dimension reduction is possible through the graphon technique. This leads to modeling the arrays of baseline values and regression coefficient as

\[
\begin{align*}
(\mu_0)_{j,k} &= \mu(\xi_j, \xi_k), \\
(\pi_0)_{j,k} &= \pi(\xi_j, \xi_k), \\
(\lambda_0)_{j,k} &= \lambda(\xi_j, \xi_k), \\
(\chi_l)_{j,k} &= \chi_l(\delta_j, \delta_k), & l = \text{MCI, AD, M, Age}, \\
(\beta_l)_{j,k} &= \beta_l(\delta_j, \delta_k), & l = \text{MCI, AD, M, Age}, \\
(\nu_l)_{j,k} &= \nu_l(\delta_j, \delta_k), & l = \text{MCI, AD, M, Age},
\end{align*}
\]  

(2.4)

where, with an abuse of notations, $\mu$, $\pi$, $\lambda$, $\chi_{\text{MCI}}$, $\chi_{\text{AD}}$, $\chi_{\text{M}}$, $\chi_{\text{Age}}$, $\beta_{\text{MCI}}$, $\beta_{\text{AD}}$, $\beta_{\text{M}}$, $\beta_{\text{Age}}$, $\lambda_{\text{MCI}}$, $\lambda_{\text{AD}}$, $\lambda_{\text{M}}$ and $\lambda_{\text{Age}}$ are smooth functions on the unit square $[0,1]^2$ and symmetric in their arguments, and $\xi_1, \ldots, \xi_J$ and $\delta_1, \ldots, \delta_J$ are latent variables taking values in the unit interval.

The reason for choosing two separate sets of latent variables is to distinguish between fixed and sparse main effect.

Additional dimension reduction is possible through sparsity considerations. Most edges are affected by covariates in the same way, implying that edge-subject interaction is constant, except for a few edges. For an array represented through a graphon, such a sparsity will occur if most values of the latent variables assume (some fixed) benchmark value, which can be conveniently taken as the midpoint $1/2$ of the unit interval. In this case for most of the edges, the estimated value becomes a function of $(1/2, 1/2)$. 


2.3 Prior specification and posterior computation

2.3.1 Prior specification

To proceed with a nonparametric Bayesian analysis, we put prior distributions on the smooth functions appearing in the graphon representation through basis expansion in tensor products of B-splines, and on the coefficients of the basis expansion. The coefficients can be arranged in the form of a square matrix. The symmetry of the matrices of coefficients ensures symmetry of the resulting functions in its arguments as required by graphon functions. Given other set of parameters and values of the random effects, (independent) normal prior on the coefficients of the tensor products of B-splines will lead to conjugacy in the normal regression model for the width, allowing a simple and fast posterior updating rule. In the binary regression model for the connection probability, normal prior still leads to conjugacy using the data augmentation technique of Albert and Chib (1993). Since no conjugacy is possible for the Poisson regression for the number of connections, Metropolis-Hastings is applied. Alternatively, adaptive rejection sampling can be applied to obtain posterior updates. Thus a spike and slab density for the latent variables with a spike at $1/2$ should be able to catch the desired sparsity. We however only induce approximate sparsity though a well-peaked symmetric beta density in place of a point mass at $1/2$, which is computationally more convenient. On the distribution $G$ of the random effects, we put a Dirichlet process scale mixture of zero mean normal prior (see Chapter 5 of Ghosal and van der Vaart (2017)).

More specifically, the prior can be completely described by the following set of relations:

(i) Graphon functions:

\[
\mu(\xi_j, \xi_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \theta_{1,mm'}B_m(\xi_j)B_{m'}(\xi_k),
\]
\[
\pi(\xi_j, \xi_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \theta_{2,mm'}B_m(\xi_j)B_{m'}(\xi_k),
\]
\[
\lambda(\xi_j, \xi_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \theta_{3,mm'}B_m(\xi_j)B_{m'}(\xi_k),
\]
and for \( l = \text{MCI}, \text{AD}, \text{M}, \)

\[
\chi_l(\delta_j, \delta_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \gamma_{1l, mm'} B_m(\delta_j) B_{m'}(\delta_k),
\]

\[
\beta_l(\delta_j, \delta_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \gamma_{2l, mm'} B_m(\delta_j) B_{m'}(\delta_k),
\]

\[
\nu_l(\delta_j, \delta_k) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \gamma_{3l, mm'} B_m(\delta_j) B_{m'}(\delta_k),
\]

where \( \theta_{t, mm'} = \theta_{t, m'm} \) for all \( t = 1, 2, 3, \) and \( \gamma_{tl, mm'} = \gamma_{t, m'm} \) for all \( t = 1, 2, 3, l = \text{MCI}, \text{AD}, \text{M}; \) and that

(a) graphon coefficients: For some chosen \( a > 0, \)

\[ \theta_{t, mm'} \overset{\text{ind}}{\sim} \text{N}(0, a^2), \quad \gamma_{tl, mm'} \overset{\text{ind}}{\sim} \text{N}(0, a^2), \quad 1 \leq m \leq m' \leq K, \]

for \( t = 1, 2, 3, l = \text{MCI}, \text{AD}, \text{M}; \)

(b) latent variables:

\[ \xi_1, \ldots, \xi_J \overset{\text{ind}}{\sim} \text{Un}(0, 1), \quad \delta_1, \ldots, \delta_J \overset{\text{ind}}{\sim} q \text{Un}(0, 1) + (1-q) \text{Be}(M, M), \]

the mixing probability \( q \) is given a default uniform prior \( \text{Un}(0, 1) \) and \( M > 0 \) is chosen big; here \( \text{Un} \) stands for the uniform distribution and \( \text{Be} \) for the beta distribution.

(ii) Random effects distribution: For \( t = 1, 2, 3 \) and \( i = 1, \ldots, n, \)

\[ \eta_{ti} | \tau_{ti} \overset{\text{ind}}{\sim} \text{N}(0, \tau_{ti}^2), \quad \tau_{ti}^2 \overset{\text{ind}}{\sim} G_t, \quad t = 1, 2, 3, \quad G_t \overset{\text{ind}}{\sim} \text{DP}(\alpha_t \text{IG}(b_1, b_2)), \]

where \( \text{DP} \) stands for the Dirichlet process, \( \text{IG} \) for the inverse-gamma distribution and the precision parameter \( \alpha_t \) of the Dirichlet process is given a gamma prior \( \alpha_t \sim \text{Ga}(c_1, c_2). \)

(iii) Error variance: \( \sigma^{-2} \sim \text{Ga}(d_1, d_2). \)
2.3.2 Posterior updating

Introduce a latent variable $I_j$ the indicator of the Un(0,1) component of the distribution of $\delta_j$, $j = 1, \ldots, J$. Now the log-likelihood is given by

$$C = \sum \exp \left\{ \sum_{m,m'} \left[ \theta_{3,m,m'} B_m(\xi_j) B_{m'}(\xi_k) + \sum_l \gamma_{3,l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} + \eta_{3i} \right] \right\}$$

$$+ \sum_{i,j,k} N_{ijk} \left\{ \sum_{m,m} \left[ \theta_{1,m,m'} B_m(\xi_j) B_{m'}(\xi_k) + \sum_l \gamma_{1,l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} + \eta_{1i} \right] \right\}$$

$$- \frac{1}{2\sigma^2} \sum_{i,j,k} N_{ijk} \log W_{ijk} - \sum_{m=1}^K \sum_{m'=1}^{M} \left( \theta_{2,m,m'} B_m(\xi_j) B_{m'}(\xi_k) \right)$$

$$+ \sum_{l} \gamma_{2,l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} + \eta_{2i} \right\}$$

$$+ \sum_{i,j,k} \left( 1 - I(N_{ijk} = 0) \right) \log \left( 1 - \Phi \left( \sum_{m,m'} \left( \theta_{2,m,m'} B_m(\xi_j) B_{m'}(\xi_k) \right) \right) \right.$$

$$+ \sum_{l} \gamma_{2,l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} + \eta_{2i} \right\)$$

$$- \frac{1}{2a^2} \sum_{m \leq m'} (\theta_{1,m,m'}^2 + \theta_{2,m,m'}^2 + \theta_{3,m,m'}^2)$$

$$- \frac{1}{2a^2} \sum_{m \leq m'} \sum_{l} (\gamma_{1,m,m'}^2 + \gamma_{2,m,m'}^2 + \gamma_{3,m,m'}^2)$$

$$+ \log \left( (1 - I_j) \delta_j^{M-1} (1 - \delta_j)^{M-1} \Gamma(M)^2 / \Gamma(2M) + I_j \right)$$

$$+ I_j \log q + (1 - I_j) \log(1 - q) - (nJ^2/2 + d1 - 1) \log \sigma^2 - d2/\sigma^2,$$

where $C$ invoies only hyperparameters $a, M, K, b_1, b_2, c_1, c_2, d_1, d_2, q$ and the observations, but not the parameters of the model.

Posterior updates are explained here. For notational convenience we define $\theta_1 = (\theta_{1,m,m'} : 1 \leq m \leq m' \leq K)$, $\gamma_{1l} = (\gamma_{1l,m,m'} : 1 \leq m \leq m' \leq K, l = MCI,AD,M)$, $\theta_2 = (\theta_{2,m,m'} : 1 \leq m \leq m' \leq K)$, $\gamma_{2l} = (\gamma_{2l,m,m'} : 1 \leq m \leq m' \leq K, l = MCI,AD,M)$, $\theta_3 = (\theta_{3,m,m'} : 1 \leq m \leq m' \leq K)$, $\gamma_{3l} = (\gamma_{3l,m,m'} : 1 \leq m \leq m' \leq K, l = MCI,AD,M)$, $\xi = (\xi_1, \ldots, \xi_K)$, $\delta = (\delta_1, \ldots, \delta_K)$, $\eta = (\eta_{11}, \ldots, \eta_{1n})$, $\theta = (\eta_{21}, \ldots, \eta_{2n})$, $\eta = (\eta_{31}, \ldots, \eta_{3n})$, $\tau_1 = (\tau_{11}, \ldots, \tau_{1n})$, $\tau_2 = (\tau_{21}, \ldots, \tau_{2n})$, $\tau_3 = (\tau_{31}, \ldots, \tau_{3n})$, $\sigma$. Let $A$ be a symmetric matrix of dimension $K \times K$, vec($A$) be the vector stacking columns of $A$ and uni($A$) = \{ $A_{ij} : 1 \leq i \leq j \leq K$ \} be the vector
of unique entries of $A$. We can construct matrix $B$ of dimension $(\frac{K}{2}) \times K$ such that $\text{vec}(A)' = \text{uni}(A)'B$. Let $\psi_{jk} = (B_m(\xi_j)B_n(\xi_k) : 1 \leq m \leq m' \leq K)$ and $\psi_{jk}' = (B_m(\delta_j)B_n(\delta_k) : 1 \leq m \leq m' \leq K)$. To perform data augmentation technique for block updating parameters in the probit regression part, we introduce latent variable $L_{ijk}$. Here we use $\text{Id}_l$ to denote identity matrix of dimension $l$.

- Updating $\sigma^2$: Generate a sample from the inverse gamma distribution with parameters $(d_1 + nJ^2/2)$ and $(d_2 + \sum_{i,j,k} N_{ijk}[\log W_{ijk} - \sum_{m=1}^{K} \sum_{m'=1}^{K} (\theta_{1,mm'}B_m(\xi_j)B_m'(\xi_k) + \sum_l \gamma_{1l,mm'}B_m(\delta_j)B_m'(\delta_k)Z_{il}) + \eta_{1i}]^2)$.

- Updating $\theta_1$: Generate a sample from the multivariate normal distribution with mean $M_{\theta_1} = V_{\theta_1}B(\sum_{i,j,k} \psi_{jk}N_{ijk}(\log W_{ijk} - \sum_l \gamma_{1l}B\psi_{jk}'))$ and variance $V_{\theta_1} = [\sum_{i,j,k} N_{ijk}\psi_{jk}')(\sum_{i,j,k} N_{ijk}\psi_{jk})/\sigma^2 + \text{Id}_{K(K+1)/2}/a^2]^{-1}$.

- Updating $\gamma_{1l}$: Generate a sample from the multivariate normal distribution with mean $M_{\gamma_{1l}} = V_{\gamma_{1l}}B(\sum_{i,j,k} \psi_{jk}N_{ijk}(\log W_{ijk} - \sum_{v\neq l} \gamma_{1v}B\psi_{jk}'))$ and variance $V_{\gamma_{1l}} = [\sum_{i,j,k} N_{ijk}Z_{il}\psi_{jk}')(\sum_{i,j,k} N_{ijk}Z_{il}\psi_{jk})/\sigma^2 + \text{Id}_{K(K+1)/2}/a^2]^{-1}$, $M_{\gamma_{1l}} = V_{\gamma_{1l}}B(\sum_{i,j,k} \psi_{jk}N_{ijk}(\log W_{ijk} - \sum_{v\neq l} \gamma_{1v}B\psi_{jk}'))$. Generate $\gamma_{1l}$ from Multivariate normal with mean $M_{\gamma_{1l}}$ and covariance matrix $V_{\gamma_{1l}}$.

- Updating $L_{ijk}$: The posterior distribution of $L_{ijk}$ is

$$N(\sum_{m,m'} (\theta_{2,mm'}B_m(\xi_j)B_m'(\xi_k) + \sum_l \gamma_{2l,mm'}B_m(\delta_j)B_m'(\delta_k)Z_{il}) + \eta_{2i}, 1),$$

truncated to $[0, \infty)$ if $N_{ijk} > 0$, but truncated to $(-\infty, 0)$ if $N_{ijk} = 0$.

- Updating $\theta_2$: Generate a sample from the multivariate normal distribution with mean $M_{\theta_2} = V_{\theta_2}B(\sum_{i,j,k} \psi_{jk}(L_{ijk} - \sum_l \gamma_{2l}B\psi_{jk}'))$ and variance $V_{\theta_2} = [\sum_{j,k} n\psi_{jk}')(\sum_{j,k} n\psi_{jk}) + \text{Id}_{K(K+1)/2}/a^2]^{-1}$.

- Updating $\gamma_{2l}$: Generate a sample from the multivariate normal distribution with mean $M_{\gamma_{2l}} = V_{\gamma_{2l}}B(\sum_{i,j,k} \psi_{jk}(L_{ijk} - \sum_{v\neq l} \gamma_{2v}B\psi_{jk}'))$ and variance $V_{\gamma_{2l}} = [\sum_{j,k} nZ_{il}\psi_{jk}')(\sum_{j,k} nZ_{il}\psi_{jk})/\sigma^2 + \text{Id}_{K(K+1)/2}/a^2]^{-1}$.

- Updating $\theta_{3,mm'}$: Execute the following steps of the Metropolis-Hastings algorithm:
  - generate $\epsilon_{3,mm'}$ from $N(0, s_3)$ for some suitably tuned $s_3 > 0$ (see below);
  - update $\theta_{3,mm'}$ to $\theta_{3,mm'}^* = \theta_{3,mm'} + \epsilon_{3,mm'}$ with probability

$$P_{a,3,mm'} = \min\left\{\frac{L(m, m', \theta_{3,mm'}^*)}{L(m, m', \theta_{3,mm'})}, 1\right\},$$
where \( \log L(m, m', x) \) is given by

\[
- \sum_{i,j,k} \exp \left\{ \sum_{v,v' \neq m,m'} \left[ \theta_{3,vv'} B_v(\xi_j) B_{v'}(\xi_k) \right] + \sum_l \gamma_{3,vv'} B_v(\delta_j) B_{v'}(\delta_k) Z_{il} \right\} \\
+ x B_m(\xi_j) B_{m'}(\xi_k) + \eta_{3i} \right\} \\
+ \sum_{i,j,k} N_{ijk} \left[ \sum_{v,v' \neq m,m'} \left( \theta_{3,vv'} B_v(\xi_j) B_{v'}(\xi_k) \right) + \sum_l \gamma_{3,vv'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} \right] \\
+ x B_m(\xi_j) B_{m'}(\xi_k) + \eta_{3i} \right].
\]

- Updating \( \gamma_{3l,mm'} \): Execute the following steps of the Metropolis-Hastings algorithm:

  - generate \( \epsilon_{3l,mm'} \) from \( N(0, s_3) \) for some suitably tuned \( s_3 > 0 \) (see below);
  - update \( \gamma_{3l,mm'} \) to \( \gamma_{3l,mm'}^* \) with probability

\[
P_{\alpha,3l,mm'} = \min \left\{ \frac{L_1(l, m, m', \gamma_{3l,mm'}^*)}{L_1(l, m, m', \gamma_{3l,mm'}^*)}, 1 \right\},
\]

where \( \log L_1(l, m, m', x) \) is given by

\[
- \sum_{i,j,k} \exp \left\{ \sum_{v,v' \neq m,m'} \left[ \theta_{3,vv'} B_l(\xi_j) B_{v'}(\xi_k) \right] + \sum_l \gamma_{3,vv'} B_v(\delta_j) B_{v'}(\delta_k) Z_{il} \right\} \\
+ x B_m(\delta_j) B_{m'}(\delta_k) + \theta_{3,mm'} B_m(\xi_j) B_{m'}(\xi_k) + \eta_{3i} \right\} \\
+ \sum_{i,j,k} N_{ijk} \left[ \sum_{v,v' \neq m,m'} \left( \theta_{3,vv'} B_v(\xi_j) B_{v'}(\xi_k) \right) + \sum_l \gamma_{3,vv'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il} \right] \\
+ x B_m(\delta_j) B_{m'}(\delta_k) + \theta_{3,mm'} B_m(\xi_j) B_{m'}(\xi_k) + \eta_{3i} \right].
\]

- Updating \( \delta_j \): Execute the following steps of the Metropolis-Hastings algorithm:

  - generate \( U_{1j} \) from \( \text{Be}(B_1, B_1) \);
  - update \( \delta_j \) to \( \delta_j^* = \delta_j U_{1j} / \{ \delta_j U_{1j} + (1 - \delta_j)(1 - U_{1j}) \} \) with probability

\[
P_{\alpha,\delta_j} = \min \left\{ \frac{L_2(j, \delta_j^*) f(\delta_j^*|\delta_j)}{L_2(j, \delta_j) f(\delta_j^*|\delta_j)}, 1 \right\},
\]

where \( L_2(j, x) \) denotes likelihood at \( \delta_j = x \) (keeping other parameters fixed) and

\[
f(\delta_j^*|\delta_j) = \frac{\delta_j (1 - \delta_j) B_1^{-1} (1 - \delta_j) + \delta_j (1 - \delta_j) B_1^{-1} (1 - \delta_j) \delta_j^*}{B(B_1, B_1)} \frac{\delta_j^* (1 - \delta_j^*)}{(\delta_j + \delta_j^* - 2\delta_j \delta_j^*)^2};
\]
here \( B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b) \) is the beta function, and \( B_1 > 0 \) is to be tuned suitably (see below).

- Updating \( \xi_j \): Execute the following steps of the Metropolis-Hastings algorithm:
  - generate \( U_{2j} \) from \( \text{Be}(B_2, B_2) \);
  - update \( \xi_j \) to \( \xi_j^* = \xi_j U_{2j}/\{\xi_j U_{2j} + (1 - \xi_j)(1 - U_{2j})\} \) with probability
    \[
    P_{a,\xi_j} = \min\left\{ \frac{L_3(j, \xi_j^*) f_2(\xi_j^*|\xi_j) \cdot 1}{L_3(j, \xi_j) f_2(\xi_j^*|\xi_j)}, 1 \right\},
    \]
    where \( L_3(j, x) \) denotes likelihood at \( \xi_j = x \) (keeping other parameters fixed),
    \[
    f(\xi_j^*|\xi_j) = \frac{(\xi_j(1-\xi_j^*)(\xi_j^*+\xi_j^*-2\xi_j^*)\Gamma(B_1(B_2, B_2))+(\xi_j^*(1-\xi_j^*))^{B_2-1}}{\Gamma(B_2, B_2)(\xi_j+\xi_j^*-2\xi_j^*)^2}.
    \]
    and \( B_2 > 0 \) is to be tuned suitably (see below).

- Updating \( \eta_i \): Generate a sample from the multivariate normal distribution with mean \( M_{\eta_i} = V_{\eta_i}(\varepsilon_{i1}, \ldots, \varepsilon_{in}) \), where \( \varepsilon_{1i} = \sum_{jk} N_{i,j,k} \log L_{i \cdot j \cdot k} - \sum_{m=1}^{K} \sum_{m'=-1}^{K} (\theta_{m,mm'} B_{m}^{(\xi_j)} B_{m'}^{(\xi_k)} + \sum_{l \in L_{mm'}} B_{m}^{(\delta_j)} B_{m'}^{(\delta_k)} Z_{il})), i = 1, \ldots, n, \) and covariance
    \[
    V_{\eta_i} = \text{diag}(\sum_{jk} N_{i,j,k}, \ldots, \sum_{jk} N_{n,j,k}/\sigma^2 + \text{diag}(1/\tau_{i1}^2, \ldots, 1/\tau_{in}^2))^{-1}.
    \]

- Updating \( \eta_2 \): Generate a sample from the multivariate normal distribution with mean \( M_{\eta_2} = V_{\eta_2}(\varepsilon_{21}, \ldots, \varepsilon_{2n}) \), where \( \varepsilon_{2i} = \sum_{jk} L_{i \cdot j \cdot k} - \sum_{m=1}^{K} \sum_{m'=-1}^{K} (\theta_{2,mm'} B_{m}^{(\xi_j)} B_{m'}^{(\xi_k)} + \sum_{l \in L_{mm'}} B_{m}^{(\delta_j)} B_{m'}^{(\delta_k)} Z_{il})), i = 1, \ldots, n, \) and covariance
    \[
    V_{\eta_2} = \text{diag}(J^2, \ldots, J^2 + \text{diag}(1/\tau_{21}^2, \ldots, 1/\tau_{2n}^2))^{-1}.
    \]

- Updating \( \eta_3 \): Execute the following steps of the Metropolis-Hastings algorithm:
  - generate \( \varepsilon_{3,i} \) from \( N(0, s_{3i}) \) for some suitable tuned \( s_3 > 0 \);
  - update \( \eta_3 \) to \( \eta_3^* = \eta_3 + \varepsilon_{3,i} \) with probability
    \[
    P_{a,\eta_3} = \min\left\{ \frac{L_4(i, \eta_3^*)}{L_4(i, \eta_3)}, 1 \right\},
    \]
where \( \log L_4(i, x) \) is given by

\[
- \sum_{j,k} \exp\left\{ \sum_{m,m'} [\theta_3_{m,m'} B_m(\xi_j) B_{m'}(\xi_k) + \sum_l \gamma_3_{l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il}] + x \} \right\}
+ \sum_{j,k} N_{ijk} \left\{ \sum_{m,m'} [\theta_3_{m,m'} B_m(\xi_j) B_{m'}(\xi_k) + \sum_l \gamma_3_{l,m,m'} B_m(\delta_j) B_{m'}(\delta_k) Z_{il}] + x \right\} \].

- Updating \( \tau_{ti} \): Execute the following steps of the Pólya urn scheme:
  - generate \( r \) with \( P(r = l) = q_{t,il}/\{\sum_k q_{t,ik}\} \), where
    \[
    q_{t,ij} \propto \begin{cases} 
    \frac{1}{\tau_{ti}} \exp(-\eta_{ti}^2/(2\tau_{ti}^2)), & \text{if } j \neq i, j \geq 1 \\
    \alpha_t \{\tau - 0.5 - b_1/2 - 1 \exp(-\eta_{ti}^2/(2\tau_{ti}^2))\}, & \text{if } j = i;
    \end{cases}
    \]
  - if \( r = i \), sample \( \tau_{ti}^2 \) from \( \text{Ga}(-0.5 - b_1/2, \eta_{ti}^2/2 + b_2) \), else set \( \tau_{ti} = \tau_{tr} \).
- Updating of \( q \): Generate \( q \) from \( \text{Be}(\sum_j I_j + 1, J - \sum_j I_j) \).

2.3.3 Tuning

In the above Metropolis-Hastings steps, we need to tune \( s_3, s_{31}, s_{32}, s_{33}, B_1, B_2 \) and \( s_{3\eta} \) (these are defined in the supplementary material) to achieve good acceptance rates. We automatically adjust those values after every 500 iterations. The standard deviations are reduced (respectively, increased) to increase (respectively, to reduce) the acceptance rate of the Metropolis-Hastings moves. On the other hand, \( B_1 \) and \( B_2 \) are increased (respectively, reduced) to increase (respectively, to reduce) the acceptance rate of the corresponding parameters. If \( U_j = 0.5 \), update will be same as the current value. So, the distribution of \( U_j \) should have a high concentration at 0.5 to make a local move from the current state. A higher value of \( B \) will induce a proposed move in the close vicinity of the original position while a smaller value will lead to a proposed move to a farther location. Tuning the values of \( B_1 \) and \( B_2 \) can help maintain a desirable proportion of accepted moves. The number of B-spline basis functions \( (K) \) is tuned via a grid search over a sequence of values in the range 5-10. For each possible values of \( K \), we generate 10 sets of latent variables. For each set of latent variables, we can fit a simple linear regression to estimate the B-spline coefficients and calculate average AIC over all the sets of latent variables. Based on these AIC values, we pick the \( K \) with lowest AIC value or the smallest value after which there is not much improvement in AIC.
2.4 Simulation

In this section, we study the performance of the proposed Bayesian method in comparison with ANCOVA. For simplicity, we do not consider zero inflation or the random effect in the data generating process as well as in the model, that is, we consider the following simplified analog of (2.1):

\begin{align*}
N_{ijk} &\sim \text{Poisson}(e^{\lambda_{ijk}}), \\
\log W_{ijk} &= \mu_{ijk} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0,\sigma^2/N_{ijk}), \\
\mu_{ijk} &= \mu_{0,jk} + Z_i^t \chi_{jk}, \quad \lambda_{ijk} = \lambda_{0,jk} + Z_i^t \nu_{jk}.
\end{align*}

(2.5)

We consider two simulation settings. In the first setting, the actual data generating process follows the graphon model for dimension reduction given by (2.4). This is the well-specified case. In the second set-up, the data are generated from the scheme in (2.5) with $\chi_{jk}$ and $\nu_{jk}$ sparse around a non-zero real number. This is the misspecified case. The constructions of these matrices are described below.

We consider $n = 10, 25, 50, 100, 200, 500$ subjects for well-specified case when parameters follow (2.4) and $n = 30, 50, 100, 200, 500$ subjects for misspecified case when parameters do not have any restriction for $J = 20$ nodes.

**Data generation in the well-specified case:**

- The latent variables $\xi_1, \ldots, \xi_8$ are generated from Un(0,1).
- For other set of latent variables: $\delta_i = 0.5$ if $i \neq 6$ and $\delta_6 = 0.7$.
- Each element of B-spline coefficients $\theta_1, \theta_2, \theta_3$ and $\gamma_{1l}, \gamma_2, \gamma_{3l}$ for $l = (\text{MCI}, \text{AD}, \text{M})$ are generated from $N(0,1)$ with $K = 8$ terms in the expansion for each argument.
- With the above values we can construct $\lambda_{ijk}$ and $\mu_{ijk}$ and generate $N_{ijk}$ and $W_{ijk}$ for all $1 \leq j \leq k \leq J$.

**Data generation in the misspecified case:**

- The elements of intercept $\mu_0 = \{\mu_{0,jk} : j \leq k\}$ and $\lambda_0 = \{\lambda_{0,jk} : j \leq k\}$ are generated from $N(0,1)$.
- Entries of each of the components in (2.3) will be divided into two groups for fixed effect and effect different from fixed effect with probability $(0.9, 0.1)$. 
Then we assign fixed effect $-2$ and different values were generated from $N(-3, 0.5)$ for $\chi_{MCI}$ and $\nu_{MCI}$, correspondingly $-1$ and $N(0, 0.5)$ for $\chi_{AD}$ and $\nu_{AD}$, $2.5$ and $N(0, 0.5)$ for $\chi_M$ and $\nu_M$ and lastly $0$ and $N(1.5, 0.5)$ for $\chi_{Age}$ and $\nu_{Age}$.

With the above values we can construct $\lambda_{ijk}$ and $\mu_{ijk}$ and generate $N_{ijk}$ and $W_{ijk}$ for $1 \leq j \leq k \leq J$ and $i = 1, \ldots, n$.

Reproducibility: All the data generation can be reproduced by setting seed at 1 in the ‘R’ statistical programming language. Thus, the true set of matrices is identical across sample sizes and replications.

We have performed 50 replications for each cases. We collect 5000 MCMC samples after burning in 5000 initial samples.

Choice of hyperparameters: We choose the hyperparameters $a = 10$, $M = 10$, $b_1 = b_2 = 0.1$ and $c_1 = c_2 = 10$.

For ANCOVA estimation, we use the weighted least squares technique for the normal model and the generalized linear regression for Poisson regression model for the Poisson link function. We provide table for squared bias and variance of the estimates for the slope parameters for largest sample size in a table. We present a comparison plot of squared bias, variance and MSE of the estimates across different sample sizes in Figure 2.3 and Table 2.1.

For misspecified case as well we provide squared bias, variance and MSE of the estimates for the largest sample size 500 in Table 2.2. There is also a comparison plot of bias square, variance and MSE of estimates across different sample sizes in Figure 2.5.

<table>
<thead>
<tr>
<th></th>
<th>Bayes</th>
<th></th>
<th>ANCOVA</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Square bias</td>
<td>Variance</td>
<td>MSE</td>
<td>Square bias</td>
<td>Variance</td>
<td>MSE</td>
</tr>
<tr>
<td>$\chi_{MCI}$</td>
<td>0.01</td>
<td>0.0001</td>
<td>0.0101</td>
<td>0.02</td>
<td>0.0015</td>
<td>0.0215</td>
</tr>
<tr>
<td>$\chi_{AD}$</td>
<td>0.00</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.02</td>
<td>0.0013</td>
<td>0.0213</td>
</tr>
<tr>
<td>$\chi_M$</td>
<td>0.00</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.00</td>
<td>0.0012</td>
<td>0.0015</td>
</tr>
<tr>
<td>$\chi_{Age}$</td>
<td>0.01</td>
<td>0.0000</td>
<td>0.0100</td>
<td>0.03</td>
<td>0.0002</td>
<td>0.0302</td>
</tr>
<tr>
<td>$\nu_{MCI}$</td>
<td>0.00</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.01</td>
<td>0.0003</td>
<td>0.0103</td>
</tr>
<tr>
<td>$\nu_{AD}$</td>
<td>0.01</td>
<td>0.0000</td>
<td>0.0100</td>
<td>0.03</td>
<td>0.0002</td>
<td>0.0302</td>
</tr>
<tr>
<td>$\nu_M$</td>
<td>0.00</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.0002</td>
<td>0.0012</td>
</tr>
<tr>
<td>$\nu_{Age}$</td>
<td>0.01</td>
<td>0.0000</td>
<td>0.0100</td>
<td>0.02</td>
<td>0.0000</td>
<td>0.0200</td>
</tr>
</tbody>
</table>

In Figure 2.2, we see in well specified case, our Bayesian method identifies the true structure.
Figure 2.2: Heatmap of truth along with Bayesian methods and ANCOVA estimates for sample size 500 in the well specified case with 20 nodes.
Figure 2.3: Plot of squared bias and variance for different parameters against sample sizes for Bayesian method and ANCOVA estimates in the well-specified case.

for all the cases. From Figure 2.3 we conclude that proposed Bayesian method performs better than ANCOVA after sample size 25. In Figure 2.4, for misspecified case we observe both of the Bayesian and ANCOVA estimates are in similar color range with the truth for most of the cases. So, both of the methods could estimate the fixed effect for most of the parameters. In Figure 2.5 we can compare the two estimates quantitatively. The two methods show comparable performance as sample size increases for all of the eight parameters.

2.5 Real data analysis

We analyze a real dataset with 100 individuals, collected from ADNI. A demographic summary of the data is provided in Table 2.3. The baseline subject is a female subject of average age with no cognitive impairment. The plot of the estimates, obtained from the proposed Bayesian method are given in Figure 6.

For each predictor variable, we identify the edges which are affected most significantly by the disease. Each predictor has an unknown overall effect for all the edges, which may be considered as the reference value. To find that unknown overall effect we need to test for $83^2$ hypotheses for an estimated matrix $A_{83 \times 83}$, $A_{i,j} = \vartheta, i, j = 1, \ldots, 83$ simultaneously. We follow a procedure
Figure 2.4: Heatmap of truth along with Bayesian methods and ANCOVA estimates for sample size 500 in the misspecified case with 20 nodes.
similar to the one used in White and Ghosal (2014). We fit a Gaussian mixture model (GMM) for the marginal density of $A_{i,j}$, $\Psi(a) = \sum_{k=1}^{K} \omega_k \phi(a, \mu_k, \sigma_k)$. We choose the value of $k$ between 3 to 5 depending on the AIC value after fitting the model. The R package \texttt{mixtools} is used for estimation of GMM. Then based on the highest posterior probability component i.e. the $l$ for which $\omega_l$ is maximum in the mixing vector $\omega$, we find the significant edges and that highest posterior probability component $\mu_l$ is the representative value for the estimated matrix. The posterior probability for each edge $A_{ij}$ is defined by the fraction $\frac{\omega_l \phi(A_{ij}, \mu_l, \sigma_l)}{\Psi(A_{ij})}$. The number of significant edges is obtained based on 5% tolerance level of the posterior probabilities of different edges i.e. the fraction from the previous line. There are total 3486 many edges including self edges.

Most significant 50 edges are plotted for each parameter of interest. The edge-widths are proportional to the level of significance, more width means more significant. The colored circles are the different regions and their names are mentioned in the legend. These plots are in 2.6, 2.7 and 2.8.

In the Figure 2.6 among 50 most significant edges there are many self-edges. A significant self-edge infers that the corresponding region has a significant effect on the connectome structure. On the contrary, most of the significant edges in Figure 2.7 and 2.8 are inter edges.
Table 2.2: Comparison of estimation accuracy for $n = 500$ in misspecified case

<table>
<thead>
<tr>
<th></th>
<th>Bayes</th>
<th>ANCOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Square bias</td>
<td>Variance</td>
</tr>
<tr>
<td>$\chi_{MCI}$</td>
<td>0.40</td>
<td>0.0080</td>
</tr>
<tr>
<td>$\chi_{AD}$</td>
<td>0.15</td>
<td>0.0070</td>
</tr>
<tr>
<td>$\chi_{M}$</td>
<td>0.40</td>
<td>0.0001</td>
</tr>
<tr>
<td>$\chi_{Age}$</td>
<td>0.06</td>
<td>0.0005</td>
</tr>
<tr>
<td>$\nu_{MCI}$</td>
<td>0.08</td>
<td>0.0003</td>
</tr>
<tr>
<td>$\nu_{AD}$</td>
<td>0.07</td>
<td>0.0002</td>
</tr>
<tr>
<td>$\nu_{M}$</td>
<td>0.99</td>
<td>0.0004</td>
</tr>
<tr>
<td>$\nu_{Age}$</td>
<td>0.07</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Table 2.3: Demographic table

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>No cognitive impairment (NC)</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Alzheimer disease (AD)</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Mild cognitive impairment (MCI)</td>
<td>14</td>
<td>26</td>
</tr>
<tr>
<td>Average Age</td>
<td>71.80</td>
<td>75.15</td>
</tr>
<tr>
<td>(6.74)</td>
<td>(7.04)</td>
<td></td>
</tr>
</tbody>
</table>

between two different regions. The significant edges for these two plots are similar. Both of the two are related to the presence or absence of edges or number of edges. Most significant edges for covariates related to disease states involve edges related to memory management regions banks of the superior temporal sulcus, entorhinal, and transversetemporal. In the network plots for covariate ‘Age’ there are above mentioned regions along with edges related to cuneus. It also involves medial orbito frontal. For the covariate ‘Gender’ most of the significant edges are related to the regions responsible for memory management and visual processing.

2.6 Large-sample Properties

We consider an abstract setting a fixed $J$ number of nodes, which need not be specifically restricted to the problem of connectome. The observations follow the data generating process in (2.1) and parameters are given by (2.2). Instead of being specific about predictors like in 2.3, we consider that there are a fixed number $d$ of covariates denoted by $Z_i = (Z_{i1}, \ldots, Z_{id})'$, $i = (1, \cdots, n)$, where $n$ is the sample size. The asymptotics is in terms of $n \to \infty$. Then regression coefficients $\chi_{jk}, \beta_{jk}, \nu_{jk}$ are $d$-dimensional. For notational convenience, we write $j = (j, k)$ and append all the fixed effect parameters in one array of parameters and define $\chi^* = (\mu_{0,j}, \chi_j), \beta^*(\pi_{0,j}, \beta_j)$ and $\nu^* = (\lambda_{0,j}, \nu_j)$. Each of $\chi^*, \beta^*, \nu^*$ will be $p = d(J^2)$ dimensional. We
Figure 2.6: Effect of different covariates on connection width, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge.
consider for each pair \((i, j : i = 1, \ldots, n; j = 1, \ldots, \binom{J}{2})\) that there are \(p\) dimensional covariates \(X_{ij}\) such that \(X_{ij} = \left(1, Z_i'\right) \otimes e_j\), where \(e_j\) is vector of length \(\binom{J}{2}\) with one at \(j\)th position and \(\otimes\) stands for the Kronecker product. Posterior consistency is established in Kim and Kim (2011) for mixed effect model with binary data \(P(Y_{ij} = 1 | X_{ij}, Z_{ij}) = H(X_{ij}'\beta + Z_{ij}'b_i)\) where \(\beta = (\beta_1, \ldots, \beta_p)'\) and \(X_{ij} = (X_{ij1}, \ldots, X_{ijp})'\) are \(p\)-dimensional vectors standing respectively for fixed effects and the corresponding regressors, \(b = (b_1, \ldots, b_q)'\) and \(Z_{ij} = (Z_{ij1}, \ldots, X_{ijq})'\) are \(q\)-dimensional vectors standing respectively for random effects covariates with distribution function \(F\) and the corresponding regressors. Let \(H\) be a known increasing and symmetric link function. Then we can rewrite our model as

\[
\begin{align*}
\log W_{ij} &= X_{ij}'\gamma^* + \eta_1 + \epsilon_{ij}, \\
\Xi_{ij} &\sim \text{Bin}(1, H(X_{ij}'\beta^* + \eta_2)), \\
N_{ij}\{|\Xi_{ij} = 1\} &\sim \text{Poisson}(e^{\lambda_{ij}}), \\
\lambda_{ij} &= X_{ij}'\nu^* + \eta_3, \\
\eta_{1i} &\sim \text{iid } F_1, \\
\eta_{2i} &\sim \text{iid } F_2, \\
\eta_{3i} &\sim \text{iid } F_3, \\
\end{align*}
\]

here \(H\) is a known link function which is increasing and symmetric. Let the distributions of \(\eta_1\), \(\eta_2\) and \(\eta_3\) be \(F_1\), \(F_2\) and \(F_3\) respectively, which are considered as unknown. The whole parameter

Figure 2.7: Effect of covariates on connection-probability, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge.
Figure 2.8: Effect of covariates on number of connections, each circle denotes different cortical brain regions and edge-widths are proportional to the degree of significance of that edge.
is denoted by $\theta = (\chi^*, \beta^*, \nu^*, F_1, F_2, F_3)$. Note that in our setup $q = 1$ and $Z_{i1}^i = 1$ for all $i$.

We proceed like the posterior consistency result of Kim and Kim (2011) with appropriate modifications, and show that the arguments extend to cover the normal and Poisson regression mixed effect regression models as well. More specifically, their posterior consistency result is established by verifying that the true parameter lies in the Kullback-Leibler support of the prior and constructing uniformly exponentially consistent tests for testing the true point null against a complement of an arbitrary sized neighborhood of the truth and applying the general theory of posterior consistency; see Ghosal and van der Vaart (2017) for details on the general theory. Using standard arguments, we establish the prior positivity condition for each of the three — binary regression, Poisson regression and normal regression model. To construct the required tests, we proceed like Kim and Kim (2011) who considered mixed effects binary regression models and constructed tests using the arguments given in Amewou-Atisso et al. (2003). However they assumed that the whole real line is the support for the first continuous covariate, which is inconvenient for our application. Under symmetry of the link function $H$, we construct the required test without this assumption. We then show that for the normal mixed effect regression model part, the required test is automatically obtained from that in the binary part by viewing binary observations as arising from data-coarsening of normal observations. For the Poisson regression part, we construct the required tests directly using ideas used for the binary regression model.

Let the parameter space for $F_1, F_2$ and $F_3$ be denoted by $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ respectively. Thus the parameter space for $\theta$ is $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \times \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$. We consider a prior $\tilde{\Gamma}_1$ for $F_1, \tilde{\Gamma}_2$ for $F_2$ and $\tilde{\Gamma}_3$ for $F_3$ independently, and independent of $F_1, F_2$ and $F_3$, we consider priors $\alpha_1$ for $\chi^*$, $\alpha_2$ for $\beta^*$ and $\alpha_3$ for $\nu^*$ respectively. Let $\chi_0, \beta_0$ and $\nu_0$ be the true values of $\chi^*, \beta^*$ and $\nu^*$ respectively and $F_0^1, F_0^2$ and $F_0^3$ be the true random effects distributions of $\eta_1, \eta_2$ and $\eta_3$ respectively. Call $(\chi_0, \beta_0, \nu_0, F_0^1, F_0^2, F_0^3)$ to be $\theta_0$. We make the following assumptions.

Notation: A set $Q_n$ is (asymptotically) non-null if $\liminf \frac{|Q_n|}{n} > 0$, where $|.|$ denotes cardinality of a set.

- (RE). Random effect distribution: The random effects $\eta_{1i}, \eta_{2i}$ and $\eta_{3i}$ have mean zero and symmetric symmetric distribution.

- (SG). Subgaussian tail: The true distribution $F_0^i$ has subgaussian tail: $F_0^i(z : |z| > t) \lesssim \exp(-bt^2)$ for some constant $b > 0$.

- (Supp). Support of the prior: The true distributions $F_0^i \in \text{supp}(\Gamma_i), i = 1, 2, 3$, and $\chi^0 \in \text{supp}(\alpha_1), \beta_0 \in \text{supp}(\alpha_2)$ and $\nu_0 \in \text{supp}(\alpha_3)$, where supp stands for the topological support of a probability distribution, with respect to the weak topology on the space of probability measures and with respect to the Euclidean distance on $\mathbb{R}^k$.  

27
• (CA1) For all \( j, k \), there exists an \( \epsilon_0 \) such that \( \{ i : X_{ikj} > \epsilon_0 \} \) and \( \{ i : X_{ikj} < -\epsilon_0 \} \) are non-null.

• (CA2) Consider, for any \( k, j \)

\[
P_{n,k,j} = \{ i : 1 \leq i \leq n, X_{ikj} > \epsilon_0 \},
\]

\[
P'_{n,k,j} = \{ i : 1 \leq i \leq n, X_{ikj} < -\epsilon_0 \},
\]

\[
A_{n,k,j} = \{ i : 1 \leq i \leq n, X'_{ik,-j}^{-} \beta_0^{-} \leq 0 \).
\]

We assume \( A_{n,k,j} \cap P'_{n,k,j}, A_{n,k,j} \cap P_{n,k,j}, A_{n,k,j} \cap P'_{n,k,j} \) and \( A_{n,k,j} \cap P_{n,k,j} \) are all non-null.

CA2 is sufficient. For our model if CA2 holds for a fixed \( j \), then CA1 and CA2 only for that \( j \) are together sufficient.

• The link function \( H \) is symmetric and increasing.

**Theorem 2.1.** Under the assumed conditions, the posterior for \( \theta \) is consistent at \( \theta_0 \).

The first assumption (RE) is very standard assumption for random effect when there is a intercept term in the model. The assumption of symmetry is crucial since we drop the assumption that the whole real line is the support of the first continuous covariate made in Kim and Kim (2011). A prior that complies with the symmetry restriction is a random scale mixture of normal prior. For this prior, \( F = \int \phi_\sigma(\cdot)dG(\sigma) \) and the mixing distribution \( G \) is given a prior distribution with large weak support. In particular, if \( G \) is a Dirichlet process prior, the resulting prior is a Dirichlet process scale mixture of normal (DPSMN) prior. We used this prior in our methodology, but for posterior consistency we allow much more general priors. Our only requirements are that the true distribution of the random effects belongs to the weak support of the prior and has finite mean. The random scale mixture a prior generates symmetric strongly unimodal densities, and hence in order to meet the support requirement, the corresponding true distribution \( F_0 \) must also be of the form of a scale mixture of normal \( F_0 = \int \phi_\sigma dG \). Then the conditions hold if \( G_0 \) is contained in the weak support of the prior for \( G \) and \( \int \sigma dG(\sigma) < \infty \) a.s. with respect to the prior. The latter condition is implied by \( \int \sigma dG^*(\sigma) < \infty \), where \( G^* = EG \). For the DPSMN prior, if the true mixing distribution is contained in the support of the base measure of the Dirichlet process, the first condition holds and the finiteness of the mean can be ensured under the
much weaker condition $\int \log(1 + \sigma)dG^*(\sigma) < \infty$; see Ghosal and van der Vaart (2017) for the details. The condition of finite mean can be ensured easily. For the DPSNM prior, writing $F = \int \phi_{\sigma}(\cdot)dG(\sigma)$ and $G \sim DP(M, G_0)$, finiteness of the mean of $F$ is ensured with probability one if $\int \sigma dG_0(\sigma) < \infty$. Since $G_0 = N(0, 1)$, above conditions hold. However as we put DPSMN prior, much weaker condition $\int \log(1 + \sigma)dG_0(\sigma) < \infty$ is sufficient. The condition (Supp) that the true parameter values are in the supports of the corresponding prior distributions is a basic requirement for posterior consistency. The subgaussianity condition (SG) on the true distributions of random effects allows transition from weak support to Kullback-Leibler support, which is again vital for posterior consistency; see Chapter 6 of Ghosal and van der Vaart (2017).

In the methodology we actually used a graphon function to reduce dimension of the regression coefficients for each pair of regions. Clearly, then the prior is concentrated on a lower dimensional space, and hence to satisfy the support requirements, the truth must have a similar graphon based structure. Once the support condition is ensured, posterior consistency of the resulting Bayesian procedure can be derived as a corollary from Theorem 2.1. For our model

$$
\chi = ((\mu_0), \langle \chi_{\text{MCI}} \rangle, \langle \chi_{\text{AD}} \rangle, \langle \chi_{\text{M}} \rangle, \langle \chi_{\text{Age}} \rangle),
$$

$$
\beta = ((\pi_0), \langle \beta_{\text{MCI}} \rangle, \langle \beta_{\text{AD}} \rangle, \langle \beta_{\text{M}} \rangle, \langle \beta_{\text{Age}} \rangle),
$$

$$
\nu = ((\lambda_0), \langle \nu_{\text{MCI}} \rangle, \langle \nu_{\text{AD}} \rangle, \langle \nu_{\text{M}} \rangle, \langle \nu_{\text{Age}} \rangle).
$$

The parameter for each pair follows the graphon structure in (2.4) in reduced parameter space. In the abstract setting we can describe this as

$$
\chi = \zeta(\mu_0, \chi_{\text{MCI}}, \chi_{\text{AD}}, \chi_{\text{M}}, \chi_{\text{Age}}, \xi, \delta),
$$

$$
\beta = \zeta(\pi_0, \beta_{\text{MCI}}, \beta_{\text{AD}}, \beta_{\text{M}}, \beta_{\text{Age}}, \xi, \delta),
$$

$$
\nu = \zeta(\lambda_0, \nu_{\text{MCI}}, \nu_{\text{AD}}, \nu_{\text{M}}, \nu_{\text{Age}}, \xi, \delta),
$$

where the function $\zeta$ is continuous is its arguments and matrices are replaced by the corresponding graphon functions. Let the parameter space for graphon functions be $\mathcal{L} = \text{class of real-valued smooth continuous functions on } [0, 1]^2$. After this reparametrisation the new parameter will be $\tilde{\theta} = (\mu_0, \chi_{\text{MCI}}, \chi_{\text{AD}}, \chi_{\text{M}}, \chi_{\text{Age}}, \pi_0, \beta_{\text{MCI}}, \beta_{\text{AD}}, \beta_{\text{M}}, \beta_{\text{Age}}, \lambda_0, \nu_{\text{MCI}}, \nu_{\text{AD}}, \nu_{\text{M}}, \nu_{\text{Age}}, \xi, \delta)$. The parameter space of $\tilde{\theta}$ is $\mathcal{L}^{15} \times [0, 1]^2$. Let $\tilde{\theta}_0$ be the null value of $\tilde{\theta}$. All the prior assumptions will continue to hold except for the support condition of the prior of the fixed effects (Supp.). New support condition is

- (RSupp) The true distributions $F_i^0 \in \text{supp}(\Gamma_i)$, $i = 1, 2, 3$ and the support of the prior of the graphon functions are real-valued, smooth and continuous functions and contain the
true graphon functions.

**Corollary 2.2.** Under the support assumption of prior of \( \theta \) with \((\text{Supp})\) assuption replaced by \((R\text{Supp})\), the posterior for \( \tilde{\theta} \) is consistent at \( \tilde{\theta}_0 \).

### 2.6.1 Proof

We use the consistency result, given in theorem 2 of Amewou-Atisso et al. (2003). To proceed with this consistency result, we show prior positivity of the parameters and construct exponentially consistent tests. The whole likelihood can be split into three parts for the three model. It suffices to show prior positivity for the three models separately (i) normal, (ii) binomial and (iii) Poisson model.

**Lemma 2.3.** Consider a kernel function \( K(x; v, \eta) \), which is continuous in \( v \) and \( \eta \). For a fixed parameter \( v_0 \) and distribution \( F_0 \) of \( \eta \) if for a sequence \( (v_n, F_n) \) such that \( F_n \rightarrow F_0 \) and 
\[
\sup_{\eta \in \mathbb{R}} \|K(x; v_n, \eta) - K(x; v_0, \eta)\| \rightarrow 0,
\]
then we have 
\[
\int K(x; v_n, \eta) dF_n(\eta) \rightarrow \int K(x; v_0, \eta) dF_0(\eta).
\]
(Here \( \rightarrow \) denotes convergence.)

**Proof.** We can show
\[
\| \int K(x; v_n, \eta) dF_n(\eta) - \int K(x; v_0, \eta) dF_0(\eta) \| \leq \int \| K(x; v_n, \eta) - K(x; v_0, \eta) \| dF_n(\eta) + \int \| K(x; v_0, \eta) \| (dF_n(\eta) - dF_0(\eta)) \|
\]
This inequality proves the statement in the lemma. \( \square \)

**Prior positivity**

(i) **Normal model**

Let \( W_i^* = (\log W_{i1}, \ldots, \log W_{ip})' \), \( N_i = (N_{i1}, \ldots, N_{ip})' \), \( X_i = (X_{i1}, \ldots, X_{ip})' \) and \( D_i = \sum_j N_{ij} \). Then we can write

\[
p_{i,\theta}(W_i^*|x_i) = \int_{\mathbb{R}} \prod_j \frac{1}{\sqrt{2\pi\sigma_j / \sqrt{N_{ij}}}} \phi\left( \frac{\log W_{ij} - X_{ij}\chi - \eta_{i1}}{\sigma / \sqrt{N_{ij}}} \right) dF_1(\eta_{i1}).
\]

Defining \( \bar{\sigma} = \max_j \sigma_j / \sqrt{N_{ij}} \) and \( \sigma = \min_j \sigma_j / \sqrt{N_{ij}} \) and \( \nu_j = \sum_k |X_{ijk}| \), we obtain

\[
p_{i,\theta}(W_i^*|x_i) \geq \int_{\mathbb{R}} \prod_j \frac{1}{\sqrt{2\pi\bar{\sigma}}} \phi\left( \frac{\log W_{ij} - X_{ij}\chi - \eta_{i1}}{\bar{\sigma}} \right) dF_1(\eta_{i1})
\]
\[
\geq \int_{-B_1}^{B_1} \frac{1}{2\sqrt{2\pi\bar{\sigma}}} \phi\left( \frac{\sum_j \log(W_{ij}) + M_2 \sum_j \nu_j + B_1}{\bar{\sigma}} \right).
\]

30
Let \( T_i = |\sum_j \log W_{ij}|. \) For some positive constant \( c \) depending on \( B_1 \), we have that

\[
p_i,\theta_0(W^*_i|x_i)/p_i,\theta(W^*_i|x_i) \lesssim \exp(cT_i^2).\]

This implies that for some \( \delta' > 0 \),

\[
E_{F_1}\left(\frac{p_i,\theta_0(W^*_i|x_i)}{p_i,\theta(W^*_i|x_i)}\right)^{\delta'} < \infty. \tag{2.6}
\]

For two probability densities \( f \) and \( g \), define the Hellinger distance \( d_H \) by \( d_H^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2 \), the Kullback-Leibler divergence \( K(f, g) = \int f \log(f/g) \) and the Kullback-Leibler variation \( V(f, g) = \int f \log^2(f/g) \). Applying Theorem 5 of Wong and Shen (1995), from (2.6), we obtain

\[
K(p_i,\theta(W^*_i|x_i), p_i,\theta_0(W^*_i|x_i)) \lesssim \epsilon_1^2 \log \frac{1}{\epsilon_1},
\]

\[
V(p_i,\theta(W^*_i|x_i), p_i,\theta_0(W^*_i|x_i)) \lesssim \epsilon_1^2 \log^2(1/\epsilon_1),
\]

whenever \( \epsilon_1 = d_H(p_i,\theta(W^*_i|x_i), p_i,\theta_0(W^*_i|x_i)) < 1/2 \). Thus to verify the condition on Kullback-Leibler support, we only need to verify that Hellinger neighborhoods of \( p_i,\theta_0(W^*_i|x_i) \) have positive prior probabilities, or equivalently \( L_1 \)-neighbors since the Hellinger distance induces the same neighborhood system as the \( L_1 \)-distance on probability densities. In view of Lemma 2.3, the condition is ensured by the assumed support condition (Supp).

(ii) Binomial model

Prior positivity follows from the argument in Kim and Kim (2011). They showed that if \( \beta \) converges to \( \beta^0 \) and \( F_2 \) converges weakly to \( F_2^0 \), then Kullback-Leibler divergence and variation (write these explicitly) tend to zero.

(iii) Poisson model

We can write the probability mass function of the observation \( N_i \) as

\[
p_i,\theta(N_i|x_i) = C \int \prod_j \left( \exp \left( -\eta_{3i} e^{X_{ij}'\nu} \right) \left( e^{\eta_{3i} e^{X_{ij}'\nu}} \right)^{N_{ij}} \right) dF_3(\eta_{3i})
\]

\[
= C \int \exp \left( -\eta_{3i} \sum_j e^{X_{ij}'\nu} + \eta_{3i} \sum_j N_{ij} \right) \times \prod_j \left( e^{X_{ij}'\nu} \right)^{N_{ij}} dF_3(\eta_{3i}), \tag{2.7}
\]

where \( C^{-1} = \prod_j N_{ij}! \) is free of the parameters.

By applying Jensen’s inequality and then using the obvious fact that \( N_{ij} \) for any \( j \) is less than or equal to its sum over the argument, we can write \( \prod_j (e^{X_{ij}'\nu})^{N_{ij}} \) as

\[
\left( \exp \left( \sum_j \frac{N_{ij} X_{ij}'\nu}{\sum_j N_{ij}} \right) \right)^{N_{ij}} \leq \left( \frac{\sum_j N_{ij} e^{X_{ij}'\nu}}{\sum_j N_{ij}} \right)^{\sum_j N_{ij}} \leq \left( \sum_j e^{X_{ij}'\nu} \right)^{\sum_j N_{ij}}.
\]
Substituting above inequality in (6) and observing that the integrand is maximized at \( \eta_{3i} = \log(\sum_j N_{ij}/\sum_j e^{X_{ij} \nu_j}) \), we obtain the upper bound for (2.7) as

\[
p_i,\theta(N_i|x_i) \leq C\left(\sum_j N_{ij}\right)^{\sum_j N_{ij}} \exp(-\sum_j N_{ij}) = CD_i^{D_1}e^{-D_i},
\]

where \( D_i = \sum_j N_{ij} \).

Letting \( v_{ij} = |X_{ij}| \) and \( \max_j |\nu_j| = M_3 \), it follows that \( p_i,\theta(N_i|x_i) \) is lower bounded by

\[
\frac{C}{2} \int_{-B_3}^{B_3} \exp \left( -e^{\eta_{3i}} \sum_j e^{v_{ij}M_3} + \eta_{3i} \sum_j N_{ij} \right) \left( \sum_j e^{-v_{ij}M_3} \right) \sum_j N_{ij} dF_3(\eta_{3i}),
\]

which leads to the estimate \( p_i,\theta_0(N_i|x_i)/p_i,\theta(N_i|x_i) \leq D_i^{D_1} \). Observe that if \( N \) is a random variable distributed as Poisson with parameter \( \lambda > 0 \), then for any \( \delta < 1 \), \( \text{E} \exp(\delta N \log N) < \infty \).

This holds because \( \log n! \geq n \log n - n \) and that \( \sum_{n=1}^{\infty} \exp(\delta n \log n + n \log \lambda - n \log n + n) < \infty \) whenever \( \delta < 1 \). Therefore by Theorem 5 of Wong and Shen (1995), we obtain the bounds

\[
K(p_i,\theta(N_i|x_i),p_i,\theta_0(N_i|x_i)) \lesssim \epsilon_2^2 \log \frac{1}{\epsilon_2},
\]

\[
V(p_i,\theta(N_i|x_i),p_i,\theta_0(N_i|x_i)) \lesssim \epsilon_2^2 \left( \log \frac{1}{\epsilon_2} \right)^2,
\]

where \( \epsilon_2 = d_H(p_i,\theta(N_i|x_i),p_i,\theta_0(N_i|x_i)) < 1/2 \). By Lemma 2.3 and the assumed condition (Supp), the condition on the Kullback-Leibler support is verified.

**Test construction**

(iii) *Normal model*

If we reduce the data into binary observations by defining \( S_{ij} = I(W_{ij} \geq 0) \). Then \( S_{ij} \) will follow a probit model. If we can construct tests for this reduced data, test construction for the whole data will continue to hold. Test construction for binary data is given below for a symmetric increasing link function. This construction will hold for probit model as well.

(iii) *Binomial model*

Tests for the binary data with symmetric increasing link function are constructed below.

Let,

\[
T_1^i(\Delta_1) = \{ (\beta^*, F_2) : |\beta^*_i - \beta^0_i| > \Delta_1 \},
\]
\[ T^l_1(\Delta_1, \Delta_l) = \{(\beta^*, F_2) : |\beta^*_1 - \beta^*_1| \leq \Delta_1, |\beta^*_l - \beta^*_l| > \Delta_l, l = 2, \ldots, p \} \]

\[ T^{p+1}_1(\Delta_1, \ldots, \Delta_{p+1}) = \{(\beta^*, F_2) : |\beta^*_1 - \beta^*_1| \leq \Delta_1, l = 1, \ldots, p \text{ and } d_w(F_2, F^0_2) > \Delta_{p+1} \} \]

Consider, \( g(x) = E_{F_2}[H(x + \eta_2)] - E_{F_2}[H(x + \eta_2)] \).
\( K_n = \{ i : 1 \leq i \leq n, g(X'_{i1} - \beta^0_1) \leq 0 \} \).
\( P_n = \{ i : 1 \leq i \leq n, X_{i11} \geq 0 \} \).
\( A_n = \{ i : 1 \leq i \leq n, X'_{i11} \beta^0_1 \leq 0 \} \).

Notation: A set \( Q_n \) is (asymptotically) non-null if \( \liminf \frac{|Q_n|}{n} > 0 \), where \(|.| \) denotes cardinality of a set.

We have \( g(\infty) = g(-\infty) = 0 \) and \( g(x) = -g(-x) \) because \( H(x) + H(-x) = 1 \). This implies \( g(0) = 0 \). We have \( H''(x) = -H''(-x) \), this follows \( g''(x) = -g''(-x) \). We have \( g''(0) = 0 \), so \( x = 0 \) is a point of inflection. If \( F_2 \not= F^0_2 \), we would find an interval \([-a, a] \) such that \( g(x) \neq 0 \) for \( x \in [-a, 0] \). There are two possibilities \( g(x) \) and \( x \) are of same sign or of opposite signs. Both of the two cases are equivalent. Let at \( x = \pm t \), \( g'(x) = 0 \). We show test construction for the case where \( g(x) \) and \( x \) are of same sign. So, we have \( g(x) > 0 \) for \( x \in (0, a) \) and \( g(x) < 0 \) for \( x \in [-a, 0) \). We have, \( E_{F_2, \beta}(I_{i1}) - E_{F^0_2, \beta^0}(I_{i1}) = E_{F_2}H(X'_{i1} \beta + \eta_2) - E_{F_2}H(X'_{i1} \beta^0 + \eta_2) + g(X'_{i1} \beta^0) \). We can create difference in mean if either \( X'_{i1} \beta \leq X'_{i1} \beta^0, X'_{i1} \beta \leq X_{i11} \beta^0 \) or \( X'_{i1} \beta > X'_{i1} \beta^0, X_{i1} \beta > X_{i11} \beta^0 \). To summarize difference in mean will exists when \( X'_{i1} \beta - X_{i1} \beta^0 + X_{i11} (\beta_1 - \beta^0) > 0 \). Assumption CA2 will ensure \( K_n \cap P^l_n, K^c_n \cap P^c_n \) and \( K_n \cap P_n \) are non-null.

For \( \beta_1 - \beta^0_1 > \Delta_1 \), \( (X'_{i1} \beta_1 - X_{i1} \beta^0_1 + X_{i11} (\beta_1 - \beta^0_1)) (X'_{i1} \beta^0) > (X'_{i1} \beta_1 - X_{i1} \beta^0_1 + X_{i11} \Delta_1) (X'_{i1} \beta^0) \), when \( X_{i11} (X'_{i1} \beta^0) \) is positive which is true as \( K_n \cap P^c_n \) and \( K_n \cap P_n \) are non-null.

Above inequality will be greater than zero for the case \( X'_{i1} \beta_1 > X_{i1} \beta^0_1 \) with elements from \( K^c_n \cap P_n \) and for \( X'_{i1} \beta_1 < X_{i1} \beta^0_1 \) we choose elements from \( K_n \cap P^c_n \). The other possibility is \( \beta_1 - \beta^0_1 < -\Delta_1 \). Then we have, \( (X'_{i1} \beta_1 - X_{i1} \beta^0_1 + X_{i11} (\beta_1 - \beta^0_1)) (X'_{i1} \beta^0) > (X'_{i1} \beta_1 - X_{i1} \beta^0_1 - X_{i11} \Delta_1) (X'_{i1} \beta^0) \) with elements from \( K_n \cap P_n \) and \( K^c_n \cap P^c_n \). This time we choose \( K^c_n \cap P^c_n \) for \( X'_{i1} \beta_1 > X_{i1} \beta^0_1 \) case and \( K_n \cap P_n \) for \( X'_{i1} \beta_1 < X_{i1} \beta^0_1 \).

For \( T^l_1 \), the difference in mean is created when

\[ \{ i : |X_{i11}| < |(X'_{i1} \beta_1 - X_{i1} \beta^0_1) - |\Delta_1 X_{i11}|/\Delta_1 | \} \]

is non-null. We can choose \( \Delta_1 \) accordingly depending on \( \Delta_1 \) such that above set is non-null.
(iii) Poisson model
Let,
\[ T_2^1(\Delta_1) = \{ (\nu^*, F_3) : |\nu_1^* - \nu_0^*| > \Delta_1 \}, \]
\[ T_2^l(\Delta_1, \Delta_l) = \{ (\nu^*, F_3) : |\nu_1^* - \nu_0^*| \leq \Delta_1, |\beta_1^* - \beta_0^*| > \Delta_l, l = 2, \ldots, p, \]
\[ T_2^{p+1}(\Delta_1, \ldots, \Delta_{p+1}) = \{ (\nu^*, F_3) : |\nu_1^* - \nu_0^*| \leq \Delta_1, l = 1, \ldots, p \text{ and } d_{we}(F_3, F_3^0) > \Delta_{p+1} \}. \]

Test construction for \( T_2^1 \) is discussed here. We have \( E_{F_3, \nu}[N_{ij}] - E_{F_3^0, \nu}[N_{ij}] = [\exp(X'_{ij} \nu + c_1) - \exp(X'_{ij} \nu^0 + c_0)] \), where, \( c_1 = \log(E_{F_3}[\exp(\eta_{ih})]) \) and \( c_0 = \log(E_{F_3^0}[\exp(\eta_{ih})]) \). This is negative or positive if \( (X'_{ij}(\nu - \nu^0) + c_1 - c_0) \leq 0 \). Now the construct is same as previous case. For the case \( X'_{i1-1}\beta_1 - X'_{i1-1}\beta_0 + (c_1 - c_0) > 0 \), we consider elements from the set \( P_n \) or \( P_n^c \), depending on the hypothesis \( \beta_1 - \beta_0 > \Delta_1 \) or \( \beta_1 - \beta_0 < -\Delta_1 \). Other case will be similar. Test construction for \( T_2^p \) will be similar to the case in \( T_2^1 \).

For \( T_1^{p+1} \), we can follow the part of the proof from Kim and Kim (2011). Since random effect does not have any co-efficient, we can use similar steps from that paper and we do not need to bound the random effects for our case. If we had co-efficients to the random effects, we could have used the sets \( K_n \) or \( K_n \) and previous statistics to construct the test. But that is not required in this case. We can directly follow steps from Kim and Kim (2011).

For the last part \( T_2^{p+1} \), define \( C_F(l) = \int \exp(l + Jb) dF_3(b) \).

According to the model, \( D_i \sim \text{Poisson}(\sum_j X'_{ij} \nu + J_{ih}) \). So, \( E_{F_3}(D_i|X_i) = \int \exp(\sum_j X_{ij} \nu + Jb) dF_3(b), \)

\[ E_{F_3}(D_i|X_i) - E_{F_3^0}(D_i|X_i) = \int \exp(\sum_j X'_{ij} \nu + Jb) dF_3(b) - \int \exp(\sum_j X'_{ij} \nu^0 + Jb) dF_3^0(b) \]
\[ > \int \exp(\sum_j X'_{ij} \nu^0 - \sum_j X'_{ij} \Delta_f + Jb) dF_3(b) - \int \exp(\sum_j X'_{ij} \nu^0 + Jb) dF_3^0(b) \]
\[ = \exp(\sum_j X'_{ij} \nu^0) \left( \int \exp(- \sum_j X'_{ij} \Delta_f + Jb) dF_3(b) - \int \exp(Jb) dF_3^0(b) \right) \]
\[ = \exp(\sum_j X'_{ij} \nu^0) \left( C_{F_3}(\sum_j X'_{ij} \Delta_f) - C_{F_3^0}(0) \right) \]
\[ > \exp(\sum_j X'_{ij} \nu^0) \left( C_{F_3}(\sum_j |X'_{ij}| \Delta_f) - C_{F_3^0}(0) \right). \]
Observe, here \( m = - \sum_j |X_i^j| \Delta f \) is always negative.

**Lemma 2.4.** There exists \( \tau_1 > 0 \) such that \( \sup_{l \in \mathbb{R}} |C_{F_3}(l) - C_{F_3^0}(l)| > \tau \) for all \( F_3 \) with \( d(F_3, F_3^0) > \Delta_{p+1} \).

**Proof.** Suppose there is no \( \tau > 0 \). Then we can construct a sequence of \( F_3 \), namely \( F_{3,n} \) such that \( \sup_{l \in \mathbb{R}} |C_{F_3}(l) - C_{F_3^0}(l)| < 1/n \) and \( d(F_3, F_3^0) > \Delta_{p+1} \). Then \( C_{F_3,n} \) converges in distribution to \( C_3 \). And from proof of Lemma 4.2 in Kim and Kim (2011), \( F_{3,n} \) converges to \( F_3^0 \). That is a contradiction.

For \( l \leq 0 \exp(l) \leq 1 \), the mean value theorem gives us \( |C_{F_3}(l_1) - C_{F_3}(l_2)| \leq |l_1 - l_2| \), for \( l_1 \leq 0, l_2 \leq 0 \).

Now, for a given \( F_3 \) with \( d(F_3, F_3^0) > \Delta_{p+1} \), \( l^* = \text{argmin}_{l \in \mathbb{R}} |C_{F_3}(l) - C_{F_3^0}(l)| \).

Now, we can have \( C_{F_3}(l^*) - C_{F_3^0}(l^*) > \tau \) or \( < -\tau \).

For first case take \( l^* < 0 \) such that \( |m - l^*| \leq \psi \), then

\[
C_{F_3}(l^*) - C_{F_3^0}(l^*) \leq C_{F_3}(m) - C_{F_3^0}(m) + 2\psi \leq C_{F_3}(m) - C_{F_3^0}(0) - m + 2\psi,
\]

and in the other case,

\[
C_{F_3}(l^*) - C_{F_3^0}(l^*) \geq C_{F_3}(m) - C_{F_3^0}(m) - 2\psi \geq C_{F_3}(m) - C_{F_3^0}(0) + m - 2\psi
\]

We need to choose \( \Delta_f \) accordingly, to get \( \tau + m - 2\psi > 0 \). This creates difference in the means which ensures existence of exponentially consistent test by Lemma 7.3.1 in Ghosh and Ramamoothi (2003).

This ends the proof of posterior consistency.

**Proof of the corollary**

After the dimension reduction, test constructions will remain the same. Only Kullback-leibler prior positivity needs to be verified. We have shown if \( (\chi, \beta, \nu) \) converge to \( (\chi^0, \beta^0, \nu^0) \) and corresponding random effect distributions converge to true distributions, Kullback leibler converges to zero. In dimension reduction set up, \( ||\chi - \chi^0||_2 \leq ||\mu - \mu^0||_\infty + \sum_l ||\chi_l - \chi^0_l||_\infty + ||\xi - \xi^0||_2 + ||\delta - \delta^0||_2, ||\beta - \beta^0||_2 \leq ||\pi - \pi^0||_\infty + \sum_l ||\beta_l - \beta^0_l||_\infty + ||\xi - \xi^0||_2 + ||\delta - \delta^0||_2, ||\nu - \nu^0||_2 \leq ||\lambda - \lambda^0||_\infty + \sum_l ||\nu_l - \nu^0_l||_\infty + ||\xi - \xi^0||_2 + ||\delta - \delta^0||_2 \) for \( l = \text{MCI, AD, M} \) with \( \xi \) and \( \delta \) are the latent variables and \( \mu, \chi_l, \pi, \beta_l, \lambda, \nu_l \) are the co-efficient functions. If the co-efficient functions are close to the true functions and the latent variables are close to the truth, then \( (\chi, \beta, \nu) \) is close to \( (\chi^0, \beta^0, \nu^0) \). Thus Kullback leibler converges to zero. This completes the proof for reduced dimension set up.
2.7 Conclusions and discussion

We study the effects of some common measurable covariates on the human brain connectome. Our work extends statistical inference on graph structure from the two sample problem (Tang et al. (2017)) to a more general regression modeling framework. We propose regression models to explain the extent of connections between different cortical brain regions. In this setup, traditional techniques of separately regressing connections, between each pair of regions and the covariates are not appropriate due to missingness at several edges of the connectome. We solve this problem by using graphon functions to reduce the dimension of the parameter space through a fewer number of fundamental parameters and developed a Bayesian method to estimate those. Subject inhomogeneity is incorporated through random effects and the distributions of the random effects are estimated by using Dirichlet process scale mixture of normal (DPSMN) prior. For well-specified case, for a sample of size more than 25 the proposed Bayesian method performs better than the ANCOVA (see Figure 2.3).

Previous studies have demonstrated that the entorhinal cortex is one of the earliest and most significant areas of pathologic involvement and neurodegeneration in AD (Van Hoesen et al. (1991)). Interestingly, our model demonstrates that structural connections in the right and left entorhinal cortex are significantly changed in cases of MCI and AD (see Figures 2.6 (a), (b) and 2.7 (a), (b)). Recent studies have suggested that the core pathologies of AD may propagate their effects based on the structural connectivity, with tau either directly spreading from the temporal lobe to other areas of the brain along white matter connections, or causing impaired signaling from the temporal lobe to connected regions. It is also thought that increased beta amyloid may accelerate the effects of tau, and possibly even the spread of tau outside of the temporal lobe (Khan et al. (2014)). Our results support the suggestion that abnormalities in white matter connectivity between the entorhinal cortex and other cortical regions in the brain are important in MCI and AD. Note that the prominence of significant self-edges, which in general are shorter connections, may be related to the tractography algorithm itself, which tend to extract shorter edges more consistently. The longer the connection the more likely that the tractography algorithm will terminate early due to preset termination parameters, as demonstrated in multiple previous structural and functional connectivity studies of AD (Stam (2014)).
Chapter 3

High dimensional Single Index Bayesian Modeling of the Brain Atrophy over time

This chapter’s material is a joint work with Subhashis Ghosal and Kingshuk Roy Choudhury.

3.1 Introduction

Alzheimer disease (AD) is a progressive neurodegenerative disease that affects approximately 5.5 million people in the United States and about 30 million people worldwide. It is believed to have a prolonged preclinical phase initially characterized by the development of silent pathologic changes when patients appear to be clinically normal, followed by mild cognitive impairment (MCI) and then dementia (AD) (Petrella (2013)). Apart from its manifestation in the impairment of cognitive abilities, disease progression also produces a number of structural changes in the human brain, which includes the deposition of amyloid protein and the shrinkage or atrophy for certain regions of the brain over time (Thompson et al. (2003)). Previous studies have shown that the rate of brain atrophy is significantly modulated by a number of factors, such as gender, age, baseline cognitive status and most markedly, allelic variants in the APOE gene (Hostage et al. (2014)). In this chapter, we examine if any other genes are also implicated in modulating the rate of brain atrophy. This analysis represents a technical challenge because the genomic data is high dimensional and needs to be incorporated in a model for longitudinal progression of brain volumes measured in multiple parts of the brain.

Our study uses the data collected by Alzheimer Disease Neuroimaging Initiative (ADNI). In this dataset, volumes of thirteen disjoint brain regions are recorded over six visits after every six
months. With age these volumes change and Alzheimer disease makes these changes much more pronounced. Genetic variations also have impacts on the configuration of the brain. Factors such as gender, age etc. are also responsible for these changes. In this chapter, we propose a model for the volume of different brain regions to quantify the effects of different factors on the volume of brain regions. The regions, we studied here, are depicted in the Figure 3.1. This image is obtained from Ahveninen et al. (2012).

Figure 3.1: Anatomic parcellation of cortical surface from different angles showing brain regions used for analysis.

We consider volumetric measurements of thirteen disjoint brain regions that are recorded over time along with that for volume of the grey matter. At a given time point, the volume of a particular region will depend on the initial configuration and rate of change of that region. We consider two sets of unknown functions of covariates to model these initial configurations and rates of changes for different regions with two inputs, one from a high dimensional SNP and another from a low-dimensional vector other covariates. These functions will look like $\{a_{0,j}(X'\beta, Z'\eta)\}$ for the initial configuration and $\{a_{1,j}(X'\beta, Z'\eta)\}_{1\leq j \leq 14}$ for the rate of change in $j$-th region; here $X$ and $Z$ are high and low dimensional covariates respectively. These functions are modeled by a bivariate single index model and a finite random series prior is put on the function based on tensor products of B-splines with appropriate prior distribution on the coefficients. We reparametrize the coefficients of the covariates in the single index to ease the estimation of this model. After the reparametrization, due to high dimensionality of the SNP
data we carefully put prior for a desired shrinkage in the estimate. To incorporate the effect of time for analyzing our dataset, we consider an increasing function of time into the model. This is also estimated nonparametrically with a finite random series prior on the function based on B-spline with appropriate prior on the coefficients.

Apart from modeling volume of brain regions over time in terms of SNP and external covariates and obtaining Bayes estimates, the proposed method develops an estimation scheme for a general high dimensional single index model. Estimation for high dimensional single index model is addressed in Zhu and Zhu (2009), Yu and Ruppert (2002), Wang and Zhu (2017), Peng and Huang (2011), Radchenko (2015) and Luo and Ghosal (2016). All of them used $\ell_1$ penalty and worked out an optimization technique to get the estimates. In the Bayesian framework, Antoniadis et al. (2004) used Fishervon Mises prior on the directional vector. This can not be easily modified for high dimensional case as then we need a prior which favors larger number of zeros in the unit vector. That will make the prior too complex to use. Another paper addressing sparse Bayesian single index model estimation is Alquier and Biau (2013). Even though their method is theoretically attractive, due to high computational complexity it is difficult to implement when the number of covariates is very high. In Wang (2009), they developed a sparse Bayesian estimation for sparse single index model using reversible jump Markov chain Monte Carlo technique which puts a lot of computational burden in estimating the model parameters. These methods are extremely difficult to implement for extreme high dimensional cases. For affordable computation using our method, we provide an R package for estimation in single index model with inputs both in high and low dimensional setup.

The rest of the chapter is arranged in the following manner. The next section discusses the dataset and modeling in more detail. In Section 3.3, we describe the prior on different parameters of the model. Section 3.4 describes posterior computation in this setup. We provide a simulation study comparing the proposed Bayesian procedure with its linear counterpart in Section 3.5. We study posterior rate of contraction of the model in Section 3.6 under the asymptotic regime where the number of individuals goes to infinity but the number of time points where measurements are taken and the number of regions are fixed. The posterior consistency result justifies the use of the proposed Bayesian procedure from a frequentist perspective. In Section 3.7 we present the results from a real data analysis. Section 3.8 concludes the chapter with some further remarks.

### 3.2 Data description and modeling

Data used in the preparation of this article were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary
goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to predict the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). In the ADNI dataset, the grey matter part of the brain is divided into thirteen disjoint regions. The volume of these regions and the whole brain are recorded over time for \( n = 748 \) individuals. The number of visits is not uniform across individuals but varies between 1 to 6. The volume data of \( J = 14 \) components of brain over \( T_i \) many time points for the \( i-th \) individual, for \( i = 1, \ldots, 748 \) is collected where \( 1 \leq T_i \leq 6 \).

Apart from the volumetric measurements, we also have high dimensional SNP data and data on some other covariates for each individual. The other covariates are gender, disease state, age and allele 2 and 4 of APOE gene. Except for the covariate age, all other low-dimensional covariates are categorical. Since the disease status has three states: NC (no cognitive impairment), MCI (mild cognitive impairment) and AD (Alzheimer disease). We consider two dummy variable \( Z^{AD} \) and \( Z^{MCI} \) respectively standing for the onset of MCI and AD, setting NC at the baseline. Similarly, the dummy variable \( Z^M \) indicating male gender is introduced setting females at the baseline. Also, we introduce \( Z^{APOE,2}, Z^{APOE,4} \) standing for alleles 2 and 4 for the two alleles APOEallele2 and APOEallele4 together setting allele 3 as the baseline for each of the two cases. Let \( Z = (Z^{MCI}, Z^{AD}, Z^M, \text{Age}, Z^{APOE,2}, Z^{APOE,4}) \) stand for the whole vector of covariates. The continuous variable, Age, is standardized. Along with that, we have SNP data for each individual. In ADNI, the subjects were genotyped using participants the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) was used. We have a set of 620,901 SNP and copy number variation (CNV) markers. The APOE gene has been the most significant gene in GWAS of Alzheimer disease. The corresponding SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip. At the time of participant enrollment APOE genotyping was performed and included in the ADNI database. The two SNPs (rs429358, rs7412) define the epsilon 2, 3, and 4 alleles and therefore were genotyped using DNA extracted by Cogenics from a 3 mL aliquot of EDTA blood. These alleles are considered separately in the study.

With time, different brain regions change differently. We study the effects of different attributes to these changes. For every individual, the volume of a brain region on a particular visit should primarily depend on the volume of that region at the zeroth visit and the rate of change of volume for that region with time. These components are not uniform across individuals or regions. Hence, it will be logical to consider that the baseline volume, recorded at the initial visit and the rate of change as a function of the region and the individual covariates. As the geometry of brain structure is complicated, we don’t assume a form of standard spatial dependence between measurements across brain regions. Thus we need two sets \( \{(a_{0,j}(\cdot), a_{1,j}(\cdot))_{1 \leq j \leq 14}\} \) of
functions for modeling volume at the initial visit and the rate of change for the \( j \)-th region. These functions are unknown and are modeled nonparametrically. For nonparametric regression problems, single-index models provide a lot of flexibility in estimation and interpretation of the results. Hence, we adopt the bivariate single-index model with two inputs for high-dimensional and low-dimensional covariates separately for easy interpretation and computational efficiency. The effect of time is captured through an unknown increasing function \( F_0(\cdot) \), which is bounded in \([0, 1]\). This is also modeled nonparametrically.

Thus the data generating process can be represented through the following specification

\[
Y_{ijt} = F(i, j, t) + \epsilon_{ijt}, \epsilon_{ijt} \sim N(0, \sigma^2), \quad (3.1)
\]

\[
F(i, j, t) = a_{0,j}(X_i^t \beta, Z_{t,i}^j \eta) - a_{1,j}(X_i^t \beta, Z_{t,i}^j \eta) F_0(t),
\]

where \( Y_{ijt} \) is the volume of the \( j \)-th brain region for the \( i \)-th individual at the \( t \)-th time point in logarithmic scale, \( X_i \) is high-dimensional SNP expression of length \( p \) for the \( i \)-th individual and \( a_{0,j}(\cdot), a_{1,j}(\cdot), F_0(\cdot) \) are all unknown functions satisfying the condition that \( F_0 \) is continuous, monotone increasing function form \([0, 1]\) onto \([0, 1]\); and \( N \) stands for normal distribution. For identifiability of the functions along with the parameters \( \beta \) and \( \eta \), we assume that \( \|\beta\|_2 = 1 \), \( \|\eta\|_2 = 1 \) and that the first non zero coefficients of \( \beta \) and \( \eta \) are positive; here \( \|\cdot\|_2 \) denotes \( L_2 \)-norm of a vector. The biggest challenge for estimation in this model is the high dimensionality of \( \beta \). We normalize the covariates for each individual i.e. \( X_i \) and \( Z_{t,i} \) for each combination of \((t, i)\) such that the norm is one. This is to make the inputs of the functions bounded between \([-1, 1]\].

We view \( Z_i \) as an abstract covariate of dimension \( k \). Due to the high dimensionality of \( \beta \), we propose a sparse estimation scheme. First we reparametrize the two unit vector \( \beta = (\beta_1, \ldots, \beta_p) \) and \( \eta = (\eta_1, \ldots, \eta_k) \) to their respective polar forms which allows us to work with Euclidean spaces. In the polar setup, for \( s \leq p - 1 \), \( \beta_s = \prod_{l=1}^{s-1} \sin \theta_l \cos \theta_s \), and \( \beta_p = \prod_{l=1}^{p-1} \sin \theta_l \) where \( \{ \theta = (\theta_1, \ldots, \theta_{p-1}) \} \) is the polar angle corresponding to the unit vector \( \beta \). Here \( \theta_s \in [0, \pi] \) for \( s \neq (p - 1) \) and \( \theta_{p-1} \in [0, 2\pi] \). Similarly, let \( \alpha \) be the polar angle corresponding to \( \eta \). Then for \( s \leq k - 1 \), \( \eta_s = \prod_{l=1}^{s-1} \sin \alpha_l \cos \alpha_s \) and \( \eta_k = \prod_{l=1}^{k-1} \sin \alpha_l \).

### 3.3 Prior specification

In the nonparametric Bayesian setup described above, we induce prior distributions on the smooth functions \( a_{0,j}^d \) and \( a_{1,j}^d \) in (3.1) through basis expansions in tensor products of B-splines and prior on the corresponding coefficients. Given other parameters in this setup, a normal prior distribution on the coefficients of the tensor products of B-splines will lead to conjugacy and faster sampling via Gibbs sampling scheme. An inverse gamma prior on \( \sigma^2 \) is an obvious choice due to conjugacy and faster sampling. We also put a B-spline series prior on the smooth
increasing function of time $F_0(\cdot)$. The coefficients for this function would be increasing in the index of the basis functions and lie in $(0, 1]$. To put prior on an increasing sequence, we introduce a set of latent variables of size equal to number B-spline coefficients. Then the B-spline coefficients would be normalized cumulative sum of those latent variables. Other two parameters $\beta$ and $\eta$ are reparametrized into their polar coordinate system. The parameter space of the polar angles will be a hyperrectangle. It will be easier to put prior on polar angles. To estimate using the sparsity of $\beta$, we need to carefully put a shrinkage prior on the polar angles. A polar angle of $\pi/2$ will ensure that the corresponding coordinate in the unit vector equal to zero. When there is sparsity in the unit vector, most of polar angles will become $\pi/2$. Thus a spike and slab prior on the polar angle with the spike at $\pi/2$ should be able to capture sparsity in the corresponding unit vector. Last polar angle has both at 0 and $\pi/2$, due to the structure of last and penultimate coordinate of a unit vector.

The priors are described in detail below:

(i) Intercept and slope functions:

$$a_0(x, y) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{0,mm'}^j B_m(x)B'_m(y),$$

$$a_1(x, y) = \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{1,mm'}^j B_m(x)B'_m(y),$$

-the function coefficients: For some chosen $a > 0$,

$$\lambda_{t,mm'}^j \sim N(0, a^2), \quad \lambda_{1,mm'}^j \sim N(0, a^2), \quad 1 \leq m, m' \leq K.$$

(ii) The function of time:

$$F_0(x) = \sum_{m=1}^{K} \lambda_mB_m(x),$$

-for the above function we need $\lambda_1 = 0$ and $\lambda_K = 1$. For estimation of the coefficients we introduce a new set of latent variables $(\delta_1, \ldots, \delta_{K-1})$

$$\delta_i \sim \text{Un}(0, 1), \quad \lambda_{m+1} = \frac{\sum_{i=1}^{m} \delta_i}{\sum_{i=1}^{K-1} \delta_i},$$

for $m = 1, \ldots, K-1$ and here Un stands for the uniform distribution. This approach to model increasing co-efficients was earlier used by Das and Ghosal (2017).

(iii) Error variance: $\sigma^{-2} \sim \text{Ga}(d_1, d_2)$, where Ga stands for the gamma distribution.
(iv) Polar angles $\alpha$ of $\eta$: We put independent $\text{Un}(0, \pi)$ prior on $\alpha_i$ for $i = 1, \ldots, (k - 2)$ and $\alpha_{k-1} \sim \text{Un}(0, 2\pi)$

(v) Polar angles $\theta$ of $\beta$: We reparametrize the polar angles as $\theta_i' = 1 - \frac{\min(\theta_i, \pi - \theta_i)}{\pi/2}$ for $1 \leq i \leq (p - 2)$. And $\theta_p'' = \frac{\theta_{p-1}}{\pi/2} - 1(\theta_{p-1} > \pi/2) - 1(\theta_{p-1} > \pi) - 1(\theta_{p-1} > 3\pi/2)$. Then we would reduce it further to $\theta_i' = \frac{4}{\pi} \min(\theta_i'', \pi/2 - \theta_i'')$. Then we put a spike and slab prior $\theta_i' \sim (1 - \gamma)\text{Be}(M_1, M_2) + \gamma \text{Un}(0, 1)$, where $\text{Be}$ stands for the Beta distribution.

To incorporate for first non-zero coefficient to be positive, the support for the polar angles up to the first non-zero coefficient in $\beta$ or $\eta$ is set to $[0, \pi/2]$. The spike distribution on $\theta$ looks like Figure 3.2. The first plot is for the first $p - 2$ angles and the second plot is for the last polar angle.

Model selection: Polar angles with posterior probability more than 0.5 are considered in the model.

![Figure 3.2: Spike distribution with $M_1 = 0.01$ and $M_2 = 10$.](image)
3.4 Computation

Introduce a latent variable $I_i$ for the indicator of the Un(0,1) component of the distribution of $\theta_i^t$, $l = 1, \ldots, p - 1$. Now the log-likelihood is given by

\[
C - \sum_{i,j,t} \frac{1}{2\sigma^2} (Y_{ijt} - \lambda_{0,mm'}B_{m'}(\sum_{s=1}^{p-1} X_{is} \prod_{l=1}^{s-1} \sin \theta_l \cos \theta_s ) \\
+ X_{ip} \prod_{l=1}^{p-1} \sin \theta_l) B_{m'}(\sum_{s=1}^{k-1} Z_{is} \prod_{l=1}^{s-1} \sin \alpha_l \cos \alpha_s + Z_{ik} \prod_{l=1}^{k-1} \sin \alpha_l )
\]

\[
+ \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{1,mm'} B_{m'}(\sum_{s=1}^{p-1} X_{is} \prod_{l=1}^{s-1} \sin \theta_l \cos \theta_s ) + X_{ip} \prod_{l=1}^{p-1} \sin \theta_l) B_{m'}(\sum_{s=1}^{k-1} Z_{is} \prod_{l=1}^{s-1} \sin \alpha_l \cos \alpha_s )
\]

\[
+ Z_{ik} \prod_{l=1}^{k-1} \sin \alpha_l \sum_{m=1}^{M} \sum_{i=1}^{K-1} \delta_i B_{m+1}(t)^2 - \sum_{m,n,j} (\lambda_{0,mm'}^{j} + (\lambda_{1,mm'}^{j})^2 ) \delta m^2, \eta^2
\]

where $C$ involves only the hyperparameters $a, M_2, M_1, K, d_1, d_2$ and the observations but not parameter of the model.

All the B-spline coefficients and $\sigma$ are updated using the conjugacy structure.

Posterior updates are described here. For notational convenience, we define $\lambda_{0,mm'}^{j} = \{\lambda_{0,mm'}^{j} : 1 \leq m, m' \leq K\}$, $\lambda_{1,mm'}^{j} = \{\lambda_{1,mm'}^{j} : 1 \leq m, m' \leq K\}$, $\delta = (\delta_1, \ldots, \delta_{K-1})$, $\theta = \{\theta_i : 1 \leq i \leq (p - 1)\}$ and $\alpha = \{\alpha_i : 1 \leq i \leq (k - 1)\}$,

\[
X_{ij}' = \sum_{s=1}^{p-1} X_{is} \prod_{l=1}^{s-1} \sin(\theta_l) \cos(\theta_s ) + X_{ip} \prod_{l=1}^{p-1} \sin(\theta_l )
\]

\[
Z_{ij}' = \sum_{s=1}^{k-1} Z_{is} \prod_{l=1}^{s-1} \sin(\alpha_l ) \cos(\alpha_s ) + Z_{ik} \prod_{l=1}^{k-1} \sin(\alpha_l )
\]

$\psi_i = \{B_m(X_{ij}' \beta)B_{m'}(Z_{ij}' \eta) : 1 \leq m, m' \leq K\}$. Below we use $\text{Id}_l$ to denote identity matrix of dimension $l$.

- **Updating $\sigma$:** Posterior density of $\sigma$ is inverse gamma with parameters $d_1, d_2$.
- **Updating $\lambda_{0,mm'}^{j}$:** The posterior variance ($V_{0,j}$) and posterior mean ($M_{0,j}$) are respectively
\[ V_{0,j} = 2(\sigma^{-2}\sum_i T_i \psi_i \psi_i' + a^{-2} \text{Id}_K)^{-1} \text{ and} \]
\[ M_{0,j} = V_{0,j}\left[ \sum_i \psi_i \sum_{t=0}^{T_i} (Y_{ijt} + \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{i,m,m'} B_m(X_i' \beta) B_{m'}(Z_i' \eta) \frac{1}{\sum_{i=1}^{K-1} \delta_i} \sum_{m=1}^{K-1} \sum_{i=1}^{m} \delta_i B_{m+1}(t)) \right]. \]

- **Updating \( \lambda_{ij} \):** The posterior variance \((V_{1,j})\) and posterior mean \((M_{1,j})\) are respectively
\[ V_{1,j} = 2(\sigma^{-2}\sum_i \psi_i \psi_i' (\sum_{t=1}^{T_i} \frac{1}{\sum_{k=1}^{K} \delta_k} \sum_{m=1}^{K} \sum_{i=1}^{m} \delta_i B_{m+1}(t))^2 + \text{Id}_K/a^2)^{-1} \text{ and } M_{1,j} = V_{1,j}\sum_i \psi_i \sum_{t=0}^{T_i} (Y_{ijt} - \sum_{m=1}^{K} \sum_{m'=1}^{K} \lambda_{0,m,m'} B_m(X_i' \beta) B_{m'}(Z_i' \eta)) \sum_{i=1}^{K-1} \delta_i \sum_{m=1}^{K-1} \sum_{i=1}^{m} \delta_i B_{m+1}(t) \].

- **Updating \( \theta_i \):** Sample by the Metropolis-Hastings algorithm
  - Generate \( U_{1i} \) from \( \text{Be}(B_1, B_1) \);
  - Update \( \theta_i \) to \( \theta_i^* = \begin{cases} \pi \theta_i U_{1i}^{i \in (\pi - \theta_i)(1 - U_{1i})} & \text{if } i < p - 1, \\ 2\pi \theta_i U_{1i}^{i \in (2\pi - \theta_i)(1 - U_{1i})} & \text{if } i = p - 1. \end{cases} \)

The conditional density of this update is given by
\[ f(\theta_i^*|\theta_i) = \frac{(\theta_i'^*(1 - \theta_i'^*)) B_{i-1}(1 - \theta_i'^*(1 - \theta_i'^*)) B_{i-1}}{B(B_1, B_1)} \theta_i'^*(1 - \theta_i'^*) \]
where \( \theta_i' \) is as in item (v) of the prior specification part.

The acceptance probability in Metropolis-Hastings algorithm is then given by
\[ P_{\alpha, \theta_i} = \min\left\{ \frac{L_1(i, \theta_i^*) f(\theta_i^*|\theta_i)}{L_1(i, \theta_i) f(\theta_i|\theta_i)}, 1 \right\}; \]
where \( L_1(i, x) \) denotes the likelihood for \( \theta_i = x \). We need to tune \( B_1 \) to achieve a good acceptance rate. This approach is due to Roy et al. (2017).

- **Updating \( \alpha_i \):** Sampling by the Metropolis-Hastings algorithm
  - Generate \( U_{2i} \) from \( \text{Be}(B_2, B_2) \)
  - Update \( \alpha_i \) to \( \alpha_i^* = \begin{cases} \pi \alpha_i U_{2i}^{i \in (\pi - \alpha_i)(1 - U_{2i})} & \text{if } i < k - 1, \\ 2\pi \alpha_i U_{2i}^{i \in (2\pi - \alpha_i)(1 - U_{2i})} & \text{if } i = k - 1. \end{cases} \)

The conditional density of this update is given by
\[ f_1(\alpha_i^*|\alpha_i) = \frac{(\zeta_i'^*(1 - \zeta_i'^*)) B_{i-1}(1 - \zeta_i'^*(1 - \zeta_i'^*)) B_{i-1}}{B(B_2, B_2)} \zeta_i'^*(1 - \zeta_i'^*) \]
\[ \frac{(\zeta_i + \zeta_i'^* - 2\zeta_i^2)}{(\zeta_i + \zeta_i^* - 2\zeta_i^2)^2} \]
where \( \zeta_i = \alpha_i / \pi \), \( \zeta^*_i = \alpha^*_i / \pi \) for \( i < k - 1 \) and \( \zeta_{k-1} = \alpha_{k-1} / 2\pi \), \( \zeta^*_{k-1} = \alpha^*_{k-1} / 2\pi \).

The acceptance probability in the Metropolis-Hastings algorithm is then given by

\[
P_{\alpha, \alpha_i} = \min \left\{ \frac{L_2(i, \alpha_i^*) f_1(\alpha_i | \alpha_i^*)}{L_2(i, \alpha_i f_1(\alpha_i^* | \alpha_i)), 1} \right\};
\]

here \( L_2(i, x) \) denotes the likelihood for \( \alpha_i = x \). We need to tune \( B_2 \) to achieve a good acceptance rate.

- **Updating \( \delta \):**
  - Generate from \( U_3 \) of length \( K - 1 \), each from \( \text{Be}(B_3, B_3) \);
  - Update \( \delta_j \) to \( \delta^*_j = \frac{\delta_j U_{3j} + (1 - \delta_j)(1 - U_{3j})}{\delta_j + U_{3j} - 2\delta_j U_{3j}} \).

  Elementary calculations give us the conditional density of this update as

\[
f(\delta^* | \delta) = \prod_{j=1}^{K-1} \frac{\delta_j (1-\delta^*_j)^{B_3-1} (1 - \frac{\delta_j (1-\delta^*_j)}{\delta_j + U_{3j} - 2\delta_j U_{3j}})^{B_3-1} \delta^*_j (1-\delta^*_j)^2}{\delta_j + U_{3j} - 2\delta_j U_{3j}}.
\]

  The acceptance probability in Metropolis-Hastings algorithm is then given by

\[
P_{\alpha, \delta_j} = \min \left\{ \frac{L_3(\delta^*) f(\delta | \delta^*)}{L_3(\delta) f(\delta^* | \delta)}, 1 \right\};
\]

  here \( L_3(x) \) denotes the likelihood for \( \delta = x \). We need to tune \( B_3 \) to achieve a good acceptance rate.

Due to the large size of the vector \( \theta \), one can consider updating in chunks. For ultra-high dimension, we updated \( \theta \) in chunks of size 30 coordinates. From each post burn-in sample we get the angles, that should be included in the model according to the latent indicator variable \( I \). The default value of each polar angle is considered to be \( \pi/2 \). This way if an angle is not in the model, in the rectangular system, the corresponding co-ordinate will be zero.

### 3.5 Simulation

We compare our model with the following simplified linear model:

\[
\log(Y_{ijt}) = X_i' \beta + Z_i \eta + \gamma_{1,j} - (X_i' \beta + Z_i \eta + \gamma_{2,j})^t + e_{ijt}. \tag{3.2}
\]

In the above model, \( \beta \) is a sparse vector and all other parameters are unpenalized. The performance of these two models are compared based on MSE values on test set. For sample sizes 200,
500 and 1000, we gather the MSE values corresponding to those models both for well-specified and misspecified cases. We use half of the sample for training and the remaining half for testing. Among other parameters we consider 13 regions in total, 5 time points and vary the value of $p$ as 5000, 10000 and 20000. We include one case for ultra high dimension of $p = 100000$ and sample size 200. We set $M_1 = 0.1$ and tune $M_2$ for different cases to ensure a good acceptance rate and desired model size (sum of $q_i$’s) across MCMC samples. The results are summarized below for 30 replications and 1000 post burn-in samples. The number of basis functions for spline is different across different sample sizes. For the ultra-high dimensional case with $p = 100000$ and $n = 200$, we consider 10 replications.

Data generation for the non-linear case:

- Generate a data matrix $X$ with elements coming from Bernoulli distribution with success probability of $i^{th}$ row as $p_i$.
- Generate $p_i$’s from the standard uniform distribution.
- Generate all the elements of the matrix $Z$ from $N(0, 1)$.
- Generate the sparse vector $\beta$ with 5% elements non-zero. Positions for non-zero elements are chosen first at random by sampling $p/20$ elements from total $p$ positions. Here $p$ is the length of $\beta$. The non zero elements are generated from mixture distribution of two normals $N(2, 1)$ and $N(-1, 1)$.
- Set the value of $\eta$ to $(1, -2, 4.3, 10, -8)$.
- Construct the intercept and slopes from functions through tensor products of B-splines with 11 basis functions in each direction, the coefficients for the intercept functions from $N(2, 1)$ and for slopes from $N(-1, 1)$.
- Normalize each row of $X$ and $Z$ along with $\beta$ are $\eta$ to the unit norm.

After generating the design matrix, the data $Y_{ijt}$ is generated from $N(a_{0,j}(X_i^t \beta, Z_i^t \eta) - a_{1,j}(X_i^t \beta, Z_i^t \eta) F_0(t), 1)$.

Data generation for the linear case:

In this case, the true model is the one given in (3.2). All the steps for generating $X$, $Z$, $\beta$ and $\eta$ are similar as above. The coefficients $\gamma_{1,j}$’s and $\gamma_{2,j}$’s are generated from $N(0, 1)$. After generating the design matrix, the data $Y_{ijt}$ is generated from $N(X_i^t \beta + Z_i^t \eta + \gamma_{1,j} - (X_i^t \beta + Z_i^t \eta + \gamma_{2,j}) t, 1)$.

For data generation, we set the error standard deviation $\sigma_0 = 1$. We split the whole data into two equal parts for training and testing. We compare the performance of our method with LASSO (Tibshirani (1996)) and SCAD (Fan and Li (2001)).
on testing dataset based estimates of parameters from training dataset. We use the R package \texttt{glmnet} for LASSO and \texttt{ncvreg} for SCAD. To fit our model we vary the number of B-spline basis function as 8 for 200 case, 11 for 500 case, and 14 for 1000 case. These values are considered based on the asymptotic rate of contraction of the number of B-spline basis functions.

Table 3.1: For non-linear case

<table>
<thead>
<tr>
<th>Total sample size</th>
<th>Dimension of $\beta$ $(p)$</th>
<th>SIM MSE</th>
<th>LASSO MSE</th>
<th>SCAD MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>5000</td>
<td>1.28</td>
<td>8.92</td>
<td>1.45</td>
</tr>
<tr>
<td>200</td>
<td>10000</td>
<td>1.43</td>
<td>8.86</td>
<td>1.50</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>1.45</td>
<td>9.03</td>
<td>1.61</td>
</tr>
<tr>
<td>500</td>
<td>5000</td>
<td>1.14</td>
<td>7.22</td>
<td>1.41</td>
</tr>
<tr>
<td>500</td>
<td>10000</td>
<td>1.19</td>
<td>7.82</td>
<td>1.46</td>
</tr>
<tr>
<td>500</td>
<td>20000</td>
<td>1.33</td>
<td>8.44</td>
<td>1.50</td>
</tr>
<tr>
<td>1000</td>
<td>5000</td>
<td>1.01</td>
<td>6.28</td>
<td>1.41</td>
</tr>
<tr>
<td>1000</td>
<td>10000</td>
<td>1.07</td>
<td>6.7</td>
<td>1.45</td>
</tr>
<tr>
<td>1000</td>
<td>20000</td>
<td>1.15</td>
<td>7.33</td>
<td>1.50</td>
</tr>
<tr>
<td>200</td>
<td>100000</td>
<td>1.14</td>
<td>7.32</td>
<td>1.43</td>
</tr>
</tbody>
</table>

Table 3.2: For linear case

<table>
<thead>
<tr>
<th>Total sample size</th>
<th>Dimension of $\beta$ $(p)$</th>
<th>SIM MSE</th>
<th>LASSO MSE</th>
<th>SCAD MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>5000</td>
<td>1.36</td>
<td>1.04</td>
<td>1.10</td>
</tr>
<tr>
<td>200</td>
<td>10000</td>
<td>1.50</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>200</td>
<td>20000</td>
<td>1.67</td>
<td>1.12</td>
<td>1.03</td>
</tr>
<tr>
<td>500</td>
<td>5000</td>
<td>1.31</td>
<td>1.03</td>
<td>1.07</td>
</tr>
<tr>
<td>500</td>
<td>10000</td>
<td>1.47</td>
<td>1.04</td>
<td>1.07</td>
</tr>
<tr>
<td>500</td>
<td>20000</td>
<td>1.61</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>1000</td>
<td>5000</td>
<td>1.26</td>
<td>1.02</td>
<td>1.08</td>
</tr>
<tr>
<td>1000</td>
<td>10000</td>
<td>1.37</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>1000</td>
<td>20000</td>
<td>1.61</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>200</td>
<td>100000</td>
<td>1.12</td>
<td>1.02</td>
<td>1.05</td>
</tr>
</tbody>
</table>

From the Table 3.1 and 3.2, we infer that the performance of our Bayesian method based on the high dimensional single index model is always better than the LASSO and the SCAD for the non-linear case. For linear case, it is competitive with linearity based methods like the
LASSO or the SCAD. This is natural as the LASSO or the SCAD use more precise modeling information which on semiparametric methods cannot compete on.

3.6 Large-sample Properties

We examine large sample properties of the proposed Bayesian procedure for the model in this section. We have observations for fixed $J$ number of regions and $T$ many time points. We show posterior consistency in the asymptotic regime of increasing sample size and increasing dimension ($p$) of SNPs. We need to establish some preliminary results before getting into the consistency results. The first result relates the Euclidean distance between two unit vectors with the corresponding polar angles.

**Lemma 3.1.** Let $\beta_1$ and $\beta_2$ are two unit vectors and $\theta_1$ and $\theta_2$ are their corresponding polar co-ordinates. Then $\|\beta_1 - \beta_2\|^2 \leq 4\|\theta_1 - \theta_2\|^2$

The proof is given in the Section 3.10.

We have the following set up in the model:

$$ Y_{ijt} = F(i, j, t) + \epsilon_{ijt}, \quad \epsilon_{ijt} \sim N(0, \sigma^2), $$

$$ F(i, j, t) = a_j^0(X_i^T \beta, Z_i^T \eta) - a_j^1(X_i^T \beta, Z_i^T \eta)F_0(t). $$

Here $i = 1, \ldots, N$, $j = 1, \ldots, J$ and $t = 1, \ldots, T_i$. Let $\Pi$ denote a probability measure on $\mathcal{P}$, as a subset of all probability measures. For $P, Q \in \mathcal{P}$, let

$$ K(p, q) = P \log \frac{P}{q}, \quad V(p, q) = P \log^2 \frac{P}{q}. $$

We also use Hellinger distance $d_H(P, Q) = \sqrt{\int (\sqrt{P} - \sqrt{Q})^2}.$

Let $P_{0i}$ denote the true distribution with density $p_{0i}$ for $i$-th individual. For each individual $i$, we consider the observation $Y_i$ of length $J \times T_i$ and $Y_i \sim \text{MVN}(\mu_i, \sigma^2 I_{J \times T_i})$; here $\mu_i = \{F(i, j, t) : j = 1, \ldots, J$ and $t = 1, \ldots, T_i\}$; here MVN stands for multivariate normal distribution.

For $p_i = \text{MVN}(\mu_1^i, \sigma_1^2 I_k)$ and $q_i = \text{MVN}(\mu_2^i, \sigma_2^2 I_k)$, we have

$$ K(p_i, q_i) = k \log \left( \frac{\sigma_2}{\sigma_1} \right) - \frac{1}{2} \left[ k - \frac{(\mu_1^i - \mu_2^i)'(\mu_1^i - \mu_2^i)}{\sigma_2^2} - k \frac{\sigma_1^2}{\sigma_2^2} \right], $$

$$ V(p_i, q_i) = \frac{k}{2} \left( \frac{\sigma_1^2}{\sigma_2^2} - 1 \right)^2 + \frac{k}{2} \frac{(\mu_1^i - \mu_2^i)'(\mu_1^i - \mu_2^i)}{\sigma_2^4} \sigma_1^2, $$

$$ d_H^2(P, Q) = 1 - \left( \frac{2\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} \right)^k \exp \left( - \frac{(\mu_1 - \mu_2)'(\mu_1 - \mu_2)}{4(\sigma_1^2 + \sigma_2^2)} \right). $$

49
Lemma 3.2. In the above setup we have $d^2_H(P,Q) \leq \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2} + (\mu_1-\mu_2)'(\mu_1-\mu_2)$ and for some $C > 2$ and $\mu_1, \mu_2$ close enough, $Cd^2_H(P,Q) \geq \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2} + \frac{(\mu_1-\mu_2)'(\mu_1-\mu_2)}{2(\sigma_1^2+\sigma_2^2)}$.

The proof is in the Section 3.10.

We denote $\Theta = (\beta, \eta, \sigma_0, a_1, F_0)$ as set of parameters. Let $\Theta_1 = (\beta_1, \eta_1, \sigma_0, a_{0,1}, a_{1,1}, F_{0,1})$ and $\Theta_2 = (\beta_2, \eta_2, \sigma_2, a_{0,2}, a_{1,2}, F_{0,2})$ be two such sets of parameters. We use $\|\cdot\|_\infty$ to denote the supremum norm.

Upper bound for the Hellinger distance

We can show for two functions approximated with $D$ many B-spline basis functions along each direction, $\|a_{0,2}^j - a_{0,2}^j\|_\infty^2 \leq \|\phi_{0,2}^j - \phi_{0,2}^j\|_\infty^2$, where $\phi_{0,1}$ and $\phi_{0,2}$ are the coefficients to the B-spline expansion. Similarly $\|a_{1,1}^j - a_{1,2}^j\|_\infty^2 \leq \|\phi_{1,1}^j - \phi_{1,2}^j\|_\infty^2$, where $\phi_{0,1}$ and $\phi_{0,2}$ are the corresponding B-spline coefficients. Lastly, $\|F_{0,1} - F_{0,2}\|_\infty^2 \leq \|\kappa_{0,1} - \kappa_{0,2}\|_\infty^2$ with $\kappa_{0,1}$ and $\kappa_{0,2}$ as corresponding B-spline coefficients. We can show that

$$\frac{1}{n} \sum_{i=1}^n d^2_H(p_{i1}, p_{i2}) \leq \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2} + 2 \sum_j (\|\phi_{0,1}^j - \phi_{0,2}^j\|_\infty + (\phi_{1,1}^j)^2) \left(\|\beta_0 - \beta_1\|_2^2 \lambda_1 + (\|\eta_0 - \eta_1\|_2^2 \lambda_2)ight)
+ 2 \|\kappa_{0,1} - \kappa_{0,2}\|_\infty^2 (\max_j (a_{1,1}^j)^2). \quad (3.3)$$

The proof of (3.3) is in the Section 3.10.

Lower bound for the Hellinger distance

By Lemma 2, we obtain that as $\sum_i d^2_H(p_{i1}, p_{i2}) \rightarrow 0, \sigma_1 \rightarrow \sigma_2$ and $\mu_{i1} \rightarrow \mu_{i2}$ for each $i$. As $F_{0,1}(0) = 0$ and $F_{0,2}(0) = 0$, we have $a_{0,1}^j(X_{1}^{i}\beta_1, Z_{1}^{i}\eta_1) \rightarrow a_{0,2}^j(X_{1}^{i}\beta_2, Z_{1}^{i}\eta_2)$. On the other hand $F_{0,1}(1) = 1$ and $F_{0,2}(1) = 1$ imply $a_{1,1}^j(X_{1}^{i}\beta_1, Z_{1}^{i}\eta_1) \rightarrow a_{1,2}^j(X_{1}^{i}\beta_2, Z_{1}^{i}\eta_2)$. This implies $F_{0,1} \rightarrow F_{0,2}$ pointwise at the given time points.

The following two inequalities can be used to get the lower bound in terms of difference of the functions:

$$(g + h)^2 + g^2 = 0.5(0.5((2g + h)^2 + h^2) + (g + h)^2 + g^2) \geq 0.25h^2 + 0.5g^2, \quad (3.4)$$

$$(g_1h_1 - g_2h_2)^2 + (g_1 - g_2)^2 = (g_1(h_1 - h_2) + h_2(g_1 - g_2))^2
+ h_2^2(g_1 - g_2)^2 + (1 - h_2^2)(g_1 - g_2)^2
\geq g_1^2(h_1 - h_2)^2 + (g_1 - g_2)^2. \quad (3.5)$$
On applying the two inequalities in (6.2) and (6.3) repeatedly, we get

\[
\frac{1}{n} \sum_{i,j,t} (F_{\Theta_1}(i,j,t) - F_{\Theta_2}(i,j,t))^2 \\
\geq \bar{T} C_1 \sum_{i,j} (a_{0,1}^j(X_i' \beta_1, Z_i' \eta_1) - a_{0,2}^j(X_i' \beta_2, Z_i' \eta_2))^2 \\
+ \bar{T} C_2 \sum_{i,j} (a_{1,1}^j(X_i' \beta_1, Z_i' \eta_1) - a_{1,2}^j(X_i' \beta_2, Z_i' \eta_2))^2 \\
+ \frac{1}{n} \sum_{i} T_i \sum_{t=1}^{T_i} C_3 \min_j (a_{1,1}^j)^2 (F_{0,1}(t) - F_{0,2}(t))^2.
\]

for some positive constants \(C_1, C_2\) and \(C_3\).

From the identifiability conditions on the single index model we conclude that as the functional values come close, the parameters and functions themselves become close as the number of observations goes to infinity. On the other hand \(F_{0,1}(\cdot)\) approaches \(F_{0,2}(\cdot)\) for all those points where there is an observation.

We make the further assumptions on the unknown functions:

- **Assumption 1.** The true functions are Lipschitz continuous.

- **Assumption 2.** There exists a large number \(M > 0\), such that for all \(j = 1, \ldots, J\) the functions \(a_{1,0}^j < M\).

- **Assumption 3.** There exists \(L_1 > 0, L_2 > 0\) such that \(|(a_{0,0}^j)'(\cdot)| < L_1\) and \(|(a_{1,0}^j)'(\cdot)| < L_2\), where \((a_{0,0}^j)'(\cdot)\) and \((a_{1,0}^j)'(\cdot)\) denote the first derivatives of \(a_{0,0}(\cdot)\) and \(a_{1,0}(\cdot)\) respectively.

Let \(\Theta_0 = (\beta_0, \eta_0, \sigma_0, a_{0,0}, a_{1,0}, F_{0,0})\) be the true set of parameters. For alternating \(\Theta_1\), we get

\[
\frac{1}{n} \sum_{i,j,t} (F_{\Theta_0}(i,j,t) - F_{\Theta_1}(i,j,t))^2 \\
\leq \bar{T} \left[ 2 \sum_j \|a_{0,0}^j - a_{0,1}^j\|^2 + 4J(L_1 + L_2)(\|\beta_0 - \beta_1\|^2 \lambda_1 + \|\eta_0 - \eta_1\|^2 \lambda_2) \\
+ 2JM^2\|F_{0,0} - F_{0,1}\|_\infty + 2 \sum_j \|a_{1,0}^j - a_{1,1}^j\|^2 \right].
\]

The proof of (3.6) is given in the Section 3.10.
Note the following relations

\[
\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} (\mu_0^i - \mu_1^i)(\mu_0^i - \mu_1^i) &= \frac{1}{n} \sum_{i,j,t} (F_{\Theta_0}(i,j,t) - F_{\Theta_1}(i,j,t))^2, \\
\frac{1}{n} \sum_{i=1}^{n} K(p_{i0}, p_{i1}) &= J_0 \log(\sigma_1^2/\sigma_0^2) - \frac{1}{2} J_0 - \frac{1}{n} \left( \sum_{i=1}^{n} (\mu_0^i - \mu_1^i)(\mu_0^i - \mu_1^i) \right) - J_0 \sigma_0^2/\sigma_1^2.
\end{align*}
\]

Except for \(\theta\), all other components are fixed dimensional. Positive prior probability for those cases can easily be verified. Let \(s\) be a binary array of length \(p - 1\) which has 1 at places where \(\theta_i\)s are not equal to the null values. Null values are \(\pi/2\) for initial \(p - 2\) angles and 0 or \(\pi/2\) for last angle. If \(\theta_{p-1}\) is not among those \(s\) angles, it may be zero as well. Let \(s_0\) denotes the true value of \(s\). We can represent \(\theta = \frac{\pi}{2}(1-s) + s \cdot \theta\) if all the null values of \(\theta\) are \(\pi/2\) where \(s \cdot \theta\) denotes element wise multiplication of \(s\) and \(\theta\). Thus we obtain using the fact that each polar angle is at most \(2\pi\),

\[
\|\theta_0 - \theta_1\|^2 \leq \left( \frac{\pi}{2} \right)^2 \|s_1 - s_0\|^2 + \|s_1 \cdot \theta_1 - s_0 \cdot \theta_0\|^2 \\
\leq \left( \frac{\pi}{2} \right)^2 \|s_1 - s_0\|^2 + (2\pi)^2 \|s_0 - s_1\|^2 + \|s_0 \cdot \theta_1 - s_0 \cdot \theta_0\|^2.
\]

(3.7)

Let for a given dimension \(D\), B-spline basis functions be \(\psi_{D,1}, \ldots, \psi_{D,D} : X \rightarrow [0,1]\) on a domain \(X\), a bounded convex set in \(\mathbb{R}^d\). Let us define the matrix \(\Phi\) of dimension \(2J \times D^2\) of which the first \(J\) rows are \(\{\phi_{0,0}^1, \ldots, \phi_{0,0}^d\}\) and the last \(J\) rows are \(\{\phi_{1,0}^1, \ldots, \phi_{1,0}^d\}\).

Priors on each row of \(\Phi\) and \(\kappa\) are multivariate normals and the supremum norm is used to induce a metric and \(\Pi(D = j) = b_1 \exp(-b_2 j (\log j)^{b_3})\) where \(0 \leq b_3 \leq 1\). Therefore,

**Lemma 3.3.** For large \(j\), we can show \(\exp(-b_2 j (\log j)^{b_3}) \asymp \Pi(D = j) \leq \Pi(D \geq j) \asymp \exp(-b_2 j (\log j)^{b_3})\)

**Proof.** The first relation trivially follows from the definition of \(\Pi(D = j)\). For the other part we can show \(\Pi(D = j+1) = g(j)\Pi(D = j)\), where \(g(j) = \exp(-b_2 (j+1) (\log(j+1))^{b_3} + b_2 j (\log j)^{b_3})\).

Now for large enough \(d\) we have \(g(m+1) < g(m) < 1\) for all \(m > j\) and \(g(\infty) = 0\),

\[
\Pi(D \geq j) = \sum_{l=j}^{\infty} \Pi(D = l) \leq (1 - g(j))^{-1} \Pi(D = j).
\]

For large \(j\), thus we have \(\Pi(D \geq j) \asymp \exp(-b_2 j (\log j)^{b_3})\).

For each row of \(\Phi\) and \(\kappa\), following assertions hold as we have multivariate normal prior on these vectors.
Lemma 3.4. For a given vector of functions \( L_0 \) that for the sieves of the form \( \mathcal{W} \) the above forms.

Proof. We have total 2 functions in the model. Let \( L_0 \) be the list of true functions i.e. \( L_0 = \{a_j^0, a_j^1, F_0 : j = 1, \ldots J\} \). We need to modify the Lemma 10.20 from page 290 of Ghosal and van der Vaart (2017) for this case. Due to Lemma 3.3, and (A1), (A2) and (A3) the required support conditions on prior for Lemma 10.20 hold.

Let \( d \) be the metric on the above vector of functions such that \( d(L_0, (\Phi D^2, \kappa D^2)) = \max\{\|a_j^0 - \Phi D^2\|, \|a_j^1 - \Phi D^2\| \} \). To show posterior contraction, we construct a sieve for the parameter space. We consider the sieves of the form \( \mathcal{W} \mathcal{D}_{n, \kappa} = \{\Phi D^2, \kappa D^2 : \kappa \leq D_n, \|\kappa\| \leq M_n \} \).

Lemma 3.4. For a given vector of functions \( L_0 \), there exists a positive integer \( D_n \) and \( \kappa D_n \) for all \( i = 1, \ldots, J \) and \( \kappa \) such that for \( \epsilon_n/\tilde{D}_n \) sufficiently small and \( \epsilon_n \leq D_n^{3/2} M_n \), we have,

\[
\Pi((\Phi D^2, \kappa D^2) : d(L_0, (\Phi D^2, \kappa D^2)) \leq \epsilon_n) \geq A(D_n) \left( \frac{\epsilon_n}{\tilde{D}_n} \right)^{r_1 D_n^2} \left( \frac{\epsilon_n}{\sqrt{D_n}} \right)^{r_2 D_n^2},
\]

\[
\log N(\epsilon_n, \mathcal{W} \mathcal{D}_{n, \kappa}, d) \leq J D_n^2 \log(3D_n^2 M_n/\epsilon_n) + D_n \log(3D_n^{3/2} M_n/\epsilon_n),
\]

where \( A(j) = \exp(-b_2 j (\log j)^{b_3}) \).

Proof. The proof of the above Lemma is similar to that of Lemma 10.20 in Ghosal and van der Vaart (2017). Only for the first assertion we need to use the fact that our metric is a supremum norm over an array of functions. Using the results in (A1), (A2), (A3) and Lemma 3.3, we get the above forms.

Using Lemma 3.4, we can now modify Theorem 10.21 from page 291 of Ghosal and van der Vaart (2017) to get the contraction rates.
Theorem 3.5. Let $\epsilon_n \geq \bar{\epsilon}_n$ be sequences of positive numbers with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, and let $D_n, \bar{D}_n \geq 3, M_n > 0$ be such that, for some sequence $b_n \to \infty$,

$$\inf_{\Phi, \kappa} d(L_0, (\Phi', \psi_{D_n^2}, \kappa', \psi_{\bar{D}_n})) \leq \bar{\epsilon}_n, \quad (3.8)$$

$$\bar{D}_n \log \bar{D}_n \geq b_n \log \bar{D}_n - \epsilon_n \leq n\epsilon_n^2, \quad (3.9)$$

$$D_n^2 \log \left(\frac{D_n M_n}{n \epsilon_n}\right) \leq n\epsilon_n^2, \quad (3.10)$$

$$D_n \log D_n - b_n \epsilon_n^2, \quad (3.11)$$

$$\log D_n + b_n \epsilon_n^2 \leq M_n^2. \quad (3.12)$$

Let the prior on $\kappa$ satisfy, for some $H_n$,

$$\Pi(\kappa : \|\kappa - \kappa_0\| < \bar{\epsilon}_n) \geq e^{-n\epsilon_n^2}, \quad (3.13)$$

$$\log_+ \text{diam}(H_n) \leq n\epsilon_n^2, \quad (3.14)$$

$$\Pi(\kappa \not\in H_n) \leq e^{-b_n \epsilon_n^2}. \quad (3.15)$$

Assume that the root averaged squared Hellinger distance $(d_n)$ of $\Theta_1, \Theta_2 \in W_{D_n, M_n} \times H_n$ and for average Kullback-Leibler divergence of $\Theta_0, \Theta_1$ satisfy the conditions, for some $a, \epsilon > 0$

$$n^{-\epsilon} d_n(\Theta_1, \Theta_2) \lesssim d(\Theta_1, \Theta_2)^a + \|\kappa_1 - \kappa_2\|^a, \quad (3.16)$$

and for $\Theta_0$ and $\Theta_1$ sufficiently close,

$$\frac{1}{n} \sum_{i=1}^n K(p_{i|0}, p_{i|1}) \lesssim d^2(\Theta_0, \Theta_1) + \|\kappa_1 - \kappa_2\|^2. \quad (3.17)$$

Then $P_{\Theta_0}^n \Pi_n(d_n(\Theta, \Theta_0) > K_n \epsilon_n) \to 0$, for every $K_n \to \infty$, where $P_{\Theta_0}^n$ is the joint distribution of the vector of observations.

Proof of the above theorem can follow that of Theorem 10.21 in Ghosal and van der Vaart (2017). This Theorem extends the result of previous theorem of Ghosal and van der Vaart (2017) for an array of functions.

We just need to verify the above conditions and get the contraction rate.
For $\Theta_1, \Theta_2 \in W_{D_n, M_n} \times \mathcal{H}_n$, using the result in (3.3) we get

$$d_n = \frac{1}{n} \sum_{i=1}^{n} d^2_H(p_{i1}, p_{i2})$$

$$\leq \frac{1}{2} n^{-2/5}(\sigma_1 - \sigma_2)^2 + 2 \sum_j [\| \phi_{0,1}^j - \phi_{0,2}^j \|_\infty^2 + \| \phi_{1,1}^j - \phi_{1,2}^j \|_\infty^2]$$

$$+ 8M_n^2 D_n^3(\| \theta_1 - \theta_2 \|_2^2 \lambda_1 + \| \alpha_1 - \alpha_2 \|_2^2 \lambda_2) + 2D_n M_n^2 \| \kappa_0 - \kappa_0 \|_\infty^2$$

$$\leq \frac{1}{2} n^{-2/5}(\sigma_1 - \sigma_2)^2 + 2 \sum_j [\| \phi_{0,1}^j - \phi_{0,2}^j \|_\infty^2 + \| \phi_{1,1}^j - \phi_{1,2}^j \|_\infty^2]$$

$$+ 8M_n^2 D_n^3 \left( \left( \frac{17}{4} \pi^2 \| s_1 - s_2 \|_2^2 + \| s_1 \cdot \theta_1 - s_1 \cdot \theta_2 \|_2^2 \right) \lambda_1 + \| \alpha_1 - \alpha_2 \|_2^2 \lambda_2 \right)$$

$$+ 2D_n M_n^2 \| \kappa_0 - \kappa_0 \|_\infty^2.$$  (3.18)

For $X \sim \text{Be}(M_1, M_2)$ with $M_1 < 1 \leq M_2$, using the transformation $y = x/(1 - \epsilon)$

$$P(X > \epsilon) = \frac{1}{B(M_1, M_2)} \int_{\epsilon}^{1} x^{M_1 - 1} (1 - x)^{M_2 - 1} dx$$

$$= \frac{1}{B(M_1, M_2)} \int_{0}^{1 - \epsilon} (1 - x)^{M_1 - 1} x^{M_2 - 1} dx$$

$$= \frac{(1 - \epsilon)^{M_2}}{B(M_1, M_2)} \int_{0}^{1} (1 - (1 - \epsilon)y)^{M_1 - 1} y^{M_2 - 1} dy,$$

$$\leq (1 - \epsilon)^{M_2}. \quad \text{(3.19)}$$

The last inequality is due to $(1 - (1 - \epsilon)y)^{M_1 - 1} \leq (1 - y)^{M_1 - 1}$ as $M_1 < 1$.

We have

$$P(s_{1i} = 1) = \frac{\gamma}{\gamma + (1 - \gamma) \frac{\Gamma(M_1 + M_2)}{\Gamma(M_1) \Gamma(M_2)} \Gamma(\theta'_i) M_1 - 1 (1 - \theta'_i) M_2 - 1},$$

here we define $\theta'_i = 1 - \frac{\min(\theta_i - \pi - \theta_i)}{\pi/2}$ for $1 \leq i \leq (p - 2)$ and $\theta''_p-1 = \frac{\theta_p-1}{\pi/2} - 1(\theta_p-1 > \pi) - 1(\theta_p-1 > 3\pi/2)$. Then we would reduce it further to $\theta''_p-1 = \frac{\pi}{3} \min(\theta''_p, \pi/2 - \theta''_p-1)$. This reparametrization is only required because the beta density is defined in $[0, 1]$.

We have $i$th polar angle in the model if and only if for an decreasing function $\Psi$,

$$P(s_{1i} = 1) > 0.5$$

$$\iff \frac{\gamma}{1 - \gamma} > \frac{\Gamma(M_1 + M_2)}{\Gamma(M_1) \Gamma(M_2)} \Gamma(\theta'_i) M_1 - 1 (1 - \theta'_i) M_2 - 1,$$

$$\iff \theta'_i > \Psi(\frac{\gamma}{1 - \gamma}).$$
The last inequality is deduced from the fact that $M_1 < 1 \leq M_2$ and thus the beta density of $\theta^i_1$ is decreasing in $[0, 1]$. Any upper bound on density will correspond to lower bound on the parameter $\theta^i$. As $M_2$ increases in (3.19), $P(\theta^i > \epsilon)$ decreases to zero for each $\epsilon$ and thus chances of $P(s_{1i} = 1) > 0.5$' goes to zero for most of the $i$'s for fixed $\gamma$. The prior would make most of the entries in $s_1$ and $s_2$ zero. To achieve $\|s_1 - s_2\|^2$ small, $s_1$ necessarily needs to coincide with $s_2$. We note that $s_1 \cdot \theta_1 - s_1 \cdot \theta_2$ is in a low dimensional space due to the element wise multiplication of $\theta_1$ and $\theta_2$ with $s_1$.

We need to vary $M_2$ with $n$. Thus we consider $M_{2n}$. From (3.19), we obtain $(1 - \epsilon_n)^{M_{2n}} = o(1/p)$ which implies $M_{2n} > \log_p/\epsilon_n$.

Using the fact that $\log(1 + x) \leq |x|$, we obtain from (3.6) for $\Theta_0$ and $\Theta_1$ sufficiently close,

$$
\frac{1}{n} \sum_{i=1}^{n} K(p_{i0}, p_{i1})
\leq J\bar{T}\left|\frac{\sigma_2^2 - \sigma_0^2}{\sigma_0^2}\right| + \bar{T} \left[2 \sum_{j} (\|\phi_{0,0}^{j} - \phi_{0,1}^{j}\|_{\infty}^2 + \|\phi_{1,0}^{j} - \phi_{1,1}^{j}\|_{\infty}^2) + 8J(L_1 + L_2)(\|\theta_0 - \theta_1\|_{2}^2 \lambda_1 + \|\alpha_0 - \alpha_1\|_{2}^2 \lambda_2) + 2JM^2 \|\kappa_{0,0} - \kappa_{0,1}\|_{\infty}^2]\right]
$$

$$
\leq J\bar{T}\left|\frac{\sigma_3^2 - \sigma_0^2}{\sigma_0^2}\right| + \bar{T} \left[2 \sum_{j} (\|\phi_{0,0}^{j} - \phi_{0,1}^{j}\|_{\infty}^2 + \|\phi_{1,0}^{j} - \phi_{1,1}^{j}\|_{\infty}^2) + 8J(L_1 + L_2)(\left\|\frac{17}{4}\pi^2 s_1 - s_0\right\|_{2}^2 + \|s_0 \cdot \theta_1 - s_0 \cdot \theta_0\|_{2}^2 \lambda_1 + \|\alpha_0 - \alpha_1\|_{2}^2 \lambda_2) + 2JM^2 \|\kappa_{0,0} - \kappa_{0,1}\|_{\infty}^2]\right].
$$

Thus (3.16) and (3.17) from Theorem 3.5 are verified. Let $\iota$ be the regularity level of the true function in each direction. Then for B-spline approximation of the true functions $w_{10}$ (univariate) and $w_{20}$ (bivariate) with $D$ number of basis functions in each direction we have $\inf_{\zeta} \|w_{10} - \zeta \psi_D\| \lesssim D^{-\iota}$ for univariate functions and $\inf_{\zeta} \|w_{20} - \zeta \psi_D\| \lesssim D^{-\iota}$ for bivariate functions. The conditions (3.9)–(3.12) change accordingly such that right hand side of the inequalities in Lemma 3.4 becomes of the order $n\epsilon_n^2$.

**Theorem 3.6.** Under the assumed conditions on prior and true functions, the posterior contraction rate of the proposed single index model is

$$
\max \left\{ n^{-\iota/(\iota+1)} (\log n)^{1/(\iota+1)} b_3 - b_3/2 + 1, (n/\log b_3/2 n)^{-1/(\iota+1)} \right\},
$$

where $b_3 = b_3 \vee 1$.

**Proof.** In our case $\inf_{\Phi, \kappa} d(\mathcal{L}_0, (\Phi' \psi_{D_n}^2, \kappa' \psi_{D_n}^2)) \lesssim D_n^{-\iota}$. Combining with (3.10), we obtain that
\[ D^{-i} \leq \varepsilon_n \text{ and } D_n^2 \log n \leq n \bar{\epsilon}_n^2. \] Thus to satisfy (3.8)–(3.12) we choose
\[
\bar{D}_n \asymp \left( n / \log \tilde{b}_3 / 2 \right)^{1/(i+1)},
\]
\[
\bar{\varepsilon}_n \asymp \left( n / \log \tilde{b}_3 / 2 \right)^{-i/(i+1)},
\]
\[
\varepsilon_n \gg n^{-i/(i+1)} \left( \log n \right)^{i/(i+1)} \tilde{b}_3 - b_3 / 2 + 1,
\]
so that
\[
D_n \gg n^{1/(i+1)} \left( \log n \right)^{i/(i+1)} \tilde{b}_3 - b_3.
\]
Here the nuisance parameter set is \( \kappa \). Conditions (3.13)–(3.15) for the prior on \( \kappa = \sigma, \alpha, s, \theta \) of the Theorem 3.5 will automatically hold for above rates. We also need from (3.19) that \( M_2n > \log p / \left[ \left( n / \log \tilde{b}_3 / 2 \right)^{-i/(i+1)} \right] \).

### 3.7 Real data analysis

#### 3.7.1 Modification of the model for real data application

##### 3.7.1.1 Incorporating random effect and region wise varying effect

As the data is a longitudinal time series, it is reasonable to add a subject specific random effect \( \tau_i \) in the model and vary the effect of the low dimensional covariates region-wise. The new modified model will then become
\[
Y_{ijt} = F(i, j, t) + \tau_i + \epsilon_{ijt}, \epsilon_{ijt} \sim N(0, \sigma^2),
\]
\[
F(i, j, t) = a_0^j(X_i^j \beta, Z_i^j \eta_j) - a_1^j(X_i^j \beta, Z_i^j \eta_j) F_0(t).
\] (3.21)

We can take successive difference within a pair of individual, region over time and another successive differences within each individual over regions. These will not involve the random effect. Consistency of the posterior distribution given this reduced data will follow from the results we showed in Section 2.6. By plugging-in those estimates form the reduced model into the original model, we can consistently estimate random effects.

**Prior on random effect**

We put a Dirichlet process scale mixture of normal prior on the random effect distribution.
3.7.1.2 Region wise varying effect with no SNP

To compare the results with Hostage et al. (2014), we also fit the following model without the SNPs,

\[ Y_{ijt} = F(i, j, t) + \tau_i + \epsilon_{ijt} \sim N(0, \sigma^2), \]
\[ F(i, j, t) = a_j^0(Z_i\eta_j) - a_j^1(Z_i^\prime\eta_j)F_0(t). \]  

(3.22)

3.7.1.3 Corresponding Linear Model

We compare the performance of our above non-linear models with following linear model,

\[ Y_{ijt} = H(i, j, t) + \tau_i + \epsilon_{ijt} \sim N(0, \sigma^2), \]
\[ H(i, j, t) = \varphi_j^0 + \varphi_j^0Z_i, M + \varphi_j^0Z_i, AD + \varphi_j^0NCZ_i, NC + \varphi_j, Allele4Z_i, Allele4 + \varphi_j, Allele2Z_i, Allele2 + \varphi_j, AgeZ_i, Age + \varphi_j, AD, Allele2Z_i, AD, Allele2 + \varphi_j, AD, Allele4Z_i, AD, Allele4 + \varphi_j, NC, Allele2Z_i, NC, Allele2 + \varphi_j, NC, Allele4Z_i, NC, Allele4 - [\varphi_j^1 + \varphi_j^1Z_i, M + \varphi_j^1Z_i, AD + \varphi_j^1NCZ_i, NC + \varphi_j, Allele4Z_i, Allele4 + \varphi_j, Allele2Z_i, Allele2 + \varphi_j, AgeZ_i, Age + \varphi_j, AD, Allele2Z_i, AD, Allele2 + \varphi_j, AD, Allele4Z_i, AD, Allele4 + \varphi_j, NC, Allele2Z_i, NC, Allele2 + \varphi_j, NC, Allele4Z_i, NC, Allele4]t. \]  

(3.23)

We have data for total thirteen brain regions along with the whole brain over time for 748 individuals. For each individual, we have covariate information which is summarized in Table 3.3. The standard deviations for age in each group are mentioned in the bracket. The baseline subject for our analysis is a female individual with average age and no cognitive impairment.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>No cognitive impairment (NC)</td>
<td>99</td>
<td>114</td>
</tr>
<tr>
<td>Alzheimer disease (AD)</td>
<td>84</td>
<td>94</td>
</tr>
<tr>
<td>Mild cognitive impairment (MCI)</td>
<td>122</td>
<td>235</td>
</tr>
<tr>
<td>APOEgene allele2</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>APOEgene allele4</td>
<td>150</td>
<td>223</td>
</tr>
<tr>
<td>Average age</td>
<td>73.51</td>
<td>74.60</td>
</tr>
<tr>
<td></td>
<td>(6.67)</td>
<td>(6.80)</td>
</tr>
</tbody>
</table>
We first fit the model in (3.22) and the following linear model in (3.23) in accordance with Hostage et al. (2014) with same set of covariates and interactions between APOE and disease states. Then we compare the prediction MSE. Prediction error gives us the predictive performance and fitted relative MSE helps to judge the reliability of inference. We consider 17 basis B-spline functions for univariate and 17\(^2\) basis functions for bivariate cases.

Linear model gives the prediction error 3.83 whereas that in our non-linear model improves to 0.8 which is around 80% improvement. The model in (3.21) with SNPs better the prediction error to 0.6 which is around 25% improvement.

For prediction error, we divide the whole dataset into training (Tr) and testing (Te). We use stratified type sampling using each subject-region pair as strata so that training will have all the individuals that belong to the testing set. This is important for prediction with a random effect in the model. The formula for prediction error will be \( \frac{1}{|Te|} \sum_{(i,t) \in Te} (Y_{ijt} - \hat{Y}_{ijt})^2 \); here \(|Te|\) denotes total number of elements in test set, Te.

To obtain the marginal effect of the covariates on atrophy we regress the estimated slope function on the covariates. Regression coefficient corresponding to a covariate is analogous to average partial derivative of the function with respect to the covariate. For binary covariate, the partial derivative with respect to \( Z_k \) is to be replaced by a difference to calculate an effect for region \( j \):

\[
\left\{ (a_1^j (X_i^\beta, Z_i^\eta_j))|Z_{ik}=1, Z_{im}=0, \forall m \neq k - a_1^j (X_i^\beta, Z_i^\eta_j)|Z_{im}=0, \forall m \right\} \frac{dF_0(t)}{dt}.
\]

Above expression is derived from the partial derivative of our model in (3.22) with respect to time. Similarly for interaction between \( Z_k \) and \( Z_l \), the expression will become

\[
\left\{ (a_1^j (X_i^\beta, Z_i^\eta_j))|Z_{ik}=1, Z_{il}=1, Z_{im}=0, \forall m \neq k - a_1^j (X_i^\beta, Z_i^\eta_j)|Z_{ik}=1, Z_{im}=0, \forall m \right. \\
\left. - a_1^j (X_i^\beta, Z_i^\eta_j)|Z_{il}=1, Z_{im}=0, \forall m + a_1^j (X_i^\beta, Z_i^\eta_j)|Z_{im}=0, \forall m \right\} \frac{dF_0(t)}{dt}.
\]

For continuous covariate, we regress the partial derivative on the covariate to get the marginal main effect. Estimated parameters related to atrophy are given below in Table 3.4 to Table 3.17 for the non-linear model for different regions, and in Table 3.18 for the linear model corresponding to the whole brain. Here positive values mean atrophy.

We map the significant SNPs from our analysis to corresponding gene ids using the R packages biomaRt and rsnps. We tune the parameters in the model to get twenty most significant SNPs. We find the corresponding gene for eleven of those. Their average effects on atrophy are also estimated and it is recorded in Table 3.19 along with p-values.

### 3.8 Conclusions and discussion

We fit a bivariate single index model to capture the volumetric change of different cortical regions in the human brain. There are both high and low-dimensional covariates as input to the
Table 3.4: Estimates of covariates for Total Brain for slope from non-linear model

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0085</td>
<td>0.0006</td>
<td>13.2069</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0080</td>
<td>0.0009</td>
<td>8.5275</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0008</td>
<td>0.0013</td>
<td>0.6491</td>
<td>0.5163</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0012</td>
<td>0.0005</td>
<td>2.2546</td>
<td>0.0242</td>
</tr>
<tr>
<td>MCI:time</td>
<td>−0.0045</td>
<td>0.0008</td>
<td>−5.8255</td>
<td>0.0000</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0081</td>
<td>0.0011</td>
<td>7.3063</td>
<td>0.0000</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0004</td>
<td>0.0003</td>
<td>1.4899</td>
<td>0.1364</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0043</td>
<td>0.0011</td>
<td>3.9985</td>
<td>0.0001</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0072</td>
<td>0.0012</td>
<td>5.7506</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>0.0025</td>
<td>0.0020</td>
<td>1.2259</td>
<td>0.2204</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0001</td>
<td>0.0032</td>
<td>−0.0375</td>
<td>0.9701</td>
</tr>
</tbody>
</table>

Table 3.5: Estimates for Ventricles for slope

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>−0.0531</td>
<td>0.0043</td>
<td>−12.4436</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0346</td>
<td>0.0062</td>
<td>5.5875</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0646</td>
<td>0.0083</td>
<td>7.7687</td>
<td>0.0000</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0108</td>
<td>0.0035</td>
<td>3.1423</td>
<td>0.0017</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0133</td>
<td>0.0051</td>
<td>2.5784</td>
<td>0.0100</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0380</td>
<td>0.0074</td>
<td>5.1490</td>
<td>0.0000</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0004</td>
<td>0.0017</td>
<td>0.2557</td>
<td>0.7982</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>−0.0199</td>
<td>0.0070</td>
<td>−2.8182</td>
<td>0.0049</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>−0.0292</td>
<td>0.0083</td>
<td>−3.5336</td>
<td>0.0004</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0285</td>
<td>0.0133</td>
<td>−2.1435</td>
<td>0.0322</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0648</td>
<td>0.0210</td>
<td>−3.0828</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

unknown functions determining initial configuration and rate of change of different regions. To tackle the high dimensional covariate within a single index model, we provide a new technique to assign sparse prior in this chapter. Posterior consistency results are also established. An ‘R’ package is developed based on this chapter which can be found in https://github.com/royarkaprava/SIMBayes. This can be used to fit high and low dimensional single index models. This package can fit a single index model with only one input or two inputs with at most one in shrinkage consideration. The function betaupdate of this package is used to generate MCMC samples for polar angles of $\beta$ and $\eta$ of our model.

In our results on the real dataset, we find allele 4 of APOE gene, Alzheimer disease state, and their interaction as significant covariates for almost all the cases. The fact that allele 4 of APOE gene is significant was established in Hostage et al. (2014). They used a linear model, similar
Table 3.6: Estimates for Left Hippocampus for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0053</td>
<td>0.0011</td>
<td>4.8678</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0094</td>
<td>0.0016</td>
<td>5.9956</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0031</td>
<td>0.0021</td>
<td>1.4513</td>
</tr>
<tr>
<td>Gender:time</td>
<td>−0.0023</td>
<td>0.0009</td>
<td>−2.6260</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0073</td>
<td>0.0013</td>
<td>5.6415</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0038</td>
<td>0.0019</td>
<td>2.0409</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0003</td>
<td>0.0004</td>
<td>0.6419</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0003</td>
<td>0.0018</td>
<td>0.1548</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0076</td>
<td>0.0021</td>
<td>3.6578</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0099</td>
<td>0.0034</td>
<td>−2.9336</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0017</td>
<td>0.0053</td>
<td>−0.3210</td>
</tr>
</tbody>
</table>

Table 3.7: Estimates for Right Hippocampus for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0044</td>
<td>0.0014</td>
<td>3.0999</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0068</td>
<td>0.0021</td>
<td>3.3176</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>−0.0005</td>
<td>0.0028</td>
<td>−0.1961</td>
</tr>
<tr>
<td>Gender:time</td>
<td>−0.0021</td>
<td>0.0011</td>
<td>−1.8073</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0121</td>
<td>0.0017</td>
<td>7.0573</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0039</td>
<td>0.0025</td>
<td>1.5869</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0006</td>
<td>0.0006</td>
<td>1.0891</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0050</td>
<td>0.0023</td>
<td>2.1432</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0062</td>
<td>0.0027</td>
<td>2.2402</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0190</td>
<td>0.0044</td>
<td>−4.3067</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0014</td>
<td>0.0070</td>
<td>−0.2000</td>
</tr>
</tbody>
</table>

to the model in (3.23). But in our estimates of parameters for the linear case in Table 3.18, APOE allele 4 is not significant. Such is the case for several other regions in the linear case. But in our non-linear setup, it is significant for the whole brain as well as for all other important regions as given in the region-wise tables in Supplementary materials. For some cases, few of other covariates are also significant like mild cognitive impairment (MCI) disease state, and gender. We get some other significant genes which are mentioned in Table 3.19. We observe that main effect of APOE allele 4 is more than each of the main effects of these other genes. Except for MND1 and LRBA, rest of the significant genes were found to be associated with the Alzheimer Disease before. These genes are mentioned in following two websites which contains all the known genes associated with AD, http://amp.pharm.mssm.edu/Harmonizome/gene_set/alzheimer+disease/GWASdb+SNP-Phenotype+Associations and http://amp.pharm
<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>−0.0628</td>
<td>0.0050</td>
<td>−12.4490</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0458</td>
<td>0.0073</td>
<td>6.2527</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0663</td>
<td>0.0098</td>
<td>6.7491</td>
<td>0.0000</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0064:time</td>
<td>0.0041</td>
<td>1.5800</td>
<td>0.1142</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0070</td>
<td>0.0061</td>
<td>1.1562</td>
<td>0.2477</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0472</td>
<td>0.0087</td>
<td>5.4115</td>
<td>0.0000</td>
</tr>
<tr>
<td>Age:time</td>
<td>−0.0011</td>
<td>0.0021</td>
<td>−0.5146</td>
<td>0.6068</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>−0.0263</td>
<td>0.0083</td>
<td>−3.1526</td>
<td>0.0016</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>−0.0406</td>
<td>0.0098</td>
<td>−4.1578</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0329</td>
<td>0.0157</td>
<td>−2.0900</td>
<td>0.0367</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0716</td>
<td>0.0249</td>
<td>−2.8801</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

Thus these gene have shown significant effect of brain activity. Now, we get these genes significant in our atrophy study.

### 3.9 Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer Association; Alzheimer Drug Dis-
Table 3.10: Estimates for Left Medial Temporal for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0042</td>
<td>0.0009</td>
<td>4.5576</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0049</td>
<td>0.0013</td>
<td>3.6554</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0032</td>
<td>0.0018</td>
<td>1.7819</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0012</td>
<td>0.0007</td>
<td>1.5881</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0061</td>
<td>0.0011</td>
<td>5.4561</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0044</td>
<td>0.0016</td>
<td>2.7614</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.1820</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0000</td>
<td>0.0015</td>
<td>0.0070</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0048</td>
<td>0.0018</td>
<td>2.6679</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0021</td>
<td>0.0029</td>
<td>−0.7264</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0004</td>
<td>0.0045</td>
<td>−0.0802</td>
</tr>
</tbody>
</table>

Table 3.11: Estimates for Right Medial Temporal for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0083</td>
<td>0.0011</td>
<td>7.8945</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0055</td>
<td>0.0015</td>
<td>3.6024</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0010</td>
<td>0.0021</td>
<td>0.4745</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0009</td>
<td>0.0009</td>
<td>1.1006</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0022</td>
<td>0.0013</td>
<td>1.7644</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0076</td>
<td>0.0018</td>
<td>4.1755</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.3162</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0033</td>
<td>0.0017</td>
<td>1.9000</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0059</td>
<td>0.0020</td>
<td>2.9146</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0019</td>
<td>0.0033</td>
<td>−0.5662</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>0.0001</td>
<td>0.0052</td>
<td>0.0202</td>
</tr>
</tbody>
</table>

dcovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the
study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The first author would like to thank Mr. Kushal Kumar Dey for helping him how to build an ‘R’ package.

### 3.10 Proofs

**Proof of Lemma 3.1**

**Proof.** Let $J_{\beta,\theta}$ be the Jacobian matrix for transformation from $\beta$ to $\theta$. Then from multivariate mean value theorem, we have $\|\beta_1 - \beta_2\|_2 \leq \sup_{\beta,\theta} \|J_{\beta,\theta}\|_{sp} \|\theta_1 - \theta_2\|_2$, where $\|\cdot\|_{sp}$ denotes
Table 3.14: Estimates for Left Fusiform for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p–value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0125</td>
<td>0.0012</td>
<td>10.8219</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0099</td>
<td>0.0017</td>
<td>5.8934</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>−0.0037</td>
<td>0.0022</td>
<td>−1.6363</td>
<td>0.1019</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0025</td>
<td>0.0009</td>
<td>2.7219</td>
<td>0.0065</td>
</tr>
<tr>
<td>MCI:time</td>
<td>−0.0006</td>
<td>0.0014</td>
<td>−0.4046</td>
<td>0.6858</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0102</td>
<td>0.0020</td>
<td>5.1188</td>
<td>0.0000</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0001</td>
<td>0.0005</td>
<td>0.1823</td>
<td>0.8554</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0045</td>
<td>0.0019</td>
<td>2.3821</td>
<td>0.0173</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0086</td>
<td>0.0022</td>
<td>3.8403</td>
<td>0.0001</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>0.0049</td>
<td>0.0036</td>
<td>1.3715</td>
<td>0.1703</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>0.0053</td>
<td>0.0057</td>
<td>0.9349</td>
<td>0.3499</td>
</tr>
</tbody>
</table>

Table 3.15: Estimates for Right Fusiform for slope

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p–value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0070</td>
<td>0.0009</td>
<td>7.9958</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0040</td>
<td>0.0013</td>
<td>3.1873</td>
<td>0.0015</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0002</td>
<td>0.0017</td>
<td>0.1286</td>
<td>0.8977</td>
</tr>
<tr>
<td>Gender:time</td>
<td>−0.0003</td>
<td>0.0007</td>
<td>−0.3875</td>
<td>0.6984</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0028</td>
<td>0.0011</td>
<td>2.7002</td>
<td>0.0070</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0058</td>
<td>0.0015</td>
<td>3.8476</td>
<td>0.0001</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0003</td>
<td>0.0004</td>
<td>0.8653</td>
<td>0.3869</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>−0.0004</td>
<td>0.0014</td>
<td>−0.2745</td>
<td>0.7837</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0041</td>
<td>0.0017</td>
<td>2.4033</td>
<td>0.0163</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0029</td>
<td>0.0027</td>
<td>−1.0650</td>
<td>0.2870</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>0.0005</td>
<td>0.0043</td>
<td>0.1248</td>
<td>0.9007</td>
</tr>
</tbody>
</table>

spectral norm of a matrix. Let us write \( J_{\beta, \theta} = JL_{\beta, \theta} + JC_{\beta, \theta} \), where \( JL_{\beta, \theta} \) corresponds to the lower triangular part of \( J_{\beta, \theta} \) and \( JC_{\beta, \theta} \) is matrix with only one band of elements above the diagonal. Then we can write

\[
\|J_{\beta, \theta}\|_{sp} = \sup_{\|x\| \leq 1} \|J_{\beta, \theta}x\| \leq \sup_{\|x\| \leq 1} (\|JL_{\beta, \theta}x\| + \|JC_{\beta, \theta}x\|) \leq 1 + 1 = 2,
\]

because \( JL_{\beta, \theta} \) is lower triangular with maximum diagonal entry one and the maximum element of \( JC_{\beta, \theta} \) is also one.

\[
\text{Proof of Lemma 3.2}
\]

\[
\text{Proof. For } m_1 = 2\sigma_1\sigma_2/(\sigma_1^2 + \sigma_2^2) \text{ and } m_2 = (\mu_1 - \mu_2)'(\mu_1 - \mu_2)/(4(\sigma_1^2 + \sigma_2^2)), \quad d_2^2(P, Q) =
\]
Table 3.16: Estimates for Left Entorhin for slope

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0003</td>
<td>0.0010</td>
<td>0.3323</td>
<td>0.7397</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0114</td>
<td>0.0014</td>
<td>7.9188</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0031</td>
<td>0.0019</td>
<td>1.6251</td>
<td>0.1043</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0053</td>
<td>0.0008</td>
<td>6.6727</td>
<td>0.0000</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0054</td>
<td>0.0012</td>
<td>4.5014</td>
<td>0.0000</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0037</td>
<td>0.0017</td>
<td>2.1663</td>
<td>0.0304</td>
</tr>
<tr>
<td>Age:time</td>
<td>−0.0004</td>
<td>0.0004</td>
<td>−0.9482</td>
<td>0.3431</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0058</td>
<td>0.0016</td>
<td>3.5491</td>
<td>0.0004</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0108</td>
<td>0.0021</td>
<td>5.1571</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0028</td>
<td>0.0031</td>
<td>−0.9120</td>
<td>0.3618</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0030</td>
<td>0.0049</td>
<td>−0.6087</td>
<td>0.5428</td>
</tr>
</tbody>
</table>

Table 3.17: Estimates for Right Entorhin for slope

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.0068</td>
<td>0.0014</td>
<td>4.7268</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.0107</td>
<td>0.0021</td>
<td>5.1571</td>
<td>0.0000</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>0.0050</td>
<td>0.0028</td>
<td>1.7943</td>
<td>0.0729</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.0042</td>
<td>0.0012</td>
<td>3.6146</td>
<td>0.0003</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.0167</td>
<td>0.0017</td>
<td>9.6401</td>
<td>0.0000</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.0030</td>
<td>0.0025</td>
<td>1.2302</td>
<td>0.2187</td>
</tr>
<tr>
<td>Age:time</td>
<td>0.0002</td>
<td>0.0006</td>
<td>0.2899</td>
<td>0.7719</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.0015</td>
<td>0.0024</td>
<td>0.6315</td>
<td>0.5278</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.0090</td>
<td>0.0028</td>
<td>3.2332</td>
<td>0.0012</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>−0.0145</td>
<td>0.0045</td>
<td>−3.2515</td>
<td>0.0012</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.0044</td>
<td>0.0071</td>
<td>−0.6283</td>
<td>0.5298</td>
</tr>
</tbody>
</table>

\[1 - m_1^{k/2} \exp(-m_2).\] Note that \((1 - x) \leq \exp(-x)\) for all real \(x\) and \(\exp(-x) \leq (1 - x/2)\) for positive \(x \leq 1\). Also for \(t \geq 2\), \((1 - m_1^t) = 1 - ((m_1 - 1) + 1)^t \leq C_1(1 - m_1)\) as \(0 \leq m_1 \leq 1\), where \(C_1\) can be taken as the sum of the coefficients of the expansion \(((m_1 - 1) + 1)^t\). \(\square\)
Table 3.18: Estimates of covariates for Total Brain for slope from linear model

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0.013</td>
<td>0.001</td>
<td>2155</td>
<td>13.208</td>
<td>0.000</td>
</tr>
<tr>
<td>APOEallele4:time</td>
<td>0.000</td>
<td>0.001</td>
<td>2155</td>
<td>0.047</td>
<td>0.962</td>
</tr>
<tr>
<td>APOEallele2:time</td>
<td>−0.001</td>
<td>0.002</td>
<td>2155</td>
<td>−0.249</td>
<td>0.804</td>
</tr>
<tr>
<td>Gender:time</td>
<td>0.003</td>
<td>0.001</td>
<td>2155</td>
<td>3.077</td>
<td>0.002</td>
</tr>
<tr>
<td>MCI:time</td>
<td>0.006</td>
<td>0.001</td>
<td>2155</td>
<td>4.811</td>
<td>0.000</td>
</tr>
<tr>
<td>AD:time</td>
<td>0.014</td>
<td>0.002</td>
<td>2155</td>
<td>6.969</td>
<td>0.000</td>
</tr>
<tr>
<td>Age:time</td>
<td>−0.002</td>
<td>0.000</td>
<td>2155</td>
<td>−4.663</td>
<td>0.000</td>
</tr>
<tr>
<td>APOEallele4:MCI:time</td>
<td>0.004</td>
<td>0.002</td>
<td>2155</td>
<td>2.689</td>
<td>0.007</td>
</tr>
<tr>
<td>APOEallele4:AD:time</td>
<td>0.004</td>
<td>0.002</td>
<td>2155</td>
<td>2.105</td>
<td>0.035</td>
</tr>
<tr>
<td>APOEallele2:MCI:time</td>
<td>0.001</td>
<td>0.003</td>
<td>2155</td>
<td>0.292</td>
<td>0.770</td>
</tr>
<tr>
<td>APOEallele2:AD:time</td>
<td>−0.003</td>
<td>0.006</td>
<td>2155</td>
<td>−0.530</td>
<td>0.596</td>
</tr>
</tbody>
</table>

Table 3.19: SNPs selected after variable selection from non-linear model along with gene names and main effect on slope

<table>
<thead>
<tr>
<th>SNP name</th>
<th>Gene name</th>
<th>Overall</th>
<th>Main Effect on Atrophy</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1    rs9990174</td>
<td>SLC6A1</td>
<td>0.002</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>2    rs999784</td>
<td>LRBA</td>
<td>0.002</td>
<td>1e-04</td>
<td></td>
</tr>
<tr>
<td>3    rs9998327</td>
<td>KCNIP4</td>
<td>1e-04</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>4    rs9998456</td>
<td>ADGRL3</td>
<td>3e-04</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>5    rs9998709</td>
<td>SORBS2</td>
<td>0.002</td>
<td>5e-04</td>
<td></td>
</tr>
<tr>
<td>6    rs9999069</td>
<td>MND1</td>
<td>2e-04</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>7    rs999915</td>
<td>LPAR3</td>
<td>2e-04</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>8    rs9999448</td>
<td>SHROOM3</td>
<td>0.004</td>
<td>1e-08</td>
<td></td>
</tr>
<tr>
<td>9    rs999981</td>
<td>SORCS3</td>
<td>0.003</td>
<td>4e-08</td>
<td></td>
</tr>
<tr>
<td>10   rs9999820</td>
<td>NPY2R</td>
<td>8e-04</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>11   rs999985</td>
<td>CWF19L2</td>
<td>0.002</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>
To obtain (3.3), we can show the following inequalities which will imply the result trivially:

\[
\sum_{i,j,t} (F_{\Theta_1}(i,j,t) - F_{\Theta_2}(i,j,t))^2 \\
\leq 2 \sum_{i,j,t} [ (a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{0,2}^i(X_i'\beta_2, Z_i'\eta_2))^2 + (a_{1,1}^i(X_i'\beta_1, Z_i'\eta_1)F_{0,1}(t) \\
- a_{1,2}^i(X_i'\beta_2, Z_i'\eta_2)F_{0,2}(t))^2],
\]

(3.24)

\[
(a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{0,2}^i(X_i'\beta_2, Z_i'\eta_2))^2 \\
= (a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{0,1}^i(X_i'\beta_2, Z_i'\eta_2) + a_{0,1}^i(X_i'\beta_2, Z_i'\eta_2) - a_{0,2}^i(X_i'\beta_2, Z_i'\eta_2))^2 \\
\leq 2 \|a_{0,1}^i - a_{0,2}^i\|^2 + 2(a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{0,1}^i(X_i'\beta_2, Z_i'\eta_2))^2,
\]

(3.25)

\[
\frac{1}{n} \sum_{i=1}^n (a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{0,1}^i(X_i'\beta_2, Z_i'\eta_2))^2 \\
\leq \sum_k (\phi_{0,1k}^i)^2 D^2(\|\beta_0 - \beta_1\|_2^2 \lambda_1 + \|\eta_0 - \eta_1\|_2^2 \lambda_2),
\]

(3.26)

\[
\frac{1}{n} \sum_{i=1}^n (a_{1,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{1,1}^i(X_i'\beta_2, Z_i'\eta_2))^2 \\
\leq \sum_k (\phi_{1,1k}^i)^2 D^2(\|\beta_0 - \beta_1\|_2^2 \lambda_1 + \|\eta_0 - \eta_1\|_2^2 \lambda_2),
\]

(3.27)

\[
(a_{0,1}^i(X_i'\beta_1, Z_i'\eta_1)F_{0,1}(t) - a_{1,2}^i(X_i'\beta_2, Z_i'\eta_2)F_{0,2}(t))^2 \\
\leq (a_{0,1}^i(X_i'\beta_2, Z_i'\eta_2))^2(F_{0,1}(t) - F_{0,2}(t))^2 + (a_{1,1}^i(X_i'\beta_1, Z_i'\eta_1) \\
- a_{1,2}^i(X_i'\beta_2, Z_i'\eta_2))^2F_{0,2}(t)^2 \\
\leq (a_{1,1}^i(X_i'\beta_1, Z_i'\eta_1) - a_{1,2}^i(X_i'\beta_2, Z_i'\eta_2))^2 \\
+ (F_{0,1}(t) - F_{0,2}(t))^2 \max_j (a_{1,j}^i(X_i'\beta_j, Z_i'\eta_2))^2;
\]

(3.28)

above, we have used \(n^{-1}z'\sum_{i=1}^n X_iX_i'z \leq \lambda_1 z'z\), where \(\lambda_1\) is the maximum eigenvalue of \(n^{-1}\sum_{i=1}^n X_iX_i'\) and \(z'n^{-1}\sum_{i=1}^n Z_iZ_i'z \leq \lambda_2 z'z\), where \(\lambda_2\) is the maximum eigenvalue of \(n^{-1}\sum_{i=1}^n Z_iZ_i'\).

Proof of (3.6)

For \(\Theta_0\) using Lipshitz condition followed by the bound \(z'n^{-1}\sum_{i=1}^n X_iX_i'z \leq \lambda_1 z'z\), where \(\lambda_1\) is the maximum eigenvalue of \(n^{-1}\sum_{i=1}^n X_iX_i'\) and \(z'n^{-1}\sum_{i=1}^n Z_iZ_i'z \leq \lambda_2 z'z\), where \(\lambda_2\) is the
maximum eigenvalue of $n^{-1} \sum_{i=1}^{n} Z_i Z_i'$ gives

$$\frac{1}{n} \sum_{i=1}^{n} (a_{0,0}^j(X_i'\beta_0, Z_i'\eta_0) - a_{0,1}^j(X_i'\beta_1, Z_i'\eta_1))^2$$

$$\leq 2\|a_{0,0}^j - a_{0,1}^j\|_2^2 + 2 \frac{1}{n} \sum_{i=1}^{n} (a_{0,0}^j(X_i'\beta_0, Z_i'\eta_0) - a_{0,0}^j(X_i'\beta_1, Z_i'\eta_1))^2$$

$$\leq 2\|a_{0,0}^j - a_{0,1}^j\|_2^2 + 2L_1\|\beta_0 - \beta_1\|_2^2 \lambda_1 + 2L_1\|\eta_0 - \eta_1\|_2^2 \lambda_2,$$

and

$$\frac{1}{n} \sum_{i=1}^{n} (a_{1,0}^j(X_i'\beta_0, Z_i'\eta_0) F_{0,0}(t) - a_{1,1}^j(X_i'\beta_1, Z_i'\eta_1) F_{0,1}(t))^2$$

$$\leq M^2\|F_{0,0} - F_{0,1}\|_\infty^2 + \frac{1}{n} \sum_{i=1}^{n} (a_{1,0}^j(X_i'\beta_0, Z_i'\eta_0) - a_{1,1}^j(X_i'\beta_1, Z_i'\eta_1))^2$$

$$\leq 2M^2\|F_{0,0} - F_{0,1}\|_\infty^2 + 2\|a_{1,0}^j - a_{1,1}^j\|_\infty^2 + 2L_2\|\beta_0 - \beta_1\|_2^2 \lambda_1 + 2L_2\|\eta_0 - \eta_1\|_2^2 \lambda_2.$$

Thus using above two relations, we obtain (3.6).
Chapter 4

The Product Independent Gaussian Process Prior for Continuous, Sparse and Piecewise Smooth Signal Estimation

This chapter’s material is a joint work with Ana-Maria Staicu, Brian Reich, Joe Guiness and Russell T. Shinohara.

4.1 Introduction

Modeling a continuous, sparse and piecewise smooth signal for high-dimensional data poses several challenges such as: 1) complex spatial dependence of the data exhibits piecewise smoothness in the signal, 2) the signal is expected to be simultaneously sparse and continuous; here sparsity is defined in terms of the number of non-zero smooth pieces that define the signal and 3) the dimension of such signal is often very large. In a frequentist framework, lasso-type penalization (Tibshirani, 1996) cannot ensure smooth changes from zeros to non-zero subregions. Using fused lasso (Tibshirani et al., 2005), one can ensure both sparsity and smoothness in the estimation. But the approach imposes huge computational demand. In a Bayesian framework, parameter sparsity is modeled using the traditional spike and slab prior (Mitchell and Beauchamp, 1988), the horseshoe prior (Carvalho et al., 2010), normal-gamma prior (Griffin and Brown, 2010), double-Pareto prior (Armagan et al., 2013) or Dirichlet-Laplace prior (Bhattacharya et al., 2015). However, none of these priors ensures a smooth spatial structure. In the context of high dimensional data, this adds computational challenges as well.
We review some of the works on sparse and spatially smooth parameter estimation for different image regressions. For image-on-scalar regression, Yan and Liu (2017) and Chen et al. (2016) tackled a similar problem. The first paper considers a Laplacian type penalty and the second paper introduces a fused SCAD type penalty to account simultaneously for spatial smoothness and sparsity. For studying functional magnetic resonance imaging (fMRI) studies Zhang et al. (2016) and Musgrove et al. (2017) considered similar regression models. In their method of estimation, they considered a spike and slab prior to induce sparsity and considered spatial smoothness for the selection parameter. However, the approach does not ensure that the estimated signal is smooth. In scalar-on-image regression, there is limited work on sparse and piece-wise smooth signal estimation. Goldsmith et al. (2014) and Li et al. (2015) proposed priors that account separately for spatial dependence and sparsity. For the same problem, Wang and Zhu (2017) proposed a penalty based on total variation. The spatial dependence is still not fully incorporated in this approach. In Kang et al. (2016), the proposed soft-thresholded Gaussian process prior account for both spatial dependence and sparsity simultaneously, but due to the thresholding this method suffers in terms of power. In the context of image response and image predictors, there is a more limited research [see Morris et al. (2011); Jog et al. (2013); Sweeney et al. (2013)]. Noh and Park (2010), Tang et al. (2013) considered a varying coefficient model that accounts for sparsity but not smoothness. To the best of our knowledge, only Boehm-Vock et al. (2015) and Jhuang et al. (2018) consider both smoothness and sparsity for image-on-image regression. They capture spatial dependence using copulas, which is computationally expensive for large datasets.

In this chapter, we propose a novel prior which can be used to estimate continuous, sparse and piecewise smooth functions. The proposed prior has a high mass around zero, thus creating sparsity in the estimation, and a smooth a covariance kernel to ensure a large support for the spatially varying function. To handle the heavy computational burden associated with this kind of prior, we propose to use a transformation that reduces the dimensionality of the data. Specifically we use the fast Fourier transformation (FFT) that not only has data-reduction advantages, but also decorrelates the stationary part of the response. The FFT algorithm requires regularly spaced input data. In reality, the datasets are not often on a regular grid. Thus we propose a fast imputation technique to transform the data into a regular grid. If the dimension of the dataset is manageable for computation in the spatial domain, one can exploit the conjugacy structure of our prior to get the full conditional distribution of parameters given the error process is Gaussian. Based on several metrics, we analyze the performance of our prior with respect to commonly used Gaussian process (GP) prior in different linear image regressions with signals that are sparse, piecewise smooth and continuous.

We organize remainder of the chapter as follows. In the next section, we describe the image-
on-scalar regression model along with the new sparse prior process. We discuss the usage of
our new prior to other image regression models in Section 4.3. In Section 4.4, we describe other
computational aspects that we use for faster computation. In Section 4.5, we provide several
simulation results to evaluate the performance of this new prior for different image regression
models extensively. We apply our method to a longitudinal magnetic resonance imaging (MRI)
data in Section 4.6 and end with some concluding remarks in Section 4.7.

4.2 The modeling framework

Our research is motivated by a longitudinal study of multiple sclerosis (MS) via magnetic
resonance imaging (MRI) images. We thus introduce our main ideas for the case when we
have images collected at multiple time points for a single subject. Specifically, let
\( Y_i(v) \) be the intensity of \( i \)-th image collected at time \( t_i \) and at location \( v \), which is a 3 dimensional voxel of
the MRI images for a MS subject. We consider the following linear image-on-scalar regression
model

\[
Y_i(v) = \alpha(v) + t_i \beta(v) + E_i(v),
\]

(4.1)

where \( \alpha(v) \) is a spatially-varying intercept function and \( \beta(v) \) is an unknown continuous function
that quantifies the effect of time, it is assumed that \( \beta(\cdot) \) is in addition piecewise smooth and
sparse. The error \( E_i(v) \) is a spatially-correlated mean-zero error process, independent across
visits. We propose to model \( \beta(v) \) as a product of independent Gaussian processes (GPs). We
formally describe its properties and the error process in the remainder of this section.

4.2.1 PING process

Let \( \beta_1(v), \ldots, \beta_q(v) \) be \( q \) independent and identically distributed GPs with mean \( E\{\beta_j(v)\} = 0 \),
variance \( V\{\beta_j(v)\} = 1 \), and covariance kernel \( \operatorname{Cov}\{\beta_j(v), \beta_j(v')\} = K(v, v') \) for \( j = 1, \ldots, q \).
The zero-mean Product of INdependent Gaussian (PING) stochastic process is defined as the
point-wise product of independent Gaussian processes (GPs), \( \beta(v) = \sigma \beta_1(v) \cdot \beta_2(v) \cdot \ldots \cdot \beta_q(v) \)
where \( \sigma > 0 \) is a scale parameter. The stochastic process \( \beta = \{\beta(v) : v \in V\} \) constructed in
this way is denoted \( \beta \sim \text{PING}(q, \sigma^2, K) \).

4.2.1.1 Properties of the marginal distribution

We first discuss the distribution of the PING process at a single location \( v \). The theoretical
properties of the marginal distribution of \( \beta(v) \) have been studied by Stojanac et al. (2017) and
we revisit them here for completeness. The marginal density function \( f_q(x) \) for the product
of $q$ standard normals is given by $f_q(x) = \frac{1}{(2\pi)^{n/2}} G_{0,q}^q(\frac{x^2}{2}) |0\rangle$, where $G(\cdot)$ denotes the Meijer G-function (Stojanac et al., 2017). The $k^{th}$ marginal moment is $E\{\beta(v)^k\} = [(k-1)!!]^q$ where $n!!$ is the product of all numbers from 1 to $n$ that have the same parity i.e. whether the number is odd or even as $n$. The density is unimodal and symmetric about zero; thus, all the odd-order moments are zero. The variance is $V\{\beta(v)\} = \sigma^2$. The marginal kurtosis is equal to $3^q - 3$ which is an increasing function of $q$. As a result, the marginal density has thicker tail and sharper peak at zero for larger $q$. This is depicted in the first row of Figure 4.1.

4.2.1.2 Properties of the bivariate distribution

Next, we study the bivariate properties of the PING process at a pair of locations $v_1$ and $v_2$. From the construction of the PING process with $q$ components, this bivariate distribution is in fact the distribution of the product of $q$ bivariate normals. Simple calculations show that its mean is $E\{\beta(v_j)\} = 0$ for $j = 1, 2$, and its covariance is $\text{Cov}\{\beta(v_1), \beta(v_2)\} = \sigma^2 K^q(v_1, v_2)$, implying a correlation coefficient that is smaller than the correlation of each individual Gaussian components and that further decays with the number of components, $q$. In particular, if $K(\cdot)$ is the powered exponential correlation kernel, $K(v_1, v_2) = \exp\{-(\|v_1 - v_2\|/\rho)^\nu\}$, the PING covariance is $\exp\{-(\|v_1 - v_2\|/\rho)^\nu\}$. Therefore, while the covariance decreases with $q$ for a fixed kernel function, strong spatial correlation can be maintained for large $q$ by simply increasing the parameter $\rho$ with $q$. Thus the smoothness of the product process remain the same with the individual components for these powered exponential cases. This remedy of separating sparsity and spatial dependence should hold for other kernel functions too. To quantify the shrinkage properties, we study the kurtosis of this product distribution. Kurtosis of a multivariate random variable $Z$ of dimension $p$ with mean $\mu_z$ and covariance matrix $\Sigma_Z$ is defined as $E[(Z - \mu_Z)^T \Sigma_Z^{-1} (Z - \mu_Z)]^2 - p(p+2)$ (Mardia, 1970). The kurtosis of a general product of bivariate normal random variable is summarized by the following theorem.

**Theorem 4.1.** Let $Z_1, \ldots, Z_q$ be such that $Z_i \sim \text{BVN}(0, 0, \sigma_{i1}^2, \sigma_{i2}^2, \rho)$ for $i = 1, \ldots, q$ and $P_q = Z_1 \odot \cdots \odot Z_q$. The mean and the covariance matrix of $P_q$ are $E(P_q) = 0$ and

$$\text{Cov}(P_q) = \begin{bmatrix} \prod_{i=1}^q \sigma_{i1}^2 & \rho \prod_{i=1}^q \sigma_{i1} \sigma_{i2} \\ \\
\rho \prod_{i=1}^q \sigma_{i1} \sigma_{i2} & \prod_{i=1}^q \sigma_{i2}^2 \end{bmatrix}. $$

Then the kurtosis is $\text{Kurt}(q, m) = \frac{2 \times 3^q}{(1-m)^q} [1 + 2(\frac{1+2m}{3})^q + (\frac{1+2m}{3})^q - 4m^q] - 8$, with $m = \rho^2$. Also $\text{Kurt}(q, m)$ increases with $q$.

Here “BVN” stands for bivariate normal and $\odot$ denotes the element wise product. We allow varying variances for individual components as the kurtosis does not depend on the variances. See Section 3.10 for details. The “increasing” property of the kurtosis results from application
of arithmetic and geometric means inequality. The distribution of $\beta(v)$ for two locations $v_1$ and $v_2$ under PING process has the above mentioned properties with $\sigma_{i1} = \sigma_{i2} = 1$ for $i > 1$. Since it is a unimodal symmetric distribution, higher kurtosis suggests a heavier tail and higher peakedness at zero as $q$ increases; Figure 4.1 depicts the joint density function of $\beta(v_1)$ and $\beta(v_2)$ for $q = 1, q = 3$ and $q = 5$, and for different correlations. In this plot, we observe that the mass at zero increases with $q$ while they share same covariance structure with unit variance term. Figure 4.2 illustrates conditional density of one location $s$ given $\beta(s')$. The conditional density at one location tends to have shorter peak as the value at other location moves away from zero. Also conditional densities tend to be more positively skewed as we condition on higher values for the other location.

### 4.2.1.3 Multivariate properties

Let $P_q = Z_1 \otimes \ldots \otimes Z_q$ be random vector of length $p$ where $Z_1 \sim \text{MVN}(0, \Sigma_1)$ and $Z_2, \ldots, Z_q \overset{ind}{\sim} \text{MVN}(0, \Sigma_2)$ and $\Sigma_2$ has diagonal entries equal to one. Then we have $E(P_q'P_q) = \text{trace}(\Sigma_1)$ for all $q$ and $\Sigma_2 \otimes \ldots \otimes (q\text{-times}) \otimes \Sigma_2 \rightarrow I_p$, as $q$ increases to infinity where $I_p$ is the $p \times p$ identity matrix. The distribution of $\beta(v)$ for a finite set of locations has the above mentioned properties as $\beta_2(v), \ldots, \beta_q(v)$ have the same covariance kernel with one total variance. In general, calculations of the kurtosis in multivariate setup is very difficult to obtain. But we can prove the following result.

**Theorem 4.2.** The multivariate kurtosis of $P_q$ increases with $q$.

This can be proved using the method of induction and Theorem 4.1. The proof is in the Section 3.10.

In summary for $q = 1$, the PING process is the standard GP and as $q$ increases mass near zero and tail probabilities increase with maintaining the smoothness properties of the original GP. Therefore, the PING process is an attractive model for sparse and smooth signal.

### 4.2.2 Error distribution and Matérn

Next, we discuss modeling the error process $E_i(v)$ in (4.1). To account for both large and small scale spatial deviations of $Y_i$ we consider the following decomposition, similar to Reich et al. (2018) of $E_i(v)$, 

$$E_i(v) = \sum_{j=1}^{J} Z_j(v) \gamma_{ij} + \epsilon_i(v), \quad (4.2)$$

where the first term is a linear combination of known basis functions $Z_j$'s and $\gamma_{ij}$ are unknown coefficients for $i$-th visit and $j$-th basis and captures the large-scale deviation. The second term
Figure 4.1: Comparison of Gaussian density with PING-3 and PING-5 densities.
Figure 4.2: Conditional density plot of $\beta_2$ given different values of $\beta_1$ for different correlations and components for bivariate realization $(\beta_1, \beta_2)$ at two locations.
\( \epsilon_i \) is intended to capture small scale deviations. We assume that \( \epsilon_i \) as well as the intercept (\( \alpha \)) and each component of the PING prior (\( \beta_k \)'s that comprise the PING prior) are mean-zero GPs with stationary and isotropic Matérn covariance function as follows:

\[
\text{Cov}(\epsilon_i(v), \epsilon_i(v')) = C(\theta) = \sigma^2 I(v = v') + \tau^2 \mathcal{M}_\nu \left( \frac{\|v - v'\|}{\phi} \right),
\]

\( (4.3) \)

\[
\text{Cov}(\alpha(v), \alpha(s')) = C(\theta_0) = \sigma^2_0 I(v = v') + \tau^2_0 \mathcal{M}_{\nu_0} \left( \frac{\|v - v'\|}{\phi_0} \right),
\]

\( (4.4) \)

\[
\text{Cov}(\beta_k(v), \beta_k(v')) = C(\theta_1) = \tau^2_1 \mathcal{M}_{\nu_1} \left( \frac{\|v - v'\|}{\phi_1} \right).
\]

\( (4.5) \)

where \( \mathcal{M}_\nu(h) = \frac{2^{1-\nu}}{\Gamma(\nu)}(3h\sqrt{\nu})^\nu K_\nu(3h\sqrt{\nu}) \) and \( K \) is the modified Bessel function of the second kind. The Matérn covariance has four parameters \( \theta = (\sigma^2, \tau^2, \phi, \nu) \), that represent the variance of the non-spatial error (nugget), the variance of the spatial process (partial sill), the spatial range and the smoothness of the correlation function respectively. We remove the nugget variance from each component of PING to ensure smoothness.

The large-scale spatial structure is described by \( J \) random-effect covariates \( \{Z_1, \ldots, Z_J\} \). Among many different choices for \( Z_j \)'s, we consider outer product of B-spline basis functions. We assume the random effects are normally distributed, i.e., \( \gamma_i = (\gamma_{i1}, \ldots, \gamma_{iJ})^T \sim \text{Normal}(0, \Sigma) \), where \( \Sigma \) is the \( J \times J \) covariance matrix. The nonstationary component of the covariance is

\[
NS(s, s') = \sum_{j=1}^J \sum_{l=1}^J Z_j(v)Z_l(s')\Sigma_{jl}.
\]

\( (4.6) \)

Then the overall covariance becomes sum of \( (4.3) \) and \( (4.6) \).

### 4.3 Extension to other image regression models

The model in the previous section is designed for image-on-scalar regression. Our sparse prior can easily be adopted for other image regressions as described below.

#### 4.3.1 Image-on-image regression

Consider the case of a linear image-on-image regression model (see for example Gelfand et al. (2003); Morris et al. (2011); Jog et al. (2015, 2017); Hazra et al. (2017))

\[
Y_i(v) = \alpha(v) + \sum_{j=1}^p X_{ij}(v)\beta_j(v) + E_i(v),
\]

\( (4.7) \)
where $Y_i$ is the image response; $X_{ij}$’s are the $j$-th $p$ image predictors for subject $i$. Here $\alpha(\cdot)$ is an unknown intercept as before and $\beta_j(\cdot)$ are spatially varying piecewise smooth and sparse covariate effects, and $E_i(\cdot)$ is the error process.

We put the PING prior on each of $\beta_j$ for sparse and smooth estimation. This gives local variable selection as the subset of the covariates with beta shrunk towards zero changes with $s$. In the simulation study of Section 2.4, we investigate the performance of such approach.

### 4.3.2 Scalar-on-image regression

Finally consider the case of a scalar-on-image regression model (see Wang and Zhu (2017); Kang et al. (2016); Goldsmith et al. (2014); Li et al. (2015)). This model is

$$Y_i = \sum_{j=1}^{n} X_i(v_j) \beta(v_j) + \epsilon_i,$$

(4.8)

where $Y_i$ is the scalar response and $X_i$ is an image with $n$ spatial locations for subject $i$. Here $\beta(\cdot)$ are spatially varying piecewise smooth and sparse covariate effect, and $\epsilon_i$ is the error which follows $N(0, \sigma^2)$. We again put a PING prior on $\beta$ for sparse and smooth estimation and its performance is studied Section 2.4.

### 4.4 Computational details

For small and moderate datasets, standard Markov chain Monte-Carlo (MCMC) algorithms apply to the PING model and computation is straightforward. One advantage of the PING prior is the element of the $j$-th component $(\beta_j(v_1), \ldots, \beta_j(v_n))$ have multivariate Gaussian full conditional distribution given the other $(q - 1)$ GPs, and thus Gibbs steps can be used to update the PING process parameters. For large $n$ however, these updates become slow and we use spectral methods, described in the remainder of this section.

#### 4.4.1 The model in spectral domain

Similar to Reich et al. (2018), we partially decorrelate the data by transforming it into the spectral domain. Let us denote spectral representation of the processes $Y_i(v), \alpha(v), \beta(v), \beta_k(v), X_i(v), Z_j(v)$ and $\epsilon_i(v)$ as $\tilde{Y}_i(\omega), \tilde{\alpha}(\omega), \tilde{\beta}(\omega), \tilde{\beta}_k(\omega), \tilde{X}_i(\omega), \tilde{Z}_j(\omega)$ and $\tilde{\epsilon}_i(\omega)$ for frequency $\omega \in \mathcal{F} \subset \mathbb{R}^3$. Since discrete Fourier transformation (DFT) preserves linearity, the spatial model in
(3.1) in the spectral domain can be written as

\[
\tilde{Y}_i(\omega) = \tilde{\alpha}(\omega) + \tilde{\beta}(\omega) + \sum_{j=1}^{J} \tilde{Z}_j(\omega) \gamma_{ij} + \tilde{E}_i(\omega),
\]

(4.9)

\[
\tilde{\beta}(\omega) = \tilde{\beta}_1(\omega) \ast \tilde{\beta}_2(\omega) \ast \ldots \ast \tilde{\beta}_q(\omega).
\]

(4.10)

The notation * denotes convolution. The Gaussian process \( \alpha(v), \beta_k(v) \) and \( \epsilon_i(v) \) are stationary and defined over a discrete spatial domain. In order to avoid computationally expensive Bessel function and spectral aliasing calculations, we use the quasi Matern spectral density (Guinness and Fuentes, 2016), which mimics the flexibility of the Matern spectral density for \( \tilde{\alpha}(\omega), \tilde{\beta}_k(\omega) \) and \( \tilde{\epsilon}_i(\omega) \),

\[
\lambda(\omega|\theta) = (\sigma^2, \tau^2, \phi, \nu) = \sigma^2 + \tau^2 \left[ \frac{1}{\theta^2} + h(\omega) \right]^{-\nu-d/2},
\]

(4.11)

where \( d \) is the dimension, \( \omega \in [0, 2\pi]^d \) and \( h(\omega) = \sum_{j=1}^{d} \sin(\omega_j/2)^2 \). More specifically,

\[
\tilde{E}_i(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta)),
\]

(4.12)

\[
\tilde{\alpha}(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta_0)),
\]

\[
\tilde{\beta}_k(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta_1)),
\]

where \( \tilde{\lambda}(\omega|\theta) = \lambda(\omega|\theta)/2 \) if \( \omega \in \{0, \pi \}^3 \) and \( \tilde{\lambda}(\omega|\theta) = \lambda(\omega|\theta) \) otherwise. All the parameters in \( \theta \) have the same interpretation as in (4.3). For \( \tilde{\beta}_k \), the nugget variance is zero.

### 4.4.2 Imputation method

In practice, data are often not collected on a complete regular grid and thus the response \( Y_i(v) \) is missing at many locations \( v \in \mathcal{V} \) if we transform it into a regular grid. Also in brain imaging, the data is on a complete grid, but there are missing values outside of the skull. Each spectral element \( Y_i(\omega) \) is a function of \( Y_i(v) \) for all \( v \in \mathcal{V} \), and thus spectral methods require complete data. Missing values are handled naturally in a Bayesian context within a Gibbs sampler that draws the missing values from their conditional distribution given the observed data and the other parameters. Because imputation is applied during each MCMC iteration to account for imputation uncertainty, this step must be computationally efficient.

Let mean function including the random effects, accounting for the non-stationary part of the error process, be

\[
\mu_i(v) = \alpha(v) + X_i(v)\beta(v) + \sum_{j=1}^{J} Z_j(v)\gamma_{ij},
\]
and define $Y_{i1}$ to be the vector of observed data for subject $i$ and $Y_{i2}$ to be the vector representing the missing values. Likewise, let $\mu_{i1}$ and $\mu_{i2}$ be the vectors of means. The conditional distribution of $(Y_{i1}, Y_{i2})$ given all of the other parameters is

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix}_{| \text{rest}} \sim N\left( \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

and thus the conditional distribution of $Y_{i2}$ given $Y_{i1}$ and the rest of the parameters is normal with mean $\mu_{i2} + \Sigma_{21}\Sigma_{11}^{-1}(Y_{i1} - \mu_{i1})$ and covariance $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

For large datasets directly sampling from this distribution is infeasible. The limiting computational task in computing the conditional mean is solving a linear system with $\Sigma_{11}$. Since $\Sigma_{11}$ is symmetric and positive definite, this can be achieved with a preconditioned conjugate gradient (PCG) algorithm (Golub and Van Loan, 2012), an iterative method for solving the linear system $\Sigma_{11}a = b$. The goal of iterative linear solvers is to generate a sequence $a_1, a_2, \ldots$ that converges to $a = \Sigma_{11}^{-1}b$. The algorithms generally require us to compute $\Sigma_{11}a_k$ at each iteration $k$ to check for convergence and to generate the next vector in the sequence, and thus the algorithms are fast when this forward multiplication can be computed quickly. In this case, forward multiplications with $\Sigma_{11}$ can be computed in $O(n \log n)$ time and $O(n)$ memory with circulant embedding algorithms (Wood and Chan, 1994), as can the forward multiplication with $\Sigma_{21}$. This is because $\Sigma_{11}$ and $\Sigma_{21}$ can be embedded in the larger circulant matrix $\Sigma$, that is,

$$\Sigma \begin{bmatrix} a_k \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} a_k \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{11}a_k \\ \Sigma_{21}a_k \end{bmatrix},$$

and fast Fourier transform can be exploited to compute the forward multiplication with the (nested block) circulant matrix $\Sigma$, since (nested block) circulant matrices are diagonalizable by the ($d$-dimensional) discrete Fourier transform. The preconditioned conjugate gradient algorithm uses an approximate inverse of $\Sigma_{11}$, called a preconditioner, to encourage the sequence $a_k$ to converge to $a$ is a small number of iterations.

Completing the imputation step requires us to simulate a residual vector with covariance matrix $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. To accomplish this, we first simulate a vector $(\varepsilon_{i1}, \varepsilon_{i2})$ with mean zero and covariance as in (4.13), which is again efficient with circulant embedding. Then we form and the residual $\varepsilon_{i2} - \Sigma_{21}\Sigma_{11}^{-1}\varepsilon_{i1}$, which has the desired and can be computed in the same fashion as the conditioned mean. Further computation details for the conditional draws can be found in Stroud et al. (2016) and Guinness and Fuentes (2016).
### 4.4.3 Sampling

The FFT transformation of the PING process parameters is the convolution of frequencies as in (4.10). Conducting a full conditional Gibbs update, even in the spectral domain, is computationally expensive. The existing Metropolis techniques for joint update of large coefficient vectors, such as the gradient adjusted Metropolis Hastings (Roberts and Rosenthal, 1998) or Hamiltonian Monte Carlo (Duane et al., 1987) are very slow mixing. Thus we introduce this new sampling technique. Our proposed sampling technique has a Gibbs step based on an approximated model and then a function of the Gibbs update is treated as a candidate for the Metropolis step corresponding to the original model. We explain the technique for two-component case, and this can be generalized for any number of components, \(q\). Let \(\beta(v) = \beta_1(v) \cdot \beta_2(v)\) where \(\beta_1\) and \(\beta_2\) are GPs, then the model in (4.1) can be re-written as,

\[
Y_i(v) = \alpha(v) + \beta_1(v)\beta_2(v)t_i + E_i(v). \quad (4.14)
\]

Denote the estimated values at the \(N\)-th stage of the MCMC iteration as \(Y^N\) (samples using PCG), \(\alpha^N\), \(\beta_1^N\) and \(\beta_2^N\). We can calculate the error at \(N\)-th stage as \(E^N = Y^N - \alpha^N - \beta_1^N\beta_2^N t\).

We can rewrite our model in (4.14) as

\[
\frac{Y_i(v)}{\beta_2(v)} = \frac{\alpha(v)}{\beta_2(v)} + \beta_1(v)t_i + E_i(v)(\frac{1}{\beta_2(v)} - 1) + E_i(v). \quad (4.15)
\]

Except for \(\beta_1(v)\) and the last \(E_i(v)\), all other values are replaced by the ones from the \(N\)th step. Then it would look like,

\[
\frac{Y_i^N(v)}{\beta_2^N(v)} = \frac{\alpha^N(v)}{\beta_2^N(v)} + \beta_1^N(v)t_i + E_i^N(v)(\frac{1}{\beta_2^N(v)} - 1) + E_i(v). \quad (4.15)
\]

We now perform a Gibbs sampling to get update, \(\beta_1^u\), of \(\beta_1\) from the model (4.15) in spectral domain given the values from \(N\)th step. Now we consider a Metropolis step corresponding to the original model in (4.14). The candidate for this step is \(\beta_1^N + c\frac{\beta_1^u - \beta_1^N}{\|\beta_1^u - \beta_1^N\|_2}\); here \(c\) acts as a tuning parameter and \(\|\cdot\|\) denotes the \(\ell_2\) norm, defined as \(\|\beta\|_2^2 = \int_{v \in V} \beta^2(v)dv\). Smaller values of \(c\) generate higher acceptance rate and vice versa. In our simulation, we adjust \(c\) to maintain an acceptance rate of around 0.6 for this scheme to ensure good mixing. For number of components more than two, while updating one of the components we consider element-wise product of all other components as one component. Then it reduces to a two component case. Among the Matérn parameters, except for total variance all other parameters are updated using Metropolis sampling and total variances are updated from their posterior inverse gamma distributions.
We use this method for all image-on-scalar regressions in this paper. For the simulated image-on-image and scalar-on-image regressions in Section 2.4 the datasets are small and we use Gibbs sampling for the PING process parameters. For larger problems, the Metropolis scheme, explained above could also be adapted to image-on-image and scalar-on-image regressions.

4.5 Simulation results

In this section we present simulation results for all three regression setups, namely image-on-scalar regression, Image-on-Image regression and Scalar-on-Image regression. We compare the results in terms of MSE, power and Type 1 error for different levels of signal to noise ratios (SNR). For MSE we consider overall MSE as well MSE at the locations where the true value is zero.

4.5.1 Image-on-scalar regression

Here we consider the image-on-scalar regression model in Section 4.2 for images of dimension $20 \times 20 \times 20$ with at 20 visits. The model is

$$Y_i(s) = \alpha(s) + t_i \beta(s) + e_i(s), \quad (4.16)$$

here $s \in \{1, \ldots, 20\}^3$ with $i = 1, \ldots, 20$ and $t_i$'s are 20 equidistant points such that $\sum_i t_i = 0$ and $\sum_i t_i^2 = 20$ obtained by standardizing the times 1 : 20. The slope $\beta(s)$ has the structure, plotted in Figure 4.3. The true signal is zero for most of the spatial locations but has subregions that are non-zero. Let, $d_1 = (6, 14, 6), d_2 = (6, 10, 14), d_3 = (14, 6, 14), d_4 = (14, 14, 14)$ and $d_5 = (6, 6, 6)$

$$\kappa(s) = 2[\exp(-4\|s - d_1\|_2^2/20) + \exp(-1.5\|s - d_2\|_2^2/20) + \exp(-4\|s - d_3\|_2^2/20) + \exp(-4\|s - d_4\|_2^2/20) + \exp(-4\|s - d_5\|_2^2/20).$$

Then mathematically the true beta is $\beta(s) = \kappa(s)\text{ if } \kappa(s) \geq 0.1$ and $\beta(s) = 0$ otherwise. The error process $e_i(s)$ is assumed to be GP with stationary Matérn covariance function. The true reparametrized Matérn parameters for intercept process $\alpha(s)$ are $(1, 0.95, 10, 1)$ and last three parameters for error are $(0.90, 10, 1)$. The reparametrization is explained below.

We put Gaussian process prior with Matérn covariance function on the intercept process $\alpha(s)$ and PING prior (Section 4.2.1) on the slope. We represent $\theta = (\sigma^2, \tau^2, \phi, \nu), \theta_0 = (\sigma_0^2, \tau_0^2, \phi_0, \nu_0)$ and $\theta_k = (\sigma_k^2, \tau_k^2, \phi_k, \nu_k)$ as the Matérn parameters for the error, intercept and $k$-th component in PING process prior on slope respectively. We reparametrize the Matérn parameter $\theta = (\sigma^2, \tau^2, \phi, \nu)$ to $\theta' = (\vartheta^2, \zeta^2, \phi, \nu)$ as $\vartheta^2 = (\sigma^2 + \tau^2)$ and $\zeta^2 = \frac{\tau^2}{\sigma^2 + \tau^2}$. Here $\vartheta^2$ is called the total variance. Similarly we reparametrize $\theta_k$ to $\theta'_k$. The total variance of error is set at 0.09, 0.017 and 0.009 to achieve different SNRs which are mentioned in the Table 4.1. All these results, compiled in Table 4.1 are based on 50 replications.
The priors are described in detail below

- Total variance: $\vartheta^{-2}, \vartheta_0^{-2}, \vartheta_1^{-2} \sim \text{Gamma}(0.1, 0.1)$ and we set $\vartheta_2 = \cdots = \vartheta_q = 1$.

- Other spatial parameters:
  - For error and intercept: $\text{logit}(\zeta), \log \phi, \log \nu \sim \text{N}(0,1)$ and $\text{logit}(\zeta_0), \log \phi_0, \log \nu_0 \sim \text{N}(0,1)$.
  - For PING process: We set $\zeta_1 = \cdots = \zeta_q = 1$ (as nugget variance is zero), $\phi_1 = \cdots = \phi_q = \phi'$ (say) and $\nu_1 = \cdots = \nu_q = \nu'$ (say). Then $\log \phi', \log \nu' \sim \text{N}(0,1)$.

From the values in the Table 4.1, we infer that for lower SNR, more components in the PING process prior leads to better estimation. Figure 4.4 compares the estimates for one slice of the 3-D slope across different methods along with the true $\beta(s)$. Gaussian process prior overestimates the regions where the true value is zero as shown Figure 4.4. This results in higher Type 1 error and higher MSE for locations where the true value is zero. Here all methods have high power.

### 4.5.2 Image-on-image regression

We consider image-on-image regression model as in Section 4.3 on data collected over 100 locations, selected at random in $[0, 1]^2$ with $i = 1, \ldots, 20$ observations at each location. The spatially varying 10 predictors ($X$’s) are generated using the reparametrized Matérn parameters, generated randomly. First a random vector of four elements are generated from $\text{N}(0, 1)$. We exponentiate first, third and fourth element and take inverse logit transformation of the second element to get those reparametrized Matérn parameters for each predictor. The model is

$$ Y_i(s) = \alpha(s) + \sum_{j=1}^{10} X_{ij}(s) \beta_j(s) + e_i(s), \quad (4.17) $$

$\beta_j(s) = 0$ for $j = 1, \ldots, 5$, and other slope parameters are sparse and piecewise smooth. The error process $e_i(s)$ is assumed to be GP with stationary Matérn covariance function, independent over $i$. Rest of those five $\beta$’s have the structures, plotted in Figure 4.5. These are zero at most of the locations with some non-zero subregions. We discuss how these are generated. With out loss of generality, we discuss how $\beta_6$ is generated. We divide the whole $[0, 1]^2$ space into a $50 \times 50$ grid. Then we generate a random number $h$ in $\{1, 2, 3\}$. Then we generate $h$ set of co-ordinates in $[0, 1]^2$. Let these be $u_1, \ldots, u_h$. Let us define $\kappa(s) = \sum_{i=1}^{h} 2 \exp(-3\|s - 50u_i\|^2/50)$. Then $\beta_6(s) = \kappa(s)$ if $\kappa(s) \geq 0.1$ and $\beta_6(s) = 0$ otherwise. The true reparametrized Matérn parameters for intercept are $(1, 0.95, 10, 1)$ and last three parameters for error are $(0.9, 10, 1)$. The total variance of error is set to 0.5664, 0.1133 and 0.0566 to achieve different SNRs which
Figure 4.3: 3-D image of the true slope $\beta(s)$ for the image-on-scalar regression simulation study.
Table 4.1: Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type 1 error for the slope of the image-on-scalar simulation with different SNRs for Gaussian, PING 3 and PING 5 as choices of prior

<table>
<thead>
<tr>
<th>Fitted Model</th>
<th>Gaussian</th>
<th>PING 3</th>
<th>PING 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total MSE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.22 x 10^{-2}</td>
<td>3.08 x 10^{-2}</td>
<td>0.23 x 10^{-2}</td>
<td></td>
</tr>
<tr>
<td>(0.25 x 10^{-3})</td>
<td>(13.09 x 10^{-3})</td>
<td>(1.42 x 10^{-3})</td>
<td></td>
</tr>
<tr>
<td><strong>MSE for $\beta = 0$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.49 x 10^{-3}</td>
<td>4.34 x 10^{-5}</td>
<td>1.25 x 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>(3.61 x 10^{-4})</td>
<td>(0.38 x 10^{-4})</td>
<td>(5.38 x 10^{-4})</td>
<td></td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>0.97</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>0.97</td>
<td>0.70</td>
<td>0.98</td>
</tr>
<tr>
<td><strong>Type 1 error</strong></td>
<td>0.063</td>
<td>0.001</td>
<td>0.0</td>
</tr>
</tbody>
</table>

| **Total MSE** | 9.63 x 10^{-4} | 12.07 x 10^{-4} | 4.19 x 10^{-4} |
| (1.19 x 10^{-4}) | (11.44 x 10^{-4}) | (0.49 x 10^{-4}) |
| **MSE for $\beta = 0$** |          |        |        |
| 14.95 x 10^{-4} | 3.84 x 10^{-4} | 4.10 x 10^{-4} |
| (2.35 x 10^{-4}) | (1.87 x 10^{-4}) | (0.56 x 10^{-4}) |
| **Power** | 1 | 1 | 1 |
| **Coverage** | 0.88 | 0.98 | 0.99 |
| **Type 1 error** | 0.219 | 0 | 0 |

| **Total MSE** | 7.58 x 10^{-4} | 3.79 x 10^{-4} | 2.25 x 10^{-4} |
| (8.17 x 10^{-5}) | (3.09 x 10^{-5}) | (2.64 x 10^{-5}) |
| **MSE for $\beta = 0$** |          |        |        |
| 11.94 x 10^{-4} | 2.49 x 10^{-4} | 2.25 x 10^{-4} |
| (2.11 x 10^{-4}) | (9.57 x 10^{-5}) | (3.06 x 10^{-5}) |
| **Power** | 1 | 1 | 1 |
| **Coverage** | 0.85 | 0.98 | 0.99 |
| **Type 1 error** | 0.272 | 0 | 0 |
Figure 4.4: Plot of the estimates for the slope of the image-on-scalar regression simulation with different signal to noise ratios (SNRs) for Gaussian, Product of independent Gaussian with 3 components (PING 3) and PING 5 as choices of prior.
are mentioned in the table. We report MSE, power, Type 1 error and coverage averaged over $\beta$. All these results, compiled in Table 4.2 are based on 50 replications.

The priors for the Matérn parameters are same as previous section. The only exception is with computation. Since the dimension of this dataset was manageable for Gibbs sampling of the intercept and slope parameters, we do not use spectral methods here.

In Table 4.2, we see that the PING process prior always gives better estimate in terms of MSE. The GP prior overestimates the regions where the true value is zero. This results in higher Type 1 error for GP prior. Here all methods have similar power.

![Fig4.5](image.png)

**Figure 4.5:** Plots of the truth for last five regression coefficients (first five were zero) from second simulation corresponding to image-on-image regression.

### 4.5.3 scalar-on-image regression

We closely replicate the simulation from (Kang et al., 2016) with 100 observations. For each observation, there is a two-dimensional image $X_i$ of dimension $20 \times 20$ with an exponential
Table 4.2: Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type 1 error for the slope of the image-on-image simulation with different SNRs for Gaussian, PING 3 and PING 5 as choices of prior

<table>
<thead>
<tr>
<th>SNR</th>
<th>Metric</th>
<th>Fitted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gaussian</td>
</tr>
<tr>
<td>1</td>
<td>Total MSE</td>
<td>$3.69 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>MSE for $\beta = 0$</td>
<td>$2.91 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(4.25 \times 10^{-4})$</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.020</td>
</tr>
<tr>
<td>5</td>
<td>Total MSE</td>
<td>$9.71 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>MSE for $\beta = 0$</td>
<td>$8.82 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(11 \times 10^{-5})$</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.029</td>
</tr>
<tr>
<td>10</td>
<td>Total MSE</td>
<td>$5.36 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>MSE for $\beta = 0$</td>
<td>$5.00 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(5.88 \times 10^{-5})$</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.033</td>
</tr>
</tbody>
</table>
covariance structure having range parameter 3. The model is

\[ Y_i \overset{\text{ind}}{\sim} N(\sum_{j,k=1}^{20} X_{ijk} \beta_{jk}, \sigma^2), \]  

(4.18)

Here the coefficient \( \beta = ((\beta_{jk})) \) is a matrix of dimension 20 \( \times \) 20. The true \( \beta \) is generated in such a way that it has five peaks. Let, \( d_1 = (4, 16), d_2 = (16, 4), d_3 = (4, 4), d_4 = (16, 16) \) and \( d_5 = (10, 10) \) and \( \kappa(s) = \sum_{i=1}^{5} 2 \exp(-20||s - d_i||^2/50) \). Then mathematically the true beta is \( \beta(s) = \kappa(s) \) if \( \kappa(s) \geq 0.1 \) and \( \beta(s) = 0 \) otherwise. Only in this set up we have number of observations much less than number of parameters to be estimated. We report MSE, power, type 1 error and coverage in estimation of the slope \( \beta \) matrix. All these results, compiled in Table 4.3, are based on 50 replications.

We put PING prior on the slope \( \beta \). We represent \( \theta_k = (\sigma^2_k, \tau^2_k, \phi_k, \nu_k) \) as the Matérn parameters for \( k \)-th component in PING process prior on slope. We shall reparametrize the Matérn parameter as before. We consider three choices of \( \sigma^2 \) in generating the data, 0.1, 1 and 1.5.

The prior for \( \sigma^{-2} \) is Gamma(0.1, 0.1). Rest of parameters have same prior as previous subsections. We consider low rank approximation for each component of PING prior on \( \beta \) while comparing the performance with soft-thresholded Gaussian process (STGP) which is proposed in (Kang et al., 2016). Instead of Matérn, we consider conditionally autoregressive prior for each coefficient corresponding to low rank approximations of the components in PING prior.

In Table 4.3, we see that the estimates from the PING process prior are superior to those of the GP in terms of MSE. In Figure 4.6 the estimated parameters from PING are less noisy than the Gaussian estimates. Since it is an \( n < p \) problem, we can see improvement in all metrics for the PING prior.

While comparing with STGP method (Kang et al., 2016), we consider low rank approximation of each component of PING. Due to this, all the results change for each cases of PING. Thus we compile the comparison results with STGP in a separate table. In Table 4.4, we compare MSE, power, type 1 error and coverage in estimation of the slope \( \beta \) matrix between STGP method and PING 8 components. We consider 8 after cross-validating over a grid of number of components. For this comparison, we consider low rank approximation of (Kang et al., 2016) both for the STGP and each component of PING 8. These results are based on 50 replications. We can see that the estimates from PING process prior are far better than STGP in terms of overall MSE, power and coverage but the MSE at the subregion where the truth is zero. In STGP method, the estimates are made zero beyond a certain threshold. Thus this kind of results are indeed expected.

89
Figure 4.6: Comparison plot of the estimates for the slope of the scalar-on-image simulation with different true variances for Gaussian, PING 3 and PING 5 as choices of prior.
Table 4.3: Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and Type1 error for the slope of the scalar-on-image regression model with different true variances for Gaussian, PING 3 and PING 5 as choices of prior

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>Metric</th>
<th>Fitted Model</th>
<th>Gaussian</th>
<th>PING 3</th>
<th>PING 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total MSE</td>
<td>0.02090</td>
<td>0.00031</td>
<td>0.00050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000979)</td>
<td>(0.000017)</td>
<td>(0.000043)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>MSE for $\beta = 0$</td>
<td>0.01275</td>
<td>0.00018</td>
<td>0.00020</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000438)</td>
<td>(0.000014)</td>
<td>(0.000018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.0176</td>
<td>0.0037</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.94</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.94</td>
<td>0.99</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total MSE</td>
<td>0.0256</td>
<td>0.0033</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00171)</td>
<td>(0.00077)</td>
<td>(0.00151)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MSE for $\beta = 0$</td>
<td>0.0166</td>
<td>0.0011</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00112)</td>
<td>(0.00032)</td>
<td>(0.00036)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.01424</td>
<td>0.00068</td>
<td>0.00102</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.86</td>
<td>0.91</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.96</td>
<td>0.99</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total MSE</td>
<td>0.0301</td>
<td>0.0062</td>
<td>0.0088</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0018)</td>
<td>(0.0021)</td>
<td>(0.0028)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>MSE for $\beta = 0$</td>
<td>0.0201</td>
<td>0.0021</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00132)</td>
<td>(0.00066)</td>
<td>(0.00082)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.00814</td>
<td>0.00034</td>
<td>0.00034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.79</td>
<td>0.75</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4: Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power and Type 1 error for the slope of the scalar-on-image simulation with different true variances for soft-thresholded Gaussian process (STGP) and PING 8 as choices of prior.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>Metric</th>
<th>STGP</th>
<th>PING 8 components</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Total MSE</td>
<td>0.002864</td>
<td>0.000262</td>
</tr>
<tr>
<td></td>
<td>(0.001259)</td>
<td>(0.000084)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE for $\beta = 0$</td>
<td>0.000027</td>
<td>0.000117</td>
</tr>
<tr>
<td></td>
<td>(0.000086)</td>
<td>(0.000020)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.867744</td>
<td>1.000000</td>
</tr>
<tr>
<td></td>
<td>Type 1 error</td>
<td>0.109622</td>
<td>0.066541</td>
</tr>
</tbody>
</table>

| 1        | Total MSE  | 0.003810   | 0.000685          |
|          | (0.001540) | (0.000283) |
|          | MSE for $\beta = 0$ | 0.000079 | 0.000315          |
|          | (0.000128) | (0.000092) |
|          | Power      | 0.858622   | 0.999048          |
|          | Type 1 error | 0.104044 | 0.056101          |

| 1.5      | Total MSE  | 0.004220   | 0.001303          |
|          | (0.001570) | (0.000588) |
|          | MSE for $\beta = 0$ | 0.000104 | 0.000482          |
|          | (0.000149) | (0.000182) |
|          | Power      | 0.851933   | 1.000000          |
|          | Type 1 error | 0.101367 | 0.061761          |

Figure 4.7: Variation in MSE of estimated slope for varying number of components $q$ for scalar-on-image simulation.
4.6 Application to longitudinal MRI data

Next, we turn to the study of multiple sclerosis (MS) using MRI images. Each subject was scanned approximately once per month over several hospital visits. There are individuals who were scanned over 3 years. We focus on the set of images of a single subject. Using a 1.5T GE scanner with clinically optimized scanning parameters whole-brain magnetization transfer ratio (MTR), T1 weighted, and fluid attenuation inversion recovery (FLAIR) volumes were acquired. All the modalities were interpolated to a voxel size of $1mm^3$ yielding images of dimension $182 \times 218 \times 182$. We use normalized FLAIR images in our study. All images were registered longitudinally and across the modalities and rigidly aligned to the Montreal Neurological Institute standard space (Fonov et al., 2009). (Sweeney et al., 2016) has a complete description of the study along with the acquisition parameters. Our preliminary investigation seems to indicate that the image intensity varies linearly with time. Let $Y_i(v)$ denotes the image intensity at a 3 dimensional voxel $v$ of $i$-th image at time $t_i$, which denotes number of days passed between $i$-th and the first visit of a single subject. In general $v$ is used to denote voxel. We normalize the time covariate $t_i$ and set the image of the first visit as the baseline. We consider the following model from Section 4.2

$$Y_i(v) = \alpha(v) + \beta(v)t_i + E_i(v),$$

where $\alpha(v)$ is the spatially varying intensity image at baseline visit and $\beta(v)$ quantifies the brain regions that are deteriorated over time due to MS. In such a short period of time, healthy brain is not expected to change much. It is assumed that few regions are changing due to MS. Thus $\beta(v)$ is expected to be sparse. But due to spatial dependence in the brain, it is also desired to be piecewise smooth and continuous.

For the error process we consider the non-stationary covariance model as discussed in Section 4.2.2. We use GP prior with Matérn covariance function for $\alpha$. Further details of the model specifications are: $\gamma_i \sim N(0, \Sigma)$ where $\Sigma \sim IW(J + 0.1, \frac{c}{J+0.1}I_J)$ and $c^{-1} \sim$ Gamma$(0.1,0.1)$, where “IW” stands for inverse Wishart, $J = 6^3 = 216$ B-spline basis functions. We reduce the dimensionality of the images to $91 \times 109 \times 91$ using resize function of imager package of R; the reduced images preserve the overall structure of the original images. The time of the visits is roughly every month. We normalize the time vector. We present the analysis for one MS subject in the study. We provide the plots of real data of the middle cross section for their first 12 visits in Figure 4.8.

We use both the proposed method with the signal modeled using the PING process, and with the signal modeled via a GP. We consider both PING with three and five components and Gaussian prior for the slope $\beta$ and compare the estimates in terms of prediction MSE. We consider the data for first 11 time points to estimate the model parameters and based on
that calculate prediction MSE for 12-th time point. The estimates are based on 5000 post-burn MCMC samples after 5000 burn-in. We sample the values in the image outside of the brain using the techniques of Section 4.4.2 after each 30 iterations. There is not much difference in prediction MSE using PING-3 and PING-5. The PING priors slightly improve the MSE for Gaussian prior from 1.22 to 1.18. Figure 4.9 compares the estimates from Gaussian, PING-3 and PING-5 are provided for the middle cross section of which the real data plot is given in Figure 4.8. In these plots the affected regions are better highlighted in PING estimates than Gaussian.

4.7 Conclusion and discussion

We propose a new class of prior, entitled the PING prior, for estimating spatially sparse and smooth signals. We analyze the performance of our prior in different kinds of image regressions, namely image-on-scalar, image-on-image and scalar-on-image. We develop techniques to tackle huge dimensional datasets by transforming into spectral domain. Our simulations show that this new prior work better than Gaussian prior for all the image regressions. Due to inherent conjugacy structure of the components in spatial domain given that the error process is Gaussian, one can use various other kind of covariance kernels as well for each component instead of the ones, used in this chapter.

Our simulation results suggest that PING priors give better estimates than Gaussian at the locations where true value is zero. This results in lower Type 1 error for PING. All of the methods have high power for both image-on-scalar and image-on-image regression models. For the scalar-on-image model they even have better power than Gaussian along with lower Type 1 error and MSEs. The versatility in application of this prior is well studied in the simulation section of this chapter. In the real data, there is little improvement in prediction MSE from Gaussian to PING. But the disease affected areas are more easily distinguishable in PING estimates due to shrinkage.

4.8 Proofs

Proof of the Theorem 1:

Let $\Sigma_{P_q}$ be the covariance matrix of $P_q$

$$\Sigma_{P_q} = \begin{bmatrix}
\prod_{i=1}^{q} \sigma_{i1}^2 & \rho^q \prod_{i=1}^{q} \sigma_{i1} \sigma_{i2} \\
\rho^q \prod_{i=1}^{q} \sigma_{i1} \sigma_{i2} & \prod_{i=1}^{q} \sigma_{i2}^2
\end{bmatrix}$$
Figure 4.8: Data for each visit of the middle slice of the MRI image for 12 visits of the selected subject.
Figure 4.9: Estimated slope $\beta(s)$ of the middle slice using different priors along with the color scale.

and $\Sigma_{P_q}^{-1} =$

$$
\prod_{i=1}^{q} \sigma_{i1}^2 \sigma_{i2}^2 (1 - \rho^2 q^2)^2 \left[
\begin{array}{cc}
\prod_{i=1}^{q} \sigma_{i2}^2 & -\rho^2 \prod_{i=1}^{q} \sigma_{i1} \sigma_{i2} \\
-\rho^2 \prod_{i=1}^{q} \sigma_{i1} \sigma_{i2} & \prod_{i=1}^{q} \sigma_{i1}^2
\end{array}
\right]
$$

Kurtosis of $P_q$ is defined by $E(P_q^{T} \Sigma_{P_q}^{-1} P_q)^2 - 8$. Here $P_q = (P_{q,1}, P_{q,2})$ Let $w_{11} = \prod_{i=1}^{q} \sigma_{i2}^2$, $w_{12} = -\rho^2 \prod_{i=1}^{q} \sigma_{i1} \sigma_{i2}$, $w_{22} = \prod_{i=1}^{q} \sigma_{i1}^2$ and $D = \prod_{i=1}^{q} \sigma_{i1}^2 \sigma_{i2}^2 (1 - \rho^2 q^2)^2$

Using the notations, defined above we get $E(P_q^{T} \Sigma_{P_q}^{-1} P_q)^2 = D^2 E(w_{11}^2 P_{q,1}^4 + w_{22}^2 P_{q,2}^4 + 2w_{11} w_{22} P_{q,1}^2 P_{q,2}^2 + 4w_{11}^2 P_{q,1}^2 P_{q,2}^2 + 4w_{12} w_{21} P_{q,1}^3 P_{q,2} + 4w_{12} w_{22} P_{q,1} P_{q,2}^3)$

We have, $E(X_{i1}^4) = 3\sigma_{i1}^4$, $E(X_{i1}^2 X_{i2}^2) = \sigma_{i1}^2 \sigma_{i2}^2 (1 + 2\rho^2)$, $E(X_{i1}^3 X_{i2}) = 3\rho \sigma_{i1}^3 \sigma_{i2}$, $E(X_{i1} X_{i2}^3) = 3\rho \sigma_{i1} \sigma_{i2}^3$.

Using above results we get,
Hence the above inequality will hold if \(2 \times 3^q + 4(1 + 2r^2)r^2q + 2(1 + 2r^2)q - 8 \times 3^q \rho^{2q} = 2 \times 3^q \left[1 + 2 \left(\frac{1 + 2r^2}{3}\right)^q \right] + \left(\frac{1 + 2r^2}{3}\right)^q - 4 \rho^{2q}\), where \(m = \rho^2\).

For \(q = 1\) we have \(2 \times 3^q = 8\), which is the kurtosis for bivariate normal and 2 \(\times 3^q\) increases with \(q\). This fraction \(T = \frac{1}{1 - m^2} [1 + 2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q - 4m^q]\) approaches to 1 as \(q\) approaches to \(\infty\) for each \(m\). For \(q = 1\), this is 4/3. If we can show this fraction \(4/3 \geq T \geq 1\) for 0 < \(m \leq 1\) and \(q \geq 1\), we are done since then the overall kurtosis will increase for every unit increase in \(q\).

We first simplify the above inequality in the following few steps:

\[\frac{1}{1 - m^2} [1 + 2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q - 4m^q] \geq 1 \iff 1 + 2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q - 4m^q \geq (1 - m^q)^2 \iff 2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q \geq 2m^q + m^{2q}\]

By applying AM-GM inequality in LHS, we get \(2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q \geq 2m^{5q/3} + m^{2q/3}\). Hence the above inequality will hold if \(2m^{5q/3} + m^{2q/3} \geq 2m^q + m^{2q/3} \iff 2m^{2q/3} + m^{q/3} \geq 2 + m^q\).

We have \(1 + m^{2q/3} \geq 2m^{2q/3} \Rightarrow (1 + m^{2q/3})(1 - m^{2q/3}) \geq 2m^{2q/3}(1 - m^{2q/3}) \Rightarrow 1 - m^{4q/3} \geq 2m^{2q/3} - 2m^q \Rightarrow m^{-q/3} - m^q \geq 2 - 2m^{2q/3} \Rightarrow 2m^{2q/3} + m^{-q/3} \geq 2 + m^q\).

To show \(T \leq 4/3\) we need \(3(1 + 2 \left(\frac{1 + 2m}{3}\right)^q + \left(\frac{1 + 2m}{3}\right)^q - 4m^q) \leq 4(1 - m^q)^2 \iff 3 \left(\frac{1 + 2m}{3}\right)^q (2m^q + 1) \leq (2m^q + 1)^2 \iff 3 \left(\frac{1 + 2m}{3}\right)^q \leq 2m^q + 1\). Since \(m^q\) is a convex function we have \(\left(\frac{1 + 2m}{3}\right)^q \leq m^q\).
This completes the proof.

We can further comment on the behavior of \( \frac{1}{(1-m^q)^2} \left[ 1 + 2 \left( \frac{(1+2m)m}{3} \right)^q + \frac{(1+2m)^q}{3} - 4m^q \right] \) as \( m \) varies from 0 to 1. It is 1 at \( m = 0 \) and the limiting value of this quantity as \( m \) tends to 1 is 1 as well. Between [0,1] this fraction is always positive and its maxima depends on the value of \( q \).

If we have varying correlation coefficient i.e. \( X_i \sim \text{BVN}(0, 0, \sigma_{11}^2, \sigma_{12}^2, \rho_i) \), the kurtosis will become \( \frac{2 \times 3^q}{(1-\prod_{i=1}^{q} m_i)^2} \left[ 1 + 2 \left( \prod_{i=1}^{q} \left( \frac{1+2m_i}{3} \right) \right) - 4 \prod_{i=1}^{q} m_i \right] - 8 \) with \( m_i = \rho_i^2 \). By taking \( m \) as the geometric mean of \( m_1, \ldots, m_q \) after applying AM-GM inequality in the proof similar result can be proved for this case.


