BAYESIAN VARIABLE SELECTION AND FUNCTIONAL DATA ANALYSIS: APPLICATION TO BRAIN IMAGING

By

Asish Kumar Banik

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2019

ABSTRACT

BAYESIAN VARIABLE SELECTION AND FUNCTIONAL DATA ANALYSIS:
APPLICATION TO BRAIN IMAGING

By

Asish Kumar Banik

High-dimensional statistics is one of the most studied topics in the field of statistics. The most interesting problem to arise in the last 15 years is variable selection or subset selection. Variable selection is a strong statistical tool that can be explored in functional data analysis. In the first part of this thesis, we implement a Bayesian variable selection method for automatic knot selection. We propose a spike-and-slab prior on knots and formulate a conjugate stochastic search variable selection for significant knots. The computation is substantially faster than existing knot selection methods, as we use Metropolis-Hastings algorithms and a Gibbs sampler for estimation. This work focuses on a single nonlinear covariate, modeled as regression splines. In the next stage, we study Bayesian variable selection in additive models with high-dimensional predictors. The selection of nonlinear functions in models is highly important in recent research, and the Bayesian method of selection has more advantages than contemporary frequentist methods. Chapter 2 examines Bayesian sparse group lasso theory based on spike-and-slab priors to determine its applicability for variable selection and function estimation in nonparametric additive models.

The primary objective of Chapter 3 is to build a classification method using longitudinal volumetric magnetic resonance imaging (MRI) data from five regions of interest (ROIs). A functional data analysis method is used to handle the longitudinal measurement of ROIs, and the functional coefficients are later used in the classification models. We propose a Pólya-gamma augmentation method to classify normal controls and diseased patients based on functional MRI measurements. We obtain fast-posterior sampling by avoiding the slow and complicated Metropolis-Hastings algorithm. Our main motivation is to determine the important ROIs that have the highest separating power to classify our dichotomous response. We compare the sensitivity, specificity,

and accuracy of the classification based on single ROIs and with various combinations of them. We obtain a sensitivity of over 85% and a specificity of around 90% for most of the combinations.

Next, we work with Bayesian classification and selection methodology. The main goal of Chapter 4 is to employ longitudinal trajectories in a significant number of sub-regional brain volumetric MRI data as statistical predictors for Alzheimer's disease (AD) classification. We use logistic regression in a Bayesian framework that includes many functional predictors. The direct sampling of regression coefficients from the Bayesian logistic model is difficult due to its complicated likelihood function. In high-dimensional scenarios, the selection of predictors is paramount with the introduction of either spike-and-slab priors, non-local priors, or Horseshoe priors. We seek to avoid the complicated Metropolis-Hastings approach and to develop an easily implementable Gibbs sampler. In addition, the Bayesian estimation provides proper estimates of the model parameters, which are also useful for building inference. Another advantage of working with logistic regression is that it calculates the log of odds of relative risk for AD compared to normal control based on the selected longitudinal predictors, rather than simply classifying patients based on cross-sectional estimates. Ultimately, however, we combine approaches and use a probability threshold to classify individual patients. We employ 49 functional predictors consisting of volumetric estimates of brain sub-regions, chosen for their established clinical significance. Moreover, the use of spike-and-slab priors ensures that many redundant predictors are dropped from the model.

Finally, we present a new approach of Bayesian model-based clustering for spatiotemporal data in chapter 5 . A simple linear mixed model (LME) derived from a functional model is used to model spatiotemporal cerebral white matter data extracted from healthy aging individuals. LME provides us with prior information for spatial covariance structure and brain segmentation based on white matter intensity. This motivates us to build stochastic model-based clustering to group voxels considering their longitudinal and location information. The cluster-specific random effect causes correlation among repeated measures. The problem of finding partitions is dealt with by imposing prior structure on cluster partitions in order to derive a stochastic objective function.

Copyright by
ASISH KUMAR BANIK
2019

I dedicate this dissertation to my late grandparents for their endless love, support and blessings

ACKNOWLEDGEMENTS

I would like to first thank my supervisor Dr. Tapabrata Maiti for his continuous support, encouragement and assistance. His expertise and experience helped me to formulate the research problems and kept me motivated all the time. I am also grateful to the Department of Statistics and Probability for providing me the opportunity to grow my knowledge over years.

Next, I would like to thank Dr. Alla Sikorskii. Dr. Sikorskii supported me as a Research Assistant with College of Nursing for over four years. I have received a great deal of support from her to build my expertise around clinical trials research. I would also like to thank Professor Dr. Andrew Bender for providing us the brain image data and helped me to understand the nitty gritty details of human brain mechanisms. His research questions motivated me to build various statistical methods for this dissertation.

My sincerest gratitude to other committee members Dr. R.V. Ramamoorthi and Dr. Seungik Baek for providing important suggestions and invaluable advice. I also would like to thank all Department of Statistics professors for their valuable lessons through various courses.

Last but not least, I would like to devote my deepest gratitude to my parents. They believed in me like nobody else. I would like to thank my brother and sister-in-law for encouraging throughout writing this dissertation. My final gratitude is for my friends and peers. I always received much needed important suggestions from them through out my research journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Variable selection

Statistical modeling with large numbers of predictors is one of the most studied topics in the field of statistics. The most interesting problem to arise in the last 15 years is variable selection or subset selection. Statisticians are excited about variable selection for various reasons. Most often, when a model contains a large number of predictors, we are not satisfied with prediction accuracy, interpretation, over-fitting, estimation, or testing of regression coefficients, among others. Numerous variable selection methods have been proposed in the Bayesian and non-Bayesian literature, such as forward and backward stepwise selection, shrinkage methods, principal components regression, and partial least squares. The problem of choosing an optimal model from a subset of all possible models has led to a variety of algorithms and methodology building. The following discusses Bayesian variable selection and its application to various nonparametric regression problems.

1.1.1 Bayesian Variable Selection and Spike-and-Slab Prior

We start by defining the setup,

$$y = \sum_{i=1}^{p} X_i \beta_i + \epsilon$$

where $p \gg n$ and y is a $n \times 1$ vector. It is mathematically intuitive that the parameter vector β can only be estimated by n observations. If β is sparse in terms of l_0 norm i.e. $||\beta||_0^0 = |\{i : \beta_i \neq 0\}| < n$, then we can achieve reasonably good estimates. Various authors worked on this problem. *Miller, 2002*[89] used a spike-and-slab prior with a spike on the zero coefficient and a slab part on the estimable coefficients. A spike-and-slab prior was introduced by [43] where β_i follows a normal mixture distribution:

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2)$$

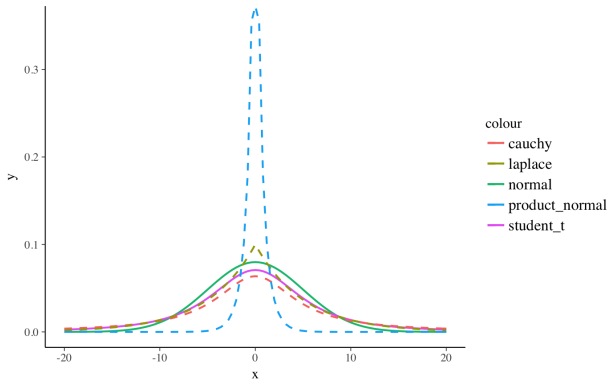


Figure 1.1: spike-and-slab Prior[a]

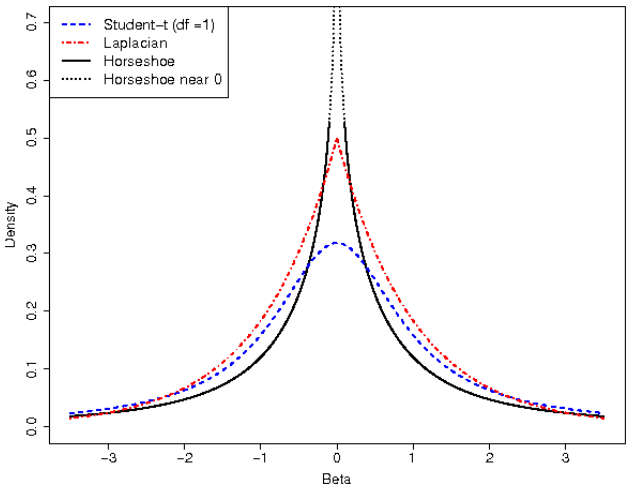


Figure 1.2: Horseshoe prior[a]

[a] Plot obtained from on-line resources

where γ_i is a latent variable that controls the number of significant coefficients or model size. The motivation for the frequently used Dirac-distributed spike-and-slab prior came from the above structure. We observe a point mass on the spike part when τ_i^2 tends to 0. Model complexity is a required property. The degree of sparseness can be set based on an optimality criterion. *Burnham and Anderson, 2004*[3] mentioned the best predictive ability as an optimal criterion in their book, "Model Selection and Multi-model Inference." As we have noted $P(\gamma_i = 1)$ controls the sparsity of the model; *George & McCulloch, 1993*[43] proposed to use a Bernoulli distribution with 0.5 probability. Bayesian variable selection is not limited to spike-and-slab prior. For example, *Liang et al., 2008* [78] used a mixture of g-priors for variable selection. *Carvalho et al., 2010*[18] introduced the horseshoe prior as a global-local shrinkage. The horseshoe is another popular prior choice for variable selection where a half-cauchy prior is used to estimate significant predictors. We have seen that this method penalizes less on the strong signals and parameters are estimated with fewer constraints. However, it comes with a computation cost, as in most cases it is difficult to determine a proper posterior distribution. *Casella and Moreno, 2006*[21] presented an objective Bayesian criterion for variable selection. In general, the literature on variable selection is vast, but we focused on spike-and-slab prior in this study.

The application of a spike-and-slab prior can be extended to various problems, such as nonparametric function estimation or curve estimation. *Smith and Kohn, 1996*[115] applied a similar approach for univariate nonparametric regression. We closely studied their work for function estimation. Although their prior structure does not directly resemble a spike-and-slab setup, they also used an indicator variable γ, such that

$$\beta_\gamma \sim N(0, c\sigma^2(X_\gamma^T X_\gamma)^{-1}), \;\; p(\gamma_i) = \pi_i$$

where X_γ is the design matrix with columns whose regression coefficients are nonzero. Once we find $p(\gamma|y)$, it is easy to sample using Gibbs sampling. *Smith and Kohn, 1996*[115] used this approach to select significant knots from an unknown function $f(x)$ approximated by cubic spline. Later, *Ishwaran and Rao, 2005*[57] started reshaping the spike-and-slab prior. They described the following setup in their paper:

$$\begin{aligned}
(Y_i/x_i, \beta, \sigma^2) &\overset{ind}{\sim} N(x_i'\beta, \sigma^2), \quad (i = 1, ..., n) \\
(\beta/\gamma) &\sim N(\mathbf{0}, \Gamma), \\
\gamma &\sim \pi(d\gamma), \\
\sigma^2 &\sim \mu(d\sigma^2),
\end{aligned}$$

The above formulation is a general setup, where Γ is a diagonal matrix. The authors developed continuous bimodal priors which have separate spike and slab parts. *Malsiner-Walli and Wagner*[82] performed a thorough revision of spike-and-slab priors. They argued that the selected indicator variables are independent conditional on the prior inclusion probability, but marginal dependence is not a logical condition. Instead, they used an individual inclusion probability for each regression coefficient.

1.1.2 Penalization Methods

Variable selection has been extensively studied in the non-Bayesian literature too. Lasso (*Tibshirani, 1996* [120]) is the most popular variable selection technique, initially developed for linear

regression models. If $p > n$, the ordinary least square estimator is not unique and *Tibshirani* introduced the idea of l_1 penalization on regression coefficients,as formulated below:

$$\hat{\beta}_{lasso}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right)$$

where $\lambda \geq 0$ is the penalty parameter. This estimator can drag many coefficients towards zero, depending on the sparsity of the design matrix and the extent of the penalization. The solution is a convex optimization problem and easy to implement. Lasso can perform variable selection in the sense that it can provide an exact zero solution for some components. On the other hand, another old and popular penalization method, ridge regression, shrinks regression coefficients by penalizing on the size. The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta}_{ridge}(\lambda) = \arg\min_{\beta} \left(\frac{1}{n} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2 \right)$$

Similarly λ, the complexity parameter, controls the amount of shrinkage. Ridge regression performs proportional shrinkage, and lasso truncates the coefficient at zero. To understand the relationship, we use the famous picture of the lasso and ridge regression solution path for two parameters in Figure 1.3. The solution path has elliptical contours centered around OLS estimates. The solution for ridge is the disk, while for lasso it is the diamond. The solution is obtained when the elliptical contour first hits the constraint region. As the disk has corners, lasso has one solution equal to zero. For p > 2, we can have more solutions exactly equal to zero. In case of an orthonormal design matrix, we can have the explicit solutions for lasso and ridge. On the other hand, lasso uses soft-thresholding, as depicted in Figure 1.4.

Lasso and Ridge regression both use Bayes estimates with different priors. *Tibshirani, 1996* [120] suggested that Lasso estimates can be obtained if the regression parameters have Laplace or double-exponential priors. *Park and Casella, 2008*[96] provided an easy-to-compute Gibbs sampler

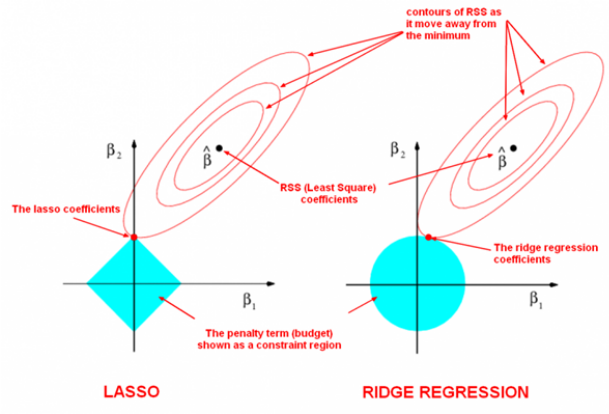


Figure 1.3: Lasso vs Ridge solution Path[a]

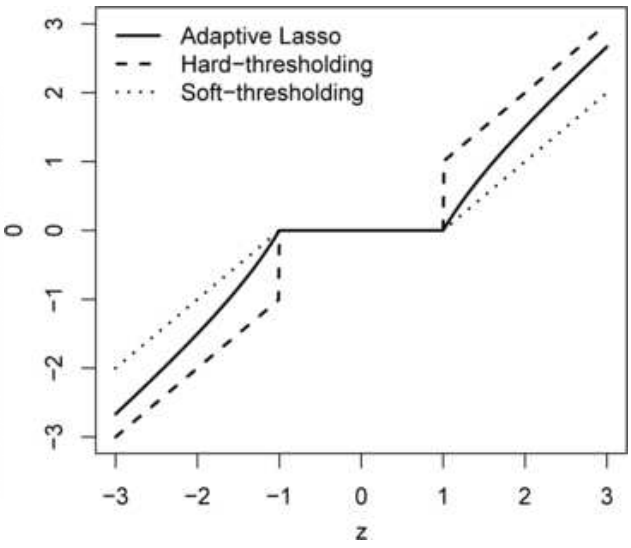


Figure 1.4: Threshold functions for orthonormal design[a]

[a] Plot obtained from on-line resources

for Bayesian lasso by introducing a scale mixture of normal priors on regression coefficients.

$$\beta|\sigma^2, \tau_1^2, .., \tau_p^2 \sim N_p(0, \sigma^2 D_\tau)$$

$$D_\tau = diag(\tau_1^2, .., \tau_p^2)$$

$$\tau_1^2, .., \tau_p^2 \sim \prod_{i=1}^{p} \frac{\lambda^2}{2} \exp(-\lambda^2 \tau_i^2/2) d\tau_i^2$$

$$\pi(\sigma^2) \sim \frac{1}{\sigma^2}$$

Although this method helps us computationally, we do not obtain exact zero estimates from the posterior median solution. Figure 1.5 comes from *Park and Casella, 2008*'s paper, where it is clearly visible that by using lasso, we can have a zero solution, in contrast to Bayesian lasso or ridge. Bayesian lasso provides very small estimates for insignificant predictors, and one needs to impose one more level of threshold to drop redundant variables from the model. This problem can be solved using a point mass spike-and-slab prior on regression coefficients. *Xu and Ghosh, 2015*[128] used this phenomenon for Bayesian Group Lasso , which we discuss in next section.

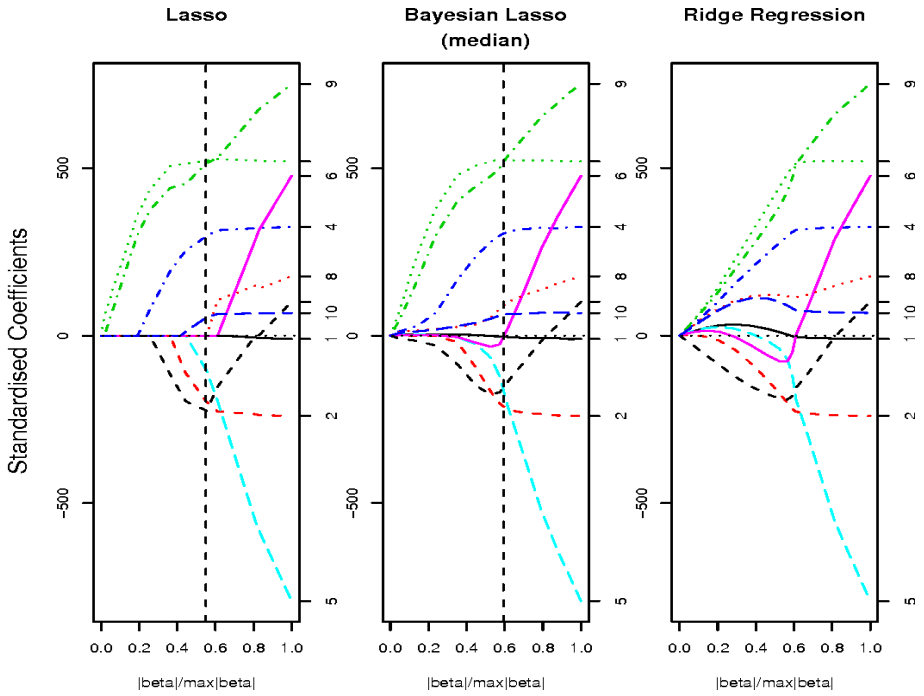


Figure 1.5: comparison between Lasso, Bayesian Lasso and Ridge regression solution paths[a]

[a] Plot obtained from on-line resources

1.1.3 Bayesian Group Lasso

Grouping structures among predictors occurs in many situations, such as when a model has categorical variables or functional predictors. The categorical predictors are represented by a group of dummy variables and, in the case of functional predictors, sometimes a group of basic functions used in the model. Gene expression data forms groups in the form of gene pathways, and these groups have a natural correlation structure. The lasso solution is not entirely satisfactory in such scenarios, as it only selects the individual predictors instead of the whole group. It also depends on how the dummy variables are orthogonalized. To deal with this problem *Yuan et al.* [132] introduced 'Group Lasso' which can select important grouped variables as a whole. Consider this model-

$$\mathbf{Y}_{n\times 1} = \sum_{g=1}^{G} \mathbf{X}_g \beta_g + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ follows normal distribution and β_g is the g-th grouped predictor among G total group variables. The group size can vary. *Yuan et al.* [132] proposed group lasso, which is an extension

of lasso, to handle grouped variables. The group lasso solution for linear regression is defined as

$$\min_{\beta_g} ||\mathbf{Y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g|| + \lambda \sum_{g=1}^{G} ||\beta_g||_2$$

In this dissertation, we deal with a generalized linear model with binary response. Hence, we are interested in how lasso and group lasso have been developed for logistic regression. The lasso penalty can be applied to logistic regression or multinomial logistic regression. *Genkin et al.* [42] applied l_1 penalty to their sparse model and used a Laplace prior to avoid overfitting. Instead of the residual sum of squares, they used a negative log-likelihood criterion and MAP estimates of the regression parameters. The first introduction of group lasso penalty for logistic regression model was observed in *Kim et al.* [66], where they proposed a gradient descent algorithm for estimation. *Meier et al.* [87] presented group lasso for a "large p" which minimizes the negative log-likelihood function with group penalty. The group lasso solution in their paper is defined as follows:

$$\min_{\beta_g} \; - \left[\sum_{i=1}^{n} y_i \sum_{g=1}^{G} \mathbf{X}_{i,g} \beta_g - \log(1 + \exp(\sum_{g=1}^{G} \mathbf{X}_{i,g} \beta_g)) \right] + \lambda \sum_{g=1}^{G} \sqrt{m_g} ||\beta_g||_2$$

where m_g is the size of β_g. They propose a block co-ordinate gradient descent algorithm to find the solution to this minimizing problem.

Despite the more common use of frequentist group lasso methods, Bayesian approaches may be more appropriate for the same problem. This is because the estimates of the standard lasso estimators do not provide meaningful standard errors that can be used for hypothesis testing or constructing confidence intervals. The lasso estimator has a complex limit distribution and is complicated to implement (*Knight et al.* [68],*Chatterjee et al.*[23]). On the other hand, a Bayesian prior-based formulation of lasso can provide reliable standard errors of estimates by obtaining the MAP estimators. In his seminal Lasso paper, *Tibshirani*[120] stated that the posterior mode with an independent double exponential prior for the regression coefficient is the same as a lasso estimator. Later, *Park and Casella*[96] developed a highly efficient Gibbs sampler for Bayesian lasso by introducing scale mixture priors for regression coefficients. The regression coefficients

follow the normal prior, and the variance components of the normal distribution follow the Gamma prior, giving us a highly efficient way of posterior sampling. Based on *Park and Casella's*[96] work, *Kyung et al.*[69] provided a Bayesian hierarchical model for Bayesian group lasso. As in penalized linear regression with grouped variables, the conditional prior β/σ^2 can be written as

$$\pi(\beta/\sigma^2) \propto \exp\left\{-\frac{\lambda}{\sigma}\sum_{g=1}^{G}||\beta_g||_2\right\}$$

Kyung et al.[69] expressed this same prior as

$$\begin{aligned}\beta_g/\tau_g^2, \sigma^2 &\sim N_{m_g}(\mathbf{0}, \tau_g^2\sigma^2 I_{m_g}) \\ \tau_g^2 &\sim Gamma(\frac{m_g+1}{2}, \frac{\lambda^2}{2}), \quad g = 1, .., G\end{aligned}$$

Li et al. 2015[75] worked with a high-dimensional varying-coefficient model equipped with Bayesian group lasso. They tried to obtain group lasso estimators as posterior mode estimates of multivariate i.i.d. Laplace priors on regression parameters.

Although it is highly computationally efficient, a significant disadvantage of *Park and Casella's*[96] normal mixture prior setup is that it does not ensure an exact zero solution for the regression coefficients. This is also the case if we obtain the posterior mean or median from the prior setup by *Kyung et al.*[69] – it will not provide exact 0 estimates for β_g's. Therefore, to impose sparsity in group level, *Xu et al.* [128] proposed a multivariate zero-inflated mixture prior for each β_g. The following is a hierarchical structure with an independent spike-and-slab prior for each β_g:

$$\begin{aligned}Y|X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I) \\ \beta_g|\tau_g^2, \sigma^2 &\sim (1-\pi_0)N_{m_g}(0, \sigma^2\tau_g^2 I_{m_g}) + \pi_0\delta_0(\beta_g), \quad g = 1, .., G \\ \tau_g^2 &\sim Gamma\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), \quad g = 1, .., G \\ \sigma^2 &\sim IG(\alpha, \gamma) \\ \pi_0 &\sim Beta(a, b)\end{aligned}$$

where $\delta_0(\beta_g)$ denotes point mass at $\mathbf{0}$. The mixing probability π_0 can be defined as a function of the number of predictors to impose more sparsity as the feature size increases. The choice of λ is very critical to control sparsity. Large values of λ will produce biased estimates and very small λ values will impose diffuse distribution for the slab part. *Xu et al.* [128] mentioned an empirical Bayes approach to estimate λ. Due to the intractability of marginal likelihood, they proposed a Monte Carlo EM (Expectation Maximization) algorithm for the estimation of λ. Moreover, they have shown theoretically and numerically that the median thresholding of posterior β_g samples provides exact zero estimates for insignificant group predictors.

The class of spike-and-slab zero-inflated mixture prior was first introduced by *Mitchell et al.*[91]. It is highly useful for variable selection. In the zero-inflated mixture prior, the slab part assumes some known distribution at nonzero coefficients, and the spike part has point mass at zero coefficients. Later *George et al.* ([43],[44]) used a zero-inflated normal mixture prior to build Gibbs sampling for variable selection. *Narisetty et al.*[94] worked with shrinking and diffusing priors. The predictors with zero or very small coefficients have variance tending to 0, as those coefficients reach point mass at zero (spike), and the active predictors' variance reaches infinity as a diffused prior (slab). On a different note, *Lykou et al.*[79] worked with Bayesian lasso variable selection; they assumed independent double exponential distribution for regression coefficients and concentrated on the shrinkage parameter λ. Bayes factors criteria are used to choose the selection vector and the shrinkage parameter. *Zhang et al.*[134] generalized this prior using Dirac spike-slab where each coefficient group follows a normal distribution, as follows:

$$\beta_g|\gamma_g, \sigma^2, \tau_g^2 \sim \gamma_g N(0, \sigma^2 D_{\tau_g}) + (1 - \gamma_g)\delta_0(\beta_g)$$

where $D_{\tau_g} = diag(\tau_{g1}^2, .., \tau_{gm_g}^2)$ and γ_g follows a Bernoulli distribution. The more interesting advancement in their work is the simultaneous selection of groups and the members within those groups. In addition, they incorporated group serial correlations with Bayesian fused lasso technique for within-group selection. *Xu et al.* [128] also proposed an algorithm for bi-level selection in group lasso with a two-way sparsity assumption: among grouped predictors and within selected groups.

The main application of this kind of selection arises in genetic association studies. For instance, when not all genetic variations in a selected gene are responsible for a disease, it is necessary to incorporate bi-level selection.

1.2 Bayesian logistic regression

Classification using longitudinal data can be challenging with a large number of predictors. The first significant approach to handle longitudinal predictors is to consider each multiple-occasion observation as a single function that is observed over a time interval. Functional predictors have a high correlation between adjacent measurements, and the observational space is high-dimensional. The number of predictors required for estimation often exceeds the number of observations, which introduces the problem of dimensionality. A regression framework is often the most suitable to model all possible longitudinal effects across ROIs, where the proposed method will select the important predictors. Moreover, many biomedical studies have shown that a limited number of specific brain regions or ROIs are essential for AD classification. Thus, dimension reduction techniques can be applied, and classification can be limited to the reduced feature set. *Zhu et al.* [136] proposed a method for classification and selection of functional predictors: first, functional principle component scores are calculated for each functional predictor; then, the functional principle component scores are used for the classification of each individual observation. They proposed using Gaussian priors for selection and developed a hybrid Metropolis-Hastings/Gibbs sampler algorithm. Although the method reported in the present study was inspired by this method, we developed a simple Gibbs sampler where MCMC samples are drawn from standard distributions. We also focused on applying penalized regression for dimension reduction. In the Bayesian variable selection literature, the spike-and-slab prior has widespread applications due to its superior selection power. *George et al.* ([43],[44]) initially proposed that each coefficient β can be modeled either from the "spike" distribution, where most of its mass is concentrated around zero, or from the "slab" distribution, which resembles a diffuse distribution. Instead of imposing the spike-and-slab prior directly on regression coefficients, *Ishwaran et al.* [57] introduced a method in which they place

a spike-and-slab prior on the variance of Gaussian priors. The Bayesian variable selection methods also include different Bayesian regularization methods, such as Bayesian lasso [96], Bayesian Group Lasso, and Bayesian elastic net [76]. We opted for spike-slab Bayesian group lasso which is extensively discussed in *Xu et al.* [128]. The group structure among the coefficients in our model comes from functional smoothing of the coefficients, and group lasso facilitates the selection of the important functional predictors. Thus, our proposed method uses the idea of Bayesian variable selection in a generalized functional linear model with binary responses.

We model binary response $y_i \in \{0, 1\}$ $(i = 1, .., n)$ and predictors with logistic regression. We know that posterior sampling of logistic regression coefficients is difficult due to the model's complicated likelihood function. In addition, we assume a Gaussian prior for regression coefficients, which makes the likelihood function analytically inconvenient. Full posterior sampling of parameters requires candidate density following the Metropolis-Hastings algorithm. Bayesian inference for a probit model is comparatively easier [2]. Different sampling algorithms motivated by *Albert et al.'s* work [2] have been proposed. For instance, *Holmes et al.* [54] presented an indirect sampling method by introducing auxiliary variables for binary and multinomial regression. Later, more methods based on latent variables for logistic regression were advanced by *Frühwirth-Schnatter et al.* [41], *Gramacy et al.* [45] and *Polson et al.* [97]. Among all these works, *Polson et al.'s* algorithm is the most interesting to us due to its ease of computational implementation and its sampling efficiency. Our aim is to avoid the complex Metropolis algorithm while sampling from posterior distributions of regression coefficients.

Holmes et al. [54] introduced an auxiliary variable to avoid the conditional non-conjugacy for

updating β. The prior structure they used is

$$y_i = \begin{cases} 1, & \text{if } z_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$z_i = \mathbf{x}_i\beta + \epsilon_i$$

$$\epsilon_i \sim N(0, \lambda_i)$$

$$\lambda_i = (2\psi_i)^2$$

$$\psi_i \sim Kolmogorv - Smirnov\ distribution$$

$$\beta \sim Normal\ distribution$$

The above prior is interesting because the marginal likelihood $\mathcal{L}(\beta|data)$ is same as the likelihood for logit model. One can achieve a conjugate full conditional distribution of β given data, but the conditional distribution of $\pi(\lambda_i|z_i, \beta)$ does not have any standard form. One has to use a complicated rejection sampling method to sample for conditional λ_i. Hence, adding an auxiliary variable does not give us significant computational improvement compared to using the Metropolis-Hastings algorithm. *Frühwirth-Schnatter et al.* [41] addressed the problem with the same approach, but instead of using a single auxiliary variable they used a two-stage augmentation method. In the first stage, they assumed the existence of a latent variable, where the binary response variable is conditional on the sign of the auxiliary variable. They assumed the error part of the model to follow a type I extreme value distribution, which is non-normal density. Then, in the second stage of data augmentation, this non-normal error distribution was approximated by the mixture of the normal distribution. Finally, they obtained a multivariate normal distribution for the posterior of β. Among all the popular methods, pólya-gamma augmentation is most interesting to us, and it is easy to apply due to the availability of the R package.

1.3 Functional data smoothing

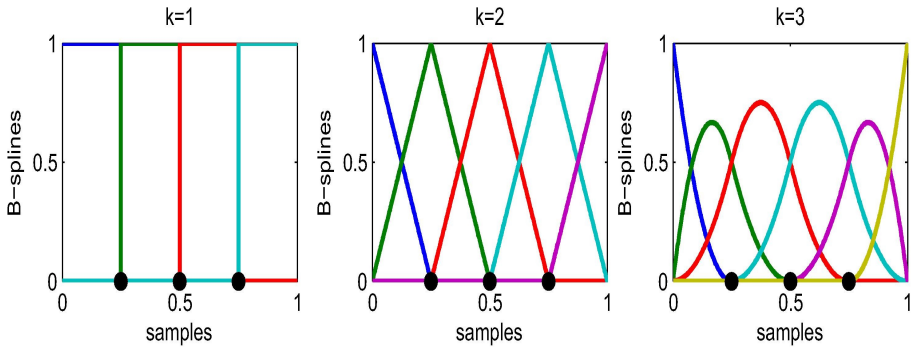
Here we discuss some of the key features of functional data or functional observations in the regression context. Throughout this thesis, we use linear combinations of basis functions to rep-

resent functions. The pivotal philosophy of functional data analysis is to consider observed data functions as single observations. Measurement error is a frequent problem in functional observations. In the following, we discuss the simplest case of nonparametric regression problem: we observe $(x_i, y_i)_{i=1,..,n}$ with $y_i = f(x_i) + \sigma\epsilon_i$ where x_i's follow a Uniform (0,1) distribution (i.i.d). To estimate $\hat{f}(x)$, we must minimize $\left[E\int_0^1(\hat{f}(x) - f(x))^2dx\right]$. We need basis functions to estimate the unknown function. Basis functions have mathematical properties that help statisticians to approximate any function by taking a linear combination. Some popular basis functions are the Haar basis, Fourier basis, Spline basis, wavelets, and polynomial bases, among others. In general, basis functions have the property of orthonormality. Now, if we have a set of basis functions ϕ_j in $L_2(0,1)$ then we can define $f_J(x) = \frac{1}{n}\sum_{j=0}^{J}\theta_j\phi_j(x)$ where $\theta_j = \int_0^1 \phi_j(x)dx$. We plug in the estimator $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} y_i\phi_j(x_i)$ such that $E(\hat{\theta}_j) = \theta_j$. The choice of basis function is crucial in functional data analysis. We focus on B-spline basis functions in this thesis, and we discuss its properties in the following.

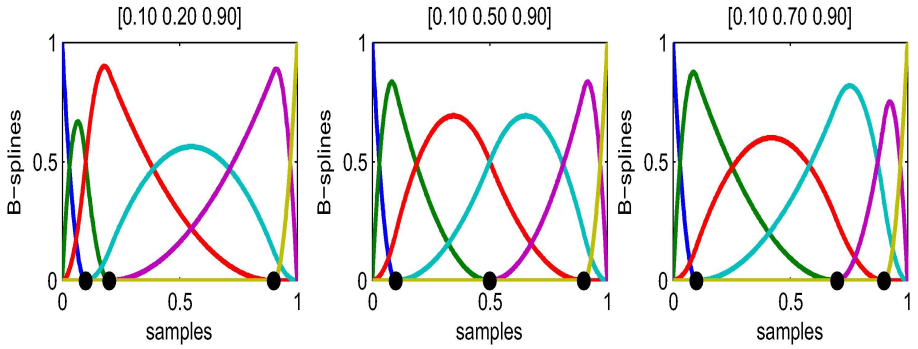
Spline functions are most often used to approximate non-periodic functional data. Splines are easy to handle computationally and provide better results than polynomial approximations. One can approximate most of the functions using a moderate number of basis functions. The first step in constructing splines in the input space is to divide the space into breakpoints, which are called knots. Let $\epsilon_0 < \epsilon_1$ and $\epsilon_K < \epsilon_{K+1}$ be two binary knots which can be defined as the range of the domain over which we approximate our function. Now, we define:

- $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_M \leq \epsilon_0$
- $\tau_{j+M} = \epsilon_j, j = 1,..,K$
- $\epsilon_{K+1} \leq \tau_{K+M+1} \cdots \leq \tau_{K+2M}$

Although the additional knots are defined outside of boundary points, in practice they are all equal to ϵ_0 and ϵ_{K+1} respectively. We define $B_{i,k}(x)$ as the i^{th} B-spline basis function of order 'k' with



(a) Fixed vector of free knots, $\delta = [0.25\ 0.50\ 0.75]$, varying B–spline order



(b) Fixed B–spline order, $k = 3$, varying vector of free knots

Figure 1.6: B-spline smoothing under varying orders and knots[a]
[a] Plot obtained from online resources

knot sequence $\boldsymbol{\tau}$ and they are constructed as follows:

$$B_{i,1}(x) = 1, \quad if \tau_i \leq x \leq T_{i+1}, \quad for\ i = 1, .., K + 2M - 1$$
$$B_{i,k}(x) = \frac{x - \tau_i}{\tau_{i+k-1} - \tau_i} B_{i,k-1}(x) + \frac{\tau_{i+k} - x}{\tau_{i+k} - \tau_{i+1}} B_{i+1,k-1}(x), \quad for\ i = 1, .., K + 2M - k$$

To understand the construction we obtained Figure 1.6 from a paper by *Dertimanis et al., 2018*[30]. Now we can replace the general $\phi_j(x)$ basis functions with $B_{i,k}(x)$ basis functions, and a function $f(x)$ can be approximated as $f(x) = \sum_{j=1}^{K} \beta_j B_j(x)$. The least square method can be used to estimate unknown βs.

Knot selection is another domain of research that statisticians have already explored. Automatic knot selection can be achieved using model selection criteria such as cross validation and Mallow's C_p in a linear regression framework. The aim is to choose a set of knots from a large number of

candidate knots such that optimization criteria are satisfied. The number of all possible models increases as the cardinality of candidate knots does. Some works to mention on this topic are by (*Smith and Kohn, 1996*[115],*Denison et al., 1997*[24],*DiMatteo et al., 2001*[24]). *Smith and Kohn, 1996*[115], *Denison et al., 1997*[24] utilized Bayesian variable selection to build MCMC methods for function estimation.

The problem of univariate function estimation can be extended to function selection in additive models. Many authors are interested in selecting significant functions for $y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i$ when $p \gg n$. *Scheipl et al., 2012*[106] proposed a normal-mixture-of-inverse gamma distributions by introducing similar indicator variable used in spike and slab prior. Their setup is for a generalized linear model with continuous response replaced by link functions. *Scheipl et al., 2013*[107] mentioned more than one method in their paper, including using penalized likelihood and smoothness priors, and an indicator selection approach after transforming functions into a linear combination of centered B-spline expansion. *Lan Xue*[129] proposed a penalized polynomial spline method for a simultaneous model estimation and variable selection in additive models. He explored spline SCAD (Smoothly Clipped Absolute Deviation) penalty as a regularization method. Although *Fan et al., 2015*[33] did not engage in variable selection in their work, they did utilize a penalized least squares optimization technique to efficiently deal with high-dimensional functional predictors. They minimized a loss function with a penalized l_2 norm of functions to find estimated functions. *McLean et al., 2013*[85] explored both MCMC and variational Bayes algorithms to fit a functional generalized additive model. *Huang et al.*[56] tackled the same problem in a conventional non-Bayesian variable selection approach by applying adaptive group lasso. To start, they also decomposed functional predictors as a linear combination of B-spline basis functions and later penalized the unknown coefficients as groups. The Bayesian group lasso we built is inspired by their work, but we deal with binary response that is equivalent to a generalized linear model.

1.4 Bayesian model based clustering

Cluster analysis is the modern data mining tool used to group or segment objects into clusters based on similarity measurements. Many available algorithms have been built based on a particular criterion that objects within a cluster are more similar to each other than to objects that belong to other clusters. Some popular clustering methods are K-means, agglomerative clustering, hierarchical clustering, and model-based clustering. We focus more on model-based clustering in this chapter. A disadvantage of hierarchical clustering or K-means methods is that they are model-free heuristic methods. Model-based clustering is an alternative to provide an option to formulate a model while grouping objects. Model-based clustering was first introduced *Banfield and Raftery, 1993*[6], who assumed that observations come from multivariate normal distribution and that maximum likelihood yields the estimates . *Fraley and Raftery, 2002*[37] introduced the finite mixture models as a formal setting for model based clustering. Let $y_1, .., y_n$ are independently distributed p-dimensional observations from a K-component mixture distribution-

$$f(\boldsymbol{y}; \boldsymbol{\tau}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f_k(y_i|\theta_k) \tag{1.1}$$

Here τ_k is the probability that a particular object belongs to k^{th} cluster and θ_k parametrizes the density f_k. In general, f_k is multivariate normal density where $f_k \equiv MVN(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$. The EM algorithm helps us to get the estimates of unknown parameters. In a later study, *Melnykov and Maitra, 2010*[88] provided a thorough reference for model-based clustering. A point of subjectivity that is always attached to cluster analysis is choosing the number of clusters. Similar to K-means, model-based clustering needs a specified number of clusters at the start of the algorithm. *Fraley and Raftery, 1998*[36] tried to answer this question using model selection criteria. They suggested studying BIC over all possible cluster numbers and choosing the one with lowest BIC. Another study to mention is that of *Yeung et al., 2001* [131], who applied model-based clustering to group gene expression data. Over time, researchers included other factors inside the model-based approach, such as variable selection[101] and penalizing the parameter space θ[95]. All the aforementioned studies have been done in a conventional statistical framework without using any prior information

on the cluster structure, the number of clusters, or unknown parameters (τ, θ). Only Bayesian model selection has been used to select optimal number of clusters. *Handcock et al., 2007*[46] described two methods to estimate unknown parameters: first, a two-stage maximum likelihood method, and second, a Bayesian estimation method. In the Bayesian part, regression parameters have a multivariate normal prior, and the mixture probability τ has a Dirichlet prior with MCMC estimation. Furthermore, *Fraley and Raftery, 2007*[38] discussed the problems of using the EM algorithm and where it can fail to converge. Instead of MLE, they preferred to work with the MAP estimator with conventional priors on mean and covariance parameters, such as MVN on mean and an Inverse-Wishart distribution on a covariance matrix. *Medvedovic et al., 2004* [86] proposed a contrasting formulation of a Bayesian mixture-based clustering algorithm to group gene expression data with replicates. Interestingly, the Poisson-Dirichlet process is the most logical and convenient way of assuming priors on the number of clusters and mixture probability if one does not want to limit the number of clusters when starting the algorithm [70].

The methods and literature described above have various advantages and can be applied to different complicated problems faced by researchers. Here we should reiterate the problem with which we are dealing: we aim to build a Bayesian model-based clustering method with the same white matter data used for spatiotemporal modeling discussed in Section 5.1. We have longitudinal voxel-wise white matter measurements for healthy aging subjects, and we would like to group voxels in homogeneous regions considering the longitudinal and spatial information. To this end, we review the literature for spatiotemporal clustering methods. Authors in the computer science literature have examined this topic extensively. *Kalnis et al., 2005*[65] dealt with a highly complicated problem of detecting clusters among moving objects that change locations over time. For instance, they clustered trajectories and mining movement patterns for a group of migrating animals or a convoy of cars moving in a city. The complexity of this problem is manifold, as we are simplifying our situation by assuming that spatial locations of voxels are fixed over time. *Kisilevich et al., 2009*[67] described a detailed study of spatiotemporal clustering on trajectories and provided

in-depth research development on this topic.

1.5 ADNI Longitudinal Data

Alzheimer's disease (AD) is the most common form of age-related neurodegeneration and dementia. More than 25 million people in the world are currently affected by dementia, with most suffering from AD, and around 5 million new cases occur every year. The number of people with dementia is anticipated to double every 20 years [99]. Some suggest that earlier diagnosis and intervention may offer the greatest potential for treatment, making the early detection of AD of the utmost clinical importance. AD symptoms are diagnosed via clinical neuropsychological and cognitive measures, including the Clinical Dementia Rating (CDR) scale. In addition to other standardized clinical neuropsychological measures, such as the Ray Auditory Verbal Learning Task or the Mini Mental Status Examination (MMSE), biomarkers from MRI and PET imaging are also used in diagnostic classification. Based in part on these scores, patients are diagnosed or classified into one of three primary categories: cognitively normal (CN), mild cognitive impairment (MCI), and AD. In this dissertation, we demonstrate a method to differentiate AD patients from CN persons via the differential nonlinear longitudinal trajectories in regional brain volumes. Various high-dimensional classification and regression methods have been proposed for biostatistical applications to neurological diagnostic methods for early-stage AD detection. The present study focuses on applying volumetric MRI data at the region of interest (ROI) level, and limiting the included features to more clinically established sub-regions of brain volumes in AD and aging. Unlike commonly used Bayesian classification methods, we emphasize the selection of the longitudinal trajectories of brain volumes as predictors for classification of patients as AD vs. CN.

MRI is a valuable tool for in vivo assessment of brain biomarkers related to disease progression. Longitudinal patterns of brain atrophy follow a common pattern, with prominent volumetric loss in established and canonical regions, such as the hippocampus or the entorhinal cortex, as well as ventricular and sulcal expansion. The currently available, fully processed ADNI longitudinal MRI data includes cortical and white matter (WM) parcellation, surface area, and cortical thickness of

different brain regions. Instead of using all possible regional measurements, we focus on volumetric longitudinal changes in a more theoretically guided and limited number of brain sub-regions to classify patients. We obtained nearly 115 different sub-regions' volume measurements from ADNI 1, ADNI 2, & ADNI GO. *Leung et al.* [74] compared MCI converters and MCI nonconverters considering longitudinal hippocampal volumes, and showed that MCI converters have higher rates of hippocampal atrophy. Reduced hippocampal volume has been proposed to diagnose AD earlier than clinical diagnosis [40]. We believe that measuring volumetric change over multiple occasions rather than at a single time-point will provide a more reliable estimate of quantitative change. In this dissertation, we concentrate on the selection of key brain regions that differ in longitudinal change between AD patients and CN older adults using a Bayesian variable selection method. In addition to prior findings related to the selection of volumetric MRI measurements, there are also numerous studies involving the selection of functional predictors. The key advantage of using longitudinal MRI data is that it captures the atrophy rates with time as patients' progress through multiple disease stages. FreeSurfer software is able to measure accurate brain volumes and cortical thickness, and the ADNI website has 1.5T structural MRI data with parcellation performed by FreeSurfer.

The pivotal contribution of this dissertation is that it explores functional data analysis with Bayesian variable selection. We extensively refer to the methodologies described in this introduction in the development of later chapters. Chapter 2 introduces the applicability of spike-and-slab priors for knot selection in function estimation. The research interest in function estimation is still prevalent in the statistics literature. Spike-and-slab group lasso, which was discussed earlier, works well for group selection and outperforms other methodologies. Chapter 4 presents extensive application of Bayesian group lasso to function selection. Moreover, it helps to establish some statistical properties. In addition, Bayesian classification is another field that is discussed later in this dissertation. The aim of this introduction was to help the reader better understand the perspective of this dissertation.

CHAPTER 2

NONPARAMETRIC FUNCTION ESTIMATION USING BAYESIAN VARIABLE SELECTION

Bayesian variable selection for nonparametric function estimation is a challenging research field. We have tried to explore it with a large number of functional predictors in an additive model. We started with the knot selection problem for a single function estimation.

2.1 Univariate function estimation

The literature on function estimation is quite old and has already established excellent mathematical properties. Bayesian methods have worked well for estimating functions with a penalized spline. However, we propose a new approach to function estimation by assuming uncertainty with knots. We describe our methodology in the following.

2.1.1 Introduction

Consider a very simple regression problem with data observed as $(y_i, X_i)_{i=1,..,n}$ where X is a single covariate and Y is $n \times 1$ response variable. We regard a nonlinear model as a potential setup, such that the mean $E(y|X)$ varies nonlinearly with the predictor. Then, we can replace the conventional regression problem of $y_i = X_i\beta + \epsilon_i$ with

$$y_i = f(X_i) + \epsilon_i, \;\; i = 1, .., n$$

where $f(.)$ belongs to some class of nonlinear functions. There are numerous methods available to estimate the unknown function, one of which is basis function expansion. This single covariate has a nonlinear component with respect to response, which is modeled as a regression spline basis and can be written as

$$f(x) = \sum_{j=1}^{K} \beta_j \phi_j(x) \tag{2.1}$$

where $\phi_j(x)$ are some orthonormal basis functions. Now, if we limit our study only for spline functions, the value 'K' depends on the number of internal knots we are using. Let us assume that we have $\tau_1, .., \tau_K$ internal knots to approximate the function. Then, equation (1.1) can be written with p^{th} degree spline model as

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 X + \cdots + \beta_p X^p + \sum_{j=1}^{K} \beta_{pj} (X - \tau_j)_+^p \\ &= \sum_{j=1}^{K+p+1} \beta_j \phi_j(x) \end{aligned}$$

here $(X - \tau_j)_+ = \max(0, (x - \tau_j))$ and as $(X - \tau_j)_+^p$ has $p - 1$ continuous derivatives, the smoothness of the spline function increases with a value of p. We work with cubic spline bases here.

To start the statistical problem, we assume that we have a large number of candidate knots, and we would like to build a data-driven algorithm that automatically selects important knots to accurately approximate a given function. Knots' locations are another point of interest. Various model selection criteria have been used before, such as cross-validation and Mallow's C_p etc. The number of possible models increases as the cardinality of candidate knots expands. There is the potential to over-fit the data if we use a large number of knots. One advantage of using spline bases is that the design matrix for the linear model is sparse and we can explore the variable selection methods in computation. We examine Bayesian variable selection as a potential method for knot selection regarding this problem.

Our method implicitly assumes that f is well approximated by a linear combination of spline basis functions with some number of knots. In practice, we assume that f can be represented by a linear combination of basis functions. This class of basis splines is large and approximates any locally smooth function arbitrarily well. We note that our approach determines which knots are effective in estimating the spline function, and we closely follow the works on variable selection in linear regression models by *George and McCulloch, 1993,1997*[43][44]. These authors first

introduced the spike-and-slab prior with the following setup:

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2)$$

where γ_i is a latent variable that controls the number of significant coefficients or model size. *Smith and Kohn ,1996*[115] implemented another method of g-priors on the regression coefficients and applied this prior in univariate function estimation. On the other hand *DiMatteo et al., 2001*[31] quantified two unknown quantities, such as number of knots and location of knots with Poisson prior and Dirichlet prior, respectively.

We focus on *George and McCulloch, 1993,1997*[43][44]'s approach in this section, as the conventional spike-and-slab prior is not considered for function estimation. The Gibbs sampler is computationally feasible and easy to implement when the model has a large number of predictors. Once we integrate corresponding variance and regression coefficients, we show that the Bayesian variable selection convergence rate is much faster than with other approaches.

2.1.2 Variable selection using Spike-Slab Prior

A spike-and-slab model is often represented by a Bayesian hierarchical model. The most conventional definition is given below:

$$\begin{aligned}
(Y_i/x_i, \beta, \sigma^2) &\overset{ind}{\sim} N(x_i'\beta, \sigma^2), \quad (i = 1, ..., n) \\
(\beta/\gamma) &\sim N(\mathbf{0}, \Gamma), \\
\gamma &\sim \pi(d\gamma), \\
\sigma^2 &\sim \mu(d\sigma^2),
\end{aligned}$$

where $\mathbf{0}$ is a K-dimensional zero vector, Γ is the $K \times K$ diagonal matrix $\text{diag}(\gamma_1, ..., \gamma_K)$, π is the prior measure for $\gamma = (\gamma_1, ..., \gamma_K)^t$ and μ is the prior measure for σ^2.

George and McCulloch(1993)[43] first developed the most widely used spike-and-slab model version. It identifies zero and nonzero β_i's by using zero-one indicator variables γ_i, assuming a

scale mixture of two normal distributions:

$$(\beta_i/\gamma_i) \overset{ind}{\sim} (\mathbf{1}-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2), \quad i = 1,..,K$$

$\tau_i^2 > 0$ is some suitably small value, while $c_i > 0$ is some suitably large value. $\gamma_i = 1$ represents the β_i's that are significant, and the variances of these coefficients are large, with larger posterior β_i values. The opposite occurs when $\gamma_i = 0$. The prior hierarchy for $\boldsymbol{\beta}$ is completed by assuming a prior for γ_i. When τ_i^2 tends to zero we provide more masses on 0 which operates as the prior for insignificant βs. The prior distribution for the regression coefficients then can be written as

$$(\beta_i/\gamma_i) \overset{ind}{\sim} (\mathbf{1}-\gamma_i)I_0 + \gamma_i N(0,\nu^2) \quad (*)$$

with I_0 point mass at 0 coefficients, where ν^2 is the limit for $c_i^2\tau_i^2$ when τ_i^2 tends to zero with a large enough c_i^2. We use prior (*) to select significant knots from candidate knots.

2.1.2.1 Model description and Prior

We start with the simple linear model:

$$y = X\beta + \varepsilon$$

where y is the $n \times 1$ vector of observations. X is the $n \times r$ design matrix,
$\varepsilon \sim N(0, \sigma^2 I_n)$ is the error vector, and $\beta = (\beta_1, ..., \beta_r)'$ is the $rx1$ vector of regression coefficients. Let γ be the $rx1$ vector of indicator variable with i-th element γ_i such that:

$$\gamma_i = \left\{ \begin{array}{ll} 1, & \text{when } \beta_i \neq 0 \\ 0, & \text{when } \beta_i = 0 \end{array} \right\}$$

Given γ, let β_γ consist of all the nonzero coefficients of β and let X_γ be the columns of X corresponding to particular components of γ that are equal to one. We then make the following prior assumptions:

1. Given γ and σ^2, the prior for $\beta_\gamma \sim N(0, c\sigma^2(X_\gamma' X_\gamma)^{-1})$, where c is a positive scale factor. For our simulation study we used c=100. The above formulation of βs is identical to the spike-and-slab

prior defined in equation (*).For large values of c, the prior for β_γ contains very little information about β_γ compared to the likelihood. We take the prior variance of β_γ proportional to $\sigma^2(X'_\gamma X_\gamma)^{-1}$ which makes the Gibbs sampler very fast. We also think this prior contains the associative relationships of βs.

2. The γ_i are assumed to be a priori independent with

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i \quad (i = 1, ..., r)$$

$p(\boldsymbol{\gamma}/\mathbf{Y}) \propto p(\mathbf{Y}/\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ contains information relevant to variable selection. We started with $p_i = \frac{1}{2}$ which justifies no bias towards model size.

3. Finally, for σ^2 we use the inverse gamma conjugate prior

$$p(\sigma^2/\boldsymbol{\gamma}) \sim IG(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2),$$

We note that ν_γ and λ_γ may depend on $\boldsymbol{\gamma}$ to incorporate dependence between $\boldsymbol{\beta}$ and σ^2. For our practical example, we choose $\nu_\gamma \equiv 0$ and $\lambda_\gamma \equiv 0$ which gives us a commonly used prior for σ^2: $p(\sigma^2/\boldsymbol{\gamma}) \propto \frac{1}{\sigma^2}$.

2.1.2.2 The Gibbs Sampler

The primary advantage of applying a conjugate hierarchical setup is that it enables analytical marginalizing of γ with respect to β_γ and σ^2 from
$p(\beta_\gamma, \sigma^2, \gamma/\mathbf{Y}) \propto p(\mathbf{Y}/\beta_\gamma, \sigma^2)p(\beta_\gamma/\sigma^2, \gamma)p(\sigma^2)p(\gamma)$. For a given γ, let $q_\gamma = \sum_{i=1}^{r} \gamma_i$ be the number of nonzero elements of β and

$$S(\gamma) = y'y - \frac{c}{1+c} y'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma y$$

Then,

$$p(y/\gamma) \propto \int_{\sigma} \left\{ \int_{\beta} p(y/\beta_{\gamma}, \sigma^2) p(\beta_{\gamma}/\sigma^2)\, d\beta_{\gamma} \right\} p(\sigma^2)\, d\sigma^2$$
$$\propto (1+c)^{-\frac{q_{\gamma}}{2}} S(\gamma)^{-\frac{n}{2}};$$

β_{γ} is integrated out as a normal integral and σ^2 is integrated out as an inverse gamma integral. The posterior distribution of γ is

$$p(\gamma/y) \propto p(y/\gamma)p(\gamma)$$
$$\propto (1+c)^{-q_{\gamma}/2} S(\gamma) \prod_{i=1}^{r} p_i^{\gamma_i} (1-p_i)^{1-\gamma_i}$$
$$\equiv g(\boldsymbol{\gamma}),$$

The marginalizing constant can be obtained by computing $g(\gamma)$ for all γ values. To find an efficient posterior sample, we do not immediately need normalizing constant.

Gibbs Sampler:

We will start with an initial sequence of $\gamma^0 = (\gamma_1^0, .., \gamma_r^0)$ and then we draw each member of γ from the Bernoulli distribution with probability:

$$p(\gamma_i = 1 | y, \gamma_{j \neq i}) = \frac{1}{1+h},\ where$$

$$h = \frac{1-\pi_i}{\pi_i}(1+c)^{\frac{1}{2}} \left(\frac{S(\gamma^1)}{S(\gamma^0)} \right)^{\frac{n}{2}},\ where$$

$\gamma^1 = (\gamma_1, .., \gamma_{i-1}, \gamma_i = 1, .., \gamma_r)$ and $\gamma^0 = (\gamma_1, .., \gamma_{i-1}, \gamma_i = 0, .., \gamma_r)$. We repeat this step for a warm up period and sampling period until the sequence converges.

For large r, the above method takes a long time to execute, and computation complexity is $O(r^3)$. However, MCMC chain still can still be used to find high-probability γ values. We can construct

easy MCMC algorithms to simulate a Markov chain

$$\gamma^1, ..., \gamma^K$$

which converges in distribution to $p(\gamma/y)$.

Metropolis-Hastings algorithms:

Another way of sampling from $g(\gamma)$ is to use Metropolis-Hastings (MH) algorithms. We start with a proposal density $q(\gamma^0, \gamma^1)$, which is also called candidate density. For each γ^0 value, $q(\gamma^0, \gamma^1)$ is a probability distribution over γ^1 values. Once we determine the candidate proposal density $q(\gamma^0, \gamma^1)$, the below algorithm populates posterior sample of γ:

1. Generate γ^{new} with probability $q(\gamma^0, \gamma^1)$.
2. Set $\gamma^{(j+1)} = \gamma^{new}$ with probability

$$\alpha^{MH}(\gamma^{(j)}, \gamma^{new}) = min\left\{\frac{q(\gamma^{new}, \gamma^{(j)})}{q(\gamma^{(j)}, \gamma^{new})}\frac{g(\gamma^{new})}{g(\gamma^{(j)})}, 1\right\}$$

$$= 0, \quad \text{Otherwise}$$

Under weak conditions on $q(\gamma^0, \gamma^1)$, the sequence obtained by this algorithm will be a Markov chain that converges to $p(\gamma/y)$.

In case of symmetric proposal density, Metropolis-Hastings becomes simple Metropolis algorithm, i.e. if $q(\gamma^0, \gamma^1)$ is symmetric. Then the above α^{MH} simplifies to

$$\alpha^{M}(\gamma^{(j)}, \gamma^{new}) = min\left\{\frac{g(\gamma^{new})}{g(\gamma^{(j)})}, 1\right\}$$

One of the simple proposal density is

$$q(\gamma^0, \gamma^1) = \frac{1}{r}, \quad \text{if} \quad \sum_{i=1}^{r} |\gamma_i^0 - \gamma_i^1| = 1$$

Which alters the above MH algorithm-

1. Generate γ^{new} by changing one element of $\gamma^{(j)}$.

2. Set $\gamma^{(j+1)} = \gamma^{new}$ with probability $\alpha^M(\gamma^{(j)}, \gamma^{new})$. Otherwise, $\gamma^{(j+1)} = \gamma^{(j)}$.

This algorithm was proposed in *Raftery, Madigan and Hoeting (1993)*[53].

From the above algorithm, we estimate γ with the highest probability of occurrence. We can use the estimated $\hat{\gamma}$ to obtain the estimates of γ and σ^2. If we plug this $\hat{\gamma}$ in the model and select only those covariates for which the γ_is are 1, estimation of β and σ^2 becomes much easier. This is a huge advantage of using the spike-and-slab prior, and it is the main motivation behind using it.

2.1.2.3 Updating β and σ^2

We use *George and McCulloch's (1993)*[43] SSVS procedure to simulate a full parameter sequence. This method is based on the Gibbs sampler. At each step, we use the full conditional distribution of β and σ^2 to generate new samples.

Let $\hat{\gamma}$ be the final posterior estimate of γ:

$$p(y/\beta_{\hat{\gamma}}, \sigma^2)\, p(\beta_{\hat{\gamma}}/\sigma^2)\, p(\sigma^2) \propto \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp(-\frac{1}{2\sigma^2}(y - X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(y - X_{\hat{\gamma}}\beta_{\hat{\gamma}}))$$

$$\times \left(\frac{1}{\sigma^2}\right)^{q_{\hat{\gamma}}/2} \exp(-\frac{1}{2c\sigma^2}(X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(X_{\hat{\gamma}}\beta_{\hat{\gamma}})) \times \left(\frac{1}{\sigma^2}\right)$$

The full conditional for $\beta_{\hat{\gamma}}$ is multivariate normal with mean $A^{-1}X_{\hat{\gamma}}\boldsymbol{y}$ and variance $\sigma^2 A^{-1}$, where $A = (1 + \frac{1}{c})X'_{\hat{\gamma}}X_{\hat{\gamma}}$. Symbolically,

$$p(\beta_{\hat{\gamma}}/y, \sigma^2) \sim N(A^{-1}X_{\hat{\gamma}}\boldsymbol{y}, \sigma^2 A^{-1})$$

The full conditional distribution of σ^2 is inverse gamma: The terms in the joint distribution involving σ^2 are

$$\left(\sigma^2\right)^{-(\frac{n}{2}+\frac{q_{\hat{\gamma}}}{2}+1)} \exp(-\frac{1}{2\sigma^2}((y - X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(y - X_{\hat{\gamma}}\beta_{\hat{\gamma}}) + \frac{1}{c}(X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(X_{\hat{\gamma}}\beta_{\hat{\gamma}}))$$

so σ^2 is conditionally inverse gamma with shape parameter $\frac{n}{2} + \frac{q_{\hat{\gamma}}}{2}$ and scale parameter $\frac{1}{2}((y - X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(y - X_{\hat{\gamma}}\beta_{\hat{\gamma}}) + \frac{1}{c}(X_{\hat{\gamma}}\beta_{\hat{\gamma}})'(X_{\hat{\gamma}}\beta_{\hat{\gamma}}))$.

The Gibbs sampler cyclically samples from the distributions of $\boldsymbol{\beta}$ and σ^2 conditional on the current values of the another parameter. Finally, we obtain the estimates of β and σ^2 from the posterior median.

2.1.3 Regression Splines

Let's assume,

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, ..., n$$

where y_i is the i-th response, error part ϵ_i follows $N(0, \sigma^2)$, and $f(x)$ is the unknown smooth function that needs to be estimated. Under the suitable smoothness assumption, we propose to approximate $f(x)$ using a normalized B-spline basis. One of the research aims involved here is to quantify the location and the number of knots. Badly placed knots will definitely miss the function's properties and provide bad approximations. On the other hand, if we use a large number of knots, we will definitely have low bias but very high variance. The most convenient way to approach this problem is to introduce variable selection. One can assume that there are initially a large number of candidate knots and can then select a small set of significant knots using variable selection (e.g., *Friedman and Silverman, 1989*[39]). Now we can visualize the problem as a potential variable selection problem, as columns of the design matrix correspond to individual knots. This is what guides us to use the spike-and-slab prior. In a univariate case, we place knots in at least a quarter of the positions, depending on the number of observations, but we try to stay below 40 knots. We

want to ensure that placed knots capture the curvature of functions. For our simulation study, we place knots after every fourth sorted observation.

2.1.3.1 Simulated example

In this section we studied performance of Bayesian nonparametric estimators using simulated data. We used these functions in the simulation:

1. $f_1(x) = \sin(2x) + 2\exp(-16x^2)$
2. $f_2(x) = (\phi(x, 0.15, 0.05) + \phi(x, 0.6, 0.2))/4$

which is a nonlinear function. A well-judged nonparametric kernel-based method can estimate these functions. We approximate these functions with a cubic B-spline basis.

One hundred response observations are generated from $N(f(x), \sigma^2)$ with $x_i \sim U(0,1)$ and $e_i \sim N(0, \sigma^2)$. The choice of σ is made such that the signal to noise ratio $\frac{sd(f)}{sd(\epsilon)}$ stays around 3. The number of initial knots is directly proportional to the number of observations such that when n increases, so does the number of knots. For our simulation, we start with $\frac{n}{4}$ number of , which implies that as n increases, there is more sparsity. We place knots at every fourth position, which generates a total of r=28 columns in the design matrix X with a B-spline of degree=3. The Gibbs sampler is run on 1,600 iterations, with the sampling starting after the 100th iteration.

If $\hat{f}(x)$ is the estimated unknown function, then we need some statistic to measure accuracy of estimation. The integrated squared error (ISE)=$\int_x \{f(x) - \hat{f}(x)\}^2 dx$ over the unit interval is the statistic on which we focus. We created 400 equally spaced grids $z_i = i/400, \quad i = 1, ..., 400$ with

$$ISE = \frac{1}{400} \sum_{i=1}^{400} \left(f(z_i) - \hat{f}(z_i) \right)^2$$

We repeat the simulation 100 times. ISE results are shown in the below table and compared to the kernel method of estimation.

The simulation result with Bayesian variable selection show better performance than kernel

Table 2.1: Log(ISE) comparison with Kernel methods of approximation

	r	Bayes Mode log(ISE) Mean (sd)	Kernel log(ISE) Mean (sd)
$f_1(x)$	28	-2.7 (0.05)	-1.28 (0.03)
$f_2(x)$	28	-0.92 (0.02)	-0.49 (0.01)

estimation. We use Gaussian kernel with 0.10 bandwidth. Next, we illustrate the above method using 'Chloride' data from *Bates and Watts, 1988*[8].

2.1.3.2 Chloride concentration Data

The Chloride data contains 54 observations of concentrations of chloride taken over an interval of time. We implement the same variable selection algorithm for knot selection discussed above, and we compare it to fixed knot basis approximation and linear regression estimates, as shown in Figure 1.7. We can state that B-spline basis approximation with four internal knots works better than the Bayesian variable selection method. We obtain 0.23 of mean log(ISE) and posterior mode selects 11 basis columns out of 17 basis functions.

2.1.3.3 Discussion:

The main drawback of this method is its sensitivity to the initial choice of basis. We need to account for the uncertainty in selection based on the properties of the corresponding basis function. Here we work with cubic B-spline basis functions, which are assumed to be smooth. However, this assumption is not always valid when dealing with spike functions. In these situations, one requires prior information or needs to work with the Bayesian penalized spline method. Hence, prior information should be given on both the basis choice and the corresponding basis coefficients.

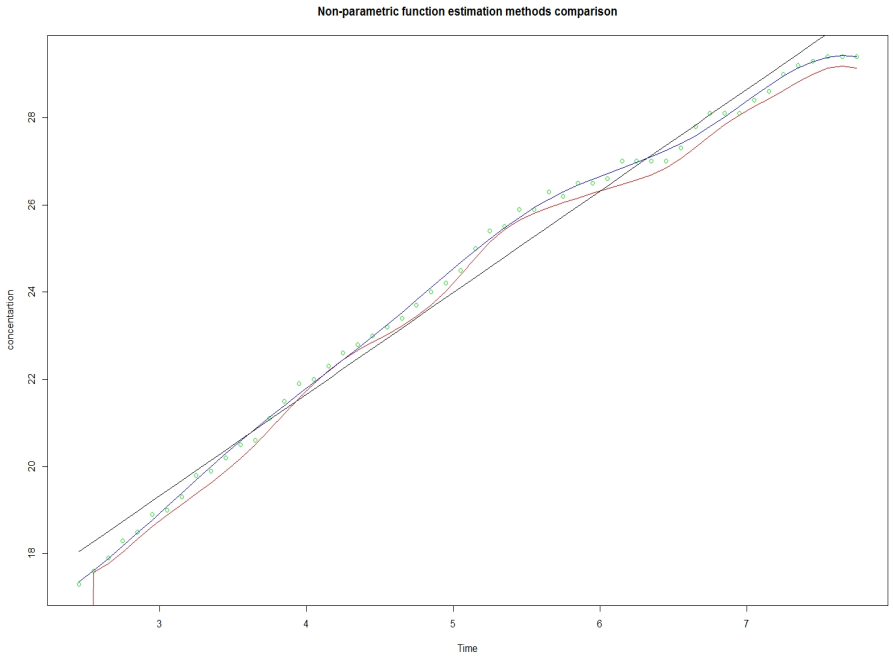


Figure 2.1: Nonparametric function estimation comparison between variable selection, linear regression, and fixed knot approximation

2.2 Functional additive model estimation with bi-level selection

We examine a different problem in this section: we seek to explore how Bayesian variable selection can perform in function selection in additive models.

2.2.1 Introduction

Modern-day life sciences develop applied questions regarding technological innovations for the selection of influential components in regression models. The statistics literature is full of feature selection methods and their corresponding theories in linear models where the number of predictors is very large compared to the sample size. The literature is rich in terms of variable selection that has been developed for generalized linear models, hazard rate models, and so forth. On the other hand, the selection of significant functions in additive models is very recent. Consider the nonparametric additive model

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i,$$

where μ is the intercept term, X_{ij} is i^{th} observation of j^{th} covariate X_j, f_js are unknown functions and ϵ_i is an unobserved random variable with mean 0 and variance σ^2.

The questions we aim to answer are: Are all $f_j(X_j)$'s significant? And based on the significant ones, can we develop an algorithm that handles the feature selection and estimates the true functions simultaneously? We assume that not all $f_j(X_j)$'s are important and that we are dealing with a much larger number of functions than the sample size, i.e . p is larger than n. We further assume that the number of true functions in a model is still less than n. The estimation of true functions is an important component of this article, whether the true functions are linear or nonlinear.

Both frequentist and Bayesian statisticians deal with model selection or function selection in additive models. Most of the ideas have come from variable selection in regression models with a large number of predictors. The most popular methods of variable selection are lasso [Tibshirani, 1996][120], SCAD penalty [Fan and Li, 2001][32], adaptive LASSO [Zou, 2006][138],

group LASSO [Yuan and Lin,2006][132]. Bayesian variable selection methods, particularly for linear models, can be divided into two segments: one based on the spike-and-slab prior, and another that introduces selection indicators with the regression coefficients. George and McCulloch [1193,1997][43][44] proposed the basic idea that each coefficient β_j's can be modeled either from the "spike" distribution, where most of its mass is concentrated around zero, or from the "slab" distribution, which is like a diffuse distribution. The introduction of selection indicators for variable selection is a closely related idea, where we decide whether a coefficient is in the model or not by using selection indicators. Smith and Kohn [1996][115] developed this notion by adding selection indicators with the columns of the design matrix which eventually selects the significant covariates. Furthermore, unlike George and McCulloch's spike-and-slab, which focuses on β coefficient, Ishwaran and Rao [2005][57] placed a spike-and-slab prior on the variance of Gaussian priors.

The two ideas of Bayesian variable selection also extend to group selection. In their paper, Ghosh and Xu [2015][128] described how spike-and-slab priors can be used to select groups in covariates as a form of Bayesian group lasso. Our method of Bayesian function selection implements Ghosh and Xu's [2015][128] ideas. First, we represent each function in our model by B-spline basis function coefficients with a subsequently large number of knots placed in each function. Next, we introduce two-stage spike-and-slab priors to find the significant functions and the significant knots for true functions. In both situations, we assume that the number of functions in the model is large, that they are sparse (large p, small n), and that the number of knots introduced to estimate the functions are also sparse. Chen et al. [2016][24] published a paper on Bayesian sparse group selection using selection indicators, which is an extension of variable selection in grouping structures.

In the next section, we describe the construction of functions in terms of B-spline basis functions, and the application of Bayesian sparse group lasso with a spike-and-slab prior for selecting significant functions and estimating in "large p, small n" settings. The posterior median thresholding, as shown in Ghosh and Xu [2015][128], estimates both the null functions and their members

as zero. As we use a large number of internal knots to estimate the functions, the spike-and-slab prior with median thresholding estimates the insignificant knots as zero based on the structure of the true functions. Section 1.2.3 presents simulation studies to evaluate the performance of our method.

2.2.2 Bayesian Sparse group Lasso in nonparametric additive models

Suppose we have data (Y, X_j), $j = 1, .., p$; and $X_j \in [a, b]$ with $a, b < \infty$. To hold the identifiable conditions for functions f_j, we assume that f_j's are centered around zero, i.e. $E f_j(X_j) = 0,\ 1 \leq j \leq p$. For the model $Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i$, after centering covariates and responses, the intercept term can be dropped for simplicity. Corresponding nonlinear components can be transformed into a linear regression model with spline coefficients. Thus, the problem of detecting a significant function in the additive model is equivalent to selecting groups of variables in a linear regression setup with a predefined grouping structure. We hence define each function f_j through basis functions.

We define the construction of B-splines as follows. Let d denote the degree of the B-spline, which implies the order as $d+1$. Our next assumption is that a sequence of knots is placed in each function where the number of knots is large but does not exceed the sample size; as sample size increases, so does the number of knots. To define the sequence of knots, let $a = \xi_0 < \xi_1 < \cdots \xi_K < \xi_{K+1} = b$ be a partition of $[a, b]$ into K sub intervals. In addition, define d knots $\xi_{-d} = \xi_{-d+1} = .. = \xi_{-1} = \xi_0$ and another set of d knots $\xi_{K+1} = \xi_{K+2} = .. = \xi_{K+d+1}$. The B-spline basis functions are defined as

$$B_{i,1}(x) = \begin{cases} 1, & \xi_i \leq x < \xi_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$$B_{i,d+1}(x) = \frac{x - \xi_i}{\xi_{i+d} - \xi_i} B_{i,d}(x) + \frac{\xi_{i+d+1} - x}{\xi_{i+d+1} - \xi_{i+1}} B_{i+1,d}(x),$$

for $j = -d, .., K$. With the use of additional knots we will get precisely $K + d + 1$ basis functions. Thus for any function f_j, defined on the function space $\mathscr{F}_j$, can be written as

$$f_j(x) = \sum_{k=1}^{m_n} \beta_{jk} B_{jk}(x)$$

and the vector $\mathbf{f}_j$ of function can be expressed as

$$\mathbf{f}_j = \mathbf{X}_j \boldsymbol{\beta}_j$$

with basis function values $B_{jk}(X_{ij})$ as elements of the design matrix $\mathbf{X}$. Finally the model can be written as

$$\mathbf{Y} = \sum_{j=1}^{p} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}_j$'s are the grouped coefficients corresponding to the basis functions $\mathbf{X}_j$. The error $\boldsymbol{\epsilon}$ has mean 0 and variance σ^2.

We now introduce the method of Bayesian sparse group selection with a spike-and-slab prior. The importance of this method is that we assume sparsity at a number of functions and also at a number of knots within a function. Simon et al. [2013][113] proposed sparse group lasso to produce exact 0 coefficients at the group level and within a group. The sparse group lasso estimator of β is given by

$$\min_{\beta} \left(||Y - \sum_{g=1}^{G} X_g \beta_g||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 \sum_{g=1}^{G} ||\beta_g||_2 \right).$$

A corresponding prior can be constructed as

$$\pi(\beta) \propto \exp\left\{ - \lambda_1 ||\beta||_1 - \lambda_2 \sum_{g=1}^{G} ||\beta_g||_2 \right\},$$

which can be expressed as a scale mixture of normals. However, to select variables both at group level and within a group, Ghosh and Xu [2015][128] developed a hierarchical spike-and-slab prior structure that shrinks coefficients to exactly 0 with posterior median thresholding. Two sets of spike-and-slab distributions, one at group level and another at individual level with a posterior

median estimator, have great variable selection and prediction performance.

We describe the model specifications exactly same as Ghosh and Xu [2015][128] described in their paper. The coefficients $\boldsymbol{\beta}_j$'s are reparameterized to handle two different degrees of sparsity. $\boldsymbol{\beta}_j = \boldsymbol{V}_j{}^{\frac{1}{2}}\boldsymbol{b}_j$ where $\boldsymbol{V}_j{}^{\frac{1}{2}} = diag(\tau_{j1}, .., \tau_{jm_j}), \tau_{jk} \geq 0,\ j = 1, .., p;\ k = 1, .., m_j$. To select variables at the group stage, the first set of spike-and-slab prior is introduced.

$$\boldsymbol{b}_j \overset{ind}{\sim} (1 - \pi_0)N_{m_j}(0, I_{m_j}) + \pi_0\delta_0(\boldsymbol{b}_j), \quad j = 1, .., p$$

When $\tau_{jk} = 0$, $\boldsymbol{\beta}_{jk}$ drops from the model even though $\boldsymbol{b}_{jk} \neq 0$. τ_{jk} controls the magnitude of the elements of $\boldsymbol{\beta_j}$. Hence to choose variables within a group Ghosh and Xu [2015][128] placed another spike-and-slab prior with τ_{jk}'s:

$$\tau_{jk} \overset{ind}{\sim} (1 - \pi_1)N^+(0, s^2) + \pi_1\delta_0(\tau_{jk}),\ \ j = 1, .., p;\ k = 1, .., m_j$$

where $N^+(0, s^2)$ is a normal distribution $N(0, s^2)$ truncated below 0 with mean $\sqrt{\frac{2}{\pi}}s$ and variance s^2. The error variance σ^2 follows a Inverse Gamma distribution with shape α and scale γ:

$$\sigma^2 \sim InverseGamma(\alpha, \gamma).$$

Ghosh and Xu [2015][128] set up an hierarchical setup to decide the values for hyper-parameters π_0, π_1.

$$\pi_0 \sim Beta(a_1, a_2), \quad \pi_1 \sim Beta(c_1, c_2).$$

For s^2, a conjugate inverse gamma prior is placed,

$$s^2 \sim InverseGamma(1, t).$$

We can update 't' with a monte carlo EM algorithm (Casella, 2001[20]; Park and Casella, 2008[96]). For k^{th} EM update,

$$t^{(k)} = \frac{1}{E_{t^{(k-1)}}[\frac{1}{s^2}|\boldsymbol{Y}]},$$

where the posterior mean of $\frac{1}{s^2}$ can be obtained through the mean of Gibbs samples with $(k-1)^{th}$ iteration.

Gibbs Sampler:

Let $\boldsymbol{\beta}_{(j)}$ denote $\boldsymbol{\beta}$ vector without jth group.

$$\boldsymbol{\beta}_{(j)} = (\boldsymbol{\beta}_1^T, .., \boldsymbol{\beta}_{j-1}^T, \boldsymbol{\beta}_{j+1}^T, .., \boldsymbol{\beta}_p^T)^T.$$

$\mathbf{X}_{(j)}$ is the corresponding design matrix to $\boldsymbol{\beta}_{(j)}$. Similarly $\boldsymbol{\beta}_{(jk)}$ denote the whole set of vector of coefficients without kth element corresponding jth group:

$$\boldsymbol{\beta}_{(jk)} = (\boldsymbol{\beta}_{11}, .., \boldsymbol{\beta}_{1m_n}, .., \boldsymbol{\beta}_{j1}, .., \boldsymbol{\beta}_{j\overline{k-1}}, \boldsymbol{\beta}_{j\overline{k+1}}, .., \boldsymbol{\beta}_{jm_n}, .., \boldsymbol{\beta}_{p1}, .., \boldsymbol{\beta}_{pm_n})$$

The corresponding full conditional posterior distributions are:

$$\boldsymbol{b}_j | rest \sim l_j \delta_0(\boldsymbol{b}_j) + (1 - l_j) N_{m_j}(\mu_j, \Sigma_j),$$

where $l_j = P(\boldsymbol{b}_j = 0 | rest)$,

$$l_j = \frac{1}{\pi_0 + (1-\pi_0)|\Sigma_j|^{\frac{1}{2}} \exp\left\{\frac{1}{2\sigma^4}||\Sigma_j^{\frac{1}{2}} \boldsymbol{V}_j^{\frac{1}{2}} \boldsymbol{X}_j^T (\boldsymbol{Y} - \boldsymbol{X}_{(j)} \boldsymbol{V}_{(j)} \boldsymbol{b}_{(j)})||_2^2\right\}},$$

with

$$\mu_j = \frac{1}{\sigma^2} \Sigma_j \boldsymbol{V}_j^{\frac{1}{2}} \boldsymbol{X}_j^T (\boldsymbol{Y} - \boldsymbol{X}_{(j)} \boldsymbol{V}_{(j)} \boldsymbol{b}_{(j)}), \;\; and \;\; \Sigma_j = (I_{m_n} + \frac{1}{\sigma^2} \boldsymbol{V}_j^{\frac{1}{2}} \boldsymbol{X}_j^T \boldsymbol{X}_j \boldsymbol{V}_j^{\frac{1}{2}})^{-1}$$

Full conditional posterior of τ_{jk} is a spike-and-slab distribution:

$$\tau_{jk} | rest \sim q_{jk} \delta_0(\tau_{jk}) + (1 - q_{jk}) N^+(u_{jk}, v_{jk}^2), \;\; j = 1, .., p; \;\; k = 1, 2, .., m_n, m_n + 1, .., p * m_n,$$

where

$$u_{jk} = \frac{1}{\sigma^2} v_{jk}^2 (\boldsymbol{Y} - \boldsymbol{X}_{(jk)} \boldsymbol{\beta}_{(jk)})^T \boldsymbol{X}_{jk} \boldsymbol{b}_{jk}, \;\; v_{jk}^2 = (\frac{1}{s^2} + \frac{1}{\sigma^2} \boldsymbol{X}_{jk}^T \boldsymbol{X}_{jk} \boldsymbol{b}_{jk}^2)^{-1}$$

and $q_{jk} = P(\tau_{jk} = 0 | rest)$,

$$q_{jk} = \frac{1}{\pi_1 + 2(1-\pi_1)(s^2)^{-\frac{1}{2}} (v_{jk}^2)^{\frac{1}{2}} \exp\left\{\frac{u_{jk}^2}{2v_{jk}^2}\right\} \left[\Phi(\frac{u_{jk}}{v_{jk}})\right]}.$$

$$\sigma^2 | rest \sim Inverse\ Gamma\left(\frac{n}{2} + \alpha, \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \gamma\right).$$

With Conjugate Beta and Inverse Gamma prior, the subsequent posteriors are:

$$\pi_0|rest \sim Beta(\#(\boldsymbol{b}_j = 0) + a_1, \#(\boldsymbol{b}_j \neq 0) + a_2),$$

$$\pi_1|rest \sim Beta(\#(\tau_{jk} = 0) + c_1, \#(\tau_{jk} \neq 0) + c_2)$$

$$s^2|rest \sim Inverse\ Gamma\left(1 + \frac{1}{2}\#(\tau_{jk} = 0), t + \frac{1}{2}\sum_{j,k}\tau_{jk}^2\right)$$

2.2.3 Simulated Example:

We apply Ghosh and Xu's [2015] bi-level Bayesian sparse group selection method in function selection and estimation for two simulated examples.

Example 1.

We define 4 functions-

- $f_1(x) = 5x$
- $f_2(x) = 3(5x-1)^2$
- $f_3(x) = \frac{10\sin(2\pi x)}{2-\sin(2\pi x)}$
- $f_4(x) = 6(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3)$

This generating model is the same as Huang et al.'s [2010][] Example 1. We use p=20 and n=100. We employ $\frac{n}{4}$ as the number of internal knots for each function, so that in total we have $\frac{pn}{4}$ number of coefficients. Again, our assumption is those $\frac{n}{4}$ internal knots are sparse, and we will select important knots from them. We use B-spline basis functions of degree=3 with $\frac{n}{4}$ numbers of internal knots. We have $f_5(x) = f_6(x) = .. = f_{20}(x) = 0$ as null function.

We generate the covariates as $x_j = (w_j + t * u)/(1+t)\ \ for\ j = 1,2,3,4$ and $x_j = (w_j + t * v)/(1+t)\ \ for\ j = 5,..,20$. $(w_j, u, v) \overset{ind}{\sim} N^+(0,1)$ and t controls the amount of correlation. The

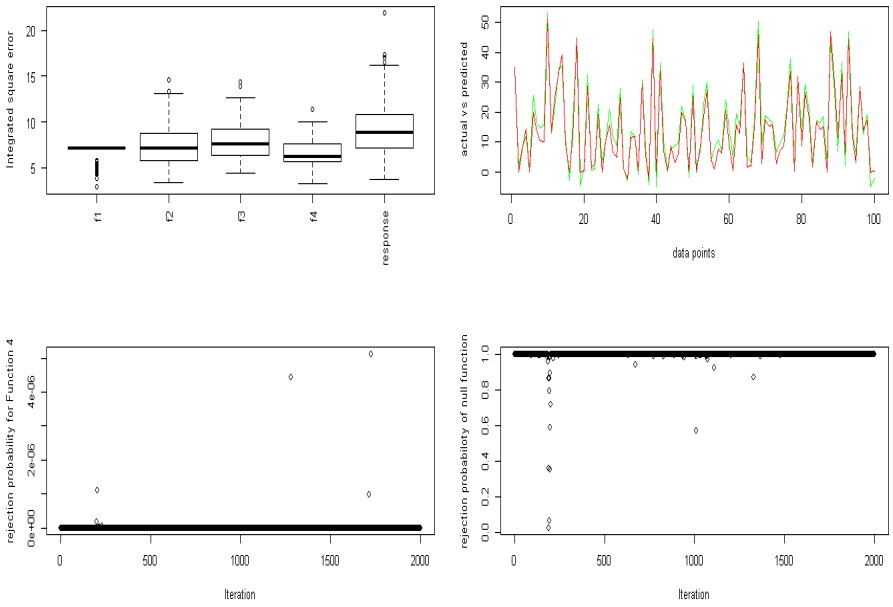


Figure 2.2: Example 1 simulation plot

covariates from zero component and non-zero components are independent. We work with t=0,0.7 and a signal to noise ratio as 9.

We run this simulation setup 100 times and examine the selection performance of the method. We compute the number of true positives and true negatives for 100 repetitions. For our model, we first have four models as true functions and the rest as null functions. Our first function is detected 20% of the time, and the other three functions are correctly detected 100% of the time. In contrast, the null functions are not selected in a model a single time in our 100 repetitions. We compute the ISE for each true function and plot them in a boxplot for 100 repetitions, with the error computed for overall response as well. We have a few of other plots above, such as actual vs. predicted plots for response, and plots of rejection probabilities for true functions and null functions. All the results above are found in the setup where predictors are not correlated.

In our next step, we run the same simulation setup with predictors that are correlated. The response is generated by additive models with the same four functions mentioned above and the same signal-to-noise ratio, but we set the correlation between covariates as 0.7. We run this setup

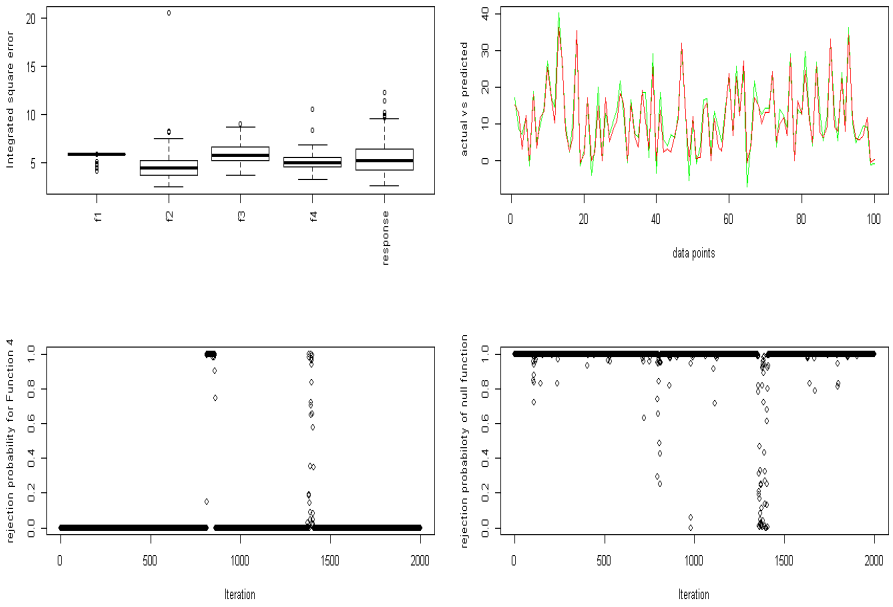


Figure 2.3: Example 1 simulation plot

for 100 repetitions and study the results. Out of 100 repetitions, the selection of the first function is very low at 6%, while the other three true functions are selected 100% of the iterations. On the other hand, one null function come as significant once that is the selection percentage is 1% for that null function. Below are the same plots that were given for the non-correlated setup:

Example 2.

In 2nd example we took another four different functions defined as:

- $f_1(x) = 3x$,
- $f_2(x) = x + \frac{(2x-2)^2}{5.5}$,
- $f_3(x) = -x + \pi \sin(\pi x)$,
- $f_4(x) = 0.5x + 15\phi(2(x-0.2)) - \phi(x+0.4)$, where $\phi()$ is a standard normal density function.

Again, in this example we use p=20 and n=100. We utilize $\frac{n}{4}$ as the number of internal knots for each function, so that in total we have $\frac{pn}{4}$ number of coefficients. Again, our assumption is that

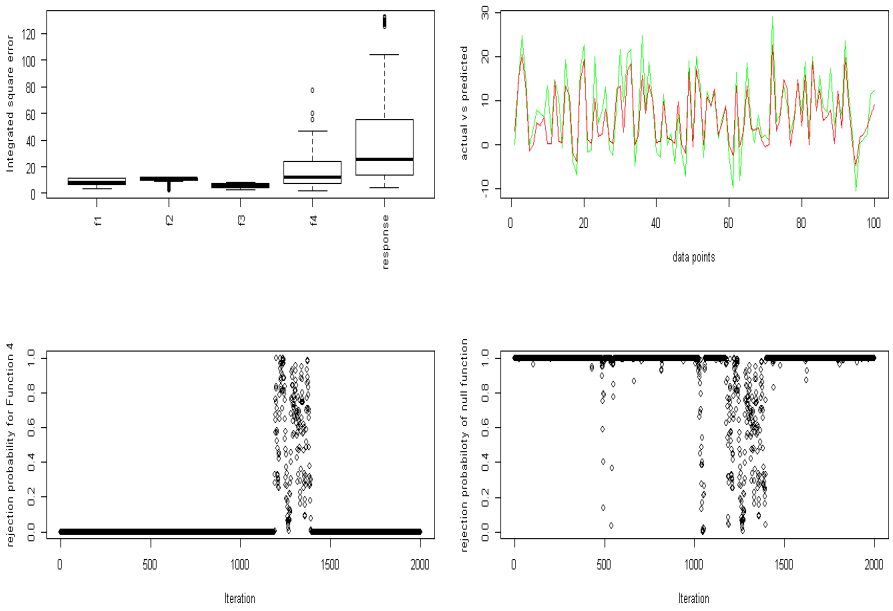


Figure 2.4: Example 2 simulation plot

those $\frac{n}{4}$ internal knots are sparse for functional estimation. We use the B-spline basis functions constructed from predictors. The predictors are generated independently from Uniform (-2,2), and the signal-to-noise ratio is kept at 5. The simulation setup runs for 100 repetitions. Out of 100 repetitions, the first four true functions are selected 75%, 28%, 85%, and 92% of the iterations, respectively, whereas the null functions are not selected a single time in 100 repetitions. A few plots are included below to provide a better picture.

In the next step, we generate the covariates from AR(1) process with $\rho = 0.7$. In the case of correlated covariates, the relationship between predictors and response is still of theoretical interest. We repeat the simulation for the same set of functions mentioned above and gather outputs. We run this setup for 100 repetitions and study the results. Out of 100 repetitions, the selection percentages of the functions are 97%, 50%, 13%, and 97%, respectively. The selection of false positives is zero, except for one null function that is significant with 1% appearances among the null functions we've used. No null functions we use. No null function is selected in the 100 repetitions. Below are the same plots as in the uncorrelated scenario.

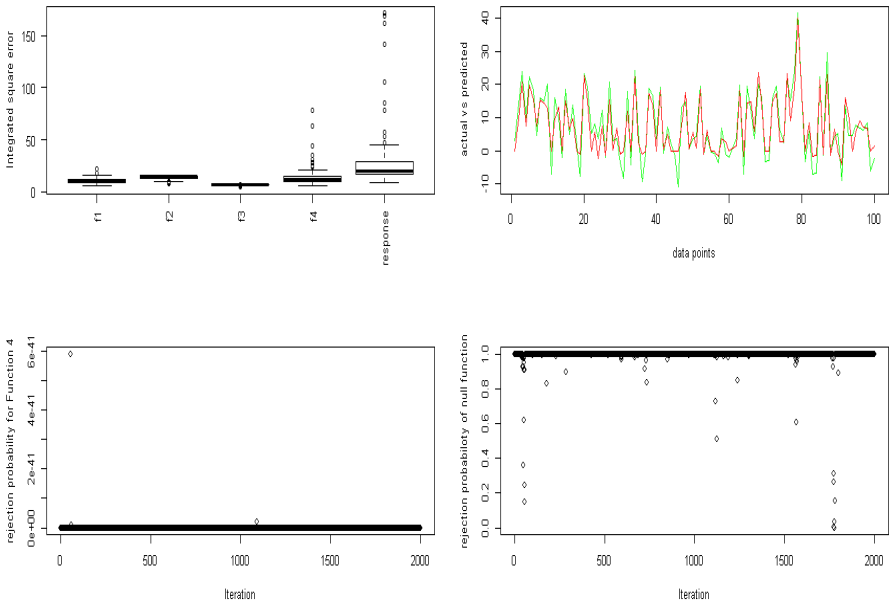


Figure 2.5: Example 2 simulation plot

2.2.4 Conclusion:

We explored an application of Bayesian variable selection for a sparse additive model. The motivation behind applying bi-level selection for function selection and estimation is to find true nonzero functions and estimate them by knot selection. Our idea worked for a small number of functions. If we have 20 functions and 4 of them are significant, the proposed algorithm performs satisfactorily. However, the initial problem was to apply this to a large number of functional spaces. Therefore, more research is needed to draw a final conclusion.

CHAPTER 3

BAYESIAN CLASSIFICATION OF ALZHEIMER'S DISEASE STAGES FROM LONGITUDINAL VOLUMETRIC MRI DATA

3.1 Introduction

Alzheimer's disease (AD) is the most frequent neurodegenerative and age-related form of dementia. AD patients suffer loss of memory and difficulty with speech, and over time become unable to perform daily tasks such as bathing, dressing, eating, and using the bathroom. AD is becoming a significant societal and financial burden among elderly people. As many treatments are being developed and evaluated, it is important to be able to determine early and accurately which individuals are relatively more likely to progress clinically. Based on signs and symptoms, physicians usually track AD using the Clinical Dementia Rating (CDR), and subjects are classified in three states: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). AD accounts for between 60% and 80% of dementia cases. Brain MRI scans provide useful information regarding dementia and its progression. Many group studies based on volumetric regions of interest (ROI)([105],[118]), voxel-based morphometry ([124],[125]) or group comparison of cortical thickness ([55],[84]) have shown that many brain regions like the enthorinal cortex, the hippocampus, lateral and inferior temporal structures, and the anterior and posterior cingulate are responsible for brain atrophy in AD. The atrophy patterns vary in different disease stages over different regions. According to *Leung et al.* [74] MCI converters have higher decaying rates of brain volume than nonconverters do. Hence, hippocampal atrophy using MRI is a marker of AD pathology.

MRI provides scientifically accurate and easy-to-collect data on brain sub-regional volumes, surface area, and cortical thickness. Besides measuring the cross-sectional volume of hippocampal or other regions' structural changes, there is great interest in longitudinal brain volume changes. The

greatest advantage of using longitudinal data is that it can capture a high correlation between observations and time trends, and it can track the rate of changes over time. A prominent phenomenon of significant hippocampal atrophy over time has been observed in AD groups ([7],[49],[60],[119]), whereas healthy aging control groups do not have such high rates of atrophy. In this chapter, we focus on brain regional volumetric changes over time for five ROIs measured by MRI scans: the hippocampus (H), entorhinal cortex (EC), middle temporal cortex (MTC), fusiform gyrus (FG), and the whole brain (WB). The key aim of this chapter is to apply a novel Bayesian classification method for classifying patients into disease groups by observing their longitudinal brain regional changes. Throughout this chapter, we work with dichotomous target variables: AD and CN.

A number of high-dimensional classification and regression methods have been developed for the classification of patients at different stages of disease, and also for the prediction of future clinical changes in MCI patients. Several new machine learning algorithms are used to deal with high-dimensional data, such as support vector machines (SVMs) and linear discriminant analysis. Different approaches have been proposed in various papers ([50],[80],[90],[100]) to classify patients with Dementia from other stages of disease. *Zhang et al.* [133] applied a multi-kernel SVM for classification of patients using a longitudinal feature selection method. In terms of classification between MCI and AD patients, they achieved 78.4% accuracy, 79% sensitivity, and 78% specificity. *Lee et al. 2016* [72] used logistic regression with fused lasso regularization to predict the conversion from MCI to Alzheimer's. *Seixas et al.* [108] developed a Bayesian network model that accounted for a combination of expert knowledge and data-oriented modeling. Writers assumed a known structured Bayesian network model and estimated parameters using the EM algorithm.

In this chapter, we develop a Bayesian classification method using longitudinal volumetric data. This data was obtained from the ADNI database (`http://www.loni.ucla.edu/ADNI`). Longitudinal MRI data has variability in the number of observations for each subject. However, we assume that the longitudinal volumetric measurements are functional observations, and we consider

patients with at least three data points. Longitudinal observations of each subject are smoothed using basis splines even if only a few observations are available. Finally, a Bayesian classification method is employed to classify individuals between AD and CN. Our classification method is unique and easy to implement, and it works with functional predictors. In the Bayesian literature, very few classification methods have been developed using functional predictors. In this context, we want to mention *Zhu et al.* [136], who worked with functional predictors for classification. They developed a selection method to obtain important functional predictors for a binary model. Later, they proposed two MCMC algorithms for posterior sampling, which are Metropolis-Hastings/Gibbs sampler hybrids. We propose the Bayesian Pólya-gamma augmentation method for classification; it is easy to implement using the Gibbs sampler algorithm. Functional predictors are smoothed by cubic basis splines, and basis coefficients are used as predictors in the classification model. We obtain very good results of classification in terms of sensitivity, specificity, and accuracy. This chapter is divided into five sections. Section 3.2 discusses the functional smoothing and classification method developed for the analysis. In Section 3.3, we explain the data in detail. Section 3.4 presents the numerical results and some of the key points we extract from our analysis. Section 3.5 concludes with a short discussion.

3.2 Methodology

In this section, we develop a supervised classification technique to handle functional predictors.

3.2.1 Smoothing functional data

Classification with functional data is a challenge. Functional predictors have high correlation between observations from the same individual. The sole purpose of using functional predictors is to catch the time trend present in the data. Therefore, we take an approach to smooth the parametric curves with a cubic basis spline. After we smooth a particular longitudinal functional curve, the resulting coefficient vector can be considered as a predictor for the next stage of classification. To avoid the complexity of the problem, we use data from patients who have at least three time period

observations, such that smoothed curves are comparable. [61] applied a similar approach to obtain the estimates of a generalized linear model with functional predictors. However, his approach was frequentist, whereas ours is Bayesian.

Let us assume that we observe n patients with their functional observations where each patient has p functions (co-variates). Let $x_{ij}(t)$ be the j-th function observed at time point t from the i-th patient. Let T be the compact domain of $x_{ij}(t)$ and $x_{ij}(t) \in \mathcal{L}^2[T]$. With the functional predictors $x_{ij}(t)$ we assume that we have binary response variable y_i which takes value 0 and 1. Therefore a logistic regression equation would look like-

$$\log\left\{\frac{P(y_i = 1|x_{i1}, .., x_{ip})}{1 - P(y_i = 1|x_{i1}, .., x_{ip})}\right\} = \sum_{j=1}^{p} \int_T x_{ij}(t)\beta_j(t)dt \tag{3.1}$$

To fit the discrete observations $x_{ij}(t)$ we assume that, at any given time t, instead of $x_{ij}(t)$, we observe $X_{ij}(t)$ where

$$x_{ij}(t) = X_{ij}(t) + e(t)$$

where $e(t)$ is a zero-mean Gaussian process. We use a basis function expansion for $X_{ij}(t)$ of the form

$$X_{ij}(t) = \sum_{k=1}^{q} c_{ijk}\phi_k^j(t) = \mathbf{c}'_{ij}\phi^j(t)$$

where $\phi^j(t)$ is the q-dimensional spline basis at time t for jth function, $\mathbf{c}_{ij}$ the q-dimensional spline coefficients for the jth predictor from ith patient. We used ordinary least square estimates for estimating spline coefficients. A simple linear smoother is obtained by minimizing the least squares criterion $||x_{ij} - \Phi\mathbf{c}_{ij}||^2$ as

$$\hat{\mathbf{c}}_{ij} = (\Phi'\Phi)^{-1}\Phi' x_{ij} \tag{3.2}$$

Once the orthonormal basis coefficients have been estimated, we can combine (1) and (2) by

plugging $\hat{x}_{ij}(t)$ in (1), which gives-

$$\log\left\{\frac{P(y_i=1|x_{i1},..,x_{ip})}{1-P(y_i=1|x_{i1},..,x_{ip})}\right\} = \sum_{j=1}^{p}\int_T \hat{\mathbf{c}}'_{ij}\phi(t)\beta_j(t)dt$$
$$= \sum_{j=1}^{p}\hat{\mathbf{c}}'_{ij}\boldsymbol{\beta}_j$$
$$= \mathbf{c}'_i\boldsymbol{\beta} \tag{3.3}$$

where $\beta_j^T = \int_T \beta_j(t)\phi^{j^T}(t)dt$, the coefficient vector for the jth functional predictor. Here $\mathbf{c}_i$ vector has first element as 1 and rest of the spline coefficients for ith patient and $\boldsymbol{\beta}$ contains intercept of the model as first element.

Functional principal component (FPC) analysis is also a popular method that can be applied here. Instead of least square basis estimates, one can work with FPC scores for classification. [93] worked on functional modeling that extends the applicability of FPC analysis for longitudinal data. Specifically, when we have few repeated and irregularly observed data points, FPC scores can be used. In our functional smoothing method, we expand the functional observation with spline basis functions and use the basis coefficients for classification. On the other hand, the same intuition can be applied for FPC scores. For functional component analysis, we assume that longitudinal observations are observations from a smooth random function $X(t)$ mean function $\mu(t) = E\ X(t)$ and covariance function $G(s,t) = cov(X(s), X(t))$. The covariance function can be represented as $G(s,t) = \sum_{k=1}^{\infty}\lambda_k\phi_k(s)\phi_k(t)$ where ϕ_k's are eigenfunctions and λ_k's are eigenvalues. Then the underline process can be written as:

$$X(t) = \mu(t) + \sum_{k=1}^{\infty}\xi_k\phi_k(t),$$

where ξ_k's are frequently referred to as FPC scores. These scores can be used later in the classification model. We do not work with infinite numbers of scores; instead, the above sum is approximated with a finite K that explains the majority of the variance in functional observations. For most cases, the first two FPC scores are enough to build a good classification model. *Zhu et al.* [136] also

used FPC scores in their classification model; they chose functional predictors with significant FPC scores.

There are relatively few counterparts to functional PCA in the Bayesian literature. *Behseta et al.* [9] proposed Bayesian FPCA using two methods, one with a random-coefficient model and another with a hierarchical Gaussian process model. They found better performance for the hierarchical model. Later, in 2008, *van der Linde* [121] introduced a variational algorithm for one-parameter exponential families to obtain approximate Bayesian inference in functional PCA. Recently, *Suarez and Ghosal* [117] proposed a prior structure on the covariance function of functional observations. Their model simultaneously smooths functional observations while estimating principal components. In this chapter, we work with a basis spline smoothing method due to its ease of implementation in statistical software. In R, we have the *splines* package, which fits cubic basis splines on longitudinal data with equally placed knots. We do not investigate any findings using FPC scores instead of basis spline coefficients, as our main focus is on the classification algorithm and basis spline coefficients work very well for our classification model.

3.2.2 Classification using Pólya-Gamma Augmentation

In this section, we present the relationship between predictors and outcome using a logistic regression model. The binary data is denoted as $y_i \in \{0, 1\}$ $(i = 1, .., n)$. Now, posterior sampling of logistic regression coefficients is difficult due to the model's complicated likelihood function. The assumption of the Gaussian prior for regression coefficients is highly important, but the full posterior distribution of regression coefficients becomes analytically inconvenient. In comparison to the logistic model, it is computationally easier to execute Bayesian inference using a probit model [2]. Different sampling algorithms have been proposed [2]. *Holmes et al.* [54] developed an indirect sampling method by introducing auxiliary variables for binary and multinational regression. Later, more methods based on latent variables for logistic regression were reported by *Frühwirth-Schnatter et al.* [41], *Gramacy et al.* [45] and *Polson et al.* [97]. Among all these works,

Polson et al.'s algorithm is most interesting to us due to its ease of computational implementation as well as its sampling efficiency. Our aim is to avoid the complex Metropolis algorithm while sampling from posterior distributions of regression coefficients.

A vast body of literature is concerned with the analysis of Bayesian logistic models. As mentioned earlier, *Holmes et al.* [54] proposed using an auxiliary variable to avoid the conditional non-conjugacy for updating β. The prior structure they used was:

$$y_i = \begin{cases} 1, & \text{if } z_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$z_i = \mathbf{x}_i\beta + \epsilon_i$$

$$\epsilon_i \sim N(0, \lambda_i)$$

$$\lambda_i = (2\psi)^2$$

$$\psi_i \sim Kolmogorv - Smirnovdistribution$$

$$\beta \sim Normaldistribution$$

The above prior is interesting because the marginal likelihood $\mathcal{L}(\beta|data)$ is the same as the likelihood for the logit model. However, the main disadvantage of using this prior structure is that although we obtain a conjugate full conditional distribution of β given data, the conditional distribution of $\pi(\lambda_i|z_i, \beta)$ does not have any standard form. One must use a complicated rejection sampling method to sample for conditional λ_i. Hence, adding an auxiliary variable does not give us significant computational improvement compared to using the Metropolis-Hastings algorithm. *Frühwirth-Schnatter et al.* [41] addressed this problem with the same approach, but instead of a single auxiliary variable they used a two-stage augmentation method. In the first stage, they assumed the existence of a latent variable, where the binary response variable was conditional on the sign of the auxiliary variable. The error part of the model was assumed to follow a type I extreme value distribution that had non-normal density. Then, in the second stage of data augmentation, they approximated this non-normal error distribution using the mixture of the normal

distribution. Finally, they obtained a multivariate normal distribution for the posterior of β. Among all popular methods, pólya-gamma augmentation is most interesting to us and it's easy to apply due to availability of the R package.

We now discuss *Polson et al.'s* [97] algorithm in detail. They showed how a Gaussian variance mixture distribution with a Pólya-gamma mixing density can approximate logit likelihood. We start with defining Pólya-gamma density-

Random variable $X \sim PG(b, c)$, a Pólya-gamma distribution with parameters $b > 0$ and $c \in \mathfrak{R}$, if

$$X \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + \frac{c^2}{4\pi^2}},$$

where $g_k \sim Gamma(b, 1)$ are independent gamma random variables and $\overset{d}{=}$ indicates equality in distribution.

In their work, *Polson et al.* [97] showed that Bernoulli likelihoods parametrized by log-odds can be represented as mixtures of Gaussians with respect to Pólya-gamma distribution. Assume we have latent variables ω such that $\omega \sim PG(b, 0)$ distribution, which is infinite sum of gammas: $\omega \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-\frac{1}{2})^2}$. Now if $\omega \sim PG(b, 0)$ below equality can be proved:

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\frac{\omega\psi^2}{2}} p(\omega) d\omega \tag{3.4}$$

where $\kappa = a - \frac{b}{2}$.

Motivated by *Polson et al.*'s [97] integral result, we construct a Bayesian prior formulation targeted to handle binary logistic regression. Equation 3.3 has a Bernoulli likelihood function with logit link. We proposed a Normal prior for $\boldsymbol{\beta}$ vector in equation (3.3). To derive our Gibbs sampler

, we introduce a latent variable ω. Our prior set up is-

$$
\begin{aligned}
y_i|\mathbf{c}_i, \boldsymbol{\beta} &\sim Bernoulli(\frac{e^{\mathbf{c}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{c}_i'\boldsymbol{\beta}}}),\ i = 1, .., n \\
\omega_i &\sim PG(1, 0),\ i = 1, .., n \\
\boldsymbol{\beta} &\sim Normal(b, B)
\end{aligned} \tag{3.5}
$$

Gibbs Sampler:

The likelihood for ith observation is

$$
\begin{aligned}
L_i(\boldsymbol{\beta}) &= \frac{(e^{\mathbf{c}_i'\boldsymbol{\beta}})^{y_i}}{1+e^{\mathbf{c}_i'\boldsymbol{\beta}}} \\
&\propto e^{\kappa_i \mathbf{c}_i'\boldsymbol{\beta}} \int_0^\infty \exp\left\{-\frac{\omega_i(\mathbf{c}_i'\boldsymbol{\beta})^2}{2}\right\} p(\omega_i)d\omega
\end{aligned}
$$

where $\kappa_i = y_i - \frac{1}{2}$. We combine the likelihood function with $\boldsymbol{\beta}$ prior, given $\omega = (\omega_1, .., \omega_n)$:

$$
\begin{aligned}
p(\boldsymbol{\beta}|\omega, \mathbf{y}, \mathbf{c}) &\propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\left\{\kappa_i \mathbf{c}_i'\boldsymbol{\beta} - \frac{\omega_i(\mathbf{c}_i'\boldsymbol{\beta})^2}{2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}(z - \mathbf{C}\boldsymbol{\beta})^T \Omega(z - \mathbf{C}\boldsymbol{\beta})\right\} \\
&\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - b)^T B^{-1}(\boldsymbol{\beta} - b)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{C}_\omega(\mathbf{C}^T\Omega z + B^{-1}b))^T \mathbf{C}_\omega^{-1}(\boldsymbol{\beta} - \mathbf{C}_\omega(\mathbf{C}^T\Omega z + B^{-1}b))\right\}
\end{aligned}
$$

where $z = (\frac{\kappa_1}{\omega_1}, .., \frac{\kappa_n}{\omega_n})$, $\Omega = diag(\omega_1, .., \omega_n)$ and $\mathbf{C}_\omega = (\mathbf{C}^T\Omega\mathbf{C} + B^{-1})^{-1}$.

Finally, posterior samples for regression coefficients can be computed with Pólya-gamma augmentation. The steps are

$$
\begin{aligned}
(\omega_i|\boldsymbol{\beta}) &\sim PG(1, \mathbf{c}_i^T\boldsymbol{\beta}) \\
(\boldsymbol{\beta}|y, \omega) &\sim N(\mathbf{C}_\omega(\mathbf{C}^T\Omega z + B^{-1}b), \mathbf{C}_\omega)
\end{aligned} \tag{3.6}
$$

We used 0.5 probability threshold to construct our response variable Y as dichotomous $\{1, 0\}$. Once we obtain the posterior samples of β, we can calculate the predicted probability for each patient. We use the posterior median estimate for β - that is, we calculate the median of the posterior samples of β and use those estimates for validation in the test data set. Finally, we need to decide a threshold to classify each patient into either the AD or the CN group. The classification method is implemented in the R package *BayesLogit*. The sampler is highly efficient for sampling from Pólya-gamma distribution with a positive parameter as one. In 2013, *Choi et al.* [26] published a paper in which they showed that "the Pólya-gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic". This result is very crucial, as it guarantees that Monte Carlo averages of posterior samples follow the central limit theorem. This is a strong reason to adopt this method for our analysis.

3.3 Data Description

The MRI data used in the analysis of this thesis was collected from the Alzheimer's disease Neuroimaging Initiative (ADNI) website. We consider five main longitudinal predictors as potential markers of differences between AD patients and normal aging persons. AD patients and normal aging persons both have brain region atrophy, and our aim is to find the essential markers related to brain atrophy that causes dementia after a certain age. We have longitudinal volumetric measurements of the hippocampus (H), entorhinal cortex (EC), middle temporal cortex (MTC), fusiform gyrus (FG), and the whole brain (WB). We believe that these are the potential predictors that can distinguish the AD patients from normal aging patients.

From the ADNI database, we have information on a total of 1,279 patients diagnosed as either AD or CN over their visits for tests. We discard five patients who changed their disease status from CN to AD over their longitudinal measurements. At the start of our analysis, we decide to keep patients who have at least three data points as longitudinal measurements. After filtering for number of data points and excluding patients whose predictor-related information is missing, 528 patients

are included in the study, divided into two groups: AD (n=206) and CN (n=322). The rest of the patients have the status of mild cognitive impairment (MCI). All patients went through an initial clinical evaluation to measure their baseline scores for different tests, such as the Mini-Mental State Examination (MMSE) and the 11-item Alzheimer's disease Assessment Scale–Cognitive Subscale (ADAS11). In addition, patients underwent a baseline structural MRI scan, giving us baseline measurements of all the predictors of interest. In addition, at the start of the study, subjects provided apolipoprotein E (APOE) genotyping information. Data was collected from patients' visits at specific time points, such as after 6, 12, 18, 24, and 36 months. Patients' visits were irregular, and we do not have uniformity over differences between consecutive measurements. We start by comparing the baseline measurements between the AD and the CN group, as shown in Table 1. Although age distribution and gender ratio are not significantly different between these two groups, the other three features, namely baseline MMSE, ADAS11 score, and APOE proportion, show significant differences between the two disease classes at the 5% level of significance. This implies that those three features can be used as strong predictors for classifying patients. Furthermore, we examine the patterns and spread for the number of visits for patients. For all the patients we consider, the minimum observed time span is 10 months (only three observations were collected), while the longest is 10 years. In terms of number of visits or data points, there is a minimum of three data points due to our threshold, and the maximum is 11. The median number of data points is around 4. Table 2 provides a detailed breakdown of this information with respect to the AD and the CN group.

In the spaghetti plot in Figure 1, we plot the predictor variables for each subject segmented by their disease status. In addition, we create another plot for patients' MMSE scores. The mean values for each group are highlighted in the plots. In the first five plots, we can see that the hippocampus and other important variables have much higher values for the CN group than for dementia patients. We can visually observe the MMSE score patterns, although in Table 1 we see much higher MMSE scores for normal patients. The goal is to predict the probability of a patient belonging to either of these groups with measured MRI scans.

Table 3.1: Patients Baseline Characteristics

	AD	CN	p-value
n	206	322	
Age (Mean ± sd)	73.53 ± 7.67	74.15 ± 5.67	0.325
Gender (F/M)	97/109	155/167	0.814
MMSE (Mean ± sd)	24.38 ± 2.5	29.07 ± 1.13	<0.0001
ADAS11 (Mean ± sd)	16.8 ± 6.42	5.85 ± 2.87	<0.0001
APOE (+/-)	143/63	85/237	<0.0001

Notes: Comparison of Baseline Age, Gender ratio, MMSE score, ADAS11 score and APOE ratio between AD and CN groups

Table 3.2: Data Characteristics

	n	Span (Months)			#Data Points (Nobs)		
		Max	Min	Median	Max	Min	Median
AD	206	71	10	24	7	3	4
CN	322	120	11	36	11	3	4

* duration of patients' visits and number of visits available for each group

3.4 Application results

This section presents the numerical results of applying the method proposed in Section 2 to ADNI data. We first deal with a single potential covariate, "hippocampus," as a representation of a single ROI's volume. Section 4.1 lists the results we obtain using the hippocampus as a single predictor. We repeat the same procedure in Section 4.2 by listing all classification results using single ROIs. Section 4.3 presents the results regarding combinations of the ROIs to determine the best model with the greatest classification power. We work with five potential predictors and a significant number of combinations of five ROIs.

3.4.1 Classification using Hippocampus

In this section, we consider the volumetric MRI data of the hippocampus to demonstrate the applicability of the proposed method with a single predictor.

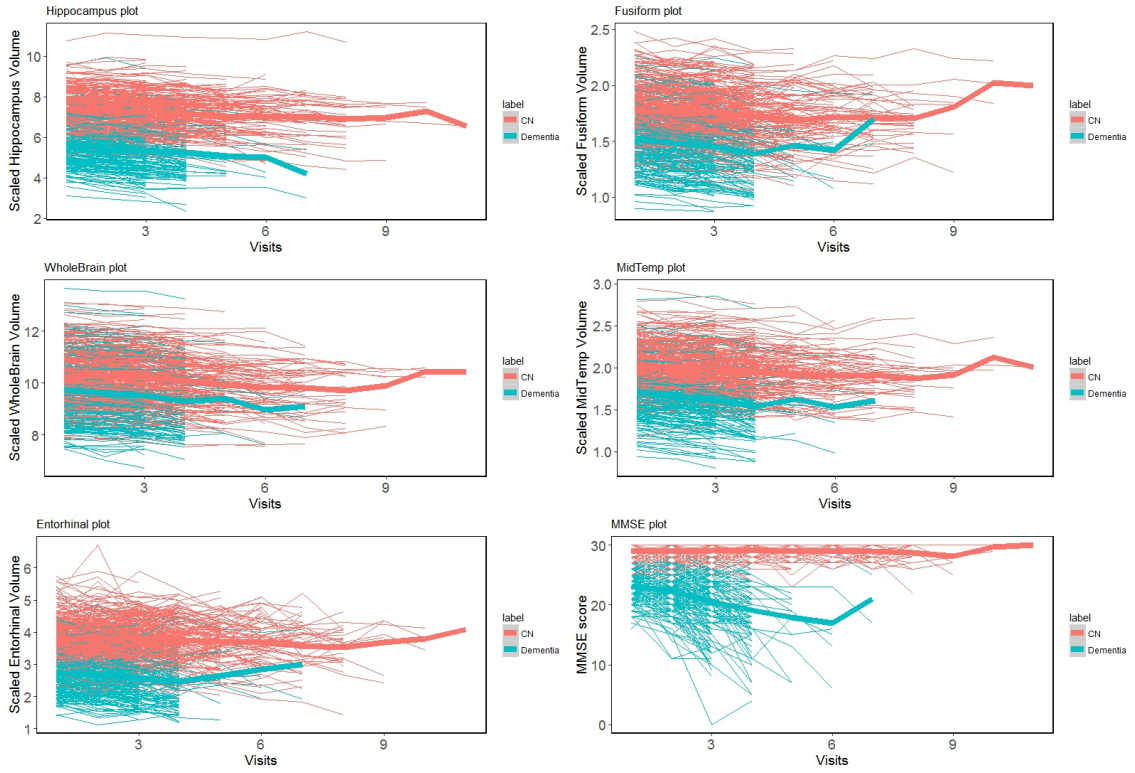


Figure 3.1: Longitudinal volume of ROIs for dementia patients and normal controls. The first five plots (plots in the first column and the top two plots in the second column) are scaled ROIs of interest. The last plot has MMSE scores of patients. On the X-axis, we have the number of visits for patients. Blue lines are for the dementia group and orange lines are for the normal group. Thin lines represent each patient's data, and thick lines are the pooled mean for the AD and CN groups.

We start our analysis by smoothing the longitudinal observed values of the hippocampus. A simple least squares approximation is sufficient, as we assume that the residuals of the true curve are independently and identically distributed with mean 0 and constant variance. We use the cubic B-spline basis functions for spline smoothing of the observed hippocampus volume. Four internal knots are used for spline smoothing with intercept, which gives us eight basis functions. We want to ensure that the smoothed estimated curve yields a good fit to each patient's observed curve. As we do not have a large number of data points for each patient, we do not consider controlling over fitting of our estimated curve. Figure 4 shows the observed hippocampus volume of patients in both group for all visits. The black points represent the mean values of the smoothed curves for all patients at each time point or visit. The smooth curve seems to track the variation in the volumetric

measurements accurately. Besides least squares smoothing, functional principle component scores can also be used for this analysis. At the start of the analysis, we divide the data set into two parts: three-quarters of the patients (457) are reserved as a training data set, and the rest (152) are kept for testing. For each patient in the training data set, we gather the basis coefficients used as predictors for classification. For classification, we generate β prior from a multivariate normal distribution with a mean vector as 0 and a diagonal covariance matrix with diagonal elements as 100. The large diagonal numbers for the covariance matrix represent the absence of any prior knowledge for β and indicate a non-informative prior. The pólya-gamma augmentation method is repeated for 10,000 iterations, and the initial 5,000 iterations are considered a burn-in period for the Monte Carlo Markov Chain. Figure 5 shows the predicted probability plot, with a red horizontal line representing a probability value equal to 0.5. A probability threshold of 0.5 is used to classify patients as either AD or CN. Patients with more than a 0.5 probability are tagged as AD patients. Figure 4c presents a boxplot of the β coefficients estimated from Pólya-gamma augmentation. The coefficients are very small and mostly close to 0, except for the intercept, which is not shown in the plot.

We apply the method described in Section 3 to the training data and obtain the estimated β coefficients. We use functional smoothing on the hippocampus volume test data set. Using both estimated coefficients and basis matrices for this data set, we obtain the predicted probability. To check the robustness of our classification method, we repeat the procedure 100 times and examine the corresponding sensitivity, specificity, accuracy, and area under the ROC curve (AUC). Table 3 lists the results for all important predictors we consider when used as single ROIs in the model, including sensitivity, specificity, accuracy, and AUC mean with standard deviation. We obtain 82% sensitivity, 88% specificity, and 86% accuracy for the classification using the Hippocampus, which gives us strong confidence in the validity of our methods. The corresponding ROC curve is also shown in Figure 3.

3.4.2 Classification with Single ROI

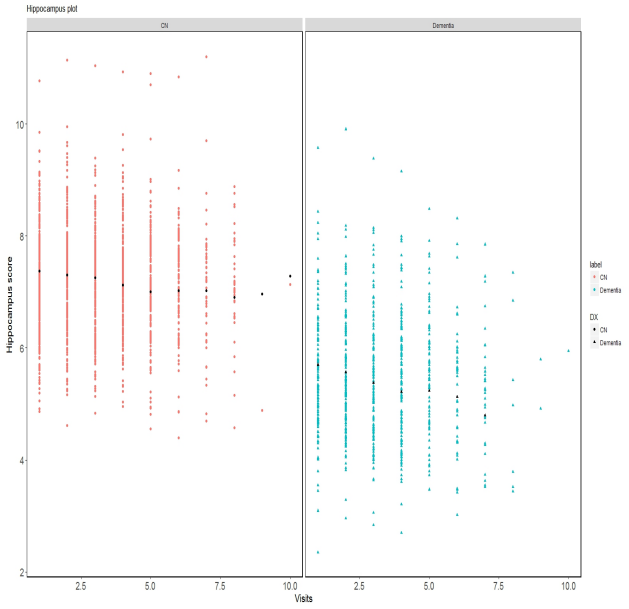
In the previous section, we showed the applicability of our proposed method using the hippocampus as a single ROI. We now evaluate the classification performance if we consider all five ROIs: H, WB, EC, FG, and MTC. We repeat our proposed procedure 100 times and calculate sensitivity, specificity, accuracy, and AUC with one-quarter of the test data set. We set the threshold probability to 0.5 for all cases. Sensitivity, specificity, accuracy, and AUC mean with standard deviation are listed in Table 3.

Table 3.3: Classification performance using single ROI

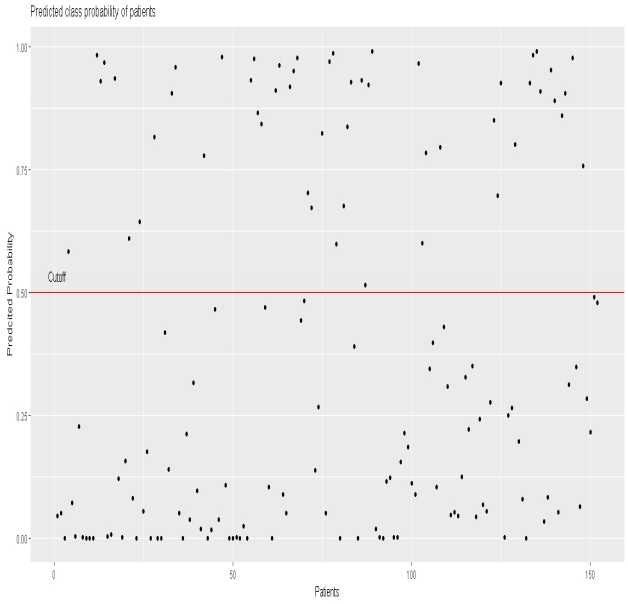
	sensitivity	**specificity**	**accuracy**	**AUC**
Hippocampus (Mean ± sd)	0.82 ± 0.01	0.88 ± 0.01	0.86 ± 0.001	0.92 ± 0.02
WholeBrain (Mean ± sd)	0.75 ± 0.01	0.74 ± 0.01	0.75 ± 0.001	0.83 ± 0.03
Entorhinal Cortex (Mean ± sd)	0.80 ± 0.01	0.90 ± 0.001	0.86 ± 0.01	0.92 ± 0.02
Fusiform gyrus (Mean ± sd)	0.73 ± 0.01	0.87 ± 0.01	0.81 ± 0.01	0.89 ± 0.03
Middle Temporal Cortex (Mean ± sd)	0.75 ± 0.01	0.87 ± 0.01	0.83 ± 0.01	0.89 ± 0.03

Notes: Sensitivity is the proportion of correct AD predictions; specificity is the proportion of correct CN predictions; AUC is area under the ROC curve. The mean and standard deviation are based on 100 repeated results in test data sets. The probability threshold is 0.5.

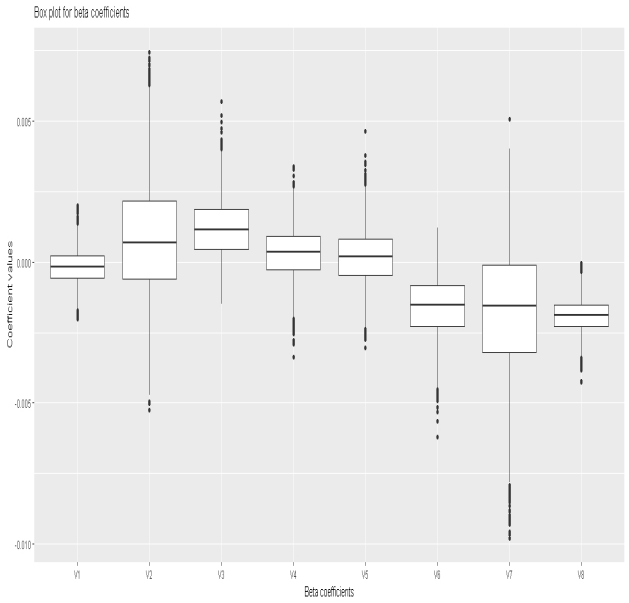
It is evident from Table 3 that single ROIs are critical for distinguishing dementia patients from normal aging people. We obtain an accuracy rate of around 0.80 for single ROIs, and AUC is around 0.9. These are excellent statistics. We also examine the sensitivity and specificity using single ROIs. A high sensitivity of 0.80 gives us confidence in our proposed method and implies that volumetric MRI data of single ROI does have strong classification power.



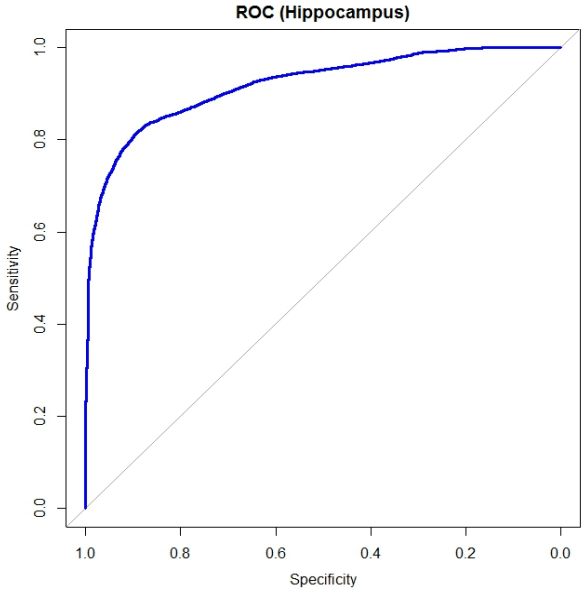
(a) Mean smoothed Hippocampus scores



(b) Predicted class probability



(c) Box plot for posterior median estimates of model coefficients using Hippocampus



(d) ROC curve for Hippocampus model

Figure 3.2: In (a), the colored points represent hippocampus values for each time point, and the black dot signifies the mean value of the smoothed curve for that time point. Orange is for CN and blue is for AD. In (b), the points are the predicted class probability for patients in the test data set. (c) presents the spread of the estimated regression coefficients from the Bayesian logistic model in 100 repeated runs. (d) shows the ROC curve after classifying patients using the hippocampus only.

3.4.3 Classification using combination of ROIs

In the final stage of our analysis, we consider all combinations of ROIs. We first check the classification performance of our model using combinations of five ROIs, and we later include the baseline MMSE, ADAS11 score, and APOE status. Table 1 showed that patients had significant differences in MMMSE and ADAS11 scores at the beginning of the study. We include these variables to make our model stronger in terms of classification power. We describe our analysis in the following. Before applying our method directly to the data, we make one change in the volumetric MRI ROIs: we normalize all the volumetric ROIs by dividing them by the intracranial volume (ICV) to bring each patient's ROI measurements to the same level. For example, to normalize the ROI volume, we consider each patient's ICV for each time point and divide the ROI volume by the corresponding ICV for each patient at each time point. We then calculate the basis coefficients from longitudinal measurements of volumetric ROIs and use those coefficients, including baseline MMSE, ADAS11, and APOE, in the Bayesian logistic model for each subject. The same pólya-gamma augmentation is used to avoid complicated Metropolis-Hastings algorithm. The following is the model for classification with $p_i = P(y_i = 1|x_{ij}(t)$ using the combination of H, WB, EC, FG and MTC:

$$\begin{aligned} logit(p_i) =& \beta_0 + \boldsymbol{\beta}_1 \mathbf{c}_i^H + \boldsymbol{\beta}_2 \mathbf{c}_i^{WB} + \boldsymbol{\beta}_3 \mathbf{c}_i^{EC} + \boldsymbol{\beta}_4 \mathbf{c}_i^{FG} + \boldsymbol{\beta}_5 \mathbf{c}_i^{MTC} \\ &+ \beta_6 MMSE_i + \beta_7 ADAS11_i + \beta_8 APOE_i, \quad i = 1, ..., n \end{aligned} \tag{3.7}$$

We use this model with our normalized data to classify patients into two groups: AD and CN. We obtain astonishingly better results than what we observed using single ROIs. The classification power of the model is very high, with 93

3.4.4 Conclusion

We collected results for 26 different combinations of ROIs, excluding the results of single ROIs. We list the important points and conclusions below based on those results:

- Overall, we obtain 80% sensitivity, 90% specificity, and 85% accuracy using single ROIs as

potential predictors. This implies that each individual ROI has strong classification power in differentiating AD patients from normal aging people. Among the five potential volumetric MRIs, the whole brain yields the lowest sensitivity, specificity, and accuracy, while the hippocampus performs best.

- To compare different model results using combinations of ROIs, we start with the findings obtained from the combination of two ROIs. The overall sensitivity moves to around 80% and the specificity to around 90%. The hippocampus and fusiform gyrus together have the best performance. We do not use the normalized volume of ROIs for these models.

- In comparing results between different combinations, we mostly emphasize sensitivity, as it indicates the percentage of patients correctly identified as Alzheimer's patients. We achieve more than 80% sensitivity for most of the combinations. Furthermore, we obtain more than 85% sensitivity for the majority of combinations of three ROIs. Specifically, **H+WB+FG** has 91% sensitivity and 84% accuracy.

- Next, when using combinations of four ROIs, *H+WB+EC+FG* has 0.91 sensitivity and 0.79 specificity. However, *H+WB+EC+MTC* has sensitivity of 0.77 and 0.89 specificity, indicating that the inclusion of the middle temporal cortex (MTC) in the model and the exclusion of the fusiform gyrus (FG) decreases sensitivity and increases specificity.

- We obtain the best result when using all five ROIs with baseline MMSE, ADAS11, and APOE in the model. This model achieves 96% specificity and 95% accuracy, which is high compared to other combinations. The addition of MMSE and ADAS11 scores in the model definitely increases the performance of method. However, the five potential ROIs together yield 81% sensitivity, 77% specificity, and 79% accuracy.

3.5 Discussion

We have proposed a Bayesian classification method motivated by a practical problem in the domain of Alzheimer's disease. The differentiation of Alzheimer's patients based on their MRI

inputs is crucial in medical science. The importance of this problem is two-fold: first, doctors need to know the main factors involved with dementia in aging patients; and second, a computationally feasible and less time-consuming classification method is proposed here. Bayesian logistic regression with functional predictors is itself a challenge due to its analytically complicated likelihood function. The Metropolis-Hastings algorithm is a popular tool used by many researchers to tackle this problem, but the choice of candidate density is highly subjective and convergence rates are not great. In this situation, Pólya-gamma augmentation provides us with sampling efficiency that can easily be integrated into hierarchical Bayesian modeling.

One more challenging question arises when working with functional predictors: What about the selection of predictors? In our study, we work with five important ROIs and a few baseline score variables. This problem can be extended to a situation in which one has multiple functional predictors and some of them are redundant. It is crucial in modern Bayesian theories to find a proper variable selection method for classification which works with functional predictors as well. In the context of Alzheimer's disease, not all MRI measurements over time are important for predicting AD patients. In our next study, we will build a data-driven method to provide a feasible solution to this problem.

In this chapter, we applied a classification method to distinguish AD patients from healthy controls. However, in the Alzheimer's disease domain, the differentiation of MCI converters (MCI-c) from MCI nonconverters (MCI-nc) is more interesting. The main problem we faced while dealing with this situation was the low ratio of MCI-c to MCI-nc. The data set was sparse, and we did not find a significant number of disease cases in the test data on which we could validate our method. In addition, most of the baseline MCI patients had very few data points in their longitudinal measurements.

Besides the points mentioned above, our method performs very well for the given data. The

best classification model is

H+WB+EC+FG+MTC+MMSE_bl+ADAS11_bl+APOE_bl with sensitivity=0.93, specificity=0.96, accuracy=0.95 and AUC=0.98. This is almost an oracle classifier. Our Bayesian logistic model uses very few functional predictor variables and performs much better than other exiting methods. Our method prefers a highly cost-effective and less time-consuming data collection process, as fewer MRI measurements are needed for analysis. In summary, the method proposed in this chapter shows good performance in distinguishing Alzheimer's patients from normal aging controls.

Acknowledgment

We would like to thank Yingjie Li[77], email:liyingj1@stt.msu.edu, for providing us with the data used in this study

CHAPTER 4

BAYESIAN PENALIZED MODEL FOR CLASSIFICATION AND SELECTION OF FUNCTIONAL PREDICTORS USING LONGITUDINAL MRI DATA FROM ADNI

4.1 Introduction

The research literature on applied mathematical approaches and classification methods using longitudinal MRI data has seen massive growth over the past decade. Among the broad range of methods applied with variable degrees of success, several warrant mention. *Misra et al.* [90] implemented a high-dimensional pattern recognition method to baseline and longitudinal MRI scans to predict conversion from MCI to AD over a 15-month period. *Zhang et al.* [133] used a multi-kernel SVM for classification of patients between MCI and AD, achieving 78.4% accuracy, 79% sensitivity, and 78% specificity. *Lee et al.* [72] applied logistic regression in predicting conversion from MCI to Alzheimer's, using fused lasso regularization to select important features. *Seixas et al.* [108] proposed a Bayesian network decision model for detecting AD and MCI which considered the uncertainty and causality behind different disease stages. Their Bayesian network used a blended effect of expert knowledge and data-oriented modeling, and the parameters were estimated using an EM algorithm. *Adaszewski et al.* [1] employed classical group analyses and automated SVM classification of longitudinal MRI data at the voxel level. *Arlt et al.* [4] studied the correlation between the test scores over time with fully automated MRI-based volume at the baseline. However, few studies to date have developed methods that increase the sensitivity, accuracy, and specificity of classification in AD diagnosis or progression to more than 80%.

Classification using longitudinal data can be a challenge with a large number of predictors. The first significant approach to handle longitudinal predictors is to consider each multiple-occasion observation as a single function observed over a time interval. Functional predictors have a high correlation with adjacent measurements, and the observational space is high-dimensional. The

number of predictors required for estimation often exceeds the number of observations, thus introducing the problem of dimensionality. A regression framework is frequently the most suitable to model all possible longitudinal effects across ROIs, where the proposed method will select the important predictors. Moreover, many biomedical studies have shown that a limited number of specific brain regions or ROIs are essential for AD classification. Thus, dimension reduction techniques can be applied, and classification can be limited to the reduced feature set. *Zhu et al.* [136] advanced a method for classification and selection of functional predictors that entails calculation of functional principle component scores for each functional predictor, followed by the use of these scores to classify each individual observation. They proposed using Gaussian priors for selection and created a hybrid Metropolis-Hastings/Gibbs sampler algorithm. Although the method reported in the present study is inspired by this method, we develop a simple Gibbs sampler where MCMC samples are drawn from standard distributions. We also focus on applying penalized regression for dimension reduction. In the Bayesian variable selection literature, the spike-and-slab prior has widespread applications due to its superior selection power. *George et al.* ([43],[44]) initially proposed that each coefficient β can be modeled either from the "spike" distribution, where most of its mass is concentrated around zero, or from the "slab" distribution, which resembles a diffuse distribution. Instead of imposing the spike-and-slab prior directly on regression coefficients, *Ishwaran et al.* [57] introduced a method in which they placed a spike-and-slab prior on the variance of Gaussian priors. The Bayesian variable selection methods also include different Bayesian regularization methods, such as Bayesian Lasso [96], Bayesian Group Lasso, Bayesian elastic net [76]. We employ a Bayesian group lasso algorithm blended with a spike-and-slab prior obtained from *Xu and Ghosh, 2015*[128]. The group structure among coefficients in our model comes from functional smoothing of the coefficients, and group lasso facilitates the selection of the important functional predictors. Thus, our proposed method takes the idea of Bayesian variable selection to a generalized functional linear model with binary responses.

The fundamental challenge of this work is to perform logistic regression in a Bayesian frame-

work while using a large number of functional predictors. The direct sampling of regression coefficients from the Bayesian logistic model is difficult due to its complicated likelihood function. In high-dimensional scenarios, selection of predictors becomes crucial with the introduction of either a spike-and-slab prior, non-local priors, or horseshoe priors. For all such priors, the full posterior distribution of regression coefficients is analytically inconvenient. We obtain the Pólya-gamma augmentation method with priors proposed by *Xu and Ghosh, 2015*[128], which yields full conditional samples from standard distributions. Our aim is to avoid the complications of Metropolis-Hastings and to develop an easily implementable Gibbs sampler. In addition, Bayesian estimation provides proper estimates of the model parameters, which are also useful for building inference. The key advantage of this method is that it calculates the log of odds of AD with respect to CN based on the selected longitudinal predictors. Moreover, we use a probability threshold for classifying individual patients to validate our modeling performance. We obtained the data used in the dissertation from the ADNI server. The volumetric MRI brain data includes parcellated sub-regions of the whole brain, with separate subdivisions for the left and right hemispheres. Volumetric measurements of brain sub-regions across multiple occasions over time demonstrate differential patterns of brain atrophy between AD patients and normal aging people. Because not all brain regions are as closely related to AD, the redundant features derived from the unrelated brain regions can be removed by limiting the selection to brain sub-regions important to classification. The problem of identifying important brain sub-regions from a large number of functional predictors or longitudinal measurements is far from simple. Various variable selection methods have been designed for single-time-point data with respective target variables. We apply a Bayesian variable selection method to select longitudinal features or functional predictors for our data set. We work with 49 functional predictors consisting of longitudinal volumetric measurements in different sub-regional brain ROIs. The use of the spike-and-slab prior ensures that a large number of redundant predictors are dropped from the model. The ROI sub-regions selected by our method will be helpful for future studies to detect the progression of dementia.

The chapter is organized as follows. In section 2, we introduce Bayesian variable selection with a spike-and-slab prior. Section 3 discusses functional smoothing of the longitudinal predictors. In Section 4, we introduce our methodology and algorithm for simultaneous selection and classification. Theoretical properties and consistency results are shown in Section 5. We then discuss the application results with simulated data and real data in Sections 6 and 7. Finally, Section 8 examines another potential modeling approach, and Section 9 covers the overall development and limitations of the methodology.

4.2 Bayesian Variable selection

We will briefly discuss about Bayesian variable selection below:

4.2.1 Spike-Slab prior

A Bayesian model with a spike-and-slab prior can be constructed as follows:

$$
\begin{aligned}
(Y_i/x_i, \beta, \sigma^2) &\overset{ind}{\sim} N(x_i'\beta, \sigma^2), \quad (i = 1, ..., n) \\
(\beta/\gamma) &\sim N(\mathbf{0}, \Gamma), \\
\gamma &\sim \pi(d\gamma), \\
\sigma^2 &\sim \mu(d\sigma^2),
\end{aligned}
$$

where $\mathbf{0}$ is a p-dimensional zero vector, Γ is the p x p diagonal matrix diag$(\gamma_1, ..., \gamma_p)$, π is the prior measure for $\gamma = (\gamma_1, ..., \gamma_p)^t$ and μ is the prior measure for σ^2. *Ishwaran et al.* [57] proposed this setup and developed optimal properties based on the prior choice of (β/γ).

A popular version of the spike-and-slab model, introduced by *George et al.* ([43],[44]), identifies zero and non-zero β_i's by using zero-one indicator variables γ_i and assuming a scale mixture of two normal distributions:

$$
(\beta_i/\gamma_i) \overset{ind}{\sim} (\mathbf{1} - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2), \quad i = 1, .., p
$$

The value for $\tau_i^2 > 0$ is some suitably small value, while $c_i > 0$ is some suitably large value. $\gamma_i = 1$ represents the β_i's which are significant, and these coefficients have large posterior hypervariances and large posterior β_i values. The opposite occurs when $\gamma_i = 0$. The prior hierarchy for $\boldsymbol{\beta}$ is completed by assuming a prior for γ_i. When τ_i^2 tends to zero we provide more masses on 0, as the prior for insignificant βs. The prior distribution for the regression coefficients can then be written as:

$$(\beta_i/\gamma_i) \overset{ind}{\sim} (\mathbf{1} - \gamma_i)I_0 + \gamma_i N(0, \nu^2)$$

with I_0 point mass at 0 coefficients; and ν^2 is the limit for $c_i^2\tau_i^2$ when τ_i^2 tends to zero and c_i^2 is large enough.

4.2.2 Bayesian Group lasso

We discussed extensively about Bayesian Group Lasso in introduction. The form of Bayesian Group lasso we extensively worked with initiated in *Xu and Ghosh, 2015*[128]. A multivariate zero-inflated mixture prior can bring sparsity in group level which is elaborately discussed in *Xu and Ghosh, 2015*[128]. The following hierarchical structure with independent spike-and-slab prior for each β_g:

$$\begin{aligned}
Y|X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I) \\
\beta_g|\tau_g^2, \sigma^2 &\sim (1 - \pi_0)N_{m_g}(0, \sigma^2\tau_g^2 I_{m_g}) + \pi_0\delta_0(\beta_g), \quad g = 1, .., G \\
\tau_g^2 &\sim Gamma\left(\frac{m_g + 1}{2}, \frac{\lambda^2}{2}\right), \quad g = 1, .., G \\
\sigma^2 &\sim IG(\alpha, \gamma) \\
\pi_0 &\sim Beta(a, b)
\end{aligned}$$

where $\delta_0(\beta_g)$ denotes point mass at $\mathbf{0}$. The mixing probability π_0 can be defined as a function of the number of predictors to impose more sparsity as the feature size increases. The choice of λ is very critical for *Xu and Ghosh's* prior setup. Large values of λ produce biased estimates, while very

small λ values impose diffuse distribution for the slab part. *Xu and Ghosh, 2015*[128] mentioned an empirical Bayes approach to estimate λ. Due to intractability of marginal likelihood, they proposed a Monte Carlo EM algorithm for the estimation of λ. Moreover, they showed theoretically and numerically that the median thresholding of posterior β_g samples provides exact zero estimates for insignificant group predictors.

4.3 Functional smoothing for longitudinal data

Classification with the selection of significant functional predictors is challenging. Researchers commonly observe high correlation values between functional predictors. In this dissertation, we work with the assumptions of independence between predictors; hence, later we propose a corresponding prior in the coefficient space.The main advantage of using functional predictors is that it allows us to measure time trends present in data . We start our methodology by smoothing functional observations using a cubic basis spline. We restrict our data set to patients with at least four time period observations, such that smoothed curves are comparable. *Gareth M. James* [61] used a similar approach to obtain the estimates of a generalized linear model with functional predictors.

Let us assume that we observe n patients with their functional observations and each patient has p functions. We assume that not all p functional observations are important. Let $x_{ij}(t)$ be the jth function observed from the ith patient. Let T be the compact domain of $x_{ij}(t)$ and $x_{ij}(t) \in \mathcal{L}^2[T]$. With the functional predictors $(x_{i1}(t), .., x_{ip}(t))$, we assume that we have binary response variable y_i which takes value 0 and 1. We also assume that the predictors have been centered in this work, so that we can ignore the intercept term. Therefore, we have the following logistic regression equation:

$$\log\left\{\frac{P(y_i = 1|x_{i1}, .., x_{ip})}{1 - P(y_i = 1|x_{i1}, .., x_{ip})}\right\} = \sum_{j=1}^{p} \int_T x_{ij}(t)\beta_j(t)dt \tag{4.1}$$

Next, we construct an orthonormal basis $\phi_k(t)$ that can be used to decompose the functional

predictors and the corresponding logistic regression coefficients, such as

$$x_{ij}(t) = \sum_{k=1}^{q} c_{ijk}\phi_k(t), \quad \beta_j(t) = \sum_{k=1}^{q} \beta_{jk}\phi_k(t)$$

where c_{ijk} and β_{jk} are the coefficients of $x_{ij}(t)$ and $\beta_j(t)$ with respect to the k^{th} orthonormal basis $\phi_k(t)$. For notational convenience, we denote the basis coefficients as β_{jk}. These are different than the functional coefficients $\beta_j(t)$. We use cubic basis splines as the orthonormal basis for our simulation examples and real data applications. Hence, the choice of q completely depends on the number of internal knots used in basis spline constructions. The j^{th} component in equation (1) can thus be written as

$$\int_T x_{ij}(t)\beta_j(t)dt = \sum_{k=1}^{q} c_{ijk}\beta_{jk} = \mathbf{c}'_{ij}\boldsymbol{\beta}_j \tag{4.2}$$

To fit the discrete observations $x_{ij}(t)$, we assume that, at any given time t, instead of $x_{ij}(t)$, we observe $X_{ij}(t)$:

$$x_{ij}(t) = X_{ij}(t) + e(t)$$

where $e(t)$ is a zero-mean Gaussian process. We use the same basis function expansion for $X_{ij}(t)$ of the form

$$X_{ij}(t) = \sum_{k=1}^{q} c_{ijk}\phi_k(t) = \mathbf{c}'_{ij}\phi(t)$$

where $\phi(t)$ is the q-dimensional spline basis at time t for jth function, $\mathbf{c}_{ij}$ the q-dimensional spline coefficients for the jth predictor from ith patient. We use ordinary least square estimates for estimating spline coefficients. A simple linear smoother is obtained by minimizing the least squares criterion $||x_{ij} - \boldsymbol{\Phi}\mathbf{c}_{ij}||^2$ as

$$\hat{\mathbf{c}}_{ij} = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}' x_{ij} \tag{4.3}$$

Once the orthonormal basis coefficients have been estimated, we can combine (1), (2) and (3)

by plugging $\hat{x}_{ij}(t)$ in (1), which yields

$$\log\left\{\frac{P(y_i = 1|x_{i1},..,x_{ip})}{1 - P(y_i = 1|x_{i1},..,x_{ip})}\right\} = \sum_{j=1}^{p}\int_T \hat{\mathbf{c}}'_{ij}\phi(t)\beta_j(t)dt$$
$$= \sum_{j=1}^{p}\hat{\mathbf{c}}'_{ij}\boldsymbol{\beta}_j$$
$$= \mathbf{c}'_i\boldsymbol{\beta} \tag{4.4}$$

where $\boldsymbol{\beta}_j^T = \int_T \beta_j(t)\phi^T(t)dt$, the coefficient vector for the jth functional predictor. Here, $\mathbf{c}_i$ vector has its first element as 1 and rest of the spline coefficients for i^{th} patient, and $\boldsymbol{\beta}$ contains intercept of the model as its first element. We use no intercept form for our real data and simulation application where $\mathbf{c}'_i = (\hat{c}_{i1},..,\hat{c}_{ip})'$ does not have first element as 1 and $\boldsymbol{\beta}^{pqx1} = (\boldsymbol{\beta}_1^{Tqx1},..,\boldsymbol{\beta}_p^{Tqx1})^T$ has group structure with each group size=q. Our selection method drops the redundant β's and will select the important coefficient groups.

Functional principal component (FPC) analysis is another popular method that can be applied here. Instead of least square basis estimates, one can work with FPC scores for classification. *Zhu et al.* [136] also used FPC scores in their classification model, and they selected the functional predictors whose FPC scores were significant. *Müller* [93] extended the applicability of FPC analysis for modeling longitudinal data. Specifically, FPC scores can be used when we have few repeated and irregularly observed data points. In our functional smoothing method, we expanded the functional observation with spline basis functions and used the basis coefficients for classification. The same intuition can also be applied for FPC scores. For functional component analysis, we assume that longitudinal observations are from a smooth random function $X(t)$ and its mean function is $\mu(t) = E\ X(t)$ and covariance function $G(s,t) = cov(X(s), X(t))$. The covariance function can be represented as $G(s,t) = \sum_{k=1}^{\infty}\lambda_k\phi_k(s)\phi_k(t)$ where ϕ_k's are eigenfunctions and λ_k's are eigenvalues. Then, the underline process can be written as:

$$X(t) = \mu(t) + \sum_{k=1}^{\infty}\xi_k\phi_k(t),$$

where ξ_k's are frequently referred to as FPC scores. These scores can be used later in the classification model. We do not work with an infinite number of scores; instead, the above sum is

approximated with a finite K that explains the majority of the variance in functional observations. For most cases, the first two FPC scores are enough to build a good classification model. In this chapter, we work with the basis spline smoothing method due to its ease of implementation in statistical software. In R, we have the *splines* package, which fits cubic basis splines on longitudinal data with equally placed knots. We do not investigate any findings using FPC scores instead of basis spline coefficients, as our main focus is on the classification algorithm, and basis spline coefficients work very well for our classification model.

4.4 Simultaneous Classification of binary response with selection of functional predictors

4.4.1 Classification using Pólya-Gamma Augmentation

In the following, we discuss *Polson et al.'s* [97] algorithm; these authors showed how a Gaussian variance mixture distribution with a Pólya-gamma mixing density can approximate logit likelihood. We start by defining Pólya-gamma density-

Random variable $X \sim PG(b, c)$, a Pólya-gamma distribution with parameters $b > 0$ and $c \in \mathfrak{R}$, if

$$X \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + \frac{c^2}{4\pi^2}},$$

where $g_k \sim Gamma(b, 1)$ are independent gamma random variables and $\stackrel{d}{=}$ means equality in distribution.

Polson et al.'s [97] main result parametrized the log-odds of logistic likelihood as mixtures of Gaussian with respect to Pólya-gamma distribution. The fundamental integral result, which is easily integrated into the Gaussian prior hierarchy is that, for $b > 0$ -

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\frac{\omega\psi^2}{2}} p(\omega) d\omega \tag{4.5}$$

where $\kappa = a - b/2$ and $\omega \sim PG(b, 0)$. The introduction of latent variables $(\omega_1, .., \omega_n)$ later helped us in deriving conjugate posterior distribution. R package *BayesLogit* has an efficient algorithm to sample from Pólya-gamma distribution and it was proposed by *Windle et al.,2014*[126].

4.4.2 Selection using Bayesian Group Lasso

As we discussed in the Section 2.2, *Meier et al.*[87] developed group lasso for logistic regression in a frequentist setup. In our model, we have p number of functional predictors $(x_{i1}(t), .., x_{ip}(t)$ with binary response $y_i \in \{0, 1\}$, each group has q levels. We can write our model as-

$$\log\left\{\frac{P(y_i = 1|x_{i1}(t), .., x_{ip}(t))}{1 - P(y_i = 1|x_{i1}(t), .., x_{ip}(t))}\right\} = \sum_{j=1}^{p} \mathbf{c}'_{ij}\boldsymbol{\beta_j}$$
$$= \eta_\beta(c_i)$$

According to *Meier et al.*[87] method, the logistic group lasso estimator with basis spline coefficients would look like

$$\hat{\beta}_{GL} = \min_{\beta}\left\{-l(\beta) + \lambda \sum_{j=1}^{p} \sqrt{q}||\beta_j||_2\right\}$$

where $l(\beta) = \sum_{i=1}^{n}(y_i\eta_\beta(c_i) - \log(1 + \exp\{\eta_\beta(c_i)\})$ is the log-likelihood function.

Before moving on to our proposed Bayesian method, we want to mention a similar model presented by *Zhu et al.* [136]: they used latent variables for Bayesian logistic regression, and FPC scores represented the functional predictors. They proposed a normal prior for the concatenation coefficients, which is the same as our coefficients β_js.

Now, motivated by *Polson et al.*'s [97] integral result, we construct a Bayesian prior formulation targeted to handle binary logistic regression. Equation 4.4 has a Bernoulli likelihood function with logit link. We propose a spike-and-slab prior motivated by *Xu and Ghosh, 2015*[128] with a zero-inflated mixture prior, which helps us in selecting the important group coefficients. As previously described, we introduce latent variables $(\omega_1, .., \omega_n)$ to take advantage of the integral

identity described in equation (5). Our prior setup is

$$
\begin{aligned}
y_i|c_i,\beta &\sim \ Bernoulli\left(\frac{\exp(c_i^T\beta)}{1+\exp(c_i^T\beta)}\right), \quad i=1,..,n \\
\omega_i &\sim PG(1,0), \quad i=1,..,n \\
\beta_j|\tau_j^2,\pi_0 &\sim (1-\pi_0)N_q(0,\tau_j^2 I_q)+\pi_0\delta_0(\beta_j), \quad j=1,...,p \\
\tau_j^2|\lambda^2 &\sim Gamma(\frac{q+1}{2},\frac{\lambda^2}{2}), \quad j=1,...,p \\
\pi_0 &\sim Beta(a,b)
\end{aligned} \tag{4.6}
$$

Gibbs Sampler:

The likelihood for ith observation is:

$$
\begin{aligned}
L_i(\beta) &= \frac{(e^{c_i^T\beta})^{y_i}}{1+e^{c_i^T\beta}} \\
&\propto \exp\{\kappa_i c_i^T\beta\}\int_0^\infty \exp\left\{-\frac{\omega_i(c_i^T\beta)^2}{2}\right\}p(\omega_i)d\omega_i, \quad from\ equation\ (5)
\end{aligned}
$$

where $\kappa_i = y_i - 0.5$ and $\omega_i \sim PG(1,0)$. If we consider all n independent observations, given ω_i we can write the joint likelihood as-

$$
\begin{aligned}
\prod_{i=1}^n L_i(\beta|\omega_i) &= \prod_{i=1}^n \exp\left\{\kappa_i c_i^T\beta - \frac{\omega_i(c_i^T\beta)^2}{2}\right\} \\
&= \exp\left\{\frac{\omega_i}{2}(c_i^T\beta-\frac{\kappa_i}{\omega_i})^2\right\} = \exp\left\{-\frac{1}{2}(z-C\beta)^T\Omega(z-C\beta)\right\}
\end{aligned}
$$

where $z = (\frac{\kappa_1}{\omega_1},..,\frac{\kappa_n}{\omega_n})$ and $\Omega = diag(\omega_1,..,\omega_n)$.

Next, we combine the likelihood function with β prior, given $\omega = (\omega_1, .., \omega_n)$:

$$
\begin{aligned}
p(\beta, \tau^2, \pi_0 | Y, C, \omega) \propto & \exp\left\{-\frac{1}{2}(z - C\beta)^T \Omega (z - C\beta)\right\} \\
& \times \prod_{j=1}^{p} \left[(1-\pi_0)(\tau_j^2)^{-\frac{q}{2}} \exp\{-\frac{1}{2\tau_j^2}\beta_j^T \beta_j\} I_{(\beta_j \neq 0)} + \pi_0 \delta_0(\beta_j) \right] \\
& \times (\lambda^2)^{\frac{q+1}{2}} (\tau_j^2)^{\frac{q+1}{2}-1} e^{-\frac{\lambda^2 \tau_j^2}{2}} \\
& \times \pi_0^{a-1} (1-\pi_0)^{b-1}
\end{aligned}
$$

Due to the introduction of Pólya-gamma augmentation, we can derive a block Gibbs sampler with a posterior distribution of β_j's. The same method is derived in *Xu and Ghosh, 2015*[128] for continuous Y in linear model setup. The blocks Gibbs sampler was introduced by *Hobert et al.*[52]. To build this sampler, we start with some notations. Let $\beta_{(j)}$ denotes the β vector without j-th group,

$$\beta_{(j)} = (\beta_1^T, .., \beta_{j-1}^T, \beta_{j+1}^T, .., \beta_p^T)^T$$

and the corresponding design matrix can be written as:

$$C_{(j)} = (C_1, .., C_{j-1}, C_{j+1}, .., C_p)$$

C_j is the corresponding design matrix for β_j.

When $\beta_j \neq 0$:

$$
\begin{aligned}
p(\beta_j | rest) & \propto \exp\left\{-\frac{1}{2}(z - C_{(j)}\beta_{(j)} - C_j\beta_j)^T \Omega (z - C_{(j)}\beta_{(j)} - C_j\beta_j)\right\} \\
& \times \exp\left\{-\frac{1}{2\tau_j^2}\beta_j^T \beta_j\right\} \\
& \propto \exp\left\{-\frac{1}{2}\left[\beta_j^T (C_j^T \Omega C_j + \frac{1}{\tau_j^2} I_q)\beta_j - 2(z - C_{(j)}\beta_{(j)})^T \Omega C_j \beta_j\right]\right\} \\
& \propto \exp\left\{-\frac{1}{2}(\beta_j - A_j)^T B_j (\beta_j - A_j)\right\}
\end{aligned}
$$

where, $B_j = (C_j^T \Omega C_j + \frac{1}{\tau_j^2} I_q)$ and $A_j = B_j^{-1} C_j^T \Omega (z - C_{(j)}\beta_{(j)})$ Hence the posterior full conditional of β_j is a spike-and-slab distribution,

$$(\beta_j | rest) \sim (1 - l_j) N_q(A_j, B_j^{-1}) + l_j \delta_0(\beta_j), \quad j = 1, .., p \tag{4.7}$$

where $l_j = p(\beta_j = 0 | rest)$. Now we will find the probability l_j :

$$\begin{aligned}
l_j &= p(\beta_j = 0 | rest) \\
&= \frac{p(\beta_j = 0, y | C, \omega, \tau_j^2, \pi_0)}{\int_{\beta_j \neq 0} p(\beta_j, y | C, \omega, \tau_j^2, \pi_0) d\beta_j} \\
&= \frac{p(y | \beta_j = 0, C, \omega, \tau_j^2, \pi_0) p(\beta_j = 0 | \tau_j^2, \pi_0)}{p(y | \beta_j = 0, C, \omega, \tau_j^2, \pi_0) p(\beta_j = 0 | \tau_j^2, \pi_0) + \int_{\beta_j \neq 0} p(y | \beta_j \neq 0, C, \omega, \tau_j^2, \pi_0) p(\beta_j \neq 0 | \tau_j^2, \pi_0) d\beta_j} \\
&= \frac{M \pi_0}{M \pi_0 + N(1 - \pi_0)}
\end{aligned}$$

where $\pi_0 = p(\beta_j = 0 | \tau_j^2, \pi_0)$,

$$M = p(y | \beta_j = 0, C, \omega, \tau_j^2, \pi_0) = \exp\left\{ -\frac{1}{2} (z - C_{(j)}\beta_{(j)})^T \Omega (z - C_{(j)}\beta_{(j)}) \right\}$$

$$\begin{aligned}
N &= \int_{\beta_j \neq 0} p(y | \beta_j \neq 0, C, \omega, \tau_j^2, \pi_0) d\beta_j \\
&= \int_{\beta_j \neq 0} \exp\left\{ -\frac{1}{2} (z - C\beta)^T \Omega (z - C\beta) \right\} (2\pi\tau_j^2)^{-\frac{q}{2}} e^{-\frac{\beta_j^T \beta_j}{2\tau_j^2}} d\beta_j \\
&= M \times \int_{\beta_j \neq 0} \exp\left\{ -\frac{1}{2} \left[\beta_j^T (C_j^T \Omega C_j + \frac{1}{\tau_j^2} I_q) \beta_j - 2\beta_j^T C_j^T \Omega (z - C_{(j)}\beta_{(j)}) \right] \right\} (2\pi\tau_j^2)^{-\frac{q}{2}} d\beta_j \\
&= M \times (\tau_j^2)^{-\frac{q}{2}} \exp\left\{ \frac{1}{2} A_j^T B_j A_j \right\} \int_{\beta_j \neq 0} (2\pi)^{-\frac{q}{2}} \exp\left\{ -\frac{1}{2} (\beta_j - A_j)^T B_j (\beta_j - A_j) \right\} d\beta_j \\
&= M \times (\tau_j^2)^{-\frac{q}{2}} \exp\left\{ \frac{1}{2} A_j^T B_j A_j \right\} |B_j|^{-\frac{1}{2}}
\end{aligned}$$

Hence,

$$l_j = \frac{\pi_0}{\pi_0 + (1 - \pi_0)(\tau_j^2)^{-\frac{q}{2}} |B_j|^{-\frac{1}{2}} \exp\left\{ \frac{1}{2} A_j^T B_j A_j \right\}} \tag{4.8}$$

The posterior full conditional distributions of other parameters are stated below, and the derivations of the posteriors are described in appendix.

$$\left(\frac{1}{\tau_j^2}|rest\right) \sim \begin{cases} Inverse - Gamma(\frac{q+1}{2}, \frac{\lambda^2}{2}), & \text{if } \beta_j = 0 \\ Inverse - Gaussian(\frac{\lambda}{||\beta_j||_2}, \lambda^2), & \text{if } \beta_j \neq 0 \end{cases} \tag{4.9}$$

for all $j = 1,,.,p$. Let, G_j define whether a certain group is selected or not

$$G_j = \begin{cases} 1, & \text{if } \beta_j \neq 0 \\ 0, & \text{if } \beta_j = 0 \end{cases}$$

Then,

$$(\pi_0|rest) \sim Beta\left(p - \sum_{j=1}^{p} G_j + a, \sum_{j=1}^{p} G_j + b\right) \tag{4.10}$$

We will sample our augmented variables $\omega = (\omega_1, .., \omega_n)$ using the posterior samples of β:

$$(\omega_i|\beta) \sim PG(1, \mathbf{c}_i'\beta), \quad i = 1, .., n \tag{4.11}$$

Finally, we are left with the values of λ. λ is the most crucial parameter for our model and should be treated carefully. A large λ shrinks most of the group coefficients towards zero and produces biased estimates. In our real data analysis, we try to control the λ value by assigning a different range of values. *Xu and Ghosh, 2015*[128] proposed a Monte Carlo EM algorithm for estimating λ. The following is the k^{th} EM update for λ from their paper-

$$\lambda^{(k)} = \sqrt{\frac{p(q+1)}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}\left[\tau_j^2|y\right]}}$$

The expected value of $\tau_j^2|y$ for binary response y is intractable. In other words, this expected value can be calculated by taking mean of posterior samples of τ_j^2.

4.5 Median thresholding and Theoretical properties

4.5.1 Marginal Prior for $\boldsymbol{\beta}_j$:

We first study the marginal priors of $\boldsymbol{\beta}_j$'s to examine the theoretical properties of the Bayesian group lasso estimators. We aim to establish the connection between $\boldsymbol{\beta}_j$ group priors and existing

Group Lasso penalization methods. We integrate out τ_j^2 from $\boldsymbol{\beta}_j$ priors. The marginal priors for $\boldsymbol{\beta}_j$'s are calculated based on *Xu and Ghosh, 2015*[128] work with extension to binary response instead of continuous response. For $\boldsymbol{\beta}_j \neq \mathbf{0}$:

$$
\begin{aligned}
p(\beta_j/\pi_0) &\propto \int_{\tau_j^2} p(\beta_j/\tau_j^2, \pi_0) p(\tau_j^2) d\tau_j^2 \\
&\propto \int_0^\infty (1-\pi_0)(\tau_j^2)^{-\frac{q}{2}} \exp\left\{-\frac{1}{2\tau_j^2}\beta_j^T\beta_j\right\} (\lambda^2)^{\frac{q+1}{2}} (\tau_j^2)^{\frac{q+1}{2}-1} \exp\left\{-\frac{\lambda^2}{2}\tau_j^2\right\} d\tau_j^2 \\
&\propto (1-\pi_0)(\lambda^2)^{\frac{q+1}{2}} \exp\left\{-\lambda||\beta_j||_2\right\} \int_0^\infty (\alpha_j^2)^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\frac{\beta_j^T\beta_j}{\alpha_j^2}\left[\alpha_j^2 - \frac{\lambda}{||\beta_j||_2}\right]^2\right\} d\alpha_j^2 \\
&\propto (1-\pi_0)\left(\lambda^2\right)^{\frac{q}{2}} \exp\left\{-\lambda||\beta_j||_2\right\}
\end{aligned}
$$

where $\alpha_j^2 = \frac{1}{\tau_j^2}$. The marginal prior for $\boldsymbol{\beta}_j$'s are also spike-slab with with point mass at 0 and the slab part consists of a Multi-Laplace distribution which same as the one considered in Bayesian group lasso (*Kyung et al.,2010*[69]) or matches with penalization mentioned in Bayesian Adaptive Lasso (*Leng et al.,2014*[73]).

$$
\beta_j/\pi_0 \;\sim\; (1-\pi_0)M - Laplace\left(\mathbf{0}, \frac{1}{\lambda}\right) + \pi_0\delta_0(\boldsymbol{\beta}_j) \tag{4.12}
$$

Combining spike and slab both, the components facilitates variable selection at group level and shrinks the coefficients of the selected groups.

4.5.2 Median thresholding as posterior estimates

We previously discussed obtaining the selected group coefficient estimation through median thresholding of the MCMC sample. *Xu and Ghosh, 2015*[128] generalized the median thresholding proposed by *Johnstone et al.,2004*[64] for multivariate spike-and-slab prior. *Johnstone et al.,2004*[64] showed median thresholding, under a spike-and-slab prior for normal means, has some desirable properties. In this section, we generalize this idea to a binary classification problem and show that the posterior median estimator serves as group variable selection by obtaining a zero coefficient for the redundant groups. We further demonstrate the posterior median as a soft thresholding estimator

that is consistent in model selection and has an optimal asymptotic estimation rate.

Focusing on only one group, then *Xu and Ghosh, 2015*[128] proposed the following theorem on Median thresholding:

$$\boldsymbol{Z}_{mx1} \sim f(\boldsymbol{z}-\boldsymbol{\mu})$$
$$\boldsymbol{\mu} \sim \pi_0\delta_0(\boldsymbol{\mu}) + (1-\pi_0)\gamma(\boldsymbol{\mu})$$

where $\boldsymbol{Z}$ is an m-dimensional random variable, and $\gamma(.)$ and $f(.)$ are both density functions for m-dimensional random vectors. $f(\boldsymbol{t})$ is maximized at $\boldsymbol{t}=0$. Let $Med(\mu_i|\boldsymbol{z})$ denote the marginal posterior median of μ_i given data. By definition,

$$c = \frac{\int f(-\nu)\gamma(\nu)d\nu}{f(\mathbf{0})} \leq \frac{\int f(\mathbf{0})\gamma(\nu)d\nu}{f(\mathbf{0})} = 1$$

Theorem 1: Suppose $\pi_0 > \frac{c}{1+c}$, then there exists a threshold $t(\pi_0) > 0$, such that when $||z||_2 < t$,

$$Med(\mu_i|z) = 0, \quad for\ any\ 1 \leq i \leq m$$

Next, we focus on our problem setup. If we assume β_j follows a Gaussian prior, $\beta_j \sim N(0, B_j)$ and the design matrix satisfies the condition $C_j^T \Omega C_{(j)} = 0$. Then the posterior estimates of $\beta_j|rest$ is:

$$\hat{\beta}_j = \beta_j|rest \sim N(\mu_j, \Sigma_j)$$
$$\Sigma_j = (C_j^T \Omega C_j + B_j^{-1})^{-1}$$
$$\mu_j = \Sigma_j C_j^T \Omega z$$

According to theorem 1, assuming $\pi_0 > \frac{c}{1+c}$, then there exists $t(\pi_0) > 0$, such that the marginal posterior median of β_{jk} under prior (6) satisfies

$$Med(\beta_{jk}|\hat{\beta}_j) = 0 \quad for\ any \quad 1 \leq k \leq q$$

when $||\hat{\beta}_j||_2 < t$. We can interpret this result in the context of the same explanation provided by *Xu and Ghosh, 2015*[128]: the median estimator of the j-th group of regression coefficients is zero when the norm of the posterior estimates under any other prior distribution is less than a certain threshold.

Posterior Median as soft thresholding:

We assume that $C_j^T \Omega C_j = nI_q$ and C matrix is group wise Orthogonal with $C_j^T \Omega C_{(j)} = 0$. We are considering the model defined in (6) with fixed $\tau_{j,n}^2$ and it depends on n. In this set-up, the posterior distribution of β_j will be similar to the one derived in the previous section:

$$\beta_j|C, y, \omega \ \sim \ (1 - l_j)N_q\left(\frac{1}{n}(1 - D_{j,n})C_j^T \Omega z, \frac{1}{n}(1 - D_{j,n})I_q\right) + l_j\delta_0(\beta_j),$$

where $D_{j,n} = \frac{1}{1+n\tau_{j,n}^2}$ and,

$$l_j = \frac{\pi_0}{\pi_0 + (1 - \pi_0)(1 + n\tau_{j,n}^2)^{-\frac{q}{2}} \exp\left\{\frac{1}{2n}(1 - D_{j,n})||C_j^T \Omega z||_2^2\right\}}$$

Then, the marginal posterior distribution for $\beta_{jk}(1 \leq k \leq q)$ conditional on the observed data is a spike-and-slab distribution,

$$\beta_{jk}|C, y, \omega \ \sim \ l_j\delta_0(\beta_{jk}) + (1 - l_j)N\left(\frac{1}{n}(1 - D_{j,n})C_{jk}^T \Omega z, \frac{1}{n}(1 - D_{j,n})\right)$$

where C_{jk} is the k-th vector of the C_jth group matrix. The corresponding soft thresholding estimator is

$$\hat{\beta}_{jk} \ = \ Med(\beta_{jk}|C, y, \omega) \ = \ sgn\left(C_{jk}^T \Omega z\right)\left(\frac{1}{n}(1 - D_{j,n})|C_{jk}^T \Omega z| - \frac{1}{\sqrt{n}}Q_j\sqrt{1 - D_{j,n}}\right)_+$$

where z_+ is the positive part of z and $Q_j = \Phi^{-1}\left(\frac{1}{2(1-\min(\frac{1}{2}, l_j))}\right)$. Our results also follow *Xu and Ghosh, 2015*[128]'s work to show the soft thresholding. One should especially note that the term $D_{j,n}$ depends on $\tau_{j,n}^2$ which controls the shrinkage factor.

Oracle Property:

Let $\boldsymbol{\beta}^0, \boldsymbol{\beta_j}^0, \beta_{jk}^0$ be the true values $\boldsymbol{\beta}, \boldsymbol{\beta_j}, \beta_{jk}$, respectively. The index vector of true model is $\mathcal{A} = \left(I(||\boldsymbol{\beta_j}||_2 \neq 0),\ j = 1, .., p\right)$, and the index vector model selected by certain thresholding estimator $\hat{\boldsymbol{\beta}}_j$ is $\mathcal{A}_n = \left(I(||\hat{\boldsymbol{\beta}}_j||_2 \neq 0),\ j = 1, .., p\right)$. Model selection consistency is attained if and only if $\lim_n P(\mathcal{A}_{n\to\infty} = \mathcal{A}) = 1$.

Theorem 2: Assume the following design exists, $C_j^T \Omega C_j = nI_q$. Suppose $\sqrt{n}\tau_{j,n}^2 \to \infty$ and $\log(\tau_{j,n}^2)/n \to 0$ as $n \to \infty$, for $j = 1, .., p$, then the median thresholding estimator has oracle property, that is, varaible selection consistency,

$$\lim_{n\to\infty} P(\mathcal{A}_n^{Med} = \mathcal{A}) = 1$$

The proof follows same as steps as the proof of Theorem 4 in *Xu and Ghosh, 2015*[128].

4.5.3 Posterior Consistency:

In this section, we conduct a theoretical investigation regarding the convergence of the group lasso estimator model to the true model. To show model consistency, we refer to the results and theorems mentioned in the paper titled "On the consistency of Bayesian variable selection for high dimensional binary regression and classification" by *Jiang,2006*[62]. In this paper, the author setup Bayesian variable selection similar to *Smith et al.,1996*[115] by introducing a selection indicator vector $\gamma = (\gamma_1, .., \gamma_p)$ where $\gamma_i = 0/1$. The corresponding prior setup is as follows:

$$\begin{aligned} y &= X\beta + \epsilon \\ \beta_\gamma &\sim N(0, c\sigma^2 \left(X_\gamma^T X_\gamma)^{-1}\right) \\ \gamma_i &\sim Bernoulli(\pi),\ \ i = 1, .., p \\ (\sigma^2|\gamma) &\sim 1/\sigma^2 \end{aligned}$$

We can establish a direct connection between our model and the above penalized regression. We re-parametrize the groups coefficient vector $\boldsymbol{\beta}_j = \gamma_j \boldsymbol{b}_j$ where $\gamma_j, j = 1, .., p$ is the selection indicator

0/1 valued. As in section 5.1 we have shown the marginal prior of β_j follows a Multi-Laplace distribution, we can place a Bernoulli prior in γ_j,

$$\begin{aligned} b_j|\lambda &\sim \ Multi - Laplace(0, \frac{1}{\lambda}) \\ \gamma_j &\sim \ Bernoulli(1-\pi_0), \ \ j = 1, .., p \end{aligned} \tag{4.13}$$

The marginal prior distribution of β_j is same as in equation (12).

Next, we study the asymptotic results as $n \to \infty$. Let y be the binary response and $\vec{c}$ is the corresponding basis coefficients for any given subject. Let the true model be of the form $\mu_o(c) = \frac{e^{\sum_{j=1}^{p_n} c_j^T \beta_j}}{1+\sum_{j=1}^{p_n} c_j^T \beta_j} = \psi(\sum_{j=1}^{p_n} c_j^T \beta_j)$, β_j is a $qX1$ vector with $p_n(\uparrow n)$ number of group vectors present in the model. As described by *Jiang,2006*[62], we assume that the data dimension satisfies $1 \prec p_n$ and $\log(p_n) \prec n$, where $a_n \prec b_n$ represents $a_n = o(b_n)$, or $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. We assume sparsity of the regression coefficients on the group level, i.e. $\lim_{n\to\infty} \sum_{j=1}^{p_n} ||\beta_j||_2 < \infty$, which implies that only a limited number of group coefficients are nonzero. We further assume $||c_j||_2 \leq 1, \ j = 1, .., p_n$ for simplicity.

We assume n i.i.d. observations. $D^n = (\vec{c}_1, .., \vec{c}_{p_n}, y_i)_{i=1}^n$ and $f_0 = \mu_0^y (1-\mu_0)^{1-y}$. Before we move forward with the results, we define the posterior estimator of the true density f_0 as-

$$\hat{f}_n(y, c) = \sum_{\gamma} \int_{\beta_\gamma} f(y, c|\gamma, \beta_\gamma) \pi_n(\beta_\gamma, \gamma|D^n) d\beta_\gamma$$

and we define the posterior estimate of μ_0 as

$$\hat{\mu}_n(c) = \sum_{\gamma} \int_{\beta_\gamma} \psi(c_\gamma^T \beta_\gamma) \pi_n(\beta_\gamma, \gamma|D^n) d\beta_\gamma.$$

We define the classifier as $\hat{C}_n(c) = I[\hat{\mu}_n(c) > 0.5]$, so that $\hat{C}_n(c)$ will be the validation tool for our algorithm's performance.

Next we define consistency using *Jiang's,2006*[62] description of density function, and measure the distance between two density functions with Hellinger distance $d_H(f, f_0) = \sqrt{\int \int (\sqrt{f} - \sqrt{f_0})^2 dxdy}$.

The below definitions are quoted from *Jiang's,2006*[62] article.

Definition 1: "Suppose D^n is i.i.d. sample based on density f_0. The posterior $\pi_n(.|D^n)$ is asymptotically consistent for f_0 over Hellinger neighborhood if for any $\epsilon > 0$,

$$\pi_n \left[f : d_H(f, f_0) \leq \epsilon | D^n \right] \xrightarrow{P} 1, \ \ as\ n \to \infty \ \ \ (Density\ Consistency)$$

" Next we define consistency in classification from *Jiang,2006*[62] paper in terms of how the misclassification error $E_{D^n} P[\hat{C}_n(c) \neq y|D^n]$ approaches the minimal error $P[C_0(c) \neq y]$, where $C_0(c) = I[\mu_0(c) > 0.5]$.

Definition 2: "Let $\hat{B}_n(c)$ be a classification rule obtained based on the observed data D^n. If $\lim_{n\to\infty} E_{D^n} P[\hat{B}_n(c) \neq y|D^n] = P[C_0(c) \neq y]$, then $\hat{B}_n(c)$ is called a consistent classification rule."

Combining Proposition 1 and Proposition 3 from *Jiang,2006*[62], under conditions I, S, and L, density consistency directly implies classification consistency. The proof follows by checking conditions I, S, and L from Jiang's(2006) paper[62], since our prior satisfies his prior setup. To have density consistency and classification consistency for posterior estimates, we need to check whether our prior setup follows Jiang's conditions. The motivation for the proof and the technique of checking conditions to establish the theorem were discussed in theses *Atreyee Majumder, 2017* [81] and *Guiling Shi, 2017*[111].

Condition I: (On inverse link function ψ) "Denote $w(u)$ as the log odds function $w(u) = \log[\psi(u)/(1-\psi(u))]$. The derivative of the log odds $w'(u)$ is continuous and satisfies the following boundaries condition when the size of the domain increases: $sup_{|u|\leq C}|w'(u)| \leq C^q$ for some $q \geq 0$, for all large enough C."

Condition S: (For prior π_n on small approximation set.) "There exists a sequence r_n increasing to infinity as $n \to \infty$, such that for any $\eta > 0$, and $\sum_{j\notin\gamma(r_n)} ||\beta_j||_2 \prec \epsilon_n^2$, we have

$\pi_n[\gamma = \gamma(r_n)] > e^{-cn\epsilon_n^2}$ and $\pi_n[\beta_\gamma \in M(r_n, \eta)|\gamma = \gamma(r_n)] > e^{-cn\epsilon_n^2}$, for all large enough n."

Condition L: (For prior π outside a large region) "There exist some $\bar{r}_n = o(n/lnp_n), \bar{r}_n \in [1, p_n]$, and some C_n satisfying $C_n^{-1} = o(1)$ and $lnC_n = o(n/\bar{r}_n)$, such that for some $c > 0$, $\pi_n[|\gamma| > \bar{r}_n] \leq \exp(-cn\epsilon_n^2)$, and $\pi_n\left(\bigcup_{j:\gamma_j=1} \left[||\beta_j||_2 > C_n\right] |\gamma\right) \leq \exp(-cn\epsilon_n^2)$ for all $|\gamma| \leq \bar{r}_n$, for all large enough n."

We checked for the conditions; corresponding proofs are in the appendix.

4.6 Simulation results

We assess the performance of our proposed simultaneous classification and selection methodology with simulated data sets. We apply our method to both simulated and real data. We compare the results from our Bayesian method with those from a frequentist group lasso selection method for binary response. To the best of our knowledge, no other Bayesian method reported in the literature is as convenient and efficient as the presently proposed method. The following section reports the method testing by creating three different examples with varying numbers of predictors. We generate a binary response with simulated functional predictors; there are a significant number of inessential predictors.

Example 1: We first generate functional predictors $x_{ij}(t)$ using a 10-dimensional Fourier basis $\phi_0(t) = 1$ and $\phi_k(t) = \sqrt{2}\cos(k\pi t)$, $k = 1, .., 9$, adding an error term. We work with a similar simulation set up mentioned in *Fan et al.*[33], as Fan's model setup is also based on functional predictors. We generate our predictors as follows:

$$x_{ij}(t_k) = \boldsymbol{\phi}(t_k)^T \boldsymbol{\theta}_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2), \quad \boldsymbol{\theta}_{ij} \sim N_{10}(0, I)$$

where $\boldsymbol{\phi}(t_k) = (\phi_0(t_k), \phi_1(t_k), .., \phi_{10}(t_k))'$. We take $\sigma = 0.5$ and we generate 200 i.i.d observations using 20 functional predictors. Each predictor is observed at 50 time points, and time points are equally distributed between 0 and 1. $\boldsymbol{\theta}_{ij}$ and ϵ_{ijk} are independently sampled. It is easier to understand the set up notationally as 'i' varies from 1 to 200, 'j' varies from 1 to 20 and 'k' varies from

1 to 50. We construct a cubic basis spline on $(0 = t_1, .., t_{50} = 1)$ with four internal knots equally spaced at 20%, 40%, 60% and 80% quantiles. We use R-package '*splines*' and the '*bs*' function to construct the basis matrix $\boldsymbol{\phi}$. With 4 internal knots, plus intercept and degree=3, we end up having eight columns in the basis matrix for each predictor, i.e. q=8. To validate classification and selection performance, we use 75% of the observations as training data, and the remaining 25% for testing purposes. We repeat this process 100 times to limit sampling bias in data and concatenate all results considering the 100 repetitions.

Example 2 and 3: In example 2, we increase the number of predictors from 20 to 50 while maintaining 200 observations with 50 time points for each observation. The functional predictor generation in Example 3 follows the same method as in Example 1, but generates 500 observations with 100 functional predictors and 20 time points for each observation. We use three internal knots to smooth the predictors.

In both cases, we chose the second and final predictor, i.e. $\beta_2(t)$and $\beta_p(t)$ as non-zero, and the rest of the coefficients are zero. We generate the binary response $y \in (0, 1)$ from a Bernoulli distribution using the set of pre-assigned β. In all of the examples, 75% of the data is used for training and 100 repetitions are used to normalize sampling bias. We obtain 20,000 Gibbs samples, and the first one-third of these samples are discarded as a burn-in period. All the parameter estimates are obtained using the remaining samples. As *Xu and Ghosh, 2015*[128] showed that median thresholding gives exact 0 estimates for the redundant group coefficients, we apply a posterior median on posterior samples to obtain β estimates. We choose a=1,b=1 as the initial parameter values for the prior distribution of π_0 and $\boldsymbol{\beta} = \mathbf{0}$ is used as the initial choice for the first iteration. Although we have p number of functional predictors, the number of coefficients we need to estimate is p*q. In Example 2, we have p=50, and with four internal knots for each function we obtain q=8. Hence, the number of coefficients we need to estimate is 400 using 200 observations. From this perspective, our algorithm is applicable to "large p, small n" conditions. The simulation results are presented

below.

Example1 Results: We obtain a 100% true positive rate and a 0% false positive rate in terms of selection, i.e. the two nonzero coefficients are captured in all 100 iterations. Moreover, none of the predictors that originally had zero coefficients are selected. In terms of classification, our method shows 97% sensitivity, 93% specificity, 95% accuracy, and AUC=0.99. Below are the rejection probability plots for $\beta_2(t)$ and $\beta_1(t)$, of which the first is nonzero and the second is zero in the true model. In addition, we plot the posterior median estimates of the coefficient function with respect to its true values. The ROC curve establishes the differentiating power of our method.

Example 2 and 3 Results: In Example 2, we obtain a 100% true positive rate and a 0.73% false positive rate out of 100 repetitions, with 97% sensitivity and 95% specificity. In Example 3, we achieve a 100% true positive rate and a 0% false positive rate with 98% sensitivity and 97% specificity. We compare our simulation results with those of frequentist group lasso for logistic regression for all the setups above. Our methodology yields the best results in terms of classifying subjects into the right class, far exceeding frequentist group lasso. Although the frequentist group lasso approach successfully identifies the true significant predictors for the model, it also selects many redundant functional predictors that have zero effect on the true model. The false selection of predictors in the model is very high compared to that of our algorithm. The table below summarizes the numerical results of all three aforementioned examples, with comparisons to frequentist group lasso for logistic regression.

4.7 Application on ADNI MRI data

This section reports the results of the application of our proposed method to ADNI data. The MRI data used in all analyses was downloaded from the ADNI database (`http://www.adni-info.org/`). The fundamental goal of ADNI is to develop a large, standardized neuroimaging database

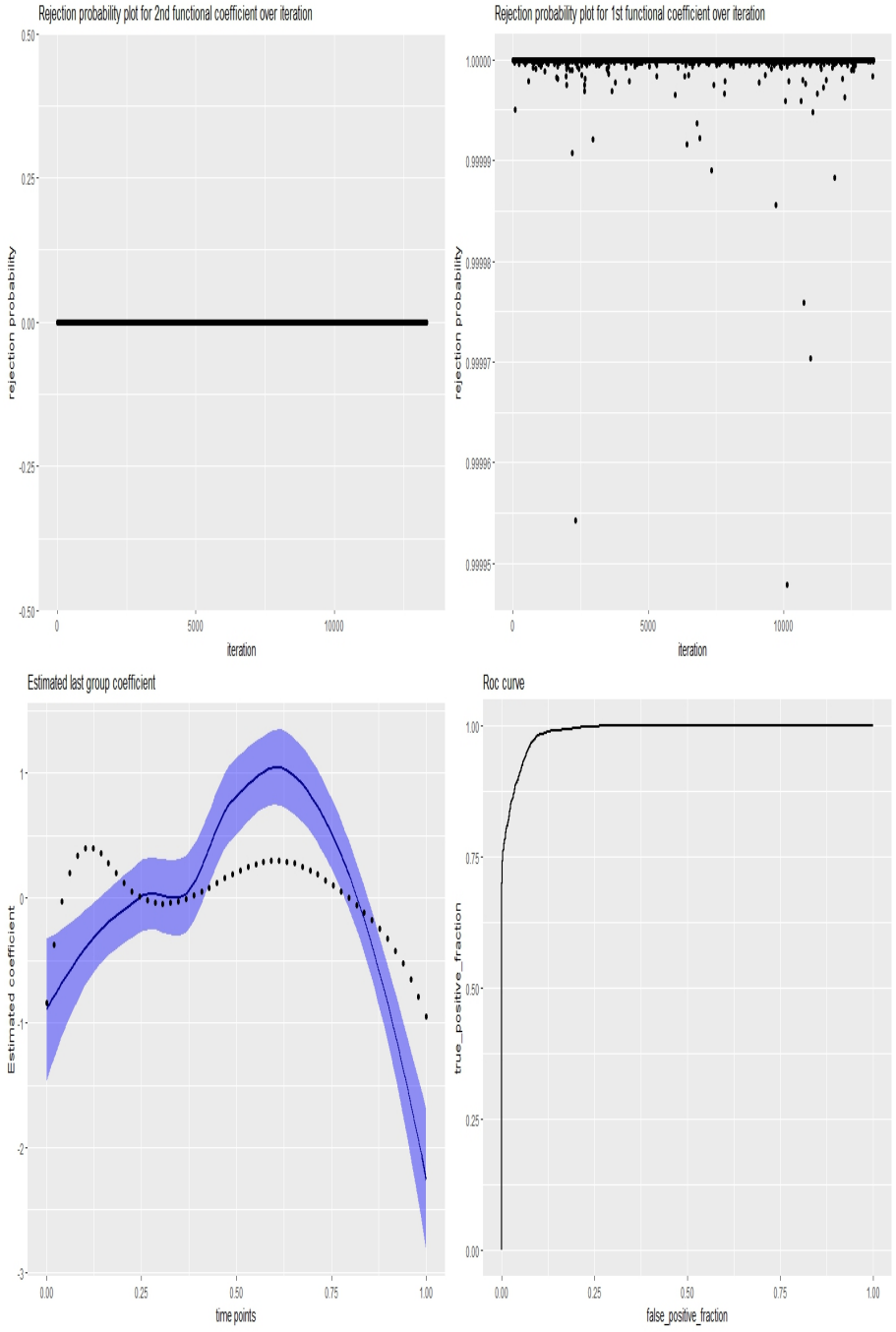


Figure 4.1: Plots based on Example 1

Table 4.1: Classification and selection performance Table

	Bayesian Classification with Bayesian Group Lasso				
	Sensitivity	Specificity	TPR	FPR	-2 Log likelihood
Example 1 n=200 p=20 t=50	0.97 (0.01)	0.93 (0.01)	1 (0)	0 (0)	3.65
Example 2 n=200 p=50 t=50	0.97 (0.01)	0.95 (0.01)	1 (0)	0.0073 (0.05)	0.219
Example 3 n=500 p=100 t=20	0.98 (0.001)	0.97 (0.01)	1 (0)	0 (0)	6.87
	Logistic regression with frequentist group lasso				
	Sensitivity	Specificity	TPR	FPR	-2 Log likelihood
Example 1 n=200 p=20 t=50	0.92 (0.01)	0.86 (0.01)	1 (0)	0.114 (0.04)	69.23
Example 2 n=200 p=50 t=50	0.81 (0.01)	0.88 (0.01)	1 (0)	0.34 (0.05)	53.66
Example 3 n=500 p=100 t=20	0.91 (0.001)	0.94 (0.001)	1 (0)	0.05 (0.02)	190.23

Notes: Simulation result comparisons between Bayesian and Frequentist methods

with strong statistical power for research on potential biomarkers in AD incidence, diagnosis, and disease progression. ADNI data available at this time include three projects: ADNI-1, ADNI-GO, and ADNI-2. Starting in 2004, ADNI-1 collected prospective data on cognitive performance, brain structure, and biochemical changes every 6 months. Participants in ADNI-1 included 200 CN, 200 MCI, and 400 AD patients. Then, starting in 2009, ADNI-GO continued the longitudinal study of

the existing patients from ADNI-1 and established a new cohort that included early MCI patients, who were enrolled to identify biomarkers manifesting at earlier stages of the disease. ADNI-GO and ADNI-2 together contain additional MRI sequences plus perfusion and diffusion tensor imaging. The volumetric estimation for our data set was performed using FreeSurfer by the UCSF/SF VA Medical Center.

Considerable research has been conducted to develop automatic approaches for patient classification into different clinical groups, with many ADNI studies identifying ROIs associated with different disease stages. A support vector machine (SVM) is a primary tool utilized in many studies to evaluate the patterns in training data sets and to create classifiers to identify new patients. *Fan et al.*[34] used neuroimaging data to create a structural phenotypic score reflecting brain abnormalities associated with AD. In classifying AD vs CN, a positive score in their framework identified AD-like structural brain patterns. Their classifier obtained 94.3% accuracy in AD vs CN, although their approach used only left and right whole brain volumes as potential predictors. Some researchers have used Bayesian statistical methods in studying Alzheimer's data. *Shen et al.*[110] employed a sparse Bayesian learning method, which they named automatic relevance determination (ARD) and predictive ARD, to classify AD patients. This method outperformed an SVM classifier. *Yang et al.*[130] proposed a data-driven approach to the automatic classification of MRI scans based on disease stages. Their methodology was broadly divided into two parts. First, they extracted the potentially classifying features from normalized MRI scans using independent component analysis. Next, the separated independent coefficients were applied for the SVM classification of patients. In contrast to this approach, our proposed method selects important components and classifies patients simultaneously. Moreover, we consider multiple brain sub-regions to identify those potential regions whose longitudinal trajectories are specifically related to AD. Another seminal paper by *Jack et al.*[58] used MRI-based measurements of hippocampal volume to assess the future risk of conversion from MCI to AD. A bivariate model included hippocampal volume and other factors like age and APOE genotype, but only hippocampal volume was identified as significant. *Wang*

et al.[123] employed a functional modeling approach using Haar wavelets and lasso regularization to find ROIs in voxel-level data. In that approach, large Haar wavelet coefficients were related to most important features, with a sparse structure among redundant features. The majority of these methods are based on SVM classification, which often uses kernel-based methods for functional smoothing. *Casanove et al.*[19] utilized a penalized logistic regression approach, and they calculated estimates using coordinate-wise descent optimization techniques from the GLMNET library. Similarly, our method employs penalized logistic regression with group lasso penalty. However, our approach differs in its use of both functional predictors and a custom algorithm developed in-house.

We consider the longitudinal volume of various brain regions, such as the Para hippocampal gyrus, cerebellar cortices, entorhinal cortex, fusiform gyrus, and precuneus, among many others. Although the accessed ADNI data set includes corresponding volume, surface area, and cortical thickness information, we work with only the volume information to acquire uniformity over longitudinal predictors. Because the brain is divided into right and left hemispheres, the data includes sub-regional brain volumes for both hemispheres. Our main objective is to identify the brain subregions whose volumetric trajectories can differentiate AD patients from the normal aging control group. As mentioned in the introduction, dementia is associated with widespread brain atrophy, although the time course and magnitude of shrinkage varies across regions.

The initial sample includes 761 patients' data from the ADNI database, classified as AD, MCI, or CN throughout their visits for the study. We exclude all patients classified as MCI, and any AD or CN patients whose diagnostic status changed over time. This is because our model assumes that response does not depend on time. Of the remaining patients, we include those with data from at least four longitudinal measurement occasions. This yields 296 patients who have at least four data points and unchanging diagnoses of either AD or CN. The final sample is composed of 174 AD patients and 122 normally aging controls. All patients underwent a thorough initial

Table 4.2: Patients Baseline Characteristics

	AD	CN	p-value
n	174	122	
Age (Mean ± sd)	74.76 ± 7.23	75.61 ± 5.45	0.25
Gender (F/M)	69/105	55/67	0.35
MMSE (Mean ± sd)	25.43 ± 2.40	28.96 ± 1.17	<.0001
ADAS11 (Mean ± sd)	14.78 ± 5.44	5.95 ± 2.94	<.0001
APOE (+/-)	119/55	30/90	<.0001

Notes: Comparison of Baseline Age, Gender ratio, MMSE score, ADAS11 score and APOE ratio between AD and CN groups

clinical evaluation to measure baseline cognitive and medical scores, including MMSE, the 11-item Alzheimer Cognitive Subscale (ADAS11), and other standardized neuropsychological tests. In addition, at baseline, APOE genotyping information was obtained from patients. Longitudinal structural MRI scans were parcellated into sub-regional brain volumetric measurements. Our initial model includes 49 sub-regional brain volumes chosen by *Dr. Andrew Bender*, based on knowledge of the extant literature regarding atrophy patterns in AD. Although these 49 sub-regions are not assumed to change in uniform magnitude, the direction of change over time is hypothesized to be consistent (i.e., shrinking). Thus, the model includes 49 longitudinal predictors that we consider as functional predictors. We assume that not all predictors are potential candidates for classifying patients, and that the sparse assumption is valid. However, because some patients' visits were irregular, we do not have an equal number of time points across patients. We start by comparing the baseline measurements between the AD and CN groups, as shown in Table 1.

In the next stage, we smooth the longitudinal trajectories for the observed volumes of all brain sub-regions. A simple least squares approximation is sufficient, as we assume that the residuals of the true curve are independently and identically distributed with mean 0 and have constant variance. We use the cubic B-spline basis functions for spline smoothing of observed volumes. Three internal knots are used for spline smoothing with intercept, which gives us seven basis functions. We seek to ensure that the smoothed estimated curve is a good fit for each patient's observed curve. As

we do not have a large number of data points for each patient, we do not consider controlling for potential overfitting of our estimated curve. Besides least squares smoothing, functional principle component scores can also be used for this analysis.

Prior to analysis, we scale the brain volumes to the corresponding patient's brain ICV measurement by fitting a simple regression to adjust volume measurements for individual brain volume changes. The aim is to remove systematic variation in brain volumes due to differences in physical size. The formula we use is $ROI_{adj} = ROI_{vol} - \beta_0(ICV - ICV_{mean})$, where β_0 is the regression coefficient by regressing ROI_{vol} on ICV [103] [59]. We adjust or correct the volumes using the above method for each gender group: male and female. Next, we scale the corrected volumes between 0 and 1 to bring all brain regions onto the same scale. We then divide the data set into two parts: two-thirds of the patients are reserved for the training data set (n=198), and the rest are kept for testing (n=98). We gather the basis coefficients for each patient in the training data set and use them as predictors for classification. We initialize choice of β with all zero to start iterations. The π_0 probability has a Beta(a,b) distribution with a and b both set up as 1. As a first step, we examine λ using Pólya-Gamma transformation of our sample with a spike-slab penalty on the training data. After estimating λ, we evaluate the remainder of the algorithm on the training data with 30,000 MCMC samples. The first one-third of observations are left out as a burn-in period. We propose a spike-and-slab prior on the β coefficient, which transforms into posterior estimates of zero for most the functional predictors. We run our model 100 times with different training samples to nullify sampling bias in the training and test data. In the 100 iterations, the model does not consistently or uniformly select many of the brain sub-regions; therefore, we choose the brain regions that frequently appear as significant in each iteration. The median thresholding selects the left hippocampus, left lateral orbitofrontal cortex, and left posterior cingulate gyrus with 100% probability. Other brain regions that are selected as important are the right Para hippocampal gyrus, left caudate nucleus, left medial orbitofrontal cortex, left putamen, left superior temporal gyrus, left thalamus, right hippocampus, and right middle temporal gyrus. In Figure 2, we plot the brain

volume changes of the left hippocampus, left lateral orbitofrontal cortex, and left posterior cingulate gyrus over time. Orange and green signify the normal aging and dementia group, respectively. The bold thick line represents the mean curve for the corresponding group. The plot shows that there are significant differences in volume between the groups, and our model identifies these regions as significant. In Figure 3, we plot the acceptance probability of the MCMC sample for the left hippocampus and left lateral orbitofrontal brain regions.

The method classifies patients into the correct group with 77% accuracy. We achieve 72% sensitivity, 85% specificity, and a corresponding AUC of 0.87. We use the median predicted probability from the training sample as the threshold for classification validation. We also test the classification by adding clinical measurements such as the ADAS11 (11-item Alzheimer Cognitive subscale), MMSE scores, "CDRSB," "RAVLT immediate," and "RAVLT forgetting," measured over time. In this classification, we initially select longitudinal brain volumes that are significant, and then we add the clinical variables. We achieve very high classification measures of 97% accuracy, 97% sensitivity, and 98% specificity. If we ignore the MMSE score and run the model with the rest of the functional predictors, we observe similar classification accuracy. In all scenarios, we model diseased patients as 1 and CN as 0 for the interpretation of classification sensitivity/specificity.

In addition to finding functional models of longitudinal trajectories in sub-regional brain volumes to differentiate between the AD and normal groups, we also apply our method for MCI converters vs MCI nonconverters. We select patients who entered the study as MCI, and we assign the label of MCI nonconverter (MCI-nc) to those who did not transition to AD across all measurement occasions and a label of MCI converter (MCI-c) for any who did transition to AD. The total subsample includes 163 patients who were either MCI-c or MCI-nc. We use three-quarters of the patients to train our model. We note the significant brain ROIs that are selected after 100 iterations. Among the selected ROIs that contribute to classification are the right posterior cingulate gyrus, right superior parietal cortex, right thalamus, right isthmus cingulate gyrus, right fusiform

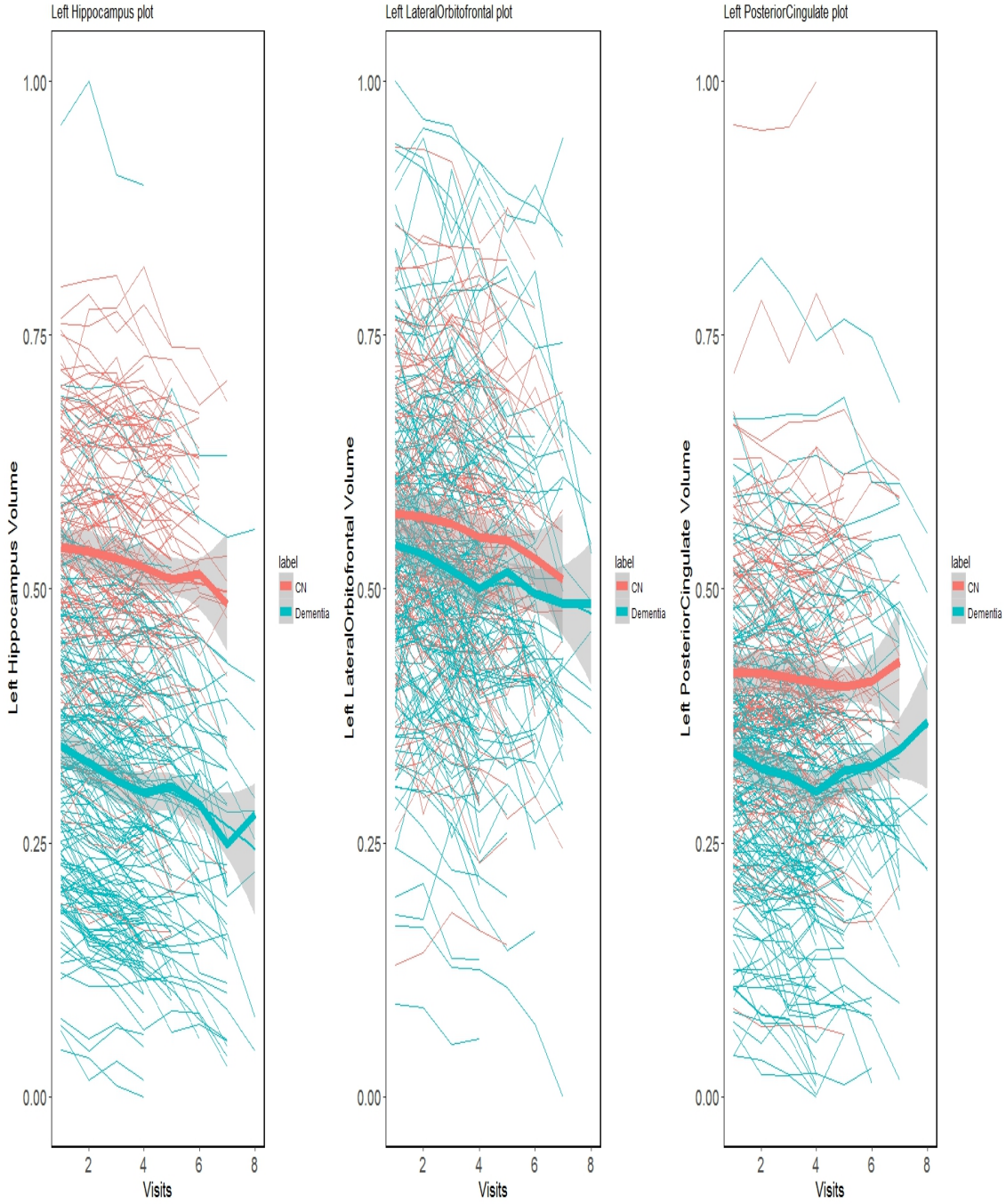


Figure 4.2: Brain volume changes of Left Hippocampus, Left Lateral Orbitofrontal cortex, and Left Posterior Cingulate over time for Normal and Dementia patients

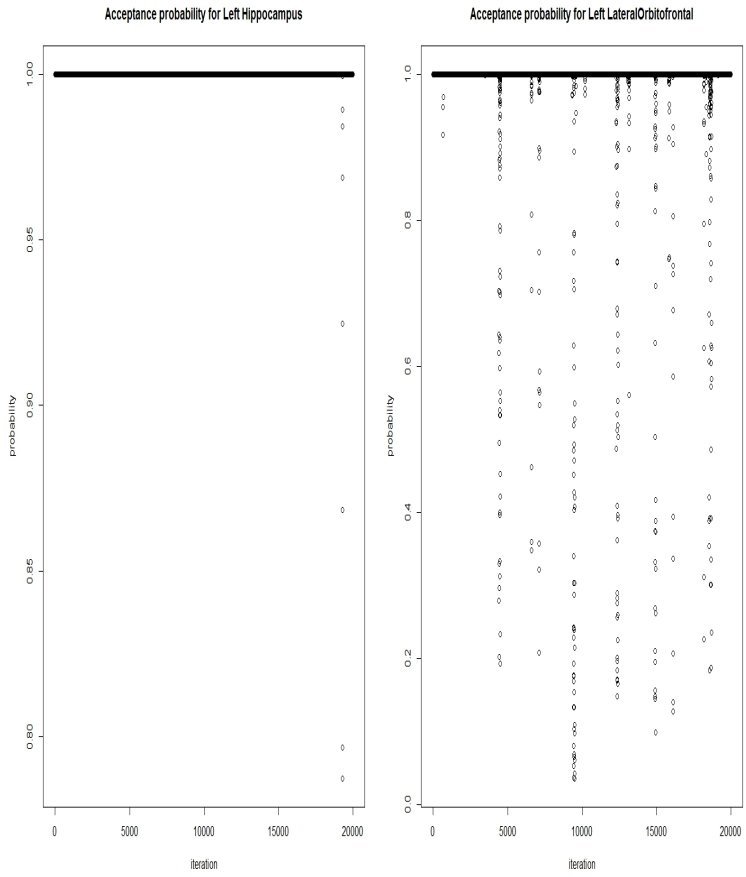
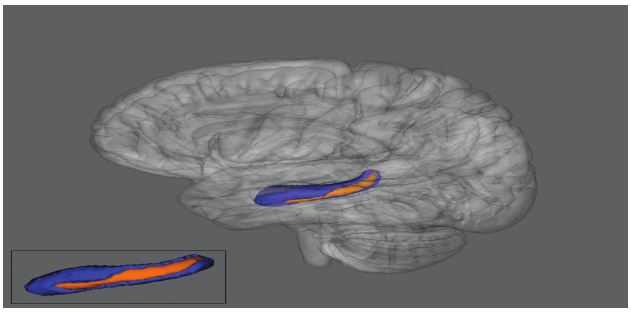


Figure 4.3: Acceptance probability of MCMC sample for Left Hippocampus and Left-Lateral Orbitofrontal brain regions

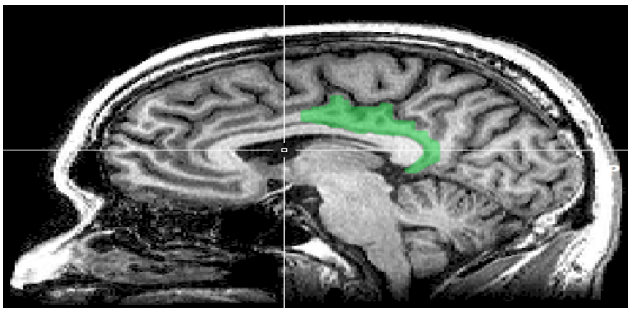
gyrus, left thalamus, and left precuneus. However, the classification performance is not as good as compared to the previous model: 62% accuracy and 0.66 AUC. The biological explanation for this result is critical to acknowledge. The mean difference of functional predictors between MCI-c vs MCI-nc is not significant for segmenting patients. Moreover, we also neglect some time points' data for this set of patients.

4.8 An alternate modeling Proposal:

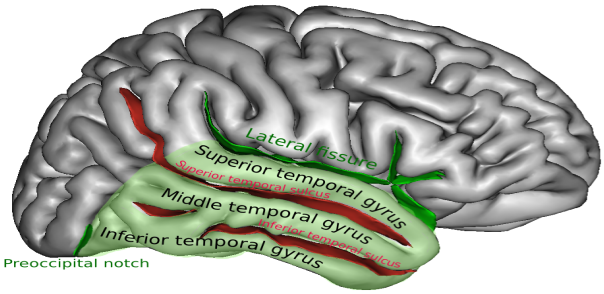
It is always important in penalization setups to protect strong signal estimations. We should control over-shrinkage of the selected group coefficients with lesser bias. Hence, we propose a horseshoe prior (*Carvalho et al.,2009*[17]) within the spike-slab structure. This setup provides a heavy-tailed distribution for the slab part. We propose a normal prior for the slab part with



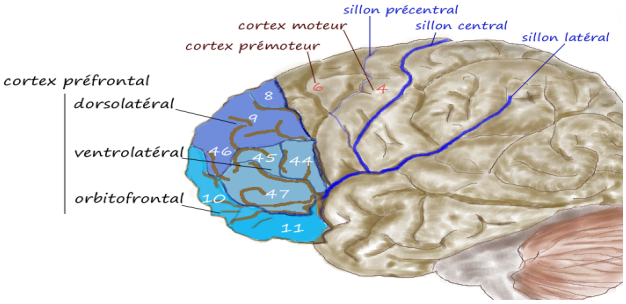
(a) Left Hippocampus[a]



(b) Posterior Singulate[a]



(c) Middle Temporal Gyrus[a]



(d) Left lateral orbitofrontal cortex[a]

Figure 4.4: Pictorial representation of selected brain ROI's discriminating diseased group from normal control

[a] Plot obtained from on-line resources

two hyper-parameters controlling local and global shrinkage. We compare the corresponding penalization on group coefficients with our main proposed penalization from the previous section. We contrast the results of applying both methods on the AD patients. Below, we report an alternative exploratory setup:

$$\begin{aligned}
\beta_j/\lambda_{j1},..,\lambda_{jq},\tau_j,\pi_0 &\sim (1-\pi_0)N_q(\mathbf{0},\tau_j^2\boldsymbol{D}_{\boldsymbol{\Lambda_j}}) + \pi_0\delta_0(\beta_j), \quad j=1,..,p \\
\lambda_{jk} &\sim C^+(0,1), \quad k=1,..,q \\
\tau_j &\sim C^+(0,1), \quad j=1,..,p
\end{aligned} \tag{4.14}$$

where $\boldsymbol{D_{\Lambda_j}} = diag(\lambda_{j1}^2, .., \lambda_{jq}^2)$. We can also develop an efficient Gibbs sampler to update the model parameters from the corresponding full conditional distributions:

$$\begin{aligned}
\beta_j/rest &\sim (1-l_j)N_q(\bar{\beta}_j, \Sigma_j) + l_j\delta_0(\beta_j), \quad j = 1, .., p \\
\Sigma_j^{-1} &= \left(C_j^T \Omega C_j + \frac{1}{\tau_j^2} \boldsymbol{D_{\Lambda_j}^{-1}} \right) \\
\bar{\beta}_j &= \Sigma_j C_j^T \Omega \left(z - C_{(j)}\beta_{(j)} \right) \\
l_j &= \frac{\pi_0}{\pi_0 + (1-\pi_0)(\tau_j^2)^{-\frac{q}{2}} |\Sigma_j^{-1}|^{-\frac{1}{2}} \exp\left\{ \frac{1}{2}\bar{\beta}_j (\Sigma_j^{-1})^T \bar{\beta}_j \right\}}
\end{aligned} \tag{4.15}$$

It is also important to sample the global and local scale parameters λ_{jk}'s and τ_j's efficiently. *Polson et al.,2014*[98] provided an efficient slice sampling scheme which can acquired for our setup. We define $\alpha_{jk} = \lambda_{jk}^{-2}$, and sample $u_{jk}/\alpha_{jk} \sim Uniform\left(0, \frac{1}{1+\alpha_{jk}}\right)$ and next we sample from $\alpha_{jk}/u_{jk} \sim Exp(2\tau_j^2/\beta_{jk}^2) I(\alpha_{jk} < \frac{1-\alpha_{jk}}{\alpha_{jk}})$.

4.9 Discussion:

This chapter discusses the use of Bayesian group lasso penalization combined with Pólya-Gamma augmentation to build a simultaneous classification and selection method. The Bayesian spike-and-slab prior helps in identifying functional parameters generated from longitudinal trajectories of multiple brain ROIs, and discriminates the patient group from normal controls. The inclusion of Pólya-Gamma augmentation helps avoid the Metropolis-Hastings algorithm or the incorporation of other expensive sampling algorithms related to latent variables. We consider the longitudinal brain ROI volume measurements as functional predictors, and the cubic basis splines smooth the curves over time. The next steps include using those smoothed functional predictors as discriminating inputs with sparsity assumptions among them.

Our method performs very well on simulated data sets, outperforming available frequentist methods. Furthermore, our method is applied on a data set that has a large number of predictors. We have one more strong assumption for the functional predictors: namely, that they are independent of

each other. As our prior setup does not account for any correlation structure among predictors, this method will not work well for highly correlated functional predictors. To handle this incompetency, we also proposed an alternative prior structure in Section 4.8, but we did not compare its results to those of our original method. This will serve as the basis for further research.

CHAPTER 5

BAYESIAN SPATIOTEMPORAL CLUSTERING MODEL FOR ANALYZING WHITE MATTER DATA

5.1 Spatiotemporal Linear mixed effects modeling for the distribution of cerebral white matter on MRI scans

5.1.1 Introduction

Bernal-Rusiel et al., 2013[12] introduced a linear mixed effects (LME) model for mass-univariate analysis of longitudinal neuroimaging data. They tried to exploit the spatial structure of the cortical thickness in the brain's sub-regions using a spatiotemporal setup. Before that, *Bernal-Rusiel et al., 2012* [11] used a similar approach with an LME model but for longitudinal modeling; they built a base for the extension of space information. We seek to develop a similar spatiotemporal model with a different research question and a completely distinct application. Spatiotemporal modeling has not been studied extensively in the literature regarding longitudinal white matter (WM) changes for healthy aging with diffusion tensor imaging (DTI). *Bender et al. 2016*[10] compared age -related changes across brain regions and identified the influences of age, vascular risk at the baseline on WM diffusion. They conducted a seven-year follow-up study to establish that diffusion properties in association with WM tracts deteriorated with age. Due to the limited study of spatiotemporal association with changes in WM integrity obtained by fractional anisotropy (FA), we work with data from healthy middle-aged and older adults. Some authors have reported evidence of brain region shrinkage with aging (*Raz and Rodrigue, 2006*[104];*Raz and Kennedy, 2009*[102]). In contrary, very little is known about age-related changes in structural properties of cerebral WM, which varies at different rates in various parts of the brain. Diffusion tensor imaging provides principal directions from diffusion tensor eigenvalues, one of which is fractional anisotropy. According to *Montag and Reuter, 2015*[92], "Fractional anisotropy (FA) is a scalar value between zero and one that describes the degree of anisotropy of a diffusion process. A value of zero means

that diffusion is isotropic, i.e. it is unrestricted (or equally restricted) in all directions. A value of one means that diffusion occurs only along one axis and is fully restricted along all other directions."

The literature on longitudinal neuroimaging analysis is vast; in the following, we mention some of the significant studies that have been done in the last few years. *Asami et al., 2011*[5] studied longitudinal loss of gray matter among schizophrenia patients. They established a clinical correlation with gray matter decay. They used a high-dimensional warping and individualized baseline-rescan to measure longitudinal volume changes within subjects, and they performed a comparison with time-varying manual ROI analysis on the same subjects. *Chetelat et al., 2005*[25] found the highest rates of atrophy in the anterior, inferior, middle, and medial parts of the temporal lobes in their study of the dynamics of gray matter changes for patients transitioning from mild cognitive impairment (MCI) to Alzheimer's disease (AD). On the other hand, *Fjell et al., 2009*[35] worked with healthy aging subjects, similar to those in our experiment. They found that changes in brain structure can occur within one year and that atrophy accelerates with the advancement of age. Although longitudinal studies have yielded some novel discoveries, they have not answered all questions regarding brain image data. Moreover, most statistical methods are not enough to capture the real system. *Bernal-Rusiel et al., 2012* [11] implemented an LME model to analyze brain sub-regions' cortical thickness in Alzheimer's patients. They performed hypothesis testing of the regression coefficients, calculated the sample size, and derived statistical power. In longitudinal studies, researchers broadly focus on two methodologies. The first is the repeated measure of variance and secondly the cross-sectional study after aggregated data. We previously conducted a separate study with longitudinal data by assuming brain volume changes as a function of time, and we used estimated functions as potential predictors for classification; the manuscript is under review.

The aim of this article is to apply an LME model for longitudinal data where space information is considered as an additional effect. The LME method captures the covariance structure of brain voxels properly for serial measurements. It also has the ability to impute missing responses for any

time point. We work with a balanced data case, but the extension of the methodology to unbalanced data is trivial. We obtain our data in the "nifty" image format, and each pixel of the image has location coordinates. We want to find regions in the brain that have homogeneous properties, which will help us to track region-wise longitudinal changes instead of voxel-wise analysis. Several previous studies have examined longitudinal changes for voxel-level data (*Zipunnikov et al., 2011*[137], *Zhang et al., 2009*[135], *Skup et al., 2012*[114], *Shinohara et al., 2011*[112]). However, most of these works employ voxel-based analysis and do not account for spatial information. *Bernal-Rusiel et al., 2013*[12] apply their work to a different problem with cortical thickness measurements, but their covariance structure works nicely for our WM DTI data as well. As our motivation is to work with homogeneous brain regions and to track changes in fractional anisotropy WM measurements in relation to subjects' demographic and clinical factors, we work with the K-means algorithm. We consider each voxel's location, measurements, and predictive response values as clustering attributes for voxel-wise analysis. The homogeneous regions offer a scope to build a spatiotemporal model separately for each cluster. This strategy has several advantages. Firstly, this is a highly parsimonious way of building models, such that the number of estimable parameters are much lesser than voxel-wise models. Secondly, voxel-level models do not consider spatial structure in the data, which implies that estimators are less informative with lower statistical power. Finally, we are able to identify brain regions behaving similarly in terms of WM atrophy as the healthy aging people age.

This chapter is organized as follows. In part 1 of Section 2, we describe voxel-wise model building, estimation of fixed effect parameters, and temporal covariance structure with respect to subjects' time-variant random effects. We calculate residual values for each voxel used as an attribute in the clustering algorithm. We describe the agglomeration method in part 2, and finally, we apply the spatiotemporal mixed effect model in part 3. We perform separate inference in part 3 for region-wise reasoning. Section 3 then describes the WM data and how we obtained it from subjects. Results are described in Section 4, while a final discussion is presented in Section 5.

5.1.2 Methods

5.1.2.1 Voxel-wise linear mixed effect model

We start modeling with a voxel-wise LME model fitting with WM MRI data. In a longitudinal study, the outcome variable for, for instance, the fractional anisotropy or cerebral WM in our data is measured repeatedly for the same individual for a number of times. The aim is to measure changes in the patients' response over time and their association with demographic and biological factors, such as gender, hypertension, MMSE score, and starting age for the study. We fix a spatial location such as for a fixed particular voxel, we have the leeway to apply a single LME independently for that particular voxel. One advantage of this modeling is that it can assess the within-subject changes or correlation across different time points. Voxel-wise LME models have been used in various studies. For example, *Bowman and Kilts, 2003*[16] implemented a mixed effect model to analyze PET data. They acheived improved estimation and inference by considering various covariance structures to handle correlation among individuals' repeated scans. *Bernal-Rusiel et al., 2013*[12] worked with a voxel-wise mixed effect model; they analyzed longitudinal brain MRI data from ADNI. The following are other significant papers on these methods: (*Delaloye et al., 2008*[28], *Lau et al., 2008*[71], *Shaw et al., 2008*[109]). A longitudinal study in a spatiotemporal setup offers a distinctive understanding; for example, it reflects the temporal trajectory of an underlying non-stationary continuous process. The correlation between pairs of repeated measures is directly proportional to the distance between consecutive time points, and between-subject variations are not constant, which gives us scope to assume a subject-specific covariance structure.

In general, we observe three different variabilities in the covariance structure for voxel-specific longitudinal measurements. A particular voxel's WM measurements have (i) between-subject variability, (ii) within-subject changes over time, and, finally, (iii) a measurement error uniform with the other two variations. Now, the voxel-wise LME model imposes structure on the covariance matrix by introducing random effects. In totality, all the above variations are handled with this

structure except the measurement error, although we believe that measurement error has a direct influence on the covariance between longitudinal measurements within a subject. The mean trajectory of WM is modeled with subject-specific fixed effects and time-variant random effects. In this research article, we consider working with balanced, normally distributed data. Now, let Y_i be the $T \times 1$ vector of serial univariate measurement for a particular voxel for subject i, where T is the number of time points. X_i is the $T \times p$ subject specific design matrix built with fixed effects (variables such as gender, hypertension indicator, MMSE score, Base Age, CESD score etc.). Let Z_i be the $T \times q$ design matrix of random effects (e.g. systolic and diastolic pressure, measurement time etc.). We consider time-variant predictors as the random effects and the intercept as the fixed effect. Hence, we can write the model as follows:

$$
\begin{aligned}
Y_i &= X_i\beta + Z_i u_i + \epsilon_i, \quad i = 1, .., M \\
u_i &\sim N_q(0, \sigma_r^2 D_r) \\
\epsilon_i &\sim N_T(0, \sigma^2 I_T)
\end{aligned}
\tag{5.1}
$$

We define

$$
D_r = \begin{bmatrix}
1 & \rho & \dots \rho^{q-1} \\
\rho & 1 & \dots \rho^{q-2} \\
\vdots & \vdots & \ddots \\
\rho^{q-1} & \rho^{q-2} & \dots 1
\end{bmatrix}
$$

and $X_i = \mathbf{1}_T \otimes \vec{x}_i^T$. The components of u_i control the deviation from the population mean for i-th subject. As we mentioned above, this model structure handles the variability present in longitudinal data, and ϵ_i controls the remaining measurement error. Finally, we assume that we have M number of subjects/patients.

The subject specific mean of Y_i given u_i is $E(Y_i|u_i) = X_i\beta + Z_i u_i$; thus, the vector u_i with fixed effect provides subject specific coefficients. The β regression coefficients are the same for all individuals. The initial motivation to write about the single voxel longitudinal model was to find

the temporal covariance matrix for i^{th} subject, and

$$Cov(Y_i) = \Sigma_i = \sigma_r^2 Z_i D_r Z_i^T + \sigma^2 I_T \tag{5.2}$$

We can find the unbiased estimates of the unknown parameters β, σ_r, ρ, σ by maximizing the log-likelihood function.

Parameter Estimation: In the following, we shortly describe the method to estimate the unknown parameters mentioned above, as we use the voxel-wise estimation to segment the whole brain into homogeneous regions. We explain the segmentation method in the next section. Now, given the normally distributed WM assumptions, the vector of measurements for a particular subject's fixed voxel is as follows:

$$Y_i \sim N(X_i\beta, \Sigma_i)$$

If we have the estimates of $\hat{\sigma_r}, \hat{\rho}, \hat{\sigma}$, the closed-form solution of β is based on MLE and given as:

$$\hat{\beta} = \left(\sum_{i=1}^{M} X_i^T \hat{\Sigma_i}^{-1} X_i\right)^{-1} \sum_{i=1}^{M} X_i^T \hat{\Sigma_i}^{-1} Y_i$$

Once we maximize the restricted log-likelihood we can have unbiased estimates of $\hat{\sigma_r}, \hat{\rho}, \hat{\sigma}$ (*Verbeke and Molenberghs, 2000*[122]):

$$l_{ReML} = \frac{1}{2}\sum_{i=1}^{M} \log|\Sigma_i^{-1}| - \frac{1}{2}\sum_{i=1}^{M}(Y_i - X_i\hat{\beta})^T \Sigma_i^{-1}(Y_i - X_i\hat{\beta}) - \frac{1}{2}\log\left|\sum_{i=1}^{M} X_i^T \Sigma_i^{-1} X_i\right|$$

The solution of the estimates can be obtained by using either the expectation maximization (EM) algorithm, the Newton-Raphson algorithm, or Fisher's scoring algorithm. We employ Fisher's scoring algorithm to obtain the parameter estimates, and we need some initial estimates to run the algorithm.

5.1.2.2 Clustering voxels into homogeneous regions using K-means algorithm

Our key aim is to build a spatiotemporal model for homogeneous regions in the brain. Homogeneous regions or clusters contain voxels that are similar in terms of cerebral WM, location in

the brain, and quantitative relationship with the subjects' demographic/biological factors. Once we obtain homogeneous regions in the brain from the clustering algorithm, we can assume that temporal variance is shared among all voxels present in a particular region. The K-means algorithm we implement is a universal method and completely data-driven. Our goal is to partition the voxels into k-uniform clusters $C_1, C_2, .., C_k$ such that i) $C_i \cap C_j = \phi$ for $i \neq j$ and ii) elements of C_i are homogeneous based on some conditions. We now list the conditions we consider to run the K-means algorithm.

We first calculate the coefficient of variation of a particular voxel based on all the longitudinal observations present for all subjects. In our study, we have $M * T$ number of cerebral WM measurements for each voxel, and we calculate $CV^v = \frac{sd^v}{mean^v}$ for each voxel. We use the CV and mean measurement for each voxel in the K-means algorithm. Next, we calculate the residuals for each voxel by fitting the mixed-effect model described in the last section. Once we estimate the parameters for the temporal covariance matrix and fixed effect regression coefficients, it is easy to compute residuals for each voxel. In addition, we have one more piece of directional information for FA measurements over voxels. This information helps us to collate voxels whose WM changes are in a uniform direction. Finally, we calculate the polar coordinates of the three-dimensional voxel locations from their Cartesian coordinates using the following formula:

$$r = \sqrt{x^2 + y^2 + z^2},\ \theta = \tan^{-1}(\frac{y}{x}),\ \phi = \cos^{-1}(\frac{z}{r})$$

The K-means algorithm starts by defining a dissimilarity matrix as an input. In our data, we have measurements x_{ij} for $i = 1, 2, .., N$ and $m = 1, 2, .., m$ where N is the number of active voxels in the brain with positive WM measurements and m is the number of attributes considered for a particular voxel. We use five attributes to cluster the voxels. We define the dissimilarity matrix between voxel i and i' as

$$D(x_i, x_{i'}) = \sum_{j=1}^{m} d_j(x_{ij}, x_{i'j})$$

Most of the common choices are Euclidean distances: $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ or Manhattan

distance: $|x_{ij} - x_{i'j}|$ or based on correlation:

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}'_i)}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}'_i)^2}}$$

As shown in *Hastie et al., 2009*[47], if the attributes are standardized, then $\sum_j (x_{ij} - x_{i'j})^2 \propto 2(1 - \rho(x_i, x_{i'}))$. Hence, we use Euclidean distance to implement K-means clustering.

5.1.2.3 Cluster based Spatiotemporal linear mixed effect models

Spatiotemporal models that consider a temporal covariance matrix and that are shared among spatial locations have been studied in the functional neuroimaging literature for a long time. *Bowman, 2007*[14] described a framework for spatiotemporal modeling by defining a functional distance metric in the model to perform voxel-wise inference. The distance metric reflected bonding between voxels based on functionality rather than proximity. The spatiotemporal model estimated temporal and spatial correlations in a given region (ROI). Later, *Bowman et al., 2008*[15], proposed a spatial Bayesian hierarchical model for analyzing functional neuroimaging data. They built the hierarchical model after anatomical parcellation of the brain consisting of 116 regions. *Derado et al., 2012*[29] also worked with a multivariate Bayesian hierarchical model, where a Gaussian prior was imposed on a voxel-specific population-level mean parameter separately for a baseline and a follow-up session. Next, the wishart prior on the inverse covariance matrix of subject specific measurements for a particular voxel provides estimators. *Woolrich et al., 2004*[127] presented a full Bayesian modeling approach to handle spatiotemporal noise modeling and haemodynamic response function (HRF) modeling. They implemented a simultaneously specified autoregressive model in which the spatial noise structure would remain the same over time. Although this method precisely estimated the parameters and grasped the spatiotemporal dependency, it was important to model the spatial covariance matrix among those locations of voxels present in a particular cluster. Moreover, the proposed voxel-wise model or prior did not account for the inter-voxel correlation. The authors assumed that responsive voxels are spread over the whole brain and that the temporal covariance matrix is a scaled version of a global covariance structure matrix. We aim to model our cerebral

WM data with a similar assumption as in *Bernal-Rusiel et al., 2013*[12], for two main reasons. First, we believe that the temporal covariance structure is likely to be different in different segments of the brain and that it needs to be modeled separately for each cluster of voxels. Second, inter-voxel correlations depend on the location of the voxels; for example, voxels in close proximity share similar behavior at different stages in various disease processes. The spatiotemporal covariance of the cerebral WM for a particular brain segment is a product of both the spatial covariance of the voxels present in that region and the temporal covariance of a particular patient whose parameters are different for different regions. The spatial structure presented in *Bernal-Rusiel et al., 2013*[12] is inspired from *Bowman, 2007*[14]. *Bowman, 2007*[14] proposed the alternate strategy of using a spatiotemporal model to estimate spatial and temporal correlations in a given ROI. We assume a similar spatial covariance matrix captured through a parametric matrix that models the dependency as an exponential transformation of the distance between voxels given a region/cluster.

Our assumption is that a temporal covariance structure is shared among voxels within a homogeneous region obtained from the K-means algorithm. The parametric covariance structure that controls the dependency between voxels in a region solely depends on the distance between those voxels. We consider a homogeneous parcellation of the brain consisting of $g = 1, .., G$ regions. Let g be the region/segment/cluster in which we are interested and v_g be the number of voxels present in the g^{th} cluster. Let Y_{ig} be the $(T * v_g) \times 1$ vector of measurements for WM for i^{th} subject. As we defined in section (2.1), we have M number of patients, with each patient having T time points of observations (balanced case). We can expand the Y_{ig} vector as-

$$Y_{ig} = \begin{bmatrix} Y_{ig1} \\ Y_{ig2} \\ \vdots \\ Y_{igv_g} \end{bmatrix}$$

where Y_{igv} is the $T \times 1$ vector measurements of v^{th} voxel for patient i. We assumed similar

covariance matrix used by *Bernal-Rusiel et al., 2013*[12] for Y_{ig} as-

$$cov(Y_{ig}) = W_{ig} = G_g \otimes \Sigma_{ig} = \begin{bmatrix} G_{g11}\Sigma_{ig} & G_{g12}\Sigma_{ig} & \dots G_{g1v_g}\Sigma_{ig} \\ G_{g21}\Sigma_{ig} & G_{g22}\Sigma_{ig} & \dots G_{g2v_g}\Sigma_{ig} \\ \vdots & \vdots & \vdots \\ G_{gv_g1}\Sigma_{ig} & G_{gv_g2}\Sigma_{ig} & \dots G_{gv_gv_g}\Sigma_{ig} \end{bmatrix}$$

where $\otimes$ is the Kronecker product, and $\Sigma_{ig} = \sigma^2_{rg} Z_i D_{rg} Z_i^T + \sigma^2_g I_T$ from equation (2) is the temporal covariance matrix for i^{th} patient in g^{th} cluster. G_g is the $v_g \times v_g$ spatial covariance matrix that accounts for the correlation between voxels present in the g^{th} cluster. Similarly, we used the spatial structure proposed in *Bernal-Rusiel et al., 2013*[12] which is empirically useful:

$$G_g = \begin{bmatrix} 1 & \exp^{-a_g d_{12} - b_g d^2_{12}} & \dots & \exp^{-a_g d_{1v_g} - b_g d^2_{1v_g}} \\ \exp^{-a_g d_{21} - b_g d^2_{21}} & 1 & \dots & \exp^{-a_g d_{2v_g} - b_g d^2_{2v_g}} \\ \vdots & \vdots & \vdots & \vdots \\ \exp^{-a_g d_{v_g1} - b_g d^2_{v_g1}} & \exp^{-a_g d_{v_g2} - b_g d^2_{v_g2}} & \dots & 1 \end{bmatrix}$$

where $a_g, b_g \geq 0$ are unknown parameters and d_{jk} is the Euclidean distance between voxels j & k in the g^{th} cluster. Hence, the joint distribution of the Y_{igv} vector of WM measurements in g^{th} cluster follows:

$$Y_{ig} \sim N(X_{ig}\beta_{ig}, W_{ig}), \tag{5.3}$$

where $X_{ig} = I_{v_g} \otimes X_i$ is $(v_g * T) \otimes (v_g * p)$ matrix and

$$\beta_g = \left(\beta_{g11}, \dots, \beta_{g1p}, \beta_{g21}, \dots, \beta_{gv_g1}, \dots, \beta_{gv_gp}\right)^T$$

is a $(v_g * p) \times 1$ stacked vector of fixed effects which gives separate fixed effect estimates for each voxel present in g^{th} cluster. Finally, we will work with restricted log-likelihood and Fisher's Scoring method to estimate the parameters associated to covariance matrix W_{ig} i.e. $(\sigma_{gr}, \rho_g, \sigma_g, a_g, b_g)$. The restricted log-likelihood is:

$$l = \frac{1}{2}\sum_{i=1}^{M} \log|W_{ig}| - \frac{1}{2}\sum_{i=1}^{M}(Y_{ig} - X_{ig}\hat{\beta}_g)^T W_{ig}^{-1}(Y_{ig} - X_{ig}\hat{\beta}_g) - \frac{1}{2}\log|\sum_{i=1}^{M} X_{ig}^T W_{ig}^{-1} X_{ig}|,$$

where $\hat{\beta}_g = \left[\sum_{i=1}^{M} X_{ig}^T \hat{W}_{ig}^{-1} X_{ig}\right]^{-1} \sum_{i=1}^{M} X_{ig}^T \hat{W}_{ig}^{-1} Y_{ig}$ is the least square estimator. One more advantage of segmenting the brain or clustering homogeneous voxels together is that it reduces the number of parameters to estimate. In the situation of voxel-wise model building, we are required to estimate parameters for the covariance matrix for each voxel, whereas in this situation we estimate $(\sigma_{gr}, \rho_g, \sigma_g, a_g, b_g)$ only for G number of regions.

Fisher's scoring algorithm starts by calculating partial derivatives and expected information matrix for each iteration to update parameter estimates:

$$\theta^{(k+1)} = \theta^{(k)} + I^{-1}(\theta^{(k)}) \frac{\delta l}{\delta \theta}|_{\theta=\theta^{(k)}}$$

Let's define $\theta_g = (\sigma_{gr}, \rho_g, \sigma_g)$ and $s_g = (a_g, b_g)$. Next, define

$$M_{igj} = \frac{\delta \Sigma_{ig}}{\delta \theta_{gj}},\ H_g = \sum_{i=1}^{M} X_i^T \Sigma_{ig}^{-1} X_i,\ F_{gk} = \frac{\delta G_g}{\delta s_{gk}}$$

$$\hat{\beta}_g = \sum_{i=1}^{M} \left(I_{v_g} \otimes H_g^{-1} X_i^T \Sigma_{ig}^{-1}\right) Y_{ig},\ \ and\ r_{ig} = Y_{ig} - X_{ig}\hat{\beta}_g$$

Then the partial derivatives of the restricted log-likelihood functions are:

$$\begin{aligned}
\frac{\delta l}{\delta \theta_{gj}} &= -\frac{v_g}{2} \sum_{i=1}^{M} \left(tr(M_{igj}\Sigma_{ig}^{-1}) + \frac{1}{2} r_{ig}^T (G_g^{-1} \otimes \Sigma_{ig}^{-1} M_{igj} \Sigma_{ig}^{-1}) r_{ig}\right) \\
&\quad + \frac{v_g}{2} tr(H_g^{-1} \sum_{i=1}^{M} X_i^T \Sigma_{ig}^{-1} M_{igj} \Sigma_{ig}^{-1} X_i) \\
\frac{\delta l}{\delta s_{gk}} &= \frac{1}{2} \sum_{i=1}^{M} r_{ig}^T \left(G_g^{-1} F_{gk} G_g^{-1} \otimes \Sigma_{ig}^{-1}\right) r_{ig} - \frac{(M * T - p)}{2} tr(F_{gk} G_g^{-1})
\end{aligned}$$

The individual entries of Fisher's Information matrix for g^{th} cluster is described below:

$$\begin{aligned}
I_g(\theta_{gj}, \theta_{gk}) &= \frac{v_g}{2} \left(\sum_{i=1}^{M} tr(\Sigma_{ig}^{-1} M_{igj} \Sigma_{ig}^{-1} M_{igk}) - tr(H_g^{-1}(2Q_{gjk} - P_{gj} H_g^{-1} P_{gk}))\right) \\
I_g(\theta_{gj}, s_{gk}) &= \frac{1}{2} tr(F_{gk} G_g^{-1}) \left(\left(\sum_{i=1}^{M} tr(\Sigma_{ig}^{-1} M_{igj})\right) + tr(H_g^{-1} P_{gj})\right) \\
I_g(s_{gj}, s_{gk}) &= \frac{(M * T - p)}{2} tr(F_{gj} G_g^{-1} F_{gk} G_g^{-1}),
\end{aligned}$$

Table 5.1: Subject Characteristics

Baseline variables	Women Mean (sd)	Men Mean (sd)	p-value
Age	62.60 (8.51)	63.62 (9.61)	0.825
MMSE	28.50 (0.92)	29.12 (1.35)	0.30
% Hypertension	37.5%	25%	1
Systolic BP	133.75 (9.4)	128.81 (7.8)	0.27
Diastolic BP	78.65 (7.5)	76.93 (8.5)	0.67
ED	15.5 (3.11)	17.87 (2.8)	0.13
% Smoker	0%	12.5%	1

p-values computed for two-sample t-test or test of association, sd=standard deviation

where $Q_{gjk} = \sum_{i=1}^{M} X_i^T \Sigma_{ig}^{-1} M_{igj} \Sigma_{ig}^{-1} M_{igk} \Sigma_{ig}^{-1} X_i$, and $P_{gj} = \sum_{i=1}^{M} X_i^T \Sigma_{ig}^{-1} M_{igj} \Sigma_{ig}^{-1} X_i$ After we input all the necessary terms in the iteration parameters do converge after sixth or seventh step.

5.1.3 WM Data

The subjects in this study were from the Midwestern part of the United States. They were no younger than 50 during their first visit. All subjects had three longitudinal measurements, simplifying the problem to a balanced case. In total, 16 patients participated. Subjects were screened for history of psychiatric and neurological disorders or any heart disease, whether hypertension or not. They were then screened for cognitive impairment using the Mini Mental Status Examination (MMSE) for symptoms of depression. The sample comprised 16 healthy subjects (50% women) who were between 50 and 80 years old (mean=62.6, sd=8.51 years). The men and women did not differ in age, MMSE, hypertension, or systolic and diastolic pressure. We present the demographic and clinical characteristics in Table 1; the corresponding p-value confirms that there was no gender bias in the sample. Moreover, the proportion of smoking, regular exercise, and diagnosed hypertension did not differ based on gender.

MRI scans were obtained at three separate time points (T3, T2, and T3), and the DTI sequence

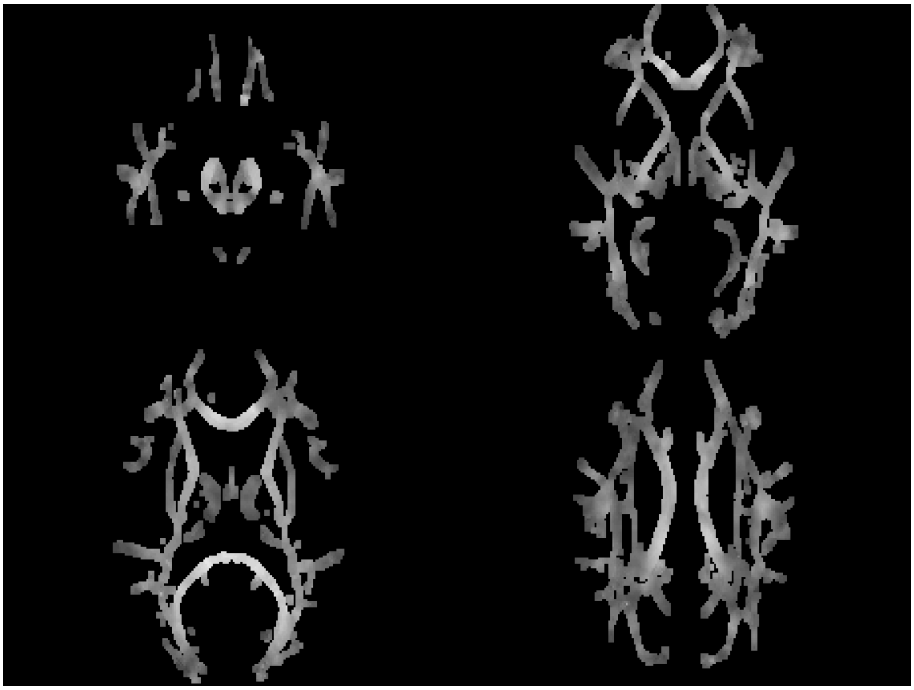


Figure 5.1: Change of WM across Z-axis from top-left to bottom-right Z-slice 15,30,45,60

was initiated on the first occasion (T1). The mean differences between each time point are: mean T1-T2=15.56, SD=1.03 months; mean T2-T3=15.06, SD=2.8 months. The subjects were scanned using a 1.5T Siemens Magnetom Sonata scanner. One image has a voxel size $1.8 \times 1.8 \times 3\ mm^3$. In total we have $182 \times 218 \times 182$ slices of 'nifti' image with 48 number time points (#patients $\times$ time points). We have plenty of extraneous slices in our data which are the outer side of skull and extra-cranial tissues. In order to process the data we need to remove those empty slices. We used neurobase::dropEmptyImageDimensions function to reduce the dimension with reduced proportion as $121 \times 157 \times 100$. The voxel-wise measurement changes across the z-axis and Figure 5.1 has a representation of how image changes if we move from head to toe.

The voxel-wise measurement changes across the z-axis. Figure 5.3 shows the orthographic view of the reduced "nifti" image in three different planes for a single time point, as well as the voxels whose WM measurements are greater than 0.5. This gives us a scope to investigate brighter regions and voxels with lower WM intensity. We are more interested in the WM analysis and in studying WM distribution. After masking the first time point's image, we examine the density plot of the data. Figure 5.2 shows that WM measurements follow approximately a bell shaped distribution,

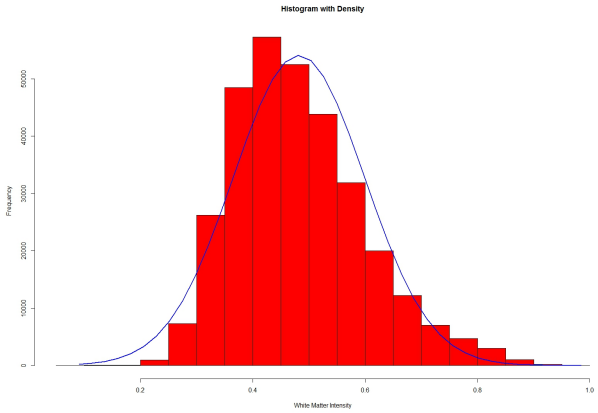


Figure 5.2: white matter histogram with density plot

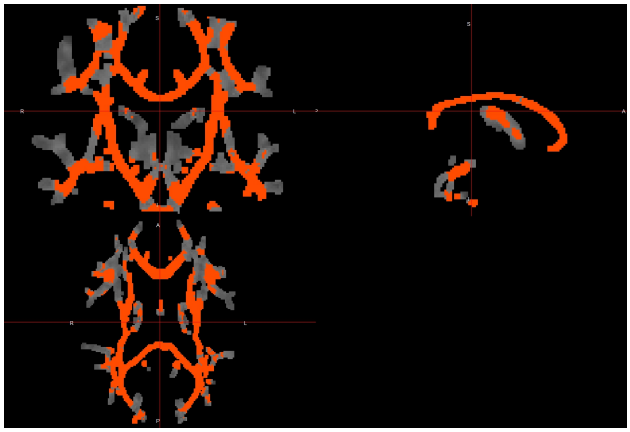


Figure 5.3: white matter measurement greater than 0.5

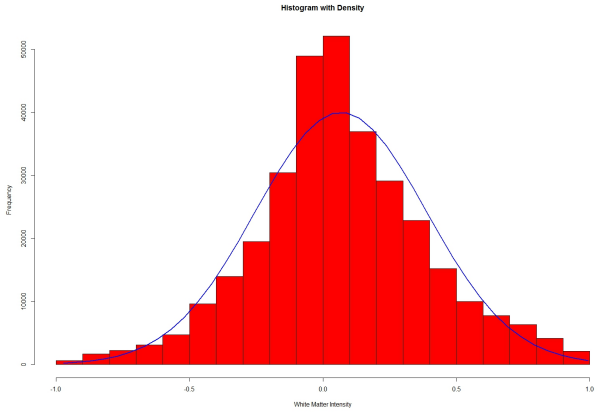


Figure 5.4: voxels' directional density

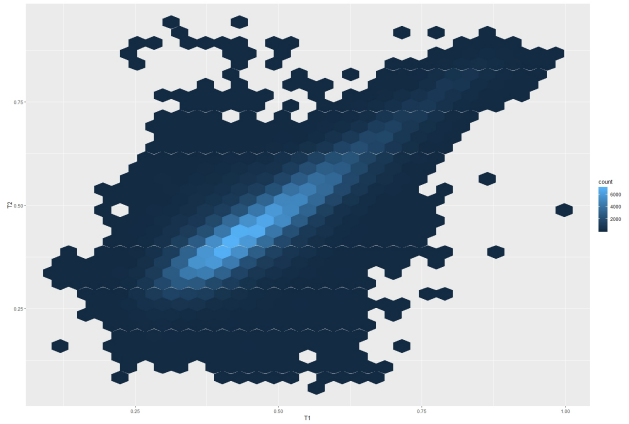


Figure 5.5: 2-dimensional distributions of T1 & T2 imaging sequence

with values between 0.2 and 1. This finding helps us to assume a Gaussian distribution for the response variable. We also want to compare image files for different time points and to calculate the correlation between consecutive occurrences. Figure 5.5 presents a two-dimensional distribution of T1 and T2 imaging sequences against each other. We also have another nifti object, which contains the directional information about FA measurement changes. The directional values range from -1 to 1, as plotted in Figure 5.4. Voxels with close directional values have similar properties regarding change in WM/FA values.

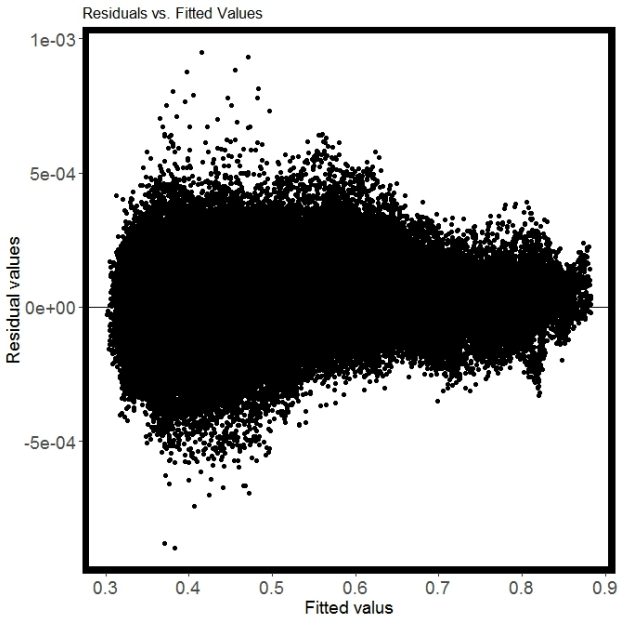


Figure 5.6: Residual vs predicted plot

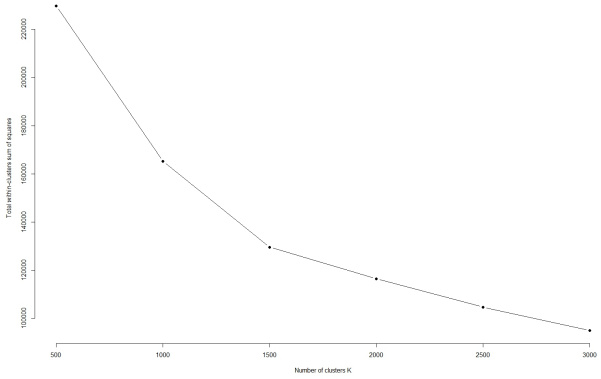


Figure 5.7: Within sum of squares by number of clusters

5.1.4 Results

We start our analysis with voxel-wise model fit. We work with a linear mixed effect model, where we use eight factors as fixed effects and two time-variant variables as random effects. The following fixed effects are considered: gender, hypertension indicator, MMSE score, and baseline age, systolic pressure, diastolic pressure, ED, and smoking indicators. We have few longitudinal observations for each patient, which limits us to the use of two time-variant random effects, such as systolic and diastolic pressure, in the model. We consider the intercept as the fixed effect. Coefficients are estimated using Fisher's scoring method, where each voxel has separate estimates for $(\beta, \sigma^2, \sigma_r^2, \rho)$ and a best predictor (BP) for $\hat{u}_i$. Our key aim is to calculate residuals for each voxel to use as attributes for the clustering algorithm. We check the randomness of the residuals with respect to the voxel-wise predicted response for each patient and each time point. The plot in Figure 5.6 shows no pattern in residual values.

We start our clustering algorithm by calculating each voxel's coefficient of variations, mean, residuals, and directional values, and the polar coordinates of the three-dimensional voxel location.

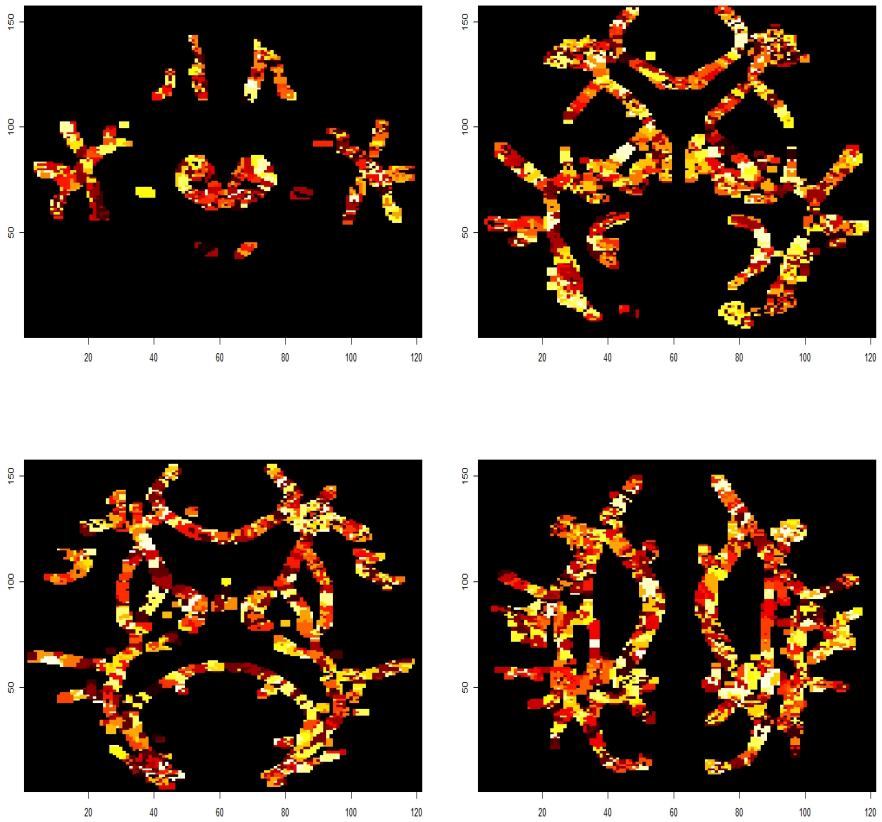


Figure 5.8: Segmentation of brain into similar regions

We then use the K-means algorithm to segment the brain into 3,000 clusters with a maximum cluster size of 258 voxels. We check the within sum of squares over the number of clusters in Figure 5.7, and 3,000 is the optimal point for minimum WSS (within sum of squares) to have a significant number of voxels present. The average cluster size is around 105 voxels. Once we obtain each voxel's cluster number, i.e. the cluster to which it belongs, we feed the corresponding number to that voxel as an input. We visually examine how voxels are clustered together in each hemisphere. Note that, voxels in same region have four major attributes: mean WM, variation in WM over time, residual values, and location of the voxels. We plot the same brain slices in Figure 5.8 as we used in Figure 5.1 (z-axis: 15, 30, 45, 60) with cluster position numbers. Based on the visual representation, we can claim that voxels close to similar WM belong to an analogous region.

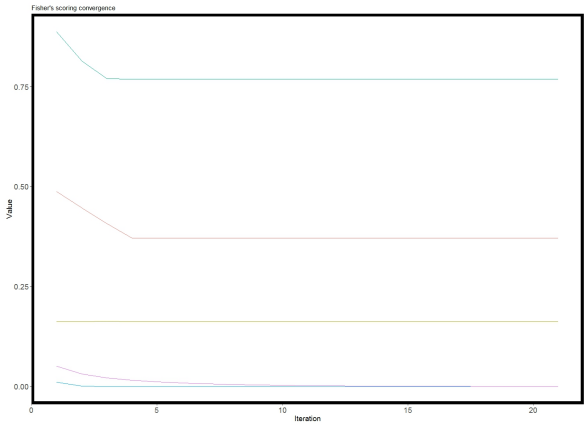


Figure 5.9: Fisher's scoring method convergence

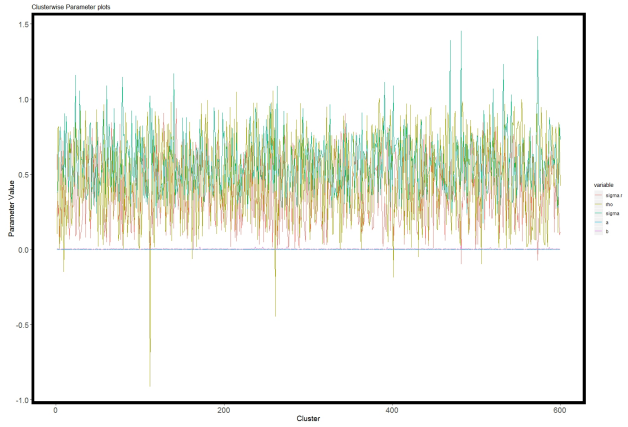


Figure 5.10: Estimated parameters plot for all clusters

In the next step, we work with spatiotemporal linear mixed effects with the same set of predictors used in voxel-wise longitudinal models. However, this time we consider all the voxels present in a cluster. Firstly, we estimate parameters $(\sigma_r^2, \rho, \sigma^2, a, b)$ related to random effects, measurement error, and the spatial covariance matrix that controls the spatiotemporal covariance matrix. We use Fisher's scoring algorithm to estimate these parameters, and all parameters converge within 20 iterations. We inspect the convergence of parameters for cluster 1 in Figure 5.9. Our motivation behind segmenting the brain is to find regions whose voxels are similar to each other and the clusters that differ from each other. To study how much variability we observe from one cluster to another, we plot the estimated parameters for all 3,000 clusters in Figure 5.10. Visually, we can identify that the correlation estimates differ from one cluster to another.

Finally, we discuss our pivotal interest in estimating β components. As we have mentioned before, $Y_{ig} \sim N(X_{ig}\beta_{ig}, W_{ig})$ and $\hat{\beta}_g = \left[\sum_{i=1}^{M} X_{ig}^T \hat{W}_{ig}^{-1} X_{ig}\right]^{-1}$ $\sum_{i=1}^{M} X_{ig}^T \hat{W}_{ig}^{-1} Y_{ig}$ is the least square estimator, where β_g is a $(v_g * p) \times 1$ vector. This implies that for a particular cluster, we will obtain v_g set of $\beta(p \times 1)$ estimators, where the first component is the intercept and and the rest of the components belong to the corresponding fixed effect. Note that v_g is the number of voxels present in g^{th} cluster. Similar to the previous approach, we study the histograms of beta components over different clusters. We examine two separate figures, one

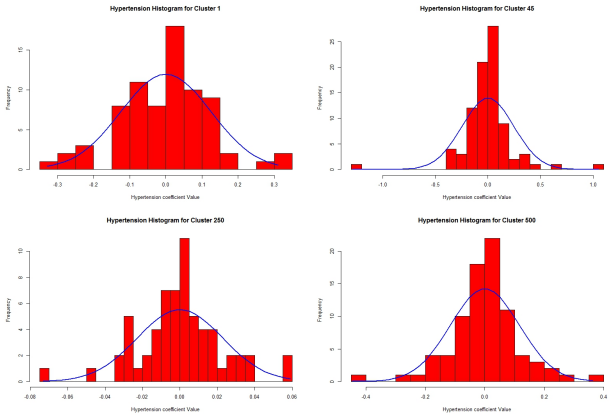


Figure 5.11: Histogram of Intercepts

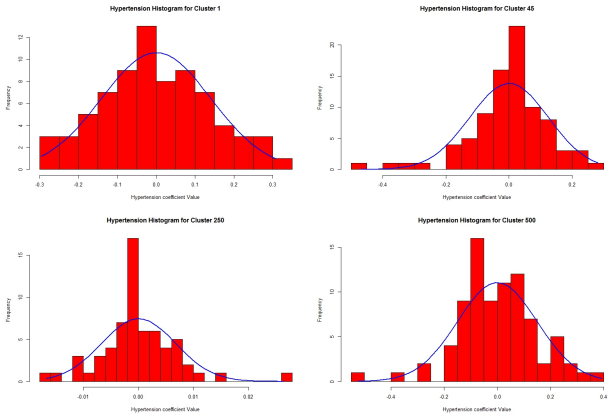


Figure 5.12: Histogram of Gender coefficients

for the intercept and another for the gender coefficient, in Figures 5.11 and 5.12, respectively. We choose clusters 1, 45, 250, and 500 to inspect variability. From the plots, it is clear that voxel-wise estimates are different among clusters, implying voxel variability over regions.

5.1.5 Discussion

The linear mixed effect model is a powerful tool that has been used here to capture spatiotemporal structures hidden in voxel-level data. *Bernal-Rusiel et al., 2013*[12] proposed this method for analyzing cortical thickness in brain sub-regions. LME provides a flexible approach to capture the longitudinal changes and spatial dependence. This method always works better than voxel-wise longitudinal studies and statistical power increases. To the best of our knowledge, this method has never before been used to analyze longitudinal WM changes in healthy aging subjects. The implications of the whole setup can be divided into three parts. i) First, we find the longitudinal changes of WM in each voxel, and we identify how subjects' age, gender, hypertension, MMSE score, and systolic and diastolic pressure relate to these repeated measures. ii) In the next stage, we find the brain regions that behave similarly with respect to the WM measurements, voxel locations, and voxel-wise model coefficients. This gives us homogeneous regions in the brain to build more parsimonious models. iii) Finally, we develop an individual model of each of these regions, where

the covariance matrix is a bi-product of the longitudinal and spatial covariance matrix. We observe that fixed-effects regression coefficients, estimated variance, and covariance components of random effects vary from one cluster to another. These give us a scope to study longitudinal changes in these homogeneous regions rather than in a voxel-wise study.

The main motivation for building this linear mixed effects spatiotemporal model was to work with Bayesian model-based clustering with some stochastic objective function. In Part A, we used the K-means algorithm to segment the brain, but we were dissatisfied with the limitations of this method. One such limitation is that we needed to fix the number of clusters a priori, while this is most often unknown. Researchers study the within sum of squares to choose an optimal number of clusters, but it does not have any probabilistic implications. An advantage of the stochastic objective function is that it incorporates dependence among voxels through spatial and temporal structures within a cluster. We elaborately formulate the method in Part B.

5.2 Bayesian Model based clustering in application to Spatiotemporal data

5.2.1 Introduction

Cluster analysis is a modern data mining tool used to group or segment objects into clusters based on similarity measurements. Many available algorithms have been built based on the particular criterion that objects within a cluster are more similar to each other than to objects belonging to other clusters. Some popular clustering methods are K-means, agglomerative clustering, hierarchical clustering, and model-based clustering. We focus on model-based clustering in this chapter. A disadvantage of hierarchical clustering and K-means is that they are model-free heuristic methods. Model-based clustering is an alternative that provides the option to formulate a model while grouping objects. Model-based clustering was first introduced by *Banfield and Raftery, 1993*[6], who assumed that observations come from multivariate normal distribution and that maximum likelihood yields the estimates. *Fraley and Raftery, 2002*[37] introduced the finite mixture models as a formal setting for model-based clustering. Let $y_1, .., y_n$ be independently distributed p-dimensional observations from a K-component mixture distribution:

$$f(\boldsymbol{y}; \boldsymbol{\tau}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f_k(y_i|\theta_k) \tag{5.4}$$

Here, τ_k is the probability that a particular object belongs to k^{th} cluster, and θ_k parametrize the density f_k. In general, f_k is multivariate normal density, where $f_k \equiv MVN(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$. EM algorithm helps us to obtain the estimates of unknown parameters. *Melnykov and Maitra, 2010*[88] provided a thorough reference for model-based clustering. One point of subjectivity that is always attached to cluster analysis is how to decide the number of clusters. Similar to K-means, model-based clustering needs a specified number of clusters at the start of the algorithm. *Fraley and Raftery, 1998*[36] tried to address this question by using model selection criteria. They suggested examining BIC over all possible cluster numbers and choosing the one with the lowest BIC. In another notable study, *Yeung et al., 2001* [131] applied model-based clustering to group gene expression data. Over time, researchers have included other factors in the model-based approach,

such as variable selection [101] and penalizing the parameter space θ[95]. All the aforementioned studies were performed in a conventional statistical framework, without using any prior information on the cluster structure, the number of clusters, or unknown parameters (τ, θ). As an exception, Bayesian model selection has been used to select an optimal number of clusters. *Handcock et al. 2007*[46] described two methods to estimate unknown parameters. They first used a two-stage maximum likelihood method, and then they proposed a Bayesian estimation method. In Bayesian part, parameters have a multivariate normal prior, and the mixture probability τ has Dirichlet prior with MCMC estimation. Moreover, *Fraley and Raftery, 2007*[38] published a paper in 2007 in which they discussed the problems with using the EM algorithm and where it can fail to converge. Instead of MLE, they preferred to work with a MAP estimator with conventional priors on mean and covariance parameters, such as MVN on the mean and an Inverse-Wishart distribution on the covariance matrix. *Medvedovic et al., 2004* [86] proposed a contrasting formulation of a Bayesian mixture-based clustering algorithm for grouping gene expression data with replicates. Interestingly, the Poisson-Dirichlet process is the most logical and convenient way of assuming priors on the number of clusters and mixture probability, if one does not want to limit the number of clusters at the start of the algorithm [70].

The methods and literature described above have various advantages and can be applied to different complicated problems with which researchers deal. Here, we should reiterate the problem we are addressing. We aim to build a Bayesian model-based clustering method with the same WM data used for spatiotemporal modeling in Section 5.1. We have longitudinal voxel-wise WM measurements for healthy aging subjects, and we want to group voxels into homogeneous regions considering longitudinal and spatial information. To this end, we review some of the literature regarding spatiotemporal clustering methods. The field of computer science has worked with this topic extensively. *Kalnis et al., 2005*[65] investigated a highly complicated problem of detecting clusters among moving objects that changed locations over time. For example, they clustered trajectories and mined movement patterns for a group of migrating animals or a convoy of cars moving in

a city. The complexity of this problem is manifold, but we simplify our situation by assuming that the spatial locations of voxels are fixed over time. *Kisilevich et al., 2009*[67] described a detailed study of spatiotemporal clustering on trajectories and provided in-depth research development on this topic.

In Section 5.1.2.2, we apply K-means clustering of group voxels into homogeneous regions without any model assumption. In advancement, we seek to build a clustering algorithm that satisfies the clustering criteria and simultaneously helps us to estimate the regression coefficients and variance components for each cluster. Section 5.1 presents a two-stage procedure of clustering objects and then a fitting spatiotemporal model for each cluster. To blend these two processes together and work closely with the Bayesian clustering structure, we explore the method described by *Booth et al., 2008*[13]. They derived an objective function to cluster objects and a stochastic search process to find the posterior distribution for the number of clusters. We also take advantage of the parsimonious representation of the mean via regression. Similar ideas were explored in studies published before *Booth et al., 2008*'s paper came [51] [48]. In Section 5.2.2, we discuss the general algorithm of our proposed methodology and model building with posterior derivations. Section 5.2.3 concerns the stochastic search method to optimize the objective function.

5.2.2 Methodology :

5.2.2.1 The Ewens-Pitman Prior on Cluster partitions :

The basic clustering objective is easy to convey. Let us assume we have n voxels in the brain observed for M subjects over T time points. We want to group these voxels into clusters that are independent of subject variation, and the algorithm should consider spatial and temporal covariance structures together. Voxels belonging to the same cluster are similar to each other whereas differs widely with members belong to other clusters. As stated in *Booth et al., 2008*[13], we want to

find an objective function $\pi : \mathscr{P}_n \rightarrow \mathscr{R}^+$, where $\mathscr{P}_n$ is all possible partitions of the n objects. Partitions must be non-empty sets that do not overlap each other. π assigns a score to each partition measuring the overall achievement. Equivalently, we need to optimize the objective function π so that we can obtain the partition with the highest score. That partition will be the best partition given all circumstances. In addition, we use the same linear mixed effect model as the one described in Section 5.1.2.3, which considers the spatiotemporal dependence among voxels with other fixed, random effects. Instead of assuming objects coming from equation (5.4) with fixed K groups, it is more realistic to assume that there is an unknown partition ω with $c = c(\omega)$ as the number of clusters. We can denote the clusters as $C_1, .., C_c$ and equation (5.4) can be written as-

$$f(\boldsymbol{y}|\theta_1, ..\theta_c, \omega) = \prod_{k=1}^{c(\omega)} \prod_{i \in C_k} f(y_i|\theta_k) \tag{5.5}$$

with $C_i \cap C_j = \Phi$ for $i \neq j$. The introduction of ω leads to uncertainty in the number of clusters or a probabilistic component attached to each partition. Once we find estimated $\hat{\omega}$, we can fix the cluster partition with the fullest confidence and find estimated $\hat{\theta}$ for each cluster.

Equation (5.5) draws a general picture of the clustering proposal. Spatiotemporal data needs a separate covariance structure; we work with the covariance structure mentioned in *Bernal-Rusiel et al., 2013*[12]. We used WM data in this covariance structure in Section 5.1. We believe that this structure accurately captures the dependence between voxels, as well as temporal changes among time points to model spatiotemporal data in a non-functional way. The objective function that finalizes clusters' partition is the posterior distribution $\pi(\omega|\boldsymbol{y})$, hence we need prior distribution on ω. The optimization of $\pi(\omega|\boldsymbol{y})$ is highly computationally challenging work, especially in spatiotemporal data. As stated in *Booth et al., 2008*[13] if there are n data points to cluster, the total number of all possible partitions is called the *Bell number*[116] and is denoted by $B(n) = \#\mathscr{P}_n$. $B(n)$ increases rapidly with n. The table below gives an idea of the relationship between 'n' and $B(n)$. We have

n	1	2	3	4	5	6	7	8	9	10	40	100
B(n)	1	2	5	15	52	203	877	4140	21147	115975	1.6×10^{35}	4.8×10^{115}

around 316,000 voxels to cluster, which is computationally challenging. We also need a stochastic search algorithm to optimize the objective function. We are still working on building this algorithm.

We now focus on the prior distribution of ω. The posterior distribution $\pi(\omega|\boldsymbol{y})$ is calculated using $\pi(\omega)$ and $\pi(\theta|\omega)$ with integrating out nuisance parameter θ from the likelihood. It should be noted that estimates of θ are cluster specific and should depend on ω. We explore the same prior used by *Booth et al., 2008*[13] for $\omega \in \mathscr{P}_n$,

$$\pi(\omega) = \frac{\Gamma(m)m^{c(\omega)}}{\Gamma(n+m)} \prod_{k=1}^{c(\omega)} \Gamma(n_k), \;\; c(\omega) = 1,..,n, \;\; \omega \in \mathscr{P}_n \tag{5.6}$$

where $n_k = \#(C_k)$ is the number of objects in cluster C_k, $n = \sum_{k=1}^{c(\omega)} n_k$ is total number of objects in study, and $m > 0$ is a parameter. This distribution was first used by *Crowley, 1997*[27]. This prior is also known as "**The Ewens-Pitman Prior**"[22]. *McCullagh and Yang, 2006*[83] presented detailed mathematical derivation of this prior. Moreover when $c(\omega) \to \infty$ this prior is also called as **Chinese Restaurant Process**. Parameter 'm' controls the number of clusters to be formed, and it increases proportionally. When $m \to 0$, we can achieve a single cluster with all objects gathered into it. Similarly, for $m \to \infty$, number of clusters reaches 'n', which is the same as the number of objects. As mentioned in *Booth et al., 2008*[13], the expected number of clusters to be formed if $\omega \sim \pi(\omega)$ is-

$$E\left[c(\omega)\right] = m \sum_{i=0}^{n-1} \frac{1}{m+i}$$

This prior has two other interesting properties:

- **Exchangeability**: If two partitions ($\in \mathscr{P}_n$) have same $c(\omega)$ and same $n_1, n_2, .., n_{c(\omega)}$ then they have exact same probability under $\pi(\omega)$.

- **Consistency**: Prior probability for a partition based on n objects ($\omega \in \mathscr{P}_n$) doesn't depend on if $(n+1)^{th}$ element enters into study, i.e. for $\omega^* \in \mathscr{P}_n$ and $S \subset \mathscr{P}_{n+1}$, $\sum_{\omega \in S} \pi_{n+1}(\omega) = \pi_n(\omega^*)$ is satisfied.

Next we will define our assumed model and prior for $\pi(\theta|\omega)$. We will calculate posterior of ω as-

$$\pi(\omega|y) \propto \int_{\theta} f(y|\theta,\omega)\pi(\theta|\omega)\pi(\omega)d\theta$$

5.2.2.2 Spatiotemporal Linear Mixed Effects model:

In this part, we focus on the same linear mixed effect model as the one used in Section 5.1.2.3. Let us assume that $\boldsymbol{y_{ij}}$ is the i^{th} voxel's j^{th} subject's measurement observed over T time points, and for fixed ω, $\boldsymbol{y_{ij}} \in C_k = 1, 2, .., n_k$ the k^{th} cluster. Then,

$$\begin{aligned} \boldsymbol{y_{ij}} &= X_j\beta_k + Z_j u_i + \epsilon_{ij},\ for\ i = 1, .., n_k\ and j = 1, .., M \\ u_i &\sim N_q(0, \lambda\sigma_k^2 I_q) \\ \epsilon_{ij} &\sim N_T(0, \sigma_k^2 I_T) \end{aligned} \tag{5.7}$$

$X_j^{T\times p}$ is j^{th} subject's fixed effect design matrix, $Z_j^{T\times q}$ is j^{th} subject's time-variant random effect matrix, β_k is a p-dimensional coefficient vector for k^{th} cluster, u_i is the i^{th} voxel's random effect associated with j^{th} patient, and ϵ_{ij} is the measurement error. Once we aggregate a single voxel's measurement for all M subjects, we obtain

$$\begin{aligned} \boldsymbol{y_i} &= X\beta_k + Zu_i + \epsilon_i,\ i = 1, .., n_k \\ cov(\boldsymbol{y_i}) &= \sigma_k^2\left(\lambda ZZ^T + I_{MT}\right) = \sigma_k^2\Sigma \end{aligned} \tag{5.8}$$

Σ contains the temporal covariance matrix, and we can assume that Z is block-wise orthogonal, such that each patient has an independent random effect. Once we consider that all n_k voxels belong to k^{th} cluster, we get $\boldsymbol{y_k}^{(n_k MT)\times 1} = (\boldsymbol{y_1}^T, .., \boldsymbol{y_{n_k}}^T)^T$, we write covariance matrix as a product of spatial and temporal covariance matrices

$$cov(\boldsymbol{y_k}) = \sigma_k^2 W_k = \sigma_k^2 (G_k \otimes \Sigma) = \sigma_k^2 \begin{bmatrix} G_{k11}\Sigma & G_{k12}\Sigma & \dots G_{k1n_k}\Sigma \\ G_{k21}\Sigma & G_{k22}\Sigma & \dots G_{k2n_k}\Sigma \\ \vdots & \vdots & \vdots \\ G_{kn_k1}\Sigma & G_{kn_k2}\Sigma & \dots G_{kn_kn_k}\Sigma \end{bmatrix}$$

where $\otimes$ is the Kronecker product, and $\Sigma = \left(\lambda ZZ^T + I_{MT}\right)$ from equation (5.8) is the temporal covariance matrix for all patients' voxels present in the k^{th} cluster. G_k is the $n_k \times n_k$ spatial covariance matrix that accounts for the correlation between voxels present in the k^{th} cluster. The spatial structure mentioned below and on which we worked in previous sections was proposed by *Bernal-Rusiel et al., 2013*[12]. It is empirically useful:

$$G_k = \begin{bmatrix} 1 & \exp^{-a_k d_{12} - b_k d_{12}^2} & \dots & \exp^{-a_k d_{1n_k} - b_k d_{1n_k}^2} \\ \exp^{-a_k d_{21} - b_k d_{21}^2} & 1 & \dots & \exp^{-a_k d_{2n_k} - b_k d_{2n_k}^2} \\ \vdots & \vdots & \vdots & \vdots \\ \exp^{-a_k d_{n_k 1} - b_k d_{n_k 1}^2} & \exp^{-a_k d_{n_k 2} - b_k d_{n_k 2}^2} & \dots & 1 \end{bmatrix}$$

where $a_k, b_k \geq 0$ are unknown parameters and d_{ij} is the Euclidean distance between voxels i & j in the k^{th} cluster. Hence, the joint distribution of the $\boldsymbol{y_k}$ vector of WM measurements in k^{th} cluster is as follows:

$$\boldsymbol{y_k} \sim N(X^* \beta_k, \sigma_k^2 W_k), \tag{5.9}$$

where $X^* = \vec{1}_{n_k} \times X$ is $(n_k * MT) \otimes p$ matrix. For time being, we assume that W_k is known. W_k has three unknown parameters λ, a_k, b_k. (a_k, b_k) can be plugged in from our previous study, and we estimate λ from a data-driven method. Our primary focus is on parameters $\theta_k = (\beta_k, \sigma_k^2)$, and a convenient non-informative prior on $(\beta_k, \log(\sigma_k))$ is,

$$\pi(\beta_k, \sigma_k^2 | \omega) \propto \frac{1}{\sigma_k^2}$$

5.2.2.3 Objective function derivation:

Once we decide on the prior distributions, it is easy to derive posterior distributions. We determine the first posterior distribution for $\pi(\beta_k | \sigma_k^2, \boldsymbol{y_k}, \omega)$ conditional on σ_k^2 & ω. Next, we

ascertain the marginal posterior distribution of $\pi(\sigma_k^2|\boldsymbol{y_k}, \omega)$.

$$\begin{aligned}\left(\beta_k|\sigma_k^2, \boldsymbol{y_k}, \omega\right) &\sim N(\hat{\beta}_k, \sigma_k^2 V_{\beta_k}),\ where \\ \hat{\beta}_k &= \left(X^{*^T} W_k^{-1} X^*\right)^{-1} X^{*^T} W_k^{-1} \boldsymbol{y_k} \\ V_{\beta_k} &= \left(X^{*^T} W_k^{-1} X^*\right)^{-1}\end{aligned} \tag{5.10}$$

Marginal posterior distribution of σ_K^2 is

$$\begin{aligned}\left(\sigma_k^2|\boldsymbol{y_k}, \omega\right) &\sim Inv - \chi^2\left((n_k * MT) - p, s_k^2\right)\ where \\ s_k^2 &= \frac{1}{(n_k * MT) - p}\left(\boldsymbol{y_k} - X^*\hat{\beta}_k\right)^T W_k^{-1}\left(\boldsymbol{y_k} - X^*\hat{\beta}_k\right)\end{aligned} \tag{5.11}$$

To find $\pi(\omega|\boldsymbol{y})$, we need to integrate out β_k and σ_k^2 so that we can find

$$\begin{aligned}\pi(\boldsymbol{y}_k|\omega) &\propto \int_{\sigma_k^2}\int_{\beta_k} f(\boldsymbol{y}_k|\beta_k, \sigma_k^2, \omega)\pi(\beta_k, \sigma_k^2|\omega)d\beta_k d\sigma_k^2 \\ &\propto \Gamma\left(\frac{(n_k * MT) - p}{2}\right)\left(\frac{1}{2}\left(\boldsymbol{y}_k - X^*\hat{\beta}_k\right)^T W_k^{-1}\left(\boldsymbol{y}_k - X^*\hat{\beta}_k\right)\right)^{-\frac{(n_k*MT)-p}{2}}\end{aligned} \tag{5.12}$$

Finally, we product the above function over all clusters and multiply it with the prior distribution on cluster partitions $\pi(\omega)$. This provides us our long-desired objective function,

$$\begin{aligned}\pi(\omega|\boldsymbol{y}) &\propto \pi(\omega)\prod_{k=1}^{c(\omega)}\pi(\boldsymbol{y}_k|\omega) \\ &\propto \frac{\Gamma(m)m^{c(\omega)}}{\Gamma(n+m)}\prod_{k=1}^{c(\omega)}\Gamma(n_k)\Gamma((n_k * MT - p)/2)\left(\frac{1}{2}\left(\boldsymbol{y}_k - X^*\hat{\beta}_k\right)^T W_k^{-1}\left(\boldsymbol{y}_k - X^*\hat{\beta}_k\right)\right)^{-\frac{(n_k*MT)-p}{2}}\end{aligned} \tag{5.13}$$

The posterior function $\pi(\omega|\boldsymbol{y})$ is logically reasonable. The residual quadratic form calculates within the sum of squares in a cluster and prefers partitions with lower WSS. Furthermore, this posterior objective function prefers large homogeneous regions for large n_k . It also measures lack of fit from the linear mixed effect model we have assumed. Optimizing the objective function is challenging work.

CHAPTER 6

FUTURE WORK

The dissertation is focused on Bayesian variable selection and its application to brain image data. We extensively worked with Bayesian model based clustering in order to segment brain in similar regions. Optimization of the objective function would be our future works. As we have discussed earlier, the total number of possible combinations for 316 thousands voxels would be in billions. It's implausible to optimize the function for such a large space. In order to utilize this objective function we decided to reduce our sample space. We will find a large number of partitions using K-means described in section 5.1.2.2 with different cluster numbers. Once we finalize these partitions the above derived objective function will work as criterion to pick the best partition among selected large number of partitions.

We focused on the Metropolis-Hastings based on biased random walk algorithm described in *Booth et al., 2008*[13]. If we assume each partition as a node of an undirected graph and edges are created based on connection from one node to another. As authors described, from a particular partition if we move one element of a cluster and place in another cluster and obtain a new partition then these two partitions have an edge. Although following this logic we won't be able to obtain symmetric candidate density for Metropolis-Hastings algorithm. Our future work would be to propose and alternative graph creation which will provide symmetric candidate proposal. There is also a potential problem with our reduced sample space approach. As we've mentioned, we will subset our space with large (not infinite) number of partitions. There is a possibility that two separate partitions won't have any edges or they have no connection in terms of voxel distribution. We need more brainstorming in order to handle this situation. All these potential computational problems will be answered in our future work.

APPENDIX

APPENDIX

APPENDIX

A.1 Chapter 3: Tables for Classification performance using combination of ROIs

Table A1: Combination two ROIs

	sensitivity	specificity	accuracy	AUC
H + WB (Mean ± SD)	0.87 ±0.02	0.85 ±0.01	0.86 ±0.01	0.92 ±0.02
H + EC (Mean ± SD)	0.86 ±0.01	0.88 ±0.01	0.87 ±0.01	0.93 ±0.02
H + FG (Mean ± SD)	0.86 ±0.02	0.88 ±0.02	0.87 ±0.001	0.94 ±0.02
H + MTC (Mean ± SD)	0.84 ±0.02	0.90 ±0.01	0.88 ±0.01	0.93 ±0.02
WB + EC (Mean ± SD)	0.81 ±0.02	0.90 ±0.01	0.87 ±0.01	0.92 ±0.03
WB + FG (Mean ± SD)	0.77 ±0.03	0.84 ±0.01	0.81 ±0.01	0.89 ±0.03
WB + MTC (Mean ± SD)	0.78 ±0.01	0.88 ±0.01	0.84 ±0.01	0.90 ±0.02
EC + FG (Mean ± SD)	0.84 ±0.02	0.89 ±0.001	0.87 ±0.01	0.92 ±0.03
EC + MTC (Mean ± SD)	0.82 ±0.04	0.89 ±0.01	0.86 ±0.01	0.92 ±0.02
FG + MTC (Mean ± SD)	0.81 ±0.03	0.84 ±0.01	0.83 ±0.01	0.91 ±0.03

Notes: Sensitivity is proportion of correct AD prediction, Specificity is proportion of correct CN prediction, AUC is area under ROC curve. The mean and standard deviation are based on 100 repeated results on test data sets. Probability threshold is 0.5.

Table A2: Combination three ROIs

	sensitivity	specificity	accuracy	AUC
H+WB+EC (Mean ± sd)	0.86±0.02	0.86±0.02	0.86±0.01	0.91±0.05
H+WB+FG (Mean ± sd)	0.91±0.02	0.80±0.02	0.84±0.01	0.91±0.10
H+WB+MTC (Mean ± sd)	0.87±0.05	0.86±0.03	0.87±0.02	0.93±0.07
H+EC+FG (Mean ± sd)	0.85±0.03	0.87±0.02	0.86±0.001	0.93±0.03
H+EC+MTC (Mean ± sd)	0.83±0.02	0.90±0.01	0.87±0.01	0.93±0.03
H+FG+MTC (Mean ± sd)	0.86±0.02	0.87±0.01	0.86±0.01	0.93±0.02
WB+EC+FG (Mean ± sd)	0.84±0.04	0.86±0.01	0.85±0.02	0.91±0.03
WB+EC+MTC (Mean ± sd)	0.85±0.02	0.89±0.01	0.88±0.01	0.94±0.02
WB+FG+MTC (Mean ± sd)	0.82±0.02	0.86±0.03	0.84±0.02	0.91±0.03
EC+FG+MTC (Mean ± sd)	0.86±0.02	0.85±0.01	0.85±0.01	0.92±0.03

Notes: Sensitivity is proportion of correct AD prediction, Specificity is proportion of correct CN prediction, AUC is area under ROC curve. The mean and standard deviation are based on 100 repeated results on test data sets. Probability threshold is 0.5.

Table A3: Combination four ROIs

	sensitivity	specificity	accuracy	AUC
H+WB+EC+FG (Mean ± sd)	0.91±0.01	0.79±0.04	0.84±0.02	0.92±0.06
H+WB+EC+MTC (Mean ± sd)	0.77±0.04	0.89±0.03	0.84±0.02	0.89±0.10
H+WB+FG+MTC (Mean ± sd)	0.74±0.08	0.90±0.02	0.84±0.03	0.85±0.14
H+EC+FG+MTC (Mean ± sd)	0.85±0.02	0.84±0.02	0.84±0.02	0.91±0.05
WB+EC+FG+MTC (Mean ± sd)	0.85±0.03	0.86±0.02	0.86±0.01	0.92±0.03
H+WB+EC+FG+MTC* (Mean ± sd)	0.81±0.01	0.77±0.01	0.79±0.01	0.87±0.02

Notes: Sensitivity is proportion of correct AD prediction, Specificity is proportion of correct CN prediction, AUC is area under ROC curve. The mean and standard deviation are based on 100 repeated results on test data sets. Probability threshold is 0.5.
* Normalized volumes were used for this model run

A.2 Chapter 4: Posterior consistency proofs

Let us define $B(r_n) = sup_{\gamma=\gamma(r_n)} Ch(G_{r_n}^{-1})$ and $\bar{B}(r_n) = sup_{\gamma=\gamma(r_n)} Ch(G_{r_n})$ where $G_{r_n} = diag(\tau_1^* I_q, .., \tau_{r_n}^* I_q)$. $B(r_n)$ is the largest eigenvalues of G_{r_n} and $D(R) = 1+R.sup_{|h|\leq R}|a^{'}(h)|sup_{|h|\leq R}|\psi(h)|$. Let $\epsilon_n \longrightarrow (0, 1]$ with $n\epsilon_n^2 \succ 1$ and assuming the below conditions hold which come from *Jiang, 2007*[63] paper:

Conditions:

$$i)\ p_n \log(1/\epsilon_n^2) \prec n\epsilon_n^2$$

$$ii)\ p_n \log(p_n) \prec n\epsilon_n^2$$

$$iii)\ p_n \log(D(\frac{p_n}{\lambda_n}\bar{B}(r_n)n\epsilon_n^2)) \succ n\epsilon_n^2$$

$$iv)\ r_n \succ p_n$$

$$v)\ r_n \log(\bar{B}(r_n)n) \succ n\epsilon_n^2 \ \ and \ \ \Delta(r_n) \succ n\epsilon_n^2$$

$$vi) \log(\frac{r_n}{p_n}) \leq -\frac{4n\epsilon_n^2}{p_n}$$

Proof of Condition I: If $\psi(u) = e^u/(1+e^u)$, then

$$w(u) = \log[\psi(u)/(1-\psi(u))] = u$$
$$\implies w'(u) = 1$$
$$\implies |w'(u)| \leq C^q$$

Proof of Condition S: The proof starts with defining set and notations used in condition S. Let r_n be a large integer > 0 and η is small > 0, then

$$S(r_n, \eta) = \left\{(\gamma, \beta_\gamma)) : \gamma = \gamma(r_n), \beta_\gamma \in M(r_n, n)\right\}$$
$$M(r_n, n) = \left\{(b_1, .., b_{r_n})^T : b_j \in \beta_j \pm \frac{n\epsilon_n^2}{r_n},\ j = 1, .., r_n\right\}$$

Here r_n is the model size and $\gamma(r_n) = (1, 2, .., r_n, 0, ...)$ is an increasing sequence whose first r_n components take value 1.

Let, $1 \prec r_n \prec \min(p_n, n/\log(p_n))$ and $\sum_{j=1}^{\infty} ||\beta_j||_2 < \infty$.

$$\pi_n \left[\beta_\gamma \in \beta_j \pm \frac{n\epsilon_n^2}{r_n} | \gamma = \gamma(r_n)\right] \geq \prod_{j=1}^{r_n} \left[\frac{(\lambda_n^2)^{(q/2)}}{(2\pi)^{(q-1)/2}} \exp\left(-\lambda_n \sqrt{\bar{\beta}_j^T \bar{\beta}_j}\right) \left(\frac{n\epsilon_n^2}{r_n}\right)\right]$$

where $\bar{\beta}_j$ is some intermediate value which achieves the minimum density over $(\beta_j \pm \frac{n\epsilon_n^2}{r_n})_{j\in\gamma(r_n)}$. Then ,

$$\lambda_n \sum_{j=1}^{r_n} \sqrt{\bar{\beta}_j^T \bar{\beta}_j} \leq C_1 B(r_n)$$

as $\sum_{j=1}^{r_n} \sqrt{\bar{\beta}_j^T \bar{\beta}_j} \leq \lim_{n\to\infty} \sum_{j=1}^{p_n} \sqrt{\bar{\beta}_j^T \bar{\beta}_j} + \frac{n\epsilon_n^2}{r_n}$ is bounded. In addition we can show that,

$$\prod_{j=1}^{r_n} \frac{(\lambda_n^2)^{(q/2)}}{(2\pi)^{(q-1)/2}} \geq \exp(-C_2 r_n - C_3 r_n \log(\bar{B}(r_n))$$

where $\bar{B}(r_n) = sup_{\gamma=\gamma(r_n)} Ch(G_{r_n})$. Therefore,

$$\pi_n \left[\beta_\gamma \in \beta_j \pm \frac{n\epsilon_n^2}{r_n} | \gamma = \gamma(r_n)\right] \geq \exp\left(-C_2 r_n - C_3 r_n \log(\bar{B}(r_n) - C_1 B(r_n) - r_n \log(\frac{r_n}{n\epsilon_n^2})\right)$$
$$\geq \exp(-cn\epsilon_n^2)$$

To prove the prior condition: Let $\bar{r}_n$ such that $r_n < \bar{r}_n \leq p_n$ & $\bar{r}_n \prec n/ln(p_n)$. For our model we have placed $\pi_{0,n} \sim Beta\ distribution$ which is equivalent way of proposing Bernoulli distribution on $\gamma = \gamma(r_n)$ where $\gamma(r_n) \sim Bernoulli(\pi_{0,n})$ [115].

Now,

$$ln\pi_n = r_n ln\pi_{0,n} + (p_n - r_n) ln(1 - \pi_{0,n})$$

if $r_n \approx p_n \lambda_n$ then for $\pi_{0,n} = r_n/p_n$ small and $1 \prec r_n \prec \min(p_n, n/lnp_n)$

$$\implies ln\pi_n \geq -r_n lnp_n > -cn\epsilon_n^2 \ \ for\ large\ n$$
$$\implies \pi_n[\gamma = \gamma(r_n)] > \exp(-cn\epsilon_n^2)$$

Satisfying condition (S).

Proof of Condition L: Let us assume $D(R) = 1 + R.sup_{|h|\leq R}|a^{'}(h)|sup_{|h|\leq R}|\psi(h)|$ and there exists some C_n such that

$$\bar{r}_n ln(\frac{1}{\epsilon_n^2}) \prec n\epsilon_n^2$$

$$\bar{r}_n ln(p_n) \prec n\epsilon_n^2$$

$$\bar{r}_n lnD(\bar{r}_n C_n) \prec n\epsilon_n^2$$

then,

$$\pi_n(|\gamma| > \bar{r}_n) = \pi_n(|\gamma| = p_n) = (\frac{r_n}{p_n})^{p_n}$$
$$\implies ln(\pi_n(|\gamma| > \bar{r}_n) = p_n ln(\frac{r_n}{p_n}) \leq -cn\epsilon_n^2$$
$$\implies \pi_n(|\gamma| > \bar{r}_n) \leq e^{-cn\epsilon_n^2}$$

Next,

$$\pi_n(||\beta_j||_2 > t|\gamma) \propto \int_t^\infty e^{-\lambda_n\sqrt{\beta_j^T\beta_j}}\, d\beta_j$$
$$\leq \frac{1}{\lambda_n} e^{-\lambda_n t}$$

If $t = C_n = \frac{cn\epsilon_n^2}{\lambda_n}$ and $n\epsilon_n^2 > 1$, then

$$\frac{1}{\lambda_n} e^{-\lambda_n t} \leq e^{-cn\epsilon_n^2},\ as\ \ \lambda_n \geq 1$$

Satisfying condition (L).

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Stanisław Adaszewski, Juergen Dukart, Ferath Kherif, Richard Frackowiak, and Bogdan Draganski. How early can we predict alzheimer's disease using computational anatomy? *Neurobiology of aging*, 34(12):2815–2826, 2013.

[2] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[3] DR Anderson and K Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, page 63, 2004.

[4] Sönke Arlt, Ralph Buchert, Lothar Spies, Martin Eichenlaub, Jan T Lehmbeck, and Holger Jahn. Association between fully automated mri-based volumetry of different brain regions and neuropsychological test performance in patients with amnestic mild cognitive impairment and alzheimer's disease. *European archives of psychiatry and clinical neuroscience*, 263(4):335–344, 2013.

[5] Takeshi Asami, Sylvain Bouix, Thomas J Whitford, Martha E Shenton, Dean F Salisbury, and Robert W McCarley. Longitudinal loss of gray matter volume in patients with first-episode schizophrenia: Dartel automated analysis and roi validation. *Neuroimage*, 59(2):986–996, 2012.

[6] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[7] J Barnes, RI Scahill, C Frost, JM Schott, MN Rossor, and NC Fox. Increased hippocampal atrophy rates in ad over 6 months using serial mr imaging. *Neurobiology of aging*, 29(8):1199–1203, 2008.

[8] Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. 1988.

[9] Sam Behseta, Robert E Kass, and Garrick L Wallstrom. Hierarchical models for assessing variability among functions. *Biometrika*, 92(2):419–434, 2005.

[10] Andrew R Bender, Manuel C Völkle, and Naftali Raz. Differential aging of cerebral white matter in middle-aged and older adults: a seven-year follow-up. *Neuroimage*, 125:74–83, 2016.

[11] Jorge L Bernal-Rusiel, Douglas N Greve, Martin Reuter, Bruce Fischl, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66:249–260, 2013.

[12] Jorge L Bernal-Rusiel, Martin Reuter, Douglas N Greve, Bruce Fischl, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. Spatiotemporal linear mixed effects

modeling for the mass-univariate analysis of longitudinal neuroimage data. *Neuroimage*, 81:358–370, 2013.

[13] James G Booth, George Casella, and James P Hobert. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139, 2008.

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

[15] F DuBois Bowman, Brian Caffo, Susan Spear Bassett, and Clinton Kilts. A bayesian hierarchical framework for spatial modeling of fmri data. *NeuroImage*, 39(1):146–156, 2008.

[16] F DuBois Bowman and Clinton Kilts. Modeling intra-subject correlation among repeated scans in positron emission tomography (pet) neuroimaging data. *Human brain mapping*, 20(2):59–70, 2003.

[17] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009.

[18] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[19] Ramon Casanova, Benjamin Wagner, Christopher T Whitlow, Jeff D Williamson, Sally A Shumaker, Joseph A Maldjian, and Mark A Espeland. High dimensional classification of structural mri alzheimer's disease data based on large scale regularization. *Frontiers in neuroinformatics*, 5:22, 2011.

[20] George Casella. Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

[21] George Casella and Elias Moreno. Objective bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167, 2006.

[22] George Casella, Elías Moreno, F Javier Girón, et al. Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613–658, 2014.

[23] Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[24] Ray-Bing Chen, Chi Hsiang Chu, Shinsheng Yuan, and Ying Nian Wu. Bayesian sparse group selection. *Journal of Computational and Graphical Statistics*, 25(3):665–683, 2016.

[25] G Chetelat, B Landeau, F Eustache, F Mezenge, F Viader, V De La Sayette, B Desgranges, and J-C Baron. Using voxel-based morphometry to map the structural changes associated with rapid conversion in mci: a longitudinal mri study. *Neuroimage*, 27(4):934–946, 2005.

[26] Hee Min Choi, James P Hobert, et al. The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

[27] Evelyn M. Crowley. Product partition models for normal means. *Journal of the American Statistical Association*, 92(437):192–198, 1997.

[28] Christophe Delaloye, Guenael Moy, Fabienne De Bilbao, K Weber, S Baudois, Sven Haller, Aikaterini Xekardaki, Alessandra Canuto, U Giardini, K-O Lövblad, et al. Longitudinal analysis of cognitive performances and structural brain changes in late-life bipolar disorder. *International journal of geriatric psychiatry*, 26(12):1309–1318, 2011.

[29] Gordana Derado, F DuBois Bowman, Lijun Zhang, and Alzheimer's Disease Neuroimaging Initiative. Predicting brain activity using a bayesian spatial model. *Statistical methods in medical research*, 22(4):382–397, 2013.

[30] Vasilis K Dertimanis, Minas D Spiridonakos, and Eleni N Chatzi. Data-driven uncertainty quantification of structural systems via b-spline expansion. *Computers & Structures*, 207:245–257, 2018.

[31] Ilaria Dimatteo, Christopher R. Genovese, and Robert E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

[32] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[33] Yingying Fan, Gareth M James, Peter Radchenko, et al. Functional additive regression. *The Annals of Statistics*, 43(5):2296–2325, 2015.

[34] Yong Fan, Nematollah Batmanghelich, Chris M Clark, Christos Davatzikos, Alzheimer's Disease Neuroimaging Initiative, et al. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4):1731–1743, 2008.

[35] Anders M Fjell, Kristine B Walhovd, Christine Fennema-Notestine, Linda K McEvoy, Donald J Hagler, Dominic Holland, James B Brewer, and Anders M Dale. One-year brain atrophy evident in healthy aging. *Journal of Neuroscience*, 29(48):15223–15231, 2009.

[36] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[37] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[38] Chris Fraley and Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2):155–181, 2007.

[39] Jerome H. Friedman and Bernard W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21, 1989.

[40] Thomas Frodl, Annette Schaub, Sandra Banac, Marketa Charypar, Markus Jäger, Petra Kümmler, Ronald Bottlender, Thomas Zetzsche, Christine Born, Gerda Leinsinger, et al. Reduced hippocampal volume correlates with executive dysfunctioning in major depression. *Journal of Psychiatry and Neuroscience*, 31(5):316, 2006.

[41] Sylvia Frühwirth-Schnatter and Rudolf Frühwirth. Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, 51(7):3509–3528, 2007.

[42] Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[43] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[44] Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

[45] Robert B Gramacy, Nicholas G Polson, et al. Simulation-based regularized logistic regression. *Bayesian Analysis*, 7(3):567–590, 2012.

[46] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[47] Trevor J. Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning*. 2009.

[48] Nicholas A Heard, Christopher C Holmes, and David A Stephens. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, 101(473):18–29, 2006.

[49] WJP Henneman, JD Sluimer, J Barnes, WM Van Der Flier, IC Sluimer, NC Fox, Ph Scheltens, H Vrenken, and F Barkhof. Hippocampal atrophy rates in alzheimer disease added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009.

[50] Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, Alzheimer's Disease Neuroimaging Initiative, et al. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *Neuroimage*, 48(1):138–149, 2009.

[51] David B Hitchcock, George Casella, and James G Booth. Improved estimation of dissimilarities by presmoothing functional data. *Journal of the American Statistical Association*, 101(473):211–222, 2006.

[52] James P Hobert and Charles J Geyer. Geometric ergodicity of gibbs and block gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67(2):414–430, 1998.

[53] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[54] Chris C Holmes, Leonhard Held, et al. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

[55] Xue Hua, Suh Lee, Igor Yanovsky, Alex D Leow, Yi-Yu Chou, April J Ho, Boris Gutman, Arthur W Toga, Clifford R Jack Jr, Matt A Bernstein, et al. Optimizing power to track brain degeneration in alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an adni study of 515 subjects. *Neuroimage*, 48(4):668–681, 2009.

[56] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313, 2010.

[57] Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[58] Clifford R Jack, Ronald C Petersen, Yue Cheng Xu, Peter C O'Brien, Glenn E Smith, Robert J Ivnik, Bradley F Boeve, Stephen C Waring, Eric G Tangalos, and Emre Kokmen. Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1397, 1999.

[59] Clifford R Jack, Ronald C Petersen, Yuecheng Xu, Peter C O'Brien, Glenn E Smith, Robert J Ivnik, Eric G Tangalos, and Emre Kokmen. Rate of medial temporal lobe atrophy in typical aging and alzheimer's disease. *Neurology*, 51(4):993–999, 1998.

[60] Clifford R Jack, Ronald Carl Petersen, Y Xu, PC O'brien, Glenn E Smith, Robert J Ivnik, Bradley F Boeve, Eric George Tangalos, and E Kokmen. Rates of hippocampal atrophy correlate with change in clinical status in aging and ad. *Neurology*, 55(4):484–490, 2000.

[61] Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

[62] Wenxin Jiang. On the consistency of bayesian variable selection for high dimensional binary regression and classification. *Neural computation*, 18(11):2762–2776, 2006.

[63] Wenxin Jiang et al. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511, 2007.

[64] Iain M Johnstone, Bernard W Silverman, et al. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[65] Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. On discovering moving clusters in spatio-temporal data. In *International Symposium on Spatial and Temporal Databases*, pages 364–381. Springer, 2005.

[66] Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise sparse regression. *Statistica Sinica*, pages 375–390, 2006.

[67] Slava Kisilevich, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. Spatio-temporal clustering. In *Data mining and knowledge discovery handbook*, pages 855–874. Springer, 2009.

[68] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[69] Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[70] John W Lau and Peter J Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.

[71] Jonathan C Lau, Jason P Lerch, John G Sled, R Mark Henkelman, Alan C Evans, and Barry J Bedell. Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of alzheimer's disease. *Neuroimage*, 42(1):19–27, 2008.

[72] Sang Han Lee, Alvin H Bachman, Donghyeon Yu, Johan Lim, Babak A Ardekani, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting progression from mild cognitive impairment to alzheimer's disease using longitudinal callosal atrophy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2:68–74, 2016.

[73] Chenlei Leng, Minh-Ngoc Tran, and David Nott. Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244, 2014.

[74] Kelvin K Leung, Josephine Barnes, Gerard R Ridgway, Jonathan W Bartlett, Matthew J Clarkson, Kate Macdonald, Norbert Schuff, Nick C Fox, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer's disease. *Neuroimage*, 51(4):1345–1359, 2010.

[75] Jiahan Li, Zhong Wang, Runze Li, and Rongling Wu. Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The annals of applied statistics*, 9(2):640, 2015.

[76] Qing Li, Nan Lin, et al. The bayesian elastic net. *Bayesian analysis*, 5(1):151–170, 2010.

[77] Yingjie Li. *High Dimensional Classification for Spatially Dependent Data with Application to Neuroimaging*. Michigan State University, 2018.

[78] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[79] Anastasia Lykou and Ioannis Ntzoufras. On bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 23(3):361–390, 2013.

[80] Benoît Magnin, Lilia Mesrob, Serge Kinkingnéhun, Mélanie Pélégrini-Issac, Olivier Colliot, Marie Sarazin, Bruno Dubois, Stéphane Lehéricy, and Habib Benali. Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri. *Neuroradiology*, 51(2):73–83, 2009.

[81] Atreyee Majumder. *Variable Selection in High-Dimensional Setup: A Detailed Illustration Through Marketing and MRI Data*. Michigan State University, 2017.

[82] Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. *arXiv preprint arXiv:1812.07259*, 2018.

[83] Peter McCullagh and Jie Yang. Stochastic classification models. In *International Congress of Mathematicians*, volume 3, page 72, 2006.

[84] CR McDonald, LK McEvoy, L Gharapetian, C Fennema-Notestine, DJ Hagler, D Holland, A Koyama, JB Brewer, AM Dale, Alzheimer's Disease Neuroimaging Initiative, et al. Regional rates of neocortical atrophy from normal aging to early alzheimer disease. *Neurology*, 73(6):457–465, 2009.

[85] Mathew W. McLean, Fabian Scheipl, Giles Hooker, Sonja Greven, and David Ruppert. Bayesian functional generalized additive models with sparsely observed covariates. *arXiv preprint arXiv:1305.3585*, 2013.

[86] Mario Medvedovic, Ka Yee Yeung, and Roger Eugene Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

[87] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[88] Volodymyr Melnykov, Ranjan Maitra, et al. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010.

[89] Alan Miller. *Subset selection in regression*. Chapman and Hall/CRC, 2002.

[90] Chandan Misra, Yong Fan, and Christos Davatzikos. Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad: results from adni. *Neuroimage*, 44(4):1415–1422, 2009.

[91] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[92] Christian Montag and Martin Reuter. Internet addiction: Neuroscientific approaches and therapeutical interventions. 2015.

[93] HANS-GEORG MÜLLER. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005.

[94] Naveen Naidu Narisetty, Xuming He, et al. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

[95] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164, 2007.

[96] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[97] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[98] Nicholas G Polson, James G Scott, and Jesse Windle. The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733, 2014.

[99] Chengxuan Qiu, Miia Kivipelto, and Eva von Strauss. Epidemiology of alzheimer's disease: occurrence, determinants, and strategies toward intervention. *Dialogues in clinical neuroscience*, 11(2):111, 2009.

[100] Olivier Querbes, Florent Aubry, Jérémie Pariente, Jean-Albert Lotterie, Jean-François Démonet, Véronique Duret, Michèle Puel, Isabelle Berry, Jean-Claude Fort, Pierre Celsis, et al. Early diagnosis of alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047, 2009.

[101] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[102] Naftali Raz and Kristen M Kennedy. A systems approach to the aging brain: Neuroanatomic changes, their modifiers, and cognitive correlates. In *Imaging the aging brain*, pages 43–70. Oxford University Press, 2009.

[103] Naftali Raz, Ulman Lindenberger, Karen M Rodrigue, Kristen M Kennedy, Denise Head, Adrienne Williamson, Cheryl Dahle, Denis Gerstorf, and James D Acker. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex*, 15(11):1676–1689, 2005.

[104] Naftali Raz and Karen M Rodrigue. Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, 30(6):730–748, 2006.

[105] H Rusinek, Y Endo, S De Santi, D Frid, W-H Tsui, S Segal, A Convit, and MJ de Leon. Atrophy rate in medial temporal lobe during progression of alzheimer disease. *Neurology*, 63(12):2354–2359, 2004.

[106] Fabian Scheipl, Ludwig Fahrmeir, and Thomas Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012.

[107] Fabian Scheipl, Thomas Kneib, and Ludwig Fahrmeir. Penalized likelihood and bayesian function selection in regression models. *AStA Advances in Statistical Analysis*, 97(4):349–385, 2013.

[108] Flávio Luiz Seixas, Bianca Zadrozny, Jerson Laks, Aura Conci, and Débora Christina Muchaluat Saade. A bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment. *Computers in biology and medicine*, 51:140–158, 2014.

[109] Philip Shaw, Noor J Kabani, Jason P Lerch, Kristen Eckstrand, Rhoshel Lenroot, Nitin Gogtay, Deanna Greenstein, Liv Clasen, Alan Evans, Judith L Rapoport, et al. Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience*, 28(14):3586–3594, 2008.

[110] Li Shen, Yuan Qi, Sungeun Kim, Kwangsik Nho, Jing Wan, Shannon L Risacher, Andrew J Saykin, et al. Sparse bayesian learning for identifying imaging biomarkers in ad prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–618. Springer, 2010.

[111] Guiling Shi. *Bayesian Variable Selection: Extensions of Nonlocal Priors*. Michigan State University. Statistics, 2017.

[112] Russell T Shinohara, Ciprian M Crainiceanu, Brian S Caffo, and Daniel S Reich. Longitudinal analysis of spatiotemporal processes: a case study of dynamic contrast-enhanced magnetic resonance imaging in multiple sclerosis. 2011.

[113] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[114] Martha Skup, Hongtu Zhu, and Heping Zhang. Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. *Biometrics*, 68(4):1083–1092, 2012.

[115] Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

[116] Richard P. Stanley. *Enumerative Combinatorics: Volume 1*. 1986.

[117] Adam J Suarez, Subhashis Ghosal, et al. Bayesian estimation of principal components for functional data. *Bayesian Analysis*, 12(2):311–333, 2017.

[118] Tero Tapiola, Corina Pennanen, Mia Tapiola, Susanna Tervo, Miia Kivipelto, Tuomo Hänninen, Maija Pihlajamäki, Mikko P Laakso, Merja Hallikainen, Anne Hämäläinen, et al. Mri of hippocampus and entorhinal cortex in mild cognitive impairment: a follow-up study. *Neurobiology of aging*, 29(1):31–38, 2008.

[119] Paul M Thompson, Kiralee M Hayashi, Greig I De Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, Michael S Hong, David H Herman, David Gravano, David M Doddrell, et al. Mapping hippocampal and ventricular change in alzheimer disease. *Neuroimage*, 22(4):1754–1766, 2004.

[120] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[121] Angelika Van Der Linde. A bayesian latent variable approach to functional principal components analysis with binary and count data. *AStA Advances in Statistical Analysis*, 93(3):307–333, 2009.

[122] Geert Verbeke and Geert Molenberghs. A model for longitudinal data. *Linear mixed models for longitudinal data*, pages 19–29, 2000.

[123] Xuejing Wang, Bin Nan, Ji Zhu, and Robert Koeppe. Regularized 3d functional regression for brain image data via haar wavelets. *The annals of applied statistics*, 8(2):1045, 2014.

[124] Jennifer L Whitwell, Scott A Przybelski, Stephen D Weigand, David S Knopman, Bradley F Boeve, Ronald C Petersen, and Clifford R Jack Jr. 3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease. *Brain*, 130(7):1777–1786, 2007.

[125] Jennifer L Whitwell, Maria M Shiung, SA Przybelski, Stephen D Weigand, David S Knopman, Bradley F Boeve, Ronald C Petersen, and CR Jack. Mri patterns of atrophy associated with progression to ad in amnestic mild cognitive impairment. *Neurology*, 70(7):512–520, 2008.

[126] Jesse Windle, Nicholas G Polson, and James G Scott. Sampling polya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.

[127] Mark William Woolrich, Mark Jenkinson, J Michael Brady, and Stephen M Smith. Fully bayesian spatio-temporal modeling of fmri data. *IEEE transactions on medical imaging*, 23(2):213–231, 2004.

[128] Xiaofan Xu, Malay Ghosh, et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

[129] Lan Xue. Consistent variable selection in additive models. 2009.

[130] Wenlu Yang, Xinyun Chen, Hong Xie, and Xudong Huang. Ica-based automatic classification of magnetic resonance images from adni data. In *Life System Modeling and Intelligent Computing*, pages 340–347. Springer, 2010.

[131] Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[132] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[133] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS one*, 7(3):e33182, 2012.

[134] Lin Zhang, Veerabhadran Baladandayuthapani, Bani K Mallick, Ganiraju C Manyam, Patricia A Thompson, Melissa L Bondy, and Kim-Anh Do. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620, 2014.

[135] Xiaoxi Zhang, Tim Johnson, Rod Little, and Yue Cao. Longitudinal image analysis of tumor/brain change in contrast uptake induced by radiation. 2009.

[136] Hongxiao Zhu, Marina Vannucci, and Dennis D Cox. A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics*, 66(2):463–473, 2010.

[137] Vadim Zipunnikov, Sonja Greven, Haochang Shou, Brian S. Caffo, Daniel S. Reich, and Ciprian M. Crainiceanu. Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *Ann. Appl. Stat.*, 8(4):2175–2202, 12 2014.

[138] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.