HIGH-DIMENSIONAL VARIABLE SELECTION FOR SPATIAL REGRESSION AND COVARIANCE ESTIMATION

By

Siddhartha Nandy

A DISSERTATION

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

Statistics – Doctor of Philosophy

2016

ABSTRACT

HIGH-DIMENSIONAL VARIABLE SELECTION FOR SPATIAL REGRESSION AND COVARIANCE ESTIMATION

By

Siddhartha Nandy

Spatial regression is an important predictive tool in many scientific applications and an additive model provides a flexible regression relationship between predictors and a response variable. Such a model is proved to be effective in regression based prediction. In this article, we develop a regularized variable selection technique for building a spatial additive model. We find that the approaches developed for independent data do not work well for spatially dependent data. This motivates us to propose a spatially weighted ℓ_2- error norm with a group LASSO type penalty to select additive components for spatial additive models. We establish the selection consistency of the proposed approach where a penalty parameter depends on several factors, such as the order of approximation of additive components, characteristics of the spatial weight and spatial dependence, etc. An extensive simulation study provides a vivid picture of the impacts of dependent data structures and choices of a spatial weight on selection results as well as the asymptotic behavior of the estimates. We also investigate the impact of correlated predictor variables. As an illustrative example, the proposed approach is applied to lung cancer mortality data over the period of 2000-2005, obtained from Surveillance, Epidemiology, and End Results Program by the National Cancer Institute, U.S.

Providing a **b**est **l**inear **u**nbiased **p**redictor (BLUP) is always a challenge for a non-repetitive, irregularly spaced, spatial data. The estimation process as well as prediction involves inverting an $n \times n$ covariance matrix, which computationally requires $O(n^3)$. Studies showed the potential observed process covariance matrix can be decomposed into two additive matrix components, measurement error and an underlying process which can be non-stationary. The non-stationary component is often assumed to be fixed but low rank.

This assumption allows us to write the underlying process as a linear combination of fixed numbers of spatial random effects, known as **f**ixed **r**ank **k**riging (FRK). The benefit of smaller rank has been used to improve the computation time as $O(nr^2)$, where r is the rank of the low rank covariance matrix. In this work we generalize FRK, by rewriting the underlying process as a linear combination of n random effects, although only a few among these are actually responsible to quantify the covariance structure. Further, FRK considers the covariance matrix of the random effect can be represented as product of $r \times r$ cholesky decomposition. The generalization leads us to a $n \times n$ cholesky decomposition and use a group-wise penalized likelihood where each row of the lower triangular matrix is penalized. More precisely, we present a two-step approach using group LASSO type shrinkage estimation technique for estimating the rank of the covariance matrix and finally the matrix itself. We investigate our findings over a set of simulation study and finally apply to a rainfall data obtained on Colorado, US.

ACKNOWLEDGMENTS

I would like to take this opportunity, to extent my appreciation towards both my advisors, Prof. Chae-Young Lim and Prof. Tapabrata Maiti for their guidance, encouragement and challenge me to thrive in whatever way I was successful. I would like to specifically thank Prof. Lim for her time in reading the draft and suggesting my changes all through-out. I would also like to thank my other committee members who gave me some positive feedbacks and suggestions regarding future work. I would also like to thank my friends and family for the support they gave me during my times of need.

PREFACE

The work in this thesis can be summed up into a general problem of selecting fixed and random effects component in a mixed-effect prediction model, with a Gaussian error, where the predicting variable is observed only on a spatial location grids. These grids could be irregular as well. The spatial locations site are in a two dimensional (2D) or three dimensional (3D) space in our applications. Although, the final goal is to successfully detect both fixed and random effects component simultaneously, the first step toward solving this problem by achieving each separately is quite challenging. It still requires specific attention compared to some existing methods, which can be looked as special situation of our method.

This spatial data structures appears in abundance, if one is interested in modeling climate parameters or interested in understanding brain connectivity using functional magnetic resonance imaging signal. Climate parameters are supposed to vary over both latitude and longitude, although often altitudes are also considered. Hence for climate problems deal with either 2D or 3D in space. On the other hand, brain acitivity varies over points on 3D surface. All our efforts are towards certain challenges in reproducing a feasible parsimonious model. An obvious step is to decompose the response variable into two components, one mean function, and the other spatially dependent Gaussian error.

The mean function has a general additive model structure, and the number of covariates grows as fast as exponential of the number of sites in the study. The dependence structure of the Gaussian error can either be stationary, or non-stationary and anisotropy. The case of stationary covariance allows the number of parameter required to model can be controlled and we use covariance functions viz. Exponential (Exp), Matérn (Mat), Inverse Multi - quadratic (Inv MQ), and Gaussian (Gauss). The case of non-stationary and anisotropy covariance requires a different approach as now each and every location requires multiple parameters and this case can often be translated to a random effects model.

So one of our chapters consider stationary covariance function and hence it talks about

selecting the most appropriate set of fixed effects components and that only. While the other chapter with a non-stationary covariance talks about selection from mixed effects model but, we keep our focus of selection only to the random effects components. Once we are able to combine the two methods of selection we will be simultaneously estimating gerenal additive mean function and non-stationary covariance function of Gaussian spatial process defined on a spatial surface.

Since, the numbers of covariates in the general additive mean function, grows exponentially with sample size, high - dimensional variable selection has its own challenges (Huang, *et.al.* (2010)). Additionally, we have a spatially dependent Gaussian error model which requires special attention. This thesis holds proof of our effort to extend Huang's work to our setup of spatial dependent error model. We start by considering a stationary covariance structure for spatial dependent error model and try to achieve a parsimonious mean model. A more elaborated picture is depicted through chapter 1.

A relevant conclusion of our work at this point is, one can bypass the problem of estimating the stationary covariance function if the sole interest is in selecting the true positive components, as long as we use a weighted least squares by inverse of certain stationary covariance matrix with both short and long range dependence. It also holds documentation about the fact that even identity matrix as a choice of the weight matrix for least squares is an improvement over Huang's work as per under - selection of false positive components in the additive model. This is possible since the penalty parameter is still higher compared to Huang's choice.

Suppose the problem of interest is now not in reducing the numbers of covariates anymore rather, it is in estimating the covariance function. Also, a more general solution to the problem of estimating the covariance matrix is obviously for non-stationary and anisotropic covariance. This generalization can be overcome by representing the Gaussian error as sum of two independent Gaussian processes, one is a linear combination of spatial random effects vector, and the other is measurement error with finite variance. The spatial random effects

vector is assumed to have a non-stationary covariance dependence but of not full rank. The factors involved in computing the linear combination are bi-variate spline functions.

This above decomposition of the overall Gaussian error in to two independent Gaussian process, allows the overall model error to have a Gaussian process with a very special covariance structure. The covariance matrix of the overall model error is a full rank matrix, ideally the computation time of inverting it should be of the order of cubic power of dimension of the square matrix. But the above decomposition allows us to use **S**herman-**M**orisson-**W**oddbury (SMW) identity on inverse of matrices. Along with SMW identity we exploit the fact that the random effects vector has a low rank covariance structure. In 2008 Cressie *et.al*. exploited this property and considered if the true value of the low rank non-stationary component is known, then one can invert the covariance matrix with a controlled computation time.

As it turns out that the spatial random effects component, which is responsible of the non-stationary component in covariance matrix, can be represented as a linear combination of several spatial basis vectors. The weights corresponding to a particular spatial location depends on the distance between the spatial location and some basis knot locations. The knots play the role of uniformity in many senses. The knowledge of necessary and sufficient of knot locations translates to the simpler version of Cressie's work in 2008. This technique has also been succesfully implimented by Nychka *et.al* 2015 in their work on multi-resolution kriging.

Our contribution to this direction is a data driven approach to estimate the number and positions of these knot locations. This problem can be rewritten as selection problem of spatial random effects components. A more detailed scrutiny convinced us to penalize the Gaussian likelihood by a group wise LASSO penalty to overcome the curse of dimensionality. Although, there has been quite extensive research in using group LASSO or ℓ_1/ℓ_2- penalty in reducing factors that are used to model mean, this problem is quite unique and requires special attention for several reasons.

The rest of the thesis is organized as follows. Chapter 1 discusses a mathematical

introduction of the high dimensional models used in this work. It also throws light on why this problem of dimension reduction should be addressed separately. It breaks down the overall problem of reducing dimension of both mean and covariance function in few steps and finally discuss how to combine the picture. Chapter 2 discusses the result on dimension reduction of mean function while the covariance funcion is hold fixed at a low dimension but unkown. It introduces the algorithm for estimating the rank of the non-stationary covariance matrix. It also provides some theoritical findings justifying the consistency of the parameters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

For many statistical problems arising in spatially observed data, it is straight forward to consider the underlying process $Y = \{Y(s); s \in S\}$, is a spatially varying process. It is also often assumed to have a mean function which is based on the available set of covariates. Let us denote the mean function by $m(s)$. If we center the underlying process $Y(s)$ with $m(s)$ the residual, $\epsilon(s)$ is often broadly assumed to have Gaussian error structure or in some more general notion of having a Gaussian tail structure. We will discuss the dependence structure of this spatial process elaborately in this work. Therefore the following model could be the basic of this research.

$$Y(s) = m(s) + \epsilon(s). \tag{1.0.1}$$

1.1 Additive model building

Consider that the mean function $m(s)$ is modeled using J covariates denoted as, $\{X(s) = (X_1(s), \cdots, X_J(s))\}$, and each of these variables are spatially varying processes too. In a simpler version a linear relation is assumed between each of these $X_j(s)$'s and $Y(s)$. This work will be generalizing the linear relation to a broaer scenario of any possible unknown functions, which is refered as general additive mean function. This is a non-parametric version of the linearity assumption. The non-parametric nature of functional regression makes the mean prediction complicated to start with which requires special attention and we will following some pre-proved standards from the literature for this complicacy. So, the mean function $m(s)$ can be represented as the following additive model,

$$Y(s) = \mu + \sum_{j=1}^{J} f_j(X_j(s)) + \epsilon(s), \tag{1.1.1}$$

with, μ being the overall mean. Now if we want to put some light on the spatially dependence structure of $\epsilon(s)$, we differentiate the problem in to two cases, either with stationary

covariance structure or, with non-stationary and anisotropic covariance structure. Although the final goal is to get a joint estimation of mean and covariance function, it is not a straightforward extension of combining mean and covariance functions estimation. So we will take one step at a time. The technique used for selecting additive components is shown to be generalizing over its independent Gaussian error counterpart.

We also consider a situation, where the number of covariates, J is increasing in n and can even grow exponentially with the number of subjects used to estimate the model. In a spatial study the number of subject, is the number of location sites where we have observed the process $Y(s)$. The dimension of the covariance matrix, is increasing with the number of location considered under study. This is a serious issue, but first goal is to identify those variables which are relevant in making a prediction of the response variable. We should keep in mind that here $J > n$, hence the standard variable selection techniques fall apart and, so we too shall explore some recent advancement in penalized optimizations. We will also point out the challenges in deciding the optimal penalty parameter specific to our problem.

The concept of penalized optimization have two major components, one is the statistical likelihood which indicates the distribution of the error process or the nature of error norm. We are interested in minimizing the likelihood subject to a constraint. The second component is the constraint and also called the penalty component. The penalty component is driven and motivated from the fact that number of parameters in the model is larger than the number of spatial locations observed for the study. We want to penalize our original likelihood and shrink some of the component in our additive model to zero. Over the last decade research on penalized optimization flourished due to abundance of high dimensional problem in various fields of applied science.

The Gaussian error $\epsilon = (\epsilon(s); s \in S)$ in the model, in equation (1.1.1) has a spatial dependence. For future notational purposes let us say $\epsilon \sim N(\mathbf{0}, \Sigma)$. From the perspective of estimating mean or selecting variables in a prediction model, we can use the following

penalized optimization function,

$$Q(f_1,\ldots,f_J,\lambda_n) = \left(Y - \sum_{j=1}^{J} f_j\left(X_j\right)\right)' \Sigma^{-1} \left(Y - \sum_{j=1}^{J} f_j\left(X_j\right)\right) + \lambda_n \sum_{j=1}^{J} p(f_j) \quad (1.1.2)$$

We know the technique to control these general additive models are to approximate each of these functions f_j using spline representation. The correspoding structure gives rise to a parametric formulation that allows us to use a very celebrated technique called group **L**east **A**bsolute **S**hrinkage and **S**election **O**perator(LASSO). We will skip and leave detailed discusssion on penalized optimization for rest of the thesis. Joint estimation of mean and the dependence structure or even considering a spatially dependent error, but avoid estimating the covariance function while using techniques like group LASSO complicates the problem in various spectrum.

Equation (1.1.2) gives a flavor of penalized optimization of weighted ℓ_2- norm, where the weights are proportional to the inverse of the dependence structure of the Gaussian error process. The challenges of estimating Σ while detecting important additive components are both theoritical and algorithmic. So we will keep the joint estimation out of the picture for a while and deal each problem separetly. There are two alternative ways to model this dependence structure, one where we assume that dependence between two location site solely depends on the distance between the two sites. The covariance function in this case is a parametric function of the distance this technique of modeling the covariance function in the literature has be refered as stationary covariances.

Another alternative and a more general model is when we can relax the assumption of dependence through distance between location sites by introducing the class of non-stationary and anistropic covariance functions. The next chapter, chapter 2 is on selecting the necessary non-zero components out of all the J covariates under study with out going into the complications of estimating Σ. To start with we shall restrict ourseleves to the class of stationary covariance functions, *i.e.* $\Sigma = \left(\left(\sigma_t(\left|s-s'\right|; s, s' \in S)\right)\right)$ and $\left|s-s'\right|$ is the measure of distance between s and s'. On the other hand chapter 3 considers the a general non-stationary

covariance function.

Although Σ is unknown, in our next chapter the covariance function refered above as $\sigma_t(\cdot)$, belongs to either of a class of parametric covariance functions viz. Exponential, Gaussian, Inverse Multi-quadratics or Matérn. We succesfully defended the idea that selection of components among all the covariates in our additive model is robust in the choice of the precise form of the true covariance function. We made significant amount of research about how the penalty parameter of our spatially dependent model should differ as compared to its independent counterpart.

Instead of using Σ, the true covariance matrix in (1.1.2) we use a different spatial weight matrix $W = \left(\left(\sigma_w(\left|s - s'\right|; s, s' \in S)\right)\right)$, which an user can choose based on an exploratory analysis of data. Henceforth we optimized,

$$Q(f_1, \ldots, f_J, \lambda_n) = \left(Y - \sum_{j=1}^{J} f_j\left(X_j\right)\right)' W^{-1} \left(Y - \sum_{j=1}^{J} f_j\left(X_j\right)\right) + \lambda_n \sum_{j=1}^{J} p(f_j) \quad (1.1.3)$$

1.2 Estimating non-stationary covariance

As mentioned our next goal is to explore how to estimate the non-stationary covariance function. The complication is two fold. First, joint estimation of both mean and covariance for a Gaussian process works better if the overall likelihood is optimized rather minimizing just the prediction error like (1.1.2). Second, the computation time of the likelihoood or the gradient of the optimization function for a non-stationary matrix is of cubic order of the dimension of the square matrix. To east out let us reduce the burden of selection of additive components. Chapter 3 introduces the technique of using spatial random effects models and multi-resolution knots to model the non-stationary covariance matrix overcoming both complications.

Spatial random effects modeling allows us to incorporate another indepdent additive Gaussian component $\pi(s)$ along with the stationary component $\epsilon(s)$. For more simplicity

we consider $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbb{I})$ *i.e.*, the stationary covariance function takes non-zero value only when distance is zero. The objective of using spatial random effects model is to capture the covariance between the response variable Y at two different location s and s' through,

$$cov\left(Y(s), Y(s')\right) = \begin{cases} R(s)'_{1\times r}\Omega_{r\times r}R(s')_{r\times 1} & \text{if } s \neq s' \\ R(s)'_{1\times r}\Omega_{r\times r}R(s')_{r\times 1} + \sigma^2 & \text{if } s = s'. \end{cases} \tag{1.2.1}$$

In spatial literature the parameter σ is called nugget effect. Note the above representation requires a valid quadratic form multiplication, where the length of the vector $R(s)'$ and the dimension of the positive definite matrix Ω should coincide and is denoted by r. The magnitude of r plays an important role in reducing computation time of inverse of the overall covariance matrix $\sigma^2\mathbb{I} + R\Omega R'$. The choice of r has been always a challenge. To understand the nature of the parameter r, an alternative representation of (1.0.1) as follows,

$$\begin{aligned} Y(s) &= m(s) + \pi(s) + \epsilon(s) \\ &= m(s) + R(s)\alpha + \epsilon(s), \end{aligned} \tag{1.2.2}$$

where $\alpha \sim N_r(\mathbf{0}, \Omega)$ is assumed, and used in the last chapter. A magnified look of the representation, $\pi(s) = R(s)'\alpha$, gives

$$\pi(s) = \sum_{k=1}^{r} R_k(s)\alpha_k \tag{1.2.3}$$

where $R(s) = (R_1(s), R_2(s), \ldots, R_r(s))'$ weights for $r-$ components of the spatial random effects. The varibility of the random effect α, Ω is a positive definite matrix. So we can use a cholesky representation $\Omega = \Phi\Phi'$. Before we go into the details of how we are proposing to estimate both parameters r and Φ let me remind the nature of $m(s)$ is relaxed from a general additive models and use just a linear model. We even disregard the need to selecting the variables assuming $J < n$. To start with a simpler model one can even choose $m(s) \equiv 0$.

We introduce an alternative spatial random effects model and rewrite equation (1.2.3) as,

$$\pi(s) = R(s)\alpha \asymp \widetilde{R}(s)\widetilde{\alpha} = \sum_{m=1}^{M}\sum_{j=1}^{\ell_m} \widetilde{R}_{j(m)}(s)'\widetilde{\alpha}_{j(m)} \tag{1.2.4}$$

where $\widetilde{R}(s) = (\widetilde{R}_{1(1)}(s), \ldots, \widetilde{R}_{\ell_1(1)}(s); \ldots; \widetilde{R}_{1(M)}(s), \ldots, \widetilde{R}_{\ell_M(M)}(s))'$. There are M resolutions, and m^th-resolution has $\ell(m)$ number of knots such that $\ell_1 + \ldots + \ell_M = L$ (say). The technique of using multiple resolution is defended by several in the field (Cressie et.al. (2008)) and is named as multi-resolution kriging by Nychka et.al. (2015). We keep aside the details on how to choose the number of resolution, and number of knots per resolution optimal for the study under consideration for chapter 3.

It is assumed that, $\widetilde{\alpha} = \left(\widetilde{\alpha}_{1(1)}, \widetilde{\alpha}_{2(1)}, \ldots, \widetilde{\alpha}_{\ell_M(M)}\right) \sim N_L\left(\mathbf{0}, \widetilde{\Omega}\right)$, with $\widetilde{\Omega}$ being a $L \times L$ positive semi-definite matrix with r non-zero eigen values. This also gives us the freedom to write $\widetilde{\Omega} = \widetilde{\Phi}\widetilde{\Phi}'$ where $\widetilde{\Phi}$ is a $L \times L$ lower triangular matrix with rank r. To summarize chapter 3, we propose a technique to estimate which r among these L knots are effective to estimate the structure of the non-stationary covariance structure. If we note the cholesky structure

$$\widetilde{\Phi} = \begin{bmatrix} \widetilde{\varphi}_{11} & 0 & \cdots & 0 \\ \widetilde{\varphi}_{21} & \widetilde{\varphi}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\varphi}_{L1} & \widetilde{\varphi}_{L2} & \cdots & \widetilde{\varphi}_{LL} \end{bmatrix} = \begin{bmatrix} \widetilde{\varphi}_{(1)} \\ \widetilde{\varphi}_{(2)} \\ \vdots \\ \widetilde{\varphi}_{(L)} \end{bmatrix},$$

we can infer that variance of j^{th} component in the random effect corresponding to the j^{th} row of the cholesky matrix. We propose the idea of selecting a component out of the random vector $\widetilde{\alpha}$ with non-zero variances by estimating the $\widetilde{\Phi}$ based on which of the rows of the estimated matrix, $\widehat{\widetilde{\Phi}}$. Since each row corresponds to one component we propose a group wise penalized likelihood maximization technique by maximizing the following negative log-likelihood,

$$\begin{aligned} Q_n(\widetilde{\Phi}, \sigma^2, \tau_n, \psi) = n\,\mathbf{T}\mathrm{r}\left(\Xi_0\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right)^{-1}\right) &+ \log\det\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right) \\ &+ \tau_n \left\|\widetilde{\Phi}^{vec}_{Fullset}\right\|_{2,1,\psi}, \end{aligned} \tag{1.2.5}$$

where $\Xi_0 = XX'/n$ is the empirical covariance matrix.

Rest of the thesis is organized as follows. The next chapter under the heading. 'Additive Model Building for Spatial Regression' discusses the issues and challenges of selecting variables in a general additive model in a high dimensional scenario. It contains details theoritical and extensive simulations studies on different situations. Following chapter under the heading, 'Estimating Non-stationary Spatial Covariance Matrix using Multi-resolution Knots' throws light on the problem of estimating a non-stationary covariance matrix. It discusses a very well posed problem of numerical efficiency while using likelihood based estimation of a large covariance matrix. This deals with finding inverse or determinant of large covariance matrices.

CHAPTER 2

ADDITIVE MODEL BUILDING FOR SPATIAL REGRESSION

It is important yet fairly challenging to identify important factors that explain certain phenomena such as climate change, economic volatility, ecological dynamics and disease progresses, etc. In such applications, spatially dependent data are often observed and spatial regression is a natural tool for data analysis. An additive model provides a flexible regression relationship and is proven to be effective for regression based prediction. The models developed for independent data are often statistically inefficient in this context. Thus, the statistical models dealing with spatially dependent data have received considerable attention over the last few decades.

A common feature for spatial data is spatial dependence among sampling sites. Generally, we assume that the dependence between two data points at two sampling sites decreases as the distance between two sites increases. At each sampling location, we observe a quantity of interest (response variable) and additional information (covariates or predictors) that could affect the response. A natural approach to identify contributing factors that influence response variable is selection of covariates in a regression set up, commonly known as the variable selection. Among many variable selection techniques, the regularization technique (e.g., Tibshirani, 1996) received huge attention in recent years.

The current literature on variable selection concentrates heavily on regression models for independent observations. The methods without any adaptation are not expected to work well for spatially dependent data. Further, theoretical justification for spatial data requires special attention and so true for the variable selection. There are studies on variable selection for time series data. Typical time series data can be viewed as a special case of spatial lattice data by reducing the dimension of an observation domain to one and assuming the data are observed evenly over time. Wang *et* al. (2007) studied selection of regression coefficients and autoregressive order via LASSO for regression models with autoregressive

errors. Nardi and Rinaldo (2011) considered LASSO for autoregressive process modeling. Hsu *et* al. (2008) applied LASSO to select the subset for vector autoregressive processes. Xu *et* al. (2012) studied variable selection for autoregressive models with infinite variance. While these approaches deal with a specific dependence structure (autoregressive structure), Gupta (2012) investigated variable selection for weakly dependent time series data under a linear regression model.

Variable selection or model selection for spatial data is relatively new. Hoeting *et* al. (2006) derived Akaike's Information Criterion (AIC) for a geostatistical model and used it for selecting explanatory variables under spatial correlation. Huang and Chen (2007) introduced model selection criterion with generalized degrees of freedom for selecting a spatial prediction model. Huang *et* al. (2010a) considered a spatial LASSO for selecting covariates and spatial neighborhoods with a known spatial dependence structure but no theoretical investigation was made. Wang and Zhu (2009) considered penalized least squares for geostatistical data with various penalty functions and investigated their theoretical properties.

In likelihood based approaches, Zhu *et* al. (2010) considered selection of spatial linear models together with a spatial neighborhood structure for spatial lattice data using a penalized maximum likelihood method with an adaptive LASSO penalty. Chu *et* al. (2011) investigated variable selection for spatial linear models for geostatistical data using a penalized maximum likelihood method. They considered an approximated penalized maximum likelihood approach with a tapered spatial covariance function. Reyes *et* al. (2012) extended the approach by Zhu *et* al. (2010) for spatial-temporal lattice models. For spatial binary data, Fu *et* al. (2013) considered selection in autologistic regression models using a penalized pseudolikelihood.

For flexible relationship between the response and the regressor variables, we consider an additive model with spatially dependent error. Consider $\{Y(s); s \in \mathbb{R}^d\}$ be a spatial process on $\mathbb{R}^d$ and $\{X(s) = (X_1(s), \cdots, X_J(s)); s \in \mathbb{R}^d\}$ be a J-dimensional vector which can be stochastic, if necessary. We consider a spatial additive model given in (2.1.1) with

the overall mean μ and f_j being unknown functions describing the relation between Y and X_j. We assume the error process is a mean zero stationary Gaussian random field with covariance function, $\delta(h)$. J could be larger than the sample size. Our objective is to select the 'effective' f_j's.

For selection and estimation of nonlinear components, f_j, Huang *e*t al. (2010b) proposed adaptive group LASSO for the additive model (2.1.1) *b*ut with independent errors. Their work is based on spline approximation of non-linear components which led to rewrite the mean of (2.1.1) as a linear regression model with spline coefficients. Hence the problem of selecting a component in an additive model is transformed into selecting groups of variables in a linear regression with predefined group members. Meier *e*t al. (2009) also considered variable selection for high dimensional additive models using spline approximation but with a sparsity-smoothness penalty, which controls both sparsity as well as smoothness in spline approximation. Several other works on the selection of additive models or additive nonparametric regression comprise of Antoniadis and Fan (2001), Lin and Zhang (2006), Ravikumar *e*t al. (2009) and Lai *e*t al.(2012), etc. These works assumed independent error distributions. For geo-additive regression models, Kneib *e*t al. (2009) considered a variable selection method using a penalized spline approach for breeding bird communities data. However, no theoretical justification was discussed. To the best of our knowledge, there is no work on spatial additive model selection with theoretical justification.

First, we empirically examined a group LASSO approach developed for independent data to select nonzero components in an additive model when the errors are spatially dependent. For the additive model (2.1.1), we considered $J = 10$ with two true nonzero components, $f_1(x) = \sin(x)$ and $f_2(x) = x$. We considered $m \times m$ unit square lattices with $m = 6, 12, 24$. For the spatial errors, we used Gaussian distribution with an exponential covariance function, $\delta(h) = \exp(-\rho|h|)$ and $\rho = 0.5$. We generated 400 datasets. Since the spatial dependence can also be captured by a function of the location in the mean, we also investigated two different intercepts in addition to the additive component model (2.1.1): one is a constant

m	Constant		A function of the location	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
6	5.64	1.74	3.61	1.23
12	6.61	1.60	4.31	1.62
24	7.22	1.38	5.91	2.01

Table 2.1 Average and Standard deviation for the number of selected covariates using 400 datasets from the exponential covariance function, $\delta(h) = exp(-\rho|h|)$ with $\rho = 0.5$. The true number of nonzero components is 2. $m \times m$ unit square lattices are considered.

and the other is a nonparametric function of the location using spline approximation. For a nonparametric function of location $s = (s_1, s_2)$, we considered an additive structure in terms of s_1 and s_2. Table 2.1 shows the average and standard deviation of the number of selected components when the independence approach is used. The regularization parameter (or penalty parameter) is chosen as recommended by Huang *e*t al. (2010b). The number of selected covariates is much larger than the true number of nonzero components for both cases at various sample sizes. Although a function of location as a mean component helps improve the selection results in this simulation study, it is not satisfactory. The performance did not improve even with a larger sample size. This leads us to investigate a variable selection method suitable for spatial additive models. The comprehensive simulation study is given in Section 2.3.

For a spatial additive model, we maintain the mean structure of the independent data group LASSO approach, that is, we use the idea of approximating f_j by a linear combination of spline functions and a group LASSO penalty for sparsity. Then, we introduce a spatial weight matrix in the objective function. The choice of weight matrix is motivated from the concept of weighted least squares. The dependence in the random error is compensated by the spatial weight matrix. Our assymptotic result supports well-conditioned weight matrix. At this point we would like to point out our readers that, identity matrix is a perfect choice of well-conditioned matrix. Although, it should be noted that choice of identity matrix as weight in the penalized weighted least squares, for our findings is not same as, case of using

penalizing ordinary least squares as in case of Huang *e*t al. (2010b).

We develop asymptotic theory for selection consistency of nonzero components in the additive model under spatially dependent Gaussian errors. The spatial dependence structures that we assumed here are common in modeling spatial data and valid for a wide range of applications. Variable selection is often sensitive to the choice of a penalty parameter. We found out that the theoretical lower bound for the penalty parameter depends on spatial dependence as well as a spatial weight matrix. The case of identity matrix as weight in penalized weighted least squares gives rise to a different form of the lower bound for penalty parameter. We demonstrate a method for selecting a penalty parameter and a spatial weight matrix guided by our theory and the existing practice in variable selection.

The rest of this chapter is organized as follows. Section 2.1 describes the proposed approach for selecting and estimating nonzero components in an additive model with spatially dependent errors. Section 2.2 discusses the main theoretical results for asymptotic properties of our proposed estimators. Section 2.3 contains simulation results along with a real data example for illustration. Finally, we make some concluding remarks in section 2.4. Proofs of all theorems are provided in the appendix. Proofs of related lemmas, extension of theoretical results and additional simulation results are given in the supplementary material. All numerical study was performed by the code written in the statistical software **R**. The example code is available at `stt.msu.edu/users/maiti/PublicationFiles/simulation_code.pdf`.

2.1 Method for selecting components in spatial additive models

We consider the following spatial additive model:

$$Y(s) = \mu + \sum_{j=1}^{J} f_j(X_j(s)) + \epsilon(s), \qquad \forall s \in \mathbb{R}^d \tag{2.1.1}$$

where μ is the overall mean, f_j is an unknown function describing the relation between Y and X_j, and $\{\epsilon(s); s \in \mathbb{R}^d\}$ is a zero mean stationary Gaussian random field with a covariance

function, $\delta(h)$. J could be larger than the sample size. Suppose that $(Y(s), X(s))$ are observed at n different locations lying in sampling region $D_n \subset \mathbb{R}^d$. Let S be the set of sampling locations. We use a spline approximation of f_j in the additive models.

$$f_j(X_j(s)) \approx f_{nj}(X_j(s)) := \sum_{l=1}^{m_n} \beta_{jl} \mathbb{B}_l(X_j(s)), \quad \text{for } j = 1, \ldots, J, \tag{2.1.2}$$

where $\mathbb{B}_l(\cdot)$s are normalized B-spline bases, and β_{jl}s are called control points [Schumaker (2007)] . The approximation (2.1.2) is based on the theory which states that every smooth function can be uniquely represented by a linear combination of B-splines. Then, the model (2.1.1) is approximated as

$$Y(s) = \mu + \sum_{j=1}^{J} \sum_{l=1}^{m_n} \beta_{jl} \mathbb{B}_l(X_j(s)) + \pi(s), \quad \text{for } s \in S, \tag{2.1.3}$$

where $\pi(s) = \epsilon(s) + \theta(s)$ with $\theta(s) = \sum_{j=1}^{J} (f_j(X_j(s)) - f_{nj}(X_j(s)))$ and also define $\theta = (\theta(s); s \in S)'$. The model (2.1.3) can be written in a matrix form. $Y = \mu + \mathbb{B}\beta + \pi$, where $Y = (Y(s), s \in S)'$, $\beta = (\beta_1', \beta_2', \ldots, \beta_J')'$, $\mathbb{B}$ is the design matrix constructed by spline functions and $\pi = (\pi(s), s \in S)'$.

For $v = (v_1', v_2', \ldots, v_J')'$, where v_j's are vectors, and $\psi = (\psi_1, \psi_2, \ldots, \psi_J)'$, we define a weighted ℓ_1/ℓ_2-norm, $\|v\|_{2,1,\psi} = \sum_{j=1}^{J} \psi_j \|v_j\|_2$, where $\|\cdot\|_2$ is the ℓ_2-norm of a vector. This is a weighted ℓ_1-norm of $(\|v_1\|_2, \cdots, \|v_J\|_2)$ with a weight vector, ψ. Then, we propose the following weighted ℓ_1/ℓ_2-penalized least-squares objective function, weighted by spatial weight matrix Σ_W:

$$Q_n(\beta, \lambda_n) = (Y - \mu - \mathbb{B}\beta)' \Sigma_W^{-1} (Y - \mu - \mathbb{B}\beta) + \lambda_n \|\beta\|_{2,1,\psi_n}, \tag{2.1.4}$$

where λ_n is a regularization parameter, $\psi_n = (\psi_{n1}, \psi_{n2}, \ldots, \psi_{nJ})'$ is a suitable choice of a weight vector for the penalty term and β is the $(Jm_n \times 1)$ dimensional vector of control points introduced in (2.1.2). Σ_W is a known positive definite spatial weight matrix where more weights are given if two locations are closer and vice-versa. For example, we can construct Σ_W using some commonly used spatial covariance functions. Choosing Identity for the

spatial weight matrix, (2.1.4) is a unweighted objective function for dependent data additive model with a group LASSO penalty. Our theoritical findings covers both the weighted and unweighted objective function, although the choice of penalty function and the rate of convergence changes accordingly. The vartiation in choice of penalty is discussed elaborately in subsection 2.2.1.

We allow the possibility that the regularization parameter, λ_n, and the weight vector, ψ_n, can depend on the sample size, n. To avoid an identifiability issue, we assume that $\mathbb{E}f_j(X_j) = 0, \forall 1 \leq j \leq J$, which leads us to assume

$$\sum_{l=1}^{m_n} \beta_{jl} \mathbb{B}_l(X_j(s)) = 0, \qquad \forall 1 \leq j \leq J,\ s \in S. \tag{2.1.5}$$

Combining (2.1.4) and (2.1.5), we have the unconstrained objective function given by

$$Q_n(\beta, \lambda_n) = (Y^c - \mathbb{B}^c\beta)' \Sigma_W^{-1} (Y^c - \mathbb{B}^c\beta) + \lambda_n \|\beta\|_{2,1,\psi_n}, \tag{2.1.6}$$

where $Y^c = (Y^c(s), s \in S)' = (Y(s) - \bar{Y}, s \in S)'$ with $\bar{Y} = \frac{1}{n}\sum_{s\in S} Y(s)$ and $\mathbb{B}^c = (\mathbb{B}_1^c, \mathbb{B}_2^c, \ldots, \mathbb{B}_J^c)$ is the design matrix with a n by m_n matrix $\mathbb{B}_j^c$. Each row of $\mathbb{B}_j^c$ is $(\mathbb{B}_1^c(X_j(s)),$ $\ldots, \mathbb{B}_{m_n}^c(X_j(s)))$ with $\mathbb{B}_l^c(X_j(s)) = \mathbb{B}_l(X_j(s)) - \frac{1}{n}\sum_{s'\in S} \mathbb{B}_l(X_j(s'))$.

As the first step, we obtain an estimate of β by minimizing the objective function, $Q_{n1}(\beta, \lambda_{n1}) := Q_n(\beta, \lambda_{n1})$ with $\psi_{nj} = 1$, for all $j = 1, \ldots, J$. We call this estimate as a group LASSO (gL) estimate, $\hat{\beta}_{gL}(\lambda_{n1})$, and the corresponding objective function as the gL objective function. Note that $\|\beta\|_{2,1,\mathbf{1}} = \sum_{j=1}^J \|\beta_j\|_2$. To improve the selection, we use the following updated weights from $\hat{\beta}_{gL}$,

$$\psi_{nj} = \begin{cases} 1/\|\hat{\beta}_{gL,j}\|_2 & \text{if } \|\hat{\beta}_{gL,j}\|_2 > 0, \\ \infty & \text{if } \|\hat{\beta}_{gL,j}\|_2 = 0. \end{cases} \tag{2.1.7}$$

in an objective function which is called an adaptive group LASSO (AgL) objective function. That is, we define an AgL objective function, $Q_{n2}(\beta, \lambda_{n2}) := Q_n(\beta, \lambda_{n2})$ with ψ_n given in (2.1.7). The estimate from this updated objective function, $\hat{\beta}_{AgL}(\lambda_{n2}) = \arg\min_\beta Q_{n2}(\beta, \lambda_{n2})$, as a function of λ_{n2} is referred as an adaptive group LASSO estimate.

We define $\infty \times 0 = 0$, so components not selected by the gL method are not included for the AgL method. Also, due to the nature of weights in the penalty term, the AgL objective function puts higher penalty on components with smaller ℓ_2-norm and lower penalty on components with larger ℓ_2-norm of the gL estimates. Hence, the components with larger gL estimates have higher chance to be selected in the final model. Finally, the AgL estimates for μ and f_j are given by

$$\hat{\mu} \;=\; \bar{Y} = \frac{1}{n}\sum_{s\in S} Y(s) \quad \text{and} \quad \hat{f}_{AgL,j}(X_j(s)) = \sum_{l=1}^{m_n} \hat{\beta}_{AgL,jl}\mathbb{B}_l(X_j(s))$$

for all $j = 1, \ldots, J$, respectively.

Computationally efficient objective function: The objective function, (2.1.6), involves Σ_W^{-1} which may cause computational complicacy particularly for *l*arge n. To avoid such situation, we reformulate (2.1.6) using Cholesky decomposition of Σ_W^{-1}. Let $\Sigma_W^{-1} = LL'$, where L is a lower triangular matrix. Then, (2.1.6) can be rewritten as,

$$Q_n(\beta, \lambda_n) = (Z^c - \mathbb{D}^c\beta)'(Z^c - \mathbb{D}^c\beta) + \lambda_n \|\beta\|_{2,1,\psi_n}, \tag{2.1.8}$$

where $Z^c = L'Y^c$ and $\mathbb{D}^c = L'\mathbb{B}^c$ with $\mathbb{D}_j^c = L'\mathbb{B}_j^c$. Then, the objective function becomes the one with no spatial weight matrix with a new response variable Z so that we can adopt an available algorithm for a group LASSO method with independent errors. Note that we used a known spatial weight matrix so that Cholesky decomposition to get L is done only once.

2.2 Main Theoretical Results

In this section, we present results on asymptotic properties of the gL and AgL estimators introduced in Section 2.1. We start with introducing some notations. Let A_0 and A_* be the sets of zero and nonzero components, respectively. Without loss of generality, we consider $A_* = \{1, 2, \ldots, q\}$ and $A_0 = \{q+1, \ldots, J\}$. Also, we assume that there exists $\tilde{A}_0$ that satisfies $\sum_{j\in\tilde{A}_0} \|\beta_j\|_2 \le \eta_1$ for some $\eta_1 \ge 0$ and let $\tilde{A}_* = \{1, 2, \ldots, J\} \setminus \tilde{A}_0$. Existence of $\tilde{A}_0$ is referred to as generalized sparsity condition (GSC) [Zhang and Huang (2008)].

First, we consider the selection and estimation properties of a gL estimator. Let $\tilde{A}_\beta$ be the index set of nonzero gL estimates for β_j and $\hat{A}_f$ be the index set of nonzero gL estimators for f_j. Necessary assumptions to study asymptotic properties are given in Assumption 2.2.

(H 1) Among J covariates, the number of nonzero components, q, is fixed and there exists a constant $k_f > 0$ such that $\min_{1 \le j \le q} \|f_j\|_2 \ge k_f$.

(H 2) There exists $v > 0$ such that $\min_{s \neq s' \in S} \|s - s'\|_2 > v$, where $\|\cdot\|_2$ for a vector is the ℓ_2-norm so that $\|s - s'\|_2$ is the Euclidean distance between s and s'.

(H 3) The random vector $\epsilon = \{\epsilon(s), s \in S\} \sim \text{Gaussian}(0, \Sigma_T)$, where Σ_T is constructed by a stationary covariance function $\delta_T(h)$ which satisfies $\int_{D_n} \delta_T(h) dh < \infty$. $D_n \subset \mathbb{R}^d$ is the sampling region that contains the sampling locations S. Without loss of generality, we assume that the origin of $\mathbb{R}^d$ is in the interior of D_n and D_n is increasing with n.

(H 4) $f_j \in \mathcal{F}$ and $\mathbb{E} f_j(X_j) = 0$ for $j = 1, \ldots, J$, where

$$\mathcal{F} = \left\{ f \mid \left| f^{(k)}(s) - f^{(k)}(t) \right| \le C \left| s - t \right|^\nu, \, \forall s, t \in [a, b] \right\}$$

for some nonnegative integer k and $\nu \in (0, 1]$. Also suppose that $\tau = k + \nu > 1$.

(H 5) The covariate vector X has a bounded continuous density function $g_j(x)$ of X_j on a bonded domain $[a, b]$ for $j = 1, \cdots, J$.

(H 6) $m_n = O(n^\gamma)$ with $1/6 \le \gamma = 1/(2\tau + 1) < (1 - \alpha)1/3$.

(H 7) Σ_W is constructed by a stationary covariance function $\delta(h)$ that satisfies the same condition as $\delta_T(h)$ in (H 3) and $\kappa(\Sigma_W) = \kappa(\Sigma_W^{-1}) \le M$ for some $M < \infty$, where κ is the condition number of a matrix.

The assumption (H 1) indicates that we need strong signals for nonzero components to distinguish them from the noise. The assumption (H 2) implies that we consider increasing-domain asymptotics [Stein (1999)] for our large sample properties, which is a common sampling assumption for the asymptotic theory of spatial statistics.

The assumption (H 3) specifies distributional assumption of spatial error and its spatial dependence. Commonly available stationary spatial covariance models satisfy integrable assumption. For example, popular spatial covariance functions such as Exponential, Maérn, Gaussian covariance functions are all integrable. For the explicit expression of such covariance functions, please see the supplementary material. We assume that D_n contains the origin to have spatial lag h and observation location s on the same spatial domain. The stationary spatial covariance function $\delta(h)$ gives a marginal variance at $h = 0$ (i.e. $s = s'$) and decreases toward zero as $\|h\| \to \infty$. The condition on $\delta(h)$ in the assumption (H 3) becomes meaningful by assuming D_n contains the origin. For example, the condition does not guarantee whether $\delta(h)$ is integrable if we do not assume that D_n contains the origin.

The assumption (H 4) considers f_j is in the class of functions defined on $[a,b]$ such that the k^{th} derivative satisfies the Lipschitz condition of order ν and zero expectation condition is needed to avoid an identifiability issue.

The assumption (H 5) is needed to have spline approximation for additive components. The assumption (H 6) is related to the number of B-spline bases to approximate additive components. The parameter γ in assumption (H 6) controls the smoothness of additive components, where imposing an upper and lower bound to γ implies that those functions can be neither too smooth nor too wiggly. If functions are too smooth, then it would be hard to detect those distinctively from the overall mean, whereas, if the functions are too wiggly, then it would be hard to detect those distinctively from the random noise.

The assumption (H 7) implies that a spatial weight matrix, Σ_W, is a well-conditioned matrix. Note that we consider a stationary spatial covariance function to construct a spatial weight matrix. Under the assumption (H 2), one can see that the smallest eigenvalue of the spatial weight matrix constructed by several spatial covariance functions is bounded away from zero [see, e.g. Wendland (2005)]. On the other hand, the largest eigenvalue of a spatial weight matrix is related to the norm of the matrix. By the Geršgorin's theorem [Horn and Johnson (1985)], we can show that $\rho_{\max}(\Sigma_W) \leq \max_j \sum_k |\Sigma_{W,jk}| = \max_j \sum_k \delta(s_j - s_k)$,

where $\Sigma_{W,jk}$ is the (j,k)-th entry of Σ_W. This is bounded by a finite constant which is independent of n for the spatial weight matrix from stationary integrable covariance functions.

Now, we introduce the consistency result for the gL estimator.

Theorem 1. *Suppose that conditions in Assumption 2.2 hold, and if*

$$\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha}m_n \log(Jm_n)}$$

for a sufficiently large constant C. Then, we have

(a) $\sum_{j=1}^{J} \|\hat{\beta}_{gL,j} - \beta_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L)m_n^3\log(Jm_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda_{n1}^2}{n^2}\right)$,

(b) If $\frac{m_n^2\lambda_{n1}^2}{n^2} \longrightarrow 0$ *as* $n \longrightarrow \infty$, *all the nonzero components* $\beta_j, 1 \leq j \leq q$ *are selected with probability (w.p.) converging to* 1.

The spatial dependence of the data contributes to the theoretical lower bound of the penalty parameter and the convergence rate by an additional m_n term compared to the independent data case. Recall that m_n is the number of spline basis functions for approximating the f_j's. This additional m_n comes from the bound of the expected value for a function of spatially dependent error (see, Lemma **??**). The spatial weight matrix also contributes to the theoretical lower bound of the penalty parameter and the convergence rate via the maximum eigenvalue of L. Recall that L is the Cholesky decomposition component of Σ_W^{-1}. This implies that spatial dependence of the data and a spatial weight matrix affect the convergence rate and the selection of components via the choice of the penalty parameter. All these additional quantities make the lower bound of the penalty parameter for spatial additive models larger compared to the independent data case. A larger lower bound reduces overestimation even in case of $\Sigma_W = \mathbb{I}$ due to the additional m_n from spatial dependence. This is also observed in the simulation study.

To prove Theorem 1, we need to control spline approximation of f_j as well as the spatial dependence in the error terms. The boundedness of the number of selected components is critical to prove the theorem. These are investigated as lemmas in Appendix 2.5 and used in

the proof of the theorem. Although the approach to prove the asymptotic properties stated above for spatial additive models is similar to the one for additive models with independent errors, details are different due to spatial dependence as well as a spatial weight matrix in the proposed method. Proofs of Theorem 1 as well as subsequent theorems are given in the Appendix 2.5.

Next theorem provides consistency in terms of the estimated nonlinear components $\hat{f}_j$.

Theorem 2. *Suppose that conditions in Assumption 2.2 hold and if*

$$\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha}m_n\log(Jm_n)}$$

for a sufficiently large constant C. Then,

(a) $\|\hat{f}_{gL,j} - f_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L)m_n^2\log(Jm_n)}{n^{1-\alpha}} + \frac{1}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau}} + \frac{4m_n\lambda_{n1}^2}{n^2}\right)$ *for* $j \in \tilde{A}_\beta \cup A_*$, *where* $\tilde{A}_\beta$ *is the index set of nonzero gL estimates for* β_j,

(b) If $\frac{m_n\lambda_{n1}^2}{n^2} \longrightarrow 0$ *as* $n \longrightarrow \infty$, *all the nonzero components* $f_j, 1 \le j \le q$ *are selected w.p. converging to* 1.

Note that by (a) and (b) of Theorem 2 with $\lambda_{n1} = O(\rho_{\max}(L)\sqrt{n^{1+\alpha}m_n\log(Jm_n)})$ and $m_n = O(n^\gamma)$ with $1/6 \le \gamma = 1/(2\tau+1) < (1-\alpha)1/3$ (assumption (H6)), we have,

(i) $\|\hat{f}_{gL,j} - f_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L)n^{2\gamma}\log(Jm_n)}{n^{1-\alpha}}\right)$, for $j \in \tilde{A}_\beta \cup A_*$ and,

(ii) If $\frac{\rho_{\max}^2(L)\log(J)}{n^{1-\alpha-2\gamma}} \longrightarrow 0$ as $n \longrightarrow \infty$ then, with probability converging to 1, all the nonzero components $f_j, 1 \le j \le q$ are selected.

Hence, we can infer that, the number of additive components, J, can be as large upto,

$$\exp\left(o\left(\rho_{\min}^2(L)\,n^{(1-\alpha-2\gamma)-2\gamma}\right)\right).$$

In particular, if we need second order differentiability for the functions f_j, then $\tau = 2$ implies $\gamma = 1/5$ and in this case J can be as large as $\exp\left(o\left(\rho_{\min}^2(L)\,n^{(3/5-\alpha)3/5}\right)\right)$. Keeping L fixed, we can see that J increases exponentially in n.

Next, we state the additional assumptions for the asymptotic properties of the AgL estimator.

(K 1) The initial estimators, $\widehat{\beta}_{gL,j}$, are r_n-consistent :

$$r_n \max_{1\leq j\leq J} \left\|\widehat{\beta}_{gL,j} - \beta_j\right\|_2 = O_p(1), \quad \text{as} \quad r_n \longrightarrow \infty,$$

and there exists a constant $k_b > 0$ such that

$$\mathbb{P}(\min_{j\in A_*} \|\hat{\beta}_{gL,j}\|_2 \geq k_b b_{n1}) \longrightarrow 1$$

where $b_{n1} = \min_{j\in A_*} \|\beta_j\|_2 \asymp m_n^{1/2}$, where for two positive sequences a_n and b_n, $a_n \asymp b_n$ if there exists a_1 and a_2 such that $0 < a_1 < a_n/b_n < a_2 < \infty$.

(K 2)

$$\frac{\sqrt{\rho^2_{\max}(L)\, n^{1+\alpha} m_n \log(s_n m_n)}}{\lambda_{n2} r_n} + \frac{n^2}{\lambda^2_{n2} r_n^2 m_n} + \frac{\lambda_{n2} m_n}{n} = o(1)$$

where $s_n = J - |A_{**}|$. A_{**} is the set of indices that correspond to the components in the additive model which are correctly selected by the AgL approach. Mathematical definition is given in the proof of Theorem 3.

Note that (K 1) ensures the availability of a r_n-consistent estimator provided certain regularity conditions are satisfied. Also one can observe that under assumptions (H 1) to (H 7), a suitable choice of r_n is $\left(\frac{\rho^2_{\max}(L) m_n^3 \log(Jm_n)}{n^{1-\alpha}}\right)^{-1/2}$. Then, (K 2) can be replaced by (K 2)′ given as,

(K 2)′

$$\frac{\lambda_{n1}\sqrt{m_n}}{\lambda_{n2}} + \frac{\lambda_{n2} m_n}{n} = o(1).$$

Detail derivation is given in the supplementary material.

For the selection consistency of the AgL estimator, we introduce "$\hat{\beta}_{AgL} =_0 \beta$" which means that $sgn_0(\|\hat{\beta}_{AgL,j}\|_2) = sgn_0(\|\beta_j\|_2)$ for all j, where $sgn_0(\|x\|_2) = 1$ if $\|x\|_2 > 0$ and

$= 0$ if $\|x\|_2 = 0$. Define $J_0 = |A_* \cup \{j : \|\hat{\beta}_{AgL,j}\|_2 > 0\}|$. Note that J_0 is bounded by a finite number *w.p.* converging to 1 by Theorem 1.

Theorem 3. *Suppose that conditions in Assumptions 2.2 and 2.2 are satisfied. Then,*

(a) $\mathbb{P}(\hat{\beta}_{AgL} =_0 \beta) \longrightarrow 1$,

(b) $\sum_{j=1}^{q} \|\hat{\beta}_{AgL,j} - \beta_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^3 \log(J_0 m_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right)$.

By Theorem 3, we can show that the proposed AgL estimator of β can separate out true zero components for the spatial additive model. Similar to Theorem 1, we have additional quantities on the right hand side of the expression in Theorem 3 (b), which are due to the spatial dependence in errors as well as use of a spatial weight matrix. Since J_0 is bounded by a finite number *w.p.* converging to 1, the convergence rate of the AgL estimator of β is faster than the gL estimator of β. Next theorem shows that estimated components for f_js in the spatial additive model using the AgL estimator of β can identify zero components consistently. Also, the theorem provides the convergence rate for the estimated components.

Theorem 4. *Suppose that conditions in Assumptions 2.2 and 2.2 are satisfied. Then,*

(a) $\mathbb{P}(\|\hat{f}_{AgL,j}\|_2 > 0, j \in A_* \text{ and } \|\hat{f}_{AgL,j}\|_2 = 0, j \notin A_*) \longrightarrow 1$,

(b) $\sum_{j=1}^{q} \|\hat{f}_{AgL,j} - f_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^2 \log(J_0 m_n)}{n^{1-\alpha}} + \frac{1}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau}} + \frac{4m_n\lambda_{n2}^2}{n^2}\right)$.

The upper bounds of the convergence rates in Theorems 1 – 4 show that the convergence rates are slower than those for the independent data case, which is not surprising given that we are dealing with dependent data. Also, our theoretical results show we need an improved lower bound of the penalty parameter for spatial additive models, which is critical since the penalty parameter is sensitive to the selection results in practice. This is supported by the simulation study in the next section where we can clearly see worse performance when we blindly apply the approach developed for independent data to select non-zero components in a spatial additive model.

We assumed that each additive component shares the same smoothness by (H 4) in Assumption 2.2. We can extend our results to allow different levels of smoothness for additive

components without much difficulty. Necessary changes in Assumption 2.2 are

(H 4)′ $f_j \in \mathcal{F}_j$ and $\mathbb{E}f_j(X_j) = 0$ for $j = 1, \ldots, J$, where

$$\mathcal{F}_j = \left\{ f \,|\, \left| f^{(k_j)}(s) - f^{(k_j)}(t) \right| \leq C\,|s-t|^{\nu_j}\,,\ \forall s,t \in [a,b] \right\}$$

for some nonnegative integer k_j and $\nu_j \in (0,1]$. Also suppose that $\tau_j = k_j + \nu_j > 1$, and,

(H 6)′ $m_{nj} = O(n^{\gamma_j})$ with $1/6 \leq \gamma_j = 1/(2\tau_j + 1) < (1-\alpha)1/3,\ \forall j = 1, 2, \ldots, J$, where m_{nj} is the number of B-spline bases (or knots) to approximate the j^{th} additive component.

The revised theorems and lemmas are provided in the supplementary material, where $m_n = \max_{j=1,\ldots,J} m_{nj}$. Results are similar except a few changes due to the introduction of m_{nj}. In practice, we standardize the range and variability of all additive components and use the same number of knot points for each component which we choose as m_n. Note that the largest number of knot points m_n can cover all smooth additive components. Since the order of m_n also has to be between $\left[n^{1/6}, n^{(1-\alpha)1/3}\right)$ according to the assumption (H 6)′, we use this bound as a guide line to choose m_n. The suggestion of using the same number of knot points for each component, in practice, is also suggested by Hastie and Tibshirani (1990).

2.2.1 Selection of a penalty parameter and a spatial weight matrix

The selection result is sensitive to the choice of a penalty parameter (or regularization parameter). In addition to the penalty parameter, the proposed approach for a spatial additive model requires to choose a spatial weight matrix as well. Theoretical results only provide the lower bound of the penalty parameter which involves the information of a spatial weight matrix through $\rho_{\max}(L)$. The approach to find an optimal penalty parameter in the penalized methods for independent data can not be applied directly to our setting due to the spatial weight matrix. Complete theoretical investigation for finding an optimal value is interesting, however, beyond the scope of this article. Also, theoretically obtained optimal

choice of the penalty parameter is not feasible in practice since it is only valid asymptotically and it often depends on unknown nuisance parameters in the true model [Fan and Tang (2013)]. Thus, we demonstrate below a practical way of selecting the penalty parameter guided by theoretical results derived in this article.

We assume a spatial weight matrix is constructed by a class of stationary spatial covariance functions controlled by a parameter, ρ, for simplicity. We call ρ a spatial weight parameter. Example classes of spatial covariance functions that satisfy the assumption (H 7) to construct a spatial weight matrix are Gaussian covariance function and inverse multiquadratic function. The explicit expression of these functions are given in the supplementary material. The selection problem is then reduced to select a spatial weight parameter. To choose a penalty parameter together, we consider a theoretical lower bound of the penalty parameter as our penalty parameter value at a given value of ρ so that one parameter (a spatial weight parameter) controls both regularization level and spatial weight. Finally, we adopt generalized information criterion (GIC) [Nishii (1984) and Fan and Tang (2013)] as a measure to choose a spatial weight parameter. That is, we find ρ that minimizes $GIC(\lambda_n(\rho)) = \log(RSS) + df_{\lambda_n}\frac{\log\log(n)\log(J)}{n}$. In practice, we consider a sequence of ρ and choose the one that minimizes GIC. While experimenting with different information criteria and comparing with the existing cross validation criterion suggested by Yuan and Lin (2006), we noticed that we cannot have an initial least square estimator when $J > n$. Thus, we define degrees of freedom (df_{λ_n}) as the total number of estimated non-zero components, i.e. $df_{\lambda_n} = \hat{q}_{\lambda_n} m_n$ where $\hat{q}_{\lambda_n}$ is the active set of selected variables.

For the independent data case, Huang *et* al. (2010b) suggested extended Bayesian information criterion (EBIC) to choose a penalty parameter, which requires to choose an additional parameter (ν in their paper). We found that a smaller value compared to the suggested value for the additional parameter works better in our setting from the simulation study. Given the sensitivity of EBIC with this additional parameter, we instead recommend to use GIC, which does not have any additional parameter.

There are two places where a spatial covariance model is considered. One is for modeling a spatial dependence of the data and the other one is for constructing a spatial weight matrix in the objective function. Our theory shows that the method is valid for a class of underlying spatial distributions that satisfy the condition (**H** 3) and for a class of spatial weight matrices that satisfy the condition (**H** 7). However, some spatial covariance models that satisfy (**H** 3) may not satisfy (**H** 7). In this regard, our approach is more general as it covers the case that the true spatial covariance matrix as a spatial weight matrix as long as both conditions (**H** 3) and (**H** 7) are valid. Further, our objective is not to estimate the true covariance matrix and our method does not require to estimate the true covariance matrix to select additive components.

2.3 Numerical investigation

2.3.1 Simulation study

In this section, we present a simulation study to illustrate our theoretical findings. We consider $S = \{(s_1, s_2), s_i, s_j = 1, \cdots, m, \}$ with $m = 6, 12, 24$ which makes sample sizes, $n = 36, 144, 576$, respectively and we consider $J = 15, 25, 35$. We have $q = 4$ nonzero components which are $f_1(x) = 5x$, $f_2(x) = 3(2x-1)^2$, $f_3(x) = 4\sin(2\pi x)/(2-\sin(2\pi x))$,$f_4(x) = 0.6\sin(2\pi x) + 1.2\cos(2\pi x) + 1.8\sin^2(2\pi x) + 2.4\cos^3(2\pi x) + 3.0\sin^3(2\pi x)$ and the remaining components are set to zero. That is, $f_j(x) \equiv 0$ for $j = 5, \cdots, J$. The covariates are $X_j = (W_j + tU)/(1+t)$, for $j = 1, \cdots, J$, where W_j and U are i.i.d. from Uniform[0, 1]. Note that correlation between X_j and X_k is given as $t^2/(1+t^2)$, therefore, with an increase in t, the dependence among covariates increases. In the simulation study, we consider $t = 1$ and $t = 3$. Note that the nonzero components, $f_1, \cdots, f_4$ given above are taken from Huang *et* al. (2010b) which are originally introduced by Lin and Zhang (2006).

We assume that the error process follows a stationary mean zero Gaussian process with a spatial covariance function, $\delta(h)$. To investigate selection performance of the proposed

method by the type of spatial dependence of the process, we consider three different covariance models: Exponential, Matérn and Gaussian covariance functions. Exponential and Gaussian covariance functions have two parameters, σ^2 and ρ and Matérn covariance function involves one more parameter, ν. The expression of such spatial covariance functions are given in the supplementary material. For simplicity, we set the variance, $\sigma^2 = 1$. For an exponential covariance function, we consider $\rho = 0.5$ and 1. For a Matérn covariance function, we consider $\nu = 3/2$ and $5/2$ and for each of these cases we have chosen ρ to be 1.5 and 2.5. For a Gaussian covariance function, we consider $\rho = 1.5$ and 2.5. The parameter ρ controls how fast the covariance function decays as the distance $\|h\|$ increases, thus, the level of spatial dependence within the covariance model. Given specified f_js and spatial dependence structure of the error process described above, we generate 100 data sets for each case.

Three covariance functions are characterized by mean square differentiability of the process. A Gaussian process with an exponential covariance function is continuous in a mean squared sense while a Gaussian process with a Gaussian covariance function is infinitely differentiable. As an intermediate, a Gaussian process with a Matérn covariance function is $\lceil \nu - 1 \rceil$ times differentiable in a mean squared sense, where $\lceil x \rceil$ is the smallest integer larger than or equal to x [Stein (1999)]. The mean square differentiability is related to the smoothness of the processes, that is, the local behavior of the processes, thus, we can also investigate selection performance in view of local property of the processes by considering different types of covariance functions.

Once the data are generated, the selection performance of the proposed method is examined under several choices of spatial weight matrix. In particular, we considered two classes: Gaussian and Inverse Multiquadratic functions. In addition, we also considered identity matrix and the true covariance function of the underlying process as a spatial weight matrix for comparison. When a spatial weight matrix is controlled by a spatial weight parameter, we applied the approach introduced in the section 2.2.1 to select both a spatial

weight parameter and a penalty parameter. When we consider no spatial weight matrix, we do not have a spatial weight parameter to control. In this case, we considered a sequence of the penalty parameter values around the theoretical lower bound of λ_{n1} and choose the one that minimizes GIC.

We also implemented the method by Huang *et* al. (2010b) which was developed for independent data for comparison. We refer this approach as 'independent approach'. Following the standard practice, we computed the average and standard deviation of True Positive (TP) and False Positive (FP). TP is the number of additive components that are correctly selected, FP is the number of additive components that are falsely selected. In our simulation setting, the desired values of TP and FP are $4(=q)$ and 0, respectively.

Table 2.2 shows the selected results for the case with $t=1$ when generating X_j, $m=6,24$ and a selected set of true correlation parameter values. Complete simulation results are given in the supplementary material. To see the applicability of our method for the case where the dependence between covariates increases, we include results corresponding to $t=3$ in the supplementary material as well.

The first row (where Spatial Weight is 'None(Indep)') in each of the covariance model block in Table 2.2 corresponds to independent data approach. This clearly indicates overestimation of FP components. By looking at TP, one may think the result is good for the independent approach, but with the expense of larger FP, the method is actually selecting many more components than the truth. The trend remains even when sample size increases. This is expected as discussed before. The following rows are results from various choices of spatial weight matrices. Our method successfully reduced overestimation. Even when the spatial weight matrix is an identity matrix under dependent error models, our method still reduces overestimation of selected components compared to the independent approach. The result persists for various sample sizes (m), covariance models we considered, choices of spatial weight matrices and of course a large number of covariates (J).

When the true covariance model is exponential or Matérn and we used the true co-

J	Cov Model	Spatial Weights	m=6 GLASSO		m=24 GLASSO	
			True Positive	False Positive	True Positive	False Positive
15	Exp(0.5)	None(Indep)	3.74(0.48)	3.63(1.83)	4(0)	0.55(0.77)
		I	3.17(0.82)	2.02(1.34)	4(0)	0(0)
		Gauss	3.01(0.85)	1.76(1.24)	4(0)	0(0)
		InvMQ	2.94(0.93)	1.58(1.44)	4(0)	0(0)
		True	1.66(1.02)	0.29(0.56)	4(0)	0(0)
	$Mat_{3/2}(2.5)$	None(Indep)	3.8(0.45)	3.63(1.86)	4(0)	0.34(0.57)
		I	3.26(0.77)	1.87(1.5)	4(0)	0.05(0.22)
		Gauss	2.98(0.85)	1.64(1.34)	4(0)	0.01(0.1)
		InvMQ	2.74(0.86)	1.5(1.18)	4(0)	0(0)
		True	0.8(0.68)	0.09(0.32)	4(0)	0(0)
	$Mat_{5/2}(2.5)$	None(Indep)	3.82(0.41)	3.59(2.08)	4(0)	0.58(0.88)
		I	3.22(0.76)	2.06(1.55)	4(0)	0.04(0.2)
		Gauss	2.91(0.89)	1.72(1.35)	4(0)	0(0)
		InvMQ	2.97(0.89)	1.67(1.33)	4(0)	0(0)
		True	0.37(0.56)	0.04(0.24)	4(0)	0(0)
	Gauss(1.5)	None(Indep)	3.76(0.49)	3.97(2.06)	4(0)	0.59(0.81)
		I	3.23(0.75)	2.07(1.24)	4(0)	0.03(0.17)
		Gauss	3.02(0.82)	1.86(1.25)	4(0)	0.01(0.1)
		InvMQ	2.76(0.79)	1.58(1.17)	4(0)	0(0)
		True	3.3(0.73)	2.32(1.55)	4(0)	0.12(0.33)
35	Exp(0.5)	None(Indep)	3.38(0.74)	6.22(2.39)	4(0)	1.49(1.55)
		I	2.62(0.94)	3.15(2.14)	4(0)	0.04(0.2)
		Gauss	2.39(0.98)	2.59(1.84)	4(0)	0.02(0.14)
		InvMQ	2.22(1.04)	2.38(1.79)	4(0)	0(0)
		True	1.21(0.95)	0.48(0.78)	4(0)	0(0)
	$Mat_{3/2}(2.5)$	None(Indep)	3.48(0.67)	5.79(2.32)	4(0)	1.25(1.37)
		I	2.71(0.9)	2.94(1.75)	4(0)	0.1(0.39)
		Gauss	2.57(0.92)	2.59(1.6)	4(0)	0.01(0.1)
		InvMQ	2.25(0.9)	2.24(1.68)	4(0)	0(0)
		True	0.52(0.58)	0.12(0.41)	4(0)	0(0)
	$Mat_{5/2}(2.5)$	None(Indep)	3.33(0.74)	5.59(2.19)	4(0)	1.34(1.51)
		I	2.72(1)	3.06(1.97)	4(0)	0.1(0.33)
		Gauss	2.53(1.04)	2.57(1.81)	4(0)	0.01(0.1)
		InvMQ	2.25(0.97)	2.17(1.68)	4(0)	0(0)
		True	0.34(0.57)	0.05(0.26)	4(0)	0(0)
	Gauss(1.5)	None(Indep)	3.15(0.88)	6.37(1.95)	4(0)	1.78(1.54)
		I	2.44(1.02)	3.35(1.98)	4(0)	0.08(0.27)
		Gauss	2.24(0.95)	2.83(1.61)	4(0)	0.02(0.14)
		InvMQ	2.18(0.9)	2.2(1.41)	4(0)	0(0)
		True	2.78(0.87)	3.66(1.74)	4(0)	0.17(0.45)

Table 2.2 Monte Carlo Mean (Standard dev.) for the selected number of nonzero covariates using 100 datasets under both Independent and Dependent setup using a spatially weighted group LASSO algorithm

variance model as a spatial weight matrix, the proposed method did not perform well in terms of TP for small sample size. One possible reason is that the exponential and Matérn covariance functions in the spatial weight matrices produced larger maximum eigenvalues of L for small sample. For example, when $\text{Mat}_{5/2}(2.5)$ is used for a spatial weight, the corresponding maximum eigenvalue of L is 20.32 for $m = 6$ while the maximum eigenvalue of L is 1.23 for Gauss(1.5) as a spatial weight. A larger maximum eigenvalue of L makes a larger penalty parameter, so less components are selected. However, the performance improves as m increases. For small size ($m = 6$), inverse multiquadratic spatial weights tend to underestimate so that the TP is lower compared to other spatial weight matrix choices, but improving as m increases. Gaussian spatial weight matrix maintains similar level of TP while FP is reduced compared to the independent approach. Thus, we recommend to use a Gaussian spatial weight matrix, in particular for small sample size.

Results for increased dependence among covariates ($t = 1$ to $t = 3$) show that there is an increase (overestimation) of FP. Please see tables in section 6 of the supplementary material. This is somewhat expected since strong dependence between covariates may hinder the selection power of the variable selection approaches, in turn, results in selection of more components. However, our approach performs comparatively better than the independent approach.

2.3.2 Real data example

We considered lung cancer mortality data over the period of 2000-2005, obtained from Surveillance, Epidemiology, and End Results Program (SEER, www.seer.cancer.gov) by the National Cancer Institute, U.S. as an illustrative example. The SEER data can be accessed by submitting a signed SEER Research Data Agreement. The data is avaiable at, `seer.cancer.gov/data/access.html`. The SEER data includes incidence or mortality of cancers and associated variables for U.S. counties. We considered the southern part of Michigan which constitute of 68 counties.

We applied Tukey's transformation (e.g., Cressie and Chan (1989)) to age-adjusted lung cancer mortality rates used as the response variable. We included 20 covariates obtained from the SEER database which are originally from the U. S. Census Bureau. We also added $PM_{2.5}$ (Particulate matter smaller than 2.5 micrometers) obtained from the U.S. EPA's National Emission Inventory (NEI) database (`www.epa.gov/air/data/emisdist.html?st~MI~Michigan`). Since emission data in this website is available for the years 2001 and 2002, we considered the average of $2001 - 2002$ emission data. The unit is tons per county.

For our analysis, we scaled each of our predictor variables to $[0, 1]$. We considered a Gaussian covariance function for the spatial weight matrix and a sequence of ρ value around the estimated ρ, obtained by fitting an empirical variogram. Then, we applied the selection approach introduced in Section 2.2.1. Selected variables for both group LASSO and adaptive group LASSO algorithms under independent and dependent error are presented in Table 2.3.

Our method of variable selection has a strict sense of selecting variables in the sense of dropping more variables. For this data example, our approach (adaptive group LASSO case) dropped two more variables among the variables selected by the independent approach. The selected components, 'Poverty' and 'Move different state', do not seem to be related to lung cancer mortality rates directly but one may think 'Poverty' can be a proxy of more relevant covariate to lung cancer mortality rates. For example, a study shows smoking is more prevalent in lower socio-economic groups of society [Haustein (2006)]. Thus, one can think of 'Poverty' as a proxy of tobacco use. Although a variable 'Move different state' was kept, our approach at least dropped a few more irrelevant variables compared to the independent approach.

To explore more on those covariates dropped by the proposed approach but selected by the independent approach, we fit a multiple linear regression model. Although we considered non-linear relationship in spatial additive models in this paper, simple linear regression can provide initial assessment about the result. We present outputs of linear regression model

ID		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Independent	**GLASSO**	×	✓	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	✓	✓
	AGLASSO	×	✓	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	✓	✓
Dependent	**GLASSO**	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓
	AGLASSO	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×

Table 2.3 Comparing the two methods (Independent approach and Dependent approach) for the real data example using Group LASSO and Adaptive Group LASSO Algorithm. Variable description for ID is 1: Population Mortality, 2: Poverty, 3: PM25, 4: Urban, 5: Nonwhite, 6: Never Married, 7: Agriculture, 8: Unemployment, 9: White Collar, 10: Higher Highschool, 11: Age more than 65, 12: Age less than 18, 13: Crowding, 14: Foreign born, 15: Language isolation, 16: Median household income, 17: Same house no migration, 18: Move same County, 19: Move same State, 20: Move different State, 21: Normalized cost of living

Variable	**E**stimate	**S**td. Error	**t** value	**p**-value
Intercept	0.43958	0.07400	5.940	1.41e-07
Poverty	0.42467	0.12887	3.295	**0**.00163
Unemployment	0.01086	0.14638	0.074	0.94107
Move same State	-0.01837	0.10624	-0.173	0.86331
Move diff State	-0.15474	0.10379	-1.491	**0**.14106
Normalized cost of living	0.02715	0.07346	0.370	0.71298

Table 2.4 Coefficient estimates(standard error) and corresponding p-values obtained from Linear regression using R of the variables selected under independent error assumption

with five covariates in Table 2.4. This shows that our approach selected two most significant variables based on p-values (bold) while independent approach selected some insignificant variables.

2.4 Discussion

We established a method of selecting non-zero components in spatial additive models using a two-step adaptive group LASSO type approach and a spatial weight in the ℓ_2-error term. The consistency results allow the number of additive components to increase exponentially with the sample size. The theory showed that the lower bound of the penalty parameter involves the spatial weight matrix. Thus, we considered an approach of choosing a spatial weight matrix together with a penalty parameter that works well in practice. Furthermore, our theoretical result implies that the proposed variable selection method still

works with different convergence rate and different lower bound of the penalty parameter, compared to the independent data approach, when an identity matrix is used as a spatial weight matrix while the observed data has a stationary dependence. Indeed, the simulation results showed superiority of our approach in this case compared to the straight use of independent data approach.

In the gL objective function, we introduced a spatial weight matrix in the ℓ_2 error term, which looks like the generalized least square. If we use the covariance matrix of the true process, it is the form of a generalized least squares. Note, then the problem becomes simultaneous estimation of mean and variance, which is beyond the scope of this paper as our goal is mean selection, not covariance estimation.

The condition for the spatial covariance function in the assumption (H 3) can be extended to $\int_{D_n} \delta(h)dh = O(n^{\alpha})$ for some $\alpha \in [0,1)$. Here $\alpha = 0$ corresponds to the current assumption (H 3). By allowing $0 < \alpha < 1$, we can include spatial covariance models of long-memory processes, so that the theorems under this assumption cover a broader class of spatial covariance functions. The theoretical lower bound of the penalty parameter and the convergence rate will be modified as α is introduced. We provide revised lemmas, theorems and related discussion in the supplementary material for this extended setting.

Spatial covariance models that correspond to $\alpha = 0$ still have parameters that control spatial dependence. Our current theoretical results only bring an additional m_n into the convergence rate and the theoretical lower bound of the penalty parameter when we consider spatial dependence. This is due to the bound that we used in proving Lemma 3. If we can find a tighter bound which depends on those covariance parameters, it would have helped understanding the role of those covariance parameters better in variable selection of spatial additive models.

2.5 Proofs of theorems

Before proving the theorems stated in Section 2.2, we introduce some notations and lemmas. Note that for $f \in \mathcal{F}$, there exists $f_n \in \mathcal{S}_n$ such that $\|f - f_n\|_2 = O(m_n^{-\tau})$, where $\|\cdot\|_2$ for a function is defined as $\|f\|_2 = \sqrt{\int_a^b f^2(x)\mathrm{d}x}$ and $\mathcal{S}_n$ is the space spanned by B-spline bases [e.g. see Huang *e*t al. (2010b)]. Then, define the centered $\mathcal{S}_{nj}^0$ by,

$$\mathcal{S}_{nj}^0 = \left\{ f_{nj} : f_{nj} = \sum_{l=1}^{m_n} b_{jl}\mathbb{B}_l^c(x), (b_{j1}, \cdots, b_{jm_n}) \in \mathbb{R}^{m_n} \right\}, 1 \le j \le J. \tag{2.5.1}$$

Recall that $\mathbb{B}_l^c(x) = \mathbb{B}_l(x) - \frac{1}{n}\sum_{s' \in S}\mathbb{B}_l(X_j(s'))$, which depend on X_j. Also, for the purpose of emphasizing the fact that, π, ϵ and θ depends on sample size n we will use suffix n for each of these three quantities in our proofs.

Recall that $A_0 = \{j : f_j(x) \equiv 0, 1 \le j \le J\}$ and $A_* = \{1, 2, \ldots, J\} \setminus A_0$ so that A_0 and A_* are the sets of zero and nonzero components, respectively. We introduced $\tilde{A}_0$ as the index set that satisfies $\sum_{j \in \tilde{A}_0} \|\beta_j\|_2 \le \eta_1$ for some $\eta_1 \ge 0$ and let $\tilde{A}_* = \{1, 2, \ldots, J\} \setminus \tilde{A}_0$. Without loss of generality, we can assume $A_0 \subseteq \tilde{A}_0$ so that $\tilde{A}_* \subseteq A_*$ and $q^* := |\tilde{A}_*| \le q$. For any subset $A \subset \{1, 2, \ldots, J\}$, define $\beta_A = (\beta_j, j \in A)'$ and $\Omega_A = \mathbb{D}_A^{c\prime}\mathbb{D}_A^c/n$, where $\mathbb{D}_A^c = L'\mathbb{B}_A^c$ and $\mathbb{B}_A^c = (\mathbb{B}_j^c, j \in A)$ is the sub-design matrix formed by the columns indexed in the set A. Note that $\beta_{A_0} \equiv \mathbf{0}$. Also, denote the minimum and maximum eigenvalues and the condition number of a matrix M by $\rho_{\min}(M)$, $\rho_{\max}(M)$ and $\kappa(M)$, respectively.

Lemma 1 (Lemma 1 in Huang et al. (2010b))**.** *Suppose that $f \in \mathcal{F}$ and $\mathbb{E}f(X_j) = 0$. Then under (H 4) and (H 5) in Assumption 2.2, there exists an $f_n \in \mathcal{S}_{nj}^0$ such that*

$$\|f_n - f\|_2 = O_p\left(m_n^{-\tau} + \sqrt{\frac{m_n}{n}}\right). \tag{2.5.2}$$

Particularly, under the choice of $m_n = O\left(n^{\frac{1}{2\tau+1}}\right)$, we have

$$\|f_n - f\|_2 = O_p\left(m_n^{-\tau}\right) = O_p\left(n^{-\frac{\tau}{2\tau+1}}\right). \tag{2.5.3}$$

Lemma 2. *Suppose that $|A|$ is bounded by a fixed constant independent of n and J. Let $h_n \asymp m_n^{-1}$. Then under (H 4) and (H 5) in Assumption 2.2, with probability converging to one,*

$$\rho_{\min}(\Sigma_W^{-1})d_1 h_n \leq \rho_{\min}(\Omega_A) \leq \rho_{\max}(\Omega_A) \leq \rho_{\max}(\Sigma_W^{-1})d_2 h_n \tag{2.5.4}$$

Additionally under (H 7), (2.5.4) *becomes,*

$$c_1 h_n \leq \rho_{\min}(\Omega_A) \leq \rho_{\max}(\Omega_A) \leq c_2 h_n \tag{2.5.5}$$

where d_1, d_2, c_1 and c_2 are some positive constants.

***P**roof.* One can follow the proof of the Lemma 3 in Huang *e*t al. (2010b) but after observing that,

$$\rho_{\min}(\Sigma_W^{-1})\left(\frac{\mathbb{B}_A^{c\prime}\mathbb{B}_A^c}{n}\right) \leq \Omega_A = \frac{\mathbb{D}_A^{c\prime}\mathbb{D}_A^c}{n} \leq \rho_{\max}(\Sigma_W^{-1})\left(\frac{\mathbb{B}_A^{c\prime}\mathbb{B}_A^c}{n}\right)$$

which gives (2.5.4). By (H 7), the well-conditioned property of Σ_W^{-1}, we have (2.5.5). □

Lemma 3. *Define M_n be a non-negative definite matrix of order n and,*

$$T_{jl} = \left(\frac{m_n}{n}\right)^{\frac{1}{2}} a_{jl}' M_n \epsilon \qquad \forall 1 \leq j \leq J, 1 \leq l \leq m_n \tag{2.5.6}$$

where $a_{jl} = (\mathbb{B}_l^c(X_j(s)), s \in S)'$ and $T_n = \max_{\substack{1\leq j\leq J \\ 1\leq l\leq m_n}} |T_{jl}|$. Then, under assumptions (H 2) to (H 5) in Assumption 2.2,

$$\mathbb{E}(T_n) \leq C_1 \rho_{\max}(M_n)\sqrt{(m_n \log(Jm_n))O(n^\alpha)}, \tag{2.5.7}$$

for some $C_1 > 0$.

***P**roof.* Since $\epsilon \sim \text{Gaussian}(\mathbf{0}, \Sigma_T)$, $T_{jl} \sim \text{Gaussian}(0, \frac{m_n}{n} a_{jl}' M_n \Sigma_T M_n' a_{jl})$. Therefore we can use maximal inequalities of sub-Gaussian random variables [van der Vaart and Wellner (1996), Lemmas 2.2.1 and 2.2.2]. Let $\|\cdot\|_\phi$ be the Orlicz norm, defined by $\|X\|_\phi = \inf\{k \in (0,\infty) \,|\, \mathbb{E}(\phi(|X|/k)) \leq 1\}$. Then, conditional on $\{X_j(s), s \in S, 1 \leq j \leq J\}$, we have

the following

$$\begin{aligned}
&\Big\| \max_{1\le j\le J, 1\le l\le m_n} |T_{jl}| \Big| X_j(s), s\in S, 1\le j\le J \Big\|_{\phi_2} \\
&\quad \le K\sqrt{\frac{m_n}{n}\log(1+Jm_n)} \max_{1\le j\le J,1\le l\le m_n} \| \, |a'_{jl}M_n\epsilon| \Big| X_j(s), s\in S, 1\le j\le J\|_{\phi_2} \\
&\quad \le K\sqrt{\frac{m_n}{n}\log(Jm_n)} \max_{1\le j\le J,1\le l\le m_n} \sqrt{a'_{jl}M_n\Sigma_T M'_n a_{jl}},
\end{aligned}$$

where $K>0$ is a generic constant and $\phi_p(x)=e^{x^p}-1$. Now taking expectation with respect to $\{X_j(s), s\in S, 1\le j\le J\}$ on both sides of the above inequality,

$$\begin{aligned}
&\| \max_{1\le j\le J,1\le l\le m_n} |T_{jl}| \|_{\phi_2} \\
&\quad \le K\sqrt{\frac{m_n}{n}\log(Jm_n)}\, \mathbb{E}\left(\max_{1\le j\le J,1\le l\le m_n} \sqrt{a'_{jl}M_n\Sigma_T M'_n a_{jl}} \right) \\
&\quad = K\sqrt{\frac{m_n}{n}\log(Jm_n)}\, \mathbb{E}\left(\sqrt{\max_{1\le j\le J,1\le l\le m_n} a'_{jl}M_n\Sigma_T M'_n a_{jl}} \right) \\
&\quad \le K\rho_{\max}(M_n)\sqrt{\frac{m_n}{n}\log(Jm_n)} \sqrt{\mathbb{E}\left(\max_{1\le j\le J,1\le l\le m_n} a'_{jl}\Sigma_T a_{jl} \right)}. \qquad (2.5.8)
\end{aligned}$$

Since $\mathbb{B}^c_l(x)$ are normalized B-splines, we have

$$\begin{aligned}
&\mathbb{E}\left(\max_{1\le j\le J,1\le l\le m_n} \sum_{s\in S}\sum_{s'\in S} \mathbb{B}^c_l(X_j(s))\delta(s-s')\mathbb{B}^c_l(X_j(s')) \right) \\
&\qquad \le 4\sum_{s\in S}\sum_{s'\in S}\delta(s-s') \\
&\qquad \le K\sum_{s\in S}\int_{h\in CD_n}\delta(h)dh \\
&\qquad \le Kn\int_{h\in CD_n}\delta(h)dh. \qquad (2.5.9)
\end{aligned}$$

for some $K, C>0$.

From (2.5.8) and (2.5.9),

$$\begin{aligned}
\| \max_{1\le j\le J,1\le l\le m_n} |T_{jl}| \|_{\phi} \quad &\le K\rho_{\max}(M_n)\sqrt{m_n\log(Jm_n)}\sqrt{\int_{h\in CD_n}\delta(h)dh} \\
&\le K\rho_{\max}(M_n)\sqrt{m_n\log(Jm_n)O(n^{\alpha})}.
\end{aligned}$$

Finally, (2.5.7) follows from $\|X\|_{L^1}\le C\|X\|_{L^2}\le\|X\|_{\phi_2}$, where $\|X\|_{L^p}=(\mathbb{E}(|X|^p))^{1/p}$. □

Before we delve into the proof of the theorems, let us define and summarize some of index sets we will be using. Recall that $A_0 = \{j : f_j \equiv 0, 1 \le j \le J\}$. For an index set $\tilde{A}_1$ that satisfies $\tilde{A}_\beta = \{j : \|\hat{\beta}_{gL,j}\|_2 > 0\} \subseteq \tilde{A}_1 \subseteq \tilde{A}_\beta \cup \tilde{A}_*$, we consider the following sets:

	"Large" $\|\beta_j\|_2$ (i.e. $\tilde{A}_*$)	"Small" $\|\beta_j\|_2$ (i.e. $\tilde{A}_0$)
$\tilde{A}_1$	$\tilde{A}_3$	$\tilde{A}_4$
$\tilde{A}_2 = \tilde{A}_1^c$	$\tilde{A}_5$	$\tilde{A}_6$

We can deduce some relations from the above table $\tilde{A}_3 = \tilde{A}_1 \cap \tilde{A}_*$, $\tilde{A}_4 = \tilde{A}_1 \cap \tilde{A}_0$, $\tilde{A}_5 = \tilde{A}_1^c \cap \tilde{A}_*$, $\tilde{A}_6 = \tilde{A}_2 \cap \tilde{A}_0$, and hence we have $\tilde{A}_3 \cup \tilde{A}_4 = \tilde{A}_1$, $\tilde{A}_5 \cup \tilde{A}_6 = \tilde{A}_2$, and $\tilde{A}_3 \cap \tilde{A}_4 = \tilde{A}_5 \cap \tilde{A}_6 = \phi$. Also, let $|\tilde{A}_1| = q_1$. For an index set $\hat{A}_1$ that satisfies $\hat{A}_f = \{j : \|\hat{f}_{gL,j}\|_2 > 0\} \subseteq \hat{A}_1 \subseteq \hat{A}_f \cup A_*$,

	"Large" f_j (i.e. A_*)	"Small" f_j (i.e. A_0)
$\hat{A}_1$	$\hat{A}_3$	$\hat{A}_4$
$\hat{A}_2 = \hat{A}_1^c$	$\hat{A}_5$	$\hat{A}_6$

We have a similar set of relations from the above which is $\hat{A}_3 = \hat{A}_1 \cap A_*$, $\hat{A}_4 = \hat{A}_1 \cap A_0$, $\hat{A}_5 = \hat{A}_1^c \cap A_*$, $\hat{A}_6 = \hat{A}_2 \cap A_0$, and hence we have $\hat{A}_3 \cup \hat{A}_4 = \hat{A}_*$, $\hat{A}_5 \cup \hat{A}_6 = \hat{A}_2$, and $\hat{A}_3 \cap \hat{A}_4 = \hat{A}_5 \cap \hat{A}_6 = \phi$.

To prove Theorem 1, we need boundedness of $|\tilde{A}_\beta|$, which is given in the following lemma.

Lemma 4. *Under the Assumption 2.2 with* $\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha} m_n \log(Jm_n)}$ *for a sufficiently large constant C, we have* $|\tilde{A}_\beta| \le M_1 |A_*|$ *for a finite constant* $M_1 > 1$ *with w.p. converging to* 1.

P*roof.* Along with considering the approximation error for spline regression, we also have to take care of the dependence structure of a Gaussian random vector ϵ according to $(H\,3)$. To emphasize the dependence on n, we denote write ϵ_n instead of ϵ and similar notation for others as well. Recall $\pi_n = \epsilon_n + \theta_n$, where $\theta_n = (\theta_n(s); s \in S)'$ with $\theta_n(s) = \sum_{j=1}^J (f_j(X_j(s)) - f_{nj}(X_j(s)))$. Note that $\|\theta_n\|_2 = O(n^{1/2} q^{1/2} m_n^{-\tau}) = O(q^{1/2} n^{1/(4\tau+2)})$ by Lemma 1 since

$m_n = O(n^{1/(2\tau+1)})$. Define $\lambda_{n,J} = 2\sqrt{K\rho^2_{\max}(L) m_n n^{1+\alpha}\log(Jm_n)}$ for some $K > 0$ and $\lambda_{n1} \geq \max\{\lambda_0, \lambda_{n,J}\}$, where $\lambda_0 = \inf\{\lambda : M_1(\lambda)q^* + 1 \leq q_0\}$ for some finite $q_0 > 0$ and $M_1(\lambda) > 1$, which will be specified later in the proof. Without loss of generality, we will assume the infimum of an empty set to be ∞. That is, if $\{\lambda : M_1(\lambda)q^* + 1 \leq q_0\}$ is an empty set, it implies that $\lambda_{n1} = \lambda_0 = \infty$ and which in turn implies that we drop all the components in our additive model, i.e. $|\tilde{A}_\beta| = 0$. So part (i) is trivial in this case and hence for the rest of the proof we will assume $\{\lambda : M_1(\lambda)q^* + 1 \leq q_0\}$ is a non-empty set.

First, define a new vector U_k such that $U_k = \mathbb{D}_k{}^{c\prime}(Z^c - \mathbb{D}^c\hat{\beta}_{gL})\big/\lambda_{n1}$ for $k = 1, \cdots, J$. By **K**arsuh-**K**uhn-**T**ucker (KKT) conditions of the optimization problem for $Q_n(\beta, \lambda_n)$ with the solution $\hat{\beta}_{gL}$, we have

$$U_k \begin{cases} = \dfrac{\hat{\beta}_{gL,k}}{\|\hat{\beta}_{gL,k}\|_2} & \text{if } \|\hat{\beta}_{gL,k}\|_2 > 0, \\ \leq \mathbf{1} & \text{if } \|\hat{\beta}_{gL,k}\|_2 = 0. \end{cases} \tag{2.5.10}$$

Then, the norm of U_k is

$$\|U_k\|_2 \begin{cases} = 1 & \text{if } \|\hat{\beta}_{gL,k}\|_2 > 0, \\ \leq 1 & \text{if } \|\hat{\beta}_{gL,k}\|_2 = 0. \end{cases} \tag{2.5.11}$$

Now we introduce the following quantities.

$$\begin{aligned} x_r &= \max_{|A|=r} \max_{\|U_k\|_2=1, k\in A, B\subset A} \left|\pi_n' w_{A|B}\right|, \quad \text{and} \\ x_r^* &= \max_{|A|=r} \max_{\|U_k\|_2=1, k\in A, B\subset A} \left|\epsilon_n' w_{A|B}\right|, \end{aligned}$$

where $w_{A|B} = W_{A|B}/\left\|W_{A|B}\right\|_2$ with $W_{A|B} = \left(\mathbb{D}^c_A(\mathbb{D}^c_A{}'\mathbb{D}^c_A)^{-1}\lambda_{n1}Q'_{BA}Q_{BA}U_A - (\mathbb{I} - P_A)\mathbb{D}^c\beta\right)$. For $B \subset A$, Q_{BA} is the matrix corresponding to the selection of variables in B from A, *i.e.* $Q_{BA}\beta_A = \beta_B$. $P_A = \mathbb{D}^c_A\Omega^{-1}_A\mathbb{D}^c_A{}'/n$, $U_A = (U_k; k \in A)'$. By the triangle inequality and

Cauchy-Schwarz inequality, for some set A with $|A| = r > 0$, we have

$$\begin{aligned}\left|\pi_n' w_{A|B}\right| &\leq \left|\epsilon_n' w_{A|B}\right| + \|\theta_n\|_2 \\ &\leq \left|\epsilon_n' w_{A|B}\right| + K_2 q n^{1/(4\tau+2)} \\ &\leq \left|\epsilon_n' w_{A|B}\right| + K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}} \end{aligned} \tag{2.5.12}$$

where the last inequality holds for a sufficiently large n for some $K_1 > 0$. By introducing the following sets,

$$\begin{aligned}\Omega_{r_0} &= \left\{(\mathbb{D}^c, \pi_n); x_r \leq 2K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}}, \forall r \geq r_0\right\}, \quad \text{and} \\ \Omega_{r_0}^* &= \left\{(\mathbb{D}^c, \epsilon_n); x_r^* \leq K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}}, \forall r \geq r_0\right\},\end{aligned}$$

we can show $\mathbb{P}\left\{(\mathbb{D}^c, \pi_n) \in \Omega_{r_0}\right\} \geq \mathbb{P}\left\{(\mathbb{D}^c, \epsilon_n) \in \Omega_{r_0}^*\right\}$ for any $r_0 > 0$ since we have

$$x_r \leq x_r^* + \|\theta_n\|_2 \leq x_r^* + K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}} \tag{2.5.13}$$

by recalling the definitions of x_r and x_r^* and (2.5.12).

Now, we want to show that $\mathbb{P}\left\{(\mathbb{D}^c, \pi_n) \in \Omega_{q_1}\right\} \to 1$ implies $|\tilde{A}_\beta| \leq M_1 |\tilde{A}_*| = M_1 q^*$ for some finite $M_1 > 1$, which completes the proof since $q^* \leq q$. Before proving this claim, we first show $\mathbb{P}\left\{(\mathbb{D}^c, \epsilon_n) \in \Omega_{q_1}^*\right\} \to 1$, which implies $\mathbb{P}\left\{(\mathbb{D}^c, \pi_n) \in \Omega_{q_1}\right\} \to 1$. We start with the following:

$$1 - \mathbb{P}\left\{(\mathbb{D}^c, \epsilon_n) \in \Omega_{q_1}^*\right\} \tag{2.5.14}$$

$$\begin{aligned}&\leq \sum_{r=0}^{\infty} \mathbb{P}\left(x_r^* > K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}}\right) \\ &\leq \sum_{r=0}^{\infty} \binom{J}{r} \mathbb{P}\left(|w_{A|B}' \epsilon_n| > K_1 \sqrt{\frac{(rm_n \vee m_n)\rho_{\max}^2(L) n^\alpha m_n \log(Jm_n)}{\rho_{\max}(\Omega_{\tilde{A}_1})}}\right),\end{aligned} \tag{2.5.15}$$

where $|A| = r$. Since $w'_{A|B}\epsilon_n \sim \text{Gaussian}(0, w'_{A|B}\Sigma_T w_{A|B})$, (2.5.15) becomes

$$\begin{aligned}
&\leq 2\sum_{r=0}^{\infty}\binom{J}{r}\exp\left(-0.5K_1^2\frac{(rm_n \vee m_n)\rho_{\max}^2(L)n^{\alpha}m_n\log(Jm_n)}{(w'_{A|B}\Sigma_T w_{A|B})\rho_{\max}(\Omega_{\tilde{A}_1})}\right)\\
&\leq 2\sum_{r=0}^{\infty}\binom{J}{r}\exp\left(-0.5K_1^2\frac{(rm_n \vee m_n)\rho_{\max}^2(L)n^{\alpha}m_n\log(Jm_n)}{\rho_{\max}(\Sigma_T)\rho_{\max}(\Omega_{\tilde{A}_1})}\right)\\
&= 2\sum_{r=0}^{\infty}\binom{J}{r}(Jm_n)^{-0.5K_1^2(rm_n\vee m_n)\rho_{\max}^2(L)n^{\alpha}m_n\Big/\rho_{\max}(\Sigma_T)\rho_{\max}\left(\Omega_{\tilde{A}_1}\right)}.
\end{aligned} \tag{2.5.16}$$

Let $K_n = 0.5K_1^2 m_n^2\rho_{\max}^2(L)\,n^{\alpha}\Big/\left(\rho_{\max}(\Sigma_T)\,\rho_{\max}\left(\Omega_{\tilde{A}_1}\right)\right)$. Then (2.5.16) becomes,

$$\begin{aligned}
&= 2(Jm_n)^{-K_n} + 2\sum_{r=1}^{\infty}\binom{J}{r}(Jm_n)^{-rK_n}\\
&\leq 2(Jm_n)^{-K_n} + 2\sum_{r=1}^{\infty}\frac{1}{r!}\left(\frac{J}{(Jm_n)^{K_n}}\right)^r\\
&= 2(Jm_n)^{-K_n} + 2\exp\left(\frac{J}{(Jm_n)^{K_n}}\right) - 2.
\end{aligned} \tag{2.5.17}$$

Define $\|\Sigma_T\|_1 = \max_{s\in S}\sum_{s'\in S}\sigma_{s,s'}$ and note that $\|\Sigma_T\|_1 \asymp \int_{h\in D_n}\delta(h)dh = O(n^{\alpha}1)$. Therefore by using the fact, $\frac{1}{\sqrt{n}}\|\Sigma_T\|_1 \leq \rho_{\max}(\Sigma_T) \leq \sqrt{n}\|\Sigma_T\|_1$ and $\rho_{\max}(\Sigma_W^{-1}) \leq \rho_{\max}(L)\rho_{\max}(L') = \rho_{\max}^2(L)$, we have

$$K_n \geq c_1\frac{0.5K_1^2 m_n^3 n^{\alpha}}{\sqrt{n}\|\Sigma_T\|_1} \asymp 0.5c_1K_1^2\sqrt{n^{6\gamma-1}},$$

and $K_n \longrightarrow \infty$ by (H 6). This shows (2.5.17) goes to zero as $n \to \infty$.

To show $\mathbb{P}\left\{(\mathbb{D}^c, \pi_n) \in \Omega_{q_1}\right\} \to 1$ implies $|\tilde{A}_{\beta}| \leq M_1|\tilde{A}_*| = M_1 q^*$, let

$$V_{1j} = \frac{\Omega_{\tilde{A}_1}^{-\frac{1}{2}}Q'_{j1}U_{\tilde{A}_j}\lambda_{n1}}{\sqrt{n}}\ \text{, for } j = 1,3,4,$$

and

$$u = \frac{\mathbb{D}^c_{\tilde{A}_1}\Omega_{\tilde{A}_1}^{-1/2}V_{14}/\sqrt{n} - \omega_2}{\|\mathbb{D}^c_{\tilde{A}_1}\Omega_{\tilde{A}_1}^{-1/2}V_{14}/\sqrt{n} - \omega_2\|_2},$$

where, for simplicity in notations, $Q_{kj} = Q_{\tilde{A}_k\tilde{A}_j}$ is the matrix corresponding to the selection of variables in $\tilde{A}_k$ from $\tilde{A}_j$ and $\omega_2 = (\mathbb{I} - P_{\tilde{A}_1})\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2} = (\mathbb{I} - P_{\tilde{A}_1})\mathbb{D}^c\beta$. We can show

that, $V_{11} = V_{14} + V_{13}$ and $Q'_{31}Q_{31} + Q'_{41}Q_{41} = \mathbb{I}_{m_n|\tilde{A}_1|}$ due to the fact that $\tilde{A}_3 \cup \tilde{A}_4 = \tilde{A}_1, \tilde{A}_3 \cap \tilde{A}_4 = \phi$ and hence $\beta'_{\tilde{A}_3} Q_{31} + \beta'_{\tilde{A}_4} Q_{41} = \beta_{\tilde{A}_1}$. Since $q_1 = |\tilde{A}_1| = |\tilde{A}_3| + |\tilde{A}_4|$ and $|\tilde{A}_3| \leq q^*$, $|\tilde{A}_4| \geq (q_1 - q^*)$. Then, we have the following lower bound for L_2-norm of V_{14},

$$\|V_{14}\|_2^2 \geq \frac{\lambda_{n1}^2 \|Q'_{41} U_{\tilde{A}_4}\|_2^2}{n\rho_{\max}(\Omega_{\tilde{A}_1})} = \frac{\lambda_{n1}^2 \|Q'_{41} Q_{41} U_{\tilde{A}_1}\|_2^2}{n\rho_{\max}(\Omega_{\tilde{A}_1})} = \frac{\lambda_{n1}^2 m_n |\tilde{A}_4|}{n\rho_{\max}(\Omega_{\tilde{A}_1})} \geq B_1 \frac{(q_1 - q^*)^+}{q^*}, \quad (2.5.18)$$

with $B_1 = \dfrac{\lambda_{n1}^2 m_n q^*}{n\rho_{\max}\left(\Omega_{\tilde{A}_1}\right)}$. From (2.5.18), we have

$$|\tilde{A}_\beta| \leq |\tilde{A}_1| = q_1 \leq (q_1 - q^*)^+ + q^* \leq q^* \frac{\|V_{14}\|_2^2}{B_1} + q^* = \left(\frac{\|V_{14}\|_2^2}{B_1} + 1\right) q^*. \quad (2.5.19)$$

Thus, to show $|\tilde{A}_\beta| \leq M_1 q^*$ for some finite $M_1 > 1$, we need to show $\frac{\|V_{14}\|_2^2}{B_1} + 1 \leq M_1$ for some finite $M_1 > 1$.

We start with an upper bound of $\|V_{14}\|_2^2 + \|\omega_2\|_2^2$. Since

$$\begin{aligned} \|V_{14}\|_2^2 + \|\omega_2\|_2^2 &= V'_{14}V_{14} + \|\omega_2\|_2^2 = V'_{14}(V_{11} - V_{13}) + \|\omega_2\|_2^2 \\ &\leq V'_{14}V_{11} + \|V_{14}\|_2 \|V_{13}\|_2 + \|\omega_2\|_2^2, \end{aligned} \quad (2.5.20)$$

we find upper bounds for $V'_{14}V_{11}$, $\|V_{14}\|_2\|V_{13}\|_2$ and $\|\omega_2\|_2^2$, respectively. First,

$$\begin{aligned} V'_{14}V_{11} &= \frac{\lambda_{n1}^2}{n} U'_{\tilde{A}_4} Q_{41} \Omega_{\tilde{A}_1}^{-1} U_{\tilde{A}_1} \\ &= \frac{\lambda_{n1}^2}{n} U'_{\tilde{A}_4} Q_{41} \Omega_{\tilde{A}_1}^{-1} \left(\mathbb{D}^{c\prime}_{\tilde{A}_1} \left(Z^c - \mathbb{D}^c \hat{\beta}_{gL}\right) / \lambda_{n1}\right) \\ &= \frac{\lambda_{n1}}{n} U'_{\tilde{A}_4} Q_{41} \Omega_{\tilde{A}_1}^{-1} \mathbb{D}^{c\prime}_{\tilde{A}_1} \left(\mathbb{D}^c_{\tilde{A}_1} \beta_{\tilde{A}_1} + \mathbb{D}^c_{\tilde{A}_2} \beta_{\tilde{A}_2} + \pi_n - \mathbb{D}^c_{\tilde{A}_1} \hat{\beta}_{gL,\tilde{A}_1}\right) \\ &= \lambda_{n1} U'_{\tilde{A}_4} Q_{41} \left((\beta_{\tilde{A}_1} - \hat{\beta}_{gL,\tilde{A}_1}) + \frac{\Omega_{\tilde{A}_1}^{-1}}{n} (\mathbb{D}^{c\prime}_{\tilde{A}_1} \mathbb{D}^c_{\tilde{A}_2}) \beta_{\tilde{A}_2} + \frac{\Omega_{\tilde{A}_1}^{-1}}{n} \mathbb{D}^{c\prime}_{\tilde{A}_1} \pi_n\right) \\ &= \lambda_{n1} U'_{\tilde{A}_4} (\beta_{\tilde{A}_4} - \hat{\beta}_{gL,\tilde{A}_4}) + \lambda_{n1} U'_{\tilde{A}_4} Q_{41} \left(\frac{\Omega_{\tilde{A}_1}^{-1}}{n} (\mathbb{D}^{c\prime}_{\tilde{A}_1} \mathbb{D}^c_{\tilde{A}_2}) \beta_{\tilde{A}_2} + \frac{\Omega_{\tilde{A}_1}^{-1}}{n} \mathbb{D}^{c\prime}_{\tilde{A}_1} \pi_n\right) \\ &\leq \lambda_{n1} \sum_{k \in \tilde{A}_4} \|\beta_k\|_2 + \frac{\lambda_{n1} U'_{\tilde{A}_4} Q_{41} \Omega_{\tilde{A}_1}^{-1} (\mathbb{D}^{c\prime}_{\tilde{A}_1} \mathbb{D}^c_{\tilde{A}_2}) \beta_{\tilde{A}_2}}{n} + \frac{\lambda_{n1} U'_{\tilde{A}_4} Q_{41} \Omega_{\tilde{A}_1}^{-1} \mathbb{D}^{c\prime}_{\tilde{A}_1} \pi_n}{n}, \end{aligned} \quad (2.5.21)$$

where the last inequality is based on $\left|U'_{\tilde{A}_4}\beta_{\tilde{A}_4}\right| \leq \sum_{k\in\tilde{A}_4}\left|U'_k\beta_k\right| \leq \sum_{k\in\tilde{A}_4}\|\beta_k\|_2$ and

$$U'_{\tilde{A}_4}\hat{\beta}_{gL,\tilde{A}_4} = \sum_{k\in\tilde{A}_4\cap\tilde{A}_\beta} U'_k\hat{\beta}_{gL,k} \geq 0.$$

For $\|V_{14}\|_2\|V_{13}\|_2$, we have

$$\|V_{14}\|_2\|V_{13}\|_2 \leq \|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}} \tag{2.5.22}$$

from the definition of V_{13}. For $\|\omega_2\|_2^2$,

$$\begin{aligned}
\|\omega_2\|_2^2 &= \|(I-P_{\tilde{A}_1})\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2}\|_2^2 \\
&= \beta'_{\tilde{A}_2}\mathbb{D}^{c\prime}_{\tilde{A}_2}(I-P_{\tilde{A}_1})\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2} \\
&= \beta'_{\tilde{A}_2}\left(n\Omega_{\tilde{A}_2}\beta_{\tilde{A}_2} - \frac{1}{n}\mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2}\right) \\
&\leq \beta'_{\tilde{A}_2}\left(\lambda_{n1}D_{\tilde{A}_2} - \mathbb{D}^{c\prime}_{\tilde{A}_2}\pi_n - \mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}\left(\beta_{\tilde{A}_1} - \hat{\beta}_{gl,\tilde{A}_1}\right) - \frac{1}{n}\mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2}\right) \\
&= \beta'_{\tilde{A}_2}\left(\lambda_{n1}D_{\tilde{A}_2} - \mathbb{D}^{c\prime}_{\tilde{A}_2}\pi_n - \frac{\lambda_{n1}}{n}\mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}U_{\tilde{A}_1} + \frac{1}{n}\mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\pi_n\right) \\
&= \lambda_{n1}\beta'_{\tilde{A}_2}D_{\tilde{A}_2} - \omega'_2\pi_n - \frac{\lambda_{n1}}{n}U'_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2},
\end{aligned}$$

where the inequality is from

$$\mathbb{D}^{c\prime}_{\tilde{A}_2}\mathbb{D}^c_{\tilde{A}_1}(\beta_{\tilde{A}_1} - \hat{\beta}_{gL,\tilde{A}_1}) + n\Omega_{\tilde{A}_2}\beta_{\tilde{A}_2} + \mathbb{D}^{c\prime}_{\tilde{A}_2}\pi_n \leq \lambda_{n1}D_{\tilde{A}_2}. \tag{2.5.23}$$

In (2.5.23), D_A is a $0-1$ vector whose k^{th} entry is $I(\|\hat{\beta}_{k,gL}\|_2 = 0)$, where $I(A)$ is the indicator function for a set A. The inequality between vectors is defined entry-wise. Note that (2.5.23) holds due to (2.5.11).

Since $V_{14} \perp\!\!\!\perp \omega_2$, we have

$$\|\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1}_{\tilde{A}_1}Q'_{41}U_{\tilde{A}_4}\lambda_{n1}/n - \omega_2\|_2^2 = \|\mathbb{D}^c_{\tilde{A}_1}\Omega^{-1/2}_{\tilde{A}_1}V_{14}/\sqrt{n} - \omega_2\|_2^2 = \|V_{14}\|_2^2 + \|\omega_2\|_2^2$$

so that

$$\left(\frac{\lambda_{n1}}{n}U'_{\tilde{A}_4}Q_{41}\Omega^{-1}_{\tilde{A}_1}\mathbb{D}^{c\prime}_{\tilde{A}_1} - \omega'_2\right)\pi_n = (\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}(u'\pi_n)$$

using the definition of u. Then, this implies

$$\begin{aligned}\|\omega_2\|_2^2 &\leq \lambda_{n1}\beta'_{\tilde{A}_2}D_{\tilde{A}_2} - \frac{\lambda_{n1}}{n}U'_{\tilde{A}_1}\Omega_{\tilde{A}_1}^{-1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^{c}_{\tilde{A}_2}\beta_{\tilde{A}_2} \\ &\quad +(\|V_{14}\|_2^2+\|\omega_2\|_2^2)^{1/2}(u'\pi_n) - \frac{\lambda_{n1}}{n}U'_{\tilde{A}_4}Q_{41}\Omega_{\tilde{A}_1}^{-1}\mathbb{D}^{c\prime}_{\tilde{A}_1}\pi_n. \qquad (2.5.24)\end{aligned}$$

Combining (2.5.21) and (2.5.24), we have

$$\begin{aligned}&V'_{14}V_{11}+\|\omega_2\|_2^2 \\ &\leq \lambda_{n1}\sum_{k\in\tilde{A}_4}\|\beta_k\|_2 - \frac{\lambda_{n1}U'_{\tilde{A}_3}Q_{31}\Omega_{\tilde{A}_1}^{-1}(\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^{c}_{\tilde{A}_2})\beta_{\tilde{A}_2}}{n} + \lambda_{n1}\beta'_{\tilde{A}_2}D_{\tilde{A}_2} \\ &\quad +(\|V_{14}\|_2^2+\|\omega_2\|_2^2)^{1/2}|u'\pi_n| \\ &= \lambda_{n1}\sum_{k\in\tilde{A}_4}\|\beta_k\|_2 - V'_{13}\Omega_{\tilde{A}_1}^{-1/2}(\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^{c}_{\tilde{A}_2})\beta_{\tilde{A}_2}/\sqrt{n} + \lambda_{n1}\beta'_{\tilde{A}_2}D_{\tilde{A}_2} \\ &\quad +2\left((\|V_{14}\|_2^2+\|\omega_2\|_2^2)^{1/2}/2\right)|u'\pi_n| \\ &\leq \lambda_{n1}\sum_{k\in\tilde{A}_4}\|\beta_k\|_2 + \|V_{13}\|_2\|\Omega_{\tilde{A}_1}^{-1/2}(\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^{c}_{\tilde{A}_2})\beta_{\tilde{A}_2}\|_2/\sqrt{n} + \lambda_{n1}\sum_{k\in\tilde{A}_2}\|\beta_k\|_2 \\ &\quad +\frac{(\|V_{14}\|_2^2+\|\omega_2\|_2^2)}{4}+|u'\pi_n|^2, \qquad (2.5.25)\end{aligned}$$

where the inequality is by the Cauchy-Schwarz inequality, triangle inequality and $2ab \leq a^2+b^2$. Then, by (2.5.22) and (2.5.25),

$$\begin{aligned}\|V_{14}\|_2^2+\|\omega_2\|_2^2 &\leq V'_{14}V_{11}+\|V_{14}\|_2\|V_{13}\|_2+\|\omega_2\|_2^2 \\ &\leq \lambda_{n1}\sum_{k\in\tilde{A}_4}\|\beta_k\|_2 + \|V_{13}\|_2\|\Omega_{\tilde{A}_1}^{-1/2}(\mathbb{D}^{c\prime}_{\tilde{A}_1}\mathbb{D}^{c}_{\tilde{A}_2})\beta_{\tilde{A}_2}\|_2/\sqrt{n} \\ &+ \lambda_{n1}\sum_{k\in\tilde{A}_2}\|\beta_k\|_2 + \frac{(\|V_{14}\|_2^2+\|\omega_2\|_2^2)}{4}+|u'\pi_n|^2 \\ &+ \|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}. \qquad (2.5.26)\end{aligned}$$

Since $\{(\mathbb{D}^c,\pi_n)\in\Omega_{q_1}\}$ implies $|u'\pi_n|^2 \leq (x_{q_1})^2 \leq \frac{(q_1 m_n \vee m_n)\lambda_{n1}^2}{4n\rho_{\max}(\Omega_{\tilde{A}_1})} = \frac{q_1 m_n \lambda_{n1}^2}{4n\rho_{\max}(\Omega_{\tilde{A}_1})} = \frac{1}{4}\frac{q_1}{q^*}B_1$, we can show $|u'\pi_n|^2 \leq \frac{1}{4}\frac{q_1}{q^*}B_1 \leq \frac{1}{4}(\|V_{14}\|_2^2+B_1)$ by using (2.5.18). Along with $\|V_{13}\|_2 \leq$

$\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}$, (2.5.26) becomes

$$
\begin{aligned}
&\leq \lambda_{n1}\sum_{k\in\tilde{A}_4}\|\beta_k\|_2+\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\|\Omega_{\tilde{A}_1}^{-1/2}\mathbb{D}_{\tilde{A}_1}^{c\prime}\mathbb{D}_{\tilde{A}_2}^{c}\beta_{\tilde{A}_2}\|_2/\sqrt{n}+\lambda_{n1}\sum_{k\in\tilde{A}_2}\|\beta_k\|_2\\
&+\frac{(\|V_{14}\|_2^2+\|\omega_2\|_2^2)}{4}+\frac{(\|V_{14}\|_2^2+B_1)}{4}+\|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\\
&=\lambda_{n1}\sum_{k\in\tilde{A}_5}\|\beta_k\|_2+\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\|P_{\tilde{A}_1}^{1/2}\mathbb{D}_{\tilde{A}_2}^{c}\beta_{\tilde{A}_2}\|_2+\lambda_{n1}\sum_{k\in\tilde{A}_4\cup\tilde{A}_6}\|\beta_k\|_2\\
&+\frac{\|V_{14}\|_2^2}{2}+\frac{\|\omega_2\|_2^2}{4}+\frac{B_1}{4}+\|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\\
&\leq \lambda_{n1}\sum_{k\in\tilde{A}_5}\|\beta_k\|_2+\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\|P_{\tilde{A}_1}^{1/2}\mathbb{D}_{\tilde{A}_2}^{c}\beta_{\tilde{A}_2}\|_2+\lambda_{n1}\eta_1\\
&+\frac{\|V_{14}\|_2^2}{2}+\frac{\|\omega_2\|_2^2}{4}+\frac{B_1}{4}+\|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}},
\end{aligned}
$$

where the equality comes from $\tilde{A}_2=\tilde{A}_5\cup\tilde{A}_6$ with $P_{\tilde{A}_1}^{1/2}=\Omega_{\tilde{A}_1}^{-1/2}\mathbb{D}_{\tilde{A}_1}^{c\prime}/\sqrt{n}$ and the last inequality is due to GSC. This result gives

$$
\begin{aligned}
\frac{1}{2}\|V_{14}\|_2^2+\frac{3}{4}\|\omega_2\|_2^2 \leq \lambda_{n1}\sum_{k\in\tilde{A}_5}\|\beta_k\|_2+\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\|P_{\tilde{A}_1}^{1/2}\mathbb{D}_{\tilde{A}_2}^{c}\beta_{\tilde{A}_2}\|_2+\lambda_{n1}\eta_1\\
+\frac{B_1}{4}+\|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}} \qquad (2.5.27)
\end{aligned}
$$

from which we have

$$
\begin{aligned}
\|V_{14}\|_2^2 \quad &\leq 2\left(\frac{1}{2}\|V_{14}\|_2^2+\frac{3}{4}\|\omega_2\|_2^2\right)\\
&\leq 2\lambda_{n1}\sum_{k\in\tilde{A}_5}\|\beta_k\|_2+2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}\|P_{\tilde{A}_1}^{1/2}\mathbb{D}_{\tilde{A}_2}^{c}\beta_{\tilde{A}_2}\|_2+2\lambda_{n1}\eta_1\\
&\quad+\frac{B_1}{2}+2\|V_{14}\|_2\lambda_{n1}\sqrt{\frac{m_n|\tilde{A}_3|}{n\rho_{\min}(\Omega_{\tilde{A}_1})}}.
\end{aligned}
$$

Note that the largest possible $\tilde{A}_1$ contains all the "large" $\|\beta_j\|_2$ then $\tilde{A}_5 = \phi$ so that $\tilde{A}_2 = \tilde{A}_6$, $\sum_{k\in\tilde{A}_5}\|\beta_k\|_2 = 0$ and $P^{1/2}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_2}\beta_{\tilde{A}_2} = P^{1/2}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_6}\beta_{\tilde{A}_6}$. Finally using $\|P^{1/2}_{\tilde{A}_1}\mathbb{D}^c_{\tilde{A}_6}\beta_{\tilde{A}_6}\|_2 \leq \max_{A\subset\tilde{A}_0}\|\sum_{k\in A}\mathbb{D}^c_k\beta_k\|_2$ since $\tilde{A}_6 \subset \tilde{A}_0$, we have the following inequality.

$$\|V_{14}\|_2^2 \leq 2\eta_2\sqrt{B_2} + 2\lambda_{n1}\eta_1 + \frac{B_1}{2} + 2\sqrt{B_2}\|V_{14}\|_2, \tag{2.5.28}$$

where $\eta_2 = \max_{A\subset\tilde{A}_0}\|\sum_{k\in A}\mathbb{D}^c_k\beta_k\|_2$ and $B_2 = \dfrac{\lambda^2_{n1}m_nq^*}{n\rho_{\min}\left(\Omega_{\tilde{A}_1}\right)}$. Hence after using the fact $2\sqrt{B_2}\|V_{14}\|_2 = 2\left(\sqrt{2B_2}\right)\left(\|V_{14}\|_2/\sqrt{2}\right) \leq 2B_2 + \|V_{14}\|_2^2/2$, we obtain an upper bound for $\|V_{14}\|_2^2$ from (2.5.28),

$$\|V_{14}\|_2^2 \leq B_1 + 4\lambda_{n1}\eta_1 + 4\eta_2\sqrt{B_2} + 4B_2$$

which, when combined with (2.5.19), implies

$$\left|\tilde{A}_\beta\right| \leq M_1q^*,$$

where $M_1 = M_1(\lambda_{n1}) = 2 + 4r_1 + 4r_2\sqrt{C_{12}} + 4C_{12}$ with

$$r_1 = r_1(\lambda_{n1}) = \left(\frac{c_2\eta_1 n}{q^*m_n\lambda_{n1}}\right), r_2 = r_2(\lambda_{n1}) = \left(\frac{c_2\eta_2^2 n}{q^*m_n\lambda^2_{n1}}\right)^{1/2} \text{ and } C_{12} = \frac{c_2}{c_1}.$$

Note that $M_1(\lambda_{n1})$ is a decreasing function in λ_{n1}.

If $\eta_1 = 0$ which is referred to as a narrow-sense sparsity condition, then $r_1 = r_2 = 0$ and hence $M_1(\lambda_{n1}) = 2 + 4C_{12} < \infty$. Note that since we are assuming $\lambda_0 < \infty$, we implicitly assume that $(2 + 4C_{12})q^* + 1 \leq q_0$ holds. In general, as long as η_1 and η_2 satisfy that $\eta_1 \leq \left(\frac{C_1q^*}{c_2}\right)\left(\frac{m_n\lambda_{n1}}{n}\right)$ and $\eta_2^2 \leq \left(\frac{C_2q^*}{c_2}\right)\left(\frac{m_n\lambda^2_{n1}}{n}\right)$ for some finite C_1 and C_2, we will have $r_1 \leq C_1$ and $r_2 \leq C_2$, which gives $M_1(\lambda_{n1}) \leq (2 + 4C_1 + 4C_2\sqrt{C_{12}} + 4C_{12}) < \infty$. Thus, we complete the proof.

□

PROOF OF THEOREM 1

Part (a): Since $\hat{\beta}_{gL}$ is the group LASSO estimate by minimizing $Q_{n1}(\beta, \lambda_{n1})$, for any β,

$$\|Z^c - \mathbb{D}^c \hat{\beta}_{gL}\|_2^2 + \lambda_{n1}\|\hat{\beta}_{gL}\|_{2,1,\mathbf{1}} \;\leq\; \|Z^c - \mathbb{D}^c \beta\|_2^2 + \lambda_{n1}\|\beta\|_{2,1,\mathbf{1}}. \tag{2.5.29}$$

Let $A_2 = \{j : \|\beta_j\|_2 > 0 \text{ or } \|\hat{\beta}_{gL,j}\|_2 > 0\}$, where $\beta_{A_2} = (\beta_j, j \in A_2)$ and $\hat{\beta}_{gL,A_2} = (\hat{\beta}_{gL,j}, j \in A_2)$. Recall that $\|\beta_{A_2}\|_{2,1,\mathbf{1}} = \sum_{j\in A_2}\|\beta_j\|_2$ and $\|\hat{\beta}_{gL,A_2}\|_{2,1,\mathbf{1}} = \sum_{j\in A_2}\|\hat{\beta}_{gL,j}\|_2$. By (2.5.29), we have

$$\|Z^c - \mathbb{D}^c_{A_2}\hat{\beta}_{gL,A_2}\|_2^2 + \lambda_{n1}\|\hat{\beta}_{gL,A_2}\|_{2,1,\mathbf{1}} \;\leq\; \|Z^c - \mathbb{D}^c_{A_2}\beta_{A_2}\|_2^2 + \lambda_{n1}\|\beta_{A_2}\|_{2,1,\mathbf{1}}. \tag{2.5.30}$$

Let $\vartheta_{A_2} = Z^c - \mathbb{D}^c_{A_2}\beta_{A_2}$ and $\zeta_{A_2} = \mathbb{D}^c_{A_2}(\hat{\beta}_{gL,A_2} - \beta_{A_2})$. Using the fact that $\|a-b\|_2^2 = \|a\|_2^2 - 2a'b + \|b\|_2^2$ and $Z^c - \mathbb{D}^c_{A_2}\hat{\beta}_{gL,A_2} = Z^c - \mathbb{D}^c_{A_2}\beta_{A_2} - \mathbb{D}^c_{A_2}(\hat{\beta}_{gL,A_2} - \beta_{A_2})$, we can rewrite (2.5.30) such that

$$\|\zeta_{A_2}\|_2^2 - 2\vartheta'_{A_2}\zeta_{A_2} \leq \lambda_{n1}\|\beta_{A_2}\|_{2,1,\mathbf{1}} - \lambda_{n1}\|\hat{\beta}_{gL,A_2}\|_{2,1,\mathbf{1}} \leq \lambda_{n1}\sqrt{|A_2|}\,\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2. \tag{2.5.31}$$

Subsequent steps will be to bound $\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2^2$. First, we have

$$2|\vartheta'_{A_2}\zeta_{A_2}| \leq 2\|\vartheta^*_{A_2}\|_2\|\zeta_{A_2}\|_2 = 2\left(\sqrt{2}\|\vartheta^*_{A_2}\|_2\right)\left(\frac{\|\zeta_{A_2}\|_2}{\sqrt{2}}\right) \leq 2\|\vartheta^*_{A_2}\|_2^2 + \frac{\|\zeta_{A_2}\|_2^2}{2}, \tag{2.5.32}$$

where the last inequality is based on the fact that, $2ab \leq a^2 + b^2$ and $\vartheta^*_{A_2}$ is the projection of ϑ_{A_2} to the span of $\mathbb{D}^c_{A_2}$. Now by combining (2.5.31) and (2.5.32), we have

$$\|\zeta_{A_2}\|_2^2 \leq 4\|\vartheta^*_{A_2}\|_2^2 + 2\lambda_{n1}\sqrt{|A_2|}\,\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2. \tag{2.5.33}$$

On the other hand, we have

$$\begin{aligned}\|\zeta_{A_2}\|_2^2 &= (\hat{\beta}_{gL,A_2} - \beta_{A_2})'\mathbb{D}^{c\prime}_{A_2}\mathbb{D}^c_{A_2}(\hat{\beta}_{gL,A_2} - \beta_{A_2}) \geq n\rho_{\min}\left(\frac{\mathbb{D}^{c\prime}_{A_2}\mathbb{D}^c_{A_2}}{n}\right)\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2^2 \\ &= n\rho_{\min}(\Omega_{A_2})\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2^2 \geq nc_1h_n\|\hat{\beta}_{gL,A_2} - \beta_{A_2}\|_2^2,\end{aligned} \tag{2.5.34}$$

where the last inequality is from Lemma 2. Combining (2.5.33) and (2.5.34), we have

$$
\begin{aligned}
&nc_1h_n\|\hat{\beta}_{gL,A_2}-\beta_{A_2}\|_2^2 \\
&\quad\le 4\|\vartheta^*_{A_2}\|_2^2+2\lambda_{n1}\sqrt{|A_2|}\|\beta_{A_2}-\hat{\beta}_{gL,A_2}\|_2 \\
&\quad= 4\|\vartheta^*_{A_2}\|_2^2+2\left(\frac{\lambda_{n1}\sqrt{2|A_2|}}{\sqrt{nc_1h_n}}\right)\left(\|\beta_{A_2}-\hat{\beta}_{gL,A_2}\|_2\sqrt{\frac{nc_1h_n}{2}}\right) \\
&\quad\le 4\|\vartheta^*_{A_2}\|_2^2+\frac{\lambda_{n1}^2 2|A_2|}{nc_1h_n}+\|\beta_{A_2}-\hat{\beta}_{gL,A_2}\|_2^2\frac{nc_1h_n}{2} \quad \text{so that} \\
&\|\hat{\beta}_{gL,A_2}-\beta_{A_2}\|_2^2 \le \frac{8\|\vartheta^*_{A_2}\|_2^2}{nc_1h_n}+\frac{4\lambda_{n1}^2|A_2|}{n^2c_1^2h_n^2}. \qquad (2.5.35)
\end{aligned}
$$

Let $\vartheta(s)$ be the entry of ϑ_{A_2}. Then, we can rewrite $\vartheta(s)=\epsilon(s)+(\mu-\bar{Y})+f_0(X(s))-f_{nA_2}(X(s))$, where $f_0(X(s))=\sum_{j=1}^J f_j(X_j(s))$ and $f_{nA_2}(X(s))=\sum_{j\in A_2} f_{nj}(X_j(s))$. Note that $|\mu-\bar{Y}|^2=O_p(n^{-1})$. Let $\epsilon^*_{A_2}$ be the projection of ϵ_n to the span of $\mathbb{D}^c_{A_2}$, that is, $\epsilon^*_{A_2}=(\mathbb{D}^{c\prime}_{A_2}\mathbb{D}^c_{A_2})^{-1/2}\mathbb{D}^{c\prime}_{A_2}\epsilon_n$. Then, we have

$$
\begin{aligned}
\|\vartheta^*_{A_2}\|_2^2 \quad &\le \|\epsilon^*_{A_2}\|_2^2+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \qquad (2.5.36)\\
&= \|(\mathbb{D}^{c\prime}_{A_2}\mathbb{D}^c_{A_2})^{-1/2}\mathbb{D}^{c\prime}_{A_2}\epsilon_n\|_2^2+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \\
&\le \frac{\|\mathbb{D}^{c\prime}_{A_2}\epsilon_n\|_2^2}{nc_1h_n}+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \qquad \text{(By Lemma 2)} \\
&\le \frac{1}{nc_1h_n}\max_{A:|A|\le|A_2|}\|\mathbb{D}^{c\prime}_{A}\epsilon_n\|_2^2+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \\
&\le \frac{1}{nc_1h_n}|A_2|\,m_n\max_{1\le j\le J,1\le l\le m_n}\left|\mathbb{D}^{c\prime}_{jl}\epsilon_n\right|^2+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \\
&= \frac{1}{c_1h_n}|A_2|\max_{1\le j\le J,1\le l\le m_n}\left|\left(\frac{m_n}{n}\right)^{1/2}a'_{jl}L\epsilon_n\right|^2+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}) \\
&= O_p\left(\frac{|A_2|\rho^2_{\max}(L)\,m_n\log(Jm_n)}{c_1h_n}\right)+O_p(n^\alpha 1+|A_2|\,nm_n^{-2\tau}), \qquad (2.5.37)
\end{aligned}
$$

where the last equality is by By Lemma 3 using $M_n=L$. The part (a) follows by combining (2.5.35) and (2.5.37) since $|A_2|$ is bounded by Lemma 4.

Part (b): If $\frac{m_n^2\lambda_{n1}^2}{n^2}\longrightarrow 0$ then, by the condition $\lambda_{n1}>C\rho_{\max}(L)\sqrt{n^{1+\alpha}m_n\log(Jm_n)}$, we have $\frac{\rho^2_{\max}(L)m_n^3\log(Jm_n)}{n^{1-\alpha}}\longrightarrow 0$, also note that, $1=\rho_{\max}(\mathbb{I})=\rho_{\max}(\Sigma_W^{-1}\Sigma_W)\le\rho_{\max}(\Sigma_W^{-1})$

$\rho_{\max}(\Sigma_W) \leq \rho^2_{\max}(L)\rho_{\max}(\Sigma_W)$ therefore $\rho^2_{\max}(L) \geq \rho_{\min}\left(\Sigma_W^{-1}\right)$ and hence,

$$\begin{aligned}\left\{\frac{\rho^2_{\max}(L)\, m_n^3 \log(Jm_n)}{n^{1-\alpha}}\right\} &= \left\{\rho^2_{\max}(L)\, m_n^2 \log(Jm_n)\right\}\frac{m_n}{n^{1-\alpha}} \\ &\geq \left\{\rho_{\min}\left(\Sigma_W^{-1}\right) m_n^2 \log(Jm_n)\right\}\frac{m_n}{n^{1-\alpha}} \\ &\geq \left\{C m_n^2 \log(Jm_n)\right\}\frac{m_n}{n^{1-\alpha}}, \end{aligned} \tag{2.5.38}$$

where C is a generic constant. Since L.H.S of (2.5.38) goes to 0 and $m_n^2 \log(Jm_n) \longrightarrow \infty$, $\frac{m_n}{n^{1-\alpha}} \longrightarrow 0$. Similarly,

$$\left\{\frac{\rho^2_{\max}(L)\, m_n^3 \log(Jm_n)}{n^{1-\alpha}}\right\} \tag{2.5.39}$$

$$= \left\{\frac{\rho^2_{\max}(L)\, m_n^{2\tau+2} \log(Jm_n)}{n^{1-\alpha}}\right\}\frac{1}{m_n^{2\tau-1}} = \left\{\frac{\rho_{\min}\left(\Sigma_W^{-1}\right) n^{\frac{2\tau+2}{2\tau-1}} \log(Jm_n)}{n^{1-\alpha}}\right\}\frac{1}{m_n^{2\tau-1}}$$

$$\geq \left\{\rho_{\min}\left(\Sigma_W^{-1}\right) n^{\alpha} \log(Jm_n)\right\}\frac{1}{m_n^{2\tau-1}} \geq \left\{C n^{\alpha} \log(Jm_n)\right\}\frac{1}{m_n^{2\tau-1}} \tag{2.5.40}$$

Since L.H.S of (2.5.40) goes to 0 and $n^{\alpha}\log(Jm_n) \longrightarrow \infty$, $\frac{1}{m_n^{2\tau-1}} \longrightarrow 0$. Thus, we have part (b) of Theorem 1.

PROOF OF THEOREM 2

The part (a) is from the fact that

$$c_* m_n^{-1} \|\hat{\beta}_{gL,j} - \beta_j\|_2 \leq \|\hat{f}_{gL,j} - f_j\|_2 \leq c^* m_n^{-1} \|\hat{\beta}_{gL,j} - \beta_j\|_2$$

for some $c_*, c^* > 0$. The part (b) is from the part (a).

PROOF OF THEOREM 3

Part (a): Recall that, by the KKT conditions, a necessary and sufficient condition for $\hat{\beta}_{AgL}$ is

$$\begin{cases} \mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}^c \hat{\beta}_{AgL}) = \lambda_{n2}\eta_{nj}\dfrac{\hat{\beta}_{AgL,j}}{2\|\hat{\beta}_{AgL,j}\|_2}, \text{ when } \|\hat{\beta}_{AgL,j}\|_2 > 0, \\ \|\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}^c \hat{\beta}_{AgL})\|_2 \leq \lambda_{n2}\eta_{nj}/2, \text{ when } \|\hat{\beta}_{AgL,j}\|_2 = 0. \end{cases} \tag{2.5.41}$$

Let $A_{**} = A_* \cap \{j : \|\hat{\beta}_{AgL,j}\|_2 > 0\}$. Define

$$\hat{\beta}_{A_{**}} = (\mathbb{D}^c_{A_{**}}{}'\mathbb{D}^c_{A_{**}})^{-1}(\mathbb{D}^c_{A_{**}}{}'Z^c - \lambda_{n2} v_n), \tag{2.5.42}$$

where $v_n = (v_{nj}, j \in A_{**})$ with $v_{nj} = \eta_{nj}\hat{\beta}_j/(2\|\hat{\beta}_j\|_2)$. Then, we have $\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}}) = \lambda_{n2}\eta_{nj}\dfrac{\hat{\beta}_j}{2\|\hat{\beta}_j\|_2}$, for $j \in A_{**}$. If we assume $\|\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}})\|_2 \leq \lambda_{n2}\eta_{nj}/2$ for all $j \notin A_{**}$, then (2.5.41) holds for $(\hat{\beta}'_{A_{**}}, 0')$, so that $\hat{\beta}_{AgL} = (\hat{\beta}'_{A_{**}}, 0')$ since $\mathbb{D}^c\hat{\beta} = \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}}$. If $\|\beta_j\|_2 - \|\hat{\beta}_j\|_2 < \|\beta_j\|_2$ for all $j \in A_{**}$, then $\hat{\beta}_{A_{**}} =_0 \beta_{A_{**}}$ so that we have $\hat{\beta}_{AgL} =_0 \beta$.

Therefore, we can have the following inequalities,

$$\begin{aligned}
\mathbb{P}(\hat{\beta}_{AgL} \neq_0 \beta) \quad & \leq \mathbb{P}(\|\beta_j\|_2 - \|\hat{\beta}_j\|_2 \geq \|\beta_j\|_2, \exists j \in A_{**}) \\
& + \mathbb{P}(\|\mathbb{D}_j^{c\prime}(Z - \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}})\|_2 > \lambda_{n2}\eta_{nj}/2, \exists j \notin A_{**}) \\
& \leq \mathbb{P}(\|\hat{\beta}_j - \beta_j\|_2 \geq \|\beta_j\|_2, \exists j \in A_{**}) \\
& + \mathbb{P}(\|\mathbb{D}_j^{c\prime}(Z - \mathbb{D}_{A_*}^c \hat{\beta}_{A_{**}})\|_2 > \lambda_{n2}\eta_{nj}/2, \exists j \notin A_{**}), \qquad (2.5.43)
\end{aligned}$$

where the last inequality is from $\|\beta_j\|_2 > 0$ for $j \in A_{**}$. First, we show

$$\mathbb{P}(\|\hat{\beta}_j - \beta_j\|_2 \geq \|\beta_j\|_2, \exists j \in A_{**}) \to 0. \qquad (2.5.44)$$

To show (2.5.44), it is sufficient to show that $\max_{j \in A_{**}} \|\hat{\beta}_j - \beta_j\|_2 \to 0$ in probability since $\|\beta_j\|_2 > 0$ for $j \in A_{**}$. Define $T_{nj} = (\mathbb{O}_{m_n}, \ldots, \mathbb{O}_{m_n}, \mathbb{I}_{m_n}, \mathbb{O}_{m_n}, \ldots, \mathbb{O}_{m_n})$ be a $m_n \times qm_n$ matrix with $\mathbb{I}_{m_n}$ is in the j^{th} block, where $\mathbb{O}_{m_n}$ be $m_n \times m_n$ matrix of zeros and $\mathbb{I}_{m_n}$ be an $m_n \times m_n$ identity matrix. From (2.5.42), $\hat{\beta}_{A_{**}} - \beta_{A_{**}} = n^{-1}\Omega_{A_{**}}^{-1}(\mathbb{D}_{A_{**}}^{c\prime}\epsilon_n + \mathbb{D}_{A_{**}}^{c\prime}\theta_n - \lambda_{n2}v_n)$. Thus, if $j \in A_{**}$, we have $\hat{\beta}_j - \beta_j = n^{-1}T_{nj}\Omega_{A_{**}}^{-1}(\mathbb{D}_{A_{**}}^{c\prime}\epsilon_n + \mathbb{D}_{A_{**}}^{c\prime}\theta_n - \lambda_{n2}v_n)$. By triangle inequality,

$$\begin{aligned}
& \|\hat{\beta}_j - \beta_j\|_2 \\
& \leq \frac{\|T_{nj}\Omega_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{c\prime}\epsilon_n\|_2}{n} + \frac{\|T_{nj}\Omega_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{c\prime}\theta_n\|_2}{n} + \frac{\lambda_{n2}\|T_{nj}\Omega_{A_{**}}^{-1}v_n\|_2}{n}. \qquad (2.5.45)
\end{aligned}$$

We show each term on the right hand side in (2.5.45) goes to zero in probability. For the

first term,

$$\begin{aligned}
&\max_{j\in A_{**}} n^{-1}\|T_{nj}\Omega_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{c\prime}\epsilon_n\|_2 \\
&\leq \frac{1}{n\rho_{\max}(\Omega_{A_{**}})}\|\mathbb{D}_{A_{**}}^{c\prime}\epsilon_n\|_2 \leq \frac{\sqrt{|A_{**}|}}{n^{1/2}\rho_{\max}(\Omega_{A_{**}})}\sqrt{\max_{\substack{j\in A_{**}\\ 1\leq l\leq m_n}}\frac{m_n}{n}|\mathbb{D}_{jl}^{c}{}'\epsilon_n|^2} \\
&= O_p\left(\sqrt{\frac{\rho_{\max}^2(L)m_n^3\log(|A_{**}|m_n)}{n^{1-\alpha}}}\right) \longrightarrow 0 \qquad (2.5.46)
\end{aligned}$$

where the last equality holds by Lemma 3 and (2.5.46) holds by assumptions (H 6) and (H 7). For the second term,

$$\begin{aligned}
\max_{j\in A_{**}} n^{-1}\|T_{nj}\Omega_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{c\prime}\theta_n\|_2 \qquad &\leq n^{-1/2}\|\Omega_{A_{**}}^{-1}\|_2\|n^{-1}\mathbb{D}_{A_{**}}^{c\prime}\mathbb{D}_{A_{**}}^{c}\|_2^{1/2}\|\theta_n\|_2 \\
&\leq n^{-1/2}\rho_{\min}^{-1}(\Omega_{A_{**}})\rho_{\max}^{1/2}(\Omega_{A_{**}})O_p(n^{1/(4\tau+2)}) \\
&= O_p\left(n^{1/(2\tau+1)-1/2}\right) \longrightarrow 0, \qquad (2.5.47)
\end{aligned}$$

where (2.5.47) holds by assumption (H 6). For the third term, we first find an upper bound for $\|v_n\|_2^2$,

$$\begin{aligned}
\|v_n\|_2^2 &= \frac{1}{2}\sum_{j\in A_{**}}\eta_{nj}^2 = \frac{1}{2}\sum_{j\in A_{**}}\|\hat{\beta}_{gL,j}\|_2^{-2} = \frac{1}{2}\sum_{j\in A_{**}}\frac{\|\beta_j\|_2^2 - \|\hat{\beta}_{gL,j}\|_2^2 + \|\hat{\beta}_{gL,j}\|_2^2}{\|\hat{\beta}_{gL,j}\|_2^2\|\beta_j\|_2^2} \\
&= \frac{1}{2}\sum_{j\in A_{**}}\frac{\|\beta_j\|_2^2 - \|\hat{\beta}_{gL,j}\|_2^2}{\|\hat{\beta}_{gL,j}\|_2^2\|\beta_j\|_2^2} + \frac{1}{2}\sum_{j\in A_{**}}\|\beta_j\|_2^{-2} \leq Ck_b^{-2}b_{n1}^{-4}\|\hat{\beta}_{gL,A_{**}} - \beta_{A_{**}}\|_2^2 + qb_{n1}^{-2} \\
&= O_p(k_b^{-2}b_{n1}^{-4}r_n^{-1} + qb_{n1}^{-2}) = O_p(k_n^2), \qquad (2.5.48)
\end{aligned}$$

where C is a generic constant. Then, we have

$$\begin{aligned}
&\max_{j\in A_{**}} n^{-1}\lambda_{n2}\|T_{nj}\Omega_{A_{**}}^{-1}v_n\|_2 \leq n^{-1}\lambda_{n2}\rho_{\min}^{-1}(\Omega_{A_{**}})\|v_n\|_2 = O_p(n^{-1}\lambda_{n2}\rho_{\min}^{-1}(\Omega_{A_{**}})k_n) \\
&= O_p(n^{-1}\lambda_{n2}(r_n^{-1/2} + m_n^{1/2}) = O_p\left(\frac{\lambda_{n2}m_n^{1/2}}{n}\right) \longrightarrow 0, \qquad (2.5.49)
\end{aligned}$$

where (2.5.49) is implied by assumption (K 2). Therefore by combining (2.5.46), (2.5.47) and (2.5.49), we have (2.5.44). Now, we show

$$\mathbb{P}(\|\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_*}^{c}\hat{\beta}_{A_{**}})\|_2 > \lambda_{n2}\eta_{nj}/2, \exists j \notin A_{**}) \to 0. \qquad (2.5.50)$$

As $\eta_{nj} = \|\hat{\beta}_{gL,j}\|_2^{-1} = O_p(r_n)$ for $j \notin A_{**}$, instead of (2.5.50) it is sufficient to show,

$$\mathbb{P}(\|\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}})\|_2 > \lambda_{n2} r_n/2, \exists j \notin A_{**}) \to 0. \tag{2.5.51}$$

For $j \notin A_{**}$,

$$\begin{aligned}
&\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c \hat{\beta}_{A_{**}}) \\
&\quad = \mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c ({\mathbb{D}_{A_{**}}^c}' \mathbb{D}_{A_{**}}^c)^{-1} {\mathbb{D}_{A_{**}}^c}' Z^c + \lambda_{n2} n^{-1} \mathbb{D}_{A_{**}}^c \Omega_{A_{**}}^{-1} v_n) \\
&\quad = \mathbb{D}_j^{c\prime} H_n Z^c + \lambda_{n2} n^{-1} \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}}^c \Omega_{A_{**}}^{-1} v_n \\
&\quad = \mathbb{D}_j^{c\prime} H_n \mathbb{D}^c \beta + \mathbb{D}_j^{c\prime} H_n \epsilon_n + \mathbb{D}_j^{c\prime} H_n \theta_n + \lambda_{n2} n^{-1} \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}}^c \Omega_{A_{**}}^{-1} v_n \\
&\quad = \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}^c}^c \beta_{A_{**}^c} + \mathbb{D}_j^{c\prime} H_n \epsilon_n + \mathbb{D}_j^{c\prime} H_n \theta_n + \lambda_{n2} n^{-1} \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}}^c \Omega_{A_{**}}^{-1} v_n \\
&\quad = \mathbb{D}_j^{c\prime} \mathbb{D}_{A_* \cap A_{**}^c}^c \beta_{A_* \cap A_{**}^c} + \mathbb{D}_j^{c\prime} H_n \epsilon_n + \mathbb{D}_j^{c\prime} H_n \theta_n \\
&\quad + \lambda_{n2} n^{-1} \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}}^c \Omega_{A_*}^{-1} v_n,
\end{aligned} \tag{2.5.52}$$

where $H_n = \mathbb{I} - P_{A_{**}}$, the first equality is by replacing $\hat{\beta}_{A_{**}}$ with the expression as in (2.5.42) and the fourth equality is because H_n is the projection matrix on to A_{**}^c.

By (2.5.52), the left hand side of (2.5.51) can be bounded above by

$$\begin{aligned}
&\mathbb{P}\left(\|\mathbb{D}_j^{c\prime}(Z^c - \mathbb{D}_{A_{**}}^c \hat{\beta}_{AgL,A_{**}})\|_2 > \lambda_{n2} r_n/2, \exists j \notin A_{**}\right) \\
&\quad \le \mathbb{P}\left(\|\mathbb{D}_j^{c\prime} \mathbb{D}_{A_* \cap A_{**}^c}^c \beta_{A_* \cap A_{**}^c}\|_2 > \lambda_{n2} r_n/8, \exists j \notin A_{**}\right) \\
&\qquad + \mathbb{P}\left(\|\mathbb{D}_j^{c\prime} H_n \epsilon_n\|_2 > \lambda_{n2} r_n/8, \exists j \notin A_{**}\right) + \mathbb{P}\left(\|\mathbb{D}_j^{c\prime} H_n \theta_n\|_2 > \lambda_{n2} r_n/8, \exists j \notin A_{**}\right) \\
&\qquad + \mathbb{P}\left(\|\lambda_{n2} n^{-1} \mathbb{D}_j^{c\prime} \mathbb{D}_{A_{**}}^c \Omega_{A_*}^{-1} v_n\|_2 > \lambda_{n2} r_n/8, \exists j \notin A_{**}\right)
\end{aligned} \tag{2.5.53}$$

so that we find upper bounds of the four terms in (2.5.53). For the first term, we have

$$\begin{aligned}
\max_{j \notin A_{**}} \|\mathbb{D}_j^{c\prime} \mathbb{D}_{A_* \cap A_{**}^c}^c \beta_{A_* \cap A_{**}^c}\|_2 \quad &\le n \max_{j \notin A_{**}} \|n^{-1/2} \mathbb{D}_j^{c\prime}\|_2 \|n^{-1/2} \mathbb{D}_{A_* \cap A_{**}^c}^c\|_2 \|\beta_{A_* \cap A_{**}^c}\|_2 \\
&= O_p(n \rho_{\max}^{1/2}(\Omega_{A_{**}^c}) \rho_{\max}^{1/2}(\Omega_{A_{**}}) m_n^{1/2}) = O_p(n m_n^{-1/2}).
\end{aligned}$$

Then, we have, for some generic constant C,

$$\begin{aligned}
&\mathbb{P}(\|\mathbb{D}_j^{c\prime}\mathbb{D}^c_{A_*\cap A^c_{**}}\beta_{A_*\cap A^c_{**}}\|_2 > \lambda_{n2}r_n/8, \exists j \notin A_{**}) \\
&\qquad \leq \mathbb{P}(\max_{j\notin A_{**}} \|\mathbb{D}_j^{c\prime}\mathbb{D}^c_{A_*\cap A^c_{**}}\beta_{A_*\cap A^c_{**}}\|_2 > C\lambda_{n2}r_n/8) \\
&\qquad \leq \mathbb{P}(nm_n^{-1/2} > C\lambda_{n2}r_n/8) \\
&\qquad = \mathbb{P}\left(\frac{nm_n^{-1/2}}{\lambda_{n2}r_n} > C/8\right) \longrightarrow 0, \qquad (2.5.54)
\end{aligned}$$

where (2.5.54) holds by assumption (K 2). For the second term, let $s_n = J - |A_{**}|$. Since $\rho_{\max}(H_n) = \rho_{\max}(\mathbb{I} - P_{A_{**}}) = 1 - \rho_{\min}(P_{A_{**}})$ and $P_{A_{**}}$ is a non-negative definite matrix, $\rho_{\max}(H_n) \leq 1$. By Lemma 3 with $M_n = LH_n$, and using the face that $\rho_{\max}(LH_n) \leq \rho_{\max}(L)\rho_{\max}(H_n)$, we have,

$$\begin{aligned}
\mathbb{E}\left(\max_{j\notin A_{**}} n^{-1/2}\|\mathbb{D}_j^{c\prime}H_n\epsilon_n\|_2\right) &= \mathbb{E}\left(\max_{j\notin A_{**}} n^{-1/2}\sqrt{\sum_{l=1}^{m_n}|\mathbb{D}_{jl}^{c}{}'H_n\epsilon_n|^2}\right) \\
&\leq \mathbb{E}\left(\max_{\substack{j\notin A_{**}\\ 1\leq l\leq m_n}} \left(\frac{m_n}{n}\right)^{1/2}|a_{jl}'LH_n\epsilon_n|\right) \\
&= O(\sqrt{\rho^2_{\max}(L)n^{\alpha}m_n log(s_n m_n)}). \qquad (2.5.55)
\end{aligned}$$

Thus, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}(\|\mathbb{D}_j^{c\prime}H_n\epsilon_n\|_2 > \lambda_{n2}r_n/8, \exists j \notin A_{**}) &\leq \mathbb{P}(\max_{j\notin A_{**}} n^{-1/2}\|\mathbb{D}_j^{c\prime}H_n\epsilon_n\|_2 > Cn^{-1/2}\lambda_{n2}r_n/8) \\
&\leq O\left(\frac{\sqrt{\rho^2_{\max}(L)n^{1+\alpha}m_n log(s_n m_n)}}{C\lambda_{n2}r_n}\right) \\
&\longrightarrow 0, \qquad (2.5.56)
\end{aligned}$$

where C is a generic constant and (2.5.56) holds by assumption (K 2). For the third term,

$$\begin{aligned}
\max_{j\notin A_{**}} \|\mathbb{D}_j^{c\prime}H_n\theta_n\|_2 &\leq n^{1/2}\max_{j\notin A_{**}} \|n^{-1}\mathbb{D}_j^{c\prime}\mathbb{D}_j^c\|_2^{1/2}\|H_n\|_2\|\theta_n\|_2 \\
&= O(n\rho^{1/2}_{\max}(\Omega_{A^c_{**}})m_n^{-\tau}) = O(nm_n^{-\tau-1/2}).
\end{aligned}$$

Therefore, for some generic constant C,

$$\begin{aligned}\mathbb{P}(\|\mathbb{D}_j^{c\prime}H_n\theta_n\|_2 > \lambda_{n2}r_n/6, \exists j \notin A_{**}) &\leq \mathbb{P}(\max_{j\notin A_{**}} \|\mathbb{D}_j^{c\prime}H_n\theta_n\|_2 > C\lambda_{n2}r_n/8) \\ &\leq \mathbb{P}(nm_n^{-\tau-1/2} > C\lambda_{n2}r_n/8) = \mathbb{P}\left(\frac{n}{\lambda_{n2}r_n m_n^{(2\tau+1)/2}} > C/8\right) \\ &\longrightarrow 0, \end{aligned} \tag{2.5.57}$$

where (2.5.57) is implied by assumption (K 2). Finally, using (2.5.48) we have

$$\begin{aligned}&\max_{j\notin A_{**}} \|\lambda_{n2}n^{-1}\mathbb{D}_j^{c\prime}\mathbb{D}_{A_{**}}^{c}\Omega_{A_{**}}^{-1}v_n\|_2 \\ &\leq \lambda_{n2}\max_{j\notin A_{**}} \|n^{-1/2}\mathbb{D}_j^{c\prime}\|_2\|n^{-1/2}\mathbb{D}_{A_{**}}^{c}\Omega_{A_{**}}^{-1/2}\|_2\|\Omega_{A_{**}}^{-1/2}\|_2\|v_n\|_2 \\ &= O_p(\lambda_{n2}\rho_{\max}^{1/2}(\Omega_{A_{**}^c})\rho_{\min}^{-1/2}(\Omega_{A_{**}})k_n) \\ &= O_p\left(\lambda_{n2}(m_n^{-1}r_n^{-1/2} + m_n^{-1/2})\right).\end{aligned}$$

Then, we have, for some generic constant C,

$$\begin{aligned}&\mathbb{P}(\|\lambda_{n2}n^{-1}\mathbb{D}_j^{c\prime}\mathbb{D}_{A_{**}}^{c}\Omega_{A_{**}}^{-1}v_n\|_2 > \lambda_{n2}r_n/8, \exists j \notin A_{**}) \\ &\quad\leq \mathbb{P}(\max_{j\notin A_{**}} \|\lambda_{n2}n^{-1}\mathbb{D}_j^{c\prime}\mathbb{D}_{A_{**}}^{c}\Omega_{A_{**}}^{-1}v_n\|_2 > C\lambda_{n2}r_n/8) \\ &\quad\leq \mathbb{P}\left(\lambda_{n2}(m_n^{-1}r_n^{-1/2} + m_n^{-1/2}) > C\lambda_{n2}r_n/8\right) \\ &\quad= \mathbb{P}\left(\frac{m_n^{-1}r_n^{-1/2} + m_n^{-1/2}}{r_n} > C/8\right) \longrightarrow 0, \end{aligned} \tag{2.5.58}$$

where (2.5.58) holds since $r_n, m_n \to \infty$. Hence by combining (2.5.56), (2.5.57) and (2.5.58), (2.5.50) follows.

Part (b): Denote $\eta_* = \max_{j\in A_*} 1/\|\beta_j\|_2$. Let $A_{***} = A_* \cup \{j : \|\hat{\beta}_{AgL,j}\|_2 > 0\}$. Note that $J_0 = |A_{***}|$. Define $\vartheta_{A_{***}} = Z^c - \mathbb{D}_{A_{***}}^{c}\beta_{A_{***}}$ and denote $\vartheta_{A_{***}}^*$ and $\epsilon_{A_{***}}^*$ be the projections of $\vartheta_{A_{***}}$ and ϵ_n to the span of $\mathbb{D}_{A_{***}}^{c}$. Then, in a similar way as in the part (b) of Theorem

1, we can show

$$\begin{aligned}\|\vartheta^*_{A_{***}}\|_2^2 &\leq \|\epsilon^*_{A_{***}}\|_2^2 + O_p(n^\alpha 1 + |A_{***}|\, n m_n^{-2\tau}) \\ &= \|(\mathbb{D}^{c\prime}_{A_{***}}\mathbb{D}^{c}_{A_{***}})^{-1/2}\mathbb{D}^{c\prime}_{A_{***}}\epsilon_n\|_2^2 + O_p(n^\alpha 1 + |A_{***}|\, n m_n^{-2\tau}) \\ &= O_p\left(\frac{|A_{***}|\, n^\alpha \rho^2_{\max}(L) m_n \log(J_0 m_n)}{c_1 h_n} + n^\alpha 1 + |A_{***}|\, n m_n^{-2\tau}\right). \quad (2.5.59)\end{aligned}$$

In a similar way to get (2.5.35), we can also show

$$\|\hat{\beta}_{AgL,A_{***}} - \beta_{A_{***}}\|_2^2 \;\leq \frac{8\|\vartheta^*_{A_{***}}\|_2^2}{n c_1 h_n} + \frac{4\lambda^2_{n2}|A_{***}|\eta_*}{n^2 c_1^2 h_n^2}, \quad (2.5.60)$$

thus, by (2.5.59) and (2.5.60), we obtain the part (b) of Theorem 3.

PROOF OF THEOREM 4

The proof would go similar as Theorem 2.

2.6 Few theoretical extensions

2.6.1 Results for extension to different variability of each additive components

In this section, we provide revised lemmas and theorems when we allow different levels of smoothness for additive components. We first extend the definition of $\mathcal{S}^0_{nj}$ as follows:

$$\mathcal{S}^0_{nj} \;=\; \left\{ f_{nj} : f_{nj} = \sum_{l=1}^{m_{nj}} b_{jl}\mathbb{B}^c_l(x), (b_{j1}, \cdots, b_{jm_{nj}}) \in \mathbb{R}^{m_{nj}} \right\}, 1 \leq j \leq J. \quad (2.6.1)$$

With new assumptions, Lemmas 1 and 3 and Theorems 2 and 4 are changed as follows.

Lemma 1′. Suppose that $f \in \mathcal{F}_j$ and $\mathbb{E}f(X_j) = 0$. Then, under (H 4)′ and (H 5), there exists an $f_n \in \mathcal{S}^0_{nj}$ such that

$$\|f_n - f\|_2 = O_p\left(m_{nj}^{-\tau} + \sqrt{\frac{m_{nj}}{n}}\right). \quad (2.6.2)$$

Particularly, under the choice of $m_{nj} = O(n^{\frac{1}{2\tau_j+1}})$, we have

$$\|f_n - f\|_2 = O_p\left(m_{nj}^{-\tau_j}\right) = O_p\left(n^{-\frac{\tau_j}{2\tau_j+1}}\right). \quad (2.6.3)$$

Lemma 3′. Define M_n be a non-negative definite matrix of order n and,

$$T_{jl} = \left(\frac{m_{nj}}{n}\right)^{\frac{1}{2}} a'_{jl} M_n \epsilon \qquad \forall 1 \le j \le J, 1 \le l \le m_{nj}, \tag{2.6.4}$$

where $a_{jl} = (\mathbb{B}^c_l(X_j(s)), s \in S)'$ and $T_n = \max_{\substack{1\le j\le J \\ 1\le l\le m_{nj}}} |T_{jl}|$. With new Assumption 1,

$$\mathbb{E}(T_n) \le C_1 \rho_{\max}(M_n) \sqrt{(m_n \log(J m_n)) O(n^{\alpha})}, \tag{2.6.5}$$

for some $C_1 > 0$ and $m_n = \max_{j=1,2,\ldots,J} m_{nj}$.

Theorem 2 ′. With new Assumption 1 and $\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha} m_n \log(J m_n)}$ for a sufficiently large constant C,

(a) $\|\hat{f}_{gL,j} - f_j\|_2^2 = O_p\left\{\left(\frac{\rho^2_{\max}(L) m_n^3 \log(J m_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda^2_{n1}}{n^2}\right)/m_{nj}\right\}$ for $j \in \tilde{A}_\beta \cup A_*$, where $\tilde{A}_\beta$ is the index set of nonzero gL estimates for β_j,

(b) If $\frac{m_n^2 \lambda^2_{n1}}{m_{nj} n^2} \longrightarrow 0$ as $n \longrightarrow \infty$ for $1 \le j \le q$, all the nonzero components $f_j, 1 \le j \le q$ are selected w.p. converging to 1.

Theorem 4 ′. With new Assumptions 1 and 2,

(a) $\mathbb{P}(\|\hat{f}_{AgL,j}\|_2 > 0, j \in A_* \text{ and } \|\hat{f}_{AgL,j}\|_2 = 0, j \notin A_*) \longrightarrow 1$,

(b) $\|\hat{f}_{AgL,j} - f_j\|_2^2 = O_p\left\{\left(\frac{\rho^2_{\max}(L) m_n^3 \log(J_0 m_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda^2_{n2}}{n^2}\right)/m_{nj}\right\}$
$\forall j \in A_*$.

2.6.2 Results for extension to a long range dependence

We extend the assumption (H 3) to cover a broader class of spatial covariance functions. The assumptions (H 6) and (K 2) are adjusted accordingly as well.

(H 3)* The random vector $\epsilon = \{\epsilon(s), s \in S\} \sim \text{Gaussian}(0, \Sigma_T)$, where $\Sigma_T = ((\sigma_{s,s'}))_{s,s' \in S}$

with $\sigma_{s,s'} = \delta(s-s')$ and $\delta(h)$ is a covariance function such that $\int_{D_n} \delta(h)dh = O(n^{\alpha})$ for some $\alpha \in [0,1)$. $D_n \subset \mathbb{R}^d$ is the sampling region that contains the sampling locations S. Without loss of generality, we assume that the origin of $\mathbb{R}^d$ is in the interior of D_n and D_n is increasing with n.

(H 6)* $m_n = O(n^{\gamma})$ with $1/6 \leq \gamma = 1/(2\tau+1) < (1-\alpha)1/3$.

(K 2)*

$$\frac{\sqrt{\rho^2_{\max}(L)\, n^{1+\alpha} m_n \log(s_n m_n)}}{\lambda_{n2} r_n} + \frac{n^2}{\lambda^2_{n2} r_n^2 m_n} + \frac{\lambda_{n2} m_n}{n} = o(1)$$

where $s_n = J - |A_{**}|$.

Lemmas and theorems are then updated in the following way.

Lemma 3*. Define M_n be a non-negative definite matrix of order n and,

$$T_{jl} = \left(\frac{m_n}{n}\right)^{\frac{1}{2}} a'_{jl} M_n \epsilon \qquad \forall 1 \leq j \leq J, 1 \leq l \leq m_n \tag{2.6.6}$$

where $a_{jl} = (\mathbb{B}^c_l(X_j(s)), s \in S)'$ and $T_n = \max_{\substack{1 \leq j \leq J \\ 1 \leq l \leq m_n}} |T_{jl}|$. Then, under assumptions (H 2), (H 3)*, (H 4) and (H 5),

$$\mathbb{E}(T_n) \leq C_1 \rho_{\max}(M_n) \sqrt{(m_n \log(J m_n)) O(n^{\alpha})}, \tag{2.6.7}$$

for some $C_1 > 0$.

Lemma 4*. Under the Assumption 1 with updated (H 3)* and (H 6)*
and with $\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha} m_n \log(J m_n)}$ for a sufficiently large constant C, we have $|\tilde{A}_{\beta}| \leq M_1 |A_*|$ for a finite constant $M_1 > 1$ with w.p. converging to 1.

Theorem 1.* Suppose that conditions in Assumption 1 with updated (H 3)* and (H 6)* hold and if $\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha} m_n \log(J m_n)}$ for a sufficiently large constant C. Then, we have

(a) $\sum_{j=1}^{J} \|\hat{\beta}_{gL,j} - \beta_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^3 \log(Jm_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda_{n1}^2}{n^2}\right)$,

(b) If $\frac{m_n^2\lambda_{n1}^2}{n^2} \longrightarrow 0$ as $n \longrightarrow \infty$, all the nonzero components $\beta_j, 1 \leq j \leq q$ are selected with probability (w.p.) converging to 1.

Theorem 2.* Suppose that conditions in Assumption 1 with updated (H 3)* and (H 6)* hold and if $\lambda_{n1} > C\rho_{\max}(L)\sqrt{n^{1+\alpha} m_n \log(Jm_n)}$ for a sufficiently large constant C. Then,

(a) $\|\hat{f}_{gL,j} - f_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^2 \log(Jm_n)}{n^{1-\alpha}} + \frac{1}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau}} + \frac{4m_n\lambda_{n1}^2}{n^2}\right)$ for $j \in \tilde{A}_\beta \cup A_*$, where $\tilde{A}_\beta$ is the index set of nonzero gL estimates for β_j,

(b) If $\frac{m_n\lambda_{n1}^2}{n^2} \longrightarrow 0$ as $n \longrightarrow \infty$, all the nonzero components $f_j, 1 \leq j \leq q$ are selected w.p. converging to 1.

Theorem 3.* Suppose that conditions in Assumptions 1 and 2 with updated (H 3)*, (H 6)* and (K 2)* are satisfied. Then,

(a) $\mathbb{P}(\hat{\beta}_{AgL} =_0 \beta) \longrightarrow 1$,

(b) $\sum_{j=1}^{q} \|\hat{\beta}_{AgL,j} - \beta_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^3 \log(J_0 m_n)}{n^{1-\alpha}} + \frac{m_n}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right)$.

Theorem 4.* Suppose that conditions in Assumptions 1 and 2 with updated (H 3)*, (H 6)* and (K 2)* are satisfied. Then,

(a) $\mathbb{P}(\|\hat{f}_{AgL,j}\|_2 > 0, j \in A_* \text{ and } \|\hat{f}_{AgL,j}\|_2 = 0, j \notin A_*) \longrightarrow 1$,

(b) $\sum_{j=1}^{q} \|\hat{f}_{AgL,j} - f_j\|_2^2 = O_p\left(\frac{\rho_{\max}^2(L) m_n^2 \log(J_0 m_n)}{n^{1-\alpha}} + \frac{1}{n^{1-\alpha}} + \frac{1}{m_n^{2\tau}} + \frac{4m_n\lambda_{n2}^2}{n^2}\right)$.

The updated theorems show that the lower bound of the penalty parameter as well as the convergence rate are affected by α. More specifically, introduction of α increases the order in the lower bound of the penalty parameter and the order of the convergence

rate is decreased (slower convergence rate) with α. Note that α does not fully characterize a spatial dependence structure but it gives some information on the level of spatial dependence such that $0 < \alpha < 1$ implies a long-range dependence. For any integrable stationary spatial covariance model, $\alpha = 0$ and this is the case for most practical situations. If $0 < \alpha < 1$, one might consider estimating α for calculation of the lower bound of the penalty parameter. There are some literature which provide how to estimate long-range parameters for random fields [e.g. Anh and Lunney (1995), Boissy *e*t al. (2005)], but they are limited since a specific class of random fields or a parametric model is assumed. Estimation of α has its own interest but we do not pursue it since our focus is on variable selection.

CHAPTER 3

ESTIMATING NON-STATIONARY SPATIAL COVARIANCE MATRIX USING MULTI-RESOLUTION KNOTS

For most of statistical prediction problems, obtaining a BLUP is very crucial and generally modeling and estimating the mean does the trick. Although estimation of the underlying process covariance is instrumental for spatial BLUP also known as kriging. The concept of kriging was first introduced by D.G.Krige, a South African mining engineer (Cressie, 1990) and Matheron in 1962 coined the term to honor Krige. Kriging is a very popular tool used in earth climate modeling and environmental sciences. It uses quantification of spatial variability through covariance function and solving the standard kriging equation is often numerically cumbersome, and involves inversion of a $n \times n$ covariance matrix. With large n, which is quite reasonable for real data observed on global scale since computation cost increases with cubic power of the dimension n, spatial BLUP becomes challenging.

Hence, there have been several efforts to achieve a computationally feasible estimate. The foremost challenge of estimating covariance for a spatial set up arises due to absence of repetition. This may seem absurd if we realize this situation as a multivariate extension of computing variance from one observation. As odd as may it sound, the trick is to consider a specific sparsity structure for the covariance matrix under study. The covariance matrix is sparse when the covariance function is of finite range and due to sparsity the computation cost to invert a $n \times n$ matrix reduces considerably.

Before we delve in to the discussion of our contribution we would like to put forward a few other attempts to estimate large covariance matrices through literature review. In 1997 Barry and Pace used symmetric minimum degree algorithm when $n = 916$ for kriging. Rue and Tjelmeland (2002) approximated Σ^{-1} to be sparse precision matrix of a Gaussian Markov random field wrapped on a torus. For larger n, the first challenge in applying kriging is, increase in condition number of the covariance matrix, which plays a major role

in building up the computation time and makes the kriging equation numerically unstable. On the other hand, to handle computational complexity, Kaufman *et.al.* (2008), introduced the idea of covariance tapering which sparsify the covariance matrix element wise to approximate the likelihood. Some other worth mentioning efforts in tapering are Furrer *et.al.* (2012), Stein (2013) e.t.c. Covariance tapering gains immense computational stability, keep interpolating property and also have asymptotic convergence of the taper estimator. But tapering is restricted only to isotropic covariance structure and the tapering radius needs to be determined.

Another alternative method, FRK was introduced by Cressie & Johannesson (2008). Unlike tapering, FRK is applicable to a more flexible class of non-stationary covariance matrix, and also reduces computational cost of kriging to $O(n)$. For many non-stationary covariance model like ours, the observed process covariance matrix can be decomposed into two additive matrix components. The first is a measurement error modeled as white noise. While the second is an underlying process which can be non-stationary covariance structure and is often assumed to be fixed but low rank. The underlying process can be represented as a linear combination of r_n random effects. For FRK r_n plays the role of rank of the non-stationary component, is considered to be known r and fixed over n. In this work we would like to relax this assumption by allowing r_n changing over n.

Our goal in this paper is to achieve a data driven approach for finding the rank r_n. To do so let us assume even though there are unknown r_n random effects used to represent the underlying process, what if we start with some numbers of random effects and as we proceed, our algorithm will direct us toward a data driven value for r_n? Once we figure out that the dispersion matrix of this n dimensional random effect can be decomposed into cholesky factor, a closer look will teach us that dropping or selecting a particular random effect boils down to zero or non-zero row in the corresponding cholesky matrix. We consider a penalized likelihood approach where we penalize ℓ_2- norm within each row of the cholesky matrix and ℓ_1- norm between two different rows of the cholesky matrix.

The low rank non-stationary covariance matrix is decomposed, using a basis components (not necessarily orthogonal) and another component is dispersion matrix of random effects vector. The basis component depends primarily on the choice of the class of basis function and number of knot points. FRK recommends that the choice of basis function should be multi-resolutional, more precisely they used a local bi-square functions. This use of locally multi-resolutional knots has also been proved quite useful in the literature of kriging for large spatial data sets (Nychka (2015)) other than FRK. The choice of number of knot points and their positions is always crucial. The number of knot points is directly related to r_n, the number of random effects component. The foremost challenge in applying our method is choice of effective numbers of knot points necessary to construct the basis function under study.

Although our initial objective in this work is to provide a way to estimate the non-zero random effects and finally the covariance matrix, but like any other statistical prediction problem we shall be extending our findings in presence of covariates. Peng and Wu (2010), proved that condition number of the covariance matrix also increases with increase in input variables. To handle numerical instability, Peng and Wu (2010), suggested the idea of regularized kriging, which is a simple modification in the method of estimation. Unlike kriging, regularized kriging optimizes regularized or penalized likelihood. At this stage we have not considered dimension reduction challenges while extending our findings in presence of covariates but, for future studies, this can be a non-trivial and worthwhile extension.

A recent study on limitations of low rank kriging (Stein (2015)) shows an approximation in which observations are split into contiguous blocks and assumes independence across these blocks. It provides a much better approximation to the data likelihood than a low rank approximation requiring similar memory and calculations. It also shows that Kullback-Leibler divergence for low rank approximation is not reduced as much as it should have been in few settings. On the contrary the divergence is considerably reduced if there is a block structure. Keeping this in mind, and considering the fact that selections of knots work better

under multi-resolution setup, we consider the knots by superimposing resolutions.

Under some sensible assumptions this work will motivate our readers to the idea of existence of a consistent covariance estimator of the spatial process using a low rank modeling, whose estimation has not been discussed before in any literature to the best of our knowledge. We will discuss the practical implications of our assumption later but, we still like to point out that without loss of generality we considered, the location knots for the bi-variate spline matrix are ordered in a specific way such that the true structure has the first r_n non-zero rows and rest $n - r_n$ zero rows. We also discuss how our findings fit in the situations of limitations of low rank kriging (Stein (2015)). To avoid further mathematical details here, this part of the comparison is in discussion section 3.4.

All kinds of approximation of the covariance function introduced so far, has a motive to reduce the computational cost. Most of these existing methods fail to capture both large scale (long-range) and small scale (short-range) dependence. However tapering captures small scale dependence and, low rank techniques captures large scale dependence. A new method is discussed using adding these two components (Sang and Huang 2012). We would like to point out our readers that however worthwhile this method of combining both low rank and tapering may look, this paper provides a more sound theoretical approach to support our algorithm and findings. Although estimation of low rank covariance matrix has it's limitations, the method has not always been criticized, rather well established in several situations by various authors. Most of the interesting work in this field, can be classified in two broad classes: statistics and machine learning. Among many others in the field of statistics we think, Fan and Li (2012), Banerjee *et.al.* (2012), Tzeng and Huang (2015) e.t.c. are worth mentioning. On other the hand, the field of machine learning focuses on developing algorithms where, Frieze *et.al.* (2004), Achlioptas and McSherry (2007), Journée *et.al.* (2010) are quite reasonable to browse through. Based on these literatures it is obviously worthwhile to contribute our time and to come up with a theoretical justification behind the possibility of low rank covariance matrix estimation.

Even when we keep the rank fixed, for a very large data set (order of tens of thousands to hundreds of thousands), kriging can be quite impossible and ad hoc local kriging neighborhoods are used (Cressie (1993)). Some recent developments include Nychka *et.al.* (1996; 2002), Furrer *et.al.* (2006) and many more. Among other alternative methods, some worth discussing are Radial basis interpolation functions (B*ü*hlmann, (2004)), inverse distance weighting (Shepard, (1968)) or regression-based inverse distance weighting used by Joshep and Kang (2009) which is a fast interpolator and overcomplete basis surrogate method (Chen, Wang, and Wu (2010)). Surrogate basis representation is similar to lattice kriging (Nychka (2015)) where the basis functions are boundedly supported and over complete. But lattice kriging considers sparsity in the precision matrix through bounded basis function matrix and a parametric neighborhood matrix whereas we are considering sparsity in the covariance matrix through low rank factorization and Cholesky decomposition of the low rank covariance matrix.

The rest of this paper is organized as follows. In Section 3.1, we explain the proposed approach for selecting and estimating nonzero rows (rank) and the corresponding low rank covariance matrix. Following which in section 3.2 we present the block coordinate descent algorithm for block wise convex regularizing functions. Section 3.3 contains simulation results along with a real data example. We make some concluding remarks in section 3.4.

3.1 Methodology for estimating a non-stationary low rank covariance matrix

To vividly understand the methodology we need to explain two ideas, arranging bivariates knots in multi-resolution setup, and group LASSO penalty to estimate parameters having an inherent group structure. Both the ideas are seperately useful in solving two different problems. The arrangement of bivariates knots in multi-resolution setup plays an important role in indexing of spatial domain. As we all know unlike time series analysis,

where the time domain is easily arranged in chronological fashion, in a spatial domain from kD and $k \geq 2$, it is quite impossible to arrange the location sites in any chronological pattern. Whereas, the group LASSO penalization plays the role in solving the problem of high dimensionality of the parameter space, which depends on the number of location sites we consider in our study.

Group LASSO penalization (Yuan *et.al.* (2006)), or the concept of using ℓ_1/ℓ_2 - penalty (Bühlmann *et.al.* (2011)) has been well established in the context of selecting varibles if it is believed that there exists an inherent group structure in the parameter space. But it has not been quite clear how such approach is applied to estimating rank of a low-rank matrix and estimating the matrix itself. In this section, we propose an ℓ_1/ℓ_2 - penalized approach in estimating the low rank non-stationary covariance matrix as an extension of FRK. The goal of FRK is to reduce computation cost of inversion of a matrix from cubic to linear in sample size while allowing non-stationarity. To explain the difference between FRK and our method, we introduce the following mathematical notations. Since the ideas are interlinked, I would like to present both using the same notations and hence the following notational book keeping.

Consider a spatial process, $Y = \{Y(s); s \in \mathscr{S}\}$, perturbed with measurement error $\epsilon = \{\epsilon(s); s \in \mathscr{S}\}$ and let $X = \{X(s); s \in \mathscr{S}\}$ be the process for the observables where ϵ is a Gaussian process with mean 0 and $var\,(\epsilon(s)) = \sigma^2 v(s) \in (0, \infty), s \in \mathscr{S}$, for $\sigma^2 > 0$ and $v(\cdot)$ known. $\mathscr{S}$ is a spatial domain of interest. In general, the underlying process Y has a mean structure, $Y(s) = Z(s)'\beta + \pi(s)$, for all $s \in \mathscr{S}$ where, $\pi = \{\pi(s); s \in \mathscr{S}\}$ follows a Gaussian distribution with mean 0, $0 < var\,(\pi(s)) < \infty$, for all $s \in \mathscr{S}$, and a non-stationary spatial covariance function $cov\left(\pi(s), \pi(s')\right) = \sigma(s, s')$, for all $s, s' \in \mathscr{S}$. Also $Z = \{Z(s); s \in \mathscr{S}\}$ represents known covariates and β is the vector of unknown coefficients. Combining the underlying process and the measurement error, we have

$$X(s) = Y(s) + \epsilon(s) = Z(s)'\beta + \pi(s) + \epsilon(s) \qquad \forall s \in \mathscr{S}. \tag{3.1.1}$$

The process $X(\cdot)$ is observed only at a finite number of spatial locations $S_n = \{s_1, s_2, \ldots, s_n\} \subset \mathscr{S}$. We allow S_n to be any irregular lattice in d-dimension with cardinality n. Note that the covariance function $\sigma(s, s')$ has to be a positive definite function on $\mathbb{R}^n \times \mathbb{R}^n$. In practice, we often consider $\sigma(s, s')$ as a stationary covariance function, but in this paper we want to keep it general and allow the possibility of it being non-stationary.

Recall, the spatial random effects $\pi(s)$ in model can be represented in $R(s)'\alpha$, where $R(s) = (R_1(s), R_2(s), \ldots, R_r(s))'$. Hence the parameter r is a number of knots necessary to approximate the function $\pi(s)$. In a recent study on similar kriging estimation by Stein (2015) it came out to light, that the number of knots, r should be depending on the number of location points, n. It also discusses few challenges one might have with this technique. And, we will be comparing our findings with that of his. Henceforth, we will also be using r_n to denote the number of knots necessary to approximate $\pi(s)$. We capture the spatial information through basis functions (*Cressie, N. et al. (2008)*),

$$R(s) = (R_1(s), R_2(s), \ldots, R_{r_n}(s))', \qquad \forall s \in S_n$$

and for a positive definite $r_n \times r_n$ matrix Ω, we have a model for our covariance function $\sigma(s, s')$ as,

$$\sigma(s, s') = R(s)'\Omega R(s'), \qquad \forall s, s' \in S_n. \tag{3.1.2}$$

The above modeling can be viewed as a consequence of writing $\pi(s) = R(s)'\alpha$, where α is an r_n-dimensional vector with $var(\alpha) = \Omega$. The model for $\pi(\cdot)$ is often refered to as *s*patial random-effects(**S**RE) model. Define matrix R with $R(s_i)'$ as the i^{th} row and correspondingly, $\Sigma = R\Omega R'$, where $\mathscr{R}(\Sigma) \leq \mathscr{R}(\Omega) = r_n$, where $\mathscr{R}(\cdot)$ is the rank of a matrix.

Note that, SRE representation *i.e.*, linear combination of the random effects vector α of r_n components can be looked upon as approximating a function defined over the sampling region. While, r_n is the number of knot locations used to approximate the underlying function, $\sigma(s, s')$, the challenge is also to provide these locations. Henceforth, our effort is to come up with a methodology for estimating the parameter r_n along with positions of

these knots. The arrangement of knots in this work, can be called 'multi-resoultion' knots and recently been proved to be effective in an alternative model for covariance function by Nychka *et.al.*, (2015).

To understand the idea it is better to present it using the following set of figures. For simplicity we will be considering our points to be lying in a $2-$dimensional space. Figure 3.1 is the scatterplot of all the location sites. The horizontal axis corresponds to longitudes whereas the vertical axis corresponds to latitudes. Each and every locations are indexed by a pair of $(latittude, longitude)$. Unlike in a time series study here we can not order these paired locations. We can define a two dimensional domain that contains all the point. If we scale each axes with same length on each direction, with out loss of generality the domain containing all the points can be a square or a cube or a hyper-cube based on the dimension under study.

Let us divide the two sides of the square in equal halves. Hence we will have four equal squares with one forth area of the whole domain. The positions of the centroids of these four quadrants are the locations for the first four knots. Let's call collection of these four locations as 1^{st} resolution knots. If we repeat the same dyadic break of each of the 1^{st} resolution squares, we get sixteen squares. Centroids of these sixteen squares comprises the second resolution. Figure 3.3 points out all twenty knots of first and second resolutions. So, for a $d-$dimensional space k^{th} resolution has $(2^d)^k$ knots. In any given study let's say we have sampling sites from a $d-$dimensional space, for some M, we will have,

$$\sum_{k=0}^{M-1}(2^d)^k + 1 \leq \text{ sample points } \leq \sum_{k=0}^{M}(2^d)^k,$$

and choose M to be the number of resolutions. So, for $2-$dimensional space, there will be 4 knots in 1^{st} resolution, 16 knots in 2^{nd} resolution, 64 knots in 3^{rd} resolution, and so forth. Let L be the cumulative sum of all these knots. We start with all together L basis for each site,

$$\widetilde{R}(s) = (\widetilde{R}_{1(1)}(s), \ldots, \widetilde{R}_{\ell_1(1)}(s); \ldots; \widetilde{R}_{1(M)}(s), \ldots, \widetilde{R}_{\ell_M(M)}(s))',$$

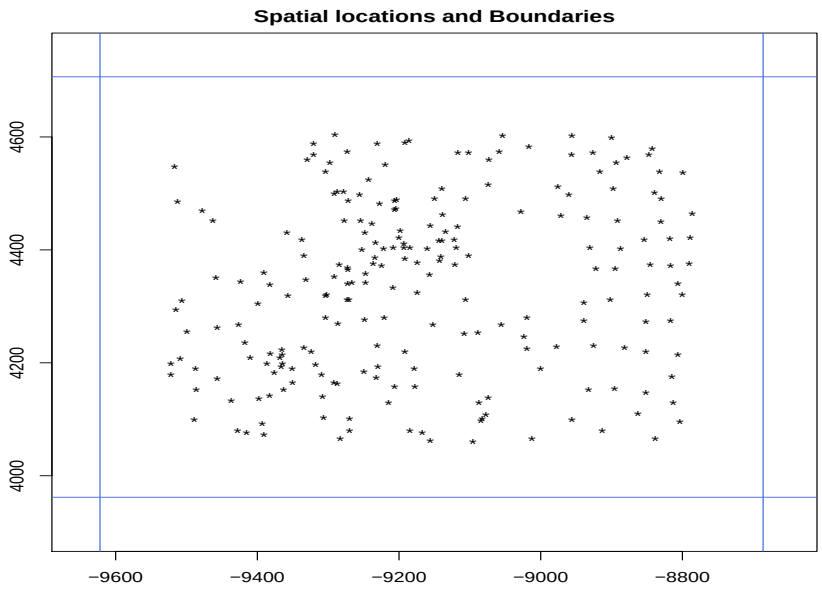

Figure 3.1 Locations sites in the study

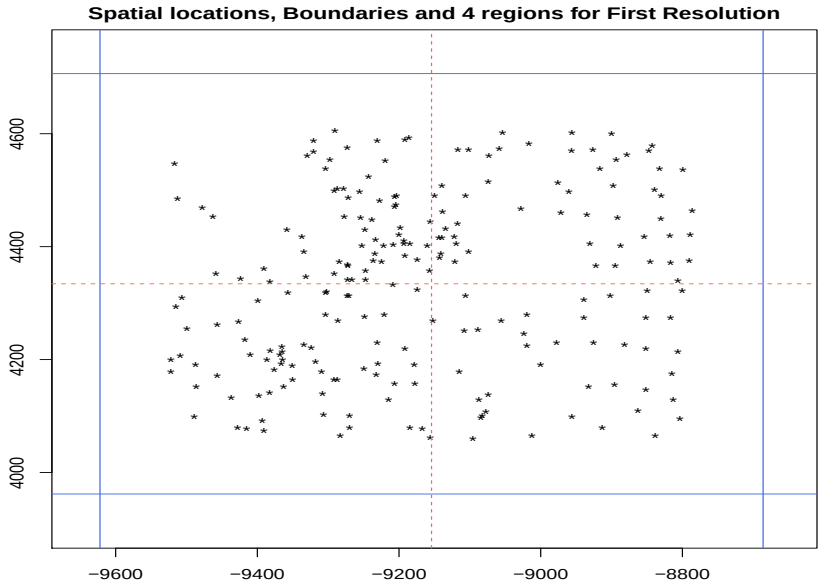

Figure 3.2 First resolution overlayed on location sites in the study

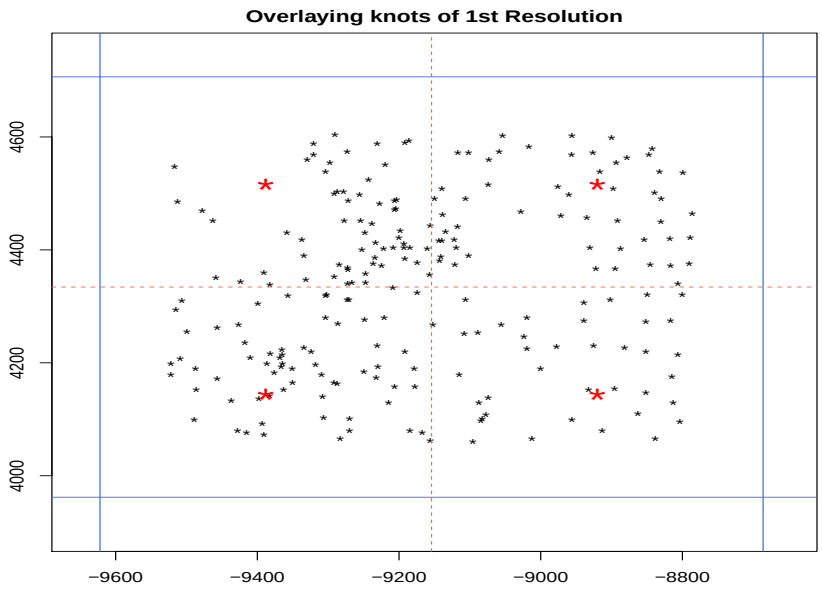

Figure 3.3 Second resoltions overlayed on location sites in the study

Figure 3.4 resolutions overlayed on locations sites in the study

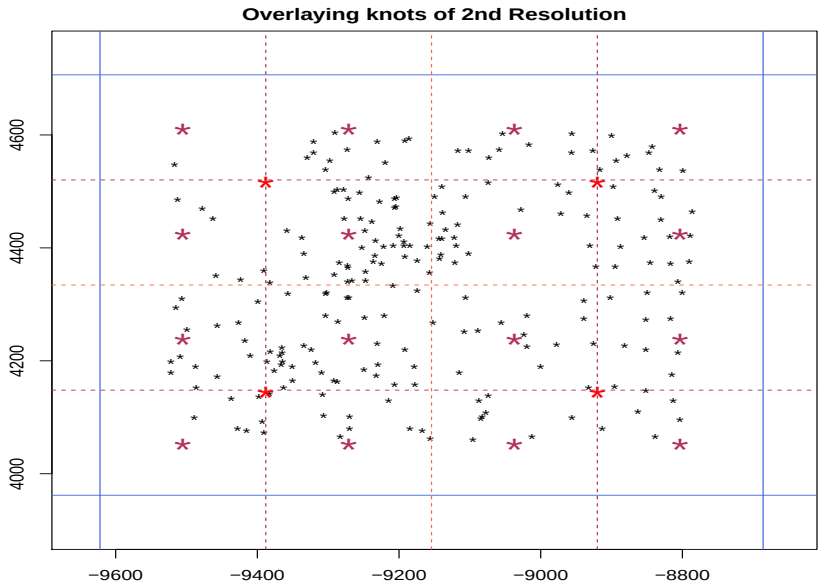

Figure 3.5 Third resolutions overlayed on locations sites in the study

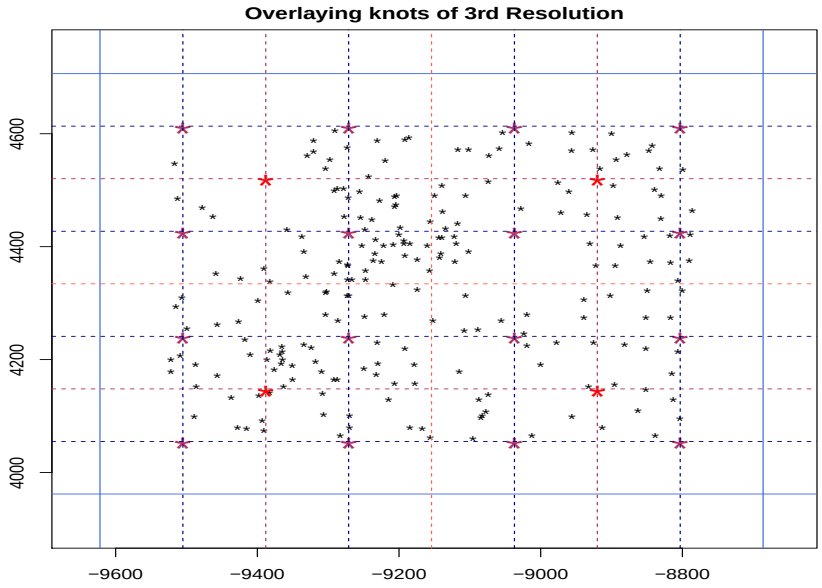

Figure 3.6 Three resolutions overlayed on locations sites in the study

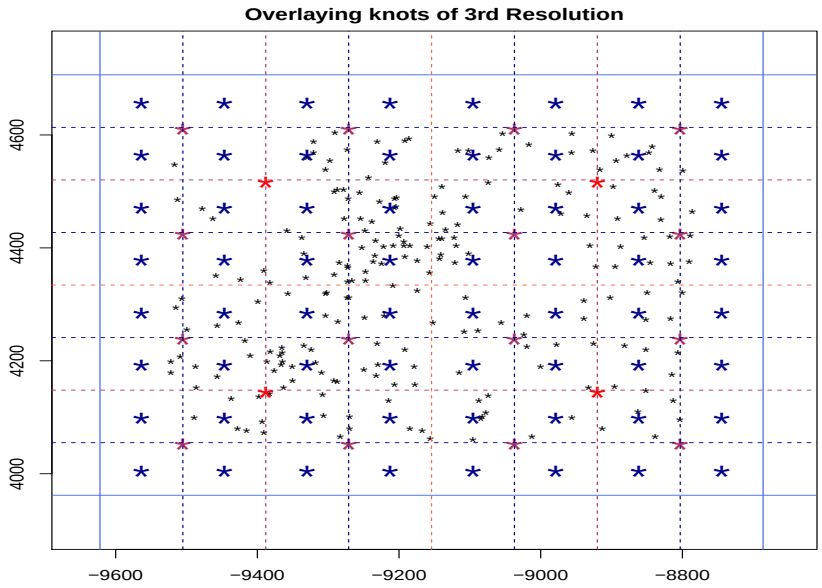

Figure 3.7 Three resolutions overlayed on locations sites in the study

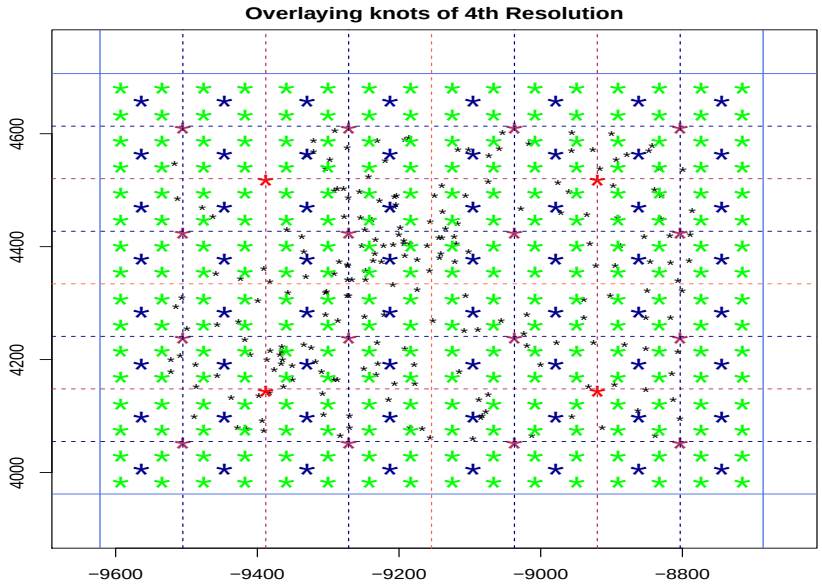

Figure 3.8 Three resolutions overlayed on locations sites in the study

where, $\ell_1 + \ldots + \ell_M = L$ (say). The basis vector $\widetilde{R}(s)$ for any location s has L components, each denoted as $\widetilde{R}_{i(k)}(s)$. Here $\widetilde{R}_{i(k)}(s)$ is the basis component corresponding to the i^{th} knot of k^{th} resolution. The matrix $\widetilde{R}$ is constructed using either 'Local bi-square' or 'Wendland' basis functions. If s denotes the spatial location points, and $u_{i(k)}$ be the i^{th} knot location for k^{th} resolution we define distance between location site s and knot location $u_{i(k)}$ as,

$$d_{i(k)}(s) = d\left(s, u_{i(k)}\right) = \frac{\left\|s - u_{i(k)}\right\|}{\theta_\ell},$$

where, $\theta_k = 1.5\times$(shortest distance between knot points of k^{th} resolution). Finally, local Bi-square basis function is defined as,

$$\widetilde{R}_{i(k)}(s) = R\left(d\left(s, u_{i(k)}\right)\right) = \begin{cases} \left(1 - d_{i(k)}(s)^2\right)^2 \text{ if, } d_{i(k)}(s) \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

and, Wendland basis function is defined as,

$$\widetilde{R}_{i(k)}(s) = R\left(d\left(s, u_{i(k)}\right)\right) = \begin{cases} \left(1 - d_{i(k)}(s)\right)^6 \left(35 d_{i(k)}(s)^2 + 18 d_{i(k)}(s) + 3\right) \text{ if, } d_{i(k)}(s) \leq 1 \\ 0 \text{ otherwise.} \end{cases}.$$

Note that, both these basis functions are bounded. Although by introducing multi-resolution knot arrangements we are bringing in more parameters and we are cursed with dimensionality we will be exploiting boundedness to acheive sparsity. The objective here, is to obtain which r_n among these L knots are necessary to model the non-stationary covariance function. In a situation, where the sample points are distributed uniformly over the sample domain, we select the first r_n basis rows, and we drop the rest to obtain R. But even, in practice for irregularly spaced data, we might hope that we would able to select r_n out of all the L knots.

As mentioned earlier, multi-resolution knots plays an important role in defining the underlying indexing system even though the original location points are hard to order, it comes with the curse of dimensionality. To present the curse of dimensionality we are using

Resolution Index i	Range of Sample point	
	$2-dimensional\ space$	$3-dimensional\ space$
1	(1-4)	(1-8)
2	(5-20)	(9-72)
3	(21-84)	(73-584)
4	(85-340)	(585-4680)
5	(341-1364)	(4681-37448)
6	(1365-5460)	(37449-299592)

Table 3.1 Number of knots necessary for every resolution

table 3.1, where the first column represents the resolution, and the other two columns are for two and three dimensional spaces respectively. The numbers in the brackets $(a-b)$ means, if the study has n sample points, with $(a \leq n \leq b)$ then we need b knots. One the other hand the number of resolution will be corresponding to the number appearing in the first column.

So, for our covariance function $\sigma(s, s')$ we start with the model,

$$\sigma(s, s') = \widetilde{R}(s)'\widetilde{\Omega}\widetilde{R}(s'), \qquad \forall s, s' \in S_n \tag{3.1.3}$$

Similar to the equation (3.1.2), one can easily see that the equation (3.1.3) is a consequence of writing $\pi(s) = \widetilde{R}(s)'\widetilde{\alpha}$, where $\widetilde{\alpha}$ is an $L-$dimensional vector with $var\,(\widetilde{\alpha}) = \widetilde{\Omega}$. Hence this expression of random effect $\pi(s)$ in (3.1), gives us

$$X(s) = Z(s)'\beta + \widetilde{R}(s)'\widetilde{\alpha} + \epsilon(s) \qquad \forall s \in \mathscr{S}. \tag{3.1.4}$$

For simplicity, let us first present our method for the case $Z\beta = \mathbf{0}$. Let us now explain the relation between two versions of random effects or covariance model and, the method used to reduce this dimensionality cost. Ideally, Ω is a sub-matrix of $\widetilde{\Omega}$ with $\mathscr{R}\left(\widetilde{\Omega}\right) = \mathscr{R}(\Omega)$ such that,

$$\begin{pmatrix} \Omega & \mathbb{O}_{r_n \times (L-r_n)} \\ \mathbb{O}_{(L-r_n)\times r_n} & \mathbb{O}_{(L-r_n)\times(L-r_n)} \end{pmatrix} = \widetilde{\Omega} = \widetilde{\Phi}\widetilde{\Phi}' = \begin{pmatrix} \Phi\Phi' & \mathbb{O}_{r_n \times (L-r_n)} \\ \mathbb{O}_{(L-r_n)\times r_n} & \mathbb{O}_{(L-r_n)\times(L-r_n)} \end{pmatrix}, \tag{3.1.5}$$

where, $\Phi_{r_n \times r_n}$ is the $r_n \times r_n$ matrix from the Cholesky decomposition of Ω, which is the unique lower triangular matrix. In practice, it may be necessary that we reorder the columns

in our basis matrix $\widetilde{R}$ to achieve the above structure. This reordering can be taken care of by introducing a permutation matrix, explained in the appendix. So, for the rest of the discussion we will consider $\Sigma = \widetilde{R}\widetilde{\Omega}\widetilde{R}'$. Since the rank is unknown, we propose to start with all L rows non-zero. Our proposed method allows us to select non-zero rows of $\widetilde{\Phi}$, which eventually captures all information required to retrieve Σ.

We drop a row from $\widetilde{\Phi}$ if and only if all the elements in that row are smaller than some preset value. The innovation of this work, is to exploit the group structure in the rows of lower triangular matrix Φ. This has two significances of this work. First, is observe, as

To do this, we consider a group wise penalization. Such shrinkage equation will have a similar nature of block-wise optimization. Denote $\widetilde{\varphi}'_{(j)} = (\widetilde{\varphi}_{j1}, \widetilde{\varphi}_{j2}, \ldots, \widetilde{\varphi}_{jj}, 0, 0, \ldots, 0) = (\widetilde{\varphi}'_j, 0, 0, \ldots, 0)$ to be the j^{th} row of $\widetilde{\Phi}$, where the number of zeros in the j^{th} row is $L-j$.

Define $\widetilde{\Phi}^{vec}_{Fullset} = (\widetilde{\varphi}'_1, \widetilde{\varphi}'_2, \ldots, \widetilde{\varphi}'_L)$ to be a row-wise vector representation of the lower triangular part of the matrix $\widetilde{\Phi}$. For a weight vector $\psi = (\psi_1, \psi_2, \ldots, \psi_L)'$, we define a weighted ℓ_1/ℓ_2-norm, $\left\|\widetilde{\Phi}^{vec}_{Fullset}\right\|_{2,1,\psi} = \sum_{j=1}^{L} \psi_j \left\|\widetilde{\varphi}_{(j)}\right\|_2$, where $\|\cdot\|_2$ is the ℓ_2-norm of a vector. So, we propose the following weighted ℓ_1/ℓ_2-penalized log likelihood function,

$$Q_n(\widetilde{\Phi}, \sigma^2, \tau_n, \psi) = X'\Xi^{-1}X + \log\det\Xi + \tau_n \left\|\widetilde{\Phi}^{vec}_{Fullset}\right\|_{2,1,\psi}, \tag{3.1.6}$$

where τ_n is the regularization parameter, $\psi_n = (\psi_{n1}, \psi_{n2}, \ldots, \psi_{nL})'$ is a suitable choice of a weight vector in the penalty term and $\Xi = \sigma^2\mathbb{I} + \Sigma$. We allow the penalty parameter, τ_n, and the weight vector, ψ_n, to depend on the sample size n. Now using the above covarince modeling for Σ *i.e.* $\Sigma = \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'$, (3.1.6) can be rewritten as,

$$\begin{aligned} Q_n(\widetilde{\Phi}, \sigma^2, \tau_n, \psi) = n\,\mathbf{T}\mathrm{r}\left(\Xi_0\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right)^{-1}\right) &+ \log\det\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right) \\ &+ \tau_n \left\|\widetilde{\Phi}^{vec}_{Fullset}\right\|_{2,1,\psi}, \end{aligned} \tag{3.1.7}$$

where $\Xi_0 = XX'/n$ is the scaled empirical covariance matrix. One can observe that the length of nonzero components in each row of $\widetilde{\Phi}$ is varying since it is a lower triangular matrix and hence ideally we should put varying penalty quantity for each row of the matrix.

One way to handle this problem is to rescale the j^{th}column of $\widetilde{R}$ by $1/\psi_{nj}$. So we define $\widetilde{R}^{\star}$ with the j^{th} column equal to $1/\psi_{nj}$ times the j^{th} column of $\widetilde{R}$, and accordingly we define $\widetilde{\Phi}^{\star}$ with the j^{th} row equal to ψ_{nj} times the j^{th} row of $\widetilde{\Phi}$ which leads to $\widetilde{R}\widetilde{\Phi} = \widetilde{R}^{\star}\widetilde{\Phi}^{\star}$. This transformation helps us to achieve invariance in $\widetilde{\Sigma} = \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'$ for adaptive group LASSO. Therefore the optimization problem in (3.1.7) boils down to an unweighted ℓ_1/ℓ_2-penalized log likelihood function,

$$Q_n(\widetilde{\Phi}, \sigma^2, \tau_n, \mathbf{1}) = \mathbf{T}\mathrm{r}\left(\Xi_0\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right)^{-1}\right) + \log\det\left(\sigma^2\mathbb{I} + \widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right) + \tau_n \left\|\widetilde{\Phi}^{vec}_{Fullset}\right\|_{2,1,\mathbf{1}}. \tag{3.1.8}$$

We want to restrict our search over the space of lower triangular matrices with absolutely bounded elements and bounded σ. Let us denote our search space by $\mathscr{P}_0^N$ introduced in (**??**), where $N = \frac{n(n+1)}{2} + 1$, the total number of parameters, is an increasing function of n. Observe that with this rescaling, magnitude of our spatial basis matrix $\widetilde{R}$ will change over n which could imply that the largest or smallest eigenvalues of $\widetilde{R}$ are not fixed for the varying sample size. As a choice for ψ_{nj}, we consider $\psi_{nj} = 1/j$, *i.e.* the j^{th} row, $\widetilde{\varphi}_{(j)}$, is scaled down by its number of nonzero components. Define, $\left(\widehat{\widetilde{\Phi}}_{gL}(\tau_n), \hat{\sigma}^2\right) = \arg\min_{\mathscr{P}_0^N} Q_n(\widetilde{\Phi}, \sigma^2, \tau_n, \mathbf{1})$. Based on $\widehat{\widetilde{\Phi}}_{gL}$, $\hat{\sigma}^2$ and $\widetilde{R}$, we can obtain $\widehat{\Xi}_{gL} = \hat{\sigma}^2\mathbb{I} + \widetilde{R}\widehat{\widetilde{\Phi}}_{gL}\widehat{\widetilde{\Phi}}'_{gL}\widetilde{R}'$ and $\hat{r}_n = \mathscr{R}\left(\widehat{\widetilde{\Phi}}_{gL}\right)$.

3.2 Block Coordinate Descent algorithm with Proximal update

In this section we will present an cost-effective algorithm for the optimization problem posed in (3.1.8). We have a block-wise function, blocks being the rows of a lower triangular matrix $\widetilde{\Phi}$, along with a group LASSO type penalty, groups corresponding to each block. There has been few significant efforts behind building efficient algorithm to minimize a penalized likelihood. Although group wise penalization is not a completely different ball

game, it still requires some special attention, which exploits the group structure and considers penalizing ℓ_2-norm of each group.

We will be using a **B**lock **C**oordiante **D**escent (BCD) method for a block multi-convex function under regularizing constraints,

$$\min_{x\in\mathcal{X}}\left\{F(x_1,x_2,\ldots,x_n)=f(x_1,x_2,\ldots,x_n)+\sum_{j=1}^{n}r_j(x_j)\right\}, \tag{3.2.1}$$

where x is decomposed into n blocks and $r_j(x_j)$ is the regularizing constraint for the j^{th} block. On comparing (3.2.1) with (3.1.8) we can see that, in our case we have n blocks, $\mathcal{X}$ is the collection of lower triangular matrices of the form, $\widetilde{\Phi}$, $F(\widetilde{\Phi})=Q_n(\widetilde{\Phi},\tau_n)$ with,

$$f(\widetilde{\varphi}_{(1)},\widetilde{\varphi}_{(2)},\ldots,\widetilde{\varphi}_{(n)})=\mathbf{T}\mathrm{r}\left(\Xi_0\left(\sigma^2\mathbb{I}+\widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right)^{-1}\right)+\log\det\left(\sigma^2\mathbb{I}+\widetilde{R}\widetilde{\Phi}\widetilde{\Phi}'\widetilde{R}'\right) \tag{3.2.2a}$$

$$r_j\left(\widetilde{\varphi}_{(j)}\right)=\tau_n\frac{\left\|\widetilde{\varphi}_{(j)}\right\|_2}{j} \tag{3.2.2b}$$

To ease the computation we use **M**atrix determinant lemma and **S**herman-Morisson-Woddbury matrix indentiy. We follow "**p**rox-linear" algorithm (Xu & Yin (2013)) where the update for $\widetilde{\varphi}_{(j)}$ in the k^{th} step is denoted by $\widetilde{\varphi}^k_{(j)}$ and is given by,

$$\widetilde{\varphi}^k_{(j)}=\arg\min_{\widetilde{\varphi}_{(j)}}\left\{\left\langle\hat{\mathrm{g}}^k_j,\widetilde{\varphi}_{(j)}-\widehat{\widetilde{\varphi}}^{k-1}_{(j)}\right\rangle+\frac{L^{k-1}_j}{2}\left\|\widetilde{\varphi}_{(j)}-\widehat{\widetilde{\varphi}}^{k-1}_{(j)}\right\|^2_2+r_j\left(\widetilde{\varphi}_{(j)}\right)\right\},\ \forall j\ \&\ k \tag{3.2.3}$$

where the extrapolated point $\widehat{\widetilde{\varphi}}^{k-1}_{(j)}$ is given as $\widehat{\widetilde{\varphi}}^{k-1}_{(j)}=\widetilde{\varphi}^{k-1}_{(j)}+\omega^{k-1}_i\left(\widetilde{\varphi}^{k-1}_{(j)}-\widetilde{\varphi}^{k-2}_{(j)}\right)$, with $\omega^{k-1}_i\geq 0$ is the extrapolation weight, $\hat{\mathrm{g}}^k_j=\bigtriangledown f^k_j\left(\widehat{\widetilde{\varphi}}^{k-1}_{(j)}\right)$ and,

$$f^k_j\left(\widetilde{\varphi}_{(j)}\right)\stackrel{def}{=}f\left(\widehat{\widetilde{\varphi}}^k_{(1)},\widehat{\widetilde{\varphi}}^k_{(2)},\ldots,\widehat{\widetilde{\varphi}}^k_{(j-1)},\widetilde{\varphi}_{(j)},\widehat{\widetilde{\varphi}}^{k-1}_{(j+1)},\ldots,\widehat{\widetilde{\varphi}}^{k-1}_{(s)}\right),\forall\, j\&\ k.$$

The second term on the right hand side, is added on the contrary to standard block coordinate descent algorithm, to make sure that the k^{th} update is not too far from the $(k-1)^{th}$ update in L_2 sense. Before we can do that we need to prove block multi-convexity (lemma **??**) and Lipschitz continuity (lemma **??**) of $f(\varphi_1,\varphi_2,\ldots,\varphi_n)$ and $\bigtriangledown f^k_j\left(\widetilde{\varphi}_{(j)}\right)$ respectively.

Generally, L^{k-1}_j is some constant greater that zero, and plays the role similar to the penalty parameter τ_n in $r_j(\widetilde{\varphi}_{(j)})$, so if the k^{th} update is too far from $(k-1)^{th}$ update in

L_2-sense, our objective would be to penalize it more and control it, so unlike standard group LASSO problem, here we have to take care of two penalty parameters rather than just one. So, we have a more challenging problem to solve, but if scaled properly one can chose constant L_j^{k-1} as a scalar multiplie of τ_n. Let us introduce a new quantity $\eta = L_j^{k-1}/\tau_n$, which is used to explain the rest our algorithm and this is refered to as a dual parameter for our optimization method.

To update (3.2.3) we use the fact that if, r_j can be represented as an indicator function of convex set $\mathcal{D}_j$, *i.e.* $r_j = \delta_{\mathcal{D}_j}(\widetilde{\varphi}_{(j)}) = \mathbb{I}\left(\widetilde{\varphi}_{(j)} \in \mathcal{D}_j\right)$, then $\widetilde{\varphi}_{(j)}^k = \mathcal{P}_{\mathcal{D}_j}\left(\widehat{\widetilde{\varphi}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)$, where $\mathcal{P}_{\mathcal{D}_j}$ is the projection to the set $\mathcal{D}_j$. Since for a group wise LASSO penalty, $r_j\left(\widetilde{\varphi}_{(j)}\right) = \tau_n \left\|\widetilde{\varphi}_{(j)}\right\|_2 / j$, which is equivalent to say that we need our pre-penalized update $\widehat{\widetilde{\varphi}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}$, first scaled down by its length j, and then project it on a surface of the sphere with radius η. And for our group wise LASSO penalty, we define our component wise scaled projection function is, $\mathcal{P}_{\mathcal{D}_j}(t) = \mathrm{sgn}(t)\max(\sqrt{|t|/j} - \eta, 0)$. So the update rule (3.2.3) can be simplified and the following can be used component wise to obtain the j^{th} row,

$$\widetilde{\varphi}_{(j)}^k = \mathrm{sgn}\left(\widehat{\widetilde{\varphi}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)\left(\sqrt{\mathrm{abs}\left(\widehat{\widetilde{\varphi}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)/j} - \eta\right)_+, \ \forall j \ \& \ k \tag{3.2.4}$$

where all the above functions defined on the vector $\widehat{\widetilde{\varphi}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}$ are used component wise. Define the corresponding lower triangular matrix as $\widetilde{\Phi}^k = $ row-bind$(\widetilde{\varphi}_{(1)}^{k\prime}, \widetilde{\varphi}_{(2)}^{k\prime}, \cdots, \widetilde{\varphi}_{(n)}^{k\prime})$ and now let us present the working algorithm for our optimization and following which we also provide a small modification in situations where a subsequent extrapolated update does not reduces the optimizing functional value.

[1] **I**nitialization: $\widetilde{\Phi}^{-1}$ and $\widetilde{\Phi}^0$ lower triangular matrices as first two initial roots with no zero rows **P**refix: $\eta > 0$ and $\epsilon > 0$ prespecified $k = 1, 2, 3, \ldots$ $j = 1, 2, 3, \ldots, n$

$\widehat{\widetilde{\varphi}}_{(j)}^k \longleftarrow$ using (3.2.4) Lower triangular matrix $\widehat{\widetilde{\Phi}}^k$

$\widetilde{\Phi}^{-1} \longleftarrow \widetilde{\Phi}^0$ and $\widetilde{\Phi}^0 \longleftarrow \widehat{\widetilde{\Phi}}^k$ $j = 1, 2, \ldots, n$

$temp_j \longleftarrow \left\|\widehat{\widetilde{\varphi}}_{(j)}^k - \widehat{\widetilde{\varphi}}_{(j)}^{k-1}\right\|_2$ $\max temp < \lambda$ **b***reak* and go to line 18 Go back to line 4 with $k = k+1$ Lower triangular matrix $\widehat{\widetilde{\Phi}}^k$

(M 1) In case of $Q\left(\widetilde{\Phi}^{k}\right) \geq Q\left(\widetilde{\Phi}^{k-1}\right)$ we modify the above algorithm by redoing the k^{th} iteration with $\widehat{\widetilde{\varphi}}_i^{k-1} = \widetilde{\varphi}_i^{k-1}$, *i.e.*, with out extrapolation.

3.3 Numerical investigation

3.3.1 Simulation study

We follow spatial sampling design with an increasing domain asymptotic framework where sample sizes increases in proportion to the area of the sampling region. So we consider $m \times m$ square lattices where $m = 20, 25, 30, 35$ which makes sample sizes $n = 400, 625, 900$, respectively. For each choice we need to consider some true value of $\mathscr{R}(\Sigma)$, rank of Σ, for different n we choose $\mathscr{R}(\Sigma) = 30, 35, 40$. We generate our error term from a mean zero and nonstationary Gaussian process from a covariance function given by (3.1.3) and we consider different choices of $\widetilde{R}(s)$ for example **R**adial **B**asis **F**unction (RBF), **W**endland **B**asis **F**ucntion (WBF), **F**ourier **B**asis **F**unction (FBF) etc. The data has been generated from model (3.1) for all possible combination of m, $\mathscr{R}(\Sigma)$ and $\widetilde{R}(s)$, we generate n data points. From summarizing all the simulation results we believe that the method starts to work better for larger n.

If one considers a dyadic break of the two dimensional spatial domain, and pick centers of each of the regions as their knot points, then the first resolution will have 2^2 knots, second resolution will have 2^4 knots, *i.e.* the k^{th} resolution will have 2^{2k} knot points. We have applied the concept of reversible jumps into our algorithm by considering a starting value of the number of effective knot points. For example lets say we start by considering all the knot points from the first two resolutions effective. After every iteration, we let our model to change by either dropping one of the knots which might have considered to be important earlier or selecting one of the knots which has not been considered to be important earlier.

Lattice size (s)	Local bi-square Basis Function $r_n = 30$	$r_n = 35$	$r_n = 40$	Wendland Basis Function $r_n = 30$	$r_n = 35$	$r_n = 40$
20	**2**8 (1.81)	**2**8 (1.66)	**3**4 (1.60)	**2**9 (1.56)	**3**1 (1.05)	**3**8 (1.05)
25	**3**0 (1.49)	**3**0 (1.26)	**3**6 (1.02)	**3**1 (1.01)	**3**2 (0.91)	**3**9 (1.09)
30	**3**1 (1.05)	**3**2 (0.89)	**4**0 (1.09)	**3**0 (0.91)	**3**4 (0.14)	**4**0 (0.59)
35	**3**0 (0.91)	**3**4 (0.88)	**4**2 (1.05)	**3**0 (0.34)	**3**5 (0.25)	**4**1 (0.29)

Table 3.2 **M**ean (Standard Devation) of 200 Monte Carlo simulations for rank estimation of the nonstationary covariance matrix Σ

3.3.2 Real data examples

The data set we used is part of a group of R data sets for monthly min-max temperatures and precipitation over the period $1895-1997$. It is a subset extracted from the more extensive US data record described in at (www.image.ucar.edu/Data/US.monthly.met). Observed monthly precipitation, min and max temperatures for the conterminous US $1895-1997$. We have taken a subset of the stations in Colorado. Temperature is in degrees C and precipitation is total monthly accumulation in millimeters. Note that minimum (maximum) monthly tempertuare is the mean of the daily minimum (maximum) temperatures. A rectagular lon-lat region $[-109.5, -101] \times [36.5, 41.5]$ larger than the boundary of Colorado comprises approximately 400 stations. Although there are additional stations reported in this domain, stations that only report preicipitation or only report temperatures have been excluded. In addition stations that have mismatches between locations and elevations from the two meta data files have also been excluded. The net result is 367 stations that have colocated temperatures and precipitation. We have used minimum temperature data as the observed process to apply our method and obtain the image plots below.

3.4 Discussion

Our work is quite significant from several perspective, although we would like to point out that it gives a dimension reduction perspective of estimation of low rank covariance matrix. As mentioned earlier Cressie & Johannesson, (2008) pointed out the benefit of

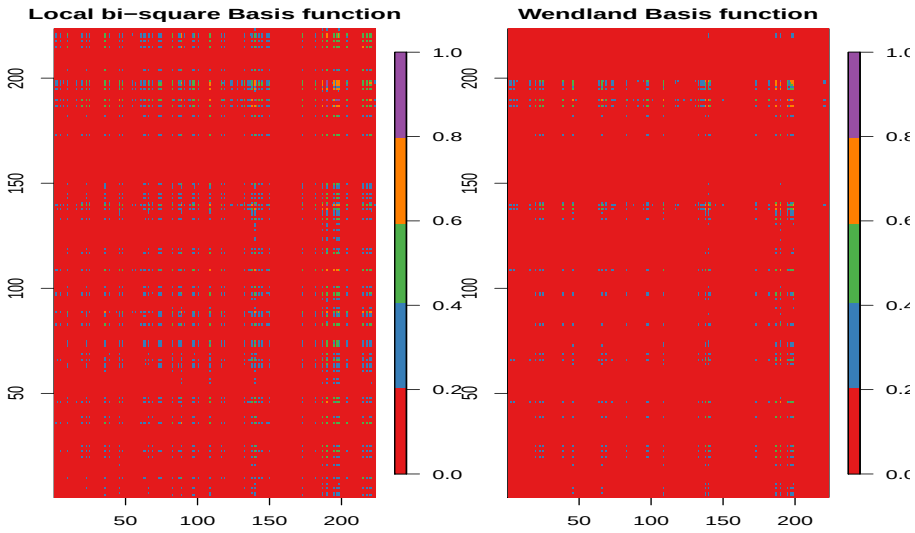

Figure 3.9 Quantile Image plot of $\widehat{\Xi}_{gL}$, the estimated covariance matrix of the observed process

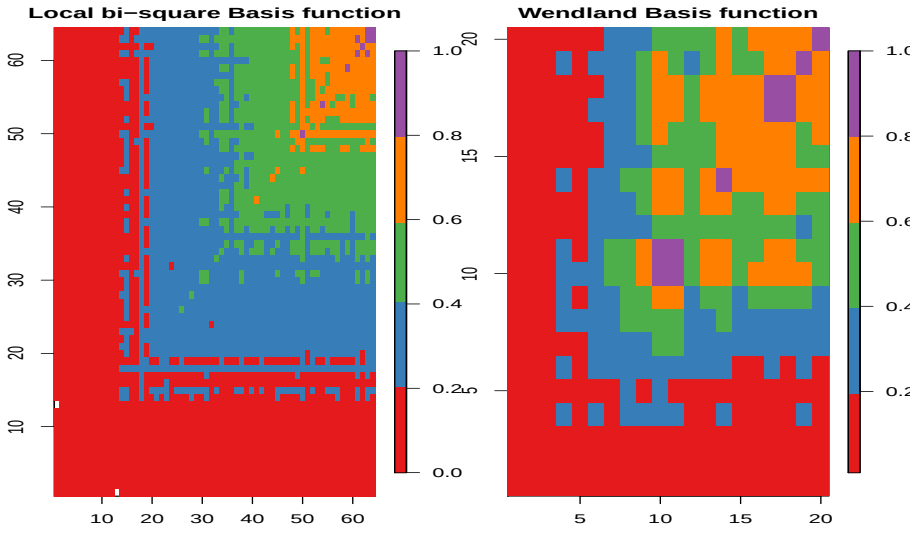

Figure 3.10 Quantile Image plot of $\widehat{\Phi}_{gL}$ estimated covariance matrix of the random effects vector

assuming a fixed but lower rank than the actual dimension of the covariance matrix. They pointed out that inversion time of $n \times n$ covariance matrix, which is $O(n^3)$ can now be reduced to $O(nr^2)$, where r is assumed to be the known fixed rank. A previous knowledge about the value of r is quite hard to believe and our contribution is to figure out a relevant way to get around this. Although at this point we do not claim that we are able to provide an unbiased estimate of rank, but our result does provide consistent estimate of the covariance matrix along with linear model parameters. We also extended the work by Cressie & Johannesson, in the sense that our method allows one to assume that r can vary over n, the sample size, more precisely $r = r_n$ it can increase in a polynomial of n.

Now let us compare our finding with another recent study (Stein, 2015), which provides some examples and discusses scenarios where appoximating a true full rank covariance matrix Ψ_0 with a matrix $\Psi_1 = \rho^2\mathbb{I} + \Upsilon$, where Υ is a low rank matrix, does not reduces the Kulback-Liebler divergence considerably. As necessary, interesting and relevant this may sound, we would like to point out dissimilarties. Firstly, unlike any full rank covariance matrix Ψ_0, we assume true covariance matrix has the structure $\Psi_0 = \rho^2\mathbb{I} + \Upsilon$ and our approach estimates Ψ_0 through estimates of ρ and Υ. Using the concept of capturing spatial dependence through a set of basis functions (Cressie & Johannesson, 2008) our model is further specified by considering the low rank compnent as, $\Upsilon = \widetilde{S}\widetilde{K}\widetilde{S}'$, where $\widetilde{K}$ is a $n \times n$ matrix of rank r_n. As mentioned earlier r_n is a polynomial in n, we would like to refer our readers to assumption (A 1) which says $r_n = Dn^{\gamma} + O(1)$, with $D > 0$ and $\gamma < 2/(15 + 11\alpha)$ with $\alpha > 0$. Although one might feel the necessity of estimating the nuisance parameter α. But let us point out the fact that our results works for any value of $\alpha > 0$. Even if we choose $\alpha = \alpha_n \longrightarrow 0$, $\gamma < 1/7.5$. This implies our finding covers Case 3 and a subset of Case 2 in Stein (2015). In the paper by Stein (2015) it is been pointed out KL divergence do not reduces sufficiently enough under a very special situation of stationary periodic process on line, which can be extended to be a process on surface, although can be quite challenging even for stationary periodic process. On the contrary our finding provides theoritical justification of consistent

estimation of $\Psi_0 = \rho^2 \mathbb{I} + \widetilde{S}\widetilde{K}\widetilde{S}'$ in a more general set up.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Achlioptas, D., & McSherry, F. (2007), "Fast computation of low-rank matrix approximations," *Journal of the ACM*, 54(2), 9.

[2] Anh, V. V. and Lunney, K. E. (1995), "Parameter estimation of random fields with long-range dependence," *Mathematical and computer modelling*, 21(9), 67-77.

[3] Arbenz, P., & Drmac, Z. (2002), "On positive semidefinite matrices with known null space," *SIAM Journal on Matrix Analysis and Applications*, 24(1), 132-149.

[4] Antoniadis, A. and Fan, J. (2001), "Regularization of wavelet approximation (with discussion)," *Journal of the American Statistical Association*, 96, 939-967.

[5] Bai, Z. D., Yin, Y. Q. (1993), "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix," *The annals of Probability*, 1275-1294.

[6] Boissy, Y., Bhattacharyya, B. B., Li, X. and Richardson, G.D. (2005), "Parameter estimates for fractional autoregressive spatial processes", *The Annals of Statistics*, 33, 2553-2567.

[7] Banerjee, A., Dunson, D. B., & Tokdar, S. T. (2012), "Efficient Gaussian process regression for large datasets," *Biometrika*, ass068.

[8] Bühlmann, P., & Van De Geer, S. (2011), "Statistics for high-dimensional data: methods, theory and applications," *Springer Science & Business Media*.

[9] Cressie, N., (1990), "The origins of kriging," *Mathematical geology*, 22(3), 239-252.

[10] Cressie, N. (1993), "Statistics for spatial data," *Wiley series in probability and statistics. Wiley-Interscience New York*, 15, 16.

[11] Cressie, N., Johannesson, G. (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 209-226.

[12] Chu, T., Zhu, J. and Wang, H. (2011), "Penalized maximum likelihood estimation and variable selection in geostatistics," *The Annals of Statistics*, 39(5), 2607-2625.

[13] Cressie, N. and Chan, N. H. (1989), "Spatial modeling of regional variables," *Journal of the American Statistical Association*, 84(406), 393-401.

[14] Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2014), "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets," *arXiv preprint* arXiv:1406.7343.

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

[16] Fan, Y., Li, R. (2012), "Variable selection in linear mixed effects models," *Annals of statistics*, 40(4), 2043.

[17] Frieze, A., Kannan, R., & Vempala, S. (2004), "Fast Monte-Carlo algorithms for finding low-rank approximations," *Journal of the ACM*, 51(6), 1025-1041.

[18] Furrer, R., Genton, M. G., & Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15(3) 502-523.

[19] Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M. and Zhu, J. (2013), "On Estimation and Selection of Autologistic Regression Models via Penalized Pseudolikelihood," *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 429-449.

[20] Gupta, S. (2012), "A note on the asymptotic distribution of LASSO estimator for correlated data," *Sankhya A*, 74, 10-28.

[21] Hastie, T. J. and Tibshirani, R. J. (1990)," Generalized additive models," *CRC Press.*

[22] Haustein, K. O. (2006), "Smoking and poverty," *European Journal of Preventive Cardiology*, 13(3), 312–318.

[23] Hoeting, J. A., Davis, R. A., Merton, A. A. and Thompson, S. E. (2006), "Model selection for geostatistical models," *Ecological Applications*, 16(1), 87-98.

[24] Horn, R. A. and Johnson, C. A. (1985), "Matrix analysis", *Cambridge University Press.*

[25] Hsu, N-J. Hung, H-L. and Chang, Y-M. (2008), "Subset selection for vector autoregressive processes using Lasso," *Computational Statistics and Data Analysis*, 52, 3645-3657.

[26] Huang, H. C. and Chen, C. S. (2007), "Optimal geostatistical model selection," *Journal of the American Statistical Association*, 102(479), 1009-1024.

[27] Huang, H. C., Hsu, N. J., Theobald, D. M. and Breidt, F. J. (2010a), "Spatial Lasso with applications to GIS model selection," *Journal of Computational and Graphical Statistics*, 19(4), 963-983.

[28] Huang, J., Horowitz, J. L. and Wei, F. (2010b), "Variable selection in nonparametric additive models," *The Annals of Statistics*, 38, 2282-2313.

[29] Journée, M., Bach, F., Absil, P. A., & Sepulchre, R. (2010), "Low-rank optimization on the cone of positive semidefinite matrices," *SIAM Journal on Optimization*, 20(5), 2327-2351.

[30] Kneib, T., Hothorn, T. and Tutz, G. (2009), "Variable selection and model choice in geoadditive regression models," *Biometrics*, 65(2), 626-634.

[31] Lai, R.C.S., Huang, H. and Lee, T.C.M. (2012), "Fixed and random effects selection in nonparametric additive mixed models," *Electronic Journal of Statistics*, 6, 810-842.

[32] Lin, Y. and Zhang, H. (2006), "Component selection and smoothing in multivariate nonparametric regression," *The Annals of Statistics*, 34, 2272-2297.

[33] Marshall, A. W., & Olkin, I. (1979), "Theory of Majorization and its Applications," *Academic, New York,* 16, 4-93.

[34] Meier, L., Van De Geer, S. and Bühlmann, P. (2009), "High-dimensional additive modeling," *The Annals of Statistics*, 37, 3779-3821.

[35] Nardi, Y. and Rinaldo, A. (2011), "Autoregressive process modeling via the Lasso procedure," *Journal of Multivariate Analysis*, 102, 528-549.

[36] Nishii, R. (1984), "Asymptotic properties of criteria for selection of variables in multiple regression," *The Annals of Statistics*, 12, 758-765.

[37] Nychka, D., Ellner, S., Haaland, P. D., & O'Connell, M. A. (1996), "FUNFITS, data analysis and statistical tools for estimating functions," *Raleigh: North Carolina State University.*

[38] Nychka, D., Wikle, C., & Royle, J. A. (2002), "Multiresolution models for nonstationary spatial covariance functions," *Statistical Modelling*, 2(4), 315-331.

[39] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., & Sain, S. (2015), "A multiresolution gaussian process model for the analysis of large spatial datasets," *Journal of Computational and Graphical Statistics*, 24(2), 579-599.

[40] Peng, C., Wu, C. F. J., (2013), "On the choice of nugget in kriging modeling for deterministic computer experiments," *Journal of Computational and Graphical Statistics*, 23(1), 151-168.

[41] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), "Sparse additive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030.

[42] Reyes, P. E., Zhu, J. and Aukema, B. H. (2012), "Selection of Spatial-Temporal Lattice Models: Assessing the Impact of Climate Conditions on a Mountain Pine Beetle Outbreak," *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 508-525.

[43] Schott, J. R. (2005), "Matrix analysis for statistics."

[44] Simon, B., (1979), "Trace ideals and their applications (Vol. 35)," *Cambridge: Cambridge University Press.*

[45] Stein, M. L. (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, 22(4), 866-885.

[46] Stein, M. L. (2014), " Limitations on low rank approximations for covariance matrices of spatial data. Spatial Statistics," *Spatial Statistics*, 8 (2014): 1-19.

[47] Schumaker, L. (2007), "Spline functions: basic theory," *Cambridge University Press*

[48] SEER (Census 2000), " Surveillance, Epidemiology, and End Results" *SEER, cf.* - www.seer.cancer.gov

[49] Stein, M. L. (1999), "Interpolation of Spatial Data", Springer.

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

[51] Tzeng, S., Huang, H. C. (2015), "Non-stationary Multivariate Spatial Covariance Estimation via Low-Rank Regularization," *Statistical Sinica*, 26, 151-172.

[52] Van der Vart, A. W. and Wellner, J.A. (1996), "Weak Convergence and Emperical Processes:With Application to Statistics," *Springer, New York.* MR 13851671.

[53] Wang, H. and Zhu, J. (2009), "Variable selection in spatial regression via penalized least squares," *The Canadian Journal of Statistics*, 37, 607-624.

[54] Wang, H., Li, G. and Tsai, C-L. (2007), "Regression coefficient and autoregressive order shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, 69, 63-78.

[55] Wendland, H. (2005). " Scattered data approximation" (Vol. 17). *Cambridge University Press*.

[56] Xu, G., Xiang, Y., Wang, S. and Lin, Z. (2012), "Regularization and variable selection for infinite variance autoregressive models", *Journal of Statistical Planning and Inference*, 142, 2545-2553.

[57] Xu, Y., Yin, W. (2013), "A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, 6(3), 1758-1789.

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

[59] Zhang, C. and Huang, J. (2008), "The sparsity and bias of the Lasso selection in high-dimensional linear regression," *The Annals of Statistics*, 46, 1567-1594.

[60] Zhao, Y. B. (2012), "An approximation theory of matrix rank minimization and its application to quadratic equations," *Linear Algebra and its Applications*, 437(1), 77-93.

[61] Zhu, J., Huang, H.-C., and Reyes, P.E. (2010), "On selection of spatial linear models for lattice data", *Journal of the Royal Statistical Society Series B*, 72, 389-402.