DISSERTATION

PARAMETRIC AND SEMIPARAMETRIC MODEL ESTIMATION AND SELECTION IN GEOSTATISTICS

Submitted by
Tingjin Chu
Department of Statistics

In partial fulfillment of the requirements
for the Degree of Doctor of Philosophy
Colorado State University
Fort Collins, Colorado
Summer 2012

Doctoral Committee:

Advisor: Haonan Wang
Co-Advisor: Jun Zhu

Mary Meyer
J. Rockey Luo

ABSTRACT

PARAMETRIC AND SEMIPARAMETRIC MODEL ESTIMATION AND SELECTION IN GEOSTATISTICS

This dissertation is focused on geostatistical models, which are useful in many scientific disciplines, such as climatology, ecology and environmental monitoring. In the first part, we consider variable selection in spatial linear models with Gaussian process errors. Penalized maximum likelihood estimation (PMLE) that enables simultaneous variable selection and parameter estimation is developed and for ease of computation, PMLE is approximated by one-step sparse estimation (OSE). To further improve computational efficiency particularly with large sample sizes, we propose penalized maximum covariance-tapered likelihood estimation ($\mathrm{PMLE_T}$) and its one-step sparse estimation ($\mathrm{OSE_T}$). General forms of penalty functions with an emphasis on smoothly clipped absolute deviation are used for penalized maximum likelihood. Theoretical properties of PMLE and OSE, as well as their approximations $\mathrm{PMLE_T}$ and $\mathrm{OSE_T}$ using covariance tapering are derived, including consistency, sparsity, asymptotic normality, and the oracle properties. For covariance tapering, a by-product of our theoretical results is consistency and asymptotic normality of maximum covariance-tapered likelihood estimates. Finite-sample properties of the proposed methods are demonstrated in a simulation study and for illustration, the methods are applied to analyze two real data sets.

In the second part, we develop a new semiparametric approach to geostatistical modeling and inference. In particular, we consider a geostatistical model with additive components, where the covariance function of the spatial random error is not pre-specified and thus flexible. A novel, local Karhunen-Loève expansion is developed and a likelihood-based method devised for estimating the model parameters.

In addition, statistical inference, including spatial interpolation and variable selection, is considered. Our proposed computational algorithm utilizes Newton-Raphson on a Stiefel manifold and is computationally efficient. A simulation study demonstrates sound finite-sample properties and a real data example is given to illustrate our method. While the numerical results are comparable to maximum likelihood estimation under the true model, our method is shown to be more robust against model misspecification and is computationally far more efficient for larger sample sizes. Finally, the theoretical properties of the estimates are explored and in particular, a consistency result is established.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Haonan Wang, my co-advisor Dr. Jun Zhu and my committee members Dr. Mary Meyer and Dr. J. Rockey Luo. Without their supports on this dissertation, It would be impossible for me to finish this dissertation. I especially want to thank my advisors for supervising my statistical research, such as research philosophy and approaches, as well as on non-statistical research, such as paper writing. I also want to thank for Dr. Jennifer Hoeting for providing the California lizards data.

I would also like to thank the department for the financial support during my Master and Ph.D studies in the past five years. Moreover, I would like to thank faculties for the statistical educations, both in theory and in practice, and the fellow students for the friendship.

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Geostatistics

Geostatistical models are widely used to analyze spatial data in many scientific disciplines, such as climatology, ecology and environmental monitoring. For example, at weather stations across Colorado, precipitation data have been collected over time. An interesting topic is to investigate the relationship between precipitation and covariates such as elevation, slope, aspect, and other satellites imagery information, and which covariates have a significant effect on precipitation. Moreover, it is of interest to quantify precipitation in areas where there are no data. For this type of spatial data, geostatistical models provide a useful tool set to help researchers to address the scientific questions of interest.

A geostatistical model has one significant difference from more traditional statistical models in that observations are spatially correlated. Suppose that there is rain at weather station A, then at a nearby weather station B, rain is quite likely. However, at weather station C far away from weather station A, it is less likely to have rain. In statistical terms, precipitation at weather stations A and B is more correlated than that at weather stations A and C. This correlation is related to the spatial closeness of these weather stations such that the correlation between two observations is determined by the locations of the corresponding observations, which is also known as spatial correlation.

For a random process $\varepsilon(\boldsymbol{s})$ in a spatial domain of interest $R \in \mathbb{R}^d$, the covariance between $\varepsilon(\boldsymbol{s}_1)$ and $\varepsilon(\boldsymbol{s}_2)$, $\mathrm{cov}\{\varepsilon(\boldsymbol{s}_1), \varepsilon(\boldsymbol{s}_2)\}$, is used to measure the spatial correlation between $\varepsilon(\boldsymbol{s}_1)$ and $\varepsilon(\boldsymbol{s}_2)$, where $\boldsymbol{s}_1, \boldsymbol{s}_2 \in R$. For modeling a spatial process, second-order stationarity and isotropy are often assumed. A random process $\varepsilon(\boldsymbol{s})$ is second-order stationarity if it satisfies the conditions

$$E\{\varepsilon(\boldsymbol{s})\} = \mu, \text{ for all } \boldsymbol{s} \in R,$$

$$Var\{\varepsilon(\boldsymbol{s})\} = \text{constant} < \infty,$$

and

$$\text{cov}\{\varepsilon(\boldsymbol{s}_1), \varepsilon(\boldsymbol{s}_2)\} = \gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2), \text{ for all } \boldsymbol{s}_1, \boldsymbol{s}_2 \in R.$$

That is, for a second-order stationary process, the expected value stays the same at different locations, and the covariance is only a function of distance, and direction between two locations. Moreover, if $\gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2)$ is only a function of distance $\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$, then the random process $\boldsymbol{\varepsilon}(\boldsymbol{s})$ is said to be isotropic (Chapter 2.3 in Cressie 1993).

There are various ways to model the covariance function $\gamma(\cdot)$. A commonly-used family of covariance functions is the Matérn class with covariance functions in the form of

$$\gamma(d) = \sigma^2(1-c)\{2^{\nu-1}\Gamma(\nu)\}^{-1}\left(2\nu^{1/2}d/r\right)^{\nu}\mathcal{K}_\nu\left(2\nu^{1/2}d/r\right),$$

where $\sigma^2 > 0$ is a variance, $c \in [0,1]$ is a nugget proportion such that $c\sigma^2$ is the nugget effect representing variation at small lag distance, $r > 0$ is a range parameter controlling the rate of autocorrelation decay with lag distance, $\nu > 0$ is a shape parameter controlling the smoothness of the spatial process, and $\mathcal{K}_\nu(\cdot)$ is a modified Bessel function of the second kind of order ν.

The exponential covariance function, $\gamma(d) = \sigma^2(1-c)\exp(-d/r)$, is commonly used and is a special case of of the Matérn class with $\nu = 1/2$. Another well-known covariance function is the Gaussian covariance function $\gamma(d) = \sigma^2(1-c)\exp(-d^2/r^2)$. Although it does not belong to the Matérn class, the Matérn covariance function converges to Gaussian covariance function, as $\nu \to \infty$.

1.2 Topics and Our Approaches

In this dissertation, we will consider a spatial linear model for a spatial process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in R\}$:

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T \boldsymbol{\beta} + \varepsilon(\boldsymbol{s}), \tag{1.1}$$

where $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))^T$ is a $p \times 1$ vector of covariates at location $\boldsymbol{s}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. Here, the error process $\{\varepsilon(\boldsymbol{s}) : \boldsymbol{s} \in R\}$ is used to model the spatial correlation, and is assumed to be a second-order stationary and isotropic Gaussian process with mean zero and a covariance function $\gamma(\boldsymbol{s}, \boldsymbol{s}')$, where $\boldsymbol{s}, \boldsymbol{s}' \in R$. We focus on three main topics for spatial linear model: parameter estimation, variable selection, and spatial prediction (also known as Kriging) at unsampled locations.

First, for parameter estimation, maximum likelihood estimates are often used. Let the response variable be $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_N))^T$ at N sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ be an $N \times p$ design matrix of covariates, $\boldsymbol{\Gamma} = [\gamma(\boldsymbol{s}_i, \boldsymbol{s}_{i'})]_{i,i'=1}^N$ be the covariance matrix for $\boldsymbol{y}$. The log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\Gamma}; \boldsymbol{y}, \boldsymbol{X}) &= -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}| \\ &\quad -(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{1.2}$$

The maximum likelihood estimate (MLE) can be obtained by maximizing (1.2).

One challenge for parameter estimation is the computational burden to invert the covariance matrix $\boldsymbol{\Gamma}$. Since $\boldsymbol{\Gamma}$ is an $N \times N$ matrix, the computational complexity for the inversion of $\boldsymbol{\Gamma}$ is $O(N^3)$, which is time consuming, if not infeasible, when sample size becomes larger. One way to overcome this challenge is to approximate the covariance matrix by a sparse matrix and take advantage of the fast computing algorithm to invert the sparse matrix (Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009). The method is known as covariance tapering, such that if the distance between two observations is large, their correlation should be small. Even if we treat these correlations as zero, the parameter estimates will still be relatively accurate. Based on this idea, a tapering function will be applied to re-scale the covariance matrix to obtain a tapered covariance matrix, which is a sparse matrix and thus faster to compute.

A second challenge for parameter estimation is that the underlying covariance structure is often unknown *a priori*. Thus, we take a nonparametric approach, namely, Karhunen-Loève expansion, to model the error process. By incorporating this nonparametric form into the log-likelihood function, parameter estimates can be obtained. However, for a geostatistical model, there is often one realization of the random field and the resulting estimate is not consistent. To overcome this difficulty, we introduce a novel local Karhunen-Loève expansion and consequently, consistent regression parameter estimates and covariance function estimates are obtained.

Second, variable selection is considered for identifying the best subsets among all possible subsets of covariates. For linear regression with independent error, variable selection has been widely studied, and various methods are available. One popular variable selection technique is stepwise selection procedure, such as forward selection and backward elimination (Draper and Smith, 1998). This method is carried out by an automatic procedure and fast to compute. However, its theoretical property is hard to understand. Another way is to use information discrepancy-based methods, such as Kolmogorov-Smirnov, Kullback-Leibler, or Hellinger discrepancy (Linhart and Zucchini, 1986), which is often time consuming when the number of covariates is large.

Recently, penalized methods have been developed for variable selection, which enable parameter estimation and variable selection simultaneously. However, in geostatistics, little has been studied about such penalized methods, especially from a theoretical perspective. In this dissertation, we focus on spatial linear models and establish consistency, asymptotical normality and oracle property of penalized maximum likelihood estimates for smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li, 2001) under certain conditions. First, we define the penalized log-likelihood function $Q(\boldsymbol{\beta}, \boldsymbol{\Gamma}; \boldsymbol{y}, \boldsymbol{X})$ as

$$Q(\boldsymbol{\beta}, \boldsymbol{\Gamma}; \boldsymbol{y}, \boldsymbol{X}) = \ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) - N \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{1.3}$$

where $p_\lambda(\cdot)$ is the known penalty function, and in this dissertation, we focus on SCAD penalty function.

By maximizing (1.3), variable selection and parameter estimation are obtained simultaneously. However, the inversion of $\boldsymbol{\Gamma}$ is also needed in order to maximize (1.3) and we face the similar computational burden as parameter estimation. One way to address this computational burden is covariance tapering and the performance of variable selection under covariance tapering method will be investigated.

Last, for spatial prediction (Kriging), we consider best linear unbiased predictor (BLUP) at unsampled locations. We begin with simple Kriging. Let $\boldsymbol{\varepsilon}(\boldsymbol{s})$ be a spatial process with $E\varepsilon(\boldsymbol{s}) = 0$, for all $\boldsymbol{s} \in R$. For an unsampled location $\boldsymbol{s}_0$, the BLUP is $\widehat{\varepsilon}(\boldsymbol{s}_0) = \boldsymbol{w}\boldsymbol{\varepsilon}$, where $\boldsymbol{w} = \boldsymbol{c}_0^T \boldsymbol{\Gamma}_0^{-1}$, $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_N))^T$, $\boldsymbol{c}_0$ is an $N \times 1$ vector whose ith component is $\gamma(\boldsymbol{s}_0, \boldsymbol{s}_i)$ and $\boldsymbol{\Gamma}_0$ is the variance-covariance for $\boldsymbol{\varepsilon}$. For universal Kriging in spatial linear model (1.1), the BLUP is

$$\widehat{y}(\boldsymbol{s}_0) = \boldsymbol{x}^T(\boldsymbol{s}_0)\widehat{\boldsymbol{\beta}} + \boldsymbol{c}_0^T \boldsymbol{\Gamma}_0^{-1}(\boldsymbol{y} - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\Gamma}_0^{-1}\boldsymbol{y}$ (section 3.4.5 in Cressie 1993).

Since $\boldsymbol{c}_0$, $\boldsymbol{\Gamma}_0$ are unknown in practice, and estimates $\widehat{\boldsymbol{c}}_0$ and $\widehat{\boldsymbol{\Gamma}}_0$ are needed. This is part of the parameter estimation problem as mentioned above. It is worth mentioning that if the sample size is large, covariance tapering method is needed in order to compute $\boldsymbol{\Gamma}_0^{-1}$ efficiently. Moreover, if the underlying covariance structure is unknown, the covariance function estimates from local Karhunen-Loève expansion may be more suitable.

The reminder of the dissertation is organized as follows. In Chapter 2, the covariance function of the error process is assumed to be parametric, and penalized maximum covariance-tapered likelihood is proposed for simultaneous parameter estimation and variable selection. In Chapter 3, we take a semiparametric approach for spatial linear model to obtain regression parameter estimates and covariance function estimates. These parameter estimates are then used in spatial prediction and variable selection. Summary and future work will be included in Chapter 4.

Chapter 2

REGULARIZED APPROACH TO VARIABLE SELECTION [1]

2.1 Introduction

Geostatistical models are popular tools for the analysis of spatial data in many disciplines. It is often of interest to estimate model parameters based on data at sampled locations and perform spatial interpolation (also known as Kriging) of a response variable at unsampled locations within a spatial domain of interest (Cressie, 1993; Stein, 1999; Schabenberger and Gotway, 2005). In addition, a practical issue that often arises is how to select the best model or a best subset of models among many competing ones (Hoeting et al., 2006). Here we focus on selecting covariates in a spatial linear model, which we believe is a problem that is underdeveloped in both theory and methodology despite its importance in geostatistics. The spatial linear model for a response variable under consideration has two additive components: a fixed linear regression term and a stochastic error term. We assume that the error term follows a Gaussian process with mean zero and a covariance function that accounts for spatial dependence. Our chief objective is to develop a set of new methods for the selection of covariates and establish their asymptotic properties. Moreover, we devise efficient algorithms for computation, making these methods feasible for practical usage.

[1] Part of this chapter is based on Tingjin Chu's Master Thesis and the paper "Penalized Maximum Likelihood Estimation and Variable Selection in Geostatistics" published in the Annals of Statistics, 39, 2607-2625.

For linear regression with independent errors, variable selection has been widely studied in the literature. The more traditional methods often involve hypothesis testing such as F-tests in a stepwise selection procedure (Draper and Smith, 1998). An alternative approach is to select models using information discrepancy such as Kolmogorov-Smirnov, Kullback-Leibler, or Hellinger discrepancy (Linhart and Zucchini, 1986). In recent years, penalized methods are becoming increasingly popular for variable selection. For example, Tibshirani (1996) developed a least absolute shrinkage and selection operator (LASSO), whereas Fan and Li (2001) proposed a nonconcave penalized likelihood method with smoothly clipped absolute deviation (SCAD) penalty. Efron et al. (2004) devised least angle regression (LARS) algorithms, which allow computing all LASSO estimates along a path of its tuning parameters at a low computational order. More recently, Zou (2006) improved LASSO and the resulting adaptive LASSO enjoys the oracle properties as SCAD, in terms of selecting the true model. Zou and Li (2008) proposed one-step sparse estimation in the nonconcave penalized likelihood approach, which retains the oracle properties and utilizes LARS algorithms.

For spatial linear models in geostatistics, in contrast, statistical methods for principled selection of covariates are limited. Hoeting et al. (2006) suggested Akaike's information criterion (AIC) with a finite-sample correction for variable selection. Like information-based selection in general, computation can be costly especially when the number of covariates and/or the sample sizes are large. Thus, these authors considered only a subset of the covariates that may be related to the abundance of orange-throated whiptail lizard in southern California, in order to make it tractable to evaluate their AIC-based model selection. Huang and Chen (2007) developed a model selection criterion in geostatistics, but for the purpose of Kriging rather than selection of covariates. Further, Wang and Zhu (2009) proposed penalized least squares (PLS) for a spatial linear model where the error process is assumed to be strong mixing without the assumption of Gaussian process. This method includes spatial autocorrelation only indirectly in the sense that the objective function involves a sum of squared errors ignoring spatial dependence. A spatial block bootstrap is then used to account for spatial dependence when estimating the variance of PLS estimates.

Here we take an alternative, parametric approach and assume that the errors in the spatial linear model follow a Gaussian process. Our main innovation here is to incorporate spatial dependence directly into a penalized likelihood function and achieve greater efficiency in the resulting penalized maximum likelihood estimates (PMLE). Unlike computation of PLS estimates which is on the same order as ordinary least squares estimates, however, penalized likelihood function for a spatial linear model will involve operations of a covariance matrix of the same size as the number of observations. Thus the computational cost can be prohibitively high as the sample size becomes large. It is essential that our new methods address this issue. To that end, we utilize one-step sparse estimation (OSE) and LARS algorithms in the computation of PMLE to gain computational efficiency. In addition, we explore covariance tapering, which further reduces computational cost by replacing the exact covariance matrix with a sparse one (Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009). We establish the asymptotic properties of both PMLE and OSE, as well as their covariance-tapered counterparts. As a by-product, we establish new results for covariance-tapered MLE which, to the best of our knowledge, have not been established before and can be of independent interest.

The remainder of the chapter is organized as follows. In Section 2.2, we develop PMLE that enables simultaneous variable selection and parameter estimation, as well as an approximation of the PMLE by one-step sparse estimation to enhance computational efficiency. For further computational improvement, we consider penalized maximum covariance-tapered likelihood estimation ($\mathrm{PMLE_T}$) and its one-step sparse estimation ($\mathrm{OSE_T}$) in Section 2.3. We establish asymptotic properties of PMLE and OSE in Section 2.4 and those of $\mathrm{PMLE_T}$ and $\mathrm{OSE_T}$ under covariance tapering in Section 2.5. In Section 2.6, finite-sample properties of the proposed methods are investigated in a simulation study and for illustration, the methods are applied to analyze two real data sets. A brief summary and discussion is given in Section 2.7, whereas technical proofs and details are given in Section 2.8.

2.2 Penalized Maximum Likelihood

2.2.1 Spatial Linear Model and Maximum Likelihood Estimation

For a spatial domain of interest R in $\mathbb{R}^d$, we consider a spatial process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in R\}$ such that

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta} + \varepsilon(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))^T$ is a $p \times 1$ vector of covariates at location $\boldsymbol{s}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. We assume that the error process $\{\varepsilon(\boldsymbol{s}) : \boldsymbol{s} \in R\}$ is a Gaussian process with mean zero and a covariance function

$$\gamma(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) = \mathrm{cov}\{\varepsilon(\boldsymbol{s}), \varepsilon(\boldsymbol{s}')\}, \tag{2.2}$$

where $\boldsymbol{s}, \boldsymbol{s}' \in R$ and $\boldsymbol{\theta}$ is a $q \times 1$ vector of covariance function parameters.

Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ denote N sampling sites in R. Let $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_N))^T$ denote an $N \times 1$ vector of response variables and $\boldsymbol{x}_j = (x_j(\boldsymbol{s}_1), \ldots, x_j(\boldsymbol{s}_N))^T$ denote an $N \times 1$ vector of the jth covariate with $j = 1, \ldots, p$, at the N sampling sites. Further, let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ denote an $N \times p$ design matrix of covariates and $\boldsymbol{\Gamma} = [\gamma(\boldsymbol{s}_i, \boldsymbol{s}_{i'}; \boldsymbol{\theta})]_{i,i'=1}^N$ denote an $N \times N$ covariance matrix. In this chapter, we consider general forms for the the covariance matrix $\boldsymbol{\Gamma}$ and describe suitable regularity conditions in Sections 2.4 and 2.5. By (2.1) and (2.2), we have

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}). \tag{2.3}$$

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ denote a $(p + q) \times 1$ vector of model parameters consisting of both regression coefficients $\boldsymbol{\beta}$ and covariance function parameters $\boldsymbol{\theta}$. By (2.3), the log-likelihood function of $\boldsymbol{\eta}$ is

$$\begin{aligned} \ell(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) &= -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}| \\ &\quad -(1/2)(\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}). \end{aligned} \tag{2.4}$$

Let $\widehat{\boldsymbol{\eta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\eta}}\{\ell(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X})\}$ denote the maximum likelihood estimate (MLE) of $\boldsymbol{\eta}$.

For a real-valued function $f(\boldsymbol{a})$ where $\boldsymbol{a} = (\boldsymbol{a}_1^T, \boldsymbol{a}_2^T)^T$, $\boldsymbol{a}_i \in \mathbb{R}^{r_i}$, $r_i \geq 1$, $r_1 + r_2 = r$, and $i = 1, 2$, let $f'(\boldsymbol{a}) = \partial f(\boldsymbol{a})/\partial \boldsymbol{a}$ denote an $r \times 1$ vector of first-order derivatives with respect to $\boldsymbol{a}$ and $f'(\boldsymbol{a}_i) = \partial f(\boldsymbol{a})/\partial \boldsymbol{a}_i$ denote an $r_i \times 1$ vector of first-order derivatives with respect to $\boldsymbol{a}_i$, $i = 1, 2$. Let $f''(\boldsymbol{a}) = \partial^2 f(\boldsymbol{a})/\partial \boldsymbol{a}\partial \boldsymbol{a}^T$ denote an $r \times r$ matrix of second-order derivatives with respect to $\boldsymbol{a}$ and $f''(\boldsymbol{a}_i, \boldsymbol{a}_j) = \partial^2 f(\boldsymbol{a})/\partial \boldsymbol{a}_i \partial \boldsymbol{a}_j^T$ denote an $r_i \times r_j$ matrix of second-order derivatives with respect to $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$, $i, j = 1, 2$.

From (2.4), we have $\ell'(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$ and the kth element of $\ell'(\boldsymbol{\theta})$ is $-(1/2)\text{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) - (1/2)(\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Gamma}^k(\boldsymbol{y} - \boldsymbol{X\beta})$, where $\boldsymbol{\Gamma}_k = \partial\boldsymbol{\Gamma}/\partial\theta_k$ and $\boldsymbol{\Gamma}^k = \partial\boldsymbol{\Gamma}^{-1}/\partial\theta_k = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}$ for $k = 1, \ldots, q$. Moreover, $\ell''(\boldsymbol{\beta}, \boldsymbol{\beta}) = -\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X}$, the kth column of $\ell''(\boldsymbol{\beta}, \boldsymbol{\theta})$ is $\boldsymbol{X}^T\boldsymbol{\Gamma}^k(\boldsymbol{y} - \boldsymbol{X\beta})$, and the (k, k')th entry of $\ell''(\boldsymbol{\theta}, \boldsymbol{\theta})$ is $-(1/2)\Big\{\text{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{kk'} + \boldsymbol{\Gamma}^k\boldsymbol{\Gamma}_{k'}) + (\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Gamma}^{kk'}(\boldsymbol{y} - \boldsymbol{X\beta})\Big\}$, where $\boldsymbol{\Gamma}_{kk'} = \partial^2\boldsymbol{\Gamma}/\partial\theta_k\partial\theta_{k'}$ and $\boldsymbol{\Gamma}^{kk'} = \partial^2\boldsymbol{\Gamma}^{-1}/\partial\theta_k\partial\theta_{k'} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'} + \boldsymbol{\Gamma}_{k'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{kk'})\boldsymbol{\Gamma}^{-1}$ for $k, k' = 1, \ldots, q$. Since $E\{-\ell''(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \boldsymbol{0}$, the information matrix of $\boldsymbol{\eta}$ is $\boldsymbol{I}(\boldsymbol{\eta}) = \text{diag}\{\boldsymbol{I}(\boldsymbol{\beta}), \boldsymbol{I}(\boldsymbol{\theta})\}$, where

$$\boldsymbol{I}(\boldsymbol{\beta}) = E\{-\ell''(\boldsymbol{\beta}, \boldsymbol{\beta})\} = \boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \tag{2.5}$$

and the (k, k')th entry of

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\{-\ell''(\boldsymbol{\theta}, \boldsymbol{\theta})\} \tag{2.6}$$

is $t_{kk'}/2$ with $t_{kk'} = \text{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'}) = \text{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{k'})$; see Mardia and Marshall (1984).

2.2.2 Penalized Maximum Likelihood Estimation

We define a penalized log-likelihood function as

$$Q(\boldsymbol{\eta}) = \ell(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) - N \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.7}$$

where $\ell(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X})$ is the log-likelihood function defined in (2.4) and $p_\lambda(\cdot)$ is a prespecified penalty function with a tuning parameter λ. We let $\widehat{\boldsymbol{\eta}}_{\text{PMLE}} = \arg\max_{\boldsymbol{\eta}}\{Q(\boldsymbol{\eta})\}$ denote the penalized maximum likelihood estimate (PMLE) of $\boldsymbol{\eta}$.

Let $\boldsymbol{\phi}(\boldsymbol{\beta}) = (p'_\lambda(|\beta_1|)\text{sgn}(\beta_1), \ldots, p'_\lambda(|\beta_p|)\text{sgn}(\beta_p))^T$ and $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \text{diag}\{p''_\lambda(|\beta_1|), \ldots, p''_\lambda(|\beta_p|)\}$. Then $Q'(\boldsymbol{\beta}) = \ell'(\boldsymbol{\beta}) - N\boldsymbol{\phi}(\boldsymbol{\beta})$ and $Q'(\boldsymbol{\theta}) = \ell'(\boldsymbol{\theta})$. Moreover, $Q''(\boldsymbol{\beta}, \boldsymbol{\beta}) = \ell''(\boldsymbol{\beta}, \boldsymbol{\beta}) - N\boldsymbol{\Phi}(\boldsymbol{\beta})$, $Q''(\boldsymbol{\beta}, \boldsymbol{\theta}) = \ell''(\boldsymbol{\beta}, \boldsymbol{\theta})$, and $Q''(\boldsymbol{\theta}, \boldsymbol{\theta}) = \ell''(\boldsymbol{\theta}, \boldsymbol{\theta})$. Thus $E\{-Q''(\boldsymbol{\eta})\} = \text{diag}\{\boldsymbol{I}(\boldsymbol{\beta}) + N\boldsymbol{\Phi}(\boldsymbol{\beta}), \boldsymbol{I}(\boldsymbol{\theta})\}$, where $\boldsymbol{I}(\boldsymbol{\beta})$ and $\boldsymbol{I}(\boldsymbol{\theta})$ are given in (2.5) and (2.6).

For penalty functions, we mainly consider smoothly clipped absolute deviation (SCAD) defined as

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta|, & \text{if} \quad |\beta| \le \lambda, \\ \lambda^2 + (a-1)^{-1}(a\lambda|\beta| - \beta^2/2 - a\lambda^2 + \lambda^2/2), & \text{if} \quad \lambda < |\beta| \le a\lambda, \\ (a+1)\lambda^2/2, & \text{if} \quad |\beta| > a\lambda \end{cases} \tag{2.8}$$

for some $a > 2$ (Fan, 1997). For iid error in standard linear regression, variable selection and parameter estimation under the SCAD penalty are shown to possess three desirable properties: unbiasedness, sparsity and continuity (Fan and Li, 2001). For spatial linear regression (2.1), these properties continue to hold for SCAD penalty following arguments similar to those in Wang and Zhu (2009).

To compute PMLE under the SCAD penalty, Fan and Li (2001) proposed a locally quadratic approximation (LQA) of the penalty function and a Newton-Raphson algorithm. Although fast, a drawback of the LQA algorithm is that once a regression coefficient is shrunk to zero, it remains to be zero in the remainder iterations. More recently, Zou and Li (2008) developed a unified algorithm to improve computational efficiency, which, unlike LQA algorithm, is based on locally linear approximation (LLA) of the penalty function. Moreover, Zou and Li (2008) proposed one-step LLA estimation that approximates the solution after just one iteration in a Newton-Raphson-type algorithm starting at the MLE. We extend this one-step LLA estimation to approximate PMLE for the spatial linear model as follows.

Algorithm 1. At the initialization step, we let $\boldsymbol{\eta}^{(0)} = \widehat{\boldsymbol{\eta}}_{\text{MLE}}$ with $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{\text{MLE}}$ and $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\text{MLE}}$. Then, we update $\boldsymbol{\beta}$ by maximizing

$$Q^*(\boldsymbol{\beta}) = -(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - N \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j| \tag{2.9}$$

with respect to $\boldsymbol{\beta}$, where the first term is from (2.4) evaluated at $\boldsymbol{\theta}^{(0)}$ and the second term is an LLA of the penalty function in (2.7). The resulting one-step sparse estimate (OSE) of $\boldsymbol{\beta}$ is denoted as $\widehat{\boldsymbol{\beta}}_{\text{OSE}}$. Although not necessary, we may update $\boldsymbol{\theta}$ by maximizing (2.4) evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\text{OSE}}$ with respect to $\boldsymbol{\theta}$. The resulting OSE of $\boldsymbol{\theta}$ is denoted as $\widehat{\boldsymbol{\theta}}_{\text{OSE}}$. We let $\widehat{\boldsymbol{\eta}}_{\text{OSE}} = (\widehat{\boldsymbol{\beta}}_{\text{OSE}}^T, \widehat{\boldsymbol{\theta}}_{\text{OSE}}^T)^T$ denote the OSE of $\boldsymbol{\eta}$, which approximates $\widehat{\boldsymbol{\eta}}_{\text{PMLE}}$. As we will show in Section 2.4, by using MLE as the initial values, consistency of $\widehat{\boldsymbol{\eta}}_{\text{OSE}}$ is ensured.

Finally, two tuning parameters, λ and a, in the SCAD penalty (2.8) need to be estimated. For computational ease, we fix $a = 3.7$ as recommended by Fan and Li (2001). To determine λ, we use Bayesian information criterion (BIC); see Wang et al. (2007b). In particular, let

$$\widehat{\sigma}^2(\lambda) = N^{-1}\{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T \boldsymbol{\Gamma}\{\widehat{\boldsymbol{\theta}}(\lambda)\}^{-1}\{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}, \tag{2.10}$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\theta}}(\lambda)$ are the PMLE obtained for a given λ and let

$$\text{BIC}(\lambda) = N \log\{\widehat{\sigma}^2(\lambda)\} + k(\lambda)\log(N), \tag{2.11}$$

where $k(\lambda)$ is the number of non-zero regression coefficients (Wang et al., 2007a). Thus an estimate of λ is $\widehat{\lambda} = \text{arg min}_\lambda\{\text{BIC}(\lambda)\}$.

2.3 Penalized Maximum Covariance-Tapered Likelihood

When computing the OSE in Algorithm 1, the initial parameter values are set to the MLE. It is well-known that computation of MLE for a spatial linear model is of order N^3 and can be very demanding when the sample size N increases (Cressie, 1993). There are various approaches to alleviating the computational cost. Here we consider covariance tapering, which could effectively reduce our computational cost in practice. Furrer et al. (2006) considered tapering for Kriging and demonstrated that not only tapering enhances computational efficiency but achieves asymptotically optimality in terms of mean squared prediction errors under infill asymptotics. For parameter estimation via maximum likelihood, Kaufman et al. (2008) established consistency of tapered MLE, whereas Du et al. (2009) established the asymptotic distribution, also under infill asymptotics. However, both Kaufman et al. (2008) and Du et al. (2009) focused on the parameters in the Matérn family of covariance functions and did not consider estimation of the regression coefficients. In contrast, our primary interest is in the estimation of regression coefficients and we investigate the asymptotic properties under increasing domain asymptotics, which to the best of our knowledge, have not been established in the literature before. We will discuss infill asymptotics in the concluding Section 2.7.

Recall that $\mathbf{\Gamma} = [\gamma(\boldsymbol{s}_i, \boldsymbol{s}_{i'})]_{i,i'=1}^N$ is the covariance matrix of $\boldsymbol{y}$. Assuming second-order stationarity and isotropy, we let $\gamma(d) = \gamma(\boldsymbol{s}, \boldsymbol{s}')$, where $d = \|\boldsymbol{s} - \boldsymbol{s}'\|$ is the lag distance between two sampling sites $\boldsymbol{s}$ and $\boldsymbol{s}'$ in R. Let $K_{\mathrm{T}}(d, \omega)$ denote a tapering function, which is an isotropic autocorrelation function when $0 < d < \omega$ and 0 when $d \geq \omega$, for a given threshold distance $\omega > 0$. Compactly supported correlation functions can be used as the tapering functions (Wendland, 1995). For example,

$$K_{\mathrm{T}}(d, \omega) = (1 - d/\omega)_+, \tag{2.12}$$

where $x_+ = \max\{x, 0\}$, in which case the correlation is 0 at lag distance greater than the threshold distance ω. Let $\mathbf{\Delta}(\omega) = [K_{\mathrm{T}}(d_{ii'}, \omega)]_{i,i'=1}^N$ denote an $N \times N$ tapering matrix. Then a tapered covariance matrix of $\mathbf{\Gamma}$ is defined as $\mathbf{\Gamma}_{\mathrm{T}} = \mathbf{\Gamma} \circ \mathbf{\Delta}(\omega)$, where $\circ$ denotes the Hadamard product (i.e., elementwise product).

We approximate the log-likelihood function by replacing $\mathbf{\Gamma}$ in (2.4) with the tapered covariance matrix $\mathbf{\Gamma}_{\mathrm{T}}$ and obtain a covariance-tapered log-likelihood function

$$\begin{aligned} \ell_{\mathrm{T}}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) &= -(N/2)\log(2\pi) - (1/2)\log|\mathbf{\Gamma}_{\mathrm{T}}| \\ &\quad -(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \mathbf{\Gamma}_{\mathrm{T}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{2.13}$$

We let $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}} = \arg\max_{\boldsymbol{\eta}}\{\ell_{\mathrm{T}}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X})\}$ denote the maximum covariance-tapered likelihood estimate (MLE$_{\mathrm{T}}$) of $\boldsymbol{\eta}$.

Let $\mathbf{\Gamma}_{k,\mathrm{T}} = \partial\mathbf{\Gamma}_{\mathrm{T}}/\partial\theta_k = \mathbf{\Gamma}_k \circ \mathbf{\Delta}(\omega)$, $\mathbf{\Gamma}_{\mathrm{T}}^{k} = \partial\mathbf{\Gamma}_{\mathrm{T}}^{-1}/\partial\theta_k = \mathbf{\Gamma}^k \circ \mathbf{\Delta}(\omega)$, $\mathbf{\Gamma}_{kk',\mathrm{T}} = \partial^2\mathbf{\Gamma}_{\mathrm{T}}/\partial\theta_k\partial\theta_{k'} = \mathbf{\Gamma}_{kk'} \circ \mathbf{\Delta}(\omega)$, $\mathbf{\Gamma}_{\mathrm{T}}^{kk'} = \partial^2\mathbf{\Gamma}_{\mathrm{T}}^{-1}/\partial\theta_k\partial\theta_{k'} = \mathbf{\Gamma}^{kk'} \circ \mathbf{\Delta}(\omega)$ denote covariance-tapered version of $\mathbf{\Gamma}_k$, $\mathbf{\Gamma}^k$, $\mathbf{\Gamma}_{kk'}$, and $\mathbf{\Gamma}^{kk'}$, respectively. From (2.13), $\ell'_{\mathrm{T}}(\boldsymbol{\beta}) = \boldsymbol{X}^T\mathbf{\Gamma}_{\mathrm{T}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ and the kth element of $\ell'_{\mathrm{T}}(\boldsymbol{\theta})$ is $-(1/2)\mathrm{tr}(\mathbf{\Gamma}_{\mathrm{T}}^{-1}\mathbf{\Gamma}_{k,\mathrm{T}}) - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\mathbf{\Gamma}_{\mathrm{T}}^{k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Moreover, $\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\beta}) = -\boldsymbol{X}^T\mathbf{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{X}$, the kth column of $\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is $\boldsymbol{X}^T\mathbf{\Gamma}_{\mathrm{T}}^{k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, and the (k, k')th entry of $\ell''_{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\theta})$ is $-(1/2)\{\mathrm{tr}(\mathbf{\Gamma}_{\mathrm{T}}^{-1}\mathbf{\Gamma}_{kk',\mathrm{T}} + \mathbf{\Gamma}_{\mathrm{T}}^{k}\mathbf{\Gamma}_{k',\mathrm{T}}) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\mathbf{\Gamma}_{\mathrm{T}}^{kk'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}$. Since $E\{-\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \mathbf{0}$, the covariance-tapered information matrix of $\boldsymbol{\eta}$ is $\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\eta}) = \mathrm{diag}\{\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}), \boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\theta})\}$, where $\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}) = E\{-\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\beta})\} = \boldsymbol{X}^T\mathbf{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{X}$ and the (k, k')th entry of $\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\theta}) = E\{-\ell''_{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\theta})\}$ is $t_{kk',\mathrm{T}}/2$ with $t_{kk',\mathrm{T}} = \mathrm{tr}(\mathbf{\Gamma}_{\mathrm{T}}^{-1}\mathbf{\Gamma}_{k,\mathrm{T}}\mathbf{\Gamma}_{\mathrm{T}}^{-1}\mathbf{\Gamma}_{k',\mathrm{T}}) = \mathrm{tr}(\mathbf{\Gamma}_{\mathrm{T}}\mathbf{\Gamma}_{\mathrm{T}}^{k}\mathbf{\Gamma}_{\mathrm{T}}\mathbf{\Gamma}_{\mathrm{T}}^{k'})$.

Now, the penalized log-likelihood function (2.7) can be approximated by

$$Q_{\mathrm{T}}(\boldsymbol{\eta}) = \ell_{\mathrm{T}}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) - N\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.14}$$

where $\ell_{\mathrm{T}}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X})$ is a covariance-tapered log-likelihood function as defined in (2.13). Moreover, we let $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE_T}} = \arg\max_{\boldsymbol{\eta}}\{Q_{\mathrm{T}}(\boldsymbol{\eta})\}$ denote the penalized maximum covariance-tapered likelihood estimates (PMLE$_{\mathrm{T}}$). In the following, we again use one-step LLA estimation to approximate PMLE$_{\mathrm{T}}$, as in Algorithm 1.

Algorithm 2. At the initialization step, we let $\boldsymbol{\eta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$ with $\boldsymbol{\beta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\mathrm{MLE_T}}$ and $\boldsymbol{\theta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{MLE_T}}$. We then update $\boldsymbol{\beta}$ by maximizing

$$Q_{\mathrm{T}}^{*}(\boldsymbol{\beta}) = -(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{T}\boldsymbol{\Gamma}_{\mathrm{T}}(\boldsymbol{\theta}_{\mathrm{T}}^{(0)})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - N\sum_{j=1}^{p} p_{\lambda}'(|\beta_{j\mathrm{T}}^{(0)}|)|\beta_{j}| \qquad (2.15)$$

with respect to $\boldsymbol{\beta}$, where the first term is from (2.13) and the second term is an LLA of the penalty function in (2.14). The resulting one-step sparse estimate (OSE) of $\boldsymbol{\beta}$ is denoted as $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}$. We may also update $\boldsymbol{\theta}$ by maximizing (2.13) with respect to $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}$. The resulting OSE of $\boldsymbol{\theta}$ is denoted as $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}$. We let $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_T}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}^{T}, \widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}^{T})^{T}$ denote the $\mathrm{OSE_T}$ of $\boldsymbol{\eta}$, which approximates $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE_T}}$.

Although we focus on maximum likelihood estimation, the methodology here can be extended to utilize restricted maximum likelihood (REML) estimation. A REML estimator of $\boldsymbol{\theta}$ is obtained by minimizing

$$\begin{aligned}\ell_{\mathrm{rl}}(\boldsymbol{\theta}) \;&=\; \{(N-p)/2\}\log(2\pi) - (1/2)\log(|\boldsymbol{X}^{T}\boldsymbol{X}|) + (1/2)\log|\boldsymbol{\Gamma}(\boldsymbol{\theta})| \\ &\quad +(1/2)\log\{|\boldsymbol{X}^{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})\boldsymbol{X}|\} + (1/2)\boldsymbol{y}^{T}\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{y},\end{aligned}$$

where $\boldsymbol{\Pi}(\boldsymbol{\theta}) = \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\{\boldsymbol{X}^{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\}\boldsymbol{X}^{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}$. Since REML estimates are consistent (Cressie and Lahiri, 1993), we may readily modify Algorithm 1 by replacing MLE with REML estimates as the initial values.

It is worth mentioning an alternative covariance-tapered with log-likelihood function (Kaufman et al., 2008),

$$\begin{aligned}\ell_{\mathrm{T2}}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{X}) \;&=\; -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}_{\mathrm{T}}| \\ &\quad -(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{T}\{\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} \circ \boldsymbol{\Delta}(\omega)\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).\end{aligned} \qquad (2.16)$$

If the alternative covariance tapering is used in Algorithm 2, we obtain $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_{T2}}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_{T2}}}^{T}, \widehat{\boldsymbol{\theta}}_{\mathrm{OSE_{T2}}}^{T})^{T}$. The estimates of parameters, especially the range parameter, tend to be more accurate, but require more time to compute $\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} \circ \boldsymbol{\Delta}(\omega)$ than $\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}$. Following Kaufman et al. (2008), we refer to this tapering as type-2 tapering.

For choosing ω in practice, we adopt an approach suggested by Kaufman et al. (2008). First, a pilot estimate $\widehat{\boldsymbol{\eta}}_p$ is obtained from a suitable subset of the data, and then an estimated variance of $\widehat{\boldsymbol{\theta}}_p$ is computed from $\boldsymbol{I}_{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_p)^{-1}$. In addition, the computing time for evaluating $\ell_{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_p; \boldsymbol{y}, \boldsymbol{X})$ is recorded. As ω increases, the performance of parameter estimates improves, but it takes more computing time; and vice versa. These two criteria need to be balanced when selecting a reasonable ω among various choices of threshold. An illustrative example is given in Appendix E of this chapter.

2.4 Asymptotic Properties of PMLE and OSE

2.4.1 Notation and Assumptions

We let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ denote the true regression coefficients, where without loss of generality $\boldsymbol{\beta}_{10}$ is an $s \times 1$ vector of nonzero regression coefficients and $\boldsymbol{\beta}_{20} = \mathbf{0}$ is a $(p-s) \times 1$ zero vector. Let $\boldsymbol{\theta}_0$ denote the vector of true covariance function parameters.

We consider the asymptotic framework in Mardia and Marshall (1984) and let n denote the stage of the asymptotics. In particular, write $R_n = R$, $N_n = N$, and $\lambda_n = \lambda$. Furthermore, we define $a_n = \max_{1 \le j \le p}\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$ and $b_n = \max_{1 \le j \le p}\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$. Also, let $\boldsymbol{\phi}_n(\boldsymbol{\beta}) = \boldsymbol{\phi}(\boldsymbol{\beta})$ and $\boldsymbol{\Phi}_n(\boldsymbol{\beta}) = \boldsymbol{\Phi}(\boldsymbol{\beta})$, both evaluated at λ_n. For all other quantities that depend on n, the stage n will be in either the left superscript or the right subscript.

Recall that ${}^n t_{kk'} = \mathrm{tr}({}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_k{}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_{k'})$. Let $\mu_1 \le \cdots \le \mu_{N_n}$ denote the eigenvalues of ${}^n\boldsymbol{\Gamma}$. For $l = 1, \ldots, N_n$, let μ_l^k denote the eigenvalues of ${}^n\boldsymbol{\Gamma}_k$ such that $|\mu_1^k| \le \cdots \le |\mu_{N_n}^k|$ and let $\mu_l^{kk'}$ denote the eigenvalues of ${}^n\boldsymbol{\Gamma}_{kk'}$ such that $|\mu_1^{kk'}| \le \cdots \le |\mu_{N_n}^{kk'}|$.

For an $N_n \times N_n$ matrix $\boldsymbol{A} = (a_{ij})_{i,j=1}^{N_n}$, the Frobenius, max, and spectral norm are defined as $\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^{N_n}\sum_{j=1}^{N_n} a_{ij}^2\right)^{1/2}$, $\|\boldsymbol{A}\|_{\max} = \max\{|a_{ij}| : i, j = 1, \ldots, N_n\}$, and $\|\boldsymbol{A}\|_s = \max\{|\mu_l(\boldsymbol{A})| : l = 1, \ldots, N_n\}$, where $\mu_l(\boldsymbol{A})$ is the lth eigenvalue of $\boldsymbol{A}$.

The following regularity conditions are assumed for Theorems 2.4.1–2.4.2.

(A.1) For $\boldsymbol{\theta} \in \Omega$ where Ω is an open subset of $\mathbb{R}^q$ such that $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, the covariance function $\gamma(\cdot, \cdot; \boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ with continuous second-order derivatives and is positive definite in the sense that, for any $N_n \geq 1$ and $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_n}$, the covariance matrix $\boldsymbol{\Gamma} = [\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})]_{i,j=1}^{N_n}$ is positive definite.

(A.2) There exist positive constants C, C_k and $C_{kk'}$, such that $\lim_{n\to\infty} \mu_{N_n} = C < \infty$, $\lim_{n\to\infty} |\mu_{N_n}^k| = C_k < \infty$, $\lim_{n\to\infty} |\mu_{N_n}^{kk'}| = C_{kk'} < \infty$ for all $k, k' = 1, \ldots, q$.

(A.3) For some $\delta > 0$, there exist positive constants D_k, $D_{kk'}$ and $D_{kk'}^*$ such that (i) $\|{}^n\boldsymbol{\Gamma}_k\|_F^{-2} = D_k N_n^{-1/2-\delta}$ for $k = 1, \ldots, q$; (ii) Either $\|{}^n\boldsymbol{\Gamma}_k + {}^n\boldsymbol{\Gamma}_{k'}\|_F^{-2} = D_{kk'} N_n^{-1/2-\delta}$ or $\|{}^n\boldsymbol{\Gamma}_k - {}^n\boldsymbol{\Gamma}_{k'}\|_F^{-2} = D_{kk'}^* N_n^{-1/2-\delta}$ for any $k \neq k'$.

(A.4) For any $k, k' = 1, \ldots, q$, (i) ${}^n a_{kk'} = \lim_{n\to\infty}\{{}^n t_{kk'}({}^n t_{kk} {}^n t_{k'k'})^{-1/2}\}$ exists and $\boldsymbol{A}_n = ({}^n a_{kk'})_{k,k'=1}^q$ is nonsingular; (ii) $\left|{}^n t_{kk} {}^n t_{k'k'}^{-1}\right|$ and $\left|{}^n t_{k'k'} {}^n t_{kk}^{-1}\right|$ are bounded.

(A.5) The design matrix $\boldsymbol{X}$ has full rank p and is uniformly bounded in max norm with $\lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \mathbf{0}$.

(A.6) There exists a positive constant C_0, such that $\|{}^n\boldsymbol{\Gamma}^{-1}\|_s < C_0 < \infty$.

(A.7) For $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Omega$, $N_n^{-1}\boldsymbol{I}_n(\boldsymbol{\beta}) \to \boldsymbol{J}(\boldsymbol{\beta})$ and $N_n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}) \to \boldsymbol{J}(\boldsymbol{\theta})$ as $n \to \infty$.

(A.8) $a_n = O(N_n^{-1/2})$ and $b_n \to 0$ as $n \to \infty$.

(A.9) There exist positive constants c_1 and c_2 such that, when $\beta_1, \beta_2 > c_1\lambda_n$, $|p''_{\lambda_n}(\beta_1) - p''_{\lambda_n}(\beta_2)| \leq c_2|\beta_1 - \beta_2|$.

(A.10) $\lambda_n \to 0, N_n^{1/2}\lambda_n \to \infty$ as $n \to \infty$.

(A.11) $\liminf_{n\to\infty} \liminf_{\beta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\beta) > 0$.

(A.2), (A.3)(i), (A.4)(i) and (A.5) are assumed in Mardia and Marshall (1984). (A.1) and (A.5) are standard assumptions for MLE, whereas (A.2), (A.3)(i), (A.4)(i), and (A.6) ensure smoothness, growth, and convergence of the information matrix

(Mardia and Marshall, 1984). Together with (A.7), they yield a central limit theorem of $\ell'(\boldsymbol{\eta})$ and convergence in probability of $\ell''(\boldsymbol{\eta})$. For establishing Theorems 2.4.1–2.4.2, only the parts (i) of (A.3) and (A.4) are used. Moreover, the implicit asymptotic framework is increasing domain, where the sample size N_n grows at the increase of the spatial domain R_n (Mardia and Marshall, 1984). Finally, (A.8)–(A.11) are mild regularity conditions regarding the penalty function and are sufficient for Theorems 2.4.1–2.4.2 to hold (Fan and Li, 2001) and (Fan and Peng, 2004).

2.4.2 Consistency and Asymptotic Normality of PMLE

Theorem 2.4.1. *Under (A.1)–(A.9), there exists, with probability tending to one, a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}$ *of* $Q(\boldsymbol{\eta})$ *defined in (2.7) such that* $\|{}^n\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2} + a_n)$. *If, in addition, (A.10)–(A.11) hold, then* ${}^n\widehat{\boldsymbol{\eta}} = ({}^n\widehat{\boldsymbol{\beta}}_1^T, {}^n\widehat{\boldsymbol{\beta}}_2^T, {}^n\widehat{\boldsymbol{\theta}}^T)^T$ *satisfies*

(i) Sparsity: ${}^n\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *with probability tending to 1.*

(ii) Asymptotic normality:

$$\begin{aligned}
&N_n^{1/2}\{\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}\left[{}^n\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}^{-1}\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10})\right]\\
&\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10})),\\
&N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}),
\end{aligned}$$

where $\boldsymbol{J}(\boldsymbol{\beta}_{10})$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ *consist of the first* $s \times s$ *upper-left submatrix of* $\boldsymbol{J}(\boldsymbol{\beta}_0)$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)$*, respectively.*

Theorem 2.4.1 establishes the asymptotic properties of PMLE. Under (A.1)–(A.9), there exists a local maximizer converging to the true parameter at the rate $O_p(N_n^{-1/2} + a_n)$. Since $a_n = O(N_n^{-1/2})$ from (A.8), the local maximizer is root-N_n consistent. As shown in Fan and Li (2001), the SCAD penalty function satisfies (A.8)–(A.11) by choosing an appropriate tuning parameter λ_n. Therefore, by Theorem 2.4.1, PMLE under the SCAD penalty possesses the sparsity property and asymptotic normality. Moreover, when the sample size N_n is sufficiently large, $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ will be close to zero. That is, performance of the PMLE is asymptotically as efficient as the MLE of $\boldsymbol{\beta}_1$ when knowing $\boldsymbol{\beta}_2 = \mathbf{0}$. The arguments above hold for other penalty functions such as L_q penalty with $q < 1$, but not $q = 1$.

2.4.3 Consistency and Asymptotic Normality of OSE

Theorem 2.4.2. *Suppose that the initial value ${}^n\boldsymbol{\eta}^{(0)}$ in Algorithm 1 satisfies ${}^n\boldsymbol{\eta}^{(0)} - \boldsymbol{\eta}_0 = O_p(N_n^{-1/2})$. For the SCAD penalty, under (A.1)–(A.7) and (A.10), the OSE ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{OSE}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}}^T, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}}^T)^T$ satisfies*

(i) Sparsity: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE}} = \mathbf{0}$ with probability tending to 1.

(ii) Asymptotic normality:

$$\begin{aligned} N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}} - \boldsymbol{\beta}_{10}) &\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10})^{-1}) \\ N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}} - \boldsymbol{\theta}_0) &\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}) \end{aligned}$$

where $\boldsymbol{J}(\boldsymbol{\beta}_{10})$ consists of the first $s \times s$ upper-left submatrix of $\boldsymbol{J}(\boldsymbol{\beta}_0)$.

Theorem 2.4.2 establishes the asymptotic properties of OSE such that the OSE is sparse and asymptotically normal under the SCAD penalty. The OSE for $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}$ has the same limiting distribution as PMLE and thus achieves the same efficiency. In fact, Theorem 2.4.2 holds for another general class of penalty functions such that $p'_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$ where $p'(\cdot)$ is continuous on $(0, \infty)$, and there is some $\alpha > 0$ such that $p'(\beta) = O(\beta^{-\alpha})$ as $\beta \to 0+$ (Zou and Li, 2008). Following similar arguments for SCAD penalty in our Theorem 2.4.2 and those in Zou and Li (2008), it can be shown that, if $N_n^{(1+\alpha)/2}\lambda_n \to \infty$ and $N_n^{1/2}\lambda_n \to 0$, Theorem 2.4.2 continues to hold. In practice, we set the initial value ${}^n\boldsymbol{\eta}^{(0)}$ to be the MLE ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}}$ as it satisfies the consistency condition.

2.5 Asymptotic Properties under Covariance Tapering

2.5.1 Notation and Assumptions

In order to establish the asymptotic properties under covariance tapering, we continue to assume (A.1)–(A.11). We now restrict our attention to a second-order stationary error process in $\mathbb{R}^2$ with an isotropic covariance function $\gamma(d)$, where $d \geq 0$ is lag distance. We also assume that the distance between any two sampling sites is greater than a constant (Mardia and Marshall, 1984). As for the tapering function, we consider (2.12).

Let $\gamma_k(d) = \partial\gamma(d)/\partial\theta_k$, $\gamma_{kk'}(d) = \partial^2\gamma(d)/\partial\theta_k\partial\theta_{k'}$, for $k, k' = 1, \ldots, q$. Two additional regularity conditions are assumed for Theorems 2.5.2–2.5.3.

(A.12) $0 < \inf_n\{\omega_n N_n^{-1/2}\} \leq \sup_n\{\omega_n N_n^{-1/2}\} < \infty$, where $\omega_n = \omega$ is the threshold distance in the tapering function (2.12).

(A.13) There exists a nonincreasing function γ_0 with $\int_0^\infty u^2\gamma_0(u)du < \infty$ such that $\max\{|\gamma(u)|, |\gamma_k(u)|, |\gamma_{k,k'}(u)|\} \leq \gamma_0(u)$ for all $u \in (0, \infty)$ and $1 \leq k, k' \leq q$.

From (A.12), the threshold distance ω_n is bounded away from 0 and grows at the rate of $N_n^{1/2}$. The condition in (A.13) has to do with the covariance function. It can be shown that they hold for some of the commonly-used covariance functions such as the Matérn class. Details are given in Appendix D of this chapter.

2.5.2 Consistency and Asymptotic Normality of PMLE$_{\mathrm{T}}$

Proposition 2.5.1. *Under (A.1)–(A.7) and (A.12)–(A.13), the MLE*$_{\mathrm{T}}$ $^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$ *is asymptotically normal with*

$$N_n^{1/2}(^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}} - \boldsymbol{\eta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\eta}_0)^{-1}).$$

Proposition 2.5.1 establishes the asymptotic normality of MLE$_{\mathrm{T}}$. In particular, MLE and MLE$_{\mathrm{T}}$ have the same limiting distribution. This implies that, under the regularity conditions, covariance-tapered MLE achieves the same efficiency as MLE. Thus, in Algorithm 2 for computing the OSE$_{\mathrm{T}}$, we may set the initial parameter values to $^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$.

Theorem 2.5.2. *Under (A.1)–(A.9) and (A.12)–(A.13), there exists, with probability tending to one, a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}}$ *of* $Q_{\mathrm{T}}(\boldsymbol{\eta})$ *defined in (2.14) such that* $\|{}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2} + a_n)$. *If, in addition, (A.10)–(A.11) hold, then* ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{T}}^T, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{T}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{T}}^T)^T$ *satisfies*

(i) Sparsity: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{T}} = \mathbf{0}$ *with probability tending to 1.*

(ii) Asymptotic normality:

$$\begin{aligned}
&N_n^{1/2}\{\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}\left[{}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{T}} - \boldsymbol{\beta}_{10} + \{\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}^{-1}\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10})\right] \\
&\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10})), \\
&N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{T}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}),
\end{aligned}$$

where $\boldsymbol{J}(\boldsymbol{\beta}_{10})$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ *consist of the first* $s \times s$ *upper-left submatrix of* $\boldsymbol{J}(\boldsymbol{\beta}_0)$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)$*, respectively.*

In Theorem 2.5.2, PMLE$_{\mathrm{T}}$ is shown to be consistent, sparse, and asymptotically normal. In particular, PMLE$_{\mathrm{T}}$ has the same asymptotic distribution as PMLE in Theorem 2.4.1. That is, PMLE$_{\mathrm{T}}$ achieves the same efficiency and oracle property as PMLE asymptotically, yet in the mean time is more computationally efficient.

2.5.3 Consistency and Asymptotic Normality of OSE$_{\mathrm{T}}$

Theorem 2.5.3. *Suppose that the initial value* ${}^n\boldsymbol{\eta}_{\mathrm{T}}^{(0)}$ *in Algorithm 2 satisfies* ${}^n\boldsymbol{\eta}_{\mathrm{T}}^{(0)} - \boldsymbol{\eta}_0 = O_p(N_n^{-1/2})$. *For the SCAD penalty function, under (A.1)–(A.7), (A.10) and (A.12)–(A.13), the OSE*$_{\mathrm{T}}$ ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_T}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE_T}}^{\mathrm{T}}, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE_T}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}^T)^T$ *satisfies*

(i) Sparsity: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE_T}} = \mathbf{0}$ *with probability tending to 1.*

(ii) Asymptotic normality:

$$\begin{aligned}
N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE_T}} - \boldsymbol{\beta}_{10}) &\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10})^{-1}) \\
N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}} - \boldsymbol{\theta}_0) &\xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})
\end{aligned}$$

where $\boldsymbol{J}(\boldsymbol{\beta}_{10})$ *consists of the first* $s \times s$ *upper-left submatrix of* $\boldsymbol{J}(\boldsymbol{\beta}_0)$.

Theorem 2.5.3 establishes the asymptotic properties of $\mathrm{OSE_T}$ under the SCAD penalty. In particular, $\mathrm{OSE_T}$ achieves the same limiting distribution as OSE of $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}$ in Theorem 2.4.2 and thus the same efficiency. Furthermore, similar to Theorem 2.4.2, Theorem 2.5.3 holds for the class of penalty functions such that $p'_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$ where $p'(\cdot)$ is continuous on $(0, \infty)$, and there is some $\alpha > 0$ such that $p'(\beta) = O(\beta^{-\alpha})$ as $\beta \to 0+$, provided that $N_n^{(1+\alpha)/2}\lambda_n \to \infty$ and $N_n^{1/2}\lambda_n \to 0$.

2.6 Numerical Examples

2.6.1 Simulation Study

We now conduct a simulation study to investigate the finite-sample properties of OSE and $\mathrm{OSE_T}$. The spatial domain of interest is assumed to be a square $[0, l]^2$ of side lengths $l = 5, 10, 15$. The sample sizes are set to be $N = 100, 400, 900$ for $l = 5, 10, 15$, respectively, with a fixed sampling density of 4. For regression, we generate seven covariates that follow standard normal distributions with a cross-covariate correlation of 0.5. The regression coefficients are set to be $\boldsymbol{\beta} = (4, 3, 2, 1, 0, 0, 0)^T$. We standardize the covariates to have mean 0 and variance 1 and standardize $\boldsymbol{y}$ to have mean 0. Thus, there will be no intercept in the vector of regression coefficients $\boldsymbol{\beta}$. For spatial dependence, we generate error terms that follow a zero-mean stationary and isotropic Gaussian process. A commonly-used family of covariance function is the Matérn class in the form of $\gamma(d) = \sigma^2(1-c)\{2^{\nu-1}\Gamma(\nu)\}^{-1}\left(2\nu^{1/2}d/r\right)^{\nu}\mathcal{K}_\nu\left(2\nu^{1/2}d/r\right)$, where $\sigma^2 > 0$ is the error variance, $c \in [0, 1]$ is a nugget proportion such that $c\sigma^2$ is the nugget effect representing variation at small lag distance, $r > 0$ is a range parameter controlling the rate of autocorrelation decay with lag distance, $\nu > 0$ is a shape parameter controlling the smoothness of the Gaussian process, and $\mathcal{K}_\nu(\cdot)$ is a modified Bessel function of the second kind of order ν. An exponential covariance function $\gamma(d) = \sigma^2(1-c)\exp(-d/r)$ is a special case of the Matérn class with shape parameter $\nu = 1/2$. We let $\sigma^2 = 9$, $c = 0.2$, $r = 1$, and $\nu = 1/2$. For each choice of sample size N, a total of 100 data sets are simulated.

For each simulated data set, we compute OSE and OSE$_{\mathrm{T}}$ using Algorithms 1 and 2. For OSE$_{\mathrm{T}}$, we consider different threshold values for covariance tapering $\omega = l/2^k$ for $k = 1, 2, \ldots$. We present only the case of $\omega = l/4$ to save space. Our methods are compared against several alternatives. Of particular interest is OSE under a standard linear regression where spatial autocorrelation is unaccounted for in the penalized loglikelihood function. This would be akin to PLS under SCAD in Wang and Zhu (2009) and will be referred to as OSE$_{\mathrm{Alt1}}$. In addition, we modify the initialization step of both Algorithms 1 and 2 by using MLE under the true model which is unknown but assumed to be known. This is an attempt to evaluate the effect of starting values and will be referred to as OSE$_{\mathrm{Alt2}}$. Last, we consider a benchmark case, referred to as OSE$_{\mathrm{Alt3}}$, where the true model is assumed to be known and MLE of the nonzero regression coefficients and the covariance function parameters are computed. Our OSE and OSE$_{\mathrm{T}}$ will be compared against this benchmark to evaluate the oracle properties.

For each choice of sample size N, we first compute the average numbers of correctly (C0) and incorrectly (I0) identified zero-valued regression coefficients from OSE $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE}}$ and OSE$_{\mathrm{T}}$ $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}$, as well as those from OSE$_{\mathrm{Alt1}}$ and OSE$_{\mathrm{Alt2}}$. The true number of zero-valued regression coefficients is 3 as is assumed in OSE$_{\mathrm{Alt3}}$. Then, we compute means of the nonzero-valued OSE $\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}}$ and OSE$_{\mathrm{T}}$ $\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE_T}}$, as well as the corresponding covariance function parameters $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}$. We estimate a standard deviation (SD) for each parameter estimate using the information matrix formulas in Sections 2.2 and 2.3. The true SD is approximated by the median of the sample SD (SDm) of the 100 parameter estimates. The results are given in Tables 2.1–2.3.

In terms of variable selection, C0 tends to the true value 3 and I0 tends to 0, as the sample size N increases, for OSE, OSE$_{\mathrm{T}}$, OSE$_{\mathrm{Alt1}}$, and OSE$_{\mathrm{Alt2}}$. When the sample size is relatively small ($N = 100$), OSE$_{\mathrm{Alt2}}$ has the best performance with the largest C0 and smallest I0, reflecting the effect of starting values in Algorithm 1. But it is not practical, as we do not know what the true model is in actual data analysis.

Table 2.1: Simulation Results for $N = 100$

Method	Truth	OSE	OSE_{T}	OSE_{Alt1}	OSE_{Alt2}	OSE_{Alt3}
C0	3	2.79	2.84	2.84	2.95	3.00
I0		0.06	0.10	0.32	0.06	0.00
β_1	4.00	4.01	4.03	4.17	4.01	4.01
SD		0.28	0.29	0.39	0.27	0.27
SDm		0.26	0.27	0.36	0.26	0.26
β_2	3.00	3.04	3.03	3.08	3.04	3.03
SD		0.30	0.30	0.41	0.30	0.29
SDm		0.25	0.26	0.36	0.25	0.25
β_3	2.00	1.94	1.97	2.00	1.94	1.93
SD		0.29	0.31	0.50	0.28	0.28
SDm		0.25	0.26	0.36	0.26	0.26
β_4	1.00	1.02	1.03	0.78	1.03	1.02
SD		0.35	0.40	0.55	0.33	0.26
SDm		0.24	0.24	0.26	0.24	0.26
r	1.00	0.79	6.31	–	0.83	0.84
SD		0.54	2.14	–	0.57	0.57
SDm		0.48	17.65	–	0.51	0.51
c	0.20	0.16	0.23	–	0.17	0.17
SD		0.12	0.13	–	0.12	0.12
SDm		0.11	0.19	–	0.11	0.11
σ^2	9.00	7.96	7.14	7.74	8.03	8.03
SD		2.28	1.53	2.06	2.36	2.36
SDm		2.21	4.79	1.16	2.28	2.28

The average number of correctly identified 0 coefficients (C0), average number of incorrectly identified 0 coefficients (I0), mean, standard deviation (SD), and median estimated standard deviation (SDm) under OSE, OSE_{T}, OSE_{Alt1}, OSE_{Alt2}, and OSE_{Alt3} for sample size $N = 100$.

Table 2.2: Simulation Results for $N = 400$.

Method	Truth	OSE	OSE_T	OSE_Alt1	OSE_Alt2	OSE_Alt3
C0	3	2.97	2.97	2.97	2.98	3.00
I0		0.00	0.00	0.01	0.00	0.00
β_1	4.00	3.98	3.98	3.98	3.99	3.99
SD		0.14	0.14	0.20	0.14	0.14
SDm		0.13	0.13	0.19	0.13	0.13
β_2	3.00	3.02	3.03	3.03	3.02	3.02
SD		0.14	0.14	0.21	0.13	0.13
SDm		0.13	0.13	0.19	0.13	0.13
β_3	2.00	2.01	2.01	2.01	2.01	2.01
SD		0.12	0.12	0.17	0.12	0.12
SDm		0.13	0.13	0.19	0.13	0.13
β_4	1.00	0.99	1.00	0.96	1.00	1.00
SD		0.12	0.12	0.26	0.12	0.12
SDm		0.13	0.13	0.19	0.13	0.13
r	1.00	0.90	2.87	–	0.90	0.90
SD		0.29	4.08	–	0.29	0.29
SDm		0.25	5.24	–	0.25	0.25
c	0.20	0.19	0.29	–	0.19	0.19
SD		0.06	0.07	–	0.06	0.06
SDm		0.05	0.11	–	0.05	0.05
σ^2	9.00	8.70	8.25	8.71	8.70	8.70
SD		1.39	1.00	1.37	1.39	1.39
SDm		1.29	2.95	0.63	1.29	1.29

The average number of correctly identified 0 coefficients (C0), average number of incorrectly identified 0 coefficients (I0), mean, standard deviation (SD), and median estimated standard deviation (SDm) under OSE, OSE_T, OSE_Alt1, OSE_Alt2, and OSE_Alt3 for sample size $N = 400$.

Table 2.3: Simulation Results for $N = 900$.

Method	Truth	OSE	OSE_T	OSE_Alt1	OSE_Alt2	OSE_Alt3
C0	3	3.00	3.00	3.00	3.00	3.00
I0		0.00	0.00	0.00	0.00	0.00
β_1	4.00	4.00	4.01	4.03	4.00	4.00
SD		0.10	0.10	0.13	0.10	0.10
SDm		0.09	0.09	0.13	0.09	0.09
β_2	3.00	3.01	3.01	2.99	3.01	3.01
SD		0.08	0.08	0.12	0.08	0.08
SDm		0.09	0.09	0.13	0.09	0.09
β_3	2.00	1.98	1.99	1.98	1.98	1.98
SD		0.08	0.08	0.11	0.08	0.08
SDm		0.09	0.09	0.13	0.09	0.09
β_4	1.00	1.00	1.00	1.01	1.00	1.00
SD		0.09	0.09	0.13	0.09	0.09
SDm		0.09	0.09	0.13	0.09	0.09
r	1.00	0.94	1.44	–	0.94	0.94
SD		0.17	0.50	–	0.17	0.17
SDm		0.17	0.40	–	0.17	0.17
c	0.20	0.19	0.25	–	0.19	0.19
SD		0.04	0.04	–	0.04	0.04
SDm		0.04	0.04	–	0.04	0.04
σ^2	9.00	8.80	8.50	8.80	8.80	8.80
SD		0.90	0.74	0.87	0.90	0.90
SDm		0.85	1.15	0.42	0.85	0.85

The average number of correctly identified 0 coefficients (C0), average number of incorrectly identified 0 coefficients (I0), mean, standard deviation (SD), and median estimated standard deviation (SDm) under OSE, OSE_T, OSE_Alt1, OSE_Alt2, and OSE_Alt3 for sample size $N =$ 900.

OSE_{Alt1} assuming no spatial dependence in the regression model seems to over-shrink the regression coefficients. While C0 = 2.84 is close to 3 under OSE_{Alt1}, I0 = 0.32 is also large, compared to our OSE and OSE_{T}. Between OSE and OSE_{T}, it appears that C0 is slightly better, but I0 is slightly worse for OSE_{T} than OSE.

In terms of estimation of the nonzero regression coefficients, both accuracy and precision improve as the sample size N increases, for all five OSE cases considered here. While the accuracy is similar between OSE_{Alt1} and our OSE and OSE_{T}, a striking feature is the larger SD of OSE_{Alt1} when compared with our OSE and OSE_{T}, for all three sample sizes $N = 100, 400, 900$. This suggests that, by including spatial dependence directly in the penalized likelihood function, we gain statistical efficiency in parameter estimation. For the small sample size ($N = 100$), SD based on information matrix without accounting for spatial dependence appears to underestimate the true variation estimated by SDm. Furthermore, the SD's of OSE and OSE_{T} tend to those in the benchmark case OSE_{Alt3} as the sample size increases, confirming the oracle properties in Sections 2.4 and 2.5. For 100 simulations, it takes about 1 second, 30 seconds, and 4 minutes per simulation for saple sizes N = 100, 400, 900, respectively.

Based on these simulation results, it may be tempting to consider using OSE_{Alt1} to select variables and then OSE_{Alt3} for parameter estimation when the sample size is reasonably large, as a means of saving computational time. We contend that this is not necessary, as our OSE or OSE_{T} enables variable selection and parameter estimation simultaneously, at the similar computational cost. Moreover, in practice, it is not always clear how large a sample size at hand really is, as an effective sample size is influenced by factors such as the strength of spatial dependence in the error process.

In addition, we investigate the effect of thresholding. We discuss our conclusions without showing the numeral results to save space. We observe that, as the threshold distance ω decreases, the covariance matrix becomes more sparse and thus the computation is faster. However, OSE_{T} is closer to OSE_{Alt1} which ignores spatial dependence and the variation of $\mathrm{OSE_T}$ of the regression coefficients increases. Conversely, as the threshold distance ω increases, $\mathrm{OSE_T}$ gets closer to OSE. There indeed is a tradeoff between computation efficiency and statistical efficiency for finite sample sizes.

In Table 2.4, we consider several modifications of our method for the sample size $N = 100$ following the suggestions of the reviewers. The first modification is to continue one-step approximation until convergence; the results are reported in the column labeled PMLE. It can been seen that the resulting estimates are very close to those from OSE algorithm. The second modification is to replace maximum likelihood estimation with restricted maximum likelihood estimation for covariance parameters; the results are reported in the column called REML-OSE. With this change, there is some improvement in the estimation of $\boldsymbol{\theta}$, but for the estimation of $\boldsymbol{\beta}$, the results are similar. Moreover, we implement type-2 tapering as described in (2.16). The results are reported in the column named $\mathsf{OSE_{T2}}$. Note that the estimation of the range parameter has greatly improved under type-2 tapering. In the last column of Table 2.4, we investigate the effect of number of covariates by increasing p from 7 to 20, where the true number of non-zero regression coefficients is 4. The estimates are found to be reasonably accurate and efficient. The sparsity can also be achieved despite the increasing number of covariates.

Finally, in Table 2.5, we investigate the robustness of our method against the misspecification of the underlining covariance structure. We generate data using Gaussian covariance function $\gamma(d) = \sigma^2(1-c)e^{-d^2/r^2}$ with $\sigma^2 = 9$, $r = 1$ and $c = 0.2$ where sample size $N = 100$. We fit the generated data using both the exponential and Gaussian covariance functions. It can be seen that, in the misspecified case, the exponential model still yields good results.

2.6.2 Data Examples

The first data example consists of January precipitation (inches per 24-hour period) on the log scale from 259 weather stations in the state of Colorado (Reich and Davis,

Table 2.4: Simulation Results of PMLE, REML-OSE, OSE_{T2} and OSE_{20}.

Method	PMLE	REML-OSE	OSE_{T2}	OSE_{20}
C0	2.83	2.83	2.87	14.68
I0	0.06	0.08	0.15	0.10
β_1	4.01	4.01	4.05	4.02
SD	0.28	0.29	0.31	0.31
SDm	0.26	0.27	0.30	0.26
β_2	3.04	3.06	3.04	3.05
SD	0.30	0.30	0.32	0.30
SDm	0.25	0.26	0.29	0.25
β_3	1.93	1.93	1.93	1.98
SD	0.29	0.29	0.33	0.29
SDm	0.25	0.26	0.29	0.25
β_4	1.03	1.02	1.02	0.96
SD	0.34	0.37	0.44	0.41
SDm	0.24	0.24	0.25	0.23
r	0.79	0.84	0.68	0.82
SD	0.54	0.58	1.57	0.57
SDm	0.49	0.52	3.52	0.49
c	0.16	0.18	0.12	0.14
SD	0.12	0.12	0.16	0.12
SDm	0.11	0.12	0.16	0.10
σ^2	7.97	8.11	7.52	8.00
SD	2.26	2.36	1.98	2.34
SDm	2.22	2.28	2.15	2.27

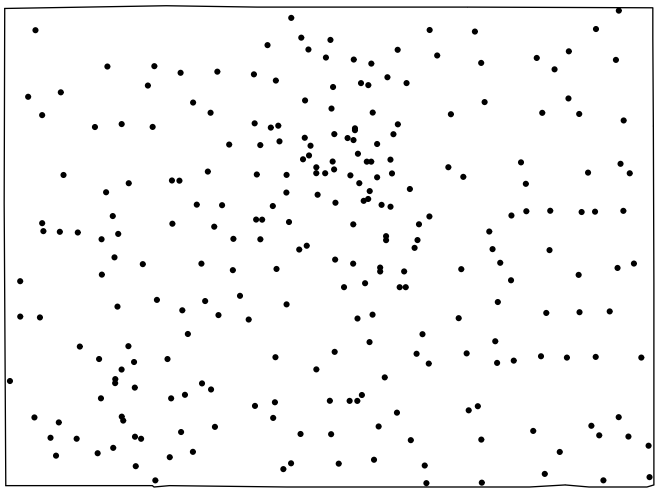
The average number of correctly identified 0 coefficients (C0), average number of incorrectly identified 0 coefficients (I0), mean, standard deviation (SD), and median estimated standard deviation (SDm) under PMLE, REML-OSE, OSE_{T2} and OSE_{20} for sample size $N = 100$.

Table 2.5: Simulation Results for the Misspecified Case.

Method	Gaussian	Exponential
C0	2.80	2.87
I0	0.01	0.01
β_1	3.96	3.98
SD	0.22	0.25
SDm	0.20	0.21
β_2	3.05	3.02
SD	0.22	0.22
SDm	0.20	0.21
β_3	2.00	1.98
SD	0.25	0.26
SDm	0.20	0.21
β_4	1.01	1.01
SD	0.27	0.29
SDm	0.20	0.21
r	0.98	1.19
SD	0.19	0.51
SDm	0.15	0.67
c	0.19	0.06
SD	0.08	0.08
SDm	0.06	0.06
σ^2	8.14	8.69
SD	2.48	3.13
SDm	2.34	3.38

The average number of correctly identified 0 coefficients (C0), average number of incorrectly identified 0 coefficients (I0), mean, standard deviation (SD), and median estimated standard deviation (SDm) under Gaussian and exponential spatial covariance models for sample size $N = 100$.

Figure 2.1: Map of Locations of 259 Sampling Sites.



Map of locations of 259 sampling sites in the Colorado precipitation data.

2008) as shown in Figure 2.1. Candidate covariates are elevation, slope, aspect, and seven spectral bands from a MODIS satellite imagery (B1M through B7M). It is of interest to investigate the relationship between precipitation and these covariates.

We first fit a spatial linear model with an exponential covariance function via maximum likelihood. The parameter estimates and their standard errors in Table 2.6 suggest that the regression coefficients for elevation, B1M, B4M, B6M, and B7M are possibly significant. Among the covariance function parameters, of most interest is the range parameter, which is significantly different from zero. This indicates that there is spatial autocorrelation among the errors in the linear regression. Our OSE method selects elevation and B4M, and shrink all the other regression coefficients to zero. The covariance function parameter estimates are close to the MLE. For comparison, we fit a standard linear regression with iid errors and the corresponding

Table 2.6: Colorado Precipitation Data

Terms	MLE	SD	OSE	SD	OSE_{Alt1}	SD
Regression coefficients						
Elevation	0.305	0.055	0.228	0.054	0.195	0.044
Slope	0.016	0.026	–	–	0.035	0.040
Aspect	-0.004	0.022	–	–	0.032	0.034
B1M	0.214	0.157	–	–	–	–
B2M	0.058	0.064	–	–	–	–
B3M	0.017	0.109	–	–	–	–
B4M	-0.404	0.183	-0.089	0.034	-0.264	0.045
B5M	0.043	0.089	–	–	–	–
B6M	-0.162	0.116	–	–	–	–
B7M	0.172	0.098	–	–	–	–
Covariance function parameters						
Range	0.967	0.368	1.043	0.417	–	–
Nugget	0.183	0.061	0.196	0.064	–	–
σ^2	0.287	0.067	0.304	0.074	0.289	0.026

Regression coefficient estimates and standard deviations (SD) using maximum likelihood (MLE) and one-step sparse estimation (OSE) under a spatial linear model with an exponential covariance function for the Gaussian error process, as well as OSE under a standard linear model with iid errors (OSE_{Alt1}).

Table 2.7: California Lizards Data

	MLE	SD	OSE	SD		MLE	SD	OSE	SD
ELEVATION	0.04	0.24	–	–	PTREE2	-0.12	0.09	–	–
CHAPARRAL	0.09	0.12	–	–	PGRASS2	0.27	0.15	–	–
COVER	0.06	0.12	–	–	POTHER	0.02	0.09	–	–
SAND	0.31	0.08	0.14	0.07	LL	-0.13	0.11	–	–
ANT$_1$	-0.20	0.11	-0.10	0.08	CRY	0.14	0.08	–	–
ANT$_2$	-0.03	0.09	–	–	CS	0.03	0.08	–	–
BARE ROCK	0.01	0.08	–	–	ORG	-0.18	0.15	–	–
SLOPE	0.00	0.08	–	–	MOS	-0.19	0.10	–	–
ASPECT	0.02	0.08	–	–	ARGEN	-0.06	0.12	–	–
CANOPYHT	0.10	0.13	–	–	HARV	0.07	0.10	–	–
SHRUBHT	-0.02	0.12	–	–	NHARVNEST	0.01	0.08	–	–
HERBHT	0.04	0.18	–	–	CARP	-0.03	0.07	–	–
CSS2	0.20	0.11	–	–	CANHTCAT	0.13	0.11	–	–

Regression coefficient estimates and standard deviations (SD) using maximum likelihood (MLE) and one-step sparse estimation (OSE) under a spatial linear model with an exponential covariance function for the Gaussian error process.

OSE$_{\text{Alt1}}$ selects slope and aspect in addition to elevation and B4M. However, the regression coefficients for slope and aspect do not appear to be significant.

In addition, we apply our method to the whiptail lizard data as described in Section 2.1. There are 148 sites, and the response variable is the abundance of lizards at each site. There are 26 covariates regarding location, vegetation, flora, soil and ants. Hoeting et al. (2006) considered only 6 covariates after a separate prescreening procedure, and selected 2 covariates in their final model. In this chapter, we consider all 26 covariates simultaneously, and interestingly reach the same final model. The parameter estimates can be found in Table 2.7. For detailed description, see Hoeting et al. (2006) and Hollander et al. (1994).

2.7 Summary and Discussion

In summary, we have proposed a penalized method for simultaneous variable selection and parameter estimation in a spatial linear model. We have also developed one-step sparse estimation and its counterpart under covariance tapering to approximate the penalized parameter estimates and gained computational efficiency. Furthermore, we have established asymptotic properties of the parameter estimates and their approximations, showing consistency, sparsity, and asymptotic normality. Finite-sample properties have been examined via a simulation study and we have found that, with direct incorporation of spatial autocorrelation in the penalized likelihood function, the accuracy of variable selection and the precision of parameter estimates improve over penalized methods that do not directly account for spatial dependence.

Furthermore, we have adopted here essentially increasing domain asymptotics. An alternative would be the infill asymptotics. Unlike our increasing domain asymptotic framework where the density of sampling sites is bounded and the spatial domain of interest grows to infinity, the spatial domain of interest is fixed in an infill asymptotic framework and the sampling density tends to infinity. Many of the theoretical results focused on MLE of the parameters in the Matérn family of covariance functions under infill (Zhang, 2004; Kaufman et al., 2008; Du et al., 2009), while much less appears to be known regarding the estimates of the regression coefficients for $d \geq 2$ dimensions. It would be interesting to investigate penalized maximum likelihood under infill asymptotics. We leave this and other possible extensions for future investigation.

2.8 Appendices: Technical Details

For ease of notation, we suppress n in ${}^{n}t_{kk'}$, ${}^{n}a_{kk'}$, ${}^{n}\boldsymbol{\Gamma}$, $\boldsymbol{I}_n$, $\boldsymbol{A}_n$, ${}^{n}\widehat{\boldsymbol{\eta}}$, ${}^{n}\widehat{\boldsymbol{\beta}}$ and ${}^{n}\widehat{\boldsymbol{\theta}}$.

2.8.1 Appendix A: Asymptotic Properties of PMLE and OSE

Lemma 1. *Under (A.1)—(A.7), for any given $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, we have, as $n \to \infty$,*

$$N_n^{-1/2}\ell'(\boldsymbol{\eta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\eta})), \quad N_n^{-1}\ell''(\boldsymbol{\eta}) \xrightarrow{P} -\boldsymbol{J}(\boldsymbol{\eta}),$$

where $\boldsymbol{J}(\boldsymbol{\eta}) = diag\{\boldsymbol{J}(\boldsymbol{\beta}), \boldsymbol{J}(\boldsymbol{\theta})\}$.

Proof. Let $\boldsymbol{W}(\boldsymbol{\eta}) = \boldsymbol{I}(\boldsymbol{\eta})^{-1/2}\ell'(\boldsymbol{\eta})$ and $\boldsymbol{V}(\boldsymbol{\eta}) = \boldsymbol{I}(\boldsymbol{\eta})^{-1/2}\ell''(\boldsymbol{\eta})\boldsymbol{I}(\boldsymbol{\eta})^{-1/2}$. Then, $N_n^{-1/2}\ell'(\boldsymbol{\eta}) = \{N_n^{-1/2}\boldsymbol{I}(\boldsymbol{\eta})^{1/2}\}\boldsymbol{W}(\boldsymbol{\eta})$ and $N_n^{-1}\ell''(\boldsymbol{\eta}) = \{N_n^{-1/2}\boldsymbol{I}(\boldsymbol{\eta})^{1/2}\}\boldsymbol{V}(\boldsymbol{\eta})\{N_n^{-1/2}\boldsymbol{I}(\boldsymbol{\eta})^{1/2}\}$. By Theorem 1 of Sweeting (1980), under (A.1)–(A.6), we have

$$\boldsymbol{W}(\boldsymbol{\eta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_{p+q}), \quad \boldsymbol{V}(\boldsymbol{\eta}) \xrightarrow{P} -\boldsymbol{\mathcal{I}}_{p+q}.$$

Thus, by (A.7) and Slusky's theorem, we have the results of Lemma 1. □

Remark. Lemma 1 establishes the asymptotic behavior of the first-order and the second-order derivatives of the log-likelihood function $\ell(\boldsymbol{\eta})$, scaled by $N_n^{-1/2}$ and N_n^{-1}, respectively. In addition, by Theorem 2 of Mardia and Marshall (1984), $\widehat{\boldsymbol{\eta}}_{\text{MLE}}$ is consistent and asymptotically normal with $\|\widehat{\boldsymbol{\eta}}_{\text{MLE}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2})$ and $N_n^{1/2}(\widehat{\boldsymbol{\eta}}_{\text{MLE}} - \boldsymbol{\eta}_0) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\eta}_0)^{-1})$. Moreover, for a random vector $\boldsymbol{\eta}^*$, such that $\|\boldsymbol{I}(\boldsymbol{\eta})^{1/2}(\boldsymbol{\eta}^* - \boldsymbol{\eta})\| = O_p(1)$, by Theorem 2 of Mardia and Marshall (1984), we have $N_n^{-1}\ell''(\boldsymbol{\eta}^*) \xrightarrow{P} -\boldsymbol{J}(\boldsymbol{\eta})$. These results will be used repeatedly in the proof of Theorems 2.4.1 and 2.4.2.

Proof of Theorem 2.4.1.

Proof. Let $\xi_n = N_n^{-1/2} + a_n$. To establish $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2} + a_n)$, it suffices to show that, for a given constant $\epsilon > 0$, there is a constant C such that, for a sufficiently large n, we have

$$P\left\{\sup_{\|\boldsymbol{u}\|=C} Q(\boldsymbol{\eta}_0 + \xi_n\boldsymbol{u}) < Q(\boldsymbol{\eta}_0)\right\} \geq 1 - \epsilon, \tag{2.17}$$

where $\boldsymbol{u} \in \mathbb{R}^{p+q}$ (Fan and Li, 2001).

Since $p_{\lambda_n}(0) = 0$, we have

$$Q(\boldsymbol{\eta}_0 + \xi_n\boldsymbol{u}) - Q(\boldsymbol{\eta}_0) \quad \leq \quad \ell(\boldsymbol{\eta}_0 + \xi_n\boldsymbol{u}) - \ell(\boldsymbol{\eta}_0) - N_n\sum_{j=1}^{s}\Big\{p_{\lambda_n}(|\beta_{j0} + \xi_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\Big\},$$

where the penalty terms now involve only the s nonzero regression coefficients. By Taylor's expansion, we obtain

$$\ell(\boldsymbol{\eta}_0 + \xi_n\boldsymbol{u}) - \ell(\boldsymbol{\eta}_0) = \xi_n\ell'(\boldsymbol{\eta}_0)^T\boldsymbol{u} - (1/2)N_n\xi_n^2\boldsymbol{u}^T\boldsymbol{J}(\boldsymbol{\eta}_0)\boldsymbol{u}\{1 + o_p(1)\}. \tag{2.18}$$

From Lemma 1 under (A.1)–(A.7), we have $N_n^{-1/2}\ell'(\boldsymbol{\eta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\eta}_0))$. Thus, $\ell'(\boldsymbol{\eta}_0) = O_p(N_n^{1/2})$ and the first term of (2.18) is of order $O_p(N_n^{1/2}\xi_n)$. For a sufficiently large C, the second term dominates the first term in (2.18). Furthermore, by Taylor's expansion and (A.8)–(A.9), the term $N_n \sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0}+\xi_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}$ is bounded by $N_n\xi_n a_n s^{1/2}\|\boldsymbol{u}\| + N_n\xi_n^2 b_n\|\boldsymbol{u}\|^2$, which is again dominated by the second term of (2.18). Thus (2.17) holds for a sufficiently large C.

We now establish the sparsity property (i) by showing that, with probability tending to 1, for any given $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\theta}}$ that satisfy $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\| = O_p(N_n^{-1/2})$ and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$, we have

$$Q\begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix} = \min_{\|\widehat{\boldsymbol{\beta}}_2\| \le C N_n^{-1/2}} Q\begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

for a sufficiently small $\epsilon_n = CN_n^{-1/2}$, which is implied by

$$\frac{\partial Q(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})}{\partial \beta_j} < 0 \quad \text{for} \quad \hat{\beta}_j \in (0, \epsilon_n) \quad \text{and} \quad \frac{\partial Q(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})}{\partial \beta_j} > 0 \quad \text{for} \quad \hat{\beta}_j \in (-\epsilon_n, 0), \quad (2.19)$$

for $j = s+1, \ldots, p$. This argument is an extension from that used in Lemma 1 of Fan and Li (2001) for iid errors.

By Taylor's expansion of (2.19), we have

$$\begin{aligned} \frac{\partial Q(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})}{\partial \beta_j} &= \frac{\partial \ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})}{\partial \beta_j} - N_n p'_{\lambda_n}(|\hat{\beta}_j|)\mathrm{sgn}(\hat{\beta}_j) \\ &= \frac{\partial \ell(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}{\partial \beta_j} + \sum_{j'=1}^{p} \frac{\partial^2 \ell(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)}{\partial \beta_j \partial \beta_{j'}}(\widehat{\beta}_{j'} - \beta_{j'0}) + \sum_{k=1}^{q} \frac{\partial^2 \ell(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)}{\partial \beta_j \partial \theta_k}(\widehat{\theta}_k - \theta_{k0}) \\ &\quad - N_n p'_{\lambda_n}(|\hat{\beta}_j|)\mathrm{sgn}(\hat{\beta}_j), \end{aligned}$$

where $\boldsymbol{\beta}^* = a\widehat{\boldsymbol{\beta}} + (1-a)\boldsymbol{\beta}_0$ and $\boldsymbol{\eta}^* = b\widehat{\boldsymbol{\theta}} + (1-b)\boldsymbol{\theta}_0$ for some $a, b \in (0,1)$. By Lemma 1, we have $N_n^{-1/2}\frac{\partial \ell(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}{\partial \beta_j} = O_p(1)$, $N_n^{-1}\frac{\partial^2 \ell(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)}{\partial \beta_j \partial \beta_{j'}} = \{\boldsymbol{J}(\boldsymbol{\beta}_0)\}_{jj'} + O_p(1)$ and $N_n^{-1}\frac{\partial^2 \ell(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)}{\partial \beta_j \partial \theta_k} = O_p(1)$. Since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(N_n^{-1/2})$ and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$, we have

$$\frac{\partial}{\partial \beta_j} Q(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) = -N_n\Big\{p'_{\lambda_n}(|\hat{\beta}_j|)\mathrm{sgn}(\hat{\beta}_j) + O_p(N_n^{-1/2})\Big\}.$$

By (A.10)–(A.11), the sign of the derivative is determined by sgn($\hat{\beta}_j$) and thus (2.19) holds.

Now, we show the asymptotic normality (ii). The PMLE $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\theta}}^T)^T$ satisfies

$$\left.\frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} = \mathbf{0} \quad \text{and} \quad \left.\frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} = \mathbf{0}.$$

Thus, we have

$$\begin{aligned} \mathbf{0}_{s\times 1} = -\boldsymbol{U}_1 \ell'(\widehat{\boldsymbol{\eta}}) - N_n \boldsymbol{\phi}(\widehat{\boldsymbol{\beta}}_1) &= -\boldsymbol{U}_1 \Big[\ell'(\boldsymbol{\eta}_0) + \{\ell''(\boldsymbol{\eta}_0) + o_p(1)\}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)\Big] \\ &\quad -N_n \Big[\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})\{1 + o_p(1)\}\Big], \end{aligned}$$

where $\boldsymbol{U}_1 = [\boldsymbol{\mathcal{I}}_{s\times s}, \mathbf{0}_{s\times(p-s+q)}]$. By Lemma 1, we have $N_n^{-1/2}\boldsymbol{U}_1\ell'(\boldsymbol{\eta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10}))$. Thus, by Slusky's theorem,

$$N_n^{1/2}\{\boldsymbol{J}(\boldsymbol{\beta}_{10})+\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}\Big[\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+\{\boldsymbol{J}(\boldsymbol{\beta}_{10})+\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}^{-1}\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10})\Big] \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_{10})).$$

Similarly, with $\boldsymbol{U}_2 = [\mathbf{0}_{q\times p}, \boldsymbol{\mathcal{I}}_{q\times q}]$, we have

$$\mathbf{0}_{q\times 1} \;=\; -\boldsymbol{U}_2\ell'(\widehat{\boldsymbol{\eta}}) = \boldsymbol{U}_2\Big[\ell'(\boldsymbol{\eta}_0) + \{\ell''(\boldsymbol{\eta}_0) + o_p(1)\}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)\Big].$$

and $N_n^{1/2}\boldsymbol{J}(\boldsymbol{\theta}_0)\Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0))$. This completes the proof of Theorem 2.4.1. □

Proof of Theorem 2.4.2.

Proof. From Lemma 1 under (A.1)–(A.7), we have $\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$, where $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\text{MLE}}$. Consider the asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\text{OSE}} = \boldsymbol{\beta}_0 + N_n^{-1/2}\widehat{\boldsymbol{u}}$, where $\widehat{\boldsymbol{u}}$ is the maximizer of $Q^*(\boldsymbol{\beta}_0 + N_n^{-1/2}\boldsymbol{u}, \boldsymbol{\theta}^{(0)}) - Q^*(\boldsymbol{\beta}_0, \boldsymbol{\theta}^{(0)})$. Note that

$$\begin{aligned} &Q^*(\boldsymbol{\beta}_0 + N_n^{-1/2}\boldsymbol{u}, \boldsymbol{\theta}^{(0)}) - Q^*(\boldsymbol{\beta}_0, \boldsymbol{\theta}^{(0)}) \\ = \;& -(1/2)\{\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta}_0 + N_n^{-1/2}\boldsymbol{u})\}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}\{\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta}_0 + N_n^{-1/2}\boldsymbol{u})\} \\ &+(1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0) - N_n\sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)(|\beta_{0j} + N_n^{-1/2}u_j| - |\beta_{0j}|) \\ = \;& -(1/2)N_n^{-1}\boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}\boldsymbol{X}\boldsymbol{u} + N_n^{-1/2}\boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0) \\ &-N_n\sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)(|\beta_{0j} + N_n^{-1/2}u_j| - |\beta_{0j}|) \\ \equiv \;& Q_1 + Q_2 + Q_3. \end{aligned}$$

By Lemma 1, we have

$$N_n^{-1}\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}\boldsymbol{X} = N_n^{-1}\frac{\partial^2\ell(\boldsymbol{\beta}_0,\boldsymbol{\theta}^{(0)})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \xrightarrow{P} -\boldsymbol{J}(\boldsymbol{\beta}_0).$$

By Taylor's expansion and Lemma 1,

$$\begin{aligned} N_n^{-1/2}\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0) &= N_n^{-1/2}\frac{\partial\ell(\boldsymbol{\beta}_0,\boldsymbol{\theta}^{(0)})}{\partial\boldsymbol{\beta}} \\ &= N_n^{-1/2}\frac{\partial\ell(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\beta}} + N_n^{-1/2}\frac{\partial\ell(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta}^{(0)}-\boldsymbol{\theta}_0) \\ &\xrightarrow{D} \boldsymbol{W}^T\boldsymbol{J}(\boldsymbol{\beta}_0), \end{aligned}$$

where $\boldsymbol{\theta}^* = a\boldsymbol{\theta}^{(0)} + (1-a)\boldsymbol{\theta}_0$ for some $a \in (0,1)$ and $\boldsymbol{W} \sim N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}_0)^{-1})$. Thus, we have

$$\begin{aligned} Q_1 = -(1/2)N_n^{-1}\boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}\boldsymbol{X}\boldsymbol{u} &\xrightarrow{D} -(1/2)\boldsymbol{u}^T\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{u}, \\ Q_2 = N_n^{-1/2}\boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(0)})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0) &\xrightarrow{D} \boldsymbol{W}^T\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{u}. \end{aligned}$$

Let $\boldsymbol{W} = (\boldsymbol{W}_1^T, \boldsymbol{W}_2^T)^T$ and $\boldsymbol{u} = (\boldsymbol{u}_1^T, \boldsymbol{u}_2^T)^T$. By arguments similar to Zou and Li (2008) for iid errors, under the penalty functions SCAD and (A.10), we have $Q_3 \xrightarrow{P} 0$ if $\boldsymbol{u}_2 = \boldsymbol{0}$; and ∞ otherwise. Thus,

$$\begin{aligned} &Q^*(\boldsymbol{\beta}_0 + N_n^{-1/2}\boldsymbol{u}, \boldsymbol{\theta}^{(0)}) - Q^*(\boldsymbol{\beta}_0, \boldsymbol{\theta}^{(0)}) \\ &\xrightarrow{D} \begin{cases} -(1/2)\boldsymbol{u}_1^T\boldsymbol{J}(\boldsymbol{\beta}_1)\boldsymbol{u}_1 + \boldsymbol{W}_1^T\boldsymbol{J}(\boldsymbol{\beta}_{10})\boldsymbol{u}_1; & \boldsymbol{u}_2 = \boldsymbol{0} \\ \infty; & \text{otherwise,} \end{cases} \end{aligned}$$

which has a unique maximum at $\boldsymbol{u}_1 = \boldsymbol{W}_1$ and $\boldsymbol{u}_2 = \boldsymbol{0}$. Applying arguments in Knight and Fu (2000), we have $\widehat{\boldsymbol{u}}_1 \xrightarrow{D} \boldsymbol{W}_1$ and $\widehat{\boldsymbol{u}}_2 \xrightarrow{P} \boldsymbol{0}$.

Furthermore, the asymptotic normality of $\widehat{\boldsymbol{\theta}}_{\text{OSE}}$ can be shown as in the proof of our Theorem 2.4.1. Indeed, with $\boldsymbol{U}_2 = [\boldsymbol{0}_{q\times p}, \boldsymbol{\mathcal{I}}_{q\times q}]$, we have

$$\boldsymbol{0}_{q\times 1} \;=\; -\boldsymbol{U}_2\ell'(\widehat{\boldsymbol{\eta}}_{\text{OSE}}) = \boldsymbol{U}_2\Big[\ell'(\boldsymbol{\eta}_0) + \{\ell''(\boldsymbol{\eta}_0) + o_p(1)\}(\widehat{\boldsymbol{\eta}}_{\text{OSE}} - \boldsymbol{\eta}_0)\Big]$$

and thus, $N_n^{1/2}\boldsymbol{J}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{\text{OSE}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\theta}_0))$. □

2.8.2 Appendix B: Asymptotic Properties of PMLE$_\mathrm{T}$ and OSE$_\mathrm{T}$

Let $|A|$ denote the cardinality of a discrete set A. Let $\mu_{1,\mathrm{T}} \leq \cdots \leq \mu_{N_n,\mathrm{T}}$ denote the eigenvalues of tapered covariance matrix $\mathbf{\Gamma}_\mathrm{T}$. Let $\mu^k_{l,\mathrm{T}}$ denote the eigenvalues of $\mathbf{\Gamma}_{k,\mathrm{T}}$ such that $|\mu^k_{1,\mathrm{T}}| \leq \cdots \leq |\mu^k_{N_n,\mathrm{T}}|$ and let $\mu^{kk'}_{l,\mathrm{T}}$ denote the eigenvalues of $\mathbf{\Gamma}_{kk',\mathrm{T}}$ such that $|\mu^{kk'}_{1,\mathrm{T}}| \leq \cdots \leq |\mu^{kk'}_{N_n,\mathrm{T}}|$. For a matrix $\boldsymbol{A}$, we let $\mu_{\min}(\boldsymbol{A})$ denote the minimum eigenvalue of $\boldsymbol{A}$. Also, recall that $t_{kk',\mathrm{T}} = \mathrm{tr}(\mathbf{\Gamma}^{-1}_\mathrm{T}\mathbf{\Gamma}_{k,\mathrm{T}}\mathbf{\Gamma}^{-1}_\mathrm{T}\mathbf{\Gamma}_{k',\mathrm{T}})$.

Lemma 2. *Under (A.12)—(A.13), we have*

$$(i)\|\mathbf{\Gamma}-\mathbf{\Gamma}_\mathrm{T}\|_\infty = O(N_n^{-1/2}); (ii)\|\mathbf{\Gamma}_k-\mathbf{\Gamma}_{k,\mathrm{T}}\|_\infty = O(N_n^{-1/2}); (iii)\|\mathbf{\Gamma}_{kk'}-\mathbf{\Gamma}_{kk',\mathrm{T}}\|_\infty = O(N_n^{-1/2}).$$

Proof. We show (i) in detail and omit details for (ii) and (iii), as similar arguments can be applied. Let $A_n = \{1, \ldots, N_n\}$ denote the indexes of N_n sampling sites. Write $\|\mathbf{\Gamma} - \mathbf{\Gamma}_\mathrm{T}\|_\infty$ as

$$\|\mathbf{\Gamma} - \mathbf{\Gamma}_\mathrm{T}\|_\infty = \max\left\{\sum_{i'\in A_n} f_n(d_{ii'}) : i \in A_n\right\},$$

where $f_n(d_{ii'}) = \gamma(d_{ii'})\{1 - (1 - d_{ii'}/\omega_n)_+\}$ is the difference between the covariance function and the tapered covariance function at lag distance $d_{ii'}$ between sampling sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$, for a given threshold distance ω_n. By (A.13)(a), there exists constant D such that $d\gamma(d)$ reaches maximum at D and by (A.12), for a sufficiently large n, $\omega_n > D$. Thus, for any $i \in A_n$,

$$\sum_{i'\in A_n} f_n(d_{ii'}) = \sum_{i'\in A_{1n}} f_n(d_{ii'}) + \sum_{i'\in A_{2n}} f_n(d_{ii'}) + \sum_{i'\in A_{3n}} f_n(d_{ii'}),$$

where $A_{1n} = \{i' : d_{ii'} \leq D\}$, $A_{2n} = \{i' : D < d_{ii'} \leq \omega_n\}$, and $A_{3n} = \{i' : d_{ii'} > \omega_n\}$.

Since the lag distance between any two sampling sites is greater than a constant, the sampling density of any subset of $R_n \in \mathbb{R}^2$ is bounded by a constant, say ρ. Thus, $|A_{1n}| \leq \pi\rho D^2$. By (A.12) and (A.13)(a), it follows that, for some $C_1 > 0$,

$$\sum_{i'\in A_{1n}} f_n(d_{ii'}) \leq \pi\rho D^2 \max_{i'\in A_{1n}} f_n(d_{ii'}) \leq \pi\rho D^3\gamma(D)/\omega_n \leq C_1 N_n^{-1/2}.$$

Next, let $B_m = \{i' : mh < d_{ii'} \leq (m+1)h\}$, where h is independent of n. Thus, $|B_m| \leq (2m+1)\rho\pi h^2$. It can be shown that, $A_{2n} \subset \bigcup_{m=\lfloor (D/h) \rfloor}^{\lfloor (\omega_n/h) \rfloor + 1} B_m$ and $A_{3n} \subset \bigcup_{m=\lfloor (\omega_n/h) \rfloor + 1}^{\infty} B_m$, where $\lfloor \cdot \rfloor$ denotes the floor function. Moreover,

$$\sum_{i' \in A_{2n}} f_n(d_{ii'}) \leq \sum_{m=\lfloor (D/h) \rfloor}^{\lfloor (\omega_n/h) \rfloor + 1} (2m+1)\pi\rho h^2 \max_{i' \in B_m} f_n(d_{ii'}) \leq \omega_n^{-1}\pi\rho \sum_{m=0}^{\infty} 3mh^2 \max_{i' \in B_m} \{d_{ii'}\gamma(d_{ii'})\}.$$

As $h \to 0$, we have $\sum_{i' \in A_{2n}} f_n(d_{ii'}) \leq \omega_n^{-1}\pi\rho \int_0^{\infty} 3u^2\gamma(u)du \leq C_2 N_n^{-1/2}$ for some $C_2 > 0$.

Similarly, we have

$$\sum_{i' \in A_{3n}} f_n(d_{ii'}) \leq \sum_{m=\lfloor (\omega_n/h) \rfloor}^{\infty} (3m)\pi\rho h^2 \max_{i' \in B_m} \gamma(d_{ii'}) \leq C_3 N_n^{-1/2}$$

for some $C_3 > 0$.

Combining the three inequalities above, we have $\sum_{i' \in A_n} f_n(d_{ii'}) \leq (C_1 + C_2 + C_3)N_n^{-1/2}$, for all $i \in A_n$. □

Remark. Lemma 2 establishes that the order of the difference between the covariance matrix $\mathbf{\Gamma}$ and the tapered covariance matrix $\mathbf{\Gamma}_{\mathrm{T}}$ is $N_n^{-1/2}$, as well as that of the first-order and the second-order derivatives of the covariance matrices. These results are used when establishing Lemma 3.

Lemma 3. *Under (A.1)–(A.4), (A.6), and (A.12)–(A.13), we have*

(C.1) $\lim_{n\to\infty} \mu_{N_n,\mathrm{T}} = C < \infty$, $\lim_{n\to\infty} |\mu_{N_n,\mathrm{T}}^{k}| = C_k < \infty$, $\lim_{n\to\infty} |\mu_{N_n,\mathrm{T}}^{kk'}| = C_{kk'} < \infty$ *for any* $k, k' = 1, \ldots, q$.

(C.2) For $k = 1, \ldots, q$, $\|\mathbf{\Gamma}_{k,\mathrm{T}}\|_F^{-2} = O(N_n^{-1/2-\delta})$, *for some* $\delta > 0$.

(C.3) $\|\mathbf{\Gamma}_{\mathrm{T}}^{-1}\|_s < C_0 < \infty$.

(C.4) For any $k, k' = 1, \ldots, q$, $a_{kk',\mathrm{T}} = \lim\{t_{kk',\mathrm{T}}(t_{kk,\mathrm{T}}t_{k'k',\mathrm{T}})^{-1/2}\}$ *exists and is equal to* $a_{kk'} = \lim\{t_{kk'}(t_{kk}t_{k'k'})^{-1/2}\}$. *That is,* $\boldsymbol{A}_{\mathrm{T}} = (a_{kk',\mathrm{T}})_{k,k'=1}^{q} = \boldsymbol{A} = (a_{kk'})_{k,k'=1}^{q}$ *and is nonsingular.*

Proof. First, we show (C.1). Since $|\|\boldsymbol{\Gamma}\|_s - \|\boldsymbol{\Gamma}_{\mathrm{T}}\|_s| \le \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_{\mathrm{T}}\|_s \le \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_{\mathrm{T}}\|_\infty \to 0$ by (i) of Lemma 2, we have $\lim_{n\to\infty} \|\boldsymbol{\Gamma}\|_s = \lim_{n\to\infty} \|\boldsymbol{\Gamma}_{\mathrm{T}}\|_s$. Thus, $\lim_{n\to\infty} \mu_{N_n,\mathrm{T}} = \lim_{n\to\infty} \|\boldsymbol{\Gamma}_{\mathrm{T}}\|_s = \lim_{n\to\infty} \|\boldsymbol{\Gamma}\|_s = \lim_{n\to\infty} \mu_{N_n} = C < \infty$, by (A.2). By similar arguments, $\lim_{n\to\infty} |\mu^k_{N_n,\mathrm{T}}| = C_k < \infty$, $\lim_{n\to\infty} |\mu^{kk'}_{N_n,\mathrm{T}}| = C_{kk'} < \infty$.

Next, we show (C.2). By (A.3)(i), $\|\boldsymbol{\Gamma}_k\|_F^{-2} = O(N_n^{-1/2-\delta})$ for some $\delta > 0$. Also, $\|\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{k,\mathrm{T}}\|_F \le N_n^{1/2} \|\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{k,\mathrm{T}}\|_\infty = O(1)$, by (ii) of Lemma 2. Thus,

$$\|\boldsymbol{\Gamma}_{k,\mathrm{T}}\|_F \quad \ge \quad \|\boldsymbol{\Gamma}_k\|_F - \|\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{k,\mathrm{T}}\|_F = O(N_n^{1/4+\delta/2}) \quad \text{and} \quad \|\boldsymbol{\Gamma}_{k,\mathrm{T}}\|_F^{-2} = O(N_n^{-1/2-\delta}).$$

Now, we show (C.3). Since $\boldsymbol{\Delta}(\omega)$ and $\boldsymbol{\Gamma}$ are both semi-positive definite and the diagonal elements of matrix $\boldsymbol{\Delta}(\omega)$ are 1's, we have

$$\mu_{\min}(\boldsymbol{\Gamma}) \le \mu_{\min}\{\boldsymbol{\Gamma} \circ \boldsymbol{\Delta}(\omega)\} = \mu_{\min}(\boldsymbol{\Gamma}_{\mathrm{T}})$$

(Chapter 5, Horn and Johnson (1991)). Thus, by (A.6),

$$\|\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\|_s = \mu_{\min}(\boldsymbol{\Gamma}_{\mathrm{T}})^{-1} \le \mu_{\min}(\boldsymbol{\Gamma})^{-1} = \|\boldsymbol{\Gamma}^{-1}\|_s < C_0 < \infty.$$

Finally, we show (C.4). We first show that $|t_{kk',\mathrm{T}} - t_{kk'}| = O(N_n^{1/2})$. Note that

$$\begin{aligned}
|t_{kk',\mathrm{T}} - t_{kk'}| &= |\mathrm{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k,\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}}) - \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'})| \\
&\le |\mathrm{tr}\{(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1})\boldsymbol{\Gamma}_{k,\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}}\}| + |\mathrm{tr}\{\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_{k,\mathrm{T}} - \boldsymbol{\Gamma}_k)\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}}\}| \\
&+ |\mathrm{tr}\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1})\boldsymbol{\Gamma}_{k',\mathrm{T}})| + |\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_{k',\mathrm{T}} - \boldsymbol{\Gamma}_{k'})\}| \\
&\equiv G_1 + G_2 + G_3 + G_4.
\end{aligned}$$

For $G_1 = |\mathrm{tr}\{(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1})\boldsymbol{\Gamma}_{k,\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}}\}| \le N_n \|\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1}\|_s \|\boldsymbol{\Gamma}_{k,\mathrm{T}}\|_s \|\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\|_s \|\boldsymbol{\Gamma}_{k',\mathrm{T}}\|_s$, the last three norms are of order $O(1)$ by (C.1) and (C.3). Further, by (A.6) and Lemma 2, $\|\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1}\|_s = \|\boldsymbol{\Gamma}^{-1}\|_s^2 \|\boldsymbol{\Gamma}_{\mathrm{T}} - \boldsymbol{\Gamma}\|_s (1 - \|\boldsymbol{\Gamma}^{-1}\|_s \|\boldsymbol{\Gamma}_{\mathrm{T}} - \boldsymbol{\Gamma}\|_s)^{-1} = O(N_n^{-1/2})$ (Stewart, 1990). Thus, we have $G_1 = O(N_n^{1/2})$. By similar arguments, $G_i = O(N_n^{1/2})$, for $i = 2, 3, 4$.

We then show that $a_{kk',\mathrm{T}} = a_{kk'}$. For any k and k', by (A.3)(ii), either $\|\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_{k'}\|_F^{-2} = O(N_n^{-1/2-\delta})$ or $\|\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{k'}\|_F^{-2} = O(N_n^{-1/2-\delta})$. Without loss of generality, we assume the first condition. Also, $2(a_{kk',\mathrm{T}} - a_{kk'}) = \lim_{n\to\infty}(2t_{kk',\mathrm{T}}t_{kk,\mathrm{T}}^{-1/2}t_{k'k',\mathrm{T}}^{-1/2} - 2t_{kk'}t_{kk}^{-1/2}t_{k'k'}^{-1/2}) = \lim_{n\to\infty}(H_1 - H_2 - H_3)$, where

$$\begin{aligned}
& 2t_{kk',\mathrm{T}}t_{kk,\mathrm{T}}^{-1/2}t_{k'k',\mathrm{T}}^{-1/2} - 2t_{kk'}t_{kk}^{-1/2}t_{k'k'}^{-1/2} \\
= & \Big[\Big\{(2t_{kk',\mathrm{T}} + t_{kk,\mathrm{T}} + t_{k'k',\mathrm{T}})t_{kk,\mathrm{T}}^{-1/2}t_{k'k',\mathrm{T}}^{-1/2} - (2t_{kk'} + t_{kk} + t_{k'k'})t_{kk}^{-1/2}t_{k'k'}^{-1/2}\Big\} \\
& \quad - \Big(t_{kk,\mathrm{T}}^{1/2}t_{k'k',\mathrm{T}}^{-1/2} - t_{kk}^{1/2}t_{k'k'}^{-1/2}\Big) - \Big(t_{k'k',\mathrm{T}}^{1/2}t_{kk,\mathrm{T}}^{-1/2} - t_{k'k'}^{1/2}t_{kk}^{-1/2}\Big)\Big] \equiv H_1 - H_2 - H_3.
\end{aligned}$$

It is straightforward to verify that $H_1 = (2t_{kk'} + t_{kk} + t_{k'k'})t_{kk}^{-1/2}t_{k'k'}^{-1/2}\left(H_{11}H_{12}^{-1} - 1\right)$ where

$$\begin{aligned}
H_{11} &= 1 + \{2t_{kk',\mathrm{T}} + t_{kk,\mathrm{T}} + t_{k'k',\mathrm{T}} - (2t_{kk'} + t_{kk} + t_{k'k'})\}(2t_{kk'} + t_{kk} + t_{k'k'})^{-1}, \\
H_{12} &= \left\{1 + (t_{kk,\mathrm{T}} - t_{kk})t_{kk}^{-1}\right\}^{1/2}\left\{1 + (t_{k'k',\mathrm{T}} - t_{k'k'})t_{k'k'}^{-1}\right\}^{1/2}.
\end{aligned}$$

Since $2t_{kk'} + t_{kk} + t_{k'k'} = \mathrm{tr}\{\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_{k'})\boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_{k'})\} \geq \mu_{N_n}^{-2}\|\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_{k'}\|_F^2$ and $|2t_{kk',\mathrm{T}} + t_{kk,\mathrm{T}} + t_{k'k',\mathrm{T}} - (2t_{kk'} + t_{kk} + t_{k'k'})| \leq 2|t_{kk',\mathrm{T}} - t_{kk'}| + |t_{kk,\mathrm{T}} - t_{kk}| + |t_{k'k',\mathrm{T}} - t_{k'k'}| = O(N_n^{1/2})$, we have $H_{11} = 1 + O(N_n^{-\delta})$. Since $t_{kk} = \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) \geq \mu_{N_n}^{-2}\|\boldsymbol{\Gamma}_k\|_F^2$, we also have $H_{12} = 1 + O(N_n^{-\delta})$. In addition, $(2t_{kk'} + t_{kk} + t_{k'k'})t_{kk}^{-1/2}t_{k'k'}^{-1/2}$ is bounded, since $\left|t_{kk}t_{k'k'}^{-1}\right|$ and $\left|t_{k'k'}t_{kk}^{-1}\right|$ are bounded by (A.4)(ii). Thus, we have $H_1 \to 0$, as $n \to \infty$. For H_2, we have

$$H_2 = t_{kk}t_{k'k'}^{-1}\left[\frac{\left\{1 + (t_{kk,\mathrm{T}} - t_{kk})t_{kk}^{-1}\right\}^{1/2}}{\left\{1 + (t_{k'k',\mathrm{T}} - t_{k'k'})t_{k'k'}^{-1}\right\}^{1/2}} - 1\right] = t_{kk}t_{k'k'}^{-1}\left\{\frac{1 + O(N_n^{-\delta})}{1 + O(N_n^{-\delta})} - 1\right\} \to 0.$$

Similarly, $H_3 \to 0$. Thus, $a_{kk',\mathrm{T}} = a_{kk'}$ and the matrix $\boldsymbol{A}_\mathrm{T}$ is identical to the matrix $\boldsymbol{A}$, which is nonsingular by (A.4)(i). □

Remark. (C.1)–(C.4) are the covariance tapering counterparts of (A.2), (A.3)(i), (A.4)(i), and (A.6). Together with (A.5), they yield Proposition 2.5.1. In fact, Lemmas 2 and 3 hold for other tapering functions such as truncated polynomial functions of d/ω with constant term equal to 1 when $d < \omega$, and 0 otherwise (Wendland, 1995). Furthermore, (A.12) can be weakened to $0 < \inf_n\{\omega_n N_n^{-1/2+\tau}\} \leq \sup_n\{\omega_n N_n^{-1/2+\tau}\} < \infty$, with $\tau < \min\{1/2, \delta\}$.

Lemma 4. *Under (A.1)–(A.7) and (A.12)–(A.13), for any given $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, we have*

$$N_n^{-1/2}\ell'_{\mathrm{T}}(\boldsymbol{\eta}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\eta})) \quad \textit{and} \quad N_n^{-1}\ell''_{\mathrm{T}}(\boldsymbol{\eta}) \xrightarrow{P} -\boldsymbol{J}(\boldsymbol{\eta}),$$

where recall that $\boldsymbol{J}(\boldsymbol{\eta}) = diag\{\boldsymbol{J}(\boldsymbol{\beta}), \boldsymbol{J}(\boldsymbol{\theta})\}$.

Proof. From Lemma 3, we have (C.1)–(C.4). Together with (A.5), the regularity conditions of Theorem 1 of Sweeting (1980) hold. Thus, we have

$$\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\eta})^{-1/2}\ell'_{\mathrm{T}}(\boldsymbol{\eta}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\mathcal{I}}_{p+q}) \quad \text{and} \quad \boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\eta})^{-1/2}\ell''_{\mathrm{T}}(\boldsymbol{\eta})\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\eta})^{-1/2} \xrightarrow{P} \boldsymbol{\mathcal{I}}_{p+q}.$$

By Slusky's theorem, it suffices to show that $N_n^{-1}\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\eta}) \to \boldsymbol{J}(\boldsymbol{\eta})$.

Note that $\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}) - \boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1})\boldsymbol{X}$. By Lemma 2 and (A.5),

$$\|N_n^{-1}\{\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}) - \boldsymbol{I}(\boldsymbol{\beta})\}\|_{\max} \leq \|\boldsymbol{\Gamma}_{\mathrm{T}}^{-1} - \boldsymbol{\Gamma}^{-1}\|_{\infty}\|\boldsymbol{X}\|_{\max}^2 = O(N_n^{-1/2})\|\boldsymbol{X}\|_{\max}^2 = O(N_n^{-1/2}).$$

Thus, $N_n^{-1}\{\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}) - \boldsymbol{I}(\boldsymbol{\beta})\} \to 0$. By (A.7), we have $N_n^{-1}\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\beta}) \to \boldsymbol{J}(\boldsymbol{\beta})$. Furthermore, the (k, k')th entry of $N_n^{-1}\{\boldsymbol{I}_{\mathrm{T}}(\boldsymbol{\theta}) - \boldsymbol{I}(\boldsymbol{\theta})\}$ is $(2N_n)^{-1}(t_{kk',\mathrm{T}} - t_{kk'})$, which tends to zero as shown in the proof of (C.4) in Lemma 3. □

Remark. Lemma 4 establishes the asymptotic behavior of the first-order and the second-order derivatives of the covariance-tapered log-likelihood function $\ell_{\mathrm{T}}(\boldsymbol{\eta})$. The rates of convergence and the limiting distributions are the same as those for the log-likelihood function. As in Lemma 1, it follows that MLE$_{\mathrm{T}}$ $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$ is consistent and asymptotically normal, as is given in Proposition 2.5.1. These results will be used to establish Theorems 2.5.2 and 2.5.3 and play the same role as Lemma 1 when showing Theorems 2.4.1 and 2.4.2.

Proof of Proposition 2.5.1.

Proof. From Lemma 3, (C.1)–(C.4) are satisfied. Together with (A.5), the regularity conditions of Theorem 2 of Mardia and Marshall (1984) hold. Thus the result in Proposition 2.5.1 follows. □

Proof of Theorem 2.5.2.

Proof. The proof of Theorem 2.5.2 is similar to that of Theorem 2.4.1. The main differences are that the parameter estimates $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE}}$, log-likelihood function $\ell(\boldsymbol{\eta})$, and penalized log-likelihood $Q(\boldsymbol{\eta})$ are replaced with their covariance-tapered counterparts $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE}_{\mathrm{T}}}$, $\ell_{\mathrm{T}}(\boldsymbol{\eta})$, and $Q_{\mathrm{T}}(\boldsymbol{\eta})$, respectively. Furthermore, we replace the results from Lemma 1 with those from Lemma 4, which holds due to Lemma 2–3 under the additional assumptions (A.12)–(A.13). □

Proof of Theorem 2.5.3.

Proof. The proof of Theorem 2.5.3 is similar to that of Theorem 2.4.2, but we replace the parameter estimates $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE}}$, log-likelihood function $\ell(\boldsymbol{\eta})$, and $Q^*(\boldsymbol{\beta})$ with their covariance-tapered counterparts $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE}_{\mathrm{T}}}$, $\ell_{\mathrm{T}}(\boldsymbol{\eta})$, and $Q^*_{\mathrm{T}}(\boldsymbol{\beta})$, respectively. As before, we replace the results from Lemma 1 with those from Lemma 4, where the additional conditions (A.12) and (A.13) are assumed and Lemma 2–3 are applied. □

2.8.3 Appendix C: Remarks on Assumptions (A.2), (A.3), (A.6) and (A.7)

Assumption (A.2) is the same as that in Mardia and Marshall (1984). In fact, it can be relaxed by replacing lim with lim sup. We consider a one-dimensional grid with the exponential covariance function $\gamma(u) = \sigma^2 e^{-u/r}$. It is easy to see that $\boldsymbol{\Gamma} = [\sigma^2 e^{-|i-i'|/r}]_{i,i'=1}^N$, $\boldsymbol{\Gamma}_1 = [e^{-|i-i'|/r}]_{i,i'=1}^N$, $\boldsymbol{\Gamma}_2 = [\sigma^2 |i-i'| r^{-2} e^{-|i-i'|/r}]_{i,i'=1}^N$, $\boldsymbol{\Gamma}_{1,1} = [0]_{i,i'=1}^N$, $\boldsymbol{\Gamma}_{1,2} = [|i-i'| r^{-2} e^{-|i-i'|/r}]_{i,i'=1}^N$ and $\boldsymbol{\Gamma}_{2,2} = [-2\sigma^2(|i-i'| r^{-3} e^{-|i-i'|/r} + |i-i'|^2 r^{-4} e^{-|i-i'|/r})]_{i,i'=1}^N$.

Since the spectral radius of any matrix is less than L_1-norm of the matrix, it suffices to show that L_1-norm of the above matrices are bounded.

Assuming that parameters r and σ^2 are in some closed set. That is, there exist constants r_1, r_2, σ_1^2 and σ_2^2, such that all pairs $(r, \sigma^2) \in [r_1, r_2] \times [\sigma_1^2, \sigma_2^2]$. We have

$$
\begin{aligned}
&\|\mathbf{\Gamma}\|_1 \leq 2\sigma^2 \sum_{i=0}^{\infty} \rho^i = 2\sigma^2/(1-\rho) < 2\sigma_2^2/(1-e^{-r_1}),\\
&\|\mathbf{\Gamma}_1\|_1 \leq 2 \sum_{i=0}^{\infty} \rho^i = 2/(1-\rho) < 2/(1-e^{-r_1}),\\
&\|\mathbf{\Gamma}_2\|_1 \leq 2\sigma^2 r^{-2} \sum_{i=0}^{\infty} i\rho^i = 2\sigma^2 r^{-2}\rho/(1-\rho)^2 < 2\sigma_2^2 r_1^{-2}/(1-e^{-r_1})^2,\\
&\|\mathbf{\Gamma}_{1,1}\|_1 = 0,\\
&\|\mathbf{\Gamma}_{1,2}\|_1 \leq 2r^{-2} \sum_{i=0}^{\infty} i\rho^i = 2r^{-2}\rho/(1-\rho)^2 < 2r_1^{-2}/(1-e^{-r_1})^2,\\
&\|\mathbf{\Gamma}_{2,2}\|_1 \leq 4\sigma^2 r^{-3} \sum_{i=0}^{\infty} i\rho^i + 2\sigma^2 r^{-4} \sum_{i=0}^{\infty} i^2\rho^i = 4r^{-3}\rho/(1-\rho)^2 + 3r^{-4}(\rho+\rho^2)/(1-\rho)^3\\
&< (4r_1^{-3} + 6r_1^{-4})/(1-e^{-r_1})^3,
\end{aligned}
$$

where $\rho = e^{-1/r}$.

For (A.3), $\|\mathbf{\Gamma}_1\|_F^2 \geq N$, $\|\mathbf{\Gamma}_2\|_F^2 = 2r^{-4}\sum_{i=1}^{N}(Ni^2\rho^{2i} - i^3\rho^{2i})$ $\geq 2r^{-4}\{N\rho^2 - 24\rho(1-\rho)^{-1} - 48\rho^2(1-\rho)^{-2} - 32\rho^3(1-\rho)^{-3} - 6\rho^4(1-\rho)^{-4}\} > 2r_2^{-4}\{Ne^{-r_2} - 110e^{-r_1}(1-e^{-r_1})^{-4}\}$ and $\|\mathbf{\Gamma}_1 + \mathbf{\Gamma}_2\|_F^2 \geq N$. Therefore, (A.3) is satisfied with $\delta = 1/2$.

For (A.6), from Brockwell and Davis (1991), the smallest eigenvalue of $\mathbf{\Gamma}$,

$$
\begin{aligned}
\mu_1 &\geq (2\pi)^{-1}\sigma^2 \sum_{i=-\infty}^{\infty} e^{-i/r}e^{-ii\omega}\\
&= (2\pi)^{-1}\sigma^2(1-\rho^2)/(1-2\rho cos(\omega)+\rho^2) \geq (2\pi)^{-1}\sigma^2(1-\rho^2)/(1+\rho)^2.
\end{aligned}
$$

Therefore, $\|\mathbf{\Gamma}^{-1}\|_s \leq (2\pi)\sigma^{-2}(1+\rho)^2/(1-\rho^2) \leq (2\pi)\sigma_1^{-2}(1+e^{-r_1})^2/(1-e^{-2r_1})$. Moreover, in Section 4 of Mardia and Marshall (1984), it was shown that under some mild assumptions, a more general family of covariance functions satisfy these assumptions.

Assumption (A.7) follows directly from Zou and Li (2008) and Wang and Zhu (2009).

2.8.4 Appendix D: Remark on Assumption (A.13)

Here we show that $\gamma(u) = \sigma^2(1-c)\Gamma(\nu)^{-1}(ru/2)^{\nu}2K_{\nu}(ru)$ from the Matérn class of covariance function satisfies (A.13).

Observe that

$$\partial\gamma(u)/\partial\sigma^2 = (1-c)\Gamma(\nu)^{-1}(ru/2)^{\nu}2K_{\nu}(ru),$$

$$\partial\gamma(u)/\partial c = -\sigma^2\Gamma(\nu)^{-1}(ru/2)^{\nu}2K_{\nu}(ru),$$

$$\partial\gamma(u)/\partial r = \sigma^2\Gamma(\nu)^{-1}r^{-1}(ru/2)^{\nu}\{2\nu K_{\nu}(ru) - 2c\nu K_{\nu}(ru) - ruK_{\nu+1}(ru) + cruK_{\nu}(ru)\},$$

$$\partial^2\gamma(u)/(\partial\sigma^2)^2 = 0,$$

$$\partial^2\gamma(u)/\partial\sigma^2\partial c = -\Gamma(\nu)^{-1}(ru/2)^{\nu}2K_{\nu}(ru),$$

$$\partial^2\gamma(u)/\partial\sigma^2\partial r = \Gamma(\nu)^{-1}r^{-1}(ru/2)^{\nu}\{2\nu K_{\nu}(ru) - 2c\nu K_{\nu}(ru) - ruK_{\nu+1}(ru) + cruK_{\nu}(ru)\},$$

$$\partial^2\gamma(u)/\partial c^2 = 0,$$

$$\partial^2\gamma(u)/\partial c\partial r = \sigma^2\Gamma(\nu)^{-1}r^{-1}(ru/2)^{\nu}\{-2\nu K_{\nu}(ru) + ruK_{\nu}(ru)\},$$

$$\partial^2\gamma(u)/\partial r^2 = \sigma^2\Gamma(\nu)^{-1}r^{-2}(ru/2)^{\nu}\{4\nu^2K_{\nu}(ru) - 2\nu K_{\nu}(ru) - 2\nu ruK_{\nu+1}(ru) + r^2u^2K_{\nu}(ru) + ruK_{\nu+1}(ru)\}.$$

Note that the covariance function and its first-order and second-order partial derivatives are linear combinations of a Bessel function of u times a polynomial of u. In order to prove (A.13), it suffices to show that, for $a, b > 0$,

(i) $\int_0^{\infty} u^{b+2}K_a(u)du < \infty$.

(ii) $x^{1/2}\int_x^{\infty} u^{b+1}K_a(u)du < \infty$, as $x \to \infty$.

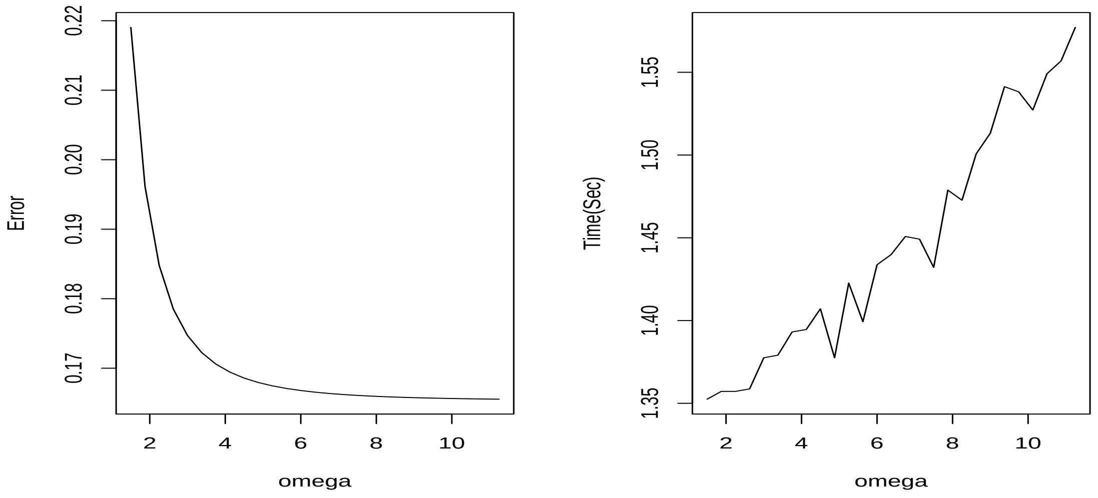
(iii) $u^bK_a(u) \to 0$, as $u \to \infty$.

Since $K_a(u) \propto e^{-u}u^{-1/2}\{1+O(1/u)\}$ when $|u| \to \infty$, there exists C and M such that $K_a(u) \le Me^{-u}u^{-1/2}(1+C/u)$, when $|u| > C$. Thus, for (i)

$$\begin{aligned}\int_0^{\infty} u^{b+2}K_a(u)du &= \int_0^{C} u^{b+2}K_a(u)du + \int_C^{\infty} u^{b+2}K_a(u)du \\ &\le \int_0^{C} u^{b+2}K_a(u)du + \int_0^{\infty} Mu^{b+3/2}e^{-u}(1+C/u)du.\end{aligned}$$

The first term is bounded because the limit of the integral is bounded. The second term $\int_0^{\infty} Mu^{b+3/2}e^{-u}(1+C/u)du = \int_0^{\infty} Mu^{b+3/2}e^{-u}du + \int_0^{\infty} CMu^{b+1/2}e^{-u}du = M(b+3/2)\Gamma(b+3/2) + CM(b+1/2)\Gamma(b+1/2)$ is also bounded.

Figure 2.2: Plots of Error and Time versus Omega.



For (ii),

$$x^{1/2}\int_x^\infty u^{b+1}K_a(u)du \quad = \quad x^{1/2}\int_{min\{x,C\}}^{C} u^{b+1}K_a(u)du + x^{1/2}\int_{max\{C,x\}}^{\infty} u^{b+1}K_a(u)du,$$

when $x \to \infty$, the first term tends to 0, and the second term is $x^{1/2}\int_x^\infty u^{b+1}K_a(u)du \leq x^{1/2}\int_x^\infty Mu^{b+1/2}e^{-u}du + x^{1/2}\int_x^\infty CMu^{b-1/2}e^{-u}du$. To show that it is bounded, it suffices to show that $x^{1/2}\int_x^\infty u^k e^{-u}du$ is bounded, where $k \geq b + 1/2$ is an integer. Since $\int_x^\infty u^k e^{-u}du = P(x)e^{-x}$, where $P(x)$ is a polynomial of x, $x^{1/2}\int_x^\infty u^k e^{-u}du$ is bounded and in fact, tends to 0.

For (iii), $u^b K_a(u) \leq Mu^{b-1/2}(1+C/u)e^{-u}$, when $|u| > C$. Therefore, $u^b K_a(u) \to 0$, as $u \to \infty$.

2.8.5 Appendix E: The Choice of ω

We following the empirical rule suggested by Kaufman et al. (2008). For a sample size $N = 900$, we use 20% of the data for estimation and obtain $\widehat{\boldsymbol{\beta}}_p = (3.87, 3.32, 2.00, 1.28, -0.01, -0.22, -0.18)^T$ and $\widehat{\boldsymbol{\theta}}_p = (0.97, 0.21, 8.32)^T$. Then we compute the estimated error and record time of evaluating $\ell_{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_p; \boldsymbol{y}, \boldsymbol{X})$, as shown in Figure 2.2. From the plots above, one reasonable choice is $\omega = 6$. We further

apply our method to the entire data set, and obtain the final estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}} = (4.22, 3.05, 1.95, 0.97, 0, 0, 0)^T$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}} = (1.21, 0.27, 8.15)^T$. Without tapering, the estimates are
$\widehat{\boldsymbol{\beta}}_{\mathrm{OSE}} = (4.22, 3.05, 1.95, 0.97, 0, 0, 0)^T$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}} = (1.00, 0.23, 8.34)^T$. Note that these estimates are quite close to each other.

Chapter 3

LOCAL KARHUNEN-LOÈVE EXPANSION

3.1 Introduction

Geostatistics are used in many scientific studies that involve analysis of spatially correlated data in a spatial domain (see, e.g., Cressie, 1993; Stein, 1999). A geostatistical model, in its general form, is a random field for an attribute of interest such that the random field is a stochastic process over a continuous index within the spatial domain. Based on geostatistical data sampled at point locations, statistical inference about the geostatistical model can be drawn. The main purpose of this chapter is to develop a novel semiparametric approach to spatial modeling and statistical inference that accounts for spatial dependence in a robust manner and carries out the computation efficiently.

For a spatial linear model, Mardia and Marshall (1984) considered maximum likelihood estimates (MLE) of the model parameters and established asymptotic properties of the MLE under regularity conditions. The computational complexity of these MLEs for a sample of size N, however, is on the order of N^3, making the computation demanding for large N (see, e.g., Cressie, 1993). To reduce such a computational burden, various methods based on approximations have been developed. One such method, covariance tapering, rescales the spatial correlation function by a weight function of the distance between two locations, effectively truncating the spatial correlation to zero when the distance exceeds a certain threshold. The resulting tapered covariance matrix as an approximation of the true covariance matrix is sparse and thus fast to compute at an appropriately chosen threshold (see, e.g., Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009; Chu et al., 2011). Alternatively, Caragea and Smith (2007) partitioned the spatial domain into blocks and approximated the likelihood function by an estimating function that separates variability within blocks and between blocks. In a so-called small-block case, the spatial processes in different small blocks are assumed to be independent, giving rise to a block-diagonal covariance matrix that is also fast to compute. Furthermore, the theoretical properties of the small-block method were established under certain conditions.

The aforementioned methods, however, assume a parametric form for the spatial covariance function, upon which the performance of statistical inference and the asymptotic results hinge. In contrast, semiparametric modeling offers an attractive alternative, as the spatial covariance function does not need to be pre-specified. The corresponding approach tends to be more flexible and potentially more robust against model misspecification (see, e.g., Im et al., 2007; Cressie and Johannesson, 2008; Zhang and Wang, 2010). For example, Cressie and Johannesson (2008) considered a flexible family of nonstationary spatial covariance functions and developed a fixed-rank Kriging. In particular, the true covariance function is assumed to be from a finite expansion of basis functions, such as splines, in a certain sequence. The covariance function was estimated by a method of moment using an empirical covariance matrix under the Frobenius norm, which was then used for Kriging. Im et al. (2007) considered the frequency domain and used B-splines to model the spectral density function, from which the covariance function is derived using the Hankel transform. The MLEs of the model parameters are computed using simulated annealing. Despite the added model flexibility, the aforementioned methods primarily focus on spatial interpolation and there appears to be little or no theoretical backing. Thus, it is of interest to develop innovative semiparametric methods for inference in general and to explore their theoretical properties in geostatistics.

In this chapter, we aim to develop a new semiparametric approach to geostatistical modeling and inference. In particular, we consider a geostatistical model with additive components, namely, a fixed mean possibly in the form of linear regression and Gaussian random errors. The spatial covariance function is left unspecified and thus flexible, enhancing the robustness against model misspecification. A novel local Karhunen-Loève expansion is developed to approximate the spatial random error. In addition, we devise a likelihood-based method for estimating the model parameters and drawing inference. The computational algorithm developed utilizes Newton-Raphson on a Stiefel manifold recently developed by Peng and Paul (2009) and the existing computational method for linear mixed models (see, e.g., Pinheiro and Bates, 2000). Our approach applies to estimation of regression coefficients, selection of covariates, and nonparametric estimation of the covariance function, in addition to spatial interpolation. Unlike Cressie and Johannesson (2008) who assumed a low-rank type representation to be the true underlying model, our method does not make such an assumption and offers a principled approach to approximate the true, unspecified spatial covariance function. Further, while we approximate the likelihood function by employing a technique similar to the small-block idea, our method does not assume a parametric form for the spatial covariance function as Caragea and Smith (2007). Finally, although more model flexibility is attained, it becomes substantially more challenging to establish the theoretical properties of semiparametric methods, an issue that is often not pursued in the existing literature. Here, we make an attempt to establish some theoretical result and in particular, the consistency of likelihood-based estimates of regression coefficients and spatial covariance function.

The remainder of the chapter is organized as follows. In Section 3.2, we describe a general geostatistical model and a local Karhunen-Loève expansion for the spatial random error. In Section 3.3, we develop a likelihood-based method for parameter estimation and a modification of the estimation to increase accuracy and numerical stability. Spatial interpolation and model selection are also implemented based on the parameter estimates. In Section 3.4, a simulation study is given to investigate the finite-sample properties of the inference in comparison with several alternative approaches, as well as a real data example. We establish the consistency of the estimates in Section 3.5. The technical proof and more simulation results are in Section 3.6.

3.2 Random Field Model

Let R be a spatial domain of interest in $\mathbb{R}^d$, where $d \geq 1$ denotes the dimension of space. The following model for a random field $\{y(\boldsymbol{s}) : \boldsymbol{s} \in R\}$ is considered:

$$y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \varepsilon_1(\boldsymbol{s}) + \varepsilon_2(\boldsymbol{s}), \tag{3.1}$$

where $\mu(\boldsymbol{s})$ is an unknown mean function of location $\boldsymbol{s}$. Furthermore, the error $\varepsilon_1(\cdot)$ is assumed to be a stationary Gaussian process with mean zero and a covariance function $\gamma(\boldsymbol{s}-\boldsymbol{s}')$, where $\boldsymbol{s}, \boldsymbol{s}' \in R$. The second error term $\varepsilon_2(\cdot)$ is assumed to be i.i.d. $N(0, \sigma^2)$ and independent of $\varepsilon_1(\cdot)$. That is, the random field $y(\cdot)$ is decomposed into three additive components: a large-scale trend $\mu(\cdot)$, a small-scale spatial variation $\varepsilon_1(\cdot)$, and a measurement error $\varepsilon_2(\cdot)$; see Cressie (1993) for more details.

3.2.1 Local Karhunen-Loève Expansion

Assume that the spatial domain R is compact and the error process $\varepsilon_1(\cdot)$ is square integrable over R. The Karhunen-Loève expansion of $\varepsilon_1(\boldsymbol{s})$ can be expressed as

$$\varepsilon_1(\boldsymbol{s}) = \sum_{j=1}^{\infty} \bar{\xi}_j \bar{\varphi}_j(\boldsymbol{s}), \quad \boldsymbol{s} \in R,$$

where $\{\bar{\xi}_j : j = 1, 2, \ldots\}$ is a sequence of independent random variables and $\bar{\xi}_j \sim N(0, \bar{\lambda}_j)$, with variances $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq 0$. Furthermore, $\{\bar{\varphi}_j(\cdot) : j = 1, 2, \ldots\}$ is a sequence of orthonormal eigenfunctions over R such that $\int_R \bar{\varphi}_j(\boldsymbol{s})\bar{\varphi}_{j'}(\boldsymbol{s})d\boldsymbol{s} = 1$ when $j = j'$ and zero otherwise. For a recent review of the Karhunen-Loève expansion, see Adler and Taylor (2007) and the references therein.

For a random field, the application of the Karhunen-Loève expansion is limited. There is usually only one realization of the random field and consequently, the variances λ_k cannot be estimated consistently. To circumvent this issue, we introduce a notion of *local Karhunen-Loève expansion.*

First, we assume that the spatial domain R can be partitioned into K compact subdomains with identical shape, namely, $R_1, \ldots, R_K$. Denote $R_k = R_1 + \boldsymbol{v}_k$, for a d-dimensional vector $\boldsymbol{v}_k$. Restricting the error $\varepsilon_1(\cdot)$ to each of the K subdomains gives rise to K error processes that are identically distributed, but not independent of each other, due to the stationarity of the error process $\varepsilon_1(\cdot)$.

Next, we apply the Karhunen-Loève expansion to the error process within each subdomain. More specifically, we have

$$\varepsilon_1(\boldsymbol{s}) = \sum_{j=1}^{\infty} \xi_{j,k}\varphi_{j,k}(\boldsymbol{s}), \quad \boldsymbol{s} \in R_k. \tag{3.2}$$

Here, for a fixed k, $\{\xi_{j,k}\}_{j=1}^{\infty}$ is a sequence of independent random variables such that $\xi_{j,k} \sim N(0, \lambda_j)$, with variances $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. For a given j, $\{\xi_{j,k}\}_{k=1}^{K}$ are identically, although not independently, distributed across all subdomains. Moreover, $\{\varphi_{j,k}(\cdot)\}_{j=1}^{\infty}$ is a sequence of orthonormal eigenfunctions on subdomain R_k. More importantly, as a direct consequence of stationarity of $\varepsilon_1(\cdot)$ over the domain R, for any given j and any $\boldsymbol{s} \in R_k$, $\varphi_{j,k}(\boldsymbol{s}) = \varphi_{j,1}(\boldsymbol{s}-\boldsymbol{v}_k)$; that is, the orthonormal eigenfunctions $\varphi_{j,k}(\boldsymbol{s})$ are the same across all K subdomains up to a constant shift.

In (3.2), the equivalence is defined in the L^2 sense. In practice, however, we approximate the error process $\varepsilon_1(\cdot)$ expressed in an infinite series by a finite sum. That is, we let $\varepsilon_1(\boldsymbol{s}) \approx \sum_{j=1}^{J} \xi_{j,k}\varphi_{j,k}(\boldsymbol{s}) = \sum_{j=1}^{J} \xi_{j,k}\varphi_{j,1}(\boldsymbol{s} - \boldsymbol{v}_k)$, for $\boldsymbol{s} \in R_k$.

3.2.2 Approximation of the Eigenfunctions

An essential component of implementing the local Karhunen-Loève expansion is the computation of the eigenfunctions $\varphi_{j,k}(\cdot)$, which ideally can be obtained by solving integral equations. That is, λ_k and $\varphi_{j,k}(\cdot)$ can be found by solving $\int_{R_k} \gamma(\boldsymbol{s} - \boldsymbol{s}')\varphi_{j,k}(\boldsymbol{s}')d\boldsymbol{s}' = \lambda_j\varphi_{j,k}(\boldsymbol{s})$, for $j = 1, 2, \ldots$ and $\boldsymbol{s}, \boldsymbol{s}' \in R_k$. In general, however, such eigenfunctions cannot be expressed explicitly except for certain special cases. Here, we propose to approximate these eigenfunctions by a set of known orthonormal basis functions.

Let $\boldsymbol{\phi}_1(\boldsymbol{s}) = (\phi_{1,1}(\boldsymbol{s}), \ldots, \phi_{M,1}(\boldsymbol{s}))^T$ be an M-dimensional vector of orthonormal basis functions on R_1. We propose to approximate the eigenfunctions $\boldsymbol{\varphi}_1(\boldsymbol{s}) = (\varphi_{1,1}(\boldsymbol{s}), \ldots, \varphi_{J,1}(\boldsymbol{s}))^T$ from the family $\{\boldsymbol{B}^T\boldsymbol{\phi}_1(\boldsymbol{s}) : \boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}_J\}$, where $\boldsymbol{B}$ is an $M \times J$ coefficient matrix and $\boldsymbol{I}_J$ is a $J \times J$ identity matrix. Suppose that an element from this approximating family, say $\boldsymbol{B}^{*T}\boldsymbol{\phi}_1(\boldsymbol{s})$, provides an adequate approximation of $\boldsymbol{\varphi}_1(\boldsymbol{s})$. Then, on the subdomain R_k, the eigenfunctions $\boldsymbol{\varphi}_k(\boldsymbol{s}) = (\varphi_{1,k}(\boldsymbol{s}), \ldots, \varphi_{J,k}(\boldsymbol{s}))^T$ can be well approximated by $\boldsymbol{B}^{*T}\boldsymbol{\phi}_k(\boldsymbol{s})$, where $\boldsymbol{\phi}_k(\boldsymbol{s}) = \boldsymbol{\phi}_1(\boldsymbol{s} - \boldsymbol{v}_k)$.

Combining the truncated local Karhunen-Loève expansion and the eigenfunction approximation, we have

$$y(\boldsymbol{s}) \approx \mu(\boldsymbol{s}) + \boldsymbol{\phi}_k(\boldsymbol{s})^T\boldsymbol{B}\boldsymbol{\xi}_k + \varepsilon_2(\boldsymbol{s}), \quad \boldsymbol{s} \in R_k, \tag{3.3}$$

where $\boldsymbol{\xi}_k = (\xi_{1,k}, \ldots, \xi_{J,k})^T$ is a J-dimensional vector of random variables such that $\boldsymbol{\xi}_k \sim N(\boldsymbol{0}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \mathrm{Var}(\boldsymbol{\xi}_k) = \mathrm{diag}\{\lambda_1, \ldots, \lambda_J\}$. That is, the error process $\varepsilon_1(\cdot)$ is approximated by a sum of independently distributed Gaussian random variables, which has substantial computational advantages, as we will demonstrate later.

For the choice of basis functions, we consider the orthonormalized cubic B-spline basis for $d = 1$ and orthonormalized radial basis function for $d \geq 2$ (Buhmann, 2003). In particular, the radial basis function is defined as $g(c\|\boldsymbol{s} - \boldsymbol{\kappa}_m\|)$, where g is a prespecified continuous function, $\boldsymbol{\kappa}_m$ is a knot point, and $c > 0$ is a constant. Commonly used choices for g include $g(h) = h^2\log(h)$, which leads to thin-plate splines, and $g(h) = e^{-h^2}$, which results in Gaussian radial splines. In practice, the vector of basis functions can be orthonormalized.

3.3 Statistical Inference

3.3.1 Constrained Likelihood-based Estimation

Henceforth, we will restrict our attention to the case of model (3.1) with a linear trend. That is, $\mu(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}$, where $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))^T$ is a p-dimensional vector of covariates at location $\boldsymbol{s}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional vector of regression coefficients. Under this setting, model (3.3) resembles a linear mixed model, but is subject to constraints due to the orthonormality of the basis functions. Consequently, standard statistical methods for estimating the parameters of a linear mixed model are not directly applicable. Peng and Paul (2009) considered a similar problem for functional data, and implemented a manifold version of the Newton-Raphson method to optimize a likelihood-based criterion with such constraints. In addition, Paul and Peng (2009) established the consistency of the resulting estimates under the assumption that there are independent replicates per subject, which generally does not hold for geostatistical data. Here, we develop a new estimation procedure as follows.

Suppose there are N sampling locations in the spatial domain R. Let $\{\boldsymbol{s}_{k,i} : i = 1, \ldots, n_k\}$ denote the sampling locations in subdomain R_k and thus, $\sum_{k=1}^{K} n_k = N$. Let $\boldsymbol{X}_k = (\boldsymbol{x}(\boldsymbol{s}_{k,1}), \ldots, \boldsymbol{x}(\boldsymbol{s}_{k,n_k}))^T$ denote an $n_k \times p$ design matrix of the covariates and $\boldsymbol{\Phi}_k = (\boldsymbol{\phi}_k(\boldsymbol{s}_{k,1}), \ldots, \boldsymbol{\phi}_k(\boldsymbol{s}_{k,n_k}))^T$ denote an $n_k \times M$ matrix of $\boldsymbol{\phi}_k(\cdot)$ evaluated at the sampling locations in the subdomain R_k. Moreover, let $\boldsymbol{y}_k = (y(\boldsymbol{s}_{k,1}), \ldots, y(\boldsymbol{s}_{k,n_k}))^T$ denote an n_k-dimensional vector of responses in R_k and $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T)^T$ denote an N-dimensional vector of responses in R, such that $\boldsymbol{\Sigma}_{0k} = \mathrm{Var}(\boldsymbol{y}_k)$ is the true covariance matrix of $\boldsymbol{y}_k$ and $\boldsymbol{\Sigma}_0 = \mathrm{Var}(\boldsymbol{y})$ is the true covariance matrix of $\boldsymbol{y}$.

Based on model (3.3), the corresponding approximating covariance matrix of $\boldsymbol{\Sigma}_{0k}$ is $\boldsymbol{\Sigma}_k = \boldsymbol{\Phi}_k^T \boldsymbol{B} \boldsymbol{\Lambda} \boldsymbol{B}^T \boldsymbol{\Phi}_k + \sigma^2 \boldsymbol{I}_{n_k}$. By ignoring the dependence among $\boldsymbol{y}_k$ in different subdomains, $\boldsymbol{\Sigma}_0$ can be approximated by a block-diagonal matrix $\boldsymbol{\Sigma}_{\mathrm{KL}} = \mathrm{diag}\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$. Consequently, up to an additive constant, the negative log-likelihood function can be approximated as,

$$L_K(\boldsymbol{\beta}, \sigma^2, \boldsymbol{B}, \boldsymbol{\Lambda}) = (2K)^{-1} \sum_{k=1}^{K} \left\{ (\boldsymbol{y}_k - \boldsymbol{X}_k \boldsymbol{\beta})^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_k - \boldsymbol{X}_k \boldsymbol{\beta}) + \log |\boldsymbol{\Sigma}_k| \right\}. \quad (3.4)$$

Note that, (3.4) provides a better approximation to the true negative log-likelihood function as the correlation of the observations between subdomains becomes weaker;

see Section 3.5 for further discussion. Let $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Lambda}})$ denote the estimates obtained from minimizing (3.4).

To carry out the minimization of (3.4), the following iterative algorithm is conducted. First, for a given $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Lambda})$, we minimize (3.4) with respect to $\boldsymbol{B}$ subject to the constraint $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}_J$. Here, we implement a Newton-Raphson type algorithm on a Stiefel manifold, which utilizes the intrinsic Riemannian geometric structure of such manifold (Peng and Paul, 2009). Next, given $\boldsymbol{B}$, we minimize (3.4) with respect to $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Lambda}$. In the second step, for a fixed $\boldsymbol{B}$, (3.4) is the log-likelihood function of a linear mixed model and thus, its minimization is straightforward (see, e.g., Pinheiro and Bates, 2000).

3.3.2 Pre-tapered and Tapered Estimates

The block-diagonal matrix $\boldsymbol{\Sigma}_{\mathrm{KL}}$ approximates the true covariance matrix by ignoring the dependence of observations between subdomains. Such an approximation has great computational advantages; however, our numerical studies suggest that the amount of error can be large. Here, for a stationary isotropic random field, we propose a novel approach to recover some of the accuracy loss caused by the approximation of the true covariance matrix.

Recall that, for subdomain R_k, the estimated covariance matrix is $\widehat{\boldsymbol{\Sigma}}_k = \boldsymbol{\Phi}_k^T\widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{B}}^T\boldsymbol{\Phi}_k + \widehat{\sigma}^2\boldsymbol{I}_{n_k}$. First, we obtain an estimated mean covariance function $\widehat{\gamma}_1(h)$ from $\widehat{\boldsymbol{\Sigma}}_k$ for spatial lag $h \in [0, D]$, where D is the maximum spatial lag within the subdomain. We take an average of covariance function estimates over the lag distances that are within a small neighborhood Δh of a given spatial lag h. For example, for $h = 0.1$ and $\Delta h = 0.05$, $\widehat{\gamma}_1(0.1)$ is the average of covariances between sampling locations that are 0.1 ± 0.05 distance apart. By taking a local mean, we observe, in our numerical examples, that the inversion of the estimated covariance matrix is more stable.

Since $\widehat{\gamma}_1(h)$ is not guaranteed to be positive definite, we implement a further transformation proposed by Hall and Patil (1994). In particular, let $\psi(\theta) = \int \exp(i\theta h)\widehat{\gamma}_1(h)dh$, where $\theta \in \mathbb{R}$, and transform $\widehat{\gamma}_1(h)$ to

$$\widehat{\gamma}_2(h) = (2\pi)^{-1} \int_{\mathbb{R}} \cos(\theta h)\widehat{\psi}(\theta)d\theta, \tag{3.5}$$

where $\widehat{\psi} = \max\{\psi, 0\}$. The resulting $\widehat{\gamma}_2(h)$ will be referred to as a *pre-tapered estimate* of the covariance function. Although it is positive definite on $[0, D]$, it may not be continuous nor positive definite on $[0, \infty]$. Thus, we further adopt a tapering function $W(h, \omega)$, which is an isotropic autocorrelation function when $h \leq \omega$ and 0 when $h > \omega$ for a given threshold distance ω. Compactly supported correlation functions are often used as the tapering functions, such as $W(h, \omega) = (1 - h/\omega)I\{h \leq \omega\}$, where $I\{h \leq \omega\}$ is an indicator function (Wendland, 1995). A tapered estimate for the covariance function $\widehat{\gamma}_3(h)$ can be obtained by $\widehat{\gamma}_3(h) = \widehat{\gamma}_2(h)W(h, \omega)$, which is a positive definite covariance function over $[0, \infty]$. The corresponding estimated covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathrm{T}} = [\widehat{\gamma}_3(d_{ii'})]_{i,i'=1}^{N}$ is also positive definite, where $d_{ii'}$ is the distance between two sampling locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$. Using $\widehat{\mathbf{\Sigma}}_{\mathrm{T}}$, we then update the estimates of $\boldsymbol{\beta}$ and σ^2, denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{T}}$ and $\widehat{\sigma}^2_{\mathrm{T}}$. Since $\widehat{\mathbf{\Sigma}}_{\mathrm{T}}$ is a sparse matrix, the computation is fast even for large sample sizes.

3.3.3 Applications of Tapered Estimates

The tapered estimates $(\widehat{\boldsymbol{\beta}}_{\mathrm{T}}, \widehat{\sigma}^2_{\mathrm{T}}, \widehat{\mathbf{\Sigma}}_{\mathrm{T}})$ developed in Section 3.3.2 can be applied to spatial prediction and variable selection. To predict the random field $y(\boldsymbol{s}_0)$ at an unsampled location $\boldsymbol{s}_0$, we apply the standard approach of best linear unbiased prediction (BLUP; Section 3.4.5 in Cressie (1993)). Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ denote N sampling locations in R, $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_N))^T$ denote an N-dimensional vector of response variables, and $\boldsymbol{X}$ denote an $N \times p$ design matrix of covariates. The BLUP of the response at an unsampled location $\boldsymbol{s}_0$ is $\widetilde{y}(\boldsymbol{s}_0) = \boldsymbol{x}^T(\boldsymbol{s}_0)\widetilde{\boldsymbol{\beta}} + \boldsymbol{c}_0^T\mathbf{\Sigma}_0^{-1}(\boldsymbol{y} - \boldsymbol{X}^T\widetilde{\boldsymbol{\beta}})$, where $\boldsymbol{c}_0$ is an N-dimensional vector comprising $\mathrm{cov}\{y(\boldsymbol{s}_0), y(\boldsymbol{s}_i)\}$, $\mathbf{\Sigma}_0 = [\mathrm{cov}\{y(\boldsymbol{s}_i), y(\boldsymbol{s}_{i'})\}]_{i,i'=1}^{N}$, and $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}\mathbf{\Sigma}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\mathbf{\Sigma}_0^{-1}\boldsymbol{y}$. Both $\mathbf{\Sigma}_0$ and $\boldsymbol{c}_0$ rely on the unknown covariance function $\gamma(\cdot)$. By directly plugging the tapered estimate $\widehat{\gamma}_3(\cdot)$ into $\mathbf{\Sigma}_0$, we obtain an empirical BLUP at $\boldsymbol{s}_0$.

Next, we consider variable selection in the context of spatial linear model with the goal of determining the best subset of the covariates. Extending a penalized least squares method by Wang and Zhu (2009), Chu et al. (2011) proposed a penalized maximum likelihood approach for simultaneous parameter estimation and variable selection, which we will adopt here. In particular, a penalized log-likelihood function is defined as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}_0; \boldsymbol{y}, \boldsymbol{X}) - N \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{3.6}$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}_0; \boldsymbol{y}, \boldsymbol{X}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Sigma}_0| - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ is the log-likelihood function and $p_\lambda(\cdot)$ is a known penalty function with a regularization parameter λ. Popular choices of the penalty function include the L_1 penalty and smoothly clipped absolute deviation (SCAD; Fan, 1997; Fan and Li, 2001). Here, we will focus on the SCAD penalty function.

When $\boldsymbol{\Sigma}_0$ is replaced with $\widehat{\boldsymbol{\Sigma}}_{\mathrm{T}}$, (3.6) is equivalent to a penalized least squares problem. The corresponding optimization problem has been widely studied; see Fan and Li (2001) and Zou and Li (2008) for more details. We implement a one-step estimation procedure proposed by Zou and Li (2008) and provide an approximate solution by a Newton-Raphson type iteration starting from $\widehat{\boldsymbol{\beta}}_{\mathrm{T}}$.

The finite-sample properties of the tapered estimates in both of these applications will be investigated in a simulation study in Section 3.4.1.

3.4 Numerical Examples

3.4.1 Simulation Study

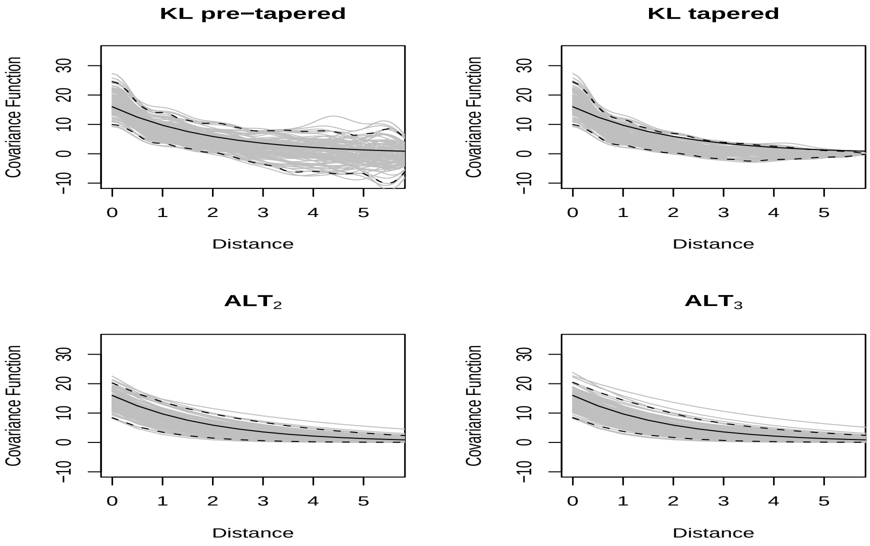
We now investigate the finite-sample properties of our proposed method using local Karhunen-Loève expansion denoted as KL. Four different scienarios are considered, which are combinations of two dimensions ($d = 1$ or 2) and two true covariance functions (exponential or not). For comparison, we consider three competing methods. The first alternative, ALT$_1$, is the ordinary least squares that ignores spatial dependence. The second alternative, ALT_2, assumes a parametric covariance function and applies maximum likelihood for parameter estimation (Mardia and Marshall, 1984). Third and last, we consider a "small block" method, ALT_3, proposed by Caragea and Smith (2007). In ALT_2 and ALT_3, we assume that the error term follows an exponential covariance function regardless of the true underlying covariance structure.

$d = 1$**, Exponential Covariance.**

Let the spatial domain be $R = [0, L]$ with $L = 30$, 60, 90. For a fixed sampling density 10, the corresponding sample size N is 300, 600 and 900, respectively. The linear regression model has seven covariates with regression coefficients $\boldsymbol{\beta} = (4, 3, 2, 1, 0, 0, 0)^T$. The covariates are generated from standard normal distributions with a cross-covariate correlation of 0.5. In addition, we standardize each covariate to have sample mean 0 and sample variance 1, and the response to have a sample mean 0. Consequently, there is no intercept in this model. For spatial dependence, we generate the error $\varepsilon_1(s)$ at the sampling location s from a zero-mean stationary and isotropic Gaussian process with an exponential covariance function $\gamma(h) = \sigma_1^2 \exp(-h/c_r)$, where σ_1^2 is a variance component and c_r is a range parameter. In addition, the measurement errors $\varepsilon_2(s)$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 2$. For KL and ALT_2, the subdomains are set to be intervals of equal length 6.

For each sample size N, we simulate 100 data sets, and for each data set, we estimate $\boldsymbol{\beta}$ using our KL method as well as the three alternatives. The mean and standard deviation of the resulting estimates are reported in Table A in Appendix D of this chapter. The results show that ALT_2 performs the best, as expected. As the sample size increases, the performances of all four estimates improve in terms of smaller biases and variances. Moreover, by accounting for spatial dependence, parameter estimation using ALT_2, ALT_3, and KL all outperforms ALT_1. The estimates from ALT_3 and KL tend to those of ALT_2, which suggests that the effect of the covariance matrix approximation becomes smaller as the sample size increases.

Figure 3.1: Estimated Covariance Functions in Scenario 1.

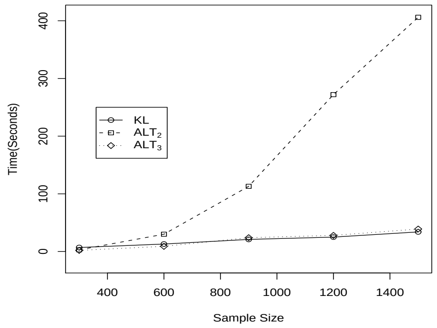


Estimated covariance functions and 95% pointwise confidence intervals using our proposed method with both pre-tapered estimates (upper-left) and tapered estimates (upper-right), maximum likelihood ALT_2 (lower-left) and a small-block method ALT_3 (lower-right). Black solid line: true exponential covariance function; grey line: estimated covariance function from each simulated data; dashed lines: pointwise confidence intervals.

The estimated covariance functions by KL are quite close to ALT_2, as illustrated in Figure 3.1. In particular, using the pre-tapered estimate $\widehat{\gamma}_2(h)$, the true covariance function falls well within the 95% pointwise confidence intervals. However, when the spatial lag increases, the pointwise confident intervals do not narrow as in ALT_2 and ALT_3, due to a nonparametric form of the error process. For the tapered estimate of covariance function $\widehat{\gamma}_3(h)$, the true covariance functions fall within 95% pointwise confidence intervals except when the distance gets close to 6, due to tapering beyond distance 6.

In Figure 3.2, the computing time for all three methods, KL, ALT_2, and ALT_3, is reported. It can be seen that, for relatively small sample sizes, the three methods take up about the same amount of time. As the sample size increases, however,

Figure 3.2: Computing Times (in seconds).



Computing times (in seconds) for KL, ALT$_2$, and ALT$_3$ versus various choices of sample size.

the computing time for ALT$_2$ increases dramatically compared with both KL and ALT$_3$ whose computing time is similar. This large difference in computing time is expected, as ALT$_2$ involves large matrix inversion, but underscores the usefulness of our KL method. Similar observations are made in the other three scenarios.

To evaluate the performance of spatial prediction, we define a mean squared prediction error (MSPE) as $n^{-1}\sum_{i=1}^{n}\{\widetilde{y}(\boldsymbol{s}_{0i}) - y(\boldsymbol{s}_{0i})\}^2$, where $\boldsymbol{s}_{01}, \ldots, \boldsymbol{s}_{0n}$ are n unsampled locations in R, $y(\boldsymbol{s}_{0i})$ is the true value and $\widetilde{y}(\boldsymbol{s}_{0i})$ is the predicted value at location $\boldsymbol{s}_{0i}$, for $i = 1, \ldots, n$. In the simulation, for each sample size N, an additional 10% observations are generated at new locations to form a test set. The mean and standard deviation of the MSPE values are reported in Table 3.3. Our KL method performs similarly to ALT$_2$ and ALT$_3$, but ALT$_1$ gives rather poor prediction. As for variable selection, KL, ALT$_2$, and ALT$_3$ perform similarly and satisfactorily, and all are slightly better than ALT$_1$, as shown in Table A in Appendix D of this chapter.

$d = 1$**, Misspecified Covariance.**

Here the setup is the same as Scenario 1 except for the spatial dependence structure. Specifically, the error process $\varepsilon_1(\boldsymbol{s})$ follows a sinusoidal covariance function: $\gamma(h) = \sigma_1^2 \sin(h/c_r)c_r/h$. Moreover, the measurement error terms $\varepsilon_2(\boldsymbol{s})$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 0.4$. The results are reported in Table B in Appendix D of this chapter.

Table 3.1: Kriging Simulation Results from Scenario 1 and 2

		Exponential Covariance		Misspecified Covariance	
N	Method	mean(MSPE)	SD(MSPE)	mean(MSPE)	SD(MSPE)
300	KL	5.83	1.52	4.88	1.30
	ALT$_1$	17.84	6.32	18.97	6.43
	ALT$_2$	5.69	1.49	4.96	1.29
	ALT$_3$	5.72	1.49	4.99	1.27
600	KL	5.70	1.20	4.81	0.98
	ALT$_1$	18.92	4.36	19.81	3.97
	ALT$_2$	5.62	1.17	4.93	0.99
	ALT$_3$	5.68	1.19	4.99	1.00
900	KL	5.61	0.84	4.76	0.72
	ALT$_1$	19.21	4.12	19.92	3.46
	ALT$_2$	5.51	0.81	4.90	0.74
	ALT$_3$	5.56	0.82	4.95	0.75

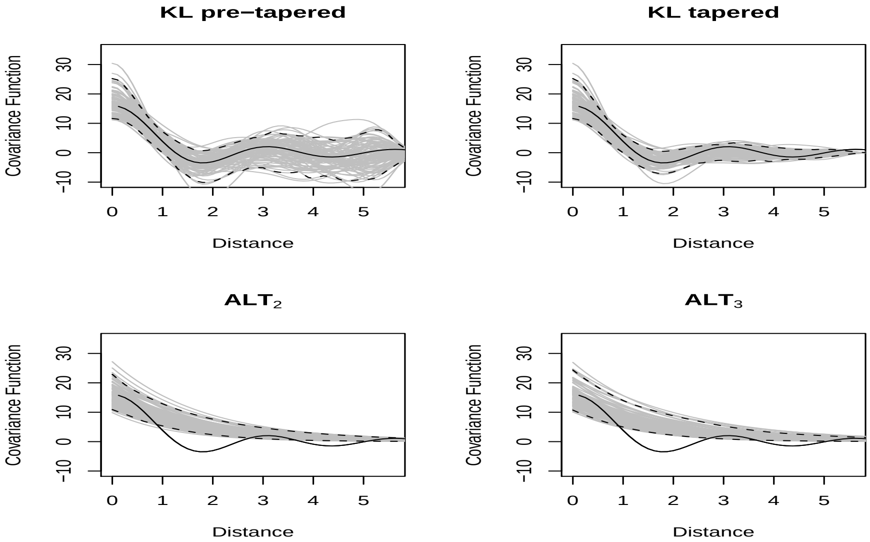
Simulation results from Scenario 1 (left panel) and 2 (right panel): Mean squared prediction error (MSPE) and standard deviation (SD) under KL, ALT$_1$, ALT$_2$, and ALT$_3$ for sample size $N = 300$, 600, 900.

Again, as the sample size increases, the estimation of all four methods improves. In addition, ALT$_2$, ALT$_3$, and KL perform better than ALT$_1$ for both parameter estimation and prediction. This suggests that it is important to consider spatial dependence, even if spatial covariance function is misspecified. However, our KL method outperforms ALT$_2$ and ALT$_3$ by providing a more robust estimate of the covariance function, as illustrated in Figure 3.3. Unlike Scenario 1, the estimated covariance functions from ALT$_2$ and ALT$_3$ are not close to the true underlying function, due to the model misspecification when applying maximum likelihood. In contrast, the performance of KL is satisfactory. Similar results regarding spatial prediction and variable selection are attained as Scenario 1.

When $d = 2$, similar conclusions can be drawn regarding parameter estimation, spatial prediction, and variable selection. To save space, those results are included in Appendix D of this chapter.

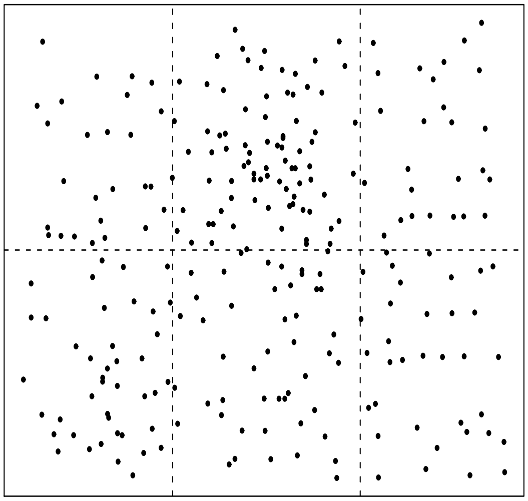
3.4.2 Data Example

Figure 3.3: Estimated Covariance Functions in Scenario 2.



Estimated covariance functions and 95% pointwise confidence intervals using our proposed method with both pre-tapered estimates (upper-left) and tapered estimates (upper-right), maximum likelihood ALT$_2$ (lower-left) and a small-block method ALT$_3$ (lower-right). Black solid line: true sinusoidal covariance function; grey line: estimated covariance function from each simulated data; dashed lines: pointwise confidence intervals.

Figure 3.4: The Domain and Subdomain of Locations of 259 Sampling Sites.



Map of locations of 259 sampling sites in the Colorado precipitation data and the subdomain (divided by dotted line) used for KL method and small block method.

Table 3.2: Precipitation Data Results under KL, ALT_1, ALT_2, and ALT_3.

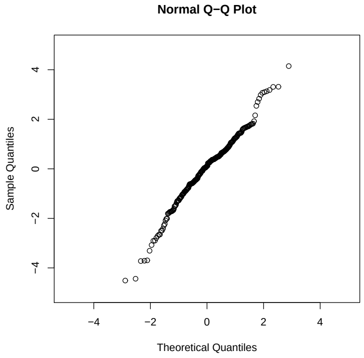
Terms	KL	SE	ALT_1	SE	ALT_2	SE	ALT_3	SE
Elevation	0.281	0.058	0.221	0.047	0.305	0.055	0.235	0.052
Slope	0.020	0.031	0.074	0.041	0.158	0.026	0.027	0.029
Aspect	0.000	0.027	0.051	0.034	-0.004	0.022	0.005	0.025
B1M	0.196	0.184	0.142	0.214	0.214	0.157	0.254	0.170
B2M	0.036	0.074	0.069	0.093	0.058	0.064	0.017	0.068
B3M	0.037	0.131	0.059	0.160	0.017	0.109	-0.015	0.112
B4M	-0.400	0.214	-0.472	0.242	-0.043	0.183	-0.381	0.199
B5M	0.090	0.105	0.155	0.137	0.043	0.089	0.115	0.098
B6M	-0.190	0.135	-0.357	0.166	-0.162	0.116	-0.212	0.124
B7M	0.158	0.116	0.241	0.150	0.172	0.098	0.121	0.110

Precipitation data: Regression coefficient estimates and standard errors (SE) under KL, ALT_1, ALT_2, and ALT_3.

The dataset consists of January precipitation (inches per 24-hour period) on the log-scale from 259 weather stations in Colorado (Reich and Davis, 2008; Chu et al., 2011), as shown in Figure 3.4. There are ten covariates of interest, including elevation, slope, aspect, and seven spectral bands from a MODIS satellite imagery (B1M through B7M). To investigate the relationship between precipitation and these covariates, we first fit a spatial linear model with an exponential covariance function via ordinary least squares, maximum likelihood, and the small block method. The parameter estimates and their standard errors in Table 3.2 suggest that the regression coefficients for elevation, B1M, B4M, B6M, and B7M are possibly significant. We then fit the data using our proposed method and the results are similar to the three alternative methods, although the MLE and the small block method appear to have slightly smaller standard errors.

Here assumptions about the spatial error process, namely, normality, stationarity and isotropy are investigated. To check the normality, we obtain normal QQ-plot for $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{e}$, where $\widehat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix from MLE method, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ is the residuals, and $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ is the regression coefficient estimates from

Figure 3.5: Normal QQ-plot for Residuals $\widehat{\mathbf{\Sigma}}^{-1/2}\boldsymbol{e}$.



MLE method. By model assumption, $\widehat{\mathbf{\Sigma}}^{-1/2}\boldsymbol{e}$ should be i.i.d normal, and Figure 3.5 shows there is no obvious violation of the normality assumption.

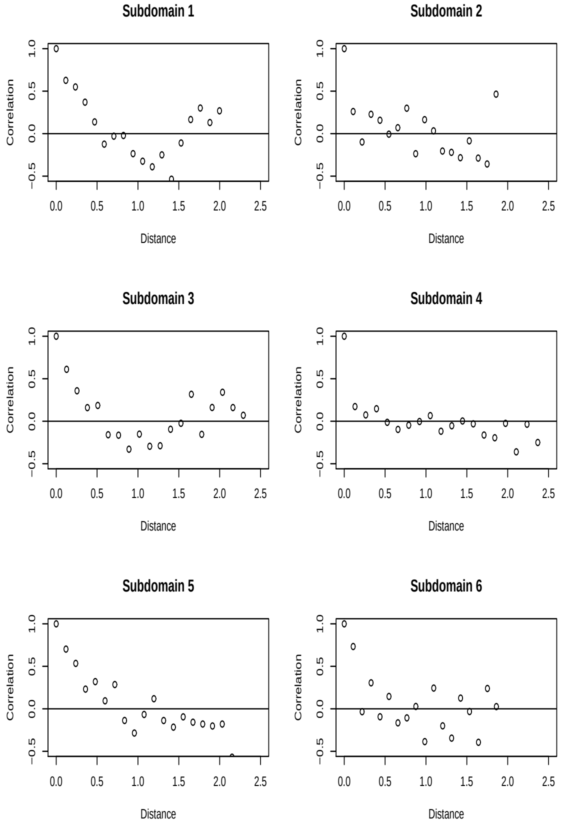
To check the assumption of stationarity, we first divide the whole domain into several small subdomains as illustrated in Figure 3.4. For each subdomain, an empirical correlation function is estimated from residuals $\boldsymbol{e}$ and the results are shown in Figure 3.6. Note that the estimated covariance function from each subdomain has the similar shape, which suggests that the assumption of stationarity holds.

Last we check the isotropy assumption by comparing the estimated correlation functions for different directions. That is, we estimate correlation functions from residuals $\boldsymbol{e}$ with angles at 0, 45, 90, and 135 degrees, as shown in Figure 3.7. Since the shape of estimated correlation functions appears similar, we conclude that there is no obvious violation for the isotropy assumption, either.

3.5 Theoretical Aspect

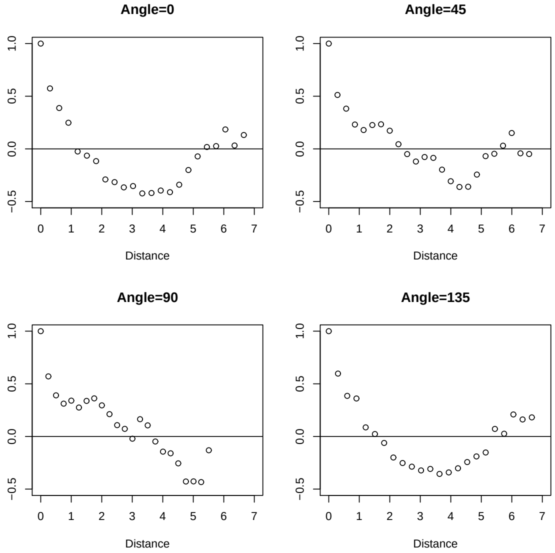
In this section, we will establish the consistency of estimates in Section 3.3.1. Recall that R denotes the compact domain of interest and N is the number of sampling locations in R. Also, λ_j is the jth largest eigenvalue in the local Karhunen-Loève expansion and n_k is the number of sampling locations in subdomain R_k. Let $\boldsymbol{\beta}_0$ denote the true parameters. We assume σ^2 is known with $\sigma^2 = 1$, without loss of generality.

Figure 3.6: Estimated Correlation Functions in Subdomains.



The estimated correlation functions in 6 different subdomains.

Figure 3.7: Estimated Correlation Functions in Different Directions.



The estimated correlation functions with directions at angles at 0, 45, 90, and 135 degrees.

Let $\lambda_i(\boldsymbol{A}_1)$ denote the ith largest eigenvalue of a square matrix $\boldsymbol{A}_1$. Furthermore, for an $n \times m$ matrix $\boldsymbol{A}_2 = (a_{ij})_{i,j=1}^{n,m}$, the Frobenius norm and L_2-norm are defined as $\|\boldsymbol{A}_2\|_F = \left(\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2\right)^{1/2}$ and $\|\boldsymbol{A}_2\|_2 = \max\{\lambda_i(\boldsymbol{A}_2^T\boldsymbol{A}_2)^{1/2} : i = 1, \ldots, m\}$, respectively. Moreover, if $\boldsymbol{A}_3$ is an $n \times l$ matrix, it holds that $\|\boldsymbol{A}_2\boldsymbol{A}_3\|_F \leq \|\boldsymbol{A}_2\|_F\|\boldsymbol{A}_3\|_F$, $\|\boldsymbol{A}_2\boldsymbol{A}_3\|_F \leq \|\boldsymbol{A}_2\|_2\|\boldsymbol{A}_3\|_F$, and $\|\boldsymbol{A}_2\boldsymbol{A}_3\|_2 \leq \|\boldsymbol{A}_2\|_2\|\boldsymbol{A}_3\|_2$.

We assume the following regularity conditions.

(A.1) There exist $0 < c_1$, c_2 $c_3 < \infty$, such that (i) $c_1 \geq \lambda_1 > \cdots > \lambda_J > \lambda_{J+1}$; (ii) $\max_{1\leq j\leq J}(\lambda_j - \lambda_{j+1})^{-1} \leq c_2$; (iii) $\lambda_1(\boldsymbol{\Sigma}_0) \leq c_3$.

(A.2) The eigenfunctions $\{\varphi_{j,1}(\cdot)\}_{j=1}^J$ are four times continuously differentiable and satisfy $\max_{1\leq j\leq J} \|\varphi_{j,1}^{(4)}(\cdot)\|_\infty \leq C_0$ for some $0 < C_0 < \infty$.

(A.3) For n_k, $\underline{n} \leq n_k \leq \overline{n}$, where $\underline{n} \geq 4$, $\overline{n}/\underline{n} = O(1)$, and $\overline{n} = \mathcal{O}(K^\kappa)$ for some $\kappa \geq 0$.

(A.4) There exist $(\boldsymbol{B}^*, \boldsymbol{\Lambda}^*)$, such that $\delta_K = \max_{1\leq k\leq K} n_k^{-1}\|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_F$ and $\overline{n}\delta_K = \mathcal{O}((M\mathrm{log}K/K)^{1/2})$, where $\boldsymbol{\Sigma}_k^* = \boldsymbol{\Phi}_k^T\boldsymbol{B}^*\boldsymbol{\Lambda}^*\boldsymbol{B}^{*T}\boldsymbol{\Phi}_k + \sigma^2\boldsymbol{I}_{n_k}$ and $\boldsymbol{B}^{*T}\boldsymbol{B}^* = \boldsymbol{I}_r$.

(A.5) There exist constants $\rho_1, d_1, d_2, K_1 > 0$, such that for $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta((\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*); \rho_1)$, $d_1 \underline{n}^2 a_K^2 < 1/K \sum_{k=1}^{K} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F^2 < d_2 \overline{n}^2 a_K^2$ for all $K \geq K_1$, where $\Theta((\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*); \rho)$ is defined in (3.11).

(A.6) For $(\varphi_{1,k}^*(\cdot), \ldots, \varphi_{J,k}^*(\cdot)) = \boldsymbol{B}^{*T}(\phi_{1,k}(\cdot), \ldots, \phi_{M,k}(\cdot))$,

$$\max_{1 \leq j \leq J} \|\varphi_{j,k}(\cdot) - \varphi_{j,k}^*(\cdot)\| \leq c_{\phi,3} M^{-4} \max_{1 \leq j \leq J} \|\varphi_{j,k}^{(4)}(\cdot)\|_\infty, \tag{3.7}$$

$$\|\boldsymbol{\Phi}_k\|^2 \leq \overline{n} c_{g,1} + c_{\phi,0}^{-1} d_\eta \{(M^{3/2} \log K) \vee (M(\overline{n} \log K)^{1/2})\}, \tag{3.8}$$

$$\|\boldsymbol{\Phi}_k\|^2 \leq c_{\phi,2} \overline{n} M, \tag{3.9}$$

where $c_{\phi,0}$, $c_{\phi,2}$, $c_{\phi,3}$, and $c_{g,1}$ are constants.

(A.7) For $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta((\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*); \rho_1)$, let $\alpha_K(\cdot)$ be the α-mixing coefficient of random variable $\mathrm{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{S}_k - \boldsymbol{\Sigma}_{0k})\}$. Then $\alpha_K(2) = o(K^{-(2+2\kappa)Mr-\eta-2})$, where $\boldsymbol{S}_k = (\boldsymbol{y}_k - \boldsymbol{X}_k \boldsymbol{\beta}_0)(\boldsymbol{y}_k - \boldsymbol{X}_k \boldsymbol{\beta}_0)^T$ and ρ_1 is defined in (A.4).

(A.8) As $K \to \infty$, $M^{-1}(K/\mathrm{log}K)^{1/9} = \mathcal{O}(1)$, $M = o((K/\mathrm{log}K)^{1/2})$, and $\overline{n}^4 M^2 \mathrm{log}K = o(K)$.

(A.9) There exist constant $C_1, C_2, N_1 > 0$, such that $C_1 \boldsymbol{I}_N \leq \boldsymbol{X}^T \boldsymbol{X}/N \leq C_2 \boldsymbol{I}_N$, for all $N \geq N_1$.

Assumptions (A.1)–(A.2) are about the spatial covariance structure and the Karhunen-Loève expansion. (A.3) is a boundedness condition for the number of sampling locations in subdomains (Paul and Peng, 2009). (A.4) assumes that there exist optimal parameters $(\boldsymbol{B}^*, \boldsymbol{\Lambda}^*)$ such that the difference of the true covariance matrix $\boldsymbol{\Sigma}_{0k}$ and optimal covariance matrix $\boldsymbol{\Sigma}_k^*$ in every subdomain tends to 0 uniformly, as the number of subdomains $K \to \infty$. Moreover, (A.4) requires that λ_{J+1} decay sufficiently fast (e.g., the expansion (3.2) of the process $\varepsilon_1(\boldsymbol{s})$ has finite J terms). In this case, the existence of the optimal parameters can be shown using the spline approximation theory, which is well-established for the one-dimensional space. (A.5) is about the properties for spline basis and (A.6) is about the properties of the fixed sampling locations. (A.7) assumes that correlations between subdomains decrease in the sense of increasing domain. (A.8) specifies the relationship between $\overline{n}$, M and K (Paul and Peng, 2009). The assumption about the design matrix $\boldsymbol{X}$ is made in (A.9).

The following Theorem 3.5.1 establishes the consistency for the estimates of the regression coefficients $\boldsymbol{\beta}$ and the spatial covariance function parameters $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$, using our proposed method.

Theorem 3.5.1. *Suppose that (A.1)–(A.9) hold, then there is a minimizer* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Lambda}})$ *of equation (3.4), such that, for* $a_K = (\overline{n}^2 M log K/K)^{1/2}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = \mathcal{O}_p(N^{-1/2}), \quad \|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F = \mathcal{O}_p(a_K), \quad \|\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}^*\|_F = \mathcal{O}_p(a_K).$$

Theorem 3.5.1 shows that for regression coefficient vector, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ converging to the true parameter $\boldsymbol{\beta}_0$ at the rate of $N^{1/2}$. For the spatial covariance function parameters $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$, the convergence is also achievable but at a slower rate of a_K. However, the convergence rate is not as slow as it might appear, especially for $\boldsymbol{B}$, since both $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$ are converging in the Frobenius norm. That is, for $\boldsymbol{B}$, the convergence is for the square sum of MJ parameters. It is also worth mentioning that the resulting estimates $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Lambda}})$ are the primary building blocks to obtain the more accurate pre-tapered and tapered estimates for covariance function in Section 3.3.2.

3.6 Appendix

3.6.1 Appendix A: Neighborhood in the Parameter Space

To maximize the approximated log-likelihood function in (3.4), a major challenge is the orthogonal constraint of the matrix parameter $\boldsymbol{B}$; that is, $\boldsymbol{B}$ is taken from the set $\mathcal{S}_{M,J} = \{\boldsymbol{A} \in \mathbb{R}^{M \times J} : \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_J\}$, which is the well-known *Stiefel manifold*. Each matrix $\boldsymbol{B}$ can be considered as a point on the manifold $\mathcal{S}_{M,J}$.

Let $\mathcal{T}_{\boldsymbol{B}}$ be the tangent space of $\mathcal{S}_{M,J}$ at the point $\boldsymbol{B}$. In particular, any element in the tangent space, $\boldsymbol{U} \in \mathcal{T}_{\boldsymbol{B}}$, can be expressed as $\boldsymbol{U} = \boldsymbol{B}\boldsymbol{A}_U + \boldsymbol{C}_U$, where $\boldsymbol{A}_U = -\boldsymbol{A}_U^T$ and $\boldsymbol{B}^T\boldsymbol{C}_U = \boldsymbol{0}$; see Edelman et al. (1998) for more details. On the manifold $\mathcal{S}_{M,J}$, the geodesic emanating from $\boldsymbol{B}$ along the direction $\boldsymbol{U}$ can be written as, for any $t \geq 0$, $\boldsymbol{G}_{\boldsymbol{B},\boldsymbol{U}}(t) = \boldsymbol{B}\boldsymbol{M}_{\boldsymbol{B},\boldsymbol{U}}(t) + \boldsymbol{Q}\boldsymbol{N}_{\boldsymbol{B},\boldsymbol{U}}(t)$, where

$$\left[\begin{array}{c} \boldsymbol{M}_{\boldsymbol{B},\boldsymbol{U}}(t) \\ \boldsymbol{N}_{\boldsymbol{B},\boldsymbol{U}}(t) \end{array}\right] = \exp\left\{t\left[\begin{array}{cc} \boldsymbol{B}^T\boldsymbol{U} & -\boldsymbol{R}^T \\ \boldsymbol{R} & \boldsymbol{0} \end{array}\right]\right\}\left[\begin{array}{c} \boldsymbol{I}_r \\ \boldsymbol{0} \end{array}\right]. \tag{3.10}$$

Here $\exp(\cdot)$ is the usual matrix exponential functional, and $\boldsymbol{Q}\boldsymbol{R}$ is the QR-decomposition of $(\boldsymbol{I}_M - \boldsymbol{B}\boldsymbol{B}^T)\boldsymbol{U}$. Note that, the function $\boldsymbol{G}_{\boldsymbol{B},\boldsymbol{U}}(t)$ is the exponential map on the manifold $\mathcal{S}_{M,J}$ at $\boldsymbol{B}$ along the direction $\boldsymbol{U}$, which essentially maps a tangent vector to a point on the manifold.

The geodesic, along with the exponential mapping, provides a useful way to define a neighborhood on the manifold. For instance, we can define the neighborhood as $\{\boldsymbol{G}_{\boldsymbol{B},\boldsymbol{U}}(t) : \text{for some small enough } t \text{ and } \boldsymbol{U}\}$. Note that the magnitudes of t and $\boldsymbol{U}$ will determine the size of the neighborhood around $\boldsymbol{B}$. For convenience, we let $t = 1$, since $\boldsymbol{G}_{\boldsymbol{B},\boldsymbol{U}}(t) = \boldsymbol{G}_{\boldsymbol{B},t\boldsymbol{U}}(1)$.

Finally, we define a *neighborhood* in parameter space, centered at $(\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*)$ and with size ρ, by

$$\begin{aligned}
& \Theta((\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*); \rho) \\
= \ & \{(\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Lambda}) : \boldsymbol{\beta} = \exp(\boldsymbol{E})\boldsymbol{\beta}_0, \boldsymbol{B} = \boldsymbol{G}_{\boldsymbol{B}^*,\boldsymbol{U}}(1), \boldsymbol{\Lambda} = \exp(\boldsymbol{D})\boldsymbol{\Lambda}^*, \\
& \text{ where } \boldsymbol{U} = \boldsymbol{B}^*\boldsymbol{A}_U + \boldsymbol{C}_U, \boldsymbol{A}_U = -\boldsymbol{A}_U^T \text{ and } \boldsymbol{B}^{*T}\boldsymbol{C}_U = \boldsymbol{0}, \\
& \boldsymbol{D} \text{ is a } r \times r \text{ diagonal matrix,} \\
& \boldsymbol{E} \text{ is a } p \times p \text{ diagonal matrix, and} \\
& \|\boldsymbol{A}_U\|_F^2 + \|\boldsymbol{C}_U\|_F^2 + \|\boldsymbol{D}\|_F^2 + \|n^{1/2}a_K\boldsymbol{E}\|_F^2 = \rho^2\right\}, \qquad (3.11)
\end{aligned}$$

This neighborhood extends the notion of restricted parameter space of $(\boldsymbol{B}, \boldsymbol{\Lambda})$ in Paul and Peng (2009), by incorporating the vector of regression coefficient $\boldsymbol{\beta}$. Paul and Peng (2009) also provided two important expansions, which will be used in our proof of Theorem 3.5.1:

$$\boldsymbol{B}^{*T}(\boldsymbol{G}_{\boldsymbol{B}^*,\boldsymbol{U}}(1)-\boldsymbol{B}^*) = \boldsymbol{B}^{*T}\boldsymbol{U}+\mathcal{O}((\|\boldsymbol{B}^{*T}\boldsymbol{U}\|_F+\|(\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{U}\|_F)\|\boldsymbol{U}\|_F), \quad (3.12)$$

$$(\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{G}_{\boldsymbol{B}^*,\boldsymbol{U}}(1) = (\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{U}+\mathcal{O}(\|(\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{U}\|_F)\|\boldsymbol{U}\|_F) \quad (3.13)$$

as $\|\boldsymbol{U}\|_F \to 0$.

3.6.2 Appendix B: Proof of Theorem 3.5.1

It suffices to show that, given $\eta > 0$, for large enough K, there exists a constant c_η, such that

$$P\left\{\inf_{(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda})\in\Theta(c_\eta a_K)} L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) > L_K(\boldsymbol{\beta}_0,\boldsymbol{B}^*,\boldsymbol{\Lambda}^*)\right\} \geq 1-\mathcal{O}(K^{-\eta}),$$

where $\Theta(c_\eta a_K) \equiv \Theta((\boldsymbol{\beta}_0,\boldsymbol{B}^*,\boldsymbol{\Lambda}^*); c_\eta a_K)$ is the neighborhood defined in Appendix A. Note that, $L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B}^*,\boldsymbol{\Lambda}^*) = \{L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda})\}$ $+\{L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B}^*,\boldsymbol{\Lambda}^*)\}$. We will quantify $L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda})$ and $L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B}^*,\boldsymbol{\Lambda}^*)$, denoted by (I) and (II), respectively.

First, we show that

$$P\left\{\inf_{(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda})\in\Theta(c_\eta a_K)} L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda}) > 0\right\} \geq 1-\mathcal{O}(K^{-\eta}). \quad (3.14)$$

It can be shown that, for any $(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)$,

$$\begin{aligned}
& 2K\{L_K(\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0,\boldsymbol{B},\boldsymbol{\Lambda})\} \\
= & \sum_{k=1}^{K}\left\{(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}) - (\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}_0)\right\} \\
= & \sum_{k=1}^{K}\left\{2(\boldsymbol{\beta}_0-\boldsymbol{\beta})^T\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}_0)\right\} + \sum_{k=1}^{K}\left\{(\boldsymbol{\beta}_0-\boldsymbol{\beta})^T\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k(\boldsymbol{\beta}_0-\boldsymbol{\beta})\right\} \\
\equiv & (I_1)+(I_2)
\end{aligned}$$

The term (I_1) has a normal distribution with mean $\mathbf{0}$ and variance $4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. Moreover, by (A.1) and the definition of $\Theta(c_\eta a_K)$, there exists a constant $c_4 > 0$, such that $4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \leq c_4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. Together with (A.9), we have $4(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \leq c_4 C_2 \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2^2$, which yields $(I_1) = \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2 \mathcal{O}_p(N^{1/2})$

Next, we consider the term (I_2). By (A.1), (A.8) and the definition of $\Theta(c_\eta a_K)$, there exists a constant $c_5 > 0$, such that $(I_2) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Sigma}_{\text{KL}}^{-1} \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \geq c_5(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$. By (A.9), we have $(I_2) \geq c_5 C_1 N \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2^2$. For a sufficient large c_η, (I_2) dominates (I_1) and thus, (3.14) follows.

For (II), we follow similar arguments in Paul and Peng (2009) and establish its uniform bound as

$$P\left\{\inf_{(\boldsymbol{\beta}_0, \boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)} L_K(\boldsymbol{\beta}_0, \boldsymbol{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*) > (c_\eta a_K)^2\right\} \geq 1 - \mathcal{O}(K^{-\eta}). \tag{3.15}$$

For any fixed $\boldsymbol{\beta}_0$, we can express (II) as

$$\begin{aligned}(II) &= K^{-1}\sum_{k=1}^{K} V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) + (2K)^{-1}\sum_{k=1}^{K} \text{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(S_k - \boldsymbol{\Sigma}_{0k})\} \\ &\quad + (2K)^{-1}\sum_{k=1}^{K} \text{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*)\} \equiv (II_1) + (II_2) + (II_3),\end{aligned}$$

where $\boldsymbol{S}_k = (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_0)(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_0)^T$ and $V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) = (1/2)\text{tr}\left\{\boldsymbol{\Sigma}_k^{-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1/2}\right\} - (1/2)\text{log}\left|\boldsymbol{I}_{n_k} + \boldsymbol{\Sigma}_k^{-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1/2}\right|$.

In Appendix B of this chapter, we bound the above three terms individually in Lemmas 6, , 7 and 8. As a consequence, we have $P\left\{L_K(\boldsymbol{\beta}_0, \boldsymbol{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*) \leq (c_\eta a_K)^2\right\} = \mathcal{O}(K^{-(2+2\kappa)MJ-\eta})$, for each point $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta_0(c_\eta a_K)$, where $\Theta_0(c_\eta a_K) \equiv \{(\boldsymbol{B}, \boldsymbol{\Lambda}) : (\boldsymbol{\beta}_0, \boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta(c_\eta a_K)\}$

Furthermore, define a restricted neighborhood in $\mathcal{S}_{M,J} \bigotimes \mathbb{R}^J$, centered at $(\boldsymbol{B}_1, \boldsymbol{\Lambda}_1)$ with size ω_K, as $\text{Ne}(\boldsymbol{B}_1, \boldsymbol{\Lambda}_1; \omega_K) = \{(\boldsymbol{B}, \boldsymbol{\Lambda}) : \|\boldsymbol{B} - \boldsymbol{B}_1\|_F^2 + \|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_1\|_F^2 \leq \omega_K^2\}$. There is a finite set in $\mathcal{S}_{M,J} \bigotimes \mathbb{R}^J$, denoted by $\mathcal{C}[\omega_K]$, such that $\bigcup_{(\boldsymbol{B}_1, \boldsymbol{\Lambda}_1) \in \mathcal{C}[\omega_K]} \text{Ne}(\boldsymbol{B}_1, \boldsymbol{\Lambda}_1; \omega_K) \supset$

$\Theta_0(c_\eta a_K)$. In fact, by a standard construction of such neighborhoods on a sphere in $\mathbb{R}^p$ $(p = MJ - J(J-1)/2)$, there exists $\mathcal{C}[\omega_K]$, in which the number of elements is of order $\max\{1, (a_K\omega_K^{-1})^p\}$.

Thus, by (A.3), for a sufficiently large K, taking $\omega_K = (\overline{n}^2 K)^{-1}$ yields $P\left\{\inf_{(\boldsymbol{B},\boldsymbol{\Lambda})\in\mathcal{C}[\omega_K]} L_K(\boldsymbol{\beta}_0, \boldsymbol{B}, \boldsymbol{\Lambda}) - L_K(\boldsymbol{\beta}_0, \boldsymbol{B}^*, \boldsymbol{\Lambda}^*) > (c_\eta a_K)^2\right\} \geq 1 - \mathcal{O}(K^{-\eta})$. Together with Lemma 5 in Appendix D of this chapter, we have (3.15). Finally, combining (3.14) and (3.15), Theorem 3.5.1 follows.

3.6.3 Appendix C: Lemmas

In this section, we present and prove lemmas which facilitate our proof of Theorem 1. In particular, Lemmas 6-8 will be used to bound the terms (II_1), (II_2) and (II_3) in the proof of Theorem 1, respectively. Lemma 5 provides a uniform bound for the quantity $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F$, which plays an important role in the proofs of Lemmas 6-8.

Lemma 5. *Under (A.1), (A.2), (A.6) and (A.8), for every* $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta_0(c_\eta a_K)$, *we have*

$$\max_{1\leq k\leq K} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F^2 \leq \left[D_1\left\{1 + D_2\{(M^{3/2}\log K/\overline{n}) \vee (M^2\log K/\overline{n})^{1/2}\}\right\}\overline{n}^2 a_K^2\right] \wedge (D_2 M\overline{n}^2 a_K^2), \tag{3.16}$$

for some constants D_1, D_2, $D_3 > 0$.

Proof. Note that $\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^* = \boldsymbol{\Phi}_k^T(\boldsymbol{B\Lambda B}^T - \boldsymbol{B}^*\boldsymbol{\Lambda}^*\boldsymbol{B}^{*T})\boldsymbol{\Phi}_k$ can be expressed as

$$\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^* = \boldsymbol{\Phi}_k^T\boldsymbol{B}^*(\boldsymbol{B}^{*T}\boldsymbol{B\Lambda B}^T\boldsymbol{B}^* - \boldsymbol{\Lambda}^*)\boldsymbol{B}^{*T}\boldsymbol{\Phi}_k + 2\boldsymbol{\Phi}_k^T\boldsymbol{B}^*\boldsymbol{B}^{*T}\boldsymbol{B\Lambda B}^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{\Phi}_k$$
$$+\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B\Lambda B}^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{\Phi}_k = (III_{1k}) + (III_{2k}) + (III_{3k}).$$

It can be seen that $\|(III_{1k})\|_F \leq \|\boldsymbol{\Phi}_k^T\boldsymbol{B}^*\|_F^2\|\boldsymbol{B}^{*T}\boldsymbol{B\Lambda B}^T\boldsymbol{B}^* - \boldsymbol{\Lambda}^*\|_F$. Note that, by (A.2) and (A.6), there exists a constant, $D_4 > 0$, such that

$$\max_{1\leq j\leq J}\|\varphi_{j,k}^*\|_\infty \leq D_4 < \infty. \tag{3.17}$$

Therefore, each entry of $\boldsymbol{\Phi}_k^T \boldsymbol{B}^*$ is bounded by D_4, and hence $\|\boldsymbol{\Phi}_k^T \boldsymbol{B}^*\|_F^2 \leq D_4^2 J n_k$. We have

$$\max_{1\leq k\leq K} \|(III_{1k})\|_F = \mathcal{O}(\overline{n}\|\boldsymbol{B}^{*T}\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^T\boldsymbol{B}^* - \boldsymbol{\Lambda}^*\|_F). \tag{3.18}$$

For (III_{2k}), we have

$$\|(III_{2k})\|_F \leq 2\|\boldsymbol{\Phi}_k^T\boldsymbol{B}^*\|_F\|\boldsymbol{B}^{*T}\boldsymbol{B}\|_2\|\boldsymbol{\Lambda}\|_2\|\boldsymbol{B}^T(\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{\Phi}_k\|_F \leq D_5\overline{n}^{1/2}\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M-\boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F, \tag{3.19}$$

for some $D_5 > 0$. Here, we use $\|\boldsymbol{B}^{*T}\boldsymbol{B}\| \leq 1$ and the definition of $\Theta(c_\eta a_K)$. Therefore,

$$\max_{1\leq k\leq K} \|(III_{2k})\|_F = \mathcal{O}(\overline{n}^{1/2}\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F). \tag{3.20}$$

Last, we have

$$\max_{1\leq k\leq K} \|(III_{3k})\|_F \leq \|\boldsymbol{\Lambda}\|\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F^2 = \mathcal{O}(\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F^2). \tag{3.21}$$

By (A.6), (A.8) and (12) in main manuscript, we have

$$\overline{n}^{-1}\max_{1\leq k\leq K} \|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F^2 \leq c_{\phi,2}M\|(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F^2 \leq c_{\phi,2}Ma_K^2(1+o(1)) = o(1).$$

Therefore, we have,

$$\max_{1\leq k\leq K} \|(III_{3k})\|_F = o(\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F). \tag{3.22}$$

By the triangle inequality, (3.18), (3.20) and (3.22), we obtain

$$\max_{1\leq k\leq K} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \max_{1\leq k\leq K} \leq \mathcal{O}(\overline{n}\|\boldsymbol{B}^{*T}\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^T\boldsymbol{B}^* - \boldsymbol{\Lambda}^*\|_F) + \mathcal{O}(\overline{n}^{1/2}\|\boldsymbol{\Phi}_k^T(\boldsymbol{I}_M - \boldsymbol{B}^*\boldsymbol{B}^{*T})\boldsymbol{B}\|_F).$$

By (A.1), (A.6), the definition of $\Theta(c_\eta a_K)$, (11) and (12) in main manuscript, the result follows.

□

Lemma 6. *Under (A.1)–(A.3), (A.5)–(A.6) and (A.8), for every* $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta_0(c_\eta a_K)$*, we have*

$$d_3 a_K^2 \leq K^{-1} \sum_{k=1}^{K} V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) \leq d_4 \overline{n}^2 a_K^2, \tag{3.23}$$

for appropriate positive constants d_3 *and* d_4*.*

Proof. Note that

$$V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) = \sum_{i=1}^{n_k} \left[\lambda_i(\boldsymbol{R}_k^*) - \log \left\{ 1 + \lambda_i(\boldsymbol{R}_k^*) \right\} \right],$$

where $\boldsymbol{R}_k^* = \boldsymbol{\Sigma}_k^{-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1/2}$.

By Taylor's expansion, we can show that, for any sufficiently small $\epsilon > 0$, there exist constants $0 < c_{1,\epsilon} < c_{2,\epsilon} < \infty$ such that, for $\|\boldsymbol{R}_k^*\|_F \leq \epsilon$,

$$c_{1,\epsilon} \|\boldsymbol{R}_k^*\|_F^2 \leq V(\boldsymbol{\Sigma}_k, \boldsymbol{\Sigma}_k^*) \leq c_{2,\epsilon} \|\boldsymbol{R}_k^*\|_F^2. \tag{3.24}$$

Straightforward matrix calculation yields

$$\frac{\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F}{\left\{ 1 + \|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F \right\}} \leq \|\boldsymbol{R}_k^*\|_F \leq \frac{\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F}{\left\{ 1 - \|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F \right\}}$$

whenever $\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F < 1$.

Next we will show that $\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F$ converges to 0 uniformly in k. Thus, by (A.5), the desired result follows.

By (3.17), (A.1) and definition of $\Theta(c_\eta a_K)$, we have

$$\|\boldsymbol{\Phi}_k^T \boldsymbol{B}^* \boldsymbol{\Lambda}^* \boldsymbol{B}^{*T} \boldsymbol{\Phi}_k\|_2 \leq \|\boldsymbol{\Phi}_k^T \boldsymbol{B}^*\|_F^2 \|\boldsymbol{\Lambda}^*\|_2 \leq D_6 \lambda_1 J \overline{n}, \text{ for } k = 1, \ldots, K,$$

for some $D_6 > 0$. Therefore, $1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_k^*) \leq \lambda_{\max}(\boldsymbol{\Sigma}_k^*) \leq 1 + D_6 \lambda_1 J \overline{n}$.

We obtain

$$(1 + D_2 \lambda_1 J \overline{n})^{-1} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F \leq \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F. \tag{3.25}$$

By Lemma 5 and (A.8), we have $\max_{1\le k\le K}\|\boldsymbol{\Sigma}_k-\boldsymbol{\Sigma}_k^*\|_F^2=o(1)$. Therefore, $\max_{1\le k\le K}\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k^*-\boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{*-1/2}\|_F=o(1)$. □

The following lemma is used to bound the quantity (II_2) in the proof of Theorem 1. This lemma is a generalization of Proposition 3 of Paul and Peng (2009) from the independent case to a more general weakly dependent case under a mild assumption (A.7). The key device used in the proof is the Bernstein inequality of an array of weakly dependent random variables; see Lemma 5.3 of Sun and Lahiri (2003).

Lemma 7. *Under (A.1)–(A.8), given any $\eta>0$, for every $(\boldsymbol{B},\boldsymbol{\Lambda})\in\Theta_0(c_\eta a_K)$, we have*

$$P\left\{\left|(2K)^{-1}\sum_{k=1}^{K}\operatorname{tr}\left\{\left(\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{\Sigma}_k^{*-1}\right)\left(\boldsymbol{S}_k-\boldsymbol{\Sigma}_{0k}\right)\right\}\right|\le d_\eta\overline{n}a_K(M\,logK/K)^{1/2}\right\}$$
$$\ge 1-\mathcal{O}(K^{-(2+2\kappa)MJ-\eta}),$$

where $\boldsymbol{S}_k=(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}_0)(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta}_0)^T$.

Proof. Let $\boldsymbol{R}_k=\boldsymbol{\Sigma}_{0k}^{1/2}(\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{\Sigma}_k^{*-1})\boldsymbol{\Sigma}_{0k}^{1/2}$. Note that

$$\boldsymbol{R}_k=(\boldsymbol{\Sigma}_{0k}^{1/2}\boldsymbol{\Sigma}_k^{-1/2})(\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_k^{*1/2})\left\{\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k-\boldsymbol{\Sigma}_k^*)\boldsymbol{\Sigma}_k^{*-1/2}\right\}(\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}^{1/2}).$$

Therefore, we have

$$\|\boldsymbol{R}_k\|_F\le\|\boldsymbol{\Sigma}_{0k}^{1/2}\boldsymbol{\Sigma}_k^{-1/2}\|_2\|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}^{1/2}\|_2\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_k^{*1/2}\|_2\|\boldsymbol{\Sigma}_k^{*-1/2}(\boldsymbol{\Sigma}_k-\boldsymbol{\Sigma}_k^*)\boldsymbol{\Sigma}_k^{*-1/2}\|_F.$$

Moreover, it can be seen that $\|\boldsymbol{\Sigma}_{0k}^{1/2}\boldsymbol{\Sigma}_k^{-1/2}\|_2\le\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_k^{*1/2}\|_2\|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}^{1/2}\|_2$. Thus, combining (3.25), we have

$$\|\boldsymbol{R}_k\|_F\ \le\ \|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}^{1/2}\|_2^2\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_k^{*1/2}\|_2^2\|\boldsymbol{\Sigma}_k-\boldsymbol{\Sigma}_k^*\|_F.$$

Next, we bound the first two terms on the right hand side of the above inequality. In fact, note that

$$\begin{aligned}\|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}^{1/2}\|_2^2 &= \|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}\boldsymbol{\Sigma}_k^{*-1/2}\|_2 \leq 1 + \|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_{0k}\boldsymbol{\Sigma}_k^{*-1/2} - 1\|_2 \\ &\leq 1 + \|\boldsymbol{\Sigma}_k^{*-1/2}\|_2\|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_2\|\boldsymbol{\Sigma}_k^{*-1/2}\|_2 \leq 1 + \|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_2.\end{aligned}$$

Similarly, we have $\|\boldsymbol{\Sigma}_k^{*-1/2}\boldsymbol{\Sigma}_k^{1/2}\|_2^2 \leq 1 + \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2$. Thus,

$$\|\boldsymbol{R}_k\|_F \leq (1 + \|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_2)(1 + \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2)\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F.$$

By (A.4), we have $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq C\overline{n}\delta_K = o(1)$ for some constant $C > 0$. In addition, by Lemma 5 and (A.8), we have $\|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*\|_F = o(1)$. Consequently, for sufficiently large K, we have $\|\boldsymbol{R}_k\| \leq 2\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F$.

Applying an array form of the Bernstein inequality, Lemma 5.3 of (Sun and Lahiri, 2003), to $\mathrm{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{S}_k - \boldsymbol{\Sigma}_{0k})\}$, which depends on both K and k, we have

$$P\left\{|(2K)^{-1}\sum_{k=1}^{K}\mathrm{tr}\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{S}_k - \boldsymbol{\Sigma}_{0k})\}| > (M\log K/K)^{1/2}(K^{-1}\sum_{k=1}^{K}\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F^2)^{1/2}\right\}$$
$$\leq \mathcal{O}\left(K^{-(2+2\kappa)MJ-\eta} + K\left\{\alpha_K(2)\right\}^{2k/(2k+1)}\right) = \mathcal{O}\left(K^{-(2+2\kappa)MJ-\eta}\right).$$

In the above inequality, the last equality is a direct consequence of (A.7). Finally, combined with (A.5), the proof is complete.

□

The following lemma is used to quantify $(II)_3$ in the proof of Theorem 1.

Lemma 8. *Under (A.1)–(A.6) and (A.8), for every* $(\boldsymbol{B}, \boldsymbol{\Lambda}) \in \Theta_0(c_\eta a_K)$, *we have*

$$(2K)^{-1}\sum_{i=1}^{K}\mathrm{tr}\left\{(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^*)\right\} = \mathcal{O}(\overline{n}a_K(M\,log K/K)^{1/2}). \tag{3.26}$$

Proof. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left| K^{-1} \sum_{i=1}^{K} \mathrm{tr}\left\{ (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1})(\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^{*}) \right\} \right| \\
\le\ & \left\{ K^{-1} \sum_{i=1}^{K} \|\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1}\|_F^2 \right\}^{1/2} \left\{ K^{-1} \sum_{i=1}^{K} \|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^{*}\|_F^2 \right\}^{1/2}.
\end{aligned}$$

We further note that

$$\left\{ K^{-1} \sum_{i=1}^{K} \|\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{*-1}\|_F^2 \right\}^{1/2} \le \max_{1\le k\le K} \|\boldsymbol{\Sigma}_k^{-1}\|_2 \|\boldsymbol{\Sigma}_k^{*-1}\|_2 \left\{ K^{-1} \sum_{i=1}^{K} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^{*}\|_F^2 \right\}^{1/2} = \mathcal{O}(\overline{n} a_K),$$

which is a direct consequence of $\|\boldsymbol{\Sigma}_k^{-1}\|_2 < 1$, $\|\boldsymbol{\Sigma}_k^{*-1}\|_2 < 1$, and (A.5). Moreover, by (A.4), we have

$$\left\{ K^{-1} \sum_{i=1}^{K} \|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^{*}\|_F^2 \right\}^{1/2} \le \max_{1\le k\le K} \|\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_k^{*}\|_F \le \overline{n}\delta_K = \mathcal{O}((M\mathrm{log}K/K)^{1/2}),$$

which completes the proof.

□

Lemma 9. *Let $(\boldsymbol{B}_1, \boldsymbol{\Lambda}_1)$ and $(\boldsymbol{B}_2, \boldsymbol{\Lambda}_2)$ be two elements of $\Theta_0(c_\eta a_K)$ satisfying $\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 + \|\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\|_F^2 \le \omega_K^2$ with $\omega_K = (\overline{n}^2 K)^{-1}$. Then under (A.1)–(A.4), (A.6) and (A.8), given $\eta > 0$, we have*

$$P\left\{ |L_K(\boldsymbol{\beta}_0, \boldsymbol{B}_1, \boldsymbol{\Lambda}_1) - L_K(\boldsymbol{\beta}_0, \boldsymbol{B}_2, \boldsymbol{\Lambda}_2)| = o(a_K^2) \right\} = \mathcal{O}(K^{-\eta-1}).$$

Proof. Note that

$$\begin{aligned}
& L_K(\boldsymbol{\beta}_0, \boldsymbol{B}_1, \boldsymbol{\Lambda}_1) - L_K(\boldsymbol{\beta}_0, \boldsymbol{B}_2, \boldsymbol{\Lambda}_2) \\
=\ & (2K)^{-1} \sum_{k=1}^{K} \mathrm{tr}\{ (\boldsymbol{\Sigma}_{1,k}^{-1} - \boldsymbol{\Sigma}_k^{*-1})(S_k - \boldsymbol{\Sigma}_{0k}) \} + (2K)^{-1} \sum_{k=1}^{K} \mathrm{tr}\{ (\boldsymbol{\Sigma}_k^{*-1} - \boldsymbol{\Sigma}_{2,k}^{-1})(S_k - \boldsymbol{\Sigma}_{0k}) \} \\
& + K^{-1} \sum_{k=1}^{K} V(\boldsymbol{\Sigma}_{1,k}, \boldsymbol{\Sigma}_{2,k}) + (2K)^{-1} \sum_{k=1}^{K} \mathrm{tr}\{ (\boldsymbol{\Sigma}_{1,k}^{-1} - \boldsymbol{\Sigma}_{2,k}^{-1})(\boldsymbol{\Sigma}_{0k} - \boldsymbol{\Sigma}_{2,k}) \} \\
=\ & (III_1) + (III_2) + (III_3) + (III_4).
\end{aligned}$$

Following a similar method as Lemma 7, we have $P(|(III_1)| = o(a_K^2)) = \mathcal{O}(K^{-\eta-1})$ and $P(|(III_2)| = o(a_K^2)) = \mathcal{O}(K^{-\eta-1})$.

For (III_3), by (3.24) and (3.25), there exists C_1, $C_2 > 0$, for large enough K, we have

$$\left|K^{-1}\sum_{k=1}^{K} V(\mathbf{\Sigma}_{1,k}, \mathbf{\Sigma}_{2,k})\right| \leq (2K)^{-1}\sum_{k=1}^{K} C_1\|\mathbf{\Sigma}_{1,k} - \mathbf{\Sigma}_{2,k}\|_F^2 \leq C_2 \max_{1\leq k\leq K}\|\mathbf{\Sigma}_{1,k} - \mathbf{\Sigma}_{2,k}\|_F^2.$$

By the triangle inequality and (A.1), simply computation reveals that

$$\begin{aligned}\|\boldsymbol{B}_1\mathbf{\Lambda}_1\boldsymbol{B}_1^T - \boldsymbol{B}_2\mathbf{\Lambda}_2\boldsymbol{B}_2^T\|_F &\leq \|(\boldsymbol{B}_1 - \boldsymbol{B}_2)\mathbf{\Lambda}_1\boldsymbol{B}_1^T\|_F + \|\boldsymbol{B}_2(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)\boldsymbol{B}_1^T\|_F + \|\boldsymbol{B}_2\mathbf{\Lambda}_2(\boldsymbol{B}_1 - \boldsymbol{B}_2)^T\|_F \\ &= \mathcal{O}(\omega_K).\end{aligned}$$

Therefore, by (A.6),we have

$$\max_{1\leq k\leq K}\|\mathbf{\Sigma}_{1,k} - \mathbf{\Sigma}_{2,k}\|_F^2 \leq \|\boldsymbol{B}_1\mathbf{\Lambda}_1\boldsymbol{B}_1^T - \boldsymbol{B}_2\mathbf{\Lambda}_2\boldsymbol{B}_2^T\|_F\|\mathbf{\Phi}_k\|_2^2 = \mathcal{O}(\overline{n}^2M^2\omega_K^2).$$

Thus, $(III_3) = o(a_K^2)$.

For (III_4), by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|(2K)^{-1}\sum_{k=1}^{K}\mathrm{tr}\{(\mathbf{\Sigma}_{1,k}^{-1} - \mathbf{\Sigma}_{2,k}^{-1})(\mathbf{\Sigma}_{0k} - \mathbf{\Sigma}_{2,k})\}\right| \leq (2K)^{-1}\sum_{k=1}^{K}\|\mathbf{\Sigma}_{1,k}^{-1} - \mathbf{\Sigma}_{2,k}^{-1}\|_F\|\mathbf{\Sigma}_{0k} - \mathbf{\Sigma}_{2,k}\|_F \\ &\leq \max_{1\leq k\leq K}\|\mathbf{\Sigma}_{1,k}^{-1} - \mathbf{\Sigma}_{2,k}^{-1}\|_F(\max_{1\leq k\leq K}\|\mathbf{\Sigma}_{0k} - \mathbf{\Sigma}_k^*\|_F + \max_{1\leq k\leq K}\|\mathbf{\Sigma}_{2,k} - \mathbf{\Sigma}_k^*\|_F) \\ = \;&\mathcal{O}(\overline{n}\omega_K(\overline{n}\delta_K + M^{1/2}\overline{n}a_K)) = o(a_K^2).\end{aligned}$$

Thus, the proof is complete. □

3.6.4 Appendix D: Additional Simulation Results

In Section 3.4 of the chapter, our proposed method (KL) is compared with three competing methods, namely ALT$_1$, ALT$_2$, and ALT$_3$, in a simulation study. Here, we report more details of the simulation results regarding regression parameter estimation, spatial prediction, and variable selection. In particular, for regression parameter estimation, the mean and standard deviation of estimated parameters are reported. For spatial prediction, the mean and standard deviation of the mean squared prediction error (MSPE) are reported. For variable selection, we report the average number of correctly identified zero-valued regression coefficients (C0) and the average number of incorrectly identified zero-valued regression coefficients (I0).

Scenario 1: $d = 1$, Exponential Covariance.

The simulation results for this scenario are reported in Table 3.3. It can be seen that, KL, ALT$_2$, and ALT$_3$ all outperform ALT$_1$ significantly. As pointed out in Section 3.4, by accounting for spatial dependence, there is noticeable improvement for regression parameter estimation, spatial prediction and variable selection. Moreover, for regression parameter estimation and variable selection, KL, ALT$_2$, and ALT$_3$ are comparable, which suggests that the estimation of $\boldsymbol{\beta}$ is not sensitive to the approximation of covariance structure. For spatial prediction, ALT$_2$ performs better than KL and ALT$_3$, as is expected.

Scenario 2: $d = 1$, Misspecified Covariance.

The simulation results for this scenario are reported in Table 3.4. Similar to Scenario 1, by considering spatial dependence, KL, ALT$_2$, and ALT$_3$ all outperform ALT$_1$ significantly, even though the covariance function is misspecified in ALT$_2$ and ALT$_3$. However, unlike Scenario 1, KL outperforms both ALT$_2$ and ALT$_3$ in spatial prediction, due to a more accurate estimate of the covariance function by KL, as shown in Figure 3. For regression parameter estimation and variable selection, KL, ALT$_2$, and ALT$_3$ perform similarly.

Scenario 3: $d = 2$, Exponential Covariance

For $d = 2$, let $R = [0, L_1] \times [0, L_2]$ in $\mathbb{R}^2$. We fix the value of L_1 at 6, and consider three different values for $L_2 = 6$, 12, 18. The sample sizes are set to be $N = 300$, 600, 900 for $L_2 = 6$, 12, 18, respectively. For regression, the large-scale trend is generated

Table 3.3: Simulation Results from Scenario 1

N	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	MSPE	C0	I0
300	KL	4.025	3.004	2.004	0.965	-0.018	0.003	-0.015	5.83	2.98	0.00
	SD	0.171	0.196	0.165	0.175	0.176	0.187	0.188	1.52		
	ALT$_1$	4.036	2.983	2.046	0.971	-0.022	0.008	-0.030	17.84	2.91	0.22
	SD	0.302	0.347	0.290	0.323	0.309	0.342	0.321	6.32		
	ALT$_2$	4.026	3.006	2.004	0.962	-0.016	0.004	-0.018	5.69	2.92	0.00
	SD	0.172	0.186	0.161	0.172	0.177	0.185	0.185	1.49		
	ALT$_3$	4.030	3.006	2.003	0.963	-0.016	0.001	-0.019	5.72	2.95	0.00
	SD	0.172	0.187	0.164	0.172	0.178	0.184	0.184	1.49		
600	KL	4.011	3.003	1.987	0.983	0.018	-0.017	0.008	5.70	2.98	0.00
	SD	0.122	0.133	0.114	0.129	0.123	0.140	0.135	1.20		
	ALT$_1$	4.023	2.994	1.985	0.967	0.029	-0.008	0.020	18.92	2.94	0.05
	SD	0.231	0.245	0.230	0.231	0.253	0.242	0.242	4.36		
	ALT$_2$	4.010	3.002	1.987	0.985	0.019	-0.018	0.008	5.62	2.96	0.00
	SD	0.124	0.129	0.115	0.129	0.126	0.138	0.137	1.17		
	ALT$_3$	4.010	3.002	1.987	0.983	0.019	-0.020	0.010	5.68	2.96	0.00
	SD	0.123	0.128	0.114	0.130	0.123	0.137	0.136	1.19		
900	KL	3.984	2.975	2.009	1.004	0.009	0.010	-0.004	5.61	2.99	0.00
	SD	0.109	0.088	0.103	0.113	0.110	0.098	0.099	0.84		
	ALT$_1$	3.967	2.981	2.009	1.039	-0.025	0.025	0.017	19.21	2.90	0.01
	SD	0.188	0.180	0.203	0.187	0.222	0.188	0.192	4.12		
	ALT$_2$	3.984	2.975	2.008	1.002	0.011	0.010	-0.003	5.51	2.99	0.00
	SD	0.108	0.088	0.102	0.114	0.108	0.096	0.100	0.81		
	ALT$_3$	3.984	2.975	2.008	1.001	0.011	0.010	-0.003	5.56	2.98	0.00
	SD	0.109	0.088	0.104	0.115	0.109	0.095	0.101	0.82		

Simulation results from Scenario 1: mean, standard deviation (SD) of regression coefficients estimates and mean squared prediction error (MSPE), average number of correctly identified zero-valued regression coefficients(C0) and average number of incorrectly identified zero-valued regression coefficients (I0) under KL, ALT$_1$, ALT$_2$, and ALT$_3$ for sample size $N = 300, 600, 900$.

Table 3.4: Simulation Results from Scenario 2

N	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_6	MSPE	C0	I0
300	KL	4.022	3.005	2.005	0.966	-0.014	0.002	-0.011	4.88	2.96	0.00
	SD	0.155	0.175	0.157	0.157	0.162	0.174	0.173	1.30		
	ALT$_1$	4.034	2.973	2.037	0.976	-0.020	0.007	0.003	18.97	2.94	0.29
	SD	0.308	0.354	0.326	0.319	0.302	0.354	0.363	6.43		
	ALT$_2$	4.023	3.005	2.003	0.963	-0.016	0.004	-0.012	4.96	2.92	0.00
	SD	0.156	0.172	0.158	0.158	0.166	0.175	0.175	1.29		
	ALT$_3$	4.026	3.005	2.002	0.964	-0.015	0.002	-0.013	4.99	2.94	0.00
	SD	0.156	0.173	0.161	0.159	0.167	0.174	0.174	1.27		
600	KL	4.008	3.003	1.988	0.984	0.017	-0.016	0.008	4.81	2.97	0.00
	SD	0.110	0.121	0.104	0.120	0.115	0.124	0.128	0.98		
	ALT$_1$	4.000	3.004	1.971	0.991	0.045	-0.014	0.008	19.81	2.94	0.06
	SD	0.226	0.233	0.221	0.243	0.264	0.275	0.241	3.97		
	ALT$_2$	4.006	3.004	1.987	0.986	0.019	-0.017	0.007	4.93	2.97	0.00
	SD	0.112	0.121	0.106	0.123	0.119	0.124	0.131	0.99		
	ALT$_3$	4.006	3.005	1.986	0.985	0.019	-0.020	0.009	4.99	2.97	0.00
	SD	0.111	0.120	0.106	0.124	0.117	0.123	0.130	1.00		
900	KL	3.985	2.975	2.008	1.001	0.012	0.011	-0.002	4.76	2.99	0.00
	SD	0.100	0.081	0.095	0.105	0.100	0.088	0.092	0.72		
	ALT$_1$	3.976	2.964	1.999	1.029	-0.024	0.026	0.035	19.92	2.92	0.02
	SD	0.216	0.174	0.218	0.207	0.226	0.183	0.190	3.46		
	ALT$_2$	3.986	2.974	2.008	0.998	0.014	0.010	-0.002	4.90	2.99	0.00
	SD	0.102	0.080	0.098	0.107	0.102	0.087	0.097	0.74		
	ALT$_3$	3.986	2.975	2.008	0.998	0.014	0.011	-0.003	4.95	2.98	0.00
	SD	0.103	0.081	0.099	0.109	0.103	0.087	0.098	0.75		

Simulation results from Scenario 2: mean, standard deviation (SD) of regression coefficients estimates and mean squared prediction error (MSPE), average number of correctly identified zero-valued regression coefficients(C0) and average number of incorrectly identified zero-valued regression coefficients (I0) under KL, ALT$_1$, ALT$_2$, and ALT$_3$ for sample size $N = 300$, 600, 900.

in the same way as the case of $d = 1$. For spatial dependence, we generate the error $\varepsilon_1(\boldsymbol{s})$ for each sampling location $\boldsymbol{s}$ from a zero-mean stationary and isotropic Gaussian process with an exponential covariance function $\gamma(h) = \sigma_1^2 \exp(-h/c_r)$. In addition, the measurement error terms $\varepsilon_2(\boldsymbol{s})$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 2$. For the ALT$_3$ and KL method, the subdomains are squares with side length 3. The results are reported in Table 3.5, and similar conclusions can be drawn as Scenario 1.

Table 3.5: Simulation Results from Scenario 3

N	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	MSPE	C0	I0
300	KL	4.017	2.974	2.024	1.004	-0.006	-0.007	-0.037	7.45	3.00	0.00
	SD	0.195	0.220	0.216	0.206	0.198	0.192	0.190	1.98		
	ALT$_1$	4.023	2.997	2.036	1.005	-0.011	-0.035	-0.065	16.08	2.88	0.12
	SD	0.300	0.319	0.312	0.265	0.297	0.299	0.267	4.99		
	ALT$_2$	4.014	2.983	2.022	1.008	-0.014	-0.011	-0.032	7.18	2.97	0.00
	SD	0.190	0.216	0.209	0.205	0.194	0.184	0.197	1.90		
	ALT$_3$	4.013	2.983	2.023	1.004	-0.011	-0.013	-0.028	7.35	2.97	0.00
	SD	0.192	0.217	0.206	0.203	0.197	0.186	0.200	1.92		
600	KL	4.021	2.995	1.978	0.990	0.018	-0.029	0.006	6.66	2.99	0.00
	SD	0.135	0.146	0.144	0.154	0.146	0.136	0.138	1.17		
	ALT$_1$	4.020	3.003	1.949	0.980	0.019	-0.006	0.036	17.64	2.96	0.04
	SD	0.260	0.240	0.213	0.252	0.196	0.219	0.207	5.46		
	ALT$_2$	4.016	2.995	1.977	0.998	0.017	-0.026	0.006	6.45	2.97	0.00
	SD	0.134	0.143	0.141	0.149	0.146	0.133	0.134	1.19		
	ALT$_3$	4.014	2.992	1.978	0.999	0.017	-0.023	0.007	6.61	2.99	0.00
	SD	0.139	0.145	0.143	0.150	0.147	0.134	0.136	1.21		
900	KL	4.004	2.966	1.992	1.008	0.002	0.009	0.008	6.67	2.99	0.00
	SD	0.119	0.125	0.127	0.114	0.107	0.110	0.102	1.02		
	ALT$_1$	4.018	2.961	1.997	0.999	0.000	0.008	0.014	17.17	2.97	0.00
	SD	0.209	0.193	0.184	0.193	0.201	0.186	0.176	3.69		
	ALT$_2$	4.001	2.966	1.995	1.008	0.002	0.008	0.010	6.50	2.98	0.00
	SD	0.117	0.124	0.126	0.107	0.107	0.110	0.099	0.96		
	ALT$_3$	4.002	2.967	1.992	1.007	0.002	0.009	0.008	6.70	2.97	0.00
	SD	0.120	0.129	0.125	0.107	0.111	0.114	0.100	1.03		

Simulation results from Scenario 3: mean, standard deviation (SD) of regression coefficients estimates and mean squared prediction error (MSPE), average number of correctly identified zero-valued regression coefficients(C0) and average number of incorrectly identified zero-valued regression coefficients (I0) under KL, ALT$_1$, ALT$_2$, and ALT$_3$ for sample size $N = 300$, 600, 900.

Scenario 4: $d = 2$, Misspecified Covariance

The setup is the same as that in Scenario 3 except for the spatial dependence structure. Specifically, the error process $\varepsilon_1(\cdot)$ follows the sinusoid covariance function

$\gamma(h) = \sigma_1^2 \sin(h/c_r)c_r/h$. Moreover, the measurement error terms $\varepsilon_2(\cdot)$ are independently generated from $N(0, \sigma_2^2)$. Let $\sigma_1^2 = 16$, $\sigma_2^2 = 4$ and $c_r = 0.4$. The results are reported in Table 3.6, and similar conclusions can be drawn as Scenario 2.

Table 3.6: Simulation Results from Scenario 4

N	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	MSPE	C0	I0
300	KL	4.008	2.986	2.018	1.010	-0.008	-0.012	-0.028	5.42	3.00	0.00
	SD	0.168	0.183	0.185	0.177	0.167	0.157	0.163	1.53		
	ALT$_1$	3.983	3.045	2.014	1.009	0.002	-0.019	-0.059	20.54	2.95	0.26
	SD	0.366	0.349	0.360	0.293	0.316	0.331	0.318	6.42		
	ALT$_2$	4.006	2.980	2.017	1.012	-0.016	-0.014	-0.016	5.83	2.97	0.00
	SD	0.177	0.198	0.183	0.189	0.178	0.157	0.186	1.58		
	ALT$_3$	4.006	2.982	2.018	1.010	-0.015	-0.014	-0.012	6.10	2.99	0.00
	SD	0.182	0.199	0.184	0.189	0.181	0.161	0.190	1.68		
600	KL	4.021	2.996	1.978	1.000	0.009	-0.024	0.004	4.70	2.99	0.00
	SD	0.118	0.120	0.123	0.123	0.128	0.114	0.114	0.94		
	ALT$_1$	4.009	2.986	1.967	0.996	0.006	-0.012	0.046	19.91	2.95	0.03
	SD	0.260	0.266	0.246	0.258	0.244	0.241	0.227	5.54		
	ALT$_2$	4.019	2.999	1.975	1.002	0.009	-0.023	0.007	5.12	2.98	0.00
	SD	0.127	0.128	0.131	0.131	0.138	0.121	0.121	0.96		
	ALT$_3$	4.017	2.997	1.975	1.004	0.009	-0.020	0.007	5.42	2.98	0.00
	SD	0.131	0.132	0.136	0.134	0.137	0.121	0.125	1.02		
900	KL	3.996	2.971	1.998	1.005	0.001	0.009	0.012	4.71	3.00	0.00
	SD	0.103	0.102	0.106	0.092	0.091	0.095	0.088	0.66		
	ALT$_1$	4.022	2.954	1.997	0.985	-0.022	0.040	0.012	19.68	2.95	0.01
	SD	0.215	0.203	0.186	0.203	0.189	0.186	0.193	3.71		
	ALT$_2$	3.994	2.967	2.000	1.006	0.003	0.005	0.014	5.17	2.98	0.00
	SD	0.107	0.109	0.110	0.093	0.097	0.101	0.090	0.73		
	ALT$_3$	3.996	2.970	1.996	1.004	0.004	0.006	0.011	5.51	2.98	0.00
	SD	0.112	0.114	0.109	0.096	0.100	0.105	0.090	0.81		

Simulation results from Scenario 4: mean, standard deviation (SD) of regression coefficients estimates and mean squared prediction error (MSPE), average number of correctly identified zero-valued regression coefficients(C0) and average number of incorrectly identified zero-valued regression coefficients (I0) under KL, ALT$_1$, ALT$_2$, and ALT$_3$ for sample size $N = 300$, 600, 900.

To choose J and M in practice, we proposed the following method. First, we increase J until $\widehat{\lambda}_J/\widehat{\lambda}_1$ is very small, e.g., $\widehat{\lambda}_J/\widehat{\lambda}_1 < 0.01$. Second, we increase M until $\widehat{\sigma}_{\mathrm{T}}^2$ decreases slowly or starts to increase. For the threshold distance ω in the tapering function, we set it to be equal or slightly smaller than D.

Chapter 4

DISCUSSION AND FUTURE WORK

4.1 Summary

In this dissertation, we have studied both parametric and semiparametric methods for parameter estimation, variable selection, and spatial prediction in geostatistics. In Chapter 2, the covariance structure of the error process is assumed to be parametric (e.g., Matern covariance function) and therefore, the likelihood function is obtained and maximum likelihood is used for parameter estimation. However, in order to save computational time, a covariance-tapered likelihood function is used instead of the likelihood function. Moreover, a penalized method is developed to carry out variable selection. Combining the two ideas above, we have proposed to maximize a penalized covariance-tapered likelihood function to conduct variable selection and parameter estimation simultaneously for a spatial linear model. We have also developed one-step sparse estimation and its counterpart under covariance tapering to approximate the penalized parameter estimates and gained computational efficiency. Furthermore, we have established asymptotic properties of the parameter estimates and their approximations, showing consistency, sparsity, and asymptotic normality. Finite-sample properties have been examined via a simulation study and we have found that, with direct incorporation of spatial autocorrelation in the penalized likelihood function, the accuracy of variable selection and the precision of parameter estimates improve over penalized methods that do not directly account for spatial dependence.

In Chapter 3, the covariance structure of the error process is not pre-specified. We have developed a nonparametric approach via Karhunen-Loève expansion to model the error process and a parametric form for the large-scale trend. That is, a principled semiparametric approach is adopted for regression parameter estimation in a spatial linear model. Taking advantage of stationarity of the error process, we developed a smoothing algorithm to further improve the accuracy of regression parameter estimation and Kriging. Our simulation study shows that the performance of the proposed method is close to the maximum likelihood method when the underlying covariance structure is correctly specified. Moreover, when the underlying covariance is misspecified, our proposed method performs better than the maximum likelihood method. Furthermore, the consistency of regression estimates has been established under certain conditions.

4.2 Future Work

For the algorithm in Section 3.3.1, Edelman et al. (1998) and Peng and Paul (2009) have investigated its convergence extensively. While the algorithm usually converges with a proper initial value and large sample size in the simulation study, no sufficient conditions for the convergence has been established theoretically. We will leave this question for future study.

Second, it is natural to extend our methods for spatial linear models to spatial-temporal linear models. Let R be the spatial domain of interest in $\mathbb{R}^d$ and T be the temporal domain of interest in $\mathbb{R}$. Consider a spatial-temporal process $\{y(\boldsymbol{s}, t) : \boldsymbol{s} \in R, t \in T\}$ to be modeled as,

$$y(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t),$$

where $\mu(\boldsymbol{s}, t)$ is an unknown mean function at location $\boldsymbol{s}$ and time t, and $\varepsilon(\boldsymbol{s}, t)$ is an error process on the spatial-temporal domain $R \times T$ (see, e.g., Stein, 2005). Although the spatial-temporal covariance function can be more complicated than the spatial covariance function, the basic idea for covariance tapering may still apply. That is, if $(\boldsymbol{s}, t)$ and $(\boldsymbol{s}', t')$ are far away from each other and their correlation is believed to be small, we can a tapering function to re-scale their covariance to zero, and obtain a sparse matrix to approximate the variance-covariance matrix. Similar to the spatial case, the computing time can be saved due to the faster inversion of the sparse matrix.

Furthermore, in this dissertation, the error process $\varepsilon(\boldsymbol{s})$ is assumed to be a stationary isotropic Gaussian process, which can be relaxed in several ways. First, stationarity can be relaxed to be local stationarity, or non-stationarity. In both case, Karhunen-Loève expansion will still work, but local Karhunen-Loève expansion not. Second, the Gaussian process can be relaxed to be other non-Gaussian process. However, for this case, Karhunen-Loève type expansion appears to be not available yet.

Finally, for the large-scale trend $\mu(\boldsymbol{s})$, we have focused on a linear trend $\boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}$ in both Chapter 2 and Chapter 3. However, other types of parametric trend can be considered for $\mu(\boldsymbol{s})$, as well as nonparametric trend by, for example, spline basis functions.

REFERENCES

Adler, R. and Taylor, J. (2007). *Random Fields and Geometry*. Springer, New York.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, second edition.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, Cambridge.

Caragea, P. and Smith, R. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98:1417–1440.

Chu, T., Zhu, J., and Wang, H. (2011). Penalized maximum likelihood estiamtion and variable selection in geostatistics. *Annals of Statistics*, 39:2607–2625.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York, revised edition.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70:209–226.

Cressie, N. and Lahiri, S. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, 45:217–233.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*. Wiley, New York, third edition.

Du, J., Zhang, H., and Mandrekar, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Annals of Statistics*, 37:3330–3361.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20:303–353.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, 32:407–499.

Fan, J. (1997). Comments on “Wavelets in statistics: A review" by A. Antoniadis. *Journal of the Italian Statistical Association*, 6:131–138.

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

Fan, J. and Peng, H. (2004). Nonconcave penlized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32:928–961.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:502–523.

Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, 99:399–423.

Hoeting, J., Davis, R., Merton, A., and Thompson, S. (2006). Model selection for geostatistical models. *Ecological Applications*, 16:87–98.

Hollander, A., F.W.Davis, and Stoms, D. (1994). *Hierarchical representations of species distribution using maps, images and sighting data. Page 71-90 in R.I Miller editor. Mapping the diversity of nature.* Chapman and Hall, London, UK.

Horn, R. and Johnson, C. (1991). *Topics in matrix analysis.* Cambridge University Press.

Huang, H.-C. and Chen, C.-S. (2007). Optimal geostatistical model selection. *Journal of the American Statistical Association*, 102:1009–1024.

Im, H. K., Stein, M. L., and Zhu, Z. (2007). Semiparametric estimation of spectral density with irregular observations. *Journal of the American Statistical Association*, 102(478):726–735.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103:1545–1555.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, New York.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 73:135–146.

Paul, D. and Peng, J. (2009). Consistency of restricted maximum likelihood estimators of principal components. *Annals of Statistics*, 37:1229–1271.

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18:995–1015.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer.

Reich, R. and Davis, R. (2008). *Lecture Notes of Quantitative Spatial Analysis.* Colorado State University, Fort Collins, Colorado.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis.* Chapman Hall, Boca Raton.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, New York.

Stein, M. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, 100:310–321.

Stewart, G. (1990). Stochastic perturbation theory. *SIAM Review*, 32:579–610.

Sun, S. and Lahiri, S. (2003). Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā: The Indian Journal of Statistics*, 68:130–166.

Sweeting, T. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics*, 8:1375–1381.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

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

Wang, H., Li, R., and Tsai, C.-L. (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568.

Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics*, 37:1–18.

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

Zhang, H. (2004). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99:250–261.

Zhang, H. and Wang, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics*, 21:290–304.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533.