STATISTICAL MODELING FOR COMPLEX DATA

BIN SUN

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

November 2019

©Bin Sun 2019

Abstract

In this dissertation, we focus on statistical modeling techniques for exploring complex data with features such as high dimensionality, nonstationary structure, heavy-tailed distributions, missing data, etc. We study four problems: dimension reduction in high-dimensional data, clarifying complex patterns in nonstationary spatial data, improving hierarchical Bayesian modeling of spatio-temporal data with staircase pattern of missing observations, and detecting change points in spatio-temporal data with outliers and heavy-tailed observations.

Sufficient dimension reduction draws a lot of attention in the last twenty years due to the largely increasing dimensions of the covariates. The semiparametric approach to dimension reduction proposed by Ma and Zhu [2012] is a novel and completely different approach to dimension-reduction problems from the existing literature. We present a theoretical result that relaxes a critical condition required by the semiparametric approach. The asymptotic normality of the estimators still maintains under weaker assumptions. This improvement increases the applicability of the semipara-

metric approach.

For spatial data, nonstationarity brings difficulties to learn the underlying processes, more specifically, to find spatial dependency using the semivariogram model. We improve the modeling technique through dimension expansion proposed by Bornn et al. [2012] by considering the correlation structure. We propose two generalized least squares methods. Both of the methods provide more accurate parameter estimations than the least squares method, which has been demonstrated through simulation studies and real data analyses.

As spatio-temporal data are usually observed over a large area and in many years, modeling spatio-temporal data is non-trivial. Missing data makes the task even more challenging. One of the problems discussed in this dissertation is to model ozone concentrations in a region in the presence of missing data. We propose a method without assumptions on the correlation structure to estimate the covariance matrix through dimension expansion method for modeling the semivariograms in nonstationary fields based on the estimations from the hierarchical Bayesian spatio-temporal modeling technique [Le and Zidek, 2006]. For demonstration, we apply the method in ozone concentrations at 25 stations in the Pittsburgh region studied in Jin et al. [2012]. The comparison of the proposed method and the one in Jin et al. [2012] are provided through leave-one-out cross-validation which shows that the

proposed method is more general and applicable.

The last problem which is also related to spatio-temporal data is to detect structural changes for spatio-temporal data with missing in the presence of outliers and heavy-tailed observations. We improve the estimation algorithm of a general spatio-temporal autoregressive (GSTAR) model proposed by Wu et al. [2017]. We propose M-estimation-based EM algorithm and change-point detection procedure. Through data examples, we compare the proposed algorithm and the proposed change-point detection procedure with the existing ones and show that our method provides more robust estimation and is more accurate in detecting change points in the presence of outliers and/or heavy-tailed observations.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisors, Professor Yuehua Wu and Professor Yuejiao Fu, for their continuous support of my Ph.D. study and research. I appreciate their patience, inspiration, motivation and immense knowledge that helped me conquer all of the difficulties and brought me the courage to take the challenges.

I would also like to thank the other members in my Ph.D. committee, Professor Xin Gao and Professor Huaiping Zhu. I am also grateful for all faculty members, staff and fellow graduate students in the Department of Mathematics and Statistics at York. I would like to thank Professor Wenzhi Yang, Professor Xiaoping Shi and Professor Baisuo Jin for their help.

This journey would not have been possible without the support of my family. They give me all the courage and strength to take the challenge to embrace the bright future. Thank you, my husband, my parents, my sons, especially my uncle and auntie who lead me to the dawn from the dark.

Contents

List of Tables

List of Figures

1 Introduction

Following the advancement in science and technology, scientific data tend to grow in both size and complexity. The growing size and complexity of data bring challenges to the field of statistics as they demand more sophisticated statistical modeling techniques. This dissertation focuses on statistical modeling techniques for exploring complex data with features such as high dimensionality, heavy-tailed distribution, nonstationary structure in the underlying process, and missing observations. In this dissertation, we study four different problems: dimension reduction in high-dimensional data, clarifying complex patterns in nonstationary spatial data, improving hierarchical Bayesian modeling of spatio-temporal data with a staircase pattern of missing observations, and detecting change points in spatio-temporal data with outliers and heavy-tailed observations. We introduce the problems in the following four sections.

1.1 Sufficient Dimension Reduction

High dimensional and complex data bring tremendous challenges to statisticians. One central problem in high-dimensional regression setting is dimension reduction. Research in Sufficient Dimension Reduction (SDR) [Cook and Weisberg, 1991] has received great attention in recent years. The SDR reduces the covariate dimension to a few linear combinations of the covariates, which contain all the regression information between the response variable and the covariates. More specifically, the goal of the SDR is to identify a few linear combinations of the covariate $\mathbf{x}$, say $\mathbf{x}^T\boldsymbol{\beta}$, to substitute the original $\mathbf{x}$ without loss of information about the response. The model is as follows

$$F(y|\mathbf{x}) = F(y|\mathbf{x}^T\boldsymbol{\beta}), \text{ for all } y \in \mathbb{R}, \tag{1.1}$$

where Y is a univariate response variable, $\mathbf{x}$ is a $p \times 1$ covariate vector, $\boldsymbol{\beta}$ is a $p \times d$ matrix, and $F(y|\mathbf{x})$ denotes the conditional distribution function of Y given $\mathbf{x}$. This model assumes that given the linear combinations $\mathbf{x}^T\boldsymbol{\beta}$, the response variable Y is statistically independent of $\mathbf{x}$ [Cook, 1998]. Note that the estimator of $\boldsymbol{\beta}$ is not unique and identifiable. The column space of $\boldsymbol{\beta}$ is called dimension reduction subspace. It is of interest to identify the intersection of all $\boldsymbol{\beta}'s$, which is the central subspace, denoted by $S_{Y|\mathbf{x}}$. It is defined as the column space of $\boldsymbol{\beta}$ which satisfies (1.1) with the smallest number of columns d [Cook, 1998]. Cook [1998] proved that the central

subspace itself is a dimension reduction subspace. If one is only interested in the mean of Y conditional on $\mathbf{x}$, the corresponding model will be

$$\mathbb{E}(y|\mathbf{x}) = \mathbb{E}(y|\mathbf{x}^T\boldsymbol{\beta}), \text{ for all } y \in \mathbb{R}. \tag{1.2}$$

Then the subspace with the smallest number of columns d is called the central mean subspace.

The SDR methodology has been developed in recent years in mainly two streams: inverse regression and non-parametric methods [Ma and Zhu, 2013]. The inverse regression methods include Sliced Inverse Regression (SIR) by Li [1991], Sliced Average Variance Estimation (SAVE) by Cook [1998], Directional Regression (DR) by Li and Wang [2007] and so on. All of the inverse regression methods require the following two assumptions on the covariates:

1. Linearity condition: $\mathbb{E}(\mathbf{x}|\mathbf{x}^T\boldsymbol{\beta})$ is a linear function of $\mathbf{x}$;

2. Constant variance condition: $\text{cov}(\mathbf{x}|\mathbf{x}^T\boldsymbol{\beta})$ is a constant matrix.

Linearity condition requires the covariate to be elliptically contoured distributed [Eaton, 1986]. To meet both of the aforementioned conditions, the covariate has to be multivariate normal distributed. The non-parametric method, Minimum Average Variance Estimation (MAVE), was first introduced by Xia et al. [2002]. Later, Xia [2007] proposed density-based MAVE (dMAVE) which concerned the central mean

subspace estimation. The non-parametric methods require the covariates to be continuous. In practice, however, it is very common to have categorical or discrete covariates in a regression problem.

To eliminate all of the stringent assumptions on the covariates, recently Ma and Zhu [2012] provided a dimension reduction method based on the semiparametric framework. The likelihood of one observation for model (1.2) can be written as

$$\eta_1(\mathbf{x})\eta_2(Y, \mathbf{x}^T\boldsymbol{\beta}),$$

where η_1 and η_2 are the probability mass function (pmf) or probability density function (pdf) of $\mathbf{x}$ and the conditional pmf/pdf of Y on $\mathbf{x}^T\boldsymbol{\beta}$. By applying the geometric tool [Bickels et al., 1993] on this particular model, Ma and Zhu [2012] derived the estimating function for the parameter of interest, $\boldsymbol{\beta}$, as follows

$$\mathbb{E}\left[\left(\mathbf{g}(Y, \mathbf{x}^T\boldsymbol{\beta}) - \mathbb{E}\left[\mathbf{g}(Y, \mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x}) - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\right] = 0,$$

for any functions $\boldsymbol{g}$ and $\boldsymbol{\alpha}$. The different choices of functions $\boldsymbol{g}$ and $\boldsymbol{\alpha}$ lead to different traditional methods SIR, SAVE, DR, and so on. The estimate from the influence function is $\sqrt{n}$-consistent and asymptotically normal distributed. The details of this result are stated in Theorem 1 of Ma and Zhu [2012].

However, the asymptotic results presented in Theorem 1 of Ma and Zhu [2012] require a set of conditions. One crucial condition requires the density functions

of $\mathbf{x}$ and $\mathbf{x}^T\boldsymbol{\beta}$ to be bounded away from 0. Such a condition narrows down the application of the method in a variety of situations. The most commonly used multivariate normal distribution, for example, does not meet this requirement. In this dissertation, we present theoretical results which allow us to relax this condition and at the same time to maintain the same asymptotic normality of the estimators.

In the literature, there are mainly two methods for relaxing the "bounded away from 0" condition. One is to add a positive constant sequence in the denominator. Fan [1993] used this technique to avoid zero in the denominator by adding n^{-2} to the denominator. The other method is the trimming method which uses the modified version of the kernel estimator in Zhu and Fang [1996]. We adopt the second method because it will theoretically relax the conditions on $\mathbf{x}$ and $\mathbf{x}^T\boldsymbol{\beta}$ in the semiparametric approaches in Ma and Zhu [2012]. In Chapter 2, we present theoretical results which allow us to relax this condition and, at the same time, to maintain the same asymptotic normality of the estimators.

1.2 Modeling Nonstationary Processes through Dimension Expansion

Spatial statistics focuses on modeling environmental processes such as agricultural output or air pollution. The goal of spatial statistics is to improve understanding

of the environmental random processes and make predictions for the locations of interest. Those environmental processes often have complex spatial features. Most of the existing spatial statistical methods for analyzing environmental processes assume that the processes are stationary [Cressie, 1993]. This assumption may be violated since the environmental processes are vulnerable to change and are easily affected by unstable environmental factors such as climate change, urban sprawl, and ozone layer depletion. We are interested in the development of the spatial statistical methods that can be applied to model nonstationary spatial random processes.

Let $\{\mathbf{Y}(\mathbf{x}) : \mathbf{x} \in \mathcal{S}\}$, $\mathcal{S} \in \mathcal{R}^d$, be an environmental random process, where $\mathbf{x}$ is a d-dimensional spatial index that varies continuously throughout the region $\mathcal{S}$. At n spatial locations denoted by $\{\mathbf{x}_i : i = 1, ..., n\}$, we observe realizations of the random process $\mathbf{Y}(\mathbf{x})$, i.e. $\{\mathbf{Y}(\mathbf{x}_i) : i = 1, ..., n\}$. We are interested in learning the spatial dependence of the process through the observed data. Semivariogram function which describes the degree of spatial dependency of an intrinsic stationary random process is a cornerstone of spatial statistics. An intrinsic stationary random process satisfies the following two conditions [Cressie, 1993]:

1. $\mathbb{E}(\mathbf{Y}(\mathbf{x})) = \mu$, for $\mathbf{x} \in \mathcal{S}$,

2. $\text{var}(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)) = 2\gamma(\mathbf{x}_i - \mathbf{x}_j)$,

where semivariogram is defined as $\gamma(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{2}\text{var}(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j))$ for two different

locations, $\mathbf{x}_i$ and $\mathbf{x}_j$, in the monitored region. Since an intrinsic stationary random process has a constant mean, Wackernagel [2003] also defined semivariogram as

$$\gamma(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{2}\mathbb{E}\left(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)\right)^2.$$

The most popular method for estimating semivariogram can be found in Matheron [1962] as following

$$\hat{\gamma}(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{2\left|\tau\right|}\sum_{\tau}\left(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)\right)^2, \tag{1.3}$$

where $|\tau|$ is the number of distinct pairs at the locations $\mathbf{x}_i$ and $\mathbf{x}_j$. A semivariogram is called isotropic, if $\gamma\left(\mathbf{x}_i - \mathbf{x}_j\right)$ is only a function of $\|\mathbf{x}_i - \mathbf{x}_j\|$, where $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the Euclidean distance d_{ij} between the two locations, that is

$$2\gamma(d_{ij}) = \mathbb{E}\left(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)\right)^2.$$

In Figure 1.1, we show two plots of two empirical semivariograms. There are 10 locations in both cases. The left side in Figure 1.1 is a nonstationary semivariogram that the correlation is not spatially dependent, and the right side is a stationary semivariogram where we can see the clear pattern of the spatial dependence.

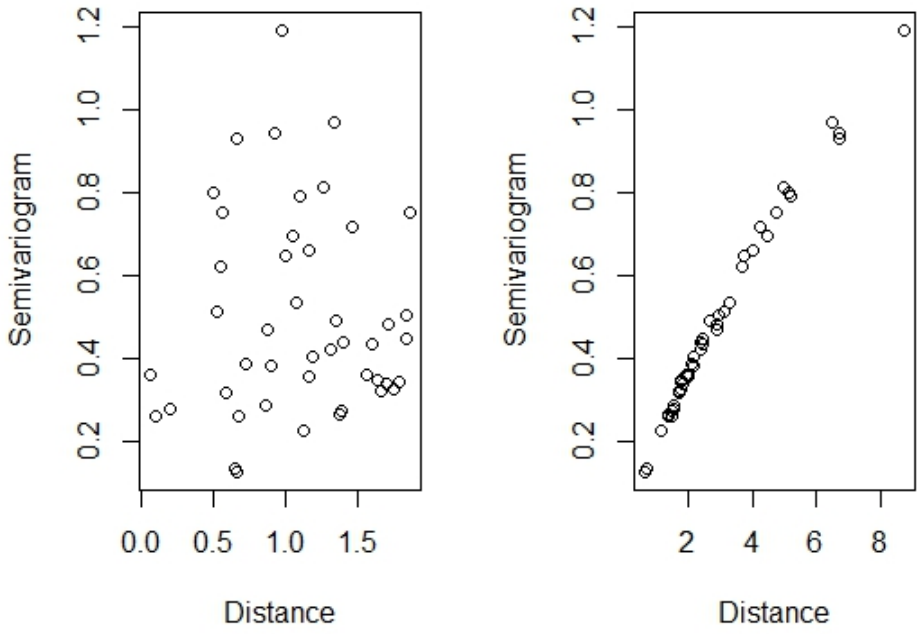

Figure 1.1: Semivariogram plots

According to the "First law of geography" [Tober, 1970], the observations are more related if their locations are closer. A stationary empirical semivariogram can be approximated by some functions. There are mainly three mathematical function forms that are used to approximate the semivariogram in applications [Cressie, 1993]. Each model is defined as the function of Euclidean distance d and some parameters.

- Exponential model

$$\gamma(d,\phi) = \begin{cases} 0 & d = 0 \\ \phi_1 + \phi_2 \left(1 - exp\left(-\frac{d}{\phi_3}\right)\right) & d \neq 0, \end{cases}$$

where $\phi = (\phi_1, \phi_2, \phi_3)^T,\ \phi_1 \geq 0,\ \phi_2 \geq 0$, and $\phi_3 \geq 0$.

- Spherical model

$$\gamma(d,\phi) = \begin{cases} 0 & d = 0 \\ \phi_1 + \phi_2 \left(\frac{3}{2}\left(\frac{d}{\phi_3}\right) - \frac{1}{2}\left(\frac{d}{\phi_3}\right)^3\right) & 0 < d < \phi_3 \\ \phi_1 + \phi_2 & d \geq \phi_3, \end{cases}$$

where $\phi = (\phi_1, \phi_2, \phi_3)^T,\ \phi_1 \geq 0,\ \phi_2 \geq 0$, and $\phi_3 \geq 0$.

- Gaussian model

$$\gamma(d,\phi) = \begin{cases} 0 & d = 0 \\ \phi_1 + \phi_2 \left(1 - exp\left(-\frac{d^2}{\phi_3^2}\right)\right) & d \neq 0, \end{cases}$$

where $\phi = (\phi_1, \phi_2, \phi_3)^T,\ \phi_1 \geq 0,\ \phi_2 \geq 0$, and $\phi_3 \geq 0$.

Bornn et al. [2012] proposed a novel approach to finding the latent dimensions over which the nonstationary fields exhibit stationarity through dimension expansion. They expanded the original field to a higher dimensional space over which the process achieves stationarity. Their idea is based on the theoretical work of Perrin

and Merring [2003] and Perrin and Schlather [2007]. In Bornn et al. [2012], the least-squares criterion does not consider the covariance structure of the empirical semivariogram, which are generally correlated. For example, assuming there are n locations, at the location $\mathbf{x}_i$, the observations of the Gaussian process $\mathbf{Y}(\mathbf{x}_i)$ at this location contributes to the calculation of the empirical semivariogram. In Chapter 3, we take consideration of the covariance structure of the empirical semivariograms and propose two generalized least-squares methods following Muller [1998]. Both methods provide more accurate latent dimensions estimation than the least-squares method. In Chapter 4, we apply the proposed method to estimate the covariance matrix for gauged and ungauged stations in modeling the spatio-temporal data.

1.3 Modeling Spatio-temporal Data with Monotone Missing Pattern

Spatio-temporal data has drawn a dramatically increasing attention due to their wide availability in many research fields including environmental study, climate change, and biology. They are usually spatially correlated and/or temporally correlated. In the literature, there are many approaches to model the spatial dependence structure as well as the temporal dependence structure in the spatio-temporal data. Examples can be found in Cressie [1993] and Cressie and Wikle [2011]. Modeling

spatio-temporal data is non-trivial since such data varies over space and time, and the interaction exists across different scales. Missing data makes the task even more challenging. As commented in Wikle et al. [1998], although we cannot escape the "curse of dimensionality", we can take advantage of recent developments in computational speed and numerical advances (e.g. Markov Chain Monte Carlo) that allow us to implement Bayesian spatio-temporal dynamical models in a hierarchical framework. Such a framework provides simple strategies for incorporating complicated spatio-temporal interactions at different stages of the models' hierarchy, and the models are feasible to be implemented for high dimensional data. Two popular hierarchical Bayesian spatio-temporal models can be found in [Wikle et al., 1998] and [Le et al., 2001]. Le et al. [2001] introduced the hierarchical Bayesian spatio-temporal modeling approach for spatio-temporal data with the monotone missing pattern. The monotone missing pattern appears when the data is reassembled in increasing order of monitoring periods, the data matrix is an ascending staircase as shown in Figure 1.2 [Le and Zidek, 2006]. In Figure 1.2, "o" represents the observed data, and "x" represents the missing data. Within each of the k blocks, the monitoring stations have the same pattern of missing data. Moreover, the numbers of missing data are in ascending order.

Note that the missingness in raw data is mostly at random and has no patterns

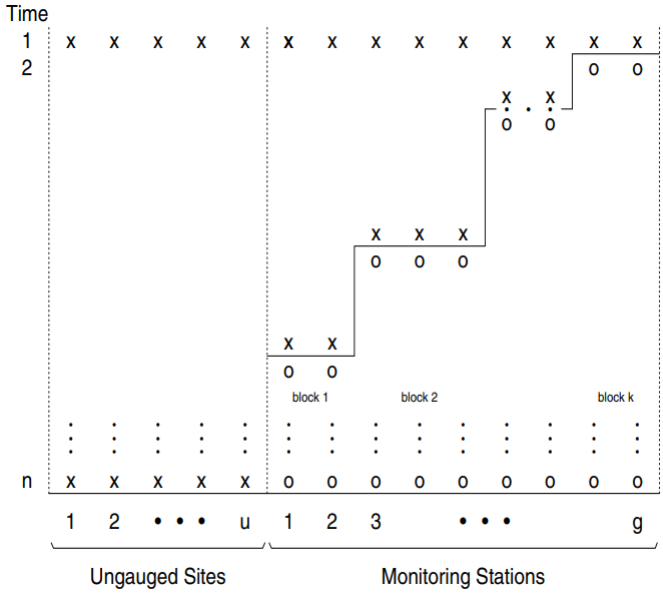

Figure 1.2: Monotone missing pattern of data

at all. When only a few observations within the blocks are missing, we can impute missing values by the predictions from regression models. Jin et al. [2012] successfully proposed a regression model to impute the missing values and then applied a hierarchical Bayesian spatio-temporal modeling technique to model the ground-level ozone concentration data in 4-consecutive summer months in the Pittsburgh region of the United States. They estimated the spatial correlation function for the gauged stations and obtained the covariance matrix for all of the stations to derive the predictive distribution. To estimate the spatial covariance matrix for all of the stations, first, they selected the generalized linear model with the quasi-Poisson family to fit

the correlation function by examing the pattern in the plot of spatial correlations based on the estimations from the hierarchical spatio-temporal modeling approach. Then, they obtained the covariance matrix using the estimated correlation. However, the generalized linear model with the quasi-Poisson family is not appropriate if there are negative correlations because it only applies to positive responses. This is a strong restriction because negative correlations are common for the ozone concentrations. Moreover, model selection only based on observed plots is naive and may cause overfitting.

In Chapter 4, we propose a method to estimate the covariance matrix through dimension expansion in the context of semivariogram modeling in nonstationary fields. For demonstration, we apply the proposed method on the same data set discussed in Jin et al. [2012]. Using the covariance matrix estimated by the proposed method on the entropy criterion in the environmental network design problem, the proposed method obtains interesting findings and the locations of the selected ungauged stations are more reasonable. We also evaluate the performance of the proposed method by leave-one-out cross-validation.

1.4 Detection of Change Points in Spatio-temporal Data

Research interest also arises on the topic to detect sudden changes occurring in spatio-temporal data over a long period. These changes could be due to exposure changes, instrument/observer changes, the implementation of government regularities and policies [Wu et al., 2015], etc. Under the framework of Bayesian approaches, Wyse et al. [2011] presented methods for analyzing multiple change-point models when dependency in the data is modeled through a hierarchical Gaussian Markov random field, and Altieri et al. [2015] proposed methods for detecting multiple change points over time in the heterogeneous intensity of a spatio-temporal point process with spatial and temporal dependence within segments. On the other hand, under the framework of maximum likelihood methods, Nappi-Choulet and Maury [2009] and Otto and Schmid [2016] introduced methods for modeling spatio-temporal or spatial data containing changes over time or space. More specifically, Nappi-Choulet and Maury [2009] proposed a hybrid method for incorporating a temporal regime switch into the spatio-temporal autoregressive model to deal with exogenous macroeconomic factors. For spatial data, Otto and Schmid [2016] proposed a test procedure to detect change points of multidimensional autoregressive processes. Their method works well to find possible structural breaks in the process that can occur at a certain distance from the predefined center. Most recently, Wu et al. [2017] proposed a

general spatio-temporal autoregressive (GSTAR) model which takes into account the effect of station surroundings, seasonality, temporal correlation among observations at the same spatial location and spatial correlation among observations from different spatial locations. The model is so multi-functional that it can also be used to detect new influences that can largely affect the measurements in the treatment area compared to the control area. However, their method is dependent on the normality assumption.

Chapter 5 studies the problem of change-point detection in spatio-temporal data with undetectable outliers and/or heavy-tailed observations. As the spatio-temporal data is usually observed over a large area and in many years, undetectable outliers can easily occur unexpectedly in any day for any small area because of measurement error or other reasons. The parameter estimation method which is given in Wu et al. [2017] may not be stable or robust. There is a great need to develop a parameter estimation method for the GSTAR model that is resistant to outliers and/or heavy-tailed observations. In the development of robust methods, M-estimation plays an important and complementary role [Huber, 1973]. We propose a robust version of EM-type algorithm, namely ${}_M$EM-type algorithm, which provides more robust estimation in the presence of outliers and/or heavy-tailed observations.

2 Semiparametric Approach to Dimension Reduction

The asymptotic results presented in Theorem 1 of Ma and Zhu [2012] required a set of conditions. One crucial condition requires the density functions of $\mathbf{x}$ and $\mathbf{x}^T\boldsymbol{\beta}$ to be bounded away from 0. Such a condition narrows down the application of the method in a variety of situations. The most commonly used multivariate normal distribution, for example, does not meet this requirement. In this dissertation, we present theoretical results which allow us to remove this condition and at the same time to maintain the same asymptotic normality of the estimators.

2.1 Main Results

In the literature, there are mainly two methods for removing the bounded away from 0 condition. One is to add a positive constant sequence in the denominator. Fan [1993] used this technique to avoid zero in the denominator by adding n^{-2} to

the denominator. The other method is the trimming method which uses a modified version of the kernel estimator in Zhu and Fang [1996]. We adopt the second method in this dissertation because it will theoretically relax the conditions on the density functions of $\mathbf{x}$ and $\mathbf{x}^T\boldsymbol{\beta}$ in the semiparametric approaches in Ma and Zhu [2012]. In the next section, we state the main results of the trimming method. The proofs are given after.

Let $\mathbf{x}$ be a $p \times 1$ covariate vector and Y a univariate response. For each $b > 0$, let $f_b = \max\{f(\mathbf{x}^T\boldsymbol{\beta}), b\}$, and $\hat{f}_b = \max\{\hat{f}(\mathbf{x}^\mathbf{T}\boldsymbol{\beta}), b\}$. Here for simplicity, we define $\boldsymbol{R}(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})}{f(\mathbf{x}^T\boldsymbol{\beta})}$. We estimate $\boldsymbol{R}(\mathbf{x}^T\boldsymbol{\beta})$ by Nadaraya-Watson kernel estimator:

$$\hat{\boldsymbol{R}}(\mathbf{x}_i^T\boldsymbol{\beta}) = \hat{\mathbb{E}}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] = \frac{\frac{1}{n-1}\sum\limits_{j\neq i} K_h(\mathbf{x}_j^T\boldsymbol{\beta}-\mathbf{x}_i^T\boldsymbol{\beta})\boldsymbol{\alpha}(\mathbf{x}_j)}{\frac{1}{n-1}\sum\limits_{j\neq i} K_h(\mathbf{x}_j^T\boldsymbol{\beta}-\mathbf{x}_i^T\boldsymbol{\beta})} = \frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}(\mathbf{x}_i^T\boldsymbol{\beta})}.$$

Now we formulate a set of weaker conditions D1 to D4 under which the asymptotic normality of the estimator of $\boldsymbol{\beta}$ still holds.

D1. The univariate kernel function $K(\cdot)$ is Lipschitz with compact support. It satisfies $\int K(u)du = 1$, $\int u^i K(u)du = 0,\ 1 \leq i \leq m-1$, $0 \neq \int u^m K(u)du < \infty$. The d-dimensional kernel function is a product of d univariate kernel functions, that is, $K_h(\boldsymbol{u}) = K(\mathbf{u}/h)/h^d = \prod\limits_{j=1}^{d} K_h(u_j)$ for $\boldsymbol{u} = (u_1, ..., u_d)^T$.

D2. Define

$$\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta}) = \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right] f(\mathbf{x}^T\boldsymbol{\beta}),$$

$$\boldsymbol{r}_2(\mathbf{x}^T\boldsymbol{\beta}) = \mathbb{E}\left[\mathbf{g}(Y, \mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right] f(\mathbf{x}^T\boldsymbol{\beta}).$$

The mth derivatives of $\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})$, $\boldsymbol{r}_2(\mathbf{x}^T\boldsymbol{\beta})$ and $f(\mathbf{x}^T\boldsymbol{\beta})$ are locally Lipschitz-continuous.

D3. The density functions, $f_{\mathbf{x}}(\mathbf{x})$ and $f(\mathbf{x}^T\boldsymbol{\beta})$, are bounded from above. Each entry in the matrices $\mathbb{E}\left[\mathbf{g}(Y, \mathbf{x}^T\boldsymbol{\beta})\mathbf{g}^T(Y, \mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]$ and $\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})\boldsymbol{\alpha}^T(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]$ is locally Lipschitz-continuous and bounded from above as a function of $\mathbf{x}^T\boldsymbol{\beta}$.

D4. As $n \to \infty$, $h \sim n^{-c_1}$, $b \sim n^{-c_2}$ with positive numbers c_1 and c_2 satisfying that $\frac{2c_2}{m} < c_1 < \min\left(\frac{1-c_2}{d+1}, \frac{1-4c_2}{d}\right)$, and $0 < c_2 < 1/4$.

Condition D1 states the regularity conditions for the kernel. Conditions D2 and D3 are concerned with the smoothness of the density functions, which are similar to the conditions C2 and C3 in Ma and Zhu [2012]. Condition D4 is the key condition to remove the bounded from below constraint on density functions of $\mathbf{x}$ and $\mathbf{x}^T\boldsymbol{\beta}$. The order of the bandwidth h and the trimming value b are defined in Condition D4. By finding the appropriate b defined in these conditions, we relax the condition of bounded from below on the density functions of the covariates. Our empirical studies

suggest that when b is small enough, the trimming method and Ma and Zhu (2012)'s method give the same result. For small to moderate sample sizes (n=50, 100, 200, 500), we suggest $b = 0.1n^{-\frac{1}{5}}$ based on our empirical studies. In the following, we summarize the main result in Theorem 1.

Theorem 1. *Under Conditions D1 to D4, the estimator $\hat{\boldsymbol{\beta}}$ obtained from the estimating equation*

$$\sum_{i=1}^{n}\left[\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})-\hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x}_i)-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\right]=0 \quad (2.1)$$

satisfies $\sqrt{n}\mathbf{A}\mathrm{vec}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\to N(\mathbf{0},\mathbf{B})$ in distribution, where $\mathbf{A}$ and $\mathbf{B}$ are defined as following

$$\mathbf{A}=\mathbb{E}\left\{\frac{\partial\left[\left(\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})-\mathbb{E}\left[\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x})-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\right]}{\partial\left\{\mathrm{vec}(\boldsymbol{\beta})\right\}^T}\right\},$$

$$\mathbf{B}=\mathrm{cov}\left\{\mathrm{vec}\left[\left(\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})-\mathbb{E}\left[\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x})-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\right]\right\},$$

where $\mathrm{vec}(\mathbf{M})$ *denotes the vector formed by concatenating the columns of* $\mathbf{M}$, $\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]=\frac{\hat{\mathbf{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}$ *and* $\hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]=\frac{\hat{\mathbf{r}}_2(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}$.

The following two lemmas are crucial results to get Theorem 1.

Lemma 1. *Assume that Conditions D1 to D4 hold. Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\}$$
$$\times\left\{\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\}=o_p(n^{-1/2}),$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\}$$
$$\times\left\{\boldsymbol{\alpha}(\mathbf{x}_i)-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\}=o_p(n^{-1/2}).$$

Lemma 2 is a modified version of Lemma 3 in Ma and Zhu [2012].

Lemma 2. *Assume that Conditions D1 to D4 hold. Let*

$$\Omega_{\boldsymbol{\beta}}=\left\{(\mathbf{x},Y,\hat{\boldsymbol{\beta}}):\mathbf{x}\in\mathbb{R}^d,Y\in\mathbb{R},\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|\leq Cn^{-1/2}\right\},$$

where $\|\cdot\|$ *is the Euclidean norm and* C *is a constant. Then there exists a basis of* $\boldsymbol{\beta}$ *of* $S_{Y|\mathbf{x}}$ *such that*

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]+\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\right|=o_p(1),$$

and

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\mathbb{E}}_b\left[\mathbf{g}(Y,\mathbf{x}^T\hat{\boldsymbol{\beta}})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right.$$
$$\left.-\mathbb{E}\left[\mathbf{g}(Y,\mathbf{x}^T\hat{\boldsymbol{\beta}})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]+\mathbb{E}\left[\mathbf{g}(Y,\mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right|=o_p(1).$$

We provide the details of proof in the following.

Proof of Lemma 1

We show the proof of the first equality since the two equations are similar. Here we relax the condition on $f_{\mathbf{x}}(\mathbf{x})$ and $f(\mathbf{x}^T\boldsymbol{\beta})$ in Ma and Zhu [2012] which requires the bounded below on both pdfs. For each $b > 0$, let $f_b(\mathbf{x}^T\boldsymbol{\beta}) = \max\{f(\mathbf{x}^T\boldsymbol{\beta}), b\}$ and $\hat{f}_b(\mathbf{x}^T\boldsymbol{\beta}) = \max\{\hat{f}(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}), b\}$. Define $\boldsymbol{R}(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})}{f(\mathbf{x}^T\boldsymbol{\beta})}$ and $\boldsymbol{R}_b(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})}{f_b(\mathbf{x}^T\boldsymbol{\beta})} = \boldsymbol{R}(\mathbf{x}^T\boldsymbol{\beta}) \cdot \frac{f(\mathbf{x}^T\boldsymbol{\beta})}{f_b(\mathbf{x}^T\boldsymbol{\beta})}$. Then

$$\hat{\boldsymbol{R}}(\mathbf{x}_i^T\boldsymbol{\beta}) = \hat{\mathbb{E}}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] = \frac{\frac{1}{n-1}\sum_{j\neq i} K_h(\mathbf{x}_j^T\boldsymbol{\beta} - \mathbf{x}_i^T\boldsymbol{\beta})\boldsymbol{\alpha}(\mathbf{x}_j)}{\frac{1}{n-1}\sum_{j\neq i} K_h(\mathbf{x}_j^T\boldsymbol{\beta} - \mathbf{x}_i^T\boldsymbol{\beta})} = \frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}(\mathbf{x}_i^T\boldsymbol{\beta})},$$

$$\hat{\boldsymbol{R}}_b(\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}.$$

We define $\varepsilon_i = \boldsymbol{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) - \mathbb{E}\left[\boldsymbol{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]$. In the following, we will show that the order of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left\{\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\} = o_p(1)$.

We expand $\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]$ as follows

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left[\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})}-\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})-f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)}{f_b^2(\mathbf{x}_i^T\boldsymbol{\beta})}\right.\\
&-\frac{\left(\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})-\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right)\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})-f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})\cdot\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}\\
&\left.+\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})-f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)^2}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})\cdot f_b^2(\mathbf{x}_i^T\boldsymbol{\beta})}\right]\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left(I_{i1}-I_{i2}-I_{i3}+I_{i4}\right).
\end{aligned}
$$

Now we show that $\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i I_{ik}\right|=o_p(1),$ for $k=2,3,4$. First we examine the case for $k=2$.

$$
\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i I_{i2}\right|=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})-f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)}{f_b^2(\mathbf{x}_i^T\boldsymbol{\beta})}\right|.
$$

By the uniform convergence of non-parametric regression [Mack and Silverman, 1982], one can get that

$$
\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{f}_b(\mathbf{x}^T\boldsymbol{\beta})-f_b(\mathbf{x}^T\boldsymbol{\beta})\right|=O_p\left(h^m+\frac{\log n}{\sqrt{nh^d}}\right),
$$

$$
\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\boldsymbol{r}}_1(\mathbf{x}^T\boldsymbol{\beta})-\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})\right|=O_p\left(h^m+\frac{\log n}{\sqrt{nh^d}}\right).
$$

Under Conditions D2 and D3, according to Chebyshev's inequality, it is easy to show

for any $\eta > 0$ and some $C_1 > 0$ and $C_2 > 0$, we have

$$\begin{aligned} P\left(\left|\frac{1}{\sqrt{n\log n}}\sum_{i=1}^{n}\varepsilon_i\right| > \eta\right) &\leq \frac{\mathbb{E}\left(\sum_{i=1}^{n}\varepsilon_i\right)^2}{\eta^2 n\log n} \\ &\leq \frac{C_1\sum_{i=1}^{n}\mathbb{E}\left(\varepsilon_i\right)^2}{\eta^2 n\log n} \\ &\leq \frac{C_2\sum_{i=1}^{n}\mathbb{E}\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{g}^T(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})\right)}{\eta^2 n\log n} \to 0,\ as\ n\to\infty. \end{aligned}$$

Therefore, we have

$$\left|\frac{1}{\sqrt{n\log n}}\sum_{i=1}^{n}\varepsilon_i\right| = o_p(1), \quad \left|\frac{1}{\sqrt{n\log n}}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right| = o_p(1).$$

By the definitions of $f_b(\mathbf{x}^T\boldsymbol{\beta})$ and $\hat{f}_b(\mathbf{x}^T\boldsymbol{\beta})$, we have $\frac{1}{f_b(\mathbf{x}^T\boldsymbol{\beta})} \leq \frac{1}{b}$, and $\frac{1}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})} \leq \frac{1}{b}$, then

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i I_{i2}\right| = O_p\left((h^m + n^{-1/2}h^{-d/2}\log n)b^{-2}\log^{1/2} n\right) = O_p(\Delta_1).$$

Next we examine the cases for $k = 3, 4$:

$$\begin{aligned} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i I_{i3}\right| &= \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\frac{\left(\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta}) - \boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right)\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta}) - f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})\cdot\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})}\right| \\ &= O_p\left[(h^{2m} + n^{-1}h^{-d}\log^2 n)b^{-2}\log^{1/2} n\right] = O_p(\Delta_2), \\ \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i I_{i4}\right| &= \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\frac{\boldsymbol{r}_1\left(\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta}) - f_b(\mathbf{x}_i^T\boldsymbol{\beta})\right)^2}{\hat{f}_b(\mathbf{x}_i^T\boldsymbol{\beta})\cdot f_b^2(\mathbf{x}_i^T\boldsymbol{\beta})}\right| \\ &= O_p\left[(h^{2m} + n^{-1}h^{-d}\log^2 n)b^{-3}\log^{1/2} n\right] = O_p(\Delta_3). \end{aligned}$$

According to the conditions $\frac{2c_2}{m} < c_1 < \min\left(\frac{1-c_2}{d+1}, \frac{1-4c_2}{d}\right)$ and $0 < c_2 < 1/4$, we have $O_p(\Delta_j) = o_p(1),\ j = 1, 2, 3$. Then we show that $\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} \varepsilon_i \left[\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})} - \frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\right]\right| = o_p(1)$. When $f(\mathbf{x}^T\boldsymbol{\beta}) > b$, then $f_b(\mathbf{x}^T\boldsymbol{\beta}) = f(\mathbf{x}^T\boldsymbol{\beta})$, we have

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})} - \frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\right| = \frac{1}{f}\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta}) - \boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right| = O_p\left(h^m + \frac{\log n}{\sqrt{nh^d}}\right).$$

Otherwise, when $f(\mathbf{x}^T\boldsymbol{\beta}) \le b$, then

$$\begin{aligned}
&\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f_b(\mathbf{x}_i^T\boldsymbol{\beta})} - \frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\right| \\
= &\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta}) - \boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{b} + \frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\left(\frac{f(\mathbf{x}_i^T\boldsymbol{\beta})}{b} - 1\right)\right| \\
\le &\frac{1}{b}\sup_{\Omega_{\boldsymbol{\beta}}}\left|\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta}) - \boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right| + \sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\left(\frac{f(\mathbf{x}_i^T\boldsymbol{\beta})}{b} - 1\right)\right|.
\end{aligned}$$

Since $\left|\frac{f(\mathbf{x}_i^T\boldsymbol{\beta})}{b} - 1\right| I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b) \le I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)$, we have

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})}\left(\frac{f(\mathbf{x}_i^T\boldsymbol{\beta})}{b} - 1\right) I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right| \le \sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})} I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right|.$$

Under Condition D3, we have $\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})\boldsymbol{\alpha}^T(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]$ bounded from above. Because $\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})\boldsymbol{\alpha}^T(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right] \ge \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\mathbb{E}^T\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]$, we have $\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]$ bounded from above. Therefore, for each $1 \le i \le n$,

$$\mathbb{E}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})} I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right| = \mathbb{E}\left|\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right| = O(b),$$

which implies $\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})} I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right| = O_p(b)$. Then under Condition D4,

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})}{f(\mathbf{x}_i^T\boldsymbol{\beta})} I(f(\mathbf{x}_i^T\boldsymbol{\beta}) \le b)\right|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\right| = O_p\left(b\log^{1/2} n\right) = o_p(1).$$

Meanwhile,

$$\frac{1}{b}\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\boldsymbol{r}}_1(\mathbf{x}_i^T\boldsymbol{\beta})-\boldsymbol{r}_1(\mathbf{x}_i^T\boldsymbol{\beta})\right|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\right| = O_p\left(\frac{h^m\log^{1/2}n}{b}+\frac{\log^{1/2}n}{b}\frac{\log n}{\sqrt{nh^d}}\right)$$

$$= o_p(1).$$

Hence we have shown $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left\{\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\} = o_p(1)$, and hence, completed the proof of Lemma 1.

Proof of Lemma 2

Because of the similarity of the two terms, we only prove the first term as follows

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]+\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\boldsymbol{\beta}\right]\right|. \quad (2.2)$$

Following Ma and Zhu [2012], we treat the nominators and denominators separately. We define $\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]$ as follows

$$\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x})|\mathbf{x}^T\hat{\boldsymbol{\beta}}\right]=\frac{\frac{1}{n}\sum_{i=1}^{n}K_h(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-\mathbf{x}^T\hat{\boldsymbol{\beta}})\boldsymbol{\alpha}(\mathbf{x}_i)}{\hat{f}_b(\mathbf{x}^T\hat{\boldsymbol{\beta}})}.$$

Therefore, (2.2) becomes

$$\sup_{\Omega_{\boldsymbol{\beta}}}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}K_h(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-\mathbf{x}^T\hat{\boldsymbol{\beta}})\boldsymbol{\alpha}(\mathbf{x}_i)}{\hat{f}_b(\mathbf{x}^T\hat{\boldsymbol{\beta}})}-\frac{\frac{1}{n}\sum_{i=1}^{n}K_h(\mathbf{x}_i^T\boldsymbol{\beta}-\mathbf{x}^T\boldsymbol{\beta})\boldsymbol{\alpha}(\mathbf{x}_i)}{\hat{f}_b(\mathbf{x}^T\boldsymbol{\beta})}\right.$$

$$\left.-\frac{\boldsymbol{r}_1(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{f(\mathbf{x}^T\hat{\boldsymbol{\beta}})}+\frac{\boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})}{f(\mathbf{x}^T\boldsymbol{\beta})}\right|.$$

Ma and Zhu [2012] showed that

$$\sup_{\Omega_{\boldsymbol{\beta}}} \Big| \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\hat{\boldsymbol{\beta}})\boldsymbol{\alpha}(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{\beta})\boldsymbol{\alpha}(\mathbf{x}_i)$$
$$-\boldsymbol{r}_1(\mathbf{x}^T\hat{\boldsymbol{\beta}}) + \boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})\Big| = O_p(h^m/\sqrt{n} + n^{-1}h^{-d-1}\log n).$$

For the denominator, we let $\boldsymbol{\alpha}(\mathbf{x}_i) = 1$ and consider $f(\mathbf{x}^T\boldsymbol{\beta})$ in the following two cases.

- **Case 1**, $f(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}) \geq b$, then $f_b(\mathbf{x}^T\boldsymbol{\beta}) = f(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})$, we have

$$\sup_{\Omega_{\boldsymbol{\beta}}} \Big| \hat{f}_b(\mathbf{x}^T\hat{\boldsymbol{\beta}}) - \hat{f}_b(\mathbf{x}^T\boldsymbol{\beta}) - f_b(\mathbf{x}^T\hat{\boldsymbol{\beta}}) + f_b(\mathbf{x}^T\boldsymbol{\beta})\Big|$$
$$= O_p(h^m/\sqrt{n} + n^{-1}h^{-d-1}\log n).$$

- **Case 2**, $f(\mathbf{x}^T\boldsymbol{\beta}) < b$, then $f_b(\mathbf{x}^T\boldsymbol{\beta}) = b$, (2.2) becomes

$$\sup_{\Omega_{\boldsymbol{\beta}}} \frac{1}{b}\Big| \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\hat{\boldsymbol{\beta}})\boldsymbol{\alpha}(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{\beta})\boldsymbol{\alpha}(\mathbf{x}_i)$$
$$-\boldsymbol{r}_1(\mathbf{x}^T\hat{\boldsymbol{\beta}}) + \boldsymbol{r}_1(\mathbf{x}^T\boldsymbol{\beta})\Big| = O_p(\frac{1}{b}h^m/\sqrt{n} + \frac{1}{b}n^{-1}h^{-d-1}\log n).$$

  Here we need $\frac{-\frac{1}{2}+c_2}{m} < c_1 < \frac{1-c_2}{d+1}$ to achieve the convergence which is insured by Condition D4. Therefore in both cases, (2.2) is of order $o_p(1)$.

Proof of Theorem 1

The left side of (2.1) can be written as

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-\hat{\mathbb{E}}_b\left[\boldsymbol{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x}_i)-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\\
=&\sum_{i=1}^{n}\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x}_i)-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\\
&+\sum_{i=1}^{n}\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\\
&+\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\left(\boldsymbol{\alpha}(\mathbf{x}_i)-\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\\
&+\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\hat{\boldsymbol{\beta}})|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right)\\
&\times\left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]-\hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right). \qquad (2.3)
\end{aligned}
$$

The first term quantity is of order $O_p(\sqrt{n})$. By Taylor's expansion, it can be expanded as

$$
\begin{aligned}
&\sum_{i=1}^{n}\left[\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\left(\alpha(\mathbf{x}_i)-\mathbb{E}\left[\alpha(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\right]\\
&+\sum_{i=1}^{n}\mathrm{dvec}\left\{\frac{\partial\mathrm{vec}\left[\left(\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})-\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\left(\alpha(\mathbf{x}_i)-\mathbb{E}\left[\alpha(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\right]}{\partial\left\{\mathrm{vec}(\boldsymbol{\beta})\right\}^T}\right\}\\
&\times\mathrm{vec}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+o_p(\sqrt{n}).
\end{aligned}
$$

Heredvec indicates that $\text{dvec}\left(\text{vec}\left(\mathbf{M}\right)\right) = \mathbf{M}$, for any matrix $\mathbf{M}$. By Central Limit Theorem, if the other three terms in (2.3) are of order $o_p(n^{1/2})$, then

$$\sqrt{n}\mathbf{A}\text{vec}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, \mathbf{B}),$$

where $\mathbf{A}$ and $\mathbf{B}$ are given in the Theorem 1 earlier. First we show the second term in (4) is of order $o_p(\sqrt{n})$. By Lemma 2, the second term becomes

$$\sum_{i=1}^{n} \left[\left(\mathbf{g}(Y_i, \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) - \mathbb{E}\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) | \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \right] \right) \left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] - \hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] \right) \right]$$
$$\times \left\{ 1 + o_p(1) \right\}.$$

By Taylor's expansion, this term asymptotically becomes

$$\sum_{i=1}^{n} \left[\left(\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) - \mathbb{E}\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) | \mathbf{x}_i^T\boldsymbol{\beta} \right] \right) \left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] - \hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] \right) \right]$$
$$\times \left\{ 1 + o_p(1) \right\}.$$

Lemma 1 indicates that the above term is of order $o_p(\sqrt{n})$ under Conditions D1 to D4. Next we turn to the third term in (2.3). By Lemma 2, the term becomes

$$\sum_{i=1}^{n} \left[\left(\mathbb{E}\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) | \mathbf{x}^T\boldsymbol{\beta} \right] - \hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) | \mathbf{x}^T\boldsymbol{\beta} \right] \right) \left(\boldsymbol{\alpha}(\mathbf{x}_i) - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \right] \right) \right]$$
$$\times \left\{ 1 + o_p(1) \right\}$$
$$= \sum_{i=1}^{n} \left[\left(\mathbb{E}\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) | \mathbf{x}^T\boldsymbol{\beta} \right] - \hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) | \mathbf{x}^T\boldsymbol{\beta} \right] \right) \right.$$
$$\left. \times \left(\boldsymbol{\alpha}(\mathbf{x}_i) - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] + \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\boldsymbol{\beta} \right] - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i) | \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \right] \right) \right] \left\{ 1 + o_p(1) \right\}.$$

Since $\mathbb{E}\left[\alpha(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]$ is locally Lipschitz-continuous, we have

$$\left|\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right| \leq c\left|\mathbf{x}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|.$$

We have assumed that $\left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right\| \leq cn^{-1/2}$, therefore

$$\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right] - \hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)$$
$$\times\left\{\boldsymbol{\alpha}(\mathbf{x}_i) - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right\} = o_p(\sqrt{n}),$$

$$\sum_{i=1}^{n}\left(\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right] - \hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)$$
$$\times\left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] - \mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right]\right) = o_p(\sqrt{n}).$$

For the last term in (2.3), by Lemma 2, it can be written as

$$\sum_{i=1}^{n}\left[\left(\mathbb{E}\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right] - \hat{\mathbb{E}}_b\left[\mathbf{g}(Y_i,\mathbf{x}_i^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}\right]\right)\right.$$
$$\left.\left(\mathbb{E}\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right] - \hat{\mathbb{E}}_b\left[\boldsymbol{\alpha}(\mathbf{x}_i)|\mathbf{x}_i^T\boldsymbol{\beta}\right]\right)\right] \times \{1+o_p(1)\}$$

which is of the order $o_p(\sqrt{n})$ under Condition D4. Finally, the proof is completed by combining all the results for the four terms in (2.3).

2.2 Trimming Parameter Selection

We suggest selecting the trimming parameter b to be $0.1n^{-\frac{1}{5}}$. We present some simulations to show how the choice of b affects the estimates. Each experiment is conducted

500 times. Let $\mathbf{x} = (X_{1,\ldots,}X_p)$. We let $p = 6$, and $\mathbf{x}$ are generated from normal population with mean zero and variance-covariance matrix $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.05^{|i-j|}$. The model is defined as follows

$$Y = (\mathbf{x}^T\beta_1)^2 + (\mathbf{x}^T\beta_2)^2 + 0.5\varepsilon,$$

where $\epsilon_i^{'}$s are independently generated from the standard normal population, $\beta_1 = (1,1,1,1,1,1)^T/\sqrt{6}$ and $\beta_2 = (1,-1,1,-1,1,-1)^T/\sqrt{6}$. The performance of the estimators is tested using the Euclidean distance between $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$, defined as the Frobenius norm of the matrix $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T$. In the simulations, the distance ranges from zero to two, and a smaller distance means a better estimate. We choose the trimming value d to be $n^{-\frac{1}{5}}$ and $0.1n^{-\frac{1}{5}}$. The results of the simulations are presented in Table 2.1 for different sizes of sample size, $n = 50,\ 100,\ 200,\ 500$. In Table 2.1, Semi-PHD refers to the semiparametric approach on PHD method in Ma and Zhu (2012), while Trimmed Semi-PHD refers to the proposed approach. The results show that when b is too large (e.g. $b = n^{-\frac{1}{5}}$), the proposed estimator performs significantly worse than Ma and Zhu [2012]. Our empirical studies suggest that when b is small enough, the trimming method and Ma and Zhu [2012] give the same result. For small to moderate sample sizes ($n = 50,\ 100,\ 200,\ 500$), we suggest $b = 0.1n^{-\frac{1}{5}}$ based on our empirical studies.

Table 2.1: Mean and standard deviation of the Euclidean distances

	n=50	n=100	n=200	n=500
Semi-PHD	0.12(0.32)	0.14(0.29)	0.11(0.26)	0.09(0.29)
Trimmed Semi-PHD $b = n^{-\frac{1}{5}}$	0.40(0.30)	0.38(0.33)	0.42(0.34)	0.39(0.30)
Trimmed Semi-PHD $b = 0.1n^{-\frac{1}{5}}$	0.12(0.32)	0.14(0.29)	0.11(0.26)	0.09(0.29)

3 Generalized Least-Squares in Modeling Nonstationary Processes

Bornn et al. [2012] proposed a novel approach to finding the latent dimensions over which the nonstationary fields exhibit stationarity through dimension expansion. They expanded the original field to a higher dimensional space over which the process achieves stationarity. Their idea is based on the theoretical work of Perrin and Merring [2003] and Perrin and Schlather [2007]. Perrin and Merring [2003] proved that any low-dimensional nonstationary random field in R^p can be viewed as a projection of a second-order stationary field in R^{2p}. Later, Perrin and Schlather [2007] proved that a Gaussian random process in R^d can be interpreted as a sample from a stationary random field in R^{d+p}, $p \geq 2$, under the moment constraints that all components of the Gaussian process have the same expectations and variances. Bornn et al. [2012] justified that for a nonstationary Gaussian process $\mathbf{Y}(\mathbf{x})$, where $\mathbf{x} \in R^d$, there exists extra dimensions $\mathbf{z} \in R^p$, $p > 0$, such that the expanded process

$\mathbf{Y}([\mathbf{x}, \mathbf{z}])$ was stationary under appropriate moment constraints. Note that $[\mathbf{x}, \mathbf{z}]$ is the concatenation of the dimensions $\mathbf{x}$ and $\mathbf{z}$. The stationary semivariogram with latent dimensions can be expressed by

$$\gamma([\mathbf{x}_i, \mathbf{z}_i] - [\mathbf{x}_j, \mathbf{z}_j]) = \frac{1}{2}\mathbb{E}\left(\mathbf{Y}([\mathbf{x}_i, \mathbf{z}_i]) - \mathbf{Y}([\mathbf{x}_j, \mathbf{z}_j])\right)^2,$$

where $[\mathbf{x}_i, \mathbf{z}_i]$ is the expanded spatial index for the ith location.

In the geostatistical study, the 2-dimensional space constituted by longitude and latitude is quite commonly recorded for the locations of interest. The nonstationary semivariogram for space with only longitude and latitude can achieve stationarity by including an extra dimension such as the elevation. To learn the latent dimensions non-parametrically from information contained within the data, Bornn et al. [2012] proposed the lasso-penalized least-squares criterion (OLS) as following

$$\left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{\text{OLS}} = \underset{\boldsymbol{\phi}, \mathbf{Z}}{\operatorname{argmin}} \sum_{i<j} \left\{\hat{\gamma}_{i,j} - \gamma_{\boldsymbol{\phi}}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right\}^2 + \lambda \sum_{k=1}^{p} \left\|\mathbf{Z}_{.k}\right\|_1, \tag{3.1}$$

where $\hat{\gamma}_{i,j}$ is the estimated semivariogram by (1.3) and $d_{i,j}([\mathbf{X}, \mathbf{Z}])$ is the Euclidean distance between the locations $[\mathbf{x}_i, \mathbf{z}_i]$ and $[\mathbf{x}_j, \mathbf{z}_j]$, $\mathbf{Z}_{.k}$ is the kth column of $\mathbf{Z}$, and $\|\cdot\|_1$ is the L_1 norm. $[\mathbf{X}, \mathbf{Z}]$ is the concatenation of the matrices $\mathbf{X}$ and $\mathbf{Z}$. The tuning parameter λ in the group lasso is used to determine the number of latent dimensions and regularize the estimation of $\mathbf{Z}$ to prevent overfitting. Note that $\gamma_{\boldsymbol{\phi}}(d_{i,j}([\mathbf{X}, \mathbf{Z}]))$ is a parametric semivariogram model with the parameter $\boldsymbol{\phi}$ for stationary fields.

Bornn et al. [2012] pointed out that their method produced similar results if any other parametric stationary semivariogram models were used.

However, the least-squares criterion (3.1) does not consider the covariance structure of the $\hat{\gamma}_{i,j},$ for $j \neq i$, which are generally correlated. For example, assuming there are n locations, at the location $\mathbf{x}_i$, the observations of the Gaussian process $\mathbf{Y}(\mathbf{x}_i)$ at this location contribute to the calculation of the $\hat{\gamma}_{i,j},$ for $j \neq i$. Following Muller [1998], we take consideration of the covariance structure of the empirical semivariograms and propose two generalized least-squares methods. Both methods estimate the latent dimensions more accurate than the least-squares method.

The remainder of the chapter is organized as follows: Section 3.1 discusses the details of the generalized least-squares fitting of the semivariogram. Section 3.2 gives the algorithms for generalized least-squares estimation. Section 3.3 provides extended simulations to show the performance of the methods. Section 3.4 presents two real data applications.

3.1 Generalized Least-Squares Methods

The crucial step of the dimension expansion approach is the lasso-penalized least-squares method to estimate the latent dimensions. Ignoring the complex covariance structure of the $\hat{\gamma}_{i,j}$ produces inefficient parameter estimation as demonstrated in

Muller [1998]. For the dimension expansion method, the generalized least-squares method is more appropriate for learning the latent dimensions. In the following, we introduce the generalized least-squares fitting criterion. First, we define an upper triangular matrix U of the form

$$U = \begin{cases} \hat{\gamma}_{i,j}, & \text{for } i \leq j, \\ 0, & \text{for } i > j. \end{cases}$$

Let $\text{vec}\left(U^T\right)$ denote the vector formed by concatenating the columns of U^T. We define $W\left([\mathbf{X}, \mathbf{Z}]\right)$ be the vector form of the distance matrix of $d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)$, for $i \leq j$. The lasso-penalized generalized least-squares criterion (GLS) is defined as follows

$$\begin{aligned} \left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{\text{GLS}} = & \underset{\boldsymbol{\phi}, \mathbf{Z}}{\text{argmin}} \left(\text{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right)^T \left[\text{cov}\left(\text{vec}\left(U^T\right)\right)\right]^{-1} \\ & \left(\text{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right) + \lambda \sum_{k=1}^{p} \left\|\mathbf{Z}_{.k}\right\|_1, \end{aligned}$$

where $\text{cov}\left(\text{vec}\left(U^T\right)\right)$ is the covariance matrix of $\hat{\gamma}_{i,j}$, for $i \leq j$, and $\gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)$ is a parametric stationary semivariogram model with parameter $\boldsymbol{\phi}$. We propose the generalized least-squares method based on Cressie [1985] to estimate the latent dimensions. In the following, we assume that $\mathbf{Y}(\mathbf{x})$ is a mean-zero Gaussian process. For demonstration, we use the exponential semivariogram model throughout our

implementations. For given $\boldsymbol{\phi}$, Cressie [1985] showed that

$$\operatorname{cov}\left(\hat{\gamma}_{i,j}, \hat{\gamma}_{i',j'}\right) = \frac{1}{2}\left[\gamma_{\boldsymbol{\phi}}\left(d_{j,i'}\right) + \gamma_{\boldsymbol{\phi}}\left(d_{i,j'}\right) - \gamma_{\boldsymbol{\phi}}\left(d_{i,i'}\right) - \gamma_{\boldsymbol{\phi}}\left(d_{j,j'}\right)\right]^2,$$

where, for example, $d_{j,i'}$ is the Euclidean distance between locations $[\mathbf{x}_j, \mathbf{z}_j]$ and $[\mathbf{x}_{i'}, \mathbf{z}_{i'}]$. As an illustration example, for a region with 4 locations, the upper triangular matrix U is defined as

$$\begin{pmatrix} \hat{\gamma}_{1,1} & \hat{\gamma}_{1,2} & \hat{\gamma}_{1,3} & \hat{\gamma}_{1,4} \\ 0 & \hat{\gamma}_{2,2} & \hat{\gamma}_{2,3} & \hat{\gamma}_{2,4} \\ 0 & 0 & \hat{\gamma}_{3,3} & \hat{\gamma}_{3,4} \\ 0 & 0 & 0 & \hat{\gamma}_{4,4} \end{pmatrix}.$$

Then $\operatorname{cov}\left(\operatorname{vec}\left(U^T\right)\right)$ is a 10×10 matrix where, for example, the covariance between $\hat{\gamma}_{1,2}$ and $\hat{\gamma}_{3,4}$ is estimated by

$$\operatorname{cov}\left(\hat{\gamma}_{1,2}, \hat{\gamma}_{3,4}\right) = \frac{1}{2}\left[\gamma_{\boldsymbol{\phi}}\left(d_{2,3}\right) + \gamma_{\boldsymbol{\phi}}\left(d_{1,4}\right) - \gamma_{\boldsymbol{\phi}}\left(d_{1,3}\right) - \gamma_{\boldsymbol{\phi}}\left(d_{2,4}\right)\right]^2. \tag{3.2}$$

However, the Euclidean distances in (3.2) are calculated based on the known dimensions for stationary processes. The distances used in dimension expansion method are based on the latent dimensions. Therefore, the covariance between $\hat{\gamma}_{i,j}$ and $\hat{\gamma}_{i',j'}$ with latent dimensions becomes

$$\begin{aligned} \operatorname{cov}(\hat{\gamma}_{i,j}, \hat{\gamma}_{i',j'}) = \frac{1}{2}\Big[&\gamma_{\boldsymbol{\phi}}\left(d_{j,i'}\left([\mathbf{X}, \mathbf{Z}]\right)\right) + \gamma_{\boldsymbol{\phi}}\left(d_{i,j'}\left([\mathbf{X}, \mathbf{Z}]\right)\right) \\ &-\gamma_{\boldsymbol{\phi}}\left(d_{i,i'}\left([\mathbf{X}, \mathbf{Z}]\right)\right) - \gamma_{\boldsymbol{\phi}}\left(d_{j,j'}\left([\mathbf{X}, \mathbf{Z}]\right)\right)\Big]^2. \end{aligned} \tag{3.3}$$

We propose the lasso-penalized generalized least-squares criterion as following:

$$\left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{\text{GLS}} = \underset{\phi,\mathbf{Z}}{\operatorname{argmin}} \left(\operatorname{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right)^T \hat{\Sigma}^{-1} \left(\operatorname{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right) + \lambda \sum_{k=1}^{p} \left\|\mathbf{Z}_{.k}\right\|_1, \tag{3.4}$$

where the entries in $\hat{\Sigma}$ is obtained from (3.3), and $\left\|\mathbf{Z}_{.k}\right\|_1$ has the same definition as in (3.1).

3.2 Algorithm

There exists a technique issue in the implementation of the above generalized least-squares fitting. The estimated covariance matrix $\hat{\Sigma}$ may not be positive definite, and/or not be invertible. There are mainly two methods to this end. One popular method is to perform the eigen-decomposition first. Then set the smallest eigenvalue to be an arbitrary small number [Knol and Berge, 1989]. This method is intuitive, however, the choice of the fixed small value can be problematic. The other method proposed by Higham [2002] guarantees that the resulting matrix is the nearest positive definite matrix by convex analysis using the Frobenius distance. In this dissertation, we adopt Higham [2002]'s approach to find the nearest positive definite matrix of $\hat{\Sigma}$.

Because the estimated covariance matrix $\hat{\Sigma}$ depends on $\boldsymbol{\phi}$ and $\mathbf{Z}$, following Cressie

[1985], Genton [1998] and Muller [1998], we apply an iterative reweighing strategy to find the estimation in the lasso-penalized generalized least-squares fitting as following:

1. Empirically estimate semivariogram $\hat{\gamma}_{i,j}$ using (1.3).

2. Set $\left(\boldsymbol{\phi}^{(0)}, \mathbf{Z}^{(0)}\right)_{GLS} = \left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{OLS}$.

3. At the kth step, calculate all the entries for the estimated covariance matrix $\hat{\Sigma}^{(k)}$, for example,

$$\begin{aligned}\operatorname{cov}\left(\hat{\gamma}_{1,2}, \hat{\gamma}_{3,4}\right)^{(k)} = \frac{1}{2}\Big[&\gamma_{\boldsymbol{\phi}^{(k)}}\left(d_{j,i'}\left([\mathbf{X}, \mathbf{Z}^{(k)}]\right)\right) + \gamma_{\boldsymbol{\phi}^{(k)}}\left(d_{i,j'}\left([\mathbf{X}, \mathbf{Z}^{(k)}]\right)\right) \\ &-\gamma_{\boldsymbol{\phi}^{(k)}}\left(d_{i,i'}\left([\mathbf{X}, \mathbf{Z}^{(k)}]\right)\right) - \gamma_{\boldsymbol{\phi}^{(k)}}\left(d_{j,j'}\left([\mathbf{X}, \mathbf{Z}^{(k)}]\right)\right)\Big]^2.\end{aligned}$$

   Calculate the inverse matrix of $\hat{\Sigma}^{(k)}$. If $\hat{\Sigma}^{(k)}$ is not invertible, use R function $nearPD$ (Higham, 2002) to obtain its nearest positive definite matrix.

4. Update $\left(\boldsymbol{\phi}^{(k+1)}, \mathbf{Z}^{(k+1)}\right)_{GLS}$ by the BFGS method [Broyden, 1979]:

$$\begin{aligned}\left(\boldsymbol{\phi}^{(k+1)}, \mathbf{Z}^{(k+1)}\right)_{GLS} = \underset{\boldsymbol{\phi}, \mathbf{Z}}{\operatorname{argmin}} &\left(\operatorname{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right)^T \left(\hat{\Sigma}^{-1}\right)^{(k)} \\ &\left(\operatorname{vec}\left(U^T\right) - \gamma_{\boldsymbol{\phi}}\left(W\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right) + \lambda \sum_{k=1}^{p} \left\|\mathbf{Z}_{\cdot k}\right\|_1.\end{aligned}$$

5. Stop if $\left\|\boldsymbol{\phi}^{(k+1)} - \boldsymbol{\phi}^{(k)}\right\|_1 + \left\|\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}\right\|_1 \leq \delta$, otherwise repeat Steps 3 and 4. In practice, we choose a small number for δ.

The simulations in the next section show that the generalized least-squares method improves accuracy. However, its computational complexity increases exponentially with a growing number of locations. Computations of the nearest positive definite matrix of $\hat{\Sigma}$ and its inverse are both computational intensive. Based on our extensive empirical study, we find that the proposed generalized least-squares criterion is not efficient when the number of locations is over 30. Therefore, we introduce an alternative method which is more computationally efficient than the GLS. We adopt the simplified covariance structure for computational simplicity which assumes the off-diagonal elements for the covariance matrix of $\hat{\gamma}_{i,j}$ are zero [Cressie, 1985]. Moreover, for a Gaussian random process $\{\mathbf{Y}(\mathbf{x}) : \mathbf{x} \in \mathcal{S}\}$, $\mathcal{S} \in \mathcal{R}^d$, $\text{var}(\hat{\gamma}_{i,j}) \simeq 2\gamma_{\boldsymbol{\phi}}^2(\|\mathbf{x}_i - \mathbf{x}_j\|)$ [Cressie, 1985]. Accordingly, we propose the lasso-penalized weighted least-squares criterion (WLS) as follows

$$\left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{WLS} = \underset{\boldsymbol{\phi},\mathbf{Z}}{\text{argmin}} \sum_{i<j} \frac{1}{\gamma_{\boldsymbol{\phi}}^2\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)} \left\{\hat{\gamma}_{i,j} - \gamma_{\boldsymbol{\phi}}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right\}^2 + \lambda \sum_{k=1}^{p} \|\mathbf{Z}_{.k}\|_1 . \tag{3.5}$$

A similar iterative reweighing algorithm to the GLS method is adopted to the WLS method. Essentially, at the kth iteration, the estimation $\left(\boldsymbol{\phi}^{(k-1)}, \mathbf{Z}^{(k-1)}\right)$ from the $(k-1)$th step are used for the weights $\gamma_{\phi}^{-2}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)$. Our empirical study suggests that its computational time is comparable with the OLS in Bornn et al. [2012].

The simulations in the next section show that this method also increases estimation accuracy compared to the OLS. We recommend this WLS method when the number of locations is more than 30.

The dimension expansion methods involve the unknown latent dimensions of the monitored locations. Due to this special feature, we propose a modified leave-one-out cross-validation method for choosing the tuning parameter λ. Here, the leave-out method means leaving the locations out. As we mentioned earlier, the observed Gaussian process $\mathbf{Y}(\mathbf{x}_i)$ at the location $\mathbf{x}_i$, $i = 1, ..., n$, contributes to obtain all of the $\hat{\gamma}_{i,j}$, for $j \neq i$. When we take the location $\mathbf{x}_i$ out, we need to predict $n-1$ semivariograms related to the location $\mathbf{x}_i$. The other problem is how to find the latent dimensions for the location $\mathbf{x}_i$. We use the thin-plate spline method to predict the latent dimensions for the location $\mathbf{x}_i$. We propose the modified Root Mean Squared Error for cross-validation ($MRMSE_{CV}$) as follows

$$MRMSE_{CV} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\hat{\gamma}_{i,j} - \gamma^*_{\hat{\boldsymbol{\phi}}_{-i}} \left(d_{i,j} \left([\mathbf{X}, \mathbf{Z}^*] \right) \right) \right)^2}, \tag{3.6}$$

where $\gamma^*_{\hat{\boldsymbol{\phi}}_{-i}} \left(d_{i,j} \left([\mathbf{X}, \mathbf{Z}^*] \right) \right)$ is the predicted semivariogram for the location $\mathbf{x}_i$. Note that $\mathbf{Z}^* = (\mathbf{z}_1, ..., \mathbf{z}_{i-1}, \mathbf{z}^*_i, \mathbf{z}_{i+1}, ..., \mathbf{z}_n)$, where $\mathbf{z}^*_i$ are the predicted latent dimensions for the location $\mathbf{x}_i$ using the thin-plate spline method. The algorithm for determining the tuning parameter λ is given in the following:

1. Choose a set of $\{\lambda_1, ..., \lambda_m\}$, for example, $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$.

2. For each λ, apply the dimension expansion method to obtain $\left(\hat{\boldsymbol{\phi}}_{-i}, \mathbf{Z}_{-i}\right)$ for the locations $1, ..., i-1, i+1, ..., n$ by taking the location $\mathbf{x}_i$ out.

3. Use the thin-plate spline method to find the function $f(\cdot)$ such that $f(\mathbf{X}_{-i}) = \mathbf{Z}_{-i}$. Predict $\mathbf{z}_i^* = f(\mathbf{x}_i)$ and obtain $d_{i,j}\left([\mathbf{X}, \mathbf{Z}^*]\right)$.

4. Obtain $\gamma^*_{\hat{\boldsymbol{\phi}}_{-i}}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}^*]\right)\right)$ using the distances $d_{i,j}\left([\mathbf{X}, \mathbf{Z}^*]\right)$ in Step 3 and $\hat{\boldsymbol{\phi}}_{-i}$ in Step 2. Calculate $MRMSE_{CV}$.

5. Choose λ corresponding to the smallest $MRMSE_{CV}$.

3.3 Simulation Studies

In this section, we consider the illustrative simulation example in Bornn et al. [2012]. The locations are simulated on a three-dimensional half-ellipsoid centered at $(0, 0, 0)$ and the projection of the first two dimensions is a disk centered at the origin. At each location, 1000 realizations of the Gaussian process $\mathbf{Y}(\mathbf{x})$ are simulated. Figure 3.1 shows the empirical semivariograms for the three-dimensional space and its projected two-dimensional space for $n = 30$. The red solid lines are fitted exponential semivariograms.

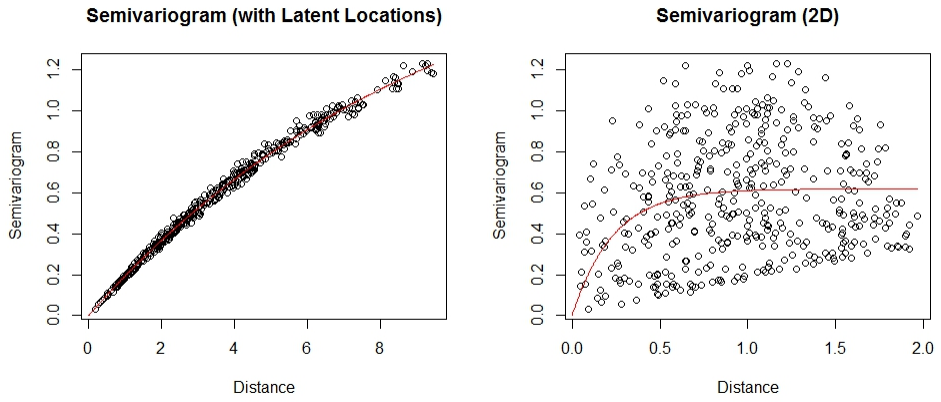

Figure 3.1: Empirical semivariogram plots of the original three dimensional space(left) and a two-dimensional projection (right)

In Figure 3.1, on the left side, the plot is the semivariograms versus the Euclidean distances based original three dimensions. On the three dimensional space, the simulated Gaussian field is stationary. The right side is the plot of semivariograms vs. Euclidean distances based on two dimensions. The red line is the fitted exponential semivariogram. The field is nonstationary with one dimension hidden. In Bornn et al. [2012], the tuning parameter λ is chosen to be 0.1 which induces that the dimension of $\mathbf{Z}$ is one. They recovered the latent dimension successfully resulting in a semivariogram that is close to the original (Figure 3.3). The contour plot of the original coordinates with the learned dimension by Bornn et al. [2012] is shown in Figure 3.2.

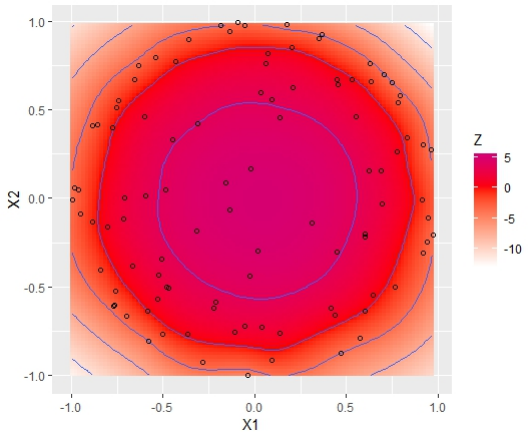

Figure 3.2: The origin coordinate with the learned dimension by OLS

Then we apply the proposed methods to find the latent dimension. The semivariogram plots with the learned dimension using three different methods are shown in Figure 3.3. The tuning parameter λ chosen for the WLS and the GLS are respectively 0.010 and 0.013 by using $MRMSE_{CV}$ in (3.6). In Figure 3.3, we see that all of the three methods can recover the true distances well. We plot the learned distances $\hat{d}_{i,j}$ among the locations with the true distances $d_{i,j}$ in Figure 3.4. The distance plots show that both GLS and WLS methods recover the locations better than the OLS method. The red line is 45 degrees from the origin. The closer of the plots to the red line, the better the learned distances $\hat{d}_{i,j}$ to the true distances $d_{i,j}$. We can see from these plots that the points in OLS depart away from the red line, while the points in GLS and WLS follow the red line closely.

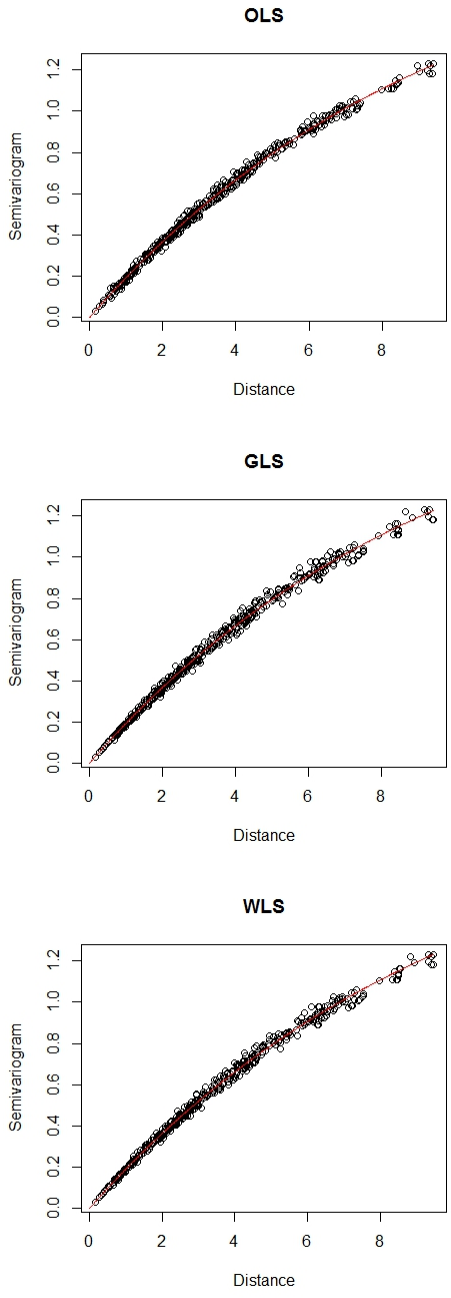

Figure 3.3: Semivariogram with learned locations

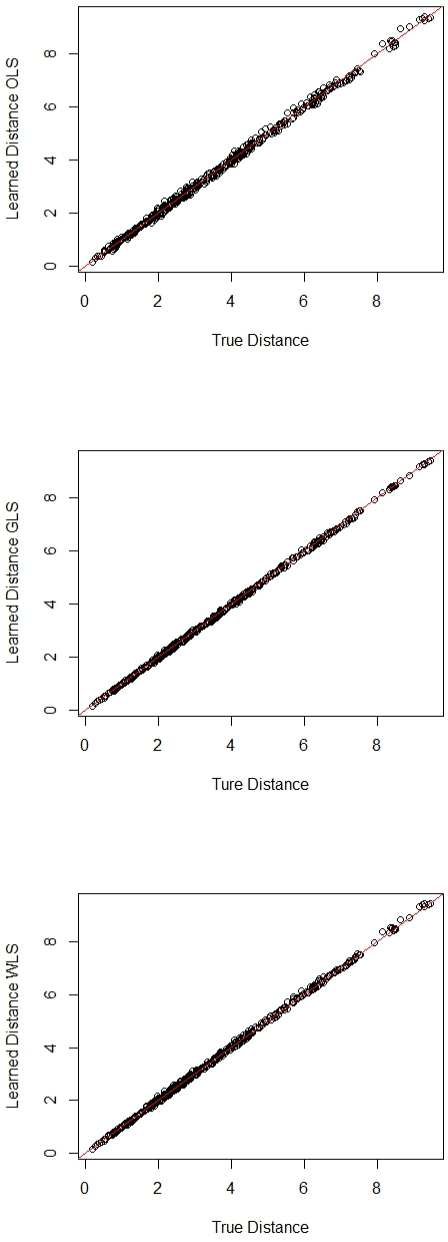

Figure 3.4: Distance plot

Next, we conduct simulations to assess the performance of the two proposed methods numerically and demonstrate the benefit of considering the covariance structure of $\hat{\gamma}_{i,j}$. In the simulations, the locations are simulated on a three-dimensional half-ellipsoid centered at $(0, 0, 0)$ and the projection of the first two dimensions is a disk centered at the origin for different numbers of locations $n = 10, 15,$ and 50. At each location, 1000 realizations of the Gaussian process $\mathbf{Y}(\mathbf{x})$ are simulated. The Sum of Squared Errors (SSE) between the true distances and the learned distances are computed to compare these three methods, i.e.

$$SSE = \sum_{i<j} (d_{i,j}[\mathbf{X}, \mathbf{Z}] - \hat{d}_{i,j}[\mathbf{X}, \mathbf{Z}])^2.$$

The boxplots of SSE based on 1000 replications are shown in Figure 3.5 for $n = 10, 15$. Throughout the simulations, we show that both of the proposed methods are better than the OLS method for $n = 10, 15$. Moreover, when the number of locations is larger than 30, we conduct some simulations to compare WLS and OLS. For $n = 50$, the results of 1000 replications are shown in Figure 3.6. Table 3.2 is the mean and standard deviation of SSE for both methods which show that WLS is better than OLS.

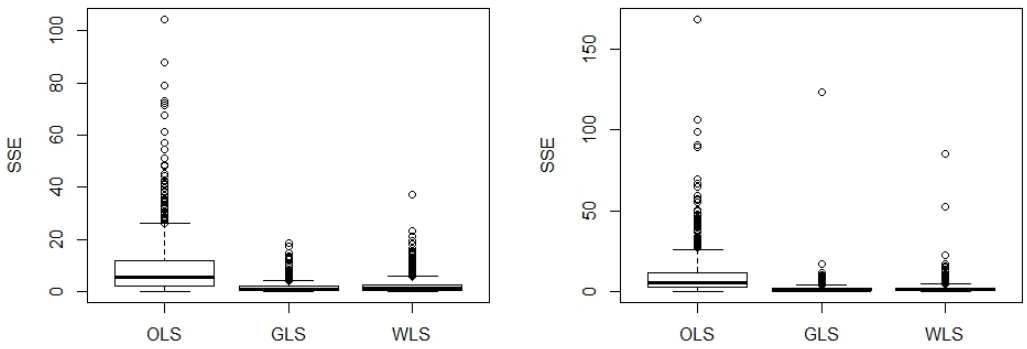

Figure 3.5: Boxplot of the SSE for $n = 10$ (left) and $n = 15$(right)

Table 3.1: Mean and standard deviation of SSE for OLS, GLS and WLS

	OLS	GLS	WLS
$n = 10$	8.84(10.89)	1.58(2.08)	2.26(3.17)
$n = 15$	9.36(12.31)	1.61(4.25)	2.06(4.01)

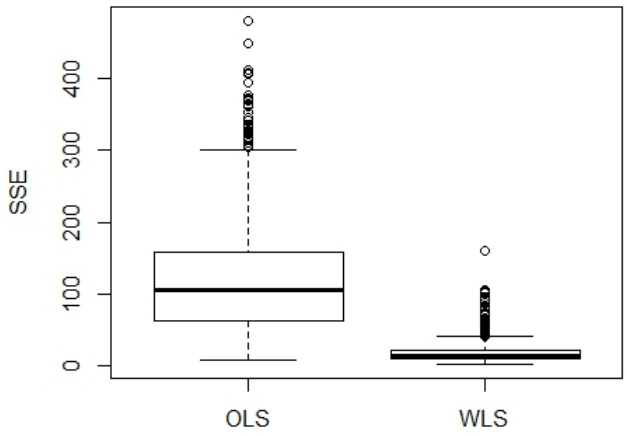

Figure 3.6: Boxplot of the SSE for $n = 50$

Table 3.2: Mean and standard deviation of SSE for OLS and WLS

	OLS	WLS
$n = 50$	120.41(79.09)	17.96(15.52)

3.4 Real Data Applications

3.4.1 Solar Radiation Data

The solar radiation data is obtained from the solar radiation monitoring network in southwestern British Columbia, Canada [Hay, 1984]. It is the daily solar radiation totals for the years 1980 to 1983 at 12 locations. The field is known to be nonstationary because of the location and elevation of Station 1 on Grouse mountain. The non-stationarity of the data was well studied in [Sampson and Guttorp, 1992] and Bornn et al. [2012]. Figure 3.7 is the plot of the empirical semivariograms versus the original locations. The points associated with Station 1 are marked as "x" in the plot. Bornn et al. [2012] uncovered the latent dimensions and through their approach, the semivariogram is closer to stationary. In Figure 3.8, the fitted exponential semivariogram is shown by the solid red line. As studied in Bornn et al. [2012], with $\lambda = 0.2$, they added two more latent dimensions. The result is shown in Figure 3.8 below. Station 1 is pushed further away with the latent dimensions and the field is closer to stationary.

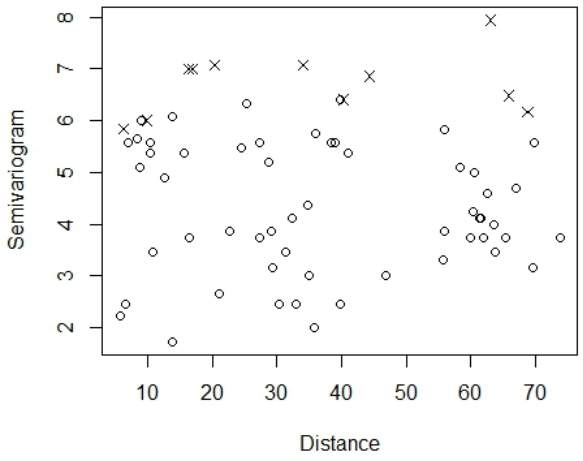

Figure 3.7: Semivariogram of the original locations

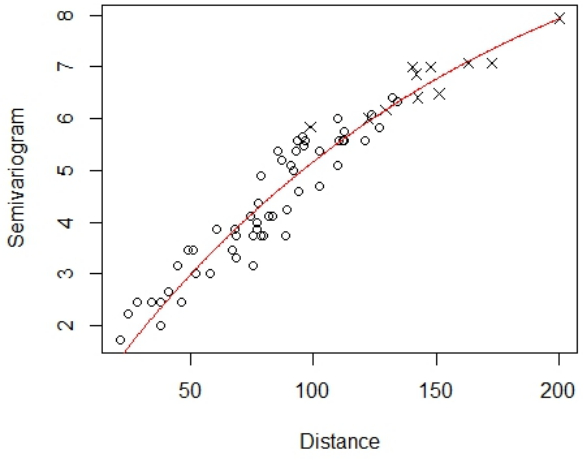

Figure 3.8: Semivariogram with learned dimensions by OLS

Now we use two proposed methods on the solar radiation data. Using the modified $MRMSE_{CV}$ in (3.6) as a criterion, we choose $\lambda = 0.014$ for WLS and $\lambda = 0.011$ for GLS. These tuning parameter values for WLS and GLS methods expand the original space to the one with two more dimensions. The results are shown in Figure 3.9.

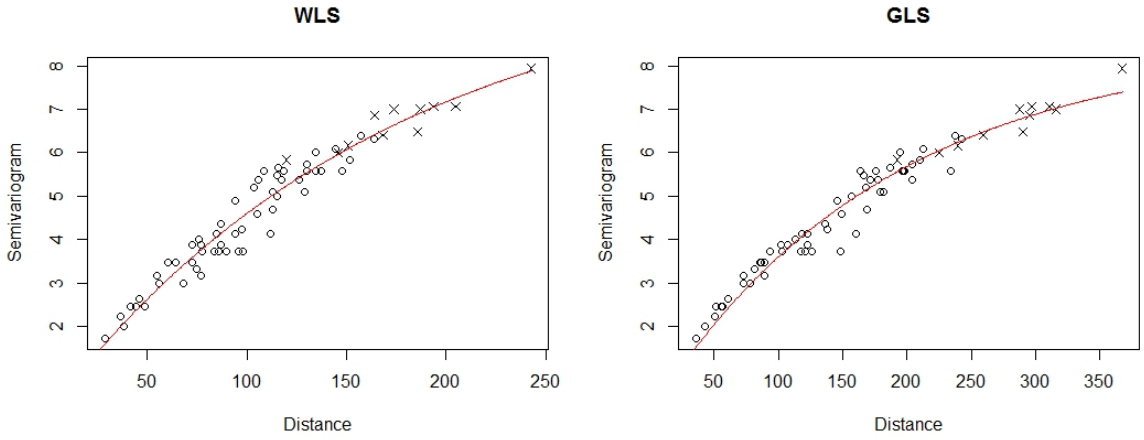

Figure 3.9: Semivariogram with learned dimensions by WLS and GLS

The field obtains stationarity by the proposed methods better than OLS. Station 1 is pushed even further with the latent dimensions, the distances of other locations are changed accordingly. We compare the fitting by SSE between the empirical semivariograms and the fitted parametric semivariograms for three methods. The SSE of OLS, WLS, and GLS are respectively 12.09, 11.45, and 10.21.

3.4.2 PM2.5 Data

Now we present the application on the logarithm of the daily average PM2.5 data set collected in 37 stations covering the Province Ontario in a region with longitude from $-74°$ to $-90°$ and latitude from $41°$ to $49°$ in Canada from 2003 to 2016. The data is obtained from the Ontario air quality archive (*www.airqualityontario.com*). Locations of the stations are shown in Figure 3.10.

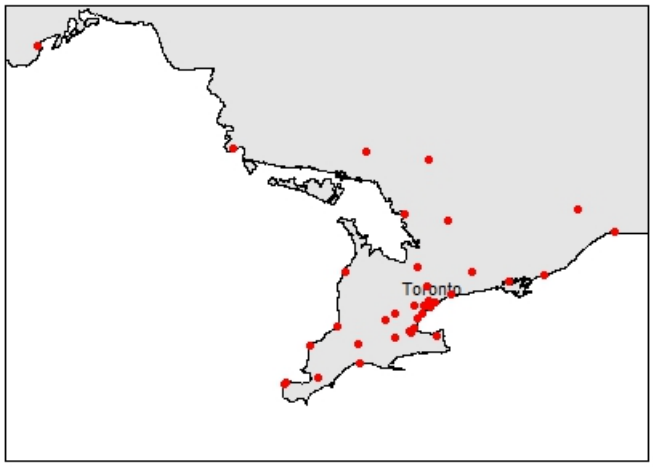

Figure 3.10: Monitoring stations of the PM2.5 data

The dimensions, longitude, latitude, and elevation are recorded for the stations. From the empirical plot of the semivariogram, Figure 3.11, we observe that the

nonstationary is mainly caused by the stations marked by "x". The stations marked as "x" in the graph are Barrie, Brampton, Belleville, Brantford, and Burlington. Interestingly, these are all the stations starting with the letter "B" monitored by the Ontario air quality archive.

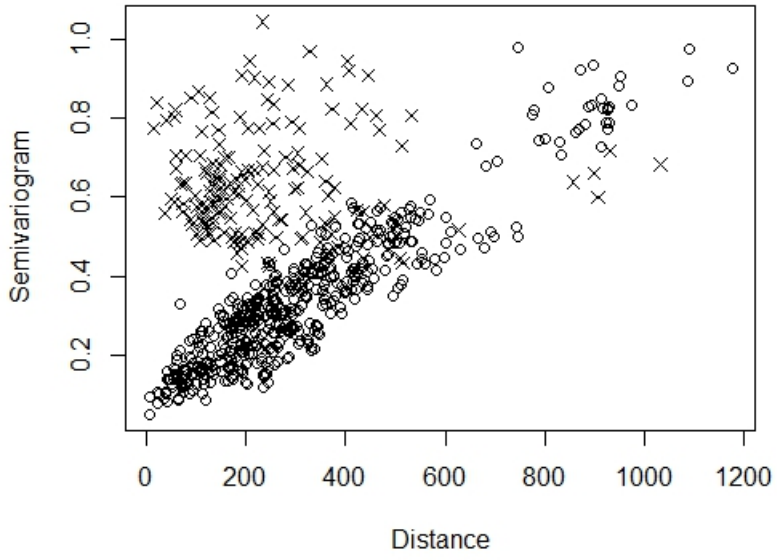

Figure 3.11: Semivariogram of the original locations

First, we choose λ for the OLS method. The leave-one-out cross-validation finds $\lambda = 0.0009$ with the smallest $MRMSE_{CV}$ and the OLS method expands two more latent dimensions. The semivariograms versus the distance with the learned two latent dimensions are shown in Figure 3.12. Next, we apply the WLS method on

the PM2.5 data. We choose $\lambda = 0.0005$ and two latent dimensions are expanded to the original space. The result is shown in Figure 3.12. All of the points are moved closer to the fitted line by using the WLS methods. The SSE between the empirical semivariograms and the fitted parametric semivariograms for OLS is 2.62 and WLS is 1.56.

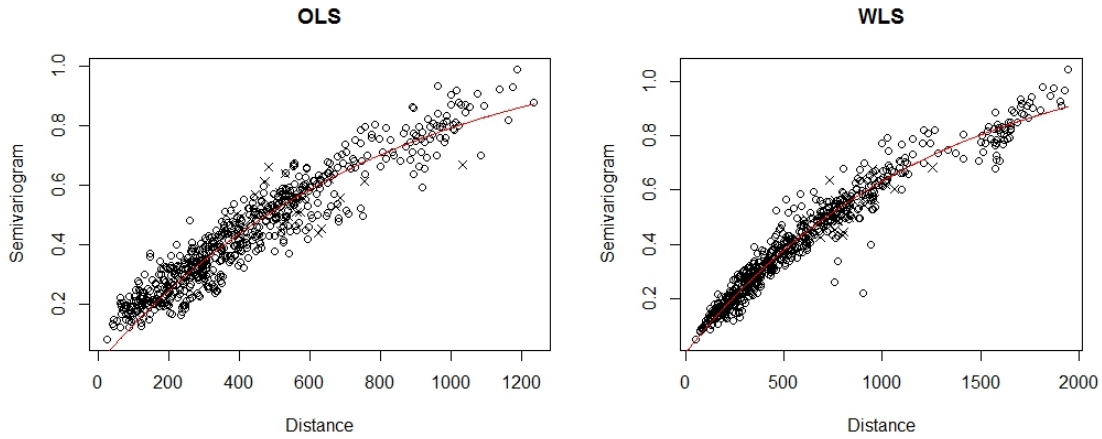

Figure 3.12: Semivariogram with two learned dimensions by OLS and WLS

4 Hierarchical Bayesian Spatio-temporal Modeling via Dimension Expansion

Ozone concentrations are the daily maximum 8-hours moving averages of hourly ozone concentration data recorded in micrograms per cubic meter, $\mu g/m^3$, which are key indicators of air quality. Monitoring the changes both spatially and temporally is very important for the assessment of air quality change, which has a great impact on our environment, society, and economy. However, modeling the ozone concentrations is not an easy task since the ozone concentrations vary over space and time with complicated spatial structures, temporal structures, and spatio-temporal interactions. Furthermore, the presence of missing data which is common at the gauged stations brings even more difficulties. Jin et al. [2012] studied the ozone concentrations within $-79°$ to $-81.5°$ longitude and $39.5°$ to $41.5°$ latitude around the Pittsburgh region ($-79.23°$, $43.39°$). All of the gauged stations have missing data in this region. They dealt with the missing problems in two steps. First, some

of the missing measurements were filled by using linear models so that the missing data had the staircase pattern. Second, they applied the hierarchical Bayesian spatio-temporal modeling on this staircase of missing data to estimate the parameters of the spatio-temporal model. They estimated the spatial correlation function for the gauged stations based on the estimations from the previous step. Next, They estimated the covariance matrix for all of the stations, then derived the predictive distribution for the ungauged sites.

In terms of the covariance matrix for all of the stations, they selected the generalized linear model with quasi-Poisson family as an appropriated spatial correlation function by examing the pattern of the plot of spatial correlations based on the hierarchical model. The generalized linear model with quasi-Poisson family is not appropriate if there exists negative correlations. This is a strong restriction because negative correlations are common for the ozone concentration data. Moreover, choosing models by exploring the observed plots is not appropriate method and may cause overfitting to the observed data. The model may be only suitable just for a particular kind of data.

In this section, we propose a method to estimate the covariance matrix through dimension expansion for modeling the semivariograms in nonstationary fields based on the estimations from the hierarchical Bayesian spatio-temporal modeling. For

demonstration, we apply the proposed method to the same data in Jin et al. [2012]. The proposed method is more general than the one used in Jin et al. [2012]. Using the covariance matrix estimated by the proposed method on the entropy criterion in the environmental network design problem, our study provides interesting findings and the locations of the selected ungauged stations are more reasonable. We also evaluate the method and compare it with Jin et al. [2012] by leave-one-out cross-validation. The results show that the proposed method provides slightly better prediction.

The chapter is arranged as follows. First, we describe the ozone concentrations in the Pittsburgh region and apply the techniques for filling missing data following Jin et al. [2012]. Then, we introduce the method to estimate the covariance matrix through dimension expansion method for modeling the semivariograms in nonstationary fields. Next, we derive spatial predictive distributions on the ungauged sites using the covariance matrix estimated by the proposed method. We also present the result of extending an environmental network. Last, we provide the model evaluation through leave-one-out cross-validation. The review of the hierarchical Bayesian spatio-temporal modeling technique [Le and Zidek, 2006] is given in Jin et al. [2012].

4.1 Ozone Concentration Data

The ozone concentrations are recorded within -79° to -81.5° longitude and 39.5° to 41.5° latitude around the Pittsburgh region for four consecutive summer months, June, July, August, and September, over the period from 1995 to 2007. There are 25 gauged stations in the region as shown in Figure 4.1. The original data set Y_0 has 25 stations and 1586 (13 years $\times$ 122 days) measurements at each station. The number of missing data in Y_0 is shown by N1.Miss in Table 4.1. We follow the steps in Jin et al. (2012) to fill some of the missing data for each station within the period of monitoring blocks using the same regression model as follows

$$\begin{aligned} y_{122(i-1)+j} &= a\sin\left(\frac{2(122(i-1)+j)\pi}{122}\right) + b\cos\left(\frac{2(122(i-1)+j)\pi}{122}\right) + c_i + \varepsilon_{122(i-1)+j} \\ &= a\sin\left(\frac{j\pi}{61}\right) + b\cos\left(\frac{j\pi}{61}\right) + c_i + \varepsilon_{122(i-1)+j}, \end{aligned} \tag{4.1}$$

for $i = 1, \cdots, 13$, and $j = 1, \cdots, 122$, where a and b are regression coefficients, c_i are the categorical factors, and $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of independently and identically distributed Gaussian random variables with mean 0 and variance σ^2. The model (4.1) assigns different means to the years with a yearly cycle of 122 days. We reexpress the 13 factors in the model via Helmert contrasts, which compare the first level of the factor with all later levels, the second level with all later levels, and so forth.

The Helmert matrix, $Z_{13\times 13}$, is defined as follows

$$Z = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 & -1 \\ 1 & 1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & 2 & \cdots & -1 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 11 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 12 \end{pmatrix}.$$

Let X, the matrix of covariates, be

$$X = \begin{pmatrix} S & Z \otimes \mathbf{1}_{122} \end{pmatrix}_{1586\times 15}, \tag{4.2}$$

where $\mathbf{1}_n = (1, 1, \cdots, 1, 1)^T_{1\times n}$ and

$$S = \begin{pmatrix} \sin(\pi/61) & \cdots & \sin(i\pi/61) & \cdots & \sin(1586\pi/61) \\ \cos(\pi/61) & \cdots & \cos(i\pi/61) & \cdots & \cos(1586\pi/61) \end{pmatrix}^T_{2\times 1586},$$

and let $\mathbf{y} = (y_1, y_2, \cdots, y_{1586})^T$, $\boldsymbol{\beta} = (a, b, d_1, \cdots, d_{13})^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{1586})^T$ denote the response variables, regression coefficient vector and error variables, respectively. The model (4.1) is written as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

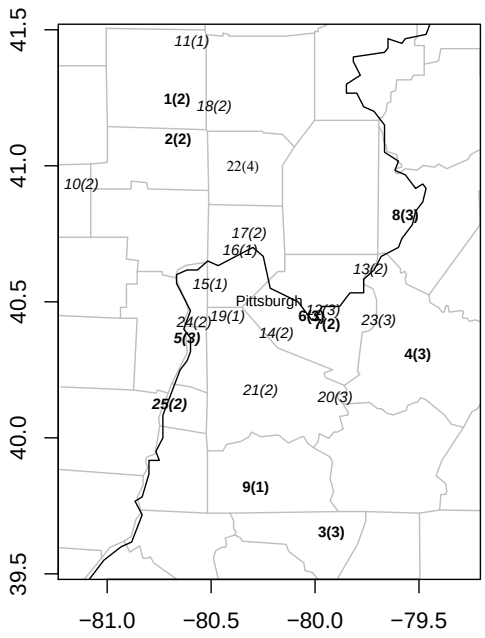

Figure 4.1: Monitoring stations in the Pittsburgh region

We fill in the missing data within the blocks by the least-squares predictions plus errors. Then we obtain a new data set Y_1. The number of missing data in Y_1 is shown in Table 4.1 by N2.Miss. Next, we follow the steps for filling the missing data by using the hierarchical Bayesian spatio-temporal modeling technique for the staircase pattern of missing data. We obtain a new data set Y_2 from Y_1 by filling in the 488 missing data at Station 5 and 25 during the end of the period of study. N3.Miss in Table 4.1 shows the number of missing data in the data set Y_2. Y_2 has a

staircase data structure because all of the missing data are located at the beginning of the period of study. Now we can use the hierarchical Bayesian model [Jin et al., 2012] to model Y_1 with the staircase structure and estimate the hyperparameters $H_g = \{V_B,\ B_0,\ (\Upsilon_{01}, H_1, \Lambda_1, \delta_1),\ \cdots,\ (\Upsilon_{0,k-1}, H_{k-1}, \Lambda_{k-1}, \delta_{k-1}),\ (\Lambda_7, \delta_7)\}$ by the EM algorithm. We put $d = 4$, $l = 15$, $n = 1586$, $k = 7$, $m_1 = 854$, $m_2 = 610$, $m_3 = 488$, $m_4 = 366$, $m_5 = 318$, $m_6 = 244$, $m_7 = 0$, $g_1 = 1$, $g_2 = 1$, $g_3 = 0$, $g_4 = 3$, $g_5 = 1$, $g_6 = 1$and $g_7 = 16$.

Table 4.1: Location of the stations and the number of missing data

ID	Class	Lon	Lat	N1.Miss	N2.Miss	N3.Miss	ID	Class	Lon	Lat	N1.Miss	N2.Miss	N3.Miss
1	2	-40.24	80.66	855	854	854	14	2	-40.38	80.18	22	0	0
2	2	-41.09	80.65	610	610	610	15	1	-40.56	80.50	13	0	0
3	3	-39.64	79.92	618	610	610	16	1	-40.68	80.35	11	0	0
4	3	-40.30	79.50	488	488	488	17	2	-40.74	80.31	4	0	0
5	3	-40.36	80.61	858	854	366	18	2	-41.21	80.48	5	0	0
6	3	-40.44	80.01	370	366	366	19	1	-40.44	80.42	16	0	0
7	2	-40.41	79.94	370	366	366	20	3	-40.14	79.90	3	0	0
8	3	-40.81	79.56	328	318	318	21	2	-40.17	80.26	1	0	0
9	1	-39.81	80.28	278	244	244	22	4	-40.99	80.34	0	0	0
10	2	-40.93	81.12	12	0	0	23	3	-40.42	79.69	5	0	0
11	1	-41.45	80.59	1	0	0	24	2	-40.42	80.58	5	0	0
12	3	-40.46	79.96	2	0	0	25	2	-40.12	80.69	488	488	0
13	2	-40.61	79.73	8	0	0							

The numbers 1, 2, 3, and 4 under Class denote agricultural, residential, commercial and industrial, respectively.

4.2 Covariance Matrix Estimation

The 100 grid boxes of a spatial resolution of latitude 0.2°× longitude 0.2° cover the Pittsburgh region. The grid points and their classes are displayed in Figure 4.4. The next task is to derive the predictive distribution for these grid points. The key step is to estimate the covariance matrix. Now, we introduce the method to estimate the covariance matrix through dimension expansion method for modeling the semivariograms in nonstationary fields based on the estimations from the hierarchical Bayesian spatio-temporal modeling. Let $\{\mathbf{Y}(\mathbf{x}) \ : \ \mathbf{x} \in \mathcal{S}\}$, $\mathcal{S} \in \mathcal{R}^d$, be an environmental random process, where $\mathbf{x}$ is a d-dimensional spatial index that varies continuously throughout the region $\mathcal{S}$. At n spatial locations denoted by $\{\mathbf{x}_i : i = 1, ..., n\}$, we observe realizations of the random process $\mathbf{Y}(\mathbf{x})$, ie., $\{\mathbf{Y}(\mathbf{x}_i) : i = 1, ..., n\}$. We are interested in learning the spatial dependency of the process through the observed data. Semivariogram function that describes the degree of spatial dependence of an intrinsic stationary random process is a cornerstone in spatial statistics. An intrinsic stationary random process satisfies the following two conditions [Cressie, 1993]:

1. $\mathbb{E}(\mathbf{Y}(\mathbf{x})) = \mu$, for $\mathbf{x} \in \mathcal{S}$,

2. $\text{var}(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)) = 2\gamma(\mathbf{x}_i - \mathbf{x}_j)$,

where semivariogram is defined as $\gamma(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{2}\text{var}\left(\mathbf{Y}(\mathbf{x}_i) - \mathbf{Y}(\mathbf{x}_j)\right)$ for two different locations, $\mathbf{x}_i$ and $\mathbf{x}_j$, in the monitored region. The estimated covariance matrix of the gauged stations $\hat{\Sigma}^{[g,g]}$ is obtained from the estimate of H_g. We estimated the semivariograms for the gauged stations $\mathbf{g}_i$ and $\mathbf{g}_j$, correspondingly, by

$$\hat{\gamma}(\mathbf{g}_i - \mathbf{g}_j) = \frac{1}{2}\hat{\text{var}}(\mathbf{Y}(\mathbf{g}_i)) + \frac{1}{2}\hat{\text{var}}(\mathbf{Y}(\mathbf{g}_j)) - \hat{\text{cov}}\left(\mathbf{Y}(\mathbf{g}_i), \mathbf{Y}(\mathbf{g}_j)\right). \tag{4.3}$$

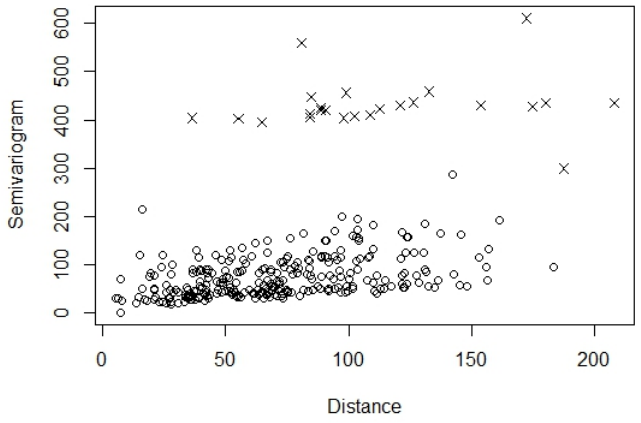

Figure 4.2: Semivariogram plot

In Figure 4.2, we notice that the estimated semivariograms related to Station 3 (marked by "x") are much higher than the other stations. We examine the location of Station 3 and notice that it is on the edge of the monitored region. Moreover, there is an airport close to this station. According to Xue et al. [1994], there was a great impact of high altitude aircraft on the ozone layer in the stratosphere. This becomes

an influential factor in modeling the ozone concentrations. Next, we introduce how this factor is considered in the modeling technique.

For a nonstationary field, Bornn et al. [2012] proposed a novel approach to finding latent dimensions over which the nonstationary fields exhibit stationarity through dimension expansion. They justified that for a nonstationary Gaussian process $\mathbf{Y}(\mathbf{x})$, where $\mathbf{x} \in R^d$, there exist extra dimensions $\mathbf{z} \in R^p$, $p > 0$, such that the expanded process $\mathbf{Y}([\mathbf{x}, \mathbf{z}])$ was stationary under appropriate moment constraints. Note that $[\mathbf{x}, \mathbf{z}]$ is the concatenation of the dimensions $\mathbf{x}$ and $\mathbf{z}$. The stationary semivariogram with latent dimensions can be expressed by

$$2\gamma([\mathbf{x}_i, \mathbf{z}_i] - [\mathbf{x}_j, \mathbf{z}_j]) = \mathbb{E}\left(\mathbf{Y}([\mathbf{x}_i, \mathbf{z}_i]) - \mathbf{Y}([\mathbf{x}_j, \mathbf{z}_j])\right)^2,$$

where $[\mathbf{x}_i, \mathbf{z}_i]$ is the expanded spatial index for the ith location. In Chapter 3, we improved the dimension expansion method by considering the covariance structure of the $\hat{\gamma}_{i,j},$ for $j \neq i$. In the data application, we adopt the lasso-penalized weighted least-squares criterion (WLS) in Chapter 3 to estimate the parameters and learn the latent dimensions as follows

$$\left(\hat{\boldsymbol{\phi}}, \mathbf{Z}\right)_{WLS} = \underset{\boldsymbol{\phi}, \mathbf{Z}}{\operatorname{argmin}} \sum_{j<i} \frac{1}{\gamma^2_{\boldsymbol{\phi}}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)} \left\{\hat{\gamma}_{i,j} - \gamma_{\boldsymbol{\phi}}\left(d_{i,j}\left([\mathbf{X}, \mathbf{Z}]\right)\right)\right\}^2 + \lambda \sum_{k=1}^{p} \left\|\mathbf{Z}_{\cdot k}\right\|_1. \tag{4.4}$$

The semivariogram plot with estimated expanded dimensions (Figure 4.3) of the gauged stations shows that the field is close to be stationary. Two extra dimensions

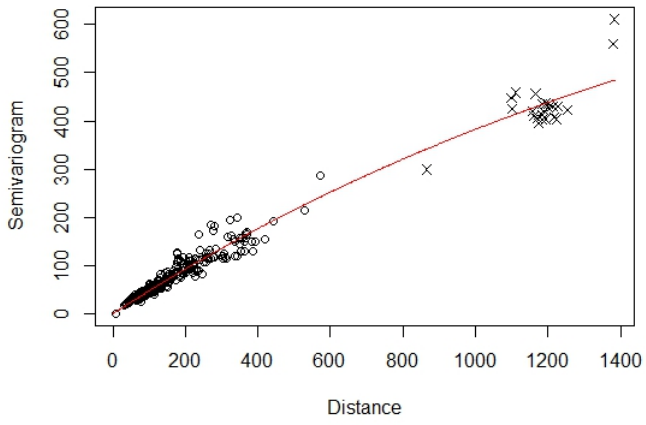

Figure 4.3: Semivariogram with learned dimensions

are added to the original coordinate with $\lambda = 0.01$. Station 3 is pushed much further with the latent dimensions. After the expanded dimensions for the gauged stations are obtained, we use the thin-plate spline method [Wabba and Wendelberger, 1980] to estimate the hidden dimensions for the ungauged sites. The semivariograms for the ungauged stations are estimated by the exponential model using the estimated parameters $\hat{\boldsymbol{\phi}}$. Next, we estimate the semivariograms γ_{s_i,s_j} between stations s_i and s_j using the exponential model based on the distances over the space composed by the original and the latent dimensions. Last, we estimate the covariance between any two sites by

$$\hat{\Sigma}_{i,j} = \hat{\text{cov}}(Y(s_i), Y(s_j)) = \frac{1}{2}\hat{\sigma}_{Y(s_i)} + \frac{1}{2}\hat{\sigma}_{Y(s_j)} - \hat{\gamma}_{s_i,s_j},$$

where $\hat{\sigma}_{Y(s_i)}$ and $\hat{\sigma}_{Y(s_j)}$ are estimates of $\sigma_{Y(s_i)}$ and $\sigma_{Y(s_j)}$ obtained by the thin-plate spline approach. Then we estimate the hyperparameters associated with the grid points $\Lambda_0,\ \tau_{00},\ H_0$ and δ_0 via

$$\hat{\delta}_0 = \frac{\hat{\hat{\delta}}_1 + \cdots + \hat{\hat{\delta}}_k}{k}, \quad \hat{H}_0 = \hat{\Lambda}^{[1,\cdots,k]}, \quad \hat{\tau}_{00} = (\hat{\Sigma}^{[g,g]})^{-1}\hat{\Sigma}^{[g,u]},$$

$$\hat{\Lambda}_0 = \frac{\hat{\hat{\delta}}_0 - u - 1}{1 + tr(\hat{\Sigma}^{[g,g]}\hat{H}_0)}(\hat{\Sigma}^{[u,u]} - \hat{\tau}_{00}^T\hat{\Sigma}^{[g,g]}\hat{\tau}_{00}),$$

where

$$\hat{\Lambda}^{[j,\cdots,k]} = \begin{pmatrix} \hat{\Lambda}_j + \hat{\tau}_{0j}^T\hat{\Lambda}^{[j+1,\cdots,k]}\hat{\tau}_{0j} & \hat{\tau}_{0j}^T\hat{\Lambda}^{[j+1,\cdots,k]} \\ \hat{\Lambda}^{[j+1,\cdots,k]}\hat{\tau}_{0j} & \hat{\Lambda}^{[j+1,\cdots,k]} \end{pmatrix} \quad , \quad j = 1, \cdots, k-1,$$

and $\hat{\Lambda}^{[k]} = \hat{\Lambda}_k$. After all of the hyperparameters in the predictive distribution are estimated, we can predict the daily ozone concentration at all the ungauged sites within the period of study by generating samples from the predictive distribution. Spatial predictive distribution at the ungauged sites is defined as follows

$$(Y^{[u]}|Y^{[g]}, H) \sim t_{n\times u}(\mu^{u|g}, \frac{\Phi^{[u|g]} \otimes \Psi^{[u|g]}}{\delta_0^*}, \delta_0^*) \tag{4.5}$$

where $\delta_0^* = \delta_0 - u + 1$, $\Psi^{[u|g]} = \Lambda_0$ $\mu^{[u|g]} = ZB_0^{[u]} + (Y^{[g]} - ZB_0^{[g]})\tau_{00}$ and $\Phi^{[u|g]} = I_n + XF^{-1}X' + (Y^{[g]} - XB_0^{[g]})H_0(Y^{[g]} - XB_0^{[g]})^T$.

4.3 Environmental Network Extension

Assume that Y has the density function f. The total reduction in uncertainty of Y can be presented by the entropy of its distribution; i.e., $H(Y) = -E[\log f(Y)/h(Y)]$, where $h(\cdot)$ is a not necessarily integrable reference density [Jaynes, 1963]. According to the predictive distribution (4.5), the total entropy $H(Y^{[u]}|Y^{[g]})$ can be defined as

$$H(Y^{[u]}|Y^{[g]}) = \frac{1}{2}\log|\Psi^{[u|g]}| + c_u(u,q), \tag{4.6}$$

where $c_u(u,q)$ is a constant depending on the degree of freedom and the dimension of the ungauged sites.

The key step in expanding an environmental network is to find appropriate ungauged sites to add to the existing network that maximizes the corresponding entropy. The optimality criterion is defined as

$$\max_{add}(\frac{1}{2}\log|\Psi^{[u|g]}|)^{add}. \tag{4.7}$$

The *add* sites, a vector of dimension u_1, are selected to maximize the entropy in (4.6). In Jin et al. [2012], the grid points $\{91, 92, 93\}$ are selected with the highest entropy 11.3774. Using the covariance matrix estimated by the proposed method, the grid points $\{41, 71, 100\}$ are selected with entropy 12.1207. This selection is more reasonable as they scatter in the region and are not crowded in the corner like $\{91, 92, 93\}$. The selected sites among 100 grid points by two methods are shown in

Figure 4.4 below.

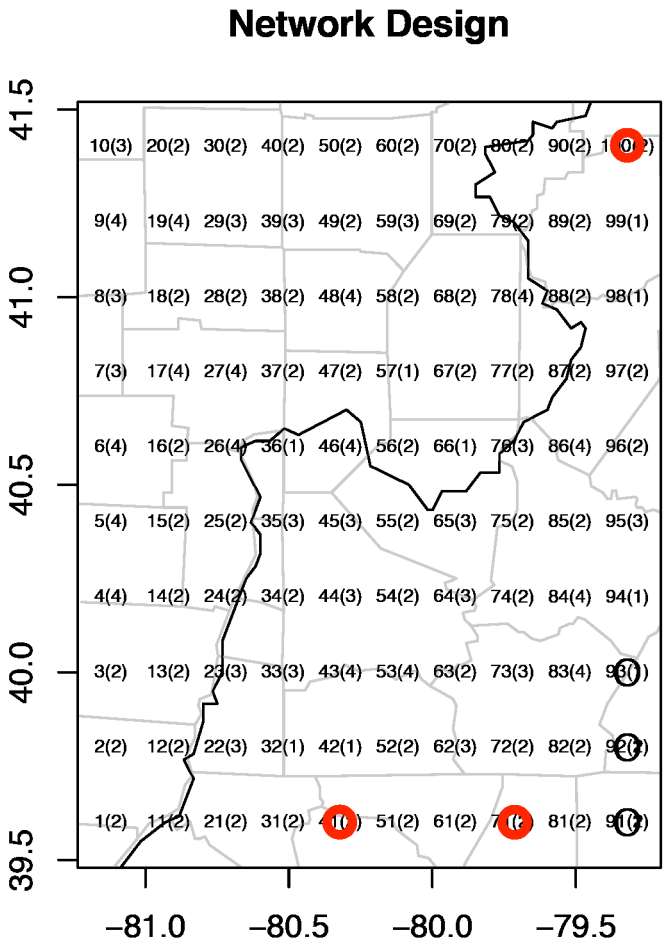

Figure 4.4: The selected sites among 100 grid points (black circled points by Jin et al. (2012), red circled points by our method)

4.4 Model Evaluation

In this section, we use the leave-one-out cross-validation to evaluate the proposed method. And we compare the proposed method with Jin et al. [2012]. We select the observations from one of the original 25 stations as the validation data, and observations in the remaining 24 stations are treated as the training data. We use the data from day 855 to 1586 from each station to evaluate the prediction because none of the stations has missing data during this period. By choosing this period, we avoid using the Bayesian hierarchical modeling technique for estimating the missing data in the training data set, which is time-consuming and not our intention for evaluating the proposed method on estimating the covariance matrix. Station 22 is excluded because it is the only industrial station in the study. For each of the 24 stations, we generate 100 samples from the predictive distribution with parameters estimated using observations from 24 stations. We compute the average of relative absolute bias (ARAB) as $\sum_{j=1}^{100} |(y_{j,i,t} - y_{i,t})/y_{i,t}|$, where $y_{j,i,t}$ is the sample generated from the predictive distributions and $y_{i,t}$ is the observation from Station i on time t. The results are given in Table 4.2.

Table 4.2: Mean and standard deviation of the average of relative absolute bias

ID	Our Method	Jin et al. (2012)	ID	Our Method	Jin et al. (2012)
1	0.0789(0.0627)	0.8134(0.0682)	13	0.1145(0.1096)	0.2003(0.1769)
2	0.1206(0.1356)	0.1221(0.1121)	14	0.1361(0.1732)	0.2211(0.2283)
3	0.8517(0.8517)	0.1572(0.1572)	15	0.1911(0.2052)	-
4	0.1756(0.1693)	-	16	0.1189(0.1179)	0.1285(0.1161)
5	0.1575(0.1731)	0.1986(0.1855)	17	0.1496(0.1594)	0.1669(0.1727)
6	0.1336(0.1513)	0.1477(0.1667)	18	0.1253(0.1154)	0.1256(0.1372)
7	0.1265(0.1563)	0.1456(0.1732)	19	0.1369(0.1272)	0.1026(0.0994)
8	0.0968(0.0804)	0.1135(0.1023)	20	0.1603(0.1598)	0.1310(0.1134)
9	0.1497(0.1104)	0.1619(0.1208)	21	0.1351(0.1154)	0.1274(0.1123)
10	0.1589(0.1796)	-	23	0.1617(0.1858)	-
11	0.6913(0.6455)	-	24	0.1286(0.1051)	-
12	0.1406(0.1409)	0.1265(0.1416)	25	0.1583(0.1701)	0.1722(0.1675)

In Table 4.2, "-" means that there is no prediction for the station because there are negative correlations and the method in Jin et al. [2012] fails to estimate the predictive distribution. The results in Table 4.2 also show that the proposed method provides slightly more accurate predictions for most of the stations. More important

is that, when there are negative correlations estimated by the hierarchical Bayesian spatio-temporal modeling technique, Jin et al. [2012] fails to estimate the covariance matrix, while the proposed method still provides accurate predictions except for Station 3. This is expected because Station 3 is an influential station as we exam the semivariograms over the expanded space. When we use observations at Station 3 as validation data, it has a great impact on estimating the covariance matrix.

5 Detection of Change Points in Spatio-temporal Data in the Presence of Outliers and Heavy-tailed Observations

Recently, Wu et al. [2017] proposed a general spatio-temporal autoregressive (GSTAR) model which takes into account the effect of station surroundings, seasonality, temporal correlation among observations at the same spatial location and spatial correlation among observations from different spatial locations. The model is multi-functional since it can also be used to detect new influences that largely affected the measurements in the treatment area compared to the control area. However, their method is dependent on the normality assumption.

As the spatio-temporal data is usually observed over a large area and in many years, undetectable outliers can easily occur unexpectedly in any day for any small area because of measurement error or other reasons. The parameter estimation method given in Wu et al. [2017] may not be stable or robust. There is a great need

to develop a parameter estimation method for the GSTAR model that is resistant to outliers and stable concerning heavy-tail distributed errors. In the development of such robust methods, M-estimation can play important and complementary roles. Thus we modify the EM-type algorithm which is given in Wu et al. [2017] by replacing the least-squares (LS) estimation by M-estimation, which is more robust in estimating parameters in the presence of outliers and/or heavy-tailed observations [Huber, 1973]. We name the modified EM-type algorithm as the $_M$EM-type algorithm. We also modify their change-point detection procedure accordingly, which is more accurate in detecting change points in the presence of outliers and/or heavy-tailed observations.

The outline of this chapter is the following. In Section 5.1, a general spatio-temporal autoregressive model is reviewed and the $_M$EM-type algorithm is presented. Then we describe the procedure for detecting change points in the treatment area via the GSTAR models. In Sections 5.2 and 5.2.2, two real data applications and related simulations are given to compare the $_M$EM-type algorithm with the original one and to compare both change-point detection procedures.

5.1 The GSTAR Model-based Procedure of Change-point Detection

In this section, we first review a specially designed EM-type algorithm to estimate the model parameters. We then give a change-point detection procedure based on the GSTAR model.

The GSTAR model in Wu et al. [2017] was given in Chapter 1. The model takes into account the effect of station surroundings, seasonality, temporal correlation among observations at the same spatial location, and spatial correlation among observations from different spatial locations while allowing the coefficients to vary over time. The GSTAR model is defined as follows

$$\begin{aligned} y_{i,T(k-1)+t} &= \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma} + c_i + \rho\sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} \qquad (5.1) \\ &- \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma} - c_l) + \varepsilon_{i,T(k-1)+t}. \end{aligned}$$

The notation in the model is explained as follows

- $y_{i,T(k-1)+t}$ is the spatio-temporal variable of interest observed at spatial location i on t^{th} day in the k^{th} year.

- $t \in \mathcal{S}$ with $\mathcal{S}$ being a set of consecutive days in a year with size T. For example, $\mathcal{S}$ could be a number of consecutive months in a year or a whole year.

- c_i's are the effects of location types taking values in $\{\tau_1, \ldots, \tau_\kappa\}$ according to different kinds of surrounding areas around the locations.

- $W = (w_{il})_{L\times L}$ is a neighbourhood matrix to describe the spatial correlation among observations collected from different spatial locations, which satisfies the conditions that $w_{il} \geq 0, w_{ii} = 0$ and $\sum_{l=1}^{L} w_{il} = 1$. The entry w_{il} of the neighbourhood matrix W represents the degree of correlation between observations collected at the spatial locations i and l, which may be chosen to be dependent on the distance between the spatial locations i and l. ρ is the spatial autoregressive parameter.

- $\boldsymbol{x}_{T(k-1)+t} = (x_{T(k-1)+t,1},\ x_{T(k-1)+t,2},\ x_{T(k-1)+t,3})'$ are explanatory variables. $x_{T(k-1)+t,1} = 1$ for all $t \in \mathcal{S}$, $(x_{T(k-1)+t,2},\ x_{T(k-1)+t,3})' = (\sin(t_j\pi/s_j),\ \cos(t_j\pi/s_j))'$ for $t \in \mathcal{S}_j$ are designed to model the seasonal cyclicities. Here $\mathcal{S}_j$, $j = 1, \ldots, J$, are J seasons in $\mathcal{S}$ with $\mathcal{S} = \cup\mathcal{S}_j$, and s_j is the number of days in the j^{th} season for $j = 1, \ldots, J$, and t_j is the number of days of t in $\mathcal{S}_j$ if t falls into the j^{th} season.

- $\boldsymbol{\beta}_{T(k-1)+t} = (\beta_{0,k,j},\ \beta_{1,k,j},\ \beta_2)'$ are regression coefficients when t falls into the j^{th} season. Note that both $\{\beta_{0,k,j}\}$ and $\{\beta_{1,k,j}\}$ vary with seasons and years.

- $\tilde{\boldsymbol{y}}_{i,T(k-1)+t} = (y_{i,T(k-1)+t-1},\ y_{i,T(k-1)+t-2}, \ldots,\ y_{i,T(k-1)+t-\iota})'$ and $\boldsymbol{\gamma} = (\gamma_1,\ \gamma_2, \ldots,\ \gamma_\iota)'$.

An autoregression term is included in the model to take into account the possible autocorrelation among observations at each location. Here ι denotes the number of autoregression terms in the model which is pre-determined in this dissertation, but may be chosen by an order selection.

The parameter set to be estimated in model (5.1) is $\mathcal{H} = \{\beta_{0,k,j},\ \beta_{1,k,j},\ j = 1,\ldots,J,\ k = 1,\ 2,\ \ldots,\ K;\ \beta_2,\ \boldsymbol{\gamma},\ \tau_1,\ \ldots,\ \tau_\kappa,\ \rho,\ \sigma^2\}$.

5.1.1 The Estimation

M-estimation is a maximum likelihood type estimation [Huber, 1973]. In the development of robust methods, M-estimation can play an important and complementary role. The well-known dispersion function for the M-estimation is the Huber's function defined as the following:

$$H(x) = \begin{cases} x^2, & if\ \ |x| \le k, \\ 2k\,|x| - k^2, & if\ \ |x| > k, \end{cases}$$

where k is a tuning constant, and usually chosen as 1.345. The EM-type algorithm given in Wu et al. [2017] used the least-squares technique. The performance of the LS estimation relies heavily on the normality assumption on the errors. Because of the complexity of spatio-temporal data, the normality assumption is easily violated in

the presence of undetectable outliers and/or heavy-tailed observations. We propose to modify it by replacing the LS technique used in the algorithm by M-estimation for estimating the GSTAR model parameters, which is more stable regardless if there are outliers and/or heavy-tailed observations in the data set.

First, we give the initial values to $\{\beta_{0,k,j},\ \beta_{1,k,j},\ j\ =\ 1,\ldots,J,\ k\ =\ 1,\ 2,\ \ldots,\ K;$ $\beta_2,\ \boldsymbol{\gamma},\ \tau_1,\ \ldots,\ \tau_\kappa,\ \rho\}$. We then carry out the following steps:

1. We calculate the mean of the available observations for each type of stations and denote them by $a_1, a_2, \ldots, a_\kappa$. We then calculate the overall mean of the available observations and denote it by a. The initial estimates of τ_q's are thus put as $\tau_q^{(0)} = a_q - a$, $q = 1, \ldots, \kappa$. Let $\bar{c} = \sum_{i=1}^{L} c_i/L$. The initial estimate of $\bar{c}$ can be obtained by $\bar{c}^{(0)} = \sum_{i=1}^{L} c_i^{(0)}/L$, where $c_i^{(0)}$ takes values in $\{\tau_1^{(0)}, \ldots, \tau_\kappa^{(0)}\}$ according to different kinds of surrounding areas around the location.

2. By averaging all equations in (5.1), we obtain that

$$\begin{aligned} \bar{y}_{T(k-1)+t} &= \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}_{T(k-1)+t} + \bar{\tilde{\boldsymbol{y}}}'_{T(k-1)+t}\boldsymbol{\gamma} + \bar{c} + \epsilon_{T(k-1)+t}, \\ &= \beta_{0,k,j} + \beta_{1,k,j}x_{T(k-1)+t,2} + \beta_2 x_{T(k-1)+t,3} + \gamma_1 \bar{y}_{T(k-1)+t-1} \\ &\quad + \ldots + \gamma_o \bar{y}_{T(k-1)+t-o} + \bar{c} + \epsilon_{T(k-1)+t}, \end{aligned} \tag{5.2}$$

where $\bar{y}_{T(k-1)+t}$ is the average of the observations on the $(T(k-1)+t)^{th}$ day of all spatial locations after removing all missing observations, $\bar{\tilde{\boldsymbol{y}}}_{T(k-1)+t} =$

$(\bar{y}_{T(k-1)+t-1}, \ldots, \bar{y}_{T(k-1)+t-o})'$ and $\epsilon_{T(k-1)+t} = \frac{1}{L}\boldsymbol{\ell}_L'(\mathcal{I}_L - \rho W)^{-1}\,\varepsilon_{T(k-1)+t}$, in which $\boldsymbol{\ell}_L = (1, 1, \ldots, 1)'_{L\times 1}$.

3. Since $\sin(\pi - \theta) = \sin(\theta)$, $\sin(\pi + \theta) = \sin(2\pi - \theta)$, $\cos(\pi - \theta) = -\cos(\theta)$, and $\cos(\pi + \theta) = -\cos(\theta)$, we can remove both the constant term and the term related to $\beta_{1,k,j}$ by the difference between two properly chosen pair of the equations given in (5.2). By doing so, we obtain

$$y^{(1)}_{T_1(k-1)+t} = \beta_2 y^{(2)}_{T_1(k-1)+t} + \boldsymbol{\gamma}\tilde{\boldsymbol{y}}^{(3)}_{T_1(k-1)+t} + \tilde{\epsilon}_{T_1(k-1)+t}, \quad t \in \mathcal{S}^{(1)}. \quad (5.3)$$

(A specific example of how to calculate $y^{(1)}_{T_1(k-1)+t}$, $y^{(2)}_{T_1(k-1)+t}$, $\tilde{\boldsymbol{y}}^{(3)}_{T_1(k-1)+t}$, $\tilde{\epsilon}_{T_1(k-1)+t}$, and $\mathcal{S}^{(1)}$ are given in the Wu et al. [2017].)

Denote $\boldsymbol{y}^{(1)} = (y^{(1)}_1, y^{(1)}_2, \ldots, y^{(1)}_{T_1K})'$, $\boldsymbol{y}^{(2)} = (y^{(2)}_1, y^{(2)}_2, \ldots, y^{(2)}_{T_1K})'$, and $\tilde{\boldsymbol{y}}^{(3)} = (\tilde{\boldsymbol{y}}^{(3)}_1, \tilde{\boldsymbol{y}}^{(3)}_2, \ldots, \tilde{\boldsymbol{y}}^{(3)}_{T_1K})'$. The M-estimates of β_2 and $\boldsymbol{\gamma}$ are given by

$$\arg\min_{\beta_2,\boldsymbol{\gamma}} H(\boldsymbol{y}^{(1)} - \beta_2\boldsymbol{y}^{(2)} - \boldsymbol{\gamma}\tilde{\boldsymbol{y}}^{(3)}),$$

which are used as the initial estimate $\beta_2^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ of β_2 and $\boldsymbol{\gamma}$ respectively.

4. We substitute β_2 and $\boldsymbol{\gamma}$ by $\beta_2^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ in model (5.2). For each year k and season j, we denote $\boldsymbol{y}^{(1)}_j = (\bar{y}_{T(k-1)+t} - \beta_2^{(0)}x_{T(k-1)+t,3} - \boldsymbol{\gamma}^{(0)}\bar{\tilde{\boldsymbol{y}}}_{T(k-1)+t} - \bar{c}^{(0)},\ t \in \mathcal{S}_j)'$, and $\boldsymbol{y}^{(2)}_j = (x_{T(k-1)+t,2},\ t \in \mathcal{S}_j)'$. We derive the M-estimates of $\beta_{0,k,j}, \beta_{1,k,j}$ for season j of the k^{th} year by

$$\arg\min_{\beta_{0,k,j},\beta_{1,k,j}} H(\boldsymbol{y}^{(1)}_j - \beta^j_{0,k}\boldsymbol{\ell}_{s_j} - \beta^j_{1,k}\boldsymbol{y}^{(2)}_j)$$

for $j = 1, \ldots, J$ respectively, where $\boldsymbol{\ell}_{s_j} = (1, 1, \ldots, 1)'_{s_j \times 1}$. Therefore, we use these least-square estimates of $\beta_{0,k,j}$ and $\beta_{1,k,j}$ as the initial estimates $\beta^{(0)}_{0,k,j}$ and $\beta^{(0)}_{1,k,j}$.

5. Set the initial value of $\rho^{(0)}$ as 0.5.

Second, we provide the $_M$EM-type Algorithm. Let $\mathcal{H}^{(m-1)} = \{\beta^{(m-1)}_{0,k,j},\ \beta^{(m-1)}_{1,k,j},\ j = 1, \ldots, J,\ k = 1,\ 2,\ \ldots,\ K,\ \beta^{(m-1)}_2,\ \boldsymbol{\gamma}^{(m-1)},\ \tau^{(m-1)}_1,\ \ldots,\ \tau^{(m-1)}_\kappa,\ \rho^{(m-1)},\ \sigma^{2(m-1)}\}$ be the set of estimates we obtained after the $(m-1)^{th}$ iteration. The $_M$EM-type algorithm has the following three steps:

1. E-step: Estimate the observation $y_{i,T(k-1)+t}$ at the m^{th} iteration by the following conditional expectation

$$\begin{aligned} & y^{(m)}_{i,T(k-1)+t} \\ = \ & E\left(y_{i,T(k-1)+t} | y^{(m-1)}_{l,T(k-1)+t}, l = 1, 2, \ldots, L, \mathcal{H}^{(m-1)}\right) \\ = \ & \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}^{(m-1)}_{T(k-1)+t} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{(m-1)} + c^{(m-1)}_i + \rho^{(m-1)} \times \\ & \sum_{l:w_{il}\neq 0}\left(y^{(m-1)}_{l,T(k-1)+t} - \boldsymbol{x}'_{T(k-1)+t}\boldsymbol{\beta}^{(m-1)}_{T(k-1)+t} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{(m-1)} - c^{(m-1)}_l\right), \end{aligned}$$

   if it is missing.

2. M-step: Obtain the estimates $\boldsymbol{c}^{(m)},\ \sigma^{2(m)},\ \rho^{(m)},\ \beta^{(m)}_2,\ \boldsymbol{\gamma}^{(m)},\ \beta^{(m)}_{0,k,j},\ \beta^{(m)}_{1,k,j},\ j = 1, \ldots, J,\ k = 1,\ \ldots,\ K$ at the m^{th} iteration sequentially as follows

(a) First derive the estimates $\left\{\tau_1^{(m)},\ \ldots,\ \tau_\kappa^{(m)}\right\}$ in the same way as we obtained the estimates $\left\{\tau_1^{(0)},\ \ldots,\ \tau_\kappa^{(0)}\right\}$. Then $\boldsymbol{c}^{(m)} = (c_1^{(m)}, c_2^{(m)}, \ldots, c_L^{(m)})$, where $c_i^{(m)}$'s take values from $\{\tau_1^{(m)},\ \ldots,\ \tau_\kappa^{(m)}\}$ based on the types of the stations.

(b) Similarly, we can remove both the constant term and the term related to $\beta_{1,k,j}$ by the difference between one properly chosen pair of the equations given in (5.1). Then we estimate σ^2 as $\sigma^{2(m)}$ by sample variances.

(c) Find the M-estimates of ρ, β_2 and $\boldsymbol{\gamma}$ after substituting σ^2 by $\sigma^{2(m)}$ to get $\rho^{(m)}$, $\beta_2^{(m)}$ and $\boldsymbol{\gamma}^{(m)}$ respectively.

(d) Substitute the estimates $\left\{\boldsymbol{c}^{(m)},\ \rho^{(m)},\ \beta_2^{(m)}, \boldsymbol{\gamma}^{(m)}\right\}$ into model (5.1) to obtain the M-estimates of $\beta_{0,k,j},\ \beta_{1,k,j}$ as $\beta_{0,k,j}^{(m)},\ \beta_{1,k,j}^{(m)}$.

3. Keep repeating the steps 1-2 until $|\boldsymbol{\gamma}^{(m)} - \boldsymbol{\gamma}^{(m-1)}| < v$, $|\beta_2^{(m)} - \beta_2^{(m-1)}| < v$, $|\beta_{0,k,j}^{(m)} - \beta_{0,k,j}^{(m-1)}| < v$ and $|\beta_{1,k,j}^{(m)} - \beta_{1,k,j}^{(m-1)}| < v$ for all k and j, where v is a predetermined small value. Then we denote $\hat{\beta}_{0,k,j} = \beta_{0,k,j}^{(m)},\ \hat{\beta}_{1,k,j} = \beta_{1,k,j}^{(m)}$, for $j = 1, \ldots, J,\ k = 1,\ 2,\ \ldots,\ K$; $\hat{\beta}_2 = \beta_2^{(m)},\ \hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{(m)}$; $\hat{\tau}_i = \tau_i^{(m)}$, for $i = 1,\ \ldots,\ \kappa$; $\hat{\rho} = \hat{\rho}^{(m)}$, and $\hat{\sigma}^2 = \hat{\sigma}^{2(m)}$.

The set of estimates we obtained is $\hat{\mathcal{H}} = \{\hat{\beta}_{0,k,j},\ \hat{\beta}_{1,k,j},\ j = 1, \ldots, J,\ k = 1,\ 2,\ \ldots,\ K,\ \hat{\beta}_2,\ \hat{\boldsymbol{\gamma}},\ \hat{\tau}_1,\ \ldots,\ \hat{\tau}_\kappa,\ \hat{\rho},\ \hat{\sigma}^2\}$.

5.1.2 The Change-point Detection Procedure

We now introduce the procedure for detecting new influences that affected the measurements in the treatment area substantially by comparing with that in the control area, which is similar to the one given in Wu et al. [2017]. We model the data collected respectively from the treatment and control areas of the region by two different GSTAR models using the algorithm proposed in the previous section. The main idea is that if new influences in the treatment area are not negligible, there should be detectable changes in the time-dependent regression coefficients in the GSTAR model for that area compared to those in the GSTAR model for the control area. A change-point detection method can be applied to the differences in regression coefficient estimates from these two areas. The M-estimation-based change-point detection procedure is described below.

1. We group the stations in the treatment area of the region into group 1 and model the spatio-temporal data collected at these stations by

$$\begin{aligned} y_{i,T(k-1)+t} &= \beta^{\mathrm{I}}_{0,k,j} + \beta^{\mathrm{I}}_{1,k,j} x_{T(k-1)+t,2} + \beta^{\mathrm{I}}_{2} x_{T(k-1)+t,3} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{I}} \\ &+ c_i + \rho^{\mathrm{I}} \sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} - \beta^{\mathrm{I}}_{0,k,j} - \beta^{\mathrm{I}}_{1,k,j} x_{T(k-1)+t,2} \\ &- \beta^{\mathrm{I}}_{2} x_{T(k-1)+t,3} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{I}} - c_l) + \varepsilon_{i,T(k-1)+t}. \end{aligned} \tag{5.4}$$

   Then we group the stations in the control area into group 2 and model the

data from these stations by

$$\begin{aligned} y_{i,T(k-1)+t} &= \beta^{\mathrm{II}}_{0,k,j} + \beta^{\mathrm{II}}_{1,k,j} x_{T(k-1)+t,2} + \beta^{\mathrm{II}}_{2} x_{T(k-1)+t,3} + \tilde{\boldsymbol{y}}'_{i,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{II}} \\ &+ c_i + \rho^{\mathrm{II}} \sum_{l=1}^{L} w_{il}(y_{l,T(k-1)+t} - \beta^{\mathrm{II}}_{0,k,j} - \beta^{\mathrm{II}}_{1,k,j} x_{T(k-1)+t,2} \\ &- \beta^{\mathrm{II}}_{2} x_{T(k-1)+t,3} - \tilde{\boldsymbol{y}}'_{l,T(k-1)+t}\boldsymbol{\gamma}^{\mathrm{II}} - c_l) + \varepsilon_{i,T(k-1)+t}. \end{aligned} \tag{5.5}$$

Note that these two models have different parameters except the effect of the station locations, c_i's.

2. First, we estimate the parameters as their initial values. Following the steps presented in section 3.2.1, we derive the station type effect $\left\{\tau_1^{(0)}, \ldots, \tau_\kappa^{(0)}\right\}$ using observations collected on stations from both groups so that the same type of stations in different groups have the same station type effect. Then, we obtain $\{\beta^{\mathrm{I}(0)}_{0,k,j},\ \beta^{\mathrm{I}(0)}_{1,k,j},\ j = 1,\ 2,\ 3,\ 4,\ k = 1,\ 2,\ \ldots,\ K, \beta^{\mathrm{I}(0)}_{2}, \boldsymbol{\gamma}^{\mathrm{I}(0)}\}$ and $\{\beta^{\mathrm{II}(0)}_{0,k,j},\ \beta^{\mathrm{II}(0)}_{1,k,j},\ j = 1,\ 2,\ 3,\ 4,\ k = 1,\ 2,\ \ldots,\ K, \beta^{\mathrm{II}(0)}_{2}, \boldsymbol{\gamma}^{\mathrm{II}(0)}\}$ for two groups of stations separately. We also set the initial values of ρ^{I} and ρ^{II}as $\rho^{\mathrm{I}(0)} = \rho^{\mathrm{II}(0)} = 0.5$.

3. We apply the $_M$EM-type algorithm proposed in section 5.1.1. In the E-step, the missing observations are filled up. In the M-step, we estimate the station type effects using data from all the stations, then estimate the other parameters sequentially for two groups of stations separately. These two steps are repeated

until convergence. We obtain the estimates $\hat{\beta}^{\mathrm{I}}_{0,k,j}, \hat{\beta}^{\mathrm{I}}_{1,k,j}$ for model (5.4) and $\hat{\beta}^{\mathrm{II}}_{0,k,j}, \hat{\beta}^{\mathrm{II}}_{1,k,j}$ for model (5.5).

4. We take the difference between these two sets of parameter estimates to obtain two sets of estimates $\{d_{0,k,j} = \hat{\beta}^{\mathrm{I}}_{0,k,j} - \hat{\beta}^{\mathrm{II}}_{0,k,j},\ j = 1,\ 2,\ 3,\ 4,\ k = 1,\ 2,\ \ldots,\ K\}$ as the difference in the intercepts of two models and $\{d_{1,k,j} = \hat{\beta}^{\mathrm{I}}_{1,k,j} - \hat{\beta}^{\mathrm{II}}_{1,k,j},\ j = 1,\ 2,\ 3,\ 4,\ k = 1,\ 2,\ \ldots,\ K\}$ as the difference in the slopes of two models. Then we apply the R package *changepoint* (Killick and Eckley 2014) to detect the possible mean shifts in $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$.

For convenience, we name the change-point detection procedure given in Wu et al. [2017] as the LS-based change-point detection procedure.

It is worth mentioning that in the above procedure $d_{0,k,j}$ and $d_{1,k,j}$ describe the effect after eliminating the effects of station types, the temporal correlation, the spatial correlation, and the randomness. Therefore, after applying the proposed procedure, the estimates $\{\hat{\beta}^{\mathrm{I}}_{0,k,j},\ \hat{\beta}^{\mathrm{I}}_{1,k,j}\}$ and $\{\hat{\beta}^{\mathrm{II}}_{0,k,j},\ \hat{\beta}^{\mathrm{II}}_{1,k,j}\}$ derived respectively from two groups of data should behave similarly if there are no new influences in the treatment area. Then there are no changes in the means of both $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$.

5.2 Data Applications

5.2.1 Ozone Concentration Data

In this section, we respectively compare the $_M$EM-type algorithm with the EM-type algorithm in Wu et al. [2017], and the M-estimation-based change-point detection procedure with the LS-based change-point detection procedure through an application and simulations on the ground-level ozone concentration data.

The data of Wu et al. [2017] includes measurements of the ground-level ozone concentration readings measured in parts per billion (ppb) from 37 monitoring stations in a region with longitude from $-80°$ to $-78.5°$ and latitude from $43°$ to $45°$ in southern Ontario over the period from 1988 to 2010. Locations of the stations are shown in Figure 5.1. Following Porter et al. [2001], the data used in the examples is the logarithm of the daily maximum 8-hour moving averages of ozone concentration. There are 36 stations. Among these 36 stations, we choose 27 stations which have been monitored for more than 5 years. On average, each station has 39.4% data missing. We let $\iota = 1$ by the pre-analysis of the data. The total number of the parameters is 194. First, we obtain the estimates of the parameters in the GSTAR model using the EM-type algorithm in Wu et al. [2017]. We name these estimates $\hat{\mathcal{H}}_{LS}$. Then the proposed $_M$EM-type algorithm is used to obtain the parameters in

GSTAR model on the same data set. We name these estimates $\hat{\mathcal{H}}_M$. We use the Euclidean distance to measure the differences as the following:

$$\left\|\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M\right\| = \sqrt{(\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M)'(\hat{\mathcal{H}}_{LS} - \hat{\mathcal{H}}_M)}.$$

The distance is 0.1015, which is small enough to show that these two methods produce almost the same parameter estimates on the same data set.

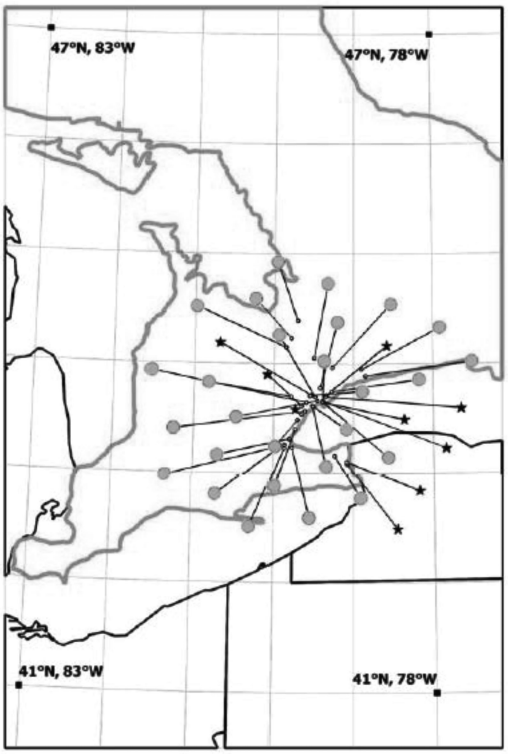

Figure 5.1: The locations of 27 stations which have data for more than 5 years are shown in circle. Data source: Regional Aquatics Monitoring Program http://www.ramp-alberta.org

Next, we show that the $_M$EM-type algorithm works well in the presence of outliers. To make the outliers reasonable, we first choose an area whose latitude is less than $43.55°N$. There are 8 stations within this area. Then we randomly pick up a day, for 9 days after this day, we expanded the log-transformed ozone concentrations by 1.5 times. In real life, this could happen for the reasons including the machine broken, unexpected activities in this area, etc. The experiment is repeated for 500 times, we recorded the Euclidean distance for both algorithms, in Table 5.1, the mean and the standard deviation (sd) of the Euclidean distance are reported. The simulation shows that outliers have less impact on the performance of parameter estimation if the proposed $_M$EM-type algorithm is used.

Table 5.1: Mean and standard deviation of the Euclidean distances

$_M$EM-type algorithm		EM-type algorithm	
mean	sd	mean	sd
0.2704	0.1325	0.4392	0.0541

Wu et al. [2017] simulated the change points under Scenario 1 in the following way. First, they separated the stations into two groups by the latitude 43.65°. Then, for each station in group 1, they added a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to each observation collected

from 1998 to 2010 to create the first change-point in 1998. They also added a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to the previously modified observations from 2008 to 2010 to create the second change-point in 2008. The results of detecting the change points by using the LS-based change-point detection procedure are shown in Figure 5.2. The right panel displays the results by using the M-estimation-based change-point detection procedure. Two sets of estimates, $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ are obtained. The plot displays the change points in $\{d_{0,k,j}\}$ (upper panel) and $\{d_{1,k,j}\}$ (lower panel) using both procedures. Figure 5.2 shows that both procedures capture the change points equally well.

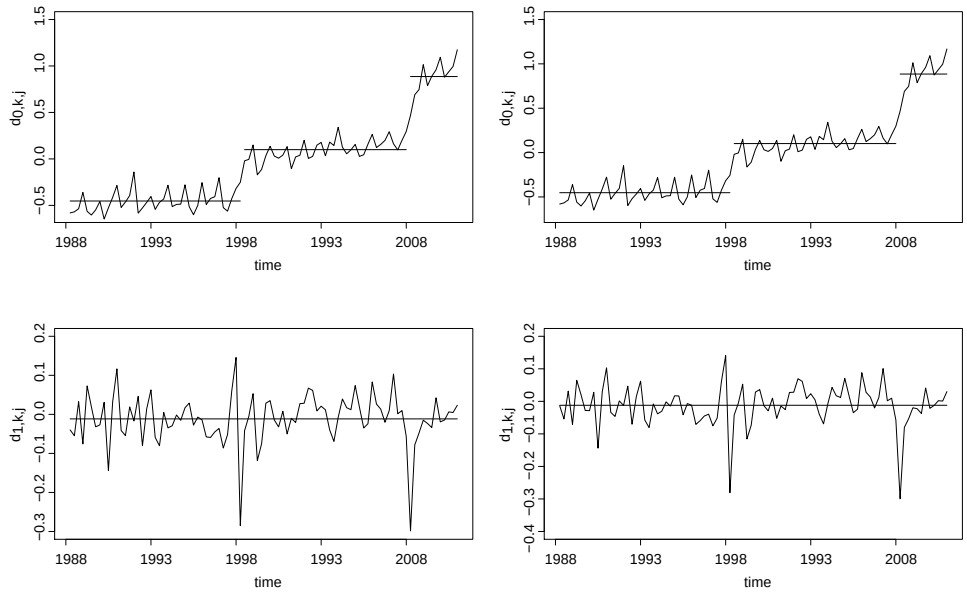

Figure 5.2: Change points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *changepoint* on the ground-level ozone concentration data. The left and right panels respectively display the results by using both LS-based and M-estimation-based change-point detection procedures.

We now modify the random number generation by changing the variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to $\sigma^2 = 1.6\tilde{\sigma}$. This modification produces a large variation in the observations after the change points. This is a reasonable scenario because if some activities are happening in a region, the observations would be more fluctuated than other times due to these activities. The M-estimation-based change-point detection procedure detects the change points at 1998 and 2008 successfully using the R package *changepoint*,

however, the LS-based method produces false change points. The results are shown in Figure 5.3, which demonstrates that the M-estimation-based change-point detection procedure is more stable than the LS-based change-point detection procedure in change-point detection in the presence of outliers and/or heavy-tailed observations.

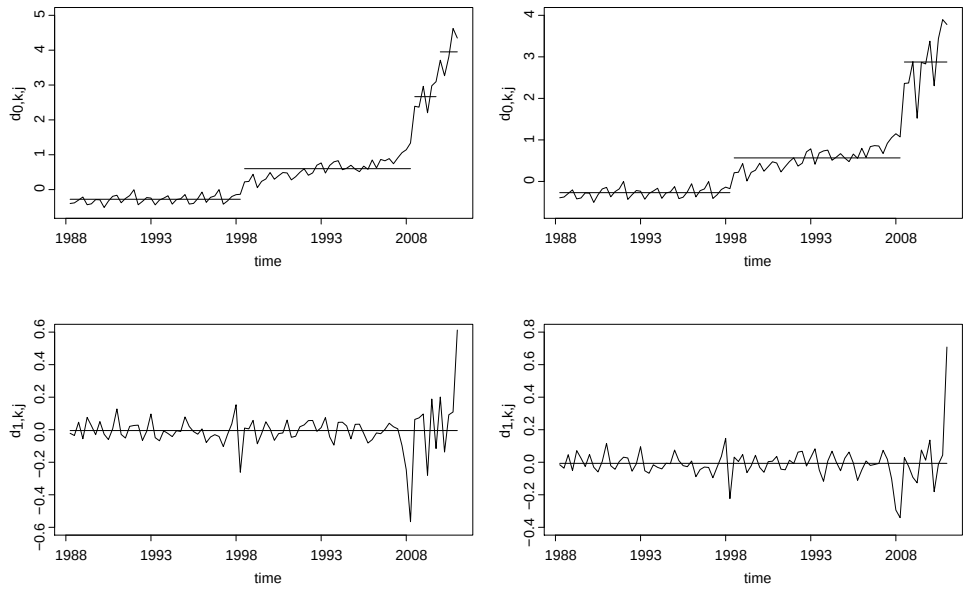

Figure 5.3: Change points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *changepoint* for heavy-tailed observations on the ground-level ozone concentration data. The left and right panels respectively display the results by using both LS-based and M-estimation-based change-point detection procedures.

5.2.2 PM2.5 Data

The data is the log of the daily average PM 2.5 data set collected in 37 stations covering the Province Ontario in a region with longitude from $-74°$ to $-90°$ and latitude from $41°$ to $49°$ in Canada from 2003 to 2016 in Chapter 3. On average, each station only has 7.5% of data missing due to the similar reasons for the missing data in the ground-level ozone concentration data in the last section. We let $\iota = 1$ by the pre-analysis of the data. The total number of the parameters is 122. The Euclidean distance of the parameter estimators obtained by the EM-type algorithm and the ${}_M$EM-type algorithm in the dissertation is 0.1023, which is small enough. We show that the ${}_M$EM-type algorithm is robust in the presence of outliers. Similarly, we first choose an area whose latitude is less than $43.50°$. There are 17 stations within this area. Then we randomly pick up a day, for 10 days after this day, we expanded the log-transformed ozone concentrations by 2 times. The experiment is repeated for 500 times, we recorded the Euclidean distance for both algorithms, in Table 5.2, the mean and the standard deviation (sd) of the Euclidean distance are reported. The simulation shows that outliers have less impact on the performance of parameter estimation if the proposed ${}_M$EM-type algorithm is used.

Table 5.2: Mean and standard deviation of the Euclidean distances

${}_M$EM-type algorithm		EM-type algorithm	
mean	sd	mean	sd
0.0322	0.0345	0.3128	0.0111

Then, we show the robustness of change-point detection of the proposed ${}_M$EM-type algorithm on the application of the PM 2.5 data set. We separate the stations into two groups by the latitude43.50°. Then, for each station in group 1, we add a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{3}{4}\tilde{\sigma}$ to each observation collected from 2008 to 2016 to create the first change-point in 2008. We also add a random number generated from the normal distribution with mean $\mu = \tilde{\sigma}$ and variance $\sigma^2 = \frac{3}{4}\tilde{\sigma}$ to the previously modified observations from 2013 to 2016 to create the second change-point in 2013. The results of detecting the change points by using the LS-based change-point detection procedure are shown in Figure 5.4. The right panel displays the results by using the M-estimation-based change-point detection procedure. Two sets of estimates, $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ are obtained. The plot displays the change points in $\{d_{0,k,j}\}$ (upper panel) and $\{d_{1,k,j}\}$ (lower panel) using both procedures. Figure 5.4 shows that both procedures capture the change points equally well.

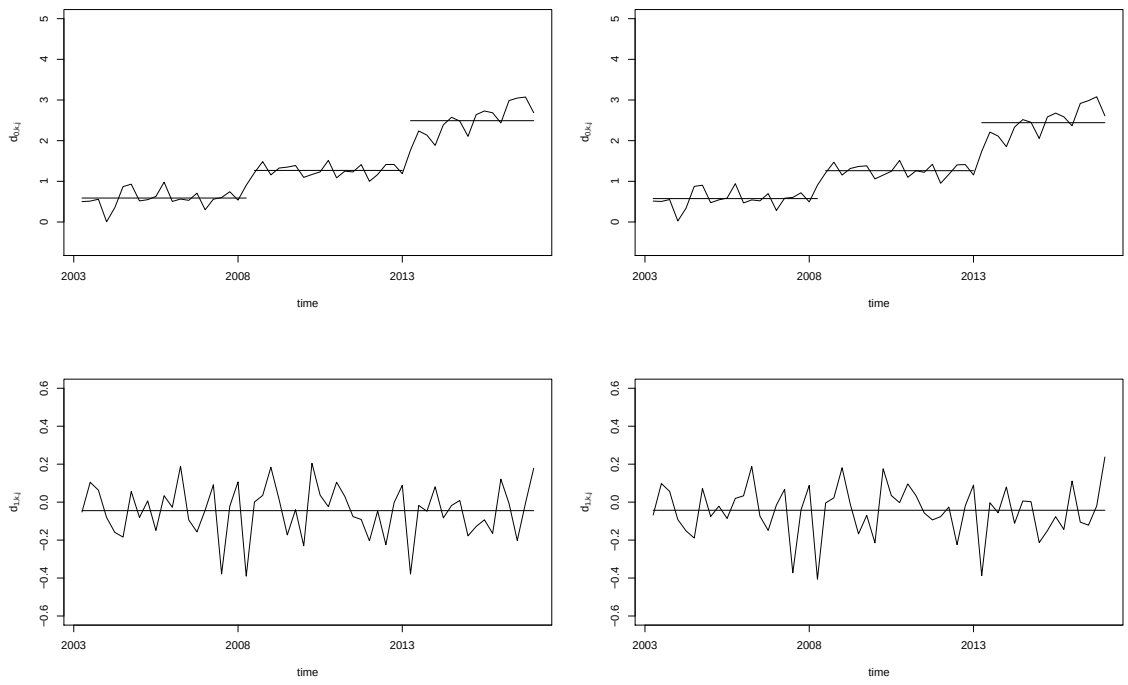

Figure 5.4: Change points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *changepoint* on the PM 2.5 data. The left and right panels respectively display the results by using both LS-based and M-estimation-based change-point detection procedures.

We now modify the random number generation by changing the variance $\sigma^2 = \frac{1}{2}\tilde{\sigma}$ to $\sigma^2 = 1.1\tilde{\sigma}$. This modification produces a large variation in the observations after the change points. This is a reasonable scenario because if some activities are happening in a region, the observations would be more fluctuated than other times due to these activities. The M-estimation-based change-point detection procedure detects

the change points at 2008 and 2013 successfully using the R package *changepoint*, however, the LS-based method produces false change points. The results are shown in Figure 5.5, which demonstrates that the M-estimation-based change-point detection procedure is more stable than the LS-based change-point detection procedure in change-point detection in the presence of outliers and/or heavy-tailed observations.

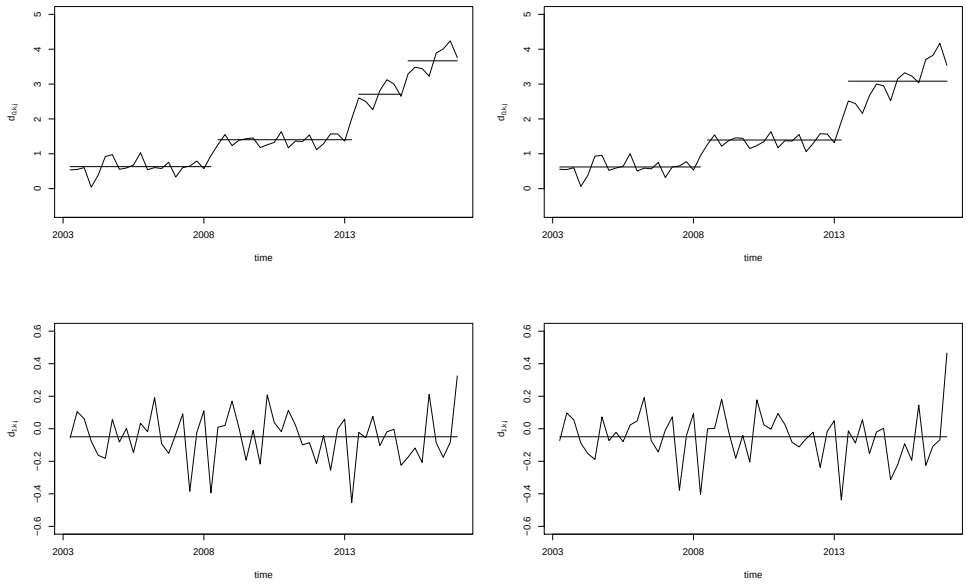

Figure 5.5: Change points in both means of $\{d_{0,k,j}\}$ and $\{d_{1,k,j}\}$ detected by using the R package *changepoint* for heavy-tailed observations on the PM 2.5 data. The left and right panels respectively display the results by using both LS-based and M-estimation-based change-point detection procedures.

6 Conclusions and Future Work

In this chapter, we summarize the contributions in this dissertation and discuss some future work.

Firstly, we have theoretically improved the method proposed by Ma and Zhu [2012] by a relaxation of the condition on the pdf of the covariates. In the implementation, the semiparametric approach needs a nonparametric estimation of the corresponding conditional expectations. Ma and Zhu [2012] used the Nadaraya-Watson estimator. It is well known that for the Nadaraya-Watson estimator, the limitation occurs when its denominator is equal to zero. We have trimmed the denominator of the Nadaraya-Watson estimator to make the estimation theoretically appropriate. In the implementation, replacing the Nadaraya-Watson estimator with other more accurate nonparametric regression estimators may improve the performance of the semiparametric approach for future research.

Secondly, we have improved the modeling technique of Bornn et al. [2012] for nonstationary processes by considering the covariance structure of the semivariograms.

We have proposed two methods of estimation. Demonstrated by the simulations, both of the proposed methods provide more accurate estimations than the original method. However, all of the methods aforementioned are restricted to Gaussian processes. In the future, we plan to extend the proposed methods to non-Gaussian processes in nonstationary fields.

Thirdly, we have modeled ozone concentrations in a region in the presence of missing data. We have derived predictive distribution using the hierarchical Bayesian spatio-temporal modeling technique [Le and Zidek, 2006] at the ungauged sites based on the covariance matrix estimated by dimension expansion method for modeling the semivariograms in nonstationary fields. Further, we have applied an entropy criterion [Jin et al., 2012] to decide whether new stations need to be added. This entropy criterion helps us solve the environmental network design problem. For demonstration, we have applied the method on ozone concentrations at 25 stations in the Pittsburgh region studied. The proposed method is more general and applicable as there is no assumption on the correlation structure among the data. For future work, the extension of the dimension expansion methods to spatio-temporal data can also be used to improve the hierarchical Bayesian spatio-temporal modeling technique.

Finally, we have improved the EM-type algorithm for the parameter estimation

of the GSTAR model by replacing the least-squares technique in the algorithm by M-estimation so that the modified algorithm provides robust estimations and is more accurate in detecting change points when data contains outliers and/or heavy-tailed observations. In the real data example, it has been shown that $_M$EM-type algorithm produces similar results for the GSTAR model as the original algorithm. In simulations, we test the robustness of the proposed methods in two different ways. First, we add some random outliers to the real data, our parameter estimates are more stable than the LS method. We test the accuracy of change-point detection. Both methods produce the same results. Second, we test the performance of the proposed method in the case when the observations are heavy-tail distributed. We increase the variance of the observations after the first change point occurs, the result shows that the LS method produces false change points, but the proposed method still successfully detects the change points with no false ones. Further investigation to find the unknown influences that cause change points in real life is valuable and interesting. We will consider it in future research.

Bibliography

L. Altieri, E. M. Scott, D. Cocchi, and J. B. Illian. A changepoint analysis of spatio-temporal point processes. *Spatial Statistics*, 14:197–207, 2015.

P. J. Bickels, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner Escobar. *Efficient and adaptive estimation for semiparametric models*. Baltimore, MD: The Johns Hopkins University Press, 1993.

L. Bornn, G. Shaddick, and J. V. Zidek. Modeling non-stationary processes through dimension expansion. *Journal of American Statistical Association*, 107:281–289, 2012.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Application*, 6:76–90, 1979.

R. D. Cook. *Regression graphics: Ideas for studying regressions through graphics*. New York: Wiley, 1998.

R. D. Cook and S. Weisberg. Discussion of 'sliced inverse regreesion for dimension reduction'. *Journal of the American Statistical Association*, 86:28–33, 1991.

N. Cressie. Fitting variogram models by weighted least squares. *Mathematical Geology*, 17(5), 1985.

N. Cressie. Statistics for spatial data. *New York: Wiley*, 1993.

N. Cressie and C. K. Wikle. *Statisitcs for spatio-temporal data.* Wiley, 2011.

M. L. Eaton. A characterization of spherical distributions. *Journal of Multivariate Analysis*, 34:439–446, 1986.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.

M. G. Genton. Variogram fitting by generalized least squares using an explicit formular for the covariance structure. *Mathematical Geology*, 30:323–345, 1998.

J. Hay. An assessment of the mesoscale variability of solar radiation at the earth's surface. *Solar Energy*, 32(3):425–434, 1984.

N. J. Higham. Computing the nearest correlation matrix-a problem from finance. *Journal of Numerical Analysis*, 22:329–343, 2002.

P. J. Huber. Robust regression. *The Annals of Statistics*, 1(5):799–821, 1973.

E. T. Jaynes. *Information theory and statistical mechanics, Statistical Physics,3, (Ed).* KW Ford. New York: Benjamin, 1963.

B. Jin, Y. Wu, and E. Chan. Hierarchical bayesian spatial-temporal modeling of regional ozone concentrations and respecitve network design. *Journal of Environmental Statistics*, 3(3), 2012.

D. L. Knol and J.F. Berge. Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, 54:53–61, 1989.

N. D. Le and J. V. Zidek. *Statistical analysis of environmental space-time processed.* Springer, 2006.

N. D. Le, W. Sun, and J. V. Zidek. Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *Canada Journal of Statistics*, 29:529–554, 2001.

B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008, 2007.

K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

Y. Ma and L. P. Zhu. A semiparametric approach on dimension reduction. *Journal of the American Statistical Association*, 107:167–179, 2012.

Y. Ma and L. P. Zhu. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 2013.

Y. P. Mack and B. W. Silverman. Weak and strong uniform consistancy of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 61:405–415, 1982.

G. Matheron. Traite de geostatitique appliquee, tome i. *Memories du Bureau de Recherches Geologiques et Minieres, NO. 14. Editions Technip, Paris*, 1962.

W. G. Muller. Least-squares tting from the variogram cloud. *Statistics and Probability Letter*, 43:93–98, 1998.

I. Nappi-Choulet and T. P. Maury. A spatiotemporal autoregressive price index for the paris office property market. *Real Estate Economics*, 37:305–340, 2009.

P. Otto and W. Schmid. Detection of spatial change points in the mean and covariances of multivariate simultaneous autoregressive models. *Biometrical Journal*, 58: 1113–1137, 2016.

O. Perrin and W. Merring. Nonstationary in $\mathcal{R}^n$ is second-order stationary in $\mathcal{R}^{2n}$. *Journal of Applied Probability*, 40(3):815–820, 2003.

O. Perrin and M. Schlather. Can any multivariate gaussian vector be interpreted as a sample from s stationary radom process? *Statistics and Probability Letter*, 77 (9):881–884, 2007.

P. Porter, S. Rao, I. Zurbenko, A. Dunker, and G. Wolff. Ozone air quality over north america: part ii-an analysis of trend detection and attribution techniques. *Journal of Air and Waste Management Association*, 51:283–306, 2001.

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119, 1992.

W. Tober. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970.

G. Wabba and J. Wendelberger. Some new mathematical methods for variational objective analysis uisng splines and cross-validation. *Monthly Weather Review*, 108:1122–1143, 1980.

H. Wackernagel. Multivariate geostatistics: An introduction with applications. *Springer*, 2003.

C. K. Wikle, L. M. Berliner, and N. Cressie. Hierachichical bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154, 1998.

Y. Wu, B. Jin, and E. Chan. Detection of changes in ground-level ozone concentrations via entropy. *Entropy*, 17:2749–2763, 2015.

Y. Wu, X. Sun, E. Chan, and S. Qin. Detecting non-negligible new influences in environmental data via a general spatio-temporial autogressive model. *British Journal of Environment & Climate Change*, 7(4):223–235, 2017.

J. Wyse, N. Friely, and H. Rue. Approximate simulation-free bayesian inference for multiple changepoint models with dependence within segments. *Bayesian Analysis*, 6(4):501–528, 2011.

Y. Xia. A contructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35:2654–2690, 2007.

Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adoptive estiamtion of dimension reduction space (with discussion). *Journal of Royal Statistical Society, Series B*, 64:363–410, 2002.

X. T. Xue, B. Guy, L. Xing, F. Pierre, G. Claire, and R. Philip. The impact of high altitude aircraft on the ozone layer in the stratosphere. *Journal of Atmospheric Chemistry*, 18:103–128, 1994.

L. X. Zhu and K. T. Fang. Asymptotic for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.