VARIABLE SELECTION IN HIGH-DIMENSIONAL SETUP: A DETAILED ILLUSTRATION THROUGH MARKETING AND MRI DATA

By

Atreyee Majumder

A DISSERTATION

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

Statistics - Doctor of Philosophy

2017

ABSTRACT

VARIABLE SELECTION IN HIGH-DIMENSIONAL SETUP: A DETAILED ILLUSTRATION THROUGH MARKETING AND MRI DATA

By

Atreyee Majumder

In the times of big data and ever growing information, variable selection is an integral part of statistical analysis. With the advancement of technology, we are able to store and access large volumes of data, only part of which is required for inference. Variable selection is a statistical technique that helps us retain valuable information while discarding everything that is non-significant.

To understand variable selection, we perform a comparative study of various popular frequentist variable selection techniques. This study analyses the difference of performance of models based on Ridge, LASSO and Elastic Net methods of penalized regression. The comparison of these methods is done for both continuous and binary outcome. We further emphasize the importance of tuning parameter selection in penalized regression models. This is done by comparing 6 different methods of tuning parameter selection for each penalized approach. The best performing method is then chosen to build statistical models for market research data of 4 varied countries. This exercise is an application of variable selection. Here, we showcase the applicability of such models in handling large information efficiently, for managerial decisions. We show how managers can leverage this technique for better resource allocation in their business decisions.

Next, we build a model for variable selection in a Bayesian setup. This is motivated by the fact that the frequentist approaches have unstable inference. Here, we analyze Alzheimer's Disease Neuroimaging Initiative (ADNI) with a Bayesian model. This is done by building a

Bayesian hierarchical model with multivariate Laplace priors in spike and slab prior style. This model is able to select a group of related variables. The frequentist counterpart of this estimator, group lasso, is also discussed. We build a classification model that is able to select the significant brain regions in Alzheimer's disease with 80% accuracy. Instead of using standard MAP thresholding, we use posterior median thresholding for variable selection. Furthermore, the consistency of this estimator is also proved.

Lastly, we build a Bayesian structured model for variable selection based on magnetic resonance imaging (MRI) data. This model is an extension of the second method but takes into account bi-level selection and spatio-temporal correlation. Voxels in brain regions have spatial correlation and repeated measurements for each voxel which brings in temporal correlation. This model is applied on a simulated functional MRI (fMRI) type data and real data. The real data detects blood oxygenation level dependent (BOLD) activation. The data is large on the account of numerous voxels present in the brain. Our method, successfully, detects the activated brain regions in the presence of a stimuli.

Thus, this thesis delves into various scenarios of variable selection with three different real data application studies. The focus is mainly on Bayesian variable selection and the use of hierarchical modeling with iterative sampling from posterior distribution in the group lasso setup. Our application of using group lasso structure to identify brain regions and voxels is an innovative approach in the context of present literature review. All of these methods have practical implication that can be used to solve relevant real world problems.

To my beloved parents Subir Kumar Majumder and Kalyani Majumder for their unconditional love and support.

ACKNOWLEDGMENTS

I am extremely obliged to be a part of the Statistics and Probability Department at Michigan State University. In my tenure of five years, I have worked with some amazing professors who have been a building block in my academic advancement. I would like to thank my PhD. advisor, Dr. Tapabrata Maiti, for his immense support, encouragement and guidance in my research. I am grateful to him for believing in my potential and giving me the opportunity of working with him.

I have been extremely fortunate to work with Dr. Alla Sikorskii for 4 years in her research projects with the College of Nursing. I would like to thank her for providing me this opportunity of learning and growth and also, for being a part of my research committee. My sincere gratitude goes to Dr. Roger Calantone for providing me with the market research dataset and collaborating on one of my projects. I would like to thank my research committee member, Dr. Gustavo de los Campos, for agreeing to be a part of my committee and sharing knowledge on Bayesian techniques. I also had the opportunity of working with Dr. David Zhu, Department of Radiology, MSU, for patiently describing details of fMRI data and helping me understand and work with brain data.

I am thankful to the Department of Statistics and Probability and College of Nursing at Michigan State University for supporting me financially through my PhD. career. My sincere thanks to all professors in the department for helping me glide through these years of research. Special thanks to Dr. Ramamoorthy and Dr. Levental for preparing us for the PhD/ qualifier courses. I have worked closely with many of my peers and would like to thank them for engaging in useful converstaions that led to the development of some useful ideas in my research.

I would like to thank my family - my parents, Mr. Subir Kumar Majumder and Mrs. Kalyani Majumder, for their love and support through the tough times. They are my pillars of strength. Last but not the least, I feel immense gratitude towards my husband, Mr. Subha Tirtha Datta, for helping me pursue my dreams. This journey is rewarding because he is a part of it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Variable Selection

Variable selection is an extremely relevant area of statistical modeling, especially in modern times. Development of high performance computing machines have paved a way for storage of large volumes of data. Today, most fields of research are able to store data on a variety of aspects. For example, a financial firm or a medical researcher may include information about subjects that may not be relevant from an apparent view but might deem to be important if explored with statistical analysis. Researchers do not want to miss out on useful information, so a lot of data is stored only to selectively use them at a later stage.

Although, the availability of large volumes of data is welcome, statisticians have to be careful to use a subset of it for valid statistical inference. In the real world, most data that is captured contains very little information for a specific outcome. In genetic data, for example, information on thousands of SNPs are available, but, only a few are able to completely characterize a disease. When building a statistical model, we want to develop a relationship that is able to explain an association between an outcome and some features with a very small error rate. In doing so, we do not wish to sacrifice the optimality of the model. One might think that inclusion of more information will reduce the error rate but, we also need to consider the complexity of this model.

Thus, building a model and selecting an optimal model have become interchangeable terms these days. Variable selection is not a new concept in statistical model building so, many procedures have existed for long. An ordinary regression equation is as follows:

$$y_i = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p, \quad i = 1, \ldots, n \tag{1.1}$$

Here, there are p independent variables which are used to build a model for predicting y. Note that, $p < n$ and all the p covariates may not be statistically significant for predicting y. We use stepwise selection procedures for selecting the significant covariates. In these procedures, we start with all variables (backward selection) or no variable at all (forward selection) and eliminate or add covariates based on their corresponding p-value. This is continued until no more covariates are eliminated or added in the model. From the description, it is evident that these methods are time cumbersome as we need to validate the significance of variables at each step until it converges. The step by step procedure may miss the optimal model and using p-values to add/drop variables may not be appropriate in cases, such as, when multicollinearity is present. Subset selection is another approach of variable selection where various candidate models are validated for optimality using a choice of metric. The problem here is that, with the increase in number of covariates, the number of candidate models increase exponentially. So, computationally this becomes a complicated problem.

In the past few years, newer approaches use penalization of the likelihood (objective) function to obtain a sparse solution. The penalty is put on the parameter space so that the parameter estimates are encouraged to move to zero if they are not statistically significant. These methods are better understood in a high-dimensional setup where $p > n$. In such a case, it is clear that ordinary regression breaks down due to issues of invertibility of the low

rank design matrix. Penalization approaches deal with this issue as well as perform variable selection, making these methods the most sought after for variable selection in this era.

1.1.1 High-Dimensional Data

High dimensional feature space arises when we have more number of feature variables than the number of observations. This scenario is commonplace in various fields of studies like social sciences, marketing research, genetic sequencing, machine learning, brain image analysis etc. For instance, in brain image data, we have tens of thousands of voxel level data and only a few responses. In genomics, hundreds of thousands of SNPs are potential covariates for a particular phenotype. In market research or social science data, we have more data on feature variables that are obtained from questionnaires, than the number of individuals who can be interviewed. All of these cases show that high-dimensional data occurs frequently. The problems that arise with high-dimensional setup should be tackled appropriately since its occurrence is unavoidable in many practical domains of research.

In the scenario of $p > n$, we can no longer pursue ordinary regression due to the non-invertibility of $X^T X$. In such a case, it is helpful to assume that our regression function lies in a low dimensional space (Fan and Lv, 2010). This can be applied by introducing the concept of sparsity. If we assume that many covariates have zero estimates then we can reduce our dimension of interest to a much smaller space.

Assuming sparsity means we believe that many of the covariates do not have significant impact on the outcome. This assumption is practical in almost all the research areas that are affected by the curse of dimensionality. In genetic studies, there is information on numerous genes but only a few are actually associated with the occurrence of a disease. The assumption of sparsity is thus valid for simpler statistical analysis and also for researchers

who would be able to make more sense from a subset of associated genes scientifically. In brain magnetic resonance imaging, the scanner collects information of the entire brain region (containing thousands of voxels), but only a few of these voxels are activated on the presence of a stimuli. Thus, including information that is not activated will mislead biologists. Therefore, assumption of sparsity makes ground for better model building as well as improved interpretation of the analysis.

Variable selection in high-dimensional setup is an important area of research. It is highly sensitive because it has real world applications. A proper statistical model based on appropriate information and consequently, its precise implications can answer a lot of questions in the medical, genetic, financial, economic, machine learning and social sciences domains. This research problem has a greater contribution to the betterment of everyday issues at large.

1.1.2 Penalized Regression Approaches

Consider the vector of response Y and the matrix of covariates X. Also, let ω be the vector of parameters and ϵ be the random error associated with the model. Assume, $\epsilon \sim N(0, \sigma_\epsilon^2)$. Let n be the sample size and p be the number of parameters in the model, then the classical regression model is:

$$Y = X\omega + \epsilon$$

The ordinary least squares method approach of estimating the parameters gives $\hat{\omega}_{ols} = (X^T X)^{-1} X^T Y$. We know that $\hat{\omega}_{ols}$ is non- estimable when $p > n$ or in the presence of collinearity. To overcome this issue we introduce the following penalized regression methods which minimizes the sum of squares of errors subject to some constraints. We assume that

the underlying feature (design) matrix is sparse with only a few predictive features. To elaborate this idea we look at the following three models that places a constraint on the parameters to induce sparsity.

Ridge regression (Hoerl and Kennard, 1970) shrinks the estimates of the coefficients towards zero. In this approach we minimize

$$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda\sum_{j=1}^{p}\omega_j^2$$

λ is the tuning parameter which controls the trade-off between bias and variance of the estimates.

The LASSO (Tibshirani, 1996) puts an L_1-penalty on the parameters. In this approach we minimize

$$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda\sum_{j=1}^{p}|\omega_j|$$

where λ is the tuning parameter. This penalized approach forces many ω to take 0 values and works consistently when $p > n$. The LASSO is an attractive tool due to simultaneous estimation and variable selection. But, it does not perform satisfactorily when multicollinearity occurs.

To improve on these methods, Zou and Hastie (2005) proposed the elastic net which includes feature of both the Ridge and the LASSO estimators. Here, we put a convex combination of the ridge and LASSO penalties on the parameters. Thus, elastic net minimizes

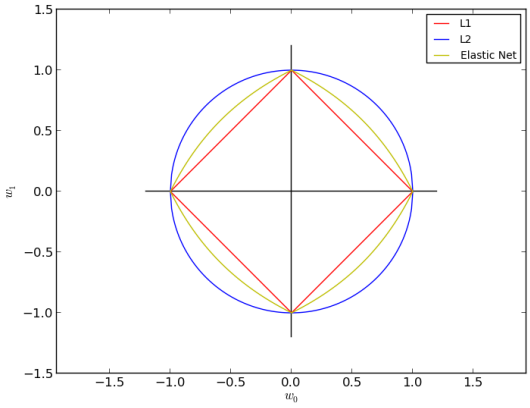
$$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda P_\alpha(\omega)$$

where λ is the tuning parameter and $P_\alpha(\omega) = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\omega_j^2 + \alpha|\omega_j|\right]$. (Zou and Hastie

(2005) called this penalty as the naive elastic net penalty and called a rescaled version elastic net penalty, but we drop this distinction here).

Figure 1.1 (Source: http://scikit-learn.sourceforge.net/0.7/modules/sgd.html) shows that the edges of LASSO and elastic net creates more opportunities for some estimates to be zero. This feature guarantees sparsity of estimates. The convex edges of ridge and elastic net encourages grouping effect. Thus, ridge penalty is good when there is multicollinearity. It does not generate a sparse model, although, it does shrink the estimates towards 0. The LASSO on the other hand generates a sparse model but fails to do grouped selection. It generally selects one variable from a group and drops the others. Since, elastic net is a convex combination of the ridge and LASSO penalty it reflects properties of both. The vertices guarantee sparsity and the convexity encourages grouping effect.

Figure 1.1: Geometry of the ridge, LASSO and elastic net ($\alpha = 0.5$)



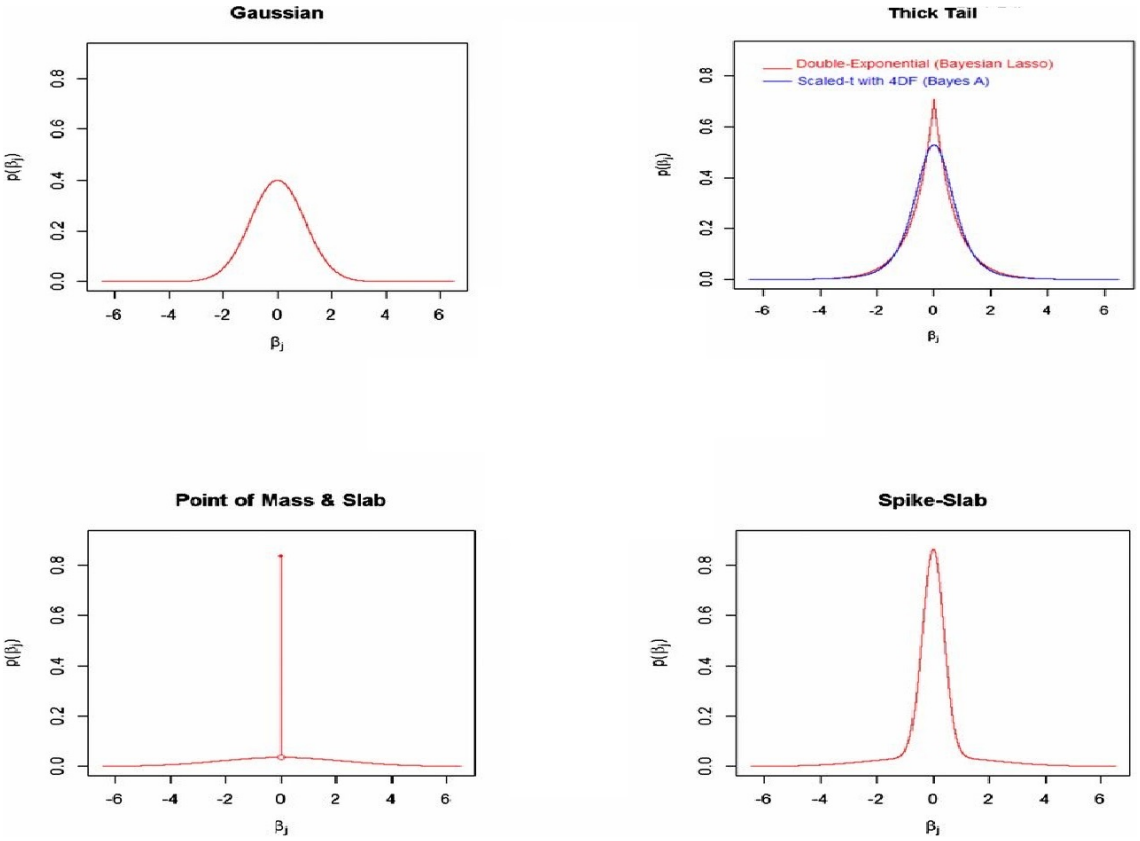
Other penalized approaches include SCAD, adaptive LASSO, group lasso etc. All of these methods use some penalty parameter, also known as tuning parameter, to induce sparsity in the model. Regularization in group lasso is introduced on a group of related variables (dummy variables or basis splines).

1.1.3 Bayesian Penalized Regression

Both frequentist and Bayesian approaches have been explored extensively for variable selection. Although, frequentist variable selection approaches are popular and useful, they provide unstable standard errors of the estimates. This makes it imperative for us to look for alternatives that are more statistically reliable. Bayesian approaches for variable selection are based on the assumption that there is *a priori* probability associated with each subset model and we select the model that has the highest posterior probability.

The key notion in variable selection is shrinkage of parameter estimates towards zero. Bayesian methods offer natural shrinkage with a proper choice of prior. Shrinkage in Bayesian literature is in terms of the estimates being shrunk towards prior belief. A careful choice of prior on the regression parameters will encourage sparsity in the model. Tibshirani (1996) noted that the posterior mode of Laplace priors on parameters will give identical estimates as that of the lasso. This has motivated (e.g., Figueiredo 2003; Bae and Mallick 2004; Yuan and Lin 2005) to use i.i.d. Laplace priors on coefficients, β's, to develop Bayesian lasso-like estimates. A normal prior is equivalent to Bayesian ridge regression where the estimates are obtained by maximizing the posterior. A spike and slab prior puts some weight on a flat density and the remaining on a spiked density thus making it a weighted mixture of two densities. The spike part encourages shrinkage and is especially helpful when number of predictors is larger than number of observations. A mixture of a point mass density and a normal density is special case of spike and slab; we will call this version spike and slab prior throughout this thesis. Figure 1.2 (de los Campos *et. al.*, 2013) shows how the prior information in Bayesian analysis naturally encourages the posterior estimates to shrink towards prior belief.

Figure 1.2: Shrinkage in four different priors



Park and Casella (2008) developed a fully Bayesian lasso by using conditional Laplace priors on β with a non-informative prior on the variance parameter to ensure unimodality. A good mixing property of normal densities with exponential density results in a Laplace prior. This property is used to formulate a hierarchical Gibb's model so we get full conditionals of all the parameters involved.

The solution path of this Bayesian posterior estimate is similar to that of the lasso estimate. In most high-dimensional setup, rarely does a model occur with very high frequency, so instead of looking at the maximum posterior probability, it is often feasible to use posterior means as the estimates of β. Posterior means do not directly give zero estimates, so we use posterior median thresholding since median is a natural thresholding estimator.

1.1.4 Frequentist Group Lasso

Variable selection is a technique of selecting an optimal model in predictive modeling. In many regression problems, we are interested in selecting feature variables that are important in predicting the response variable. The feature variables can be individual numeric variables, various levels of a categorical variable or a number of basis functions of the original measured variables. Recently proposed methods like the LASSO, SCAD etc. can efficiently perform variable selection by selecting individual feature variables. In case of an ANOVA type model where there are multiple levels of a feature variable or for an additive model where each component is a linear combination of a number of basis functions, selecting the important variable amounts to selecting all levels of the variable.

A very simple linear regression equation is of the form:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \tag{1.2}$$

Here, $\mathbf{X}$ is the design matrix whose columns are the feature variables, β is the vector of coefficients, ϵ is the error vector where each ϵ_i has a normal distribution with mean 0 and variance σ^2 and $\mathbf{Y}$ is the vector of observations.

Each feature variable in equation (1.2) can be either categorical or continuous. ANOVA is a special case where all the input variables are categorical whereas an additive model is a special case of all continuous input variables. However, the input variables could be a mixture of both numeric and categorical variables in a regression problem given by equation (1.2).

When we want to work with factor variables with G factors (groups) then we can modify our notations in equation (1.2) as follows:

$$\mathbf{Y}_{n\times 1} = \sum_{g=1}^{G} \mathbf{X}_g \beta_g + \epsilon \tag{1.3}$$

where $\epsilon_{n\times 1} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, β_g is a coefficient vector of length m_g, and $\mathbf{X}_g$ is an $n \times m_g$ covariate (feature) matrix corresponding to the factor $\beta_g, g = 1, 2 \dots G$. Let p be the total number of predictors, so $p = \sum_{g=1}^{G} m_g$. To eliminate the intercept from equation (1.3), we center response variables and each input variable so that the observed mean is 0.

The goal is to select important feature variables for accurate prediction. This amounts to selecting as well as estimating the parameter coefficients. Tibshirani (1996) proposed the LASSO method which is an attractive tool due to simultaneous estimation and variable selection. When the need for selecting a group of levels of a categorical variable or group of basis functions representing a numeric variable arise, these methods fail because they are designed to select individual feature variables and fail to select whole factors. Yuan and Lin (2006) proposed group lasso as an alternative to LASSO in terms of factor selection and also exhibit superior model selection performance.

The group lasso penalty is a hybrid of the l_1 and l_2- penalties and encourages selection at a group level. The group lasso estimate, for linear regression, minimizes

$$||Y - \sum_{g=1}^{G} X_g \beta_g||_2^2 + \lambda \sum_{g=1}^{G} ||\beta_g||_2 \tag{1.4}$$

where λ is the tuning parameter. Note that, in (1.4), when all groups have size 1 i.e. $m_1 = m_2 = \cdots = m_G = 1$, we have lasso.

1.1.5 Bayesian Group lasso

The limiting distribution of the group lasso estimator is complicated (Knight and Fu, 2000; Chatterjee and Lahiri, 2011). Thus, this estimator fails to give meaningful standard errors of the estimates which affects the statistical significance of the covariates in the chosen model. To deal with this drawback of frequentist lasso type estimators, Bayesian formulations have been developed. The Bayesian MAP estimators provide reliable standard errors for the estimates.

It is known that the lasso estimator for linear regression is equivalent to the posterior mode with independent Laplace priors on each regression coefficient. Park and Casella (2008) developed a fully hierarchical Bayesian setup for the lasso using a scale mixture prior on the regression parameters. This mixture prior results in a Laplace marginal distribution for β. This idea has been further extended to build similar fully Bayesian Hierarchical models for group lasso, fused lasso (Tibshirani *et. al.*, 2005) and the elastic net (Zou and Hastie, 2005) by Kyung *et. al.* (2010). They employ a multivariate m_g- dimensional Laplacian prior over each group of regression coefficients.

$$\pi(\beta_g) \propto \exp\Big\{ -\frac{\lambda}{\sigma}||\beta_g||_2\Big\}, \tag{1.5}$$

The classical group lasso is recovered as the MAP solution in log-space with $\frac{\lambda}{\sigma}$ having the role of a fixed Lagrangian multiplier. For a full Bayesian treatment, however, we place hyperpriors on λ and σ which lead to integrations that are analytically impossible to solve.

For finding closed form posterior distributions for all parameters we extend the hierarchical scale mixture model approach of lasso to grouped predictors. Thus, we express the prior as a scale mixture of multivariate normals over β_g with Gamma hyperpriors over the

variance hyperparameter. Specifically, with

$$\beta_g|{\tau_g}^2, \sigma^2 \sim^{ind} \mathbf{N}_{mg}\left(\mathbf{0}, {\tau_g}^2\sigma^2\mathbf{I}_{mg}\right), {\tau_g}^2 \sim^{ind} \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right) \tag{1.6}$$

the marginal distribution of β_g is of the form (1.5). This Bayesian formulation encourages shrinkage at the group level and provides comparable prediction performance with the group lasso. However, estimation of β_g by its posterior mean or median does not produce exact 0 estimates; we need to bring in the concept of sparsity here. Thus, to introduce sparsity at group level, Xu and Ghosh (2015) assumed a multivariate zero inflated mixture prior or a spike and slab prior for each β_g.

For variable selection, we want the estimates to produce exact zeroes such that they are dropped from the model. Zero inflated mixture priors are such that the slab part draws values from a known distribution and the spike part is degenerate distribution selecting zero. Xu and Ghosh (2015), further showed that median thresholding is better than using posterior mean. The spike and slab prior keeps the scale mixture prior of normals and gamma intact thus providing full conditionals. This approach is thus computationally easy and gives exact zero estimates. Narisetty and He (2014) used shrinking and diffusing priors for variable selection in a hierarchical Bayesian setup. Zero inflated mixture priors, in recent years, have been extensively utilized in Bayesian variable selection setups. George and McCulloch (1997) used zero inflated normal mixture priors in the hierarchical formulation for variable selection in a linear regression model. Chen and Dunson (2003) used a spike and slab type prior for the random effects variances in a linear setup allowing probabilistic selection of random effects.

1.1.6 Bayesian Group Lasso with Logistic Regression

So far we have talked about group lasso in a linear regression setup i.e. when the response variable has a Gaussian error. In many practical problems, we come across response values that cannot be fit into a linear model. For example, when the outcome is a binary categorical variable, count data or multi-level categorical variable then the Gaussian error assumption does not hold. In such cases we have to use generalized linear models (GLM) with various link functions. In many financial, insurance and medical data the outcome has two values thus making it a binary response variable. Since, the occurrence of binary outcome is very common in the real world we will focus on GLM with a logit link.

Since, the outcome is binary we cannot model this data with (1.1) having normal errors. Meier *et. al.* (2008) developed the logistic group lasso in a frequentist setup. Before delving into its Bayesian counterpart, let us summarize the frequentist group lasso in logistic regression setup.

Assume that we have independent and identically distributed observations $(x_i, y_i),\ i = 1, \ldots, n$, of a $p-$dimensional vector $x_i \in \mathbb{R}^p$ of G predictors and a binary response variable $y_i \in \{0, 1\}$. Each group has m_g levels. We can write $x_i = (x_{i1}^T, \ldots, x_{iG}^T)^T$. Linear logistic regression models the conditional probability $p_\beta(x_i) = \mathbb{P}_\beta(\mathbf{Y} = 1|x_i)$ by

$$\log\left\{\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right\} = \eta_\beta(x_i)$$

also known as the logit link with the link function $\eta_\beta(x_i) = \sum_{g=1}^{G} x_{ig}^T \beta_g$. The logistic group

lasso estimator, β_{GL}, is given by the minimizer of the convex function

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} ||\beta||_2$$

where $l(.)$ is the log-likelihood function i.e.

$$l(\beta) = \sum_{i=1}^{n} \big(y_i \eta_\beta(x_i) - \log[1 + \exp\{\eta_\beta(x_i)\}] \big).$$

The tuning parameter $\lambda \geq 0$ controls the amount of penalization.

Motivated by Xu and Ghosh's (2015) work, here we construct a Bayesian formulation for the logistic regression case. Here, our likelihood is Bernoulli probability mass function with a logit link. We abide by using a multivariate zero inflated mixture prior with point mass at zero and the continuous part as double exponential distribution. Since a double exponential prior on β_g can be expressed as a scale mixture of normal and Gamma priors (as in (1.6)), we use priors very similar to the linear setup.

1.2 Applications of Variable Selection

With the main focus of variable selection in our mind, we have developed some Bayesian hierarchical models for model selection by employing an efficient Gibbs' sampler. There are numerous areas in which our models can be applied; we have chosen brain imaging data to illustrate the usability of our methods. Brain imaging techniques are widely used to capture brain activity that can be used to detect various diseases related to brain deformity. We use two types of brain image data to build Bayesian hierarchical models. Although, brain image data analysis is the primary focus of this thesis, we start our illustrations with a comparative

study based on market research data.

We use market research questionnaire data to illustrate how there different variable selection methods differ in the frequentist approach. This exercise also emphasizes the need of meticulous selection of the tuning parameter and its effect on model building. We show how penalized approaches can be used to build predictive models that are relevant in a variety of areas.

1.2.1 Market Research Data: A Comparative Study

International market segmentation inevitably stands out as fundamental and crucial in global marketing strategy matching the increase in international trade. Successful targeting of customers and positioning firms' products requires insights of customer heterogeneity across borders, namely international segmentation. Moreover, international segmentation provides a strategic perspective on the balance between standardization and adaptation (Verhage, Dahringer, and Cundiff, 1989), where standardization has the economic advantages, and adaption enables firms to better satisfy specific customer needs.

In general, international market segmentation addresses heterogeneity at three levels - countries, regions, and consumers. The former two are usually defined by geography, while the last one has no aggregation level. Although geographic aggregation for international segmentation is widely accepted, it generally overlooks consumer-level heterogeneity within segments and, more importantly, "it is neither theoretically motivated nor is the managerial relevance of the segmentation variables established" (Nachum, 1994; Steenkamp and Ter Hofstede 2002). Due to effective logistic capability, mass customization, and the death of distance due to advances in consumer communications, company offerings and consumers' preferences are rarely constrained by geographic aggregation. Hence, disaggregate interna-

tional consumer segmentation offers a more effective way to examine and consider consumers' needs.

Variable inclusion in the international market segmentation domain requires a bifocal view of data employed, regardless the type of segmentation. In this study, we focus on model selection for lifestyle segmentation across countries. Additionally, we look at the very realistic situation where significant disparity exists between sample size and number of variables, often referred to as the high dimensionality problem. Variable selection is done by obtaining the maximum information required with optimal utilization of resources. Variable selection, when the number of predictors exceeds number of observations, has largely been overlooked in marketing research. In many situations, survey questions exceed the number of people interviewed. This leaves us with the problem of variable selection in a high-dimensional setup. Conventional statistical methods fail in this case. We illustrate three statistical methods that is custom designed for such scenarios. Variable selection reduces dimensionality of the wide data and selects variables most significant for segmentation.

In the problem of assessing the relation of covariates with a response variable we often face challenges such as multicollinearity and high dimensionality. Often, they occur simultaneously. The ordinary least squares method for regression fails to estimate the coefficients of a regression model in these cases due to the non-invertibility of the X^TX matrix where X is the design matrix. Typical examples of such scenarios arise in social networking sites, product reviews and market portfolio data. Penalized regression methods are getting popularity for these scenarios, see Hastie et al. (2009) and Fan and Lv (2010) for overviews among a large amount of recent literature. Some of the penalized methods used to overcome these challenges are ridge (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), SCAD penalty (Fan and Li, 2001) etc.

These methods can be used when the data has multicollinearity or high-dimensionality but they differ in terms of their variable selection efficiency or predictive power. For instance, the ridge penalty shrinks the coefficients toward zero but fails to perform variable selection. The LASSO, although helpful in variable selection, shrinks the coefficients to zero and may provide larger prediction errors. The elastic net, SCAD and adaptive LASSO address a few of these inconsistencies but none of these procedures are proved to be a generic best procedure. Although these methods are effective in dealing with the non-invertibility issue of the OLS but their performance is highly dependent on the selection of the tuning parameter.

The tuning parameter puts a penalty on the parameters of the model, affecting model selection. Performance of a model is highly sensitive on the selection of tuning parameter and one should practice cautiously while dealing with penalized procedures. We compare 6 different types of tuning parameter selection and conclude that the best way of selection is based on information criterion. Both, regularization and tuning parameter selection techniques are illustrated through simulation and real data analysis. We look at 4 countries and build separate models for each of them to address the problem of heterogeneity of geographical distances. This comparative study is useful for making managerial decisions as managers will be able to optimize their resources in terms of advertisement and consumer targeting.

1.2.2 Alzheimer's Disease Neuroimaging Initiative (ADNI) Data: Bayesian Group Lasso in Logistic Regression

Alzheimer's disease (AD) is the sixth leading cause of death in the United States. It is a form of dementia in which patients suffer loss of memory where they fail to identify people or objects, have difficulty with speech and, in later stages, are unable to perform daily life

activities like getting up from the bed or brushing their teeth. Although, it is mainly a disease of old age affecting people who are 65 or older, early onset of the disease can occur in 40 or 50 year olds in upto 5 percent of cases. AD is the most common case of dementia amounting to 60 to 80 percent of all cases. It is a progressive disease where symptoms worsen over time. AD affected patients live an average of 4 to 20 years after the symptoms become noticeable. Medical scientists are yet to find a cure for Alzheimer's but it is possible to slow down the worsening of dementia and improve lifestyles of both the affected people and their caregivers. Extensive studies are being conducted to find a treatment for AD, delay its onset or curb its advancement. More information and facts about Alzheimer's disease can be found at *www.alz.org*. According to recent studies (Leifer, 2003), early detection of AD is extremely helpful as it can be treated with novel drugs to delay AD progression.

Numerous methods have been developed for the analysis of ADNI data to identify the brain subregions that are disease related. These methods usually single out a region of interest (ROI) and perform a univariate analysis based on the chosen ROI (Luo and Nichols (2003), Grimmer *et. al.* (2009)). Univariate analysis of ROI's neglect the effect of other significant ROI's. These methods aim at analyzing each hypothesized significant ROI and then looking at multiple hypothesis where careful adjustment of multiple comparisons have to be looked at. To include all the ROI's for analysis simultaneously a regression framework seems plausible such that the model itself selects the most significant regions. However, due to high number of candidate ROI's the standard regression analysis is not possible. The good thing is, there is a medical belief that only a few ROI's are informative for characterizing AD. Thus, a dimension reduction technique such as penalized regression can be developed. The ADNI MRI database has volume, area and thickness measurements of various brain regions. Since these measurements are a direct manifestation of brain region atrophies we should consider

them as a single variable with multiple levels. Thus the number of regression parameters (brain subregions with all levels) may exceed the number of patients being studied. There are number of competitive penalized regression techniques have been developed in recent years. LASSO, perhaps the most popular technique among all (Tibshirani, 1997). However the direct use of LASSO is not appropriate in presence of multiple levels of a covariate in feature selection models. We employ, instead, a group lasso technique to build a model since it places group penalty on the parameters of a variable (feature) which makes easy selection of the whole set of volumetric measurements for an ROI. We treat different measurements of the same subregion as different levels of a covariate in this regression setup. Thus, it is easy to visualize structured correlation in the matrix of covariates (subregions) establishing the motivation of using a group lasso like method. The acuteness of Alzheimer's disease makes its early detection imperative which is why classification of a subject into healthy individuals or AD patients is of immense importance. We develop logistic regression in Bayesian setup for detection of Alzheimer's disease for getting more reliable standard error estimates.

We have obtained structural magnetic resonance imaging (MRI) data from ADNI. The structural data is the volume, surface area and thickness measurements of various brain regions like entorhinal cortex, putamen etc. on the left and right hemispheres. Since, each region has a group of measurements, we know that there is an inherent correlation among these covariates. We group these variables and the selection of a region amounts to the selection of all the corresponding measurement attributes. Thus, we can model this using a group lasso setup.

We analyze baseline data where the outcome is that a subject is a healthy control or has AD. A hierarchical Bayesian group lasso model with binary outcome is a novel method. The Bayesian group lasso formulation is motivated by Xu and Ghosh (2015). We place

independent multivariate Laplace priors on the β's and use the mixing property of normal and gamma densities to get full conditionals of all parameters involved. To ensure sufficient sparsity, we have used a spike and slab prior on β where the slab part is the normal mixing density of the multivariate Laplace prior. Instead of using maximum posterior probability technique, we employ a posterior median thresholding approach to simultaneously select and estimate the parameters. Our model is chosen by the posterior median estimates. Johnstone and Silverman (2004) showed that posterior median is a thresholding estimator under fairly general conditions and their results are generalized for multivariate spike and slab by Xu and Ghosh (2015). For the binary outcome, a logistic link is used. We have shown that the posterior median thresholding estimate is consistent. Our method gives an 80% accuracy in classifying AD from healthy controls. It has an AUC (area under curve) of 0.867 for classifying true from false. The proposed approach has also been validated with simulation studies which compares the frequentist group lasso with our Bayesian version. It is seen that the false positive rate is higher for frequentist group lasso, indicating the superiority of our method.

1.2.3 Functional Magnetic Resonance Imaging (fMRI) Data: Bayesian Spatiotemporal Inference

fMRI data is extensively used nowadays to detect and understand brain activities. Understanding of brain activity helps in understanding how the brain regions are associated with a particular disease or activity. This can reveal important insights, thus, helping medical practitioners and scientists to develop cures for diseases or explain nervous aberrations. The fMRI scans reveal various degrees of brain activity when a person is exposed to some stimuli.

Brain activity is measured by detecting changes in neural activity associated with blood flow also known as blood oxygenation level dependent activation (BOLD). This is measured by contrasting deoxygenated hemoglobin (paramagnetic substance) to oxygenated hemoglobin (diamagnetic substance) in brain regions.

The brain regions where these measurements are obtained are spatial brain volumes, known as voxels. An fMRI scan is really a picture in 3-dimension. The scans provide images of brain voxels slice-by-slice and each voxel is uniquely identified by a 3-D coordinate identifier in the 3-D space. Brain activation does not occur in all the voxels but in certain regions of the brain, affecting a few voxels only. Thus, we want to identify the voxels that are activated by some external stimuli. The fMRI scans are collected over a duration of time at certain intervals. Thus, a single scan will contain voxel level information of a subject collected at T timepoints. There are more than hundred thousand voxels in the brain and these are collected for T timepoints. So, our dataset contains number of voxel $\times$ T observation for a single subject. this is a huge dataset and our challenge is to develop a model that is able to detect the truly activated regions. This is, thus, a variable selection problem where detection of the voxels amounts to selecting variables from a large number of candidate voxels.

fMRI scan produces a highly resolved brain imaging dataset. Software like Analysis of Functional NeuroImages (AFNI) can be used to view the datasets as images. It is obvious that there exists spatial correlation among the voxels and the repeated measurements at T timepoints induce a temporal correlation as well. It is a challenge to handle all of this data at once, so wee need to come up with some modeling techniques that will be able to easily model the data. Smith *et. al.* (2003), were the first to use spatial Bayesian variable selection for fMRI data. Smith and Fahrmeir (2007) developed a method that incorporated spatial dependence by the use of Ising prior. They have considered a full-brain approach as well as

slice-by-slice analysis. This approach does not take into account the temporal dependence. Musgrove, Hughes and Eberly (2016) have developed a fully spatiotemporal Bayesian model for fMRI data. Instead of dealing with all the voxels, they have introduced a divide and rule approach. In their method, they partition the brain into several smaller parts and perform variable selection of activated voxels. This method is useful but the partitioning of the brain seems arbitrary and it is not known if a different partitioning should lead to a different result.

In a more recent work, Castruccio *et. al.* (2016) have used a multi layer approach for variable selection after considering spatio-temporal correlations. Their method is useful in handling data of the size 22 million. However, they have used region of interest information in their model to build the multi layer approach. Our data consists of voxel level data only and we introduce a novel spatio-temporal Bayesian variable selection technique by extending our group lasso idea into a bi-level selection type solution. To handle data of size 28 million, we have used a two step approach. In the first step, we perform a simple analysis without considering spatial correlation and select the "significant" voxels based on p-values. We use our Bayesian bi-level selection technique on this reduced dataset.

This work is a collaboration between the Department of Statistics and Probability and the Department of Radiology. We have obtained BOLD signal data of $64\times64\times36$ voxels over 192 timepoints. The data was obtained at the Department of Radiology on a single subject. The subject was shown two stimuli, an object and a scenery, and their BOLD signal was measured using a scanner. There is a baseline trend in the measurements but the data that we use here has been rescaled to eliminate baseline trend. The stimuli onset times are convolved with the double-gamma density.

The structure of our model is such that N voxels given a stimuli is a group. We have two groups of covariates and the covariates in each group share some correlation. So selecting a

stimuli means selecting the entire group (all voxels). However, all voxels are not activated for a given stimuli; only a few exhibit activation. So, we need to select the activated voxels within the selected stimuli. Thus, we need a second stage of covariate selection where our model should select individual voxels from the group of voxels of the selected stimuli. Our model incorporates bi-level selection in group lasso following Xu and Ghosh (2015). We place a temporal correlation on the error structure of the normal distribution of responses. We do not introduce a spatial structure through the errors, but employ it later through the prior of the voxel coefficients. Similar to the spike and slab prior on groups, we use a spike and slab prior for selecting the relevant voxels. We use an adjacency matrix for spatial dependence where we have 1 for a neighbor and 0 for non-neighbors. Note that, a 2-D space will have 4 a maximum of neighbors while a 3-D space will have 26 neighbors. We use an algorithm to build the adjacency matrix where absolute value 1 of at least one coordinates' difference of two voxels mean neighbor and absolute value the coordinates difference of two voxels greater than 1 means non-neighbor. If difference of all three coordinates is zero then we are comparing a voxel with itself.

An R package called NeuroSim is used to simulate the 2-D fMRI data. To build a close resemblance with actual fMRI data, it is important that we simulate data that is synthetic fMRI data. Our simulation results give a very high accuracy rate of selection.

Our method is a very useful application of spatio-temporal data. It has practical relevance in brain imaging data where the volume of data is large. Our Bayesian model can effectively identify the truly activated brain regions from thousands of candidate voxels. This application is extremely relevant in the field of biology and neuroimaging where BOLD signals can be used to detect brain activation, thus, leading to breakthrough revelations about the association of neuronal activity and certain diseases/ disorders or human behavior.

Chapter 2

Better Tools for Strategic Global Decision Makers Gaining Consumer Insights

2.1 Introduction

Marketing strategies are mostly based on consumer psychology and behavior. Since socio-economic conditions and cultures vary significantly across international markets, use of culture based strategies to reach out to consumers seem to be an intriguing idea. By paying attention to these cultural insights, marketers can get ahead of the curve and offer messages that anticipate changing consumer attitudes rather than simply responding to the present demands of the market. Research in markets across culture can help us better understand behavioral affiliations of consumers and help marketers launch a better campaign for their products. It is obvious that cultures vary in terms of mood of the nation, language, personal values, religion, rituals, personal preferences, social behavior and infrastructural facilities like technology and transportation. These variations can explain substantially the variation in consumer choices for many categories across countries. It is thus important for firms to develop marketing strategies that use local behavioral trends rather than a global message

across cultures. Successfully targeting customers and positioning firms' products requires subtle and deep insights into customers' cultures and tastes across borders. For instance, Kumar and Pansari, (2016) examined the importance of cultural and economic aspects of a country on customer lifetime value (CLV), where the economy of a target country directly influences customer profitability, and national culture (individualism, uncertainty avoidance, masculinity, etc.) influences customer profitability through purchase frequency and contribution margin. In this vein, national culture is significantly associated with customer behaviors such as innovativeness (Steenkamp, Ter Hofstede, and Wedel, 1999), new product development activity (Nakata and Sivakumar, 1996), word of mouth (Money, Gilly and Graham, 1998), and financial decision making (Petersen, Kushwaha, and Kumar, 2015). Customer heterogeneity is a function of numerous factors.

2.1.1 Motivation

In general, international market segmentation must adapt and expand to understand customer heterogeneity and address the issue at three levels -at the country-, region-, and consumer-level. The segments defined by countries and regions are based mainly on geography; accordingly, the culture corresponds to a specific geography. Although geographic aggregation for international segmentation is widely accepted and predominates in business practice, unfortunately it overlooks consumer-level heterogeneity within a geographic segment. More importantly, the aggregated segmentation approach does not theoretically motivate international marketing strategy decisions nor is it managerially relevant (Nachum, 1994; Steenkamp and Ter Hofstede, 2002). Because of the accelerating trend toward globalization, consumers' information, knowledge, and needs are no longer limited by location, shared language, or presumed similar culture in this era of information explosion.

Lifestyle segmentation system collects information about leisure activities, topics of interests, media profiles, personal traits and values, and thus substantially enhances the accessibility and actionability of the segments via marketing mechanisms such as promotion and advertising. A shortcut for firms can be to extract the discerning characteristics of consumers in target markets from lifestyle segments, and design corresponding marketing strategies. Thus, variable selection from a great number of indicators in lifestyle surveys helps firms to identify target customers accurately and efficiently.

Despite the fact that variable selection from lifestyle segments significantly reduces the amount of market research effort required to target customers, there are a number of challenges in making precise strategic inferences and decisions. First, lifestyle segmentation surveys usually include information as extensive and granular as possible. Hence, some information is relevant and some might not be. This happens quite often because firms tend to gather information exhaustively in the early stages of market entry. Second, although a maximum input of information should be appreciated, it is almost always accompanied by overlapping and redundant information. Thus, once consumer insights have been transformed into data, high dimensionality and multicollinearity are present in further analyses.

A traditional approach is stepwise regression, with an automatic procedure to successively include variables in the model based purely on a t-value. However, as Olusegun, Dikko, and Gulumbe (2015) recognized, the drawback of stepwise regression for the selection of variables usually comes from omitting a suppression effect. Stepwise regression for variable selection may overlook a variable that is not correlated (or weakly correlated) with the dependent variable, but which is significantly correlated with other predictor variables. Consequently, the predictive power of the model is discounted; what is worse is that omitting such a variable may present the risk of rejecting a true hypothesis as false (Pandey and Elliott, 2010).

More computationally efficient, penalized likelihood approaches-Ridge, LASSO, and elastic net are used widely in variable selection procedures. Nevertheless, the selected variables and complexity of the model rely heavily on the choice of tuning, also called regularization, parameter- λ. Typically, cross-validation and information criterion are used for selecting the optimal tuning parameter. A topic of interest is to see which model selection criterion helps achieve the optimal model. In an exercise of variable selection, the goal is to optimize model selection by balancing minimization of the mean squared prediction error (MSPE) for Gaussian models, or the negative log-likelihood (NLL) for non-Gaussian models and maintaining the sparsity closest to the true model. In practice, this is usually managed by developing models with sequential values of tuning parameters and selecting the one that achieves the optimal criterion. Information criterion approach takes into account the model fit and complexity to select the optimal tuning parameter. A potential concern of using the penalized likelihood approach is that the choice of the tuning parameter can be arbitrary (Kim et al. 2012). Fan and Tang (2012) empirically demonstrated that neither the Akaike information criterion (AIC) nor the Bayesian information criterion (BIC) is adequate to identify the true model. The concept of criteria is to put a penalty on the degrees of freedom of the model. This penalty is 2 for AIC and log(n) for BIC. One may argue that in a high-dimensional setting the number of parameters grow with increasing sample size and that there is an effect of dimensionality p on the penalty of the information criterion. We use the generalized information criterion (GIC) (Fan and Tang, 2012) with penalty log(log(n)) log(p) as a third type of information criterion.

2.2 The penalized regression models

Table 2.1 summarizes the 3 penalized likelihood methods we will compare for variable selection. It is a known fact that, ordinary regression fails in case of high-dimensionality and/or multicollinearity. In market survey data, information is collected in excess which brings along the risk of redundancy of information. Penalized or regularized likelihood methods penalize the parameter space. The penalty encourages shrinkage of estimates of the coefficients. These methods maximizes the likelihood but has a constraint attached to the function to be maximized. There is no closed form solution for this approach. A popular method is the gradient descent algorithm to find the estimates. Penalized regression simultaneously selects and estimates coefficients of the regression model.

For a generalized linear model, our model setup is as follows:

$$g(\mu) = X\omega + s$$

Here, g(.) is a link function, which links the mean (μ) of the responses to a linear combination of the covariates. The penalized likelihood method is very similar to the Gaussian case. In this setup too, we are interested in maximizing the likelihood, but instead of minimizing the sum of the square of errors we minimize the negative log-likelihood. Thus, for penalized methods, we add the penalty term to the NLL, and our objective is to minimize this.

2.2.1 Tuning parameter selection

The penalized approach selects a model of reduced dimension with a few covariates. Number of questions that are selected from the market survey questionnaire depends on the tuning

Table 2.1: Penalized Regression Methods To Be Discussed

Method	**Minimize**	**Function**
Ridge Regression (Hoerl and Kennard, 1970)	$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda\sum_{j=1}^{p}\omega_j^2$	Shrinks estimates of the coefficients toward zero.
LASSO (Tibshirani, 1996)	$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda\sum_{j=1}^{p}\lvert\omega_j\rvert$	This approach forces many ω to take 0 values.
Elastic Net (Zou and Hastie, 2005)	$\sum_{i=1}^{n}(Y_i - X_i\omega)^2 + \lambda P_\alpha(\omega)$	Works in the presence of multicollinearity, unlike LASSO.

parameter that sets the penalty on the parameters. Since, the number of questions included in the final model may play an enormous impact on a market survey financially, careful selection of the tuning parameter is of great importance. The tuning parameter reduces the degrees of freedom of the model. Thus, selecting an appropriate tuning parameter affects model selection. This directly affects the predictive performance of the selected model. A tuning parameter is generally denoted as λ; we will call it λ henceforth.

A small choice of λ selects a larger model and a large choice of λ selects a smaller model. Choice of an optimal λ is important for selecting the optimal model. A popular method of λ selection is cross-validation. Here, the data is divided into approximately equal k-sized parts. $k-1$ parts are used as training sets and kth part is used as the test dataset. A model with parameter λ is fit on the training set. The error in prediction $PE_k(\lambda) = \sum_{i \in kth\ part}(y_i - x_i\hat{\omega}^{-k})$ is computed using the test set based on $\hat{\omega}^{-k}$. For a generalized linear model, we work with the value of NLL. In the case of a binary outcome,

our NLL is $NLL(\omega) = \sum_{i=1}^{n} \Big[y_i \log(1 + \exp(-x^T\omega)) + (1 - y_i) \log(1 + \exp(x^T\omega)) \Big]$. This is repeated k times using every sub-dataset as the test and the remaining as training datasets for the same lambda. The average cross-validation error is computed. This entire procedure is repeated for every candidate value of λ and the model with the smallest average cross-validation error is selected. We know that AIC and BIC are measures of goodness of fit of a model in ordinary least squares; recently information criterion are used as a method of λ selection. The information criterion penalizes the degrees of freedom in a model thus dealing with a trade-off between goodness of fit and complexity of the model. More the number of parameters in a model better is the fit but this often leads to overfitting. The penalty term for AIC is 2 and BIC is log(n). Another kind of information criterion is the generalized information criterion (GIC) which has log(log(n)) log(p) as its penalty term. We often notice that the number of parameter grow with an increasing sample size and thus it seems reasonable to introduce the parameter size in penalty. The general formula for information criterion is:

Information Criterion = -2 log-likelihood + penalty * degrees of freedom.

The log-likelihood term attributes the goodness of fit and the degrees of freedom times penalty term attributes to the complexity of the model. A model that maximizes the likelihood is preferred but the second term brings a trade-off between model complexity. We choose that model which minimizes the information criterion.

The motivation in this chapter is to find an optimal model for a market survey dataset where the number of questions(and sub-questions) outnumber the number of consumer who participated in the survey. The chosen model will minimize the prediction error. Selection

of the model depends on λ selection. We want to find a method that consistently selects a λ value for which an optimal model is achieved. In this paper, we compare 3 different information criteria and 5, 10 and 50 fold cross-validation methods to see which one of these chooses an optimal λ consistently.

Statistical model building is a commonly used technique in market research. In marketing data, when we want to build a model where there are too many variables as opposed to too few observations (e.g. data from a survey questionnaire), these variable selection methods can be very useful. A stepwise regression type approach has its limitations (Harrell, 2013). We are looking for a statistically valid variable selection methodology. Application of a penalized regression method with careful choice of tuning parameter can be a powerful method in market research model building. The above mentioned variable selection method are not just valid for high-dimensional data. As we will see in the illustration, these methods are also valid when $p < n$. When we are trying to predict consumer behavior or segmentation this can be an efficient way of dealing with the problem. An optimal model able to predict consumer behavior or segmentation by simultaneously selecting and estimating relevant variables has multi-dimensional cost saving potential. A better model would mean better designing of market strategies in terms of a cost effective strategy. A viable model does away with non-informative variables thus saving on the cost of data collection. A good market strategy impact returns from the market which shows that the initial step of building an accurate model affects the long-term credibility of marketing strategies for a firm.

2.3 Simulation

We test our proposed approach for tuning parameter selection of penalized methods by a thorough simulation study. To demonstrate its performance and validity, we simulate data that represents four scenarios - low vs high dimensionality, and with the presence/absence of multicollinearity. These four scenarios are illustrated for both a linear case and a binary response case. It is expected that the presence of non-informative variables in a model can disturb the selection of true informative variables and skew parameter estimates. This simulation exercise compares 3 different variable selection penalized approaches and 6 different tuning parameter selection methods.

We add non-informative variables in a sequentially increasing order to the set of informative variables. To illustrate, we use several different values of p for variable selection. Consider a linear model $Y_i = X_i\beta + \epsilon_i, i = 1, \ldots, N$, where N indicates the number of observations, X_i is the design matrix for each group i and ϵ_i is the Gaussian random error with zero mean and variance of σ^2_ϵ. The overall design matrix is generated from a multivariate normal distribution with zero mean and covariance matrix Σ with pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}, \rho = 0.2$. To simulate different scenarios when multicollinearity is present or absent, we generate two different feature matrices - i) an X which does not have correlated columns; and then ii) an X with a number of correlated columns. The number of parameters p is set to vary in order to illustrate if the methods are able to reliably select true factors and covariates. We examine each of the above scenarios with increasing p starting from 5 up to 1000. X is generated to have 500 rows and 1000 columns. Our parameter vector has p entries, with the first 5 entries being 1, 2, 4, 3, 3, and the rest are all 0s. The first 360 rows of X are used as the training dataset, and the remaining 140 rows are used as the test

dataset. We validate our model here since we are testing the optimality of our models using a model metric. The error variance used to simulate Y is 0.25. The same design has been used to generate Y using a logit link generalized regression structure. Here, X, β and p are the same as the previous case - the only difference being that we used these to generate a binary response using a logit link.

For the linear setup, we find the prediction error on the test dataset and for the binary response, we obtain the NLL on the test dataset based on the estimates obtained from the training dataset, using 6 different criteria (AIC, BIC, GIC, CV-5, CV-10, and CV-50) for ridge, LASSO, and elastic net, respectively. Bootstrapping estimation is used with 100 draws, with various levels of p, the number of parameters, such that both low-dimensional and high-dimensional settings are considered. Our true model size is 5 and we demonstrate which methods are able to produce a model closest to the true model.

Tables 2.2 and 2.3 exhibit the estimation results when we use an X that does not have correlated columns, namely the absence of multicollinearity. Table 2.2 displays the scenario of low dimensionality, where the number of parameters is smaller than the number of observations (p = 50). Ridge selects all the covariates and gives the highest MSPE, compared to LASSO and elastic net. This result reflects overfitting of ridge. The 3 cross-validations, CV-5 CV-10, and CV-50, perform consistently by choosing a comparatively smaller λ. LASSO and elastic net are both able to perform variable selection. BIC and GIC both resulted in selecting a reasonable model, as the average model size is very close to the true model. AIC and CVs put a small penalty on the parameters, resulting in the selection of larger models. The standard errors of the model size are higher for AIC and CVs, indicating that they fail to consistently select a model. BIC and GIC, on the other hand, have smaller standard errors of model size. However, the MSPE is slightly smaller for AIC and CVs, mainly because AIC

and CVs result in including more variables.

Thus, BIC and GIC outperform the other selection methods indicated by reasonable MSPE and close to true model size. Given that the differences of MSPE are not significant, it shows that the extra variables included by AIC and CVs do not significantly improve the predictive power of the model. Table 2.3 exhibits the scenario of high dimensionality, where the number of parameters is greater than the number of observations (p = 700). We drop ridge since it fails to select an optimal model. A similar pattern is uncovered in that, AIC and CVs select too many variables in the model. The MSPEs yielded by BIC and GIC are slightly better, as the number of selected variables is smaller. However, the MSPE resulting from BIC and GIC are quite comparable but GIC selects the model size that is closest to the true model. It further reflects the fact that BIC and GIC compromise a trade-off between an optimal model size and a lower MSPE. The AIC curve, for higher values of p, is a decreasing function of λ; a minimum point is not achieved by the AIC curve. To maintain consistency with other criteria, we have set a limit of λ values for AIC in this study. Thus, AIC is not a good option for λ selection. Similarly, LASSO, with GIC as a tuning parameter selection, outperforms in the scenario of high dimensionality that is free from multicollinearity.

Next, we apply our proposed approach to the simulated scenarios when the first three columns of X are correlated, exhibiting multicollinearity. The first and third columns are set up with a moderate correlation [0.6] and the second and third columns are set up with a high correlation [0.9]. Similarly, we repeat the simulation for $p = 5, 50, 100, 300, 500, 700,$ and 1000. We report model results when $p = 50$ and 700 in Tables 2.4 and 2.5, respectively. The active set in the tables indicates which of the true variables the model selects. In the scenario of high dimensionality, LASSO is not able to select the true featuring variables that are correlated in neither low nor high dimensional cases, failing to identify the true model. As

expected, elastic net is able to include all true featuring variables. Therefore, in the presence of multicollinearity, elastic net demonstrates a better penalized approach. The resulting MSPE is reasonably small. Comparable to the non-multicollinearity case, BIC and GIC give higher values of λ selection. In the context of high dimensionality, AIC is the weakest selection criterion, indicated by the highest MSPE and wrongly big model size. CVs exhibit smaller MSPE but, similar to AIC, select too many featuring variables. Correspondingly, BIC and GIC yield better results regarding both model selection and reasonably low MSPE. GIC is able to identify the model closest to the true model. To conclude, in the case of high dimensionality with multicollinearity, elastic net, with GIC as a tuning parameter selection, is considered the preferred penalized method.

Table 2.2: Mean squared prediction error and model size when p=50 (without multicollinearity)

		λ	Model Size	MSPE
Ridge	AIC	0.001021388 (0.0000000001)	50	0.2893 (0.03679)
	BIC	0.003562153 (0.0005126616)	50	0.2893 (0.03679)
	GIC	0.004184590 (0.0004470174)	50	0.2893 (0.03679)
	CV-5	0.507462118 (0.4643633689)	50	0.2897 (0.03682)
	CV-10	0.415566287 (0.3319358473)	50	0.2894 (0.03684)
	CV-50	0.380492323 (0.1616254741)	50	0.2894 (0.03683)
LASSO	AIC	0.02781584 (0.008342591)	16.56 (6.92)	0.26235 (0.03299)
	BIC	0.05170223 (0.008185595)	6.84 (1.47)	0.26776 (0.03358)
	GIC	0.05460872 (0.008521452)	6.38 (1.23)	0.26902 (0.03344)
	CV-5	0.03329435 (0.006880519)	14.83 (6.15)	0.26165 (0.03324)
	CV-10	0.03121886 (0.006198419)	15.54 (5.84)	0.26165 (0.03324)
	CV-50	0.03008664 (0.007338059)	15.59 (6.04)	0.26197 (0.03336)
Elastic Net	AIC	0.04052913 (0.012860055)	22.76 (7.67)	0.26663 (0.03384)
	BIC	0.09110559 (0.014255864)	8.45 (2.20)	0.27658 (0.03486)
	GIC	0.09683863 (0.013107877)	7.70 (1.74)	0.27898 (0.03481)
	CV-5	0.05252995 (0.008847460)	19.28 (5.90)	0.26555 (0.03335)
	CV-10	0.04963789 (0.009817753)	20.23 (6.32)	0.26555 (0.03354)
	CV-50	0.04773889 (0.011280188)	20.52 (6.76)	0.26586 (0.03346)

Table 2.3: Mean squared prediction error and model size when p=700 (without multicollinearity)

		λ	Model Size	MSPE
LASSO	AIC	0.02338005 (0.007871480)	138.50 (37.02)	0.31144 (0.03719)
	BIC	0.07734480 (0.006230073)	6.96 (1.66)	0.29031 (0.0347)
	GIC	0.08481446 (0.007959283)	5.65 (0.83)	0.29667 (0.03585)
	CV-5	0.05652823 (0.008080458)	26.09 (13.50)	0.28022 (0.03396)
	CV-10	0.05387672 (0.009212344)	30.16 (16.78)	0.2800 (0.034)
	CV-50	0.05205381 (0.010043022)	32.67 (18.58)	0.28058 (0.03434)
Elastic Net	AIC	0.04130926 (0.01049730)	157.06 (28.78)	0.32671 (0.03858)
	BIC	0.15116164 (0.01347818)	9.23 (2.81)	0.32075 (0.04022)
	GIC	0.16840461 (0.01598920)	6.76 (1.42)	0.33463 (0.04147)
	CV-5	0.09941351 (0.01619864)	41.63 (18.90)	0.29747 (0.03698)
	CV-10	0.09406769 (0.01637588)	47.45 (21.04)	0.29666 (0.03747)
	CV-50	0.08952375 (0.01987736)	53.99 (27.89)	0.29788 (0.03839)

Similarly, Tables 2.6-2.9 illustrate the comparison of LASSO and elastic net on binary outcomes both with and without multicollinearity. The conclusion we draw from these results is very similar to that of the linear case. Note that NLL is lower for AIC and all cross-validation cases. Variable selection is evidently better for BIC and GIC, and more importantly, GIC offers a model closest to the true model. For the binary outcome case, BIC and GIC pay the price of getting a slightly high NLL compared to other methods. Given that our main objective is variable selection, we consistently propose to use GIC over other methods as being optimal for variable selection. One thing to note in this illustration is that AIC performs better than the linear model setup.

To summarize, LASSO and elastic net are both applicable variable selection methods for the scenario of high dimensionality that perform similarly in both a linear model and a binary outcome model. When there is no threat of multicollinearity, LASSO is considered the preferred choice as the penalized likelihood method because it selects a model closest to the true model and results in low MSPE. However, high dimensionality is often accompanied

Table 2.4: Mean squared prediction error and model size when p=50 (with multicollinearity)

		λ	Model Size	Active Set	MSPE
LASSO	AIC	0.03181344 (0.009161650)	12.42 (6.12)	3.5 (0.50)	0.26497 (0.03133)
	BIC	0.05726272 (0.009371818)	4.24 (1.26)	3.44 (0.50)	0.2668 (0.03175)
	GIC	0.05964167 (0.009658534)	4.00 (1.11)	3.42 (0.50)	0.26749 (0.03193)
	CV-5	0.03815249 (0.007085186)	10.48 (4.88)	3.53 (0.50)	0.26402 (0.03121)
	CV-10	0.03643849 (0.007736441)	10.94 (5.45)	3.56 (0.50)	0.26429 (0.03138)
	CV-50	0.03527344 (0.009265170)	11.18 (6.27)	3.58 (0.50)	0.26476 (0.03133)
Elastic Net	AIC	0.05163878 (0.01170837)	16.99 (5.43)	5 (0)	0.26807 (0.03132)
	BIC	0.10042863 (0.01673992)	7.14 (1.88)	5 (0)	0.27405 (0.03308)
	GIC	0.10884811 (0.01832204)	6.38 (1.55)	5 (0)	0.27618 (0.03268)
	CV-5	0.06252737 (0.01147031)	14.81 (4.92)	5 (0)	0.2681 (0.03082)
	CV-10	0.05944682 (0.01124177)	15.60 (4.87)	5 (0)	0.26775 (0.03123)
	CV-50	0.05569424 (0.01259223)	16.16 (5.55)	5 (0)	0.2678 (0.03113)

by multicollinearity in a variety of extents. Thus, elastic net is considered more generalizable, since it is able to select all true featuring variables when there is potential multicollinearity. With regard to selecting the tuning parameter, GIC is considered a better alternative when the goal is to select informative variables precisely and efficiently. The simulation study demonstrates that in an exercise of variable selection in high dimensionality that presents multicollinearity, elastic net, together with GIC approach, should be employed. We present a variable selection exercise with a global lifestyle segmentation in the next section.

Table 2.5: Mean squared prediction error and model size when p=700 (with multicollinearity)

		λ	Model Size	Active Set	MSPE
LASSO	AIC	0.02738359 0.02738359	117.55 (44.90)	3.53 (0.50)	0.29717 (0.03471)
	BIC	0.07876276 (0.006756074)	4.61 (1.37)	3.49 (0.50)	0.27272 (0.03042)
	GIC	0.08634825 (0.008313138)	3.62 (0.68)	3.44 (0.50)	0.27588 (0.03102)
	CV-5	0.06229525 (0.010551477)	18.16 (13.65)	3.32 (0.47)	0.26981 (0.03085)
	CV-10	0.05964375 (0.009620764)	19.6 (13.66)	3.32 (0.47)	0.26912 (0.03023)
	CV-50	0.05632187 (0.010686199)	23.73 (16.49)	3.32 (0.47)	0.26958 (0.03121)
Elastic Net	AIC	0.04487323 (0.01612084)	147.18 (42.90)	5 (0)	0.31471 (0.03655)
	BIC	0.15466768 (0.01366089)	7.37 (1.84)	5 (0)	0.29291 (0.03168)
	GIC	0.17007910 (0.01428129)	5.83 (1.01)	5 (0)	0.2999 (0.03405)
	CV-5	0.11063973 (0.01643746)	28.80 (13.61)	5 (0)	0.28104 (0.03077)
	CV-10	0.10529391 (0.01720958)	33.29 (16.64)	5 (0)	0.28127 (0.03036)
	CV-50	0.10155184 (0.01918456)	36.55 (19.12)	5 (0)	0.28189 (0.03069)

Table 2.6: Negative log-likelihood and model size when p=50 (without multicollinearity) for binary response

		λ	Model Size	NLL
LASSO	AIC	0.019 (0.001)	12.77 (2.52)	31.57 (2.70)
	BIC	0.024 (0004)	7.5 (1.67)	33.53 (3.01)
	GIC	0.025 (0.004)	6.96 (1.42)	33.05 (2.98)
	CV-5	0.01 (0.002)	22.05 (4.03)	29.69 (3.68)
	CV-10	0.009 (0.001)	23.21 (4.01)	29.59 (3.87)
	CV-50	0.01 (0001)	23.27 (3.87)	29.57 (3.86)
Elastic Net	AIC	0.02 (0.003)	23 (4.16)	33.73 (2.68)
	BIC	0.043 (0.012)	9.99 (3.69)	39.59 (3.60)
	GIC	0.049 (0.012)	8.34 (2.64)	40.08 (3.46)
	CV-5	0.01 (0.002)	34.37 (4.00)	31.91 (3.72)
	CV-10	0.01 (0.001)	35.02 (3.53)	31.81 (3.80)
	CV-50	0.009 (0.001)	35.24 (3.49)	31.8 (3.79)

Table 2.7: Negative log-likelihood and model size when p=700 (without multicollinearity) for binary response

		λ	Model Size	NLL
LASSO	AIC	0.03 (0.003)	23.67 (3.12)	37.23 (2.43)
	BIC	0.044 (0.004)	6.65 (1.51)	40.24 (2.67)
	GIC	0.05 (0.006)	5.31 (0.81)	42.34 (2.93)
	CV-5	0.015 (0.005)	68.86 (16.62)	32.64 (3.62)
	CV-10	0.015 (0.005)	69.3 (17.57)	32.74 (3.67)
	CV-50	0.015 (0.005)	68.81 (19.45)	32.82 (3.67)
Elastic Net	AIC	0.07 (0.007)	21.66 (7.13)	47.17 (2.27)
	BIC	0.099 (0.010)	7.66 (2.21)	53.02 (2.39)
	GIC	0.113 (0.012)	5.55 (1.13)	55.74 (2.78)
	CV-5	0.029 (0.009)	104.48 (25.92)	41.47 (3.01)
	CV-10	0.026 (0.009)	113.19 (27.18)	41.49 (3.15)
	CV-50	0.025 (0.010)	116.21 (31.64)	41.81 (3.43)

Table 2.8: Negative log-likelihood and model size when p=50 (with multicollinearity) for binary response

		λ	Model Size	Active Set	NLL
LASSO	AIC	0.019 (0.001)	16.47 (1.83)	3.45 (0.5)	18.45 (1.53)
	BIC	0.022 (0.003)	4.88 (1.38)	3.31 (0.46)	19.44 (1.82)
	GIC	0.023 (0.003)	4.53 (1.22)	3.28 (0.45)	19.83 (1.90)
	CV-5	0.009 (0.001)	17.4 (2.72)	3.72 (0.45)	14.21 (2.12)
	CV-10	0.009 (0.001)	17.41 (2.94)	3.71 (0.46)	14.19 (2.11)
	CV-50	0.009 (0.001)	17.26 (3.01)	3.71 (0.46)	14.21 (2.14)
Elastic Net	AIC	0.02 (0.002)	20.57 (3.08)	5 (0)	21.91 (1.72)
	BIC	0.042 (0.010)	9.52 (2.83)	5 (0)	28.53 (3.16)
	GIC	0.045 (0.009)	8.65 (2.16)	5 (0)	29.48 (2.81)
	CV-5	0.009 (0.001)	32.46 (3.02)	5 (0)	17.67 (2.39)
	CV-10	0.009 (0.001)	32.46 (3.06)	5 (0)	17.67 (2.39)
	CV-50	0.009 (0.001)	32.5 (3.00)	5 (0)	17.66 (2.39)

Table 2.9: Negative log-likelihood and model size when p=700 (with multicollinearity) for binary response

		λ	Model Size	Active Set	NLL
LASSO	AIC	0.024 (0.004)	14.95 (6.21)	3.29 (0.46)	21.26 (1.99)
	BIC	0.035 (0.004)	4.55 (1.67)	3.14 (0.35)	25.04 (2.02)
	GIC	0.04 (0.004)	3.46 (0.69)	3.11 (0.31)	26.39 (1.78)
	CV-5	0.01 (0.002)	57.88 (9.60)	3.49 (0.50)	17.25 (2.57)
	CV-10	0.01 (0.003)	56.95 (10.06)	3.49 (0.50)	17.28 (2.58)
	CV-50	0.011 (0.003)	55.97 (10.09)	3.48 (0.50)	17.34 (2.57)
Elastic Net	AIC	0.062 (0.007)	17.74 (6.28)	5 (0)	34.54 (1.87)
	BIC	0.085 (0.009)	7.07 (1.91)	5 (0)	39.4 (2.03)
	GIC	0.095 (0.009)	5.57 (0.77)	5 (0)	41.39 (2.06)
	CV-5	0.017 (0.006)	117.89 (21.31)	5 (0)	26.81 (3.08)
	CV-10	0.016 (0.006)	119.7 (20.85)	5 (0)	26.63 (3.05)
	CV-50	0.015 (0.007)	121.25 (23.50)	5 (0)	26.7 (3.18)

2.4 An Empirical Variable Selection Exercise

We perform our proposed penalized likelihood approach on a variable selection exercise of global lifestyle segments. The objective is to select characterizing features that reflect consumer heterogeneity, using data collected by an international marketing research firm specializing in consumer insights. This report collects lifestyle measures of consumers worldwide. The original response variable is defined by six global lifestyle segments: fun-seekers, intimates, creatives, altruists, devouts, and strivers. Figure 2.1 shows the distributions of the six lifestyle segments across four countries, the US, Canada, Brazil, and China. The four selected countries are expected to vary in underlying constructs. In other words, individual items may share the same underlying structure but the membership of items regarding each construct may be different across the four countries. We obtain 300 observations from random draws for each country. As expected, these reveal that the distribution of lifestyle segments is similar between the US and Canada, but this distribution differs strongly from China and Brazil. In the US and Canada, many people are fun-seekers (28.7% US and 29.7% Canada), followed by intimates (18.00% US and 20.3% Canada), creatives (15.3% US and 21.0% Canada), and strivers (15.3% US and 14.7% Canada). The US (18.0%) has more devouts than Canada (9.3%) and altruists are rare in both the US (4.7%) and Canada (5.0%). Contrarily, strivers account for more than half (58.3%) of the China sample, followed by 13.3% of altruists, 7.7% of fun seekers, 7.0% of devouts and creatives, and 6.7% of intimates. The Brazil sample exhibits a more balanced distribution and has relatively the largest percentage of intimates (22.7%), followed by devouts (21.7%), altruists (19.7%), creatives (14.7%), fun-seekers (12.0%), and strivers (9.3%).

The entire survey questionnaire includes 11 sections, each of which has a wide range in

the number of questions. We eliminated the section about evaluation of popular brands and brand involvement, because the specific brands are not available across the four countries. Following this logic, we deleted a number of items that do not apply to all four countries, but retained sections that are generalizable across countries, including mood of the nation, personal values, leisure and frequent activities, influential exposure to society, opinions about environmental issues, and demographics. We ended up having 416 explanatory variables, with 300 observations for each country sample. Thus, the data setup explicitly presents high dimensionality.

2.4.1 Characterizing Lifestyle Segments

The primary interest is to identify the featuring variables for each lifestyle segments across the four countries. Thus, we first transform the original dependent variable into six binary dependent variables, representing each type of lifestyle. According to out simulation study, we employ elastic net as the penalized likelihood method and GIC as the criterion for tuning parameter selection. The overall model size of each lifestyle segment varies from 12-15 for the US, 9-15 for Canada, 5-14 for China, and 7-15 for Brazil. As expected, personal values are the strongest indicators of lifestyle segments among the 416 explanatory variables.

The selected featuring variables characterizing lifestyle segments differ across the four countries, yet there are still degrees of mutual patterns. First, across the four countries, personal values such as pleasure are important to worldwide fun-seekers, while duty is not. Other than pleasure, US fun-seekers are into looking good; Canadian fun-seekers are into adventure and sex; Chinese fun-seekers are into romance; and Brazilian fun-seekers are into leisure, having fun, and adventure. In addition to duty, Canadian fun-seekers are not into spirituality, perseverance, and protecting the family; Chinese fun-seekers are not into

wealth; and Brazilian fun-seekers are not into spirituality. Second, the featuring variables characterizing intimates are more differential. Chinese and Brazilian intimates are fairly concentrated by one personal value, where protecting the family is important to Chinese intimates while sex is important to Brazilian intimates. It seems the concept of intimates is reflected in romantic relationships in Brazil (sex) and more broadly in China (family). Except for protecting the family, stable personal relationships are important to intimates in both the US and Canada. Interestingly, romance is important to US intimates, while it is an indicator for fun-seekers in China. US intimates don't seem to care much about status or adventure, and they would not be fine financially if they stopped working. Canadian intimates don't seem to care about status, beauty, or creativity. Third, the variables selected to characterize creatives are most differential across the four countries. US creatives advocate values such as open-mindedness, authenticity, creativity, self-reliance, and curiosity, whereas they consider wealth, pleasure, sex, tradition, and respecting ancestors not important. Except for open-mindedness and curiosity, Canadian creatives consider learning very important and status and sex not important. In addition to personal values, Canadian creatives are likely to own a stereo. Chinese creatives do not care about wealth and are more likely to purchase travel insurance. Brazilian creatives are into creativity and freedom. Fourth, US altruists appreciate perseverance and helpfulness, and are not likely to own a scanner. Canadian altruists value duty, are relatively older, and don't do fun things on a daily basis. Chinese altruists seem to be strongly affected by social indicators such as social stability, social responsibility, preserving the environment, and are more likely to make a sizable donation to a local or national organization. Brazilian altruists are characterized by the biggest number of personal values, including being in tune with nature, preserving the environment, perseverance, justice, social responsibility, helpfulness, and equality. Fifth, US devouts value

obedience, respecting ancestors, traditional gender roles, and faith, and they are interested in topics about personal finance. Similarly, Canada's devouts value tradition, obedience, and traditional gender roles. Chinese devouts value tradition and modesty, but not power, and are very interested in gardening. Brazilian devouts value spirituality and faith but do not work just to earn a living. Finally, power, wealth, and status are strong values for strivers worldwide, except that power does not seem important to Brazilian strivers. Besides, US strivers spend more time with their spouses. Canadian strivers seem to be very self-centered, since they do not care about honesty, freedom, protecting the family, equality, or friendship. However, Canadian strivers are very interested in social issues. Chinese strivers seem to care less about stable personal relationship, friendship, and respective ancestors. Tables 2.10-2.13 summarizes the selected variable and parameter estimates.

Figure 2.1: Distributions and descriptions of the six psychographic segments

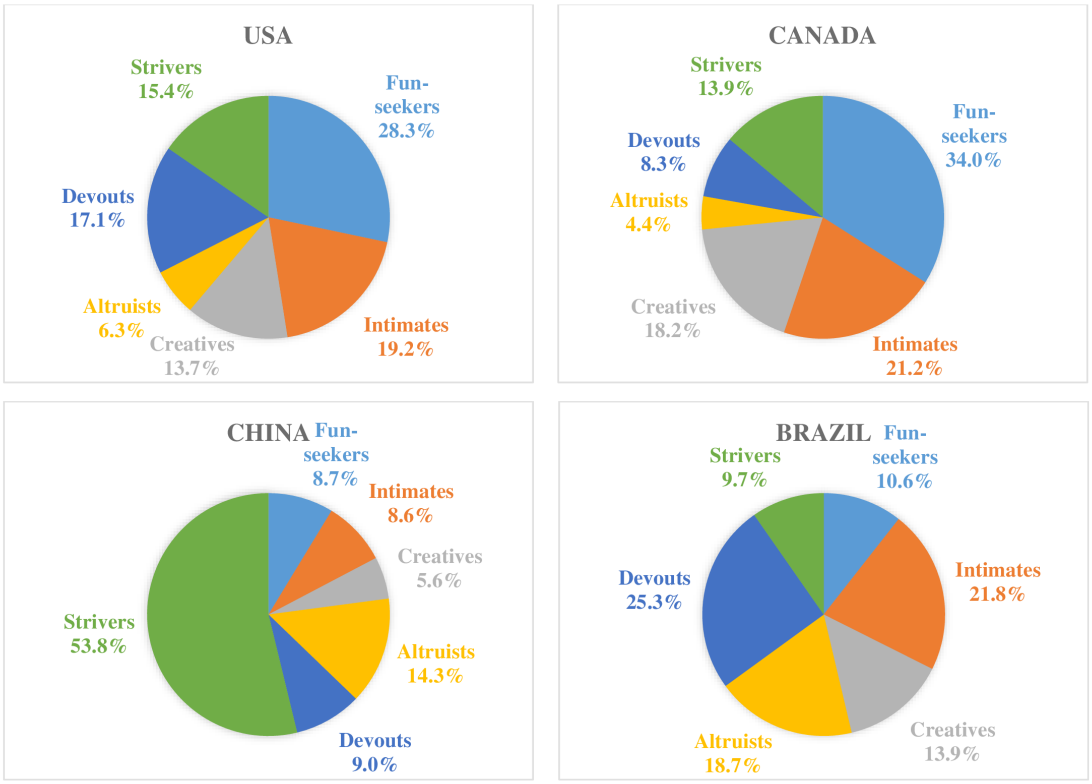


Table 2.10: Results of Variable Selection - USA

Funseekers	**Intimates**	**Creatives**	**Altruists**	**Devouts**	**Strivers**
Excitement	Adventure	Wealth	Topics very interested in: Electronic games	Topics very interested in: Personal Finance	Weekly hrs-With spouse
Spirituality	If stopped working wld be fine financially	Status	Financial services/products own/use: Hospital cash plans	Topics very interested in: Wellness, Fitness, Exercise	Environment-Things would like cos. to do: Invest in research of technologies to help preserve the envi
Honesty	Power	Open-mindedness	Financial services/products own/use: Home equity loans	Excitement	Tech Items owned: Video camera/camcorder
Authenticity	Honesty	Authenticity	Attitude toward new technology: It scares me	Friendship	Tech Items owned: Digital Home Projector
Pleasure	Status	Self- esteem	Tech Items owned: Portable Music Player (such as Walkman, Discman)	Having fun	Activities on Internet past 30 days: Watch video, broadcasts and events
Enjoying Life	Open-Mindedness	Creativity	Own inkjet/laser printer or scanner: Scanner	Sex	Activities on Internet past 30 days: Download video files
Having fun	Beauty	Self-reliance	Respondent main income earner	Tradition	Power
Adventure	Leisure	Curiosity	Leisure	Obedience	Wealth
Sex	Protecting the family	Knowledge	Spirituality	Respecting ancestors	Status
Looking good	Looking good	Public image	Preserving the environment	Traditional gender roles	Honesty
Duty	Stable personal relationships	Pleasure	Perseverance	Faith	Enduring love
Faith	Romance	Sex	Personal support	Modesty	Friendship
Excitement	Modesty	Tradition	Helpfulness		Enjoying life
Spirituality		Respecting ancestors	Equality		
Honesty		Wealth	Having fun		
Authenticity			Topics very interested in: Electronic games		
Pleasure					
Enjoying Life					

Table 2.11: Results of Variable Selection - Canada

Funseekers	**Intimates**	**Creatives**	**Altruists**	**Devouts**	**Strivers**
Have fun now - future care of itself	Three things most concerned about: Terrorism	Topics very interested in: Arts and culture	Weekly hrs-Socializing	Topics very interested in: Religion	Topics very interested in: Automobiles/Driving
Types of TV programs watch: Music video and popular music	Weekly hrs-Gardening or yard work	Environment-How feel re cos. that make effort: I don't know which companies address the environment or how	Weekly hrs-Doing fun things	Tech Items owned: Cellular/Mobile phone	Topics very interested in: Social Issues
Spirituality	Items recommended to other people: Parenting/family issues	Tech Items owned: Stereo (hi-fi)	Environment-Things would like cos. to do: Make and sell products that do not harm the environment	Tech Items owned: Portable Music Player (such as Walkman, Discman)	Most import.: brand offers good value for money
Honesty	Status	Wealth	Have fun now - future care of itself	Spirituality	Tech Items owned: Film camera
Authenticity	Beauty	Status	Age of Respondent	Traditional gender roles	Power
Perseverance	Creativity	Open-mindedness	Are you a parent?	Duty	Wealth
Protecting the family	Protecting the family	Authenticity	Excitement	Obedience	Status
Pleasure	Social responsibility	Perseverance	Preserving the environment	Traditional gender roles	Honesty
Enjoying life	Stable personal relationships	Self-reliance	Social tolerance	Patriotism	Freedom
Having fun	Enduring love	Curiosity	Adventure		Protecting the family
Live for today	Adventure	Knowledge	Duty		Equality
Adventure	Duty	Wisdom	Respecting ancestors		Romance
Sex	Obedience	Sex			Friendship
Looking good		Looking good			Enjoying life
Duty		Learning			

Table 2.12: Results of Variable Selection - China

Funseekers	Intimates	Creatives	Altruists	Devouts	Strivers
Wealth	Are you a parent?	Activ. Freq.-attend lectures-not reg study	Weekly hrs-Learning new things	Three things most concerned about: Environmental pollution	Activ. Freq.-exercise to keep fit
Protecting the family	Power	Topics very interested in: Cultures around the world	Political/societal activities in past year: Made a sizable donation to a local or national organization	Topics very interested in: Gardening	Environment-Things would like cos. to do: Support recycling or clean-up of polluted sites
Romance	Creativity	Financial services/products own/use: Life insurance or endowment products that provide a lump sum	Percentage of pre-tax monthly income save	Power	Financial services/products own/use: Credit card(s)
Duty	Protecting the family	Financial services/products own/use: Travel insurance	I spend a lot of time researching brands	Wealth	Types of TV programs watch: Programs for children
Learning	Stable personal relationships	Financial services/products own/use: Auto/car loan	Wealth	Tradition	Power
	Enduring love	Financial services/products own/use: Home equity loans	Status	Modesty	Wealth
	Pleasure	Financial services/products own/use: Overdraft facility	Honesty		Status
	Enjoying life	Financial services/products own/use: Mutual Funds or other investment product	Preserving the environment		Ambition
		Financial services/products own/use: Financial planning/advisors	Protecting the family		Honesty
		M30-Activities on Internet past 30 days: Get information related to my hobbies/interests	Social responsibility		Justice
		Wealth	Social stability		Stable personal relationships
		Material security			Friendship
		Social stability			Respecting ancestors
		Learning			

Table 2.13: Results of Variable Selection - Brazil

Funseekers	Intimates	Creatives	Altruists	Devouts	Strivers
Weekly hrs-Socializing	-Activ. Freq.-sit-down meal in a restaurant	Topics very interested in: Religion	Weekly hrs-g-Socializing	Three things most concerned about: AIDS	Three things most concerned about: Money enough to live right and pay the bills
Activ. Freq.-take pictures or photos	Ambition	Business overnight trips in past 12 mos	Health and fitness	Topics very interested in: Environmental issues	Political/societal activities in past year: Attended a political rally, speech or event
Activ. Freq.-play indoor or parlor games	Honesty	. Need help making financial decisions	Being in tune with nature	Items recommended to other people: Music	Do not like to be in debt
Topics very interested in: Science	Authenticity	Types of TV programs watch: Learning/Educational programs for adults/children	Preserving the environment	Items recommended to other people: Restaurants or places to eat	Buy whatever I want-Importance
Topics very interested in: Music	Being in tune with nature	Technology effect on relationships: Strengthens my bonds with people I care about	Perseverance	Work just to earn living	Wealth
Excitement	Creativity	Activities on Internet past 30 days: Get access to news or other up-to-the-minute information	Justice	Not person that takes risks	Status
Leisure	Knowledge	Creativity	Social responsibility	Health and fitness	Honesty
Spirituality	Wisdom	Self-reliance	Helpfulness	Spirituality	
Authenticity	Protecting the family	Freedom	Equality	Perseverance	
Protecting the family	Romance	Knowledge	Social tolerance	Self-reliance	
Justice	Friendship	Personal support	Friendship	Friendship	
Pleasure	Sex	Traditional gender roles	Pleasure	Having fun	
Having fun	Faith		Looking good	Tradition	
Adventure	Simplicity		Duty	Duty	
Duty				Faith	

2.5 Discussion

2.5.1 Theoretical Implications

The primary aim of this study is to propose an approach for variable selections in settings of high dimensionality with grouping variables. Despite the increasing phenomenon of big data, excessive consumer information becomes challenging in extracting effective and efficient consumer insights in order to properly implement market segmentation, target customers, and position advantageously. Traditional approaches of variable selection in high dimensionality may yield unstable coefficient estimation and inflated standard errors (Drolet and Morrison 2001). The family of penalized likelihood methods is advocated, but the choice of tuning parameter is fairly sensitive, since it controls the trade-off between bias and variance in the resulting parameter estimates (Hastie et al., 2009; Fan and Lv, 2010). Certain specifications of the optimal tuning parameter are challenging to quantify in practice, because the resulting tuning parameters are valid only asymptotically, and rely heavily on unknown nuisance parameters in the true model (Fan and Tang, 2013). The challenges are even amplified for variable selection in a setting of high dimensionality and multicollinearity (strong grouping effects). The proposed generalized information criterion (GIC) enables penalized likelihood methods to be flexible in choosing desired models, and, more notably, it addresses both consistent and efficient estimation. The simulation study demonstrates that our proposed approach is able to offer the best penalized likelihood approaches and tuning parameter selection in high-dimensional settings. Elastic net is considered a cutting-edge procedure in market research studies dealing with excessive and overlapping information. With the assistance of GIC, our proposed procedure is able to identify the true model in the presence of multicollinearity.

From a theoretical perspective, this study enables international market segmentation with a procedure for selecting insightful factors and/or covariates from an enormous amount of information. The study should also be able to assist uninvestigated empirical exploratory studies, which ask for simultaneously testing the relevance and importance of a large set of potential covariates and/or factors. Second, the proposed technique is flexible with multicollinearity and construct non-equivalence. In other words, our approach is able to address heterogeneity both in consumer and latent structures. Although we practice a variable selection exercise by sub-sampling among the four countries, this approach is also applicable in testing interactions for contingent and contrasting effects.

2.5.2 Managerial Implications

One of the most conspicuous applications of the proposed variable selection approach is to gain consumer insights into the presence of excessive and overlapping information, which inevitably occurs to a firm that intends to expand to a new business territory. Thanks to advanced technology, firms now are able to access extensive information and store a large repository of data. Although rich information enables managers to gain sufficient consumer insights, too many variables may camouflage the underlying structure and result in misleading interpretations for managerial implications. Consequently, inaccurate variable selection harms resource allocation and ultimately firms' profitability in the long run. This concern is particularly severe when the contexts are dynamic and diversified, e.g., among different countries as in our empirical illustration, where the variables and their associated impacts might be elementally different across contexts.

An example of contextual differentiation emerges from our findings. If a firm intends to promote a romantic product, say a high-end perfume, the communication strategy should be

quite different between the US and China markets. In the US market, the concept of romance is acknowledged more by intimates, who appreciate family and stable personal relationships. Young couples should be considered the target customers. Since they may be in a financially precarious position, a zero-interest monthly payment might be an attractive marketing tool. However, the concept of romance is more likely to be acknowledged by fun-seekers in China. Catchy marketing strategies such as tailoring scents, where consumers can select fragrance elements and combine them to customize their "own scents" would be uniquely appealing to the China market. Another interesting finding is that the perception of sex is recognized differently between Canada and Brazil. Sex is identified as the strongest value to indicate intimates in Brazil but is considered one of important values of fun-seekers in Canada. The differential perceptions of sex or being sexy motivates different messages that a firm wants to convey. Canadian fun-seekers appreciate the value of sex in a way that emphasizes adventure, while Brazilian customers may blend the concept of sex with the context of intimate relationships. The famous Victoria's Secret show is regarded as more compelling in the Canadian market than in the Brazilian market, because it is exciting and intriguing. It might be more appealing for customers in Brazil to deliver a mingled image of being sexy and intimate. Global lifestyle segments seem to be generalizable and comparable across countries, yet collating holistic lifestyle indicators may not be the most efficient and profitable way for marketing communications. Effective variable selection techniques help firms use the best characteristics to identify customers. Broadly, variable selection in international market entry almost always presents high dimensionality and multicollinearity. There is a compelling need for managers to employ tools that retain relevant and indicative information. This study provides an implementation-friendly model for selecting insightful global consumer-level information.

Chapter 3

A Bayesian Group Lasso Classification For ADNI Volumetrics Data

3.1 Introduction

Dedicated research is conducted with neuroimaging techniques for early diagnosis of Alzheimer's Disease (AD). Alzheimer's Disease Neuroimaging Initiative (ADNI) conducts multi-center case-control study of elderly people that was designed to find more sensitive and accurate methods to diagnose AD at earlier stages. AD is an older age disease in which patients develop deformities in brain structure. It is identified by loss of memory, speech inconsistensies or inability to perform daily tasks of survival. ADNI studies use brain-imaging techniques, such as positron emission tomography (PET) and magnetic resonance imaging (MRI). ADNI database has data from three phases (ADNI1, ADNI GO and ADNI 2).

Historically, studies have shown that AD causes abnormal change to brain region volumes which causes shrinkage in the hippocampal volume or reduction in its thickness or enlargement of internal ventricles. Smith *et. al.* (2012) studied structural brain alterations before mild cognitive impairment (MCI). They had previously demonstrated that volume loss in bilateral anteromedial temporal lobe is present at baseline in longitudinally followed normal subjects who later developed MCI or AD. Arlt *et. al.* (2013) believed that fully au-

tomated MRI-based volumetric measurements may serve as a biomarker for the diagnosis in patients with MCI or dementia. They concluded that fully automated MRI-based volumetry allows detection of regional grey matter volume loss that correlates with neuropsychological performance in patients with amnestic MCI or mild AD. Our objective in this chapter is to predict dementia in patients based on the volume measurements obtained from the MRI ADNI data. There is evidence of brain atrophy with increasing age but the atrophies differ significantly from normal aging to AD patients. We use the differences of brain region atrophies to distinguish subjects with or without AD. The volumetric data has brain parcellated subregions of the entire brain for the left and right hemispheres. Volume, area and thickness measurements of brain subregion is a simple way of detecting atrophied brain regions, thus the motivation of combined use of these measurements. It is believed that all these regions are not associated with dementia but only a few. Identification of a few brain regions from the large number of regions makes appropriately a dimension reduction problem.

In this article, we develop Bayesian group lasso type technique with spike and slab prior following Xu and Ghosh (2015) over other types of penalized regression because this approach presents many natural advantages. The biggest advantage is the Bayesian approach provides reliable estimates of uncertainty which can be used for statistical inference beyond feature selection. A thorough literature review has shown that the Bayesian group lasso with logistic regression model is largely overlooked. This article develops this novel method motivated by the ADNI data. Bayesian group lasso with spike and slab prior deals with feature selection (dimension reduction) in a binary outcome scenario and produces reliable estimates for regression coefficients. Unlike commonly used Bayesian variable selection methods, we propose median thresholding to make insignificant coefficients are exactly zero. Another major contribution of this paper is that we look at the brain image volumetric data at a minute

level, considering all available brain subregions mapped by FreeSurfer to include effects of all ROI's rather than looking at individual ROI's and identifying atrophied brain subregions by selecting a group of volumetric measurements of the corresponding selected ROI's. Zhang *et. al.* (2011) performed classification with MRI data based on 93 manually labeled ROI's. We use data of 116 automatically labeled ROI's (each having 4 different measurements) by FreeSurfer and analyze them together to perform a dimension reduction analysis. Zhang *et. al.* (2011) used a composite of 3 different modalities of biomarkers. Unlike Zhang *et. al.* (2011), our method provides reliable parameter estimates which can be used to calculate the log of odds or relative risk of AD based on the selected subregions instead of just classifying subjects. Group lasso encourages selection of all levels of a significant subregion and spike and slab prior on the parameter coefficients ensure that a large number of subregions which have no impact on the disease are dropped from the model. So, the proposed method selects affected brain subregions automatically from a large pool of brain subregions. Furthermore, among the selected subregions only a few subregions serve as discriminative features in the model assessed by their statistical significance. So, we are able to narrow down the regions that should be studied by scientists to stop progress of the disease or improve the quality of life of the affected individuals. Finally, we provided theoretical foundation to our proposed methodology.

This chapter is organized into 7 sections. In the next section we have reviewed the literature of group lasso and Bayesian group lasso and then in the following section we have elaborated on Bayesian group lasso in logistic regression setup. Section 3.4 shows the posterior consistency of our estimator, i.e. the model selected by the proposed method converges to the true model for sufficiently large n. In Section 3.5 we have conducted a simulation study to test the performance of the proposed method. Section 3.6 contains the

analysis on the ADNI dataset where we detail out our findings and our concluding remarks are in Section 3.7.

3.2 Group lasso

Oftentimes, we are interested in selecting a group of variables that are predictive in a model. Popular high-dimensional penalized regression techniques used for variable selection are built to select a group of related variables. For example, we may have a group of dummy variables that represent a significant categorical variable. To select this variable, it is necessary that we select all the corresponding dummy variables. LASSO fails to correctly perform this selection since it can never select all the dummy variables of a categorical variable. Similar scenarios are when we have a number of basis functions representing a function or an ANOVA model with more than one level feature variable. Group lasso, proposed by Yuan and Lin (2006), addresses this issue. It puts a hybrid l_1 and l_2- penalty on the parameter space, thus, encouraging selection or dropping of variables in groups. Note that, this is feasible since a multi-level variable's parameter is represented as an m_g- tuple vector, m_g being the number of levels for group g.

Our regression model is as follows:

$$\mathbf{Y}_{n\times 1} = \sum_{g=1}^{G} \mathbf{X}_g \beta_g + \epsilon \tag{3.1}$$

where for $g = 1, 2 \ldots G$.

$$\epsilon_{n\times 1} \sim \mathbf{N}_n(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I_n})$$

$$\beta_g^{m_g \times 1} \text{vector of coefficients}$$

$$\mathbf{X}_g^{n\times m_g} \text{design matrix}$$

If p is the total number of predictors, then $p = \sum_{g=1}^{G} m_g$. We center both the response and input variables to eliminate intercept from 1.2.

Our primary goal is to select important predictive features. The group lasso estimate, for linear regression, minimizes

$$||Y - \sum_{g=1}^{G} X_g \beta_g||_2^2 + \lambda \sum_{g=1}^{G} ||\beta_g||_2 \tag{3.2}$$

where λ is the tuning parameter.

(3.2) is an evident extension of the LASSO penalty:

$$\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.3}$$

where λ is the tuning parameter. Note that (3.3) is a special case of (3.2).

3.2.1 Bayesian Group lasso

A major issue with lasso-type estimates is that it is difficult to give satisfactory standard errors since the limit distribution of the lasso estimator is very complicated (Knight and Fu, 2000; Chatterjee and Lahiri, 2011) but the Bayesian version of lasso overcomes this by

producing reliable standard errors. Tibshirani showed that the lasso estimator for linear regression is equivalent to the posterior mode with independent Laplace priors on each regression coefficient.

Park and Casella (2008) developed a fully hierarchical Bayesian model using a scale mixture prior of normal distributions for lasso. This idea was further extended by Kyung *et.al.* (2010) to develop a general Bayesian formulation for a number of lasso variations, including the group lasso, the elastic net (Zou and Hastie, 2005) and the fused lasso (Tibshirani *et.al.*, 2005). Raman *et.al.* (2009) developed a fully Bayesian formulation of the group lasso to tackle the problem of poor variance estimates of regression coefficients. From a probabilistic perspective, the group lasso with Gaussian likelihood can be seen as a linear regression model with normal errors and a product of multivariate Laplace priors over the regression coefficients. Thus,

$$\begin{aligned} p(y|X,\beta,\sigma^2) &\propto \exp\Big\{ -||y-X\beta||^2/(2\sigma^2)\Big\} \\ &\propto (\sigma^2)^{-p/2}\exp\Big\{ -\frac{1}{2\sigma^2}(\beta-\hat{\beta})^t X^t X(\beta-\hat{\beta})\Big\} \\ &.(\sigma^2)^{-\nu/2}\exp\Big\{ -\frac{SSE}{2\sigma^2}\Big\} \end{aligned} \tag{3.4}$$

where $\hat{\beta}$ is the least squares solution, $SSE = (y-X\hat{\beta})^t(y-X\hat{\beta})$ is the sum of squared errors and $\nu = n-p$. The last equation results from "completing the squares" which is standard in Bayesian formulations. Assuming a generalized multivariate m_g- dimensional Laplacian prior over each group of regression coefficients

$$\pi(\beta_g) \propto \exp\Big\{ -\frac{\lambda}{\sigma}||\beta_g||_2\Big\}, \tag{3.5}$$

the classical group lasso is recovered as the MAP solution in log-space with $\frac{\lambda}{\sigma}$ having the role of a fixed Lagrangian mutiplier. For a full Bayesian treatment, however, we place hyperpriors on λ and σ which lead to integrations that are analytically impossible to solve. Following the hierarchical scale mixture of lasso, we extend it to group lasso regression models. Lasso employs scale mixture by using normal and Gamma densities; here we form a mixture prior with multivariate normals and Gamma hyperprior.

$$\begin{aligned} \beta_g|{\tau_g}^2, \sigma^2 &\sim^{ind} \mathbf{N}_{mg}\left(\mathbf{0}, {\tau_g}^2\sigma^2\mathbf{I}_{mg}\right), \\ {\tau_g}^2 &\sim^{ind} \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right) \end{aligned} \tag{3.6}$$

the marginal distribution of β_g is of the form 3.5. This Bayesian formulation encourages shrinkage at the group laevel and provides comparable prediction performance with frequentist group lasso. However, this approach based on estimating β_g by its posterior mean or median does not produce exact 0 estimates. Thus, to introduce sparsity at group level, we assume a multivariate zero inflated mixture prior or a spike and slab prior for each β_g.

Xu and Ghosh (2015), further showed the superiority of posterior median thresholding. The spike and slab type zero inflated prior keeps the scale mixture prior of normals and gamma so we get full conditionals making derivation of posterior distributions easier. This approach gives exact zero estimates and is easier to compute. In recent years, many studies have been conducted exploring the application of zero inflated mixture priors (see Yuan and Lin (2012), Lykou and Ntzoufras (2013), Zhang *et.al.* (2014)). Heavy tailed distributions, such as double exponential, are often used as the slab part. The slab part can be further segmented to a scale mixture of normal and gamma distributions as is done by Xu and Ghosh (2015). The following hierarchical Bayesian formulation with spike and slab prior for linear

regression (3.1) comparable to a group lasso type estimator is proposed by Xu and Ghosh (2015).

$$
\begin{aligned}
\mathbf{Y}|\mathbf{X},\beta,\sigma^2 &\sim \mathbf{N}_n\left(\mathbf{X}\beta,\sigma^2\mathbf{I}_n\right),\\
\beta_g|\sigma^2,\tau_g^2 &\sim^{\text{ind}} (1-\pi_0)\mathbf{N}_{mg}\left(\mathbf{0},{\tau_g}^2\sigma^2\mathbf{I}_{mg}\right)+\pi_0\delta_0(\beta_g), \quad g=1,2,\ldots,G,\\
{\tau_g}^2 &\sim^{ind} \text{Gamma}\left(\frac{m_g+1}{2},\frac{\lambda^2}{2}\right), \quad g=1,2,\ldots,G,\\
\sigma^2 &\sim \text{Inverse Gamma}\,(\alpha,\gamma), \quad \sigma^2>0,\\
\pi_0 &\sim \text{Beta}\,(a,b).\\
\lambda^{(k)} &= \sqrt{\frac{p+G}{\sum_{g=1}^{G}\mathbf{E}_{\lambda^{(k-1)}}[\tau_g^2|\mathbf{Y}]}}
\end{aligned}
$$

where $\delta_0(\beta_g)$ denotes a point mass as $\mathbf{0}\in\mathbb{R}^{mg}, \beta_g=(\beta_g 1\ldots\beta_g m_g)^T$. The posterior expectation of τ_g^2 will be replaced by the sample average of τ_g^2 generated in the Gibbs sampler based on $\lambda^{(k-1)}$. The value of λ should be carefully tuned. A large value of λ will overshrink the estimates while a small value will lead to overfitting. Xu and Ghosh (2015) suggested a conjugate Gamma prior can be placed on λ^2. Using an empirical Bayes approach, λ is estimated from data using marginal maximum likelihood. Since marginal maximum likelihood of λ does not have a closed form, a Monte Carlo EM algorithm (Casella, 2001; Park and Casella, 2008) can be used to estimate λ. The kth EM update for λ is given in the above setup.

3.3 Bayesian Group Lasso with Logistic Regression

Meier *et.al.* (2008) developed the logistic group lasso in a frequentist setup. Most practical situation have binary outcomes. In risk analysis we want to know if a person will default on a loan or not, in detection of diseases we want to classify subjects with or without a disease, etc. Thus, here we need to use a generalized linear model (GLM) to model our data. The use of GLM is not limited to binary data. There are numerous cases where the outcome is not binary but we need to use a GLM, such as, modelling income or count data. Since occurrence of binary outcome is very common in real life, we focus this chapter on logistic regression with a logit link.

The concept here is to maximize the likelihood function (objective function) subject to a group lasso constraint on the parameters. Refer to Section 1.1.6 for a detailed overview of frequentist group lasso in logistic regression.

Although group lasso for linear regression has been explored, a thorough literature review reveals that no work has been done on group lasso on logistic regression. Motivated by Xu and Ghosh's (2015) work, here we construct a Bayesian formulation for the logistic regression case. Our likelihood is Bernoulli probability mass function with a logit link. We abide by using a multivariate zero inflated mixture prior with point mass at zero and the continuous part as double exponential distribution. Since a double exponential prior on β_g can be

expressed as a scale mixture of normal and Gamma priors (as in (3.6)), we have.

$$
\begin{aligned}
y_i|x_i,\beta &\sim \text{Bernoulli}\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right), i=1,\ldots,n,\\
\beta_g|\tau_g^2 &\sim^{\text{ind}} (1-\pi_0)\mathbf{N}_{mg}\left(\mathbf{0},{\tau_g}^2\mathbf{I}_{mg}\right)+\pi_0\delta_0(\beta_g), \quad g=1,2,\ldots,G,\\
{\tau_g}^2 &\sim^{ind} \text{Gamma}\left(\frac{m_g+1}{2},\frac{\lambda^2}{2}\right), \quad g=1,2,\ldots,G,\\
\pi_0 &\sim \text{Beta}\,(a,b).
\end{aligned}
$$

The full posterior conditional distributions are as follows:

$$
\begin{aligned}
p(\beta,\tau^2,\pi_0|\mathbf{Y},\mathbf{X}) \propto & \prod_{i=1}^{n}\left[\left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-y_i}\right]\\
&\times\prod_{g=1}^{G}\left[(1-\pi_0)(2\pi\tau_g^2)^{-\frac{m_g}{2}}e^{-\frac{\beta_g^T\beta_g}{2\tau_g^2}}I_{[\beta_g\neq 0]}+\pi_0\delta_0(\beta_g)\right]\\
&\times\prod_{g=1}^{G}(\lambda^2)^{\frac{m_g+1}{2}}(\tau_g^2)^{\frac{m_g+1}{2}-1}e^{-\frac{\lambda^2\tau_g^2}{2}}\\
&\times\ {\pi_0}^{a-1}(1-\pi_0)^{b-1} \qquad (3.7)
\end{aligned}
$$

We can simulate an efficient block Gibbs sampler to simulate from the posterior distribution above. Details of the block Gibbs sampler is given in the appendix.

3.4 Posterior Consistency

Xu and Ghosh (2015) showed that the posterior median is an adaptive thresholding estimator for a linear regression setup. Theorem 1 in their paper gives a proof of this idea. We will extend this idea for logistic regression model numerically.

To prepare the ground for posterior consistency, we will rewrite our model using different notations just so it is alignment with the model setup of Jiang (2007). Our proof of consistency is in line with Jiang's (2007) paper so similar notations will ease understanding of the proof.

Let $D^n = \{y; X_1, \ldots, X_{P_n} : y \in \{0,1\}, X_i \in \mathcal{R}^n, i = 1, 2, \ldots, P_n\}$ denote a dataset of n observations each consisting of P_n predictors where P_n can increase with increasing n. We want to model this data using logistic regression. Let ξ_n denote a chosen (subset) model, and $|\xi_n|$ denote the model size of ξ_n. Note that, here ξ_n is the sum of all dummy variables(factor levels) of the groups that are chosen in the subset model. Let us call G^*, the number of selected groups then $G^* \leq \xi_n$. A major difference of this and Jiang's setup is multivariate $\beta_g, g = 1 \ldots G$. An interesting thing to note is that, if we express our setup in terms of the dummy variables then this layout is similar to what Jiang proposed. Thus, when the chosen model is ξ_n, we are really considering our chosen group size to be G^* and the model size as $\sum_{g=1}^{G^*} m_g = |\xi_n|$. To make the proof here in line with Jiang's (2007) paper, we express the chosen model in terms of the dummy variables rather than the groups. Clearly, this is an extension of Jiang's model since we consider a grouped structure for β's. Conditional on ξ_n, the regression coefficients

$$\beta_{\xi_n} | \tau_{\xi_n} \sim N(0, V_{\xi_n})$$

where V_{ξ_n} is a $|\xi_n| \times |\xi_n|$ covariance matrix and a function of τ_{ξ_n}. Here, $\beta_{\xi_n} = (\beta_1^{*T}, \ldots, \beta_{G^*}^{*T})$ and $\tau_{\xi_n} = (\tau_1^*, \ldots, \tau_{G^*}^*)$ denote the vector of true regression coefficients and true variance parameters respectively such that $\sum_{g=1}^{G^*} m_g = |\xi_n|$. Let $\{X_1^*, \ldots, X_{|\xi_n|}^*\} \subset \{X_1, \ldots, X_{P_n}\}$ denote the predictors chosen in model ξ_n.

Note that, $V_{\xi_n} = \begin{pmatrix} \tau_1^{*2} I_{m_1} & 0 & \cdots & 0 \\ 0 & \tau_2^{*2} I_{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_{G^*}^{*2} I_{m_{G^*}} \end{pmatrix}$

and ${\tau_g}^2 \sim^{ind} \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), \quad g = 1, 2, \ldots, G$. Let model ξ_n have the prior

$$\Pi(\xi_n|\pi_0) = \pi_0^{P_n - |\xi_n|}(1-\pi_0)^{|\xi_n|}$$

and

$$\pi_0 \sim \text{Beta}(a, b)$$

where a and b are pre- specified hyperparameters.Thus,

$$\Pi(\xi_n) = \frac{\text{Beta } (a + P_n - |\xi_n|, |\xi_n| + b)}{\text{Beta } (a, b)}$$

$$\begin{aligned} \Pi(\beta_{\xi_n}) &= \int \Pi(\beta_{\xi_n}, \tau^2\xi_n)\ d\tau^2\xi_n \\ &= \prod_{g=1}^{G^*} \int_0^\infty \Pi(\beta_{\xi_n}|\tau^2\xi_n)\Pi(\tau^2\xi_n)\ d\tau_g^2 \\ &= \prod_{g=1}^{G^*} \int_0^\infty \frac{e^{-\frac{\beta_g^T \beta_g}{2\tau_g^2}}}{(2\pi\tau_g^2)^{\frac{m_g}{2}}} (\lambda^2)^{\frac{m_g+1}{2}} (\tau_g^2)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2\tau_g^2}{2}}\ d\tau_g^2 \end{aligned}$$

Substituting $\alpha_g^2 = \frac{1}{\tau_g^2}$ we have,

$$\Pi(\beta_{\xi_n}) = \prod_{g=1}^{G^*} (\lambda^2)^{\frac{m_g}{2}} \frac{e^{-\lambda\sqrt{\beta_g^T\beta_g}}}{(2\pi)^{\frac{m_g-1}{2}}} \int_0^\infty \left(\frac{\lambda^2}{2\pi(\alpha_g^2)^3}\right)^{\frac{1}{2}} e^{-\frac{\lambda^2\beta_g^T\beta_g}{2\lambda^2\alpha_g^2}\left(\alpha_g^2 - \frac{\lambda}{\sqrt{\beta_g^T\beta_g}}\right)^2} d\alpha_g^2$$

The term in the integral is an Inverse Gaussian density
i.e. $\alpha_g^2 \sim$ Inverse Gaussian $\left(\frac{\lambda}{\sqrt{\beta_g^T\beta_g}}, \lambda^2\right)$,
thus the integrals integrate to 1 for all $g = 1, \ldots, G^*$. Therefore,

$$\Pi(\beta_{\xi_n}) = \prod_{g=1}^{G^*} (\lambda^2)^{\frac{m_g}{2}} \frac{e^{-\lambda\sqrt{\beta_g^T\beta_g}}}{(2\pi)^{\frac{m_g-1}{2}}} \tag{3.8}$$

Let ξ_n be the model obtained from the median thresholding posterior probability and ξ_n^* be the true model. We want to show that the model ξ_n converges to the true model ξ_n^* as the sample size n becomes sufficiently large. Define f^* as the true density under model ξ_n^* and f as the density proposed under model ξ_n. Hellinger distance between f and f^* is defined as

$$d(f, f^*) = \sqrt{\int\int \left(\sqrt{f} - \sqrt{f^*}\right) \nu_y(d_y)\nu_x(d_x)}.$$

To investigate posterior convergence, we formulate the following theorem based on Theorem 4 in Jiang's (2007) paper. We consider logistic regression in this paper with a density of the form $p^*(y|x) = \exp\{a(h^*)y + b(h^*) + c(y)\} \equiv f(y, h^*)$ where, $h^* = x^T\beta^*$ is the linear parameter, $a(h)$ and $b(h)$ are continuously differentiable, and $a(h)$ has non-zero derivative. The mean function $\mu^* = E(y|x) = -\frac{b'(h^*)}{a'(h^*)} \equiv \psi(x^T\beta^*) = \frac{e^{h^*}}{1+e^{h^*}}$. Thus ψ is the inverse of the logistic link function. Assume that $\lim_{n\to\infty} \sum_1^G \sqrt{\beta_g^{*T}\beta_g^*} < \infty$. Let r_n be the prior expectation of model size and for simplicity ξ be the corresponding subset model for which $|\beta| > 0, |\xi|$

is the corresponding model size. Define, $\Delta(r_n) = \inf_{\xi:|\xi|=r_n} \sum_{j:j\notin\xi} |\beta_j^*| < \infty, B(r_n) = \sup_{\xi:|\xi|=r_n} ch_1(V_\xi^{-1})$and $\bar{B}(r_n) = \sup_{\xi:|\xi|=r_n} ch_1(V_\xi)$. Let, $\tilde{B_n} = \sup_{\xi:|\xi|\leq K_n} ch_1(V_\xi)$ where K_n is the maximal model size. Let, $D(R) = 1 + R \times \sup_{|h|\leq R}|a'(h)|. \sup_{|h|\leq R}|\psi(h)|$ for any $R > 0$. Here, $ch_1(V_\xi)$ and $ch_1(V_\xi^{-1})$ are the largest eigenvalues of V_ξ and V_ξ^{-1} respectively.

Let $\epsilon_n \in (0, 1]$ for each n and $n\epsilon_n \succ 1$ and assume the following conditions hold:

Assumption1 :

A1. $K_n \log\left(\frac{1}{\epsilon_n^2}\right) \prec n\epsilon_n^2$

A2. $K_n \log(P_n) \prec n\epsilon_n^2$

A3. $K_n \log\left(D\left(K_n \frac{\tilde{B_n} n\epsilon_n^2}{\lambda_n}\right)\right) \prec n\epsilon_n^2$

A4. $r_n \prec P_n$

A5. $r_n \log \bar{B_n}(r_n) \prec n\epsilon_n^2$ and $\Delta(r_n) \prec n\epsilon_n^2$

A6. $\log\left(\frac{r_n}{P_n}\right) \leq -\frac{4n\epsilon_n^2}{P_n}$

A7. m_g is such that $\sum_{g=1}^{G^*} m_g \prec P_n, \forall g = 1 \ldots G$

We will replace λ by λ_n since λ and $\tau_g^2, g = 1, \ldots, G^*$ are dependent on n. Also, λ_n is inversely proportional to the sum of all τ_g^2's.

Theorem 1. *Assume the prior setting on 3.8 is used and the Assumption 1 hold. Let $P\{.\}$ denote the probability measure for the data D^n. Assume, $G \prec P_n$, $1 \leq \lambda_n \leq B(r_n), |x_j| \leq 1$ for all j and $\lim_{n\to\infty} \sum_1^G \sqrt{\beta_g^{*T}\beta_g^*} < \infty$ where P_n is a nondecreasing sequence in n. Also, let V_ξ be such that $\tilde{B_n} \geq 4$.*

Let ϵ_n be a sequence such that $\epsilon_n \in (0, 1]$ for each n and $n\epsilon_n^2 \succ 1$ and $\tau_g^2 < \infty, g = 1, \ldots, G^*$. Then, we have,

(i) For some $c_o > 0$,

$$\lim_{n\to\infty} P\big\{\pi\big[d(f,f^*)\le\epsilon_n|D^n\big]\ge 1-e^{-c_0 n\epsilon_n^2}\big\}=1$$

(ii) For some $C_1>0$ and for all sufficiently large n,

$$P\big\{\pi\big[d(f,f^*)>\epsilon_n|D^n\big]\ge e^{-0.5c_1 n\epsilon_n^2}\big\}\le e^{-0.5c_1 n\epsilon_n^2}$$

Proof. The proof follows by checking conditions N and O of Theorem 4 from Jiang's paper (2007) since our prior falls under the category of general prior in their paper. If we can show that our setup satisfies these two conditions under the given assumptions then by Theorem 4 in Jiang (2007) we have the posterior consistency of our spike and slab prior in this logistic regression model.

Statement of Condition (O):

Let $D(R)=1+R\times\ \sup_{|h|\le R}|a'(h)|.\ \sup_{|h|\le R}|\psi(h)|$ for any $R>0$. There exist some $C_n>0$ and some K_n satisfying $1\le K_n\le P_n$, such that

$$K_n\ln\left(\frac{1}{\epsilon_n^2}\right)\prec n\epsilon_n^2,$$
$$K_n\ln P_n\prec n\epsilon_n^2,$$
$$K_n\ln D(K_nC_n)\prec n\epsilon_n^2.$$

Furthermore, for all large enough n, the following two equations hold:

$$\pi(|\xi|>K_n)\le e^{-4n\epsilon_n^2},$$

and for all ξ such that $|\xi| \leq K_n$, for all $j \in \xi$,

$$\pi(|\beta_{gj}| > C_n|\xi) \leq e^{-4n\epsilon_n^2}.$$

Checking Condition (O) :

We have, $1 \leq K_n < P_n$.

$$\pi(|\xi| > K_n) = \pi(|\xi| = P_n) = \left(\frac{r_n}{P_n}\right)^{P_n}$$
$$\log \pi(|\xi| = P_n) = K_n \log\left(\frac{r_n}{P_n}\right) \leq -4n\epsilon_n^2 \tag{3.9}$$

(3.9) holds from assumption (A6). Thus, $\pi(|\xi| > K_n) \leq e^{-4n\epsilon_n^2}$ is checked. Now, we need to show that $\pi(|\beta_{gj}| > C_n|\xi) \leq e^{-4n\epsilon_n^2}$ holds.

$$\begin{aligned}
\pi(|\beta_{gj}| > x|\xi) &\propto \int_x^\infty e^{-\lambda_n \sqrt{\sum_{j=1}^{m_g} \beta_{gj}^2}} d\beta_{gj} \\
&\leq \int_x^\infty e^{-\lambda_n \beta_{gj}} d\beta_{gj} \\
&= -\frac{e^{-\lambda_n \beta_{gj}}}{\lambda_n}\Bigg|_x^\infty \\
&= \frac{e^{-\lambda_n x}}{\lambda_n}
\end{aligned}$$

Choose, $x = C_n = \frac{\tilde{B_n} n\epsilon_n^2}{\lambda_n}$ and $n\epsilon_n^2 \succ 1$. Then,

$$\begin{aligned}\frac{e^{-\lambda_n x}}{\lambda_n} &= \frac{e^{-\tilde{B}_n n\epsilon_n^2}}{\lambda_n} \\ &= e^{-\tilde{B}_n n\epsilon_n^2}, \text{ since } \lambda_n \geq 1, \\ &\leq e^{-4n\epsilon_n^2}\end{aligned}$$

This implies, that Condition (O) is checked. Note that, all the other conditions are also satisfied with the choice of C_n and assumptions of the theorem.

$$\pi(|\beta_{gj}| > C_n|\xi) \leq e^{-4n\epsilon_n^2}$$

Statement of Condition (N):

Assume that a sequence of (nonempty) models ξ_n exists such that, as n increases,

$$\sum_{j:j\notin\xi_n} |\beta_j^*| \prec \epsilon_n^2,$$

and for any sufficiently small $\eta > 0$, there exists N_η such that, for all $n > N_\eta$, we have

$$\pi(\xi = \xi_n) \geq e^{-\frac{n\epsilon_n^2}{8}},$$

and

$$\pi\left[\beta \in \left(\beta_j^* \pm \frac{\eta\epsilon_n^2}{|\xi_n|}\right)_{j\in\xi_n} \Big|\xi_n\right] \geq e^{-\frac{n\epsilon_n^2}{8}}$$

Checking Condition (N) :

Take the sequence of models ξ_n, such that, for each $n, \xi = \xi_n$ reached its infimum in $\Delta(r_n) = \inf_{\xi:|\xi|=r_n} \sum_{j:j\notin\xi} |\beta_j^*|$. Then, $\sum_{j\notin\xi_n} |\beta_j^*| = \Delta(r_n) \prec n\epsilon_n^2$.

For the condition on prior, $\pi\left[\beta \in \left(\beta_j^* \pm \frac{\eta\epsilon_n^2}{r_n}\right)_{j\in\xi_n} \Big|\xi_n\right]$:

$$\pi\left[\beta \in \left(\beta_j^* \pm \frac{\eta\epsilon_n^2}{r_n}\right)_{j\in\xi_n} \Big|\xi_n\right] \geq \prod_{g=1}^{G^*} \left[\frac{(\lambda_n^2)^{\frac{m_g}{2}}}{(2\pi)^{\frac{m_g-1}{2}}} e^{-\sqrt{\beta^T\beta}} \left(\frac{\eta\epsilon_n^2}{r_n}\right)\right]$$

for some intermediate value of β achieving the infimum of the density over $\left(\beta_j^* \pm \frac{\eta\epsilon_n^2}{r_n}\right)_{j\in\xi_n}$.

Note that,

$$\begin{aligned}
\lambda_n \sum_{g=1}^{G^*} \sqrt{\beta^T\beta} &\leq \sum_{g=1}^{G^*} ||\beta_g|| B(r_n) \\
&= \left(\sum_{j\in\xi_n} \sqrt{\beta_j^T\beta_j}\right) B(r_n) \\
&\leq C_1 B(r_n)
\end{aligned}$$

for some constant $C_1 > 0$, since we have assumed that $\lambda_n \leq C_1 B(r_n)$ for all large enough n and $\sum_{j\in\xi_n} \sqrt{\beta_j^T\beta_j} \leq \lim_{n\to\infty} \sum_{g=1}^{G^*} \sqrt{\beta_g^T\beta_g} + \frac{r_n\eta\epsilon_n^2}{r_n}$ is bounded.

Also note that,

$$\prod_{g=1}^{G^*} \frac{(\lambda_n^2)^{\frac{m_g}{2}}}{(2\pi)^{\frac{m_g-1}{2}}} = \frac{(\lambda_n^2)^{\frac{|\xi_n|}{2}}}{(2\pi)^{\frac{|\xi_n}{2}} - \frac{G^*}{2}}$$
$$\geq e^{-C_2 r_n + C_3 r_n \ln \bar{B}(r_n)}$$

for some constant $C_2 > 0$ and $C_3 > 0$. This is due to the fact $\frac{1}{\lambda_n} \leq \bar{B}(r_n)$ for all large enough n. Therefore,

$$\pi\left[\beta \in \left(\beta_j^* \pm \frac{\eta\epsilon_n^2}{r_n}\right)_{j\in\xi_n} \Big| \xi_n\right] \geq exp\Big[- C_2 r_n - C_3 r_n \ln \bar{B}(r_n) - C_1 B(r_n) - r_n \ln\left(\frac{r_n}{\eta\epsilon_n^2}\right) \Big]$$

This will be greater in order than any $e^{-cn\epsilon_n^2} (c > 0)$, satisfying a requirement of Condition (N), since $r_n, r_n \ln \bar{B}(r_n)$ and $B(r_n)$ are all smaller than $n\epsilon_n^2$ in order and so are

$$r_n \ln r_n \leq K_n \ln P_n$$
$$r_n \ln\left(\frac{1}{\epsilon_n^2}\right) \leq K_n \ln\left(\frac{1}{\epsilon_n^2}\right)$$

Now, consider the condition on $\pi(\xi_n)$:

$$\pi(\xi_n) = \frac{\text{Beta } (a + P_n - |\xi_n|, |\xi_n| + b)}{\text{Beta } (a, b)}$$
$$\approx \left(\frac{|\xi_n|}{P_n}\right)^{|\xi_n|} \tag{3.10}$$

(3.10) is derived by choosing $a = |\xi_n|$ and $b = 1$ and using approximations for factorial

terms. The approximation also ignores $O(1)$ terms.

Notice that ξ_n is chosen such that $|\xi_n| = r_n$ so,

$$\begin{aligned}\log \pi(\xi = \xi_n) &\sim r_n \log\left(\frac{r_n}{P_n}\right) \\ &\geq -r_n \log P_n\end{aligned}$$

and $r_n \log P_n \succ n\epsilon_n^2$.

So, $\pi(\xi = \xi_n) \geq e^{-\frac{n\epsilon_n^2}{8}}$. Thus, Condition (N) is checked.

□

3.5 Simulation

Before applying the group level Bayesian selection method on the brain image data we run a simulation study. The simulation study has the unknown parameters in control and tests the method on controlled inputs. We work on two different scenario where the first case is high-dimensional while the second case in large n small p scenario.

- Example 1:

The number of observations is 60 and there are 16 predictors each with 5 levels. Thus the number of parameters here is really 80. So, we are essentially looking at a small n large p problem here. The setup here is adapted from Example 2 of Xu and Ghosh's (2015) paper. We define the j^{th} predictor as $X_{gj} = z_g + z_{gj}$, where z_g and z_{gj} are independent standard normal variables and $g = 1, \ldots, 16, j = 1, \ldots, 5$. Thus, the predictors in a group are correlated but the predictors in different groups are independent. Assign true parameter

values as follows:

$$\beta = ((7, 6, 3, 4, 5), \mathbf{0}, (4, 5, 6, 10, 7), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, (2, 3, 4, 5, 6))$$

where $\mathbf{0}$ is a 0 vector of length 5. Use the simulated X and β to generate 60 independent Bernoulli random values using the logit link. Here, 40 observations are used to train the model and the rest are used as a test dataset.

• Example 2:

The number of observations is 100 and there are 4 predictors each with 10 levels which makes p = 40. The design matrix is generated exactly as in Example 1. Let

$$\beta = (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{2})$$

where $\mathbf{0}$ and $\mathbf{2}$ are vectors of length 10 with all elements 0 and 2 respectively. Use the simulated X and β to generate 100 independent Bernoulli random values using the logit link. 60 randomly selected rows were used as train dataset and the remaining as test data.

Hyperparameters, for both cases, a and b were both set to 1.5. 20,000 Monte-Carlo iterations were implemented. 28 bootstrapped samples were used to average out bias in estimates.

Table 3.1 summarizes the true and false positive rates and the negative log-likelihood of the two examples mentioned above. Both the methods are able to identify the true variables although the frequentist group lasso has a high false positive rate. This indicates that the group lasso tends to select more variables for an optimal tuning parameter. On the other hand, the model selected by median thresholding gives excellent result in terms of variable selection. It not only identifies the true positives correctly in all cases, it also estimates the

Table 3.1: Mean (Standard Error) True/False Positive Rate and Negative Log-Likelihood in 28 Simulations

		Bayesian Spike and Slab Group Lasso	Frequentist Group Lasso
Example 1	TPR	1.00 (0)	1.00 (0)
	FPR	0.00 (0)	0.78 (0.50)
	Neg log-likelihood	2.65 (0.53)	6.52 (0.57)
Example2	TPR	1.00 (0)	1.00 (0)
	FPR	0.00 (0)	0.32 (0.24)
	Neg log-likelihood	-2.35 (2.31)	5.45 (1.59)

unimportant values to be 0. We see that the method proposed in this paper gives a smaller negative log-likelihood indicating a better model fit. Thus, we see that in a simulated dataset, the median thresholding method is able to classify variables very well as compared to the conventional group lasso method when we have variables that have a structured correlation.

3.6 Classification of Alzheimer's Disease using ADNI MRI data

The MRI data used in this section of the paper was obtained from ADNI database. The main objective of ADNI has been to test whether serial MRI, PET, other biological biomarkers, and clinical and neuropsychological assessment can be used to detect dementia or measure its progression. Both normal aging and AD patients have brain region atrophies but it is essential to identify the abnormalities that lead to dementia. Some studies are done to study the differences of brain atrophy in these two categories of subject (Double *et.al.*, 1996). Such studies have shown that there is a significant difference in the atrophies of normal aging and AD patients so we use this idea to classify the subjects. In this paper we delve into classification of Alzheimer's disease (AD) patients from normally aging control

(CN) subjects at the baseline and estimation of parameters of selected volumetrics. The parameter estimates give us the log of odds of being AD at baseline for a subject with a given set of volumetric measurements. Thus, baseline volumetric values for AD and normal controls from ADNI dataset serve our purpose. ADNI data is collected from 2003 onwards by NIA, NIBIB, FDA and a few pharmaceutical companies as a public-private partnership. The ADNI project is a large project involving subjects across USA and Canada from more than 50 sites. This initiative was launched to develop new treatments and follow subjects through time to monitor the effectiveness of the treatments. For more information about ADNI, visit $www.adni-info.org$.

The volumetric segmentation and cortical reconstruction of the brain is done with the help of freely available software FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). An early version of the longitudinal image processing framework (Reuter et al., 2012) is used to process the sequential scans. This process does motion correction and averaging of multiple volumetric TI weighted images, removes non-brain segments, automates Talairach transformation, segments subcortical white matter (WM) and deep gray matter (GM) volumetric structures. It also automates topology correction and surface deformation of the brain. MRI data points from 1737 subjects with baseline diagnosed as Normal, MCI and AD were collected. For all subjects at each visit, structural MRI scans were acquired from 1.5T scanners for ADNI1 subjects and from 3T scanners for ADNIGO and ADNI2 subjects. MRI protocols were performed across a variety of scanners such as GE, Siemens, or Philips to ensure comparability. MRI volumes were computed using FreeSurfer by UCSF/SF VA Medical Center. ADNI1's 1.5T data was run with FreeSurfer version 4.3 and ADNIGO and ADNI2's 3T data was run with FreeSurfer version 5.1. For a detailed guide please refer to the UCSF FreeSurfer Methods documented by Hartig $et.al.(2014)$.

Many studies have been done to identify the ROI's associated with AD but using the entire brain segmentation to identify 4 different volumetric aspects of a region has not been explored. This technique includes all available subregion data in the model and identifies the subregions that are potentially associated with AD. Previous studies have isolated one or a few brain regions and used their volume measurements as a predictor in the prediction of AD or MCI from CN's (Jack *et.al.*, 1999 and Jack *et.al.*, 1997). We want the model to automatically select the atrophied regions rather than subsetting a brain region before start of the analysis. Classification of AD from CN has been done using FDG-PET scan (Herholz *et.al.*, 2002) using comparative statistical methods like t-statistics but researcher have yet to explore variable selection techniques using the entire volumetric data. Wang *et.al.*, 2014 used Haar wavelets to identify ROI's using voxel level data for dimension reduction. This method identifies ROI's successfully but does not narrow down the brain hemisphere of the ROI's. Since, our data is present for each region for the left and right hemispheres, we are able to identify the exact part of the ROI that is more significantly associated with AD. For the analysis, we use AD and CN patients to distinctively understand the difference of brain regions that cause a subject to be cognitively normal or progress to AD.

We have used the longitudinal processing data for our analysis. Due to advancement of technology in the computing area, quantitative assessment of brain volumes, obtained through volumetric MRI, are being used extensively for studies involving Alzheimer's disease. Volumetric measurements are mainly based on brain segmentation done at reliable MR centers.

The demographic characteristics of the 421 subjects are given in Table 3.2. The age and sex distribution in our dataset shows that the data is not skewed with respect to these two variables. Also, the maximum and minimum age for AD is 55.1 and 90.9 respectively and

that of CN is 59.9 and 89.6 respectively. Thus, the effect of age in the outcome has been controlled for in the ADNI dataset.

Table 3.2: Demographics of patients in ADNI data used for analyses

Category	Sex (male) Count (%)	Age Mean (SD)
AD (n = 191)	100 (52.36%)	75.27 (7.46)
CN (n = 230)	120 (52.17%)	75.86 (5.01)

We have used baseline data of 421 subjects of whom 191 have Alzheimer's disease (AD) and 230 are cognitively normal subjects. There are 72 predictors (brain regions segmented with FreeSurfer) with 4 levels each namely, volume, area, thickness average and thickness standard deviation. 46 brain regions had volume data only. The regions marked 'Unknown' and 'Undetermined' were discarded beforehand because these regions were not identified in the MR scans. We used all the remaining 116 brain regions (single and four-leveled) as predictors for dementia. The analysis to identify the (few) significant regions from the entire brain region is performed. The model identifies a unique set of brain regions significant for the classification of a binary outcomes. Our objective is to be able to select an optimal model that identifies the important brain regions for identifying the two kinds of brain cognitive functionality. Early diagnosis of dementia is very important as it can help in prompt treatment of subjects thus delaying progression of AD, oversee treatment efficiency and reduce time and costs of clinical trials.

The brain regions are segmentation of both gray matter (GM) and white matter (WM). Studies suggest that the gray matter is associated with cognitive disorders in elderly people. We keep both GM and WM to test the efficacy of our model i.e. if the model is efficient in selecting the correct brain regions. Variable selection selects the significant brain region

from a large collection of brain regions and then the model successfully classifies the subjects using the test dataset.

We have a logistic model for the two outcomes of the response variable. Around 70% of the data is used to train the model. The prior placed on the coefficients is a spike and slab type prior that encourages zero estimates for predictors which are not significant. The model is selected using median thresholding method. We run 10,000 iterations of the MCMC chain of which the first 5000 are used as burn-ins. The usual convergence diagnostics are performed.

295 subjects were randomly selected from the 421 subjects to train the model. The median thresholding model selects 29 out of the 116 brain subregions. These regions correspond to ROIs, namely, right bankssts, right pallidum, pars oper cularis, left pars orbitals, right precuneus, putamen, right anterior cingulate cortex, superior frontal, entorhinal cortex, supramarginal gyrus, right transverse temporal, left hippocampus, left inferior lateral ventricle, middle temporal gyrus, inferior temporal gyrus, left precentral gyrus, right fusiform gyrus, left parahippocampal, paracentral, third ventricle (Feng *et.al.*, 2004) and right inferior parietal. The estimates of all these regions are, however, not statistically significant. Some regions have negligible amount of contribution in the model thus making the credible interval of the feature exclude the corresponding subregion. We only keep the subregions in our model which are statistically significant as given by the corresponding credible intervals.

Figure 3.1: Some regions of interest selected by the Bayesian classification model

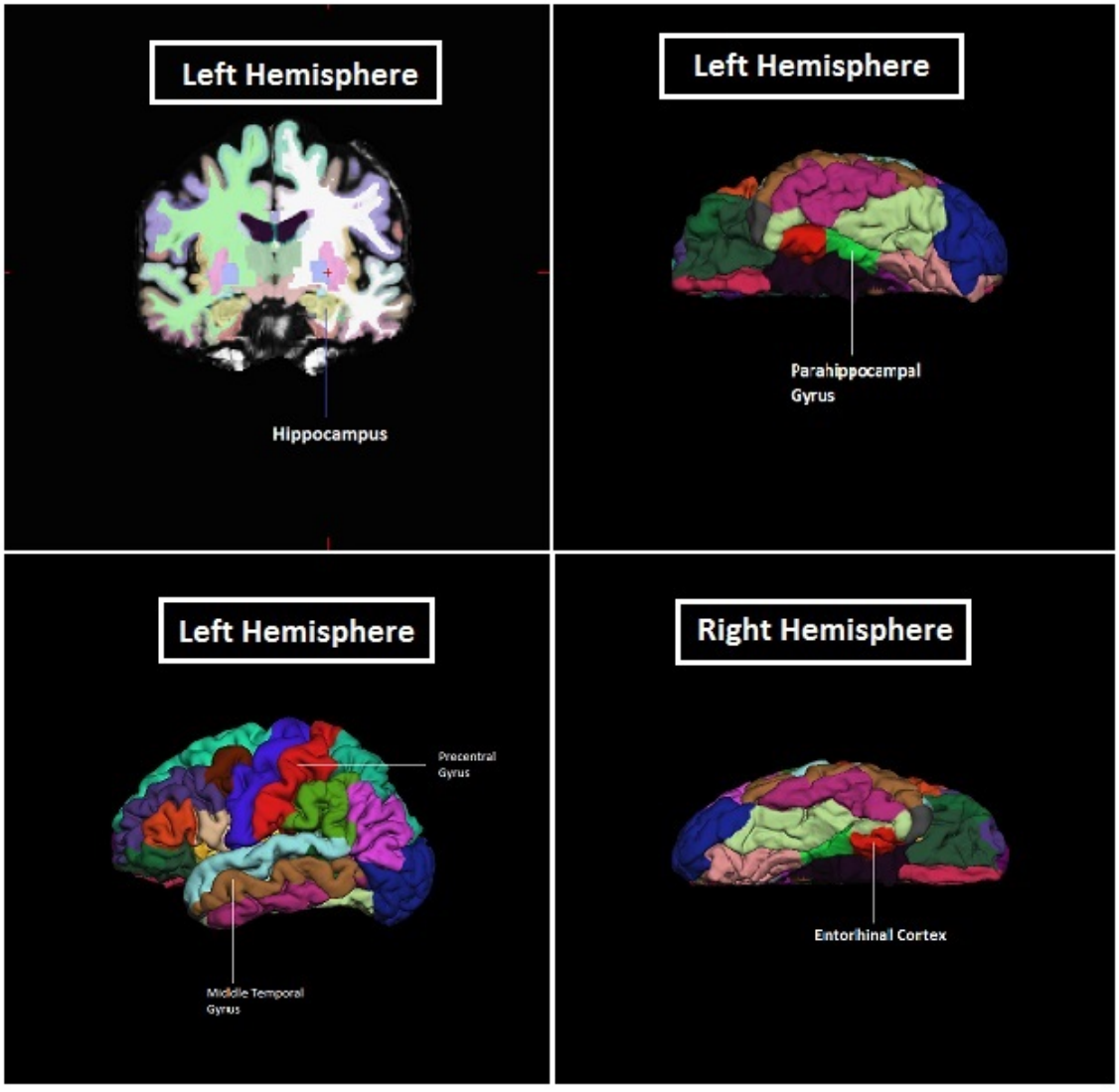


Table 3.3: Mean (Standard Error) of Parameter Estimates of Selected ROIs

ROI	Volume	Surface Area	Cortical Thickness Average	Cortical Thickness Standard Deviation
Right Entorhinal Cortex	2.97 (0.03)	1.83 (0.02)	2.14 (0.01)	1.69 (0.03)
Right Pallidum	-0.27 (0.12)	$*$	$*$	$*$
Right Pars Orbitals	-	-	-0.33 (0.16)	-
Right Precuneus	0.39 (0.05)	-0.42 (0.07)	1.0 (0.09)	-
Right Putamen	-0.21 (0.07)	$*$	$*$	$*$
Left Anterior Cingulate	-0.46 (0.12)	-0.12 (0.04)	0.15 (0.05)	0.09 (0.04)
Right Trasnverse Temporal	-	-	-0.54 (0.22)	-
Left Entorhinal Cortex	-	0.22 (0.07)	0.21 (0.07)	-0.41 (0.13)
Left Precentral Gyrus	0.25 (0.08)	0.38 (0.12)	0.45 (0.15)	0.25 (0.09)
Left Parahippocampal Gyrus	1.97 (0.11)	1.90 (0.11)	2.0 (0.11)	2.07 (0.11)
Right Bankssts	0.06 (0.02)	-	-	-
Left Middle Temporal Gyrus	1.93 (0.04)	1.57 (0.04)	1.91 (0.02)	1.63 (0.03)
Left Hippocampus	1.50 (0.02)	$*$	$*$	$*$

$*$ means these region measurements were not captured in data.
$-$ means that these regions were not statistically significant although the region was selected by median thresholding.

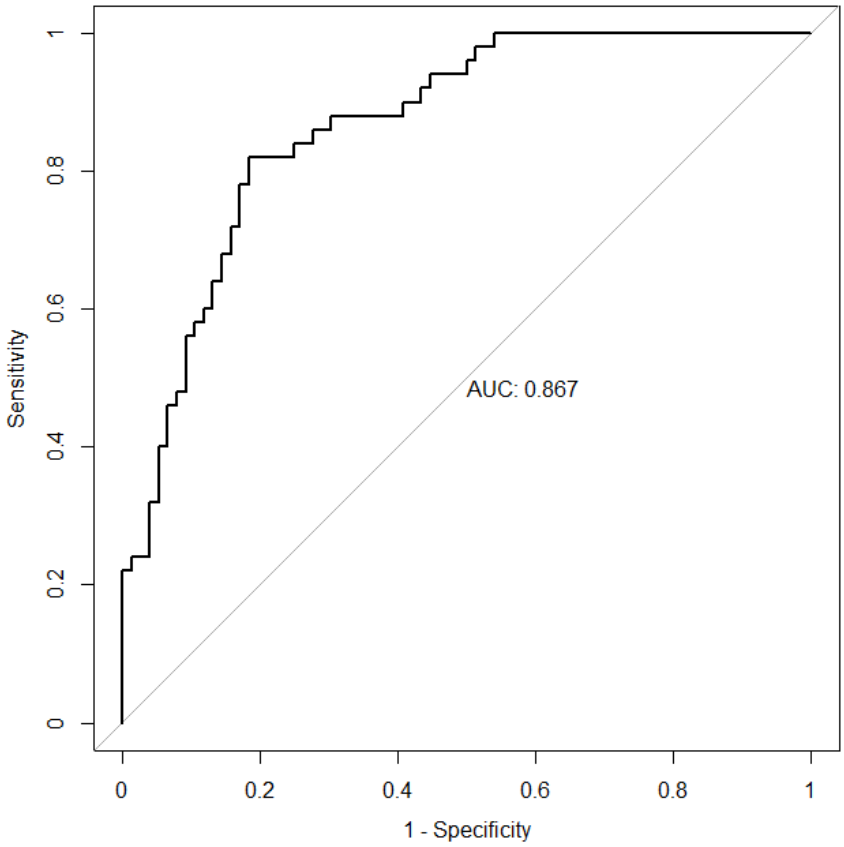
The statistically significant ROI's are given in Table 3.3. Previous studies have established the association of these regions in AD. Volumes of right entorhinal cortex are severely diminished in AD patients. The other regions selected are also coherent with relevant literature (Juottonen *et.al.*, 1998 and Galton *et.al.*, 2001). The precentral gyrus controls motor skills, middle temporal gyrus regulates semantic memory processing; hippocampus, parahippocampal and entorhinal regulate memory and navigation. Putamen, pallidum, transverse temporal and bankssts are all known to be affected by AD (Clerx *et.al.*, 2013). Recent studies have separately analyzed all these regions and found atrophies in those areas. The proposed method identifies a subset of all atrophied regions and then selects fewer regions as discriminative features for classification. The functions of the selected regions also intuitively implicate the precision of the model. Zhang *et.al.* (2011) classified AD and CN using support vector machine (SVM) thus leading to non-interpretability of the associated coefficients.

The method achieved fairly high accuracy of 80%. Cuingnate *et.al.* (2011) classified AD and CN based on ROI's but they restricted their analysis to a few selected ROI's namely the entorhinal thickness, supramarginal cortex thickness and hippocampal volume. Their sensitivity ranged from 69% to 70% whereas our method gives a sensitivity of 76%. The specificity in their study (90%) is, however, higher than ours (83%). These two studies are, however, not directly comparable except that they are both classification studies because the datasets used in these studies are different. The drawback of their method is that they pre-select a few ROI's and perform classification, unlike our method. On the other hand, the proposed method is based on statistical foundations that account and measure uncertainties due to randomness in the data set.

Our logistic model coded CN as 1 and AD as 0 so the parameter estimates should be

interpreted accordingly. Table 3.3 gives the mean parameter estimate and standard error (within parentheses) of the selected ROI's. To diagnose the accuracy of our test we build an ROC curve. An ROC (Receiver operating characteristic) curve is a plot of the sensitivity vs. 1-specificity. Figure 3.2 shows we have an area under the curve(AUC) of 0.867. This means that if we randomly draw samples for classification then our classifier will accurately classify 86.7% of the time. This AUC tells us that our classifier is, indeed, a good one.

Figure 3.2: ROC curve for our classifier



3.7 Discussion

In this chapter we propose a Bayesian approach of variable selection with spike and slab prior to identify cognitively healthy controls from Alzheimer's patients. This method uses whole brain parcellation data to classify dementia as well as interpret the association of each significant volumetric measure of a brain subregion. This technique captures the structured correlation in this type of data to retain all levels of the subregion that are disease related. The Bayesian approach guarantees better standard error estimates. Also, the median thresholding method for posterior model selection together with the use of spike and slab prior is a more efficient method than the frequentist group lasso method as shown in the simulation study. Liang *et.al.* (2013) developed a Bayesian subset selection method for generalized linear models which can select individual variables only. In their method, they place a prior on the model to perform subset selection unlike our approach of using a spike and slab prior which directly drops out variables in the many Monte Carlo iterations. Our median thresholding, as opposed to their MAP posterior probability, is able to choose the best model without comparing information criterion type quantities among several candidate models. Thus, the spike and slab prior median thresholding Bayesian group lasso has attractive properties of high dimensional variable selection and performs efficiently with structured correlated covariates. Most of the other dimension reduction techniques are unable to tackle correlated variables in variable selection.

The significant regions selected by our model is identified from a large number of subregions thus accounting for the effect of the whole brain while performing dimension reduction. Wang *et.al.* (2014) in their paper have introduced a dimension reduction technique using HAAR wavelet based on voxel level data with ADNI PET data. Their method builds on

continuous outcomes and does not perform variable selection. Our approach builds the model with the MRI brain parcellated volumetric data which are a direct indicator of dementia. This Bayesian formulation not only tackles ANOVA type dummy variables but also deals with the high dimension problem. The simulation results show that this method is effective in both low and high dimension. AUC of 0.867 shows that our model is a good classifier. Most authors have only reported the accuracy of their model but a look at AUC helps us better understand the diagnostic ability of a classifier. The greatest advantage of this method is that it considers all the subregions while building the model and efficiently drops the ones that are not disease related and, can also easily interpret the risk of dementia from the parameter estimates. This is a novel contribution to classification for Alzheimer's disease to the best of our knowledge.

Chapter 4

Bayesian Spatiotemporal Model for Detecting Voxel-level Activation in fMRI Data

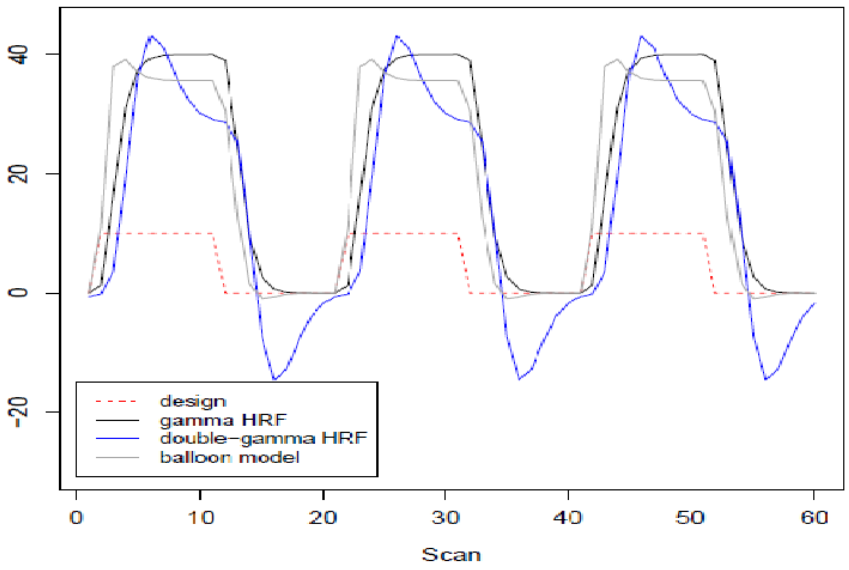
4.1 Introduction

With functional magnetic resonance imaging (fMRI), brain activation can be recorded. This is done by measuring blood oxygenation level dependent (BOLD) activation through MRI techniques. An external stimulus is administered to a subject and their BOLD MR signals are recorded at each voxel as a time series. This is useful to understand the functioning of human brain with respect to brain development, diseases related with brain functionality, neuronal activity in alcoholism etc. A single subject is administered to an external stimuli, like seeing an object or administering a physical stimuli like touching a hot surface and the MR scanner captures 3D images at every 2 to 3 secs of the entire brain. These 3D images are comprised of voxel level data on a lattice for multiple brain slices.

There are approximately 60×60 voxels in a 2D lattice with approximately 30 of such brain slices. Thus, the number of voxels is of the order of 1 million. BOLD signal responses of 1 million voxels are recorded at around 190 timepoints. Thus, we have data of the size of

20 million for a single subject making the data structured highly complicated. The stimulus is a so-called "boxcar" design where there are alternative periods of activation and rest. This stimulus is a 0-1 input and is transformed by a hemodynamic response function (HRF), Gitelman *et.al.* (2003) (See Figure 4.1, Welvaert *et.al.* (2011)). This transformation is done under the assumption that blood oxygenation is a delayed and proceeds continuously. Thus, the 0-1 boxcar stimuli is convolved with an HRF density, like double-gamma density. The transformed stimulus is the regressor in the model and the corresponding coefficient is called the "amplitude" of activation. Variable selection is equivalent to retaining the voxels when the corresponding amplitude is non-zero.

Figure 4.1: Task stimuli based on three convolution functions



The time series nature of the images occur due to voxel level images being captured at multiple timepoints. Since voxels in our brain are interconnected in an arbitrary fashion, the

presence of spatial correlation cannot be ignored. Standard fMRI analysis ignores spatial dependence and performs correlation analysis or parametric regression analysis voxelwise. The objective of fMRI data analysis is to identify the activated voxels (brain regions) that respond to the stimuli. This, clearly, is a variable selection problem where the covariates are brain voxels. We build a Bayesian variable selection model that accounts for the spatiotemporal correlation inherent in fMRI data.

Friston *et.al.* (1995) applied separate regression at each voxel to get t statistics for the activation amplitude. This popular method, however, does not account for spatial dependence. Recently, many Bayesian approaches have been developed using Gaussian MRF priors (Gossl, Auer, and Fahrmeir, 2001 and Fahrmeir and Gossl,2002) where MCMC algorithm is implemented to obtain posterior distributions. Penny, Trujillo-Barreto, and Friston (2005) used Gaussian spatial priors on regression coefficients and autoregressive coefficients of the noise process. Although, this procedure reduces computational time considerably, it replaces the true posterior with an approximate posterior. Smith and Fahrmeir (2007) introduced a spatial variable selection for fMRI data by placing an Ising prior on the activation amplitudes but they ignore the temporal correlations in the data. Recently, Musgrove *et.al.* (2016) developed a Bayesian variable selection method by parcellating the brain into several smaller units. Their method takes into account both spatial and temporal correlation. They assume a sparse areal generalized linear model ((Hughes and Haran, 2013) with spatial random effects. They use latent variables distributed as Bernoulli to indicate the presence or absence of a voxel. They perform variable selection based on the posterior probability of activation, thresholding the value at 0.8722. A major drawback of this method is that the accuracy is not validated for random parcellation of the dataset.

We develop a Bayesian variable selection method by incorporating group lasso technique

into a bilevel selection method. Our regressors are voxels that are task-based (depends on stimuli). We are interested in knowing the stimuli that activates brain voxels as well as the voxels that are activated. So, a stimuli as a regressor comprises of N voxels. In this case, selecting a stimuli is equivalent to selecting a group of covariates; where the group of covariate is the chosen stimulus at each voxel. Now, we can approach this as a group lasso problem but we know that for a chosen stimulus, all voxels may not be activated. So, we want a second level of selection where we can select the activated voxels. Thus, first we select a stimulus such that the amplitudes are non-zero and then finally choose the voxels when a second step of selection drops some amplitudes of activation indicating those voxels are not activated.

Motivated by Xu and Ghosh's (2015) bilevel variable selection, we extend their method by including spatiotemporal correlation. The large size of data presents a host of issues in data handling. We will discuss thoroughly how we have built a Gibb's sampler for efficient computation. To enable shrinkage both at group level and within a group, we use a spike and slab prior like before. To ensure sparsity we reparametrize the parameters. Park and Casella (2008) and Kyung *et.al.* (2010) have worked with two level hierarchical structure with mixture priors that has shrinkage effects but fail to produce sparse solutions since the posterior mean/ median is never exactly zero. Spike and slab ensures sparsity. Unlike selecting variables whose posterior probability is above a threshold, we incorporate posterior median thresholding for variable selection. The posterior median obtained from the posterior median of our MCMC samples tell us which amplitudes are active.

4.2 Bayesian Bi-level Variable Selection

Before we introduce the spatiotemporal model, let us have a look at the bilevel selection method when we assume that there is independence of the responses. Since the BOLD signals are normally distributed, we are concerned with linear models here. Bilevel selection in frequentist approach was proposed by Simon *et.al.* (2012) where the objective function is regularized using a combination of $l-1$ and $l-2$ penalty as follows:

$$min_{\beta}\left(\left|\left|Y-\sum_{g=1}^{G}X_g\beta_g\right|\right|_2^2+\lambda_1||\beta||_1+\lambda_2\sum_{g=1}^{G}\left|\left|\beta_g\right|\right|_2\right) \tag{4.1}$$

Thus, a MAP solution with prior of β_g in (1.2)

$\pi(\beta_g) \propto \exp\left\{-\frac{\lambda_1}{2\sigma^2}||\beta_g||_1-\frac{\lambda_2}{2\sigma^2}||\beta_g||_2\right\}, \quad g=1,\ldots,G$ is equivalent to (4.1). The coefficients are reparametrized to introduce the two types of sparsity. Let, $\beta_g = A_g^{\frac{1}{2}}b_g$, where $A_g^{\frac{1}{2}} = diag\{\tau_{g1},\ldots,\tau_{gm_g}, \quad g=1,\ldots,g; j=1,\ldots,m_g$. Here, b_g has a spike and slab prior:

$$b_g \overset{ind}{\sim} (1-\pi_o)N_N(0,\mathbf{I}_{m_g})+\pi_0\delta_0(b_g), \quad g=1,\ldots,g.$$

Thus, $A_g^{\frac{1}{2}}$ controls the magnitude of elements of β_g. Evidently, a non-zero τ_{gj} keeps β_{gj} in model. So, we place a spike and slab prior on $\tau_{gj}s$ to ensure exact sparsity.

$\tau_{jg} \overset{ind}{\sim} (1-\pi_1)N^+\left(0,s^2\right)+\pi_1\delta_0(\tau_{jg}), \quad g=1,\ldots,G; j=1,\ldots,m_g$

where $N^+\left(0,s^2\right)$ is a normal distribution truncated at 0. We place some non-informative priors n other hyperparameters , thus, forming a hierarchical model.

$\sigma_e^2 \sim$ Inverse Gamma$(\alpha=0.1,\gamma=0.1)$

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

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

t is updated with EM algorithm as was done in chapter 3. Note that, this setup is a modification of our algorithm in chapter 3 in a linear setup. We assume independence of our responses in this case. In the next section, we extend this idea to handle datasets that have spatiotemporal correlation.

4.3 Bayesian Spatiotemporal Model with Bi-level Selection

4.3.1 Model Formulation

Consider the following model:

$$y_{vt} = x_{t1}\beta_{v1} + x_{t2}\beta_{v2} + \cdots + x_{tp}\beta_{vp} + \epsilon_{vt} \tag{4.2}$$

$v = 1, \ldots, N$

$t = 1, \ldots, T$

Here, y_{vt} is the BOLD signal contrast at timepoint t and voxel v. $x_{tj}, \quad j = 1, \ldots, p$ are the HRF transformed stimuli input and β_{vj} are the activation amplitudes. We have N voxels

and T timepoints. Let $t = 1$, then

$$\boldsymbol{y_1} = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{N1} \end{pmatrix}^{N\times 1} = \begin{pmatrix} x_{11}\beta_{11}+x_{12}\beta_{12}+\cdots+x_{1p}\beta_{1p} \\ x_{11}\beta_{21}+x_{12}\beta_{22}+\cdots+x_{1p}\beta_{2p} \\ \vdots \\ x_{11}\beta_{N1}+x_{12}\beta_{N2}+\cdots+x_{1p}\beta_{Np} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{N1} \end{pmatrix}$$
$$= \begin{pmatrix} x_{11} \\ x_{11} \\ \vdots \\ x_{11} \end{pmatrix} \otimes \boldsymbol{\beta_1} + \begin{pmatrix} x_{12} \\ x_{12} \\ \vdots \\ x_{12} \end{pmatrix} \otimes \boldsymbol{\beta_2} + \cdots + \begin{pmatrix} x_{1p} \\ x_{1p} \\ \vdots \\ x_{1p} \end{pmatrix} \otimes \boldsymbol{\beta_p} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{N1} \end{pmatrix} \tag{4.3}$$

where $\boldsymbol{\beta_g} = \begin{pmatrix} \beta_{1g} \\ \beta_{2g} \\ \vdots \\ \beta_{Ng} \end{pmatrix}, g = 1, \ldots, p.$

Thus, we can write equation (4.3) as:

$$\mathbf{y_1} = \begin{pmatrix} x_{11} & 0 & \ldots & 0 \\ 0 & x_{11} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & x_{11} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{N1} \end{pmatrix} + \cdots + \begin{pmatrix} x_{1p} & 0 & \ldots & 0 \\ 0 & x_{1p} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & x_{1p} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{1p} \\ \beta_{2p} \\ \vdots \\ \beta_{Np} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{N1} \end{pmatrix}$$
$$= \sum_{g=1}^{p} X_g^1 \boldsymbol{\beta_g} + \boldsymbol{\epsilon_1}$$

where, $X_g^1 = \begin{pmatrix} x_{1g} & 0 & \ldots & 0 \\ 0 & x_{1g} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & x_{1g} \end{pmatrix}^{N\times N}, g = 1, \ldots, p$

Thus, we have,

$\boldsymbol{y_t} = \sum_{g=1}^{p} X_g^t \boldsymbol{\beta_g} + \boldsymbol{\epsilon_t} \quad , t = 1, \ldots, T$

Stacking all of this together we have,

$\boldsymbol{y_1} = \sum_{g=1}^{p} X_g^1 \boldsymbol{\beta_g} + \boldsymbol{\epsilon_1}$

$\boldsymbol{y_2} = \sum_{g=1}^{p} X_g^2 \boldsymbol{\beta_g} + \boldsymbol{\epsilon_2}$

...

...

$\boldsymbol{y_T} = \sum_{g=1}^{p} X_g^T \boldsymbol{\beta_g} + \boldsymbol{\epsilon_T}$

$$\iff \begin{pmatrix} \boldsymbol{y_1}^{N\times 1} \\ \boldsymbol{y_2}^{N\times 1} \\ \vdots \\ \boldsymbol{y_T}^{N\times 1} \end{pmatrix} = \begin{pmatrix} \sum_{g=1}^{p} X_g^1 \boldsymbol{\beta_g} \\ \sum_{g=1}^{p} X_g^2 \boldsymbol{\beta_g} \\ \vdots \\ \sum_{g=1}^{p} X_g^T \boldsymbol{\beta_g} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_T} \end{pmatrix}$$

$$\iff Y^{NT\times 1} = \begin{pmatrix} \overbrace{\begin{pmatrix} x_{11} & 0 & \dots & 0 \\ 0 & x_{11} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{11} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{N1} \end{pmatrix}}^{\text{TASK 1}} + \dots + \overbrace{\begin{pmatrix} x_{1p} & 0 & \dots & 0 \\ 0 & x_{1p} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{1p} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{1p} \\ \beta_{2p} \\ \vdots \\ \beta_{Np} \end{pmatrix}}^{\text{TASK p}} \\ \vdots \\ \begin{pmatrix} x_{T1} & 0 & \dots & 0 \\ 0 & x_{T1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{T1} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{N1} \end{pmatrix} + \dots + \begin{pmatrix} x_{Tp} & 0 & \dots & 0 \\ 0 & x_{Tp} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{Tp} \end{pmatrix}^{N\times N} \begin{pmatrix} \beta_{1p} \\ \beta_{2p} \\ \vdots \\ \beta_{Np} \end{pmatrix} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{N1} \end{pmatrix}$$

$$\iff Y = \begin{pmatrix} X_1^1 \boldsymbol{\beta_1} \\ X_1^2 \boldsymbol{\beta_1} \\ \vdots \\ X_1^T \boldsymbol{\beta_1} \end{pmatrix}^{NT\times 1} + \dots + \begin{pmatrix} X_p^1 \boldsymbol{\beta_p} \\ X_p^2 \boldsymbol{\beta_p} \\ \vdots \\ X_p^T \boldsymbol{\beta_p} \end{pmatrix}^{NT\times 1} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{N1} \end{pmatrix}$$

$$= \widetilde{X_1}\boldsymbol{\beta_1} + \widetilde{X_2}\boldsymbol{\beta_2} + \dots + \widetilde{X_p}\boldsymbol{\beta_p} + \epsilon$$

$$= \sum_{g=1}^{p} \widetilde{X_g}\boldsymbol{\beta_g} + \epsilon$$

Thus, we have

$$Y^{NT\times 1} = \sum_{g=1}^{p} \widetilde{X_g}\boldsymbol{\beta_g} + \epsilon \tag{4.4}$$

In this setup, we have p tasks and N voxels. Measurements from each voxel is taken at T timepoints. We assume a first order auto-regressive correlation between the T timepoints for each voxel.

Let, $\boldsymbol{\epsilon_t^{N\times 1}} = \begin{pmatrix} \boldsymbol{\rho_1} & 0 & \dots & 0 \\ 0 & \boldsymbol{\rho_2} & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \boldsymbol{\rho_N} \end{pmatrix} \boldsymbol{\epsilon_{t-1}^{N\times 1}} + \boldsymbol{e_t}, t = 1, \dots, T.$

Here, $\begin{pmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_T} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 \mathbb{I}_N & 0 & ... & 0 \\ 0 & \sigma_e^2 \mathbb{I}_N & ... & 0 \\ .... & .... & : & .... \\ .... & .... & : & .... \\ 0 & 0 & ... & \sigma_e^2 \mathbb{I}_N \end{pmatrix} \right)$.

Thus, $E\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_T} \end{pmatrix} = \begin{pmatrix} \sum_{g=1}^{p} X_g^1 \boldsymbol{\beta_g} \\ \sum_{g=1}^{p} X_g^2 \boldsymbol{\beta_g} \\ \vdots \\ \sum_{g=1}^{p} X_g^T \boldsymbol{\beta_g} \end{pmatrix}$ and variance-covariance matrix of Y is V.

We want to find the structure of V. Assume that $\boldsymbol{\epsilon_t}$ is a stationary process i.e. the distribution of $\boldsymbol{\epsilon_t}$ does not depend on t. Then,

$\boldsymbol{\epsilon_{it}} = \rho_i \boldsymbol{\epsilon_{it-1}} + \boldsymbol{e_{it}}, \forall i = 1 \dots N, t = 1 \dots T \qquad \boldsymbol{\epsilon_0} = 0.$

Now,

$E(\boldsymbol{\epsilon_{it}}^2) = E(\rho_i \boldsymbol{\epsilon_{it-1}} + e_{it})^2$

$= \rho_i^2 E(\boldsymbol{\epsilon_{it-1}^2}) + E(\boldsymbol{e_{it}}^2)$

$\implies E(\boldsymbol{\epsilon_{it}}^2) = \frac{\sigma_e^2}{1-\rho_i^2}, \forall i = 1 \dots N$

Similarly, $\boldsymbol{\epsilon_{it}} = \rho_i \boldsymbol{\epsilon_{it-1}} + \boldsymbol{e_{it}}$

$= \rho_i^2 \boldsymbol{\epsilon_{it-2}} + \rho_i \boldsymbol{e_{it-1}} + \boldsymbol{e_{it}}$

$\implies E(\boldsymbol{\epsilon_{it}\epsilon_{it-2}}) = E(\rho_i^2 \boldsymbol{\epsilon_{it-2}^2} + \rho_i \boldsymbol{e_{t-1}\epsilon_{it-2}} + \boldsymbol{\epsilon_{it-2}e_{it}}) = \rho_i^2 \frac{\sigma_e^2}{1-\rho_i^2}, \forall i = 1 \dots N$

In general, $E(\boldsymbol{\epsilon_{it}\epsilon_{is}}) = \rho_i^{|t-s|} \frac{\sigma_e^2}{1-\rho_i^2}, \forall i = 1 \dots N$

$$\text{Var-Cov} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_T} \end{pmatrix} = \text{Var-Cov} \begin{pmatrix} \boldsymbol{y_{11}} \\ \boldsymbol{y_{21}} \\ \vdots \\ \boldsymbol{y_{N1}} \\ \boldsymbol{y_{12}} \\ \boldsymbol{y_{22}} \\ \vdots \\ \boldsymbol{y_{N2}} \\ \vdots \\ \boldsymbol{y_{1T}} \\ \boldsymbol{y_{2T}} \\ \vdots \\ \boldsymbol{y_{NT}} \end{pmatrix} = V$$

$$= \begin{pmatrix} \begin{pmatrix} V(y_{11}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & V(y_{N1}) \end{pmatrix} & \begin{pmatrix} cov(y_{11},y_{12}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{N1},y_{N2}) \end{pmatrix} & \dots & \begin{pmatrix} cov(y_{11},y_{1T}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{N1},y_{NT}) \end{pmatrix} \\ \begin{pmatrix} cov(y_{12},y_{11}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{N2},y_{N1}) \end{pmatrix} & \begin{pmatrix} V(y_{12}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & V(y_{N2}) \end{pmatrix} & \dots & \begin{pmatrix} cov(y_{12},y_{1T}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{N2},y_{NT}) \end{pmatrix} \\ \dots & \dots & \dots & \dots \\ \begin{pmatrix} cov(y_{1T},y_{11}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{NT},y_{N1}) \end{pmatrix} & \begin{pmatrix} cov(y_{1T},y_{12}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & cov(y_{NT},y_{N2}) \end{pmatrix} & \dots & \begin{pmatrix} V(y_{1T}) & \dots & 0 \\ \vdots & & \\ 0 & \dots & V(y_{NT}) \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} \begin{pmatrix} \frac{\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{\rho_1\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{\rho_1^{T-1}\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{T-1}\sigma_e^2}{1-\rho_N^2} \end{pmatrix} \\ \begin{pmatrix} \frac{\rho_1\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{\rho_1^{T-2}\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{T-2}\sigma_e^2}{1-\rho_N^2} \end{pmatrix} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \begin{pmatrix} \frac{\rho_1^{T-1}\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{T-1}\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{\rho_1^{T-2}\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{T-2}\sigma_e^2}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{\sigma_e^2}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\sigma_e^2}{1-\rho_N^2} \end{pmatrix} \end{pmatrix}$$

$$= \sigma_e^2 \widetilde{V}$$

We will need the inverse of the variance-covariance matrix. To find the inverse let us redefine a few notations. Let $\boldsymbol{E_j}' = (\boldsymbol{\epsilon_1}, \dots, \boldsymbol{\epsilon_j}) = (\epsilon_{11}, \dots, \epsilon_{N1}, \dots, \epsilon_{1j}, \dots, \epsilon_{Nj}), j = 1, \dots, T$ and

$$A_j = \sigma_e^2 \begin{pmatrix} \begin{pmatrix} \frac{1}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{1}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{\rho_1}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{\rho_1^{j-1}}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{j-1}}{1-\rho_N^2} \end{pmatrix} \\ \begin{pmatrix} \frac{\rho_1}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{1}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{1}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{\rho_1^{j-2}}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{j-2}}{1-\rho_N^2} \end{pmatrix} \\ \dots\dots & \dots\dots & \vdots & \dots\dots \\ \dots\dots & \dots\dots & \vdots & \dots\dots \\ \dots\dots & \dots\dots & \vdots & \dots\dots \\ \dots\dots & \dots\dots & \vdots & \dots\dots \\ \begin{pmatrix} \frac{\rho_1^{j-1}}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{j-1}}{1-\rho_N^2} \end{pmatrix} & \begin{pmatrix} \frac{\rho_1^{j-2}}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{\rho_N^{j-2}}{1-\rho_N^2} \end{pmatrix} & \dots & \begin{pmatrix} \frac{1}{1-\rho_1^2} & \dots & 0 \\ \vdots & & \\ 0 & \dots & \frac{1}{1-\rho_N^2} \end{pmatrix} \end{pmatrix}$$

Write the first order auto-regressive equations as:

$$a_0 \boldsymbol{\epsilon_{it}} + a_i \boldsymbol{\epsilon_{it-1}} = \boldsymbol{e_{it}}. \tag{4.5}$$

t = 1, . . . , T where, $a_0 = 1$ and $a_i = -\rho_i$. The distribution of $\boldsymbol{E_T}$:

$$dF(\boldsymbol{E_T}) = (2\pi)^{-\frac{NT}{2}} |A_T|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{E_T}' A_T^{-1} \boldsymbol{E_T}} d\boldsymbol{E_T} \tag{4.6}$$

And, the distribution of $\boldsymbol{\epsilon_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_T}$:

$$dF(\boldsymbol{\epsilon_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_T}) = (2\pi)^{-\frac{NT}{2}} |A_1|^{-\frac{1}{2}} e^{-\frac{1}{2}\left[\boldsymbol{E_1}' A_1^{-1} \boldsymbol{E_1} + \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \boldsymbol{e_t}' \boldsymbol{e_t}\right]} d\boldsymbol{\epsilon_1} d\boldsymbol{e_2} \dots d\boldsymbol{e_T} \tag{4.7}$$

Assume, $T > 2$. Considering (4.5) as a transformation from $\boldsymbol{e_t}$ to $\boldsymbol{\epsilon_t}, t = 2, \ldots, T$ we have,

$$dF(\boldsymbol{E_T}) = (2\pi)^{-\frac{NT}{2}} |A_1|^{-\frac{1}{2}} a_0^{NT-N} e^{-\frac{1}{2}\left[\boldsymbol{E_1}' A_1^{-1} \boldsymbol{E_1} + \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \sum_{i=1}^{N} (a_0 \boldsymbol{\epsilon_{it}} + a_i \boldsymbol{\epsilon_{it-1}})^2\right]} d\boldsymbol{E_T} \tag{4.8}$$

Comparing (4.6) and (4.8):

$$(2\pi)^{-\frac{NT}{2}} |A_T|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{E_T}' A_T^{-1} \boldsymbol{E_T}}$$
$$= (2\pi)^{-\frac{NT}{2}} |A_1|^{-\frac{1}{2}} a_0^{NT-N} e^{-\frac{1}{2}\left[\boldsymbol{E_1}' A_1^{-1} \boldsymbol{E_1} + \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \sum_{i=1}^{N} (a_0 \boldsymbol{\epsilon_{it}} + a_i \boldsymbol{\epsilon_{it-1}})^2\right]}$$

Thus,

$a_0^{2N} |A_1| = a_0^{2NT} |A_T|$ and

$\boldsymbol{E_T}' A_T^{-1} \boldsymbol{E_T} = \boldsymbol{E_1}' A_1^{-1} \boldsymbol{E_1} + \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \sum_{i=1}^{N} (a_0 \boldsymbol{\epsilon_{it}} + a_i \boldsymbol{\epsilon_{it-1}})^2$

Let, $C_T = \begin{pmatrix} A_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ and $\boldsymbol{E_T}' B_T \boldsymbol{E_T} = \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \sum_{i=1}^{N} (a_0 \boldsymbol{\epsilon_{it}} + a_i \boldsymbol{\epsilon_{it-1}})^2$

$\therefore A_T^{-1} = C_T + B_T$

So, $\boldsymbol{E_T}' B_T \boldsymbol{E_T} = \frac{1}{\sigma_e^2} \sum_{t=2}^{T} \sum_{i=1}^{N} (a_0^2 \boldsymbol{\epsilon_{it}}^2 + 2 a_i a_0 \boldsymbol{\epsilon_{it-1}}' \boldsymbol{\epsilon_{it}} + a_i^2 \boldsymbol{\epsilon_{it-1}}^2)$

Now, B_T is completely known. Thus, the inverse follows from the inversion of auto-covariance matrix by Siddiqui, 1958. Let,

$$B_T^{NT \times NT} = \begin{pmatrix} B_{11} & B_{12} & \ldots & B_{1T} \\ B_{21} & B_{22} & \ldots & B_{2T} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ B_{(T-1)1} & B_{(T-1)2} & \ldots & B_{(T-1)T} \\ B_{T1} & B_{T2} & \ldots & B_{TT} \end{pmatrix}$$

where $B_{ij} = N \times N$ matrix correspond to N voxels, $i, j = 1, \ldots, T$. Assuming, $j \geq i$,

$$B_{ji} = B_{ij} = \begin{cases} \mathbf{0}^{N\times N}, & i \le T-1, i+1 < j \le T \\ \frac{1}{\sigma_e^2}\begin{pmatrix} -\rho_1 & 0 & ... & 0 \\ 0 & -\rho_2 & ... & 0 \\ \cdots & \cdots & \cdot & \cdots \\ \cdots & \cdots & \cdot & \cdots \\ 0 & 0 & ... & -\rho_N \end{pmatrix}, & i \le T-1, i \le j \le i+1 \end{cases}$$

$$B_{ii} = \begin{pmatrix} 1+\rho_1^2 & 0 & ... & 0 \\ 0 & 1+\rho_2^2 & ... & 0 \\ \cdots & \cdots & \cdot & \cdots \\ \cdots & \cdots & \cdot & \cdots \\ 0 & 0 & ... & 1+\rho_N^2 \end{pmatrix}, \quad \forall i \ne 1, T$$

$$B_{11} = \begin{pmatrix} \rho_1^2 & 0 & ... & 0 \\ 0 & \rho_2^2 & ... & 0 \\ \cdot\cdot & \cdot\cdot & \cdot & \cdot\cdot \\ \cdot\cdot & \cdot\cdot & \cdot & \cdot\cdot \\ 0 & 0 & ... & \rho_N^2 \end{pmatrix}, \quad B_{TT} = \mathbf{I}_N$$

$$\text{If, } A_T^{-1} = \begin{pmatrix} A_{11} & A_{12} & ... & A_{1T} \\ A_{21} & A_{22} & ... & A_{2T} \\ \cdots\cdots & \cdots\cdots & \cdot & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdot & \cdots\cdots \\ A_{(T-1)1} & A_{(T-1)2} & ... & A_{(T-1)T} \\ A_{T1} & A_{T2} & ... & A_{TT} \end{pmatrix}$$

where $A_{ij} = N \times N$ submatrix, $i, j = 1, \ldots, T$. We know all $A'_{ij}s$ except for i and j which is less than or equal to 1, i.e. we do not know A_{11}. Since, A_T is persymmetric, A_T^{-1} is also persymmetric. Therefore, $A_{11} = A_{TT}$. Also, $A_T^{-1} = C_T + B_T = \begin{pmatrix} A_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + B_T$ Since, $a_0^{2N}|A_1| = a_0^{2NT}|A_T|$, we have, $|A_T|^{-1} = \frac{\prod_{i=1}^N (1-\rho_i^2)}{\sigma^{2N}}$.

4.3.2 Bi-level Spatiotemporal Model for fMRI Data

(4.2) is the regression setup where we have a single brain data. The brain is divided into N voxels. p tasks(stimuli) are administered to the subject and their reactions are recorded at T time points. $\beta_g, g = 1, \ldots, p$ are the magnitude of the response to the corresponding stimuli.

We are interested in selecting a stimuli that significantly affects the response and also the voxel that is significantly associated with the corresponding task. Thus our regression setup has groups of N voxels for each stimuli. Response for this group of N voxels is present for T time points for each stimuli. Thus, (4.4) is of the group lasso form (Xu and Ghosh, 2015). Here, to select a stimuli and a voxel that are associated with the brain response accounts for bi-level selection. The first level of selection selects groups of voxel corresponding to a stimuli and then selects the voxels from among the selected stimuli. We can, thus, apply Bayesian sparse group lasso with bi-level selection. Our response is normally distributed so let us formulate the hierarchical Bayesian structure.

$$
\begin{aligned}
Y|X, \beta, \sigma_e^2, \rho &\sim N_{NT}(X\beta, V) \\
\beta_g^{N\times 1}|\tau_g, \sigma_e^2 &\sim N(0, \sigma_e^2 \Sigma_g), g = 1, \ldots, p
\end{aligned} \tag{4.9}
$$

Let $\beta_g = A_g^{\frac{1}{2}} b_g$

$b_g \overset{i.i.d.}{\sim} (1-\pi_o) N_N(0, \mathbf{I}_N) + \pi_0 \delta_0(b_g), \qquad g = 1, \ldots, p$

Here, $A_g^{\frac{1}{2}} = \begin{pmatrix} \tau_{1g} & 0 & \ldots & 0 \\ 0 & \tau_{2g} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \tau_{Ng} \end{pmatrix}$

Note that, $\tau_{jg} = 0 \implies \beta_{jg}$ is dropped out of the model even when $b_{jg} \neq 0$. This means τ_g controls a within group level selection for a selected group β_g.

In our problem, we know that there is some correlation among adjacent voxels. Consider a spatial adjacency structure among the N spatial locations that can be represented through a known spatial weight matrix $W = ((w_{ij}))$, $i, j = 1, \ldots, N$. w_{ij} is the weight corresponding to the presence or absence of adjacency between location i and j.

For the prior selection of within group, we assume a spatial cross-sectional dependence is convoluted within the covariate structure of the model. To perform a group lasso variable

selection in the above model, we need to consider a proper prior for $\beta_g, \quad g = 1, \ldots, p$ that considers the spatial (voxel) relationships among the covariates. We assume a conditional auto-regressive prior for τ as:

$\tau_{jg}|\tau_{ig} : i \neq j \sim (1 - \pi_1)N^+ \left(\sum_{i=1, i \neq j}^{N} \frac{w_{ij}}{w_{j+}} \tau_{ig}, \frac{s^2}{w_{j+}} \right) + \pi_1 \delta_0(\tau_{jg}), \qquad g = 1, \ldots, p$

$w_{j+} = \sum_{i=1}^{N} w_{ij}$

$N+ =$ folded normal towards positive side of the real line. Also,

$$
\begin{aligned}
\sigma_e^2 &\sim \text{Inverse Gamma}(\alpha = 0.1, \gamma = 0.1) \\
\pi_0 &\sim \text{Beta}(a_1, a_2) \\
\pi_1 &\sim \text{Beta}(c_1, c_2) \\
s^2 &\sim \text{Inverse Gamma}(1, t) \\
\rho &\sim \frac{1}{\sqrt{1 - \rho^2}} \text{Uniform}(-1, 1) \\
t^{(k)} &= \frac{1}{E_{t^{(k)}} \left(\frac{1}{s^2} | Y \right)}
\end{aligned}
\tag{4.10}
$$

We have an improper prior for ρ for the ease of posterior calculations. Thus, with the above model specification the joint posterior of $b, \tau^2, \sigma_e^2, \pi_0, \pi_1, \rho$ conditional on observed data is:

$$
\begin{aligned}
p(b,\tau^2,\sigma_e^2,\pi_0,\pi_1,\rho|Y,X) &= \frac{P(Y,b,\tau^2,\sigma_e^2,\pi_0,\pi_1,\rho|X)}{P(Y|X)} \\
&\propto P(Y|b,\tau^2,\sigma_e^2,\pi_0,\pi_1,\rho,X)P(b|\pi_0,\tau^2,\sigma_e^2,X)P(\sigma_e^2|\pi_1,X)P(\pi_0)P(\pi_1)P(\rho) \\
&= \frac{|V|^{-\frac{1}{2}}}{(2\pi)^{\frac{NT}{2}}} e^{-\frac{1}{2}\left(Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)^T V^{-1}\left(Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)} \\
&\times \prod_{g=1}^{p}(1-\pi_0)(2\pi)^{-\frac{N}{2}} e^{-\frac{1}{2}b_g^T b_g}\mathcal{I}_{[b_b\neq 0]} + \pi_0\delta_0(b_g) \\
&\times \prod_{g=1}^{p}\prod_{j=1}^{N}\left[(1-\pi_1)2\left(\frac{2\pi s^2}{w_{j+}}\right)^{-\frac{1}{2}} e^{-\frac{w_{j+}}{2s^2}\left[\tau_{jg}-\sum_{i\neq j=1}^{N}\frac{w_{ij}}{w_{i+}}\tau_{ig}\right]^2}\mathcal{I}_{[\tau_{jg}>0]} + \pi_1\delta_1(\tau_{jg})\right] \\
&\times (\sigma^2)^{-\alpha-1} e^{-\frac{\gamma}{\sigma^2}} \\
&\times \pi_0^{a_1-1}(1-\pi_0)^{a_2-1} \\
&\times \pi_1^{c_1-1}(1-\pi_1)^{c_2-1} \\
&\times t(s^2)^{-2} e^{-\frac{t}{s^2}} \\
&\times \frac{1}{2}\prod_{i=1}^{N}\frac{1}{\sqrt{1-\rho_i^2}}\mathcal{I}_{[-1\leq\rho_i\leq 1]}
\end{aligned}
$$

We can simulate an efficient block Gibbs sampler to simulate from the posterior distribution above. Define the following notations:

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

$$\mathbf{X}_{(g)} = (\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{g-1}, \widetilde{\mathbf{X}}_{g+1}, \ldots, \widetilde{\mathbf{X}}_p),$$

$$\beta_{(jg)} = (\beta_{11}, \beta_{21}, \ldots, \beta_{N1}, \ldots, \beta_{1g}, \ldots, \beta_{j-1g}, \beta_{j+1g}, \ldots \beta Ng, \ldots, \beta_{1p} \ldots, \beta_{Np}),$$

$$\mathbf{X}_{(jg)} = (\widetilde{x}_{11}, \ldots, \widetilde{x}_{N1}, \ldots, \widetilde{x}_{1g}, \ldots, \widetilde{x}_{j-1g}, \widetilde{x}_{j+1g} \ldots, \widetilde{x}_{Ng}, \ldots, \widetilde{x}_{1p}, \ldots, \widetilde{x}_{Np})$$

where $\mathbf{X}_g$ is the design matrix corresponding to β_g.

Detailed computation of the posterior full conditional is given below.

(**1**) We want to find the posterior distribution of b_g. Let

$$\begin{aligned} l_g = P(b_g = 0|rest) &= P(b_g = 0|Y, X, \tau_g^2, \pi_0, \pi_1, \rho, \sigma_e^2) \\ &= \frac{\pi_0 A}{\pi_0 A + (1-\pi_0)B} \end{aligned}$$

where $rem \equiv X, \tau_g^2, \pi_0, \pi_1, \rho, \sigma_e^2$,

$A = |V|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(Y - X_{(g)} A_{(g)}^{\frac{1}{2}} b_{(g)}\right) V^{-1} \left(Y - X_{(g)} A_{(g)}^{\frac{1}{2}} b_{(g)}\right)}$

and $B = |V|^{-\frac{1}{2}} \int_{b_g \neq 0} \frac{e^{-\frac{1}{2} b_g^T b_g} e^{-\frac{1}{2}\left(Y - \sum_{g=1}^p \widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)^T V^{-1} \left(Y - \sum_{g=1}^p \widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)}}{(2\pi)^{\frac{N}{2}}} db_g$

Thus we get, $l_g = \frac{\pi_0}{\pi_0 + (1-\pi_0)|\Sigma_g|^{\frac{1}{2}} e^{\frac{\mu_g^T \mu_g}{2}}}$

where, $\Sigma_g^{-1} = \mathbf{I}_N + A_g^{\frac{1}{2}} \widetilde{X}_g^T V^{-1} \widetilde{X}_g A_g^{\frac{1}{2}}$ and $\mu_g = \Sigma_g A_g^{\frac{1}{2}} \widetilde{X}_g^T V^{-1} \left(Y - X_{(g)} A_{(g)}^{\frac{1}{2}} b_{(g)}\right)$

From the posterior full conditionals we have, $b_g|rest \sim l_g \delta_0(b_g) + (1 - l_g)\mathcal{N}_N(\mu_g, \Sigma_g)$.

(**2**)

$$P(\tau_{jg}|rest) \propto e^{-\frac{1}{2}\left(Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)^T V^{-1}\left(Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right)}$$

$$\times \prod_{g=1}^{p}\prod_{j=1}^{N}\left[(1-\pi_1)2\left(\frac{2\pi s^2}{w_{j+}}\right)^{-\frac{1}{2}} e^{-\frac{w_{j+}}{2s^2}\left[\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{i+}}\tau_{ig}\right]^2}\mathcal{I}_{[\tau_{jg}>0]} + \pi_1\delta_1(\tau_{jg})\right]$$

$$\propto e^{-\frac{1}{2}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]^T V^{-1}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]}$$

$$\times (1-\pi_1)2\left(\frac{2\pi s^2}{w_{j+}}\right)^{-\frac{1}{2}} e^{-\frac{w_{j+}}{2s^2}\left[\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{i+}}\tau_{ig}\right]^2}\mathcal{I}_{[\tau_{jg}>0]}$$

$$+ e^{-\frac{1}{2}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]^T V^{-1}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]}\pi_1\delta_1(\tau_{jg})$$

Thus, we see that,

$$\tau_{jg}|rest \propto q_{jg}\delta_1\tau_{jg} + (1-q_{jg})\mathcal{N}^+(u_{jg}, v_{jg}^2), \quad g = 1,\dots p; j = 1,\dots N$$

where

$$q_{jg} = P(\tau_{jg}=0|rest)$$

$$= P(\tau_{jg}=0|Y, X, b, \pi_0, \pi_1, \rho, \sigma_e^2)$$

$$= \frac{P(Y|\tau_{jg}=0, rem1)P(\tau_{jg}=0|rem1)}{P(Y|\tau_{jg}=0, rem1)P(\tau_{jg}=0|rem1) + \int_{\tau_{jg}\neq)} P(Y|\tau_{jg}\neq 0, rem1)P(\tau_{jg}\neq 0|rem1)}$$

where $rem1 = X, b, \pi_0, \pi_1, \rho, \sigma_e^2$.

Now, $P(\tau_{jg}=0|rem1) = \pi_0$

$P(Y|\tau_{jg}=0,rem1)=e^{-\frac{1}{2}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]^T V^{-1}\left[Y-\sum_{g=1}^{p}\widetilde{X}_g A_g^{\frac{1}{2}} b_g\right]}$ and

$$\begin{aligned}
&\int_{\tau_{jg}\neq)} P(Y|\tau_{jg}\neq 0,rem1)P(\tau_{jg}\neq 0|rem1)\\
&=\int_{\tau_{jg}\neq)} e^{-\frac{1}{2}(Y-XA^{\frac{1}{2}}b)^T V^{-1}(Y-XA^{\frac{1}{2}}b)}\\
&\times(1-\pi_1)2\left(\frac{2\pi s^2}{w_{j+}}\right)^{-\frac{1}{2}} e^{-\frac{w_{j+}}{2s^2}\left[\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{i+}}\tau_{ig}\right]^2}\mathcal{I}_{[\tau_{jg}>0]}d\tau_{jg}
\end{aligned}$$

After rigorous computation, we get,

$$q_{jg}=\frac{\pi_1}{\pi_1+2(1-\pi_1)\left(\frac{s^2}{w_{j+}}\right)^{\frac{1}{2}} v_{jg} e^{-\frac{w_{j+}}{2s^2}\left(\sum_{i\neq j}\frac{w_{ij}}{w_{i+}}\tau_{jg}\right)^2+\frac{\mu_{jg}^2}{2v_{jg}^2}}\Phi\left(\frac{\mu_{jg}}{v_{jg}}\right)}$$

and $v_{jg}^2=\left(\frac{w_{j+}}{s^2}+b_{jg}^2\widetilde{x}_{jg}^T V^{-1}\widetilde{x}_{jg}\right)^{-1}$

$\mu_{jg}=v_{jg}^2\left[\frac{w_{j+}}{s^2}\sum_{i\neq j}\frac{w_{ij}}{w_{j+}}\tau_{ig}-\left(Y-X_{(jg)}\beta_{(jg)}\right)^T V^{-1}x_{jg}b_{jg}\right]$

(**3**)

$$\begin{aligned}
P(\sigma_e^2|rest)&\propto|V|^{\frac{1}{2}}e^{-\frac{1}{2}(Y-X\beta)^T V^{-1}(Y-X\beta)}(\sigma_e^2)^{-\alpha-1}e^{-\frac{\gamma}{\sigma_e^2}}\\
&\propto(\sigma_e^2)^{-\frac{NT}{2}-\alpha-1}e^{-\frac{1}{2\sigma_e^2}(Y-X\beta)^T\widetilde{V}^{-1}(Y-X\beta)-\frac{\gamma}{\sigma_e^2}}\\
&=(\sigma_e^2)^{-\left(\frac{NT}{2}+\alpha\right)-1}e^{-\frac{1}{\sigma_e^2}\left[\frac{(Y-X\beta)^T\widetilde{V}^{-1}(Y-X\beta)}{2}+\gamma\right]}
\end{aligned}$$

Therefore, $\sigma_e^2|rest \sim IG\left(\frac{NT}{2}+\alpha, \frac{(Y-X\beta)^T\tilde{V}^{-1}(Y-X\beta)}{2}+\gamma\right)$ (**4**)

$$P(\pi_0|rest) \propto \pi_0^{a_1-1}(1-\pi_0)^{a_2-1}\left[\prod_{g=1}^{p}(1-pi_0)(2\pi)^{-\frac{N}{2}}e^{-\frac{1}{2}b_g^T b_g}I_{[b_g\neq 0]}+\pi_0\delta_0(b_g)\right]$$
$$\propto \pi_0^{a_1-1}(1-\pi_0)^{a_2-1}(1-\pi_0)^{[p-(\#b_g\neq 0)]}\pi_0^{(\#b_g=0)}$$

Therefore, $\pi_0|rest \sim Beta(a_1+(\#b_g=0), a_2+p-(\#b_g\neq 0))$ (**5**)

$$P(\pi_1|rest) \propto \pi_1^{c_1-1}(1-\pi_1)^{c_2-1}(1-\pi_1)^{(\#\tau_{jg}\neq 0)]}\pi_1^{(\#\tau_{jg}=0)}$$

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

(**6**)

$$P(s^2|rest) \propto (s^2)^{-2}e^{-\frac{t}{s^2}}(s^2)^{-\frac{1}{2}(\#\tau_{jg}\neq 0)}e^{-\sum_{g=1}^{p}\sum_{j=1}^{N}\frac{w_{j+}}{2s^2}\left(\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{j+}}\tau_{ij}\right)^2}$$
$$= (s^2)^{-\left(\frac{1}{2}(\#\tau_{jg}\neq 0)+1\right)-1}e^{-\frac{1}{s^2}\left[t+\sum_{g=1}^{p}\sum_{j=1}^{N}\frac{w_{j+}}{2}\left(\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{j+}}\tau_{ij}\right)^2\right]}$$

Therefore, $s^2|rest \sim IG\left(\frac{1}{2}(\#\tau_{jg}\neq 0), t+\sum_{g=1}^{p}\sum_{j=1}^{N}\frac{w_{j+}}{2}\left(\tau_{jg}-\sum_{i\neq j}\frac{w_{ij}}{w_{j+}}\tau_{ij}\right)^2\right)$

(**7**)

We have,

$$(Y-X\beta) = \begin{pmatrix} y_{11}-\sum_{g=1}^{p}x_{1g}\beta 1g \\ y_{21}-\sum_{g=1}^{p}x_{1g}\beta 2g \\ \cdots \\ y_{N1}-\sum_{g=1}^{p}x_{1g}\beta Ng \\ y_{12}-\sum_{g=1}^{p}x_{2g}\beta 1g \\ \cdots \\ y_{N2}-\sum_{g=1}^{p}x_{2g}\beta Ng \\ \cdots \\ y_{1T}-\sum_{g=1}^{p}x_{Tg}\beta 1g \\ \cdots \\ y_{NT}-\sum_{g=1}^{p}x_{Tg}\beta Ng \end{pmatrix} = \begin{pmatrix} c_{11} \\ c_{21} \\ \cdots \\ c_{N1} \\ c_{12} \\ \cdots \\ c_{N2} \\ \cdots \\ c_{1T} \\ \cdots \\ c_{NT} \end{pmatrix}$$

Then $P(\rho_1,\ldots,\rho_N|rest) \propto (Y-X\beta)^T\widetilde{V}^{-1}(Y-X\beta)\left(\prod_{i=1}^N \frac{1}{\sqrt{1-\rho_i^2}}\mathcal{I}_{[-1\le\rho_i\le 1]}\right)$

$= (\, c_{11} \cdots c_{N1}\, c_{12} \cdots c_{N2} \cdots c_{1T} \cdots c_{NT}\,)$

$$\times \begin{bmatrix} \begin{pmatrix} 1 & \ldots & 0 \\ \dot{0} & \ldots & \dot{1} \end{pmatrix} & \begin{pmatrix} -\rho_1 & \ldots & 0 \\ \dot{0} & \ldots & -\rho_N \end{pmatrix} & 0 & \ldots & 0 \\ \begin{pmatrix} -\rho_1 & \ldots & 0 \\ \dot{0} & \ldots & -\rho_N \end{pmatrix} & \begin{pmatrix} 1+\rho_1^2 & \ldots & 0 \\ \dot{} & \dot{} & \dot{} \\ 0 & \ldots & 1+\rho_N^2 \end{pmatrix} & \begin{pmatrix} -\rho_1 & \ldots & 0 \\ \dot{0} & \ldots & -\rho_N \end{pmatrix} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \begin{pmatrix} -\rho_1 & \ldots & 0 \\ \dot{0} & \ldots & -\rho_N \end{pmatrix} & \begin{pmatrix} 1 & \ldots & 0 \\ \dot{0} & \ldots & \dot{1} \end{pmatrix} \end{bmatrix}$$

$$\times \begin{pmatrix} c_{11} \\ c_{21} \\ \cdots \\ c_{N1} \\ c_{12} \\ \cdots \\ c_{N2} \\ \cdots \\ c_{1T} \\ \cdots \\ c_{NT} \end{pmatrix} \times \left(\prod_{i=1}^N \frac{1}{\sqrt{1-\rho_i^2}}\mathcal{I}_{[-1\le\rho_i\le 1]}\right)$$

Collecting $\rho_i, \quad i=1,\ldots,N$ terms, we get,

$$-2\rho_i\left[c_{i1}c_{i2}+c_{i3}c_{i2}+c_{i4}c_{i3}+\cdots+c_{iT}c_{iT-1}\right]+\rho_i^2\left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]$$

$$\propto \left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]\times$$

$$\left[\rho_i^2-2\rho_i\frac{\left[c_{i1}c_{i2}+c_{i3}c_{i2}+c_{i4}c_{i3}+\cdots+c_{iT}c_{iT-1}\right]}{\left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]}+\frac{\left[c_{i1}c_{i2}+c_{i3}c_{i2}+c_{i4}c_{i3}+\cdots+c_{iT}c_{iT-1}\right]^2}{\left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]}\right]$$

$$\therefore \rho_i|rest \propto \frac{1}{\sigma_e^2}e^{-\frac{1}{2\sigma_e^2}\left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]\left[\rho_i-\frac{\left[c_{i1}c_{i2}+c_{i3}c_{i2}+c_{i4}c_{i3}+\cdots+c_{iT}c_{iT-1}\right]}{\left[c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2\right]}\right]^2}$$

$\rho_i|rest \sim N(\mu_{\rho_i}, v_{\rho_i}^2) \quad i=1,\ldots,N$ where

$$\mu_{\rho_i}=\frac{c_{i1}c_{i2}+c_{i3}c_{i2}+c_{i4}c_{i3}+\cdots+c_{iT}c_{iT-1}}{c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2}$$

$$v_{\rho_i}^2=\frac{\sigma_e^2}{c_{i2}^2+c_{i3}^2+\cdots+c_{iT-1}^2}$$

4.4 Computational Algorithm

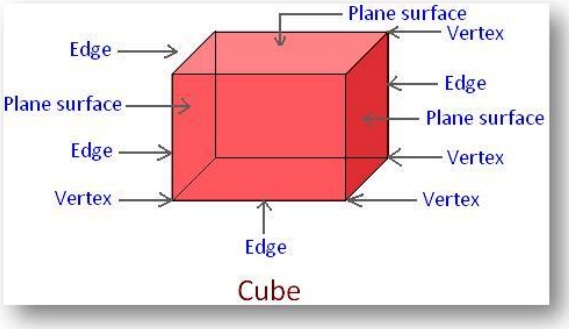
Voxel level brain data for a single subject has a very complicated structure since the size of data is huge, We have to deal with a data size of more than 20 million. Our covariance matrix is of the order of $NT \times NT$. Obtaining posterior distribution requires inversion of this matrix. It is almost impossible to deal with such large matrix inversions so we need to come up with efficient algorithms that are able to handle this data size. Computational efficiency in modern times is an inevitable step towards model building. Building a model is not sufficient if we cannot find a way of proper implementation within humanly controlled time periods.

Apart from the covariance matrix, adjacency matrix and A_g is also of high order (i.e. $N \times N$. Even storage of such matrices require a supercomputer. We have used inverse algorithms and matrix multiplication tricks to bypass the use of too many memory resources. Note that, the variance-covariance matrix in our setup is a block diagonal matrix for every $N \times N$ block. It has an autoregressive structure and as seen in the previous section, we do not need to invert the matrix using matrix algebra since we already know the elements of the inverted matrix. Thus, we build the variance-covariance matrix using the inversion of autocovariance matrix by Siddiqui, 1958 and just place all the elements in their true location. We have extensively used "Matrix" package in R for the computation. This package saves sparse matrices with "sparseMatrix" function by allocating it to a much smaller memory space. The inverse of variance-covariance matrix is sparse so we use this advantage while writing our code.

4.4.1 Creating the Adjacency Matrix

We require an adjacency matrix for specifying the adjacency of voxels in the brain. This is used when we specify weights for spatial dependence through the variance of amplitudes. For a 2D lattice, a voxel has four neighbors, sharing boundary with each of its face. For a 3D cube, a voxel has 26 neighbors. Figure 4.2 (Source: http://www.math-only-math.com/common-solid-figures.html) shows there are 8 vertices, 6 plane surfaces and 12 edges in a cube making te maximum number of neighboring voxels 26.

Figure 4.2: Illustration of neighborhood of a cube



For 2D lattice, we have two coordinates a and y. If two x-values or y-values have absolute difference of their x or y axis coordinate as 1 respectively then we call them neighbors. If the absolute difference is larger in either x or y then they are not neighbors.

In 3 dimensional setup, we have 3 coordinates x, y and z. Here, we fix a coordinate and check for the absolute difference of the other two to be 1. Then we assign weight 1 meaning neighbors to the voxel with respect to the focal voxel. This algorithm is straightforward but the adjacency matrix is huge for large voxel size. Note that, adjacency matrix is sparse as most of the elements will be 0.

4.4.2 Matrix Tricks

The largest order of matrix we need to invert in our model is $N \times N$. Since this is a variance-covariance matrix of the posterior of β_g, it is positive definite. We use Cholesky decomposition instead of traditional matrix inverse. For generating $\tau_g, \quad g = 1, \ldots, p$ we carefully perform matrix multiplication so that it saves computation time. For example, we need to evaluate $\left(Y - X_{(jg)}\beta_{(jg)}\right)^T V^{-1} x_{jg} b_{jg}$ for the posterior distribution of τ_g. Evaluating $\left(Y - X_{(jg)}\beta_{(jg)}\right)^T$ means finding NT elements N times making the computation time N^2T , so we calculate $(Y - X\beta)^T$ once which uses NT computation time. Then for N times we calculate $X_{(jg)}\beta_{(jg)}$ and subtract from previously stored value of $(Y - X\beta)^T$. This modification uses $NT + N$ computation time. Using such matrix algebra tricks we are able to bring down the Gibb's sampler run time considerably.

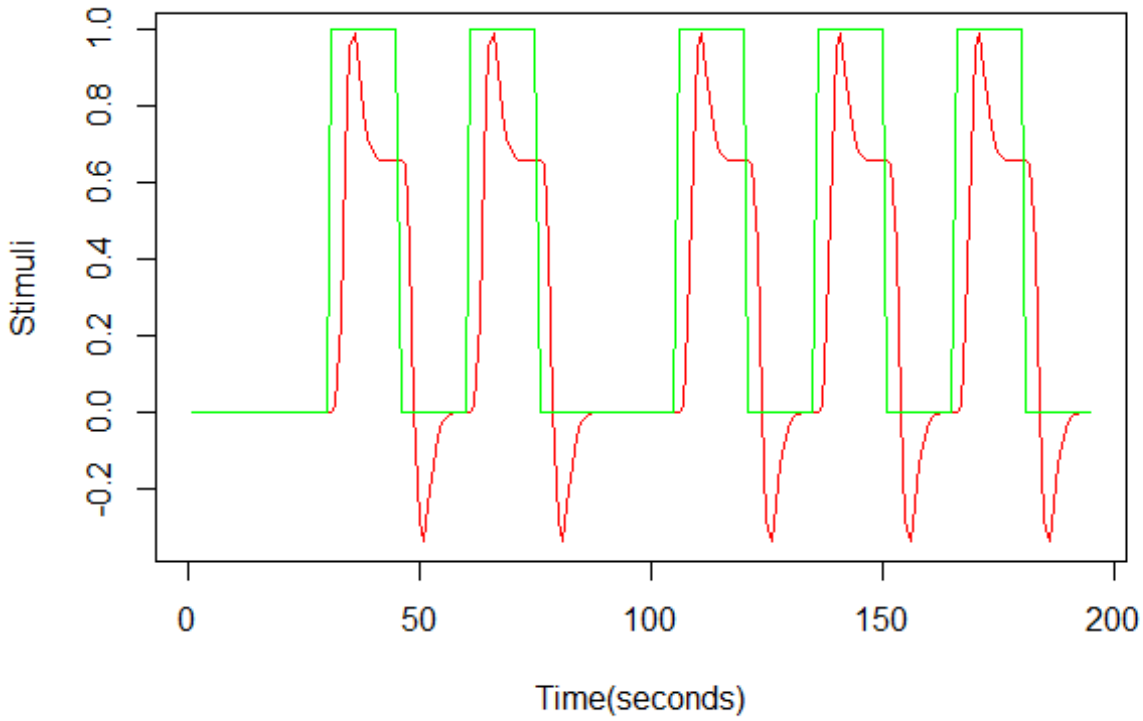
4.5 Simulation

To understand how our proposed Bayesian hierarchical, sparsity-inducing model works, we need to perform a simulation study. In both the previous two simulation studies we have simulated the datasets by assigning true parameter values and generating a model from the concerned distribution. This works fine and was easy to build since our responses were independent. Now we have dependent responses. There are ways of generating dependent responses but what is unique in this chapter is that we want a simulated dataset that represents an fMRI dataset. This is not an easy task since a fMRI dataset has very specific patterns. Welvaert *et.al.*(*2011*) created an R package called "neuRosim" that produces fMRI data replicating real fMRI data. We have used this package to generate our data in this section.

4.5.1 Generating Simulated fMRI Data

We generate a stimuli with a totaltime of 390 seconds and repetition time of 2 seconds. Thus, there are 195 time series points in the BOLD signals. The stimuli is a boxcar function which is then convolved with a double-gamma density to make it continuous. Figure 4.3 plots the simulated stimuli. The stimuli onset times are at 60, 120, 210, 270 and 330 seconds.

Figure 4.3: Stimuli convolved with double-gamma density (red) and the corresponding box-car function (green)



We generate 4 regions of activation on a 2D lattice at coordinates (10,42), (25,35), (10,10) and (40,31) with radii 3,4,3 and 3 respectively. A mixture noise is used with 0.4 and -0.25 as autocorrelation for simulating 2,500 voxels. A signal-to-noise ratio of 4.5 is used.

4.5.2 Analysis Result of Simulation Study

Left-hand panel of Figure 4.4 shows us the activated regions in our simulated 2D lattice. Our aim is to analyze the simulated dataset of 50×50 voxels on a 2D lattice for 195 timepoints. Our Bayesian hierarchical bilevel selection method is able to detect the true activated regions with a very high accuracy.

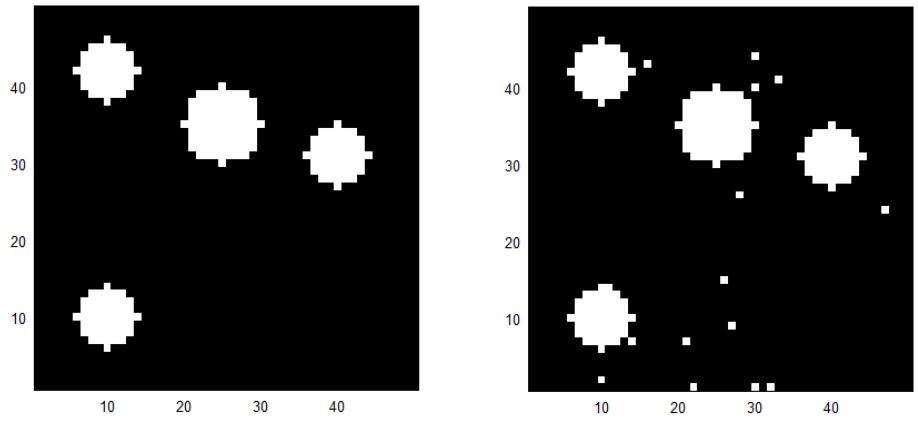
To assess the accuracy of our method we compare it to the results of Musgrove *et.al.* (2016) for the case where they have analyzed the entire brain and parcellated brain in their simulation study. The comparison is valid since we both the datasets simulated 2,500 voxels on 2D lattice with 195 timepoints. The stimuli have for both the studies have total duration and onset times to be same. Table 4.1 gives us the results. The median accuracy of our method is 99.6% which is better for both the methods of Musgrove *et.al.* (2016). The false positives are also better for our approach.

Table 4.1: Median accuracy (minimum and maximum) of correct classification and false positives over 10 simulated datasets

Method	Accuracy(%)	False Positive
Bayesian hierarchical bilevel selection	99.6 (98.9,100)	0.006 (0,0.008)
Musgrove *et.al.* (2016) Full dataset	98.7 (96.0, 99.3)	0.015 (0.007, 0.044)
Musgrove *et.al.* (2016) Parcellated dataset	99.4 (98.9, 99.7)	0.007 (0.003, 0.011)

Figure 4.4, right-hand side panel, shows us the selected voxels in white. Thus, it is evident from the results that our proposed method is an improved variable selection method. Musgrove *et.al.* (2016) have parcellated the brain to parallely run the analysis for significant time reduction. In our real data analysis section, we use a two stage procedure thus reducing our computation time significantly.

Figure 4.4: True and estimated binary map



4.6 Single-Subject fMRI Data Analysis

4.6.1 Data Acquisition

A healthy college student from Michigan State University volunteered to participate in this study and signed consent forms approved by the Michigan State University Institutional Review Board. The experiment was conducted on a 3T GE Signa HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil.

An fMRI dataset was collected on a visual stimulation condition with a scene-object fMRI paradigm. The parameters for the fMRI scan were: gradient-echo EPI, 36 contiguous 3-mm axial slices in an interleaved order, time of echo (TE) = 27.7 ms, time of repetition (TR) = 2500 ms, flip angle = 80°, field of view (FOV) = 22 cm, matrix size = 64×64, ramp sampling, and with the first four data points discarded. Each volume of images were acquired 192 times (8 min) while a subject was presented with 12 blocks of visual stimulation after an initial 10 s "resting" period. In a predefined randomized order, scenery pictures were

presented in 6 blocks and objects pictures were presented in other 6 blocks. All pictures were unique. In each block, 10 pictures were presented continuously for 25 s (2.5 s for each picture), followed with a 15 s baseline condition (a white screen with a black fixation cross at the center). The subject needed to press his/her right index finger once when the screen was switched from the baseline to picture condition. Stimuli were displayed in color in full screen on a 1024×768 32-inch LCD monitor (Salvagione Design, Sausalito, CA) placed at the back of the magnet room. The LCD subtended $10.2^\circ \times 13.1^\circ$ of visual angle. After the above functional data acquisition, high-resolution volumetric T1-weighted spoiled gradient-recalled (SPGR) images with cerebrospinal fluid suppressed were obtained to cover the whole brain with 120 1.5-mm sagittal slices, 8° flip angle and 24 cm FOV. These images were used to identify anatomical locations.

4.6.2 fMRI Data Pre-processing and Analysis

All stimulus fMRI data pre-processing and analysis for each subject were conducted with AFNI software (Cox, 1996) as described in Henderson et al. (Henderson, Zhu et al. 2011). Essentially, slice-timing correction and rigid-body motion correction were carried. Spatial blurring with a full width half maximum of 4 mm was applied to reduce random noise. Multiple linear regressions (using the "3dDeconvolve" routine in AFNI) were applied on a voxel-wise basis to find the magnitude change when each picture condition was presented, followed with general linear tests to find the statistical significances between stimulus conditions.

We apply a two stage method- first we reduce the number of voxels based on p-value analysis for each voxel, ignoring spatial correlation and then apply our method in the second stage. The first step is a data processing step since many voxel data that are captured in the

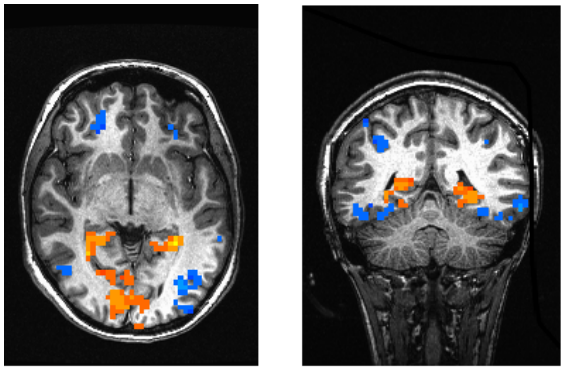
MRI scan are outside the brain and the pvalue based analysis can easily identify the non-affected voxels. We have 64×64 voxels and 36 slices of the brain. Each of these voxels were captured at 192 timepoints. This makes our data size of the order of 28 million. We have applied voxel by voxel regression analysis and retained voxels that had a p-value $< 5 \times 10^{-4}$. This reduces voxel size to 6118. Now we have data size in the order of 1 million.

In the second step, we apply our proposed Bayesian spike and slab bi-level selection to the reduced data. the subject was shown two stimuli, a scenery and an object. From the first step, we know that 6118 voxels represent the regions of interest in the brain that are activated in some way due to visual stimulation. Thus, we expect that the both the stimuli and the most activated voxels are selected by our model. Henderson *et.al.* (2007) conducted a similar study with indoor and outdoor sceneries and faces as stimuli. They conclude that there is activation in posterior parahippocampal cortex (pPHC), including parahippocampal place area(PPA) for sceneries. The faces stimuli activates fusiform gyrus and amygdala. Epstein *et.al.* (1999, 2003); Epstein and Kanwisher (1998) have shown that scenes activate PPA over faces, single objects and object arrays.

Our results show the activation of PPA in the scene - object contrast. Figure 4.5 shows the activated areas in orange. The figure illustrates that scene preferentially activates PPA over object. This finding is in lines with Epstein *et.al.* (1999, 2003) and Henderson *et.al.* (2007). Blue regions show activation by object stimuli indicating activation fusiform gyrus and amygdala.

The scene and object activation of brain regions can be looked at separately too (as in Figure 4.6). Figure 4.6 shows that both scene and object activate similar regions of interest in the brain that are traditionally related to visual activation. The coronal view specifically show that object preferentially activates the fusiform gyrus whereas scenes activate pPHC

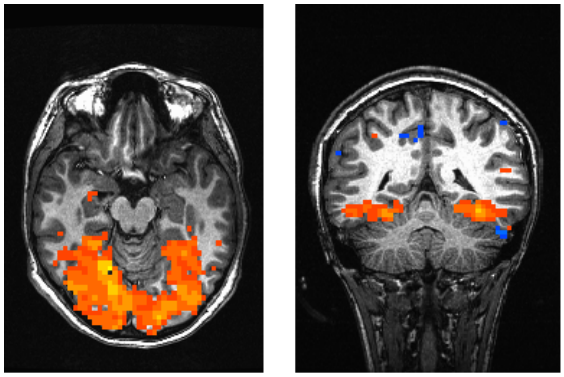
Figure 4.5: Scene minus object contrast activation



Axial view of scene-object contrast

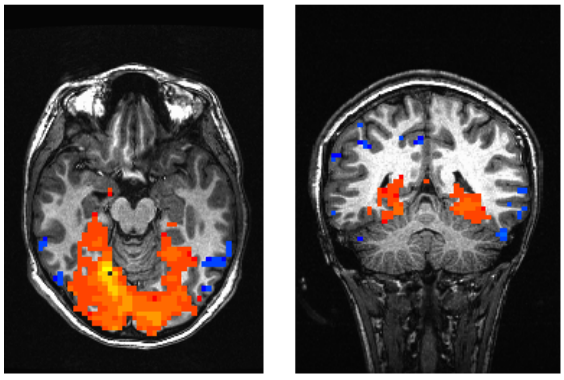
Coronal view of scene-object contrast

Figure 4.6: Scene and object activation



Axial view of object stimuli

Coronal view of object stimuli



Axial view of scene stimuli

Coronal view of scene stimuli

more than objects. From our results, we see that the activation behavior of objects is very similar to that of a faces stimuli (Henderson *et.al.* (2007)).

Our model is able to select both the stimuli that activate the regions of interest associated with visual cortex. The model does a bi-level selection in selecting not just the stimuli but the activated voxels. The selection keeps the spatial clustering of the voxels intact by choosing clustered voxels in the visual cortex. Henderson *et.al.* (2007) performed their analysis without considering any spatial correlation. Although, our results are aligned with theirs, it is difficult to directly judge a better method since the stimuli used and objective of the studies are different. Our method has an advantage because we consider spatial information which they have completely ignored.

4.7 Discussion

In this chapter we have extended our model from chapter 3 to introduce bi-level selection of covariates. The idea is to select a significant group and few significant levels within that group. This has further been convoluted with a spatiotemporal structure where we introduce temporal correlation in the likelihood and spatial correlation in the prior of β. This novel method of Bayesian hierarchical modeling is applied to a single subject fMRI data. Our goal here is to select the activated voxels in the presence of an external stimuli. Alongside the proposed variable selection method, we introduce BOLD activation in fMRI tests. We want to find how is the brain affected (activated) by a stimuli; this will lead medical professionals in understanding the neuronal activity in human brains which can further help treat various diseases or find root causes of certain human behavior.

A thorough simulation study compares our results with those of Musgrove *et.al.* (2016).

Our method is able to give a lower false positive rate indicating that our method performs better. Our simulation was conducted in a 2D setup. We extend our method to 3D real data, acquired from Department of radiology, MSU. Our method correctly identifies the activation of the visual cortex in the presence of scene and object stimuli. It shows that scene preferentially selects PPA regions of interest while objects select the fusiform gyrus and amygdala. These results align with those of Epstein *et.al.* (1999, 2003). Thus, we are able to verify that our novel approach of variable selection has a very relevant application in fMRI data. The relevance of understanding fMRI data is immense and handling of huge data sizes is a challenge. We overcome this by introducing a two step procedure. the first step filters the data voxel by voxel and then incorporates the proposed method for variable selection. Castruccio *et.al.* (2016) have used a similar 3 step approach with region of interest(ROI) information added to voxel level data. Our data does not include ROI information but does a great job selecting the truly activated voxels. thus, we use a less complicated model than Castruccio *et.al.* (2016) but a more statistically relevant model than Henderson *et.al.* (2007) to identify BOLD activation in fMRI data.

Chapter 5

Future Work

This dissertation presents three varied applications of statistical methodology in fields ranging from marketing research to brain image data. Our proposed methods are state of the art approaches for dealing with the analyzed data. Our main focus has been variable selection. When the number of parameters become larger than the number of observations then ordinary regression method fails. This has led to the development of numerous regularized modeling approaches.

In our marketing research data application, we have used frequentist style variable selection methods, namely, LASSO and elastic net. For real data analysis, we convert 6 categories of the response in 6 different sets of binary variables - one vs others. Our comparative study has been performed on linear and binary outcome. A very interesting future work will be to build a model for the multi-category response variable. A multinomial logistic regression with LASSO and/or elastic net regularization can be used.

Next, we have performed variable selection on binary response (Alzheimer's disease or normal control) in a Bayesian group lasso setup. We propose a median thresholding posterior estimator of βs and use spike and slab type prior. A logit link used for this setup works well in terms of the true and false positive rates of prediction. A high AUC under the ROC curve also indicates that our method performs competitively. For further research, it will be interesting to explore the impact of a probit link for this model setup. The logit link helps us calculate

relative risks and odd ratios for the significant ROIs, however, for another interpretation of βs, we can explore the use of a probit link in our model. There is a scope of comparing models using two different link functions but essentially, the same model. Another potential variation in the model can be approximation of intractable posterior of β by a numerical method (e.g. Laplace approximation). This proposed method has a tremendous scope of future work, given the numerous trajectories of research.

Our last application is based on a giant dataset of brain voxels in 3d image for hundreds of timepoints. We have used a spatiotemporal modeling to perform variable selection by extending our Bayesian group lasso approach to select a stimuli administered for monitoring brain activity as well brain voxels. The biggest challenge with this data is its size. We perform analysis by filtering the more active brain regions by a preliminary analysis to reduce data size. In the second step, we apply our Bayesian variable selection method. Further research should be continued to find out a method that is able to handle the entire dataset at once. This finding will be a great step ahead in today's challenge of handling big data. Our model does a great job giving a low false positive rate and is shown to better than the method of Musgrove et.al. (2016). This piece of work builds an associative model to find out the most active voxels to further identify relation of neuronal dysfunction and a disease. We can extend this model to a predictive model where one may want to predict the state of a patient given the activity in the selected voxels. This requires building a model, inclusive of variable selection phase, on a train dataset and validating it using a test dataset.

Our contribution through the novelty of application is tremendous and has carved a way for intriguing future research work.

APPENDIX

Appendix

A. Gibbs Sampler

The full posterior conditional distributions are as follows:

$$
\begin{aligned}
p(\beta, \tau^2, \pi_0 | \mathbf{Y}, \mathbf{X}) \propto & \prod_{i=1}^{n} \left[\left(\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)^{y_i} \left(\frac{1}{1 + e^{x_i^T \beta}} \right)^{1 - y_i} \right] \\
& \times \prod_{g=1}^{G} \left[(1 - \pi_0)(2\pi\tau_g^2)^{-\frac{m_g}{2}} e^{-\frac{\beta_g^T \beta_g}{2\tau_g^2}} I_{[\beta_g \neq 0]} + \pi_0 \delta_0(\beta_g) \right] \\
& \times \prod_{g=1}^{G} (\lambda^2)^{\frac{m_g+1}{2}} (\tau_g^2)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2 \tau_g^2}{2}} \\
& \times \ {\pi_0}^{a-1} (1 - \pi_0)^{b-1} \qquad (.1)
\end{aligned}
$$

Let $\beta_{(g)}$ and $X_{(g)}$ be defined as follows:

$$\beta_{(g)} = (\beta_1^T, \ldots, \beta_{g-1}^T, \beta_{g+1}^T, \ldots, \beta_G^T),$$

$$\mathbf{X}_{(g)} = (\mathbf{X}_1, \ldots, \mathbf{X}_{g-1}, \mathbf{X}_{g+1}, \ldots, \mathbf{X}_G)$$

where $\mathbf{X}_g$ is the design matrix corresponding to β_g. Detailed computation of the posterior full conditional distributions is given below: (**1**) Note that, $e^{x_i^T \beta} = e^{x_{i1}^T \beta_1 + \cdots + x_{iG}^T \beta_G}$. If, $\beta_g = 0$ then, $e^{x_i^T \beta} = e^{x_{i1}^T \beta_1 + \cdots + x_{iG}^T \beta_G . 0 + \cdots + x_{iG}^T \beta_G} = e^{x_{i(g)}^T \beta_{(g)}}$ where $x_{i(g)}$ corresponds

$\beta_{(g)}$.

$$\begin{aligned} p(\beta_g = 0|\text{rest}) &= p(\beta_g = 0|\mathbf{Y}, \mathbf{X}, \tau_g^2, \pi_0) \\ &= \frac{p(\beta_g = 0, \mathbf{Y}|\mathbf{X}, \tau_g^2, \pi_0)}{\int_{\beta_g} p(\beta_g, \mathbf{Y}|\mathbf{X}, \tau_g^2, \pi_0) d\beta_g} \\ &= \frac{A\pi_0}{A\pi_0 + B(1-\pi_0)} \end{aligned}$$

where

$$\pi_0 = p(\beta_g = 0|\tau_g^2, \pi_0)$$

$$A = p(\mathbf{Y}|\beta_g = 0, \mathbf{X}, \tau_g^2, \pi_0) = \prod_{i=1}^{n} \frac{e^{y_i x_{i(g)}^T \beta(g)}}{1 + e^{x_{i(g)}^T \beta(g)}}$$

$$\begin{aligned} B &= \int_{\beta_g \neq 0} p(\mathbf{Y}|\beta_g \neq 0, \mathbf{X}, \tau_g^2, \pi_0) d\beta_g \\ &= \int_{\beta_g \neq 0} \left[\prod_{i=1}^{n} \frac{e^{y_i x_{i(g)}^T \beta(g)}}{1 + e^{x_{i(g)}^T \beta(g)}} \right] (2\pi\tau_g^2)^{-\frac{m_g}{2}} e^{-\frac{\beta_g^T \beta_g}{2\tau_g^2}} d\beta_g \end{aligned}$$

The integral in term B is complicated so we approximate it with Monte- Carlo approximation for each $i = 1, \ldots, n$. Thus, we approximate the function $\prod_{i=1}^{n} \frac{e^{y_i x_{i(g)}^T \beta(g)}}{1+e^{x_{i(g)}^T \beta(g)}}$ by drawing $i.i.d$ samples of β_g from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\tau_g^2 \mathbf{I}_{m_g}$.

Let $l_g = p(\beta_g = 0|\text{rest})$. From the posterior full conditionals, we have that

$$\beta_g|\text{rest} \sim (1 - l_g)F + l_g \delta_0(\beta_g) \tag{.2}$$

where F is some distribution whose form is unattainable. By collecting the terms of β_g from (.1) we do not get a closed form for the distribution of $\beta_g|$rest. But, we get (.2) and use the Metropolis algorithm to draw samples from $\beta_g|$rest whenever $\beta_g \neq 0$.

The Metropolis algorithm:

The Metropolis algorithm is a method of drawing samples from a posterior distribution when the posterior distribution does not have a closed form. Suppose we have a working collection of $(\theta^{(1)}, \ldots, \theta^{(s)})$ to which we would like to add a new value $\theta^{(s+1)}$. Let us consider adding a value θ^* to the set that is in the vicinity of $\theta^{(s)}$. Let,

$$r = \frac{p(\theta^{(*)}|y)}{p(\theta^{(s)}|y)}$$

Let

$$\theta^{s+1} = \begin{cases} \theta^*, \text{with probability min(r,1)} \\ \\ \theta^s, \text{with probability 1-min(r,1)} \end{cases}$$

Here θ^* is sampled from a symmetric distribution. Sample $u \sim$ Uniform $(0, 1)$ and set $\theta^{s+1} = \theta^*$ if $u < r$, otherwise set $\theta^{s+1} = \theta^s$.

We sample θ^* from a normal distribution with a proposed mean and variance. Then the Metropolis ratio is:

$$\begin{aligned} r &= \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} \\ &= \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})} \\ &= \frac{\prod_{i=1}^{n} \text{dbinom}(y_i, \theta^*)\text{dnorm}(\theta^*, \mu, \tau)}{\prod_{i=1}^{n} \text{dbinom}(y_i, \theta^{(s)})\text{dnorm}(\theta^{(s)}, \mu, \tau)} \end{aligned}$$

Thus for our problem,

$$\log \text{ r } = \sum_{i=1}^{n} \Big[\log \text{ dbinom}(y_i, \beta_g^*) - \log \text{ dbinom}(y_i, \beta_g^{(s)})\Big] + \sum_{i=1}^{m_g} \Big[\log \text{ dnorm}(\beta_g^*) - \log \text{ dnorm}(\beta_g^{(s)})\Big]$$

Using this algorithm we approximate F in (.2) where l_g is also approximated by Monte-Carlo method. Thus, (.2) generates $\beta_g|\text{rest}$.

(**2**)

$$p(\tau_g^2|\text{rest}) \propto (\lambda^2)^{\frac{m_g+1}{2}} (\tau_g^2)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2\tau_g^2}{2}} \times \left[(1-\pi_0)(2\pi\tau_g^2)^{-\frac{m_g}{2}} e^{-\frac{\beta_g^T\beta_g}{2\tau_g^2}} I_{[\beta_g \neq 0]} + \pi_0\delta_0(\beta_g)\right]$$

If $\beta_g = 0$:

$$\begin{aligned} p(\tau_g^2|\text{rest}) &\propto (\lambda^2)^{\frac{m_g+1}{2}} (\tau_g^2)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2\tau_g^2}{2}} \pi_0\delta_0(\beta_g) \\ &\propto (\tau_g^2)^{\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2\tau_g^2}{2}} \end{aligned}$$

Letting $\tau_g^2 = \frac{1}{\alpha_g^2}$ we have,

$$p(\alpha_g^2|\text{rest}) \propto (\alpha_g^2)^{-\frac{m_g+1}{2}-1} e^{-\frac{\lambda^2}{2\alpha_g^2}}$$

If $\beta_g \neq 0$:

$$p(\tau_g^2|\text{rest}) \propto (\lambda^2)^{\frac{m_g+1}{2}}(\tau_g^2)^{\frac{m_g+1}{2}-1}e^{-\frac{\lambda^2\tau_g^2}{2}}(1-\pi_0)(2\pi\tau_g^2)^{-\frac{m_g}{2}}e^{-\frac{\beta_g^T\beta_g}{2\tau_g^2}}$$
$$\propto (\tau_g^2)^{\frac{1}{2}}e^{-\frac{1}{2}\left(\frac{\beta_g^T\beta_g}{\tau_g^2}+\lambda^2\tau_g^2\right)}$$

Letting $\tau_g^2 = \frac{1}{\alpha_g^2}$ we have,

$$p(\alpha_g^2|\text{rest}) \propto (\alpha_g^2)^{-\frac{3}{2}}e^{-\frac{\beta_g^T\beta_g}{2\alpha_g^2}\left[\alpha_g - \frac{\lambda}{\sqrt{\beta_g^T\beta_g}}\right]^2}$$

Thus,

$$\alpha_g^2|\text{rest} \sim \begin{cases} \text{Inverse Gamma } \left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right) & \text{if } \beta_g = 0 \\ \text{Inverse Gaussian } \left(\frac{\lambda}{||\beta_g||_2}, \lambda^2\right) & \text{if } \beta_g \neq 0 \end{cases}$$

(**3**)

$$p(\pi_0|\text{rest}) \propto \prod_{g=1}^{G}\left[(1-\pi_0)(2\pi\tau_g^2)^{-\frac{m_g}{2}}e^{-\frac{\beta_g^T\beta_g}{2\tau_g^2}}I_{[\beta_g\neq 0]} + \pi_0\delta_0(\beta_g)\right] \times \pi_0^{a-1}(1-\pi_0)^{b-1}$$

Let t be the number of non-zero β_g's . Define,

$$Z_g = \begin{cases} 0 & \text{if } \beta_g = 0 \\ 1 & \text{if } \beta_g \neq 0 \end{cases}$$

then $\sum_{g=1}^{G} Z_g = t$

Thus,

$$p(\pi_0|\text{rest}) \propto (1-\pi_0)^{t+b-1}\pi_0^{G-t+a-1}$$

$$\pi_0|\text{rest} \propto \text{Beta}\left(\sum_{g=1}^{G} Z_g - t + a, \sum_{g=1}^{G} Z_g + b\right)$$

(**4**) λ is the tuning parameter so a large value of λ will shrink the coefficients excessively and a small value of λ will result in a diffuse distribution. Thus, the value of λ should be carefully chosen. Aligning with Xu and Ghosh's (2015) spirit,

$$\lambda^{(k)} = \sqrt{\frac{p+G}{\sum_{g=1}^{G} E_{\lambda^{(k-1)}}[\tau_g^2|\mathbf{Y}]}},$$

Here, the posterior expectation of τ_g^2 is approximated with the sample average of τ_g^2 generated in the Gibbs sampler based on $\lambda^{(k-1)}$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Akaike, H. (1973). "Maximum likelihood identification of Gaussian autoregressive moving average models." *Biometrika,* 255-265.

[2] Albaum, G., Wiley, J., Roster, C., and Smith, S. M. (2011). "Visiting item non-responses in internet survey data collection." *International Journal of Market Research,* 53(5), 687-703.

[3] Arlt, S., Buchert, R., Spies, L., Eichenlaub, M., Lehmbeck, J. T., and Jahn, H. (2013). 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.

[4] Bae, K., and Mallick, B. K. (2004). "Gene Selection Using a Two-Level Hierarchical Bayesian Model." *Bioinformatics,* 20(18), 3423-3430.

[5] Breiman, L. (1995). "Better subset regression using the nonnegative garrote." *Technometrics,* 37(4), 373-384.

[6] Cai, S., Huang, L., Zou,J., Jing,L., Zhai,B., Ji,G., von Deneen,K.M., Ren,J., Ren, A. and Alzheimer's Disease Neuroimaging Initiative. (2015). "Changes in thalamic connectivity in the early and late stages of amnestic mild cognitive impairment: a resting-state functional magnetic resonance study from ADNI." *PloS one,* 10(2), e0115573.

[7] Casella, G. (2001). "Empirical bayes gibbs sampling." *Biostatistics,* 2(4), 485-500.

[8] Castruccio, S., Ombao, H., and Genton, M. G. (2016). "A multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data." *arXiv preprint arXiv:* 1602.02435.

[9] Chatterjee, A., and Lahiri, S. N. (2011). "Bootstrapping lasso estimators." *Journal of the American Statistical Association,* 106(494), 608-625.

[10] Chen, Z., and Dunson, D. B. (2003). "Random effects selection in linear mixed models." *Biometrics,* 59(4), 762-769.

[11] Clerx, L., Jacobs, H. I. L., Burgmans, S., Gronenschild, E. H. B. M., Uylings, H. B. M., Echavarri, C., Visser,P. J., Verhey,F. RJ. and Aalten,P. (2013). "Sensitivity of different MRI-techniques to assess gray matter atrophy patterns in Alzheimer's disease is region-specific." *Current Alzheimer research,* 10(9), 940-951.

[12] Cox, R. W. (1996). "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages." *Computers and Biomedical research,* 29(3), 162-173.

[13] Craig, C. S., and Douglas. SP (2000). "Conducting international marketing research in the twenty-first century." *International Marketing Review,* 18(1), 80-90.

[14] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M. O., Chupin,M., Benali,H., Colliot,O. and Alzheimer's Disease Neuroimaging Initiative. (2011). "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database." *Neuroimage,* 56(2), 766-781.

[15] de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). "Whole-genome regression and prediction methods applied to plant and animal breeding." *Genetics,* 193(2), 327-345.

[16] D'Elia, A., and Piccolo, D. (2005). "A mixture model for preferences data analysis." *Computational Statistics and Data Analysis,* 49(3), 917-934.

[17] Didow, N. M., Perreault, W. D., and Williamson, N. C. (1983). "A cross-sectional optimal scaling analysis of the index of consumer sentiment." *Journal of Consumer Research,* 10(3), 339-347.

[18] Dolnicar, S., and Gruen, B. (2014). "Including Don't know answer options in brand image surveys improves data quality". *International Journal of Market Research,* 56(1), 35-50.

[19] Double, K.L., Halliday, G.M., Krill, J.J., Harasty, J.A., Cullen, K., Brooks, W.S., Creasey, H. and Broe, G.A. (1996). "Topography of brain atrophy during normal aging and Alzheimer's disease." *Neurobiology of aging,* 17(4), 513-521.

[20] Drolet, A. L., and Morrison, D. G. (2001). "Do we really need multiple-item measures in service research?" *Journal of Service Research,* 3(3), 196-204.

[21] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *The Annals of Statistics,* 32(2), 407-499.

[22] Epstein, R., and Kanwisher, N. (1998). "A cortical representation of the local visual environment." *Nature,* 392(6676), 598-601.

[23] Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). "The parahippocampal place area: recognition, navigation, or encoding?" *Neuron,* 23(1), 115-125.

[24] Epstein, R., Graham, K. S., and Downing, P. E. (2003). "Viewpoint-specific scene representations in human parahippocampal cortex" *Neuron,* 37(5), 865-876.

[25] Fahrmeir, L., and Gossl, C. (2002). "Semiparametric Bayesian models for human brain mapping." *Statistical Modelling,* 2(3), 235-249.

[26] Fan, J., and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association,* 96(456), 1348-1360.

[27] Fan, J., and Lv, J. (2010). "A selective overview of variable selection in high dimensional feature space." *Statistica Sinica,* 20(1), 101.

[28] Fan, Y., and Tang, C. Y. (2013). "Tuning parameter selection in high dimensional penalized likelihood." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 75(3), 531-552.

[29] Feng, R., Wang, H., Wang, J., Shrom, D., Zeng, X., and Tsien, J. Z. (2004). "Forebrain degeneration and ventricle enlargement caused by double knockout of Alzheimer's presenilin-1 and presenilin-2." *Proceedings of the National Academy of Sciences of the United States of America,* 101(21), 8162-8167.

[30] Figueiredo, M. A. (2003). "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 25(9), 1150-1159.

[31] Friston, K., Ashburner, J., Frith, C. D., Poline, J. B., Heather, J. D., and Frackowiak, R. S. (1995). "Spatial registration and normalization of images." *Human brain mapping,* 3(3), 165-189.

[32] Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M. A., Williams, G., Antoun, N., Sahakian,B. J. and Hodges, J. R. (2001). "Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia." *Neurology,* 57(2), 216-225.

[33] George, E. I., and McCulloch, R. E. (1997). "Approaches for Bayesian variable selection." *Statistica sinica,* 339-373.

[34] Gitelman, D. R., Penny, W. D., Ashburner, J., and Friston, K. J. (2003). "Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution." *Neuroimage,* 19(1), 200-207.

[35] Gossl, C., Auer, D. P., and Fahrmeir, L. (2001). "Bayesian spatiotemporal inference in functional magnetic resonance imaging." *Biometrics,* 57(2), 554-562.

[36] Grimmer, T., Riemenschneider, M., Forstl, H., Henriksen, G., Klunk, W.E., Mathis, C.A., Shiga, T., Wester, H.J., Kurz, A. and Drzezga, A (2009). "Beta amyloid in Alzheimer's disease: increased deposition in brain is reflected in reduced concentration in cerebrospinal fluid." *Biological psychiatry,* 65(11), 927-934.

[37] Hartig, M., Truran-Sacrey, D., Raptentsetsang, S., Simonson, A., Mezher, A., Schuff, N., and Weiner, M. (2014). "UCSF FreeSurfer Methods."

[38] Hastie, T., Tibshirani, R., and Friedman, J. (2009). "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer,* 27(2), 83-85.

[39] Hawkins, D. I., and Coney, K. A. (1981). "Uninformed response error in survey research." *Journal of Marketing Research,* 370-374.

[40] Hawkins, D. I., Coney, K. A., and Jackson Jr, D. W. (1988). "The impact of monetary inducement on uninformed response error. " *Journal of the Academy of Marketing Science,* 16(2), 30-35.

[41] Henderson, J. M., D. C. Zhu and C. L. Larson (2011). "Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: An fMRI study." *Visual Cognition,* 19(7), 910-927.

[42] Henderson, J. M., Larson, C. L., and Zhu, D. C. (2007). "Cortical activation to indoor versus outdoor scenes: an fMRI study." *Experimental Brain Research,* 179(1), 75-84.

[43] Herholz, K., Salmon, E., Perani, D., Baron, J. C., Holthoff, V., Frlich, L., Schnknecht, P., Ito, K., Mielke, R., Kalbe, E. and Zndorf, G. (2002). "Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET." *Neuroimage,* 17(1), 302-316.

[44] Hoerl, A. E., and Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics,* 12(1), 55-67.

[45] Hughes, J., and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 75(1), 139-159.

[46] Jack, C.R., Petersen, R.C., Xu, Y.C., Waring, S.C., O'Brien, P.C., Tangalos, E.G., Smith, G.E., Ivnik, R.J. and Kokmen, E. (1997). "Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease." *Neurology,* 49(3), 786-794.

[47] Jack, C.R., Petersen, R.C., Xu, Y.C., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Waring, S.C., Tangalos, E.G. and Kokmen, E. (1999) "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment." *Neurology,* 52(7), 1397-1397.

[48] Jiang, W. (2007). "Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities." *The Annals of Statistics,* 35(4), 1487-1511.

[49] Johnstone, I. M., and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *Annals of Statistics,* 1594-1649.

[50] Juottonen, K., Laakso, M. P., Insausti, R., Lehtovirta, M., Pitknen, A., Partanen, K.,and Soininen, H. (1998). "Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease." *Neurobiology of aging,* 19(1), 15-22.

[51] Kim, Y., Kwon, S., and Choi, H. (2012). "Consistent model selection criteria on high dimensions." *Journal of Machine Learning Research,* 13(Apr), 1037-1057.

[52] Knight, K., and Fu, W. (2000). "Asymptotics for lasso-type estimators." *Annals of statistics,* 1356-1378.

[53] Kumar, V. (2000). "International marketing research (pp. 225-226)." *Upper Saddle River, NJ: Prentice Hall.*

[54] Kumar, V., and Pansari, A. (2016). "Competitive advantage through engagement." *Journal of Marketing Research,* 53(4), 497-514.

[55] Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis,* 5(2), 369-411.

[56] Leifer, B. P. (2003). "Early diagnosis of Alzheimer's disease: clinical and economic benefits." *Journal of the American Geriatrics Society,* 51(5s2), S281-S288.

[57] Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models" *Journal of the American Statistical Association,* 108(502), 589-606.

[58] Lpez, M.E., Brua, R., Aurtenetxe, S., Pineda-Pardo, J.., Marcos, A., Arrazola, J., Reinoso, A.I., Montejo, P., Bajo, R. and Maest, F. (2014). "Alpha-band hypersynchronization in progressive mild cognitive impairment: a magnetoencephalography study." *The Journal of Neuroscience,* 34(44), 14551-14559.

[59] Lykou, A., and Ntzoufras, I. (2013). "On Bayesian lasso variable selection and the specification of the shrinkage parameter." *Statistics and Computing,* 23(3), 361-390.

[60] Manisera, M., and Zuccolotto, P. (2014). "Modeling "don't know" responses in rating scales." *Pattern Recognition Letters,* 45, 226-234.

[61] Manisera, M., and Zuccolotto, P. (2014). "Modeling rating data with nonlinear CUB models." *Computational Statistics and Data Analysis,* 78, 100-118.

[62] Meier, L., Van De Geer, S., and Bhlmann, P. (2008). "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 70(1), 53-71.

[63] Mitchell, T. J., and Beauchamp, J. J. (1988)."Bayesian variable selection in linear regression." *Journal of the American Statistical Association,* 83(404), 1023-1032.

[64] Money, R. B., Gilly, M. C., and Graham, J. L. (1998). "Explorations of national culture and word-of-mouth referral behavior in the purchase of industrial services in the United States and Japan." *The Journal of Marketing,* 76-87.

[65] Musgrove, D. R., Hughes, J., and Eberly, L. E. (2016). "Fast, fully Bayesian spatiotemporal inference for fMRI data." *Biostatistics,* 17(2), 291-303.

[66] Nachum, L. (1994). "The choice of variables for segmentation of the international market." *International Marketing Review,* 11(3), 54-67.

[67] Nakata, C., and Sivakumar, K. (1996). "National culture and new product development: An integrative review." *The Journal of Marketing,* 61-72.

[68] Narisetty, N. N., and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics,* 42(2), 789-817.

[69] Olusegun, A. M., Dikko, H. G., and Gulumbe, S. U. (2015). "Identifying the Limitation of Stepwise Selection for Variable Selection in Regression Analysis." *American Journal of Theoretical and Applied Statistics,* 4(5), 414-419.

[70] Pandey, S., and Elliott, W. (2010). "Suppressor variables in social work research: Ways to identify in multiple regression models." *Journal of the Society for Social Work and Research,* 1(1), 28-40.

[71] Park, T., and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association,* 103(482), 681-686.

[72] Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). "Bayesian fMRI time series analysis with spatial priors." *NeuroImage,* 24(2), 350-362.

[73] Petersen, J. A., Kushwaha, T., and Kumar, V. (2015). "Marketing communication strategies and consumer financial decision making: The role of national culture." *Journal of Marketing,* 79(1), 44-63.

[74] Piccolo, D. (2003). "On the moments of a mixture of uniform and shifted binomial random variables." *Quaderni di Statistica,* 5(1), 85-104.

[75] Reuter, M., Schmansky, N. J., Rosas, H. D., and Fischl, B. (2012). "Within-subject template estimation for unbiased longitudinal image analysis." *Neuroimage,* 61(4), 1402-1418.

[76] Rokeach, M. (1973). *Value Survey.*

[77] Rutz, O. J., Trusov, M., and Bucklin, R. E. (2011). "Modeling indirect effects of paid search advertising: which keywords lead to more future visits?." *Marketing Science,* 30(4), 646-665.

[78] Schuman, H., and Presser, S. (1980). "Public opinion and public ignorance: The fine line between attitudes and nonattitudes." *American Journal of Sociology,* 85(5), 1214-1225.

[79] Schwartz, S. H. (1992). "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries." *Advances in Experimental Social Psychology,* 25, 1-65.

[80] Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics,* 6(2), 461-464.

[81] Siddiqui, M. M. (1958). "On the inversion of the sample covariance matrix in a stationary autoregressive process." *The Annals of Mathematical Statistics,* 29(2), 585-588.

[82] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). "A sparse-group lasso." *Journal of Computational and Graphical Statistics,* 22(2), 231-245.

[83] Smith, M., Ptz, B., Auer, D., and Fahrmeir, L. (2003). "Assessing brain activity through spatial Bayesian variable selection." *NeuroImage,* 20(2), 802-815.

[84] Smith, M., and Fahrmeir, L. (2007). "Spatial Bayesian variable selection with application to functional magnetic resonance imaging." *Journal of the American Statistical Association,* 102(478), 417-431.

[85] Smith, C. D., Andersen, A. H. and Gold, B. T. (2012). "Structural brain alterations before mild cognitive impairment in ADNI: validation of volume loss in a predefined antero-temporal region." *Journal of Alzheimer's Disease,* 31(s3), S49-S58.

[86] Steenkamp, J. B. E., and Ter Hofstede, F. (2002). "International market segmentation: issues and perspectives." *International Journal of Research in Marketing,* 19(3), 185-213.

[87] Steenkamp, Jan-Benedict EM, Frenkel ter Hofstede, and Michel Wedel. (1999) "A cross-national investigation into the individual and national cultural antecedents of consumer innovativeness." *The Journal of Marketing,* 55-69.

[88] Tadesse, M. G., Sha, N., and Vannucci, M. (2005)." Bayesian variable selection in clustering high-dimensional data." *Journal of the American Statistical Association,* 100(470), 602-617.

[89] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological),* 267-288.

[90] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67(1), 91-108.

[91] Verhoef, P. C., and Donkers, B. (2001). "Predicting customer potential value an application in the insurance industry." *Decision Support Systems,* 32(2), 189-199.

[92] Wang, X., Nan, B., Zhu, J., and Koeppe, R. (2014). "Regularized 3D functional regression for brain image data via Haar wavelets." *The annals of applied statistics,*8(2), 1045.

[93] Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). "neuRosim: An R package for generating fMRI data." *Journal of Statistical Software,* 44(10), 1-18.

[94] Xu, X., and Ghosh, M. (2015). "Bayesian variable selection and estimation for group lasso." *Bayesian Analysis,* 10(4), 909-936.

[95] Yuan, M., and Lin, Y. (2005). "Efficient empirical Bayes variable selection and estimation in linear models." *Journal of the American Statistical Association,* 100(472), 1215-1225.

[96] Yuan, M., and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 68(1), 49-67.

[97] Zhang, Y., Li, R., and Tsai, C. L. (2010). "Regularization parameter selections via generalized information criterion." *Journal of the American Statistical Association,* 105(489), 312-323.

[98] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and Alzheimer's Disease Neuroimaging Initiative. (2011). "Multimodal classification of Alzheimer's disease and mild cognitive impairment." *Neuroimage,* 55(3), 856-867.

[99] Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K. A. (2014). "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.

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

[101] Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association,* 101(476), 1418-1429.