DISSERTATION

STATISTICAL MODELING AND INFERENCE FOR SPATIAL AND SPATIO-TEMPORAL DATA

Submitted by

Jialuo Liu

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2019

Doctoral Committee:

Advisor: Haonan Wang

F. Jay Breidt
Piotr S. Kokoszka
Rockey J. Luo

ABSTRACT

STATISTICAL MODELING AND INFERENCE FOR SPATIAL
AND SPATIO-TEMPORAL DATA

Spatio-temporal processes with a continuous index in space and time are encountered in many scientific disciplines such as climatology, environmental sciences, and public health. A fundamental component for modeling such spatio-temporal processes is the covariance function, which is traditionally assumed to be stationary. While convenient, this stationarity assumption can be unrealistic in many situations. In the first part of this dissertation, we develop a new class of locally stationary spatio-temporal covariance functions. A novel spatio-temporal expanding distance (STED) asymptotic framework is proposed to study the properties of statistical inference. The STED asymptotic framework is established on a fixed spatio-temporal domain, aiming to characterize spatio-temporal processes that are globally nonstationary in a rescaled fixed domain and locally stationary in a distance expanding domain. The utility of STED is illustrated by establishing the asymptotic properties of the maximum likelihood estimation for a general class of spatio-temporal covariance functions, as well as a simulation study which suggests sound finite-sample properties.

Then, we address the problem of simultaneous estimation of the mean and covariance functions for continuously indexed spatio-temporal processes. A flexible spatio-temporal model with partially linear regression in the mean function and local stationarity in the covariance function is proposed. We study a profile likelihood method for estimation in the presence of spatio-temporally correlated errors. Specifically, for the nonparametric component, we employ a family of bimodal kernels to alleviate bias, which may be of independent interest for semiparametric spatial statistics. The theoretical properties of our profile likelihood estimation, including consistency and asymp-

totic normality, are established. A simulation study is conducted and corroborates our theoretical findings, while a health hazard data example further illustrates the methodology.

Maximum likelihood method for irregularly spaced spatial datasets is computationally intensive, as it involves the manipulation of sizable dense covariance matrices. Finding the exact likelihood is generally impractical, especially for large datasets. In the third part, we present an approximation to the Gaussian log-likelihood function using Krylov subspace methods. This method reduces the computational complexity from $O(N^3)$ operations to $O(N^2)$ for dense matrices and further to quasi-linear if matrices are sparse. Specifically, we implement the conjugate gradient method to solve linear systems iteratively and use Monte Carlo method and Gauss quadrature rule to obtain a stochastic estimator of the log-determinant. We give conditions to ensure consistency of the estimators. Simulation studies have been conducted to explore various important computational aspects including complexity, accuracy and efficiency. We also apply our proposed method to estimate the spatial structure of a big LiDAR dataset.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Haonan Wang, for his continuous encouragement, guidance and mentorship throughout my Ph.D. study. I would also like to thank: Dr. Jun Zhu from the Department of Statistics and Entomology at University of Wisconsin-Madison for her insightful advice, immense knowledge and limitless patience throughout our research collaboration; Dr Tingjin Chu from the Department of Mathematics and Statistics at University of Melbourne for being extremely supportive and helpful. It has been a great joy discussing research related questions with them.

I would also like to thank Dr. F. Jay Breidt, Dr. Piotr S. Kokoszka and Dr. Rockey J. Luo for being my committee members and for their constructive comments and valuable guidance. I would also like to thank my fellow graduate students, the faculty, and the staff in the Department of Statistics at Colorado State University. I am also grateful to the department for the financial support during my Ph.D. study.

Last but not least, I would like to thank my parents, Shecheng Liu and Liping Luo, for their unconditional love and support throughout the years.

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The field of spatial statistics is based on the assumption that "everything is related to everything else, but near things are more related than distant things", as stated in Tobler's First Law of Geography [Tobler, 1979]. This distinctive characteristic of spatial analysis violates the independence assumption, which is crucial in general statistics, and necessitates the development of theoretical, modeling and computational tools to model spatially-indexed dependence structures. This dissertation explores three separate but closely linked topics in spatial/spatio-temporal statistics: the asymptotic framework and random fields, semiparametric spatio-temporal modeling, and the analysis of massive spatial datasets.

The spatial data fall into three categories: geostatistical data, lattice data, or point patterns, depending on whether the spatial domain of interest is fixed and continuous, or fixed and discrete, or random [Cressie, 1993], respectively. This dissertation focuses on geostatistical models and provides a toolset including parameter estimation, interpolation and uncertainty quantification. Gaussian processes are commonly assumed for spatial/spatio-temporal modeling due to its analytical and computational tractability and flexibility [Gelfand and Schliep, 2016]. A stochastic process $Y(\boldsymbol{s})$ in the spatial domain of interest $\mathcal{R} \subset \mathbb{R}^d$ is Gaussian if any finite subset of the field locations $\{Y(\boldsymbol{s}_1,) \ldots, Y(\boldsymbol{s}_k))\}$ is multivariate normal. A Gaussian process is completely characterized by its mean and covariance and often times a stationary and isotropic covariance structure is assumed.

Recent advances in data acquisition and storage technologies have led to increasing accessibility of time-stamped spatial data. Following the same principle that "nearby" observations tend to be more alike than that are "far apart", spatio-temporal models seek to characterize processes that are spatially and temporally related simultaneously. One may consider spatio-temporal statistics as a generalization of spatial statistics by incorporating an additional time dimension and defining "nearby" in terms of both space and time. While this new dimension can promisingly improve data interpretability, the corresponding modeling procedures and model fitting tools are still in

their infancy. Recent years have witnessed the emerging literature on approaches to modeling spatio-temporal data with continuous spatial index but *discrete* temporal index [see, e.g., Huang et al., 2018, Lu et al., 2009, Stroud et al., 2001]. These approaches combine time series techniques for temporal data with geostatistical methods for spatial data and can capture nonlinearity and nonstationarity in space and/or time. For irregular time sampling, in contrast, spatio-temporal processes with continuous spatial index and *continuous* temporal index are more reasonable, yet methodologies are limited.

To derive asymptotic properties in spatial statistics, there are two mostly studied asymptotic regimes, the increasing-domain and fixed-domain asymptotics. In an increasing domain asymptotic framework, the spatial domain grows while the smallest distance among the spatial sampling locations are bounded away from zero [see, e.g., Mardia and Marshall, 1984, Cressie and Lahiri, 1993, Yao and Brockwell, 2006, Chu et al., 2011]. In an infill asymptotic framework, the spatial sampling locations get denser in a bounded spatial domain [see, e.g., Ying, 1993, Stein, 1999, Zhang, 2004, Loh, 2005]. However, asymptotic frameworks are underdeveloped for spatio-temporal data analysis, leaving the asymptotic properties of many spatio-temporal models unclear. One purpose of this dissertation is to fill the gap, to some extent, in spatio-temporal statistics by proposing a general asymptotic framework to assist the theoretical development of spatio-temporal models in a broad range of contexts.

Stationarity is a simplifying assumption that is often assumed for spatial/spatio-temporal random processes. A spatio-temporal random process $Y(\boldsymbol{s}, t)$ in the spatio-temporal domain of interest $\mathcal{R} \times \mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}^+$ is called *(weakly) stationary* if the mean $\mathrm{E}[Y(\boldsymbol{s}, t)]$ is a constant and the covariance $\mathrm{Cov}[Y(\boldsymbol{s}, t), Y(\boldsymbol{s}', t')]$ is a function of $(\boldsymbol{s} - \boldsymbol{s}', t - t')$ only. This stationarity assumption can be unrealistic in practice [Sampson and Guttorp, 1992]. For spatial data, nonstationary covariance functions including kernel-based spatial convolution and spectral-based local stationarity [see, e.g., Higdon, 1998, Fuentes, 2002, Paciorek and Schervish, 2006, Gelfand et al., 2010] have been developed. However, there is much less literature on nonstationary spatio-temporal processes, suggesting a clear demand for statistical methodology to model nonstationary spatio-temporal data.

In this dissertation, we propose a class of locally stationary spatio-temporal covariance functions that are close to stationary covariance functions locally at each location and time point but whose characteristics (e.g., covariance parameters) are changing in an unclear way as location and time change.

As for the mean structure, existing methods for spatio-temporal data analysis focus primarily on linear regression models [see, e.g., Datta et al., 2016b]. In this dissertation, we generalize these models to include a nonparametric component, which is known in the literature as *partially linear models* [Speckman, 1988, Härdle et al., 1998, Liang et al., 1999, Fan and Huang, 2005]. For independent data, profile likelihood estimation is found extremely useful for estimating partially linear models. However, the dependence structure in spatio-temporal data poses challenges for establishing the asymptotic properties of the profile likelihood estimate. While limited literature is available on semiparametric models for spatio-temporal data with continuous spatial index but *discrete* temporal index [see, e.g., Su and Jin, 2010, Sun et al., 2014, Gao et al., 2006, Lu et al., 2009], the partially linear models are underdeveloped for geostatistical models in continuous time. In this dissertation, we will establish the asymptotic properties of profile likelihood method for the partially linear model under a local stationarity condition that relaxes the stationarity assumption in traditional geostatistical models, given irregularly sampled spatio-temporal observations.

The aforementioned methodologies and models rely excessively on Gaussian processes, where likelihood-based methods are used to estimate unknown parameters and to interpolate at unobserved locations [Mardia and Marshall, 1984, Cressie and Lahiri, 1993, Chu et al., 2011]. However, the evaluation of the likelihood function is computationally cumbersome as it involves frequent inversion and determinant of large matrices. These calculations usually require $O(N^3)$ flops and $O(N^2)$ memory, where N is the number of observations, and quickly becomes infeasible as N grows. Thus, new approaches and algorithms are required to deal with very large spatial datasets. Krylov subspace methods, including the conjugate gradient method and the Lanczos algorithm, have been found very efficient in solving large sparse linear systems or dealing with large sparse

matrix eigenvalue problems. Therefore, in this dissertation, we will develop a computationally efficient method for analyzing large spatial data based on Krylov subspace methods.

The reminder of the dissertation is outlined as follows. In Chapter 2, we propose a novel spatio-temporal expanding distance (STED) asymptotic framework for studying the properties of statistical inference for spatio-temporal models and develop a new class of locally stationary spatio-temporal covariance functions. In Chapter 3, we propose a flexible spatio-temporal partially linear model and use a profile likelihood method for estimation. Chapter 4 gives a numeric approximation to the Gaussian log-likelihood function using Krylov subspace methods. Summary and future work are given in Chapter 5.

Chapter 2

Locally Stationary Spatio-temporal Processes and Inference[1]

2.1 Introduction

Spatio-temporal data are widely encountered and analyzed in many scientific disciplines, such as climatology [see, e.g., Cressie, 2018, Kuusela and Stein, 2018], environmental sciences [see, e.g., Liang et al., 2015, Porcu et al., 2018], and public health [see, e.g., Ludwig et al., 2017]. While there are a myriad of statistical modeling and methods for analyzing spatio-temporal data [see, e.g., Sherman, 2011, Cressie and Wikle, 2011], there appear to be limited tools for studying the theoretical properties of these statistical techniques. The purpose of this chapter is to fill some of this void in spatio-temporal statistics by proposing a novel asymptotic framework for data sampled in space and time.

For spatio-temporal data, spatio-temporal covariance functions have been proposed and employed to model spatio-temporal dependence. For example, Cressie and Huang [1999] and Gneiting [2002a] constructed fully parametric nonseparable spatio-temporal covariance functions using spectral density and completely monotone functions. Stein [2005] developed spatially isotropic but asymmetric spatio-temporal models by taking the derivatives of spatially isotropic fully symmetric models. These spatio-temporal covariance models assume stationarity in both space and time, which could be overly restrictive in practice. For time series data, various nonstationary models have been developed including locally stationary processes [see, e.g., Dahlhaus, 1997, Zhou and Wu, 2009, Vogt, 2012, Dahlhaus, 2012] and mixing conditions [see, e.g., Fan and Yao, 2003, Chang et al., 2015]. For spatial data, nonstationary covariance functions have also been developed, such as kernel-based spatial convolution and spectral-based local stationarity [see, e.g., Higdon, 1998,

[1]This chapter is based on a manuscript "Spatio-Temporal Expanding Distance Asymptotic Framework for Locally Stationary Processes" with Dr. Tingjin Chu, Dr. Jun Zhu and Dr. Haonan Wang.

Fuentes, 2002, Paciorek and Schervish, 2006, Gelfand et al., 2010]. More recently, Hsing et al. [2016] suggested a class of locally intrinsic stationary (LIS) covariance functions, which includes a variety of nonstationary models and has sound theoretical properties. However, there are very limited results on local stationarity for spatio-temporal processes. Nonstationary spatio-temporal processes have been considered, such as nonstationary models via spectral representation [Fuentes et al., 2008, Guinness and Fuentes, 2015] and a moving-window approach to estimating a locally stationary spatio-temporal Gaussian process [Kuusela and Stein, 2018], although the asymptotic properties of model estimation and inference are unexplored. We believe that advances are in need for studying the theoretical properties of locally stationary processes.

Asymptotic frameworks have played an important role in establishing the asymptotic properties of parameter estimates and their inference in spatial statistics. In an increasing domain asymptotic framework, the spatial domain expands while the smallest distance among the spatial sampling locations remains constant [see, e.g., Mardia and Marshall, 1984, Cressie and Lahiri, 1993, Yao and Brockwell, 2006, Chu et al., 2011]. In an infill asymptotic framework, the spatial sampling locations become denser in a fixed spatial domain [see, e.g., Ying, 1993, Stein, 1999, Zhang, 2004, Loh, 2005]. A mixed asymptotic framework has both an expanding spatial domain and denser sampling locations [see, e.g., Hall and Patil, 1994, Lahiri, 2003b, Lu and Tjøstheim, 2014]. However, asymptotic frameworks are underdeveloped for spatio-temporal statistics, leaving the asymptotic properties of many methods for nonstationary spatio-temporal data unclear. Yet, it is non-trivial to extend the existing asymptotic frameworks for spatial processes to spatio-temporal processes due to the uni-directionality of time and a lack of stationarity. For instance, in an increasing domain asymptotic framework, the "local" behavior of a covariance function is a challenge to study, whereas in an infill asymptotic framework, some of the parameters in the covariance function are not consistently estimable.

In light of these challenges, we propose a novel spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain. Let $\mathcal{R}$ denote a spatial domain of interest in $\mathbb{R}^d$ and $\mathcal{T}$ denote a temporal domain of interest in $\mathbb{R}$. Let n denote the stage of the

asymptotics while letting $\mathcal{R}_n$ and $\mathcal{T}_n$ denote the nth spatial and temporal domain, respectively, in the asymptotics, where $n \to \infty$. In the one-dimensional space ($d = 1$), the sampling locations are for example $i = 1, \ldots, n$ at the nth stage in an increasing domain asymptotic framework, but are $1/n, \ldots, (n-1)/n, 1$ in an infill asymptotic framework. In time series, a rescaled time i/n is generally assumed and not the actual time i [Fan and Yao, 2003], which we extend to a rescaled spatio-temporal domain such that both the spatial domain $\mathcal{R}_n$ and the temporal domain $\mathcal{T}_n$ are bounded.

Besides asymptotic framework, the asymptotic properties of parameter estimates also depends on the nature of spatio-temporal dependence. Here, we take the LIS framework for spatial processes as impetus [Hsing et al., 2016], and develop a class of locally stationary spatio-temporal covariance functions that vary across space and over time within the rescaled fixed spatio-temporal domain. The resulting spatio-temporal covariance functions are locally stationary in a distance expanding spatio-temporal domain; that is, they can be approximated locally by stationary covariance functions of actual distances in the spatio-temporal domain without rescaling. Such a class of spatio-temporal covariance functions is quite general and flexible as we will demonstrate. Furthermore, our proposed STED asymptotic framework is not a generalization of the mixed asymptotic framework in spatial statistics, but rather a potentially useful tool for studying the properties of statistical inference for spatio-temporal processes that are globally nonstationary in a rescaled fixed domain and locally stationary in a distance expanding domain.

The remainder of the chapter is organized as follows. We develop a class of locally stationary spatio-temporal processes in Section 2.2 and propose the STED asymptotic framework for fixed spatio-temporal domain in Section 2.3. For illustration, we consider spatio-temporal models and the theoretical properties of the corresponding maximum likelihood estimation in Section 2.4. Section 2.5 gives the technical details including theorem proofs and remarks. A simulation study is provided in Section 2.6 and generalization to a linear regression model is given in Section 2.7.

2.2 Local Stationarity in Space and Time

For the spatial domain of interest $\mathcal{R} \subset \mathbb{R}^d$ and the temporal domain of interest $\mathcal{T} \subset \mathbb{R}$, we consider a zero-mean spatio-temporal random process $\{Y(\underline{s}, t) : \underline{s} \in \mathcal{R}, t \in \mathcal{T}\}$ with a covariance function $\mathrm{Cov}(Y(\underline{s}, t), Y(\underline{s}', t')) = \gamma((\underline{s}, t), (\underline{s}', t'))$, where $\underline{s}, \underline{s}' \in \mathcal{R}$ and $t, t' \in \mathcal{T}$. To draw inference for the covariance functions, stationarity is generally assumed, which can be restrictive and may not hold in practice. Here we consider a new class of spatio-temporal covariance functions, allowing more flexibility than stationary processes to the extent of local stationarity.

We let the stage of the spatio-temporal asymptotics, n, appear as either a left superscript or a right subscript of a quantity that depends on n. We also let $\{A_n\}$ and $\{B_n\}$ denote two sequences of positive numbers and let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^d$. The following are conditions for defining locally stationary spatio-temporal covariance functions.

(LS.1) There exists a sequence of functions $g_n(\cdot, \cdot, \underline{s}, t)$ such that

$$|\gamma_n((\underline{s}, t), (\underline{s}', t')) - g_n(\underline{s}' - \underline{s}, t' - t, \underline{s}, t)| = \mathcal{O}(\|\underline{s}' - \underline{s}\| + |t' - t| + \rho_n)$$

uniformly for all $(\underline{s}, t), (\underline{s}', t') \in \mathcal{R} \times \mathcal{T}$, where $\{\rho_n\}$ is a sequence of positive numbers such that $\rho_n \to 0$ as $n \to \infty$. In addition, there exists a function g such that

$$\lim_{n\to\infty} |g_n(\underline{s}' - \underline{s}, t' - t, \underline{s}, t) - g(\underline{u}_1, u_2, \underline{s}, t)| \to 0, \text{ as } n \to \infty,$$

where $\underline{u}_1 = A_n(\underline{s}' - \underline{s})$ and $u_2 = B_n(t' - t)$.

(LS.2) Define $g(\underline{s}, t) = g(\underline{0}, 0, \underline{s}, t)$ with $g(\underline{u}_1, u_2, \underline{s}, t)$ given in (LS.1), and $g(\underline{s}, t)$ satisfies $|g(\underline{s}, t) - g(\underline{s}', t')| \leq C_1\|\underline{s} - \underline{s}'\| + C_2|t - t'|$ for all $(\underline{s}, t), (\underline{s}', t') \in \mathcal{R} \times \mathcal{T}$, where $C_1, C_2 > 0$ are constants.

(LS.3) There are two positive nonincreasing functions γ_0 and γ_1 satisfying $\int_0^\infty u^{d-1}\gamma_0(u)du < \infty$ and $\int_0^\infty \gamma_1(u)du < \infty$ such that $|\gamma_n((\underline{s}, t), (\underline{s} + \underline{u}_1/A_n, t + u_2/B_n))| \leq \gamma_0(\|\underline{u}_1\|)$

$\gamma_1(|u_2|)$ for all n and $\|\underline{u}_1\|, |u_2| \in [0, \infty)$ such that $(\underline{s}, t), (\underline{s} + \underline{u}_1/A_n, t + u_2/B_n) \in \mathcal{R} \times \mathcal{T}$.

Here, (LS.1)–(LS.2) can be viewed as a generalization of (W3)–(W4) for spatial processes in Hsing et al. [2016] to our spatio-temporal processes. In particular, (LS.1) describes *local stationarity* in the sense that the covariance function γ_n can be approximated by a function g_n, which is allowed to vary with location $\underline{s}$ and time t, rather than merely determined by spatial and temporal lags. Such approximation is adequate in a neighborhood of $(\underline{s}, t)$, provided that $(\underline{s}, t)$ and $(\underline{s}', t')$ are sufficiently close. Furthermore, the function g characterizes the limiting behavior of g_n with respect to the scaled spatial and temporal lags at the rates of A_n and B_n, respectively. (LS.2) imposes some mild restrictions on the covariance structure at the zero lag in space and time, whereas (LS.3) is a constraint on the decay rate of the covariance function in space and time.

The above definition of locally stationary covariance functions is satisfied by a variety of covariance functions. For illustration, we introduce a class of parametric covariance functions denoted as $\gamma_n((\underline{s}, t), (\underline{s}', t'); \underline{\theta})$, where $\underline{\theta}$ is a $q \times 1$ vector of parameters, which includes the following generalized spatio-temporal Matérn covariance function

$$\gamma_n((\underline{s}, t), (\underline{s}', t'); \underline{\theta}) = \begin{cases} \frac{D(\underline{s},t)D(\underline{s}',t')\sigma^2\theta_3^{d/2}2^{1-\nu}}{(\theta_1^2u_2^2+1)^\nu(\theta_1^2u_2^2+\theta_3)^{d/2}\Gamma(\nu)} m(\underline{u}_1, u_2)^\nu K_\nu \left\{m(\underline{u}_1, u_2)\right\}, & \text{if } \|\underline{u}_1\| > 0, \\ \frac{D(\underline{s},t)D(\underline{s}',t')\sigma^2\theta_3^{d/2}}{(\theta_1^2u_2^2+1)^\nu(\theta_1^2u_2^2+\theta_3)^{d/2}}, & \text{if } \|\underline{u}_1\| = 0, |u_2| > 0, \\ D(\underline{s}, t)^2\sigma^2 + \tau^2, & \text{if } \|\underline{u}_1\| = 0, |u_2| = 0, \end{cases} \tag{2.1}$$

where $\underline{\theta} = (\theta_1, \theta_2, \theta_3, \sigma^2, \tau^2)^\top$ is a vector of spatio-temporal parameters with a scaling parameter in time $\theta_1 > 0$, a scaling parameter in space $\theta_2 > 0$, a separability parameter $\theta_3 > 0$, and are two variance components σ^2 and τ^2. In addition, $\underline{u}_1 = \varrho_{1,n}(\underline{s}' - \underline{s})$ is the spatial lag scaled to the spatially expanding domain, and $u_2 = \varrho_{2,n}(t' - t)$ is the temporal lag scaled to the temporally expanding domain, where $\varrho_{1,n}$ and $\varrho_{2,n}$ are two sequences of positive real numbers. Further, $m(\underline{u}_1, u_2) = \theta_2 \left(\frac{\theta_1^2u_2^2+1}{\theta_1^2u_2^2+\theta_3}\right)^{1/2} \|\underline{u}_1\|$, $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, $\nu > 0$ is a smoothness parameter assumed to be known, and $D(\underline{s}, t)$ is a positive spatio-

temporal function such that $D(\underline{0}, 0) = 1$ and $D(\underline{s}, t)^2\sigma^2 + \tau^2$ is the variance of $Y(\underline{s}, t)$. By Cressie and Huang [1999], it can be shown that (2.1) is a positive definite function and therefore, a valid covariance function. The class of spatio-temporal covariance functions (2.1) is generally nonseparable and nonstationary. In the special case of $D(\underline{s}, t) \equiv 1$ for all $\underline{s} \in \mathcal{R}$ and $t \in \mathcal{T}$, (2.1) reduces to a class of stationary, but still nonseparable, spatio-temporal covariance functions, which was introduced by Cressie and Huang [1999]. Furthermore, (2.1) is separable only when $\theta_3 = 1$ and $D(\underline{s}, t)$ is separable for all $\underline{s} \in \mathcal{R}$ and $t \in \mathcal{T}$.

Let $\gamma_{n,k}(\cdot, \cdot; \underline{\theta}) = \partial\gamma_n(\cdot, \cdot; \underline{\theta})/\partial\theta_k$ and $\gamma_{n,kk'}(\cdot, \cdot; \underline{\theta}) = \partial^2\gamma_n(\cdot, \cdot; \underline{\theta})/\partial\theta_k\partial\theta_{k'}$ denote the first- and second-order partial derivatives of $\gamma_n(\cdot, \cdot; \underline{\theta})$, respectively, with respect to θ_k and $\theta_{k'}$ for $1 \leq k, k' \leq q$. We consider the following additional conditions for developing the locally stationary parametric spatio-temporal covariance functions.

(LS.4) The covariance function $\gamma_n(\cdot, \cdot; \underline{\theta})$ is bounded and is twice continuously differentiable with respect to $\underline{\theta}$ in an open set.

(LS.5) There exist two positive nonincreasing functions γ_2 and γ_3 with $\int_0^\infty u^{d-1}\gamma_2(u)du < \infty$ and $\int_0^\infty \gamma_3(u)du < \infty$ such that $\max\{|\gamma_{n,k}((\underline{s}, t), (\underline{s}+\underline{u}_1/A_n, t+u_2/B_n))|, |\gamma_{n,kk'}((\underline{s}, t), (\underline{s} + \underline{u}_1/A_n, t + u_2/B_n))|\} \leq \gamma_2(\|\underline{u}_1\|)\gamma_3(|u_2|)$ for all n and $\|\underline{u}_1\|, |u_2| \in [0, \infty)$ with $(\underline{s}, t), (\underline{s} + \underline{u}_1/A_n, t + u_2/B_n) \in \mathcal{R} \times \mathcal{T}$ and $1 \leq k, k' \leq q$.

In the above, (LS.4) is a standard assumption to ensure the smoothness of the covariance function, whereas (LS.5) restricts the decay rates of the first- and second-order partial derivatives of the covariance function with respect to covariance parameters by spatial lag and temporal lag.

With $A_n = \varrho_{1,n}$ and $B_n = \varrho_{2,n}$ (up to some constant scale), we establish that the generalized spatio-temporal Matérn covariance function (2.1) satisfies (LS.1)–(LS.5).

Proposition 1. *Let $D(\underline{s}, t)$ be some positive known function with $D(\underline{0}, 0) = 1$ and $|D(\underline{s}, t) - D(\underline{s}', t')| \leq \widetilde{C}_1\|\underline{s} - \underline{s}'\| + \widetilde{C}_2|t - t'|$ for all $(\underline{s}, t), (\underline{s}', t') \in \mathcal{R} \times \mathcal{T}$, where $\widetilde{C}_1, \widetilde{C}_2 > 0$ are constants. Then the generalized spatio-temporal Matérn covariance function* (2.1) *satisfies conditions (LS.1)–(LS.5).*

The proof of Proposition 1 is given in Section 2.5.1. In general, if $D(\underline{s}, t)$ is parametric and twice continuously differentiable with respect to the parameters, we can incorporate those parameters into $\underline{\theta}$ and Proposition 1 will still hold. In addition, it can be shown that a class of generalized exponential covariance functions satisfies (LS.1)–(LS.5) (see details in Section 2.5.2).

2.3 Spatio-temporal Expanding Distance Asymptotic Framework

We now develop a novel spatio-temporal asymptotic framework under which the asymptotic properties of statistical inference can be investigated. We consider a fixed spatial domain with continuous spatial indexes and a fixed temporal domain with continuous temporal indexes. Without loss of generality, we assume that the spatial domain is $\mathcal{R} = [0,1]^d$ and the temporal domain is $\mathcal{T} = [0,1]$ at all stages. We further assume that at stage n, N_n spatio-temporal sampling points are observed at $({}^n\underline{s}_1, {}^nt_1), \ldots, ({}^n\underline{s}_{N_n}, {}^nt_{N_n})$, where N_n tends to infinity as $n \to \infty$. For ease of notation, henceforth we suppress n in the left superscript of $({}^n\underline{s}_i, {}^nt_i)$.

We denote the smallest distance between the jth sampling points and the other sampling points in space and time as $\delta_{j,n} = \min\{\|\underline{s}_i - \underline{s}_j\| : 1 \le i \le N_n, \underline{s}_i \ne \underline{s}_j\}$ and $\zeta_{j,n} = \min\{|t_i - t_j| : 1 \le i \le N_n, t_i \ne t_j\}$, respectively. Let $\delta_n = \max_{1\le j\le N_n} \delta_{j,n}$ and $\zeta_n = \max_{1\le j\le N_n} \zeta_{j,n}$ denote the maximum smallest distance in space and in time, respectively. We assume that, for all n,

(A.1) $\delta_n / \min_{1\le j\le N_n} \delta_{j,n} \le c_1$,

(A.2) $\zeta_n / \min_{1\le j\le N_n} \zeta_{j,n} \le c_2$,

(A.3) $\delta_n^d A_n^d \zeta_n B_n \ge c_3$,

where c_1, c_2 and c_3 are some positive constants independent of n. Here, (A.1)–(A.2) ensure bounded mesh ratios in both the spatial and temporal domain, whereas (A.3) provides a lower bound on the minimal distances among sampling points and their nearest neighbor in space and time.

Let $\widetilde{\mathcal{R}} = A_n\mathcal{R}$ and $\widetilde{\mathcal{T}} = B_n\mathcal{T}$ denote the expanding spatial and temporal domain. For $\widetilde{\mathcal{R}}$ and $\widetilde{\mathcal{T}}$, (A.3) includes the following three sampling patterns. First, the case of $\delta_n^d A_n^d = \mathcal{O}(1)$ and $\zeta_n B_n = \mathcal{O}(1)$ corresponds to an increasing domain asymptotics in both space and time. Next, if $\delta_n^d A_n^d \to 0$ and $\zeta_n B_n \to \infty$, we have an increasing domain in time and a mixed asymptotic framework in space; that is, both the spatial domain of interest and the sampling intensity tend to infinity. Similarly, if $\delta_n^d A_n^d \to \infty$ and $\zeta_n B_n \to 0$, we have an increasing domain in space and a mixed asymptotic framework in time. We will refer to (A.1)–(A.3) as an (A_n, B_n)-rate spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain.

Together with (A.1) and (A.2), the cardinality of a neighborhood of any sampling point is decided by the minimal distances δ_n and ζ_n as well as A_n and B_n. Consequently, the sampling design is such that δ_n and ζ_n cannot decrease too fast or too slowly. That is, the number of observations in the neighborhood cannot be too few with insufficient information for parameter estimation or too many with too much redundant information. Thus, the local stationarity in space and time given in Section 2.2 is closely connected to the spatio-temporal sampling points in the STED asymptotic framework developed here.

2.4 Illustration of Theoretical Development

Consider the following spatio-temporal model,

$$y(\underline{s}, t) = \varepsilon_1(\underline{s}, t) + \varepsilon_2(\underline{s}, t), \;\; \underline{s} \in \mathcal{R}, \; t \in \mathcal{T}, \tag{2.2}$$

where $\varepsilon_1(\underline{s}, t)$ is a Gaussian spatio-temporal error process and $\varepsilon_2(\underline{s}, t)$'s are independent and identically distributed Gaussian errors with mean 0 and variance τ^2, independent of $\varepsilon_1(\underline{s}, t)$. The spatio-temporal covariance function of $y(\underline{s}, t)$ is denoted as $\gamma((\underline{s}, t), (\underline{s}', t'); \underline{\theta})$ for $(\underline{s}, t), (\underline{s}', t') \in \mathcal{R} \times \mathcal{T}$, where $\underline{\theta}$ is a $q \times 1$ vector of unknown parameters. Recall that the data are observed at N_n points $(\underline{s}_1, t_1), \ldots, (\underline{s}_{N_n}, t_{N_n})$ sampled under the STED asymptotic framework (A.1)–(A.3) in Section 2.3. Let $\underline{y} = (y(\underline{s}_1, t_1), \ldots, y(\underline{s}_{N_n}, t_{N_n}))^\top$ denote an $N_n \times 1$ vector of the response variables and let

${}^n\Gamma(\underline{\theta}) = [\gamma_n((\underline{s}_i, t_i), (\underline{s}_j, t_j); \underline{\theta})]_{i,j=1}^{N_n}$ denote an $N_n \times N_n$ covariance matrix of $\underline{y}$. For ease of notation, we omit the stage n in the left superscript of ${}^n\Gamma$. Therefore, the log-likelihood function of $\underline{\theta}$ under (2.2) is

$$\ell(\underline{\theta}) = -(N_n/2)\log(2\pi) - (1/2)\log\{\det\Gamma(\underline{\theta})\} - (1/2)\underline{y}^\top\Gamma(\underline{\theta})^{-1}\underline{y}. \tag{2.3}$$

Denote the maximizer of (2.6) as $\widehat{\underline{\theta}}_{\mathrm{MLE}}$. Next, we will establish the asymptotic properties of $\widehat{\underline{\theta}}_{\mathrm{MLE}}$ as an illustration of the STED asymptotic framework (A.1)–(A.3) under local spatio-temporal stationarity (LS.1)–(LS.5) defined in Section 2.2. The following regularity conditions are assumed.

(C.1) There exists a continuous density function $q_{\underline{s}}$ on $\mathcal{R}$ such that for any measurable set $A \subset \mathcal{R}$, $N_n^{-1}\sum_{i=1}^{N_n} 1(\underline{s}_i \in A) \to \int_A q_{\underline{s}}(\underline{s})d\underline{s}$ uniformly, as $n \to \infty$, where $1(\cdot)$ is an indicator function.

(C.2) There exists a nonnegative function $Q(t)$ with $Q(0) = 0$ and $Q(1) = 1$ such that $\sup_{t\in[0,1]}|Q_{N_n}(t) - Q(t)| = \mathcal{O}(\zeta_n)$, where $Q_{N_n}(t) = N_n^{-1}\sum_{i=1}^{N_n} 1(t_i \le t)$. In addition, $Q(t)$ has a positive continuous first-order derivative $q(t)$, which is bounded away from zero and infinity and is twice continuously differentiable with bounded derivatives.

(C.3) For some $\iota > 0$, there exist positive constants D_k such that $\|\Gamma_k\|_F^{-2} \le D_k N_n^{-1/2-\iota}$ for $k = 1, \ldots, q$.

(C.4) There exists a constant C^*, such that $\|\Gamma^{-1}\|_2 < C^* < \infty$.

(C.5) Let $t_{kk'} = \mathrm{tr}(\Gamma^{-1}\Gamma_k\Gamma^{-1}\Gamma_{k'})$, for $k, k' = 1, \ldots, q$. For sufficiently large n, $\underline{A}_n = (a_{kk'})_{k,k'=1}^q$ is nonsingular, where $a_{kk'} = \{t_{kk'}(t_{kk}t_{k'k'})^{-1/2}\}$.

(C.6) There exists a non-singular matrix $\mathcal{I}(\underline{\theta})$ which satisfies $N_n^{-1}\mathcal{J}_n(\underline{\theta}) \longrightarrow \mathcal{I}(\underline{\theta})$, as $n \to \infty$, where $\mathcal{J}_n(\underline{\theta}) = E\left\{-\frac{\partial^2\ell(\underline{\theta})}{\partial\underline{\theta}\partial\underline{\theta}^T}\right\}$.

Here, (C.1) is a condition on the spatial sampling design following Assumption (II)(i) of Lu and Tjøstheim [2014], whereas (C.2) is about the fixed sampling time points. (C.3) imposes a lower

bound on the first-order partial derivatives of the covariance matrix and (C.4) is a constraint on the smallest eigenvalue of the covariance matrix. Both (C.3) and (C.4) are requirements on the covariance function as well as the sampling design. (C.5) ensures nonsingularity in the limit and the elements of $\widehat{\underline{\theta}}$ are not asymptotically linearly dependent. (D.3) is a standard condition for information matrix, and together with (C.1)–(C.5), they yield a central limit theorem of $\frac{\partial \ell(\underline{\theta})}{\partial \underline{\theta}}$ [Chu et al., 2011]. Additionally in Section 2.5.3, we provide sufficient conditions for (C.3) and show that the generalized spatio-temporal Matérn and exponential covariance functions satisfy (C.3) under a proper sampling design.

Let $\xrightarrow{p}$ and $\xrightarrow{D}$ denote convergence in probability and in distribution, respectively, as $n \to \infty$. We first establish a result about the spatio-temporal sampling design, which is fundamental for establishing the asymptotic properties of locally stationary processes. The proof of Theorem 1 is given in Section 2.5.4.

Theorem 1. *Under (A.1)–(A.3), (LS.1)–(LS.5), and (C.1)–(C.5), we have,*

$$N_n^{-1/2} \frac{\partial \ell(\underline{\theta})}{\partial \underline{\theta}}\bigg|_{\underline{\theta}=\underline{\theta}_0} \xrightarrow{D} N(0, \mathcal{I}(\underline{\theta}_0)) \quad \textit{and} \quad N_n^{-1} \frac{\partial^2 \ell(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}^\top}\bigg|_{\underline{\theta}=\underline{\theta}_0} \xrightarrow{p} \mathcal{I}(\underline{\theta}_0).$$

Theorem 1 establishes a central limit theorem of $\frac{\partial \ell(\underline{\theta})}{\partial \underline{\theta}}$, and the convergence of $\frac{\partial^2 \ell(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}^\top}$ at the true value $\underline{\theta}_0$ for local spatio-temporal stationary covariance functions. By Theorem 1 above, Theorem 1 of Sweeting [1980], and Theorem 2 of Mardia and Marshall [1984], we have the following asymptotic results.

Theorem 2. *Under (A.1)–(A.3), (LS.1)–(LS.5), and (C.1)–(D.3), there exists, with probability tending to one, a local maximizer ${}^n\widehat{\underline{\theta}}$ of $\ell(\underline{\theta})$ such that $\|{}^n\widehat{\underline{\theta}} - \underline{\theta}_0\| = \mathcal{O}_p(N_n^{-1/2})$. Moreover, the local maximizer ${}^n\widehat{\underline{\theta}}$ is asymptotic normal; as $n \to \infty$, $N_n^{1/2}({}^n\widehat{\underline{\theta}} - \underline{\theta}_0) \xrightarrow{D} N(\underline{0}, \mathcal{I}(\underline{\theta}_0)^{-1})$.*

Theorem 2 established consistency and asymptotic normality under the proposed spatio-temporal framework, and can be used as a guideline of parameter estimation for spatio-temporal datasets. Although spatio-temporal data can be viewed as extending spatial data by adding a temporal domain, the extension is usually not straightforward and a careful examination is often needed. For

asymptotic framework, it is well known that there are three asymptotic frameworks in spatial statistics: increasing domain asymptotics, infill domain asymptotics, and mixed asymptotics. However, this division no longer works for spatio-temporal data, as mentioned in Section 2.3. In fact, if the proposed spatio-temporal framework is projected to the space domain, it can be any of the above three frameworks. To better illustrate the above point, the following example is provided.

Example 3. *For two integers a, b, and let $\lfloor a/b \rfloor$ and $\langle a/b \rangle$ denote the quotient and reminder of a divided by b. For the nth stage, let $\underline{\bar{s}}(i_1, i_2) = \left(\frac{i_1}{p_n+1}, \frac{j_2}{p_n+1}\right)$ and $\bar{t}(i_3) = \frac{i_3}{q_n+1}$, for $i_1, i_2 = 1, \ldots, p_n$, and $i_3 = 1, \ldots, q_n$. The ith spatial and temporal observation is $(\underline{s}_i, t_i) = (\underline{\bar{s}}(i_1, i_2), \bar{t}(i_3))$, where $i_1 = \langle \langle i/p_n^2 \rangle / p_n \rangle$, $i_2 = \lfloor \langle i/p_n^2 \rangle \rfloor + 1$, and $i_3 = \lfloor i/p_n^2 \rfloor + 1$, for $i = 1, \ldots, p_n^2 q_n$. Under this framework, we have $\delta_{j,n} = 1/(p_n + 1)$, $\zeta_{j,n} = 1/(q_n + 1)$, $\delta_n = 1/(p_n + 1)$, $\zeta_n = 1/(q_n + 1)$, and therefore, (A.1) and (A.2) hold. Moreover, since $p_n^2 q_n = N_n$ and $A_n^2 B_n$ is at the rate of N_n, (A.3) also holds.*

Next, we show the above spatio-temporal framework can be any of the three spatial asymptotics framework, when it is projected to the spatial domain. First, it can be calculated that the density of the projected spatial locations is p_n^2/A_n^2. In this example, we consider the case that q_n is bounded and p_n increases at the rate of $N_n^{1/2}$. If A_n is bounded and B_n increases at the rate of N_n, the resulting density is N_n and the framework is the infill domain framework. If A_n increases at the rate of $N_n^{1/2}$ and B_n is bounded, the resulting framework is the increasing domain. If A_n increases at the rate of N_n^{α} with $\alpha \in (0, 1/2)$, the resulting framework is the mixed asymptotics.

It is known that some estimates of the covariance function are not consistent under the infill asymptotics. Theorem 2 suggests that even for these datasets, when the temporal dimension is added following the proposed spatio-temporal framework, the consistency and asymptotic normality of parameter estimates can be achieved. On the other hand, if the temporal correlation is ignored, and the spatio-temporal data are treated as if they were spatial data, the resulting estimates of parameter can be inconsistent. Besides the spatio-temporal framework, it is also important to specify the proper spatio-temporal covariance functions. Theorem 2 ensures that for spatio-temporal covariance functions satisfying (LS.1)–(LS.5), consistency and asymptotic normality of

the parameter estimates are ensured, which include locally stationary covariance functions, as well as stationary covariance functions [Cressie and Huang, 1999, Gneiting, 2002a]. In particular, by Proposition 1, Theorem 2 holds for the generalized spatio-temporal Matérn covariance function. In Section 2.6, we provide a simulation study that suggests sound finite-sample properties.

2.5 Technical Details

2.5.1 Proof of Proposition 1

Proof. First, it can be seen that the generalized spatio-temporal Matérn covariance function in (2.1) is bounded and twice continuously differentiable with respect to $\underline{\theta}$; thus, (LS.4) is satisfied. Next, we will show that the generalized spatio-temporal Matérn covariance function satisfies (LS.1) and (LS.2).

For any $\underline{u}_1$ and u_2, define

$$h(\underline{u}_1, u_2) = \begin{cases} \frac{\theta_3^{d/2} 2^{1-\nu}}{(\theta_1^2 u_2^2+1)^{\nu}(\theta_1^2 u_2^2+\theta_3)^{d/2}\Gamma(\nu)} m(\underline{u}_1, u_2)^{\nu} K_{\nu}\left\{m(\underline{u}_1, u_2)\right\}, & \text{if } \|\underline{u}_1\| > 0, \\ \frac{\theta_3^{d/2}}{(\theta_1^2 u_2^2+1)^{\nu}(\theta_1^2 u_2^2+\theta_3)^{d/2}}, & \text{if } \|\underline{u}_1\| = 0. \end{cases}$$

For any $\underline{s}$ and t, let

$$\begin{aligned} &g_n(\underline{s}' - \underline{s}, t' - t, \underline{s}, t) = g(\underline{u}_1, u_2, \underline{s}, t) \\ &= \begin{cases} D(\underline{s}, t)^2 \sigma^2 h(\underline{u}_1, u_2), & \text{if } \|\underline{u}_1\| > 0 \text{ or } |u_2| > 0, \\ D(\underline{s}, t)^2 \sigma^2 + \tau^2, & \text{otherwise.} \end{cases} \end{aligned}$$

Then, $\gamma_n((\underline{s}, t), (\underline{s}, t)) = g(\underline{0}, 0, \underline{s}, t)$. For all $(\underline{s}, t), (\underline{s} + \underline{u}_1/\varrho_{1,n}, t + u_2/\varrho_{2,n}) \in \mathcal{R} \times \mathcal{T}$ with $\|\underline{u}_1\| > 0$ or $|u_2| > 0$, we have

$$\begin{aligned} &|\gamma_n((\underline{s}, t), (\underline{s}', t')) - g(\underline{u}_1, u_2, \underline{s}, t)| = D(\underline{s}, t) h(\underline{u}_1, u_2) \sigma^2 |D(\underline{s}', t') - D(\underline{s}, t)| \\ \leq & D(\underline{s}, t) h(\underline{u}_1, u_2) \sigma^2 (\widetilde{C}_1 \|\underline{s} - \underline{s}'\| + \widetilde{C}_2 |t - t'|) = \mathcal{O}(\|\underline{s} - \underline{s}'\| + |t - t'|) \end{aligned}$$

uniformly since $D(\underline{s},t)$ is bounded on $\mathcal{R}\times\mathcal{T}$ and $|h(\underline{u}_1,u_2)|\leq 1$. Thus, (LS.1) is satisfied.

For $g(\underline{s},t)$ defined in (LS.2), we have $g(\underline{s},t) = g(\underline{0},0,\underline{s},t) = D(\underline{s},t)^2\sigma^2+\tau^2$. Note that $\underline{s}' = \underline{s}+\underline{u}_1/\varrho_{1,n}$ and $t' = t+u_2/\varrho_{2,n}$, and we have $|g(\underline{s},t)-g(\underline{s}',t')| = |D(\underline{s},t)^2 - D(\underline{s}',t')^2|\sigma^2 = |D(\underline{s},t)+D(\underline{s}',t')||D(\underline{s},t)-D(\underline{s}',t')|\sigma^2 \leq |D(\underline{s},t)+D(\underline{s}',t')|(\widetilde{C}_1\|\underline{s}-\underline{s}'\|+\widetilde{C}_2|t-t'|)\sigma^2$. Thus, (LS.2) holds by adjusting the constants.

Further, we will show that the generalized spatio-temporal Matérn covariance function (2.1) satisfies (LS.3) and (LS.5). For all $(\underline{s},t),(\underline{s}',t')\in\mathcal{R}\times\mathcal{T}$, we have

$$\gamma_n((\underline{s},t),(\underline{s}',t')) \leq (\max D(\underline{s},t))^2(\sigma^2+\tau^2)h(\underline{u}_1,u_2).$$

Thus, to show (LS.3), it suffices to find $\gamma_0(\|\underline{u}_1\|)$ and $\gamma_1(|u_2|)$ to bound $h(\underline{u}_1,u_2)$. Moreover, straightforward calculation yields that all first- and second-order partial derivatives of γ_n, denoted by $\gamma_{n,k}$ and $\gamma_{n,kk'}$, can be bounded by the partial derivatives of $h(\underline{u}_1,u_2)$ up to some constant scales, which enables us to obtain $\gamma_2(\|\underline{u}_1\|)$ and $\gamma_3(|u_2|)$ in (LS.5).

In addition, we have

$$\begin{aligned}
\frac{\partial h(\underline{u}_1,u_2)}{\partial\theta_1} &= \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\partial h(0,u_2)}{\partial\theta_1}m^\nu K_\nu(m) - h(0,u_2)m^\nu K_{\nu-1}(m)\frac{\partial m}{\partial\theta_1}\right),\\
\frac{\partial h(\underline{u}_1,u_2)}{\partial\theta_2} &= -\frac{2^{1-\nu}}{\Gamma(\nu)}h(0,u_2)m^\nu K_{\nu-1}(m)\frac{\partial m}{\partial\theta_2},\\
\frac{\partial h(\underline{u}_1,u_2)}{\partial\theta_3} &= \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\partial h(0,u_2)}{\partial\theta_3}m^\nu K_\nu(m) - h(0,u_2)m^\nu K_{\nu-1}(m)\frac{\partial m}{\partial\theta_3}\right),
\end{aligned}$$

where $m = m(\underline{u}_1,u_2)$ and

$$\begin{aligned}
\frac{\partial m(\underline{u}_1,u_2)}{\partial\theta_1} &= \frac{m(\underline{u}_1,u_2)\theta_1u_2^2(\theta_3-1)}{(\theta_1^2u_2^2+\theta_3)(\theta_1^2u_2^2+1)},\\
\frac{\partial m(\underline{u}_1,u_2)}{\partial\theta_2} &= \frac{m(\underline{u}_1,u_2)}{\theta_2}, \frac{\partial m(\underline{u}_1,u_2)}{\partial\theta_3} = -\frac{m(\underline{u}_1,u_2)}{2(\theta_1^2u_2^2+\theta_3)},\\
\frac{\partial h(0,u_2)}{\partial\theta_1} &= -\frac{\theta_1u_2^2(2\nu(\theta_1^2u_2^2+\theta_3)+d(\theta_1^2u_2^2+1))}{(\theta_1^2u_2^2+1)(\theta_1^2u_2^2+\theta_3)}h(0,u_2),\\
\frac{\partial h(0,u_2)}{\partial\theta_3} &= \frac{d\theta_3^{d/2-1}\theta_1^2u_2^2}{2(\theta_1^2u_2^2+1)^\nu(\theta_1^2u_2^2+\theta_3)^{d/2+1}} = \frac{d\theta_1^2u_2^2}{2\theta_3(\theta_1^2u_2^2+\theta_3)}h(0,u_2).
\end{aligned}$$

For (LS.3), it can be seen that $m(\underline{u}_1, u_2) \leq \max\left\{\theta_2\theta_3^{-1/2}, \theta_2\right\}\|\underline{u}_1\|$. Thus, we have

$$h(\underline{u}_1, u_2) \leq \frac{\theta_3^{d/2}2^{1-\nu}\widetilde{m}(\underline{u}_1)^\nu K_\nu\left\{\widetilde{m}(\underline{u}_1)\right\}}{(\theta_1^2u_2^2+1)^\nu(\theta_1^2u_2^2+\theta_3)^{d/2}\Gamma(\nu)} \leq 1,$$

where $\widetilde{m}(\underline{u}_1) = \max\left\{\theta_2\theta_3^{-1/2}, \theta_2\right\}\|\underline{u}_1\|$. We can see that, up to some constant scales,

$$h(\underline{u}_1, u_2) \leq \left(\widetilde{m}(\underline{u}_1)^\nu K_\nu\left\{\widetilde{m}(\underline{u}_1)\right\}\right)\left(|u_2|^{-2\nu-d}\right) \equiv \gamma_0(\|\underline{u}_1\|)\gamma_1(|u_2|)$$

Here, $\gamma_0(\|\underline{u}_1\|)$ is a linear combination of a polynomial of $\|\underline{u}_1\|$ with degree ν and a modified Bessel function of the second kind and $\gamma_1(|u_2|)$ is a polynomial of $|u_2|$ with degree $-2\nu-d$.

For (LS.5), we focus on the first-order partial derivatives in detail and omit details for the second-order partial derivatives, as similar arguments can be applied. Straightforward calculation shows that the (absolute value of) partial derivatives of $h(\underline{u}_1, u_2)$ can be bounded by products of two positive functions, $\widetilde{\gamma}_2(\|\underline{u}_1\|)$ and $\widetilde{\gamma}_3(|u_2|)$. Moreover, $\widetilde{\gamma}_2(\|\underline{u}_1\|)$ is a linear combination of a polynomial of $\|\underline{u}_1\|$ with degree *at least* ν and a modified Bessel function of the second kind, and $\widetilde{\gamma}_3(|u_2|)$ is a polynomial of $|u_2|$ with degree at most $-2\nu-d$.

Since the partial derivatives of $h(\underline{u}_1, u_2)$ with respect to $\underline{\theta}$ is continuous in $\|\underline{u}_1\|$ and $|u_2|$, it is bounded if $\|\underline{u}_1\|$ and $|u_2|$ are bounded. To show (LS.3) and (LS.5), it suffices to show that, for $k, l > 0$, there exists $M > 0$ such that

(i) $\int_M^\infty u^k K_l(u)du < \infty$,

(ii) $u^k K_l(u)$ is bounded by a nonincreasing function on (M, ∞).

Since $d \geq 1$ and $k > 0$, u^{-2k-d} is bounded on (M, ∞) and $\int_M^\infty u^{-2k-d}du = M^{-2k-d+1}/(2k+d-1) < \infty$. The last two conditions hold. By the property of Bessel function, $K_l(u) \propto e^{-u}u^{-1/2}\{1+\mathcal{O}(1/u)\}$, as $|u| \to \infty$. We can find M_1, M_2 such that $K_l(u) \leq M_1e^{-u}u^{-1/2}(1+M_2/u)$, when $u \geq M_2$. So (i) holds since

$$\begin{aligned}\int_{M_2}^{\infty} u^k K_l(u)du \leq & \int_{M_2}^{\infty} M_1 u^{k-1/2} e^{-u}(1+M_2/u)du \\ \leq & 2M_1 \int_{M_2}^{\infty} u^{k-1/2}e^{-u}du < 2M_1\Gamma(k+1/2) < \infty.\end{aligned}$$

For (ii), we have $u^k K_l(u) \leq M_1 e^{-u}u^{k-1/2}(1+M_2/u) \leq 2M_1 e^{-u}u^{k-1/2}$, when $u \geq M_2$. Since $e^{-u}u^{k-1/2}$ is decreasing on $(k-1/2, \infty)$, (ii) is satisfied. □

2.5.2 Generalized Exponential Spatio-temporal Covariance Function

In this section, we show that the following exponential spatio-temporal covariance function used in a simulation study satisfies conditions (LS.1)–(LS.5). The covariance function can be written as

$$\begin{aligned}&\gamma_n((\underline{s},t),(\underline{s}',t');\underline{\theta}) \\ &= \begin{cases} D(\underline{s},t)D(\underline{s}',t')\sigma^2 \exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}, & \text{if } \|\underline{u}_1\| > 0 \text{ or } |u_2| > 0; \\ D(\underline{s},t)D(\underline{s}',t')\sigma^2 + \tau^2, & \text{otherwise,} \end{cases}\end{aligned}$$

where $\underline{\theta} = (c_s, c_t, \sigma^2, \tau^2)^\top$ is the vector of spatio-temporal parameters with the scaling parameter in space, $c_s \geq 0$, and the scaling parameter in time, $c_t \geq 0$. In addition, $\underline{u}_1 = \varrho_{1,n}(\underline{s}-\underline{s}')$ is the spatial lag scaled to the spatially expanding domain, and $u_2 = \varrho_{2,n}(t-t')$ is the temporal lag scaled to the temporally expanding domain, where $\varrho_{1,n}$ and $\varrho_{2,n}$ are positive constants. Further, $D(\underline{s},t)$ is some fixed positive spatio-temporal function with $D(\underline{0},0) = 1$. Note that $D(\underline{s},t)^2\sigma^2 + \tau^2$ is the variance of $Y(\underline{s},t)$.

By arguments similar to Section 2.5.1, we show (LS.1), (LS.2) and (LS.4). For (LS.3), we can see that, for all $(\underline{s},t),(\underline{s}',t') \in \mathcal{R}\times\mathcal{T}$,

$$\begin{aligned}\gamma_n((\underline{s},t),(\underline{s}',t')) &\leq \{\max D(\underline{s},t)\}^2(\sigma^2+\tau^2)\exp\{-\|\underline{u}_1\|/c_s\}\exp\{-|u_2|/c_t\} \\ &= \gamma_0(\|\underline{u}_1\|)\gamma_1(|u_2|).\end{aligned}$$

Here, both $\gamma_0(\|\underline{u}_1\|)$ and $\gamma_1(|u_2|)$ are nonincreasing positive functions.

Moreover, we have $\int_0^\infty e^{-u/c_s}du = 1/c_s < \infty$ and $\int_0^\infty e^{-u/c_t}du = 1/c_t < \infty$. Thus, (LS.3) is satisfied.

Further, we have

$$
\begin{aligned}
&\partial\gamma_n/\partial\tau^2 = 1_{\{\|\underline{u}_1\|=0,|u_2|=0\}},\\
&\partial\gamma_n/\partial\sigma^2 = D(\underline{s},t)D(\underline{s}',t')\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\},\\
&\partial\gamma_n/\partial c_s = D(\underline{s},t)D(\underline{s}',t')\sigma^2\|\underline{u}_1\|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}/c_s^2,\\
&\partial\gamma_n/\partial c_t = D(\underline{s},t)D(\underline{s}',t')\sigma^2|u_2|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}/c_t^2,\\
&\partial^2\gamma_n/\partial\sigma^2\partial c_s = D(\underline{s},t)D(\underline{s}',t')\|\underline{u}_1\|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}/c_s^2,\\
&\partial^2\gamma_n/\partial\sigma^2\partial c_t = D(\underline{s},t)D(\underline{s}',t')|u_2|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}/c_t^2,\\
&\partial^2\gamma_n/\partial c_s\partial c_t = D(\underline{s},t)D(\underline{s}',t')\sigma^2\|\underline{u}_1\||u_2|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}/(c_s^2c_t^2),\\
&\partial^2\gamma_n/\partial c_s^2 = D(\underline{s},t)D(\underline{s}',t')\sigma^2\|\underline{u}_1\|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}(\|\underline{u}_1\|/c_s^4 - 2/c_s^3),\\
&\partial^2\gamma_n/\partial c_t^2 = D(\underline{s},t)D(\underline{s}',t')\sigma^2|u_2|\exp\{-\|\underline{u}_1\|/c_s - |u_2|/c_t\}(|u_2|/c_t^4 - 2/c_t^3).
\end{aligned}
$$

Here, all the first- and second-order partial derivatives are continuous in $\|\underline{u}_1\|$ and $|u_2|$ and hence bounded when $\|\underline{u}_1\|$ and $|u_2|$ are bounded. In addition, they are bounded by a product of two functions $\widetilde{\gamma}_2(\|\underline{u}_1\|)$ and $\widetilde{\gamma}_3(|u_2|)$, where $\widetilde{\gamma}_2(u)$ and $\widetilde{\gamma}_3(u)$ are products of a polynomial of u and an exponential function of u. (LS.5) is satisfied, since for $k > 0$, $u^k e^{-u}$ is nonincreasing on $[k,\infty)$, and $\int_k^\infty u^k e^{-u}du < \infty$.

2.5.3 A Remark on Assumption (C.3)

In this section, we provide two sufficient conditions for (C.3). Further, we will demonstrate that, if $\theta_3 > 1$, the generalized spatio-temporal Matérn covariance function (2.1) satisfies Assumption (C.3). The two sufficient conditions are stated as follows:

(E.1) For $1 \le k \le q$, $|\gamma_{n,k}((\underline{s},t),(\underline{s}',t'))|$ satisfies one of the following two conditions: (i) $|\gamma_{n,k}((\underline{s},t),(\underline{s},t))| > 0$; (ii) For $\|\underline{u}_1\|, |u_2| \in [M,\infty)$ for some constant $M > 0$ such that

$(\underline{s},t),(\underline{s}+\underline{u}_1/A_n,t+u_2/B_n)\in\mathcal{R}\times\mathcal{T}$, we have $|\gamma_{n,k}((\underline{s},t),(\underline{s}+\underline{u}_1/A_n,t+u_2/B_n))|\geq C_3\exp(-C_4\|\underline{u}_1\|-C_5|u_2|)$ for all n, where $C_3,C_4,C_5>0$ are constants.

(E.2) (i) For any two given positive constants M_1, M_2, there exist M_1' and M_2' with $M_1<M_1'<\infty$ and $M_2<M_2'<\infty$ such that $\sum_i\sum_j 1(\|\underline{s}_i-\underline{s}_j\|\in[M_1\delta_n,M_1'\delta_n])1(|t_i-t_j|\in[M_2\zeta_n,M_2'\zeta_n])\geq C_6N_n^{1/2+\iota_1}$ for some $C_6>0$ and $\iota_1>0$. (ii) $A_n\delta_n=\mathcal{O}(b_n)$ and $B_n\zeta_n=\mathcal{O}(b_n)$, where $b_n=\log\log N_n$.

To see the sufficiency of (E.1) and (E.2), we first note that if $A_n\delta_n=\mathcal{O}(1)$ and $B_n\zeta_n=\mathcal{O}(1)$, then $\|^n\Gamma_k\|_F^2\geq CN_n^{1/2+\iota_1}$ for some $C>0$. Thus, (C.3) is satisfied with $\iota=\iota_1$. If $A_n\delta_n\to\infty$ or $B_n\zeta_n\to\infty$, by (E.1)–(E.2), we have $\|^n\Gamma_k\|_F^2\geq CN_n^{1/2+\iota_1'}$ for some $C>0$ and any ι_1' such that $0<\iota_1'<\iota_1$, so (C.3) is satisfied with $\iota=\iota_1'$.

Next, we will show that the generalized spatio-temporal Matérn covariance function (2.1) satisfies (E.1), when $\theta_3>1$. Since $\left|\frac{\partial\gamma_n((\underline{s},t),(\underline{s},t))}{\partial\sigma^2}\right|=D(\underline{s},t)^2>0$ and $\left|\frac{\partial\gamma_n((\underline{s},t),(\underline{s},t))}{\partial\tau^2}\right|=1$, $\partial\gamma_n/\partial\sigma^2$ and $\partial\gamma_n/\partial\tau^2$ satisfy (E.1)(i).

Further, we show that $\partial\gamma_n/\partial\theta_i$ satisfies (E.1)(ii) for $i=1,2,3$. Recall that for all $(\underline{s},t),(\underline{s}+\underline{u}_1/\varrho_{1,n},t+u_2/\varrho_{2,n})\in\mathcal{R}\times\mathcal{T}$ with $\|\underline{u}_1\|>0$ or $|u_2|>0$, we have

$$\begin{aligned}
\left|\frac{\partial\gamma_n}{\partial\theta_1}\right| &= \frac{D(\underline{s},t)D(\underline{s}',t')\sigma^2 2^{1-\nu}\theta_1\theta_3^{d/2}u_2^2}{\Gamma(\nu)(\theta_1^2u_2^2+1)^{\nu+1}(\theta_1^2u_2^2+\theta_3)^{d/2+1}}\left\{(\theta_3-1)m^{\nu+1}K_{\nu-1}(m)\right.\\
&\quad\left.+\left(2\nu(\theta_1^2u_2^2+\theta_3)+d(\theta_1^2u_2^2+1)\right)m^\nu K_\nu(m)\right\},\\
\left|\frac{\partial\gamma_n}{\partial\theta_2}\right| &= \frac{D(\underline{s},t)D(\underline{s}',t')\sigma^2 2^{1-\nu}\theta_3^{d/2}\{m^{\nu+1}K_{\nu-1}(m)\}}{\Gamma(\nu)\theta_2(\theta_1^2u_2^2+\theta_3)^{d/2}(\theta_1^2u_2^2+1)^\nu},\\
\left|\frac{\partial\gamma_n}{\partial\theta_3}\right| &= \frac{D(\underline{s},t)D(\underline{s}',t')\sigma^2 2^{-\nu}\theta_3^{d/2-1}\{d\theta_1u_2^2m^\nu K_\nu(m)+\theta_3m^{\nu+1}K_{\nu-1}(m)\}}{\Gamma(\nu)(\theta_1^2u_2^2+\theta_3)^{d/2+1}(\theta_1^2u_2^2+1)^\nu}.
\end{aligned}$$

Up to some constant scale, we have

$$\begin{aligned}
\left|\frac{\partial\gamma_n}{\partial\theta_1}\right| &\geq |u_2|^{-2\nu-d-2}m^{\nu+1}K_{\nu-1}(m)+|u_2|^{-2\nu-d}m^\nu K_\nu(m)\\
&\geq |u_2|^{-2\nu-d-2}\left(\theta_2\theta_3^{-1/2}\|\underline{u}_1\|\right)^{\nu+1}K_{\nu-1}(\theta_2\|\underline{u}_1\|)\\
&\quad+|u_2|^{-2\nu-d}\left(\theta_2\theta_3^{-1/2}\|\underline{u}_1\|\right)^{\nu}K_\nu(\theta_2\|\underline{u}_1\|),
\end{aligned}$$

$$
\begin{aligned}
\left|\frac{\partial \gamma_n}{\partial \theta_2}\right| &\geq |u_2|^{-2\nu-d} m^{\nu+1} K_{\nu-1}(m) \\
&\geq |u_2|^{-2\nu-d} \left(\theta_2 \theta_3^{-1/2} \|\underline{u}_1\|\right)^{\nu+1} K_{\nu-1}(\theta_2 \|\underline{u}_1\|), \\
\left|\frac{\partial \gamma_n}{\partial \theta_3}\right| &\geq |u_2|^{-2\nu-d-2} m^{\nu+1} K_{\nu-1}(m) + |u_2|^{-2\nu-d} m^{\nu} K_{\nu}(m) \\
&\geq |u_2|^{-2\nu-d-2} \left(\theta_2 \theta_3^{-1/2} \|\underline{u}_1\|\right)^{\nu+1} K_{\nu-1}(\theta_2 \|\underline{u}_1\|) \\
&\quad + |u_2|^{-2\nu-d} \left(\theta_2 \theta_3^{-1/2} \|\underline{u}_1\|\right)^{\nu} K_{\nu}(\theta_2 \|\underline{u}_1\|),
\end{aligned}
$$

since $\theta_2 \theta_3^{-1/2} \|\underline{u}_1\| \leq m(\underline{u}_1, u_2) \leq \theta_2 \|\underline{u}_1\|$. In addition, by property of Bessel function, $K_l(u) \propto e^{-u} u^{-1/2} \{1 + \mathcal{O}(1/u)\}$, as $|u| \to \infty$. We can find M_1, M_2 such that $K_l(u) \geq M_1 e^{-u} u^{-1/2}$, when $u \geq M_2$; thus, (E.1)(ii) follows.

Remark 4. *It can be seen that the generalized exponential covariance function also satisfies (E.1). Note that* $\left|\frac{\partial \gamma_n((\underline{s},t),(\underline{s},t))}{\partial \sigma^2}\right| = D(\underline{s},t)^2 > 0$ *and* $\left|\frac{\partial \gamma_n((\underline{s},t),(\underline{s},t))}{\partial \tau^2}\right| = 1$*, and (E.1)(i) holds. Next,* $\partial \gamma_n / \partial \sigma^2$*,* $\partial \gamma_n / \partial c_s$ *and* $\partial \gamma_n / \partial c_t$ *are positive when* $\|\underline{u}_1\| > 2c_s$ *and* $|u_2| > 2c_t$ *and can be written as linear combinations of products of* $|u_2|^j \exp(-a_1 |u_2|)$ *and* $\|\underline{u}_1\|^k \exp(-a_2 \|\underline{u}_1\|)$ *for* $j, k \geq 0$ *and some constants* $a_1, a_2 > 0$*. Hence, (E.1)(ii) follows.*

2.5.4 Proof of Theorem 1

Proof. To prove Theorem 1, it suffices to show that $\|\Gamma\|_2 = \mathcal{O}(1)$, $\|\Gamma_k\|_2 = \mathcal{O}(1)$ and $\|\Gamma_{kk'}\|_2 = \mathcal{O}(1)$, for all $k, k' = 1, \ldots, q$ [Mardia and Marshall, 1984]. Note that $\|A\|_2 \leq \|A\|_\infty$ for any positive definite matrix A. We only need to show that $\|\Gamma\|_\infty = \mathcal{O}(1)$, $\|\Gamma_k\|_\infty = \mathcal{O}(1)$ and $\|\Gamma_{kk'}\|_\infty = \mathcal{O}(1)$, for all $k, k' = 1, \ldots, q$.

For each i, let $\mathcal{A}_{1,i} = \{j : \|\underline{s}_i - \underline{s}_j\| \leq C_{s,n}\}$ and $\mathcal{A}_{2,i} = \{j : |t_i - t_j| \leq C_{t,n}\}$. Let $a_1 = C_{s,n}/\delta_n$, $a_2 = \delta_n A_n$, $b_1 = C_{t,n}/\zeta_n$ and $b_2 = \zeta_n B_n$. Then,

$$
\begin{aligned}
\|\Gamma\|_\infty &= \max_{1\le i\le N_n} \sum_{j\in\mathcal{A}_{1,i}\cap\mathcal{A}_{2,i}} \Gamma_{ij} + \max_{1\le i\le N_n} \sum_{j\in\mathcal{A}_{1,i}^c\cap\mathcal{A}_{2,i}} \Gamma_{ij} \\
&\quad + \max_{1\le i\le N_n} \sum_{j\in\mathcal{A}_{1,i}\cap\mathcal{A}_{2,i}^c} \Gamma_{ij} + \max_{1\le i\le N_n} \sum_{j\in\mathcal{A}_{1,i}^c\cap\mathcal{A}_{2,i}^c} \Gamma_{ij} \\
&= (I_1)+(I_2)+(I_3)+(I_4),
\end{aligned}
$$

where Γ_{ij} is the (i,j)th entry of Γ.

Denote $\mathrm{Card}(\mathcal{A})$ as the cardinality of the set $\mathcal{A}$, then

$$
\begin{aligned}
(I_1) &\le \|\Gamma\|_{\max}\cdot \mathrm{Card}(\mathcal{A}_{1,i}\cap\mathcal{A}_{2,i}) \le \mathcal{O}\left(\frac{C_{s,n}^d C_{t,n}}{\delta_n^d \zeta_n}\right) = \mathcal{O}(a_1^d b_1), \\
(I_2) &\le \mathrm{Card}(\mathcal{A}_{2,i}) \sum_{m=\lfloor\left(\frac{C_{s,n}A_n}{b}\right)\rfloor} \mathcal{O}\left(\frac{m^{d-1}b^d}{\delta_n^d A_n^d}\right) \max_{mb\le\|\underline{u}_1\|\le(m+1)b} \gamma_0(\|\underline{u}_1\|) \\
&\le \mathcal{O}\left(\frac{C_{t,n}}{\zeta_n\delta_n^d A_n^d}\right)\int_{C_{s,n}A_n}^\infty u^{d-1}\gamma_0(u)du = \mathcal{O}(b_1/a_2^d)\int_{a_1a_2}^\infty u^{d-1}\gamma_0(u)du, \\
(I_3) &\le \mathrm{Card}(\mathcal{A}_{1,i}) \sum_{m=\lfloor\left(\frac{C_{t,n}B_n}{b}\right)\rfloor} \mathcal{O}\left(\frac{b}{\zeta_n B_n}\right) \max_{mb\le|u_2|\le(m+1)b} \gamma_1(|u_2|) \\
&\le \mathcal{O}\left(\frac{C_{s,n}^d}{\delta_n^d\zeta_n B_n}\right)\int_{C_{t,n}B_n}^\infty \gamma_1(u)du = \mathcal{O}(a_1^d/b_2)\int_{b_1b_2}^\infty \gamma_1(u)du, \\
(I_4) &\le \sum_{m=\lfloor\left(\frac{C_{s,n}A_n}{b}\right)\rfloor} \mathcal{O}\left(\frac{m^{d-1}b^d}{\delta_n^d A_n^d}\right) \max_{mb\le\|\underline{u}_1\|\le(m+1)b} \gamma_0(\|\underline{u}_1\|)\times \\
&\qquad \sum_{m'=\lfloor\left(\frac{C_{t,n}B_n}{b}\right)\rfloor} \mathcal{O}\left(\frac{b}{\zeta_n B_n}\right) \max_{m'b\le|u_2|\le(m'+1)b} \gamma_1(|u_2|) \\
&\le \mathcal{O}\left(\frac{1}{\zeta_n B_n\delta_n^d A_n^d}\right)\int_{C_{s,n}A_n}^\infty u\gamma_0(u)du\int_{C_{t,n}B_n}^\infty \gamma_1(u)du \\
&= \mathcal{O}(1/a_2^d b_2)\int_{a_1a_2}^\infty u^{d-1}\gamma_0(u)du\int_{b_1b_2}^\infty \gamma_1(u)du.
\end{aligned}
$$

To show $\|\Gamma\|_\infty = \mathcal{O}(1)$, it suffices to show that

(i) $a_1^d b_1 \in [C_1, C_2]$ for some constants $C_1, C_2 > 0$,

(ii) $\mathcal{O}(1/a_1^d a_2^d)\int_{a_1a_2}^\infty u^{d-1}\gamma_0(u)du = \mathcal{O}(1)$,

(iii) $\mathcal{O}(1/b_1b_2)\int_{b_1b_2}^{\infty}\gamma_1(u)du = \mathcal{O}(1)$.

Let $C_{s,n} = 1/A_n$ and $C_{t,n} = 1/B_n$. By (A.3), $a_1^d b_1 = (\delta_n^d A_n^d \zeta_n B_n)^{-1} \leq c_3^{-1} = \mathcal{O}(1)$, the above requirements are fulfilled. By (LS.5),

$$\max\{|\gamma_{n,k}((\underline{s},t),(\underline{s}',t');\underline{\theta})|, |\gamma_{n,kk'}((\underline{s},t),(\underline{s}',t');\underline{\theta})|\} \leq \gamma_2(0)\gamma_3(0),$$

uniformly for all n and $1 < k, k' < q$. Thus, we have $\|\Gamma\|_2 = \mathcal{O}(1)$, and similar arguments can be applied to show that $\|\Gamma_k\|_2 = \mathcal{O}(1)$ and $\|\Gamma_{kk'}\|_2 = \mathcal{O}(1)$.

Together with (C.3)–(C.5) and by Theorem 1 of Sweeting [1980], we have the result of Theorem 1.

□

2.6 Simulation Study

We conduct a simulation study to investigate the finite sample performance of $\widehat{\underline{\theta}}_{\mathrm{MLE}}$ in Section 2.4. First, N_s sampling locations, $\underline{s}_1, \ldots, \underline{s}_{N_s}$, are generated within the spatial domain $[0,1]^2$. At each sampling location, we consider time points $t_1, \ldots, t_{N_t}$, where $t_i = (i - 1/2)/1000$ for $i = 1, \ldots, 1000$, and each time point has a 0.04 probability of being sampled. The spatio-temporal sampling points are generated once and remain fixed throughout the simulation study. We consider $N_s = 20, 40, 60$ sampling locations and the corresponding sample sizes are $N_n = 806, 1644, 2449$, respectively.

The spatio-temporal process $\varepsilon(\underline{s},t)$ is generated from a zero-mean Gaussian process with one of three types of covariance functions. The first type is an exponential spatio-temporal covariance function

$$\begin{aligned}&\mathrm{Cov}\{\varepsilon(\underline{s}_i,t_i),\varepsilon(\underline{s}_j,t_j)\}\\ &= \begin{cases} \sigma^2(1-c)\exp\{-\varrho_{1,n}\|\underline{s}_i-\underline{s}_j\|/c_s - \varrho_{2,n}|t_i-t_j|/c_t\}, & \text{if } i \neq j;\\ \sigma^2, & \text{if } i = j,\end{cases}\end{aligned}$$

where, σ^2 is the variance of the error process, $c \in [0, 1]$ is a nugget proportion such that $c\sigma^2$ is the nugget effect, and c_s and c_t are the positive range parameters in space and time, respectively. When there is only one spatial sampling location,, the covariance function is the same as an AR(1) model in time series. We set $\sigma^2 = 9.0,\ c = 0.2,\ c_s = 1$ and $c_t = 1$. The resulting spatio-temporal covariance function is stationary and separable in space and time, and is referred to as `COV-1`.

Table 2.1: Sample mean, sample standard deviation (SD), average information-based standard deviation (SDm) of covariance parameters with $N_n = 806, 1644, 2449$.

			$N_n = 806$			$N_n = 1644$			$N_n = 2449$		
	Para.	Truth	Mean	SD	SDm	Mean	SD	SDm	Mean	SD	SDm
`COV-1`	σ^2	9.0	8.973	0.525	0.529	9.012	0.373	0.385	8.974	0.317	0.313
	c	0.2	0.205	0.078	0.077	0.197	0.047	0.049	0.193	0.044	0.041
	c_s	1.0	1.023	0.197	0.200	1.008	0.123	0.122	0.994	0.097	0.095
	c_t	1.0	1.028	0.220	0.204	1.008	0.133	0.135	0.986	0.108	0.109
`COV-2`	σ^2	9.0	8.941	1.528	1.555	9.093	1.100	1.103	9.085	0.865	0.881
	c	0.2	0.233	0.120	0.105	0.202	0.060	0.062	0.196	0.052	0.050
	a	1.0	0.996	0.117	0.105	1.002	0.071	0.071	1.010	0.058	0.059
	b	1.0	1.001	0.139	0.135	1.005	0.089	0.088	1.008	0.069	0.070
	d	1.0	1.018	0.240	0.238	0.999	0.165	0.162	0.992	0.127	0.129
`COV-3`	σ^2	9.0	9.066	2.358	2.357	9.272	1.715	1.704	9.264	1.392	1.419
	c	0.2	0.248	0.156	0.142	0.205	0.077	0.077	0.195	0.062	0.062
	a	1.0	0.997	0.115	0.102	1.000	0.068	0.068	1.009	0.055	0.057
	b	1.0	1.001	0.134	0.131	1.004	0.087	0.084	1.008	0.067	0.068
	d	0.5	0.511	0.220	0.224	0.500	0.157	0.151	0.491	0.116	0.120
	e	0.5	0.514	0.227	0.221	0.499	0.142	0.146	0.486	0.120	0.118
	f	0.5	0.526	0.196	0.201	0.491	0.153	0.152	0.495	0.116	0.115

The second type is a generalized spatio-temporal Matérn covariance function given in (2.1). We let the smoothness parameter be $\nu = 1/2$ and the separability parameter be $\theta_3 = 1$. Then, (2.1) is simplified to

$$\begin{aligned}
&\mathrm{Cov}\{\varepsilon(\underline{s}_i, t_i), \varepsilon(\underline{s}_j, t_j)\} \\
&= \begin{cases} D(\underline{s}_i, t_i) D(\underline{s}_j, t_j) \frac{\sigma^2}{(a^2|\varrho_{2,n}(t_i - t_j)|^2 + 1)^{3/2}} \exp\{-b\varrho_{1,n}\|\underline{s}_i - \underline{s}_j\|\}, & \text{if } i \neq j; \\ D(\underline{s}_i, t_i) D(\underline{s}_j, t_j)\sigma^2 + c\sigma^2, & \text{if } i = j. \end{cases}
\end{aligned} \tag{2.4}$$

Here, σ^2 is the variance of the error process, $c \in [0, 1]$ is a nugget proportion such that $c\sigma^2$ is the nugget effect, and the range parameters in space and time are a and b, respectively. The nonstationarity of the covariance function is induced by $D(\underline{s}_i, t_i) = dt_i + 1$ with a change of variance over time. We set $\sigma^2 = 9, c = 0.2, a = 1, b = 1$, and $d = 1$. The resulting spatio-temporal covariance function is nonstationary but separable in space and time, and is referred to as `COV-2`. For the third type of covariance functions, we also consider (2.4), but let $D(\underline{s}_i, t_i) = dt_i + es_{1i} + fs_{2i} + 1$. We set $\sigma^2 = 9.0, c = 0.2, a = 1, b = 1, d = 0.5, e = 0.5$ and $f = 0.5$. The resulting spatio-temporoal covariance function is nonstationary and nonseparable in space and time, and is referred to as `COV-3`.

For each combination of the sample size N_n and the covariance function, a total of 400 simulated data sets are generated. The sample mean, sample standard deviation (SD), and averaged information matrix based standard deviation (SDm) of covariance parameters are reported in Table 2.1. For all three types of covariance functions, as the sample size increases, the sample standard deviations of parameter estimates become smaller, supporting the consistency of the parameter estimates in Theorem 2. Moreover, the sample standard deviations of parameter estimates are close to the average information-based standard deviation, as indicated by the asymptotic normality in Theorem 2. For the nonstationary covariance functions (`COV-2` and `COV-3`), the sample mean has a significant bias for the nugget proportion parameter when the number of sampling locations is $N_s = 20$, likely because the covariance functions `COV-2` and `COV-3` are more complex than `COV-1`. As the sample size increases, the bias of the nugget proportion parameter becomes smaller for both `COV-2` and `COV-3`, which suggests that in practice, a larger sample size would be needed for more complex covariance functions.

2.7 Discussions and Generalization

Consider a spatio-temporal linear regression model

$$y(\boldsymbol{s},t) = \boldsymbol{x}(\boldsymbol{s},t)^\top \boldsymbol{\beta} + \varepsilon_1(\boldsymbol{s},t) + \varepsilon_2(\boldsymbol{s},t), \;\; \boldsymbol{s} \in \mathcal{R},\; t \in \mathcal{T}, \tag{2.5}$$

where $\boldsymbol{x}(\boldsymbol{s},t) = (x_1(\boldsymbol{s},t),\ldots,x_p(\boldsymbol{s},t))^\top$ is a $p \times 1$ vector of covariates at spatial location $\boldsymbol{s}$ and time point t and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ is a $p \times 1$ vector of regression coefficients. Let $\boldsymbol{X} = [x_j(\boldsymbol{s}_i,t_i)]_{i=1,j=1}^{N_n,p}$ denote an $N_n \times p$ design matrix and let $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ denote a $(p+q) \times 1$ vector of parameters comprising both the regression coefficients $\boldsymbol{\beta}$ and the covariance function parameters $\boldsymbol{\theta}$. The log-likelihood function of (2.2) is

$$\begin{aligned} \ell_{\mathrm{reg}}(\boldsymbol{\eta}) = &-(N_n/2)\log(2\pi) - (1/2)\log\{\det\boldsymbol{\Gamma}(\boldsymbol{\theta})\} \\ &-(1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{2.6}$$

Denote the maximizer of (2.6) as $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE,reg}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{MLE,reg}}, \widehat{\boldsymbol{\theta}}_{\mathrm{MLE,reg}})$.

We use $\boldsymbol{\beta}_0$ to denote the vector of true regression coefficients and $\boldsymbol{\theta}_0$ to denote the vector of true covariance parameters. For the log-likelihood function $\ell_{\mathrm{reg}}(\boldsymbol{\eta})$, we have $\ell'_{\mathrm{reg}}(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ and $\ell''_{\mathrm{reg}}(\boldsymbol{\beta},\boldsymbol{\beta}) = -\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}$. In addition, the kth element of $\ell'_{\mathrm{reg}}(\boldsymbol{\theta})$ is $-(1/2)\cdot\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) - (1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Gamma}^k(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ and the (k,k')th entry of $\ell''_{\mathrm{reg}}(\boldsymbol{\theta},\boldsymbol{\theta})$ is $-(1/2)\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{kk'}+\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}_{k'}) - (1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Gamma}^{kk'}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. Let $\boldsymbol{\mathcal{I}}_{n,\mathrm{reg}}(\boldsymbol{\beta}) = E\{-\ell''_{\mathrm{reg}}(\boldsymbol{\beta},\boldsymbol{\beta})\}$ and $\boldsymbol{\mathcal{I}}_{n,\mathrm{reg}}(\boldsymbol{\theta}) = E\{-\ell''_{\mathrm{reg}}(\boldsymbol{\theta},\boldsymbol{\theta})\}$ be the information matrices of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. Observe that, the (k,k')th element of $\boldsymbol{\mathcal{I}}_{n,\mathrm{reg}}(\boldsymbol{\theta})$ is also $\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'})/2$, which is the same as that of $\boldsymbol{\mathcal{I}}_{n,\mathrm{zm}}(\boldsymbol{\theta})$.

Under the asymptotic framework (A.1)–(A.3) in Section 2.3, we study the theoretical properties of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE,reg}}$. Besides (LS.1)–(LS.5) and (C.1)–(C.4), we will need the following conditions.

(D.2) The design matrix $\boldsymbol{X}$ is full rank with uniformly bounded max norm. There exists a matrix $\boldsymbol{C}_X$, such that $N_n^{-1}\boldsymbol{X}^\top\boldsymbol{X} \longrightarrow \boldsymbol{C}_X$ as $n \to \infty$.

(D.3) There exist non-singular matrices $\boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\beta})$ and $\boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\theta})$ which satisfy $N_n^{-1}\boldsymbol{\mathcal{J}}_{n,{\rm reg}}(\boldsymbol{\beta}) \longrightarrow \boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\beta})$ and $N_n^{-1}\boldsymbol{\mathcal{J}}_{n,{\rm reg}}(\boldsymbol{\theta}) \longrightarrow \boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\theta})$ as $n \to \infty$.

By Theorem 1 in Section 2.4, Theorem 1 of Sweeting [1980], and Theorem 2 of Mardia and Marshall [1984], we have the following asymptotic results.

Theorem 5. *Under (LS.1)–(LS.5), (C.1)–(C.4) and (D.2)–(D.3), there exists, with probability tending to one, a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}_{\rm reg} = ({}^n\widehat{\boldsymbol{\beta}}_{\rm reg}{}^\top, {}^n\widehat{\boldsymbol{\theta}}_{\rm reg}{}^\top)^\top$ *of* $\ell_{\rm reg}(\boldsymbol{\eta})$ *such that* $\|{}^n\widehat{\boldsymbol{\eta}}_{\rm reg} - \boldsymbol{\eta}_0\| = \mathcal{O}_p(N_n^{-1/2})$. *Moreover, the local maximizer* ${}^n\widehat{\boldsymbol{\eta}}_{\rm reg}$ *is asymptotic normal; as* $n \to \infty$,

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}_{\rm reg} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\beta}_0)^{-1}) \quad \textit{and}$$
$$N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\rm reg} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\theta}_0)^{-1}).$$

Theorem 5 establishes the consistency and the asymptotic normality of ${}^n\widehat{\boldsymbol{\eta}}_{\rm reg}$. The convergence rate is again root-N_n. The model (2.2) is a special case of the regression mean model (2.5) with $\boldsymbol{\beta} = \mathbf{0}$. To compare the asymptotic variances of the estimates of $\boldsymbol{\theta}$ in Theorem 2 and Theorem 5, it is straightforward to show that $\boldsymbol{\mathcal{I}}_{\rm zm}(\boldsymbol{\theta}_0)^{-1} = \boldsymbol{\mathcal{I}}_{\rm reg}(\boldsymbol{\theta}_0)^{-1}$. That is, ${}^n\widehat{\boldsymbol{\theta}}_{\rm reg}$ has the same optimal asymptotic variance as ${}^n\widehat{\boldsymbol{\theta}}_{\rm zm}$ when $\boldsymbol{\beta}$ is assumed to be known, which suggests that the estimation of regression coefficients does not affect the asymptotic efficiency of the covariance parameter estimation.

Chapter 3

Semiparametric Method and Theory for Continuously Indexed Spatio-Temporal Processes[2]

3.1 Introduction

In this chapter, we develop new semiparametric methodology and theory for spatio-temporal processes where both space and time are continuously indexed, which often arise in many scientific disciplines [see, e.g., Porcu et al., 2018]. An illustrative data set comprises measurements of a health hazard taken in an indoor environment by both static sensors at fixed sampling locations and roving sensors at varying sampling locations over time [Ludwig et al., 2017]. The spatio-temporal sampling design is non-standard due to data irregularity and sparsity in both space and time, calling for development of novel methodology and theory.

There are multiple approaches to modeling spatio-temporal data with continuous spatial index but *discrete* temporal index. One approach is spatial time series modeling, which combines time series methods for temporal data with geostatistical methods for spatial data. For example, Stroud et al. [2001] proposed a state space model where spatial variability is captured by a locally weighted mixture of linear regressions while the regression coefficients are allowed to vary with time. Lu et al. [2009] developed a flexible class of spatio-temporal varying coefficient models and established the theoretical properties for estimation. Huang et al. [2018] proposed a nonparametric approach based on latent common factors to model the linear dependence structure of a spatio-temporal process. The aforementioned spatial time series methods can capture nonlinearity and nonstationarity in space and/or time, assuming that data are observed at irregular sampling locations but discrete and regular sampling time points. For geostatistical data observed at irregularly spaced sampling locations and sampling time points, in contrast, it is sensible to consider spatio-

[2]This chapter is based on a manuscript "Semiparametric Method and Theory for Continuously Indexed Spatio-Temporal Processes" with Dr. Tingjin Chu, Dr. Jun Zhu and Dr. Haonan Wang.

temporal processes with continuous spatial index and *continuous* temporal index, yet approaches that enable simultaneous estimation of the mean and covariance functions are limited. While the existing methods focus primarily on linear regression models [see, e.g., Datta et al., 2016b], we will develop semiparametric methods and theory for continuously indexed spatio-temporal processes with flexibility enhanced by partially linear regression for the mean function and local stationarity for the covariance function.

Partially linear models have been extensively studied in statistics [see, e.g., Engle et al., 1986, Härdle et al., 2000, Liang and Li, 2009]. For spatial areal data or spatio-temporal areal data with *discrete* spatial indexes, Su and Jin [2010] proposed a profile quasi-maximum likelihood method for spatial autoregressive partially linear models, and established some theoretical properties of the proposed method. Extending spatial autoregressive models, Sun et al. [2014] proposed a profile likelihood based estimation procedure for a semiparametric spatial dynamic model, which includes the special case of a spatial partially linear model with a nonlinear spatial trend. Compared with areal data, geostatistical models have *continuous* spatial indexes and the methods for dealing with these two types of data can be quite different. For example, Gao et al. [2006] proposed an estimation procedure based on marginal integration for geostatistical partially linear models, and the theoretical results are established. Lu et al. [2009] developed spatio-temporal varying coefficient models, which can be applied to spatio-temporal partially linear models, although the methods require densely observed data in both space and time. Overall, geostatistical partially linear models are underdeveloped for spatio-temporal data, especially when data are sparsely observed across space and over time.

For independent data, profile likelihood estimation is known to provide sound statistical properties for the partially linear model [see, e.g., Speckman, 1988, Härdle et al., 1998, Liang et al., 1999, Fan and Huang, 2005]. However, the dependence structure in spatio-temporal data poses challenges for establishing the asymptotic properties. There are two fundamental asymptotic frameworks in spatial statistics, namely, increasing-domain asymptotics and fixed-domain asymptotics. For increasing-domain asymptotics, the spatial domain is expanding as the number of observa-

tions increases [see, e.g., Mardia and Marshall, 1984, Chu et al., 2011, Cressie and Lahiri, 1993]. For fixed-domain asympotics, the spatial domain is fixed and observations are getting denser [see, e.g., Ying, 1993, Stein, 1999, Zhang, 2004, Loh, 2005]. For spatio-temporal processes, Chu et al. [2019] recently proposed a spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain, which extends the aforementioned asymptotic frameworks for spatial domain to spatio-temporal domain, and provides a flexible tool for exploring the asymptotic properties of statistical inference for spatio-temporal processes. The STED framework also paves the way for studying the local behavior of a spatio-temporal process, especially the second-order properties. Here, we will consider a locally stationary spatio-temporal covariance function, introduced by Chu et al. [2019], to study the slowly-varying second-order nonstationarity under the STED asymptotic framework. We will also establish the asymptotic properties of profile likelihood method for the partially linear model under a local stationarity condition that relaxes the stationarity assumption in the traditional geostatistics.

In addition, bandwidth selection is critical in partially linear kernel regression. For *iid* data, various methods have been studied, notably cross validation [see, e.g., Fan and Gijbels, 1996], but are known to not perform well for non-*iid* data [see, e.g., Altman, 1990, Opsomer et al., 2001, Lahiri, 2003a, De Brabanter et al., 2011]. While bandwidth selection is not straightforward for spatio-temporal models with a semiparametric mean function and a nonstationary covariance function, we propose to employ a family of bimodal kernels [De Brabanter et al., 2011]. We present a cross-validation based method with bimodal kernels to alleviate the bias in bandwidth selection due to the spatio-temporally correlated errors, which can be of independent interest for semiparametric spatial statistics.

The remainder of the chapter is organized as follows. Section 3.2 introduces the spatio-temporal model and the profile likelihood method. The asymptotic properties of the profile likelihood estimation are established in Section 3.3 under suitable regularity conditions. In Section 3.4, we discuss the choice of kernel functions and develop a procedure for bandwidth selection. Nu-

merical examples including a simulation study and the health hazard data example are given in Sections 3.5 and 3.6, respectively. Sections 3.7 contains the technical details including proofs.

3.2 Model and Estimation

3.2.1 Spatio-temporal Semiparametric Model

We consider the following spatio-temporal process for the response variable $y(\cdot, \cdot)$,

$$y(\boldsymbol{s}, t) = \boldsymbol{x}(\boldsymbol{s}, t)^\top \boldsymbol{\beta} + f(t) + \varepsilon(\boldsymbol{s}, t), \quad \boldsymbol{s} \in \mathcal{R},\ t \in \mathcal{T}, \tag{3.1}$$

where the location $\boldsymbol{s}$ is in the unit hypercube $\mathcal{R} = [0,1]^d$ for $d \geq 1$ and the rescaled time t takes values in $\mathcal{T} = [0,1]$. Here, $\boldsymbol{x}(\boldsymbol{s}, t) = (x_1(\boldsymbol{s}, t), \ldots, x_p(\boldsymbol{s}, t))^\top$ is a $p \times 1$ vector of covariates at spatial location $\boldsymbol{s}$ and time point t, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of regression coefficients, and $f(t)$ denotes a fixed nonparametric temporal function. In the special case of $\boldsymbol{\beta} = \mathbf{0}$, (3.1) has a fully nonparametric mean function. Further, the zero-mean spatio-temporal random process $\varepsilon(\boldsymbol{s}, t)$ accounts for the local variations unexplained by the mean function (i.e., trend) $\boldsymbol{x}(\boldsymbol{s}, t)^\top \boldsymbol{\beta} + f(t)$. Denote $\gamma((\boldsymbol{s}, t), (\boldsymbol{s}', t'); \boldsymbol{\theta})$ as the covariance function of $\varepsilon(\boldsymbol{s}, t)$, where $(\boldsymbol{s}, t), (\boldsymbol{s}', t') \in \mathcal{R} \times \mathcal{T}$ and $\boldsymbol{\theta}$ is a $q \times 1$ vector of covariance function parameters.

We consider N samples observed at $(\boldsymbol{s}_1, t_1), \ldots, (\boldsymbol{s}_N, t_N)$. Define the $N \times 1$ vector of the response variable as $\boldsymbol{y} = (y(\boldsymbol{s}_1, t_1), \ldots, y(\boldsymbol{s}_N, t_N))^\top$, the $N \times p$ design matrix as $\boldsymbol{X} = [x_j(\boldsymbol{s}_i, t_i)]_{i,j=1}^{N,p}$, and the $N \times 1$ vector of the errors as $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1, t_1), \ldots, \varepsilon(\boldsymbol{s}_N, t_N))^\top$. Let $\boldsymbol{f} = (f(t_1), \ldots, f(t_N))^\top$ denote the temporal function at the N sampling points. We have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\varepsilon}. \tag{3.2}$$

The $N \times N$ covariance matrix of $\boldsymbol{\varepsilon}$ is expressed as $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = [\gamma((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j); \boldsymbol{\theta})]_{i,j=1}^{N}$. Further, let $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ denote a $(p+q) \times 1$ vector of parameters comprising both the regression coefficients $\boldsymbol{\beta}$ and the covariance function parameters $\boldsymbol{\theta}$.

3.2.2 Profile Likelihood Estimation

While the likelihood principle cannot be easily adopted for semiparametric models like (3.1), here we develop a profile likelihood method for model estimation. For a given $\boldsymbol{\beta}$, let $y_i^* = y(\boldsymbol{s}_i, t_i) - \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}$ denote a partially detrended spatio-temporal process for the response variable and let $\boldsymbol{y}^* = (y_1^*, \ldots, y_N^*)^\top$ denote an $N \times 1$ vector of partially detrended spatio-temporal reponse variables. We obtain an estimate of $\boldsymbol{f}$ by local polynomial regression; that is, we minimize the following criterion, with respect to $\boldsymbol{b}_t = (b_{0,t}, b_{1,t})^\top$,

$$\sum_{i=1}^{N} \{y_i^* - b_{0,t} - b_{1,t}(t_i - t)\}^2 K_h(t_i - t), \tag{3.3}$$

where $K_h = K(\cdot/h)/h$ is a kernel function $K(\cdot)$ with a bandwidth h.

With $\boldsymbol{K}_t = \mathrm{diag}\{K_h(t_1 - t), \ldots, K_h(t_N - t)\}$ and $\boldsymbol{D}_t = (\mathbf{1}_N, \boldsymbol{d}_{1t})$, where $\mathbf{1}_N$ is an $N \times 1$ vector of 1's and $\boldsymbol{d}_{1t} = ((t_1 - t)/h, \ldots, (t_N - t)/h)^\top$, it follows from (3.1) that,

$$(\widehat{b}_{0,t}, h\widehat{b}_{1,t})^\top = \boldsymbol{\omega}(t)\boldsymbol{y}^*,$$

where $\boldsymbol{\omega}(t) = (\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t$. The resulting estimate of $\boldsymbol{f}$ is $\widetilde{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y}^* = \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, where the smoother matrix is

$$\boldsymbol{S} = \left(\boldsymbol{\omega}_1(t_1)^\top, \ldots, \boldsymbol{\omega}_1(t_N)^\top\right)^\top, \tag{3.4}$$

and $\boldsymbol{\omega}_1(t) = (1, 0)\boldsymbol{\omega}(t)$. Plugging $\widetilde{\boldsymbol{f}}$ into (3.2), we have the following approximation

$$(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{y} \approx (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.5}$$

To obtain the estimate of $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, denoted as $(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$, we propose to maximize the following profile log-likelihood function,

$$\begin{aligned}\ell(\boldsymbol{\beta},\boldsymbol{\theta}) = &-(N/2)\log(2\pi) - (1/2)\log\{\det\boldsymbol{\Gamma}(\boldsymbol{\theta})\} \\ &-(1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).\end{aligned} \tag{3.6}$$

Consequently, the estimate of $\boldsymbol{f}$ can be expressed as

$$\widehat{\boldsymbol{f}} = \boldsymbol{S}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

In addition, let $\boldsymbol{f}' = (f'(t_1),\ldots,f'(t_N))^\top$ denote an $N\times 1$ vector of the first-order derivatives of the temporal function $f(t)$ evaluated at the sampling time points $t_1,\ldots,t_N$. Minimizing (3.3) yields an estimate of $\boldsymbol{f}'$

$$\widehat{\boldsymbol{f}}' = \boldsymbol{L}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where $\boldsymbol{L} = \left(h^{-1}\boldsymbol{\omega}_2(t_1)^\top,\ldots,h^{-1}\boldsymbol{\omega}_2(t_N)^\top\right)^\top$ and $\boldsymbol{\omega}_2(t) = (0,1)\boldsymbol{\omega}(t)$.

In general, we write the estimate of $\boldsymbol{F}(t) = (f(t), hf'(t))^\top$ as $\widehat{\boldsymbol{F}}(t) = \boldsymbol{\omega}(t)(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$. In the case of spatio-temporal independence (i.e., $\boldsymbol{\Gamma} = \sigma^2\boldsymbol{I}$), the estimates of $\boldsymbol{\beta}$ and σ^2 in (3.6) can be expressed in closed form [see, e.g., Fan and Huang, 2005]. In the case of a nonparametric mean function (i.e., $\boldsymbol{\beta} = \boldsymbol{0}$), (3.6) can still be maximized to obtain the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{f}$, while the estimate of $\boldsymbol{f}'$ can be obtained by $\widehat{\boldsymbol{f}}' = \boldsymbol{L}\boldsymbol{y}$. The estimation procedure above depends on the choice of bandwidth, which will be discussed in Section 3.4.

3.3 Asymptotic Results

3.3.1 Asymptotic Framework

In this section, we will establish the theoretical properties of the method developed in the previous Section 3.2 under a spatio-temporal expanding distance (STED) asymptotic framework [Chu et al., 2019]. Denote n as the stage of the asymptotics and let $\{A_n\}$ and $\{B_n\}$ be two sequences of positive numbers. The (A_n, B_n)-rate STED asymptotic framework in a fixed spatio-temporal domain is defined as follows.

For all n, there exist positive constants c_1, c_2 and c_3, independent of n, such that

(A.1) $\delta_n / \min_{1 \le j \le N_n} \delta_{j,n} \le c_1$,

(A.2) $\zeta_n / \min_{1 \le j \le N_n} \zeta_{j,n} \le c_2$,

(A.3) $\delta_n^d A_n^d \zeta_n B_n \ge c_3$,

where $\delta_{j,n} = \min\{\|\boldsymbol{s}_i - \boldsymbol{s}_j\| : 1 \le i \le N_n, \boldsymbol{s}_i \ne \boldsymbol{s}_j\}$, $\delta_n = \max_{1 \le j \le N_n} \delta_{j,n}$, $\zeta_{j,n} = \min\{|t_i - t_j| : 1 \le i \le N_n, t_i \ne t_j\}$ and $\zeta_n = \max_{1 \le j \le N_n} \zeta_{j,n}$. We assume that the error process $\varepsilon(\boldsymbol{s}, t)$ is locally stationary. A covariance function $\gamma_n((\boldsymbol{s}, t), (\boldsymbol{s}', t'))$ is said to be *locally stationary* if there exists a sequence of functions $g_n(\cdot, \cdot, \boldsymbol{s}, t)$ such that,

$$|\gamma_n((\boldsymbol{s}, t), (\boldsymbol{s}', t')) - g_n(\boldsymbol{s}' - \boldsymbol{s}, t' - t, \boldsymbol{s}, t)| = \mathcal{O}(\|\boldsymbol{s}' - \boldsymbol{s}\| + |t' - t| + \rho_n),$$

uniformly for all $(\boldsymbol{s}, t), (\boldsymbol{s}', t') \in \mathcal{R} \times \mathcal{T}$, where $\{\rho_n\}$ is a sequence of positive numbers and $\rho_n \to 0$ as $n \to \infty$. In addition, there exists a function g such that, as $n \to \infty$,

$$\lim_{n \to \infty} |g_n(\boldsymbol{s}' - \boldsymbol{s}, t' - t, \boldsymbol{s}, t) - g(\boldsymbol{u}_1, u_2, \boldsymbol{s}, t)| \to 0,$$

where $\boldsymbol{u}_1 = A_n(\boldsymbol{s}' - \boldsymbol{s})$ and $u_2 = B_n(t' - t)$.

We use a one-dimensional (1D) toy example to illustrate the structure of the locally stationary covariance function. For locations $s \in \mathcal{R} = [0, 1]$, we construct a locally stationary covariance function by taking the product of a positive function $D(s)$ and a stationary covariance function such that $\gamma(s, s') = D(s)D(s')\exp(-d/r)$. Figure 3.1 demonstrates four covariance functions, one stationary covariance function where $D_1(s) = 1$ and three locally stationary covariance functions.

3.3.2 Asymptotic Properties

For *iid* data, the maximum profile likelihood estimate $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal [Fan and Huang, 2005]. For the spatio-temporal semiparametric model (3.1), the asymptotic properties of the maximum profile likelihood estimates ${}^n\widehat{\boldsymbol{\beta}}$ and ${}^n\widehat{\boldsymbol{\theta}}$, which maximize (3.6), will be established as follows.

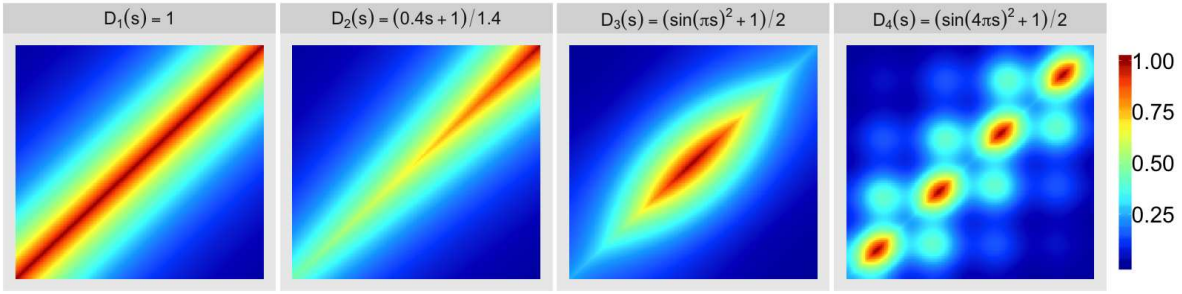

Figure 3.1: Visualization of the locally stationary correlation matrix for a 1D process with a stationary covariance function ($D_1(s) = 1$) and three nonstationary covariances ($D_2(s) = (0.4s + 1)/1.4$, $D_3(s) = \{\sin(\pi s)^2 + 1\}^2/2$ and $D_4(s) = \{\sin(4\pi s)^2 + 1\}^2/2$).

Theorem 6. *Under (C.1)–(C.14) in Section 3.7, there exists, with probability tending to one, a local maximizer* ${}^n\widehat{\boldsymbol{\eta}} = ({}^n\widehat{\boldsymbol{\beta}}^\top, {}^n\widehat{\boldsymbol{\theta}}^\top)^\top$ *of* $\ell(\boldsymbol{\eta})$ *such that* $\|{}^n\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = \mathcal{O}_p(N_n^{-1/2})$. *Moreover, the local maximizer* ${}^n\widehat{\boldsymbol{\eta}}$ *is asymptotic normal; that is, as* $n \to \infty$,

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Pi}^{-1}) \quad \textit{and} \quad N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta}_0)^{-1}).$$

Theorem 6 establishes that the estimate ${}^n\widehat{\boldsymbol{\eta}}$ is root-N_n consistent. However, the asymptotic variance of ${}^n\widehat{\boldsymbol{\beta}}$ does not converge to the information matrix (3.13). As will be seen in Section 3.7.1, if $\|\boldsymbol{\Gamma}^{-1}\|_\infty = \mathcal{O}(1)$, then $\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \succeq \boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, where $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A}-\boldsymbol{B}$ is positive semi-definite. That is, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ in partially linear model is greater than those in simple linear regression model. Following a series of lemmas in Section 3.7.2, the proof of Theorem 6 is given in Section 3.7.3.

A by-product of the proof for Theorem 6, given in Section 3.7.3, shows that $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}_0) = \lim_{n\to\infty} N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}$. Thus, we use $N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}$ as a finite sample approximation of $\boldsymbol{\Pi}$, the asymptotic variance of ${}^n\widehat{\boldsymbol{\beta}}$. In contrast, for ${}^n\widehat{\boldsymbol{\theta}}$, it can be shown that the asymptotic variance is the same as that for the case when the temporal function $f(\cdot)$ is assumed to be known.

Further, recall that $\widehat{\boldsymbol{F}}(t) = \boldsymbol{\omega}(t)(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ is the estimate of $\boldsymbol{F}(t) = (f(t), hf'(t))^\top$, where $\boldsymbol{\omega}(t) = (\boldsymbol{D}_t^\top\boldsymbol{K}_t\boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top\boldsymbol{K}_t$. The following Theorem 7 establishes the asymptotic normality of $\widehat{\boldsymbol{F}}(t)$. The proof of Theorem 7 is given in Section 3.7.4.

Theorem 7. *Suppose $f^{(3)}(t)$ is bounded. Under (C.1)–(C.14) in Section 3.7, we have, as $n \to \infty$,*

$$(N_n h)^{1/2} \left\{ \widehat{\boldsymbol{F}}(t) - \boldsymbol{F}(t) - (1/2)h^2 \begin{pmatrix} \mu_2 f''(t) \\ 0 \end{pmatrix} + o(h^2) \right\} \xrightarrow{D} N\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix} \boldsymbol{\Delta}_t \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix} \right),$$

for $t \in (0,1)$, where $\mu_k = \int_{-\infty}^{\infty} x^k K(x) dx$.

Similar to Theorems 6 and 7, we can show that, when $\boldsymbol{\beta}$ is known, the asymptotic properties of ${}^n\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{F}}(t)$ remain the same as in Theorem 7. This may be expected, since $\widehat{\boldsymbol{\beta}}$ is root-N_n consistent. Further, $\boldsymbol{\Delta}_t$ is generally unknown in practice. As suggested by (C.10), $\boldsymbol{\Delta}_t$ can be approximated by $N_n^{-1} h q(t)^{-2} \boldsymbol{k}_t^\top \boldsymbol{\Gamma} \boldsymbol{k}_t$. Since $q(t)$ represents the density of the sampling time points, we may use a kernel density method to estimate $q(t)$. An alternative is to estimate $q(t)$ by $v_{0,t}/N_n$, where $v_{0,t} = \sum_{i=1}^{N_n} K_h(t_i - t)$. Lemma 1 in Section 3.7.2 shows that such an approximation is reasonable. In the remainder of this chapter, we will refer to the former approximation as a kernel density approximation and the latter a plug-in approximation.

3.4 Selection of Kernel and Bandwidth

3.4.1 Theoretically Optimal Bandwidth

The selection of bandwidth is crucial in kernel smoothing and thus we derive a theoretically optimal bandwidth. By the results in Theorem 7, the asymptotic mean squared error (AMSE) of $\widehat{f}(t)$ is

$$\mathrm{AMSE}(t) = (1/4)h^4 \mu_2^2 f''(t)^2 + (N_n h)^{-1} (1,0) \boldsymbol{\Delta}_t (1,0)^\top,$$

and the asymptotic weighted mean integrated squared error is

$$\begin{aligned} \mathrm{AMISE}(h) &= \int_0^1 \mathrm{AMSE}(t) q(t) dt \\ &= (1/4)h^4 \mu_2^2 \int_0^1 f''(t)^2 q(t) dt + (N_n h)^{-1} \int_0^1 (1,0) \boldsymbol{\Delta}_t (1,0)^\top q(t) dt. \end{aligned}$$

Viewing the density function $q(t)$ as a weight function, we obtain an asymptotically optimal bandwidth as

$$h_{\text{opt}} = N_n^{-1/5}\mu_2^{-2/5}\left\{\frac{\int_0^1 (1,0)\boldsymbol{\Delta}_t(1,0)^\top q(t)dt}{\int_0^1 f''(t)^2 q(t)dt}\right\}^{1/5}, \tag{3.7}$$

where the convergence rate is $N_n^{2/5}$ and is the nonparametric optimal rate [Stone, 1982].

The asymptotically optimal bandwidth h_{opt} above depends on several unknown quantities: $\boldsymbol{\Delta}_t$ in the asymptotic variance of $\widehat{\boldsymbol{F}}(t)$, the density of sampling time points $q(t)$, and the second-order derivative of the temporal function $f''(t)$; thus, it is not straightforward to estimate h_{opt}. When $\boldsymbol{\Gamma} = \sigma^2\boldsymbol{I}$ (i.e., the process assumes spatio-temporal independence), $\boldsymbol{\Delta}_t$ can be expressed as $\sigma^2 q(t)^{-1}\text{diag}\{\int_{-\infty}^{\infty} K(u)^2du, \int_{-\infty}^{\infty} u^2K(u)^2du\}$. A rule of thumb for bandwidth selection in this case is available [see, e.g., Fan and Gijbels, 1996]. The idea is to plug in the estimates of σ^2 and $f''(t)$ to obtain an approximation of h_{opt}. Specifically, after a pilot global polynomial regression of degree 4 is fitted, σ^2 is estimated by the standardized residual sum of squares, and the estimate of $f''(t)$ is obtained by differentiating the resulting global fit. However, for a spatio-temporally correlated error process, the covariance matrix $\boldsymbol{\Gamma}$ needs to be estimated, and this rule of thumb is not directly applicable. Hence, a more practical bandwidth selection procedure is needed.

3.4.2 Practical Bandwidth Selection

Under model (3.1), we have

$$y^*(\boldsymbol{s},t) = f(t) + \varepsilon_1(\boldsymbol{s},t) + \varepsilon_2(\boldsymbol{s},t), \;\; \boldsymbol{s}\in\mathcal{R},\; t\in\mathcal{T}. \tag{3.8}$$

We use a leave-one-out cross-validation criterion [CV; Wasserman, 2010]. A straightforward calculation reveals that $\text{CV}(h) = N_n^{-1}\sum_{i=1}^{N_n}\left\{\frac{y_i^*-\widehat{f}(t_i)}{1-S_{ii}}\right\}^2$, where S_{ii} is the (i,i)th element of the smoother matrix $\boldsymbol{S}$. However, as will be seen in the following theorem, cross-validation is asymptotically biased in the presence of correlated errors for most commonly used kernels with $K(0)\neq 0$.

Theorem 8. *Under Assumptions (C.1)–(C.14) in Section 3.7, if there exists a sequence $C_n > 0$ such that $C_n h^{-1} \to 0$ and $1/(B_n\zeta_n)\int_{B_nC_n}^{\infty}\gamma_1(u)du \to 0$, as $n \to \infty$, then we have*

$$\begin{aligned}\mathrm{E}\{\mathrm{CV}(h)\} =& N_n^{-1}\sum_{i=1}^{N_n}\mathrm{E}\{f(t_i) - \widehat{f}^{(-i)}(t_i)\}^2 + \overline{\sigma^2} \\ & - K(0)\left\{(2/N_n)\sum_{i=1}^{N_n}\sum_{\substack{j\neq i \\ |t_j - t_i|<C_n}}\frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{b(t_i)-K(0)}\right\} + o\left(1/(N_nh)\right),\end{aligned}$$

where $\widehat{f}^{(-i)}(t_i)$ is the leave-one-out estimator with the ith observation deleted for estimation, $\mathrm{CV}(h) = N_n^{-1}\sum_{i=1}^{N_n}\left\{y_i^* - \widehat{f}^{(-i)}(t_i)\right\}^2$, $\overline{\sigma^2} = N_n^{-1}\sum_{i=1}^{N_n}\mathrm{Var}(Y_i)$ and $b(t_i) = N_nq(t_i)h(\mu_{0,t_i}\mu_{2,t_i} - \mu_{1,t_i}^2)\mu_{2,t_i}^{-1}$.

The proof of Theorem 8 is given in Section 3.7.5. Theorem 8 provides a theoretical basis for the choice of kernel functions. In practice, we propose the following procedure for the selection of bandwidth h.

(1) For a predetermined bandwidth h_0 and a kernel function K_{h_0}, obtain the estimated regression coefficients $\widetilde{\boldsymbol{\beta}}$ by the profile likelihood method.

(2) For given a kernel function, find the bandwidth h_{opt} that minimizes the cross-validation criterion

$$\mathrm{CV}(h) = N_n^{-1}\sum_{i=1}^{N_n}\left(\frac{\widetilde{y}_i^* - \widetilde{f}_i^*}{1 - S_{ii}}\right)^2, \tag{3.9}$$

where $\widetilde{y}_i^* = y_i - \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top\widetilde{\boldsymbol{\beta}}$, and $\widetilde{f}_i^*$ is the profile likelihood estimate of (3.8).

(3) Use h_{opt} and the kernel function from Step ((2)) to obtain the desired estimates of both the regression coefficients $\boldsymbol{\beta}$ and the covariance function parameters $\boldsymbol{\theta}$.

As to be illustrated in a simulation study, the estimate of $\boldsymbol{\beta}$ is not very sensitive to the choices of bandwidth and kernel function in Step ((1)). Thus, we suggest to use a pilot bandwidth to yield an underestimate of f and consequently an estimate of $\boldsymbol{\beta}$. In Steps ((2)) and ((3)), we use a bimodal kernel $K_2(u) = 2\pi^{-1/2}u^2\exp(-u^2)$; see Figure 3.2 [De Brabanter et al., 2011]. Unlike

the more commonly used kernels (e.g., Gaussian or Epanechnikov kernel), the bimodal kernel satisfies $K_2(0) = 0$, which can mitigate the influence of the spatio-temporal correlation.

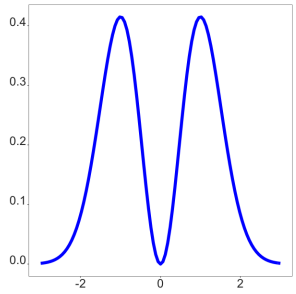

Figure 3.2: An example bimodal kernel function $K_2(u) = 2\pi^{-1/2}u^2\exp(-u^2)$.

A popular alternative to the cross-validation criterion (3.9) is the generalized cross-validation [GCV; Golub et al., 1979] criterion, in which S_{ii} is replaced by $N_n^{-1}\operatorname{tr}(\boldsymbol{S})$. For dependent data, Francisco-Fernandez and Opsomer [2005] proposed a bias-corrected generalized cross-validation criterion (GCV_c), replacing S_{ii} by $N_n^{-1}\operatorname{tr}(\boldsymbol{SR}(\boldsymbol{\theta}))$; that is,

$$\mathrm{GCV}_c(h) = \frac{\sum_{i=1}^{N_n}(\widetilde{y}_i^* - \widetilde{f}_i^*)^2}{N_n\{1 - N_n^{-1}\operatorname{tr}(\boldsymbol{SR}(\boldsymbol{\theta}))\}^2}, \tag{3.10}$$

where $\boldsymbol{R}(\boldsymbol{\theta})$ is a correlation matrix. In practice, a pilot estimate of the covariance parameter vector is required; however, the choice of such an estimate is not obvious, and would impact the overall estimation performance. To ensure the performance of parameter estimation in covariance function, for each candidate bandwidth h, we compute the corresponding estimate of $\boldsymbol{\theta}$ and obtain an estimated GCV_c criterion, denoted by GCV_{ce}. As further demonstrated in the simulation study, the results based on the cross-validation and GCV_{ce} are similar, although GCV_{ce} is computationally more expensive.

3.5 Simulation Study

3.5.1 Simulation 1: Finite sample properties

We sample N_s locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_s}$ uniformly from the spatial domain $[0,1]^2$, where $N_s = 20, 40, 60$. At these irregular sampling locations, realizations are partially missing from a temporal grid $t_1, \ldots, t_{N_t}$, where $t_i = (i-1/2)/1000$ for $i = 1, \ldots, 1000$, with a missing probability of 0.96. The space-time coordinates will remain fixed across iterations once generated.

For the regression mean function and the semiparametric mean function, the vector of regression coefficients is $\boldsymbol{\beta} = (4, 3, 2, 1)^\top$. The covariates are drawn (once) from a multivariate normal distribution with zero mean, unit variance, and a cross-covariate correlation of 0.5. Each covariate is standardized to have zero sample mean and unit sample variance. Further, the nonparametric temporal function in the semiparametric mean function is $f(t) = 2\{1 - \cos(2\pi t)\}$.

We then draw a realization from the mean zero Gaussian error process $\varepsilon(\boldsymbol{s}, t)$ using three different covariance functions. The first covariance function is an exponential spatio-temporal covariance function

$$\begin{aligned}
&\mathrm{Cov}\{\varepsilon(\boldsymbol{s}_i, t_i), \varepsilon(\boldsymbol{s}_j, t_j)\} \\
&= \begin{cases} \sigma^2(1-c)\exp\{-\varrho_{1,n}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|/c_s - \varrho_{2,n}|t_i - t_j|/c_t\}, & \text{if } i \neq j; \\ \sigma^2, & \text{if } i = j. \end{cases}
\end{aligned}$$

Here, σ^2 is the variance of $\varepsilon(\boldsymbol{s}, t)$, $c \in [0,1]$ is the proportion of random noise such that $c\sigma^2$ is the nugget effect, and c_s and c_t are the positive spatial and temporal range parameters, respectively. We take $\sigma^2 = 9.0, c = 0.2, c_s = 1$ and $c_t = 1$. This covariance function is stationary and separable and we denote it as `COV-1`.

Next, we consider a generalized spatio-temporal Matérn covariance function [Chu et al., 2019]:

$$\gamma_n((\boldsymbol{s},t),(\boldsymbol{s}',t');\boldsymbol{\theta}) =$$
$$\begin{cases} \frac{D(\boldsymbol{s},t)D(\boldsymbol{s}',t')\sigma^2\theta_3^{d/2}2^{1-\nu}}{(\theta_1^2u_2^2+1)^{\nu}(\theta_1^2u_2^2+\theta_3)^{d/2}\Gamma(\nu)}m(\boldsymbol{u}_1,u_2)^{\nu}K_{\nu}\left\{m(\boldsymbol{u}_1,u_2)\right\}, & \text{if } \|\boldsymbol{u}_1\|>0, \\ \frac{D(\boldsymbol{s},t)D(\boldsymbol{s}',t')\sigma^2\theta_3^{d/2}}{(\theta_1^2u_2^2+1)^{\nu}(\theta_1^2u_2^2+\theta_3)^{d/2}}, & \text{if } \|\boldsymbol{u}_1\|=0, |u_2|>0, \\ D(\boldsymbol{s},t)^2\sigma^2+c\sigma^2, & \text{if } \|\boldsymbol{u}_1\|=0, |u_2|=0, \end{cases} \tag{3.11}$$

where $\boldsymbol{u}_1 = \rho_{1,n}(\boldsymbol{s}'-\boldsymbol{s})$ and $u_2 = \rho_{2,n}(t'-t)$. In this covariance function, $m(\boldsymbol{u}_1,u_2) = \theta_2\left(\frac{\theta_1^2u_2^2+1}{\theta_1^2u_2^2+\theta_3}\right)^{1/2}\|\boldsymbol{u}_1\|$ and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν. Here, θ_1 and θ_2 are nonnegative range parameters of time and space respectively, $\theta_3 > 0$ is a separability parameter. The point-wise variance of $\varepsilon(\boldsymbol{s},t)$ is $D(\boldsymbol{s},t)D(\boldsymbol{s}',t')\sigma^2 + c\sigma^2$, where $c\sigma^2$ accounts for the nugget effect. The parameter ν in $K_{\nu}(\cdot)$ controls the smoothness of the covariance. If we let $\nu = 1/2$ and $\theta_3 = 1$, then (3.11) reduces to

$$\begin{aligned} &\mathrm{Cov}\{\varepsilon(\boldsymbol{s}_i,t_i),\varepsilon(\boldsymbol{s}_j,t_j)\} \\ &= \begin{cases} D(\boldsymbol{s}_i,t_i)D(\boldsymbol{s}_j,t_j)\frac{\sigma^2}{\{a^2|\varrho_{2,n}(t_i-t_j)|^2+1\}^{3/2}}\exp\{-b\varrho_{1,n}\|\boldsymbol{s}_i-\boldsymbol{s}_j\|\}, & \text{if } i\neq j; \\ D(\boldsymbol{s}_i,t_i)D(\boldsymbol{s}_j,t_j)\sigma^2+c\sigma^2, & \text{if } i=j. \end{cases} \end{aligned} \tag{3.12}$$

Similarly, σ^2 is the variance of $\varepsilon(\boldsymbol{s},t)$, $c\in[0,1]$ is the proportion of random noise such that $c\sigma^2$ is the nugget effect, and a and b are the positive temporal and spatial range parameters, respectively. Here, $D(\boldsymbol{s}_i,t_i) = dt_i+1$ varies by time, resulting in a nonstationarity covariance function. We set $\sigma^2 = 9, c = 0.2, a = 1, b = 1$ and $d = 1$. This covariance function is still separable in space and time, which we refer to as `COV-2`.

The third covariance function we considered is a slight modification of `COV-2`, with $D(\boldsymbol{s}_i,t_i) = dt_i + es_{1i} + fs_{2i} + 1$. We set $\sigma^2 = 9.0, c = 0.2, a = 1, b = 1, d = 0.5, e = 0.5$ and $f = 0.5$. This covariance function is both nonstationary and nonseparable, referred as `COV-3`.

For each combination of the sample size and the covariance function, we generate 400 simulation replicates. We also consider a special case of the semiparametric mean function where the temporal function f is assumed to be zero. As will be demonstrated later, this case will serve as a benchmark in the comparison of the estimation for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

For each simulated data set, a predetermined bandwidth $h_0 = 0.05$ is used to obtain an initial estimate of $\boldsymbol{\beta}$. The estimate of the optimal bandwidth $\widehat{h}$ is then determined by minimizing the cross-validation criterion (3.9) over a predetermined grid of bandwidth values. Given the estimated optimal bandwidth, the profile likelihood estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{f}(\cdot)$ are obtained. We further consider two variants of the GCV for determining the bandwidth in (3.9): GCV_c and GCV_{ce} as described in Section 3.4.

The profile likelihood method (PLE) results are compared with two alternative methods, namely, ALT_1 and ALT_2. In ALT_1, the parameter estimates and the estimate of the temporal function are obtained by the profile likelihood method ignoring the spatio-temporal dependence. In ALT_2, the regression coefficients $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\theta}$ are estimated by the classical maximum likelihood method assuming the temporal trend $f(\cdot)$ is known. That is, ALT_2 is essentially the maximum likelihood method under the model with the regression mean function.

To assess the performance of estimation by the different methods under the different bandwidth selection criteria, we compute the means and the standard deviations (SD) of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ from the 400 simulated data sets. We also compute the estimated standard errors of the parameters for each simulated data set based on the information matrix in Theorem 6 and report the mean estimated standard errors (SDm). For ALT_1, we use $\boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}^2 \boldsymbol{I}$ to calculate `SDm`. In addition, for the estimated temporal function $\widehat{f}$, we calculate the average squared error (ASE) for each simulated data set, defined as

$$\mathrm{ASE} = N_{\mathrm{grid}}^{-1} \sum_{i=1}^{N_{\mathrm{grid}}} \{f(t_{i,\mathrm{grid}}) - \widehat{f}(t_{i,\mathrm{grid}})\}^2,$$

where $t_{i,\mathrm{grid}} = (i - 1/2)/N_{\mathrm{grid}}$ for $i = 1, \ldots, N_{\mathrm{grid}}$ and $N_{\mathrm{grid}} = 1000$.

Finally, we generate an additional 10% new sampling locations and, at each new sampling location, new sampling time points are generated as in the simulation set-up. At these new sampling locations and time points, new observations denoted as $y_{i,\mathrm{new}}$ are generated and let $\widetilde{y}_{i,\mathrm{new}}$ denote the predicted value at the ith new sampling location and time, where $i = 1, \ldots, N_{\mathrm{new}}$, and N_{new} is the total number of new sampling locations and time points. We use the mean squared prediction error (MSPE) to evaluate the performance of the various methods as

$$\mathrm{MSPE} = N_{\mathrm{new}}^{-1} \sum_{i=1}^{N_{\mathrm{new}}} (y_{i,\mathrm{new}} - \widetilde{y}_{i,\mathrm{new}})^2,$$

The results are provided in Tables 3.5–3.6, the last two rows of which give the average values of ASE and MSPE.

As shown in Table 3.5 for the first scenario of the spatio-temporal covariance function (`COV-1`), the bandwidths chosen by the three selection criteria, CV, GCV_c and GCV_{ce}, are similar for the profile likelihood method. For parameter estimation, both the accuracy and the precision increase as the sample size increases. The empirical standard deviations are well approximated by the standard errors, supporting the information-based asymptotic variance in Theorem 6. Further, under different bandwidth selection criteria, similar ASE and MSPE values are obtained, which may not be surprising due to the similar choices of bandwidths and hence similar estimates.

For the estimation of the regression coefficients, our method PLE and the two alternative methods ALT_1 and ALT_2 have comparable estimation bias, which suggests that the accuracy of $\widehat{\boldsymbol{\beta}}$ is not sensitive to the assumption of covariance structure. However, the simulation standard deviations from ALT_1 are larger than those from PLE and ALT_2, indicating noticeable gain of statistical efficiency in the parameter estimation by accounting for spatio-temporal dependence. In addition, ALT_1 has much larger MSPE and thus poorer prediction than PLE and ALT_2. For estimating the temporal function, the ASEs for ALT_1 and PLE are similar; both decrease as the sample size increases.

When the sample size is smaller, PLE has less accuracy and precision in the estimation than ALT_2. In particular, both the standard deviations and the standard errors of the estimates from PLE are considerably larger than those of ALT_2, for all the covariance parameters except the nugget proportion c. When the sample size is larger, PLE and ALT_2 have similar estimation results. In particular, the standard deviations and the standard errors of the estimates from PLE are similar to ALT_2, supporting that the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ under the semiparametric mean function is the same as the regression mean function, as shown in Theorem 6. Moreover, ALT_2 has slightly better prediction than PLE due to possible bias in the estimation of $f(\cdot)$ in PLE.

Tables 3.1 and 3.6 show results for the second and the third scenario of the spatio-temporal covariance function, `COV-2` and `COV-3`, respectively. Similar conclusions can be drawn. Particularly, in the presence of non-separability and nonstationarity in the spatio-temporal covariance function, the finite-sample performance of the estimation for the semiparametric mean function is sound and supports the asymptotic results. The bandwidths selected by the three criteria, CV, GCV_c and GCV_{ce}, are very similar and so are the resulting estimates. Unlike `COV-1` and `COV-2`, the prediction under `COV-3` changes greatly for different sample sizes, which may be attributed to the nonstationarity in space with very different variances at different new spatial locations where the observations are predicted.

Tables 3.2–3.3 show that the regression coefficient estimates are robust against the choice of the kernel and the initial bandwidth. For a predetermined bandwidth, it can be seen that, different kernel functions in Step ((1)) of the bandwidth selection procedure yield very similar results. Moreover, those results are similar to the benchmark case when $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

As demonstrated in Theorem 8, bimodal kernels can effectively alleviate the influence of correlated errors on bandwidth selection. To see this, we compare the results from a bimodal kernel with those from a Gaussian kernel. The estimation results under the Gaussian kernel for the first scenario of spatio-temporal covariance function (`COV-1`) are given in Table 3.4. Unlike the bimodal kernel, the bandwidths selected by the three criteria, CV, GCV_c and GCV_{ce}, can be quite different. In particular, CV selects much smaller bandwidth than GCV_c and GCV_{ce}, supporting the fact that cross-validation does not handle correlation well for most commonly-used kernels with $K(0) \neq 0$. The regression coefficient estimates are similar for all three bandwidth selection criteria, which suggests that the estimation of $\boldsymbol{\beta}$ is not sensitive to choice of bandwidth. However, the estimates of covariance parameters are greatly affected by the bias in the bandwidth selection, with CV having the largest bias in parameter estimation and the largest ASE in the estimation of the temporal function, among the three criteria. As an alternative of CV in the presence of spatio-temporal correlation, GCV_c produces a much larger bandwidth, although the resulting estimates are not as accurate as those from the bimodal kernel function.

Table 3.1: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−2 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	—	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—
h	—	0.100	0.103	0.104	0.100	—	0.089	0.091	0.091	0.089	—	0.083	0.086	0.086	0.083	—
β_1	4.0	3.969	3.969	3.969	3.987	3.969	4.002	4.002	4.002	3.987	4.003	4.004	4.004	4.004	4.005	4.004
SD		0.136	0.137	0.137	0.197	0.137	0.102	0.102	0.102	0.169	0.102	0.089	0.089	0.089	0.129	0.089
SDm		0.146	0.146	0.146	0.209	0.146	0.104	0.104	0.104	0.151	0.104	0.086	0.086	0.086	0.123	0.086
β_2	3.0	3.020	3.020	3.020	3.027	3.023	3.010	3.010	3.010	3.008	3.012	2.994	2.994	2.994	2.983	2.994
SD		0.165	0.165	0.165	0.218	0.164	0.097	0.097	0.097	0.155	0.097	0.083	0.083	0.083	0.128	0.082
SDm		0.153	0.153	0.153	0.215	0.153	0.101	0.101	0.101	0.150	0.101	0.082	0.082	0.082	0.122	0.082
β_3	2.0	2.011	2.011	2.011	1.998	2.008	1.984	1.984	1.984	1.970	1.984	1.993	1.993	1.993	2.000	1.993
SD		0.146	0.146	0.146	0.201	0.146	0.100	0.100	0.100	0.136	0.100	0.088	0.088	0.088	0.130	0.088
SDm		0.152	0.152	0.152	0.209	0.152	0.106	0.106	0.106	0.154	0.106	0.082	0.082	0.082	0.121	0.082
β_4	1.0	1.003	1.003	1.003	0.983	1.004	1.002	1.002	1.002	1.021	1.002	1.002	1.002	1.002	1.000	1.002
SD		0.144	0.144	0.144	0.208	0.144	0.101	0.101	0.101	0.161	0.102	0.085	0.085	0.085	0.130	0.085
SDm		0.148	0.149	0.149	0.213	0.148	0.102	0.102	0.102	0.150	0.102	0.084	0.084	0.084	0.122	0.084
σ^2	9.0	9.075	9.073	9.071	22.751	8.934	9.198	9.199	9.199	23.316	9.094	9.163	9.159	9.159	22.736	9.085
SD		1.610	1.609	1.608	1.658	1.543	1.131	1.133	1.133	1.217	1.114	0.871	0.869	0.869	0.995	0.864
SDm		1.569	1.569	1.569	—	1.544	1.114	1.114	1.114	—	1.100	0.887	0.887	0.887	—	0.879
c	0.2	0.229	0.229	0.230	—	0.227	0.200	0.200	0.200	—	0.199	0.195	0.195	0.195	—	0.194
SD		0.121	0.121	0.121	—	0.120	0.060	0.060	0.060	—	0.061	0.051	0.051	0.051	—	0.051
SDm		0.102	0.102	0.102	—	0.103	0.061	0.061	0.061	—	0.062	0.049	0.049	0.049	—	0.050
a	1.0	0.980	0.980	0.980	—	0.996	0.991	0.991	0.991	—	1.002	1.002	1.002	1.002	—	1.010
SD		0.117	0.117	0.117	—	0.118	0.070	0.070	0.070	—	0.071	0.058	0.058	0.058	—	0.058
SDm		0.101	0.101	0.101	—	0.104	0.069	0.069	0.069	—	0.071	0.058	0.058	0.058	—	0.059
b	1.0	0.973	0.973	0.973	—	1.001	0.984	0.984	0.984	—	1.006	0.992	0.992	0.992	—	1.009
SD		0.139	0.139	0.139	—	0.140	0.089	0.089	0.089	—	0.089	0.070	0.070	0.070	—	0.070
SDm		0.131	0.131	0.131	—	0.135	0.086	0.086	0.086	—	0.087	0.069	0.069	0.069	—	0.070
d	1.0	1.020	1.020	1.020	—	1.021	0.999	0.999	0.999	—	1.000	0.991	0.992	0.992	—	0.992
SD		0.246	0.247	0.247	—	0.242	0.167	0.167	0.167	—	0.167	0.127	0.127	0.127	—	0.127
SDm		0.238	0.238	0.238	—	0.237	0.162	0.162	0.162	—	0.161	0.129	0.129	0.129	—	0.129
ASE	—	0.663	0.659	0.657	0.655	—	0.451	0.451	0.451	0.450	—	0.365	0.366	0.367	0.366	—
MSPE	—	13.454	13.456	13.456	24.114	13.429	15.646	15.645	15.646	21.682	15.609	15.629	15.629	15.629	24.123	15.598

Table 3.2: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of covariance parameters, and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ and three cases (Case II (K2+K2), Case IV (GK+K2) and the case when $\boldsymbol{\beta}$ is known in Step ((1)) of the bandwidth selection procedure) for COV−1.

Term	Case	$n_s = 20$			$n_s = 40$			$n_s = 60$		
		mean	SD	SDm	mean	SD	SDm	mean	SD	SDm
β_1	K2+K2	3.985	0.104	0.112	4.002	0.080	0.080	4.002	0.067	0.066
	GK+K2	3.985	0.104	0.112	4.002	0.080	0.080	4.002	0.067	0.066
	$\boldsymbol{\beta}$ known	3.985	0.104	0.112	4.002	0.080	0.080	4.002	0.067	0.066
β_2	K2+K2	3.017	0.123	0.116	3.009	0.075	0.078	2.996	0.065	0.065
	GK+K2	3.017	0.123	0.116	3.009	0.075	0.078	2.996	0.065	0.065
	$\boldsymbol{\beta}$ known	3.017	0.123	0.116	3.009	0.075	0.078	2.996	0.065	0.065
β_3	K2+K2	2.006	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
	GK+K2	2.006	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
	$\boldsymbol{\beta}$ known	2.006	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
β_4	K2+K2	0.996	0.110	0.115	1.002	0.077	0.079	1.000	0.065	0.065
	GK+K2	0.996	0.110	0.115	1.002	0.077	0.079	1.000	0.065	0.065
	$\boldsymbol{\beta}$ known	0.996	0.109	0.115	1.002	0.077	0.079	1.000	0.065	0.065
σ^2	K2+K2	9.111	0.569	0.546	9.120	0.387	0.395	9.055	0.325	0.319
	GK+K2	9.111	0.569	0.546	9.120	0.387	0.395	9.055	0.325	0.319
	$\boldsymbol{\beta}$ known	9.111	0.569	0.546	9.120	0.387	0.395	9.055	0.325	0.319
c	K2+K2	0.209	0.077	0.074	0.201	0.047	0.047	0.197	0.043	0.040
	GK+K2	0.209	0.077	0.074	0.201	0.047	0.047	0.197	0.043	0.040
	$\boldsymbol{\beta}$ known	0.209	0.077	0.074	0.201	0.047	0.047	0.197	0.043	0.040
c_s	K2+K2	1.090	0.224	0.213	1.052	0.134	0.127	1.027	0.103	0.099
	GK+K2	1.090	0.224	0.213	1.052	0.134	0.127	1.027	0.103	0.099
	$\boldsymbol{\beta}$ known	1.090	0.223	0.213	1.052	0.134	0.127	1.027	0.103	0.099
c_t	K2+K2	1.087	0.243	0.213	1.047	0.142	0.139	1.016	0.112	0.112
	GK+K2	1.087	0.243	0.213	1.047	0.142	0.139	1.016	0.112	0.112
	$\boldsymbol{\beta}$ known	1.087	0.243	0.213	1.047	0.142	0.139	1.016	0.112	0.112

Table 3.3: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of covariance parameters, and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ and three cases (Case I (GK+GK), Case III (K2+GK) and the case when $\boldsymbol{\beta}$ is known in Step ((1)) of the bandwidth selection procedure) for COV−1.

Term	Case	$n_s = 20$			$n_s = 40$			$n_s = 60$		
		mean	SD	SDm	mean	SD	SDm	mean	SD	SDm
β_1	GK+GK	3.985	0.104	0.112	4.003	0.080	0.080	4.002	0.067	0.066
	K2+GK	3.985	0.104	0.112	4.003	0.080	0.080	4.002	0.067	0.066
	$\boldsymbol{\beta}$ known	3.986	0.104	0.112	4.003	0.080	0.080	4.002	0.067	0.066
β_2	GK+GK	3.018	0.123	0.117	3.010	0.075	0.078	2.996	0.065	0.065
	K2+GK	3.018	0.123	0.117	3.010	0.075	0.078	2.996	0.065	0.065
	$\boldsymbol{\beta}$ known	3.018	0.123	0.117	3.010	0.075	0.078	2.996	0.065	0.065
β_3	GK+GK	2.005	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
	K2+GK	2.005	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
	$\boldsymbol{\beta}$ known	2.005	0.109	0.114	1.986	0.074	0.081	1.997	0.067	0.064
β_4	GK+GK	0.995	0.111	0.115	1.001	0.077	0.079	1.000	0.065	0.065
	K2+GK	0.995	0.111	0.115	1.001	0.077	0.079	1.000	0.065	0.065
	$\boldsymbol{\beta}$ known	0.995	0.111	0.115	1.001	0.077	0.079	1.000	0.065	0.065
σ^2	GK+GK	8.591	0.508	0.491	8.785	0.356	0.366	8.809	0.306	0.302
	K2+GK	8.591	0.508	0.491	8.785	0.356	0.366	8.809	0.306	0.302
	$\boldsymbol{\beta}$ known	8.591	0.508	0.491	8.785	0.356	0.366	8.809	0.306	0.302
c	GK+GK	0.179	0.081	0.086	0.184	0.048	0.051	0.182	0.044	0.043
	K2+GK	0.179	0.081	0.086	0.184	0.048	0.051	0.182	0.044	0.043
	$\boldsymbol{\beta}$ known	0.179	0.081	0.086	0.184	0.048	0.051	0.182	0.044	0.043
c_s	GK+GK	0.892	0.169	0.174	0.930	0.112	0.111	0.936	0.091	0.089
	K2+GK	0.892	0.169	0.174	0.930	0.112	0.111	0.936	0.091	0.089
	$\boldsymbol{\beta}$ known	0.892	0.169	0.174	0.930	0.112	0.111	0.936	0.091	0.089
c_t	GK+GK	0.880	0.193	0.179	0.929	0.122	0.125	0.928	0.099	0.103
	K2+GK	0.880	0.193	0.179	0.929	0.122	0.125	0.928	0.099	0.103
	$\boldsymbol{\beta}$ known	0.880	0.193	0.179	0.929	0.122	0.125	0.928	0.099	0.103

Table 3.4: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using Gaussian kernel for COV−1 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$n_s = 20$					$n_s = 40$					$n_s = 60$				
Method	—	PLM			ALT_1	ALT_2	PLM			ALT_1	ALT_2	PLM			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—
BW	—	0.050	0.084	0.071	0.050	—	0.050	0.076	0.067	0.050	—	0.050	0.070	0.064	0.050	—
β_1	4.0	3.985	3.985	3.985	3.998	3.985	4.003	4.003	4.003	4.001	4.002	4.002	4.002	4.002	4.003	4.002
SD		0.104	0.104	0.104	0.117	0.104	0.080	0.080	0.080	0.098	0.080	0.067	0.067	0.067	0.080	0.067
SDm		0.112	0.112	0.112	0.129	0.112	0.080	0.080	0.080	0.093	0.080	0.066	0.066	0.066	0.077	0.066
β_2	3.0	3.018	3.017	3.017	3.017	3.018	3.010	3.010	3.010	3.009	3.010	2.996	2.996	2.996	2.992	2.996
SD		0.123	0.123	0.123	0.134	0.122	0.075	0.075	0.075	0.091	0.075	0.065	0.065	0.065	0.077	0.065
SDm		0.117	0.117	0.117	0.134	0.117	0.078	0.078	0.078	0.092	0.078	0.065	0.065	0.065	0.076	0.065
β_3	2.0	2.005	2.005	2.005	2.000	2.005	1.986	1.986	1.986	1.981	1.986	1.997	1.997	1.997	2.003	1.997
SD		0.109	0.109	0.109	0.120	0.109	0.074	0.074	0.074	0.084	0.074	0.067	0.067	0.067	0.079	0.067
SDm		0.114	0.114	0.114	0.130	0.114	0.081	0.081	0.081	0.095	0.081	0.064	0.064	0.064	0.076	0.064
β_4	1.0	0.995	0.995	0.995	0.985	0.996	1.001	1.002	1.001	1.002	1.002	1.000	1.000	1.000	0.997	1.000
SD		0.111	0.111	0.111	0.130	0.110	0.077	0.077	0.077	0.091	0.077	0.065	0.065	0.065	0.079	0.065
SDm		0.115	0.115	0.115	0.131	0.115	0.079	0.079	0.079	0.093	0.079	0.065	0.065	0.065	0.076	0.065
σ^2	9.0	8.591	8.705	8.654	8.542	8.944	8.785	8.845	8.821	8.722	8.997	8.809	8.845	8.832	8.758	8.963
SD		0.508	0.521	0.533	0.504	0.525	0.356	0.362	0.367	0.347	0.372	0.306	0.306	0.310	0.301	0.317
SDm		0.491	0.502	0.497	—	0.528	0.366	0.371	0.369	—	0.384	0.302	0.304	0.303	—	0.313
c	0.2	0.179	0.187	0.183	—	0.202	0.184	0.187	0.186	—	0.196	0.182	0.185	0.184	—	0.193
SD		0.081	0.080	0.080	—	0.078	0.048	0.048	0.048	—	0.047	0.044	0.044	0.044	—	0.043
SDm		0.086	0.083	0.084	—	0.077	0.051	0.050	0.051	—	0.048	0.043	0.043	0.043	—	0.041
c_s	1.0	0.892	0.934	0.915	—	1.025	0.930	0.951	0.942	—	1.007	0.936	0.950	0.944	—	0.994
SD		0.169	0.176	0.178	—	0.198	0.112	0.113	0.116	—	0.124	0.091	0.092	0.094	—	0.098
SDm		0.174	0.182	0.178	—	0.200	0.111	0.114	0.113	—	0.121	0.089	0.090	0.090	—	0.095
c_t	1.0	0.880	0.926	0.905	—	1.029	0.929	0.950	0.942	—	1.008	0.928	0.942	0.937	—	0.987
SD		0.193	0.202	0.203	—	0.224	0.122	0.125	0.127	—	0.134	0.099	0.100	0.101	—	0.108
SDm		0.179	0.186	0.183	—	0.204	0.125	0.127	0.126	—	0.134	0.103	0.104	0.104	—	0.109
ASE	—	0.240	0.208	0.223	0.240	—	0.158	0.140	0.147	0.158	—	0.123	0.113	0.118	0.123	—
MSPE	—	6.507	6.503	6.506	8.795	6.821	7.405	7.401	7.402	8.885	7.932	6.931	6.930	6.931	8.875	7.337

Table 3.5: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE) and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−1 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	—	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—
h	—	0.079	0.081	0.082	0.079	—	0.072	0.073	0.073	0.072	—	0.068	0.069	0.069	0.068	—
β_1	4.0	3.985	3.985	3.985	3.991	3.985	4.002	4.002	4.002	3.993	4.002	4.002	4.002	4.002	4.003	4.002
SD		0.104	0.104	0.104	0.120	0.104	0.080	0.080	0.080	0.104	0.080	0.067	0.067	0.067	0.082	0.067
SDm		0.112	0.112	0.112	0.132	0.112	0.080	0.080	0.080	0.095	0.080	0.066	0.066	0.066	0.078	0.066
β_2	3.0	3.017	3.017	3.017	3.017	3.018	3.009	3.009	3.009	3.004	3.010	2.996	2.996	2.996	2.989	2.996
SD		0.123	0.123	0.123	0.140	0.122	0.075	0.075	0.075	0.095	0.075	0.065	0.065	0.065	0.079	0.065
SDm		0.116	0.116	0.116	0.136	0.117	0.078	0.078	0.078	0.094	0.078	0.065	0.065	0.065	0.077	0.065
β_3	2.0	2.006	2.006	2.006	2.003	2.005	1.986	1.986	1.986	1.982	1.986	1.997	1.997	1.997	2.001	1.997
SD		0.109	0.109	0.109	0.127	0.109	0.074	0.074	0.074	0.087	0.074	0.067	0.067	0.067	0.082	0.067
SDm		0.114	0.114	0.114	0.132	0.114	0.081	0.081	0.081	0.096	0.081	0.064	0.064	0.064	0.077	0.064
β_4	1.0	0.996	0.996	0.995	0.988	0.996	1.002	1.002	1.002	1.010	1.002	1.000	1.000	1.000	0.999	1.000
SD		0.110	0.110	0.110	0.131	0.110	0.077	0.077	0.077	0.097	0.077	0.065	0.065	0.065	0.082	0.065
SDm		0.115	0.115	0.115	0.134	0.115	0.079	0.079	0.079	0.094	0.079	0.065	0.065	0.065	0.077	0.065
σ^2	9.0	9.111	9.112	9.112	9.101	8.944	9.120	9.120	9.120	9.113	8.997	9.055	9.055	9.055	9.054	8.963
SD		0.569	0.569	0.569	0.589	0.525	0.387	0.387	0.387	0.394	0.372	0.325	0.324	0.324	0.334	0.317
SDm		0.546	0.547	0.547	—	0.528	0.395	0.395	0.395	—	0.384	0.319	0.319	0.319	—	0.313
c	0.2	0.209	0.209	0.209	—	0.202	0.201	0.201	0.201	—	0.196	0.197	0.197	0.197	—	0.193
SD		0.077	0.077	0.077	—	0.078	0.047	0.047	0.048	—	0.047	0.043	0.043	0.043	—	0.043
SDm		0.074	0.074	0.074	—	0.077	0.047	0.047	0.047	—	0.048	0.040	0.040	0.040	—	0.041
c_s	1.0	1.090	1.090	1.091	—	1.025	1.052	1.052	1.052	—	1.007	1.027	1.027	1.027	—	0.994
SD		0.224	0.224	0.224	—	0.198	0.134	0.134	0.134	—	0.124	0.103	0.103	0.103	—	0.098
SDm		0.213	0.213	0.213	—	0.200	0.127	0.127	0.127	—	0.121	0.099	0.099	0.099	—	0.095
c_t	1.0	1.087	1.088	1.088	—	1.029	1.047	1.048	1.048	—	1.008	1.016	1.016	1.016	—	0.987
SD		0.243	0.244	0.245	—	0.224	0.142	0.142	0.142	—	0.134	0.112	0.112	0.112	—	0.108
SDm		0.213	0.213	0.213	—	0.204	0.139	0.139	0.139	—	0.134	0.112	0.112	0.112	—	0.109
ASE	—	0.266	0.265	0.266	0.265	—	0.183	0.183	0.183	0.182	—	0.149	0.149	0.150	0.149	—
MSPE	—	6.501	6.501	6.501	9.245	6.484	7.403	7.403	7.403	9.169	7.386	6.934	6.933	6.933	9.090	6.918

Table 3.6: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−3 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	—	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—
h	—	0.106	0.108	0.109	0.106	—	0.095	0.098	0.099	0.095	—	0.087	0.090	0.090	0.087	—
β_1	4.0	3.964	3.963	3.963	3.984	3.963	4.003	4.003	4.003	3.984	4.003	4.005	4.005	4.005	4.004	4.005
SD		0.154	0.154	0.154	0.220	0.154	0.114	0.114	0.114	0.193	0.114	0.098	0.098	0.098	0.148	0.098
SDm		0.165	0.165	0.165	0.238	0.165	0.116	0.116	0.116	0.173	0.116	0.096	0.096	0.096	0.139	0.096
β_2	3.0	3.025	3.025	3.025	3.029	3.027	3.014	3.014	3.014	3.011	3.015	2.992	2.992	2.992	2.983	2.992
SD		0.184	0.184	0.184	0.248	0.183	0.107	0.107	0.107	0.181	0.107	0.092	0.092	0.092	0.144	0.092
SDm		0.171	0.171	0.171	0.246	0.172	0.112	0.112	0.112	0.171	0.112	0.092	0.092	0.092	0.138	0.092
β_3	2.0	2.014	2.014	2.014	1.994	2.011	1.981	1.981	1.981	1.968	1.981	1.991	1.991	1.991	2.002	1.991
SD		0.162	0.162	0.162	0.230	0.161	0.111	0.111	0.111	0.158	0.111	0.097	0.097	0.097	0.148	0.097
SDm		0.172	0.172	0.172	0.238	0.172	0.117	0.117	0.117	0.176	0.117	0.092	0.092	0.092	0.137	0.092
β_4	1.0	1.002	1.002	1.002	0.981	1.003	1.002	1.002	1.002	1.023	1.002	1.004	1.004	1.004	0.999	1.004
SD		0.163	0.163	0.163	0.238	0.163	0.114	0.114	0.114	0.177	0.114	0.096	0.096	0.096	0.145	0.095
SDm		0.168	0.168	0.168	0.242	0.168	0.112	0.112	0.112	0.172	0.112	0.093	0.093	0.093	0.137	0.093
σ^2	9.0	9.322	9.328	9.330	29.520	9.066	9.415	9.418	9.419	30.491	9.275	9.366	9.363	9.363	29.027	9.262
SD		2.418	2.417	2.418	2.115	2.374	1.757	1.758	1.756	1.552	1.728	1.409	1.408	1.408	1.228	1.393
SDm		2.397	2.398	2.398	—	2.343	1.725	1.725	1.725	—	1.700	1.431	1.431	1.431	—	1.416
c	0.2	0.238	0.238	0.238	—	0.240	0.202	0.202	0.202	—	0.201	0.193	0.193	0.193	—	0.192
SD		0.151	0.151	0.151	—	0.155	0.077	0.077	0.077	—	0.077	0.062	0.062	0.062	—	0.062
SDm		0.134	0.134	0.134	—	0.138	0.075	0.075	0.075	—	0.076	0.061	0.061	0.061	—	0.061
a	1.0	0.982	0.981	0.981	—	0.996	0.990	0.989	0.989	—	1.000	1.002	1.002	1.002	—	1.009
SD		0.115	0.115	0.115	—	0.116	0.067	0.067	0.067	—	0.068	0.055	0.055	0.055	—	0.056
SDm		0.098	0.098	0.098	—	0.101	0.066	0.066	0.066	—	0.067	0.056	0.056	0.056	—	0.056
b	1.0	0.977	0.977	0.977	—	1.001	0.986	0.986	0.985	—	1.005	0.993	0.993	0.993	—	1.008
SD		0.135	0.135	0.135	—	0.135	0.087	0.087	0.087	—	0.087	0.067	0.067	0.067	—	0.067
SDm		0.127	0.127	0.127	—	0.130	0.083	0.083	0.083	—	0.084	0.067	0.067	0.067	—	0.068
d	0.5	0.509	0.509	0.508	—	0.513	0.498	0.498	0.498	—	0.500	0.490	0.490	0.491	—	0.491
SD		0.222	0.222	0.222	—	0.221	0.158	0.158	0.158	—	0.158	0.117	0.117	0.117	—	0.116
SDm		0.222	0.222	0.222	—	0.224	0.150	0.150	0.150	—	0.150	0.120	0.120	0.120	—	0.120
e	0.5	0.502	0.501	0.501	—	0.515	0.497	0.497	0.497	—	0.500	0.483	0.483	0.483	—	0.486
SD		0.223	0.223	0.223	—	0.228	0.140	0.141	0.140	—	0.142	0.119	0.119	0.119	—	0.120
SDm		0.217	0.217	0.217	—	0.220	0.145	0.145	0.145	—	0.146	0.118	0.118	0.118	—	0.118
f	0.5	0.514	0.514	0.514	—	0.524	0.487	0.487	0.487	—	0.491	0.493	0.493	0.493	—	0.496
SD		0.193	0.193	0.193	—	0.198	0.151	0.151	0.151	—	0.153	0.116	0.116	0.116	—	0.117
SDm		0.197	0.197	0.197	—	0.200	0.151	0.151	0.151	—	0.151	0.115	0.115	0.115	—	0.115
ASE	—	0.825	0.828	0.829	0.823	—	0.577	0.577	0.578	0.576	—	0.450	0.447	0.447	0.450	—
MSPE	—	13.107	13.103	13.103	23.673	13.062	20.457	20.455	20.455	28.599	20.412	22.368	22.369	22.369	34.789	22.336

Finally, in Tables 3.5–3.6, the standard deviations and the standard errors of $\widehat{\boldsymbol{\beta}}$ from PLE are similar to those from ALT_2. This is as expected, since the design matrix in our setting does not vary by time.

3.5.2 Simulation 2: Design Matrix Varying with Time

In this section, we investigate a design matrix that varies over time. In particular, we assume that $\boldsymbol{X}$ consists of two parts $\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2$. Here, $\boldsymbol{X}_1$ is the same as the design matrix in Section 3.5 and $\boldsymbol{X}_2 = (3\boldsymbol{t}, 9\boldsymbol{t}^2, 27\boldsymbol{t}^3, 81\boldsymbol{t}^4)$, where $\boldsymbol{t} = (t_1, \ldots, t_{N_n})^\top$. The results are presented in Table 3.7.

Table 3.7: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters for sample size with $n_s = 20$ using bimodal kernel for COV-1 where design matrix varies with time.

Term	Truth	$n_s = 20$				
Method	—	PLM			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—
BW	—	0.075	0.077	0.077	0.079	—
β_1	4.0	3.988	3.988	3.988	4.003	3.986
SD		0.105	0.104	0.105	0.119	0.101
SDm		0.112	0.111	0.111	0.130	0.106
β_2	3.0	3.018	3.018	3.018	3.021	3.016
SD		0.119	0.119	0.119	0.130	0.112
SDm		0.112	0.111	0.111	0.128	0.106
β_3	2.0	2.002	2.001	2.001	1.977	2.000
SD		0.109	0.109	0.109	0.127	0.104
SDm		0.114	0.114	0.114	0.131	0.104
β_4	1.0	0.993	0.993	0.993	0.985	0.999
SD		0.082	0.081	0.081	0.069	0.032
SDm		0.087	0.086	0.086	0.072	0.034
σ^2	9.0	9.117	9.116	9.116	9.097	8.915
SD		0.570	0.570	0.569	0.585	0.526
SDm		0.547	0.547	0.547	—	0.524
c	0.2	0.209	0.209	0.209	—	0.201
SD		0.077	0.077	0.077	—	0.078
SDm		0.074	0.074	0.074	—	0.078
c_s	1.0	1.092	1.091	1.091	—	1.010
SD		0.223	0.223	0.223	—	0.194
SDm		0.214	0.214	0.214	—	0.197
c_t	1.0	1.086	1.087	1.087	—	1.012
SD		0.243	0.243	0.243	—	0.220
SDm		0.213	0.213	0.213	—	0.201

From Table 3.7, it can be seen that the standard deviations of $\widehat{\boldsymbol{\beta}}$ from PLE are larger than those from ALT_2, which indicates a loss of statistical efficiency in the estimation of $\boldsymbol{\beta}$ when the

unknown temporal function is estimated. This finding is consistent with the standard error formula in Theorem 6.

3.5.3 Simulation 3: Nonseparable and Stationary Covariance Function

In this section, we consider a nonseparable but stationary covariance function. Consider the generalized spatio-temporal Matérn covariance function given in (3.11). Let the smoothness parameter be $\nu = 1/2$. Then, (3.11) is simplified to

$$\begin{aligned}
&\mathrm{Cov}\{\epsilon(\boldsymbol{s}_i, t_i), \epsilon(\boldsymbol{s}_j, t_j)\} \\
&= \begin{cases} \frac{D(\boldsymbol{s}_i,t_i)D(\boldsymbol{s}_j,t_j)\sigma^2(1-c)\theta_3 \exp\{-\theta_2\left(\frac{\theta_1^2|\varrho_{2,n}(t_i-t_j)|^2+1}{\theta_1^2|\varrho_{2,n}(t_i-t_j)|^2+\theta_3}\right)^{1/2}\varrho_{1,n}\|\boldsymbol{s}_i-\boldsymbol{s}_j\|\}}{(\theta_1^2|\varrho_{2,n}(t_i-t_j)|^2+1)^{1/2}(\theta_1^2|\varrho_{2,n}(t_i-t_j)|^2+\theta_3)}, & \text{if } i \neq j; \\ D(\boldsymbol{s}_i, t_i)D(\boldsymbol{s}_j, t_j)\sigma^2 + c\sigma^2, & \text{if } i = j. \end{cases}
\end{aligned}$$

In addition, we let $(\sigma^2, c, \theta_1, \theta_2, \theta_3) = (9, 0.2, 1, 1, 4)$ and $D(\boldsymbol{s}, t) = 1$. The simulation results are given in Table 3.8. Note that, similar lessons as previous section can be learned.

3.5.4 Simulation 4: Choice of Kernel Functions and Initial Bandwidth

In Section 3.4, a three-step procedure is proposed for bandwidth selection. In Step ((1)), an initial choice of kernel function as well as bandwidth is needed to obtain a pilot estimate of $\boldsymbol{\beta}$. Here, we will demonstrate that the estimates of regression coefficients are not sensitive to the choice of kernel functions and initial bandwidth.

To illustrate this empirically, we choose different kernel functions in Step ((1)) and in Steps ((2))-((3)) of the bandwidth selection procedure. Accordingly, we consider four cases: (I) `GK + GK`, (II) `K2 + K2`, (III) `K2 + GK` and (IV) `GK + K2`, where `GK` is the Gaussian kernel and `K2` is the bimodal kernel as depicted in Figure 3.2. The results for Case (II) and Case (IV) are summarized in Table 3.2, and results for Case (I) and Case (III) are given in Table 3.3. As a comparison, we consider a scenario where $\boldsymbol{\beta}$ is assumed known in Step ((1)). Essentially, only Steps ((2))-((3)) are needed.

Table 3.8: Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−1 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$n_s = 20$					$n_s = 40$					$n_s = 60$				
Method	—	PLM			ALT_1	ALT_2	PLM			ALT_1	ALT_2	PLM			ALT_1	ALT_2
Criteria	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—	CV	GCV_c	GCV_{ce}	CV	—
BW	—	0.093	0.098	0.099	0.093	—	0.083	0.087	0.088	0.083	—	0.079	0.083	0.083	0.079	—
β_1	4.0	3.985	3.985	3.985	3.988	3.984	4.000	4.000	4.000	3.993	4.000	4.001	4.001	4.001	4.001	4.001
SD		0.090	0.090	0.090	0.131	0.090	0.067	0.067	0.067	0.118	0.067	0.058	0.058	0.058	0.092	0.058
SDm		0.097	0.097	0.097	0.145	0.097	0.068	0.068	0.068	0.104	0.068	0.056	0.056	0.056	0.085	0.056
β_2	3.0	3.017	3.017	3.017	3.016	3.018	3.008	3.007	3.007	3.006	3.008	2.997	2.997	2.997	2.987	2.997
SD		0.109	0.109	0.109	0.155	0.109	0.064	0.064	0.064	0.112	0.064	0.054	0.054	0.054	0.088	0.054
SDm		0.101	0.101	0.101	0.149	0.101	0.066	0.066	0.066	0.103	0.066	0.055	0.055	0.055	0.085	0.055
β_3	2.0	2.004	2.004	2.004	1.999	2.003	1.989	1.989	1.989	1.983	1.989	1.996	1.996	1.996	2.003	1.997
SD		0.095	0.095	0.095	0.143	0.095	0.064	0.064	0.064	0.097	0.064	0.057	0.057	0.057	0.096	0.056
SDm		0.099	0.099	0.099	0.145	0.099	0.069	0.069	0.069	0.106	0.069	0.054	0.054	0.054	0.084	0.054
β_4	1.0	0.998	0.998	0.998	0.989	0.998	1.002	1.002	1.002	1.007	1.002	1.001	1.001	1.001	1.000	1.001
SD		0.094	0.094	0.094	0.148	0.094	0.066	0.066	0.066	0.113	0.067	0.056	0.056	0.056	0.093	0.056
SDm		0.099	0.099	0.099	0.147	0.099	0.067	0.067	0.067	0.103	0.067	0.055	0.055	0.055	0.084	0.055
σ^2	9.0	9.235	9.232	9.232	10.955	8.990	9.228	9.227	9.228	10.989	9.034	9.142	9.141	9.140	10.888	8.984
SD		0.949	0.948	0.949	0.995	0.871	0.693	0.694	0.695	0.743	0.661	0.571	0.570	0.569	0.628	0.538
SDm		0.890	0.889	0.889	—	0.857	0.670	0.670	0.670	—	0.649	0.560	0.560	0.560	—	0.545
c	0.2	0.198	0.198	0.198	—	0.201	0.195	0.195	0.195	—	0.198	0.196	0.196	0.196	—	0.198
SD		0.045	0.045	0.045	—	0.046	0.029	0.029	0.029	—	0.029	0.024	0.024	0.024	—	0.024
SDm		0.043	0.043	0.043	—	0.045	0.028	0.028	0.028	—	0.029	0.024	0.024	0.024	—	0.024
θ_1	1.0	0.925	0.924	0.924	—	0.967	0.985	0.984	0.984	—	1.011	0.981	0.981	0.981	—	1.001
SD		0.230	0.231	0.231	—	0.208	0.136	0.136	0.136	—	0.127	0.104	0.104	0.104	—	0.101
SDm		0.217	0.218	0.218	—	0.214	0.128	0.128	0.128	—	0.130	0.103	0.103	0.103	—	0.104
θ_2	1.0	0.962	0.962	0.962	—	1.039	1.005	1.005	1.005	—	1.066	0.969	0.969	0.969	—	1.020
SD		0.421	0.420	0.421	—	0.415	0.291	0.291	0.291	—	0.286	0.231	0.231	0.231	—	0.232
SDm		0.441	0.441	0.441	—	0.473	0.285	0.285	0.285	—	0.303	0.221	0.221	0.221	—	0.232
θ_3	4.0	4.525	4.525	4.526	—	4.827	4.463	4.461	4.460	—	4.731	3.988	3.986	3.986	—	4.211
SD		3.112	3.111	3.111	—	3.093	2.223	2.225	2.226	—	2.232	1.659	1.660	1.660	—	1.699
SDm		3.816	3.816	3.816	—	4.109	2.304	2.303	2.303	—	2.462	1.634	1.634	1.634	—	1.738
ASE	—	0.525	0.522	0.524	0.523	—	0.413	0.411	0.413	0.411	—	0.360	0.355	0.355	0.358	—
MSPE	—	4.915	4.915	4.915	11.206	4.908	5.981	5.981	5.981	11.108	5.972	5.411	5.411	5.411	10.934	5.403

We notice that parameter estimates are very similar, including averages and standard deviations of parameter estimates as well as the mean values of estimated standard deviations. This finding suggests that the parameter estimates are robust against the choices of kernel function and bandwidth in Step ((1)).

On comparing Table 3.2 to Table 3.3, an interesting finding is that bimodal kernel `K2` is a better choice than the Gaussian kernel `GK` in the estimation of the covariance parameters σ^2, c, c_s and c_t. As noted in Section 3.4 and further demonstrated in Theorem 8, the bandwidth choice from Gaussian kernel is biased due to correlated errors. To see this, we further investigate the difference in bandwidth selection between Gaussian kernel and bimodal kernel. Particularly, we calculate the values of three criteria, $\mathrm{CV}(h)$, $\mathrm{GCV}_c(h)$ and $\mathrm{GCV}_{ce}(h)$, for bandwidth h taking values on a grid from 0.05 to 0.2. The results using bimodal kernel and Gaussian kernel are shown in the left panel and right panel of Figure 3.3, respectively.

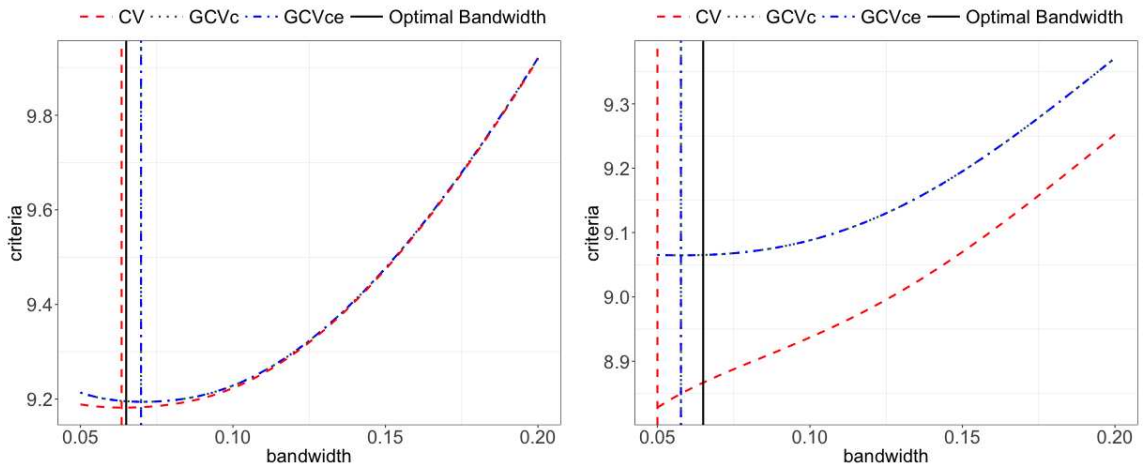

Figure 3.3: Comparison of bandwidth selection under bimodal kernel and Gaussian kernel for the first covariance function `COV-1` with $N_s = 40$. Three bandwidth selection criteria are depicted in different line types (dashed line: CV; dotted line: GCV_c; dot-dashed line: GCV_{ce}). The optimal bandwidth is represented as the vertical solid line.

In the left panel, we can see that, for bimodal kernel, $\mathrm{CV}(h)$, $\mathrm{GCV}_c(h)$ and $\mathrm{GCV}_{ce}(h)$ are very close. The bandwidths chosen by three criteria are depicted by vertical lines. As a comparison, we obtain the optimal bandwidth using (3.7) by minimizing $\mathrm{AMISE}(h)$, shown as a vertical line in

solid line type. The bandwidth from CV is very close to the optimal one, which suggests that the bimodal kernel can effectively eliminate the influence of correlated errors. Bandwidths from GCV_c and GCV_{ce} are similar, and both larger than the optimal one. In the right panel, different lesson has been learned for Gaussian kernel. We can see that bandwidths from CV, GCV_c and GCV_{ce} are very different from the optimal one. In fact, they are much smaller. It is worth mentioning that $\mathrm{GCV}_c(h)$ and $\mathrm{GCV}_{ce}(h)$ are nearly identical. For completeness, we show the full results of Gaussian kernel for `COV-1` in Table 3.4.

3.6 Data Example

To illustrate our methodology, we consider a data set collected by static sensors at fixed sampling locations in time and roving sensors traversing the spatial domain in time in an engine facility for evaluating the intensity level of noise as an occupational hazard [Lake et al., 2015, Ludwig et al., 2017]. We focus on the observations between 10:29:00 am and 11:24:00 am when all the sensors are operating. As shown in Figure 3.4, there are 56 observations, one per minute, for each of the 17 static sensor, whereas there are a total of 179 observations, at irregular time points, for the two roving sensors.

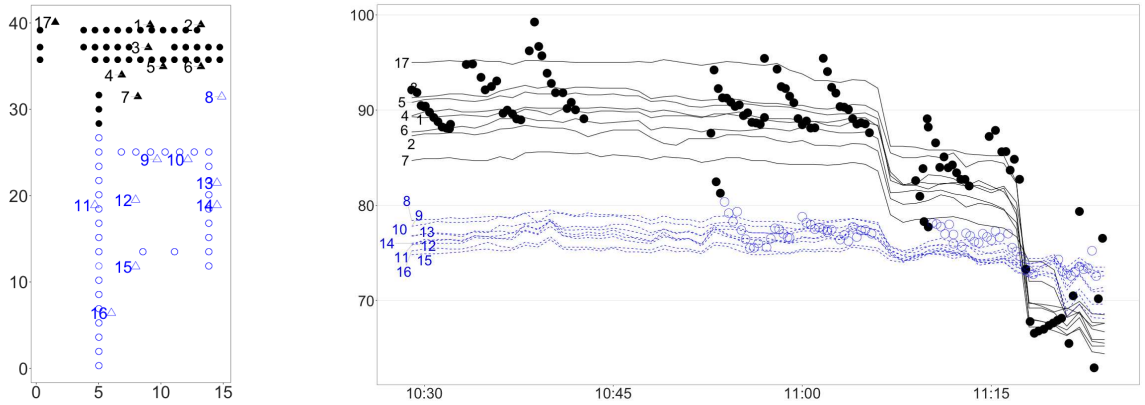

Figure 3.4: Left panel: Locations of static and roving sensors (▲: static sensors in group 1, △: static sensors in group 2, •: roving sensors closer to static sensors in group 1, and ○: roving sensors closer to static sensors in group 2). Right panel: noise intensity over time at all static and roving sensors. Here, time series for static sensors in Group 1 are shown in solid line, and those from Group 2 static sensors are shown in dashed line. In addition, measurements of roving sensors recorded near Group 1 sensors are shown in dark solid circles, otherwise, they are shown as open circles.

The left panel of Figure 3.4 shows the locations of the 17 static sensors (in triangles) and the locations of the two roving sensors (in circles). The right panel of Figure 3.4 plots the noise intensity over time at the static sensors. There appears a clear difference of noise intensity between the static sensors (#1 through #7, and #17, denoted as Group 1, solid triangles in Figure 3.4) near the upper-left corner where the active engine is located, and those further away from the active engine (#8 through #16, denoted as Group 2, as open triangles in Figure 3.4) before 11:10:00 am. Accordingly, the noise intensity measurements observed at Group 1 sensors are shown in solid lines, and those from Group 2 sensors are shown in dashed lines. In addition, the noise intensity of roving sensors are presented as circles in Figure 3.4; particularly, solid circles for the measurements near Group 1 sensors, and open circles otherwise.

We consider the semiparameteric mean function (3.12) with the generalized spatio-temporal Matérn error covariance function (3.12). More specifically, for $\boldsymbol{s} = (s_1, s_2)$, we have

$$y(\boldsymbol{s}, t) = \beta_1 s_1 + \beta_2 s_2 + f(t) + \varepsilon(\boldsymbol{s}, t),$$

where the regression is on the coordinates of the spatial location $\boldsymbol{s}$, the temporal function f is nonparametric, and the zero-mean error process $\varepsilon(\boldsymbol{s}, t)$ has the spatio-temporal covariance function (3.12). We fit three spatio-temporal covariance functions: $D_1(\boldsymbol{s}, t) = 1$ for stationarity, $D_2(\boldsymbol{s}, t) = 1 + dt$ and $D_3(\boldsymbol{s}, t) = 1 + dt + e(t - \kappa)_+$ for nonstationary. In the latter two nonstationary cases, for any fixed time point t_0, $\varepsilon(\boldsymbol{s}, t_0)$ is spatially stationary. For $D_3(\boldsymbol{s}, t)$, κ is chosen around 11:02:00 am, which is expected to capture the temporal change due to an engine shutdown.

We apply our method to analyze this data set and summarize the parameter estimates of β_1, β_2 and $\boldsymbol{\theta}$ in Table 3.9, whereas the estimated temporal function $\widehat{f}(t)$ and the pointwise 95% confidence intervals are plotted in Figure 3.5. We approximate the pointwise standard deviation of $\widehat{f}(t)$ by Theorem 7. The temporal function estimates $\widehat{f}(t)$ under the three models D_1, D_2, and D_3 are quite similar; however, the pointwise confidence interval based on D_1 is much wider than those based on D_2. For D_3, the pointwise confidence interval is much narrower than D_1 and D_2 when t is small, however it is unusually large when t is large.

Table 3.9: Selected bandwidths using bimodal kernel and corresponding parameter estimates for four covariance structures: $D_1(\boldsymbol{s},t)=1$, $D_2(\boldsymbol{s},t)=dt+1$ and $D_3(\boldsymbol{s},t)=dt+e(t-\kappa)_+ +1$. Standard errors are computed based on information matrices from Theorem 6 and given in paratheses.

	$D_1(\boldsymbol{s},t)$	$D_2(\boldsymbol{s},t)$	$D_3(\boldsymbol{s},t)$	$D_3(\boldsymbol{s},t)$ (penalized)
h	0.0193	0.0193	0.0193	0.0193
		Regression parameters		
β_1	-0.3922 (0.0820)	-0.4492 (0.0652)	-0.4872 (0.0608)	-0.4600 (0.0636)
β_2	0.3015 (0.0565)	0.4048 (0.0440)	0.4142 (0.0410)	0.3954 (0.0425)
		Covariance parameters		
σ^2	50.8840 (8.7442)	8.8096 (1.7254)	19.0058 (3.6086)	14.7628 (2.7797)
c	0.0007 (0.0001)	0.0020 (0.0005)	0.0007 (0.0002)	0.0009 (0.0002)
c_s	0.1662 (0.0040)	0.1677 (0.0037)	0.1647 (0.0035)	0.1723 (0.0038)
c_t	0.0152 (0.0025)	0.0215 (0.0033)	0.0201 (0.0031)	0.0237 (0.0036)
d	—	1.9218 (0.2647)	-0.3070 (0.1298)	0.1982 (0.1723)
e	—	—	6.5719 (0.4883)	3.0394 (0.4177)

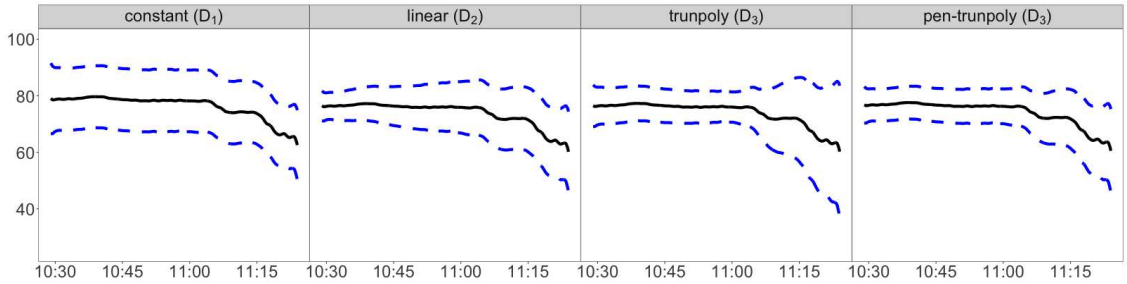

Figure 3.5: Estimated temporal function $\widehat{f}(t)$ (solid curve) and 95% pointwise confidence intervals (dash curves) by maximizing the profile-likelihood (3.6) with four covariance structures: constant $D_1(\boldsymbol{s},t)=1$; linear $D_2(\boldsymbol{s},t)=dt+1$; truncated polynomial $D_3(\boldsymbol{s},t)=dt+e(t-\kappa)_+ +1$; and maximizing a penalized profile-likelihood by adding a penalty term to (3.6) with D_3.

This finding is also reflected in Table 3.9, the estimate of the coefficient of $(t-\kappa)_+$ in D_3 (e) is unusually large. This seems like a common phenomenon in spline smoothing with truncated polynomial basis functions. To circumvent this potential issue, we consider a penalized approach [Ruppert et al., 2003]. That is, when maximizing the profile likelihood function (3.6), we consider adding an additional penalty term $-\lambda|e|$, where λ is a tuning parameter. In practice, we choose λ over a grid of λ values by minimizing the rotated residual sum of squares, defined as

$$\mathrm{RSS}(\lambda)=(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{f}})^\top \boldsymbol{R}^{-1}(\widehat{\boldsymbol{\theta}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{f}}),$$

where $\boldsymbol{R}(\boldsymbol{\theta})$ is the correlation matrix. The tuning parameter λ is chosen over a grid $\lambda=(0,2,4,\ldots,30)$. In our data analysis, $\hat{\lambda}=20$, and the resulting parameter estimates are given in the last column of Table 3.9. The resulting estimate of e is much smaller, the other estimates of e are close to each

other. The estimated standard deviation at each time point are plotted in Figure 3.6. We notice that D_3 from the penalized approach has the smallest area under the curve. As a consequence, the 95% confidence interval of the temporal function $\widehat{f}(t)$ of D_3 from the penalized approach is the narrowest compared to D_1, D_2 and D_3, as presented in the last panel of Figure 3.5.

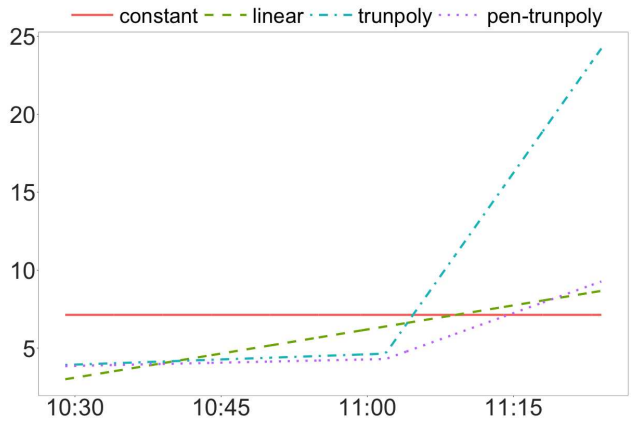

Figure 3.6: Estimated standard deviation at each time point (solid line: $D_1(\boldsymbol{s}, t)$; dashed line: $D_2(\boldsymbol{s}, t)$; dash-dotted line: $D_3(\boldsymbol{s}, t)$; dotted line: $D_3(\boldsymbol{s}, t)$ by maximizing penalized profile-likelihood function).

Finally, we consider an interpolation of the noise intensity in space and time by kriging based on D_3 with penalty. Figure 3.7 presents a dynamic evolution of the noise intensity maps over time and suggests a possible noise source in the upper-left corner with high noise intensity. There is also a sharp decrease of the noise intensity at 11:10:00 am when the engine was turned off even though all the sensors remain active, as well as a horizontal separation around $y = 30$ before 11:10:00 am, reflecting the wall that separates the facility [Ludwig et al., 2017].

3.7 Technical Details

3.7.1 Notation and Assumptions

We use $\boldsymbol{\beta}_0$ to denote the vector of true regression coefficients and $\boldsymbol{\theta}_0$ to denote the vector of true covariance parameters. We denote the log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (3.1), when $f(t)$ is known, as

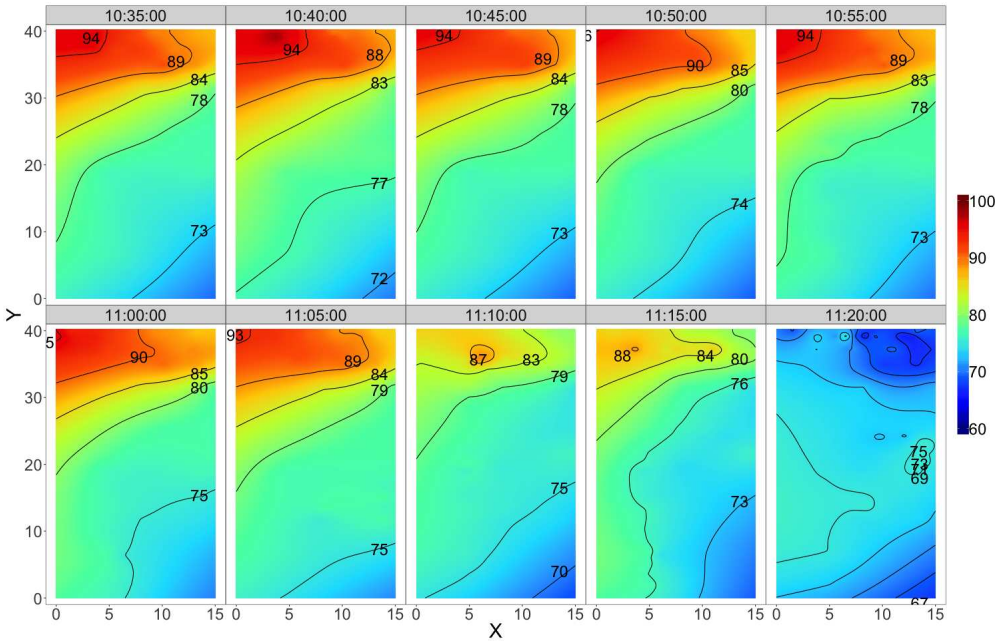

Figure 3.7: Spatio-temporal interpolation of noise intensity by kriging using estimated parameters from the last column of Table 3.9 (i.e., by maximizing penalized profile-likelihood function with truncated polynomial D_3) at 5-minute interval between 10:35:00 to 11:20:00.

$$\ell_0(\boldsymbol{\beta},\boldsymbol{\theta}) = -(N_n/2)\log(2\pi) - (1/2)\log\{\det\boldsymbol{\Gamma}(\boldsymbol{\theta})\} - (1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})^\top\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f}).$$

Let $\ell_0'(\boldsymbol{\beta}) = \partial\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\boldsymbol{\beta}$ and $\ell_0'(\boldsymbol{\theta}) = \partial\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ denote the first-order partial derivatives of $\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. For ease of notation, we suppress $\boldsymbol{\theta}$ in matrices relying on $\boldsymbol{\theta}$. For example, we write $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta})$. Then, we have $\ell_0'(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})$ and the kth element of $\ell_0'(\boldsymbol{\theta})$ is $-(1/2)\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) - (1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})^\top\boldsymbol{\Gamma}^k(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})$, 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 $k = 1,\ldots,q$.

Further, denote the second-order partial derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as $\ell_0''(\boldsymbol{\beta},\boldsymbol{\beta}) = \partial^2\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\boldsymbol{\beta}^2$, $\ell_0''(\boldsymbol{\theta},\boldsymbol{\theta}) = \partial^2\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$ and $\ell_0''(\boldsymbol{\beta},\boldsymbol{\theta}) = \partial^2\ell_0(\boldsymbol{\beta},\boldsymbol{\theta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}$. Let $\boldsymbol{\mathcal{J}}_n(\boldsymbol{\beta}) = E\{-\ell_0''(\boldsymbol{\beta},\boldsymbol{\beta})\}$ and $\boldsymbol{\mathcal{J}}_n(\boldsymbol{\theta}) = E\{-\ell_0''(\boldsymbol{\theta},\boldsymbol{\theta})\}$ denote the information matrices of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. In particular, $\ell_0''(\boldsymbol{\beta},\boldsymbol{\beta}) = -\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}$, the kth column of $\ell_0''(\boldsymbol{\beta},\boldsymbol{\theta})$ is $\boldsymbol{X}^\top\boldsymbol{\Gamma}^k(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})$, and the (k,k')th entry of $\ell_0''(\boldsymbol{\theta},\boldsymbol{\theta})$ is $-(1/2)\Big\{\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{kk'}+\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}_{k'})+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{f})^\top\boldsymbol{\Gamma}^{kk'}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-$

$\boldsymbol{f})\Big\}$, 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 $k, k' = 1, \ldots, q$.

It can be shown that $E\{\ell_0''(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \mathbf{0}$, so the information matrix of $\boldsymbol{\eta}$ is

$$\boldsymbol{\mathcal{I}}_n(\boldsymbol{\eta}) = \mathrm{diag}\{\boldsymbol{\mathcal{I}}_n(\boldsymbol{\beta}), \boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta})\},$$

where

$$\boldsymbol{\mathcal{I}}_n(\boldsymbol{\beta}) = E\{-\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta})\} = \boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \tag{3.13}$$

and the (k, k')th entry of

$$\boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta}) = E\{-\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta})\} \tag{3.14}$$

is $t_{kk'}/2$ with $t_{kk'} = \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'}) = \mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{k'})$.

For a matrix $\boldsymbol{A} = [a_{ii'}]_{i,i'=1}^{N_n}$, we let $\mu_i(\boldsymbol{A})$ denote its ith largest eigenvalue, let $\|\boldsymbol{A}\|_2 = \mu_1(\boldsymbol{A})$ denote its spectral norm, let $\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^{N_n}\sum_{i'=1}^{N_n} a_{ii'}^2\right)^{1/2}$ denote its Frobenius norm, let $\|\boldsymbol{A}\|_{\max} = \max_{i,i'}|a_{ii'}|$ denote its max norm, and let $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le N_n}\sum_{i'=1}^{N_n}|a_{ii'}|$ denote the maximum absolute column sum of the matrix.

Finally, let $\xrightarrow{P}$ denote convergence in probability and $\xrightarrow{D}$ denote convergence in distribution, as $n \to \infty$.

The theoretical properties of the methods developed in Section 3.2 are established under the following additional regularity conditions.

(C.1) The sampling sites $\{\boldsymbol{s}_i\}$ is a sequence of fixed design points on $\mathcal{R}$. In addition, there exists a continuous density function $q_{\boldsymbol{s}}$ defined on $\mathcal{R}$ such that $N_n^{-1}\sum_{i=1}^{N_n} I(\boldsymbol{s}_i \in A) \to \int_A q_{\boldsymbol{s}}(\boldsymbol{s})d\boldsymbol{s}$ uniformly for any measurable set $A \subset \mathcal{R}$.

(C.2) There exists a nondecreasing function $Q(t)$ with $Q(0) = 0$ and $Q(1) = 1$ such that (i) $\sup_{t\in[0,1]}|Q_{N_n}(t) - Q(t)| = \mathcal{O}(\zeta_n)$, where $Q_{N_n}(t) = N_n^{-1}\sum_{i=1}^{N_n} I(t_i \le t)$; (ii) its first-

order derivative function $q(t)$ is bounded away from zero and infinity and has continuous second partial derivatives.

(C.3) For $j = 1, \ldots, p$, there exists a function $g_j(\cdot)$ on $\mathcal{T}$ with a bounded second derivative satisfying

$$x_j(\boldsymbol{s}_i, t_i) = g_j(t_i) + \phi_{ij}, \quad \text{for } i = 1, \ldots, N_n,$$

where $\{\phi_{ij}\}$ is a sequence of real numbers such that

$$\lim_{n\to\infty} N_n^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Phi} = \boldsymbol{\Pi},$$

where $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{iN_n})^\top$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)$, and $\boldsymbol{\Pi}$ is a positive definite matrix. In addition, for $j = 1, \ldots, p$,

$$\limsup_{n\to\infty} (1/a_n) \max_{1\le k\le N_n} \left| \sum_{m=1}^{k} \phi_{i_m j} \right| < \infty$$

for all permutations $(i_1, \ldots, i_{N_n})$ of $(1, \ldots, N_n)$, where $a_n = N_n^{1/2} \log N_n$.

(C.4) The temporal function $f(t)$ is twice differentiable with a bounded second-order derivative on $\mathcal{T}$.

(C.5) The kernel $K(\cdot)$ is a symmetric, nonnegative, and bounded function with a compact support in $\mathbb{R}$ and with a bounded first-order derivative.

(C.6) The bandwidth h satisfies $h \to 0$, $N_n h^4 \to \infty$, $N_n h^8 \to 0$ and $\zeta_n h^{-1} \to 0$ as $n \to \infty$.

(C.7) For $k = 1, \ldots, q$, $\|\boldsymbol{\Gamma}_k\|_F^{-2} \le D_k N_n^{-1/2-\iota}$ for some $\iota > 0$ and $D_k > 0$.

(C.8) It holds that $\|\boldsymbol{\Gamma}^{-1}\|_2 < C^* < \infty$ for some constant C^*.

(C.9) $\lim_{n\to\infty} N_n^{-1} \boldsymbol{\mathcal{J}}_n(\boldsymbol{\theta}) \longrightarrow \boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta})$, where $\boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta})$ is non-singular.

(C.10) Given $t \in (0, 1)$, there exists a 2×2 matrix $\boldsymbol{\Delta}_t$, such that $(N_n^{-1} h) \boldsymbol{k}_t^\top \boldsymbol{\Gamma} \boldsymbol{k}_t \longrightarrow q(t)^2 \boldsymbol{\Delta}_t$, where $\boldsymbol{k}_t = \{K_h(t_i - t)\{(t_i - t)/h\}^{j-1}\}_{i,j=1}^{N_n,2}$ is an $N_n \times 2$ matrix.

Remarks: (C.1)–(C.2) are conditions on the spatio-temporal sampling design where observations are irregularly spaced and timed. (C.3) is a mild assumption about the relationship between the fixed design points and $\{t_i\}$ in the partially linear model, which is a generalization of Assumption 2.2 (i) in Gao and Liang [1997]. (C.4)–(C.6) are common assumptions in kernel smoothing. (C.4) ensures the smoothness of the temporal function [Liang and Li, 2009, Vogt and Linton, 2014]. (C.5) is a standard assumption for kernel functions and can be relaxed further such that $K(t)$ satisfies a Lipschitz condition $|K(t) - K(t')| \leq c\,|t - t'|$ for any $t, t' \in \mathbb{R}$ and some $c > 0$. In addition, (C.6) is a condition for the rate of bandwidth with respect to N_n and ζ_n. (C.7) assures that the first-order partial derivatives of the covariance matrix have a higher order than root-N_n. (C.8) imposes a lower bound on the smallest eigenvalue of the covariance matrix. (C.9) guarantees that the growth of information. Finally, (C.10) constrains the covariance function.

In the following proofs, we suppress n in ${}^n t_{kk'}$, ${}^n a_{kk'}$, ${}^n\boldsymbol{\Gamma}$, ${}^n\boldsymbol{\Gamma}_k$, ${}^n\boldsymbol{\Gamma}_{kk'}$, $\boldsymbol{I}_n$, $\boldsymbol{A}_n$, ${}^n\widehat{\boldsymbol{\eta}}$, ${}^n\widehat{\boldsymbol{\beta}}$ and ${}^n\widehat{\boldsymbol{\theta}}$ for ease of notation.

A Remark on Assumption (C.3)

In this section, we will show that if $\|{}^n\boldsymbol{\Gamma}^{-1}\|_\infty = \mathcal{O}(1)$, we have $\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \succeq \boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, where $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. To see this, we write

$$\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} = \boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{G} + \boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{G} + \boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$$

Since $\|{}^n\boldsymbol{\Gamma}^{-1}\|_\infty = \mathcal{O}(1)$, $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}$ is uniformly bounded elementwise. Together with (C.3), we have $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} = \mathcal{O}(N_n^{1/2}\log N_n)$.

Recall that $\lim_{n\to\infty} N_n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} = \boldsymbol{\Pi}$. Thus, $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$ is dominated by $\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, and we have

$$\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \succeq \boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi},$$

in which the equality holds if $g(\cdot) = 0$.

This result indicates that the asymptotic variances of $\widehat{\boldsymbol{\beta}}$ in the partially linear model are greater than those in the simple linear regression model.

In addition, we consider the following regularity conditions for a locally stationary process.

(C.11) Define $g(\boldsymbol{s},t) = g(\mathbf{0},0,\boldsymbol{s},t)$. Assume $g(\boldsymbol{s},t)$ satisfies $|g(\boldsymbol{s},t) - g(\boldsymbol{s}',t')| \leq C_1\|\boldsymbol{s} - \boldsymbol{s}'\| + C_2|t-t'|$ for all $(\boldsymbol{s},t),(\boldsymbol{s}',t') \in \mathcal{R}\times\mathcal{T}$, where C_1, C_2 are positive constants.

(C.12) There exist two positive nonincreasing functions γ_0 and γ_1 such that $|\gamma_n((\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n, t+u_2/B_n))| \leq \gamma_0(\|\boldsymbol{u}_1\|)\gamma_1(|u_2|)$ for all n and $\|\boldsymbol{u}_1\|, |u_2| \in [0,\infty)$ such that $(\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n, t+u_2/B_n) \in \mathcal{R}\times\mathcal{T}$. In addition, $\int_0^\infty u^{d-1}\gamma_0(u)du < \infty$ and $\int_0^\infty \gamma_1(u)du < \infty$.

(C.13) The covariance function $\gamma_n(\cdot,\cdot;\boldsymbol{\theta})$ is bounded and is twice continuously differentiable with respect to $\boldsymbol{\theta}$ in an open set.

(C.14) There exist two positive nonincreasing functions γ_2 and γ_3 such that $\max\{|\gamma_{n,k}((\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n, t+u_2/B_n))|, |\gamma_{n,kk'}((\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n, t+u_2/B_n))|\} \leq \gamma_2(\|\boldsymbol{u}_1\|)\gamma_3(|u_2|)$ for all n and $\|\boldsymbol{u}_1\|, |u_2| \in [0,\infty)$ with $(\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n, t+u_2/B_n) \in \mathcal{R}\times\mathcal{T}$ and $1 \leq k,k' \leq q$. Further, $\int_0^\infty u^{d-1}\gamma_2(u)du < \infty$ and $\int_0^\infty \gamma_3(u)du < \infty$.

3.7.2 Lemmas

In the following Lemmas 1–6, we generalize some classical results for random sampling designs [Fan and Huang, 2005] to fixed sampling design, which will be used in the proofs of Theorems 6–8.

Lemma 1. *Under Assumptions (C.2), (C.5) and (C.6), for $k \geq 0$, we have*

$$\sup_{t\in[0,1]} |v_{k,t} - N_n\mu_{k,t}q(t)| = \mathcal{O}(N_nh + N_n\zeta_nh^{-1}),$$

where $v_{k,t} = h^{-k}\sum_{i=1}^{N_n}(t_i-t)^kK_h(t_i-t)$, $K_h(t) = (1/h)K\left(t/h\right)$ and

$$\mu_{k,t} = \begin{cases} \int_{-t/h}^{\infty} x^k K(x)dx, & \text{if } t < Mh, \\ \int_{-\infty}^{\infty} x^k K(x)dx := \mu_k, & \text{if } Mh \le t \le 1 - Mh, \\ \int_{-\infty}^{(1-t)/h} x^k K(x)dx, & \text{if } t > 1 - Mh, \end{cases}$$

where $[-M, M]$ *is the compact support of* $K(\cdot)$.

Proof. For any $t \in [0, 1]$, we have

$$\begin{aligned} |v_{k,t} - N_n \mu_{k,t} q(t)| &\le \left| N_n h^{-k} \int_0^1 (z-t)^k K_h(z-t)\, d(Q_{N_n} - Q)(z) \right| \\ &\quad + \left| N_n h^{-k} \int_0^1 (z-t)^k K_h(z-t)\, dQ(z) - N_n \mu_{k,t} q(t) \right| \\ &\equiv (I_{1,1}) + (I_{1,2}). \end{aligned}$$

For $(I_{1,1})$,

$$\begin{aligned} & \left| N_n h^{-k} \int_0^1 (z-t)^k K_h(z-t)\, d(Q_{N_n} - Q)(z) \right| \\ = \; & N_n h^{-k} \Bigg| (z-t)^k K_h(z-t)\, (Q_{N_n} - Q)(z)|_0^1 \\ & - \int_0^1 (Q_{N_n} - Q)(z) \left[(z-t)^k K_h(z-t) \right]' dz \Bigg| \\ = \; & N_n h^{-2} \left| \int_0^1 (Q_{N_n} - Q)(z) k \left(\frac{z-t}{h} \right)^{k-1} K\left(\frac{z-t}{h} \right) dz \right. \\ & \left. + \int_0^1 (Q_{N_n} - Q)(z) \left(\frac{z-t}{h} \right)^{k} K'\left(\frac{z-t}{h} \right) dz \right| \\ \le \; & N_n \sup_{z \in [0,1]} |(Q_{N_n} - Q)(z)| \times \\ & \left(\int_{-M}^{M} \left| k u^{k-1} h^{-1} K(u) \right| du + \int_{-M}^{M} \left| u^k h^{-1} K'(u) \right| du \right) \\ = \; & \mathcal{O}(N_n \zeta_n h^{-1}). \end{aligned}$$

The second equality uses the fact that $(Q_{N_n} - Q)(1) = (Q_{N_n} - Q)(0) = 0$.

For $(I_{1,2})$,

$$\sup_{t \in [0,1]} \left| N_n h^{-k} \int_0^1 (z-t)^k K_h(z-t)\, dQ(z) - N_n q(t) \mu_{k,t} \right|$$

$$
\begin{aligned}
&= \sup_{t\in[0,1]} \left| N_n \int_{-\frac{t}{h}}^{\frac{1-t}{h}} u^k K(u) \left(q(t) + q'(t)uh + \frac{q''(\tilde{t})u^2h^2}{2} \right) du - N_n q(t) \mu_{k,t} \right| \\
&\leq N_n h \left\{ \sup_{t\in[0,1]} |q'(t)| \int |u^{k+1}K(u)| du + (h/2) \sup_{t\in[0,1]} |q''(t)| \int |u^{k+2}K(u)| du \right\} \\
&= \mathcal{O}(N_n h),
\end{aligned}
$$

where $\tilde{t} \in [t, t+uh]$. Thus, we have (1).

□

Lemma 2. *Under Assumptions (C.2) and (C.4)–(C.6), we have*

$$
\sup_{t\in[0,1]} |\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t)| = \mathcal{O}(h^2).
$$

Proof. First, straightforward calculation yields $\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t = \begin{pmatrix} v_{0,t} & v_{1,t} \\ v_{1,t} & v_{2,t} \end{pmatrix}$. By Lemma 1, uniformly on $[0,1]$, we have $v_{0,t} = N_n q(t)\mu_{0,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$, $v_{1,t} = N_n q(t)\mu_{1,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$ and $v_{2,t} = N_n q(t)\mu_{2,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$. In addition, notice that

$$
(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} = \left(\frac{v_{2,t}}{v_{0,t}v_{2,t} - v_{1,t}^2}, \frac{-v_{1,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} \right).
$$

Thus, we have

$$
\frac{v_{2,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} = N_n^{-1}(q(t))^{-1} \frac{\mu_{2,t}}{\mu_{0,t}\mu_{2,t} - \mu_{1,t}^2} + \mathcal{O}(N_n^{-1}h + N_n^{-1}\zeta_n h^{-1}),
$$

$$
\frac{-v_{1,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} = N_n^{-1}(q(t))^{-1} \frac{\mu_{1,t}}{\mu_{0,t}\mu_{2,t} - \mu_{1,t}^2} + \mathcal{O}(N_n^{-1}h + N_n^{-1}\zeta_n h^{-1})
$$

uniformly on $[0,1]$.

Recall that $\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t) = (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{f} - f(t)$. A Taylor's expansion yields $f(t_i) - f(t) = f'(t)(t_i - t) + 1/2 f''(\xi_i)(t_i - t)^2$, where ξ_i is between t and t_i. Thus, we have

$$
\begin{aligned}
\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t) &= (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{f} - f(t) \\
&= (1/2)(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_\xi,
\end{aligned}
$$

where $\boldsymbol{d}_\xi = (f''(\xi_1)(t_1-t)^2, \ldots, f''(\xi_{N_n})(t_{N_n}-t)^2)^\top$.

In addition, we have

$$\begin{aligned} & \sup_{t\in[0,1]} \left|(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_\xi\right| \\ \leq & \max_{x\in[0,1]} |f''(x)| \sup_{t\in[0,1]} \left(\left| \frac{v_{2,t}^2}{v_{0,t}v_{2,t}-v_{1,t}^2} \right| + \left| \frac{v_{1,t}v_{3,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \right| \right) h^2 = \mathcal{O}(h^2). \end{aligned}$$

Thus, we have the desired result.

□

Lemma 3. *Suppose that Assumptions (C.2) and (C.4)–(C.6) hold. For any random vector $\boldsymbol{\varepsilon}$ of zero mean, we have*

$$\sup_{t\in[0,1]} |\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon}| = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}.$$

Proof. For a random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{N_n})^\top$, we have

$$\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon} = (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{\varepsilon} = (I_{3,1}) - (I_{3,2}),$$

where $(I_{3,1}) = \frac{v_{2,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \sum_{i=1}^{N_n} K_h(t_i-t)\varepsilon_i$ and $(I_{3,2}) = \frac{v_{1,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \sum_{i=1}^{N_n} K_h(t_i-t)(t_i-t)h^{-1}\varepsilon_i$.

Note that $(I_{3,1}) = \frac{v_{0,t}v_{2,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \frac{\sum_{i=1}^{N_n} K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i-t)}$, by Lemma 5, we have

$$\begin{aligned} \sup_{t\in[0,1]} \left| \frac{\sum_{i=1}^{N_n} K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i-t)} \right| &= \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}, \\ \sup_{t\in[0,1]} \left| \frac{\sum_{i=1}^{N_n} (t_i-t)h^{-1} K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i-t)} \right| &= \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}. \end{aligned}$$

Therefore, using similar arguments in Lemma 2, we have $\sup_{t\in[0,1]} |(I_{3,1})| = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$ and $\sup_{t\in[0,1]} |(I_{3,2})| = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$.

□

Lemma 4. *Suppose that Assumptions (C.2) and (C.4)–(C.6) hold, we have*

$$\sup_{t\in[0,1]} \left| \widetilde{f}(t) - f(t) \right| = \mathcal{O}_p\left\{ h^2 + \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\},$$

where $\widetilde{f}(t) = \boldsymbol{\omega}_1(t)\boldsymbol{y}^* = \boldsymbol{\omega}_1(t)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

Proof. First, note that $\boldsymbol{\omega}_1(t)\mathbf{1}_{N_n} - 1 = 0$. Thus, we have $\widetilde{f}(t) - f(t) = \boldsymbol{\omega}_1(t)\left\{\boldsymbol{f} + \boldsymbol{\varepsilon} - f(t)\mathbf{1}_{N_n}\right\}$. Next, we have $\sup_{t\in[0,1]}\left|\widetilde{f}(t) - f(t)\right| \leq \sup_{t\in[0,1]}\left|\boldsymbol{\omega}_1(t)(\boldsymbol{f} - f(t)\mathbf{1}_{N_n})\right| + \sup_{t\in[0,1]}\left|\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon}\right|$. The desired result follows from Lemmas 2 and 3. □

Lemma 5. *Suppose Assumptions (C.2), (C.5) and (C.6) hold. For any random vector* $\boldsymbol{\varepsilon}$ *of zero mean, we have*

$$\sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}(t_i - t)^j h^{-j} K_h(t_i - t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i - t)}\right| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_n h}\right)^{1/2}\right\}$$

for $j = 0, 1$.

Proof. Let I_k be the interval centered at c_k with the length $l_{N_n} = \left\{\log N_n/(N_n h)\right\}^{1/2} h^{3+j}$. There exist $r_{N_n} = \lfloor l_{N_n}^{-1}\rfloor + 1$ intervals satisfying $[0,1] \subset \bigcup_{k=1}^{r_{N_n}} I_k$.

First, we have

$$\sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}(t_i - t)^j h^{-j} K_h(t_i - t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i - t)}\right| \leq (I_{6,1}) + (I_{6,2}),$$

where

$$(I_{6,1}) = \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{\sum_{i=1}^{N_n}(t_i - t)^j K_h(t_i - t)\varepsilon_i}{h^j\sum_{i=1}^{N_n} K_h(t_i - t)} - \frac{\sum_{i=1}^{N_n}(t_i - c_k)^j K_h(t_i - c_k)\varepsilon_i}{h^j\sum_{i=1}^{N_n} K_h(t_i - c_k)}\right|,$$

$$(I_{6,2}) = \max_{1\leq k\leq r_{N_n}}\left|\frac{\sum_{i=1}^{N_n}(t_i - c_k)^j h^{-j} K_h(t_i - c_k)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i - c_k)}\right|.$$

Note that

$$\begin{aligned}
&\frac{\sum_{i=1}^{N_n}(t_i - t)^j h^{-j} K_h(t_i - t)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i - t)} - \frac{\sum_{i=1}^{N_n}(t_i - c_k)^j h^{-j} K_h(t_i - c_k)\varepsilon_i}{\sum_{i=1}^{N_n} K_h(t_i - c_k)} \\
= \;& (v_{0,t})^{-1}\left[\sum_{i=1}^{N_n}\bar{K}(t, t_i, c_k)(t_i - t)^j h^{-j}\varepsilon_i\right] \\
&+(v_{0,t})^{-1}\left[\sum_{i=1}^{N_n}\{(t_i - t)^j - (t_i - c_k)^j\}h^{-j} K_h(t_i - c_k)\varepsilon_i\right]
\end{aligned}$$

$$-(v_{0,t}v_{0,c_k})^{-1}\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)\sum_{i=1}^{N_n}(t_i-c_k)^j h^{-j}K_h(t_i-c_k)\varepsilon_i,$$

where $\bar{K}(t,t_i,c_k)=K_h(t_i-t)-K_h(t_i-c_k)$. Therefore, we have

$$\begin{aligned}(I_{6,1}) \leq & \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)(t_i-t)^j h^{-j}\varepsilon_i\right]\right| \\ & + \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\{(t_i-t)^j-(t_i-c_k)^j\}h^{-j}K_h(t_i-c_k)\varepsilon_i\right]\right| \\ & + \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}v_{0,c_k}}\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)\sum_{i=1}^{N_n}(t_i-c_k)^j h^{-j}K_h(t_i-c_k)\varepsilon_i\right|.\end{aligned}$$

By Lemma 1, it can be shown that $\sup_{t\in[0,1]}\left|v_{0,t}^{-1}\right|=\mathcal{O}(N_n^{-1})$. In addition, by (C.5), for any $t\in I_k$, $|\bar{K}(t,t_i,c_k)|\leq h^{-1}\max_{x\in\mathbb{R}}|K'(x)|\left|\frac{t_i-t}{h}-\frac{t_i-c_k}{h}\right|=\mathcal{O}(h^{-2}l_{N_n})$. Therefore, we have

$$\begin{aligned}& \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)(t_i-t)^j h^{-j}\varepsilon_i\right]\right| \\ \leq & \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}|\bar{K}(t,t_i,c_k)|\max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left\{\frac{1}{v_{0,t}}\sum_{i=1}^{N_n}\left|(t_i-t)^j h^{-j}\varepsilon_i\right|\right\} \\ = & \mathcal{O}(N_n^{-1}l_{N_n}h^{-2-j})\sum_{i=1}^{N_n}|\varepsilon_i|.\end{aligned}$$

We further note that

$$\begin{aligned}& \max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\{(t_i-t)^j-(t_i-c_k)^j\}h^{-j}K_h(t_i-c_k)\varepsilon_i\right]\right| \\ = & \begin{cases}0, & \text{if } j=0, \\ \mathcal{O}(N_n^{-1}l_{N_n}h^{-1-j})\sum_{i=1}^{N_n}|\varepsilon_i|, & \text{if } j=1.\end{cases}\end{aligned}$$

Moreover, we have

$$\max_{1\leq k\leq r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}v_{0,c_k}}\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)\sum_{i=1}^{N_n}(t_i-c_k)^j h^{-j}K_h(t_i-c_k)\varepsilon_i\right|$$

$$
\begin{aligned}
&\leq \max_{1\leq k\leq r_{N_n}} \sup_{t\in I_k} \left\{ \frac{1}{v_{0,t}v_{0,c_k}} \sum_{i=1}^{N_n} |\bar{K}(t,t_i,c_k)| \right\} \times \\
&\quad \max_{1\leq k\leq r_{N_n}} \left\{ \sum_{i=1}^{N_n} |(t_i-c_k)^j h^{-j}\varepsilon_i| K_h(t_i-c_k) \right\} \\
&= \mathcal{O}(N_n^{-2})\mathcal{O}(N_n l_{N_n} h^{-2})\mathcal{O}(h^{-1-j}) \sum_{i=1}^{N_n} |\varepsilon_i| = \mathcal{O}(N_n^{-1} l_{N_n} h^{-3-j}) \sum_{i=1}^{N_n} |\varepsilon_i| .
\end{aligned}
$$

Since $\sum_{i=1}^{N_n} |\varepsilon_i| = \mathcal{O}_p(N_n)$, we have $(I_{6,1}) = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$.

For $(I_{6,2})$, let $\boldsymbol{e} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\varepsilon}$ be a sequence of *iid* $N(0,1)$, and we have $\sum_{i=1}^{N_n}(t_i-t)^j K_h(t_i-t)\varepsilon_i = h^j \boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma}^{1/2}\boldsymbol{e}$, where $\boldsymbol{k}_{t,j+1}$ is the $(j+1)$th column of $\boldsymbol{k}_t$, $j = 0, 1$. By Bernstein inequality, for any $\lambda > 0$ and $t \in [0,1]$, we have

$$
P\left(\left| \sum_{i=1}^{N_n} (t_i-t)^j K_h(t_i-t)\varepsilon_i \right| > 2\lambda v_{0,t} \left(\frac{\log N_n}{N_n h} \right)^{\frac{1}{2}} \right) < \exp\left\{ \frac{-\lambda^2 v_{0,t}^2 \frac{\log N_n}{N_n h}}{h^{2j}\boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma} \boldsymbol{k}_{t,j+1}} \right\}.
$$

In addition, we have $h^{2j}\boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma} \boldsymbol{k}_{t,j+1} \leq \|\boldsymbol{\Gamma}\|_2 \sum_{i=1}^{N_n}(t_i-t)^{2j} K_h(t_i-t)^2 \leq \|\boldsymbol{\Gamma}\|_2 \sum_{i=1}^{N_n} K_h(t_i-t)^2$. By similar arguments as in Lemma 1, we can show that, $\sup_{t\in[0,1]} \left| \sum_{i=1}^{N_n} K_h(t_i-t)^2 \right| = \mathcal{O}(N_n h^{-1})$. From Lemma 1, we also have $\inf_{t\in[0,1]} v_{0,t}^2 = \mathcal{O}(N_n^2)$, and therefore, by choosing a large enough λ, we have

$$
\sup_{t\in[0,1]} P\left(\left| \sum_{i=1}^{N_n} (t_i-t)^j K_h(t_i-t)\varepsilon_i \right| > \lambda v_{0,t} \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right) = \mathcal{O}(N_n^{-2}).
$$

Since

$$
\begin{aligned}
&P\left((I_{6,2}) > \lambda \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right) \\
&\leq \sum_{k=1}^{r_{N_n}} P\left(\left| \frac{\sum_{i=1}^{N_n}(t_i-c_k)^j K_h(t_i-c_k)\varepsilon_i}{v_{0,c_k}} \right| > \lambda \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right) \\
&= \mathcal{O}(r_{N_n} N_n^{-2}) = o(1),
\end{aligned}
$$

we have $(I_{6,2}) = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$. Thus, we have the result.

□

Lemma 6. *Suppose that Assumptions (C.2)–(C.6) hold, we have*

$$\sup_{t\in[0,1]} |\boldsymbol{\omega}(t)\boldsymbol{X}| = \mathcal{O}(1).$$

Proof. Using similar arguments as in Lemma 2, we have $v_{k,t}/(v_{0,t}v_{2,t} - v_{1,t}^2) = \mathcal{O}(N_n^{-1})$, for $k = 0, 1, 2$. The ith element of the first row of $\boldsymbol{\omega}(t)$ is $v_{2,t}/(v_{0,t}v_{2,t}-v_{1,t}^2)K_h(t_i-t)-v_{1,t}/(v_{0,t}v_{2,t}-v_{1,t}^2)K_h(t_i-t)(t_i-t)/h = \mathcal{O}(N_n^{-1}h^{-1})$. Similarly, the ith element of the second row of $\boldsymbol{\omega}(t)$ is $\mathcal{O}(N_n^{-1}h^{-1})$. Thus, $\boldsymbol{\omega}(t)\boldsymbol{\phi}_j = \mathcal{O}(N_n^{-1/2}h^{-1}\log N_n)$. Using similar arguments in Lemma 2, we obtain $(0,1)\boldsymbol{\omega}(t)\boldsymbol{g}_j - hg_j'(t) = (0,1/2)(\boldsymbol{D}_t^\top \boldsymbol{K}_t\boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top \boldsymbol{K}_t\boldsymbol{d}_{\xi,j} = \mathcal{O}(h^2)$, where ξ_i is between t and t_i and $\boldsymbol{d}_{\xi,j} = (g''(\xi_1)(t_1-t)^2, \ldots, g''(\xi_{N_n})(t_{N_n}-t)^2)^\top$. Thus, $\boldsymbol{\omega}(t)\boldsymbol{X}_j = \boldsymbol{\omega}(t)(\boldsymbol{\phi}_j + \boldsymbol{g}_j) = \mathcal{O}(1)$.

□

3.7.3 Proof of Theorem 6

Proof of Theorem 6. By Mardia and Marshall [1984], the convergence property of $\ell_0'(\boldsymbol{\beta})$, $\ell_0'(\boldsymbol{\theta})$, $\ell_0''(\boldsymbol{\beta},\boldsymbol{\beta})$, $\ell_0''(\boldsymbol{\beta},\boldsymbol{\theta})$ and $\ell_0''(\boldsymbol{\theta},\boldsymbol{\theta})$ can be established. By (C.7)–(C.8), together with proof of Theorem 1 in Chu et al. [2019], we have

$$N_n^{-1/2}\ell_0'(\boldsymbol{\theta}) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta})\right), \tag{3.15a}$$

$$N_n^{-1}\ell_0''(\boldsymbol{\theta},\boldsymbol{\theta}) \xrightarrow{p} -\boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta}). \tag{3.15b}$$

Under (C.1)–(C.10), we first show the following results

$$N_n^{-1/2}\ell'(\boldsymbol{\beta}) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Pi}\right), \tag{3.16a}$$

$$N_n^{-1/2}\ell'(\boldsymbol{\theta}) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta})\right), \tag{3.16b}$$

$$N_n^{-1}\ell''(\boldsymbol{\beta},\boldsymbol{\beta}) \xrightarrow{p} -\boldsymbol{\Pi}, \tag{3.16c}$$

$$N_n^{-1}\ell''(\boldsymbol{\theta},\boldsymbol{\theta}) \xrightarrow{p} -\boldsymbol{\mathcal{I}}_0(\boldsymbol{\theta}), \tag{3.16d}$$

$$N_n^{-1}\ell''(\boldsymbol{\beta},\boldsymbol{\theta}) \xrightarrow{p} \mathbf{0}. \tag{3.16e}$$

By Lemmas 2–3, we have

$$\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\|^2 = N_n\mathcal{O}(h^4) = \mathcal{O}(N_n h^4), \tag{3.17a}$$

$$\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2 = N_n\mathcal{O}_p\left(\frac{\log N_n}{N_n h}\right) = \mathcal{O}_p\left(\frac{\log N_n}{h}\right). \tag{3.17b}$$

Proof of (3.16a). Straightforward calculation yields

$$\begin{aligned}\ell'(\boldsymbol{\beta}) =& \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\\ =& \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon} + \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f} \equiv (I_1)+(I_2).\end{aligned}$$

For (I_1), by Assumption (C.3), we have

$$\begin{aligned}&N_n^{-1/2}\{\boldsymbol{\phi}_j+\boldsymbol{g}_j\}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon}\\ =&N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon} + N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon} + N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon}+\\ &N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon} + N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon} + N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon}\\ \equiv&(I_{11})+(I_{12})+(I_{13})+(I_{14})+(I_{15})+(I_{16}).\end{aligned}$$

For (I_{11}), it can be shown that $N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon} \xrightarrow{D} N(\boldsymbol{0}, N_n^{-1}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\phi}_j)$. In addition, we have

$$\begin{aligned}&N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2 \le N_n^{-1}\|\boldsymbol{\Gamma}^{-1}\|_2^2\,\mathrm{E}\,|\boldsymbol{\phi}_j^\top\boldsymbol{S}\boldsymbol{\varepsilon}|^2 = \mathcal{O}((\log N_n)^3 N_n^{-1}h^{-1}),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2 \le N_n^{-1}\|\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\|^2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2 = \mathcal{O}(h^3\log N_n),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon})^2 \le N_n^{-1}\|\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2 = \mathcal{O}(N_n^{-1}h^{-2}(\log N_n)^2),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon})^2 \le N_n^{-1}\|\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2 = \mathcal{O}(h^4).\end{aligned}$$

By Lemma 1 and Assumption (C.3), we have

$$\|\boldsymbol{S}\boldsymbol{\phi}_j\| = \mathcal{O}(N_n^{1/2}N_n^{-1}h^{-1}N_n^{1/2}\log N_n) = \mathcal{O}(h^{-1}\log N_n).$$

Thus, for (I_{16}),

$$N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2 \le N_n^{-1}\|\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2 = \mathcal{O}(N_n^{-1}h^{-3}(\log N_n)^3).$$

Similarly, for (I_2), we have

$$N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f} = \mathcal{O}(N_n^{-1/2}N_n^{1/2}h^2N_n^{1/2}h^2) = \mathcal{O}(N_n^{1/2}h^4),$$
$$N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f} = \mathcal{O}(N_n^{-1/2}N_n^{1/2}\log N_n h^2) = \mathcal{O}(h^2\log N_n),$$
$$N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f} = \mathcal{O}(N_n^{-1/2}h^{-1}\log N_n N_n^{1/2}h^2) = \mathcal{O}(h\log N_n).$$

Proof of (3.16b) and (3.16d). Since the kth element of $-2\ell'(\boldsymbol{\theta})$ is $\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) + (\boldsymbol{f}+\boldsymbol{\varepsilon})^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{f}+\boldsymbol{\varepsilon})$, by (3.17a)–(3.17b), we have

$$N_n^{-1/2}\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f} \le N_n^{-1/2}\|\boldsymbol{\Gamma}^k\|_2\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\|^2 = \mathcal{O}(N_n^{1/2}h^4),$$
$$N_n^{-1/2}\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon} \le N_n^{-1/2}\mathcal{O}(h^2)\mathbf{1}_{N_n}^\top\boldsymbol{\varepsilon} \xrightarrow{p} 0,$$
$$N_n^{-1/2}|\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k\boldsymbol{S}\boldsymbol{\varepsilon}| \le N_n^{-1/2}\mathcal{O}_p\left(N_n^{1/2}h^3\log N_n\right),$$
$$\mathrm{E}\left|N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^k\boldsymbol{S}\boldsymbol{\varepsilon}\right| \le N_n^{-1/2}\|\boldsymbol{\Gamma}^k\|_2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2 = \mathcal{O}_p\left(\frac{\log N_n}{N_n^{1/2}h}\right).$$

By Lemma 3, we have $N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon} = N_n^{-1/2}\mathbf{1}_{N_n}^\top\boldsymbol{\varepsilon}o_p(1) \xrightarrow{p} 0$. Therefore, $N_n^{-1/2}\boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon} \xrightarrow{D} N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon}$. Thus, we have (3.16b), and similar argument applies to (3.16d).

Proof of (3.16c). Using similar argument in the proof of (3.16a), we can show

$$N_n^{-1}\ell''(\boldsymbol{\beta}) = -N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X} \xrightarrow{p} -N_n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} = -\boldsymbol{\Pi}.$$

Proof of (3.16e). The kth column of $-\ell''(\boldsymbol{\beta},\boldsymbol{\theta})$ is

$$\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{f}+\boldsymbol{\varepsilon}).$$

The same argument as in (3.16a) can be used to show $N_n^{-1}\ell''(\boldsymbol{\beta},\boldsymbol{\theta}) \xrightarrow{p} \mathbf{0}$.

Next, we show the consistency and asymptotic normality of parameter estimates. To establish $\|\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\| = \mathcal{O}_p(N_n^{-1/2})$, it suffices to show that, for a given constant $\epsilon > 0$, there is a constant C such that, for a sufficiently large n,

$$P\left\{\sup_{\|\boldsymbol{u}\|=C}\ell(\boldsymbol{\eta}_0+N_n^{-1/2}\boldsymbol{u}) < \ell(\boldsymbol{\eta}_0)\right\} \ge 1-\epsilon, \tag{3.18}$$

where $\boldsymbol{u} \in \mathbb{R}^{p+q}$. By Taylor's expansion, we obtain

$$\ell(\boldsymbol{\eta}_0 + N_n^{-1/2}\boldsymbol{u}) - \ell(\boldsymbol{\eta}_0) = N_n^{-1/2}\ell'(\boldsymbol{\eta}_0)^\top \boldsymbol{u} - (1/2)N_n^{-1}\boldsymbol{u}^\top \ell''(\boldsymbol{\eta}_0)\boldsymbol{u}\{1 + o_p(1)\}. \tag{3.19}$$

By (3.16a)–(3.16e), we have $N_n^{-1/2}\ell'(\boldsymbol{\eta}_0) = \mathcal{O}_p(1)$ and $N_n^{-1}\ell''(\boldsymbol{\eta}_0) = \mathcal{O}_p(1)$. Therefore, for a sufficiently large C, the second term dominates the first term in (3.19), and therefore, (3.18) holds.

To further establish the asymptotic normality of $\widehat{\boldsymbol{\eta}}$, it can be shown that $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ satisfies

$$\mathbf{0} = \ell'(\boldsymbol{\eta})\big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} = \ell'(\boldsymbol{\eta}_0) + \{\ell''(\boldsymbol{\eta}_0) + o_p(1)\}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0).$$

Together with (3.16a)–(3.16e), we have Theorem 6. □

3.7.4 Proof of Theorem 7

Proof. First, we have

$$\widehat{\boldsymbol{F}}(t) - \boldsymbol{F}(t) = \boldsymbol{\omega}(t)\left\{\boldsymbol{f} + \boldsymbol{\varepsilon} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} - \boldsymbol{F}(t) = (II_1) + (II_2) - (II_3),$$

where $(II_1) = \boldsymbol{\omega}(t)\boldsymbol{f} - \boldsymbol{F}(t)$, $(II_2) = \boldsymbol{\omega}(t)\boldsymbol{\varepsilon}$ and $(II_3) = \boldsymbol{\omega}(t)\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

For (II_1), a Taylor's expansion yields

$$\begin{aligned} &f(t_i) = f(t) + f'(t)(t_i - t) + 1/2f''(t)(t_i - t)^2 + 1/6f^{(3)}(\xi_i)(t_i - t)^3 \\ &= \left(1, \frac{t_i - t}{h}\right)\boldsymbol{F}(t) + 1/2f''(t)(t_i - t)^2 + 1/6f^{(3)}(\xi_i)(t_i - t)^3, \end{aligned}$$

where ξ_i is between t and t_i. Therefore, $(II_1) = (1/2)\boldsymbol{\omega}(t)\boldsymbol{d}_2 f''(t) + \frac{1}{6}\boldsymbol{\omega}(t)\boldsymbol{d}_{3,\xi}$, where $\boldsymbol{d}_{3,\xi} = (f^{(3)}(\xi_1)(t_1 - t)^3, \ldots, f^{(3)}(\xi_{N_n})(t_{N_n} - t)^3)^\top$ and $\boldsymbol{d}_2 = ((t_1 - t)^2, \ldots, (t_{N_n} - t)^2)^\top$. Given $t \in (0, 1)$, by Lemma 1, we have

$$\boldsymbol{\omega}(t)\boldsymbol{d}_2 f''(t) = \begin{pmatrix} v_{2,t}^2 - v_{1,t}v_{3,t} \\ v_{0,t}v_{3,t} - v_{1,t}v_{2,t} \end{pmatrix} \frac{h^2 f''(t)}{v_{0,t}v_{2,t} - v_{1,t}^2} = h^2 \begin{pmatrix} \mu_2 f''(t) \\ 0 \end{pmatrix} + o(h^2).$$

Moreover, we have

$$|\boldsymbol{\omega}(t)\boldsymbol{d}_{3,\xi}| \leq \max_{x\in\mathbb{R}} \frac{4|f^{(3)}(x)|h^3}{v_{0,t}v_{2,t}-v_{1,t}^2} \left\{ \left| \begin{pmatrix} v_{2,t}v_{3,t}-v_{1,t}v_{4,t} \\ -v_{1,t}v_{3,t}+v_{0,t}v_{4,t} \end{pmatrix} \right| \right\} = \mathcal{O}(h^3).$$

Therefore, $(II_1) = h^2 \begin{pmatrix} \mu_2 f''(t) \\ 0 \end{pmatrix} + o(h^2)$.

For (II_2), let $A(\boldsymbol{\varepsilon}) = \sum_{i=1}^{N_n} K_h(t_i - t)\varepsilon_i$ and $B(\boldsymbol{\varepsilon}) = \sum_{i=1}^{N_n} K_h(t_i - t)\frac{t_i - t}{h}\varepsilon_i$, we have

$$\boldsymbol{\omega}(t)\boldsymbol{\varepsilon} = \frac{1}{v_{0,t}v_{2,t}-v_{1,t}^2} \begin{pmatrix} v_{2,t}A(\boldsymbol{\varepsilon}) - v_{1,t}B(\boldsymbol{\varepsilon}) \\ -v_{1,t}A(\boldsymbol{\varepsilon}) + v_{0,t}B(\boldsymbol{\varepsilon}) \end{pmatrix}.$$

For $t \in (0,1)$, by Lemma 1, we have $\frac{N_n v_{0,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \to \mu_2^{-1}q(t)^{-1}$, $\frac{N_n v_{1,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \to 0$ and $\frac{N_n v_{2,t}}{v_{0,t}v_{2,t}-v_{1,t}^2} \to q(t)^{-1}$. Since $\boldsymbol{\varepsilon}$ is a Gaussian process, by (C.10) and Slusky's Theorem, we have

$$(N_n h)^{1/2}\boldsymbol{\omega}(t)\boldsymbol{\varepsilon} \xrightarrow{D} N\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix} \boldsymbol{\Delta}_t \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}\right).$$

For (II_3), by Lemma 6, we have $\boldsymbol{\omega}(t)\boldsymbol{X} = \mathcal{O}(1)$. By Theorem 6, we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{O}_p(N_n^{-1/2})$. Consequently, we have $(II_3) = \mathcal{O}_p(N_n^{-1/2})$. Thus, we have Theorem 7.

□

3.7.5 Proof of Theorem 8

Proof. Let $\boldsymbol{\omega}_1^{(-i)}(t) = (1,0)\left[\left\{D_t^{(-i)}\right\}^\top K_t^{(-i)} D_t^{(-i)}\right]^{-1}\left\{D_t^{(-i)}\right\}^\top K_t^{(-i)}$, where $D_t^{(-i)}$ is the matrix of D_t with ith row deleted, and $K_t^{(-i)}$ is the matrix K_t with both ith row and column deleted. In addition, we let $\boldsymbol{y}^{*(-i)}$ denote the vector of response variables with the ith entry left out. Straightforward calculation reveals

$$\begin{aligned} \widehat{f}^{(-i)}(t_i) = \boldsymbol{\omega}_1^{(-i)}(t_i)\boldsymbol{y}^{*(-i)} &= \frac{\left(v_{2,t_i}^{(-i)} \quad -v_{1,t_i}^{(-i)}\right)}{v_{2,t_i}^{(-i)}v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2}\left\{D_t^{(-i)}\right\}^\top K_{t_i}^{(-i)}\boldsymbol{y}^{*(-i)} \\ &= \sum_{j\neq i} a_j^{(-i)}(t_i)y_j^*, \end{aligned}$$

where $a_j^{(-i)}(t_i) = \frac{\begin{pmatrix} v_{2,t_i}^{(-i)} & -v_{1,t_i}^{(-i)} \end{pmatrix}}{v_{2,t_i}^{(-i)} v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2} \left\{ D_t^{(-i)} \right\}^\top K_{t_i}^{(-i)}$, $v_{0,t_i}^{(-i)} = \sum_{j \neq i} K_h(t_j - t_i) = v_{0,t_i} - K_h(0)$, $v_{1,t_i}^{(-i)} = \sum_{j \neq i} K_h(t_j - t_i) \frac{t_j - t_i}{h} = v_{1,t_i}$ and $v_{2,t_i}^{(-i)} = \sum_{j \neq i} K_h(t_j - t_i) \left(\frac{t_j - t_i}{h} \right)^2 = v_{2,t_i}$.

By Lemma 1, we have

$$\begin{aligned} a_j^{(-i)}(t_i) =& \frac{v_{2,t_i}^{(-i)} K_h(t_j - t_i)}{v_{2,t_i}^{(-i)} v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2} - \frac{v_{1,t_i}^{(-i)} K_h(t_j - t_i)(t_j - t_i)/h}{v_{2,t_i}^{(-i)} v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2} \\ =& \frac{v_{2,t_i} K_h(t_j - t_i)}{\{v_{0,t_i} - K_h(0)\} v_{2,t_i} - v_{1,t_i}^2} - \frac{v_{1,t_i} K_h(t_j - t_i)(t_j - t_i)/h}{\{v_{0,t_i} - K_h(0)\} v_{2,t_i} - v_{1,t_i}^2} \\ =& \frac{K\left(\frac{t_j - t_i}{h}\right) - \frac{\mu_{1,t_i}}{\mu_{2,t_i}} K\left(\frac{t_j - t_i}{h}\right)\left(\frac{t_j - t_i}{h}\right)}{N_n h q(t_i)(\mu_{0,t_i}\mu_{2,t_i} - \mu_{1,t_i}^2)/\mu_{2,t_i} - K(0)} + \mathcal{O}\left(N_n^{-1} + N_n^{-1}\zeta_n h^{-2}\right). \end{aligned}$$

Thus, for the leave-one-out cross-validation (CV) score function, we have

$$\begin{aligned} &\mathrm{E}\{\mathrm{CV}(h)\} = \mathrm{E}\left[\frac{1}{N_n} \sum_{i=1}^{N_n} \{f(t_i) + \varepsilon_i - \widehat{f}^{(-i)}(t_i)\}^2 \right] \\ =& \frac{1}{N_n} \sum_{i=1}^{N_n} \mathrm{E}\{f(t_i) - \widehat{f}^{(-i)}(t_i)\}^2 + \frac{1}{N_n} \sum_{i=1}^{N_n} \mathrm{Var}(Y_i) - \frac{2}{N_n} \sum_{i=1}^{N_n} \mathrm{Cov}\{\widehat{f}^{(-i)}(t_i), \varepsilon_i\}. \end{aligned}$$

Denote $A(h) = \frac{2}{N_n} \sum_{i=1}^{N_n} \mathrm{Cov}\left\{ \widehat{f}^{(-i)}(t_i), \varepsilon_i \right\}$, we have

$$A(h) = \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{j \neq i} \frac{K\left(\frac{t_j - t_i}{h}\right)\left(1 - \frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j - t_i}{h}\right)\right)}{b(t_i) - K(0)} \mathrm{Cov}(\varepsilon_i, \varepsilon_j) + \mathcal{O}\left(N_n^{-1}\right),$$

since $\|\mathbf{\Gamma}\|_\infty = \mathcal{O}(1)$ as shown in the proof of Theorem 1 in Chu et al. [2019].

Under the asymptotic framework (A.1), and since $K(\cdot)$ has a bounded first-order derivative at the origin, we obtain

$$\begin{aligned} & \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{\substack{j \neq i \\ |t_j - t_i| \leq C_n}} \frac{K\left(\frac{t_j - t_i}{h}\right)\left(1 - \frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j - t_i}{h}\right)\right)}{b(t_i) - K(0)} \mathrm{Cov}(\varepsilon_i, \varepsilon_j) \\ =& \frac{K(0)}{b(t_i) - K(0)} \left(\frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{\substack{j \neq i \\ B_n |t_j - t_i| < C_n}} \mathrm{Cov}(\varepsilon_i, \varepsilon_j) \right) + o\left(\frac{1}{N_n h}\right). \end{aligned}$$

Note that

$$\begin{aligned}\sum_{\substack{j\neq i\\ |t_j-t_i|>C_n}}\mathrm{Cov}(\varepsilon_i,\varepsilon_j) &= \sum_{m'=\lfloor\left(\frac{B_nC_n}{b}\right)\rfloor}\mathcal{O}\left(\frac{b}{B_n\zeta_n}\right)\max_{mb\leq|u_2|\leq(m+1)b}\gamma_1(|u|)\\ &\leq \mathcal{O}\left(\frac{b}{B_n\zeta_n}\right)\int_{B_nC_n}^{\infty}\gamma_1(u)du\to 0,\end{aligned}$$

so we have

$$\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{|t_j-t_i|>C_n}\frac{K\left(\frac{t_j-t_i}{h}\right)\left(1-\frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j-t_i}{h}\right)\right)}{b(t_i)-K(0)}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=o\left(\frac{1}{N_nh}\right).$$

Therefore,

$$A(h)=K(0)\left(\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{\substack{j\neq i\\ |t_j-t_i|<C_n}}\frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{b(t_i)-K(0)}\right)+o\left(\frac{1}{N_nh}\right).$$

Thus, the desired results are shown. □

Chapter 4

Krylov Subspace Methods for Large Spatial Datasets[3]

4.1 Introduction

Advances in geographic data acquisition technologies have led to increasing prevalence of very large spatial datasets, and geostatistical models have been employed for analyzing these datasets [Cressie, 1993, Stein, 1999]. For spatial models, maximum likelihood estimation is often used, since the estimates are consistent and asymptotically normal [Mardia and Marshall, 1984]. However, evaluation of the likelihood function involves the inverse and determinant of large spatial covariance matrices. These calculations usually require $O(N^3)$ flops and $O(N^2)$ memory, where N is the number of observations. As the sample size N increases, evaluation of the likelihood function becomes time-consuming, if not infeasible. This necessitates the development of new statistical methods that are adaptable to large-scale spatial data. The purpose of this chapter is to develop a computationally efficient method based on Krylov subspace.

To address the computational challenges of large spatial data, various approaches have been proposed [Sun et al., 2012, Bradley et al., 2016]. Typically, these methods achieve computational simplification either by imposing a low-rank structure or a sparse structure on the spatial covariance matrices. Low-rank models seek approximation to the Gaussian processes on a lower dimensional subspace by fixed-rank kriging [Cressie and Johannesson, 2008], predictive process models [Banerjee et al., 2008, Finley et al., 2009] or multi-resolution approximations [Katzfuss, 2017]. Most of those methods have been shown to be linear time; however, it remains unclear whether the resulting parameter estimates are consistent when the computational cost is $O(N)$.

Sparse methods proceed from enforcing sparsity either on the precision matrices or the covariance matrices. Sparsity in precision matrices is introduced by composite likelihoods [Eidsvik et al.,

[3]This chapter is based on a joint work with Dr. Tingjin Chu, Dr. Jun Zhu and Dr. Haonan Wang.

2014, Bevilacqua and Gaetan, 2015], Markov random fields [Rue and Held, 2005], or products of lower dimensional conditional distributions based only on nearest neighbors[Vecchia, 1988, Stein et al., 2004]. Further, Datta et al. [2016a] generalized the Vecchia [1988] approach and proposed a nearest neighbor Gaussian process model for improved Bayesian inference and less computing time. In contrast, covariance tapering [Chu et al., 2011, Du et al., 2009, Furrer et al., 2006, Kaufman et al., 2008] constructs sparse covariance matrices by zeroing out correlations between observations that are far apart, using compactly supported covariance functions. However, the evaluation of the exact tapered likelihood relies on sparse Cholesky factorization, which generally needs computationally cumbersome permutations of the rows and columns of the matrix.

In general, when evaluating a log-likelihood function, the computational difficulties mainly arise from frequent operations on large covariance matrices, including the *inversion* and the *log determinant*. In this chapter, we try to alleviate computational complexity by iterative methods based on Krylov subspace. Here, we propose to approximate matrix inversion through conjugate gradient method, and the log determinant through Lanczos Quadrature method [Ubaru et al., 2017]. A major advantage of our proposed method is that the approximation error can be quantified and asymptotically goes to zero, enabling the establishment of theoretical property of parameter estimates. Furthermore, the computational complexity of the proposed method is considered for both dense and sparse covariance matrices. Particularly, the relationship between the convergence rate of parameter estimates and the computational complexity of the algorithm is established for our proposed method, which provides important guidelines for choosing the tuning parameters. For sparse matrices, the computational complexity is reduced to $O(N \log N)$, while the root-N convergence of the resulting parameter estimates can still be achieved by our proposed method.

The remainder of the chapter is outlined as follows. Section 4.2 presents our methodology to approximate the log-likelihood function and gives theoretical justifications. Specifically, Section 4.2.1 details the calculation of matrix inversion through Conjugate Gradient algorithm, and Section 4.2.2 describes the log-determinant approximation via stochastic Lanczos method. We provide quantifications of the computational complexity and insights for further improvement in

the remaining subsections. In Section 4.3, several experimental studies are conducted to investigate the computational efficiency and accuracy of our proposed method. Section 4.4 applies the methodology to Light Detection and Ranging (LiDAR) data with about 5 million observations, followed by theoretical developments in Section 4.5.

4.2 Methodology

Consider a spatial process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^d\}$, where $\mathcal{D}$ is the spatial domain of interest. For ease of exposition, we assume that the mean of the process is zero since the main computational challenges arise in estimating the covariance structure. Denote the covariance between $y(\boldsymbol{s})$ and $y(\boldsymbol{s}')$, $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$, by

$$\gamma(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) = \text{Cov}\{y(\boldsymbol{s}), y(\boldsymbol{s}')\},$$

which is assmued known up to some parameter $\boldsymbol{\theta} \in \mathbb{R}^q$. The inference on $\boldsymbol{\theta}$ is based on N observations $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_N))^\top$ collected at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N \in \mathcal{D}$. The log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}(\boldsymbol{\theta})| - (1/2)\boldsymbol{y}^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\boldsymbol{y}, \tag{4.1}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = [\gamma(\boldsymbol{s}_j, \boldsymbol{s}_{j'}, \boldsymbol{\theta})]_{j,j'=1}^N$ is the covariance matrix of $\boldsymbol{y}$. In the rest of this chapter, we suppress $\boldsymbol{\theta}$ in $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ for notational convenience.

Maximizing the log-likelihood function can be challenging, particularly when sample size is large. The bottleneck is frequent evaluation of inverse and log-determinant of covariance matrix at different parameter values. Most widely-used approaches for solving linear systems involve Cholesky decomposition, which generally requires $O(N^3)$ operations and $O(N^2)$ memory. In this section, to mitigate the computational burden, we will present a more tractable approach by approximating the log-likelihood function using Krylov subspace methods. Krylov subspace methods are known for their effectiveness in two major aspects, solving large linear systems and finding a few leading eigenpairs of large matrices, especially when the linear systems or the matrices are sparse.

4.2.1 Matrix Inversion via Conjugate Gradient Method

An efficient matrix inversion by means of Cholesky decomposition can be performed in-place, with minor additional storage. However, the storage of an $N \times N$ matrix is still demanding and may exceed the memory of a single computer for sufficiently large N. Since the log-likelihood function (4.1) does not require an explicit storage of $\boldsymbol{\Gamma}^{-1}$ but rather a vector $\boldsymbol{\Gamma}^{-1}\boldsymbol{y}$, iterative techniques only based on matrix-vector multiplications can be applied [Shewchuk et al., 1994].

A class of the iterative methods relies on the Krylov subspace. Given a square matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and a nonzero vector $\boldsymbol{v} \in \mathbb{R}^N$, for $k \geq 1$, the kth Krylov subspace generated by the pair $(\boldsymbol{A}, \boldsymbol{v})$ is defined as $\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{v}) = \text{span}\{\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{v}\}$, and $\mathcal{K}_0(\boldsymbol{A}, \boldsymbol{v}) = \{\boldsymbol{0}\}$. Consider a linear system $\boldsymbol{r}_0 = \boldsymbol{A}\boldsymbol{z}$, where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a real symmetric positive definite matrix, and $\boldsymbol{z}, \boldsymbol{r}_0 \in \mathbb{R}^N$. Let $p(\cdot)$ be the minimal polynomial of $\boldsymbol{A}$ with degree L ($L \leq N$), and hence, $p(\boldsymbol{A}) = \sum_{l=0}^{L} \zeta_l \boldsymbol{A}^l = 0$. Consequently, we have $\boldsymbol{A}^{-1}\boldsymbol{r}_0 = -\frac{1}{\zeta_0}\sum_{l=1}^{L} \zeta_l \boldsymbol{A}^{l-1}\boldsymbol{r}_0$, which suggests that the solution to the linear system, $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{r}_0$, lies in the Lth Krylov subspace $\mathcal{K}_L(\boldsymbol{A}, \boldsymbol{r}_0)$. The Krylov subspace methods seek to find a vector in the growing Krylov subpace $\mathcal{K}_l(\boldsymbol{A}, \boldsymbol{r}_0)$, $l = 1, 2, \ldots, L$, to approximate the solution. Moreover, in L steps, an exact solution can be obtained.

One commonly used Krylov subspace method is the conjugate gradient (CG) method. It is built upon a set of mutually conjugate directions $\{\boldsymbol{d}_0, \boldsymbol{d}_1, \ldots\}$ such that $\boldsymbol{d}_l^\top \boldsymbol{A}\boldsymbol{d}_{l'} = 0$ for $l \neq l'$. A nice property of the conjugate set is that each new direction $\boldsymbol{d}_l$ can be computed using only the previous direction $\boldsymbol{d}_{l-1}$ without requiring storage of all previous searching directions. Among various iterative methods based on Krylov subspace, the CG method is optimal in the sense that it minimizes the energy function $f(\boldsymbol{z}) = (1/2)\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} - \boldsymbol{r}_0^\top \boldsymbol{z}$ at each iteration. The solution at the lth iteration can be formulated as

$$\boldsymbol{z}_l = \underset{\boldsymbol{z}\in\mathcal{K}_l(\boldsymbol{A},\boldsymbol{r}_0)}{\arg\min}\ f(\boldsymbol{z}), \quad l \geq 0.$$

The detailed algorithm is outlined as follows.

Algorithm 1: Conjugate Gradient Algorithm

Initialization: $\boldsymbol{z}_0 = \mathbf{0}, \boldsymbol{d}_0 = \boldsymbol{r}_0$.

for $l = 1, 2, \ldots$ **do**

  $a_l = (\boldsymbol{r}_{l-1}^\top \boldsymbol{r}_{l-1})/(\boldsymbol{d}_{l-1}^\top \boldsymbol{A}\boldsymbol{d}_{l-1})$

  $\boldsymbol{z}_l = \boldsymbol{z}_{l-1} + a_l \boldsymbol{d}_{l-1}$

  $\boldsymbol{r}_l = \boldsymbol{r}_{l-1} - a_l \boldsymbol{A}\boldsymbol{d}_{l-1}$

  $b_l = (\boldsymbol{r}_l^\top \boldsymbol{r}_l)/(\boldsymbol{r}_{l-1}^\top \boldsymbol{r}_{l-1})$

  $\boldsymbol{d}_l = \boldsymbol{r}_l + b_l \boldsymbol{d}_{l-1}$

end

Here, $\boldsymbol{r}_l = \boldsymbol{r}_0 - \boldsymbol{A}\boldsymbol{z}_l$ is the residual at the lth iteration, which is orthogonal to $\mathcal{K}_l(\boldsymbol{A}, \boldsymbol{r}_0)$. Thus, $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{l-1}$ form an orthogonal basis of $\mathcal{K}_l(\boldsymbol{A}, \boldsymbol{r}_0)$. In fact, $\boldsymbol{d}_l$ and $\boldsymbol{r}_l$ span the same Krylov subspace, thereby the new residual $\boldsymbol{r}_l$ can be calculated using only $\boldsymbol{r}_{l-1}$ and $\boldsymbol{d}_{l-1}$. For a positive definite matrix $\boldsymbol{A}$, the sequence of solutions $\{\boldsymbol{z}_l\}$ converges to the unique solution $\boldsymbol{z}^* = \boldsymbol{A}^{-1}\boldsymbol{r}_0$ in at most N steps. Let ψ_l be the relative error measured at the lth step. It is well known that

$$\psi_l = \frac{f(\boldsymbol{z}_l) - f(\boldsymbol{z}^*)}{f(\boldsymbol{z}_0) - f(\boldsymbol{z}^*)} = \frac{\|\boldsymbol{z}_l - \boldsymbol{z}^*\|_{\boldsymbol{A}}^2}{\|\boldsymbol{z}^*\|_{\boldsymbol{A}}^2} \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^l, \tag{4.2}$$

where $\boldsymbol{z}_0 = \mathbf{0}$ is the initial solution to the linear system, $\kappa \equiv \kappa(\boldsymbol{A}) = \lambda_{\max}/\lambda_{\min}$ is the condition number of $\boldsymbol{A}$, and $\|\boldsymbol{z}\|_{\boldsymbol{A}} \equiv \boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z}$. The `CG` method is monotonically improving, since ψ_l decreases as l increases [Golub and Van Loan, 2012].

Now consider the problem of maximizing the log-likelihood function (4.1). Let $\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}}\{\ell(\boldsymbol{\theta})\}$ denote the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. Let $\boldsymbol{z} = \boldsymbol{\Gamma}^{-1}\boldsymbol{y}$, and we can rewrite (4.1) as

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}| - (1/2)\boldsymbol{y}^\top \boldsymbol{z}.$$

To circumvent the computation of matrix inverse, we can approximate $\boldsymbol{z}$ through the aforementioned `CG` method. Particularly, let $\boldsymbol{z}_l$ be an approximation of $\boldsymbol{z}$ in the lth iteration of the `CG`

algorithm. Therefore, the approximate log-likelihood function at the lth iteration can be written as

$$\widetilde{\ell}^{(l)}(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}| - (1/2)\boldsymbol{y}^\top \boldsymbol{z}_l. \tag{4.3}$$

In Theorem 9, we show the existence of a local maximizer of $\widetilde{\ell}^{(l)}(\boldsymbol{\theta}; \boldsymbol{y})$. To further study the asymptotic properties of the resulting estimator (maximizer), we consider the asymptotic framework in Mardia and Marshall [1984], and denote n as the stage of asymptotics. For our notation convenience, we suppress the subscript n in the main context, except for the theorems, and will restore it in Section 4.5 to avoid confusion.

Theorem 9. *Under (A.1)–(A.5), there exists, with probability tending to one, a local maximizer $\widehat{\boldsymbol{\theta}}^{(l)}$ of $\widetilde{\ell}^{(l)}(\boldsymbol{\theta}; \boldsymbol{y})$, such that*

$$\|\widehat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0\| = O_p\left(\max\left\{\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^{l/2}, N_n^{-1/2}\right\}\right),$$

where $\kappa_0 \equiv \sup_{\boldsymbol{\theta}\in\Omega}\kappa(\boldsymbol{\theta})$ and Ω is an open subset of $\mathbb{R}^q$ such that $\boldsymbol{\theta}_0 \in \Omega$. In particular, if $l > \frac{\sqrt{\kappa_0}+1}{2}\log N_n$, we have $\|\widehat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$.

Theorem 9 shows that the consistency of the estimate $\widehat{\boldsymbol{\theta}}^{(l)}$ is determined by the sample size N, the condition number and the number of iterations l in the CG algorithm. Given a bounded condition number, the estimate $\widehat{\boldsymbol{\theta}}^{(l)}$ achieves root-N consistency in $O(\log N)$ iterations. The proof of Theorem 9 is given in 4.5.1. In practice, to achieve better computational performance, l can be determined by some predetermined stopping criterion for each optimization iteration and hence may vary for different choices of $\boldsymbol{\theta}$.

4.2.2 Log-determinant Approximation via Stochastic Lanczos Method

Next, we will describe an approach to approximate the log-determinant of $\boldsymbol{\Gamma}$, $\log\det(\boldsymbol{\Gamma})$. Consider a real symmetric positive-definite matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, and its eigendecomposition can be written as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q}$ is the matrix whose columns are eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_N\}$ is a diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{A}$ in

ascending order. It is well known that $\log\det(\boldsymbol{A}) = \sum_{i=1}^{N}\log(\lambda_i)$. Consequently, a straightforward approach to calculate the matrix log-determinant can be carried out through the eigendecomposition, which can be costly for large matrix.

An alternative approach is to use a stochastic trace estimator of a matrix logarithm [Hutchinson, 1990]. Note that the logarithm of $\boldsymbol{A}$ can be expressed as $\log(\boldsymbol{A}) = \boldsymbol{Q}\log(\boldsymbol{\Lambda})\boldsymbol{Q}^\top$, and the eigenvalues of $\log(\boldsymbol{A})$ are $\log(\lambda_1), \ldots, \log(\lambda_N)$. Thus, we have

$$\log\det(\boldsymbol{A}) = \mathrm{tr}(\log(\boldsymbol{A})) = \sum_{i=1}^{N}\log(\lambda_i). \tag{4.4}$$

In addition, for any symmetric matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$, we have

$$\mathrm{tr}(\boldsymbol{B}) = \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{u}), \tag{4.5}$$

where $\boldsymbol{u} = (u_1, \ldots, u_N)^\top$ is a vector of independent samples from a random variable with mean 0 and variance 1. Combining (4.4) and (4.5), we can see that

$$\log\det(\boldsymbol{A}) = \mathbb{E}(\boldsymbol{u}^\top\log(\boldsymbol{A})\boldsymbol{u}). \tag{4.6}$$

Therefore, a Monte Carlo estimator of $\mathbb{E}(\boldsymbol{u}^\top\log(\boldsymbol{A})\boldsymbol{u})$ can be used to approximate $\log\det(\boldsymbol{A})$. Amongst all zero mean unit variance random variables, the Rademacher random variable is shown to achieve the minimum variance of $\boldsymbol{u}^\top\log(\boldsymbol{A})\boldsymbol{u}$ [Hutchinson, 1990]. This leads to the Hutchinson trace estimator

$$\log\det(\boldsymbol{A}) \approx N_v^{-1}\sum_{i=1}^{N_v}\boldsymbol{\chi}_i^\top\log(\boldsymbol{A})\boldsymbol{\chi}_i = NN_v^{-1}\sum_{i=1}^{N_v}\boldsymbol{u}_i^\top\log(\boldsymbol{A})\boldsymbol{u}_i, \tag{4.7}$$

where $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_{N_v}$ are independent vectors whose elements are *i.i.d.* Rademacher random variables and $\boldsymbol{u}_i = \boldsymbol{\chi}_i/\|\boldsymbol{\chi}_i\|_2$.

Evaluation of (4.7) is still computationally intractable since the requisite of eigendecomposition remains in finding $\log(\boldsymbol{A})$. To circumvent this problem, one alternative is to further approx-

imate the quadratic form $\boldsymbol{u}^\top \log(\boldsymbol{A})\boldsymbol{u}$ numerically. Note that the analytic function $\log(\cdot)$ can be approximated using the orthonormal polynomial techniques, namely, Taylor's expansions [Zhang and Leithead, 2007], Chebyshev expansions [Han et al., 2015] and their variants [Boutsidis et al., 2017]. Here, we adopt a method based on the Gaussian quadrature rule, which outperforms the aforementioned methods [Ubaru et al., 2017]. Let α_i denote the ith element of $\boldsymbol{Q}^\top \boldsymbol{u}$, we have

$$\boldsymbol{u}^\top \log(\boldsymbol{A})\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{Q} \log(\boldsymbol{\Lambda}) \boldsymbol{Q}^\top \boldsymbol{u} = \sum_{i=1}^{N} \log(\lambda_i)\alpha_i^2,$$

which can be written as a Riemann-Stieltjes integral with piecewise constant measure

$$\sum_{i=1}^{N} \log(\lambda_i)\alpha_i^2 = \int_{\lambda_1}^{\lambda_N} \log(\lambda) d\alpha(\lambda), \tag{4.8}$$

where $\alpha(\cdot)$ is defined as

$$\alpha(t) = \begin{cases} 0, & \text{if } t < \lambda_1, \\ \sum_{k=1}^{i} \alpha_k^2, & \text{if } \lambda_i \leq t < \lambda_{i+1}, \\ \sum_{k=1}^{N} \alpha_k^2, & \text{if } t \geq \lambda_N. \end{cases} \tag{4.9}$$

Furthermore, we can approximate (4.8) via Gaussian quadrature rules, with the general form given by

$$\int_{\lambda_1}^{\lambda_N} \log(\lambda) d\alpha(\lambda) \approx \sum_{i=0}^{m-1} \omega_i \log(\phi_i), \tag{4.10}$$

where $\{(\omega_i, \phi_i), i = 0, 1, \ldots, m-1\}$ ($m << N$) are the weight-node pairs of the m-point Gaussian quadrature rule and can be computed by the Lanczos algorithm [Golub and Welsch, 1969] outlined in Algorithms 2–3.

Algorithm 2: Lanczos Algorithm for Orthonormalization of the Krylov Subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{u})$.

Initialization: $\boldsymbol{q}_1 = \boldsymbol{u}/\|\boldsymbol{u}\|_2, \boldsymbol{Q}_1 = [\boldsymbol{q}_1]$, $a_1 = (\boldsymbol{q}_1)^\top \boldsymbol{A}\boldsymbol{q}_1$, $\boldsymbol{\xi}_1 = (\boldsymbol{A} - a_1\boldsymbol{I})\boldsymbol{q}_1$.

for $k = 2, \ldots, m$ **do**

  $b_k = \|\boldsymbol{\xi}_{k-1}\|_2$

  if $b_k = 0$ **then**

    return $(\boldsymbol{Q}; a_1, \ldots, a_k; b_1, \ldots, b_{k-1})$

  $\boldsymbol{q}_k = \boldsymbol{\xi}_{k-1}/b_k$; $\boldsymbol{Q}_k = [\boldsymbol{Q}_{k-1}, \boldsymbol{q}_k]$

  $a_k = (\boldsymbol{q}_k)^\top \boldsymbol{A}\boldsymbol{q}_k$

  $\boldsymbol{\xi}_k = (\boldsymbol{A} - a_k\boldsymbol{I})\boldsymbol{q}_k - b_k\boldsymbol{q}_{k-1}$

end

output: $(\boldsymbol{Q}_m; a_1, \ldots, a_m; b_1, \ldots, b_{m-1})$

The Lanczos algorithm orthonormalizes the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{u})$ and yields an $n \times m$ matrix $\boldsymbol{Q}_m$ whose columns are orthonormal bases of the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{u})$ and a $m \times m$ tri-diagonal matrix $\boldsymbol{T}_m$ with the diagonal elements being $(a_1, \ldots, a_m)$ and the sub-diagonal elements and super-diagonal elements being $(b_1, \ldots, b_{m-1})$. Denote $p(\cdot)$ as the polynomial of the smallest degree such that $p(\boldsymbol{A})\boldsymbol{u} = 0$ and let M be the degree of $p(\cdot)$, then $\boldsymbol{A}\boldsymbol{Q}_M = \boldsymbol{Q}_M\boldsymbol{T}_M$. Then, the eigenvalues of $\boldsymbol{T}_M$ are also eigenvalues of $\boldsymbol{A}$. Let $\{(\phi_k, \boldsymbol{\Phi}_k), k = 0, 1, \ldots, m-1\}$ be the eigenpairs of $\boldsymbol{T}_m$, we have

$$\boldsymbol{u}^\top \log(\boldsymbol{A})\boldsymbol{u} \approx \sum_{k=0}^{m-1} \omega_k \log(\phi_k) \text{ with } \omega_k = (\boldsymbol{e}_1^\top \boldsymbol{\Phi}_k)^2, \tag{4.11}$$

where $\boldsymbol{e}_1 \in \mathbb{R}^m$ with all elements being zero, except the first element which is equal to 1.

The log-determinant of $\boldsymbol{A}$ can then be approximated as

$$\log|\boldsymbol{A}| \approx \frac{N}{N_v} \sum_{i=1}^{N_v} \left(\sum_{k=0}^{m-1} \omega_k^{(i)} \log(\phi_k^{(i)}) \right), \tag{4.12}$$

where $\{(\phi_k^{(i)}, \omega_k^{(i)}), k = 0, \ldots, m-1\}$ are the eigenvalues and the square of the first element of the eigenvectors corresponding to the ith starting vector. The complete algorithm is outlined in Algorithm 3.

Algorithm 3: Log-Determinant Approximation by Gaussian Quadrature Rule

Input : A p.d. matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, degree m and number of starting vectors N_v.

Output: $\Xi_{m,N_v} = \frac{N}{N_v} \sum_{i=1}^{N_v} \left(\sum_{k=0}^{m-1} \omega_k^{(i)} \log(\phi_k^{(i)}) \right)$.

1 **for** $i = 1, \ldots, N_v$ **do**

2 Draw a random vector $\boldsymbol{\chi}_i$ from the Rademacher distribution as the ith starting vector

3 Calculate $\boldsymbol{T}_m^{(i)}$ through Algorithm 2 with $\boldsymbol{A} = \boldsymbol{\Gamma}$ and $\boldsymbol{v} = \boldsymbol{\chi}_i$

4 Calculate eigenpairs $(\phi_k^{(i)}, \boldsymbol{\Phi}_k^{(i)})$ of $\boldsymbol{T}_m^{(i)}$ and compute $\omega_k^{(i)} = (\boldsymbol{e}_1^\top \boldsymbol{\Phi}_k^{(i)})^2$ for $k = 0, 1, \ldots, m-1$.

Combined with the `CG` algorithm, we further approximate the log-likelihood function with

$$\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\Xi_{m,N_v} - (1/2)\boldsymbol{y}^\top \boldsymbol{z}_l, \tag{4.13}$$

where Ξ_{m,N_v} is the approximation of $\log\det(\boldsymbol{\Gamma})$ by Algorithm 3, and $\boldsymbol{z}_l$ is the solution (i.e., approximation) of $\boldsymbol{\Gamma}^{-1}\boldsymbol{y}$ at the lth iteration of the `CG` algorithm. Maximizing $\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}; \boldsymbol{y})$ yields an approximate estimate of $\boldsymbol{\theta}$. We demonstrate the existence and consistency of such estimator in the following Theorem 10.

Theorem 10. *Under (A.1)-(A.5), there exists, with probability tending to one, a local maximizer* $\widehat{\boldsymbol{\theta}}^{(l,m,N_v)}$ *of* $\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}; \boldsymbol{y})$*, such that*

$$\|\widehat{\boldsymbol{\theta}}^{(l,m,N_v)} - \boldsymbol{\theta}_0\| = O_p\left(\max\left\{ \left(\frac{\sqrt{\kappa_0} - 1}{\sqrt{\kappa_0} + 1} \right)^{l/2}, \left(\frac{\sqrt{\kappa_0} - 1}{\sqrt{\kappa_0} + 1} \right)^{m}, N_n^{-1/2} \right\} \right).$$

In the special case when $l > \frac{\sqrt{\kappa_0}+1}{2} \log N_n$ *and* $m > \frac{\sqrt{\kappa_0}}{4} \log(N_n C_1)$*, where* $C_1 = \lambda_{\max}\sqrt{\kappa_0}\log(\lambda_{\max} + \lambda_{min})$*, we have* $\|\widehat{\boldsymbol{\theta}}^{(l,m,N_v)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$.

The proof of Theorem 10 is given in 4.5.2. Theorem 10 establishes that the consistency of the estimate $\widehat{\boldsymbol{\theta}}^{(l,m,N_v)}$ is determined by the sample size N, the condition number, the number of iterations in the CG algorithm l and the order of Gaussian quadrature rule m. Given a well-conditioned covariance matrix, the estimate $\widehat{\boldsymbol{\theta}}^{(l)}$ achieves root-N consistency in $O(\log N)$ iterations. We also note that the convergence result does not depend on the number of Monte Carlo steps N_v. This is further demonstrated empirically in our simulation study. As will be seen in Section 4.3.3, the number of starting vectors N_v has negligible impact on the accuracy of log-likelihood evaluation.

4.2.3 Generalization to Spatial Linear Regression Model

The methodology in the previous sections can be readily generalized to spatial linear regression model. Consider a Gaussian process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^d\}$ such that

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^\top \boldsymbol{\beta} + \varepsilon(\boldsymbol{s}), \tag{4.14}$$

where $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))^\top$ is a $p \times 1$ vector of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of regression coefficients. The error process $\varepsilon(\boldsymbol{s})$ is assumed to have zero mean, and the covariance between $\varepsilon(\boldsymbol{s})$ and $\varepsilon(\boldsymbol{s}')$ is $\gamma(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) = \mathrm{Cov}\{\varepsilon(\boldsymbol{s}), \varepsilon(\boldsymbol{s}')\}$. For $j = 1, \ldots, p$, we write $\boldsymbol{x}_j = (x_j(\boldsymbol{s}_1), \ldots, x_j(\boldsymbol{s}_N))^\top$, where $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N \in \mathcal{D}$ are the sampling locations. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{N \times p}$ denote the design matrix, then the log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}(\boldsymbol{\theta})| - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4.15}$$

For any given $\boldsymbol{\theta}$, maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ yields the profile likelihood estimate (PLE) of $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}}_{\text{PLE}}(\boldsymbol{\theta}) = (\boldsymbol{X}^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1} \boldsymbol{y}.$$

The term $\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\boldsymbol{X}$ in the formula involves the problem of solving linear systems with the same coefficient matrix but different right-hand sides, which can also be carried out iteratively using the

CG method. However, the conjugate gradient computations are still computationally demanding, especially when the number of covariates p is large. A more computationally efficient approach is to obtain a consistent estimate of $\boldsymbol{\beta}$ separately, and then estimate $\boldsymbol{\theta}$ based on the vector of residuals using our proposed approximate likelihood approach in Sections 4.2.1–4.2.2. In the following theorem, we show that the resulting estimate of $\boldsymbol{\theta}$ is consistent and achieve the same convergence rate as shown in Theorem 10.

Theorem 11. *Under (A.1)-(A.6), for any given $\widetilde{\boldsymbol{\beta}}$ satisfying $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(N_n^{-1/2})$, there exists, with probability tending to one, a local maximizer $\widehat{\boldsymbol{\theta}}^{(l,m,N_v)}$ of $\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}; \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})$, such that the results in Theorem 10 hold.*

In practice, we propose to use the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. Note that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is root-N consistent; see 4.5.3. Therefore, our proposed two-step estimation procedure will result in a consistent estimate of $\boldsymbol{\theta}$. The proof of Theorem 11 is given in 4.5.3.

4.3 Computational Aspects

In this section, we first discuss the computational complexity of our proposed methodology in Section 4.3.1. In Section 4.3.2, we provide a fast Krylov covariance tapering method. The trade-off between accuracy and computational complexity, using different choices of m and N_v is explored in Section 4.3.3. Then, we demonstrate the performance of our proposed method in parameter estimation, using compactly supported covariances in Section 4.3.4. Finally, we compare our proposed method to competitors including covariance tapering and the NNGP method [Datta et al., 2016a] in Section 4.3.5.

4.3.1 Computational Complexity

The computational complexity of evaluating the approximate log-likelihood function (4.13) is dominated by the conjugate gradient algorithm in Section 4.2.1 and the stochastic log-determinant approximation algorithm in Section 4.2.2. As can be seen from Algorithm 1, the conjugate gradient algorithm involves only scalar product of vectors, inner products of vectors and matrix-vector

multiplications of dimension N at each iteration of the algorithm. The conjugate gradient algorithm can be performed in $O(lN^2)$ flops, where l is the number of iterations of the conjugate gradient algorithm. Theorem 9 shows that $O(\frac{\sqrt{\kappa_0}+1}{2}\log N)$ number of iterations ensures consistent estimate of the parameters. In practice, the convergence rate of the conjugate gradient method can be improved by various preconditioning methods, of which the Jacobi preconditioner [Saad, 2003] is the most efficient one in our implementation. In our simulation studies, a rather small value of iterations (less than 100) is usually sufficient with the Jacobi preconditioning technique. The log-determinant approximation algorithm involves N_v Monte Carlo runs, and each run involves Krylov subspace orthogonalization of dimension m via Lanczos algorithm and eigen decomposition of an $m \times m$ matrix. Similar to the conjugate gradient algorithm, each iteration of the Lanczos algorithm requires only basic linear algebra subroutines, hence can be performed in $O(N^2)$ steps. The eigen decomposition can be achieved in $O(m^3)$ flops, which is dominated by the Lanczos step. Therefore, the log-determinant approximation algorithm costs $O(mN_vN^2)$ flops. By Theorem 10, consistency of the parameters is guaranteed by letting m increase with N at a rate of $O(\log N)$. Thus, the log-determinant approximation algorithm has $O(N_vN^2 \log N)$ time complexity.

For dense covariance matrices, our proposed method provides a substantial improvement over the traditional method based on Cholesky decomposition, $O(N^2)$ compared to $O(N^3)$. Furthermore, our complexity analysis above reveals that we can achieve quasi-linear complexity by exploiting sparsity. Indeed, sparse matrix-vector multiplications involves only $O(\|\mathbf{\Gamma}\|_0)$ operations, which means our algorithm can be implemented in $O((mN_v + l)\|\mathbf{\Gamma}\|_0)$ flops, where $\|\mathbf{\Gamma}\|_0$ is the number of nonzero entries of $\mathbf{\Gamma}$.

In the following Section 4.3.2, we present a fast Krylov subspace method, which can achieve similar levels of performance as the classical method via Cholesky decomposition in quasi-linear time.

4.3.2 Fast Krylov Covariance Tapering

As illustrated in Sections 4.3.1, our proposed method can achieve quasi-linear complexity by exploiting sparsity. One way to introduce sparsity into computation is by covariance tapering

[Furrer et al., 2006, Kaufman et al., 2008]. Here, we propose a fast covariance tapering approach by using Krylov subspace methods. The tapered covariance function, constructed by multiplying the covariance function with a compactly supported covariance function, is a valid covariance function but with compact support [Furrer et al., 2006]. The covariance-tapered log-likelihood function is given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}_{\text{tap}}| - (1/2)\boldsymbol{y}^\top \boldsymbol{\Gamma}_{\text{tap}}^{-1}\boldsymbol{y}, \tag{4.16}$$

where $\boldsymbol{\Gamma}_{\text{tap}}$ is the tapered covariance. The tapered log-likelihood function can then be approximated by

$$\widetilde{\ell}_{\text{tap}}^{(l,m,N_v)}(\boldsymbol{\theta}; \boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\Xi_{\text{tap},m,N_v} - (1/2)\boldsymbol{y}^\top \boldsymbol{z}_{\text{tap}}^{(l)}, \tag{4.17}$$

where Ξ_{tap,m,N_v} is the stochastic lanczos estimator of $\log\det(\boldsymbol{\Gamma}_{\text{tap}})$, and $\boldsymbol{z}_{\text{tap}}^{(l)}$ is the approximation of $\boldsymbol{\Gamma}_{\text{tap}}^{-1}\boldsymbol{y}$ at the lth iteration. Maximizing (4.17) yields the approximated covariance estimates. For linear regression model, similar estimation procedure as in Section 4.2.3 can be readily applied.

4.3.3 Computational Efficiency

To evaluate the efficiency and accuracy of our proposed Krylov covariance tapering method, we consider a zero-mean Gaussian process and simulate datasets as follows. First, the locations are uniformly sampled from a two-dimensional spatial domain $[0, \sqrt{N}/2]^2$, where N is the sample size. This setting guarantees a fixed sampling density of 4. For each simulated dataset, a centered Gaussian process is generated with an exponential covariance function

$$\gamma_{\exp}(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) = \sigma^2(1-c)\exp(-\|\boldsymbol{s} - \boldsymbol{s}'\|/r), \tag{4.18}$$

where $r = 2$ is the range parameter, $c = 0.2$ is the nugget proportion such that $c\sigma^2$ is the nugget effect, $\sigma^2 = 9$ is the variance, and $\boldsymbol{\theta} = (r, c, \sigma^2)^\top$.

To assess the run time of likelihood evaluation, we simulate datasets with sample sizes roughly from $5,000$ to $250,000$. We construct the tapered covariance using a Wendland tapering kernel

[Wendland, 1995]

$$\gamma_\delta(\boldsymbol{s}, \boldsymbol{s}') = (1 - \|\boldsymbol{s} - \boldsymbol{s}'\|/\delta)_+^4 \, (1 + 4\|\boldsymbol{s} - \boldsymbol{s}'\|/\delta)\,, \tag{4.19}$$

where δ is the range parameter that controls the sparsity of the covariance matrix.

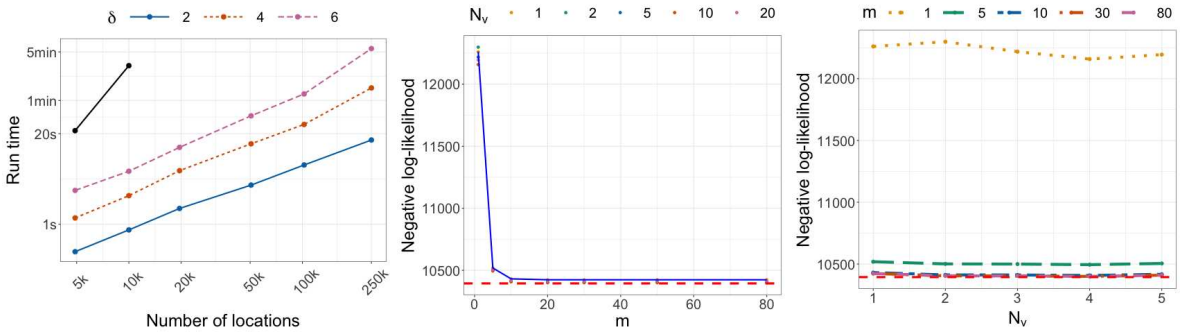

Figure 4.1: (a) Run time for a single iteration of likelihood evaluation by number of locations with different levels of sparsity. Both run time and number of locations are on a log scale. The black line indicates the run time for exact likelihood calculation using Cholesky decomposition. (b) Negative log-likelihood with increasing m, under different number of Monte Carlo iterations N_v. The connected dots in blue solid line is when $N_v = 1$. (c) Negative log-likelihood using increasing number of Monte Carlo iterations N_v and different m. In both (b) and (c), the red dashed line indicates the exact negative log-likelihood.

The accuracy of our proposed likelihood approximation method depends on the dimension of Krylov subspace m and the number of Monte Carlo samples N_v. Here, we fix the sample size N at $4,900$ and the tapering range parameter δ at 6. The effects of m is evaluated first, and results are reported in Figure 4.1b. The approximated negative log-likelihood is plotted against m with $N_v = 1, 2, 5, 10$ and 20. We connect the dots only for $N_v = 1$, as all the cases are closely overlapping each other. For m less than 5, the accuracy of likelihood calculation is relatively poor compared with the exact negative log-likelihood, shown as dashed line type. Increasing m from 1 to 5 effectively reduces the approximation error. Moreover, $m = 20$ seems to be sufficient for this experiment since the marginal improvement on accuracy is negligible beyond that. Next, we evaluate the effect of N_v, and the results are reported in Figure 4.1c. The approximated negative log-likelihood is plotted against N_v for values of m fixed at $1, 5, 10, 30$ and 80. The number of starting vectors has negligible influence on the quality of approximation. Even with one starting

vector, as long as $m \geq 10$, the approximation error is small. Based on our findings here, we will fix $m = 30$ and $N_v = 1$ in the rest of this chapter.

Figure 4.1a shows the run time (averaged over 10 replicates) for a single iteration of likelihood evaluation with sample size N. The black curve represents the time needed for exact likelihood, and the execution time is reported for values of N less than 10,000. For sparse method, $\delta = 2, 4$ and 6 are considered. As δ increases, the sparsity of covariance matrix decreases, and therefore, the computational time increases. More importantly, compared with exact likelihood, the computational time of the sparse method increases much slower than exact likelihood, as sample size increases. For $\delta = 2, 4$ and 6, the computational time increases almost linearly, which supports the computational complexity analysis in Section 4.3.1.

4.3.4 Performance of Parameter Estimation

As demonstrated in Sections 4.3.1-4.3.3, our proposed method can achieve quasi-linear complexity by exploiting sparsity. One way to introduce sparsity into computation is using correlation functions with compact support. It is not rare in real world applications that observed correlations among observations vanish beyond a certain distance, in which cases compactly supported covariances are appealing [Gneiting, 2002b]. The first purpose of this section is to benchmark our proposed method against exact likelihood method using optimized sparse Cholesky factorization with the `spam` package in `R` [Furrer and Gerber, 2008]. In addition, we demonstrate the effectiveness of the proposed estimation procedure in Section 4.2.3 by comparing it to an alternative method. We let the consistent estimate of $\boldsymbol{\beta}$ in the first step be the approximated PLE $\widetilde{\boldsymbol{\beta}}_{\text{PLE}}(\boldsymbol{\theta})$ using `CG` algorithm. We denote this alternative method as `Krylov-gls` and the one in Section 4.2.3 as `Krylov-ols`, respectively, which should cause no confusion.

For the spatial linear model in (4.14), we generate two covariates from a normal distribution with unit variance and cross-covariate correlation of 0.5. The covariates are then standardized to have mean 0 and variance 1. The regression coefficients are fixed at $\boldsymbol{\beta} = (2, 1)^\top$. The compactly supported covariance function considered is the product of the exponential covariance function (4.18) and the Wendland kernel function (4.19)

$$\gamma_{\text{csc}}(\boldsymbol{s}, \boldsymbol{s}') = \gamma_{\text{exp}}(\boldsymbol{s}, \boldsymbol{s}') \cdot \gamma_{\delta}(\boldsymbol{s}, \boldsymbol{s}') \tag{4.20}$$

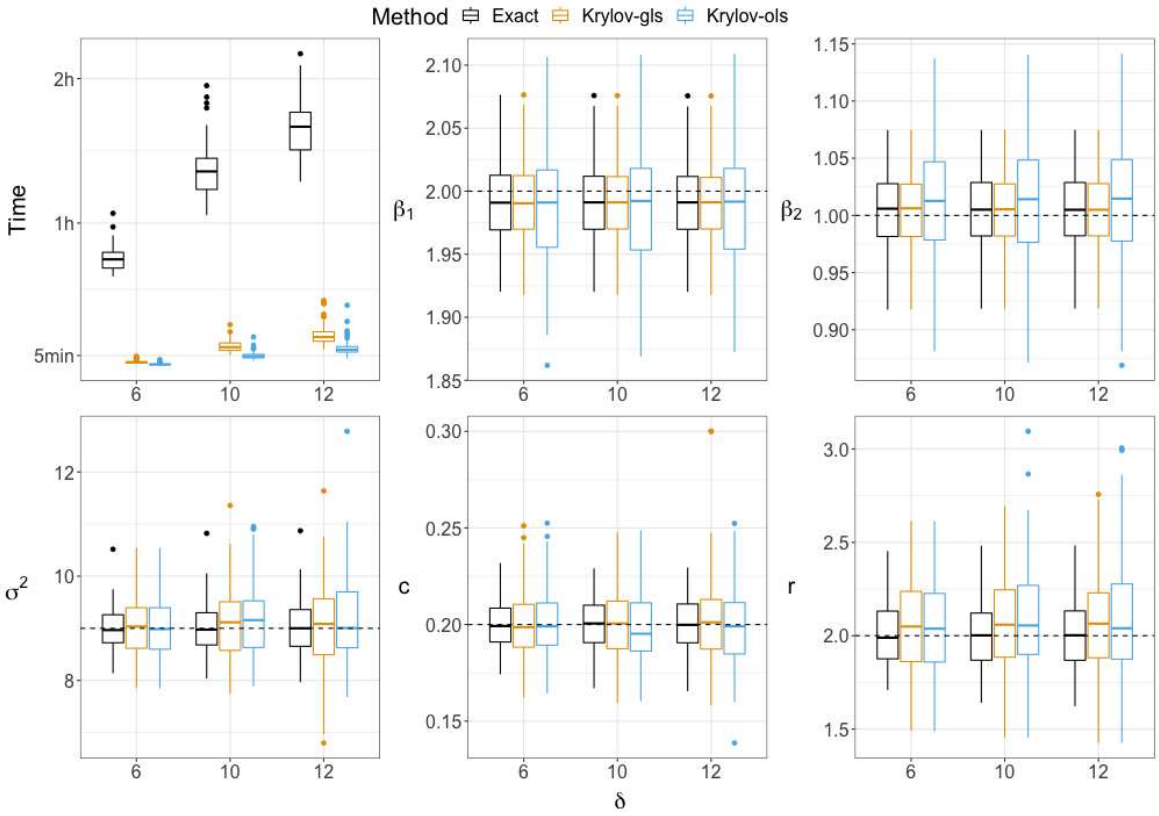

Figure 4.2: Boxplots for execution time, regression coefficients (β_1, β_2) and covariance parameters (σ^2, c, r) for $\delta \in \{6, 10, 12\}$ under maximum likelihood method (exact) and two Krylov subspace methods: `Krylov-gls` and `Krylov-ols`.

We generate 100 simulated datasets of sample size $N = 4,900$. The threshold parameter δ is set to be $6, 10, 12$, indicating different levels of sparsity. Figure 4.2 shows boxplots of both the run time of parameter estimation and the estimates. At the sparsity level $\delta = 6$, the optimizing time is reduced by a factor of approximately 20 and 33 for `Krylov-gls` and `Krylov-ols`, respectively. The simulation results show that the proposed Krylov approximation can greatly improve computational efficiency for sparse matrix. For regression coefficients, both the point estimates and variability of the regression coefficients based on `Krylov-gls` are nearly identical to the maximum likelihood estimates, for each level of sparsity considered. As expected, the `Krylov-ols` method has slightly larger bias and variance compared to another two methods,

indicating a loss of statistical efficiency associated with unconsidered correlation. For covariance parameters, the performance of `Krylov-ols` is nearly as good as `Krylov-gls`, showing that the estimation of mean trend has negligible association with covariance estimation. Both Krylov-based methods have larger variances compared to the maximum likelihood estimates, showing decreased efficiency of our method due to approximation error. However, the accuracy of parameter estimation using Krylov-based methods is comparable to the exact method, which is consistent with the conclusions in Theorem 11.

4.3.5 Comparison

To further illustrate the capabilities of our proposed fast Krylov covariance tapering method, we consider several alternative approaches for comparison. As in Section 4.3.4, we consider an alternative Krylov covariance tapering method by finding $\widetilde{\boldsymbol{\beta}}_{\text{PLE}}(\boldsymbol{\theta})$ using `CG` algorithm in the first step. We denote this alternative method and the fast Krylov covariance tapering method as `Krylov-gls` and `Krylov-ols`, respectively, which should cause no confusion. Note that, both `Krylov-gls` and `Krylov-ols` are compared to the covariance tapering method (referred as `Tapering`). In addition, we consider a state-of-art method in literature, the nearest-neighbor Gaussian process method (`NNGP`; Datta et al. [2016a]), as a competitor. The exact maximum likelihood estimation serves as a benchmark to evaluate the performance of the aforementioned methods. These approaches are implemented in `R`, using extensively the `spam` package for covariance tapering, `spNNGP` for nearest-neighbor Gaussian process method, and `spKrylov` (https://github.com/liujl93/spKrylov/) for fast Krylov covariance tapering method.

Consider two simulated examples of sample sizes $5,390$ and $11,000$ both from a Gaussian process with a linear mean function as described in Section 4.3.4 and with a Matérn covariance function defined as

$$\gamma_{\text{mat}}(\boldsymbol{s},\boldsymbol{s}';\boldsymbol{\theta}) = \sigma^2(1-c)\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{r}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{\|\boldsymbol{s}-\boldsymbol{s}'\|}{r}\right). \tag{4.21}$$

Here, $\sigma^2 = 9$ is the variance, $c = 0.2$ is the nugget proportion such that $c\sigma^2$ is the nugget effect and $r = 2$ is the scale parameter, and ν is the smooth parameter. We set $\nu = 0.5$ for the first example and $\nu = 2.5$ for the second one. In each simulated dataset, 90% of the observations are used for model fitting, and the rest are used for evaluation of predictive performance. For observations in the hold-out set, denoted as $y_{i,\text{new}}$, $i = 1, \ldots, N_{\text{new}}$, we use the mean squared prediction error (MSPE) to evaluate the prediction accuracy

$$\text{MSPE} = N_{\text{new}}^{-1} \sum_{i=1}^{N_{\text{new}}} (y_{i,\text{new}} - \widehat{y}_{i,\text{new}})^2.$$

For the `NNGP` method, the numbers of nearest neighbors are fixed at 10, 20 and 30. For tapering methods, we taper the Matérn covariance function as follows

$$\gamma_{\text{tap}}(\boldsymbol{s}, \boldsymbol{s}') = \gamma_{\text{mat}}(\boldsymbol{s}, \boldsymbol{s}') \cdot \gamma_{\delta}(\boldsymbol{s}, \boldsymbol{s}') \tag{4.22}$$

Three different levels of sparsity are considered with threshold parameter δ set to be $6, 10$ or 12. The first dataset is moderately large to accommodate exact and tapered likelihood calculations, whereas the second example highlights our superiority over the `NNGP` method with a larger sample size. Both simulations are repeated for 100 iterations, and the results are summarized in Figure 4.3 and Figure 4.4 for the first and second simulated dataset, respectively.

Our proposed approximate tapering method involves two stages of approximation. At the first stage, covariance tapering is applied, which assumes a misspecified covariance with distant location pairs forced to be independent. At the second stage, numerical approximations to the tapered likelihood function are obtained within the Krylov subspace. To illustrate the approximation capabilities of the first stage, we compare `Tapering` method to the exact maximum likelihood estimates. In terms of prediction accuracy and regression coefficient estimates, `Tapering` method demonstrates nearly indistinguishable performance with the exact MLE. On the contrary, the tapered covariance parameter estimates have larger biases, especially when δ is small, as compared to the exact method. As expected, these biases decrease as we increase the value of δ. For the second

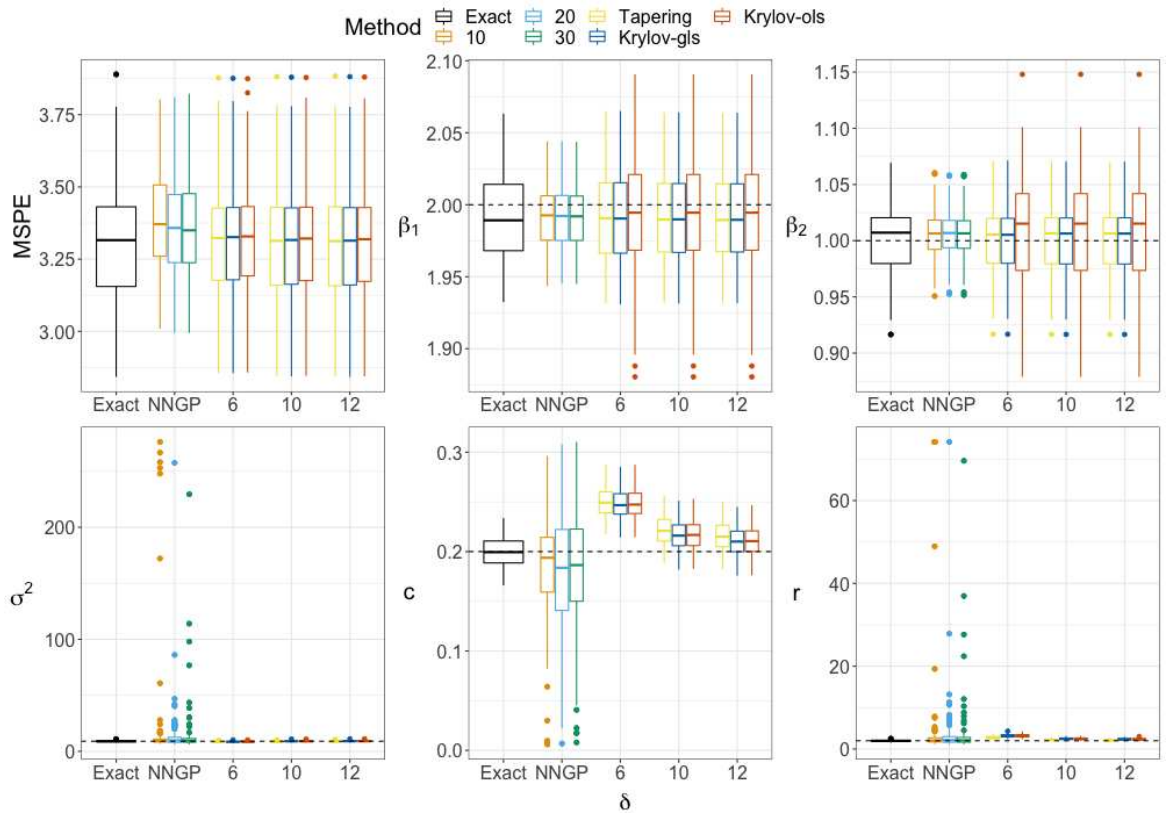

Figure 4.3: First Simulated Dataset: Boxplots for mean squared prediction error (MSPE), regression coefficients (β_1, β_2) and covariance parameters (σ^2, c, r) for $\delta \in \{6, 10, 12\}$ under maximum likelihood method (Exact), covariance tapering method (`Tapering`), Krylov covariance tapering methods (`Krylov-gls` and `Krylov-ols`) and nearest-neighbor Gaussian process model (`NNGP`). For `NNGP` method, the number of nearest neighbors are chosen to be 10, 20 and 30. For `Krylov-gls` and `Krylov-ols`, $\delta = 6, 10, 12$.

stage, both `Krylov-ols`, `Krylov-gls` exhibit almost identical estimation and prediction performance to `Tapering`, for all three cases ($\delta = 6, 10, 12$), with the exception that `Krylov-ols` yield larger variances of regression coefficients.

Finally, we compare our proposed model to `NNGP` method. Surprisingly, `NNGP` has similar accuracy but superior efficiency in terms of regression estimates, as compared to all other approaches including the exact maximum likelihood estimates. However, `NNGP` is dominated by tapering based methods in terms of both covariance estimation and model prediction. As shown in Table 4.1, consistently among all the methods considered, the point estimates of $\boldsymbol{\beta}$ are accurate and the empirical confidence intervals contain the true value. We note that `Krylov-gls` has slightly larger confidence intervals compared to others, which might due to the approximation

error in finding $\boldsymbol{\Gamma}^{-1}\boldsymbol{X}$. Both `Tapering` and Krylov subspace methods give similar accuracy and precision in the estimation of covariance parameters. However, for `NNGP` method, the covariance parameter estimates are highly unstable and inaccurate, which can also be seen from Figure 4.3. Additionally, the MSPE of `NNGP` method ranges from 3.36 to 3.38, which is much higher than that of our proposed method, ranging from 3.31 to 3.32.

Parameter	True value						$\delta = 6$			$\delta = 10$			$\delta = 12$		
Method		Percentile	Exact	NNGP(10)	NNGP(20)	NNGP(30)	Tapering	Krylov-gls	Krylov-ols	Tapering	Krylov-gls	Krylov-ols	Tapering	Krylov-gls	Krylov-ols
β_1	2	50%	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99
		2.5%	1.95	1.95	1.95	1.95	1.95	1.90	1.95	1.95	1.90	1.95	1.95	1.90	1.95
		97.5%	2.06	2.04	2.03	2.03	2.06	2.08	2.06	2.06	2.08	2.06	2.06	2.08	2.06
β_2	1	50%	1.01	1.01	1.01	1.01	1.01	1.02	1.01	1.01	1.02	1.01	1.01	1.02	1.01
		2.5%	0.95	0.96	0.96	0.96	0.94	0.90	0.95	0.95	0.90	0.95	0.95	0.90	0.95
		97.5%	1.06	1.05	1.05	1.05	1.06	1.10	1.06	1.06	1.10	1.06	1.06	1.10	1.06
σ^2	9	50%	8.91	9.12	9.75	9.44	8.22	8.53	8.52	8.61	9.01	9.01	8.70	9.08	9.08
		2.5%	7.97	6.90	6.67	6.62	7.46	7.70	7.70	7.76	8.09	8.08	7.82	8.13	8.15
		97.5%	10.18	255.76	44.58	87.92	9.06	9.44	9.44	9.61	10.13	10.13	9.75	10.26	10.26
c	0.2	50%	0.20	0.19	0.18	0.19	0.25	0.25	0.25	0.22	0.22	0.22	0.22	0.21	0.21
		2.5%	0.17	0.01	0.04	0.02	0.22	0.22	0.22	0.19	0.19	0.19	0.19	0.18	0.18
		97.5%	0.23	0.26	0.27	0.27	0.28	0.28	0.28	0.25	0.24	0.24	0.24	0.24	0.24
r	2	50%	1.95	2.06	2.15	2.09	2.73	3.11	3.11	2.21	2.46	2.46	2.13	2.34	2.34
		2.5%	1.66	1.27	1.23	1.26	2.20	2.48	2.48	1.84	2.03	2.03	1.79	1.96	1.96
		97.5%	2.43	74.16	12.28	25.16	3.59	4.15	4.16	2.76	3.08	3.08	2.64	2.91	2.91
MSPE	—	—	3.31	3.38	3.36	3.36	3.32	3.32	3.32	3.31	3.32	3.31	3.31	3.32	3.31

Table 4.1: Percentiles of estimates of regression and covariance parameters under Krylov covariance tapering methods (`Krylov-gls` and `Krylov-ols`), maximum likelihood method (Exact), nearest-neighbor Gaussian process model (`NNGP`), covariance tapering method (`Tapering`). For `NNGP` method, the number of nearest neighbors are chosen to be 10, 20 and 30. For `Krylov-gls` and `Krylov-ols`, $\delta = 6, 10, 12$.

To further illustrate the efficiency and accuracy of the `NNGP` method and Krylov method, we fit a model for the second simulated example using `NNGP` method with $10, 20, 30$ and 50 nearest neighbors and Krylov subspace methods with tapering parameter $\delta = 4, 6$ and 10. We record both the run time and MSPE of both methods in Figure 4.4. As expected, increasing the number of nearest neighbors in the `NNGP` method decreases MSPE at a sacrifice of computational efficiency. Similarly for Krylov subspace methods, a significant drop in MSPE is obtained by increasing δ from 4 to 6, followed by a plateau, showing that $\delta = 6$ provides a sufficiently close approximation to the exact method. On the contrary, the `NNGP` method with 50 nearest neighbors provides a slightly inferior alternative, but the computation becomes extremely time consuming. Additionally, the `Krylov-ols` method achieves notable computational savings over `Krylov-gls` with almost identical predictive performance, for all levels of sparsity considered.

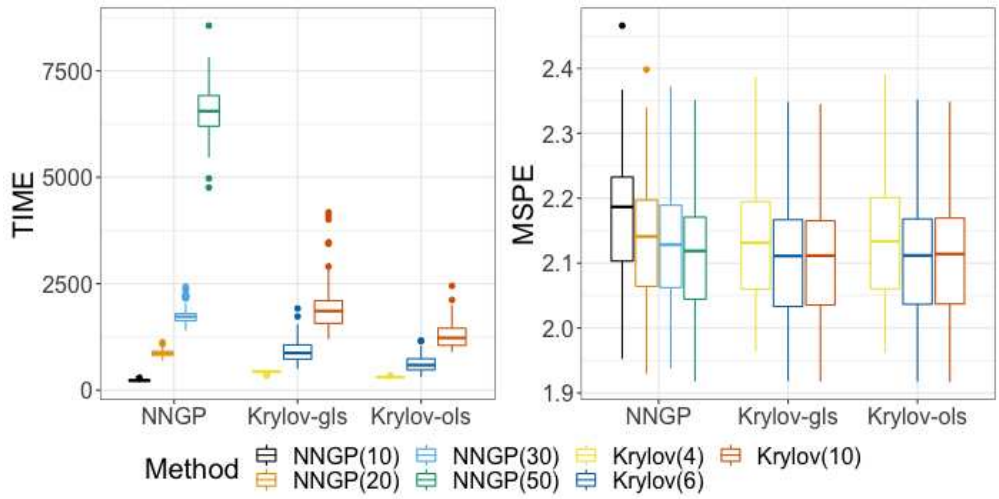

Figure 4.4: Second Simulated Dataset: Boxplots for execution time of parameter estimation and mean squared prediction error (MSPE) under Krylov covariance tapering methods (`Krylov-gls` and `Krylov-ols`) and nearest-neighbor Gaussian process model (`NNGP`). For `NNGP` method, the number of nearest neighbors are chosen to be 10, 20, 30 and 50. For `Krylov-gls` and `Krylov-ols`, $\delta = 4, 6, 10$.

4.4 Application to the LiDAR Data

In this section, we employ our methodology to the analysis of a big data set with $N = 5.025 \times 10^6$ LiDAR-based estimates of forest canopy height in west Alaska during a 2014 Tanana Inventory Unit (TIU) campaign [Cook et al., 2013, Finley et al., 2019] . The covariates of interest are: the measurements of tree cover at a spatial resolution of 30 meters and the occurrence of forest fire [Hansen et al., 2013]. The tree cover is measured in percentage for peak growing season in 2010, and the fire occurrence is encoded as 1 if the fire ever occurs within the past 20 years and 0 otherwise. As illustrated in Figure 4.5, the spatial domain is irregular. Since the dataset is randomly sampled from a full dataset of size $28,751,400$, the locations are also irregularly spaced.

To characterize the relationship of forest canopy height and the two covariates, we fit a Gaussian process model with a linear mean trend and an exponential covariance function (4.18) with parameters σ^2, c and r. As discussed in Section 4.3.2, we consider a tapered approximation of the likelihood function using the Wendland covariance function (4.19) with a threshold $\delta = 0.6$. Further, we encode the dataset to 200 blocks, with $25,000$ observations within each block, using a mean-distance-ordered blocking structure. The mean-distance-ordered blocking structure is inspired by an efficient mean-distance-ordered search algorithm for image vector quantization [Ra

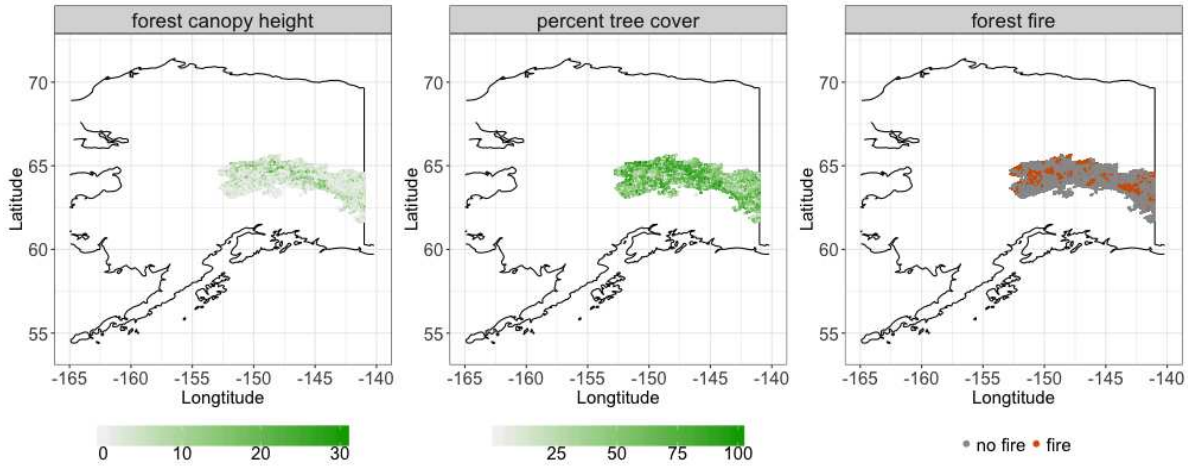

Figure 4.5: Maps for forest canopy height, tree cover and forest fire in west Alaska.

and Kim, 1993]. The idea is originated from the inequality between arithmetic mean and quadratic mean, i.e., $(a+b)^2 \leq 2(a^2+b^2)$. For two locations $\boldsymbol{s} = (s_1, s_2)$ and $\boldsymbol{s}' = (s_1', s_2')$, we define $w = \frac{s_1+s_2}{\sqrt{2}}$ and $w' = \frac{s_1'+s_2'}{\sqrt{2}}$. Then,

$$\|\boldsymbol{s} - \boldsymbol{s}'\| = \sqrt{(s_1 - s_1')^2 + (s_2 - s_2')^2} \geq \frac{|s_1 + s_2 - s_1' - s_2'|}{\sqrt{2}} = |w - w'|.$$

That is, the Euclidean distances between two locations is bounded from below by a multiple of the absolute difference of mean of coordinate vectors. The mean-distance-ordered blocking scheme is outlined as follows: (1) let $\boldsymbol{w} = (w_1, \ldots, w_N)'$ be the codevector of $\mathcal{S} = (\boldsymbol{s}_1', \ldots, \boldsymbol{s}_N')$, where $\boldsymbol{s}_i = (s_{i1}, s_{i2})'$ and $w_i = \frac{s_{i1}+s_{i2}}{\sqrt{2}}$; (2) order the dataset according to the $\boldsymbol{w}$ ordering; (3) collapse adjacent observations to subsets of (approximate) equal size.

We use both the Krylov and `NNGP` method to obtain the parameter estimates and prediction. The regression coefficients include an intercept and two slope regression coefficients, referred as β_0, $\beta_{\text{tree cover}}$ and β_{fire}. Table 4.2 shows that both Krylov method and `NNGP` method yield comparable parameter estimates and prediction. This comparable performance is also reflected in Figure 4.6. The residual map for the hold-out dataset using Krylov and `NNGP` method are almost identical. Indeed, the root mean squared prediction error (RMSPE) of the $25,000$ hold-out observations for Krylov method is slightly smaller than that achieved by the `NNGP` method.

Method	β_0	$\beta_{\text{tree cover}}$	β_{fire}	σ^2	c	r	RMSPE
Krylov	2.398	0.022	0.747	18.012	0.067	0.201	1.707
NNGP	2.429	0.022	0.54	19.455	0.053	0.183	1.709

Table 4.2: Point estimates of regression and covariance parameters and root mean squared prediction error (RMSPE) for the Krylov and NNGP methods, respectively.

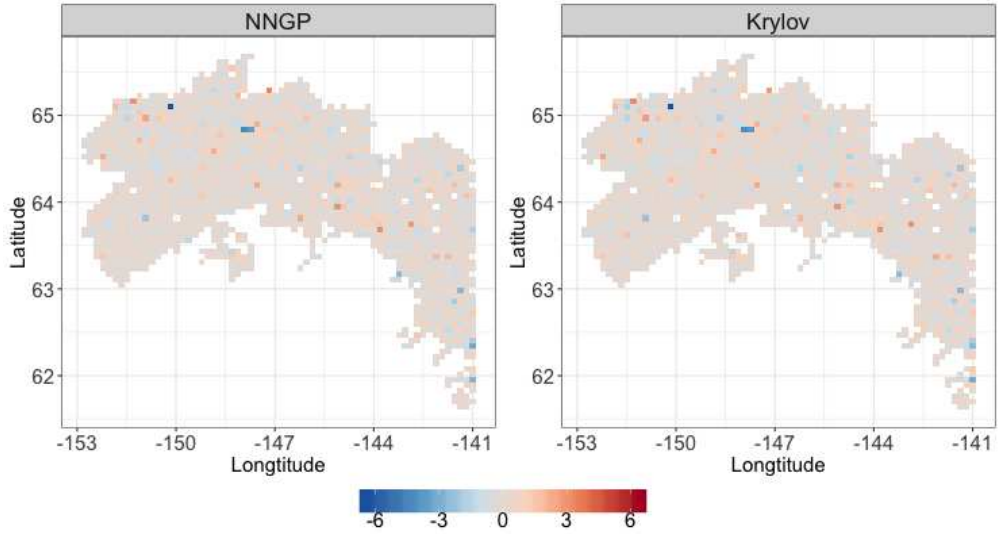

Figure 4.6: Residuals maps by the Krylov and NNGP methods, respectively.

4.5 Technical Details

Let $\boldsymbol{\theta}_0$ denote the vector of true covariance parameters. Let $\ell'(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\ell''(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ denote the first- and second-order derivatives of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, respectively. For $\iota, \iota' = 1, \ldots, q$, the ιth element of $\ell'(\boldsymbol{\theta})$ is $-(1/2)\,\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota) - (1/2)\boldsymbol{y}^\top\boldsymbol{\Gamma}^\iota\boldsymbol{y}$, where $\boldsymbol{\Gamma}_\iota = \partial\boldsymbol{\Gamma}/\partial\theta_\iota$ and $\boldsymbol{\Gamma}^\iota = \partial\boldsymbol{\Gamma}^{-1}/\partial\theta_\iota = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}$. The (ι, ι')th entry of $\ell''(\boldsymbol{\theta})$ is $-(1/2)\,\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota\iota'} + \boldsymbol{\Gamma}^\iota\boldsymbol{\Gamma}_{\iota'}) - (1/2)\boldsymbol{y}^\top\boldsymbol{\Gamma}^{\iota\iota'}\boldsymbol{y}$, where $\boldsymbol{\Gamma}^{\iota\iota'} = \partial^2\boldsymbol{\Gamma}/\partial\theta_\iota\partial\theta_{\iota'} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'} + \boldsymbol{\Gamma}_{\iota'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota - \boldsymbol{\Gamma}_{\iota\iota'})\boldsymbol{\Gamma}^{-1}$.

Let $\lambda_1 \leq \ldots \leq \lambda_N$, $|\lambda_1^\iota| \leq \ldots \leq |\lambda_N^\iota|$ and $|\lambda_1^{\iota\iota'}| \leq \ldots \leq |\lambda_N^{\iota\iota'}|$, $\iota, \iota' = 1, \ldots q$ denote the eigenvalues of $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}^\iota$ and $\boldsymbol{\Gamma}^{\iota\iota'}$, respectively. We consider the asymptotic framework in [Mardia and Marshall, 1984] and denote n as the stage of the asymptotics. Let $\boldsymbol{I}_n(\boldsymbol{\theta}) = \mathrm{E}(-\ell''(\boldsymbol{\theta}, \boldsymbol{\theta}))$ denote the information matrix of $\boldsymbol{\theta}$, then the (ι, ι')th entry of the information matrix is ${}^n t_{\iota\iota'}/2$, where ${}^n t_{\iota\iota'} = \mathrm{tr}({}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_\iota\,{}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_{\iota'})$.

The following regularity conditions are assumed.

(A.1) The covariance function $\gamma(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$.

(A.2) $\limsup_{n\to\infty} \lambda_{N_n} = c < \infty$, $\limsup_{n\to\infty} \lambda^{\iota}_{N_n} = c^{\iota} < \infty$, $\limsup_{n\to\infty} \lambda^{\iota\iota'}_{N_n} = c^{\iota\iota'} < \infty$ for some positive constants c, c^{ι} and $c^{\iota\iota'}$ and all $\iota, \iota' = 1, \ldots, q$.

(A.3) For all $k, k' = 1, \ldots, q$, ${}^{n}a_{\iota\iota'} = \lim_{n\to\infty}\{{}^{n}t_{\iota\iota'}/({}^{n}t_{\iota\iota}\,{}^{n}t_{\iota'\iota'})^{1/2}\}$ exists and ${}^{n}\boldsymbol{A} = ({}^{n}a_{\iota\iota'})_{\iota,\iota'=1}^{q}$ is a nonsingular matrix.

(A.4) $\lim_{n\to\infty} N_n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}) \to \boldsymbol{J}(\boldsymbol{\theta})$, where $\boldsymbol{J}(\boldsymbol{\theta})$ is non-singular.

(A.5) $\sup_{\boldsymbol{\theta}\in\Omega} \kappa(\boldsymbol{\theta}) \equiv \kappa_0 = O_p(1)$, where Ω is an open subset of $\mathbb{R}^q$ such that $\boldsymbol{\theta}_0 \in \Omega$.

(A.6) The fixed design matrix $\boldsymbol{X}$ satisfies $\lim_{n\to\infty} N_n^{-1}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} \to \boldsymbol{C}$, where $\boldsymbol{C}$ is some positive definite matrix.

Assumptions (A.1)–(A.3) and (A.6) are standard; see, for instance, Mardia and Marshall [1984]. Assumption (A.4) ensures the convergence of information matrix. Assumption (A.5) assumes that the condition number of the covariance matrix is uniformly bounded in $\boldsymbol{\theta}$.

4.5.1 Proof of Theorem 9

Proof. Note that

$$\begin{aligned}\ell(\boldsymbol{\theta}; \boldsymbol{y}) - \widetilde{\ell}^{(l)}(\boldsymbol{\theta}; \boldsymbol{y}) &= -(1/2)\boldsymbol{y}^{\top}(\boldsymbol{z} - \boldsymbol{z}_l) = -(1/2)\boldsymbol{z}^{\top}\boldsymbol{\Gamma}(\boldsymbol{z} - \boldsymbol{z}_l) \\ &= -(1/2)(\boldsymbol{z} - \boldsymbol{z}_l)^{\top}\boldsymbol{\Gamma}(\boldsymbol{z} - \boldsymbol{z}_l) - (1/2)\boldsymbol{z}_l^{\top}\boldsymbol{\Gamma}(\boldsymbol{z} - \boldsymbol{z}_l) \\ &= -(1/2)(\boldsymbol{z} - \boldsymbol{z}_l)^{\top}\boldsymbol{\Gamma}(\boldsymbol{z} - \boldsymbol{z}_l) \equiv -(1/2)\|\boldsymbol{z} - \boldsymbol{z}_l\|^2_{\boldsymbol{\Gamma}}.\end{aligned} \tag{4.23}$$

The second term in (4.23) vanishes since $\boldsymbol{z}_l \in \mathcal{K}_l(\boldsymbol{\Gamma}, \boldsymbol{y})$ and $\boldsymbol{\Gamma}(\boldsymbol{z} - \boldsymbol{z}_l) = \boldsymbol{y} - \boldsymbol{\Gamma}\boldsymbol{z}_l = \boldsymbol{r}_l$ is orthogonal to $\mathcal{K}_l(\boldsymbol{\Gamma}, \boldsymbol{y})$. Thus, we have

$$|\ell(\boldsymbol{\theta}; \boldsymbol{y}) - \widetilde{\ell}^{(l)}(\boldsymbol{\theta}; \boldsymbol{y})| \le \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{l} \|\boldsymbol{z}\|^2_{\boldsymbol{\Gamma}} \le \left(\frac{\sqrt{\kappa_0} - 1}{\sqrt{\kappa_0} + 1}\right)^{l} \|\boldsymbol{z}\|^2_{\boldsymbol{\Gamma}} \equiv g^{(l)}(\boldsymbol{\theta}),$$

where $\|\boldsymbol{z}\|^2_{\boldsymbol{\Gamma}} = \boldsymbol{y}^{\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{y}$.

For ease of notation, we omit $\boldsymbol{y}$ in the log-likelihood functions. Let $\delta_n = N_n^{\alpha}$, where $\alpha \in [-1/2, 0]$. First, we will show that, for a given constant $\epsilon > 0$, there is a constant C such that, for a sufficiently large n,

$$P\left\{ \sup_{\|\boldsymbol{u}\|=C} \widetilde{\ell}^{(l)}(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) < \widetilde{\ell}^{(l)}(\boldsymbol{\theta}_0) \right\} \geq 1 - \epsilon,$$

where $\boldsymbol{u} \in \mathbb{R}^q$.

$$\begin{aligned}
\widetilde{\ell}^{(l)}(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) - \widetilde{\ell}^{(l)}(\boldsymbol{\theta}_0) &= \{\ell(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) - \ell(\boldsymbol{\theta}_0)\} + \left\{\ell(\boldsymbol{\theta_0}) - \widetilde{\ell}^{(l)}(\boldsymbol{\theta_0})\right\} \\
&\quad - \left\{\ell(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) - \widetilde{\ell}^{(l)}(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u})\right\} \\
&\leq \{\ell(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) - \ell(\boldsymbol{\theta}_0)\} + g^{(l)}(\boldsymbol{\theta}_0) + g^{(l)}(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u})
\end{aligned}$$

By Taylor's expansion, we obtain

$$\ell(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) - \ell(\boldsymbol{\theta}_0) = \delta_n \ell'(\boldsymbol{\theta}_0)^\top \boldsymbol{u} - (1/2) N_n \delta_n^2 \boldsymbol{u}^\top \boldsymbol{J}(\boldsymbol{\theta}_0) \boldsymbol{u}\{1 + o_p(1)\}. \tag{4.24}$$

Under (A.1)–(A.5), using Lemma 1 in Chu et al. [2011], we have $\ell'(\boldsymbol{\theta}_0) = O_p(N_n^{1/2})$. Therefore, the first term of (4.24) is of order $O_p(N_n^{1/2}\delta_n \boldsymbol{u})$. The second term of (4.24) is at the rate of $O_p(N_n \delta_n^2 \boldsymbol{u}^\top \boldsymbol{u})$. If $\alpha = -1/2$, then for a sufficiently large C, the second term dominates the first term in (4.24) for all n. If $\alpha > -1/2$, the second term dominates the first term in (4.24) for sufficiently large n.

Since $\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1} < 1$, there exists an l such that $\delta_n \geq \left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^{l/2}$. As $\|\boldsymbol{z}\|_{\boldsymbol{\Gamma}}^2 \leq \|\boldsymbol{\Gamma}\|_2 \boldsymbol{y}^\top \boldsymbol{y} = O_p(N_n)$ by (A.2), we have

$$g^{(l)}(\boldsymbol{\theta}) = \left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l \|\boldsymbol{z}\|_{\boldsymbol{\Gamma}}^2 = O_p\left(\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l N_n\right) = O_p(N_n \delta_n^2),$$

which is dominated by the second term of (4.24), for a sufficiently large C.

In the special case when $l > \frac{\sqrt{\kappa_0}+1}{2} \log N_n$, we have $\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l = O(N_n^{-1})$. Therefore,

$$g^{(l)}(\boldsymbol{\theta}) = \left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l \|\boldsymbol{z}\|_{\boldsymbol{\Gamma}}^2 = O_p(1),$$

For $\delta_n = N_n^{-1/2}$, we have $\|\widehat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$. □

4.5.2 Proof of Theorem 10

In the following Lemmas 7–12, we provide some technical results for establishing the convergence results of the algorithms, which will be used in the proofs of Theorems 10–11.

Lemma 7. *Under Assumption (A.1), we have*

$$\mathrm{tr}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right) = \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota) = \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota)$$

and

$$\mathrm{tr}\left(\frac{\partial^2 \log \boldsymbol{\Gamma}}{\partial \theta_\iota \partial \theta_{\iota'}}\right) = \mathrm{tr}(-\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota'} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota\iota'}) = \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'} - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota\iota'}),$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are the eigenvalues of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}_\iota = \partial\boldsymbol{\Lambda}/\partial\theta_\iota$ and $\boldsymbol{\Lambda}_{\iota\iota'} = \partial^2\boldsymbol{\Lambda}/\partial\theta_\iota\partial\theta_{\iota'}$.

Proof. Consider the eigendecomposition $\boldsymbol{\Gamma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, we have

$$\begin{aligned}
\mathrm{tr}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right) &= \mathrm{tr}\left(\frac{\partial \boldsymbol{Q}}{\partial \theta_\iota}\log\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \boldsymbol{Q}\frac{\partial \log\boldsymbol{\Lambda}}{\partial\theta_\iota}\boldsymbol{Q}^\top + \boldsymbol{Q}\log\boldsymbol{\Lambda}\frac{\partial\boldsymbol{Q}^\top}{\partial\theta_\iota}\right)\\
&= \mathrm{tr}\left(\left(\boldsymbol{Q}^\top\frac{\partial\boldsymbol{Q}}{\partial\theta_\iota} + \frac{\partial\boldsymbol{Q}^\top}{\partial\theta_\iota}\boldsymbol{Q}\right)\log\boldsymbol{\Lambda} + \boldsymbol{Q}\frac{\partial\log\boldsymbol{\Lambda}}{\partial\theta_\iota}\boldsymbol{Q}^\top\right)\\
&= \mathrm{tr}\left(\boldsymbol{Q}\frac{\partial\log\boldsymbol{\Lambda}}{\partial\theta_\iota}\boldsymbol{Q}^\top\right) = \mathrm{tr}\left(\frac{\partial\log\boldsymbol{\Lambda}}{\partial\theta_\iota}\right) = \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota),
\end{aligned}$$

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota) &= \mathrm{tr}\left(\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^\top\frac{\partial\boldsymbol{Q}}{\partial\theta_\iota}\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^\top\boldsymbol{Q}\frac{\partial\boldsymbol{\Lambda}}{\partial\theta_\iota}\boldsymbol{Q}^\top + \boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{\Lambda}\frac{\partial\boldsymbol{Q}^\top}{\partial\theta_\iota}\right)\\
&= \mathrm{tr}\left(\left(\boldsymbol{Q}^\top\frac{\partial\boldsymbol{Q}}{\partial\theta_\iota} + \frac{\partial\boldsymbol{Q}^\top}{\partial\theta_\iota}\boldsymbol{Q}\right) + \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota\right) = \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota),
\end{aligned}$$

$$\begin{aligned}
\mathrm{tr}\left(\frac{\partial^2 \log \boldsymbol{\Gamma}}{\partial \theta_\iota \partial \theta_{\iota'}}\right) &= \mathrm{tr}\left(\frac{\partial}{\partial\theta_{\iota'}}\left(\frac{\partial \boldsymbol{Q}}{\partial \theta_\iota}\log\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \boldsymbol{Q}\frac{\partial \log\boldsymbol{\Lambda}}{\partial\theta_\iota}\boldsymbol{Q}^\top + \boldsymbol{Q}\log\boldsymbol{\Lambda}\frac{\partial\boldsymbol{Q}^\top}{\partial\theta_\iota}\right)\right)\\
&= \mathrm{tr}\left(\frac{\partial^2\boldsymbol{Q}}{\partial\theta_\iota\partial\theta_{\iota'}}\log\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \frac{\partial\boldsymbol{Q}}{\partial\theta_\iota}\frac{\partial\log\boldsymbol{\Lambda}}{\partial\theta_{\iota'}}\boldsymbol{Q}^\top + \frac{\partial\boldsymbol{Q}}{\partial\theta_\iota}\log\boldsymbol{\Lambda}\frac{\partial\boldsymbol{Q}^\top}{\partial\theta_{\iota'}}\right) +
\end{aligned}$$

$$
\begin{aligned}
& \operatorname{tr}\left(\frac{\partial \boldsymbol{Q}}{\partial \theta_{\iota'}} \frac{\partial \log \boldsymbol{\Lambda}}{\partial \theta_{\iota}} \boldsymbol{Q}^{\top}+\boldsymbol{Q} \frac{\partial^{2} \log \boldsymbol{\Lambda}}{\partial \theta_{\iota} \partial \theta_{\iota'}} \boldsymbol{Q}^{\top}+\boldsymbol{Q} \frac{\partial \log \boldsymbol{\Lambda}}{\partial \theta_{\iota}} \frac{\partial \boldsymbol{Q}^{\top}}{\partial \theta_{\iota'}}\right)+ \\
& \operatorname{tr}\left(\frac{\partial \boldsymbol{Q}}{\partial \theta_{\iota'}} \log \boldsymbol{\Lambda} \frac{\partial \boldsymbol{Q}^{\top}}{\partial \theta_{\iota}}+\boldsymbol{Q} \frac{\partial \log \boldsymbol{\Lambda}}{\partial \theta_{\iota'}} \frac{\partial \boldsymbol{Q}^{\top}}{\partial \theta_{\iota}}+\boldsymbol{Q} \log \boldsymbol{\Lambda} \frac{\partial^{2} \boldsymbol{Q}^{\top}}{\partial \theta_{\iota} \partial \theta_{\iota'}}\right) \\
= & \operatorname{tr}\left(-\boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota'}+\boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota \iota'}\right).
\end{aligned}
$$

Since $\log\det\boldsymbol{\Gamma} = \log\det\boldsymbol{\Lambda}$, so does their second derivatives, that is,

$$
\operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'} - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota\iota'}) = \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota'} - \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota\iota'}).
$$

□

For our notation convenience, we define the following intermediate approximation to the log-likelihood function

$$
\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}) = -(N/2)\log(2\pi) - \frac{1}{2N_v}\sum_{i=1}^{N_v} \boldsymbol{\chi}_i^{\top}\log(\boldsymbol{\Gamma})\boldsymbol{\chi}_i - (1/2)\boldsymbol{y}^{\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{y}, \tag{4.25}
$$

where $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_{N_v}$ are independent vectors whose elements are *i.i.d.* Rademacher random variables.

Lemma 8. *Under Assumption (A.1), we have*

$$
\widetilde{\boldsymbol{I}}_n^{(N_v)}(\boldsymbol{\theta}) = \boldsymbol{I}_n(\boldsymbol{\theta}), \ \textit{for} \ N_v = 1, 2, \ldots, \tag{4.26}
$$

where $\widetilde{\boldsymbol{I}}_n^{(N_v)}(\boldsymbol{\theta}) = E\{-(\widetilde{\ell}^{(N_v)})''(\boldsymbol{\theta}, \boldsymbol{\theta})\}$.

Proof.

$$
\begin{aligned}
& \mathrm{E}\left(-\frac{\partial^{2} \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta})}{\partial \theta_{\iota} \partial \theta_{\iota'}}\right)=\frac{1}{2N_v} \sum_{i=1}^{N_v} \operatorname{tr}\left(\frac{\partial^{2} \log \boldsymbol{\Gamma}}{\partial \theta_{\iota} \partial \theta_{\iota'}}\right)+\frac{1}{2} \operatorname{tr}(\boldsymbol{\Gamma}^{\iota \iota'} \boldsymbol{\Gamma}) \\
& =\frac{1}{2} \operatorname{tr}(\boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota'}-\boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda}_{\iota \iota'})-\frac{1}{2} \operatorname{tr}(\boldsymbol{\Gamma}^{\iota \iota'} \boldsymbol{\Gamma})=\frac{1}{2} \operatorname{tr}(\boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{\iota} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{\iota'})
\end{aligned}
$$

where $\boldsymbol{\Gamma}^{\iota\iota'} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_{\iota}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'} + \boldsymbol{\Gamma}_{\iota'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota} - \boldsymbol{\Gamma}_{\iota\iota'})\boldsymbol{\Gamma}^{-1}$. □

Lemma 9. *Under Assumption (A.1), we have*

$$\mathrm{Var}(\widetilde{\ell}^{(N_v)'}(\boldsymbol{\theta})) \preceq \left(1 + \frac{1}{N_v}\right) \boldsymbol{I}_n(\boldsymbol{\theta}). \tag{4.27}$$

Here, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

Proof. First, we have

$$\mathrm{Var}(\widetilde{\ell}^{(N_v)'}(\boldsymbol{\theta})) = \boldsymbol{I}_n(\boldsymbol{\theta}) + \frac{1}{4N_v}\boldsymbol{D}_n(\boldsymbol{\theta}),$$

where the (ι, ι')th element of $\boldsymbol{D}_n(\boldsymbol{\theta})$ is

$$\begin{aligned}
\boldsymbol{D}_{n,\iota\iota'}(\boldsymbol{\theta}) &= \mathrm{Cov}\left(\boldsymbol{\chi}_1^\top \frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\boldsymbol{\chi}_1, \boldsymbol{\chi}_1^\top \frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\boldsymbol{\chi}_1\right) \\
&= \sum_{i=1}^{N_n}\sum_{j=1}^{N_n}\sum_{l=1}^{N_n}\sum_{m=1}^{N_n} \mathrm{Cov}\left(\chi_{1,i}\chi_{1,j}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)_{i,j}, \chi_{1,l}\chi_{1,m}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right)_{l,m}\right) \\
&= \sum_{i\neq j} \mathrm{Cov}\left(\chi_{1,i}\chi_{1,j}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)_{i,j}, \chi_{1,i}\chi_{1,j}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right)_{i,j}\right) + \\
&\quad \sum_{i\neq j} \mathrm{Cov}\left(\chi_{1,i}\chi_{1,j}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)_{i,j}, \chi_{1,i}\chi_{1,j}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right)_{j,i}\right) \\
&= \mathrm{tr}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right) + \mathrm{tr}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}^\top\right) - 2\sum_{i=1}^{N_n}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)_{i,i}\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right)_{i,i},
\end{aligned}$$

where $\chi_{1,i}$ is the ith element of $\boldsymbol{\chi}_i$ and $\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)_{i,j}$ is the (i,j)th element of $\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)$. Further, we note that

$$\begin{aligned}
&\mathrm{tr}(\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_{\iota'}}\right)) \\
=\ &\mathrm{tr}\left\{\left(\boldsymbol{Q}_\iota \log \boldsymbol{\Lambda}\boldsymbol{Q}^\top + \boldsymbol{Q}\frac{\partial \log \boldsymbol{\Lambda}}{\partial \theta_\iota}\boldsymbol{Q}^\top + \boldsymbol{Q}\log\boldsymbol{\Lambda}\boldsymbol{Q}_\iota^\top\right)\cdot\right. \\
&\left.\left(\boldsymbol{Q}_{\iota'} \log \boldsymbol{\Lambda}\boldsymbol{Q}^\top + \boldsymbol{Q}\frac{\partial \log \boldsymbol{\Lambda}}{\partial \theta_{\iota'}}\boldsymbol{Q}^\top + \boldsymbol{Q}\log\boldsymbol{\Lambda}\boldsymbol{Q}_{\iota'}^\top\right)\right\} \\
=\ &\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota'} + \boldsymbol{Q}_\iota\log\boldsymbol{\Lambda}\boldsymbol{Q}^\top\boldsymbol{Q}_{\iota'}\log\boldsymbol{\Lambda}\boldsymbol{Q}^\top + \log\boldsymbol{\Lambda}\log\boldsymbol{\Lambda}\boldsymbol{Q}_{\iota'}^\top\boldsymbol{Q}_\iota \\
&+ \log\boldsymbol{\Lambda}\log\boldsymbol{\Lambda}\boldsymbol{Q}_\iota^\top\boldsymbol{Q}_{\iota'} + \boldsymbol{Q}\log\boldsymbol{\Lambda}\boldsymbol{Q}_\iota^\top\boldsymbol{Q}\log\boldsymbol{\Lambda}\boldsymbol{Q}_{\iota'}^\top)
\end{aligned}$$

$$= \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota'}) - \sum_{m=1}^{N_n}\sum_{l=1}^{N_n}(\log(\lambda_m)+\log(\lambda_l))^2(\boldsymbol{Q}_\iota^\top\boldsymbol{Q})_{(m,l)}(\boldsymbol{Q}_{\iota'}^\top\boldsymbol{Q})_{(m,l)}.$$

Since

$$\operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'}) = \operatorname{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_\iota\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\iota'}) + \sum_{m=1}^{N_n}\sum_{l=1}^{N_n}\left(\sqrt{\frac{\lambda_m}{\lambda_l}}-\sqrt{\frac{\lambda_l}{\lambda_m}}\right)^2(\boldsymbol{Q}_\iota^\top\boldsymbol{Q})_{(m,l)}(\boldsymbol{Q}_{\iota'}^\top\boldsymbol{Q})_{(m,l)},$$

we have

$$\operatorname{tr}\left(\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_\iota}\right)\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_{\iota'}}\right)\right) = \operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'}) - \sum_{m=1}^{N_n}\sum_{l=1}^{N_n}a_{m,l}(\boldsymbol{Q}_\iota^\top\boldsymbol{Q})_{(m,l)}(\boldsymbol{Q}_{\iota'}^\top\boldsymbol{Q})_{(m,l)}$$

where $a_{m,l} = \left(\sqrt{\frac{\lambda_m}{\lambda_l}}-\sqrt{\frac{\lambda_l}{\lambda_m}}\right)^2 + (\log(\lambda_m)+\log(\lambda_l))^2 \geq 0$ and $a_{m,l} = a_{l,m}$.

For any real numbers $u_1,\ldots,u_q$,

$$\sum_{k=1}^{q}\sum_{k'=1}^{q}\sum_{m=1}^{N_n}\sum_{l=1}^{N_n}a_{m,l}u_ku_{k'}(\boldsymbol{Q}_k^\top\boldsymbol{Q})_{(m,l)}(\boldsymbol{Q}_{k'}^\top\boldsymbol{Q})_{(m,l)} = \sum_{m=1}^{N_n}\sum_{l=1}^{N_n}a_{m,l}\left\{\sum_{k=1}^{q}u_k(\boldsymbol{Q}_k^\top\boldsymbol{Q})_{(m,l)}\right\}^2 \geq 0$$

$$\sum_{k=1}^{q}\sum_{k'=1}^{q}u_ku_{k'}\sum_{i=1}^{N_n}\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_\iota}\right)_{i,i}\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_{\iota'}}\right)_{i,i} = \sum_{i=1}^{N_n}\left\{\sum_{k=1}^{q}u_k\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_\iota}\right)_{i,i}\right\}^2 \geq 0$$

□

Lemma 10. *Under Assumption (A.1) and (A.4), we have*

$$N_n^{-1/2}\widetilde{\ell}^{(N_v)'}(\boldsymbol{\theta}) = O_p(1) \tag{4.28}$$

Proof. By Lemma 7, we have

$$\mathrm{E}\left(\frac{\partial\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta})}{\partial\theta_\iota}\right) = -\frac{1}{2}\operatorname{tr}\left(\frac{\partial\log\boldsymbol{\Gamma}}{\partial\theta_\iota}\right) + \frac{1}{2}\operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota) = 0$$

Further, by (A.4), $N_n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}) \to \boldsymbol{J}(\theta)$. By Lemma 9, $\operatorname{Var}(\widetilde{\ell}^{(N_v)'}(\boldsymbol{\theta})) \preceq \left(1+\frac{1}{N_v}\right)\boldsymbol{I}_n(\boldsymbol{\theta})$, so $\operatorname{Var}(\widetilde{\ell}^{(N_v)'}(\boldsymbol{\theta})) = O(N_n)$, □

Let $\widetilde{\boldsymbol{\theta}}^{(N_v)} = \arg\max_{\boldsymbol{\theta}}\{\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta})\}$ denote the maximizer of the approximate likelihood function (4.25). The following lemma establishes the consistency of the estimator $\widetilde{\boldsymbol{\theta}}^{(N_v)}$.

Lemma 11. *Under Assumptions (A.1)-(A.5), there exists, with probability tending to one, a local maximizer $\widehat{\boldsymbol{\theta}}^{(N_v)}$ of $\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}; \boldsymbol{y})$, such that $\|\widehat{\boldsymbol{\theta}}^{(N_v)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$.*

Proof. Let $\delta_n = N_n^{\alpha}$, where $\alpha \in [-1/2, 0]$. To establish $\|\widetilde{\boldsymbol{\theta}}^{(N_v)} - \boldsymbol{\theta}_0\| = O_p(N_n^{\alpha})$, 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

$$P\left\{\sup_{\|\boldsymbol{u}\|=C} \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{u}) < \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0)\right\} \geq 1 - \epsilon,$$

where $\boldsymbol{u} \in \mathbb{R}^q$.

Write $h(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Gamma}| + \frac{1}{2N_v}\sum_{i=1}^{N_v} \boldsymbol{\chi}_i^{\top} \log(\boldsymbol{\Gamma})\boldsymbol{\chi}_i$, then

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_\iota} = -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota) + \frac{1}{2N_v}\sum_{i=1}^{N_v} \boldsymbol{\chi}_i^{\top} \frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota} \boldsymbol{\chi}_i.$$

By Lemma 7, we have $\mathrm{E}\left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_\iota}\right) = 0$. In addition,

$$\begin{aligned}
&\operatorname{Var}\left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_\iota}\right) \leq \frac{1}{2N_v}\operatorname{tr}\left(\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)\left(\frac{\partial \log \boldsymbol{\Gamma}}{\partial \theta_\iota}\right)\right) \\
&= \frac{1}{2N_v}\left(\operatorname{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\iota\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\iota'}) - \sum_{m=1}^{N_n}\sum_{l=1}^{N_n} a_{m,l}(\boldsymbol{Q}_k^{\top}\boldsymbol{Q})_{(m,l)}(\boldsymbol{Q}_k^{\top}\boldsymbol{Q})_{(m,l)}\right) = O(N_n N_v^{-1})
\end{aligned}$$

where $a_{m,l} \geq 0$ is defined in Lemma 9. Hence $h'(\boldsymbol{\theta}) = O_p(N_n^{1/2}N_v^{-1/2})$.

By Taylor's expansion, we obtain

$$\begin{aligned}
\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{u}) - \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0) &= \ell(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{u}) - \ell(\boldsymbol{\theta}_0) + h(\boldsymbol{\theta}_0) - h(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{u}) \\
&= \delta_n\ell'(\boldsymbol{\theta}_0)^{\top}\boldsymbol{u} - \tfrac{1}{2}N_n\delta_n^2\boldsymbol{u}^{\top}\boldsymbol{J}(\boldsymbol{\theta}_0)\boldsymbol{u}\{1 + o_p(1)\} - \delta_n h'(\boldsymbol{\theta}^*)^{\top}\boldsymbol{u}, \qquad (4.29)
\end{aligned}$$

where $\boldsymbol{I}_n(\boldsymbol{\theta}) = \mathrm{E}\{-\ell''(\boldsymbol{\theta}, \boldsymbol{\theta})\}$, $N_n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}) \to \boldsymbol{J}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \delta_n\boldsymbol{u}$. Since $\ell'(\boldsymbol{\theta}_0) = O_p(N_n^{1/2})$ and $h'(\boldsymbol{\theta}^*) = O_p(N_n^{1/2}N_v^{-1/2})$, if we further assume $\delta_n = N_n^{-1/2}$, the first and third terms of (4.29) are of order $O_p(\boldsymbol{u})$. The second term of (4.29) is at the rate of $O_p(\boldsymbol{u}^{\top}\boldsymbol{u})$. Thus, for a sufficiently large C, the second term dominates other terms in (4.29). □

Using Algorithm 3, we have the following approximation of the log-likelihood function

$$\widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta};\boldsymbol{y}) = -(N/2)\log(2\pi) - (1/2)\Xi_{m,N_v} - (1/2)\boldsymbol{y}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{y}, \tag{4.30}$$

where Ξ_{m,N_v} is the approximate log-determinant of the covariance matrix.

Lemma 12. *Under Assumptions (A.1)-(A.5), there exists, with probability tending to one, a local maximizer $\widehat{\boldsymbol{\theta}}^{(m,N_v)}$ of $\widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta};\boldsymbol{y})$, such that $\|\widehat{\boldsymbol{\theta}}^{(m,N_v)} - \boldsymbol{\theta}_0\| = O_p\left(\max\left\{\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^m, N_n^{-1/2}\right\}\right)$. In particular, if $m > \frac{\sqrt{\kappa_0}}{4}\log(N_nC_1)$, where $C_1 = \lambda_{\max}\sqrt{\kappa_0}\log(\lambda_{\max}+\lambda_{min})$, we have $\|\widehat{\boldsymbol{\theta}}^{(m,N_v)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$.*

Proof. Continue to define δ_n as in Lemma 11. By Lemma 4.4 in Ubaru et al. [2017], we have $|\frac{1}{N_v}\sum_{i=1}^{N_v}\boldsymbol{\chi}_i^\top\log(\boldsymbol{\Gamma})\boldsymbol{\chi}_i - \Xi_{m,N_v}| \leq \frac{N_nC}{\rho^{2m}}$, where $\rho = \frac{\sqrt{\kappa_0}+1}{\sqrt{\kappa_0}-1}$, $C = \frac{(\lambda_{\max}-\lambda_{\min})(\sqrt{\kappa_0}-1)^2\log(\lambda_{\max}+\lambda_{\min})}{2\sqrt{\kappa_0}}$. By Assumption (A.5), there exists an α such that $\delta_n > \frac{\sqrt{C}}{\rho^m}$. Therefore,

$$\begin{aligned}
& \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0) \\
= \;& \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0) + (\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta_0}) - \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta_0})) \\
& -(\widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u})) \\
\leq \;& \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(N_v)}(\boldsymbol{\theta}_0) + O_p(N_n\delta_n^2) \\
\leq \;& \delta_n\ell'(\boldsymbol{\theta}_0)^\top\boldsymbol{u} - \frac{1}{2}N_n\delta_n^2\boldsymbol{u}^\top\boldsymbol{J}(\boldsymbol{\theta}_0)\boldsymbol{u}\{1+o_p(1)\} - O_p(N_n^{1/2}N_v^{-1/2}\delta_n\boldsymbol{u}) + O_p(N_n\delta_n^2)
\end{aligned}$$

For a sufficiently large C, the second term dominates the other terms. In particular, if $m > \frac{\sqrt{\kappa_0}}{4}\log(N_nC_1)$, $|\frac{1}{N_v}\sum_{i=1}^{N_v}\boldsymbol{\chi}_i^\top f(A)\boldsymbol{\chi}_i - \Xi_{m,N_v}| \leq \frac{N_n\lambda_{\max}\sqrt{\kappa_0}\log(\lambda_{\max}+\lambda_{\min})}{2\rho^{2m}} \leq 1$. Let $\delta_n = N_n^{-1/2}$, we have $\|\widehat{\boldsymbol{\theta}}^{(m,N_v)} - \boldsymbol{\theta}_0\| = O_p(N_n^{-1/2})$.

□

Finally, we will prove Theorem 10.

Proof of Theorem 10.

$$\begin{aligned}
& \widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0) \\
= \;& \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) - \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0) + (\widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta_0}) - \widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta_0}))
\end{aligned}$$

$$
\begin{aligned}
& -(\widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u})-\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u})) \\
\leq\ & \widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u})-\widetilde{\ell}^{(m,N_v)}(\boldsymbol{\theta}_0)+g^{(l)}(\boldsymbol{\theta}_0)+g^{(l)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u}) \\
\leq\ & \delta_n\ell'(\boldsymbol{\theta}_0)^\top\boldsymbol{u}-\frac{1}{2}N_n\delta_n^2\boldsymbol{u}^\top\boldsymbol{J}(\boldsymbol{\theta}_0)\boldsymbol{u}\{1+o_p(1)\}-O_p(N_n^{1/2}N_v^{-1/2}\delta_n\boldsymbol{u})+O_p(N_n\delta_n^2)
\end{aligned}
$$

For a sufficiently large C, the second term dominates the other terms, hence we complete the proof. □

4.5.3 Proof of Theorem 11

Remark. We will first show that under (A.2), (A.5) and (A.6), there exists, with probability tending to one, a local minimizer $\widehat{\boldsymbol{\beta}}$ of the residual sum of squares $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(N_n^{-1/2})$ [Wang and Zhu, 2009]. In fact, for the least squares estimator $\widehat{\boldsymbol{\beta}}_{\text{OLS}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, we have $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})=\boldsymbol{\beta}$ and

$$
\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\leq\|\boldsymbol{\Gamma}^{-1}\|_2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.
$$

By Assumption (A.2), (A.5) and (A.6), we have the desired results.

Next, we will prove Theorem 11.

Proof of Theorem 11. For any given $\widetilde{\boldsymbol{\beta}}$ satisfying $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(N_n^{-1/2})$, we minimize the criterion $\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})$, where $\widetilde{\boldsymbol{y}}=\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}$. Let $R_1(\boldsymbol{\theta})=\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})-\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta};\boldsymbol{y}_0)$, where $\boldsymbol{y}_0=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0$, then

$$
\begin{aligned}
\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u};\widetilde{\boldsymbol{y}})-\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0;\widetilde{\boldsymbol{y}}) &= \widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u};\boldsymbol{y}_0)-\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta}_0;\boldsymbol{y}_0) \\
&+ R_1(\boldsymbol{\theta}_0+\delta_n\boldsymbol{u})-R_1(\boldsymbol{\theta}_0)
\end{aligned}
$$

Notice that

$$
\begin{aligned}
R_1(\boldsymbol{\theta}) &= \widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})-\widetilde{\ell}^{(l,m,N_v)}(\boldsymbol{\theta};\boldsymbol{y}_0)=\widetilde{\ell}^{(l)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})-\widetilde{\ell}^{(l)}(\boldsymbol{\theta};\boldsymbol{y}_0) \\
&= (\widetilde{\ell}^{(l)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})-\ell(\boldsymbol{\theta};\widetilde{\boldsymbol{y}}))+(\ell(\boldsymbol{\theta};\boldsymbol{y}_0)-\widetilde{\ell}^{(l)}(\boldsymbol{\theta};\boldsymbol{y}_0))+(\ell(\boldsymbol{\theta};\boldsymbol{y}_0)-\ell(\boldsymbol{\theta};\widetilde{\boldsymbol{y}})) \\
&= (I_1)+(I_2)+(I_3)
\end{aligned}
$$

For (I_1), let $\widetilde{\boldsymbol{z}} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{z}}_l$ be the solution from conjugate gradient algorithm at the lth step.

$$\widetilde{\ell}^{(l)}(\boldsymbol{\theta};\widetilde{\boldsymbol{y}}) - \ell(\boldsymbol{\theta};\widetilde{\boldsymbol{y}}) = -\frac{1}{2}\|\widetilde{\boldsymbol{z}} - \widetilde{\boldsymbol{z}}_l\|^2_{\boldsymbol{\Gamma}} \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^l \|\widetilde{\boldsymbol{z}}\|^2_{\boldsymbol{\Gamma}} \leq \left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l \|\widetilde{\boldsymbol{z}}\|^2_{\boldsymbol{\Gamma}}.$$

Note that,

$$\begin{aligned}
\|\widetilde{\boldsymbol{z}}\|^2_{\boldsymbol{\Gamma}} &= (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0) + 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \\
&\quad -(\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \\
&= O_p(N_n) + O_p(1) + O(1) = O_p(N_n)
\end{aligned}$$

From proof of Theorem 9, both (I_1) and (I_2) are of order $O_p\left(\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l N_n\right)$. For (I_3), we know that

$$\begin{aligned}
\ell(\boldsymbol{\theta};\boldsymbol{y}_0) - \ell(\boldsymbol{\theta};\widetilde{\boldsymbol{y}}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \\
&= O_p(1).
\end{aligned}$$

So $R_1(\boldsymbol{\theta})$ is of order $O_p\left(\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^l N_n\right)$. By proof of Theorem 10, we have the desired results.

□

Chapter 5

Summary and Future Work

In this dissertation, we have studied both theoretical and computational aspects of spatial or spatio-temporal modeling. In Chapter 2, we propose a novel spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain for studying the properties of statistical inference for spatio-temporal models. This framework generalizes the existing asymptotic frameworks for spatial processes to spatio-temporal processes. For characterizing the spatio-temporal dependence, we propose a class of locally stationary spatio-temporal covariance functions, whose variance varies with space and time on a rescaled fixed spatio-temporal domain. This spatio-temporal rescaling approach, being a purely theoretical device, corresponds to the rescaled-time principle in time series that have the regression function depend on rescaled time [Vogt, 2012]. The resulting spatio-temporal covariance functions are locally stationary in the sense that they can be approximated locally by stationary covariance functions of actual distances in the spatio-temporal domain without rescaling. The covariance function is studied under the proposed STED asymptotic framework and the asymptotic properties of the maximum likelihood estimation is established.

In Chapter 3, we develop semiparametric method for spatio-temporal processes with continuous spatial index and *continuous* temporal index. A partially linear mean function is considered, and the covariance structure of the error process is assumed to be locally stationary. We study a profile likelihood method for simultaneous estimation of the mean and covariance functions. For estimating the nonparametric trend in the partial linear model, we propose to use biomodal kernels in cross-validation. This helps mitigate bias in bandwidth selection due to dependent error terms and can be of independent interest for semiparametric spatial statistics. The asymptotic properties of our methods, including consistency and asymptotic normality, are established under a local stationarity condition. The finite sample properties based on a simulation study further support the theory.

The previous chapters rely largely on the evaluation of the likelihood function, which involve matrix decompositions whose complexity increases as $O(N^3)$ in the sample size N. This overwhelms traditional implementations of spatial or spatio-temporal statistics, even for moderately large datasets. This "big N problem" for spatial data modeling is addressed in Chapter 4. We have provided an approximate Gaussian log-likelihood function using Krylov subspace methods to reduce computational complexity. Our proposed method reduces the computation burden from $O(N^3)$ to $O(N^2)$ for dense matrices and further to quasi-linear for sparse matrices. We have established that the parameter estimates based on the approximate likelihood is consistent under some regularity conditions. Our simulation study shows that the performance of our proposed method is comparable to the exact MLE, while the computational cost is greatly reduced.

Our methodology and theory in Chapter 2 and Chapter 3 are based on a fixed sampling design where the spatial sampling locations and the temporal sampling points are assumed to be fixed. It is possible to explore an alternative stochastic sampling design in space and/or time. In addition, it is of interest to explore alternative approaches to modeling and estimating semiparametric or fully nonparametric locally stationary covariance functions. For example, nonparametric approach via the Karhunen-Loève expansion can be applied to model the error process. Our algorithms in Chapter 4 are especially useful when the covariance matrices are sparse, where covariance tapering approaches come into play. However, the tapering procedure could result in loss of statistical efficiency. It might be useful to combine tapering method with other approximation methods including low rank approach and block-based method. For example, Romary and Desassis [2018] proposed a model to combine the covariance tapering approach with the low rank approach, where our method can be easily applied. Furthermore, our approach is a numerical approximation of the likelihood function, which can be easily extended to spatio-temporal models. In addition, our method can be further generalized to nonstationary models. We leave these topics for future research.

Bibliography

N. S. Altman. Kernel smoothing of data with correlated errors. *Journal of the American Statistican Association*, 85:749–759, 1990.

Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70: 825–848, 2008.

Moreno Bevilacqua and Carlo Gaetan. Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Statistics and Computing*, 25:877–892, 2015.

Christos Boutsidis, Petros Drineas, Prabhanjan Kambadur, Eugenia-Maria Kontopoulou, and Anastasios Zouzias. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. *Linear Algebra and its Applications*, 533:95–117, 2017.

Jonathan R Bradley, Noel Cressie, and Tao Shi. A comparison of spatial predictors when datasets could be very large. *Statistics Surveys*, 10:100–131, 2016.

Jinyuan Chang, Bin Guo, and Qiwei Yao. High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, 189:297–312, 2015.

Tingjin Chu, Jun Zhu, and Haonan Wang. Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, 39:2607–2625, 2011.

Tingjin Chu, Jialuo Liu, Jun Zhu, and Haonan Wang. Spatio-temporal expanding distance asymptotic framework for locally stationary processes. Manuscript, 2019.

Bruce Cook, Ross Nelson, Elizabeth Middleton, Douglas Morton, Joel McCorkel, Jeffrey Masek, Kenneth Ranson, Vuong Ly, Paul Montesano, et al. Nasa goddard's lidar, hyperspectral and thermal (g-liht) airborne imager. *Remote Sensing*, 5:4045–4066, 2013.

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

N. Cressie and H.-C. Huang. Classes of nonseparable spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94:1330–1340, 1999.

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

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

N. Cressie and C. K. Wikle. *Statistics for Spatio-temporal Data*. Wiley, New York, 2011.

N. Cressie, T. Shi, and E.L. Kang. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19:724–745, 2010.

Noel Cressie. Mission co_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide (with discussion). *Journal of the American Statistical Association*, 113:152–168, 2018.

Rainer Dahlhaus. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25:1–37, 1997.

Rainer Dahlhaus. Locally stationary processes. In *Handbook of Statistics*, volume 30, pages 351–413. Elsevier, Amsterdam, 2012.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016a.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, Nicholas AS Hamm, and Martijn Schaap. Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The annals of applied statistics*, 10:1286, 2016b.

Kris De Brabanter, Jos De Brabanter, Johan A K Suykens, and Bart De Moor. Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research*, 12:1955–1976, 2011.

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

Jo Eidsvik, Benjamin A Shaby, Brian J Reich, Matthew Wheeler, and Jarad Niemi. Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23:295–315, 2014.

Robert F Engle, Clive W J Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81:310–320, 1986.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapmann & Hall, 1996.

Jianqing Fan and Tao Huang. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11:1031–1057, 2005.

Jianqing Fan and Qiwei Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York, 2003.

Andrew O Finley, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53:2873–2884, 2009.

Andrew O Finley, Abhirup Datta, Bruce D Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, pages 1–14, 2019.

Mario Francisco-Fernandez and Jean D Opsomer. Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics*, 33: 279–295, 2005.

M. Fuentes, L. Chen, and J. M. Davis. A class of nonseparable and nonstationary spatial temporal covariance functions. *Environmetrics*, 19:487–507, 2008.

Montserrat Fuentes. Spectral methods for nonstationary spatial processes. *Biometrika*, 89:197–210, 2002.

R Furrer and F Gerber. spam: Sparse matrix. *R package version 0.14-1*, 2008.

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

Jiti Gao and Hua Liang. Statistical inference in single-index and partially nonlinear models. *Annals of the Institute of Statistical Mathematics*, 49:493–517, 1997.

Jiti Gao, Zudi Lu, and Dag Tjostheim. Estimation in semiparametric spatial regression. *The Annals of Statistics*, 34:1395–1435, 2006.

Alan E Gelfand and Erin M Schliep. Spatial statistics and gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104, 2016.

Alan E Gelfand, Peter J Diggle, Montserrat Fuentes, and Peter Guttorp. *Handbook of Spatial Statistics*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97:590–600, 2002a.

Tilmann Gneiting. Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2):493–508, 2002b.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Gene H Golub and John H Welsch. Calculation of gauss quadrature rules. *Mathematics of computation*, 23:221–230, 1969.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

Joseph Guinness and Montserrat Fuentes. Likelihood approximations for big nonstationary spatial temporal lattice data. *Statistica Sinica*, 25:329–349, 2015.

Peter Hall and Prakash Patil. Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, 99:399–424, 1994.

Insu Han, Dmitry Malioutov, and Jinwoo Shin. Large-scale log-determinant computation through stochastic chebyshev expansions. In *International Conference on Machine Learning*, pages 908–917, 2015.

Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, SAA Turubanova, Alexandra Tyukavina, David Thau, SV Stehman, SJ Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *Science*, 342:850–853, 2013.

W. Härdle, H. Liang, and J. T. Gao. *Partially Linear Models*. Springer, 2000.

Wolfgang Härdle, Enno Mammen, and Marlene Müller. Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93: 1461–1474, 1998.

David Higdon. A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5:173–190, 1998.

Tailen Hsing, Thomas Brown, and Brian Thelen. Local intrinsic stationarity and its inference. *The Annals of Statistics*, 44:2058–2088, 2016.

Da Huang, Qiwei Yao, and Rongmao Zhang. Krigings over space and time based on latent low-dimensional structures. *arXiv preprint arXiv:1609.06789v2*, 2018.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19:433–450, 1990.

Matthias Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017.

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

Mikael Kuusela and Michael L Stein. Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A*, 474:20180400, 2018.

S.N. Lahiri. *Resampling Methods for Dependent Data*. Springer, 2003a.

S.N. Lahiri. Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. *Sankhya, Series A*, 65:356–388, 2003b.

Kirk Lake, Jun Zhu, Haonan Wang, John Volckens, and Kirsten A Koehler. Effects of data sparsity and spatiotemporal variability on hazard maps of workplace noise. *Journal of occupational and environmental hygiene*, 12:256–265, 2015.

Hua Liang and Runze Li. Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104:234–248, 2009.

Hua Liang, Wolfgang Härdle, and Raymond J Carroll. Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics*, 27:1519–1535, 1999.

Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's PM2.5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471: 20150257, 2015.

Wei-Liem Loh. Fixed-domain asymptotics for a subclass of Matérn-type gaussian random fields. *The Annals of Statistics*, 33:2344–2394, 2005.

Z. Lu, D. J. Steinskog, D. Tjøstheim, and Q. Yao. Adaptively varying coefficient spatiotemporal models. *Journal of the Royal Statistical Society, Series B*, 71:859–880, 2009.

Zudi Lu and Dag Tjøstheim. Nonparametric estimation of probability density functions for irregularly observed spatial data. *Journal of the American Statistical Association*, 109:1546–1564, 2014.

Guilherme Ludwig, Tingjin Chu, Jun Zhu, Haonan Wang, and Kirsten Koehler. Static and roving sensor data fusion for spatio-temporal hazard mapping with application to occupational exposure assessment. *The Annals of Applied Statistics*, 11:139–160, 2017.

Kanti V Mardia and Roger J Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146, 1984.

J. Opsomer, Y. Wang, and Y. Yang. Nonparametric regression with correlated errors. *Statistical Science*, 16:134–153, 2001.

Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17:483–506, 2006.

Emilio Porcu, Alfredo Alegria, and Reinhard Furrer. Modeling temporally evolving and spatially globally dependent data. *International Statistical Review*, 86:344–377, 2018.

S-W Ra and J-K Kim. A fast mean-distance-ordered partial codebook search algorithm for image vector quantization. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 40:576–579, 1993.

Thomas Romary and Nicolas Desassis. Combining covariance tapering and lasso driven low rank decomposition for the kriging of large spatial datasets. *arXiv preprint arXiv:1806.01558*, 2018.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003.

Yousef Saad. *Iterative methods for sparse linear systems*, volume 82. siam, 2003.

Paul D Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

Michael Sherman. *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley, West Sussex, UK, 2011.

Jonathan Richard Shewchuk et al. An introduction to the conjugate gradient method without the agonizing pain, 1994.

Paul Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 50:413–436, 1988.

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

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

M.L. Stein, Z. Chi, and Welty L.J. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 66:275–296, 2004.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

Jonathan R Stroud, Peter Müller, and Bruno Sansó. Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B*, 63:673–689, 2001.

Liangjun Su and Sainan Jin. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics*, 157:18–33, 2010.

Y. Sun, B. Li, and M.G. Genton. Geostatistics for large datasets. *In advances and challenges in space-time modelling of natural events*, pages 55–77, 2012.

Yan Sun, Hongjia Yan, Wenyang Zhang, and Zudi Lu. A semiparametric spatial dynamic model. *Annals of Statistics*, 42:700–727, 2014.

TJ Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, pages 1375–1381, 1980.

Waldo R Tobler. Cellular geography. In *Philosophy in geography*, pages 379–386. Springer, 1979.

Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $tr(f(A))$ via stochastic lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38:1075–1099, 2017.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B*, 50:297–312, 1988.

Michael Vogt. Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40:2601–2633, 2012.

Michael Vogt and Oliver Linton. Nonparametric estimation of a periodic sequence in the presence of a smooth trend. *Biometrika*, 101:121–140, 2014.

Haonan Wang and Jun Zhu. Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics*, 37(4):607–624, 2009.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2010.

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

Qiwei Yao and Peter J Brockwell. Gaussian maximum likelihood estimation for ARMA models II: spatial processes. *Bernoulli*, 12:403–429, 2006.

Zhiliang Ying. Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, 21:1567–1590, 1993.

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

Yunong Zhang and William E Leithead. Approximate implementation of the logarithm of the matrix determinant in gaussian process regression. *Journal of Statistical Computation and Simulation*, 77:329–348, 2007.

Zhou Zhou and Wei Biao Wu. Local linear quantile estimation for nonstationary time series. *The Annals of Statistics*, 37:2696–2729, 2009.