VARIABLE SELECTION FOR SPATIAL DATA AND ITS APPLICATION TO NEUROIMAGING

By

Abdhi Amitabha Sarkar

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 FOR SPATIAL DATA AND ITS APPLICATION TO NEUROIMAGING

By

Abdhi Amitabha Sarkar

Ecological research, geological studies, image analysis are a few examples of high resolution spatial data where proximity describes the relationship between data points collected at various locations. Such dependencies play a vital role in modeling the data accurately to improve both its predictive capacity and parameter estimation. Rapid technological advancement has brought about an abundance of such information. To better understand this information, we are in need of feature selection techniques for spatially dependent data that can tease out relevant predictors associated with the response of interest. When the response variable at the various sites is in the form of discrete binary or count data we are faced with an added layer of complexity due to the inability of explicitly describing a joint parametric distribution. This dissertation explores the benefits of adopting a penalized quasi-likelihood approach to model a fixed number(p) or an expanding dimension(p_n) of predictor variables with regard to a discrete spatial response variable. In the past this approach has been extensively studied in longitudinal data analysis. Introducing random fields that exhibit certain ρ-mixing conditions we are able to provide some general theoretical results of the estimator obtained from the solving the penalized score equation. The oracle properties of the estimator as described by J. Fan & Li (2001) are provided, followed by an algorithm to successfully implement the method. Multiple simulation studies showcase the effectiveness of the method under covariance misspecification. We apply this technique to real data obtained from the Michigan Natural Features Inventory.

Variable selection in neuroimaging has a unique formulation that leads to selection of activated regions of a brain in Task-based fMRI. As one of the most non-invasive formats of studying an active brain, Task-based fMRI provides a unique opportunity in neuroscience to study the dynamic aspects of brain function. Crude statistical techniques such as voxel-wise regression analysis have been used in the past with some success to identify active brain regions based on the blood-oxygen-level dependent (BOLD) signal of the image. Inspired by graphical covariate models proposed for genetic data we incorporate a similar idea and expand our understanding of penalized regression of weighted least squares with a separable space-time covariance model in this setup. Two penalty terms are introduced as a result; one for selection (LASSO) and another for smoothing (Ridge-type). We explore the interpretability of the proposed model as opposed to its Bayesian counterparts, its computational feasibility and various approaches to selecting an optimal tuning parameter in the case of a Single-subject study. The description of the model and its implementation are presented with discussions about theoretical implications. Extensive simulation studies and a real data example of a human brain subject to two visual stimuli are also given to provide evidence of the capability of this method.

To the memory of my family and friends who passed away during the course of my PhD.
for reminding me to cherish every moment and always strive for something better.

ACKNOWLEDGMENTS

I would like to take this opportunity to sincerely thank Dr. Chae Young Lim who introduced me to spatial statistics and her seamless understanding of the subject was truly inspiring. Her incredible dedication to her students is what pulled us through in our research. I truly appreciate her time to work with us without any hesitation. My chair advisor Professor Tapabrata Maiti took me in at his busiest time, having had so many students under his wing, just to see me succeed. I was introduced to a more modern approach to statistics and I could not be more grateful for all the opportunities given to me.

My committee member Dr. Andrew Finley was incredibly generous in providing me with opportunities and resources during my studies. I would also like to thank Dr. Ping-shou Zhong for being a part of my committee and for our discussions.

My sincerest gratitude to Steve Pierce at CSTAT for his extensive support and opportunities to collaborate on a variety of applications. This provided me with a unique insight into how statistical methods are viewed and interpreted.

A very special mention to Andy Hufford for being incredibly patient with me and trying to improve resources for the graduate students. His advice was very accurate on many occasions and it was a real pleasure to work with him.

My friends Pei, Sneha, Atreyee, Danielle, Guiling, Siddhartha and officemates Vojtech, Jimmy, Yingjie made East Lansing home and I cannot thank them enough. A special mention to my fiancé Michael Frisby for his unwavering support during the entire process and pure faith in my abilities.

Lastly I would like to thank my parents and sister for standing by my side always.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS

1. $\leq$ denotes the positive semidefinite ordering. i.e. we write $A \leq B$ if the matrix $B - A$ is positive definite.
2. $|\cdot|$ denotes the cardinality of a set.
3. $\bigtriangledown_\beta \mathcal{K}$ is the gradient of a vector $\mathcal{K}$ w.r.t β.
4. $\circ$ signifies the Schur or Hadamard product of two matrices.
5. $\otimes$ signifies the Kronecker product of two matrices.
6. $\ddot{\mu}$ denotes the double derivative of the function μ, similarly $\dddot{\mu}$ denotes the third derivative and so on.
7. $\coloneqq$ signifies assignment or is referred to as "denoted by".
8. Tr A denotes the trace of a matrix A.
9. LHS stands for Left Hand Side of an equation.
10. $\sim$ denotes neighbors. i.e. $u \sim v$ implies u and v are adjacent voxels.
11. $\langle \cdot, \cdot \rangle$ signifies inner product of two vectors.
12. $sign(a)$ signifies -1 or $+1$ depending on whether $a < 0$ or $a > 0$ respectively.
13. $\mathbb{I}(\cdot)$ denotes the indicator operator.
14. TP (True Positives) represents the average number of correctly detected nonzero coefficients.
15. FP (False Positives) represents the average number of incorrectly detected nonzero coefficients.
16. CP gives the average empirical Coverage Probability of the 95% confidence intervals.
17. CD stands for Coordinate Descent

Chapter 1

Introduction and Literature Review

Spatial data innately incorporates dependencies based on proximity. Ignoring such dependencies in the analysis of data gives rise to inconsistent and inaccurate results as formally stated in Fitzmaurice (1995) and other sources. There exists a very intuitive notion of short-range dependence that suggests, observations at locations closer to each other are more likely to be homogeneous than observations at locations that are farther apart. A variety of examples can be seen in climate and meteorological data, geological data on Earth and outer space, demographic social/economic data and image analysis. The focus of this dissertation lies in studying discrete spatial data with binary or count-like information at different sites. Several approaches to model these dependencies in the form of latent Gaussian processes are explained in P. J. Diggle et al. (1998) with some early beginnings seen in the seventies with extensive work done by Matheron (1970) in the development of *kriging*.

A parallel rather statistical approach with auto-normal and auto-logistic models were considered using Markov Random Field by Besag (1974) with the use of the positivity condition (Hammersley & Clifford, 1971). A very comprehensive take on spatial data analysis can be seen in Cressie (1993). In chapter 6 of his book, Cressie discusses the initial pairwise-dependence approaches that were employed for exponential distributions of the response (Poisson, Gamma, Binomial etc.) and chapter 7 elaborates that the non-Gaussian case

of the maximum pseudo-likelihood estimator introduced in Besag (1975) may not have a structure to obtain an optimum solution.

P. Diggle et al. (2007) give us an overview of using the latent processes to interpolate surfaces using generalized linear geospatial models (GLGM). However lack of closed form expressions of full and marginal likelihoods leads to adopting Expectation Maximization (EM) algorithms or Bayesian methods. Bayesian techniques in the form of hierarchical mixed models have been prodigious in modeling parametric spatial models, (Banerjee et al., 2014) but heavily rely on maximum likelihood methods that suffer from model misspecification. A similar model for spatial binary data was proposed by Heagerty & Lele (1998) using a penalized composite likelihood in the form of probit models. The model specifically assumes that there exists a Gaussian spatial process with binary responses that are indicators of whether the spatial process was measured with error. Here certain score function asymptotic properties with regularity conditions formalized by Guyon (1995) are used. These regularity conditions are required in the study of GEEs in chapter 2 of this dissertation. In the context of spatial regression for discrete data not much has been explored in great detail. Wedderburn (1974) published seminal work on quasi-likelihood functions describing regression based techniques that required only parametric structures to be imposed on the first two moments of a random variable; a semi-parametric method instead a full distributional assumption in terms of a maximum likelihood (MLE) method. No closed form expressions were necessary for the objective Quasi-likelihood function. Further, conditions were cautiously provided whenever the observations are dependent, (McCullagh & Nelder, 1989, Chapter 9). Gotway & Stroup (1997) propose a quasi-likelihood method for estimation and prediction with generalizations to universal and indicator kriging but provide no theoretical justification.

The use of generalized estimating equations (GEE) has been very well developed in

longitudinal data analysis also known as growth curve analysis. The methods established in Liang & Zeger (1986) do not directly translate into the spatial setting but are modified with the assumption of a latent process as in Zeger (1988); McShane et al. (1997). Lin & Clayton (2005) provided very strong theoretical justifications for binary spatial data and asymptotics under the increasing domain framework using GEE and the quasi-likelihood. A generalization of this method was given in Lin (2008) that extends the technique for using these methods for both binary and count data also allowing the underlying covariance structure to be more flexible. A working correlation setup was incorporated into the existing methodology by Lin (2010) to accommodate covariance-misspecification.

Model selection is an incredibly practical problem especially in the initial exploratory stage of any data analysis. For estimating equations usual criterion for stepwise selection are inadequate and therefore a modified version of the Akaike information criterion was introduced by Pan (2001). Prior to which some other attempts were made by using Mallow's C_p and AIC_c using Kullback-Leibler (Hurvich & Tsai, 1995). L. Wang & Qu (2009) used the quadratic inference function which generalizes the method of moments to longitudinal data and formed a novel BIC-type model selection procedure, as well as a test for checking the goodness of fit. This method does not require either the full-likelihood or the quasi-likelihood to have explicit closed form expressions. Modern penalized regularization techniques in machine learning and statistics with faster convergence algorithms have gained immense popularity and credibility over previous techniques, especially with the success of LASSO (Tibshirani, 1996). Essentially it has been formulated to specifically cater to model complexity in the form of high-dimensional data. The review article by Bickel et al. (2006) describes methods for independent samples that are commonly adopted (regularized least squares) and succeeded in addressing complex problems. Better penalty functions refer-

ring to those satisfying oracle properties like Smoothly Clipped Absolute Deviation (SCAD, J. Fan & Li (2001)) were soon established in the form of hard-thresholding by Donoho et al. (1994), Adaptive LASSO by Zou (2006), the Dantzig selector by Candes & Tao (2007) and Minimax Concave Penalty (MCP) by Zhang et al. (2010) among some potential others. Some penalized methods for dependent samples has been investigated and developed in a time series model by H. Wang et al. (2007). The method establishes a lasso tuning parameter for both the covariate regression coefficients and the autoregressive covariance parameter. A computationally efficient algorithm for GIS model selection of neighborhood extents and patterns on lattice data was explored by H.-C. Huang et al. (2010) in the form of spatial LASSO but lacked detailed theoretical backing. In general there appears to be a gap of regularization techniques applied to dependent data. J. Zhu et al. (2010) established a penalized maximum likelihood method under spatial adaptive lasso for data that has a conditional auto-regressive (CAR) model to simultaneously select covariates of relevance and a neighborhood structure.

The beginnings of penalized quasi-likelihood methods are seen in Breslow & Clayton (1993) with the use of Laplacian methods originally formulated by Green (1987) for semi-parametric regression models with an iterated weighted least squares algorithm for correlated response. These models too incorporate dependence in the form of a random effect. However the only known spatially correlated penalized quasi-likelihood methods were explored by Dean et al. (2004). No aymptotic properties are shown and the penalty is not in the context of regularization techniques. Additionally the spatial relationships discussed are in the form of adjacency matrices or a Laplacian approximation of the integral of the quasi-likelihood under a GLMM setup.

In chapter 2 of this dissertation, GEE techniques in the similar context of longitudinal

data analysis is considered. Specifically we look at a model selection procedure for a large number of predictors associated with a discrete spatial response on a lattice. Here the within-cluster covariance has a spatial structure satisfying certain $\rho-$mixing conditions and with no replicates at each location. This chapter considers two specific cases; the first where the number of predictors p is fixed and the second where the dimension suffixed by n denotes is expanding. i.e. $p_n \to \infty$ as $n \to \infty$. The penalized quasi-likelihood score function properties are shown to posses properties in the current spatial setup that match Johnson et al. (2008) similar to results shown in Feng et al. (2016) with a working correlation that satisfies the $\rho-$mixing condition in Lin (2008). Extensive simulations under various working correlations with both count and binary data are conducted and showcased in this chapter.

Portnoy (1988) addressed a very fundamental question regarding the number of parameters required for valid statistical analysis using asymptotic results. It was shown that for exponential families applying MLE methods for estimation, p^2/n should be very small. Thus formalizing methods where the number of parameters p_n goes to infinity. In the case of expanding dimensions, we therefore require comparable selection techniques since no results have been established for this scenario with regard to GEE in the spatial context. Section 2.4 addresses this by using (Ortega & Rheinboldt, 1970, Theorem 6.3.4) given in L. Wang (2011) to prove the existence and consistency of the estimate obtained from solving the penalized quasi-likelihood estimations. Similar to L. Wang et al. (2012) we show selection consistency obtained form the estimating equations of the penalized score function. Specifically a theoretical explanation is provided that under this setup a difference in the rate at which the dimension expands varies significantly in comparison to the longitudinal setup or maximum likelihood setup (Portnoy, 1988). This has been briefly touched upon by Xie et al. (2003) under the setup n is bounded and $m \to \infty$ where m denotes the cluster of time

points repeated for every individual. We require $p_n^4.n^{-1} = o(1)$ as opposed to $p_n^2 n^{-1} = o(1)$ respectively as $n \to \infty$. In section 2.3.2 we describe the model implementation and the use of Majorize-minimize(MM) algorithm by Hunter & Li (2005). Another important aspect of regularization techniques is finding an optimal tuning parameter. Cross validation techniques tend to select over-fitted models but is widely popular and successful in methods involving independent data. Since the concept of a complete replicate guides the method, this intuition is not apparent in the context of dependent data. Therefore optimal values are obtained based on a grid search over a range of values and root mean squared error is used as the optimality criterion. The purpose of this work is to lay down foundations and explore possibilities of solving ecological or geospatial applications involving high-dimensional data. The section 2.4.2 of chapter 2 establishes theoretical results with regard to variable selection for expanding dimensions of covariates. The oracle properties and selection consistency along with characterization of the solution of the penalized equations is provided in detail. The method in its current form can handle scenarios where the true underlying number of covariates (denoted by s_n) satisfies $s_n^4 n^{-1} = o(1)$ and $p_n = o(n)$ where p_n is the total number of covariates. Two real data analyses have been performed in section 2.3.4. The first real data is obtained from the Michigan Natural Features Inventory affiliated to the State Department of Michigan and Michigan State University, to study fire disturbances in the region and obtain predictors that may have sustained such disturbances in the early 1800s. The second example is county-wise data of Lung cancer incidence in the state of Iowa obtained from SEER (Surveillance, Epidemiology and End Results program) database. Here we are interested in learning which factors are associated with high or lower rates of lung cancer incidence in particular areas of Iowa.

Image analysis comprises the processes of restoration of an image from contamination,

extraction of vital information such as bar codes, facial recognition and interpretation of images. Data from these images are obtained in different forms. For example, remote sensing by satellites, digital or chemical photography using electronic photo-detectors, exposure to film respectively or through medical imaging such as recording transmission of X-rays, electron microscope imaging, ultrasound, magnetic resonance imaging (MRI), positron emission tomography (PET), diffusion tensor imaging (DTI) etc. With the advent of computer technology into the processing of images and pattern recognition, spatial statistics has played a key role. (Cressie, 1993, Section 7.4) provides details of spatial methodologies employed in the study of images.

Neuroimaging has in the past few decades (since 1990) improved greatly in terms of data acquisition. The most noninvasive procedures of taking an image of the brain is MRI. As an application to variable selection for spatial data, the focus of chapter 3 is to identify activated regions of the brain responding to a stimulus based on a task performed. fMRI (functional magnetic resonance imaging) detects active brain parts by measuring changes in the blood oxygen level dependent (BOLD) signal. For an in-depth understanding of fMRI data, Lindquist (2008) describes the nature of the field from a statistician's perspective. In general there are two major classes of experimental designs: block designs and event-related (Task-based designs). Another version of fMRI data is in the form of resting state fMRI famously introduced by Biswal et al. (1995) investigating connectivity in the brain when the human is at resting-state (performing no task). This version of the data has received mixed critiques but has seen success in studying diseased brains compared to healthy controls. Due to its unique structure in the context of variable selection, the focus of chapter 3 of this dissertation is to propose a selection technique for Task-based fMRI.

Unlike the data we have seen so far, Task-based fMRI results in large amounts of noisy

data with a complicated spatio-temporal correlation structure. Additionally in the context of selection, the design matrix associated with the response has a handful of stimuli randomly assigned over time. The objective therein is to select only those brain regions that are responding to the stimulus provided over time. Thus there are co-efficients attached to each spatial unit of the image (known as a *voxel*) and the data is collected over time at each voxel. Section 3.1 provides details of the experiment conducted and the nature of the data collected. It is important to note that the BOLD signal is regressed against a modeled hemodynamic response function (HRF) with the intuition that the coefficient associated with the voxel quantifies activation amplitude (the magnitude of correlation) between the stimulus and the voxel over time.

In the early nineties voxel-wise regression analysis was extensively used due to it's simplicity with respect to both the implementation and interpretation of the method. However a blatant draw-back was almost immediately recognized with the realization that an extremely large number of non-independent univariate comparisons were performed, thus sizeably inflating the Type-I error. Statistical Parametric Maps (SPMs) (Friston et al., 1994) was among the earliest most well established techniques referring to the probabilistic behavior of stationary Gaussian random fields (Adler, 1981) to construct a threshold using the Euler characteristic (Worsley et al., 1992). These techniques were studied for PET scans to calculate the approximate p-value in order to find the statistically significant regions of activation. Therefore the analysis was done in two steps. The first being a voxel-wise analysis of the time series data and the second step being the spatial adjustments of the p-values using random field theory (RFT). Bayesian methods had been very successfully implemented in image restoration by Geman & Geman (1984) where they judiciously exploited the equivalence of Gibbs distribution and the Markov Random Field (MRF) to construct a maximum

a posteriori MAP estimate of the degraded image. Gössl et al. (2001) introduced a hierarchical Bayesian method that incorporated modified intrinsic autoregressive priors as a kind of stochastic interpolation of neighbors for modeling spatial dependence (Gaussian MRFs) and random walks for modeling temporal dependence. The aim of their formulation was to be able to simultaneously model spatial and temporal dependencies directly in a single step. Some improvements of spatial adaptivity were made by extending the above technique to Laplacian and Cauchy-type priors by few others.

A slightly different interpretation of the voxel selection was adopted by Smith & Fahrmeir (2007). Instead of priors on the amplitudes themselves they impose an Ising (Ising, 1925) prior on the unobservable indicator latent variables of the activation effects. Thus the selection is now based on a probability map of more likely activated areas of the brain. The Ising prior or the binary spatial MRF treats the lattice of voxels as a graph incorporating neighboring interactions to determine clusters of activated regions. This spatial Bayesian variable selection (SBVS) method empirically refined issues of over smoothing and edge effects but was not generalized to model in time dependencies due to the additional computation burden. Lee et al. (2014) integrated auto-regressive (AR) and moving average (MA) time series to the setup of Smith & Fahrmeir (2007) and rigorously applied it to a longitudinal study of Alzheimer's disease (AD). For the block-design experiment they found the setup not conducive for moving averages but implemented an Empirical Bayes approach to estimate (ρ) parameter associated with AR(1). They also compared this with models AR(2), ARMA(1,1) and MA(1) but found AR(1) to provide the best results. In general for multiple subjects with each subject having over 100,000 voxels over a 20-30 mins study, the data is unimaginable massive. Furthermore with Bayesian MCMC methods applied to the data a healthy compromise between a suitable method and computational complexity must be

sought. Musgrove et al. (2016) identify this issue and propose a partitioning of the brain optimally and then fitting models to these partitions. They show through simulations that this kind of parcellation does not exhibit unnecessary edge-effects and boundary problems but no theoretical justification is provided.

Although the probability maps based on the latent indicator process determines whether the effect of a voxel is active or inactive, it does not necessarily render a simplistic interpretation of the activation amplitude itself (co-efficient of interest). Regularization methods in bio-informatics has seen significant success with widely available data and greater computational flexibility. Infusing complex neighborhood structures that models correlated predictors using graph-structured regression covariates for variable selection in high-dimensional genomic data as seen in Li & Li (2010) is one such example. A generalization of this method is adopted by introducing sparse laplacian shrinkage (SLS, J. Huang et al. (2011)) where the LASSO penalty for sparsity is replaced by MCP for its oracle properties. In a similar vein, to study the HIV type I protease structure Xue et al. (2012) incorporated coupling coefficients using the Ising model to select the true underlying structure of interactions using the non-concave SCAD penalty. These rather non-Bayesian methods for massive data incorporate the coordinate-descent algorithm that produces remarkable computational efficiency. Some key aspects of this algorithm are touched upon based on a technical report by Tseng (1988) in section 3.6 of this dissertation. Further, Grosenick et al. (2013) provide a comprehensive report of attempts made to use of-the-shelf machine learning methods to analyze fMRI data. For a whole brain analysis with correlated data they provide a handful of variants of the Graph-constrained Elastic Net (GraphNet). These methods are proposed and implemented on whole brains of multiple subjects based on event-designs with fewer than 10 time points.

In chapter 3 of this dissertation we venture to propose a rather frequentist approach

using regularization methods discussed above by treating the 2D or 3D image as a graph and applying a method similar to that employed by Li & Li (2010) in a single subject study with fMRI block-design experiments. There exists a fundamental difference in the implementation of the method in a single subject fMRI study unlike Genomic data on multiple subjects due to the spatio-temporal setup. Replicates for each node on the graph i.e. each voxel is in the form of time-series and no other independent replicates exist. Thus the responses are directly associated with the graphical structure imposed in the smoothing penalty of Li & Li (2010). Details of the method are provided in section 3.3. Two penalty terms to a weighted least squares objective function are applied. The first penalty term is a LASSO continuous convex non-differentiable penalty for selection and the second is a smoothing penalty that uses an adjacency matrix which shares properties of a Laplacian matrix on a graph to obtain a convex penalty.

The method is demonstrated through a variety of simulation studies that uses a known spatio-temporal formulation with some examples incorporating covariance tapering (Ex: Wendland) to infuse additional sparsity in a rather dense matrix formulation. This induction of sparsity improves the computationally feasibility of the method. Additionally some variation of the coordinate descent method are like path-wise coordinate descent, warm starts, active sets and full tuning parameter search grids in section 3.3.2. We also use the block-design to our advantage and investigate the performance of the method to find optimal tuning parameters using time Cross Validation explained in 3.3.2.4. One more attempt is made at efficiently locating the tuning parameter using the empirical bayes estimate obtained for the LASSO in the Bayesian LASSO approach introduced by Park & Casella (2008). Using Zou & Hastie (2005) we can reformulate the two penalty and convert it to a LASSO regularization method and then use the empirical bayes estimate of the LASSO tuning parameter.

The empirical investigation of this method suggests that there is a need to find an adequate criteria to find optimal tuning parameters. Furthermore the extent of spatio- temporal variance may affect the selection of these parameters. We would like to in the future like to explore the theoretical properties of the estimator obtained from this method. Lastly as a real data application we consider a single subject study where an individual's brain is scanned through a series of visual stimuli. Specifically two stimuli, an object and a landscape image are shown to the individual randomly at regular intervals with some rest. The result show which portions of th brain appear to be activated by these stimuli. This study was conducted in Michigan State University, Department of Radiology by David Zhu.

Chapter 2

Variable Selection for Discrete Spatial Data using a Penalized Quasi-likelihood Approach

2.1 Model Setup

Let us consider a discrete, binary or count random field denoted by a random variable $Y = \{Y(s_1), Y(s_2), ..., Y(s_n)\}$ as an n-dimensional vector associated with sampling sites $s_1, ..., s_n$ arranged on a spatial domain in $\Re^d$, where $d \geq 2$. Let X denote the design matrix with $n \times p$ dimension. In the context of regression, the corresponding coefficient vector $\beta = (\beta_1, .., \beta_p)^T$ denotes the magnitude and direction of the impact of a predictor on the response. The mean of the response vector $E(Y) = \mu = (\mu(s_1), ..., \mu(s_n))^T$ is assumed to be related to the matrix X of p predictors through a link function $g(\cdot)$ such that $g(\mu) = X\beta$. Each x_{ij} is finite and corresponds to the j^{th} predictor at site s_i i.e. $max_{ij}|x_{ij}| < \infty$. We further impose that this link function $g(\cdot)$ is smooth over μ such that the first and second order derivatives of $g^{-1}(\cdot)$ are finite.The variance function $v(\cdot)$ of Y is assumed to be smooth such that $Var(Y(s_i)|x_i.) = v(g^{-1}((x_i^T\beta))$. Let $Cov(Y) = V(\mu)\sigma^2 = V\sigma^2$ where σ^2 is

independent of μ. By virtue of the data being collected on sites at a different spatial locations, a dependence structure in assumed for the variance of the response Y i.e. $\sigma^2 V = \Sigma^{1/2}\Gamma\Sigma^{1/2}$ where Γ is the true underlying symmetric, isotropic and positive definite spatial covariance matrix and $\Sigma = diag\{v(g^{-1}(x_1\beta))), .., v(g^{-1}((x_n\beta)))\}$ is a diagonal matrix with variance components.

In order to alleviate distribution restrictions for correlated discrete data that does not have a likelihood function that can be expressed explicitly, we use a Quasi-likelihood approach, previously used by Lin & Clayton (2005). In this chapter we explore two cases in which the design matrix X presents itself.

(Case 1) The design matrix X has a fixed number of predictors. i.e. $p < n$ and is fixed.

(Case 2) The design matrix X has a varying number of predictors that grows as the sample size increases. i.e. $p_n \to \infty$ as $n \to \infty$ but $p_n < n$.

Below is a brief review of Quasi-likelihood functions and score functions as described in (McCullagh & Nelder, 1989, Chapter 9); with regard to the structure and attributes of both cases considered above.

2.2 Penalized Quasi-likelihood Estimating Equations

The Quasi-likelihood Estimating equations for the parameter of interest β is obtained by differentiating an objective Quasi-likelihood function $\mathcal{QL}(\beta)$, resulting in solving the equations $\mathcal{U}(\beta) = 0$ where the Quasi-score function is given by,

$$\mathcal{U}(\beta) = D^T V^{-1}(Y - \mu)/\sigma^2 \tag{2.1}$$

$D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$, the partial derivative of the i^{th} mean function with respect to j^{th} regression coefficient. The derivative of the score function with respect to β that yields the gradient, $\frac{\partial^2 \mu_i}{\partial \beta_k \partial \beta_j}$ are assumed to be finite and equal the covariance of $\mathcal{U}(\beta)$. Specifically,

$$Cov(\mathcal{U}(\beta)) = D^T \Sigma^{-1/2} \Gamma \Sigma^{-1/2} D \tag{2.2}$$

$$E[\bigtriangledown_\beta \mathcal{U}(\beta)] = -D^T \Sigma^{-1/2} \Gamma \Sigma^{-1/2} D \tag{2.3}$$

For correlated observations, however there may be multiple roots that solves equation (2.1). As suggested by (McCullagh & Nelder, 1989, Section 9.3.2) to ensure that $\hat{\beta}$ is unique and globally maximizes the quasi-likelihood, we require a V^{-1} such that $\bigtriangledown_\beta \mathcal{U}(\beta)$ is symmetric in order for the integral of the score-function that describes the Quasi-likelihood to be path independent. This is satisfied by our choice of Γ explained in section 2.1.

In order to obtain asymptotic properties of the score function we require some additional conditions on the dependence structure of the matrix Γ. Let $\wedge$ denote a lattice subset of the locations. Lin (2008) defined the ρ-product mixing coefficients as a generalization of (Guyon, 1995, Page 112) as given below,

$$\rho_{k,l\,p}(m) = sup[|Cov\{\prod_{s_i \in \wedge_1} y(s_i), \prod_{s_j \in \wedge_2} y(s_j)\}| : E(|y(s)|^2) \leq 1, \\ | \wedge_1 | \leq k, | \wedge_2 | \leq l, d_p(\wedge_1, \wedge_2) \geq m] \tag{2.4}$$

where $d_p(\wedge_1, \wedge_2) = inf\{\|s_1 - s_2\|_p : s_i \in \wedge_i\}$ and $\mathcal{F}_{\wedge_i}$ is the σ algebra formed by the random variables corresponding to $\wedge_i$. Further it is assumed that $sup[E(|Y|)^{2+\delta} : s \in \wedge] < \infty$ for some $\delta > 0$. The mixing coefficient $\rho_{k,l\,p}(m)$ is defined to control the extent

of correlation between random variables at the different sites. The central limit theorem associated with these conditions are described in section 2.3.

The objective of this chapter is to construct robust statistical methodology that can simultaneously select variables and estimate the parameters of interest. Quasi-likelihood estimating equations have been proven to provide unbiased estimates of β co-efficients (Lin & Clayton, 2005). We impose a penalty to these score functions (Johnson et al., 2008), under the current model setup and study its properties.

We therefore consider obtaining an estimate for β that minimizes the objective function,

$$-\mathcal{QL}(\beta) + np_\lambda(|\beta|) \tag{2.5}$$

where $p_\lambda(|\beta|)$ is the penalty function with a tuning parameter λ. $\mathcal{QL}(\beta)$ is the quasi-likelihood function which yields the score function in equation 2.1. Thus the new penalized score function is given by,

$$\mathcal{U}^p(\beta) = \mathcal{U}(\beta) - nq_\lambda(|\beta|)sign(\beta) \tag{2.6}$$

where $q_\lambda(|\beta|) = (\partial p_\lambda(|\beta_1|)/\partial|\beta_1|sign(\beta_1), ..., \partial p_\lambda(|\beta_p|)/\partial(|\beta_p|sign(\beta_p))^T$. Therefore $\hat{\beta}$ is the solution of $\mathcal{U}^p(\beta) = 0$.

It is important to note that $\mathcal{U}(\beta)$ is a $p \times 1$ or $p_n \times 1$ dimensional vector for Case 1 and Case 2 respectively. Theoretical justifications of asymptotic properties in these scenarios vary considerably. The chapter henceforth is split in two sections addressing both cases.

2.3 Asymptotic Properties of the Penalized Quasi-Likelihood Estimator (p-fixed)

Lin (2008) developed a central limit theorem (CLT) for a random field on a lattice under L_p metrics in the increasing domain framework using the mixing conditions defined in 2.4. Under similar conditions for sampling sites $\{s_1, .., s_n\} \in \mathcal{R}^2$ on an $n = n1 \times n2$ regular grid, the consistency and asymptotic properties of the estimate obtained from solving the following system of equations are established.

$$\mathcal{U}^p(\beta) = D^T \Sigma^{-1/2} \Gamma^{-1} \Sigma^{-1/2} (Y - \mu) - n q_\lambda(|\beta|) \tag{2.7}$$

Notice that the portion of equation 2.7 associated with the penalty term is discontinuous and non-differentiable at $|\beta_j| = 0$ for some $j = \{1, ..p\}$. Therefore we study the two portions of this equation separately. Let us begin by looking at the score function that resembles equation 2.1.

Asymptotics for Score Function:

Let $S_n = \sum_{s \in \wedge_n} \{Y(s) - \mu(s)\}$ and $\sigma_n^2 = var(S_n)$ where $\wedge_n$ is a strictly increasing subsequence of lattice sets. With $\rho_{k,l;p}(m)$ as defined in 2.4,

Lemma 2.3.1. *Lin (2008)* If a random field satisfies the conditions

(i) $\rho_{1,2;p}(m) = O(m^{-d-\epsilon})$ for some $\epsilon > 0$

(ii) $\rho_{k,l;p}(m) = O(m^{-d})$ for $k + l \leq 4$

then S_n/σ_n converges in distribution to $N(0, 1)$ and $n \to \infty$.

This generalization of the CLT for weakly dependent stationary processes does not have any structural restrictions on the covariance other than (i) and (ii). It can be verified that both exponential and spherical covariances satisfy the conditions needed for CLT. Liang & Zeger (1986) have successfully adopted robust non-parametric covariance matrix estimates to replace the unknown correlation matrix in the form of a working correlation matrix and still obtained efficient unbiased estimates of the mean parameters. Therefore as long as the true underlying correlation structure as well as the misspecified correlation structure substituted in its place satisfy the ρ mixing conditions (i) and (ii) CLT holds. Thus during the implementation of the method not only can the parameters of a working correlation be misspecified, we can additionally misspecify its structure as well. Consequently, we replace the true Γ with a working correlation $\hat{\Gamma}$ that satisfies (i) and (ii). Henceforth, we investigate properties of the score function with the working correlation matrix denoted by $\mathcal{U}_w(\beta)$ as shown below,

$$\mathcal{U}_w(\beta) = D^T \Sigma^{-1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2} (Y - \mu) \tag{2.8}$$

Since $Y(s)$ follows a central limit theorem based on the above Lemma, if all elements of $D^T \Sigma^{-1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2}$ are finite, then asymptotic normality for the score function 2.8 can be established using the following assumption.

(A1) There exists a working correlation matrix $\hat{\Gamma}$ such that $n^{-1} D^T \Sigma^{-1/2} \hat{\Gamma} \Sigma^{-1/2} D \to \mathfrak{I}_0(\beta)$ as $n \to \infty$ where $\mathfrak{I}_0(\beta)$ is bounded and positive definite.

Based on the model setup in section 2.1, $\max_i |\mu(s_i)|$ and $\max_i |\mu(s_i)^{-1}|$ are finite and $\exists$ a constant $M_0 > 0$ such that $max_{ij}|D_{ij}| \leq M_0$. Also Lin (2008) showed that for a correlation matrix Γ where each element is denoted by $\gamma_{ij} = c_o||i-j||^{-d-\epsilon}$ which satisfies

the mixing conditions in 2.3.1, we obtain that each element of the inverse matrix Γ^{-1} is also bounded. Hence a constant $M_1 > 0$ may be found such that $D^T\Sigma^{-1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{-1/2}D \leq M_1 J^T\Sigma^{1/2}\Gamma\Sigma^{1/2}J$ where $J_{n\times p}$ is a matrix of 1s. Since $\sum_{j=1}^{n}\gamma_{ij} = O(1)$ we have

$$D^T\Sigma^{-1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{-1/2}D = O(n) \tag{2.9}$$

Thus we require the following assumption to establish asymptotic properties of the score function in the context of replacing the true underlying correlation with a working correlation satisfying (i),(ii) in lemma 2.3.1 and (A1).

(A2) There exists a working correlation matrix $\hat{\Gamma}$ such that

$$\tfrac{1}{n}D^T\Sigma^{-1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{-1/2}D \to \mathcal{I}_1(\beta) \text{ as } n \to \infty$$

where $\mathcal{I}_1(\beta)$ is bounded positive definite matrix. Refer to (Lin, 2010, Appendix Proof of Theorem 1) for details. Thus an adequate working correlation maybe found to satisfy condition 2.9. Let us consider an estimate $\hat{\beta}$ which is the solution to $\mathcal{U}_w(\beta) = 0$.

Lemma 2.3.2. *(Lin, 2008, Theorem 2)* Under current model setup assumptions (A1) and a working correlation satisfying (i),(ii) and equation 2.9, we get,

$$n^{-1/2}\mathcal{U}_w(\beta) \to N(0, \mathcal{I}_0(\beta)).$$

Let us now consider the asymptotic properties of the gradient matrix that is denoted by $\bigtriangledown_\beta\mathcal{U}_w(\beta)$. Specifically, $\bigtriangledown_\beta\mathcal{U}_w(\beta) = (\bigtriangledown_{\beta_1}D^T, .., \bigtriangledown_{\beta_p}D^T) \circ \Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}(Y-\mu) +$ $D^T \circ (\bigtriangledown_{\beta_1}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}), ..., \bigtriangledown_{\beta_p}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}))(Y-\mu) - D^T\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}D$. Due to the finiteness and smoothness assumptions of the link function $g(\cdot)$, it can be seen that $(\bigtriangledown_{\beta_1}D^T, .., \bigtriangledown_{\beta_p}D^T) \circ \Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}$ and $D^T \circ (\bigtriangledown_{\beta_1}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}), ..., \bigtriangledown_{\beta_p}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2})$

are both bounded vectors. A similar justification as given in (Lin, 2008, Theorem 3) with bounded vectors and the asymptotic normality of $Y(s)$ we have, a multivariate normal random variable Z^* such that $n^{-1/2}\{(\bigtriangledown_{\beta_1} D^T, .., \bigtriangledown_{\beta_p} D^T) \circ \Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}(Y-\mu) + D^T \circ (\bigtriangledown_{\beta_1}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2}), ..., \bigtriangledown_{\beta_p}(\Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2})(Y-\mu)\}$ is equivalent to $Z^* + o_p(1)$.Therefore,

$$n^{-1} \bigtriangledown_\beta \mathcal{U}_w(\beta) = n^{-1} O_p(n^{1/2}) - n^{-1} D^T \Sigma^{-1/2}\hat{\Gamma}\Sigma^{-1/2} D \tag{2.10}$$

Hence by assumption (A1) we have that,

$$n^{-1} \bigtriangledown_\beta \mathcal{U}_w(\beta) \to -\mathcal{I}_0(\beta) \text{ in probability as } n \to \infty.$$

The Inverse function theorem and an open ball argument can now be used for this score function 2.8 with a $\hat{\Gamma}$ to show that the solution $\hat{\beta}$ of the system of equations $\mathcal{U}_w(\beta) = 0$ is consistent with regard to the underlying true β_0 asymptotically (Lin, 2008, Theorem 3 Proof). We can therefore represent these results in a compact form that is exactly an assumption used in (Johnson et al., 2008, Condition C.1) for most commonly used estimating equations. The combined results of asymptotic normality of the score function, the positive definite limit in probability for the gradient function and the consistency of the $\hat{\beta}$ estimate result in the following corollary.

Corollary 2.3.2.1. There exists a positive definite matrix A, such that for a given constant $\mathcal{M} > 0$,

$$\sup_{|\beta-\beta_0| \le \mathcal{M} n^{-1/2}} |n^{-1/2}\mathcal{U}_w(\beta) - n^{-1/2}\mathcal{U}_w(\beta_0) - n^{1/2} A(\beta - \beta_0)| = o_p(1)$$

So far consistency results have been established for estimates of β that solves 2.8. However situations rise where we are interested in selecting only those relevant predictors for a relatively large number of available predictors. Incorporating a penalty term like LASSO

(Tibshirani, 1996) or SCAD (J. Fan & Li, 2001) would allow us to simultaneously select the variables and estimate them. There is a natural sparsity assumption that builds into this formulation of the penalized score function. We proceed by investigating properties of the penalty term additionally attached to 2.1.

Asymptotics for Penalty Term:

Consider the true $\beta_0 = (\beta_{01}, ..., \beta_{0p})^T$ for a fixed p. Without loss of generality, suppose $\beta_{0j} \neq 0$ for $j \leq s$ and $\beta_{0j} = 0$ for $j > s$, where s denotes the true number of non-zero βs. We now consider solving the following system of equations and investigate the relationship of between β_0 and the solution $\hat{\beta}$ obtained from solving $\mathcal{U}_w^p(\hat{\beta}) = 0$ where $\mathcal{U}_w^p(\beta)$ is given by,

$$\mathcal{U}_w^p(\beta) = D^T \Sigma^{-1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2} (Y - \mu) - n q_\lambda(|\beta|) \tag{2.11}$$

The penalty function, specifically its derivative must posses certain properties that basically ensures a consistent solution with regard to selection and estimation. Therefore we consider assumptions identical to (Johnson et al., 2008, Condition C.2).

(A3) For all nonzero θ, $lim_{n\to\infty} n^{1/2} q_{\lambda_n}(|\theta|) = 0$ and $lim_{n\to\infty} q\prime_{\lambda_n}(|\theta|) = 0$
where $q\prime_{\lambda_n}(|\theta|) \coloneqq \frac{\partial}{\partial\theta} q_{\lambda_n}(|\theta|)$

(A4) For any $M > 0$, $\lim_{n\to\infty} \inf_{|\theta| < M\sqrt{n}} q_{\lambda_n}(|\theta|) \to \infty$

Conditions (A3) particularly secures equation 2.11 whenever $\beta_{0j} \neq 0$ from being dominated by the penalty term since it vanishes i.e. $n^{1/2} q_{\lambda_n}(|\beta_{0j}|) = 0$. For $\beta_{0j} = 0$ for some $j > s$ the penalty term diverges. These properties allows us to consistently distinguish significantly large coefficients. Further, (A4) also implies that for any consistent solutions of equation 2.11 there must satisfy zero estimates i.e. $\hat{\beta}_j = 0$. An alternative and equivalent version of

consistency can be established by showing 2.22 described in section 2.4 for the case $p \to \infty$ as $n \to \infty$. Finally both these conditions ensure the oracle property of the penalty term as defined by J. Fan & Li (2001). Some examples of penalty terms that satisfy (A4) and (A4) are SCAD, Adaptive LASSO (Zou, 2006), hard-thresholding (Donoho et al., 1994) and MCP (Zhang et al., 2010). Examples of those penalty terms that do not satisfy these conditions are LASSO and Elastic-Net (EN,(Zou & Hastie, 2005)).

2.3.1 Selection Consistency and Oracle Property

Singularity at the origin:

The derivative of the penalty terms under consideration used in equation 2.11 for a given $\beta_i = 0$ is discontinuous. The sign function that results from modulus-type and indicator functions fluctuates at the origin rendering it difficult to obtain an exact zero-crossing solution. Therefore as suggested by (Johnson et al., 2008, Theorem 1a) a solution $\hat{\beta}$ is zero-crossing if,

$$\limsup_{\epsilon \to 0+} n^{-1} \mathcal{U}^p_{w_j}(\hat{\beta} + \epsilon e_j)(\mathcal{U}^p_{w_j}(\hat{\beta} - \epsilon e_j) \leq 0 \tag{2.12}$$

where e_j is the j^{th} canonical unit vector. Additionally, an approximate zero crossing estimator is defined by,

$$\limsup_{n \to \infty} \limsup_{\epsilon \to 0+} n^{-1} \mathcal{U}^p_{w_j}(\hat{\beta} + \epsilon e_j)(\mathcal{U}^p_{w_j}(\hat{\beta} - \epsilon e_j) \leq 0 \tag{2.13}$$

Consequently, the zero-crossing is an exact solution to the penalized score function whenever $\mathcal{U}^p_w$ is continuous. Using this definition we show that the system of equations maybe

solved to obtain a $\sqrt{n}$-consistent estimator with oracle properties.

Lemma 2.3.3. *Existence:* Under assumptions (i),(ii) in lemma 2.3.1, (A1)-(A4) there exists a $\sqrt{n}$ consistent approximate zero-crossing of $\mathcal{U}_w^p(\beta)$.

With an approximate zero-crossing solution we proceed to check if the estimator posses selection consistency of the true zeroes in the true sparse β co-efficient vector. Based on the sparsity of β as defined in section 2.3, we assume that there are a fixed number of true non-zero βs $s << p$. Then we show the following.

Lemma 2.3.4. *Selection Consistency:* For any $\sqrt{n}$ consistent zero-crossing solution of $\mathcal{U}_w^p(\beta)$ under assumptions (A3) and (A4),

$$\lim_{n\to\infty} P(\hat{\beta}_j = 0, j > s) = 1$$

The current estimator thus exhibits selection consistency indicating potential for quality variable selection with a working correlation. The oracle property introduced by J. Fan & Li (2001) is established for non-concave penalty functions such as SCAD that satisfies assumption (A3) and (A4). The oracle property means that the penalized estimator is asymptotically equivalent to the oracle estimator that is the ideal estimator obtained only using the true non-zero (signal) variables without subjecting it to regularization.

Theorem 2.3.5. *Oracle Property:* Under assumptions (i),(ii) and (A1)-(A4), for $s << p$, we denote $\hat{\beta}_1 = (\hat{\beta}_1, .., \hat{\beta}_s)^T$ as the estimates of the true non-zero βs denoted by $\beta_{01} = (\beta_{01}, .., \beta_{0s})^T$. Then,

$$\sqrt{n}(B_{11} + \mathcal{I}_{0_{11}})\{\hat{\beta}_1 - \beta_{01} + (B_{11} + \mathcal{I}_{0_{11}})^{-1} b_n\} \to_d N(0, V)$$

where $B_{11} = diag\{-q'_{\lambda_n}(|\beta_{01}|)sign(\beta_{01}\}$, $\mathcal{I}_{0_{11}}$ is an $s \times s$ sub-matrix of $\mathcal{I}_0$, $b_n = -(q_{\lambda_n}(|\beta_{01}|)sign(\beta_{01}), ..., q_{\lambda_n}(|\beta_{0s}|)sign(\beta_{0s}))^T$ and V is the $s \times s$ sub-matrix of $\mathcal{I}_0(\beta_0)^{-1}\mathcal{I}_1(\beta_0)\mathcal{I}_0(\beta_0^{-1})$.

Limiting conditions of assumptions (A1) and (A2) result in matrices $\mathcal{I}_0$ and $\mathcal{I}_1$. The proofs of 2.3.3, 2.3.4 and 2.3.5 are provided in the Appendix of chapter 2. The theoretical results of consistency and oracle properties of the estimator provides strong evidence of the performance of the proposed statistical methodology. We have addressed the discontinuity of the score function in this section by defining approximate zero-crossing. In general numerical methods like Newton-Raphson algorithm are used to obtain the solution for the simultaneous equations using a simple update rule. However, the penalized score function in 2.11 is also non-differentiable and a modification is therefore needed to successfully implement the method. The following section provides an update rule using the MM-algorithm (Hunter & Li, 2005) and a way to incorporate a data-driven working correlation into the rule to be used in practice.

2.3.2 Model Implementation

2.3.2.1 MM-Algorithm

Maximum likelihood estimation (MLE) have dominated the field of Statistics for a long time with a very natural theoretical intuition and efficiency. Thus they have also been very well formulated and studied. In the context of MLE, regularization methods have been extensively applied to create variable selection methods. It is known that the Quasi-likelihood functions coincide with maximum likelihood for data having distributions in the exponential family. When a penalty term is added to this setup it too faces discontinuity and non-differentiable

issues as mentioned in section 2.3.1. J. Fan & Li (2001) introduced a unified framework for the optimization using local quadratic approximations to be able to use a Newton-Raphson like update rule. This approximation for β_j with an initial starting value β_j^0 is given by,

$$[p_\lambda(|\beta_j|)]' = q_\lambda(|\beta_j|)sign(\beta_j) \approx \frac{q_\lambda(|\beta_j^0|)}{|\beta_j^0|}\beta_j$$

i.e. for $\beta_{0j} \neq 0$ and $\beta_j \approx \beta_j^0$ we have,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + \frac{q_\lambda(|\beta_j^0|)}{2|\beta_j^0|}\{\beta_j^2 - \beta_j^{0\,2}\} \tag{2.14}$$

Since 2.14 is undefined at $\beta_j^0 = 0$, the associated βs were set to 0 and the approximation is used only for the non-zero βs. Those βs that are set to 0 remain so for all of the iterations. Alternatively, Hunter & Li (2005) introduced the Majorize-Minimize (MM)- Algorithm in the context of MLE that both generalized the EM algorithm and circumvented the issue of discontinuity while generalizing the initial local quadratic approximation previously discussed as a special case of the MM-algorithm. By incorporating a perturbation to 2.14 smoothening out the denominator the function does not majorize the original piece-wise differentiable $p_\lambda(\cdot)$ but a perturbed version of it. Then they proposed that asymptotically this perturbation does not indeed affect the optimization process using a Newton-Raphson like update rule.

Similar to method implemented for MLE, the objective is to minimize the objective function 2.5. A small perturbation in the form of small $\epsilon > 0$ renders the local quadratic approximation differentiable. Therefore we construct a modified penalized score function

2.11 in the following way,

$$\mathcal{U}^p_{w,\epsilon}(\beta) = D^T\Sigma^{-1/2}\hat{\Gamma}^{-1}\Sigma^{-1/2}(Y-\mu) - nq_{\lambda,\epsilon}(|\beta|) \tag{2.15}$$

where $q_{\lambda,\epsilon}(|\beta|) = (q_\lambda(|\beta_1|+)sign(\beta_1)\frac{|\beta_1|}{\epsilon+|\beta_1|}, ..., q_\lambda(|\beta_p|+)sign(\beta_p)\frac{|\beta_p|}{\epsilon+|\beta_p|})^T$ and $q_\lambda(|\beta_j|+) = \lim_{x\to|\beta_j|+} p'_\lambda(x),\ j = 1, .., p.$ Similar to the solution for MLE provided by (Hunter & Li, 2005, Section 3.3) we can find a solution $\hat{\beta}_\epsilon$ of 2.15 that requires $p_\lambda(\cdot)$ to be piece-wise differentiable, non-decreasing, concave on $(0, \infty)$, continuous at 0 and $p'_\lambda(0+) < \infty$. Thus in order to obtain the solution one must iteratively solve,

$$\beta^{k+1}_\epsilon = \beta^k_\epsilon + [H(\beta^k_\epsilon) + nE(\beta^k_\epsilon)]^{-1}\mathcal{U}^p_{w,\epsilon}(\beta^k_\epsilon) \tag{2.16}$$

where $H(\beta^k_\epsilon) = D(\beta^k_\epsilon)^T\Sigma^{-1/2}(\beta^k_\epsilon)\hat{\Gamma}^{-1}\Sigma^{-1/2}(\beta^k_\epsilon)D(\beta^k_\epsilon)$ and
$E(\beta^k_\epsilon) = diag\{\frac{q_\lambda(|\beta^k_{1,\epsilon}|+)}{\epsilon+|\beta^k_{1,\epsilon}|}, ..., \frac{q_\lambda(|\beta^k_{p,\epsilon}|+)}{\epsilon+|\beta^k_{p,\epsilon}|}\}$
A choice for ϵ was proposed by Hunter & Li (2005) ensuring that it satisfies,

$$|\mathcal{U}^p_{w,\epsilon} - \mathcal{U}^p_w| < \tau/2$$

where τ is a very small predetermined tolerance (Ex: $\tau = 10^{-5}$). Therefore this results in,

$$\epsilon = \frac{\tau}{2np'_\lambda(0+)} min\{|\hat{\beta}^0_j| : \hat{\beta}^0_j \neq 0\} \tag{2.17}$$

Several start values of $\hat{\beta}_0$ must be employed however to ensure that the algorithm converges to a global maximum, otherwise the algorithm may oscillate near local maximas. Convergence is obtained when $\|\hat{\beta}^{k+1} - \hat{\beta}^k\| < \delta$ for some $\delta > 0$. It is important to note that the smoothening using the perturbation and update-rules do not drive down the βs to be

exactly 0, thus whenever convergence is attained, if $|\mathcal{U}_w(\hat{\beta}_j^k)| > \tau$, then $\hat{\beta}_j^k$ is set to 0.

SCAD penalty

Smoothly Clipped Absolute Deviation (SCAD, J. Fan & Li (2001)) satisfies all assumptions associated with equation 2.15, (A3)-(A4) and has been implemented in our current efforts to showcase the methodology. The explicit form of the derivative of the penalty is given by,

$$q_\lambda(\beta) = \lambda\left\{\mathbb{I}(\beta \leq \lambda) + \frac{(\mathfrak{a}\lambda - \beta)_+}{(\mathfrak{a}-1)\lambda}\mathbb{I}(\beta > \lambda)\right\} \tag{2.18}$$

A recommended value of $\mathfrak{a} = 3.7$ provided by (J. Fan & Li, 2001, Section 2.1) is used for the simulations in section 2.3.3.

2.3.2.2 Tuning Parameter Selection

In regularization methods selection of the optimal tuning parameter is extremely crucial for efficiency of the method. Cross validation (CV) techniques have proven to be fruitful in doing so. The premise of this method lies in the independence assumption of data. Spatial data disallows for such independence and therefore CV techniques seem unreliable in this paradigm. The most natural cross-validation technique that is usually used in kriging would be a jackknife-like n-fold CV estimate. However for a sequence of lambda estimates it appears to be computationally highly inefficient.

Alternative methods of using likelihood based criteria like AIC and BIC are not possible. We may have been able to calculate the modified AIC for gee by Pan (2001), however a closed form expression for the quasi-likelihood needs to be explicit is yet another barrier in using the method.Therefore we resort to a grid search of the tuning parameters over plausible values

and find the tuning parameter that satisfies an optimality criteria like minimizing prediction error. Therefore we implement the MM-algorithm with multiple starts on a sequence of tuning parameters and select the $\hat{\beta}$ with the minimum mean squared error.

2.3.2.3 Construction of Working Correlation

To implement the update rule 2.16, the working correlation $\hat{\Gamma}$ is required. Since the underlying true spatial covariance function is unknown, we use the Pearson residuals at each iteration (k). The start value $\hat{\beta}^{(0)}$ is obtained using GLM under independence. Then at each iteration of the update rule the pearson residuals is calculated and given by,

$$\hat{r}_i^{(k)} = (y(s_i) - \mu_i(\hat{\beta}^{(k)}))/\sqrt{\Sigma_{ii}(\mu(\hat{\beta}^{(k)}))}$$

The over-dispersion parameter σ^2 is then estimated using $\hat{\sigma^2}_{(k)} = \sum_{i=1}^{n}(\hat{r}_i^{(k)})^2/(n - p^{(k)})$. After which the residuals are used to fit a variogram model denoted by $\gamma(h)$ at lag h to parametric models that satisfy (i) and (ii) of lemma 2.3.1. $\hat{\Gamma}$ is then constructed by using the estimated $\hat{\gamma}(h)$ as $\hat{\Gamma}(h) = 1 - \hat{\gamma}(h)$. Fit using least squares may increase computation time significantly. Certain instabilities creep when each iteration may result in a slightly different variogram model shape. Fixing the model structure in the beginning may instill some stability in the method. Further details maybe found using this method in Feng et al. (2016) where it is employed specifically to spatial binary data.

2.3.3 Simulation

In order to investigate the effectiveness of the proposed method we consider two scenarios of responses; binary and count spatial response.

Let us first consider the binary case. Three different unit-distance grids : 15×15, 20×20

and 25×25 are considered i.e. $n = 225, 400, 625$ respectively. The number of predictors associated with the response is 20 with 3 non-zero coefficients i.e.$\beta^0_{20\times 1} = (1, -1.5, 1, 0, \ldots, 0)^T$. The corresponding design matrix $X_{20\times n}$ are generated from a Uniform(0,1) distribution. The mean of the spatial binary distribution is associated with the predictors through a *logit* link i.e. $E(y(s_i)) = \pi_i$ such that $log\frac{\pi_i}{1-\pi_i} = x_i^T\beta^0$ for $i = 1, \ldots, n$. Two spatial correlation functions are considered the power correlation $\rho^{d_{ij}}$ where $d_{ij} = \|s_i - s_j\|$ and the Matérn covariance function with parameters θ and ν. In the tables below $\tilde{\rho}$ denotes the working correlation value substituted. The binary data is generated using the archived R package *mvtBinaryEP* based on the paper by Emrich & Piedmonte (1991) that describes generating correlated binary data using tetrachoric correlation. A renewed version of the model generating function may be found in the R package *MultiOrd*. For each setup 100 data sets were simulated. The estimated mean squared error (MSE) is equal to $\frac{1}{100}\sum_{j=1}^{100}\|\hat{\beta}_j - \beta_0\|^2$. We compare results obtained from using SCAD, LASSO as the penalty function and since $p < n$ we used the score function obtained from the original $\mathcal{QL}$ method in equation 2.1.

Both tables 2.1 and 2.2 indicate that the penalized quasi-likelihood approach provides considerably better estimates compared to the quasi-likelihood approach. The choice of working correlation parameter does not have a very significant impact on estimation efficiency. Further there is a tendency for the method to over-select when there is a stronger spatial dependence. Compared to results using the SCAD penalty, the LASSO penalty provides a less sparse result yielding a more complex model with many more predictors selected but not significantly improving the MSE. Consequently, even though TP results with the LASSO penalty is slightly better, FP results is also larger than that for the case with the SCAD penalty. A noticeable improvement is observed when a misspecified parameter working correlation is incorporated in comparison to employing the method under the independence

Table 2.1: Simulation results for spatial binary data generated using the power correlation model with $\rho = 0.1$. QL represents a quasi-likelihood approach without a penalty term, PQL.LASSO represents a penalized quasi-likelihood approach with the LASSO penalty and PQL.SCAD represents a penalized quasi-likelihood approach with the SCAD penalty. PQL.LASSO.Ind and PQL.SCAD.Ind uses the identity matrix as working correlation assuming independence.

	Grid Size	MSE	TP	FP	CP
	15×15	0.10	2.70	7.31	0.85
PQL.LASSO.Ind	20×20	0.08	2.82	4.98	0.92
	25×25	0.07	2.99	3.97	0.95
	15×15	0.12	2.27	3.44	0.72
PQL.SCAD.Ind	20×20	0.11	2.60	2.37	0.88
	25×25	0.14	2.94	0.70	0.96
$\tilde{\rho} = 0.1$	Grid Size	MSE	TP	FP	CP
	15×15	0.30	2.94	14.42	–
QL	20×20	0.15	2.99	13.61	–
	25×25	0.08	3.00	12.14	–
	15×15	0.11	2.66	7.16	0.87
PQL.LASSO	20×20	0.08	2.79	4.46	0.93
	25×25	0.07	2.99	3.68	0.94
	15×15	0.12	2.23	3.17	0.69
PQL.SCAD	20×20	0.11	2.62	2.26	0.87
	25×25	0.14	2.94	0.57	0.96
$\tilde{\rho} = 0.3$	Grid Size	MSE	TP	FP	CP
	15×15	0.33	2.95	14.65	–
QL	20×20	0.15	2.99	13.53	–
	25×25	0.09	3.00	12.76	–
	15×15	0.10	2.63	6.22	0.86
PQL.LASSO	20×20	0.08	2.80	4.94	0.87
	25×25	0.06	2.99	3.54	0.93
	15×15	0.12	2.35	3.11	0.78
PQL.SCAD	20×20	0.11	2.70	1.95	0.88
	25×25	0.14	2.92	0.78	0.93

assumption i.e. using the identity matrix as the working correlation. Finally both MSE and selection results significantly improve as the sample size increases.

The proposed method with assumptions (i) and (ii) of the ρ mixing conditions in lemma 2.3.1 allow for the misspecification of parameters and the covariance structure as long as

Table 2.2: Simulation results for spatial binary data generated using the power correlation model with $\rho = 0.3$. QL represents a quasi-likelihood approach without a penalty term, PQL.LASSO represents a penalized quasi-likelihood approach with the LASSO penalty and PQL.SCAD represents a penalized quasi-likelihood approach with the SCAD penalty. PQL.LASSO.Ind and PQL.SCAD.Ind uses the identity matrix as working correlation assuming independence.

	Grid Size	MSE	TP	FP	CP
	15×15	0.10	2.58	6.59	0.91
PQL.LASSO.Ind	20×20	0.08	2.84	4.13	0.96
	25×25	0.06	2.99	4.13	0.96
	15×15	0.11	2.26	3.04	0.87
PQL.SCAD.Ind	20×20	0.10	2.67	2.12	0.90
	25×25	0.15	2.89	0.72	0.92
$\tilde{\rho} = 0.1$	Grid Size	MSE	TP	FP	CP
	15×15	0.28	2.95	14.72	–
QL	20×20	0.15	3.00	13.74	–
	25×25	0.10	3.00	12.85	–
	15×15	0.10	2.61	6.73	0.83
PQL.LASSO	20×20	0.07	2.90	4.42	0.92
	25×25	0.07	2.97	3.74	0.84
	15×15	0.11	2.16	2.86	0.78
PQL.SCAD	20×20	0.10	2.74	2.44	0.90
	25×25	0.15	2.90	0.55	0.89
$\tilde{\rho} = 0.3$	Grid Size	MSE	TP	FP	CP
	15×15	0.25	2.98	14.32	–
QL	20×20	0.13	2.98	13.39	–
	25×25	0.08	3.00	12.44	–
	15×15	0.09	2.67	5.94	0.78
PQL.LASSO	20×20	0.07	2.88	4.34	0.85
	25×25	0.06	2.99	3.64	0.93
	15×15	0.12	2.28	3.11	0.71
PQL.SCAD	20×20	0.10	2.72	1.94	0.81
	25×25	0.14	2.93	0.63	0.88

the assumptions are satisfied. Therefore in what follows, (see table 2.3) we generate spatial binary data using the Matérn correlation model,

$$C(\theta, \nu) = \frac{(\sqrt{2\nu}\theta)^{\nu} K_{\nu}(\sqrt{2\nu}\theta)}{2^{\nu-1}\Gamma(\nu)} \tag{2.19}$$

where $K_{\nu}(\cdot)$ is the modified Bessel function of second kind. The true parameters $\theta_0 = 0.7$ and $\nu_0 = 0.3$. The working correlation is substituted with the true underlying spatial correlation, one with misspecified parameters $\tilde{\theta} = 0.8$ and $\tilde{\nu} = 0.4$ and a misspecified exponential correlation structure with $\tilde{\rho} = 0.1$. The results in table 2.3 show that the performance of the method under misspecification is comparable to results obtained under the true correlation structure. The LASSO penalty still tends to over-select thus worsening the FP rate. A significant improvement in selection is observed as the same size increases in all scenarios i.e. TP rate increases and FP rate decreases as n increases.

We further explore whether multicollinearity or correlated predictors affect the performance of the proposed method. In the setup of table 2.4 we consider once again spatial binary data that is generated with the exponential covariance with $\rho = 0.3$. There is a marked breakdown of the method for significantly high correlated variables as high as 0.9.

An investigation was made as well to consider the setup which involves a mixture of correlated and independent variables with the true non-zero coefficients shared among both sets of variables. Results in table 2.5 indicate that there is slight improvement in identifying the variables thus confirming our notion that like most methods, very high multicollinearity naturally produces identifiable issues and thus may result in poorer performance.

For the second case, we consider investigating this method for count data. Similar to the binary setup we use three different unit-distance grids : 15×15, 20×20 and 25×25

Table 2.3: Simulation results when a Matérn correlation model with $\theta = 0.7$ and $\nu = 0.3$ is used to generate spatial binary data. The results are compared to misspecification of Matérn parameters $\tilde{\theta} = 0.8$ and $\tilde{\nu} = 0.4$ and misspecification of structure exponential with $\tilde{\rho} = 0.1$.

$\tilde{\theta} = 0.7$, $\tilde{\nu} = 0.3$	Grid Size	MSE	TP	FP	CP
	15×15	0.30	2.93	14.19	–
QL	20×20	0.13	2.98	13.32	–
	25×25	0.08	3.00	12.42	–
	15×15	0.10	2.78	7.58	0.84
PQL.LASSO	20×20	0.07	2.94	5.01	0.91
	25×25	0.06	2.99	3.80	0.94
	15×15	0.18	2.59	4.21	0.76
PQL.SCAD	20×20	0.15	2.71	1.38	0.77
	25×25	0.15	2.94	0.56	0.93
$\tilde{\theta} = 0.8$, $\tilde{\nu} = 0.4$	Grid Size	MSE	TP	FP	CP
	15×15	0.28	2.97	14.24	–
QL	20×20	0.14	2.98	13.24	–
	25×25	0.08	3.00	12.46	–
	15×15	0.10	2.83	7.46	0.85
PQL.LASSO	20×20	0.07	2.94	5.10	0.90
	25×25	0.06	2.99	3.88	0.95
	15×15	0.18	2.51	3.91	0.73
PQL.SCAD	20×20	0.15	2.81	1.47	0.91
	25×25	0.14	2.92	0.67	0.91
$\tilde{\rho} = 0.1$	Grid Size	MSE	TP	FP	CP
	15×15	0.31	2.98	14.29	–
QL	20×20	0.14	2.99	13.63	–
	25×25	0.09	3.00	12.43	–
	15×15	0.10	2.79	7.54	0.84
PQL.LASSO	20×20	0.07	2.95	5.39	0.93
	25×25	0.06	2.98	4.01	0.93
	15×15	0.17	2.45	2.48	0.70
PQL.SCAD	20×20	0.15	2.85	1.24	0.86
	25×25	0.15	2.86	0.71	0.89

are considered i.e. $n = 225, 400, 625$ respectively. The number of predictors associated with the response is 20 with 3 non-zero coefficients i.e.$\beta^0_{20\times 1} = (1, -1.5, 1, 0, \ldots, 0)^T$. The corresponding design matrix $X_{20\times n}$ are generated from a Uniform(0,1) distribution. The

Table 2.4: Simulation results when the covariates are correlated. Spatial binary data with a power correlation model ($\rho = 0.3$) on 20×20 grid are considered. A power correlation model was used to construct a working correlation matrix ($\tilde{\rho}$). t is the level of dependence among covariates.

LASSO	t	$\tilde{\rho}$	MSE	TP	FP
	1	0.1	0.14	2.55	6.19
	1	0.3	0.16	2.36	5.52
	3	0.1	0.19	1.72	6.75
	3	0.3	0.58	1.60	7.68
SCAD	t	$\tilde{\rho}$	MSE	TP	FP
	1	0.1	0.17	2.24	3.93
	1	0.3	0.18	1.98	3.49
	3	0.1	0.20	1.38	4.71
	3	0.3	0.20	1.28	4.31

Table 2.5: Simulation results when 10 covariates are correlated and 10 are independent. Spatial binary data with a power correlation model ($\rho = 0.3$) on 20×20 grid are considered. A power correlation model was used to construct a working correlation matrix ($\tilde{\rho}$). t is the level of dependence among covariates. $\beta_0 = (1, -1.5, 0 \ldots, 1, 0 \ldots)^T$ i.e. 2 non-zero coefficients are correlated and 1 is independent.

LASSO	t	$\tilde{\rho}$	MSE	TP	FP
	1	0.1	0.11	2.62	5.61
	1	0.3	0.12	2.62	5.59
	3	0.1	0.15	2.29	4.9
	3	0.3	0.20	2.32	5.74
SCAD	t	$\tilde{\rho}$	MSE	TP	FP
	1	0.1	0.18	2.28	2.68
	1	0.3	0.19	2.41	4.58
	3	0.1	0.25	2.13	4.63
	3	0.3	0.45	2.34	7.89

mean of the spatial Poisson distribution is associated with the predictors through a *log* link i.e. $E(y(s_i)) = \lambda_i$ such that $log(\lambda_i) = x_i^T \beta^0$ for $i = 1, \ldots, n$. The spatial correlation function considered once again is the power correlation $\rho^{d_{ij}}$ where $d_{ij} = \|s_i - s_j\|$. In study areas such as crime statistics, disease mapping spatial count data plays a very keen role. However without the existence of a known joint distribution ensuring that the moment

Table 2.6: Simulation results when the power correlation model with $\rho = 0.1$ is used to generate spatial Poisson data. The power correlation model with $\tilde{\rho}$ is used for the working correlation matrix for estimation.

$\tilde{\rho} = 0.1$	Grid Size	MSE	TP	FP
QL	15×15	0.24	2.83	14.10
	20×20	0.15	3.00	9.96
	25×25	0.07	3.00	10.74
PQL.LASSO	15×15	0.19	2.74	11.49
	20×20	0.12	3.00	7.22
	25×25	0.04	3.00	6.91
PQL.SCAD	15×15	0.24	2.95	14.31
	20×20	0.13	3.00	9.84
	25×25	0.07	3.00	10.81
$\tilde{\rho} = 0.3$	Grid Size	MSE	TP	FP
QL	15×15	0.25	2.76	13.71
	20×20	0.19	3.00	11.22
	25×25	0.07	3.00	10.84
PQL.LASSO	15×15	0.21	2.54	11.50
	20×20	0.16	3.00	9.11
	25×25	0.04	3.00	7.65
PQL.SCAD	15×15	0.24	2.87	14.52
	20×20	0.16	3.00	9.96
	25×25	0.06	3.00	10.61

conditions can hold as those assumed in the proposed method may be challenging. The review paper by Inouye et al. (2017) covers known areas of simulating correlated count data with a particular Poisson like marginal distribution using copula methods, mixed models and Gaussian random fields etc. However these methods may be able to approximate the intended model there is a fair amount of deviation that may not be easily accounted for. Therefore we resort to generating spatial Poisson data from a Poisson-Normal model with a conditional autoregressive (CAR) correlation structure for a Normal error with mean $X\beta$.

Tables 2.6 and 2.7 provide results but are not nearly as satisfactory. There is an unusual amount of selection with very high FP values. Unfortunately we attribute the failure of

Table 2.7: Simulation results when the power correlation model with $\rho = 0.3$ is used to generate spatial Poisson data. The power correlation model with $\tilde{\rho}$ is used for the working correlation matrix for estimation.

$\tilde{\rho} = 0.1$	Grid Size	MSE	TP	FP
QL	15×15	0.24	2.99	14.38
	20×20	0.11	3.00	9.69
	25×25	0.07	3.00	11.49
PQL.LASSO	15×15	0.19	2.92	12.16
	20×20	0.09	3.00	7.84
	25×25	0.05	3.00	10.74
PQL.SCAD	15×15	0.25	2.98	14.62
	20×20	0.10	3.00	10.35
	25×25	0.08	3.00	12.14
$\tilde{\rho} = 0.3$	Grid Size	MSE	TP	FP
QL	15×15	0.25	2.99	13.69
	20×20	0.13	3.00	10.68
	25×25	0.06	3.00	11.21
PQL.LASSO	15×15	0.20	2.88	11.28
	20×20	0.11	3.00	7.49
	25×25	0.03	3.00	7.35
PQL.SCAD	15×15	0.25	2.99	14.54
	20×20	0.11	3.00	9.33
	25×25	0.07	3.00	11.42

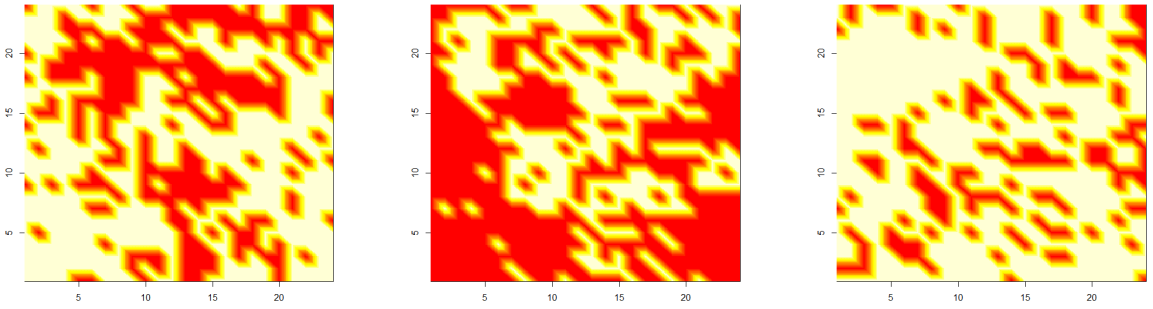
the method to the inability of generating data that satisfy the assumptions of the method, especially with respect to correlation satisfying the ρ mixing conditions.

2.3.3.1 Synthetic Analysis: Lansing Woods

To illustrate how this method may perform on real data, we use the well known Lansing woods data, originally used as application by (Lin & Clayton, 2005, Section 3). The data obtained from (Fingleton, 1986, Page 49) consists of 576 sites located on 24 X 24 grid. At each site there is a recording of whether there may be an oak, hickory or maple tree nearby in that gridded block, see Fig 2.1. Lin & Clayton (2005) tried to specifically study whether the presence of a hickory tree inhibits the presence of a maple tree. They specify that using

a conditional logistic model would be viable technique for the data set and if the spatial dependence were to be ignored then the method would asymptotically result in a chi-square test for independence (Agresti & Kateri, 2011). The estimates obtained from the using the quasi-likelihood approach for two predictors; presence of oak and hickory, with the response presence of maple lie within the the liberal estimates using a deflated chi-square test by Fingleton (1986) and the conditional binary logistic.

Figure 2.1: Lansing Woods Data on 24×24 grid showing the presence/absence of a specific tree



Maple Hickory Oak

For our purposes of implementing a variable selection technique, we considered a set of spurious covariates that include 18 arbitrary covariates generated from 9 standard normal and 9 standard uniform distributions and the two interesting indicator variables for oak and hickory. By employing a SCAD penalty and using the power correlation to construct a working correlation as described in section 2.3.2.3 we successfully discarded all artificially added variables. The estimates obtained are shown in Table 2.8.These estimates are similar to the results obtained from Lin & Clayton (2005) where the estimate for the co-efficient of the presence of Hickory (-0.26) with standard deviation (0.001), thus indicating that the method efficiently selects and estimates the parameters of interest.

Table 2.8: Lansing Woods: Estimates and 95% C.I.

Variable Name	Estimate	0.025%	0.975%
Presence of Hickory	-0.24	-0.25	-0.23
Presence of Oak	0.29	0.28	0.30

2.3.4 Real Data Analysis

In this section we discuss two examples showcasing the technique discussed above. The first example is binary response data describing the presence or absence of fire disturbances in Crawford County in Michigan, U.S. The second is an example of count data that relates to county-wise lung cancer incidence in Iowa, U.S. Results of these two studies are published in Feng et al. (2016).

The notion of distances between locations in the datasets described are computed in two alternative ways.

I Euclidean distance between the latitude and longitude coordinates in degrees.

II Geodesic distant between the latitude and longitude coordinates in kms with consideration to the curvature of the Earth's surface.

Coordinate transformation for Geodesic distance - Bearing Measurement

To convert the distances in kms with respect to the Earth's curvature, we begin by using an average radius from the center of the Earth i.e. $R = 6378.137$ kms.

Let us consider two distinct points $w_1 = (\kappa_1, \psi_1)$ and $w_2 = (\kappa_2, \psi_2)$, then we define the differences between coordinates of w_1 and w_2 which are converted to radians to be,

$\Delta Lat = (\kappa_2 - \kappa_1) * \pi/180$ and $\Delta Lon = (\psi_2 - \psi_1) * \pi/180$. Then we calculate the following,

$$a = sin(\frac{\Delta Lat}{2}) * sin(\frac{\Delta Lat}{2}) + cos(\frac{\kappa_1 * \pi}{180}) * cos(\frac{\kappa_2 * \pi}{180}) * sin(\frac{\Delta Lon}{2}) * sin(\frac{\Delta Lon}{2})$$

$$b = cos(\frac{\kappa_1 * \pi}{180}) * sin(\frac{\kappa_2 * \pi}{180}) - sin(\frac{\kappa_1 * \pi}{180}) * cos(\frac{\kappa_2 * \pi}{180}) * cos(\Delta Lon)$$
$$c = 2 * \arctan(\sqrt{\frac{a}{1-a}})$$
$$d = sin(\Delta Lon) * sin(\frac{\kappa_2 * \pi}{180})$$

We then use a, b, c, d to calculate the length denoted by $r = R * c$ and its corresponding angle, $\theta = \arctan(\frac{b}{d})$.

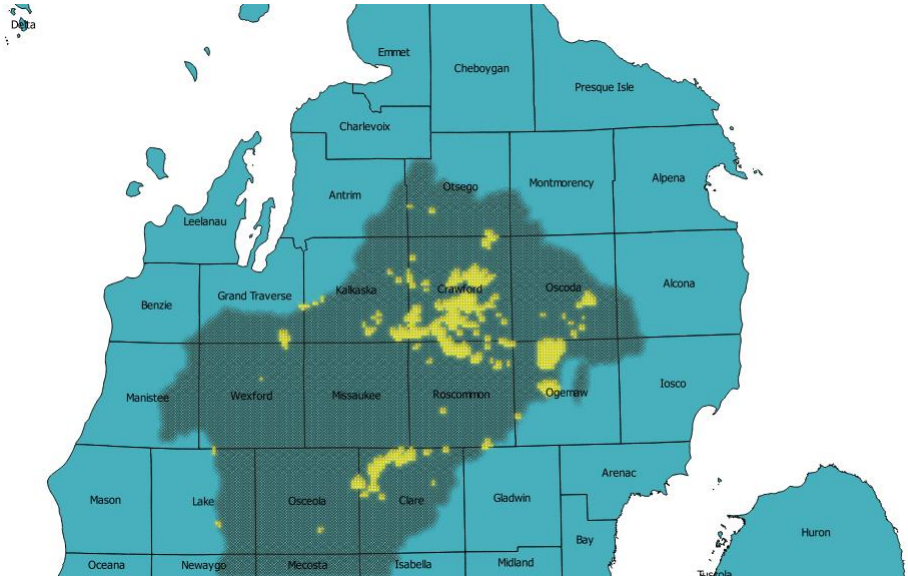
In this example consider w_1 to be the reference point, then distance r and angle θ for every other point may be obtained relative to this reference. These points are then projected into a Cartesian coordinate system such that the euclidean distance of two points in the new coordinate system is the kilometer distance between them.

2.3.4.1 Crawford County Fire Disturbances

Within the spatial framework considered so far we base the real data example on a regularly shaped political border constructed for Crawford county belonging to the high plains region of the southern peninsula of Michigan. The response variable $Y(s)$ is the presence or absence of fire disturbances in a 1.2 mile resolution with a total of 1484 locations. This data dates back to the early 1800's using satellite information and land use surveys conducted by the Michigan Natural Features Inventory associated with Michigan State University and the State Department.

The covariates or explanatory variables include several vegetation types, soil types, drainage indices and multiple topological indices at various spatial resolutions. Some preliminary interpolation techniques have been used to standardize the scale of resolution for each of these variables. To avoid issues of what is better known as separation in statistics,

Figure 2.2: Crawford County Fire Disturbance Map



we only considered those covariates that covered over 25 percent of the spatial area under investigation. Specifically, since we are looking at binary responses (presence/absence) we want to make sure that multiple combinations of the explanatory exist for both outcomes, i.e. there is sufficient variability among the covariates. We also excluded certain highly correlated vegetation types, soil types and indices (ex: correlation =0.9). After screening through these variables we had 11 covariates under consideration; three vegetation types (Jack Pine-Red Pine, mixed Coniferous swamp, Pine Barrens), one soil cover type (glacial out-wash sand and gravel and post-glacial alluvium), topographical indices (TPI) at one mile and 240 meters radius, Schaetzl's drainage index from Schaetzl (1986), water capacity of soils in each 1 mile polygon, sand cover, elevation and slope of the region. The distances between points is measured based on the geographic coordinate system (latitude and longitude) for rectangular polygons of 1.2 square miles covering the Crawford county uniformly.

After applying the variable selection method for binary responses on the presence/absence responses, imposing an power correlation structure and obtaining a working correlation as

explained in section 2.3.2.3. Two relevant variables were selected using the SCAD penalty with distances in terms of I in section 2.3.4. Table 2.9 provides the estimates and the 95% confidence intervals using the asymptotic normality obtained from the oracle property.

Table 2.9: Crawford County: Estimates and 95% C.I.

Variable Name	Estimate	0.025%	0.975%
Proportion of Pine Barrens coverage	0.0204	0.0199	0.0220
TPI 1-mile	-0.0293	-0.0318	-0.0267

Diagnostically the consulting geologist provided us evidence that the vegetation type Pine Barrens is known to be associated with presence of Fire as observed in the State of New Jersey by Givnish (1981). When compared to a regular generalized linear model for independent data it may be observed that the selected variables were indeed very statistically significant.

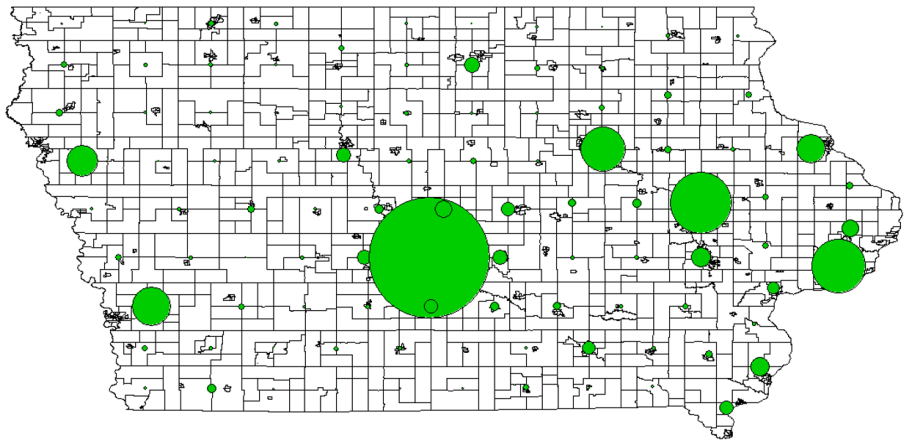
Crawford county fire disturbances (in kms) Applying the transformation II in section 2.3.4 to the coordinates of Crawford county and then performing the variable selection method with SCAD penalty as described in Section 2.3.1, exactly one variable was selected. The proportion of Pine Barrens at each location had a coefficient estimate of 0.0275 with 95% CI $(0.0248, 0.0302)$.

2.3.4.2 Lung Cancer Incidence of Counties in Iowa

In order to showcase the method with respect to spatial count data, we provide results obtained from using this selection technique to find relevant socio-economic predictors that relate to lung cancer incidence in the state of Iowa. The county-wise rate of lung cancer incidence represents a time normalized aggregate between the years 2008 and 2012. This data is combined from sources like the Iowa State Cancer Registry (SEER data, National Cancer Institute Ries et al. (2006)), State Data Center of Iowa (Census Data) and Iowa

Public Health Tracking to obtain a complete spectrum of socio-economic covariates related to each of the 99 counties in Iowa. Due to the natural regular shape of the administrative county boundaries, we used to the Geographical Coordinate System to locate the center of the counties. The distances between sites are related to these central locations. The response variable $Y(s_i)$ is the rate of incidence of lung cancer in the i^{th} county.

Figure 2.3: Lung Cancer Incidence Map of Iowa and bubble plot indicating county-wise population



The model applied to the count data was a Poisson distribution i.e. $Y(s_i) \sim Poisson(\mu_i)$ with $\mu_i = E_i e^{x_i \beta}$ where E_i is the estimated population at risk for the i^{th} county and is calculated by considering the age-adjusted risk at both the county level and population level. Details and formulae can be found in Dass et al. (2012).

Among the 33 original variables that were collected, a screening process was implemented to discard repetitive and highly correlated variables. From the 19 variables that were deemed suitable for the analysis, we applied the proposed method and used a power correlation structure based on the empirical variogram with the iterative working correlation estimation algorithm as described in section 2.3.2.3. This resulted in the selection of three variables.

Additional diagnostics using the Lung Cancer data indicated that under the independence

Table 2.10: Iowa Lung Cancer Incidence Rate: Estimates and 95% C.I.

Variable Name	Estimate	0.025%	0.975%
Percentage population below 18 years of age	-0.0806	-0.0873	-0.0739
Percentage population with education less than 9th grade	0.0652	0.0583	0.0722
Percentage population that moved within the same county	-0.0675	-0.0743	-0.0607

assumption if a generalized linear model approach is considered only the selected variables are significant. If the analysis is only performed on the selected variables the likelihood ratio test determined that the two models one with all variables present and the second with only the selected variables were statistically similar.

Iowa state County-Wise Lung Cancer Data (in kms)

Applying the geodesic transformation as shown in II in section 2.3.4 and then performing the variable selection method proposed with a SCAD penalty and a corresponding log link function for the corresponding Poisson rate data, the same three variables were selected and their estimates are shown in the table 2.11.

Table 2.11: Iowa Lung Cancer Incidence Rate: Estimates and 95% C.I.

Variable Name	Estimate	0.025%	0.975%
Percentage population below 18 years of age	-0.0925	-0.0928	-0.0922
Percentage population with education less than 9th grade	0.09404	0.0937	0.09438
Percentage population that moved within the same county	-0.0839	-0.0842	-0.0835

In the following section, we build on the existing setup to try and solve issues of selection whenever the predictor space $p_n \to \infty$ as $n \to \infty$ and create a premise to solve high-dimensional problems in the discrete spatial paradigm.

2.4 Asymptotic Properties of the Penalized Quasi-Likelihood Estimator (p - expanding dimension)

2.4.1 Score Function Asymptotics

Until now we have demonstrated the need for variable selection techniques for applications using discrete spatial data with fixed p number of covariates. A natural extension of this method is the case where the number of predictors in the design matrix X increases with the size of the data. Let us denote the dimension of the new design matrix as $p_n \times n$. These predictors are then regressed on to a response $Y(s)$ on different sites $\{s_1, ..., s_n\} \in \mathcal{R}^d$ as discussed earlier in section 2.1. In this portion we will study the asymptotics of an estimator $\hat{\beta}_n$ obtained from solving generalized estimating equations where $p_n \to \infty$ as $n \to \infty$. The parameters of interest β_n are suffixed with an n to distinguish the solution from the fixed p case. We continue to assume that the working correlation used in the score function and the true underlying spatial covariance satisfy the conditions on $\rho-$mixing. It is also important to note that no asymptotics for the score function within the spatial architecture has been established, to the best of our knowledge. Therefore we begin by first studying the properties of the score function for $p_n \to \infty$.

We characterize the score function in the following way,

$$\mathcal{U}_n(\beta_n) = \frac{1}{\sqrt{n}} X^T \Sigma^{1/2}(\beta_n) \Gamma^{-1} \Sigma^{-1/2}(\beta_n)(Y - \mu(\beta_n)) \tag{2.20}$$

Unlike the original score function seen in equation 2.1, we consider only those smooth link functions $g(\cdot)$ such that $D_{ij} = \frac{\partial \mu_i}{\partial \beta_j} = X\Sigma$. It is known that responses $Y(s)$ from a marginal canonical exponential family share this property. The corresponding score function in which a working correlation $\hat{\Gamma}$ with properties similar to those described in section 2.3 is denoted by $\hat{\mathcal{U}}_n(\beta_n)$.

Further in L. Wang et al. (2012) a decomposition is shown of the derivative of 2.20 in the supplementary material with derivative of matrix product properties specified in Pan (2002).

Lemma 2.4.1. The gradient of the score function can be decomposed as,

$$\begin{aligned} \frac{\partial \mathcal{U}_{nk}(\beta_n)}{\partial \beta_n^T} &= \mathcal{H}_{nk}(\beta_n) + \mathcal{E}_{nk}(\beta_n) + \mathcal{G}_{nk}(\beta_n) \\ &\coloneqq - \bigtriangledown_\beta \, \mathcal{U}_{nk} \end{aligned} \tag{2.21}$$

where, $\mathcal{U}_{nk}(\beta_n) = e_k^T \mathcal{U}_n(\beta_n)$, e_k is a p_n dimensional basis vector with the k^{th} element equal to 1 and

$$\begin{aligned} \mathcal{H}_{nk}(\beta_n) &= -\frac{1}{\sqrt{n}} e_k^T X^T \Sigma^{1/2}(\beta_n) \Gamma^{-1} \Sigma^{1/2}(\beta_n) X \\ \mathcal{E}_{nk}(\beta_n) &= -\frac{1}{2\sqrt{n}} e_k^T X \Sigma^{1/2}(\beta_n) \Gamma^{-1} \Sigma^{-3/2}(\beta_n) \mathcal{C}(\beta_n) \mathcal{D}(\beta_n) X \\ \mathcal{G}_{nk}(\beta_n) &= \frac{1}{2\sqrt{n}} e_k^T X \Sigma^{1/2}(\beta_n) \mathcal{D}(\beta_n) \mathcal{J}(\beta_n) X \end{aligned}$$

for $\mathcal{D}(\beta_n) = diag(\ddot{\mu}(X_1^T \beta_n), .., \ddot{\mu}(X_n^T \beta_n))$, $\mathcal{C}(\beta_n) = diag(Y_1 - \mu_1, .., Y_n - \mu_n)$ and $\mathcal{J}(\beta_n) = diag(\Gamma^{-1} \Sigma^{-1/2}(\beta_n)(Y - \mu(\beta_n)))$.

In order to show that there exists a sequence of roots $\hat{\beta}_n$ of $\hat{\mathcal{U}}_n(\beta_n) = 0$ such that

$\|\hat{\beta}_n - \beta_n\| = O_p(\sqrt{p_n/n})$ it is sufficient to show that $\forall \epsilon > 0$ there exists a constant $\Delta > 0$ such that for n sufficiently large,

$$P(\sup_{\|\beta_n - \beta_{n0}\| = \Delta\sqrt{p_n/n}} (\beta_n - \beta_{n0})^T \hat{\mathfrak{U}}_n(\beta_n) < 0) \geq 1 - \epsilon \tag{2.22}$$

This result follows from results similar to GEE in longitudinal models for clustered binary data in (L. Wang et al., 2012, Section 3). Before we formally verify 2.22 we require certain regularity conditions to hold. The absence of the summation for the score function illustrates the case of a single subject study with the cluster size going to infinity and the cluster exhibits spatial autocorrelation, in the context of longitudinal studies.

2.4.1.1 Regularity Conditions

Below are some regularity conditions that appear most commonly in longitudinal data analysis literature using score functions and GEE's in L. Wang (2011), Xie et al. (2003) and Balan et al. (2005).

(A5) $sup_{i,j}|X_{ij}| = O(\sqrt{p_n})$

(A6) β_n parameters belong to a compact subset $\mathcal{B} \subseteq \mathcal{R}_n^{p_n}$, and the true unknown parameter belongs to the interior of $\mathcal{B}$ and there exists a positive constant such that $0 < b_1 < g(\beta_{n0})$.

(A7) There exists two positive constants c_1 and c_2 for some $\alpha \in [0, 1]$ such that, $c_1 \leq \lambda_{min}(\frac{1}{n^\alpha} X^T X) \leq \lambda_{max}(\frac{1}{n^\alpha} X^T X) \leq c_2$.

(A8) Let $\mathcal{S}_n = \{\beta_n : \|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}\}$, then $\dot{\mu}(X\beta_n)$ is uniformly bounded above and below by positive constants and $\ddot{\mu}(X\beta_n)$ and $\dddot{\mu}(X\beta_n)$ are uniformly bounded by

a finite positive constant M' on $\mathcal{S}_n$.

(A9) $\|\hat{\Gamma}^{-1} - \Gamma^{-1}\| = O_p(\sqrt{p_n/n})$ where both $\hat{\Gamma}$ and Γ satisfy $\rho-$ mixing conditions and are positive definite.

(A10) Let $r = \Sigma^{-1/2}(Y - \mu)$ and $E(r^T r)/n < M^* \ \forall n > N_\epsilon$.

Assumptions (A5) and (A7) are used in the study of asymptotics of M-estimators with large p in Portnoy (1985) and imposes the condition that $X^T X$ is positive definite. Assumption (A6) ensures the minimum eigenvalue of Σ with diagonal entries that are a function of g have a positive lower bound i.e. $0 \leq b_3 \leq \lambda_{min}(\Sigma(\beta_{n0}))$. Assumption (A8) is identical to (A6) in L. Wang et al. (2012) is common in GEE literature and holds true for the log and logit link functions. Condition (A9) was shown to be true in longitudinal data with the robust non-parametric working correlation in (L. Wang, 2011, Example 2). In this case we require a similar assumption with the structure of the covariances satisfying (i) and (ii) in lemma 2.3.1.

Remark 1. It is also necessary to highlight that the asymptotics presented below are valid under the increasing domain of the spatial random field. Bachoc & Furrer (2016) under this increasing domain framework establish a lower bound on the minimum eigen value of the covariance matrix that is away from zero. This is a vital consequence that is necessary for the asymptotic results that follow. (Lin, 2008, Section 2) shows algebraically that the maximum eigen value of the inverse correlation matrix can be bounded above.

Assumption (A10) is necessary to control the sum of variance of all pairwise residuals of a naturally unbounded random variable.

In order to show the consistency of the estimator obtained from solving 2.20, below we note some common properties of matrix theory that are essential. For some, $\mathcal{C} > 0$

(Note 1) Consider a symmetric positive definite matrix A with dimension $n \times n$,

then $\lambda_{max}(A) \leq \mathcal{C} \cdot n$.

(Note 2) Whenever A is invertible, $-\frac{1}{\mathcal{C} \cdot n} \geq -\lambda_{min}(A^{-1})$

The inclusion of the working correlation and how it relates to the original score function with regard to the expanding dimension of p_n requires investigation. Therefore based on the assumptions above we have the lemma below.

Lemma 2.4.2. Under assumptions (A5)-(A10) and $p_n^2 n^{-1} = o(1)$ we have

$$\|\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})\| = O_p(n^{\alpha/2}\sqrt{p_n})$$

Similar results need to be established for terms based on the decomposition in lemma 2.4.1 with respect to the working correlation and therefore we have the following lemma.

Lemma 2.4.3. Assume the conditions (A5)-(A10). If $n^{-1}p_n^2 = o(1)$ then $\forall \Delta > 0$, for $b_n \in \mathcal{R}^{p_n}$, we show

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} \sup_{|b_n\|=1} |b_n^T[\bigtriangledown_\beta \mathcal{U}_n(\beta_n^*) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*)]b_n| = O_p(n^\alpha \sqrt{p_n})$$

Similar to the identities established in (L. Wang, 2011, Section 3) we require to show the following to establish relationships between terms in the decomposition 2.4.1 where β_n^* lies between β_n and β_{n0}.

Lemma 2.4.4. Assume the conditions (A5)-(A10). If $n^{-1}p_n^2 = o(1)$ then $\forall \Delta > 0$, we show

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} |(\beta_n - \beta_{n0})^T[\hat{\mathcal{H}}_n(\beta_n^*) - \hat{\mathcal{H}}_n(\beta_{n0})](\beta_n - \beta_{n0})| = o_p(p_n^2 n^{\alpha - 3/2})$$

Lemma 2.4.5. Assume the conditions (A5)-(A10). If $n^{-1}p_n^4 = o(1)$ then $\forall \Delta > 0$, we show

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} (\beta_n - \beta_{n0})^T [\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*) - \hat{\mathcal{H}}(\beta_n^*)](\beta_n - \beta_{n0}) = o_p(p_n)$$

Proof of the lemmas above can be found in the Appendix section of chapter 2. Based on all the notes and conditions considered above we therefore have the following theorem that provides both the existence and consistency of the estimator obtained from solving the system of equations in 2.20.

Theorem 2.4.6. *Existence and Consistency*
Assume (A5)-(A10), for $\alpha \in (0.5, 1]$ and let $p_n^4 n^{-1} = o(1)$, then $\hat{\mathcal{U}}_n(\beta_n) = 0$ has a root $\hat{\beta_n}$ such that $\|\hat{\beta_n} - \beta_{n0}\| = O_p(\sqrt{p_n/n})$.

Proof. On the set $\mathcal{S}_n = \{\beta_n : \|\beta_n - \beta_{n0}\| = \Delta\sqrt{p_n/n}\}$, we would like to show 2.22, therefore we begin by examining the Taylor expansion,

$$\begin{aligned}
(\beta_n - \beta_{n0})^T \mathcal{U}_n(\beta_n) &= (\beta_n - \beta_{n0})^T \mathcal{U}_n(\beta_{n0}) - (\beta_n - \beta_{n0})^T \bigtriangledown \mathcal{U}_n(\beta_n^*)(\beta_n - \beta_{n0}) \\
&\coloneqq I_{n1} + I_{n2} \\
\text{where, } I_{n1} &= (\beta_n - \beta_{n0})^T \hat{\mathcal{U}}_n(\beta_{n0}) + (\beta_n - \beta_{n0})^T [\mathcal{U}_n(\beta_{n0}) - \hat{\mathcal{U}}_n(\beta_{n0})] \\
&\coloneqq I_{n11} + I_{n12}
\end{aligned}$$

From lemma 2.4.2 and $n^{-1}p_n^2 = o(1)$ we have,
$|I_{n12}| \leq \|\beta_n - \beta_{n0})\| \|\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})\| = \Delta\sqrt{p_n/n} O_p(n^{\alpha/2}\sqrt{p_n}) = \Delta o_p(p_n n^{(\alpha-1)/2}) \leq \Delta o_p(p_n)$

The last inequality holds true due to $\alpha \in [0,1]$. For I_{n11} consider,

$$\begin{aligned} E(\|\hat{\mathcal{U}}_n(\beta_{n0}\|^2) &= \frac{1}{n}(Y-\mu)^T\Sigma^{-1/2}\hat{\Gamma}^{-1}\Sigma^{1/2}XX^T\Sigma^{1/2}\Gamma^{-1}\Sigma^{-1/2}(Y-\mu)) \\ &\leq \lambda_{max}(XX^T)\lambda_{max}(\hat{\Gamma}^{-2})\lambda_{max}(\Sigma)\frac{1}{n}E(r^Tr) \\ &\leq \mathrm{Tr}(X^TX)\lambda_{max}(\hat{\Gamma}^{-2})\lambda_{max}(\Sigma)\frac{1}{n}E(r^Tr) \\ &\leq Cp_n \cdot M^*n = C'np_n \end{aligned}$$

$$\implies |I_{n11}| \leq \|\beta_n - \beta_{n0}\|\|\hat{\mathcal{U}}_n(\beta_{n0})\| \leq k\Delta\sqrt{p_n/n}\sqrt{np_n} \text{ for some } k>0, \text{ then } |I_{n1}| \leq \Delta p_n.$$

Now consider I_{n2} rewritten with working correlation,

$$\begin{aligned} I_{n2} &= -(\beta_n-\beta_{n0})^T \bigtriangledown_\beta \hat{\mathcal{U}}(\beta_n^*)(\beta_n-\beta_{n0}) - (\beta_n-\beta_{n0})^T[\bigtriangledown_\beta \mathcal{U}_n(\beta_n^*) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*)](\beta_n-\beta_{n0}) \\ &\coloneqq I_{n21} + I_{n22} \end{aligned}$$

where,

$$\begin{aligned} I_{n21} &= -(\beta_n-\beta_{n0})^T \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*)(\beta_n-\beta_{n0}) \\ &= -(\beta_n-\beta_{n0})^T[\hat{\mathcal{H}}_n(\beta_{n0})](\beta_n-\beta_{n0}) - (\beta_n-\beta_{n0})^T[\hat{\mathcal{H}}_n(\beta_n^*) - \hat{\mathcal{H}}_n(\beta_{n0})](\beta_n-\beta_{n0}) \\ &- (\beta_n-\beta_{n0})^T[\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*) - \hat{\mathcal{H}}_n(\beta_n*)](\beta_n-\beta_{n0}) \\ &\coloneqq I_{n211} + I_{n212} + I_{n213} \end{aligned}$$

Evaluating each term in the following way we find,

$$\begin{aligned} I_{n211} &= -(\beta_n-\beta_{n0})^T[\mathcal{H}_n(\beta_{n0})](\beta_n-\beta_{n0}) \\ &\leq -\frac{1}{\sqrt{n}}\lambda_{min}(\Gamma^{-1})\lambda_{min}(X^TX)\|\beta_n-\beta_{n0}\|^2\lambda_{min}(\Sigma) \end{aligned}$$

$$\leq -\tilde{c}\Delta^2 n^{\alpha}\frac{p_n}{n^{3/2}}$$

Thus $|I_{n211}| = o(1)$ for $\alpha \in (0,1)$ applying assumptions (A7), (A6) and (Note 2). Further we see from lemma 2.4.4 that $|I_{n212}| = o_p(p_n^2 n^{\alpha-3/2})$ and from lemma 2.4.5 that $|I_{n213}| = o_p(p_n)$ and $I_{n22} = o_p(p_n^{3/2} n^{\alpha-1})$

For a choice of $\alpha = (0.5, 1)$ we can therefore see $(\beta_n - \beta_{n0})^T \mathcal{U}_n(\beta_n)$ on the set $\mathcal{S}_n$ is asymptotically dominated in probability by I_{n2} which is negative for a large enough Δ. ■

In order to explore the distributional behavior of the score function equation 2.20, we build up on conditions considered in the fixed p case. Therefore consider the following,

Lemma 2.4.7. If $p_n^4 n^{-1} = o(1)$ and assumptions (i),(ii), (A3)-(A10) then for $\|a\| = 1$

$$\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{U}}_n(\beta_{n0}) \coloneqq \frac{1}{n^{(1+\alpha)/2}} a^T X^T \Sigma^{1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2} (Y - \mu) \rightarrow_d N(0, a^T \mathcal{I}_2(\beta_{n0}) a)$$

as $n \to \infty$ where $\mathcal{I}_2(\beta_{n0})$ is a bounded positive definite matrix.

Sketch of Proof:

Since $Y(s)$ has a CLT from Lin (2008) as shown in section 2.3, with the assumptions of ρ mixing conditions we can show that the linear combination of the score function is indeed asymptotically normally distributed if the variance of the score function has a positive definite limit.

$$Var(\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{U}}_n(\beta_{n0})) = \frac{1}{n^{\alpha+1}} a^T X^T \Sigma^{1/2} \hat{\Gamma}^{-1} \Gamma \hat{\Gamma}^{-1} \Sigma^{1/2} X a \qquad \coloneqq a^T \hat{\mathcal{V}}_n a \qquad (2.23)$$

We can further show that,

$$\frac{1}{n^{\alpha+1}}a^T X^T \Sigma^{1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{1/2}Xa \leq \frac{1}{n^{\alpha+1}}a^T \lambda_{max}(X^T X)\Sigma^{1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{1/2}a$$
$$\leq \mathcal{K}^* \frac{1}{n}a^T \Sigma^{1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{1/2}a$$

due to assumption (A7). Condition (A1) in section 2.3 provides a positive definite limit in the special case of the identity link function. Consequently, we have,

$$\mathcal{K}^* \frac{1}{n}a^T \Sigma^{1/2}\hat{\Gamma}^{-1}\Gamma\hat{\Gamma}^{-1}\Sigma^{1/2}a \to a^T \mathcal{I}_2(\beta_{n0})a \text{ as } n \to \infty \tag{2.24}$$

where $\mathcal{I}_2(\beta_{n0})$ is a positive definite matrix. Thus we are able to establish the asymptotic distribution of score function 2.20. As a consequence we also have,

Corollary 2.4.7.1. $\frac{1}{n^{\alpha/2}}a^T \hat{\mathcal{V}}_n^{-1/2}\hat{\mathcal{U}}_n(\beta_{n0}) \sim N(0,1)$ as $n \to \infty$ for all $a \in \mathcal{R}^{p_n}$ and $\|a\| = 1$.

We would now like to establish asymptotic normality of the GEE estimator $\hat{\beta}_n$ obtained from solving 2.20. Following along with results similar to (L. Wang, 2011, Theorem 3.8) we can show that,

Theorem 2.4.8. *Asymptotic Normality* Under assumptions (i),(ii) (A1),(A5)-(A10), for $\alpha = 1$, $n^{-1}p_n^4 = o(1)$ and $\|a\| = 1$, as $n \to \infty$ we have,

$$\frac{1}{n^{\alpha/2}}a^T \hat{\mathcal{V}}_n^{-1/2}\hat{\mathcal{H}}_n(\beta_{n0})(\hat{\beta}_n - \beta_{n0}) \to_d N(0,1)$$

Proof. Consider the following,

$$\frac{1}{n^{\alpha/2}}a^T \hat{\mathcal{V}}_n^{-1/2}\hat{\mathcal{U}}_n(\beta_{n0}) = \frac{1}{n^{\alpha/2}}a^T \hat{\mathcal{V}}_n^{-1/2}\mathcal{U}_n(\beta_{n0}) + \frac{1}{n^{\alpha/2}}a^T \hat{\mathcal{V}}_n^{-1/2}[\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})] \tag{e.1}$$

$$= \frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2} \bigtriangledown_\beta \mathcal{U}_n(\beta_n^*)(\hat{\beta}_n - \beta_{n0}) + \frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})]$$

$$= \frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2} \hat{\mathcal{H}}_n(\beta_{n0})(\hat{\beta}_n - \beta_{n0}) + \frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\bigtriangledown_\beta \mathcal{U}_n(\beta_n^*) - \hat{\mathcal{H}}_n(\beta_{n0})](\hat{\beta}_n - \beta_{n0})$$

$$+ \frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})]$$

The second inequality is due to the Taylor expansion of $\mathcal{U}_n(\beta_{n0})$ where $\mathcal{U}_n(\hat{\beta_n}) = 0$. Since the LHS of equation e.1 is asymptotically normal from lemma 2.4.7, it suffices to show,

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} |\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0})](\hat{\beta}_n - \beta_{n0})| = o_p(1) \tag{T.1}$$

$$\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})] = o_p(1) \tag{T.2}$$

The proof for statements T.1 and T.2 have been redirected to the Appendix of chapter 2. ■

The theoretical properties of the score function under the expanding dimension indeed seem to preserve properties of consistency and asymptotic normality, however only under the condition that the number of predictors p_n expands very slowly with respect to n i.e. $p_n^4.n^{-1} = o(1)$. In reality, data in exploratory studies often have a large number of covariates and the goal is to extract only interesting covariates that are significantly correlated to the response. In the context of expanding variates, even though the initial number of parameters may not be very large, if one begins to include interactions of variables with enough degrees of freedom allocated, the number of covariates tends to increase with the sample size wherein only a fraction of the considered covariates may have an important effect. Thus a penalized approach where only s_n number of true non-zero coefficients associated with the predictors can be successfully identified with a natural sparsity assumption may be a more effective method.

2.4.2 Penalized Score Function Asymptotics

The method described in the rest of this section emulates established results in longitudinal data analysis by L. Wang, Zhou, & Qu, 2012 but under the current setup. We will proceed by using the above formulations specifically for $\alpha = 1$. The new set of penalized estimating equations that has a normalizing factor similar to what is seen for consistency results of expanding dimensions of parameters of a response belonging to an exponential family by Strawderman & Tsiatis (1996), is under consideration and is given by,

$$\hat{\mathfrak{U}}_n^p(\beta_n) = \frac{1}{n} X^T \Sigma^{1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2} (Y - \mu(\beta_n)) - q_{\lambda_n}(|\beta_n|) sign(\beta_n) \tag{2.25}$$

where q_{λ_n} is the derivative of the penalty function described for equation 2.6 that satisfies assumptions (A3) and (A4). The estimator $\hat{\beta}_n$ is obtained by solving the equations in 2.25. Once again we can see that $\hat{\mathfrak{U}}_n^p(\beta_n)$ is discontinuous at 0 for $\beta_{nj} = 0$ for any j. Thus we consider an approximate solution $\hat{\mathfrak{U}}_n^p(\beta_n) = o(\omega_n)$ where $\omega_n \to 0$ as $n \to \infty$. Due to the discontinuity, once again the MM-algorithm may be implemented as shown in section 2.3.2.

Let us denote the true $\beta_{n0} = (\beta_{n10}^T, \beta_{n20}^T)^T$ where without loss of generality we can assume that the first s_n coefficients are associated with β_{n10}; the truly non-zero coefficients and the remaining $p_n - s_n$ predictors associated with β_{n20} are the truly zero coefficients. Based on the results obtained in 2.4.1 we require $s_n = o(n^{1/4})$. Therefore we include one more assumption with regard to the penalty function similar to (A7) in L. Wang et al. (2012),

(A11) $\min_{1 \le j \le s_n} |\beta_{n0j}|/\lambda_n \to \infty$, $\lambda_n \to 0$, $\frac{n\lambda_n^2}{log(n)^2} \to \infty$, $s_n^4 n^{-1} \to 0$, $\frac{s_n^2 log(n)}{\lambda_n \sqrt{n}} = o(1) \implies$
$\frac{s_n^{3/2} log(n)}{\lambda_n \sqrt{n}} = o(1)$, $\frac{p_n s_n^3 log^2(n)}{n^2 \lambda_n^2} = o(1)$ and $\frac{p_n s_n^3 log^4(n)}{n^2 \lambda_n^4} = o(1)$ as $n \to \infty$.

A stronger modification of the design matrix is needed to be made in the form of the following

assumption,

(A12) X_{ij} are uniformly bounded where $i = 1, \ldots, n \quad j = 1, \ldots, p_n$.

In order to establish the selection consistency of the score function 2.25, we would like to as in (L. Wang et al., 2012, Theorem 1) show the following where the associated penalty function is the SCAD penalty (J. Fan & Li, 2001) with details provided in section 2.3.3,

Theorem 2.4.9. Under assumptions (i),(ii) in lemma 2.3.1 and (A1)-(A12) there exists an approximate solution $\hat{\beta}_n$ that solves equation 2.25 such that,

$$P\left(|\hat{\mathfrak{U}}^p_{nj}(\hat{\beta}_n)| = 0, \qquad j = 1, \ldots, s_n\right) \to 1 \tag{2.26}$$

$$P\left(|\hat{\mathfrak{U}}^p_{nj}(\hat{\beta}_n)| \leq \frac{\lambda_n}{log(n)}, \qquad j = s_n + 1, \ldots, p_n\right) \to 1 \tag{2.27}$$

as $n \to \infty$.

Properties 2.26 and 2.27 characterize the solution of the estimating equations in 2.25. This technique is an alternative approach to the zero-crossing solution defined by Johnson et al. (2008). The oracle properties now easily follow from results obtained in the un-penalized version of the score function 2.28 and can be easily derived from lemma 2.4.7 with a minor adjustment to the normalizing factor,

Corollary 2.4.9.1. Under assumptions (i), (ii), (A4)-(A12) and $\alpha = 1$ we have,
$P(\hat{\beta}_{n2} = 0) \to 1$ and $\forall a \in \mathcal{R}^{p_n}$ such that $\|a\| = 1, a^T \hat{\mathcal{V}}_n^{-1/2} \hat{\mathcal{H}}_{n1}(\beta_{n0})(\hat{\beta_n} - \beta_{n0}) \to_d N(0, 1)$ as $n \to \infty$.

Now that the oracle properties of the estimator are established we need to deduce the order in which the number of covariates p_n is increasing. It is rather realistic to suppose

a fixed dimension for the underlying true model i.e. s_n is fixed. (L. Wang et al., 2012, Remark 3) in provides the order of the λ_n and p_n. Similarly using conditions imposed in assumption (A11) we have the following remark.

Remark 2. Based on conditions assumed in (A11), the above results are applicable to $p_n = o(n^{\eta})$ where $0 < \eta < \frac{3}{2}$ restricted to $\lambda_n = O(n^{-\nu})$ where $0 < \nu < \frac{2-\eta}{4}$ satisfying all impositions in (A11).

Therefore similar to results obtained in the longitudinal case we are indeed able to utilize the proposed method for $p_n = o(n)$. However in order to truly address statistical problems of high-dimensional spatial regression with dimensions of exponential order with respect to sample size, we may require a stronger theoretical justification.

2.5 Discussion

This chapter establishes a variable selection method under the framework of discrete spatial observations. It provides foundational scope for high dimensional variable selection and bridges a gap in statistical methodology that has not been previously explored. Using the GEE formulation, well studied in longitudinal data analysis we have provided asymptotic theory of the estimators that solve penalized quasi-likelihood score equation 2.11 with rigor. We begin by considering the fixed $p < n$ scenario. Under a slightly strong assumption of ρ mixing co-efficients (i) and (ii) we are able to obtain selection consistency and oracle properties of the estimator. Simulation results in the binary response case appear efficient and perform well under covariance misspecification. To justify the technique, we investigate performance of the method under various scenarios. The synthetic example provides evidence of the method with results being validated by the real data analysis in Lin & Clayton (2005).

In general during the implementation of the proposed method and the MM-algorithm, an improvement can be made with regard to tuning parameter selection. Without the availability of a likelihood expression, regular criteria like AIC and BIC cannot be used. Based on evidence provided by J. Fan & Li (2004) and Johnson et al. (2008) for selection in semi-parametric we adopt a variation of the generalized cross-validation (GCV) statistic but on a sequential grid of tuning parameters. More efficient methods may be explored to cater to spatial dependence.

With regard to simulations related to spatial count response, there needs to be a more meticulous exploration of successfully generating multivariate Poisson models. The review article Inouye et al. (2017) provides well established methods for pair-wise dependence. Certain copula based techniques can be used to explore the method as one suggested by

Yahav & Shmueli (2012) who were successful in approximately obtaining correlated Poisson with a small bias but was restrictive in regard to the amount of correlation that can be imposed. The method for generation of binary data too suffered from being unable to incorporate a higher level of dependence.

We then proceed to establish results for expanding dimensions with a rate i.e. $(p_n^4.n^{-1} = o(1))$ which is much slower than the usual rate for longitudinal data i.e. $(p_n^2.n^{-1} = o(1))$ established by L. Wang (2011). Both the consistency and asymptotic normality of the estimator are established under certain regularity condition 2.4.1.1. What is most noticeable here is that a normalizing factor as seen in Strawderman & Tsiatis (1996) was required to achieve these results. In the spatial paradigm no such results exist unlike the ones that exist and have been used in the fixed p-case. Focusing on the goal of selection we then use a penalized approach similar to that of L. Wang et al. (2012) and obtain consistency and oracle properties of the estimator. In its current form the theory allows for $p_n = o(n)$ but we eventually seek to look at modifications of the method whenever the dimensions increase exponentially.

In order to fully study and investigate these methods we seek to perform a series of simulation studies in the future. The theoretical results suggest that these methods may be successful in the $p_n = n$ case but may require additional theoretical backing for a high dimensional setup. Therefore having formulated these basic ideas there is tremendous scope for future work of our current model setup.

APPENDIX

APPENDIX

Fixed Dimension P- Proofs

Lemma: *2.3.3*

Proof. Since $\hat{\beta}$ obtained from solving 2.11 is $\sqrt{n}$ consistent i.e. $\hat{\beta} = \beta_0 + O_p(n^{-1/2})$, for all non-zero $\beta' s$ $j = 1, .., s$ under assumption (A3) and lemma 2.3.2 we get,

$$\limsup_{\epsilon \to 0+} n^{-1/2} \mathcal{U}^p_{w_j}(\hat{\beta} \pm \epsilon e_j) = \limsup_{\epsilon \to 0+} \left\{ n^{-1/2} \mathcal{U}^p_{w_j}(\hat{\beta} \pm \epsilon e_j) - \sqrt{n} q_{\lambda_n}(|\hat{\beta} \pm \epsilon e_j|) \right\} \xrightarrow{p} 0$$

as $n \to \infty$.

Under assumption (A4) for $j > s$, $\sqrt{n} q_{\lambda_n}(\epsilon e_j)$ dominates in equation 2.11 with opposing signs for $\beta' s$ that are necessarily 0. Thus there exists a $\hat{\beta}$ an approximate zero-crossing and $\sqrt{n}$-consistent solution of $\mathcal{U}^p_w(\beta)$. ■

Lemma: *2.3.4*

Proof. (Johnson et al., 2008, Appendix), (Feng et al., 2016, Appendix)

Consider the sets in probability space $C_j = \{\hat{\beta} \neq 0\}, j = s+1, .., p$. We need to show that for any $\epsilon > 0$ there exists an $n > N_\epsilon$ such that $P(C_j) < \epsilon$. Since $\hat{\beta}$ is $\sqrt{n}$ consistent, there exists an M_ϵ for a sufficiently large n such that

$$\begin{aligned} P(C_j) &= P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2}) + P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| \geq M_\epsilon n^{-1/2}) \\ &\leq P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2}) + \frac{\epsilon}{2} \end{aligned}$$

Consider the Taylor expansion of non-penalized score function and equation 2.8 we have,

$$\begin{aligned} n^{-1/2}\mathcal{U}_w(\beta) &= n^{-1/2}\mathcal{U}_w(\beta_0) + n^{-1/2} \bigtriangledown_\beta \mathcal{U}_w(\tilde{\beta})(\hat{\beta} - \beta_0) + o(1) \\ &= n^{-1/2}\mathcal{U}_w(\beta_0) + n^{1/2}(-\mathcal{I}_0(\beta) + o_p(1))(\hat{\beta} - \beta_0) + o(1) \\ &= n^{-1/2}\mathcal{U}_w(\beta_0) - n^{1/2}\mathcal{I}_0(\beta)(\hat{\beta} - \beta_0) + o_p(1) \end{aligned}$$

Thus on the set $\{\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2}\}$ for the penalized score function we get,

$$\{n^{-1/2}\mathcal{U}_{w_j}(\beta_0) - n^{1/2}\mathcal{I}_{0j}(\beta)(\hat{\beta} - \beta_0) + o_p(1) - n^{1/2}q_{\lambda_n}(|\hat{\beta}_j|)sign(\hat{\beta}_j)\}^2 = o_p(1)$$

where $\mathcal{I}_{0j}$ is the j^{th} row of $\mathcal{I}_0$.

This implies that for a large n , there exists a M_ϵ^* such that,

$$P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2}, \sqrt{n}q_{\lambda_n}(|\hat{\beta}_j|) > M_\epsilon^*) < \epsilon/2$$

Assumption (A4) implies that if $\hat{\beta}_j \neq 0$ and $|\hat{\beta}_j| < M_\epsilon n^{-1/2}$ then for large n, $\sqrt{n}q_{\lambda_n}(|\hat{\beta}_j|) > M_\epsilon^*$. Thus $P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2}, \sqrt{n}q_{\lambda_n}(|\hat{\beta}_j|) > M_\epsilon^*) = P(\hat{\beta}_j \neq 0, |\hat{\beta}_j| < M_\epsilon n^{-1/2})$ and $P(C_j) < \epsilon$. ■

Theorem: 2.3.5 (Johnson et al., 2008, Appendix)

Proof. The corollary 2.3.2.1 for the true non-zero βs can be rewritten as,

$$n^{-1/2}\mathcal{U}_{w_1}(\beta_0) + n^{1/2}\mathcal{I}_{01}(\beta)(\hat{\beta}_1 - \beta_{01}) - n^{1/2}q_{\lambda_n}(|\hat{\beta}_1|)sign(\hat{\beta}_1) = o_p(1)$$

where $\mathcal{U}_{w_1}$ is an $s \times 1$ dimensional vector associated with the signal βs. Using Taylor expansion of $n^{1/2}q_{\lambda_n}(|\hat{\beta}_1|)sign(\hat{\beta}_1)$ we have, $n^{-1/2}\mathcal{U}_{w_1}(\beta_0) + n^{1/2}\mathcal{I}_{01}(\beta)(\hat{\beta}_1 - \beta_{01}) - n^{1/2}q_{\lambda_n}(|\beta_{01}|)sign(\beta_{01}) - n^{1/2}q'_{\lambda_n}(|\beta_{01}|)sign(\beta_{01})(\hat{\beta}_1 - \beta_{01}) = o_p(1)$

Using lemma 2.3.2 for the first s elements of the score function and re-arrangng the terms

such that $B_{11} = -diag\{q'_{\lambda_n}(|\beta_{01}|)sign(\beta_{01})\}$ is an $s \times s$ diagonal matrix and

$b_n = -q_{\lambda_n}(|\beta_{01}|)sign(\beta_{01})$ is an $s \times 1$ dimensional vector, we get the desired result.

■

Expanding Dimension P-Proofs Lemma: *2.4.2*

Proof.

$$\begin{aligned}\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0}) &= \frac{1}{\sqrt{n}} X^T \Sigma^{1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2}(Y-\mu) - \frac{1}{\sqrt{n}} X^T \Sigma^{1/2} \Gamma^{-1} \Sigma^{-1/2}(Y-\mu) \\ &= \frac{1}{\sqrt{n}} X^T \Sigma^{1/2} (\hat{\Gamma}^{-1} - \Gamma^{-1}) \Sigma^{-1/2}(Y-\mu) \\ &= \frac{1}{\sqrt{n}} X^T \Sigma^{1/2} (\hat{\Gamma}^{-1} - \Gamma^{-1}) r\end{aligned}$$

Consider,

$$\begin{aligned}E\|\mathcal{U}_n(\beta_{n0}) - \hat{\mathcal{U}}_n(\beta_{n0})\|^2 &\leq \frac{1}{n}\|\hat{\Gamma}^{-1} - \Gamma^{-1}\|^2 \lambda_{max}(X^T X)\lambda_{max}(\Sigma) E r^2 \\ &\leq \mathcal{K} n^{\alpha+1} \cdot \frac{p_n}{n}\end{aligned}$$

From (Note 1) for $\lambda_{max}(\Sigma)$ and $\lambda_{max}(X^T X)$, assumptions (A9) and (A10) and due to Chebyshev's inequality, for some $\mathcal{K} > 0$ it is proved. ■

Lemma: *2.4.3*

Proof. In order to ensure that lemma 2.4.3 is true, it is sufficient to show the following,

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} \sup_{|b_n\|=1} |b_n^T[\mathcal{H}(\beta_n^*) - \hat{\mathcal{H}}(\beta_n^*)]b_n| = O_p(n^{\alpha}\sqrt{p_n}) \tag{L.1.}$$

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} \sup_{|b_n\|=1} |b_n^T[\mathcal{E}(\beta_n^*) - \hat{\mathcal{E}}(\beta_n^*)]b_n| = O_p(n^{(\alpha-1)/2}\sqrt{p_n}) \tag{L.2.}$$

$$\sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}}\ \sup_{|b_n\|=1}|b_n^T[\mathcal{G}(\beta_n^*)-\hat{\mathcal{G}}(\beta_n^*)]b_n|=O_p(1) \tag{L.3.}$$

Discordant to the above statements L.1.and L.2., condition L.3. is obtained under $n^{-1}p_n^2 = o(1)$. Let's begin with (L.1.)

$$\begin{aligned}
|b_n^T[\mathcal{H}(\beta_n^*)-\hat{\mathcal{H}}(\beta_n^*)]b_n| &= |b_n^T[\frac{1}{\sqrt{n}}X^T\Sigma^{1/2}\Gamma^{-1}\Sigma^{1/2}X+\frac{1}{\sqrt{n}}X^T\Sigma^{1/2}\hat{\Gamma}^{-1}\Sigma^{1/2}X]b_n| \\
&= \frac{1}{\sqrt{n}}|b_n^T[X^T\Sigma^{1/2}[\Gamma^{-1}-\hat{\Gamma}^{-1}]\Sigma^{1/2}X]b_n| \\
&\leq \mathcal{B}\frac{1}{\sqrt{n}}\|b_n\|^2\lambda_{max}(\Sigma)\lambda_{max}(X^TX)\|\Gamma^{-1}-\hat{\Gamma}^{-1}\| \\
&\leq \mathcal{B}'\frac{1}{\sqrt{n}}n.n^{\alpha}\cdot\sqrt{p_n/n}
\end{aligned}$$

The third inequality is due to Cauchy-Shwartz inequality and inequalities due to the largest eigenvalues. The last inequality is due to assumption (A7),(A9) and (Note 1) for some constant $\mathcal{B}>0$.

Now consider L.2.,

$$\begin{aligned}
|b_n^T[\mathcal{E}(\beta_n^*)-\hat{\mathcal{E}}(\beta_n^*)]b_n| &= \frac{1}{2\sqrt{n}}|b_n^T[X^T\Sigma^{1/2}\Gamma\Sigma^{-3/2}\mathcal{C}\mathcal{D}(\beta_n^*)X-X^T\Sigma^{1/2}\hat{\Gamma}^{-1}\Sigma^{-3/2}\mathcal{C}\mathcal{D}(\beta_n^*)X]b_n| \\
&= \frac{1}{2\sqrt{n}}|b_n^T[X^T\Sigma^{1/2}[\Gamma^{-1}-\hat{\Gamma}^{-1}]\Sigma^{-3/2}\mathcal{C}\mathcal{D}X]b_n| \\
&\leq \mathcal{K}\frac{1}{\sqrt{n}}\lambda_{max}(\mathcal{D})\lambda_{max}(X^TX)\|b_n\|^2\lambda_{max}(\Sigma^{-1})\|\Gamma-\hat{\Gamma}^{-1}\| \\
&\leq \mathcal{K}'n^{\alpha}\cdot\sqrt{p_n}/n
\end{aligned}$$

$\lambda_{max}(\Sigma^{-1}) = 1/\lambda_{min}\Sigma \leq k$ due to assumption (A6) where $k\geq 0$. Inequalities are due to reasons similar to the proof of L.1.. We require assumption (A8) for a finite upper bound

for $\lambda_{max}(\mathcal{D})$. In order to prove L.3. consider,

$$E\|\frac{1}{2\sqrt{n}}b_n^T X^T\Sigma^{1/2}\mathcal{D}[\mathcal{J}(\beta_n^*)-\hat{\mathcal{J}}(\beta_n^*)]Xb_n\|^2$$
$$\leq \frac{1}{4n}\mathcal{K}''\|X^TX\|^2\lambda_{max}(\Sigma^{-1})[\lambda_{max}(\Gamma^{-1}-\hat{\Gamma}^{-1})]^2E(r^2)$$
$$\leq \mathcal{K}^*\frac{p_n}{4n}\frac{p_n}{n}E(r^2)$$
$$\leq \mathcal{K}'p_n^2/n$$

Note that (A9) implies $\lambda_{max}(\Gamma^{-1}-\hat{\Gamma}^{-1}) = O_p(\sqrt{p_n/n})$ since the covariance matrices are symmetric. (L. Wang, 2011, Remark 2 pg 396). Further we require assumptions (A10) for the last inequality and (A6) for the last but one inequality. From the assumption $p_n^2n^{-1} = o(1)$ we have L.3.. Lemma 2.4.3 is proved. ■

Lemma: *2.4.4*

Proof. Consider,

$$|I_{n212}| = |\beta_n-\beta_{n0})^T[\hat{\mathcal{H}}_n(\beta_n^*)-\hat{\mathcal{H}}_n(\beta_{n0})](\beta_n-\beta_{n0})|$$
$$= \frac{1}{\sqrt{n}}|(\beta_n-\beta_{n0})^T[X^T\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_n^*)X - X^T\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})X$$
$$+ X^T\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})X - X^T\Sigma^{1/2}(\beta_{n0})\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})X](\beta_n-\beta_{n0})|$$
$$\leq \frac{1}{\sqrt{n}}|(\beta_n-\beta_{n0})^TX^T[\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_n^*)-\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})]X(\beta_n-\beta_{n0})|$$
$$+ \frac{1}{\sqrt{n}}|(\beta_n-\beta_{n0})^TX^T[\Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})-\Sigma^{1/2}(\beta_{n0})\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_{n0})]X(\beta_n-\beta_{n0})|$$
$$:= I_{n2121} + I_{n2122}$$

We further evaluate the two portions separately as follows similar to (L. Wang, 2011, Supplemenatry Material Lemma 3.5),

$$\begin{aligned} I_{n2121} &= \frac{1}{\sqrt{n}}|(\beta_n - \beta_{n0})^T X^T \Sigma^{1/2}(\beta_n^*)\hat{\Gamma}^{-1}[\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]X(\beta_n - \beta_{n0})| \\ &\leq \frac{1}{\sqrt{n}}\|(\beta_n - \beta_{n0})^T X^T \Sigma^{1/2}\hat{\Gamma}^{-1/2}\|\|\hat{\Gamma}^{-1/2}[\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]X(\beta_n - \beta_{n0}\| \end{aligned}$$

by Cauchy-Schwartz inequality. Now consider the two parts of the inequality,

$$\begin{aligned} \|(\beta_n - \beta_{n0})^T X^T \Sigma^{1/2}\hat{\Gamma}^{-1/2}\|^2 &= (\beta_n - \beta_{n0})^T X^T \Sigma^{1/2}\hat{\Gamma}^{-1}\Sigma^{1/2}(\beta_n^*)X(\beta_n - \beta_{n0}) \\ &\leq \lambda_{max}(\hat{\Gamma}^{-1})\lambda_{max}(\Sigma(\beta_n^*))\|X(\beta_n - \beta_{n0}\|^2 \end{aligned}$$

$$\begin{aligned} &\|\hat{\Gamma}^{-1/2}[\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]X(\beta_n - \beta_{n0}\| \\ &\qquad \leq \lambda_{max}^{1/2}(\hat{\Gamma}^{-1})\lambda_{max}^{1/2}(\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]\|X(\beta_n - \beta_{n0}\| \end{aligned}$$

Hence,

$$\begin{aligned} |I_{n2121}| &\leq \frac{1}{\sqrt{n}}\lambda_{max}(\hat{\Gamma}^{-1})\lambda_{max}^{1/2}(\Sigma(\beta_n^*)) \max_{ij}|(\Sigma_{ij}(\beta_n^*) - \Sigma_{ij}(\beta_{n0}))|\lambda_{max}(X^T X)\|\beta_n - \beta_{n0}\|^2 \\ &\leq \frac{1}{\sqrt{n}}\sqrt{n}\max_{ij}\|X_{ij}\|\|\beta_n - \beta_{n0}\|\tilde{K}n^{\alpha}\frac{p_n}{n}\Delta^2 \\ &\leq \sqrt{p_n}\sqrt{\frac{p_n}{n}\frac{p_n}{n}}n^{\alpha}\Delta^2.K \\ &= O_p(p_n^2 n^{\alpha-3/2}) \end{aligned}$$

Using Taylor expansion we see that $\max_{ij}|(\Sigma_{ij}(\beta_n^*) - \Sigma_{ij}(\beta_{n0}))| \leq \max_{ij}\|X_{ij}\|\|\beta_n - \beta_{n0}\|$. (Note 1) and assumption (A7) are used in the above inequality for some $\tilde{K} > 0$ and $\lambda_{max}(\hat{\Gamma}^{-1}) = 1/\lambda_{min}(\hat{\Gamma})$ which is bounded due to remark 1 in section 2.4.1.1. Similarly

$I_{n2122} = O_p(p_n^2 n^{\alpha-3/2})$ ■

Lemma: *2.4.5*

Proof. Let us consider, $I_{n213} = -(\beta_n - \beta_{n0})^T[\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*) - \hat{\mathcal{H}}(\beta_n^*)](\beta_n - \beta_{n0})$

Since, $\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n^*) - \hat{\mathcal{H}}(\beta_n^*) \leq |\hat{\mathcal{E}}(\beta_n^*) - \hat{\mathcal{G}}(\beta_n^*)|$

In the same spirit as (L. Wang, 2011, Lemma 3.4, 3.5) we can show that if $p_n^2/\sqrt{n} = o(1) \implies p_n^4 n^{-1} = o(1)$ for some $b_n \in \mathcal{R}^{p_n}$ then,

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} \sup_{\|b_n\|=1} |b_n^T \hat{\mathcal{E}}(\beta_n^*) b_n| = O_p(\sqrt{n} p_n) \tag{a}$$

$$\sup_{\|\beta_n - \beta_{n0}\| \leq \Delta\sqrt{p_n/n}} \sup_{\|b_n\|=1} |b_n^T \hat{\mathcal{G}}(\beta_n^*) b_n| = O_p(\sqrt{n} p_n) \tag{b}$$

Let us begin with proof of (a). It is important to note that the results for Γ and $\hat{\Gamma}$ are identical and interchangeable in what follows. Therefore without loss of generality, consider the telescopic sum,

$$\begin{aligned}
\mathcal{E}(\beta_n^*) &= \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_n^*) \Gamma^{-1} \Sigma^{-3/2}(\beta_n^*) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&= \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) \mathcal{D}(\beta_{n0}) X \\
&- \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) \mathcal{D}(\beta_{n0}) X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) \mathcal{D}(\beta_n^*) X \\
&- \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) \mathcal{D}(\beta_n^*) X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&- \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_n^*) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&- \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_n^*) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_n^*) \Gamma^{-1} \Sigma^{-3/2}(\beta_n^*) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&\implies \mathcal{E}(\beta_n^*) = \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) \mathcal{D}(\beta_{n0}) X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) [\mathcal{D}(\beta_n^*) - \mathcal{D}(\beta_{n0}] X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) [\mathcal{C}(\beta_n^*) - \mathcal{C}(\beta_{n0})] \mathcal{D}(\beta_n^*) X \\
&+ \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} [\Sigma^{-3/2}(\beta_n^*) - \Sigma^{-3/2}(\beta_{n0})] \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&+ \frac{1}{2\sqrt{n}} X^T [\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})] \Gamma^{-1} \Sigma^{-3/2}(\beta_n^*) \mathcal{C}(\beta_n^*) \mathcal{D}(\beta_n^*) X \\
&= \mathcal{E}_1(\beta_{n0}) + \sum_{l=2}^{5} \mathcal{E}_l(\beta_n^*) \qquad \text{(E*)}
\end{aligned}
$$

Now let us examine,

$$
\begin{aligned}
E\|\mathcal{E}_1(\beta_{n0})\|^2 =& \frac{1}{4n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \mathcal{D}_{j1} \mathcal{D}_{j2} E(r_{j1} \cdot r_{j2}) \cdot \mathrm{Tr}\{X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-1} \\
& e_{j1} e_{j1}^T X X^T e_{j2} e_{j2}^T \Sigma^{-1} \Gamma^{-1} \Sigma^{1/2}(\beta_{n0}) X\}
\end{aligned}
$$

where $\mathcal{D}_j$ is the j^{th} diagonal entry and e_j denotes a unit vector whose j^{th} entry is 1 and all others 0. By Assumption (A10) we have the right side of the above inequality which is given by

$$
E\|\mathcal{E}_1(\beta_{n0})\|^2 \leq C \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \frac{1}{4n} |e_{j1}^T X X^T e_{j2} e_{j2}^T \Sigma^{-1} \Gamma^{-1} \Sigma^{1/2} X X^T \Sigma^{1/2} \Gamma^{-1} \Sigma^{-1} e_{j1}|
$$

$$\leq C \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \frac{1}{4n} \|e_{j1}^T X\| \|X^T e_{j2}\| \|e_{j2} \Sigma^{-1} \Gamma^{-1} \Sigma^{1/2} X\| \|X^T \Sigma^{1/2} \Gamma^{-1} \Sigma^{-1} e_{j1}\|$$

Since, $\|e_{j2} \Sigma^{-1} \Gamma^{-1} \Sigma^{1/2} X\| \leq C^* \operatorname{Tr}(XX^T)^{1/2} \lambda_{max}^{1/2}(\Gamma^{-1}) \lambda_{max}^{1/2}(\Sigma^{-1/2}) \leq \tilde{C}\sqrt{p_n}$ due to remark 1 and assumption (A6) we get,

$$E\|\mathcal{E}_1(\beta_{n0})\|^2 \leq \tilde{C} \frac{1}{n} p_n \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \|e_{j1}^T X\| \|X^T e_{j2}\| \leq \tilde{C}' \frac{1}{n} p_n \max_{ij} |X_{ij}|^2 = \tilde{C} p_n^2 / n$$

due to assumption (A5). $\implies sup_{\|b_n\|=1} |b_n^T \mathcal{E}_1(\beta_{n0}) b_n| = O_p(1)$ if $p_n^2/n = o(1)$ as $n \to \infty$.
Similarly for

$$\mathcal{E}_2(\beta_n^*) = \frac{1}{2\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) \mathcal{C}(\beta_{n0}) [\mathcal{D}(\beta_n^*) - \mathcal{D}(\beta_{n0})] X$$

we consider the following,

$$\begin{aligned} E\|\mathcal{E}_2(\beta_n^*)\|^2 = \frac{1}{4n} \sum_{j1=1}^{n} \sum_{j2=1}^{n} & [\mathcal{D}(\beta_n^*) - \mathcal{D}(\beta_{n0})]_{j1} [\mathcal{D}(\beta_n^*) - \mathcal{D}(\beta_{n0})]_{j2} E(r_{j1} r_{j2}) \cdot \\ & \operatorname{Tr}[X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-1} e_{j1} e_{j1}^T X X^T e_{j2} e_{j2}^T \Sigma^{-1} \Gamma^{-1} \Sigma^{1/2}(\beta_{n0}) X] \\ & \leq \mathbb{K} n p_n^2 \end{aligned}$$

By assumption (A8) we have that derivatives of the mean function are uniformly bounded above. Using assumption (A10) and all other trace results are identical to that obtained for $\mathcal{E}_1(\beta_{n0})$ we have, $sup_{\|b_n\|=1} |b_n^T \mathcal{E}_2(\beta_{n0}) b_n| = O_p(\sqrt{n} p_n)$.
Further looking at the terms of equation E*,

$$\mathcal{E}_3(\beta_n^*) = \frac{1}{\sqrt{n}} X^T \Sigma^{1/2}(\beta_{n0}) \Gamma^{-1} \Sigma^{-3/2}(\beta_{n0}) [\mathcal{C}(\beta_n^*) - \mathcal{C}(\beta_{n0})] \mathcal{D}(\beta_n^*) X$$

where, $[\mathcal{C}(\beta_n^*) - \mathcal{C}(\beta_{n0})] = diag(\mu(\beta_{n0}) - \mu(\beta_n^*)) \coloneqq \mathcal{P}_n$

Then considering,

$$E(\|\mathcal{E}_3(\beta_n^*)\|^2) = \frac{1}{4n}\sum_{j1=1}^{n}\sum_{j2=1}^{n}\mathcal{D}_{j1}(\beta_n^*)\mathcal{D}_{j2}(\beta_n^*)\,\mathrm{Tr}\{X^T\mathcal{P}_n\Sigma^{-3/2}\Gamma^{-1}\Sigma^{1/2}e_{j1}e_{j1}^T XX^T$$
$$e_{j2}e_{j2}^T\Gamma^{-1}\Sigma^{-3/2}\mathcal{P}_n X\}$$

Using Taylor expansion and for some constant $\tilde{k}' > 0$ it can be shown that,

$\mathrm{Tr}(\mathcal{P}_n^2) = \|\mu(\beta_{n0}) - \mu(\beta_n^*)\|^2 \leq \tilde{k}'\|X\|^2\|\beta_n^* - \beta_{n0}\|^2$ Thus if $p_n^2 n^{-1} = o(1)$ then,

$$E(\|\mathcal{E}_3(\beta_n^*)\|^2) \leq \frac{1}{n}\tilde{K}\max_{ij}\|X_{ij}\|^2\lambda_{max}(\Gamma^{-2})\lambda_{max}(\Sigma^{-2})\,\mathrm{Tr}(X^TX)\tilde{k}'\|X\|^2\|\beta_n^* - \beta_{n0}\|^2$$
$$\leq \tilde{K}'' p_n^4/n^2\Delta^2 = o(1)$$

Lastly consider, $|b_n^T\mathcal{E}_5(\beta_{n0})b_n|$

$$= |\frac{1}{2\sqrt{n}}\sum_{j=1}^{n}\mathcal{D}_j r_j b_n^T X[\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]\Sigma^{-1}\Gamma^{-1}e_j e_j^T Xb_n|$$
$$\leq C^*\frac{1}{\sqrt{n}}\sum_{j=1}^{n}|b_n^T X[\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})]\Sigma^{-1}\Gamma^{-1}e_j|\cdot|e_j Xb_n|$$
$$\leq C^*\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\|Xb_n\|^2\lambda_{max}(\Gamma^{-1})\lambda_{max}(\Sigma^{-1})\max_j|\Sigma^{1/2}(\beta_n^*) - \Sigma^{1/2}(\beta_{n0})|$$
$$\leq C^* n\frac{1}{\sqrt{n}}\max_{ij}\|X_{ij}\|\|\beta_n^* - \beta_{n0}\|\,\mathrm{Tr}(X^TX)$$
$$\leq \frac{1}{\sqrt{n}}nC^*\sqrt{p_n}\frac{\sqrt{p_n}}{\sqrt{n}}p_n$$
$$= C^* p_n^2$$

Similar to $\mathcal{E}_5$ we can show that $\mathcal{E}_4 = O(p_n^2)$. This completes proof of (a)

The proof of (b) has similar arguments that results in using telescopic sums and investigating the properties of each its terms as discussed above. It can therefore be verified that (b) is true. Thus $|(\beta_n - \beta_{n0})^T[\bigtriangledown_\beta \hat{\mathcal{U}}(\beta_n^*) - \hat{\mathcal{H}}(\beta_n^*)](\beta_n - \beta_{n0})| = \Delta^2 o_p(p_n)$ under the assumption $n^{-1/2}p_n^2 = o(1)$ as $n \to \infty$.

■

Theorem: *2.4.8* (L. Wang, 2011, Theorem Proof 3.8)

Proof. Here we provide the proof of statements T.1 and T.2

In order to utilize lemmas 2.4.3, 2.4.4 and 2.4.5 we re-express T.1 as follows

$$\begin{aligned}
&\sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} |\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0})](\hat{\beta}_n - \beta_{n0})| \\
&\qquad \leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} |\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n)](\hat{\beta}_n - \beta_{n0})| \\
&\qquad + \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} |\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_n)](\hat{\beta}_n - \beta_{n0})| \\
&\qquad + \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} |\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2}[\hat{\mathcal{H}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0})](\hat{\beta}_n - \beta_{n0})| \\
&\qquad := i_{n1} + i_{n2} + i_{n3}
\end{aligned}$$

Consider,

$$\begin{aligned}
i_{n1} &\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} [\frac{1}{n^{\alpha}} a^T \hat{\mathcal{V}}_n^{-1/2}(\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n))^2 \hat{\mathcal{V}}_n^{-1/2} a]^{1/2} \|\hat{\beta}_n - \beta_{n0}\| \\
&\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} \max(|\lambda_{min}(\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n))|, |\lambda_{max}(\bigtriangledown_\beta \mathcal{U}_n(\beta_n) - \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n))| \\
&\quad \frac{1}{n^{\alpha/2}} \lambda_{max}(\hat{\mathcal{V}}_n^{-1/2}) O_p(\sqrt{p_n/n})
\end{aligned}$$

Since 2.23 has a positive definite limit implies as in assumption (A3) in section 2.3 that $\lambda_{max}(\hat{\mathcal{V}}_n^{-1/2})$ has an upper bound. According to lemma 2.4.3 we have,

$$\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} \frac{1}{n^{\alpha/2}} O_p(n^{\alpha/2}) O_p(\sqrt{p_n/n})$$

Therefore $\alpha \in (0,1)$ and $p_n^2 n^{-1} = o(1)$ we get $i_{n1} = o(1)$.

Now consider the next term i_{n3},

$$\begin{aligned}
i_{n3} &\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} (\frac{1}{n^{\alpha}} a^T \hat{\mathcal{V}}_n^{-1/2} [\hat{\mathcal{H}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0})]^2 \hat{\mathcal{V}}_n^{-1/2} a)^{1/2} \|\hat{\beta}_n - \beta_{n0}\| \\
&\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} \max(|\lambda_{min}(\hat{\mathcal{H}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0}))|, |\lambda_{max}(\hat{\mathcal{H}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_{n0})| \\
&\frac{1}{n^{\alpha/2}} \lambda_{max}(\hat{\mathcal{V}}_n^{-1/2}) O_p(\sqrt{p_n/n}) \\
&\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} o_p(p_n n^{\alpha-1/2}) \frac{1}{n^{\alpha/2}} O_p(\sqrt{p_n/n})
\end{aligned}$$

Here for $p_n^4.n^{-1} = o(1)$, $\begin{cases} \alpha = 0.5 \implies i_{n2} = o(p_n^{3/2} n^{-3/4}) = o(1) \\ \alpha = 1 \implies i_{n2} = o(p_n^{3/2} n^{-1/2}) = o(1) \end{cases}$

Let us finally consider,

$$\begin{aligned}
i_{n2} &\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} (\frac{1}{n^{\alpha}} a^T \hat{\mathcal{V}}_n^{-1/2} [\bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_n)]^2 \hat{\mathcal{V}}_n^{-1/2} a)^{1/2} \|\hat{\beta}_n - \beta_{n0}\| \\
&\leq \sup_{\|\beta_n-\beta_{n0}\|\leq\Delta\sqrt{p_n/n}} \frac{1}{n^{\alpha/2}} \| \bigtriangledown_\beta \hat{\mathcal{U}}_n(\beta_n) - \hat{\mathcal{H}}_n(\beta_n) \| \lambda_{max}(\hat{\mathcal{V}}_n^{-1/2}) O_p(\sqrt{p_n/n}) \\
&\leq \frac{1}{n^{\alpha/2}} \sqrt{n} p_n O_p(\sqrt{p_n/n})
\end{aligned}$$

Therefore only with $\alpha = 1$ we have $i_{n2} = o(1)$. To complete the proof we look at T.2 for $\|a\| = 1$

and $a \in \mathcal{R}^{p_n}$ we have,

$$\begin{aligned}
&\Big[\frac{1}{n^{\alpha/2}} a^T \hat{\mathcal{V}}_n^{-1/2} [\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})]\Big]^2 \\
&\qquad = \frac{1}{n^{\alpha}} a^T \hat{\mathcal{V}}_n^{-1/2} [\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})][\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})]^T \hat{\mathcal{V}}_n^{-1/2} a \\
&\qquad \leq \frac{1}{n^{\alpha}} \lambda_{max}(\hat{\mathcal{V}}_n^{-1}) \|\hat{\mathcal{U}}_n(\beta_{n0}) - \mathcal{U}_n(\beta_{n0})\|^2 \\
&\qquad \leq \mathbb{K} p_n^2 / n^{\alpha}
\end{aligned}$$

Based on the relation 2.24 for a bounded positive definite matrix $\mathcal{I}_2(\beta_{n0})$ we can find an upper bound for $\lambda_{max}(\mathcal{V}_n^{-1})$ independent of n. Therefore $i_{n2} = o(1)$ for any $\alpha \in (0.5, 1)$ under the assumption $p_n^4 n^{-1} = o(1)$. Thus we obtain the asymptotic normality of the estimate $\hat{\beta}_n$.

■

Theorem: *2.4.9* (L. Wang et al., 2012, Properties 1,2)

Proof. To prove equation 2.26 we already have $\hat{\mathcal{U}}_{nj}(\hat{\beta}_n) = 0$ from 2.4.1. Since the penalty function considered now is a SCAD penalty as shown in 2.18, it is sufficient to show that

$$P(|\hat{\beta}_{nj}| \geq \mathfrak{a}\lambda_n, \quad j = 1, \ldots, s_n) \to 1$$

since it implies that the penalty portion of equation 2.25 too tends to 0 in probability. It is simple to see the relation, $\min_{1 \leq j \leq s_n} |\hat{\beta}_{nj}| \geq \min_{1 \leq j \leq s_n} |\beta_{n0j}| - \max_{1 \leq j \leq s_n} |\beta_{n0j} - \hat{\beta}_{nj}| \geq \min_{1 \leq j \leq s_n} |\beta_{n0j}| - \|\beta_{n10} - \hat{\beta}_{n10}\|$. Thus we have,

$$\begin{aligned}
P(\min_{1 \leq j \leq s_n} |\hat{\beta}_{nj}| > \mathfrak{a}\lambda_n) &\geq P(\min_{1 \leq j \leq s_n} |\beta_{n0j}| - \|\beta_{n10} - \hat{\beta}_{n10}\| \geq \mathfrak{a}\lambda_n) \\
&= P(\|\beta_{n10} - \hat{\beta}_{n10}\| \leq \min_{1 \leq j \leq s_n} |\beta_{n0j}| - \mathfrak{a}\lambda_n) \to 1
\end{aligned}$$

Since $\|\beta_{n10} - \hat{\beta}_{n10}\| = o_p(\sqrt{s_n/n})$ and (A11) imposes $\min_{1\leq j\leq s_n} |\beta_{n0j}|/\lambda_n \to \infty$ as $n \to \infty$. So we have 2.26.

In order to prove 2.27, we notice that for $\hat{\beta}_{nk}$, $k = s_n + 1, \ldots, p_n$ the derivative of the penalty in 2.18 is indeed 0. Thus we proceed by considering only the first term of equation 2.25 and denote it by

$$\hat{\mathcal{W}}_n(\hat{\beta_n}) \coloneqq \frac{1}{n} X^T \Sigma^{1/2} \hat{\Gamma}^{-1} \Sigma^{-1/2} (Y - \mu(\hat{\beta}_n)) \tag{2.28}$$

and $\hat{\mathcal{W}}_{nk}(\hat{\beta}) = e_k^T \hat{\mathcal{W}}_n(\hat{\beta_n})$ where e_k has the k^{th} element as 1. Therefore 2.27 is equivalent to

$$P\bigg(\max_{s_n+1\leq k\leq p_n} |\hat{\mathcal{W}}_{nk}(\hat{\beta}_n)| \leq \frac{\lambda_n}{log(n)} \bigg) \to 1$$

which can be ensured if,

$$P\bigg(\max_{s_n+1\leq k\leq p_n} |\hat{\mathcal{W}}_{nk}(\hat{\beta}_n) - \mathcal{W}_{nk}(\hat{\beta}_n)| > \frac{\lambda_n}{2log(n)} \bigg) \to 0 \tag{2.29}$$

$$P\bigg(\max_{s_n+1\leq k\leq p_n} |\mathcal{W}_{nk}(\hat{\beta}_n)| > \frac{\lambda_n}{2log(n)} \bigg) \to 0 \tag{2.30}$$

To check 2.29 let us consider,

$$\begin{aligned}
&P\bigg(\max_{s_n+1\leq k\leq p_n} n^{-1} |e_k^T X^T \Sigma^{1/2} [\hat{\Gamma}^{-1} - \Gamma^{-1}] \Sigma^{-1/2} (Y - \mu(\hat{\beta}_n))| > \frac{\lambda_n}{2log(n)} \bigg) \\
&\leq P\bigg(\max_{s_n+1\leq k\leq p_n} n^{-1} \|e_k^T X^T \Sigma^{1/2}\| \|\hat{\Gamma}^{-1} - \Gamma^{-1}\| \|\Sigma^{-1/2} (Y - \mu(\hat{\beta}_n))\| > \frac{\lambda_n}{2log(n)} \bigg) \\
&\leq P\bigg(n^{-1} \|\epsilon\| > \frac{\lambda_n}{2\sqrt{s_n} log(n)} \bigg) \\
&\leq \mathbb{C} \frac{n^{-1/2} E(\|\epsilon\|) \sqrt{s_n} log(n)}{\lambda_n \sqrt{n}} = O\big(\frac{\sqrt{s_n} log(n)}{\lambda_n \sqrt{n}} \big) = o(1)
\end{aligned}$$

where $\epsilon = \Sigma^{-1/2}(Y - \mu(\hat{\beta}_n))$ is an $n \times 1$ dimensional vector. It is easy to verify that $\max_{s_n+1 \leq k \leq p_n} \|e_k^T X^T \Sigma^{1/2}\| = O(\sqrt{n})$ and therefore we get the third inequality from assumption (A9) and Markov's inequality. The last inequality comes from (A10) and Jensen's inequality in the following way,

$$\exists \mathfrak{M} > 0, \text{such that} \frac{1}{\sqrt{n}} E\|\epsilon\| \leq \big(\frac{1}{n} E\|\epsilon\|^2\big)^{1/2} \leq \mathfrak{M}$$

Thus using assumption (A11) we have 2.26. In order to prove 2.27 let us begin by considering the Taylor expansion,

$$\mathcal{W}_{nk}(\hat{\beta}_n) = \mathcal{W}_{nk}(\beta_{n0}) + \bigtriangledown_{\beta,k} \mathcal{W}_{nk}(\beta_n^*)(\hat{\beta}_n - \beta_{n0}) + (\hat{\beta}_n - \beta_{n0})^T \frac{\partial^2 \mathcal{W}_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_n^T} (\hat{\beta}_n - \beta_{n0}) \tag{2.31}$$

where β_n^* is between $\hat{\beta}_n$ and β_{n0}. Let $\mathbb{D}_k(\beta_n) \coloneqq \frac{\partial^2 \mathcal{W}_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_n^T}$ denote the $p_n \times p_n$ matrix of double derivatives and let $\mathbb{D}_{k1}(\beta_n)$ denote the top left $s_n \times s_n$ sub matrix. Based on the fact that $\hat{\beta}_n - \beta_{n0} = (\hat{\beta}_{n1} - \beta_{n10})^T, 0^T)^T$ we can rewrite 2.31 as follows,

$$\mathcal{W}_{nk}(\hat{\beta}_n) = \mathcal{W}_{nk}(\beta_{n0}) + \bigtriangledown_{\beta,k1} \mathcal{W}_{nk}(\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10}) + (\hat{\beta}_{n1} - \beta_{n10})^T \mathbb{D}_{k1}(\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10}) \tag{2.32}$$

and similar to the earlier argument we can express the following,

$$\begin{aligned}
&P\bigg(\max_{s_n+1 \leq k \leq p_n} |\mathcal{W}_{nk}(\hat{\beta}_n)| > \frac{\lambda_n}{2log(n)}\bigg) \\
&\leq P\bigg(\max_{s_n+1 \leq k \leq p_n} |\mathcal{W}_{nk}(\beta_{n0}| > \frac{\lambda_n}{6log(n)})\bigg) + P\bigg(\max_{s_n+1 \leq k \leq p_n} |\bigtriangledown_{\beta,k1}(\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10})| > \frac{\lambda_n}{6log(n)}\bigg) \\
&+ P\bigg(\max_{s_n+1 \leq k \leq p_n} (\hat{\beta}_{n1} - \beta_{n10})^T \mathbb{D}_{k1}(\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10}) > \frac{\lambda_n}{6log(n)}\bigg)
\end{aligned}$$

$\coloneqq \mathfrak{i}_1 + \mathfrak{i}_2 + \mathfrak{i}_3$

We are required to show that each of $\mathfrak{i}_1, \mathfrak{i}_2, \mathfrak{i}_3 = o(1)$. From corollary 2.4.7.1 and equation 2.24 that has a positive definite limit for the variance of the score function for a sufficiently large n we have,

$$P\bigg(\max_{s_n+1\leq k\leq p_n} |\frac{1}{\sqrt{n}} e_k^T \hat{\mathcal{V}}_n^{-1/2} X^T \Sigma^{1/2} \hat{\Gamma}^{-1} (Y - \mu(\beta_{n0}))| > \frac{\lambda_n \sqrt{n}}{6log(n)} \hat{\mathcal{V}}_{nk}^{-1/2} \bigg) \leq exp\bigg[-\frac{1}{2} \mathbb{K} \frac{n\lambda_n^2}{log(n)^2} \bigg]$$

The inequality is obtained from a normal density (ex: $P(|X| > t) \leq e^{-t^2/2}$ for $X \sim N(0,1)$). Due to 2.24 we can find a positive lower bound on the minimum eigen value of $\hat{\mathcal{V}}_{nk}$. Therefore if $\frac{n\lambda_n^2}{log(n)^2} \to \infty$ as $n \to \infty$ from assumption (A11) we get $\mathfrak{i}_1 = o(1)$

Let us consider,

$$\begin{aligned}
\mathfrak{i}_2 &= P\bigg(\max_{s_n+1\leq k\leq p_n} | \bigtriangledown_{\beta,k1} (\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10})| > \frac{\lambda_n}{6log(n)} \bigg) \\
&= P\bigg(\max_{s_n+1\leq k\leq p_n} | \bigtriangledown_{\beta,k1} (\beta_n^*)(\hat{\beta}_{n1} - \beta_{n10})| > \frac{\lambda_n}{6log(n)}, \|\hat{\beta}_{n1} - \beta_{n10}\| \leq \Delta\sqrt{s_n/n} \bigg) \\
&+ P(\|\hat{\beta}_{n1} - \beta_{n10}\| > \Delta\sqrt{s_n/n})
\end{aligned}$$

From the consistency of $\hat{\beta}_{n1}$ in Theorem for consistency (previous section) and the decomposition of 2.4.1 with the new normalizing factor we have,

$$\begin{aligned}
\mathfrak{i}_2 &\leq P\bigg(\max_{s_n+1\leq k\leq p_n} \| \bigtriangledown_{\beta,k1} (\beta_n^*)\| > \frac{\lambda_n \sqrt{n}}{6\sqrt{s_n} log(n)} \bigg) + o(1) \\
&\leq P\bigg(\max_{s_n+1\leq k\leq p_n} \|\mathbb{H}_{nk1}(\beta_n^*)\| > \frac{\lambda_n \sqrt{n}}{12\sqrt{s_n} log(n)} \bigg) \\
&+ P\bigg(\max_{s_n+1\leq k\leq p_n} \| \bigtriangledown_{\beta,k1} (\beta_n^*) - \mathbb{H}_{nk1}(\beta_n^*)\| > \frac{\lambda_n \sqrt{n}}{12\sqrt{s_n} log(n)} \bigg) + o(1) \\
&\coloneqq \mathfrak{i}_{21} + \mathfrak{i}_{22} + o(1)
\end{aligned}$$

where $\mathbb{H}_{nk1} = (H_{nk1}, \ldots, H_{nks_n})^T$ denotes the subvector $\mathbb{H}_{nk}$ with the first s_n elements with a

normalizng factor of $\frac{1}{n}$ instead of $\frac{1}{\sqrt{n}}$ as in the original lemma. Now we can easily see from 2.4.5 the norm of $\mathfrak{i}_{22}$ is $o(s_n)$ and therefore with (A11) we get $\mathfrak{i}_{22} = o(1)$. We can re-express $\mathfrak{i}_{21}$ as follows,

$$\mathfrak{i}_{21} = P\bigg(\max_{s_n+1\leq k\leq p_n} \|\mathbb{H}_{nk1}(\beta_n^*) - \mathbb{H}_{nk1}(\beta_{n10})\| > \frac{\lambda_n\sqrt{n}}{12\sqrt{s_n}log(n)}\bigg)$$
$$+ P\bigg(\max_{s_n+1\leq k\leq p_n} \|\mathbb{H}_{nk1}(\beta_{n10})\| > \frac{\lambda_n\sqrt{n}}{12\sqrt{s_n}log(n)}\bigg)$$

The first term is $o(1)$ if $\frac{\lambda_n\sqrt{n}}{s_n^{3/2}log(n)} \to \infty$ as $n \to \infty$ from lemma 2.4.4 after modifying the normalizing factor and using assumption (A11). The second term is similar to the proof in supplementary material of Wang. To begin with we have,

$$P\bigg(\max_{s_n+1\leq k\leq p_n} \|\mathbb{H}_{nk1}(\beta_{n10})\| > \frac{\lambda_n\sqrt{n}}{12\sqrt{s_n}log(n)}\bigg) \leq P\bigg(\max_{s_n+1\leq k\leq p_n} \|\mathbb{H}_{nk1}(\beta_{n10})\|^2 > \mathfrak{C}\frac{n\lambda_n^2}{s_nlog^2(n)}\bigg)$$
$$\leq P\bigg(\max_{s_n+1\leq k\leq p_n} |\|\mathbb{H}_{nk1}(\beta_{n10}\|^2 - E\|\mathbb{H}_{nk1}(\beta_{n10})\|^2| + \max_{s_n+1\leq k\leq p_n} E\|\mathbb{H}_{nk1}(\beta_{n10})\|^2 > \mathfrak{C}\frac{n\lambda_n^2}{s_nlog^2(n)})\bigg)$$

It can be shown that $|\mathbb{H}_{nkj}(\beta_{n0})|$ is uniformly bounded using assumptions (A12),(A9) and (A8). Further, $\max_{s_n+1\leq k\leq p_n} E\|\mathbb{H}_{nk1}(\beta_{n0})\|^2 = \max_{s_n+1\leq k\leq p_n} E(\sum_{j=1}^{s_n} H_{nkj}^2(\beta_{n0})) \leq Cs_n$. Then from assumption (A11) we have the second term is $o(1)$. Consider the following for a sufficiently large n,

$$P\bigg(\max_{s_n+1\leq k\leq p_n} |\|\mathbb{H}_{nk1}(\beta_{n10})\|^2 - E\|\mathbb{H}_{nk1}(\beta_{n10})\|^2| \geq \frac{\mathfrak{C}n\lambda_n^2}{2s_nlog^2(n)}\bigg)$$
$$\leq \sum_{k=s_n+1}^{p_n} P\bigg(|\|\mathbb{H}_{nk1}(\beta_{n10})\|^2 - E\|\mathbb{H}_{nk1}(\beta_{n10})\|^2| \geq \frac{\mathfrak{C}n\lambda_n^2}{2s_nlog^2(n)}\bigg)$$
$$\leq \mathfrak{C}\sum_{k=s_n+1}^{p_n} \frac{E(\sum_{j=1}^{s_n}[H_{nkj}^2(\beta_{n0}) - E(H_{nkj}^2(\beta_{n0}))]^2 s_n^2 log^4(n)}{n^2\lambda_n^4} = O(p_n s_n^3 log^4(n)/(n^2\lambda_n^4))$$

The last inequality is from markov's inequality and assumption (A11). Therefore $\mathfrak{i}_2 = o(1)$.

To complete the proof we now focus on the last term of 2.32 and need to show $\mathfrak{i}_3 = o(1)$. Similar

to the inequalities obtained above we get,

$$
\begin{aligned}
\mathfrak{i}_3 &\leq P\bigg(\max_{s_n+1\leq k\leq p_n} |(\hat{\beta}_{n1}-\beta_{n10})^T \mathbb{D}_{k1}(\beta_n^*)(\hat{\beta}_{n1}-\beta_{n10})| > \frac{\lambda_n}{6log(n)} \bigg) \\
&\leq P\bigg(\max_{s_n+1\leq k\leq p_n} |(\hat{\beta}_{n1}-\beta_{n10})^T \mathbb{D}_{k1}(\beta_n^*)(\hat{\beta}_{n1}-\beta_{n10})| > \frac{\lambda_n}{6log(n)}, \|\hat{\beta}_{n1}-\beta_{n10}\| \leq \Delta\sqrt{s_n/n} \bigg) \\
&+ P(\|\hat{\beta}_{n1}-\beta_{n10}\| > \Delta\sqrt{s_n/n}) \\
&\leq \sum_{k=s_n+1}^{p_n} P\bigg(\mathrm{Tr}(\mathbb{D}_{k1}) > \frac{n\lambda_n}{s_n log(n)} \bigg) + o(1) \\
&\leq \mathcal{K} \sum_{k=s_n+1}^{p_n} \frac{E[\mathrm{Tr}(\mathbb{D}_{k1}^2(\beta_n^*))] s_n^2 log^2(n)}{n^2\lambda_n^2} + o(1)
\end{aligned}
$$

It can be shown that $E[\mathrm{Tr}(\mathbb{D}_{k1}^2(\beta_n^*))] = E\bigg[\sum_{j=1}^{s_n} \frac{\partial^2 \mathcal{W}_{nk}}{\partial \beta_{nj}^2}(\beta_n^*)\bigg]^2 \leq C s_n^2$. Therefore with conditions similar to L. Wang et al. (2012), if $\frac{p_n s_n^3 log^2(n)}{n^2\lambda_n^2} \to 0$ as $n \to \infty$ as in assumption (A11) we have $\mathfrak{i}_3 = o(1)$.

■

Chapter 3

Voxel Selection using Penalized Least Squares for a Separable Space-Time Covariance model

3.1 Functional Magnetic Resonance Imaging (fMRI)

Magnetic Resonance imaging (MRI) uses strong a magnetic field and pulsating radio waves to capture the structure of the brain by measuring magnetic properties of certain molecules (ex: water by exciting hydrogen nuclei) that has varying densities in different parts of the brain like white matter, gray matter, brain stem, tumors, blood vessels etc. Different pulse sequences of the radio waves can be constructed to study different tissue properties, thus making the MRI an extremely flexible and powerful clinical tool. It is also among the most non-invasive procedures unlike computed axial tomography (CAT) scan that uses radiation and PET scans which requires the subject to be injected with a radioactive label. MRI as an imaging modality is meant to take a static image of the brain and has been successful in identifying structural anomalies related to certain diseases. Functional MRI (fMRI) on the other hand is used to detect brain function and connectivity.

Primarily interested in understanding neural activity in the brain, fMRI provides an indirect insight of neural function by measuring a BOLD (blood-oxygen-level-dependent) signal which is the vascular response to high neural activity. The BOLD signal is the ratio of oxygenated to de-oxygenated hemoglobin in the blood. The fMRI scanner is therefore tuned to detect the BOLD signal changes with high spatial resolution to detect significantly active parts of the brain responding to an external stimulus. Experiments are conducted either using block designs or event-related designs where a particular subject undergoes a series of active and resting states either performing a task or responding to an environmental stimulus. During this process 3D images are taken as quickly as possible so as to detect local increases in blood oxygenation at active areas of the brain in real time. These studies are tremendously interesting to neuroscientists and psychologists alike to study brain function and behavior related aspects of the human brain. Detailed explanations of these processes and further references maybe found in the book by Ashby (2011) that also introduces regular statistical techniques that are used in studying fMRI data.

Hemodynamic Response Function:

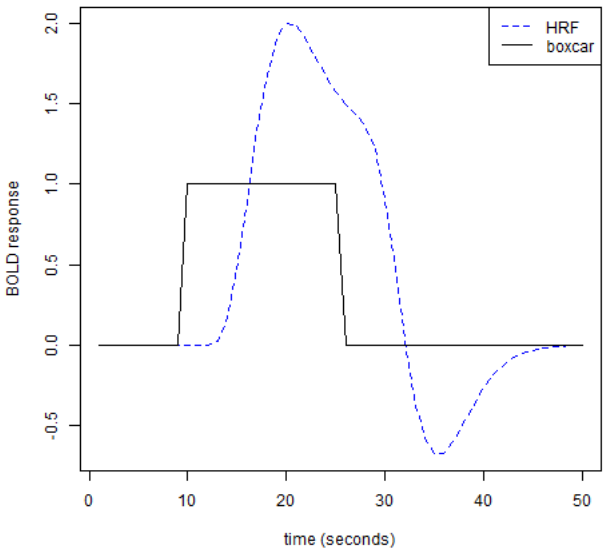
The proof-of-concept experiments indicate that through a process called *hemodynamic response* blood releases oxygen to active neurons at a greater rate than inactive neurons to meet the metabolic demands created through neural activity. Let us denote a box-car {0,1} design $s_{i,t}$ to be the i^{th} stimulus provided to a subject over time t. For an active voxel responding to the stimulus it is observed that although the BOLD signal increases with neural activation it is also sluggish. It is further hypothesized that once the neural activity stops, an oxygen debt occurs dropping the BOLD signal below baseline levels. Consequently it is necessary for the design matrix to undergo a transformation which would be a convolution of

the box-car sequence s and a hypothesized hemodynamic response function (HRF,$h(\lambda, t)$).

$$x_{i,t} = \int_0^t h(t-\tau)s_{i,\tau}d\tau$$

(Ashby, 2011, Chapter 3) discusses in detail multiple ways of estimating the HRF such as deconvolution and linear approximations. The parameters of $h(\cdot)$ are usually estimated based on a pre-processing step for the hypothesized HRF. The most common HRFs used are Poisson, Gamma or canonical (double-gamma; difference of two gamma). Below is an example of a hypothesized canonical HRF convolved with a boxcar stimulus.

Figure 3.1: A theoretical double-gamma HRF characterizing the BOLD signal of a voxel that may respond to the stimulus with some lag and undershoot over time



Thus we are interested in locating those regions that are significantly active and responding to stimuli based on how close the observed BOLD response is to the convolution of this hypothesized behavior.

3.2 Model Description

Single Subject Model:

Measurements obtained from an fMRI experiment detect brain activity associated with changes in blood flow when subjected to a stimulus repeatedly. The BOLD effect is the basis of fMRI. Here we propose a selection technique for a single subject study to detect those active parts βs (amplitudes of activation) that significantly relate to changes in blood flow incorporating both spatial and temporal dependencies and not requiring additional hypothesis testing steps.

Let $y_{v,t}$ be denote a measurement obtained from a Task-based fMRI experiment at voxel $v = 1, ..., N$ and time $t = 1, .., \mathcal{T}$. Consider the following additive model,

$$y_{v,t} = z_t^T \alpha_v + x_t^T \beta_v + \epsilon_{v,t} \tag{3.1}$$

where $Z = \{z_1, ..., z_\mathcal{T}\}$ and $Z\alpha_v$ is a baseline stimulus independent trend, $\beta_v = \{\beta_{v,1}, ..., \beta_{v,p}\}$ denotes the *amplitude* of the activation associated with all p stimuli on voxel v of the fMRI experiment. $x_t = \{x_{t,1}, .., x_{t,p}\}^T$ is the row-vector of the $\mathcal{T} \times p$ design matrix X that consists of the convolution of a boxcar stimulus and a hypothesized HRF and lastly the error term $\epsilon_{v,t}$ for time $t = 1, .., \mathcal{T}$ and voxel $v = 1, .., n$ accounts for the spatio-temporal dependence in the responses.

For simplicity and feasibility, we use a separable model to provide a structure for the spatial and temporal dependence of the responses at each voxel over time. i.e. $var(\epsilon_{v,t}) = \sigma^2$, $cov(\epsilon_{v,t}, \epsilon_{v',t'}) = \nu(|v - v'|)\rho(|t - t'|)$ where $\nu(\cdot)$ is a stationary isotropic spatial correlation model and $\rho(\cdot)$ is a stationary correlation function modeling dependence over time for $t \neq t'$

and $v \neq v'$.

In vector form let, $Y = \{y_1, .., y_n\}$ where $y_v = \{y_{v,1}, ..., y_{v,\mathcal{T}}\}$ for each voxel $v = 1, .., n$ which is a stacked vector of dimension $n\mathcal{T} \times 1$. The amplitude co-efficients β is a stacked $pn \times 1$ dimensional vector over each stimulus. Let Γ and R be correlation matrices corresponding to $\nu(\cdot)$ and $\rho(\cdot)$ respectively such that $cov(Y) = \sigma^2 Q^{-1}$ where $Q^{-1} = \Gamma \otimes R$ of dimension $n\mathcal{T} \times n\mathcal{T}$. The design matrix must be rewritten as $\bar{X} = I \otimes X$ where I is an $n \times n$ identity matrix. The equation in vector form is therefore given by,

$$Y = Z\alpha + \bar{X}\beta + \epsilon \tag{3.2}$$

where $\epsilon \sim N_{n\mathcal{T}}(0, \sigma^2 Q^{-1})$.

3.3 Estimation and Selection

Without loss of generality, let us assume that the response BOLD signal is adjusted to the stimulus independent baseline i.e. $Z\alpha = 0$. In the usual case of a Gaussian error in model 3.2, the parameters of interest β are estimated by minimizing the MLE or least squares objective function as shown below,

$$\begin{aligned} &\hat{\beta} = argmin_{\beta} S(\beta) \\ &S(\beta) \coloneqq (Y - \bar{X}\beta)^T Q (Y - \bar{X}\beta) \end{aligned} \tag{3.3}$$

To correctly identify activated voxels, we use a regularization approach to the problem by introducing certain penalty terms with the interest in simultaneously selecting and estimating the amplitude (β) associated with voxels responding significantly to the stimulus. For

simplicity, let us consider the scenario where only $\mathbf{p} = \mathbf{1}$ stimulus is used in a single-subject study. The estimated coefficients are therefore obtained by minimizing,

$$S(\lambda_1, \lambda_2, \beta) \coloneqq \frac{1}{2}(Y - \bar{X}\beta)^T Q(Y - \bar{X}\beta) + \lambda_1 \sum_{v=1}^{n} |\beta_v| + \lambda_2 \beta^T (I - MWM)\beta \tag{3.4}$$

where W is the adjacency matrix of dimension $n \times n$ based on the sampling locations such that $w_{ij} = 1$ if a voxel is a neighbor and 0 otherwise. $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters associated with selection (LASSO, l_1 penalty) and smoothing penalty terms respectively. M is a diagonal matrix with $M_{ii} = \frac{1}{\sqrt{w_{i+}}}$ where $w_{i+} = \sum_{j=1}^{n} w_{ij}$. The matrix $(I - MWM)$ is specifically,

$$(I - MWM)_{ij} = \begin{cases} 1 - \frac{1}{w_{i+}}, & \text{if } i = j \text{ and } w_{i+} \neq 0 \\ -\frac{1}{\sqrt{w_{i+} w_{j+}}}, & \text{if } i \sim j \\ 0, & \text{otherwise} \end{cases}$$

This representation is in the form of a Laplacian matrix on a Graph and thus making sure that $(I - MWM)$ is a positive semi-definite with 0 as the smallest eigenvalue and 2 as the largest eigenvalue as stated in (Li & Li, 2010, Section 2.1). Further this smoothing penalty allows us to express the penalty term in the form of sums of differences of βs in the following way,

$$\beta^T (I - MWM)\beta = \sum_{i \sim j} \Big(\frac{\beta_i}{\sqrt{w_{i+}}} - \frac{\beta_j}{\sqrt{w_{j+}}} \Big)^2 \tag{3.5}$$

Commonly seen in undirected Graph-based semi-supervised learning as in X. Zhu (2011), this formulation penalizes those amplitudes on neighboring nodes that differ widely and

have a smaller penalty for clusters that are less different. Thus incorporating a natural area activation concept.

Remark 3. The adjacency matrix considered as W in 3.6 provides weights to immediate neighbors or a pre-defined sized clusters. Some examples of adjacency matrices that can be constructed and implemented may be found in (J. Huang et al., 2011, Section 3) with regard to euclidean distances, dissimilarity measures, power adjacency, cluster analysis etc.

For known covariance structure and parameters, we can further simplify 3.4 by using a cholesky decomposition for Q (= inverse of $Cov(Y)$) i.e. $Q = LL^T$ where L is a lower-triangular matrix. Define $Y^* = L^T Y$ and $X^* = L^T \bar{X}$. Then the objective function can be rewritten as,

$$\mathbb{S}(\beta, \lambda, \alpha) = \frac{1}{2} \sum_{v=1}^{n} \sum_{t=1}^{\mathcal{T}} (y^*_{vt} - x^*_t \beta_v)^2 + \lambda P_\alpha(\beta) \tag{3.6}$$

where,

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\beta^T(I - MWM)\beta + \alpha\|\beta\|_1 = (1-\alpha)\frac{1}{2}\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{w_{i+}}} - \frac{\beta_j}{\sqrt{w_{j+}}}\right)^2 + \alpha\sum_{v=1}^{n}|\beta_v|$$

is the convex combination of the two penalty functions and $\alpha = \lambda_1/(\lambda_1 + 2\lambda_2)$ and $\lambda = (\lambda_1 + 2\lambda_2)/2n$.

Remark 4. It is important to note that this formulation is based on a single $p = 1$ stimulus. If additional stimuli are added, it affects the formulation with respect to the tuning parameters involved. We are at liberty to either restrict the objective function to two distinct parameters associated with the type of penalty or free them to have $2p$ tuning parameters in total, with no interactions associated in the penalty between stimuli or allow to incorporate

that as well. This liberal representation is currently not addressed in this dissertation.

3.3.1 Coordinate Descent Algorithm

The coordinate descent algorithm explicated by Friedman et al. (2007) is implemented to solve 3.6. For effective calculation of the derivatives of the objective function, we reconstruct 3.6 as follows,

$$\mathbb{S}(\beta, \lambda, \alpha) = \mathbb{S}_1(\beta) + \mathbb{S}_2(\beta) \tag{3.7}$$

where for a particular voxel u,

$$\mathbb{S}_1(\beta) = \frac{1}{2} \sum_{v \neq u, v=1}^{n} \sum_{t=1}^{\mathcal{T}} (y^*_{vt} - x^*_t \beta_v)^2 + \frac{1}{2} \sum_{t=1}^{\mathcal{T}} (y^*_{ut} - x^*_t \beta_u)^2 \tag{3.8}$$

$$\frac{\partial \mathbb{S}_1}{\partial \beta_u} = -\sum_{t=1}^{n} x^*_t y^*_t + \beta_u \sum_{t=1}^{\mathcal{T}} x^{*2}_t \tag{3.9}$$

and

$$\begin{aligned} \mathbb{S}_2(\beta) =& \lambda\alpha \sum_{v \neq u, v=1}^{n} |\beta_v| + \lambda(1-\alpha)\frac{1}{2} \sum_{z \sim v; z, v \neq u} \Big(\frac{\beta_z}{\sqrt{w_{z+}}} - \frac{\beta_v}{\sqrt{w_{v+}}}\Big)^2 \\ &+ \lambda\alpha|\beta_u| + \lambda(1-\alpha)\frac{1}{2} \sum_{u \sim v} \Big(\frac{\beta_u}{\sqrt{w_{u+}}} - \frac{\beta_v}{\sqrt{w_{v+}}}\Big)^2 \end{aligned} \tag{3.10}$$

$$\frac{\partial \mathbb{S}_2}{\partial \beta_u} = \lambda\alpha - \lambda(1-\alpha) \sum_{v \sim u} \frac{\beta_v}{\sqrt{w_{u+}}\sqrt{w_{v+}}} + \lambda(1-\alpha)\beta_u \tag{3.11}$$

Combining the derivatives obtained in 3.9 and 3.11,

$$\frac{\partial \mathbb{S}}{\partial \beta_u} = -\Big[\sum_{t=1}^{\mathcal{T}} x_t^* y_t^* + \lambda(1-\alpha)\sum_{v\sim u}\frac{\beta_v}{\sqrt{w_{u+}}\sqrt{w_{v+}}}\Big] + \lambda\alpha + \beta_u(\sum_{t=1}^{\mathcal{T}} x^{*2}_t + \lambda(1-\alpha)) \tag{3.12}$$

We therefore obtain the coordinate wise update form for β_u devised in (Donoho et al., 1994, Section 2.2),

$$\hat{\beta_u} \leftarrow \frac{\mathcal{S}(\sum_{t=1}^{\mathcal{T}} x_t^* y_{ut}^* + \lambda(1-\alpha)\sum_{v\sim u}\frac{\beta_v}{\sqrt{w_{u+}}\sqrt{w_{v+}}}, \lambda\alpha)}{\sum_{t=1}^{\mathcal{T}} x^{*2}_t + \lambda(1-\alpha)}. \tag{3.13}$$

The soft thresholding function $\mathcal{S}(z,\gamma)$ is defined as,

$$\mathcal{S}(z,\gamma) =: sign(z)(|z|-\gamma)_+ = \begin{cases} z-\gamma & \text{if } z>0 \text{ and } \gamma<|z| \\ z+\gamma & \text{if } z<0 \text{ and } \gamma<|z| \\ 0 & \text{otherwise.} \end{cases}$$

Previous updates of β_v are used to estimate the current coordinate β_u and this process continues until convergence is attained. In the special case of $I - MWM = I$ or more simply the Elastic Net (EN) in 3.6, Friedman et al. (2010) constructed coordinate descent algorithms for generalized linear models. The major difference between the Soft thresholding update created so far is that they do not include residual terms associated with $y_v^* - \tilde{y_v^*}$ (partial residuals) associated with the remaining voxels due to the structure of the data. Therefore in order to yield spatial influences of the amplitudes regardless of the already infused spatio-temporal covariance, the smoothing penalty structure plays a key role.

Another version of the update rule can be written in terms of the original tuning param-

eters $\lambda_1 > 0$ and $\lambda_2 > 0$ in the following way,

$$\hat{\beta_u} \leftarrow \frac{\mathcal{S}(\sum_{t=1}^{\mathcal{T}} x_t^* y_{ut}^* + \frac{2\lambda_2}{n} \sum_{v \sim u} \frac{\beta_v}{\sqrt{w_{u+}}\sqrt{w_{v+}}}, \frac{\lambda_1}{n})}{\sum_{t=1}^{\mathcal{T}} x^{*2}_t + \frac{2\lambda_2}{n}}. \tag{3.14}$$

3.3.2 Algorithm implementation

Ordinarily one can evaluate the above update rule for each fixed tuning parameter λ and α in 3.13 or λ_1 and λ_2 for 3.14 successively at each coordinate and continue until convergence is attained.

3.3.2.1 Pathwise Coordinate Descent

Friedman et al. (2007) describe the pathwise co-ordinate descent algorithm as a procedure to be applied in a way that varies the tuning parameter along a path. The solutions are computed on a decreasing sequence of values for λ given some $\alpha \in (0, 1)$, starting at the smallest value λ_{max} such that all of the co-efficients $\hat{\beta} = 0$. The scheme also exploits the notion of warm starts, leading to a more stable algorithm. Warm starts is referred to as the process of providing the smaller λ in the iterative process the start values equal to the converged estimates obtained in the previous λ considered.

It can be shown that $\hat{\beta}_j = 0, j = 1, \ldots, n$ if $\frac{1}{n\mathcal{T}}\langle x_j^*, y_j^* \rangle \leq \lambda\alpha$. Following Friedman et al. (2007) we select $\lambda_{max} = \max |\frac{1}{n\mathcal{T}}\langle x_j^*, y_j^* \rangle|$. The strategy is to select a minimum value $\lambda_{min} = \epsilon\lambda_{max}$ for a typical $\epsilon = 0.0001$, and construct a sequence of K values from λ_{min} to λ_{max} in the logarithm scale. K may be any value, depending on computational capabilities, a common value used is $K = 100$

3.3.2.2 Active Set Convergence

A slightly observation driven approach is considered in the active set convergence that may speed up the above algorithm significantly. Once the coordinate descent path is established for each combination of λ and α, usually an iterative procedure is adopted until convergence of the βs is attained. In active set convergence, after the first iteration we obtain an active set. Then the proceeding iterations are computed only for those non-zero βs the following iterations. Continuing with the next combination of λ the same procedure is employed. If the active set remains unchanged, the remaining process is considered only over the first liberal active set for λ_{max}, significantly reducing the number of iterations at coordinates that are originally considered to be zero.

Unfortunately not a very strong theoretical basis has been provided about the rate of convergence of the method but has been observed to be successful in the LASSO, EN and group LASSO setup (Meier et al., 2008).

3.3.2.3 Cholesky decomposition properties

In the current model implementation we assume that the spatio-temporal separable covariance is known. The separability is rather restrictive supposition but we hope to exploit its matrix properties in order to make the method feasible. It is in general important to understand the scope of this dataset. For fMRI data Lindquist (2008) explains that the number of slices acquired depends on how quickly the image maybe taken when the brain is excited. In approximately 2 seconds, standard scanners can image the whole brain volume of $64 \times 64 \times 30 = 122880$ voxels. In the block design setup the number of time points considered depends on the length of the experiment. Unlike event-related designs that have a signif-

icantly small number of time points (Ex: 4 to 7) but in block designs they may last upto 30 minutes or more with over 200 time points. In this setup the full covariance matrix for a single subject could exceed the dimension 24000000×24000000. Most high performance computers will be unable to store such a massive matrix in memory especially for a dense matrix with both spatial and temporal dependence.

However if this dependence is separable i.e. of the form $cov(Y) = Q^{-1} = \Gamma \otimes R$, in order to obtain the Cholesky decomposition of Q we can use the following properties for positive definite matrices using Kronecker products,

1. $(\Gamma \otimes R)^{-1} = \Gamma^{-1} \otimes R^{-1}$

2. $(\Gamma \otimes R)^{-1} = L_{\Gamma^{-1}} L_{\Gamma^{-1}}^T \otimes L_{R^{-1}} L_{R^{-1}}^T = (L_{\Gamma^{-1}} \otimes L_{R^{-1}})(L_{\Gamma^{-1}} \otimes L_{R^{-1}})^T$

3. $\Gamma^{-1} = (L_\Gamma L_\Gamma^T)^{-1} = (L_\Gamma^T)^{-1}(L_\Gamma)^{-1}$

where L is a lower triangular matrix. These properties enable us to perform matrix operations on smaller matrices based on space and time and eventually combine them and vectorize right away to obtain the weighted least squares without saving the large matrix to memory. The computational complexity of the space matrix and time matrix is dramatically smaller and using row wise Kronecker products and the properties above and multiplying it to the response and design matrix right away directly yields the weighted least squares without the need of allocating a $n\mathcal{T} \times n\mathcal{T}$ matrix. Details of these properties maybe reviewed in Schacke (2004).

3.3.2.4 Tuning Parameter Selection

To consistently identify the true model it is crucial to choose an optimal tuning parameter that ensures the true model is selected every time. Cross validation as in Chapter 2 of this

dissertation is not easily implementable due to the dependence present in a single subject; both with regard to space and time. Therefore we seek to locate the optimal value of the tuning parameters based on a criteria over a grid. Our first candidate criterion is the usual AIC based on the log-likelihood. Additionally based on the method developed by Y. Fan & Tang (2013) we adopt the GIC (generalized information criterion) technique. Under the generalized linear model(GLM) for any independent response variable of the exponential family they find a sequence a_n in such a way that the minimum GIC value corresponds to tuning parameters that are consistently close to the true model. However the fMRI data structure is fundamentally different from a regular linear model since the selection is not on the predictors but on the locations with activated voxels. In the discussion section 3.6 we address the need to modify the GIC by finding a sequence a_n such that an optimal criterion may be obtained for the current setup. As a candidate criteria we investigate whether the GIC for a regular GLM can locate the optimal tuning parameters.

Another attempt is made to use cross validation on block designs by assuming that no carry-over effects are sustained by the single subject while doing an activity repeatedly. We produce the illusion of independent replicates of block designs by splitting the different blocks over time into multiple folds. This drastically reduced the dimensions of correlation matrices used. Permutation matrices and properties of Cholesky decomposition and Kronecker products too support this formulation. However there is a computational trade-off that is involved depending on how many folds are constructed based on the blocks. The optimality of the solution may further depend on these folds. An example of its simulation results are provide in section 3.4.

The Bayesian LASSO introduced by Park & Casella (2008) provides an empirical Bayes estimate of the tuning parameter under Bayesian hierarchical setup exploiting a property of

scalar mixture normals. Based on the LASSO formulation of the EN by Zou & Hastie (2005) another attempt was made at finding optimal solutions. Certain issues automatically rose in the conducting the simulation study and are addressed in the discussion section 3.6.

3.4 Simulation Study

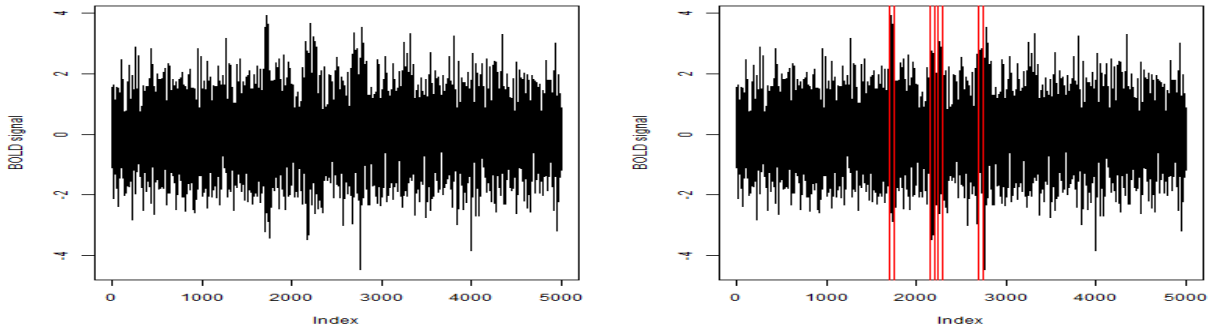
Below are a series of simulations conducted to examine the method proposed. Critical issues of tuning parameter selection arise with regard to criteria required to reach optimal solutions. Several possibilities are presented for a variety of implementations.

General Setup: Let us consider a 10×10, 15×15 and 20×20 regular grid of 100/225/400 locations each a single unit away. For a single boxcar stimulus (i.e. p=1), we construct a hemodynamic response function using the Gamma density for 50 time points. Thus the design matrix is of dimension 5000×1, 11250×1 and 20000×1 respectively. The spatial covariance function imposed on the locations is given by $\nu(|v-u|) = exp(-0.7 \cdot |v-u|)$ and the corresponding matrix produced is Γ. Further we introduce some additional sparsity by using the tapering method by Wendland (1995) that leads to a positive definite matrix. The parameters for the Wendland function used are $\theta = 15$ and $k = 2$. The temporal structure R is given two structures a tri-diagonal matrix with $\rho(|0|) = 1$ for $\rho(|t - t^{'}|) = 0.5$ if $|t - t^{'}| = 1$ or 0 otherwise and an AR(1) with $\tilde{\rho} = 0.5$. A location is selected in the grid and its closest neighbors are given an amplitude $\beta_v = 6$ while all other amplitudes are assigned a 0 as the simulations studies shown in Musgrove et al. (2016). No additional baseline trend is considered. The data is generated from a normal distribution with mean $\bar{X}\beta$ and $Q^{-1} = \Gamma \otimes R$. 100 such datasets are replicated for each setup.

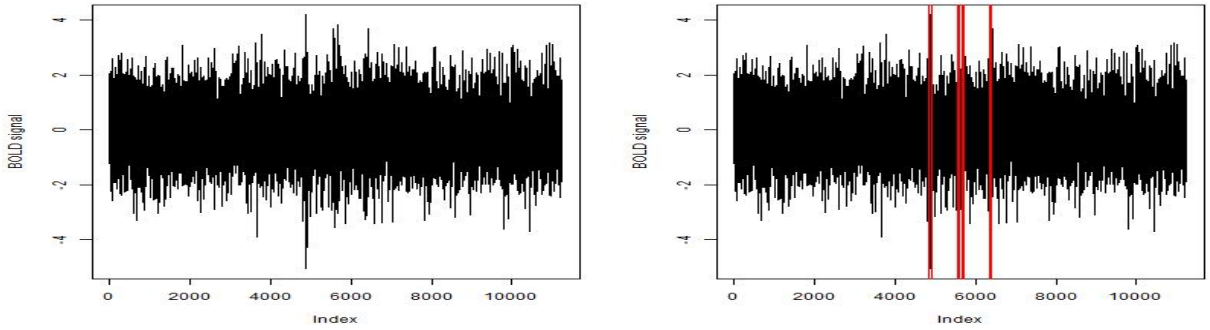
In figures 3.2 an example of the simulated data under two different time structures for

various grid sizes and 50 time points is shown. The red blocks indicate where the signal lies. We can clearly see that albeit the high signal, the data is still very noisy.

Figure 3.2: Simulated fMRI data with different grid sizes and time structures. (a) contains 5000 data points and (b) contains 11250 data points



(a) 10×10 grid with Tri-Diag Time Correlation



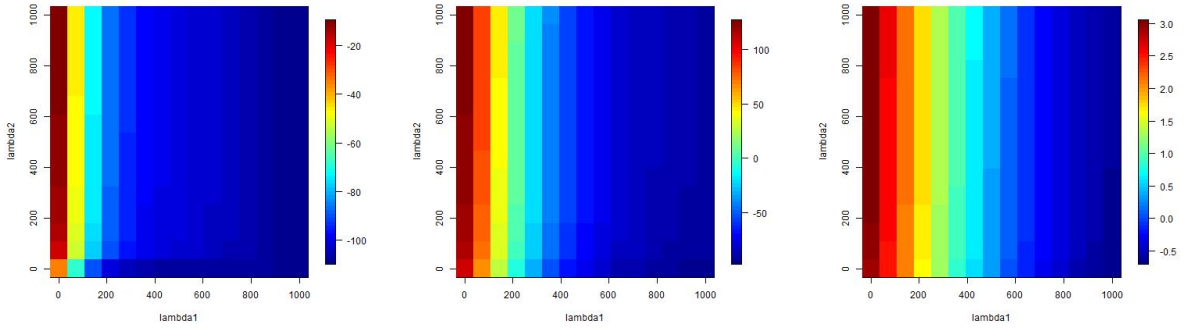
(b) 15×15 grid with AR(1) Time Correlation

To begin with, three different algorithms are used to look for optimal tuning parameters under the assumption that the underlying spatio-temporal model is known.

1. Sequential grid for λ_1 and λ_2: In the general setup considered 3.4, we first consider exploring the efficiency of the selection method over an entire grid for values of the tuning parameters λ_1 and λ_2 as indicated in equation 3.14. A range from 0.1 to 1000 is considered over 15 instances in the range. Below are image plots indicating the AIC and GIC obtained from the various values for all 3 grid types considered, averaged over 100 datasets. In Figures

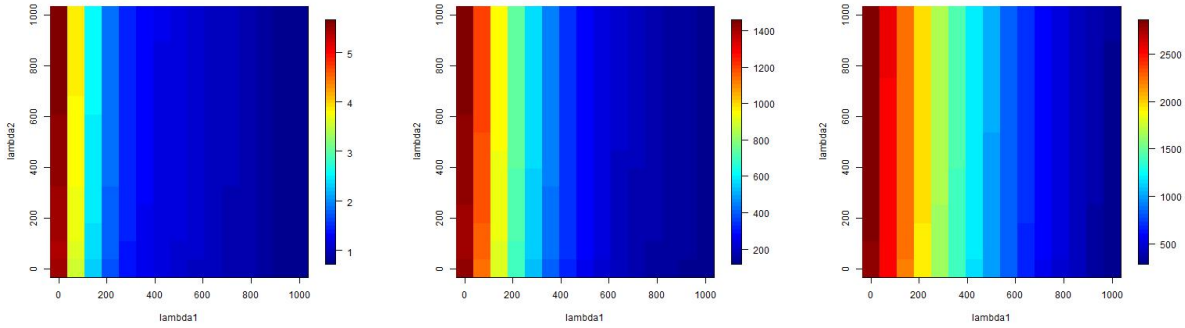
3.3, 3.4 and 3.5 it is distinctly visible that an obvious optimal combination for the tuning parameters is not apparent.

Figure 3.3: AIC for sequences of $\lambda_1 \& \lambda_2 \in (0.1, 1000)$



(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.4: GIC for sequences of $\lambda_1 \& \lambda_2 \in (0.1, 1000)$



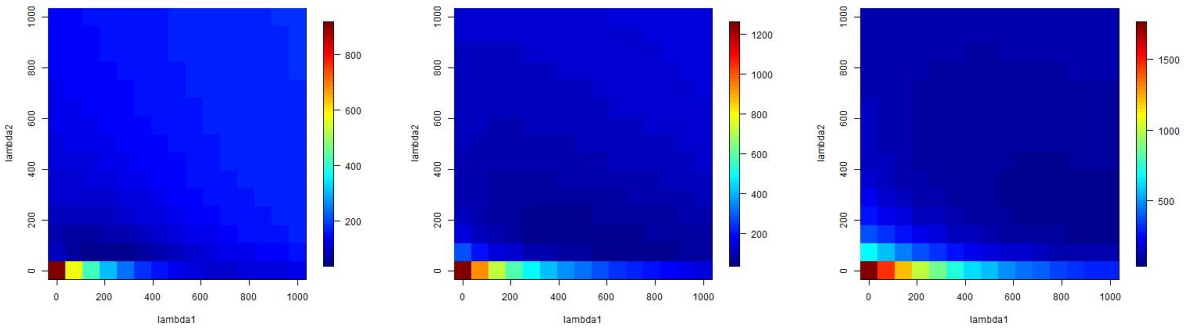
(a) 10×10 (b) 15×15 (c) 20×20

Table 3.1: Choice of Lambda sequence parameters based on MSE

Grid	Lambda_1	Lambda_2	TP	FP	MSE
10×10	142.94	71.52	5.00	31.46	45.57
15×15	500.05	142.94	5.00	29.52	45.13
20×20	1000.00	214.36	5.00	32.53	47.90

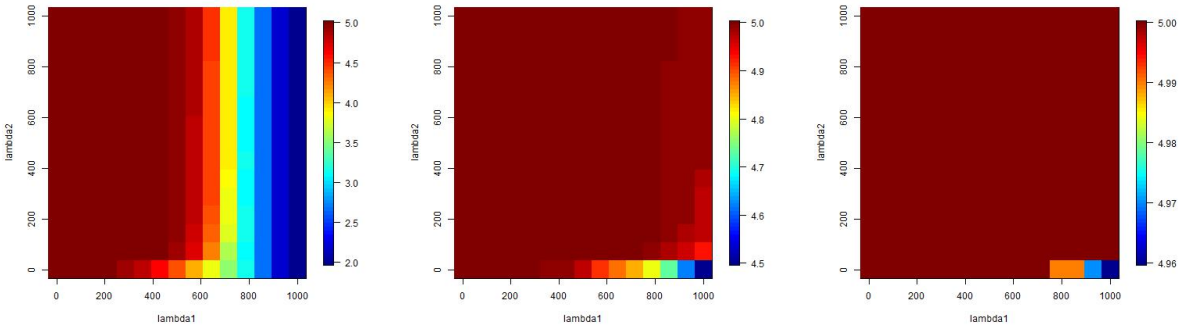
However based on the mean squared error (MSE), since the true parameters of interest β_v are known, we obtain results in Table 3.1 for each of the grids.

Figure 3.5: MSE for sequences of $\lambda_1 \& \lambda_2 \in (0.1, 1000)$



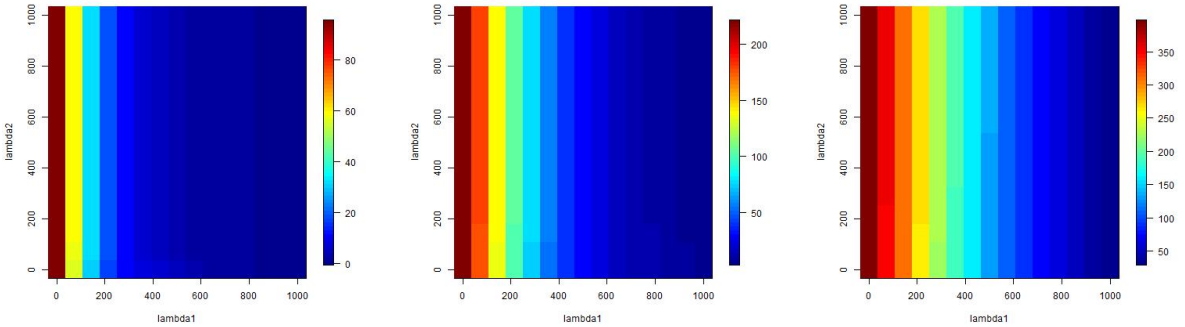
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.6: True Positives (max 5 active) for sequences of $\lambda_1 \& \lambda_2 \in (0.1, 1000)$



(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.7: False Positives (max n-5) for sequences of $\lambda_1 \& \lambda_2 \in (0.1, 1000)$

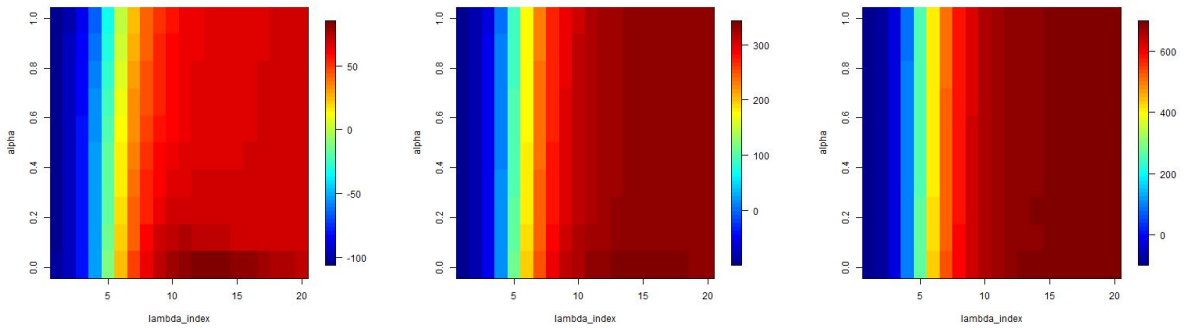


(a) 10×10 (b) 15×15 (c) 20×20

2. Pathwise Coordinate Descent For a fixed value of α a decreasing sequence of 20 values of λ is considered, where $\lambda_{max} = \max \frac{1}{n\alpha} | < \bar{X}, y > |$ and $\lambda_{min} = \epsilon \cdot \lambda_{max}$. Here

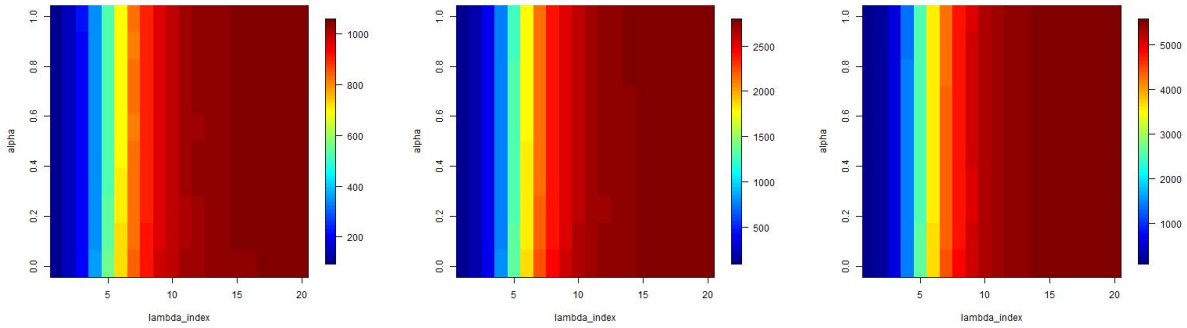
$\epsilon = 0.0001$. Therefore on a sequence of $\alpha \in (0, 1)$ varying over 10 values the selection was run repeatedly over 100 datasets. Image plots for AIC, GIC and MSE are provided in Figure 3.8, 3.9 and 3.10. In this algorithm the mean squared prediction error (MSPE) (See Fig 3.13) seemed more intuitive with regard to the tuning parameters. λ_{max} values for every dataset but did not exceed 5.

Figure 3.8: AIC for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



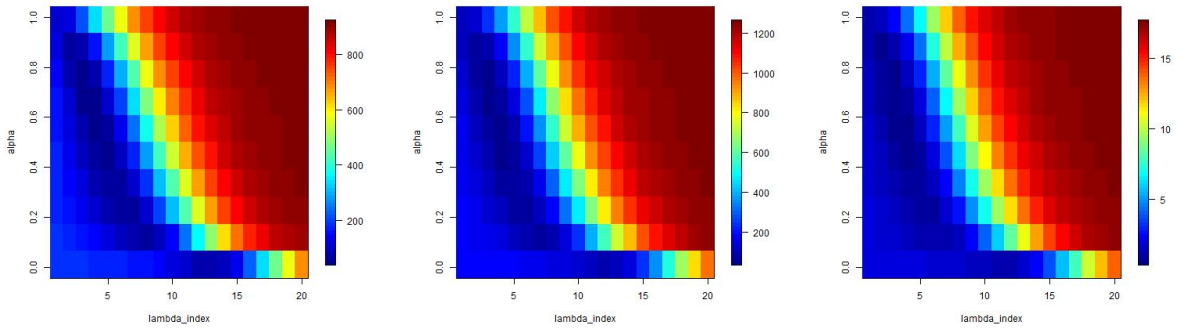
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.9: GIC for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



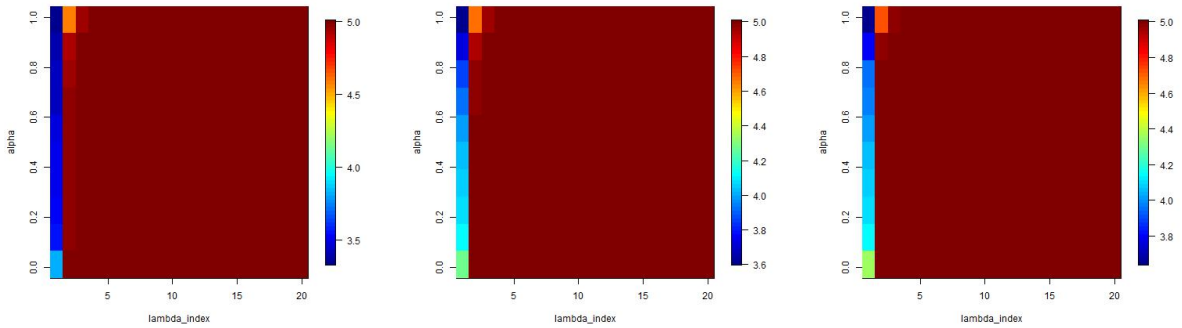
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.10: MSE (Mean squared error) for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.11: True Positives (max 5 active) for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



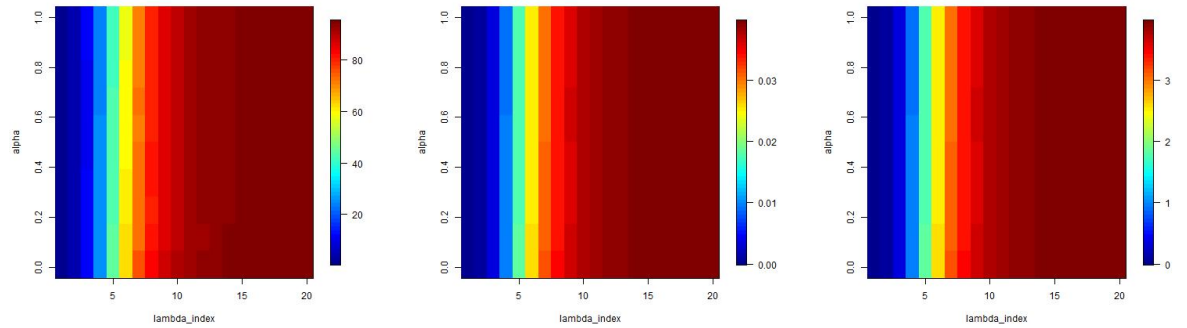
(a) 10×10 (b) 15×15 (c) 20×20

Table 3.2: Choice of Lambda sequence parameters based on MSE

Grid	Lambda index	Alpha	TP	FP	MSE
10×10	4.00	0.55	5.00	24.33	44.89
15×15	3.00	0.77	5.00	20.27	46.85
20×20	3.00	0.66	5.00	31.10	49.66

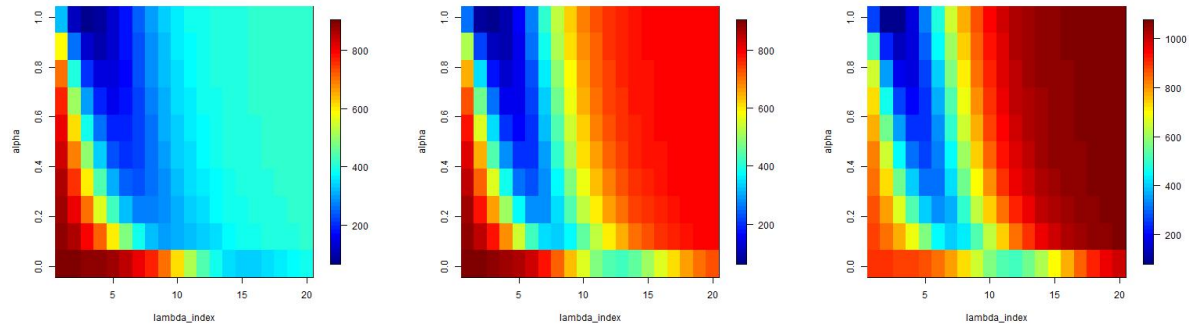
Interestingly the mean squared prediction error for the path-wise coordinate descent algorithm tends to prefer the LASSO solution at $\alpha \approxeq 1$. See Fig 3.13

Figure 3.12: False Positives (max n-5) for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



(a) 10×10 (b) 15×15 (c) 20×20

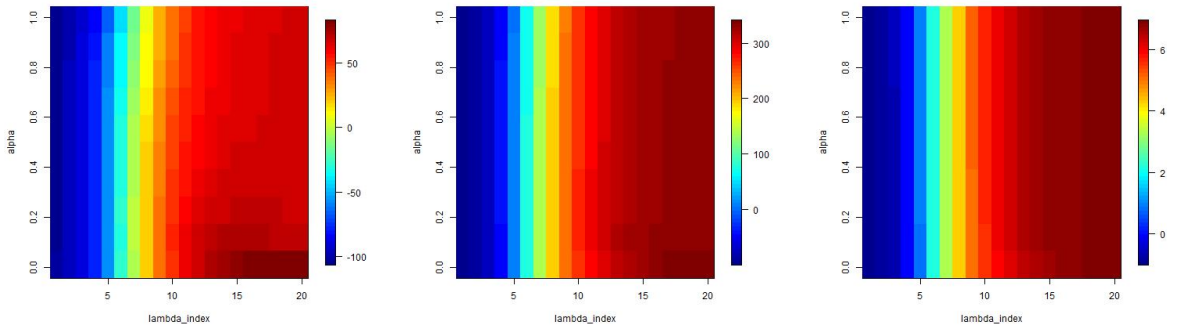
Figure 3.13: MSPE (Mean squared prediction error) for sequences of the index λ values given $\alpha \in (0, 1)$using pathwise coordinate descent



(a) 10×10 (b) 15×15 (c) 20×20

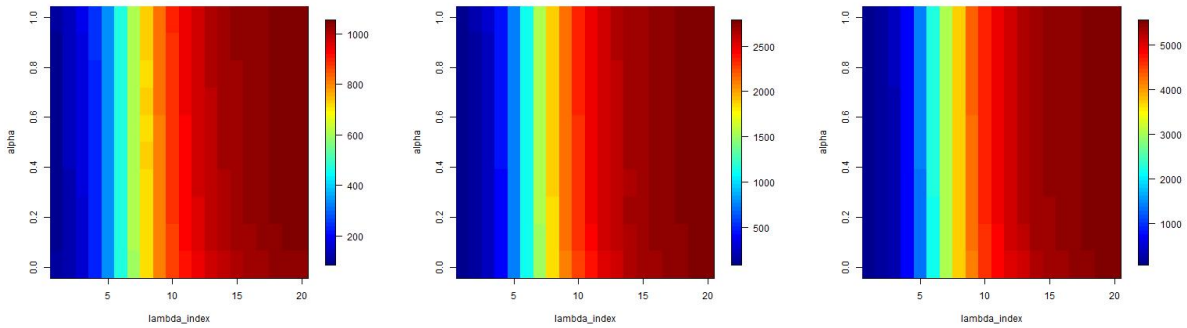
3. Active Set Convergence In active set convergence, we consider an approach similar to the path wise coordinate descent except for the consecutive sequences of λ we provide the solution of the previous iteration as a warm start. Then we proceed to do a coordinate descent only those voxels/coordinates that have not already been dropped. This results in a relatively speedy algorithm and leads to rather sparse solutions.

Figure 3.14: AIC for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



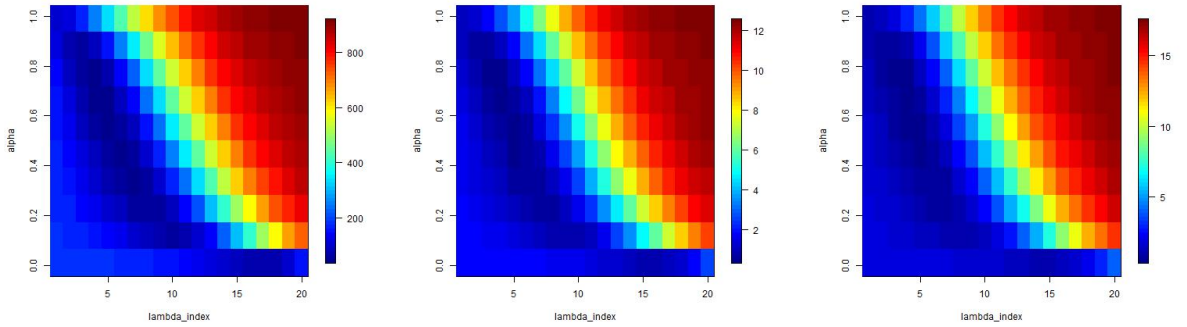
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.15: GIC for sequences of the index λ values given $\alpha \in (0, 1)$ using pathwise coordinate descent



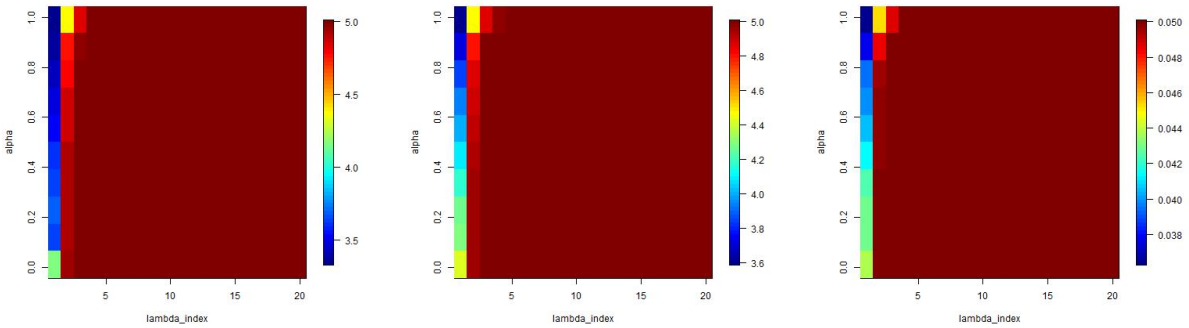
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.16: MSE (Mean squared error) for sequences of the index λ values given $\alpha \in (0, 1)$ using active sets



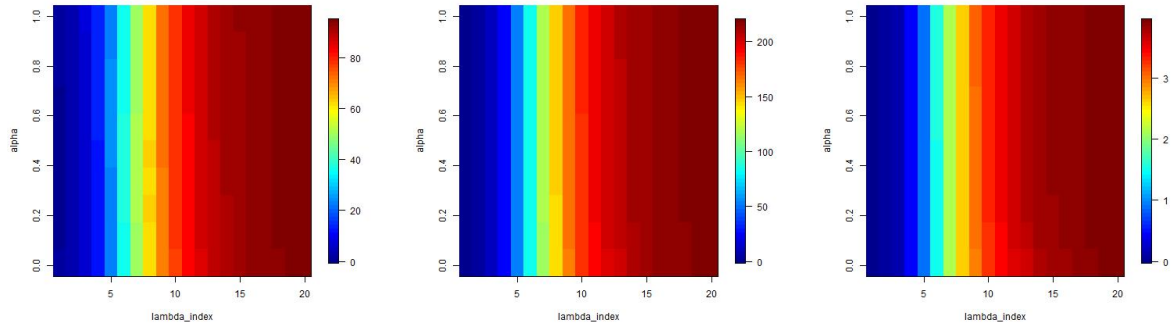
(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.17: True Positives (max 5 active) for sequences of the index λ values given $\alpha \in (0, 1)$ using active sets



(a) 10×10 (b) 15×15 (c) 20×20

Figure 3.18: False Positives (max n-5) for sequences of the index λ values given $\alpha \in (0, 1)$ using active sets

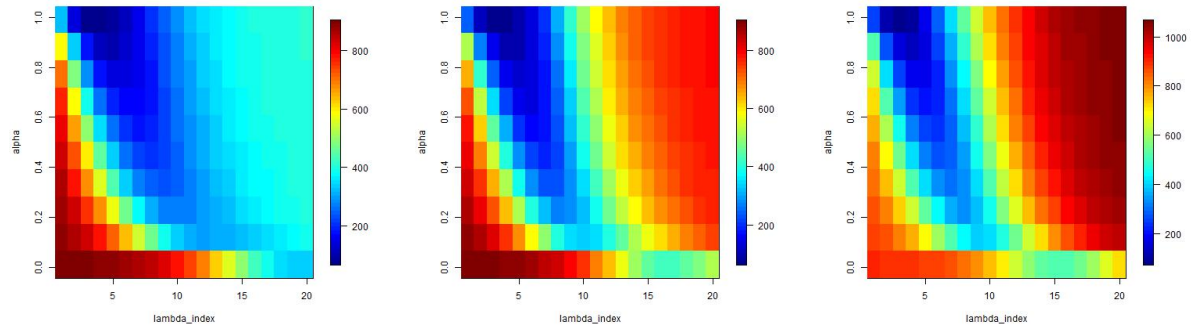


(a) 10×10 (b) 15×15 (c) 20×20

Table 3.3: Choice of Lambda sequence parameters based on MSE

Grid	Lambda index	Alpha	TP	FP	MSE
10×10	5.00	0.55	5.00	23.51	44.91
15×15	4.00	0.66	5.00	26.31	0.45
20×20	4.00	0.66	5.00	41.89	0.48

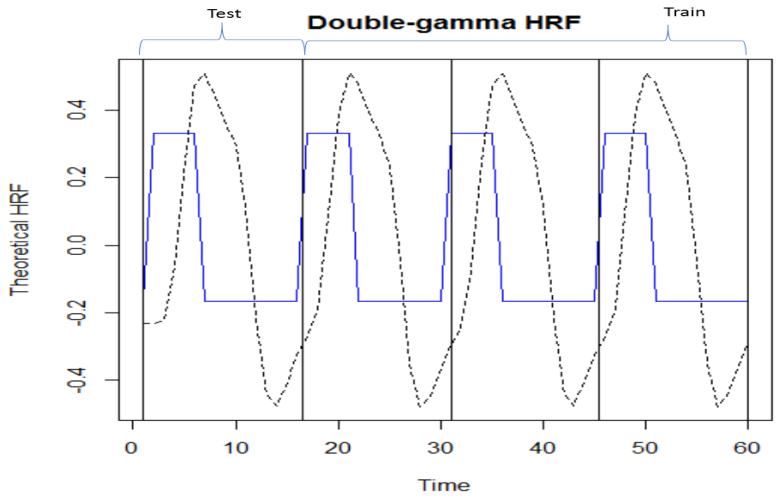
Figure 3.19: MSPE (Mean squared prediction error) for sequences of the index λ values given $\alpha \in (0, 1)$ using active sets



(a) 10×10 (b) 15×15 (c) 20×20

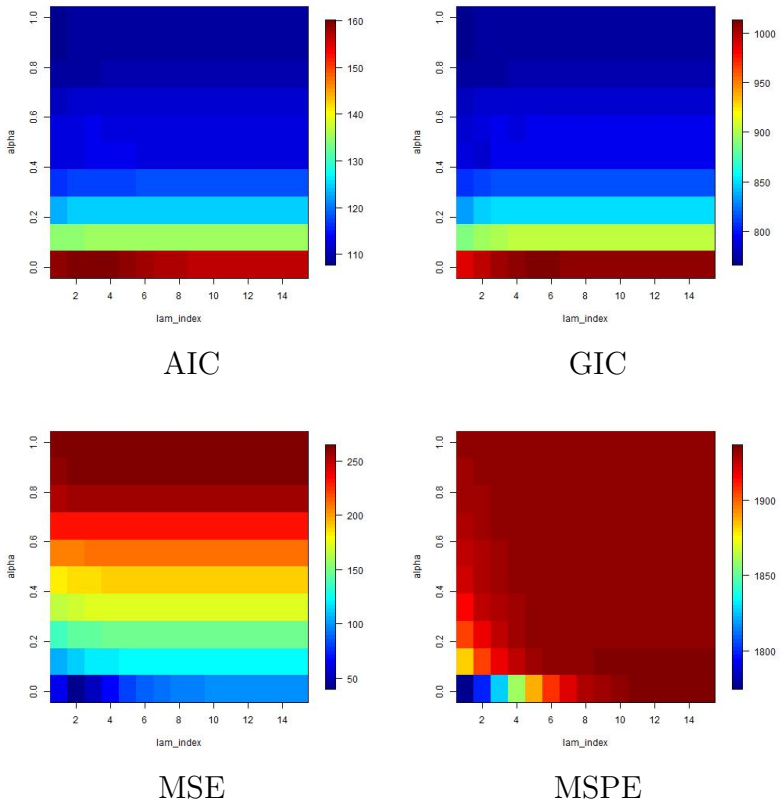
Cross Validation Regularization methods use cross validation to obtain optimal tuning parameters in big data settings with independent replicates. As explained in section 3.3.2.4 we try to exploit nature of the experimental design to produce an illusion of replicates. Assuming there are no carryover effects in these fMRI block-design experiments with respect to time let us consider the convolved double- gamma (canonical) HRF with a box-car stimulus shown in Figure 3.20.

Figure 3.20: 4-fold CV on time for design of the single subject fMRI study. Blue solid line indicates boxcar stimulus and dotted line convolved HRF



Each cycle of the experiment is considered as a fold. In the Figure 3.20 without loss of generality we can place the first as the test set and the rest as the training set and continue until each fold has been the test set once. We average different the criteria and MSPE to find optimal solutions. This is done on a single dataset, 10×10 grid. However in this analysis it is intuitively clear that the criteria considered (like AIC and GIC favor LASSO (see Figure 3.21) or the modified "Laplacian" Ridge or smoothing penalty based on MSPE and MSE which is expected when the objective is to reduce prediction error in the hope to reduce bias. The MSE and MSPE prefer the smoothing penalty as they want the least amount of bias

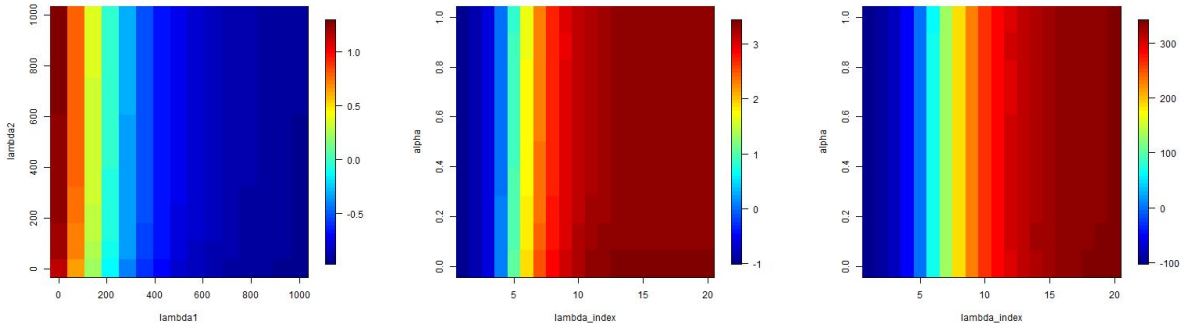
Figure 3.21: Simulation results of 4-fold cross validation on a 10×10 grid



and AIC, GIC prefer picking the LASSO penalty as the criteria is proportionate to number of voxels in the model. Thus there is no indication that CV may yield optimal solutions.

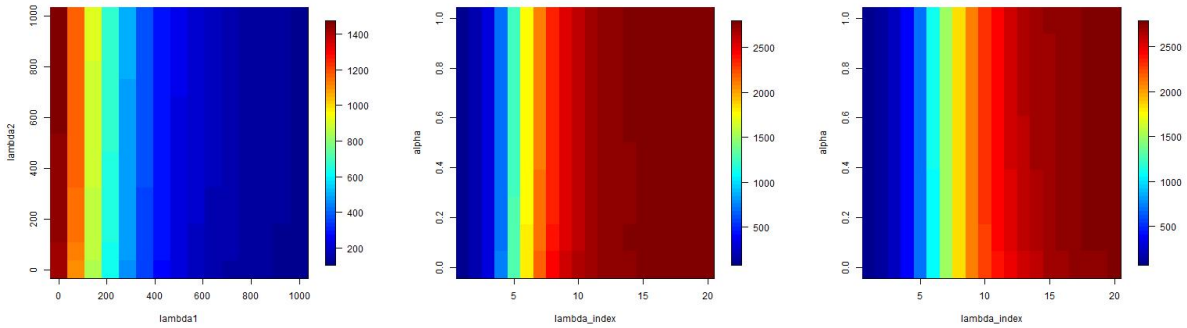
Simulation study investigating AR(1) Time-Structure Similar to the general setup provided in the simulation studies above, we wanted to investigate whether the method can be implemented successfully whenever a more commonly used and better suitable time correlation structure is used. In the current setup we look at a single pseudo image grid 15×15. The spatial covariance and the wendlend taper function is still used to simulate the data. For the time structure we generate data with a voxel vise auto-regressive AR (1) model where parameter $\rho = 0.5$. The total experiment is assumed to have 50 time points with the same HRF used for the design matrix. All three algorithms are considered and compared with each other in the image plots 3.22- 3.27.

Figure 3.22: AIC for model simulated using AR(1) Time correlation structure



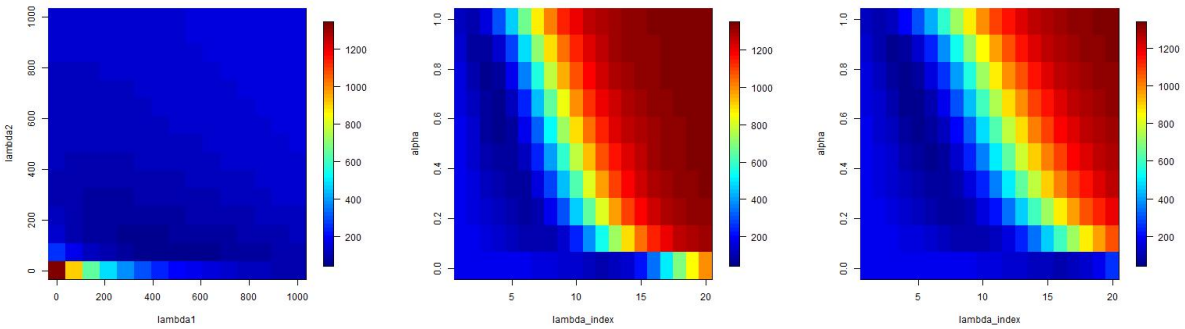
(a)Sequence CD (b) Pathwise CD (c) Active Set CD

Figure 3.23: GIC for model simulated using AR(1) Time correlation structure



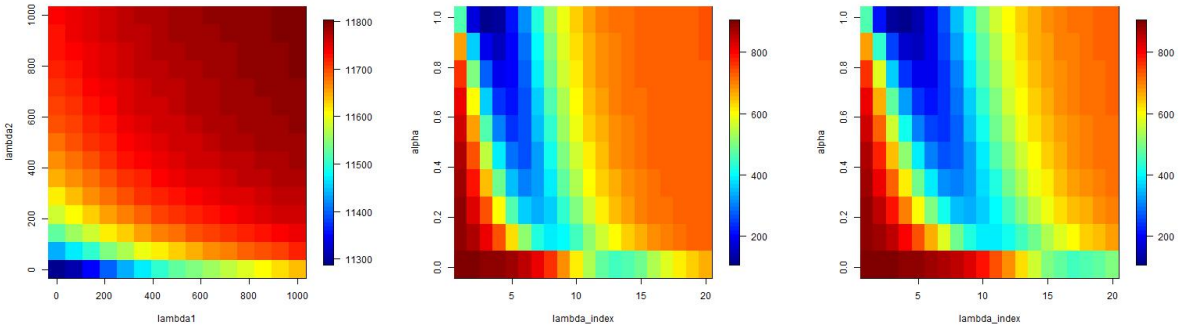
(a)Sequence CD (b) Pathwise CD (c) Active Set CD

Figure 3.24: MSE for model simulated using AR(1) Time correlation structure



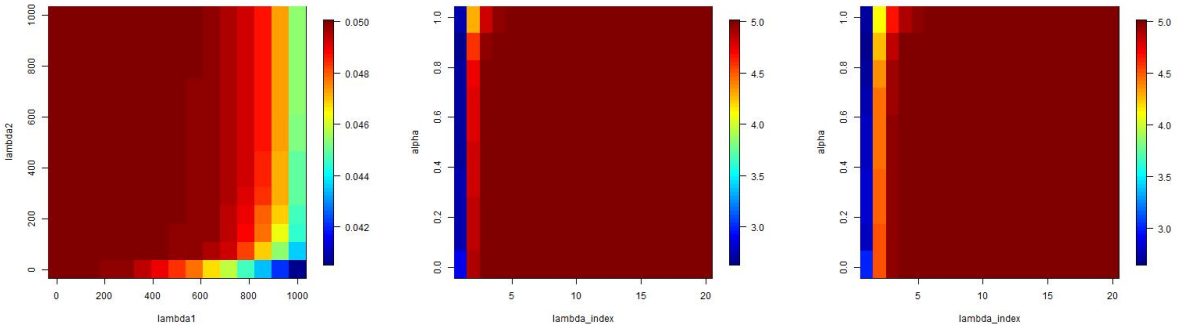
(a)Sequence CD (b) Pathwise CD (c) Active Set CD

Figure 3.25: MSPE for model simulated using AR(1) Time correlation structure



(a)Sequence CD (b) Pathwise CD (c) Active Set CD

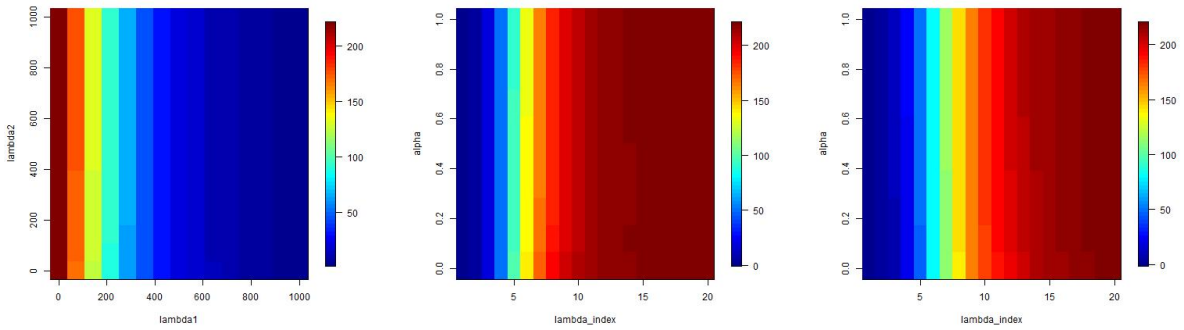
Figure 3.26: True Positives (max 5) for model simulated using AR(1) Time correlation



(a)Sequence CD (b) Pathwise CD (c) Active Set CD

The use of an AR(1) model instead of a tridiagonal time structure matrix provides very

Figure 3.27: False Positives (max n-5) for model simulated using AR(1) Time correlation



(a)Sequence CD (b) Pathwise CD (c) Active Set CD

Table 3.4: Consolidated results for AR(1) simulation with regard to all 3 algorithms

	Lambda1	Lambda2	TP	FP	MSE	
Sequence CD	500.05	71.52	4.99	21.54	52.16	
	Lambda index	Alpha	TP	FP	MSE	MSPE
Pathwise CD	3.00	0.77	4.99	17.39	53.61	113.74
Active Set CD	4.00	0.77	4.99	23.08	52.04	113.01

similar results. A noticeable increase in the computational time is experienced due to increase in the density of the matrix. A steady 10% False positive rate is observed with an almost 100% true positive selection.

3.5 Real Data Analysis

Data Acquisition

A healthy college student from Michigan State University volunteered to participate in a study on a visual stimulation condition with a scene-object fMRI paradigm. Signed consent forms approved by the Michigan State University Institutional Review Board (IRB) were obtained from the individual. The experiment was conducted on a 3T GE Signa HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil. The parameters for the fMRI scan that were collected 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 seconds "resting" period. In a predetermined randomized order, scenery pictures were presented in 6 blocks and objects pictures were presented in the other 6 blocks. All pictures were unique. In each block, 10 pictures were presented continuously for 25 seconds (2.5 s for each picture), followed with a 15 second baseline condition where the subject is defaulted to 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 on a full screen 1024×768 32-inch LCD monitor (Salvagione Design, Sausalito, CA) placed at the back of the magnet room. The LCD was subtended at a visual angle of $10.2° \times 13.1°$. 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.

fMRI Data Pre-processing

All stimulus fMRI data pre-processing were conducted with AFNI software (Cox, 1996) as described in Henderson et al. (2011). The data was detrended for artifacts and the stimulus independent baseline specifically, slice-timing correction and rigid-body motion correction were carried out. 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 by general linear tests that determine statistical significances between

stimulus conditions. Once this was completed a mask was created that discarded all of the potentially irrelevant voxels of the brain with regard to the stimuli. This significantly reduced the size of the data from $64 \times 64 \times 31 \times 192 \approx 25$ million data points to 6118 voxels with the original 192 time points ≈ 1 million.

The current model proposed assumes the underlying spatio-temporal covariance is known. In real data studies this is always untrue and the covariance needs to be estimated. This need is addressed in the discussion section 3.6 but in order to showcase the method we proceed by eye-fitting a model on a spatio-temporal variogram. The R package *spacetime* is used to fit a separable exponential models for both space and time. The model that reduced the MSE was eventually considered.

Data Analysis

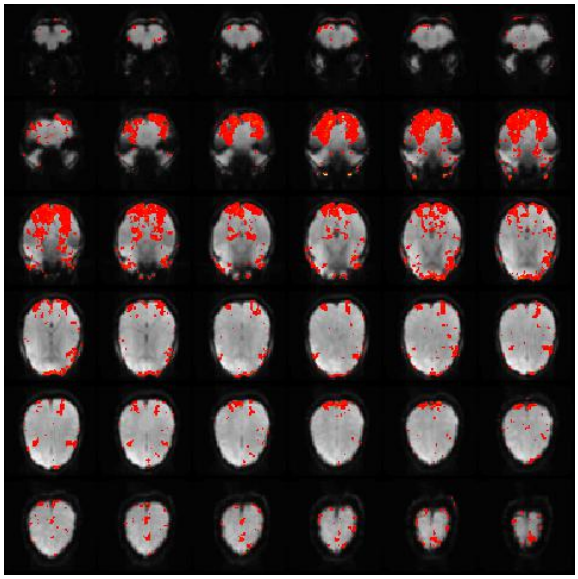
Once all of the preprocessing was done and an underlying spatio-temporal model was considered, using the Kronecker product and cholesky decomposition properties in section 3.3.2.3 we used a linear algorithm to avoid saving the Kronecker product to memory. Using this vectorized formulation we could directly calculate the transformed y^* and x^* as explained in section 3.3. As explained above, the design consists of two stimuli scene and object provided at random. Based on remark 4 we use the most conservative approach and restrict the tuning parameters to be identical. It is important to note that since only an instantaneous image was recorded for both stimuli they are required to be studied simultaneously.

Using the path-wise coordinate descent method in 3.3.2.1 we simplistically vary $\alpha \in (0, 1)$ over 10 points and obtain the sequence of λ between $\lambda_{min} = \epsilon.\lambda_{max}$ and λ_{max} for $\epsilon = 0.0001$ by selecting

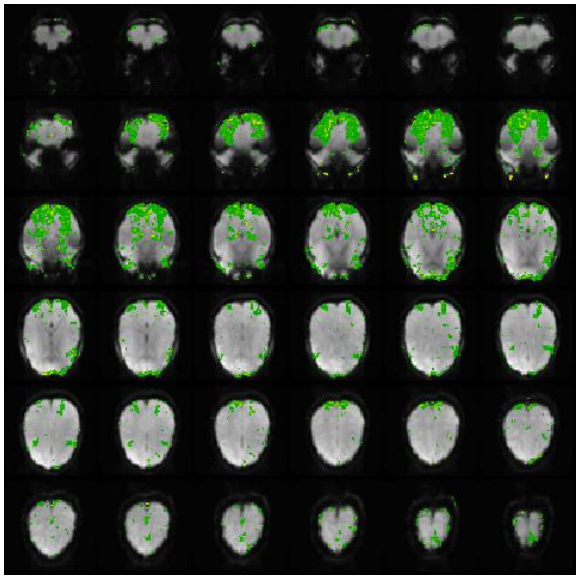
$$\lambda_{max} = \max\{|\frac{1}{n\alpha}\langle x^*_{object_j}, y^*_j\rangle|, |\frac{1}{n\alpha}\langle x^*_{scene_j}, y^*_j\rangle|\}$$

Based on all the simulation studies done above and the computational feasibility of the path-wise coordinate descent method, we provide an application of the method to a single-subject study with two visual stimuli (p=2). The results in the simulation studies above provide a range of $\alpha \in (0.5, 0.7)$ whenever the criteria for optimality is the MSE for true known coefficients. Although in actuality, the coefficients are never known the selection based on this criteria provides optimal solutions that where it takes advantage of both penalties. Thus we crudely consider an $\alpha = 0.6$ from the studies above and we look at a sequence of decreasing λs. Below are images of the 36 slices of the axial view and 64 slices of the coronal and sagittal view showing the selected voxels based on 3 choices of λ based on both stimuli (object and scene). Alternatively fixing α and varying the λ based on the stimuli has no novel effect as the intensity of both scene and object on the subject is identical. Therefore we proceed in the way described above. The results appear to coincide with the study conducted by Henderson et al. (2011). Most significantly we see activity in the visual cortex and frontal lobe of the brain. Further the method is currently unable to distinguish clearly activated regions responding to the two stimuli differently, so in the Figures below we see a fair amount of overlap.

Figure 3.28: Axial View $\lambda = 0.0019$

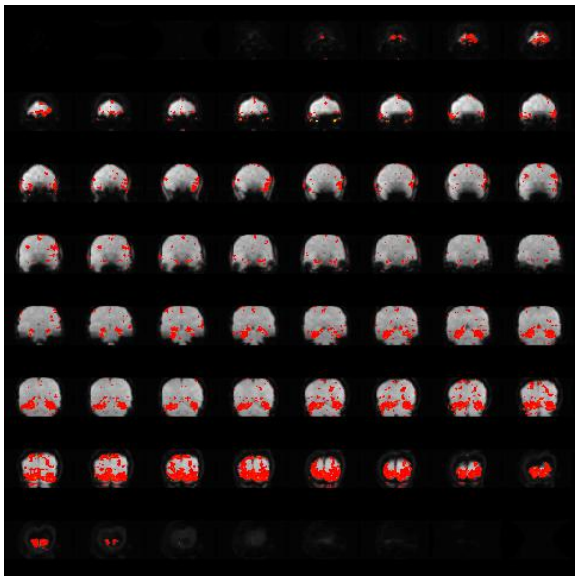


Object Stimulus

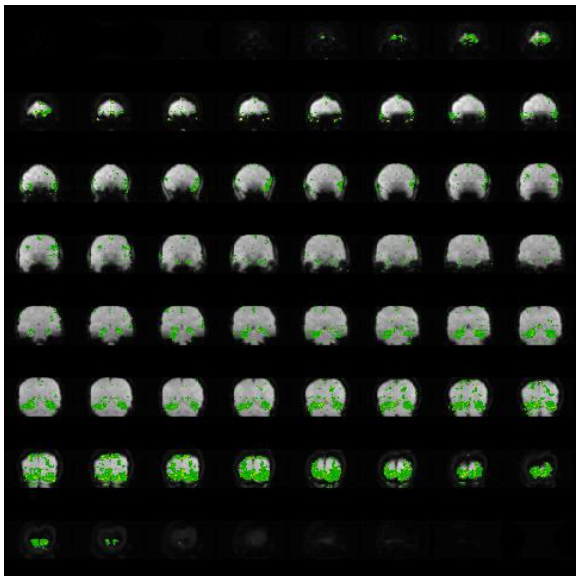


Scene Stimulus

Figure 3.29: Coronal View $\lambda = 0.0019$

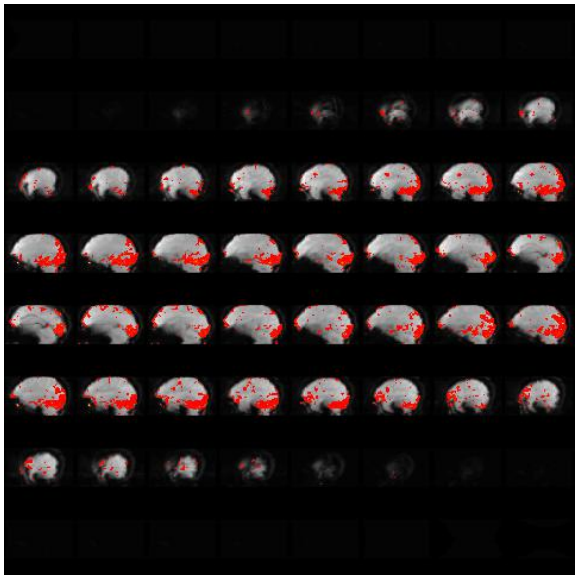


Object Stimulus

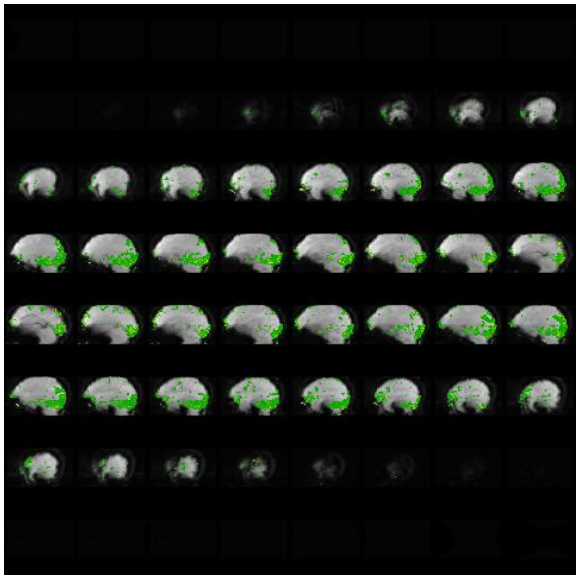


Scene Stimulus

Figure 3.30: Sagittal View $\lambda = 0.0019$

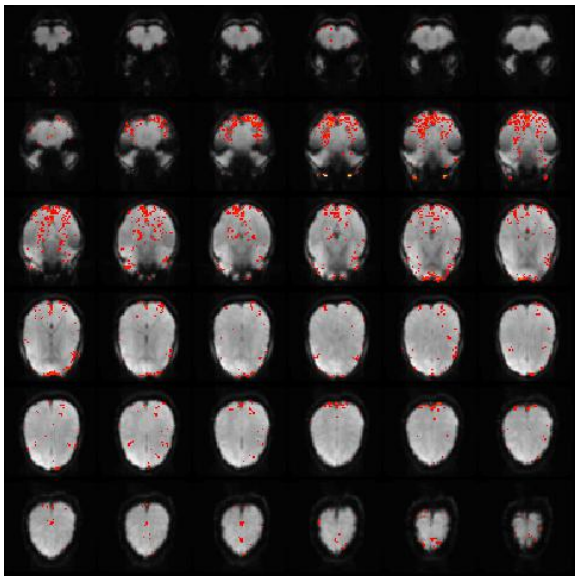


Object Stimulus

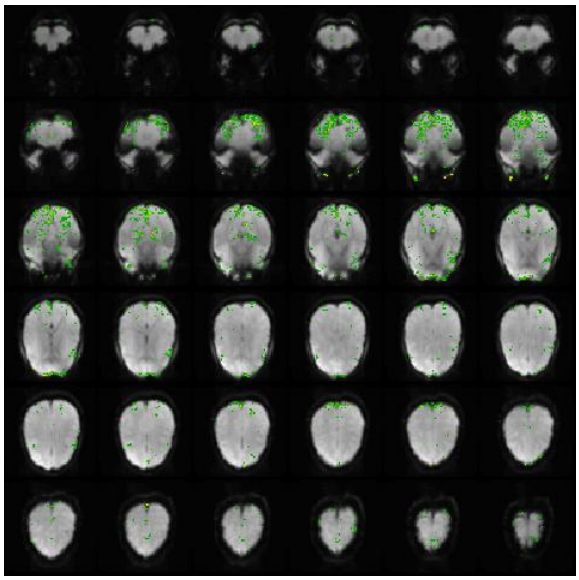


Scene Stimulus

Figure 3.31: Axial View $\lambda = 0.19$

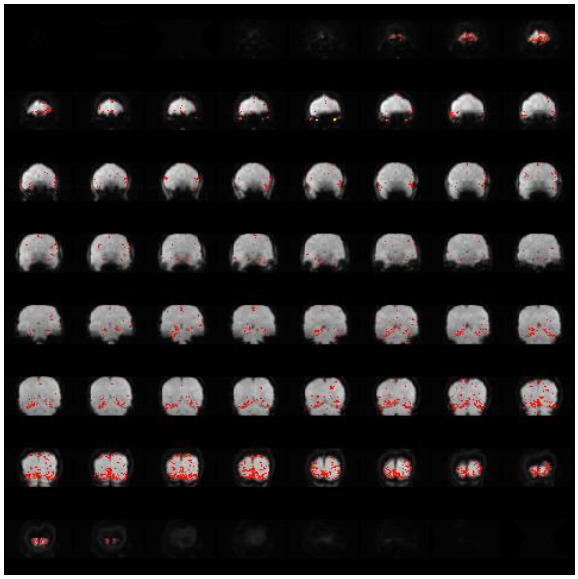


Object Stimulus

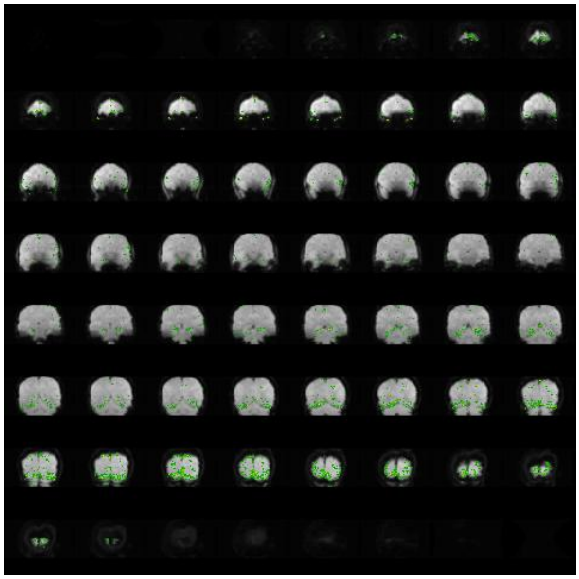


Scene Stimulus

Figure 3.32: Coronal View $\lambda = 0.19$

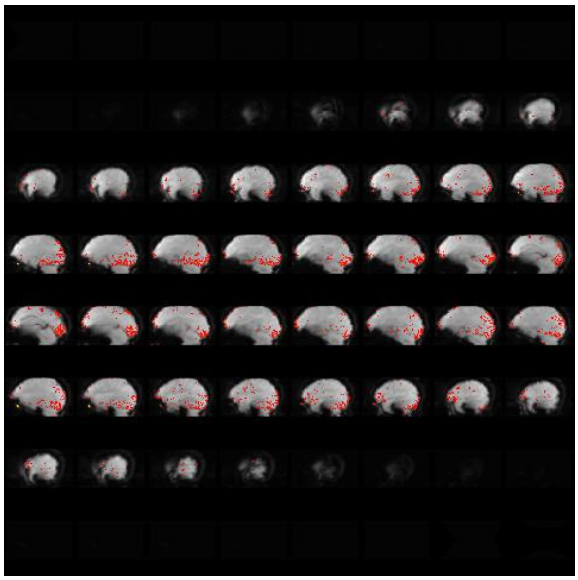


Object Stimulus

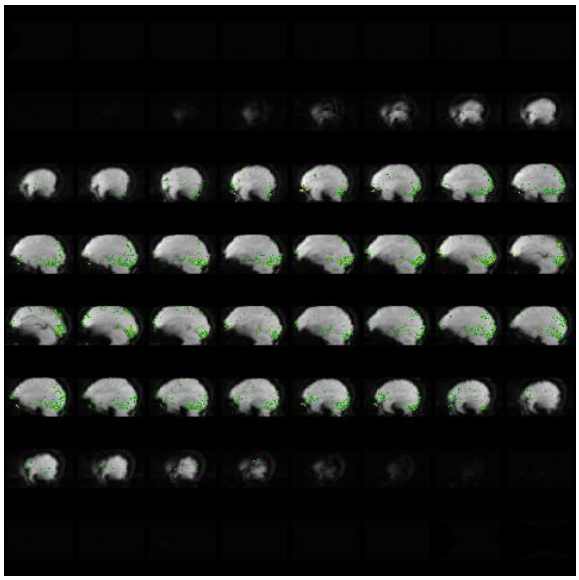


Scene Stimulus

Figure 3.33: Sagittal View $\lambda = 0.19$



Object Stimulus



Scene Stimulus

3.6 Discussion

In chapter 3 of the dissertation we empirically investigate a method proposed for selection of activated brain areas in a single subject fMRI study. As an application to variable selection and the study of regularization methods on densely dependent data, this is an area of study that requires keen exploration. Image analysis is structurally different from most data under statistical investigation. Medical imaging has vastly improved in the last decade and resources are being pooled to organize images taken as a regular diagnostic tool and improve our understanding of human anatomy.

With fMRI data in general, the massive acquisition is perhaps why statisticians refrain from applying bold methodologies and submit to crude simplistic methods. The formulation of coordinate descent like algorithms have indubitably eased the computational burden and it may be an opportunity to explore regularized methods in neuroimaging.

Under the setup of a known spatio-temporal covariance and properties of the Cholesky decomposition the method in chapter 3, sets up a variable selection routine with two penalty terms. The vectorized usually 4D structure of the data, i.e. 3D space and time series creates a matrix far too massive to computationally consider the Cholesky of order $O((n\mathcal{T})^3)$. But the separability assumption allows us to simplify the burden considerably by $O(n^3)+O(\mathcal{T}^3)$. This computation need not be repeated unless Cross Validation as described in section 3.3.2.4 as permutation matrices are not square and therefore these operations need to be repeated based on the number of folds.

A careful consideration must be made about however with regard to the implementation of coordinate descent methods for separable penalties as stated in Friedman et al. (2007). Although not addressed in J. Huang et al. (2011) and Li & Li (2010) if we look at the

the derivative of the object function 3.12 we see that the combination of penalties does not possess separability. Based on Tseng (1988) techniques that might require groups of coordinates to be updated together as in the case of the fused LASSO may yield better and more robust results, unless it can be shown otherwise.

A major aspect of the method that is currently overlooked is the estimation of this underlying spatio-temporal covariance and the separability assumption. In general there exists well developed methods for covariance estimation in the spatio-temporal paradigm using likelihood based techniques. There is scope in the future to therefore formalize a methodology. To address the latter issue of separability, we must first consider the feasibility in general. Separable assumptions simplify methods making them feasible (Ex: George & Aban (2015)) and reduce the number of parameters to be estimated dramatically. Although in the presence of non-separable relations among voxels misspecification may lead to inaccurate results, we could use ideas such as separable approximations for non-separable spatio-temporal models (Genton, 2007). This aspect is not within the scope of this dissertation and has significant future scope.

Tuning parameter selection is by far the most puzzling aspect of this setup. The observation that " *For* $\alpha \in (0, 1)$ *the method tends to be dominated by LASSO but may be an improvement over the simple lasso as it reduces any degeneracy and erratic behaviors caused by high correlation in the design*" is made in the paper by Friedman et al. (2010) about EN. We can definitely sense the LASSO domination. The coordinate descent solution for EN using Cross Validation is given by Kooij et al. (2007) in her thesis and is widely popular. Therefore in general due to these tendencies the free grid search maybe a suboptimal method and the path-wise coordinate descent method is preferable.

An interesting approach other than sequential methods or cross validation to obtain

optimal tuning parameters is estimating it. Park & Casella (2008) were able to do so in the Bayesian paradigm for the LASSO problem. Below is a conjecture following this notion but some obvious limitations exist.

3.6.1 Empirical Bayes Estimate using Bayesian LASSO

The proposed method is a generalized version that yields a special case in the form of the Elastic Net (EN). Similar to (Zou & Hastie, 2005, Section 2.2) let us consider reformulating the two penalty weighted least squares in 3.4 to a single tuning parameter LASSO objective function. Therefore given set (Y, X) and λ_1, λ_2 we define the reformulated $(\underline{Y}, \underline{X})$ by

$$\underline{Y}_{(n\mathfrak{T}+n)\times 1} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \ \underline{X}_{(n\mathfrak{T}+n)\times n} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X_{n\mathfrak{T}\times n} \\ \sqrt{\lambda_2}(I - MWM)^{1/2} \end{pmatrix} \text{ and } \underline{Q} = \underline{L}\underline{L}^T$$

where $\underline{L} = \begin{pmatrix} L & 0 \\ 0 & I \end{pmatrix}$ with L described in 3.3 such that for $\beta^* = \sqrt{1+\lambda_2}\beta$ and $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ the object function is denoted by $\mathfrak{S}^*$ and given by,

$$\mathfrak{S}^*(\gamma, \beta^*) \coloneqq (\underline{Y} - \underline{X}\beta^*)^T \underline{L}\underline{L}^T (\underline{Y} - \underline{X}\beta^*) + \gamma|\beta^*| \tag{3.15}$$

Thus $\hat{\beta}^* = argmin_{\beta^*}\mathfrak{S}^*(\gamma, \beta^*)$ subsequently $\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$.

Tibshirani (1996) interpreted the LASSO solution as the posterior mode estimates when regression estimates have an independent and identical Laplace priors. Park & Casella (2008) exploits the fact that the Laplace distribution can be represented as a scalar mixture of normals to obtain a conditionally inverse-Gaussian distribution for $1/\tau_j^2$ with parameters $\mu' = \sqrt{\frac{\gamma^2\sigma^2}{\beta_j^2}}$ and $\gamma' = \gamma^2$ where $\beta|\sigma^2, \tau_1, \ldots, \tau_n^2 \sim N_n(0, \sigma^2 D_t)$ and the variances τ^2 have a Laplace prior with hyper-parameter γ. Based on this hierarchical set up an empirical bayes

estimate was obtained of the form,

$$\gamma^{(k)} = \sqrt{\frac{2n}{\sum_{j=1}^{n} E_{\gamma^{(k-1)}}[\tau_j^2|y]}} \tag{3.16}$$

In the Bayesian paradigm a Gibbs sampler is used to evaluate the expectation in 3.16. Our primary objective is to find a suitable method that can be implemented to evaluate this tuning parameter γ in every iteration of the coordinate descent algorithm. We proceed by using the relation, $\hat{E}(\tau_j^2|y) = \sqrt{\frac{\beta_j^2}{\gamma^2}}$ for known $\sigma^2 = 1$ and $\gamma \neq 0$. A simulation study was attempted but an immediate issue was uncovered in the process. Unlike the Bayesian hierarchical setup that allows for the parameters of interest β's to have a normal prior, it is not possible to obtain 0 estimates. However with the soft-thresholding used the more number of 0 estimates leads to a higher $\gamma^{(k)}$ estimate which may further penalize the estimates in future iterations. So there is a tendency for the LASSO estimate to take the value infinity for all 0 β estimates before the estimates actually converge. Thus further investigation is necessary to find out whether such estimates lead to the selection of optimal tuning parameters.

In conclusion, the coordinate descent method for regularization has a very distinct computational benefit. However these strategies in machine learning and statistics are yet to be explored when it comes to dense dependent data. With a specific application in fMRI studies we are able to highlight a need for such methodologies and seek to theoretically justify the properties of the estimator obtained in the regularization method proposed in this chapter.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adler, R. J. (1981). The geometry of random fields. *John Wiley&Sons, Chichester*.

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.

Ashby, F. G. (2011). *Statistical analysis of fmri data*. MIT press.

Bachoc, F., & Furrer, R. (2016). On the smallest eigenvalues of covariance matrices of multivariate spatial processes. *Stat*.

Balan, R. M., Schiopu-Kratina, I., et al. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics*, *33*(2), 522–541.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

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

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician*, 179–195.

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., ... van der Vaart, A. (2006). Regularization in statistics. *Test*, *15*(2), 271–344. Retrieved from `http://dx.doi.org/10.1007/BF02607055` doi: 10.1007/BF02607055

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, *34*(4), 537–541.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9–25.

Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

Cox, R. W. (1996). Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162–173.

Cressie, N. A. (1993). Statistics for spatial data.

Dass, S. C., Lim, C. Y., & Maiti, T. (2012). Default bayesian analysis for multivariate generalized car models. *Statistica Sinica*, 231–248.

Dean, C., Ugarte, M., & Militino, A. (2004). Penalized quasi-likelihood with spatially correlated data. *Computational Statistics and Data Analysis*. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167947302003249`

Diggle, P., Ribeiro, P., & Geostatistics, M.-b. (2007). *Springer series in statistics*. Springer.

Diggle, P. J., Tawn, J., & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 299–350.

Donoho, D. L., Johnstone, I. M., et al. (1994). Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes rendus de l'Académie des sciences. Série I, Mathématique*, *319*(12), 1317–1322.

Emrich, L. J., & Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, *45*(4), 302–304.

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

Fan, J., & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*(467), 710–723.

Fan, Y., & 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.

Feng, W., Sarkar, A., Lim, C. Y., & Maiti, T. (2016). Variable selection for binary spatial regression: Penalized quasi-likelihood approach. *Biometrics*, *72*(4), 1164–1172.

Fingleton, B. (1986). Analyzing cross-classified data with inherent spatial dependence. *Geographical Analysis*, *18*(1), 48–61.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 309–317.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, *2*(4), 189–210.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, *18*(7), 681–695.

George, B., & Aban, I. (2015). Selecting a separable parametric spatiotemporal covariance structure for longitudinal imaging data. *Statistics in medicine*, *34*(1), 145–161.

Givnish, T. J. (1981). Serotiny, geography, and fire in the pine barrens of new jersey. *Evolution*, 101–123.

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

Gotway, C. A., & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 157–178.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., & Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, *72*, 304–321.

Guyon, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media.

Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices.

Heagerty, P. J., & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, *93*(443), 1099–1111.

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

Huang, H.-C., Hsu, N.-J., Theobald, D. M., & Breidt, F. J. (2010). Spatial lasso with applications to gis model selection. *Journal of Computational and Graphical Statistics*, *19*(4), 963–983.

Huang, J., Ma, S., Li, H., & Zhang, C.-H. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics*, *39*(4), 2021.

Hunter, D. R., & Li, R. (2005). Variable selection using mm algorithms. *Annals of statistics*, *33*(4), 1617.

Hurvich, C. M., & Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, 1077–1084.

Inouye, D. I., Yang, E., Allen, G. I., & Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, *9*(3).

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, *31*(1), 253–258.

Johnson, B. A., Lin, D., & Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, *103*(482), 672–680.

Kooij, A. J., et al. (2007). *Prediction accuracy and stability of regression with optimal scaling transformations*. Child & Family Studies and Data Theory (AGP-D), Department of Education and Child Studies, Faculty of Social and Behavioural Sciences, Leiden University.

Lee, K.-J., Jones, G. L., Caffo, B. S., & Bassett, S. S. (2014). Spatial bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis (Online)*, *9*(3), 699.

Li, C., & Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, *4*(3), 1498.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 13–22.

Lin, P.-S. (2008). Estimating equations for spatially correlated data in multi-dimensional space. *Biometrika*, 847–858.

Lin, P.-S. (2010). A working estimating equation for spatial count data. *Journal of Statistical Planning and Inference*, *140*(9), 2470–2477.

Lin, P.-S., & Clayton, M. K. (2005, 04). Analysis of binary spatial data by quasi-likelihood estimating equations. *Ann. Statist.*, *33*(2), 542–555. Retrieved from `http://dx.doi.org/10.1214/009053605000000057` doi: 10.1214/009053605000000057

Lindquist, M. A. (2008). The statistical analysis of fmri data. *Statistical Science*, 439–464.

Matheron, G. (1970). Random functions and their application in geology. In D. F. Merriam (Ed.), *Geostatistics: A colloquium* (pp. 79–87). Boston, MA: Springer US. Retrieved from `http://dx.doi.org/10.1007/978-1-4615-7103-2_7` doi: 10.1007/978-1-4615-7103-2_7

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC Press.

McShane, L. M., Albert, P. S., & Palmatier, M. A. (1997). A latent process regression model for spatially correlated count data. *Biometrics*, 698–706.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53–71.

Musgrove, D. R., Hughes, J., & Eberly, L. E. (2016). Fast, fully bayesian spatiotemporal inference for fmri data. *Biostatistics*, *17*(2), 291–303.

Ortega, J., & Rheinboldt, W. (1970). Iterative solution of nonlinear equations in several variables (academic, new york, 1970).

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125.

Pan, W. (2002). Goodness-of-fit tests for gee with correlated binary data. *Scandinavian Journal of Statistics*, *29*(1), 101–110.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p2/n is large; ii. normal approximation. *The Annals of Statistics*, 1403–1417.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, *16*(1), 356–366.

Ries, L. A., Harkins, D., Krapcho, M., Mariotto, A., Miller, B. A., Feuer, E. J., ... others (2006). Seer cancer statistics review, 1975-2003.

Schacke, K. (2004). On the kronecker product. *Master's thesis, University of Waterloo*.

Schaetzl, R. J. (1986). Soilscape analysis of contrasting glacial terrains in wisconsin. *Annals of the Association of American Geographers*, *76*(3), 414–425.

Smith, M., & Fahrmeir, L. (2007). Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, *102*(478), 417–431.

Strawderman, R. L., & Tsiatis, A. A. (1996). On consistency in parameter spaces of expanding dimension: an application of the inverse function theorem. *Statistica Sinica*, 917–923.

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

Tseng, P. (1988). Technical report lids-p, 1840. *Massachusetts Institute of Technology, Laboratory for Information and Decision Systems*.

Wang, H., Li, G., & Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(1), 63–78.

Wang, L. (2011). Gee analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, *39*(1), 389–417.

Wang, L., & Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(1), 177–190.

Wang, L., Zhou, J., & Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, *68*(2), 353–360.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, *61*(3), 439–447.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, *4*(1), 389–396.

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918.

Xie, M., Yang, Y., et al. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, *31*(1), 310–347.

Xue, L., Zou, H., & Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 1403–1429.

Yahav, I., & Shmueli, G. (2012). On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, *28*(1), 91–102.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, *75*(4), 621–629.

Zhang, C.-H., et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894–942.

Zhu, J., Huang, H.-C., & Reyes, P. E. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 389–402.

Zhu, X. (2011). Semi-supervised learning. In *Encyclopedia of machine learning* (pp. 892–897). Springer.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & 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.