



北京大学

## 博士研究生学位论文

题目：时间空间模型以及生存分析相  
关问题的理论及应用

姓 名：黄家盛

学 号：1501110056

院 系：数学科学学院

专 业：统计学

研究方向：函数型数据分析，非参及半参统计

导 师：姚方 教授

二〇二〇年六月



## 版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则一旦引起有碍作者著作权之问题，将可能承担法律责任。



## 摘要

在本文中，我们主要关注两方面研究：时间空间模型的理论及应用，以及因果与中介效应分析在运动学领域中的应用。首先，我们基于一个非参数时间调整方法对天津市城区的大气污染问题进行了考察。该统计方法可以有效地去除掉大气质量数据中带来的气象混淆因素。通过将该方法应用在天津市11个大气质量监测站点中的四个主要污染物（ $\text{PM}_{2.5}$ ,  $\text{SO}_2$ ,  $\text{NO}_2$ ,  $\text{O}_3$ ）数据上，我们得到了该市自2013至2016年调整均值和分位数的结果与逐年比较。结果表明未来该城市需要同时处理好 $\text{PM}_{2.5}$ 和 $\text{O}_3$ 两个污染物的治理工作。我们接下来考虑将传统的半参数回归理论方法推广到时间空间模型当中。为了使模型中的线性项取得更高的估计效率，我们在传统的kernel-profile估计方程中引入了时间空间协方差矩阵。通过对时间基底作弱相关结构的假设，该协方差阵可以弱化为一个分块对角矩阵的估计。在时间尺度上的平稳性假设下，分块对角矩阵可以用非参数二元核平滑的方法估计出来。我们通过模拟和大气污染、湖北省COVID-19肺炎数据诠释了该模型的表现。最后，我们基于运动学中心纵向研究数据，针对心脏病风险提出了一个因果与中介效应分析模型。其中首先基于前期来访数据使用一个基于生存型因变量的边缘结构化模型检验了阻力训练对发生心脏病风险的因果效应。进一步地，我们根据前期来访和后续临床试验数据提出一个包含多条纵向轨迹的多变量联合建模方法。这两个方法能够帮助我们从不同的角度挖掘出潜在的重要中介效应，并且能更加全面地分析出阻力训练对于降低心脏病风险的作用。

**关键词：**时间空间模型，时间调整，气象混淆，半参数回归，kernel-profile估计方程，弱相关，中介效应，边缘结构化模型，因果效应，多变量联合建模



# ABSTRACT

In this thesis, we mainly focus on two aspects of research: theory and applications of spatio-temporal models, and application of a causal mediation effect analysis on kinesiology. First, we investigate the air pollution problem in the urban area of Tianjin under a nonparametric temporal adjustment method, which is demonstrated on its ability to remove the meteorological confounding existed in the air quality data. By applying our method on four major pollutants ( $\text{PM}_{2.5}$ ,  $\text{SO}_2$ ,  $\text{NO}_2$ ,  $\text{O}_3$ ) from 11 air quality monitoring sites in Tianjin, we derive adjusted mean and quantile results with comparison across years (from 2013 to 2016). The results suggest that future air quality management plans in Tianjin need to be based on dual targets of  $\text{PM}_{2.5}$  and  $\text{O}_3$ . We then consider extension work of semiparametric regression methods onto spatio-temporal models. In order to achieve higher efficiency on the estimation of the linear terms in the model, spatial-temporal covariance is interpolated in the traditional kernel-profile estimation equation. Estimation of the covariance matrix can be reduced to a banded block matrix by assuming a weak dependence structure on the temporal basis. By assuming stationarity over time, the banded block matrix can be estimated fully-nonparametrically through bivariate kernel smoothing. We illustrate the performance of the proposed modeling framework and estimation methods through simulation studies and in applications to air pollution and COVID-19 data in Hubei Province. Finally, we propose a causal and mediation effect analysis on the risk of Cardiovascular Disease (CVD) based on the Aerobics Center Longitudinal Study. We first derive a marginal structural model (MSM) to examine the causal effect of Resistance Exercise (RE) on total CVD events using baseline observations under the case of a survival outcome. Then a multivariate joint modeling approach accommodating information from multiple longitudinal trajectories is proposed with data generated from both the baseline and follow-up clinical trials. These two approaches help us find the underlying potential mediation effects from different perspectives, and enables us to analysis the effect of RE on lowering CVD risk more comprehensively.

**KEYWORDS:** spatio-temporal models, temporal adjustment, meteorological confounding, semiparametric regression, kernel-profile estimation equation, weak dependence, mediation effect, marginal structural model, causal effect, multivariate joint modeling





# 目录

<b>第一章 Introduction</b>	<b>1</b>
<b>第二章 Evaluating air pollution upon a nonparametric temporal adjustment approach</b>	<b>5</b>
2.1 My contributions on this research . . . . .	5
2.2 Introduction . . . . .	5
2.3 Overview of the data . . . . .	7
2.4 Challenge on meteorological confounding . . . . .	10
2.5 Nonparametric temporal adjustment approach . . . . .	12
2.6 Main results and conclusions . . . . .	16
<b>第三章 Efficient semiparametric estimation in spatio-temporal models</b>	<b>21</b>
3.1 My contributions on this research: . . . . .	21
3.2 Overview . . . . .	21
3.3 Model setup . . . . .	24
3.4 Estimation methodology . . . . .	24
3.5 Asymptotic properties . . . . .	28
3.6 Simulation Studies . . . . .	36
3.7 Real data analysis . . . . .	38
3.7.1 Air pollution mode data in China . . . . .	38
3.7.2 COVID-19 in Hubei Province . . . . .	41
<b>第四章 Causal and mediation effect analysis</b>	<b>45</b>
4.1 My contributions on this research: . . . . .	45
4.2 Overview . . . . .	45
4.3 Data description . . . . .	48
4.4 Causal effect analysis using MSM . . . . .	50
4.5 Construction of multivariate joint modeling . . . . .	52
4.6 Numerical results . . . . .	55
4.7 Conclusion and Discussion . . . . .	58

参考文献	69
致谢	71
北京大学学位论文原创性声明和使用授权说明	73

## 第一章 Introduction

With China industrializes at a fast pace, air pollution with high concentrations of pollutants is widely existed in some parts of our country. Air pollution is known to cause serious issues on health, social and economic development. Epidemiological studies demonstrated that exposure to air pollution has significant bad effects on human health [2, 19]. Ebenstein and his collaborators found a significant negative influence on life expectancy under sustained exposure to air pollution in China's Huai River area [11]. To tackle with the notorious air pollution problem, China's State Council established a "National Ten Point Plan" in September 2013 that sets specific reduction targets in terms of annual  $PM_{2.5}$  averages for various regions of the country by Year 2017. The reduction target for Beijing-Tianjin-Hebei region is "a 25% reduction over the Year 2012 level by Year 2017". Tianjin is at the heart of the North China Plain and has experienced severe air pollution in the past few years. The underlying reason for this problem is excessive emission of pollutants from vast installations of heavy industries ranging from steel and iron making to building materials in Hebei Province surrounded Tianjin while assisted by increasing number of motor vehicles. At the same time, the dispersion condition is not good since Tianjin is surrounded by Taihang Mountain range on the west and Yan Mountain range at the north. This geography is ideal for fostering secondary generation of fine particulate matters  $PM_{2.5}$ .

Since the observed air pollution is influenced by meteorological conditions [39, 59, 65], a decline in observed concentrations may be either due to a decrease in emission or due to a favorable meteorological condition for the dispersion of pollutants. The latter one is called meteorological confounding. In order to measure the underlying temporal change in the pollution concentrations of Tianjin, we need to compare the air quality data across different years on a common temporal meteorological baseline. As demonstrated in Chapter 2, recalculating the concentrations on the meteorological baseline makes air quality statistics (both averages and quantiles) comparable over the years, and hence can effectively removes the confounding. Therefore, we develop a nonparametric temporal adjustment approach in tackling with this problem. Adjusted averages and quantiles regarding the air pollutants considered can be derived by using this method.

Furthermore, in order to derive the relationship between two air pollutants under a

particular spatial region, we consider extending the traditional semiparametric regression methods onto a spatio-temporal framework. Semiparametric regression problems have received much attention in the field of longitudinal clustered data. A typical form of semiparametric regression model for clustered data is [40]:

$$Y_{ij} = X_{ij}^T \beta + \theta(T_{ij}) + \epsilon_{ij},$$

where  $Y_{ij}$  denotes the  $j$ -th response variable and  $(X_{ij}, T_{ij})$  is the  $j$ -th covariate within the  $i$ -th cluster. In this model, the coefficient  $\beta$  is a  $p \times 1$  vector and  $\theta(\cdot)$  is an unknown function. The focus of most literatures in semiparametric regression is on the estimation of  $\beta$ . This method was additionally refined in [66, 67] where correct specification of the covariance matrix is interpolated in the model. This results in higher estimation efficiency of the parametric terms. Li improves the work on semiparametric estimation by combining the efficient semiparametric estimator with nonparametric covariance estimation using a three-stage estimation procedure [37]. This method is considered to be robust against misspecification of covariance models.

More recently, Semiparametric regression models has been widely used in spatio-temporal modeling. Related applications of semiparametric spatio-temporal regression in literatures include [6, 17, 18, 25, 75], which involved research studies on online auctions, air pollution, imaging data, interval-censored data, etc. Ma introduced semiparametric spatio-temporal covariance models with the ARMA temporal margin [45]. Cai and his collaborators investigated bayesian semiparametric model with spatially-temporally varying coefficients selection [5]. However, few studies considered the efficiency of the parametric estimators in spatio-temporal models. So in Chapter 3, I propose a novel semiparametric spatio-temporal model under weak dependence in theory [23]. By assuming  $m$ -dependence structure on the temporal basis, the estimation of the spatio-temporal covariance can be reduced to the case of a banded block diagonal covariance matrix. The estimation efficiency of the parametric counterpart after interpolating the estimated covariance matrix is proved to achieve higher efficiency compared to the case of working independence. Simulation studies and applications of the methodology on air pollution and COVID-19 are also investigated to better illustrate the effectiveness of our novel method.

In the field of kinesiology and medical science, we know that factors leading to cardiovascular disease comes from varied forms. Basic elements like gender, obesity, total cholesterol, inherited gene, exercise, healthy diet, control of blood pressure and

glucose, smoking status, alcohol use, and even air pollution are all possible factors which may increase the risk of CVD [49]. Vast amount of research has been devoted to measures on lowering CVD risk [1, 10, 51, 69]. It is universally accepted that Aerobic Exercise (AE) is effective on CVD prevention [32, 34]. More recently, cardiovascular benefits from Resistance Exercise (RE) has attracted much attention. RE is a muscle training method which helps growing strength and keeping fit [3, 57]. In addition, many professionals believe that Body Mass Index (BMI) is a crucial risk factor for CVD. It is considered to be an important mediator that links physical activity (PA) to CVD.

However, the underlying causal effect of RE on CVD risk has not been discovered before. Research on causal inference has becoming more and more important in recent years. A causal relationship is considered to possess higher significance than just association between an exposure and an outcome variable. Usually, we decompose the total effect of an exposure on the outcome into natural direct effect and natural indirect effect, and marginal structural models (MSM) are often used to estimate the marginal means in the components of natural direct and indirect effect by the ratio of mediator probability weighting approach [22, 30, 31]. In addition, mediation effects can also be analyzed through a multivariate joint modeling method, which was introduced in [48]. In Chapter 4, we will first propose a causal effect analysis based on a survival outcome instead of the frequently discussed cases regarding binary or multinomial outcomes, and then develop a multivariate joint modeling approach in analyzing the underlying potential mediation effects connecting RE with CVD risk.



## 第二章 Evaluating air pollution upon a nonparametric temporal adjustment approach

### 2.1 My contributions on this research

I am one of the three main contributors in this research work. I did data pretreatment, reorganization and analysis regarding air pollution data in Tianjin. The total amount of data for Beijing-Tianjin-Hebei region is over 44000 thousand, which is considered to be big data. Furthermore, I participate in the methodology construction work and accomplished novel methods on deriving the adjusted quantiles with R coding.

### 2.2 Introduction

Investigations on the air quality of a city are generally based on air quality monitoring data obtained from monitoring sites or satellite remote sensing data. Air quality assessments based on data from a small number of monitoring sites per city was carried out regarding China's air pollution studies in recent years. Beijing's air quality from 2010 to 2014 was assessed by using the US Embassy's  $PM_{2.5}$  data in [39]. Since modeling of hourly  $PM_{2.5}$  data in Beijing is confronted with challenges on meteorological confounding, the severe condition of  $PM_{2.5}$  was quantified by using a novel statistical approach upon adjusting  $PM_{2.5}$  w.r.t several meteorological variables. The averages and quantiles of  $PM_{2.5}$  were adjusted w.r.t the meteorological conditions in a particular time span. As the study was based on assessment instead of finding the relationship between  $PM_{2.5}$  and the weather variables, parametric or semiparametric models were not considered under this setting. The new method can help testing if  $PM_{2.5}$  levels in Beijing lowered and met the pollution reduction goal of China's State Council. This method produced a more comprehensive understanding on the pollution level which outperforms common prediction techniques such as Pollution Linked with Air-Quality and the Meteorology (PLAM) Index and the Community Multi-Scale Air Quality (CMAQ) modeling system, which do not take the uncertainties caused by confounding variables into account. Also, vast amount of data in the analysis was sufficient enough in giving accurate evaluation of the changing trend regarding air pollution in Beijing within 5 years. Results showed

that  $\text{PM}_{2.5}$  concentrations experienced a significant increase in year 2013 and 2014 in comparison to 2012. Substantial differences between the adjusted and the raw means of  $\text{PM}_{2.5}$  was discovered, which implied the necessity in implementing the statistical adjustment approach. Furthermore, two quasi-experiments were implemented on the Asia-Pacific Economic Cooperation (APEC) meeting and the annual winter heating to get deeper knowledge on the influence of  $\text{PM}_{2.5}$ , which indicated that reducing the consumption of coal-based energy was considered to be effective on reducing the impact of  $\text{PM}_{2.5}$  concentrations in Beijing.

In addition, Liang and her collaborators analyzed  $\text{PM}_{2.5}$  observations from the national pollutant monitoring stations [38]. The data reliability of  $\text{PM}_{2.5}$  data in Beijing, Shanghai, Guangzhou, Chengdu, and Shenyang was studied through using a 3-year data from January 2013. These data were collected from the U.S. diplomatic posts and the nearby Ministry of Environmental Protection sites. The research implied that the two data sources introduced provide consistent evaluation on the air quality condition in the five cities above. Standard monthly weather conditions for each of the five cities were produced to minimize the effects of the confounding variables. The results indicated that Shanghai and Guangzhou rank the highest on air quality among all the five cities, while Beijing and Chengdu attains the worst air condition. This research also suggested that geographical configuration was an extremely important factor in a city's air quality management.

Also, much attention has been paid to the composition of the pollutants [68, 73] and effects of meteorological variables. Geographical condition is another important factor that influences the pollutant's concentrations and diffusion. By analyzing  $\text{PM}_{2.5}$ , aerosol optical depth (AOD), and long-term visibility data, along with various climate and meteorological factors and the boundary layer structure, Wang and his collaborators found that the unfavorable geographical condition of Beijing-Tianjin-Hebei region limited the diffusion of pollutants [65]. Furthermore, their research showed an evidence that high humidity contributes to the aerosols and secondary transformation under high emission, leading to severe  $\text{PM}_{2.5}$  episodes in Beijing in January 2013. There were also studies based on analyzing aerosol optical depth (AOD) data from satellite remote sensing. The focus there was to calibrate the ground-level  $\text{PM}_{2.5}$  concentrations from the AOD [16, 47, 62, 64]. This approach provides a well coverage for areas where ground monitoring sites are not available. Although the calibrated  $\text{PM}_{2.5}$  from the AOD data



provide a broad spatial-temporal coverage, they are subject to relatively large calibration errors, which depend on meteorological factors and the model used in the calibration [8, 42].

The current study on air pollution has two purposes. One is to demonstrate the need and effects of the meteorological adjustment for air quality assessment. The other is to conduct air quality assessment in Tianjin based on data from both the monitoring sites and meteorological information. The latter one is designed to provide a comprehensive evaluation on the concentrations of four main air pollutants:  $\text{PM}_{2.5}$ , sulfur dioxide ( $\text{SO}_2$ ), nitrogen dioxide ( $\text{NO}_2$ ), and the ground ozone ( $\text{O}_3$ ) for 18 consecutive seasons between spring of Year 2013 and summer of 2017 in Tianjin.

Our study finds significant declines in  $\text{PM}_{2.5}$  and  $\text{SO}_2$  in Tianjin. It also reveals a significant increase in the ground ozone level at a quite alarming rate, and a static nitrogen dioxide concentration. These indicate the air quality management in Tianjin should be transformed from a sole target of  $\text{PM}_{2.5}$  to a new system with dual targets of  $\text{PM}_{2.5}$  and  $\text{O}_3$ . This new dual-targets system should have the list of the primary precursors extended to include  $\text{NO}_x$  and volatile organic compounds in addition to  $\text{SO}_2$ , which demands a new strategy in this next phase of air quality management for Tianjin.

The rest of this chapter is organized as follows. In Section 3, we give an overview of the air quality data. In Section 4, challenge on tackling with meteorological confounding is demonstrated. In Section 5, we propose the novel nonparametric temporal adjustment approach. Finally, we present main results and conclusions for Tianjin in Section 6.

## 2.3 Overview of the data

China established a national air quality monitoring network in January 2013 that provides hourly recordings on six common air pollutants. Our analysis on Tianjin is based on hourly concentrations of  $\text{PM}_{2.5}$ ,  $\text{SO}_2$ ,  $\text{NO}_2$  and the ground ozone  $\text{O}_3$  from 11 “Guokong” monitoring sites. “Guokong” means sites are directly managed by the Ministry of Environment and Protection (MEP) in data collection with instantaneous transmission to a data center in Beijing to avoid any potential local interference. Since the observed pollutants’ concentrations are ordinarily affected by the meteorological condition, we employ meteorological data at 3 weather observing stations from China Meteorological Administration (CMA) in Tianjin. The 11 air quality monitoring stations are allocated to the 3 corresponded meteorological sites in which district is belonging to.

The meteorological data consists of 6 hourly variables: air pressure (PRES), temperature (TEMP), dew point (DEWP), wind direction (W), cumulative wind velocity (CWS) and cumulative precipitation (R). The wind directions are grouped into five categories  $=\{NW, NE, SW, SE, CV\}$  with  $NW=(W,N,NW,NNW,WNW)$ ,  $NE=(NE,NNE,ENE)$ ,  $SW=(SW,SSW,WSW)$ ,  $SE=(E,S,SE,ESE,SSE)$  and CV contains both the calm and variable wind as well as the wind whose speed is less than  $0.5m/s$ . The cumulative wind velocity is defined as summation of the wind velocity since the hour that a wind direction was established. The definition of the cumulative precipitation is similar by adding up hourly precipitation amount over consecutive hours with non-zero precipitation. Any of these two variables is reset to zero whenever there is a change of wind direction or an hour without precipitation. Figure 2.1 below displays the locations of the 11 air quality monitoring stations with 3 meteorological sites in Tianjin:



图 2.1: Locations of 11 air quality monitoring stations with 3 meteorological sites in Tianjin.

The first two months of Year 2013 regarding the air quality data had high proportions of missing values, which led us to consider data from March 2013 until August 2017. The time unit of our assessment is season such as spring from March to May, summer June to August, fall September to November, and winter December to February next year. This comes from a consideration that a season in Tianjin offers quite homogeneous weather conditions and sufficient amount of data for the meteorological adjustment. As a result, we use the seasonal year from March to February next year

rather than the calendar year.

We will first give an overview on the pollution status of  $PM_{2.5}$  in Tianjin. The  $PM_{2.5}$  concentration is divided into the following 6 levels:

**Level 1 (Good):**  $PM_{2.5} \leq 35\mu g/m^3$

**Level 2 (Fine):**  $35\mu g/m^3 < PM_{2.5} \leq 75\mu g/m^3$

**Level 3 (Mild pollution):**  $75\mu g/m^3 < PM_{2.5} \leq 115\mu g/m^3$

**Level 4 (Moderate pollution):**  $115\mu g/m^3 < PM_{2.5} \leq 150\mu g/m^3$

**Level 5 (Heavy pollution):**  $150\mu g/m^3 < PM_{2.5} \leq 250\mu g/m^3$

**Level 6 (Severe pollution):**  $PM_{2.5} > 250\mu g/m^3$

Here, we set  $75\mu g/m^3$  as the maximum value of  $PM_{2.5}$  concentration with fine air quality according to the World Health Organization (WHO). Considering the improvement of air quality in Tianjin, we set  $35\mu g/m^3$  as the criteria for Level 1 in pollution. This is due to the fact that long-term exposure to  $PM_{2.5}$  concentration between  $35\mu g/m^3$  and  $75\mu g/m^3$  is still harmful to human health regarding epidemiological studies [53].

In order to gain insight on the air quality condition in Tianjin, we take average of  $PM_{2.5}$  concentrations for all the 11 air quality monitoring sites, and calculate the corresponded ratios regarding the above 6 pollution levels for each season from spring 2013 to winter 2016 based on the hourly  $PM_{2.5}$  data. The results are illustrated in Figure 2.2 below:

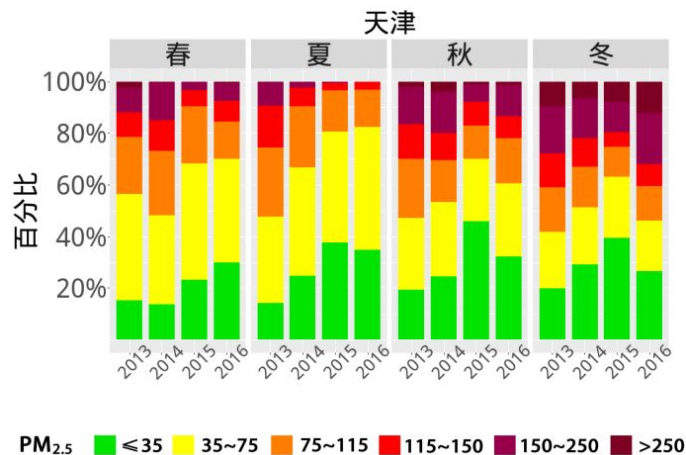


图 2.2: Ratio of 6 pollution levels in Tianjin from spring 2013 to winter 2016 based on averaged hourly  $PM_{2.5}$  data.

We see obvious difference regarding the ratio of the 6 pollution levels for different seasons, which is especially reflected in the ratio of heavy and severe pollution levels.

Heavy pollution in winter attains the highest ratio compared to the other three seasons. There is few heavy pollution in summer with mainly good, fine and light pollution status. Furthermore, great improvement on  $PM_{2.5}$  concentration is achieved for Tianjin from year 2013 to 2015, which is reflected in the continuing increase of good and fine air quality and substantially reduction regarding the ratio of heavy pollution for each season. Few heavy and severe pollution is found till summer 2016. However, the results indicate a worsen air quality condition in autumn and winter of 2016, which shows steadily increase in the ratio of heavy pollution.

For ground ozone  $O_3$ , solar radiation is a significant meteorological factor that influences its generation. We use the ultraviolet radiation with wavelengths between 200 and 440 nm, which is termed as UVB by the European Centre for Medium-Range Weather Forecasts (ECMWF). The UVB data are provided at a grid size of hourly frequency available over the study region. The process of  $O_3$  generation triggered by the solar radiation has a delayed effect. We calculate the correlation between  $O_3$  and the cumulative lagged UVBs ranging from 0 to 11h. Our analysis shows that the maximum correlation between the ozone concentration within the 8 h between 12:00 and 19:00 and the cumulative lagged UVBs was attained at lag 7 in the spring, lag 6 in the summer and the fall, lag 4 for the winter. Hence, we use these numbers of lagged cumulated UVB values and take a logarithm transform for the four seasons as a covariate in the adjustment for  $O_3$ . The weather data are from March 2010 to August 2017, three years more than the time span of air quality data. The latter one is to allow the construction of a temporally meteorological baseline at each city for the meteorological adjustments.

## 2.4 Challenge on meteorological confounding

A traditional way to compare air quality between two time periods is to calculate the raw averages of a pollutant, say  $PM_{2.5}$ , from the hourly concentrations in these two periods. As the raw  $PM_{2.5}$  concentrations are highly affected by the weather condition, these two averages are not comparable from the viewpoint of gauging on the underlying pollution emission, as the weather conditions of the two periods are not the same. To see more clearly the drawbacks of the approach, Figure 2.2 shows the average concentrations of  $PM_{2.5}$  under the five wind directions at Tianjin Beichen monitoring site, in the winters of 2015 and 2016, respectively.

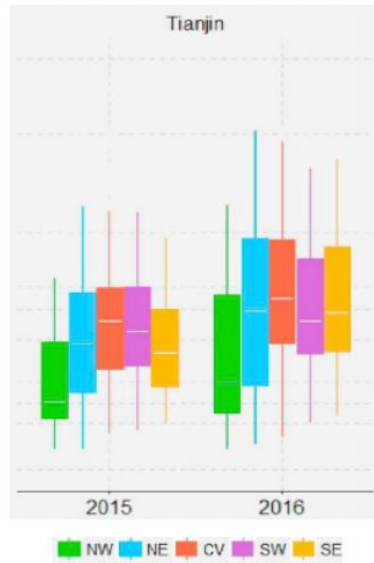


图 2.3: Boxplot of  $PM_{2.5}$  concentrations under five wind directions in Tianjin Beichen site in the winter of 2015 and 2016. Note: the white lines represent the medians.

It shows that  $PM_{2.5}$  is highly dependent on the wind direction. The concentration is the lowest under NW for Tianjin Beichen. The wind direction with low  $PM_{2.5}$  corresponds to the direction of lower emission area of Tianjin. The concentration is the highest under CV, as CV is associated with static air that is strongly associated with severe pollution. By simple calculation on the raw average of  $PM_{2.5}$  in winter 2015 and winter 2016 of Tianjin, we see that the increase of average value in winter 2016 exaggerated the increase of the underlying emission as part of the rise is due to the winter 2016 having more CV winds. This shows a confounding of the observed  $PM_{2.5}$  by wind. The confounding makes the two raw averages not comparable as far as in reflecting the underlying emission. A fairer comparison that is designed to measure the underlying emission level is to compute the potential average for winter 2016 under the 2015's wind composition, which is

$$\hat{\mu}_{2015}(2016) = \sum_{w \in \Omega} r_w(2015) \cdot \hat{\mu}_w(2016),$$

where  $\Omega$  is the set of the five wind directions, and  $r_w(2015)$  denotes the 2015's percentage of a wind direction  $w \in \Omega$ , and  $\hat{\mu}_w(2016)$  is the average concentration in winter 2016 for direction  $w$ . In the field of causal inference,  $\mu_{2015}(2016)$  is called a counterfactual or potential average.

When we compare pollution levels over several years, for instance 4 years in our current study, it is rather cumbersome to compute all pairs of potential averages. A simple method is to adjust the averages over a temporal baseline wind composition that produces only one average per season per year. Mathematically, let  $r_w(\cdot) = A^{-1} \sum_{a=1}^A r_w(a)$  be the average wind distribution based on  $A$  years' wind data. Then, average concentration pegged to the baseline for year  $a$  is

$$\hat{\mu}(a) = \sum_{w \in \Sigma} r_w(a) \cdot \hat{\mu}_w(a).$$

In addition, the observed  $\text{PM}_{2.5}$  is not only confounded by wind, but also by other meteorological variables. From both the air quality and weather data, we learn that the confounding of  $\text{PM}_{2.5}$  by dew point is the most visible one, as low (high)  $\text{PM}_{2.5}$  is highly associated with low (high) dew point. This implies  $\hat{\mu}(a)$  is confounded by dew point and other variables as it has only adjusted the wind composition.

## 2.5 Nonparametric temporal adjustment approach

As shown in the previous section, in addition to the emission, the meteorological conditions also affect the pollutant's concentrations significantly. Indeed, a favorable meteorological condition with high emission can result in low  $\text{PM}_{2.5}$  concentrations, while a static and stable weather can cause severe pollution even in a low emission regime. Therefore, instead of comparing raw concentrations, a statistical approach that can correct the meteorological confounding should be applied. Here, we consider a nonparametric model on the pollutant's concentrations, meteorological variables and emission [21].

In order to neutralize the meteorological effect, we consider a meteorological adjustment approach that was first proposed in [39] and is refined here in this study. The approach recalculates the average concentrations of a pollutant under a temporally meteorological baseline condition based on the historical meteorological data which can remove the meteorological confounding. It consists of two key components. One is estimating the nonparametric regression function for the pollutant with respect to the meteorological variables, which not only includes the wind composition, but also other related meteorological variables. The second component is to recalculate the average on a temporal baseline meteorological condition based on longer weather records.

Let  $Y_{ijt}(s)$  denote the concentration of a pollutant at hour  $t$  in season  $j$  and year  $i$  of a monitoring site  $s$ ,  $W_{ijt}(s)$  represent the wind direction which is a discrete variable and  $X_{ijt}(s)$  represent other meteorological variables, such as temperature (TEMP) and dew point (DEWP) etc. The term  $n_{ij}$  is set as the total number of observations in the season. We use the following model to quantify the relationship between pollutant concentrations and meteorological variables:

$$Y_{ijt}(s) = m_{ij}(X_{ijt}(s), W_{ijt}(s); s) + \epsilon_{ijt}(s),$$

where  $\epsilon_{ijt}(s)$  is the residual term. As the underlying emission is unobservable at the hourly frequency, it is not reflected in the model directly. However, it is implicitly embedded in the regression function  $m_{ij}(\cdot)$ . Also, part of the difference between  $m_{i_1j}(\cdot)$  and  $m_{i_2j}(\cdot)$  for two different years  $i_1$  and  $i_2$  is due to the difference of the underlying emission in the two years.

Next, we define a common probability baseline density function of the meteorological variables:

$$f_{\cdot j}(x, w; s) = \frac{1}{M} \sum_{a=1}^M f_{aj}(x, w; s),$$

where  $f_{aj}(x, w; s)$  is the probability density function of the meteorological data at season  $j$  of year  $a$  at site  $s$ , and  $M$  is the total number of years of available weather data. In our study,  $M = 7$  or  $8$  as we use data from March 2010 to August 2017 to build the weather baseline while the air quality data were from March 2013 to August 2017. The key in removing the meteorological confounding is to put the comparison under the same meteorological baseline. Hence, we define the adjusted average concentration of year  $i$  at season  $j$  with respect to the baseline probability density as  $\mu_{ij}(s) = \int \int m_{ij}(x, w; s) f_{\cdot j}(x, w; s) dx dw$ . Suppose we want to measure the difference between the average pollution between years  $i$  and  $l$ . The one based on the adjusted averages is

$$\mu_{ij}(s) - \mu_{lj}(s) = \int \int (m_{ij}(x, w; s) - m_{lj}(x, w; s)) f_{\cdot j}(x, w; s) dx dw,$$

which is solely due to the change in the two regression functions  $m_{ij}(x, w; s) - m_{lj}(x, w; s)$ . To estimate the adjusted average concentration in (3), we first estimate the function  $m_{ij}(\cdot)$  by using a nonparametric kernel regression estimator [14, 21]:

$$\hat{m}_{ij}(x, w; s) = \frac{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s)=w) Y_{ijt}(s)}{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s)=w)},$$



where  $\mathbb{1}(\cdot)$  is the indicator function for wind direction.  $K_h(z)$  is a multi-dimensional kernel function that is a product of the univariate Gaussian kernel  $k(u) = (2\pi)^{-1/2} \exp(-u^2/2)$  with bandwidths  $\{h_1, h_2, \dots, h_5\}$ . In practice, the bandwidths are chosen by the cross validation (CV) algorithm [7, 21].

By replacing baseline density with its empirical version, the adjusted average is estimated by

$$\hat{\mu}_{ij}(s) = (\sum_{a=1}^M n_{aj})^{-1} \sum_{a=1}^M \sum_{t=1}^{n_{ij}} \sum_{w \in \Omega} \hat{m}_{ij}(X, W) \mathbb{1}(W = w),$$

The adjusted average concentration of a city is obtained by averaging the adjusted averages at all monitoring sites in the city. To obtain the standard errors of the estimated adjusted average, the bootstrap re-sampling method is applied to generate repeated copies of the adjusted means. They can be obtained by using the bootstrap method which consists of two components. First we conduct the temporal block bootstrap to resample the meteorological data, followed by the regression bootstrap to generate the response variables. To illustrate the idea, we only present the case of the adjusted mean  $\mu_{ij}$ .

For a specific city, suppose there are  $S$  air quality monitoring sites. Denote  $T = \{1, 2, \dots, n_{ij}\}$  to be a particular time period of  $(Y_{ijt}(s), X_{ijt}(s), W_{ijt}(s))$  for season  $j$  of year  $i$ . Divide  $T$  into continuous blocks, each with length  $l = 12(h)$ , with  $B_1 = \{1, 2, \dots, 12\}$ ,  $B_2 = \{13, 14, \dots, 24\}$ , ...,  $B_{n_{ij}/12} = \{n_{ij} - 11, n_{ij} - 10, \dots, n_{ij}\}$ . Then, we re-sample these  $n_{ij}$  blocks with replacement, and paste them together to form a new bootstrap time series  $T^* = (t_1^*, t_2^*, \dots, t_{n_{ij}}^*)$ . For the newly-generated time span, we obtain new bootstrap samples of the meteorological variables  $\{X_{ijt}^*(s), W_{ijt}^*(s)\}$ , which are realizations of  $\{X_{ijt}(s), W_{ijt}(s)\}$  at time  $T^*$ . Given the resampled meteorological variables, we first generate the regression and variance functions  $\hat{m}_{ij}(X_{ijt}^*(s), W_{ijt}^*(s); s)$  and  $\hat{\sigma}_{ij}^2(X_{ijt}^*(s), W_{ijt}^*(s); s)$  from the following equations:

$$\begin{aligned} \hat{m}_{ij}(x, w; s) &= \frac{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w) Y_{ijt}(s)}{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w)}, \\ \hat{\sigma}_{ij}^2(x, w; s) &= \frac{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w) \hat{\epsilon}_{ijt}^2(s)}{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w)}, \end{aligned}$$

where  $\hat{\epsilon}_{ijt}(s) = Y_{ijt}(s) - \hat{m}_{ij}(X_{ijt}^*(s), W_{ijt}^*(s); s)$ . Then the re-sampled standard errors  $\{\hat{\epsilon}_{ijt}^*(s)\}$  are derived from multivariate normal distribution with mean 0 and sample covariance  $\hat{\Sigma}_{ij}$ , which is derived by:

$$\hat{\Sigma}_{ij} = (n_{ij})^{-1} (\sum_{t=1}^{n_{ij}} \hat{\epsilon}_{ijt} \hat{\epsilon}_{ijt}^T) - ((n_{ij})^{-1} \sum_{t=1}^{n_{ij}} \hat{\epsilon}_{ijt}) ((n_{ij})^{-1} \sum_{t=1}^{n_{ij}} \hat{\epsilon}_{ijt})^T,$$



where  $\hat{\epsilon}_{ijt} = \{\hat{\epsilon}_{ijt}(1), \dots, \hat{\epsilon}_{ijt}(S)\}$  are the standard residuals. Then the bootstrap realization of  $Y_{ijt}(s)$  is achieved by:

$$Y_{ijt}^*(s) = \hat{m}_{ij}(X_{ijt}^*(s), W_{ijt}^*(s); s) + \hat{\sigma}_{ij}^2(X_{ijt}^*(s), W_{ijt}^*(s); s) \cdot \hat{\epsilon}_{ijt}^*(s).$$

For the  $b$ -th bootstrap replication, repeat the calculation of the adjusted average for each site  $s$ . Then, the bootstrap sample of the adjusted mean for the city is averaged over all sites:  $\hat{\mu}_{ij}^{b*} = S^{-1} \sum_{s=1}^S \hat{\mu}_{ij}^{b*}(s)$ . Suppose the total number of bootstrap samples to be  $B$ . The variance of  $\hat{\mu}_{ij}$  is estimated by  $\frac{1}{B-1} \sum_{b=1}^B \{\hat{\mu}_{ij}^{b*} - \frac{1}{B} \sum_{b=1}^B \hat{\mu}_{ij}^{b*}\}^2$ .

In addition to the adjusted average, adjusted 90% quantiles of the concentrations can be calculated for each city in a similar way which can measure the 10% most severe pollution circumstances. We need to construct the adjusted distribution  $D_{ij}(y; s)$  of a pollutant by replacing  $m_{ij}(x, w; s)$  with  $F_{ij}(y|x, w; s) = \mathbb{E}(\mathbb{1}_{(Y_{ijt}(s) < y)} | X_{ijt}(s) = x, W_{ijt}(s) = w; s)$ . Therefore, we have

$$D_{ij}(y; s) = \int \int F_{ij}(y|x, w; s) f_{.j}(x, w; s) dx dw.$$

Similarly,  $F_{ij}(y|x, w; s)$  is estimated by the kernel smoothing method, which is given by

$$\hat{F}_{ij}(y|x, w; s) = \frac{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w) G_{h_0}(Y_{ijt}(s) - y)}{\sum_t K_h(x - X_{ijt}(s)) \mathbb{1}(W_{ijt}(s) = w)},$$

where  $G_{h_0}(z) = \int_{-\infty}^{z/h_0} g(u) du$  is the integration of univariate Gaussian kernel function.

Therefore,  $D_{ij}(y; s)$  can be estimated empirically by:

$$(\sum_{a=1}^M n_{aj})^{-1} \sum_{a=1}^M \sum_{t=1}^{n_{ij}} \sum_{w \in \Omega} \hat{F}_{ij}(y|X, W; s) \mathbb{1}(W = w).$$

Thus, for any  $q \in (0, 1)$ ,  $\hat{D}_{ij}^{-1}(q; s)$  denotes the adjusted  $q$ -th quantile concentration.

The searching method for a 90-th quantile is established by Dichotomy method. Starting with 0 and some value larger than the highest value of a pollutant, we first take average of the two most extreme values, and calculate the corresponding adjusted quantile through the formula of  $\hat{D}_{ij}(y; s)$ . If the calculated quantile  $q$  is lower than 90%, we need to search the optimal value in the interval between this average and the larger value. Otherwise we will search in the other direction. By applying this procedure for several iterations until the difference between  $q$  and 90% is lower than a value small enough, we can stop and choose the value at the final iteration as the 90% adjusted quantile.

To make a step further, we propose a new adjusted average function with respect to a meteorological variable, which we call the adjusted curve. This curve represents

the relationship between a pollutant's average concentration and a meteorological variable after properly controlling the rest meteorological variables. Often one wants to quantify the relationship between a pollutant's average concentration with respect to a meteorological variable, such as the air temperature or dew point.

Directly regressing the pollutant data on the dew point without controlling the other meteorological variables will lead to confounded regression estimation. The baseline density function of the weather variables can be used to produce the adjusted curves free of the confounding by the other variables. Let  $X^{(k)}$  be a meteorological variable that we would like to produce the adjusted curve, and  $X^{(-k)}$  represent the rest of the meteorological variables. If  $X^{(k)}$  is not the wind direction, the adjusted average curve with respect to  $X^{(k)}$  can be defined as below:

$$\mu_{ij}(X^{(k)}) = \int \int m_{ij}(x^{(k)}, x^{(-k)}, w) f_{\cdot j}(x^{(k)}, x^{(-k)}, w) dx^{(-k)} dw,$$

where the integration is carried out without the  $x^{(k)}$ . If  $X^{(k)}$  is a wind direction, then

$$\mu_{ij}(W) = \int m_{ij}(x, w) f_{\cdot j}(x, W) dx.$$

To estimate the adjusted curves, if  $X^{(k)}$  is not the wind direction, one can estimate  $\mu_{ij}(X^{(k)}, s)$  by

$$\mu_{ij}(X^{(k)}, s) = \frac{1}{N(X^{(k)}, b)} \sum_{a=1}^M \sum_{X \in A(X^{(k)}, b)} \sum_{w \in \Omega} \hat{m}_{ij}(X, W) \mathbb{1}(W = w),$$

where  $A(X^{(k)}, b)$  is a cross-section of the  $X$ -domain which has its  $k$ -th component within a  $b$  neighborhood of  $X^{(k)}$ ,  $N(X^{(k)}, b)$  denotes the number of data falling in  $A(X^{(k)}, b)$ , and  $b$  is a smoothing parameter that defines the size of the cross section. If  $X^{(k)}$  is wind direction, then

$$\mu_{ij}(W, s) = \frac{1}{N(W)} \sum_{a=1}^M \sum_{t=1}^{n_{aj}} \hat{m}_{ij}(X, W) \mathbb{1}(W = w),$$

where  $N(W)$  denotes the number of data which has wind direction  $W$ .

The adjusted average curves provide information on the pollutant's concentration with respect to the variable  $X^{(k)}$ . If one further integrate these equations with respect to the marginal baseline probability density of  $X^{(k)}$ , then  $\mu_{ij}(s)$  can be obtained.

## 2.6 Main results and conclusions

To assess the air quality improvements of Tianjin since 2013, we apply the adjustment method to remove the meteorological confounding and calculate the meteorological

adjustment concentrations for  $\text{PM}_{2.5}$ ,  $\text{SO}_2$ ,  $\text{NO}_2$  and  $\text{O}_3$  from spring 2013 to summer 2017. The results and conclusions presented here are based on the adjusted average concentrations.

### Findings on $\text{PM}_{2.5}$ and $\text{SO}_2$

$\text{PM}_{2.5}$  is one of the major urban air pollutants and can cause great harm to human health. In 2013,  $\text{PM}_{2.5}$  replaced  $\text{PM}_{10}$  as the primary air pollutant in China, and has been a key target air pollutant in the “National Ten Point Plan”. The seasonal adjacent-year comparison of  $\text{PM}_{2.5}$  in Tianjin from spring 2013 to spring 2017 is illustrated in Figure 2.4 below:

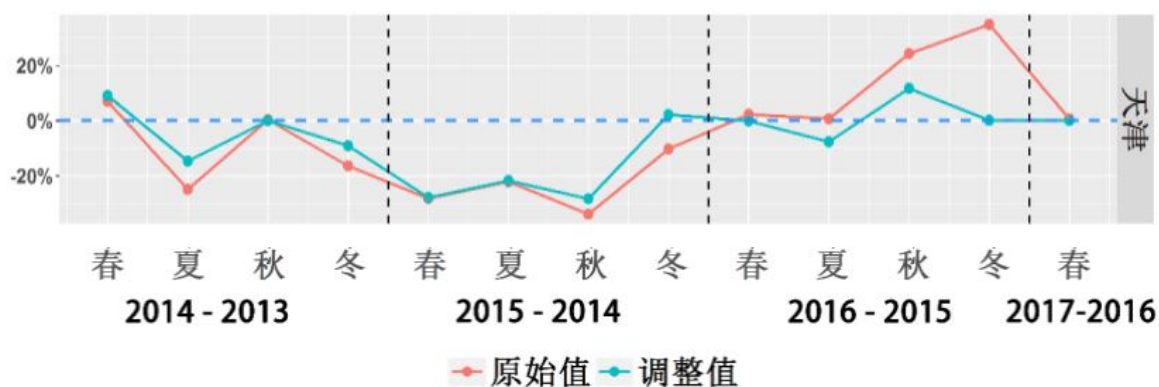


图 2.4: Seasonal adjacent-year comparison of  $\text{PM}_{2.5}$  in Tianjin from spring 2013 to spring 2017.

From Figure 2.4, we see a great reduction in the adjusted average of  $\text{PM}_{2.5}$  in Tianjin in spring, summer and autumn from year 2014 to 2015. However,  $\text{PM}_{2.5}$  concentration attained a slightly rebound in autumn 2016 compared to 2015.

Sulfur dioxide ( $\text{SO}_2$ ) is another major air pollutant which is not only harmful to human [28], but also a key gaseous precursor to  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$ . As coal has been the main source of energy for industrial and domestic use in China at large,  $\text{SO}_2$  is mainly the result of coal burning for power generation, winter heating, heavy industrial manufacturing for iron, steel and building materials. There are strong seasonality in the  $\text{PM}_{2.5}$  and  $\text{SO}_2$  concentrations with the winter achieving the highest level and summer the lowest in Tianjin. Another different feature between  $\text{PM}_{2.5}$  and  $\text{SO}_2$  is the seasonal variation that  $\text{SO}_2$  is larger than that of  $\text{PM}_{2.5}$  as represented by the winter to summer concentration ratios, being around 4. A reason for the much higher winter to summer ratio is the winter heating which burns much coal and generate excessive  $\text{SO}_2$ .

It is observed that the differences between the raw and adjusted averages were much larger in winter than the other seasons for both  $\text{PM}_{2.5}$  and  $\text{SO}_2$ , which indicates that the influence of meteorological conditions in winter was much larger than the other seasons.

The most striking aspect of the Tianjin's air-quality data was a profound decline in  $\text{SO}_2$ , representing over 60% reduction from 2013 to 2016. This decrease rate achieves the highest among all major cities in Beijing-Tianjin-Hebei region. The substantial decline in  $\text{SO}_2$  shows the significant achievement in reducing the consumption of coal by several initiatives. These initiatives include the elimination of high energy consumption and high polluting equipments, efforts to phase out coal with the natural gas in power generation and winter heating in major cities, and to increase the usage of the natural gas and electricity in the rural domestic cooking and winter heating. However, the decline in  $\text{SO}_2$  had not translated to a significant reduction in  $\text{PM}_{2.5}$  for Tianjin (with only a cumulative decrease rate lower than 20% for  $\text{PM}_{2.5}$  from year 2013 to 2016), indicating the complexity of PM generation process and the important roles of other precursors played in the process.

### **Findings on $\text{NO}_2$ and $\text{O}_3$**

Nitrogen dioxide ( $\text{NO}_2$ ) is a common air pollutant and also a precursor to nitrates and the ground level  $\text{O}_3$ . Under the strong sunlight,  $\text{NO}_2$  reacts with oxygen in the air, producing another air pollutant, ozone [60]. Different from the stratospheric ozone that protects humans from ultraviolet radiation, ground ozone causes harm to human respiratory and nerve system. We applied the same adjustment method to the  $\text{NO}_2$  and the 8-h (from 12pm to 7pm) ozone data, and calculated their seasonal adjusted averages for Tianjin. In the adjustment of the 8-h ozone, we use the logarithm of the cumulative UVB to replace PRES as the correlation between PRES and the  $\text{O}_3$  levels is small. It is seen that  $\text{NO}_2$  follows a similar seasonal pattern to  $\text{PM}_{2.5}$  and  $\text{SO}_2$  in that the concentrations are the highest in winter and the lowest in summer. The seasonal pattern for  $\text{O}_3$  is opposite that the highest levels occurs in summer and spring and the lowest in winter, which is consistent to the annual intensity cycle of the ultraviolet radiation from the sun.

Similar to  $\text{PM}_{2.5}$  and  $\text{SO}_2$ , the difference between the unadjusted and adjusted seasonal average concentration for  $\text{NO}_2$  is largest in winter, showing the impact of the meteorological condition. But for  $\text{O}_3$ , it shows that larger differences between the unadjusted and the adjusted 8-h  $\text{O}_3$  concentrations happened in the spring when it tended

to have larger variations among the key meteorological variables that affect the ozone level.

Furthermore, we see that the  $\text{NO}_2$  concentrations in Tianjin had decreased, but at much smaller rates compared to those of  $\text{PM}_{2.5}$  and  $\text{SO}_2$  (with only a cumulative decrease rate lower than 5% for  $\text{NO}_2$  from year 2013 to 2016). The situation for the 8-h  $\text{O}_3$  is a different scene: a steady regional-wide increase in the past four years. If we focus on the 8-h  $\text{O}_3$  in the spring and summer, the situation becomes even more alarming.

It is clear that the ground ozone concentration has worsened significantly in the last four and half years. Ozone has taken over  $\text{PM}_{2.5}$  as the primary pollutant in summer and spring.  $\text{NO}_x$  and the volatile organic components (VOC) constitute are known to be the important precursors of the ground ozone generation [60]. However, components of VOC are not measured by the national air quality monitoring network. The very limited reduction in  $\text{NO}_2$  over Tianjin means that there had been sustained supply of the precursors for  $\text{O}_3$  generation. Although  $\text{NO}_2$  had not increased, the regional-wide reduction in the PMs had increased the intensity of the ultraviolet radiation, which makes the oxidization of  $\text{NO}_2$  for  $\text{O}_3$  generation more efficient, indicating a negative effect of improving PM situation on  $\text{O}_3$ . This is inevitable if the  $\text{NO}_2$  and its precursors are left to be un-managed, which has been the case for Tianjin as reflected by the sluggish results of  $\text{NO}_2$ .



## 第三章 Efficient semiparametric estimation in spatio-temporal models

### 3.1 My contributions on this research:

I completed all this research under the supervision of Prof. Fang Yao. The work includes model construction, parameter estimation, simulation studies, real data analysis and necessary proofs.

### 3.2 Overview

Semiparametric regression problems have received much attention in the field of longitudinal clustered data. A common form of semiparametric regression model for clustered data is:

$$\hat{Y}_{ij} = X_{ij}^T \beta + \theta(T_{ij}) + \epsilon_{ij},$$

where the model parameters and coefficients are specified in the Introduction section (Chapter 1).

Based on the pioneering work of [40], the traditional kernel-profile estimation equation procedure was improved by correctly specifying the covariance matrix in the semiparametric regression model in order to obtain efficient estimators. The foundation work regarding this method is referred to [66]. Wang introduced a marginal nonparametric kernel regression method which accounts for within-subject correlations. The paper proposed an alternative kernel smoothing method which make proper use of correlation among observations within each cluster to estimate the nonparametric function  $\theta(\cdot)$  in the model above. The lowest variance regarding  $\beta$  can be obtained by assuming true correlation structure. The asymptotic variance uniformly lowered than that in the case of working independence. The main idea behind this novel method can be explained as follows: once an observed data  $j$  within a cluster attains its  $t$ -value in the  $h$ -neighborhood of  $t_0$  and is used to estimate  $\theta(t_0)$ , all the points in this cluster will be utilized. The contributions of all data points except  $j$  to the local estimate of  $\theta(t_0)$  are derived through the corresponded residuals. This estimation approach is primarily based on the case where numbers of observations within each cluster is finite.

A refined work of [40] on semiparametric regression in clustered data is seen in [67], which incorporated the new estimation method of [66] on the nonparametric function  $\theta(\cdot)$ . The traditional nonparametric estimator in [40] was replaced by a novel iterative kernel estimator. The parametric and nonparametric terms in the model were assumed to be dependent with each other, and the final estimate  $\beta$  was semiparametric efficient under the Gaussian case and had less variation than the estimator under working independence. It was shown that the estimator for  $\beta$  attains the semiparametric efficiency bound only if the within-cluster correlation matrix is specified correctly under the semiparametric framework. Application of this method on CD4 cell counts of HIV helps examining the efficacy of the proposed methodology. Furthermore, undersmoothing was no longer used to construct root-n consistent estimates of  $\beta$  when accounting for correlations. The numerical results suggested that the proposed method performs well in finite samples and outperforms the WI circumstance. In addition, Lin and her collaborators generalized these methods to more general repeated measures problems [41].

Modeling of covariance matrix is also important. Research work on modeling covariance in longitudinal data includes [13, 70], which used statistical methods ranging from functional linear models to quasi-likelihood and generalized variance approach. However, most of these researches assumed the observations were made on a regular time grid, and therefore are not suitable for longitudinal data collected at irregular and subject-specific times. More recently, Fan and his collaborators developed a semiparametric quasi maximum likelihood method to model and estimate the longitudinal covariance functions [13]. They studied the case in which the observation times are irregular on a continuous time interval. The quasi maximum likelihood method provides a good trade-off between model flexibility and estimation efficiency, but it relies on correctly specifying the parametric model for the correlation function. Nonparametric estimation of covariance was further studied which developed fast in the area of functional data analysis. The covariance function was modelled as a smooth function and estimated by a kernel smoother. Some recent work on this topic includes [20, 74].

A natural extension of the existing semiparametric regression models lies in combining the efficient semiparametric estimator with nonparametric covariance estimation in the framework of longitudinal clustered data. Li (2011) solved this problem by estimating the covariance matrix with bivariate kernel smoothers [37]. A three-step estimation procedure was developed to refine the estimate of  $\beta$  and  $\theta$ . This method is robust



against misspecification of covariance models. Adjustment on the kernel covariance estimator by spectral decomposition and truncation techniques was also implemented in this work which helps regularize estimation of the covariance matrix. He showed that kernel covariance estimation provided uniformly consistent estimators for the within-subject covariance matrices, and the semiparametric profile estimator with substituted nonparametric covariance was still semiparametrically efficient. The proposed estimator can achieve the semiparametric information bound through comparison to the asymptotic variance derived in [67]. This method was also applied on CD4 count data from an AIDS clinical trial. Most extensions to this article were based on change-point detection in longitudinal data, including [71, 72].

However, few studies considered the efficiency of the parametric estimators in spatio-temporal models. The error terms of the spatio-temporal models in most literatures were assumed to independent identically distributed, where no spatial correlation was considered. Therefore, a straightforward idea lies in estimating the spatio-temporal covariance matrix in order to achieve better estimation of the parametric counterparts. Since estimating the whole covariance requires massive computation and is difficult to implement, we can tackle with the spatial and temporal counterparts separately by assuming weak dependence over time. This concept was introduced and discussed in [23].

A most direct relaxation of independence is  $m$ -dependence. Suppose  $X_n$  is a sequence of elements taking values in a measurable space  $S$ . Denote  $F_k^- = \{\dots, X_{k-1}, X_k\}$  and  $F_k^+ = \{X_{k+1}, X_{k+2}, \dots\}$  as the  $\sigma$ -algebras generated by the observations up to time  $k$  and after time  $k$ . Then the sequence  $X_n$  is defined to be  $m$ -dependent if for any  $k$ , the two  $\sigma$ -algebras are independent. Under this weak assumption, the covariance matrix can be reduced to a banded block diagonal matrix which is much more convenient for estimation. We adopt a profile-kernel estimating procedure and show that by deriving a nonparametric kernel estimate of the banded spatio-temporal covariance matrix, the semiparametric estimator is asymptotically efficient and achieves the semiparametric efficiency bound.

The rest of this chapter is organized as follows. In Section 3, we introduce the framework of our semiparametric spatio-temporal model. In Section 4, specific estimation methodology is proposed under this setting. In Section 5, we derive asymptotic properties for the model parameters. In Section 6 and 7, we will illustrate the per-

formance of our novel method through simulation studies and two data examples (air pollution in Tianjin and COVID-19 data in Hubei Province).

### 3.3 Model setup

Suppose all spatial-temporal observations are made on a fixed spatial region  $S$  and time span  $T$ . The  $m$  spatial locations are denoted by  $s_1, s_2, \dots, s_m \in S$ , and the  $n$  time points are denoted as  $t_1, t_2, \dots, t_n \in T$ . For spatial location  $s_j$  at time  $t$ , we observe a response variable  $Y_t(s_j)$  and a  $p$ -dimensional covariate  $X_t(s_j)$ , where  $p$  represents the number of covariates. Mathematically, we model  $Y_t(s)$  and  $X_t(s)$  as random processes defined on a spatial region  $S$ . Then  $Y_t$  and  $X_t$  can be viewed as observations on these processes at randomly-distributed spatial locations. A semiparametric spatio-temporal model is constructed as below:

$$Y_t(s) = X_t^T(s)\beta + \theta(s) + \epsilon(s); s \in S, t \in T,$$

where  $\beta$  is an unknown coefficient vector and  $\theta(\cdot)$  is an unknown smooth function which represents the spatial main effect. We also assume the covariance of the response variable conditional on the covariates is a bivariate positive semidefinite function:

$$R(s_1, s_2) = \text{cov}\{Y_t(s_1), Y_t(s_2) | X_t(s), s \in S\} = \text{cov}\{\epsilon_t(s_1), \epsilon_t(s_2)\}; s_1, s_2 \in S.$$

We model the covariance at each time point as a bivariate nonparametric function, which is smooth except the points on the diagonal line, allowing for possible nugget effects. We can decompose the error term into two separate independent counterparts:  $\epsilon_t(s) = \epsilon_{t0}(s) + \epsilon_{t1}(s)$ , where  $\epsilon_{t0}(s)$  is a spatial process with smooth covariance function  $R_0(s_1, s_2)$  and  $\epsilon_{t1}(s)$  is a white noise process caused by measurement error. Denote  $\sigma_1^2(s) = \text{var}\{\epsilon_{t1}(s)\}$ , then we have  $R(s_1, s_2) = R_0(s_1, s_2) + \sigma_1^2(s_1)I(s_1 = s_2)$ , where  $I(\cdot)$  is an indicator function. Since correlation also exists in the temporal process within the spatial-temporal observations, we assume stationarity and weak dependence over time  $T$ . The error term  $\epsilon(s)$  is most often assumed to follow some common time series patterns, such as  $\text{AR}(p)$ .

### 3.4 Estimation methodology

As in most semiparametric regression settings [41, 66, 67], we first apply a working independence estimator, assuming that the covariance matrix at each time point is an

identity. For a given  $\beta$ , coefficients  $(\hat{\alpha}_0, \hat{\alpha}_1)$  solves the local estimating equation below:

$$\frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m S_j(s) K_{h_1}(s_j - s) \{Y_t(s_j) - [X_t(s_j)\beta + \theta(s)]\} = 0,$$

where  $S_j(s) = \{1, \frac{s_j - s}{h_1}\}$  and  $K(\cdot)$  is a kernel function. The tuning parameter  $h_1$  is the bandwidth used in the kernel function. Then we have  $\hat{\theta}(s, \beta) = \hat{\alpha}_0(s, \beta)$ , which is the kernel estimate of  $\theta(s)$ .

With the estimated spatial main effect, we estimate parameter  $\beta$  by the profile estimation equation below:

$$\frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \frac{\partial [X_t^T(s_j)\beta + \hat{\theta}(s_j, \beta)]}{\partial \beta} \{Y_t(s_j) - [X_t(s_j)\beta + \hat{\theta}(s_j, \beta)]\} = 0.$$

Iterate between these two estimating equations several times and we will get an initial estimate for  $\beta$  and  $\theta(s)$ .

In order to count down the spatial-temporal correlation within the total observations, we poll all the observations together and interpolate the covariance matrix into both the kernel and the profile estimating equations. For a given estimated covariance  $\hat{\Sigma}$ , the estimating results can then be refined by iterating between the following kernel-profile estimating equations:

$$\sum_{j=1}^m S_j(s) K_{h_2}(s_j - s) \hat{\Sigma}^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \theta(s)]\} = 0,$$

$$\sum_{j=1}^m \frac{\partial [X_t^T(s_j)\beta + \hat{\theta}(s_j, \beta)]}{\partial \beta} \hat{\Sigma}^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \hat{\theta}(s_j, \beta)]\} = 0,$$

The estimation procedure is the same as the previous working independence estimator. The main issue lies in the comparing the efficiency of  $\hat{\beta}$  before and after interpolating the estimated spatio-temporal covariance matrix in our kernel-profile estimating equations. Details in bandwidth selection is referred to [15, 54].

**Remark.** We have mainly two choices of bandwidth selection: fixed bandwidth and proportional bandwidth. Under the case of a spatial domain, fixed bandwidth refers to fixed distances, whereas proportional bandwidth means a specific proportion of neighboring spatial points considered (in consideration). When spatial locations are evenly distributed (for example: grid data), little difference exists between these two bandwidth selection methods. However, if the spatial points are irregularly spaced, proportional bandwidth usually outperforms fixed bandwidth in estimation accuracy. See the results in the Simulation and Real data analysis parts for illustration.

Since correlation exists on both spatial and temporal basis, we need to estimate the spatio-temporal covariance as a whole. However, when the temporal span  $n$  is long enough, estimation of the high-dimensional covariance matrix can be cumbersome where large amount of calculation is inevitable. To simplify this problem, we can reduce the calculation complexity by assuming weak dependence over time. The idea of weakly dependence comes from [23]. Since the assumption of independence is often too strong to be realistic in many applications, especially if data are collected sequentially over time, we can derive a most typical relaxation of independence:  $m$ -dependence. We assume correlation over time decay sufficiently fast, urging the independence between  $\sigma$ -algebras before time  $t$  and after time  $t+m$ . With this assumption, the estimation covariance matrix can be reduced to a  $m$ -banded block matrix. The tuning parameter  $m$  should be considered as a smaller order of  $n$  in estimation. In this way, the amount of estimation work can be greatly reduced since we do not need to consider the blocks outside the neighbouring  $m$  blocks for each time point  $t$ .

In addition, by assuming stationarity over time, we can estimate each block of the total covariance through bivariate local linear smoothers. We denote them as diagonal block covariance, lag-1 block covariance, lag-2 block covariance, ..., lag- $m$  block covariance, respectively. The number of blocks need to be estimated is  $n, n-1, n-2, \dots, n-m$  corresponded to each lagged block covariance. Blocks with each specific lag  $(1, 2, \dots, m)$  should be estimated separately. Furthermore, each block estimation need to be divided into two counterparts: the smooth covariance function and the nugget effect, which also requires separate estimation.

Denote  $\hat{\epsilon}_t(s_j) = Y_t(s_j) - [X_t^T(s_j)\hat{\beta} + \theta(s_j, \hat{\beta})]$ , which is the estimate of  $\epsilon_t(s_j)$ . We first estimate the smooth part of the diagonal block covariance by using bivariate local linear smoothing. Let  $\hat{R}_{00}(s_1, s_2) = \hat{\alpha}_0$ , where  $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$  minimizes:

$$\frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \sum_{j' \neq j} \{\hat{\epsilon}_t(s_j)\hat{\epsilon}_t(s_{j'}) - \alpha_0 - \alpha_1(s_j - s_1) - \alpha_2(s_{j'} - s_2)\}^2 K_{h_3}(s_j - s_1) K_{h_3}(s_{j'} - s_2).$$

Similarly, we can estimate the smooth function part of lag-1 block covariance, lag-2 block covariance, ..., and lag- $m$  block covariance with the same procedure by assuming stationarity over time. For lag- $r$  block covariance ( $1 \leq r \leq m$ ), let  $R_{r0}(s_1, s_2) = \hat{\beta}_0$ , where  $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$  minimizes:

$$\frac{1}{n-r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j' \neq j} \{\hat{\epsilon}_t(s_j)\hat{\epsilon}_{t+r}(s_{j'}) - \beta_0 - \beta_1(s_j - s_1) - \beta_2(s_{j'} - s_2)\}^2 K_{h_3}(s_j - s_1) K_{h_3}(s_{j'} - s_2)$$

According to Hall et al. (2006), we can derive the analytical form of the bivariate local linear smoothers stated (illustrated) above. Some necessary notations are needed before moving on.

Define

$$N_r = (n - r)m(m - 1),$$

$$\begin{aligned} S_{pq}^r &= \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} \left( \frac{s_j - s_1}{h_3} \right)^p \left( \frac{s_k - s_2}{h_3} \right)^q K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2) \\ &= \frac{1}{m(m-1)} \sum_{j=1}^m \sum_{j \leq k} \left( \frac{s_j - s_1}{h_3} \right)^p \left( \frac{s_k - s_2}{h_3} \right)^q K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2), \end{aligned}$$

$$R_{pq}^r = \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} \hat{\epsilon}_t(s_j) \hat{\epsilon}_{t+r}(s_k) \left( \frac{s_j - s_1}{h_3} \right)^p \left( \frac{s_k - s_2}{h_3} \right)^q K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2).$$

Denote the following terms:

$$A_1^r = S_{20}^r S_{02}^r - \{S_{11}^r\}^2,$$

$$A_2^r = S_{10}^r S_{02}^r - S_{01}^r S_{11}^r,$$

$$A_3^r = S_{01}^r S_{20}^r - S_{10}^r S_{11}^r,$$

$$B^r = A_1^r S_{00}^r - A_2^r S_{10}^r - A_3^r S_{01}^r.$$

Then we can express the smooth part estimate of lag- $r$  block covariance ( $0 \leq r \leq m$ ) as below:

$$\hat{R}_{0r}(s_1, s_2) = (A_1^r R_{00}^r - A_2^r R_{10}^r - A_3^r R_{01}^r)(B^r)^{-1}.$$

Furthermore, local linear smoothing is used to estimate the nugget effect. For the diagonal elements of  $R_{0r}(s_1, s_2)$ ,  $0 \leq r \leq m$ , let  $\hat{\sigma}^2(s) = \hat{\gamma}_0$ , where  $(\hat{\gamma}_0, \hat{\gamma}_1)$  minimizes:

$$\frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \{\hat{\epsilon}_t^2(s_j) - \gamma_0 - \gamma_1(s_j - s)\}^2 K_{h_4}(s_j - s).$$

Therefore, we obtain the estimate of the nugget effect, which is denoted by  $\hat{\sigma}^2(s)$ . Similar estimation procedures can be derived for nugget effects with the assumption of stationarity over the temporal basis. We can also obtain the analytical form the nugget effect estimator.

Define the following terms:

$$U_r = (n - r)m,$$

$$S_p^r = \frac{1}{U_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \left(\frac{s_j - s}{h_4}\right)^p K_{h_4}(s_j - s) = \frac{1}{m} \sum_{j=1}^m \left(\frac{s_j - s}{h_4}\right)^p K_{h_4}(s_j - s),$$

$$R_p^r = \frac{1}{U_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \hat{\epsilon}_t(s_j) \left(\frac{s_j - s}{h_4}\right)^p K_{h_4}(s_j - s).$$

Then we can derive the nugget effect estimate of lag- $r$  block covariance ( $0 \leq r \leq m$ ) as below:

$$\hat{\sigma}_r^2(s) = \frac{S_2^r R_0^r - S_1^r R_1^r}{S_0^r S_2^r - (S_1^r)^2}.$$

So far, the estimation of the lag- $r$  block covariance  $R_r(s_1, s_2)$  ( $0 \leq r \leq m$ ) can be derived as:

$$\hat{R}_r(s_1, s_2) = \hat{R}_{0r}(s_1, s_2) \mathbb{1}(s_1 \neq s_2) + \hat{\sigma}_r^2(s_1) \mathbb{1}(s_1 = s_2).$$

By summing up all the estimated lag- $r$  block covariance ( $0 \leq r \leq m$ ), we can derive the estimation of the spatio-temporal covariance matrix  $\hat{\Sigma}$ .

**Remark.** When estimating the spatio-temporal covariance matrix, the kernel covariance estimator is not guaranteed to be positive semidefinite, and therefore some adjustment is needed to enforce this condition. This is especially important when the sample size is relatively small, and the adjustment procedure would help by further regularizing the covariance estimator. Usually, we make a spectral decomposition of the estimated covariance, truncate the negative eigenvalue components, and interpolate the nonnegative components back into the expression of spectral decomposition. Li (2011) discusses this procedure in detail [37]. This method can effectively regularize the covariance matrices and stabilize the final estimator. Furthermore, the implementation of nugget effect estimation described in [74] is also applicable in our setting with slight modification.

### 3.5 Asymptotic properties

In order to examine whether we can obtain an efficient estimator for the parametric terms in our semiparametric spatio-temporal setting, we need to derive the asymptotic properties of the main estimators:  $\hat{\beta}$ ,  $\hat{\theta}(\cdot)$ , and  $\hat{\Sigma}$ . The main focus lies in comparing the efficiency of  $\beta$  before and after interpolating the estimated spatio-temporal covariance matrix. Recall that  $h_1$  and  $h_2$  are the bandwidths in the kernel function w.r.t the working independence estimator and the refined estimator (kernel-profile estimating

equation after incorporating the estimated covariance), respectively.  $h_3$  and  $h_4$  represent the bandwidths used in estimating the smooth and nugget effect part of the banded block diagonal covariance matrix, respectively. Before giving out the main asymptotic properties, some necessary assumptions are needed:

**Assumption 1.**  $h_1 \sim n^{-v_1}$ ,  $\frac{1}{5} < v_1 < \frac{1}{3}$  [40]

**Assumption 2.**  $h_2 \rightarrow 0$ ,  $nh_2^4 \rightarrow 0$ ,  $\frac{\log(1/h_2)}{nh_2} \rightarrow 0$  [67]

**Assumption 3.**  $mh_3^2 \rightarrow 0$ ,  $mh_4^2 \rightarrow 0$ ,  $\frac{m^2 \log(n)}{nh_3^2} \rightarrow 0$ ,  $\frac{m^2 \log(n)}{nh_4^2} \rightarrow 0$

**Lemma 1.** (Convergence rate of  $\hat{\beta}_{WI}$  and  $\hat{\theta}(s)$  under WI condition)

Denote  $\hat{\beta}_{WI}$  and  $\hat{\theta}_{WI}(s, \hat{\beta}_{WI})$  to be the estimator of the parametric and nonparametric terms under the case of working independence, respectively. Under Assumption 1, we obtain the asymptotic rate of  $\hat{\beta}_{WI} - \beta$  and  $\sup |\hat{\theta}_{WI}(s, \hat{\beta}_{WI}) - \theta(s)|$  as below:


$$\begin{aligned}\hat{\beta}_{WI} - \beta &= O_p(n^{-1/2}), \\ \sup |\hat{\theta}_{WI}(s, \hat{\beta}_{WI}) - \theta(s)| &= O_p(h_1^2 + (\frac{\log(n)}{nh_1})^{1/2}).\end{aligned}$$

This result is essential in deriving the asymptotic properties of the estimated spatio-temporal covariance matrix. Proof of Lemma 1 under this framework is similar to that in clustered data [40, 67]. The main technique used in proving the asymptotic rate of  $\sup |\hat{\theta}_{WI}(s, \hat{\beta}_{WI}) - \theta(s)|$  lies in correctly deriving the asymptotic expansion of  $\hat{f}_k(s, \hat{\beta}_{WI})$  at each iteration  $k$

**Lemma 2.** (~~Estimation rate~~ of the estimated covariance matrix)

Under Assumption 2 and the assumption of  $m$ -dependence over time, where we also assume  $m = o(n^{1/2})$ , for any two spatial locations  $s_1, s_2 \in S$ , the asymptotic rates of  $\sup |\hat{R}_0(s_1, s_2) - R_0(s_1, s_2)|$  and  $\sup |\hat{\sigma}^2(s) - \sigma^2(s)|$  are derived as:

$$\begin{aligned}\sup |\hat{R}_0(s_1, s_2) - R_0(s_1, s_2)| &= O_p(m(h_1^2 + h_3^2 + (\frac{\log(n)}{nh_1})^{1/2} + (\frac{\log(n)}{nh_3^2})^{1/2})), \\ \sup |\hat{\sigma}^2(s) - \sigma^2(s)| &= O_p(m(h_1^2 + h_4^2 + (\frac{\log(n)}{nh_1})^{1/2} + (\frac{\log(n)}{nh_4^2})^{1/2})).\end{aligned}$$

**Note:**  The rate of convergence regarding the smooth part and nugget effect are separately calculated mainly due to the difference in the estimation procedure. The smooth function part is estimated through bivariate local linear smoothing, whereas the nugget effect is estimated by local linear smoothing. Proof of Lemma 2 requires a rewriting form of  $\hat{R}_0(s_1, s_2)$  and  $\hat{\sigma}^2(s)$ , which is extensively discussed in [20]. For the smooth function part, the main idea of this proof lies in decomposition of  $\hat{R}_0(s_1, s_2)$  into bias

and variance parts. By calculating the uniform convergence rates separately and added them up, we obtain the total rate of convergence. Similar derivations are applicable for the convergence rate calculation of nugget effects.

**Proof:**

We mainly focus on the convergence rate of  $|(\hat{R}_0 - R_0)(s_1, s_2)|$ , which is the difference on the smooth part of the estimated and true covariance matrix. According to the mathematical formula we give for  $\hat{R}_{0r}(s_1, s_2)$  in Section 3.4 in this Chapter, all the components in the formula are fixed except for  $R_{pq}^r$  when the spatio-temporal domain is fixed, where  $0 \leq p, q \leq 1$ . Therefore, the only difference lies in the subtraction of term  $\hat{R}_{pq}^r$  and  $R_{pq}^r$  inside  $\hat{R}_0(s_1, s_2)$  and  $R_0(s_1, s_2)$ , respectively. To illustrate it more precisely, the following formula for  $|(\hat{R}_0 - R_0)(s_1, s_2)|$  is derived:

$$\begin{aligned} |(\hat{R}_0 - R_0)(s_1, s_2)| &= (A_1^0 R_{00}^{0*} - A_2^0 R_{10}^{0*} - A_3^0 R_{01}^{0*})(B^0)^{-1} \\ &\quad + 2 \sum_{r=1}^m (A_1^r R_{00}^{r*} - A_2^r R_{10}^{r*} - A_3^r R_{01}^{r*})(B^r)^{-1}, \end{aligned}$$

where  $A_1^r$ ,  $A_2^r$ ,  $A_3^r$  and  $B^r$  are defined in the Section 3.4 of this Chapter and

$$\begin{aligned} R_{pq}^{r*} &= \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{k \neq j}^m (\hat{\epsilon}_t(s_j) \hat{\epsilon}_{t+r}(s_k) - R_{0r}(s_1, s_2) - R_{0r}^{(1,0)}(s_j - s) - R_{0r}^{(1,0)}(s_k - s)) \\ &\quad \cdot \left(\frac{s_j - s_1}{h_3}\right)^p \left(\frac{s_k - s_2}{h_3}\right)^q K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2). \end{aligned}$$

for  $0 \leq p, q \leq 1$ . The difference  $(\hat{R}_0 - R_0)(s_1, s_2)$  is split according to banded matrix with different temporal lags, ranging from 0 to  $m$ . There are in total  $2m + 1$  terms in the summation. Thus, the main task now is to derive the convergence rates for all the terms inside the summation formula. The three most crucial parts in the derivation lies in  $R_{00}^{r*}$ ,  $R_{00}^{r*}$  and  $R_{00}^{r*}$ , where  $0 \leq r \leq m$ . Denote the following term:

$$\epsilon_{\{(t,j),(t+r,k)\}}^* = \epsilon_t(s_j) \epsilon_{t+r}(s_k) - R_{0r}(s_j, s_k),$$

then we can easily obtain that  $E(\epsilon_{\{(t,j),(t+r,k)\}}^*) = 0$ . Then we can decompose  $R_{00}^{r*}$  regarding lag- $r$  banded matrix into three parts:

$$R_{00}^{r*} = (R_{00,a}^{r*} + R_{00,b}^{r*} + R_{00,v}^{r*}) \cdot \{1 + o_p(1)\},$$



where

$$R_{00,a}^{r*} = \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} (\hat{\epsilon}_t(s_j) \hat{\epsilon}_{t+r}(s_k) - \epsilon_t(s_j) \epsilon_{t+r}(s_k)) K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2),$$

$$R_{00,b}^{r*} = \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} [R_{0r}(s_j, s_k) - R_{0r}(s_1, s_2) - R_{0r}^{(1,0)}(s_j - s) - R_{0r}^{(1,0)}(s_k - s)] \\ \cdot K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2),$$

$$R_{00,v}^{r*} = \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} \epsilon_{\{(t,j),(t+r,k)\}}^* K_{h_3}(s_j - s_1) K_{h_3}(s_k - s_2).$$

Straightforwardly, We can obtain a changed version of  $R_{00,a}^{r*}$  as:

$$R_{00,a}^{r*} = \frac{1}{N_r} \sum_{t=1}^{n-r} \sum_{j=1}^m \sum_{j \neq k} [\epsilon_t(s_j)(\hat{\epsilon}_{t+r}(s_k) - \epsilon_{t+r}(s_k)) + (\hat{\epsilon}_t(s_j) - \epsilon_t(s_j))\epsilon_{t+r}(s_k) \\ + (\hat{\epsilon}_t(s_j) - \epsilon_t(s_j))(\hat{\epsilon}_{t+r}(s_k) - \epsilon_{t+r}(s_k))].$$

By Lemma 1, we know that  $\hat{\epsilon}_t(s_j) - \epsilon_t(s_j)$  is bounded by  $O_p(h_1^2 + (\frac{\log(n-r)}{(n-r)h_1})^{1/2})$  uniformly for all  $t \in T$  and  $j$ . A similar result holds for  $\hat{\epsilon}_{t+r}(s_k) - \epsilon_{t+r}(s_k)$ . Since  $0 \leq r \leq m$  and  $m = o(n^{1/2})$ , we can derive that  $R_{00,a}^{r*}$  is of order  $O_p(h_1^2 + (\frac{\log(n)}{nh_1})^{1/2})$  uniformly for all  $s_1, s_2 \in S$  and  $0 \leq r \leq m$ .

We can further derive the order of the bias part  $R_{00,b}^{r*}$ , which is  $O_p(h_3^2 + (\frac{\log(n)}{nh_3})^{1/2})$  uniformly for  $s_1, s_2 \in S$  [37]. For the variance part  $R_{00,v}^{r*}$ , we can also attain its order by classical uniform convergence rates for bivariate kernel smoothers, which is  $O_p[(\frac{\log(n)}{nh_3})^{1/2}]$  uniformly for  $s_1, s_2 \in S$ .

With similar decomposition, we obtain that  $R_{10}^{r*}$  and  $R_{01}^{r*}$  both have order  $O_p(h_2^3 + (\frac{\log(n)}{nh_2^2})^{1/2} + h_1^2 + (\frac{\log(n)}{nh_1})^{1/2})$  uniformly for all  $s_1, s_2 \in S$ . The rates for  $S_{00}^r, S_{01}^r, S_{02}^r, S_{10}^r, S_{11}^r, S_{20}^r$  are additionally studied, which is similar to [37]. Finally, we attain the convergence rate for  $\hat{R}_0$  by organizing all these terms up and summing over all banded covariances based on lags from 0 to  $m$ , which is:

$$\sup |\hat{R}_0(s_1, s_2) - R_0(s_1, s_2)| = O_p((2m+1)(h_1^2 + h_3^2 + (\frac{\log(n)}{nh_1})^{1/2} + (\frac{\log(n)}{nh_3^2})^{1/2})) \\ = O_p(m(h_1^2 + h_3^2 + (\frac{\log(n)}{nh_1})^{1/2} + (\frac{\log(n)}{nh_3^2})^{1/2})).$$

Results on the nugget effects can be similarly produced by using one-dimensional local linear smoothing techniques:

$$\sup |\hat{\sigma}^2(s) - \sigma^2(s)| = O_p(m(h_1^2 + h_4^2 + (\frac{\log(n)}{nh_1})^{1/2} + (\frac{\log(n)}{nh_4})^{1/2})).$$

Finally, we focus on the convergence rate of  $\hat{\beta}_{rf}$ , which denotes the refined estimator of  $\beta$  after interpolating the estimated spatio-temporal covariance matrix.

**Theorem 1.** (Asymptotic normality of  $\hat{\beta}_{rf}$ )

Under Assumption 3, we have:

$$n^{1/2}(\hat{\beta}_{rf} - \beta) \xrightarrow{d} N(0, \tilde{A}^{-1}),$$

where  $\tilde{A} = E(\tilde{X}^T \hat{\Sigma}^{-1} \tilde{X}^T)$ ,  $\tilde{X} = X - \phi_{eff}(s)$ .  $\phi_{eff}(s)$  is the solution of the following Fredholm integral equation:

$$\sum_{j=1}^m \sum_{l=1}^m E\{\sigma^{jl}(X(s_l) - \phi(s_l)) | s_l = s\} f(s) = 0,$$

where  $\sigma^{jl}$  is the  $(j, l)$ -th element of  $\Sigma^{-1}$ ,  $f(\cdot)$  is the density of  $s$ .

**Note:** The main idea behind the proof lies in determining the rate of  $|\hat{\Sigma} - \Sigma|$ . The key is to show that the estimated covariance function is uniformly consistent and when it is interpolating at specific spatial locations with particular observation times and inserted into the profile local estimating equations, the estimation errors in the covariance only introduce an asymptotically negligible error into the final estimator. Thus, we can show that the kernel and profile estimating equations after interpolating estimated covariance are asymptotically equivalent to those where the true covariance matrix is used. Therefore, the error incurred by substituting for the estimated covariance matrix is asymptotically negligible. From Theorem 1, we know that  $\hat{\beta}_{rf}$  achieves the asymptotically semiparametric efficiency bound, which is more efficient than  $\hat{\beta}_{WI}$  under the case of working independence.

**Proof:**

The ultimate goal is to derive the convergence rate of  $\hat{\beta} - \beta$ . Denote

$$\delta_n^* = m(h_2^3 + h_4^2 + (\frac{\log(n)}{nh_2^3})^{1/2} + (\frac{\log(n)}{nh_4})^{1/2}),$$

which is uniformly hold for all  $s_1, s_2 \in S$ . Then we attain that  $|\hat{\Sigma}(s, t) - \Sigma(s, t)| = Op(\delta_n^*)$  according to Lemma 2 above. By simple transformation in matrix theory, we derive that

$$\hat{\Sigma}^{-1} - \Sigma^{-1} = \Sigma^{-1}(\Sigma - \hat{\Sigma})\Sigma^{-1}.$$

Therefore, we can further obtain that  $|\hat{\Sigma}^{-1}(s, t) - \Sigma^{-1}(s, t)| = Op(\delta_n^*)$ . The kernel estimating equation under a given coefficient  $\beta$  regarding the linear term can be decomposed into two parts as below:

$$\begin{aligned}
 & \sum_{j=1}^m S_j(s) K_{h_2}(s_j - s) \hat{\Sigma}^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \theta(s)]\} \\
 &= \sum_{j=1}^m S_j(s) K_{h_2}(s_j - s) \Sigma^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \theta(s)]\} \\
 & - \sum_{j=1}^m S_j(s) K_{h_2}(s_j - s) \Sigma^{-1} (\hat{\Sigma} - \Sigma) \Sigma^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \theta(s)]\} \\
 & \triangleq A_1 - A_2.
 \end{aligned}$$

The dominant term  $A_1$  above is the same as in the kernel estimation equation of [67]. The order of which is  $O_p(h_2^2 + (\frac{\log(n)}{nh_2})^{1/2})$ . The order of  $A_2$  can be seen more clearly by the following derivation:

$$\begin{aligned}
 A_2 &\propto [h_2^2 + (\frac{\log(n)}{nh_2})^{1/2}] \cdot [m(h_3^2 + h_4^2 + (\frac{\log(n)}{nh_2^3})^{1/2} + (\frac{\log(n)}{nh_4})^{1/2})] \\
 &= m(h_3^2 + h_4^2)h_2^2 + h_2^2[m(\frac{\log(n)}{nh_2^3})^{1/2} + m(\frac{\log(n)}{nh_4})^{1/2}] \\
 &+ m(h_3^2 + h_4^2)(\frac{\log(n)}{nh_2})^{1/2} + (\frac{\log(n)}{nh_2})^{1/2}[m(\frac{\log(n)}{nh_2^3})^{1/2} + m(\frac{\log(n)}{nh_4})^{1/2}] \\
 &= o_p(h_2^2 + (\frac{\log(n)}{nh_2})^{1/2}).
 \end{aligned}$$

The last equation holds due to Assumption 3. As a result,  $A_2$  is of higher order compared to  $A_1$ . So we conclude that  $\hat{\theta}(s, \beta)$  with a given  $\beta$  in the case of using an consistent estimator of spatio-temporal covariance is asymptotically equivalent to that using the true covariance matrix. Differentiate the equation above w.r.t  $\beta$  can we obtain  $\phi(t) = \frac{\partial \hat{\theta}(s, \beta)}{\partial \beta}$  is asymptotically equivalent to that found by inserting the true covariance and  $\phi(t) = \phi_{eff}(t) \cdot \{1 + o_p(1)\}$ , where  $\phi_{eff}(t)$  is the solution of the Fredholm integration equation [67]. In addition, the profile estimating function can be rewritten as:

$$\begin{aligned} & \sum_{j=1}^m \frac{\partial [X_t^T(s_j)\beta + \hat{\theta}(s_j, \beta)]}{\partial \beta} \Sigma^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \hat{\theta}(s_j, \beta)]\} \\ & - \sum_{j=1}^m \frac{\partial [X_t^T(s_j)\beta + \hat{\theta}(s_j, \beta)]}{\partial \beta} \Sigma^{-1} (\hat{\Sigma} - \Sigma) \Sigma^{-1} \{Y_t(s_j) - [X_t(s_j)\beta + \hat{\theta}(s_j, \beta)]\}. \end{aligned}$$

Similarly, the first counterpart is the dominant term, while the latter one is of higher order which is asymptotically negligible. Thus, we can derive the asymptotic distribution of  $\hat{\beta}_{rf}$  with true covariance incorporated in the equations. According to [40], we have:

$$n^{1/2}(\hat{\beta}_{rf} - \beta) = \{\mathbb{E}(\tilde{X}\Sigma^{-1}\tilde{X})\}^{-1}\{B_n + C_{1n} - C_{2n}\} + o_p(1),$$

where

$$\begin{aligned} B_n &= \frac{1}{2(nh_2^4)^{1/2}} \left[ \frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \sum_{l=1}^m \mu_t^{(1)}(s_j) (\Sigma^{-1})_t^{jl} \mu_t^{(1)}(s_l) \tilde{X}_t(s_j) \right. \\ & \quad \cdot \{b_*(s_l) + h_2 b_{*1}(s_l) + O_p(h_2^2)\} \big] (1 + o_p(1)), \end{aligned}$$

$$C_{1n} = n^{-1/2} \sum_{t=1}^n \tilde{X}^T(t) (\Sigma^{-1}) (Y_t - \mu_t),$$

$$\begin{aligned} C_{2n} &= \{n^{-1/2} \sum_{t=1}^n \sum_{j=1}^m \sum_{l=1}^m \tilde{X}_t(s_j) \mu_t^{(1)}(s_j) (\Sigma^{-1})_t^{jl} \mu_t^{(1)}(s_l) \\ & \quad \cdot (W^{-1}(s_l)) \frac{1}{n} \sum_{t'=1}^n \sum_{j'=1}^m \mu_{t'}^{(1)}(s_{j'}) (\Sigma^{-1})_{t'}^{j'j'} \\ & \quad \cdot [K_h(s_{j'} - s_l) \{ \sum_l (\Sigma^{-1})_{t'}^{j'l} (Y_{t'}(s_l) - \mu_{t'}(s_l)) \} \\ & \quad + Q_{2*}(s_l, s_{j'}) (Y_{t'}(s_{j'}) - \mu_{t'}(s_{j'})) \\ & \quad + Q_{1*}(s_l, s_{j'}) \{ \sum_l (\Sigma^{-1})_{t'}^{j'l} (Y_{t'}(s_l) - \mu_{t'}(s_l)) \} \} \} (1 + o_p(1)), \end{aligned}$$

in which the definitions of  $b_*$ ,  $Q_{1*}$ ,  $Q_{2*}$  and  $W(\cdot)$  are specifically explained in Appendix A.2 of [67]. The term  $\mu_t(s_j) = \mathbb{E}(Y_t(s_j)|X_t(s_j))$ . Since the term

$$\frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \sum_{l=1}^m \mu_t^{(1)}(s_j) (\Sigma^{-1})_t^{jl} \mu_t^{(1)}(s_l) \tilde{X}_t(s_j) \cdot \{b_*(s_l) + h_2 b_{*1}(s_l) + O_p(h_2^2)\}$$

is of higher order of  $O_p(1)$ , an  $nh_2^4$  term is needed to count down the bias in  $B_n$ . As Lemma A.1 noted in [67], we have  $B_n = o_p(1)$ .

Next, we will check if  $C_{2n}$  is of order  $o_p(1)$ . We decompose  $C_{2n}$  into 3 parts:  $C_{2n} = C_{21n} + C_{22n} + C_{23n}$ , where

$$\begin{aligned} C_{21n} &= n^{-1/2} \sum_{t'=1}^n \sum_{j'=1}^m \mu_{t'}^{(1)}(s_{j'}) (\Sigma^{-1})_{t'}^{j'j'} \\ &\quad \left\{ \frac{1}{n} \sum_{t=1}^n \sum_{j=1}^m \sum_{l=1}^m K_{h_2}(s_{j'} - s_l) \tilde{X}_t(s_j) \mu_t^{(1)}(s_j) (\Sigma^{-1})_t^{jl} \mu_t^{(1)}(s_l) W^{-1}(s_l) \right\} \\ &\quad \cdot \left\{ \sum_l (\Sigma^{-1})_{t'}^{j'l} (Y_{t'}(s_l) - \mu_{t'}(s_l)) \right\} \cdot (1 + o_p(1)), \end{aligned}$$

$$\begin{aligned} C_{22n} &= n^{-1/2} \sum_{t'=1}^n \sum_{j'=1}^m \mu_{t'}^{(1)}(s_{j'}) (\Sigma^{-1})_{t'}^{j'j'} \\ &\quad \cdot [\mathbb{E}\{\tilde{X}(s_j) (\Sigma^{-1})^{jl} W^{-1}(s_l) Q_{1*}(s_l, s)\} | s] \\ &\quad \cdot \left\{ \sum_l (\Sigma^{-1})_{t'}^{j'l} (Y_{t'}(s_l) - \mu_{t'}(s_l)) \right\} \cdot (1 + o_p(1)) \end{aligned}$$



$$\begin{aligned} C_{23n} &= n^{-1/2} \sum_{t'=1}^n \sum_{j'=1}^m \mu_{t'}^{(1)}(s_{j'}) (\Sigma^{-1})_{t'}^{j'j'} \\ &\quad \cdot [\mathbb{E}\{\tilde{X}(s_j) (\Sigma^{-1})^{jl} W^{-1}(s_l) Q_{2*}(s_l, s)\} | s] \\ &\quad \cdot (Y_{t'}(s_l) - \mu_{t'}(s_l)) \cdot (1 + o_p(1)) \end{aligned}$$

Through using Lemma A.1 in [67], we know that  $C_{22n}$  and  $C_{23n}$  are both equal to 0. Under Assumption 3, we attain that  $B_n$  and  $C_{21n}$  all both of order  $o_p(1)$ . Therefore,

$$n^{1/2}(\hat{\beta} - \beta) = \{\mathbb{E}(\tilde{X}\Sigma^{-1}\tilde{X})\}^{-1} n^{-1/2} \sum_{t=1}^n \tilde{X}^T(t) (\Sigma^{-1})(Y_t - \mu_t) \cdot (1 + o_p(1))$$

This indicates the asymptotic normality of  $\hat{\beta}_{rf}$ , which is denoted as  $n^{1/2}(\hat{\beta}_{rf} - \beta) \xrightarrow{d} N(0, \tilde{A}^{-1})$ , where  $\tilde{A} = E(\tilde{X}^T \hat{\Sigma}^{-1} \tilde{X}^T)$ .

### 3.6 Simulation Studies

In this section, we conduct two simulation studies to assess the performance of the proposed method. The simulation experiments are designed in a two-dimensional spatial region. Here, the distribution of spatial locations is categorized into two scenarios: uniformly-distributed and irregularly-designed. In the first scenario, the spatial points are uniformly placed on a  $10 * 10$  gridded square with equal distance between neighboring locations. In addition, the case for irregularly-distributed spatial locations on this square is further investigated. For both scenarios, our main focus lies in comparing the estimation efficiency of the parametric term in the semiparametric model before and after interpolating the estimated covariance matrix.

We consider the following model:

$$Y_t(s) = X_t^T(s)\beta + \theta(s) + \epsilon(s); s \in S, t \in T,$$

where  $Y_t(s)$  and  $X_t(s)$  denote the observation values of the response and covariates for spatial location  $s$  at time  $t$ . The coefficient  $\beta$  is an unknown vector and  $\theta(\cdot)$  is an unknown smooth function which represents the spatial main effect. The  $10 * 10$  square region and the temporal basis are denoted as  $S$  and  $T$ , respectively. Detailed settings for the two scenarios are illustrated below:

#### Scenario (A). Uniformly-distributed spatial region (Gridded data)

Spatial points are located uniformly on a squared region  $[1, 10] * [1, 10]$  with equal distance between neighbouring locations (100 points in total). The temporal line  $T$  ranges from 1 to 100. Spatial correlation at time  $t = 1$  follows an exponential decay from the central point (5,5) (Use the form  $\exp(-|d|/2)$  as correlation, in which  $d$  denotes the distance between two spatial points). Variance and nugget effect of the covariance matrix are set to 0.5 and 0.1, respectively. The error term follows AR(1) structure on the temporal line. Suppose two covariates  $X_1$  and  $X_2$  are included in the model, where  $X_1 \sim U[-1, 1]$  and  $X_2 \sim N(0, 1)$ . The parametric terms  $\beta_1$  and  $\beta_2$  are both set to 1. Spatial main effect  $\theta(s) = \frac{1}{10}[(x - 5)^2 + (y - 5)^2]$ .

#### Scenario (B). Irregularly-spaced spatial region

All the settings in Scenario (B) are the same as Scenario (A) except for the distribution of the spatial locations in this squared region. By decomposing the it uniformly into four subregions, we randomly allocate 40 points to sub-square  $[1, 5] * [1, 5]$ , 10 points to  $[1, 5] * (5, 10]$ , 30 points to  $(5, 10] * [1, 5]$ , 20 points to  $(5, 10] * (5, 10]$ .

The ultimate goal of these simulation studies is to examine whether efficiency gain on the estimation of  $\beta_1$  and  $\beta_2$  is achieved. This work is accomplished by comparing the variance of the estimators. Estimation results on the spatio-temporal covariance, nugget effect and spatial main effect  $\theta(s)$  are also included in the next section.

### Numerical results

In both Scenario (A) and Scenario (B), we simulate 200 sample datasets. This is done through randomly generating numbers from the distribution of  $\epsilon_t(s)$  for each sample. In both Scenario (A) and (B), we simulate 200 sample datasets. We mainly focus on the estimation of the parametric terms  $\beta_1$ ,  $\beta_2$  and the nonparametric spatial main effect  $\theta(s)$ . Table 1 and 2 show the estimation results (bias, variance) under four circumstances through fixed and proportional bandwidth selection, respectively. The first two cases use fixed and proportional bandwidth selection method under grid data, while the last two use these two bandwidth selection methods under irregularly-spaced data.

表 3.1: Estimation of parametric and nonparametric terms under four types of covariance structures in simulation studies.

	bias	WI variance	Banded bias	diagonal variance	Lag-1 bias	banded cov variance	Lag-2 bias	banded cov variance
Case 1								
$\hat{\beta}_1$	-0.007	0.021	-0.006	0.017	-0.006	0.015	0.005	0.015
$\hat{\beta}_2$	0.009	0.027	0.008	0.024	-0.008	0.022	0.007	0.022
$\hat{\theta}(s)$	0.466	1.646	0.404	1.223	0.481	1.154	0.428	0.130
Case 2								
$\hat{\beta}_1$	-0.007	0.020	-0.005	0.016	-0.006	0.014	0.004	0.014
$\hat{\beta}_2$	0.009	0.027	0.009	0.023	-0.008	0.022	0.007	0.022
$\hat{\theta}(s)$	0.461	1.644	0.386	1.221	0.485	1.129	0.444	0.121
Case 3								
$\hat{\beta}_1$	-0.009	0.039	0.008	0.034	-0.008	0.032	0.008	0.031
$\hat{\beta}_2$	0.013	0.046	0.012	0.042	-0.012	0.039	-0.011	0.038
$\hat{\theta}(s)$	0.571	2.083	0.489	1.672	0.448	1.472	0.440	1.436
Case 4								
$\hat{\beta}_1$	-0.010	0.038	0.009	0.033	-0.008	0.031	0.008	0.029
$\hat{\beta}_2$	0.011	0.045	0.011	0.041	-0.010	0.038	-0.011	0.037
$\hat{\theta}(s)$	0.555	1.963	0.538	1.666	0.440	1.424	0.445	1.412

The estimation results show that the variance of the three refined estimators are significantly smaller than those under the case of Working Indenpendce. Under the condition of uniformly-distributed grid data, the estimate of parameter  $\beta_1$  experiences a decrease of 20.0%, 30.0% and 30.0% in variance by interpolating three types of covariance structures based on the results using proportional bandwidth, and estimate of parameter

$\beta_2$  attains a decrease of 14.8%, 18.5% and 18.5% in variance meanwhile. In the case of irregularly-spaced data, the estimate of parameter  $\beta_1$  shows a decrease of 12.8%, 17.9% and 20.5% in variance by interpolating three types of covariance structures, and estimate of parameter  $\beta_2$  experiences a decrease of 8.9%, 15.6% and 17.8% in variance meanwhile. Besides, a significant decrease of variance is also seen for the estimated spatial main effect  $\theta(s)$ . Furthermore, little efficiency gain is reached by transforming from the case of lag-1 blocked diagonal matrix to that of lag-2 blocked diagonal matrix, which suggest that we already achieve high estimation efficiency through proper lag selection according to the correlation structure on the temporal basis.

**Remark.** The estimation results based on two different bandwidth selection methods show that proportional bandwidth slightly outperforms fixed bandwidth on the final estimation efficiency. Therefore, it is more advisable to use proportional bandwidth in kernel estimates whatever the distribution of spatial locations.

## 3.7 Real data analysis

### 3.7.1 Air pollution mode data in China

China has been experiencing severe air pollution in recent years. High concentration of PM<sub>2.5</sub> has caused great harm to people's health. We intend to learn about which air pollutants are highly correlated with PM<sub>2.5</sub> within representative parts of China. The air pollution dataset for this article contains grided mode data of six different pollutants (PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, NO<sub>2</sub>, O<sub>3</sub>, CO) in China and surrounding areas (including the neighbouring sea). Since PM<sub>10</sub> highly correlates with PM<sub>2.5</sub>, we only use five air pollutants instead without including PM<sub>10</sub>. Spatial location for each grid point is given by two coordinates: longitude and latitude. Time spans from 12:00am in Dec 1, 2016 to 11:00pm in Dec 31, 2016, which is hourly renewed data. In this section, we only select parts of China as a spatial region we consider in the model. Figure 3.1 shows the selected spatial region in China (with dotted red rectangular).

The rectangular region ranges to the most northern and eastern part of Beijing, and to the most western and southern part of Guizhou province with 13674 points in total. The exact longitude range is (103.600, 117.500), whereas the exact latitude range is (24.617, 41.050). To illustrate the linear relationship between PM<sub>2.5</sub> and other air pollutants under the spatio-temporal setting, we use our novel semiparametric spatio-



temporal model for analysis.



图 3.1: Map of China and selected spatial region.

Before modeling work, we use box-cox tranformation to handle the problem of heavy-tail distribution in some covariates. The new variables are denoted as:  $\text{tr}(\text{PM}_{2.5})$ ,  $\text{tr}(\text{SO}_2)$ ,  $\text{tr}(\text{NO}_2)$ ,  $\text{tr}(\text{O}_3)$ ,  $\text{tr}(\text{CO})$ . In our model, the response is  $\text{tr}(\text{PM}_{2.5})$  while the other four air pollutants are set as covariates. Parameters corresponding to these covariates are denoted as:  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  and  $\beta_4$ , respectively. The spatial main effect is denoted as  $\theta(s)$ .

Another important thing is to determine the order of the ARIMA model using the residuals calculated under the Working Independence condition. This gives light on how many lags to choose in the banded covariance. We take a sum of the residuals for all sites at each time point and determing the order through AIC and BIC criterion. For the air pollution data, AR(1) model is examined to be the best choice with AIC=688.26 and BIC=695.22. A PACF (Partial Autocorrelation Function) is used to illustrate this (see Figure 3.2).

Table 3.2 shows the final estimation results for both the parametric terms and the nonparametric spatial main effect. Both fixed and proportional bandwidth cases are considered. Similar to the simulation studies in the previous section, we consider four types of covariance structures in this setting. The two cases in the following table corresponds to estimation under fixed and proportional bandwidth selection method, respectively.

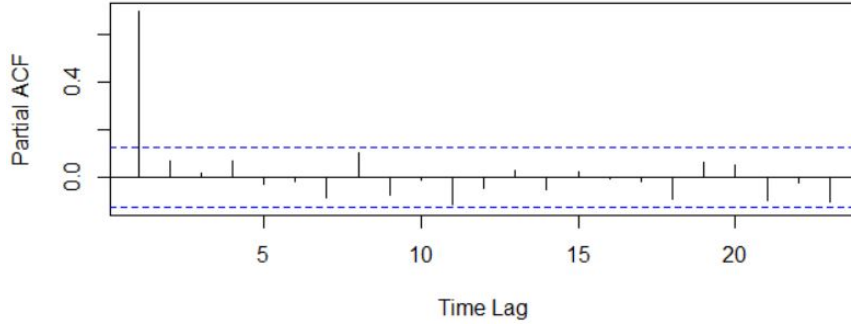


图 3.2: PACF plot: Residual sum of all spatial sites in parts of China.

表 3.2: Estimation of parametric and spatial effect under four types of covariance structures for air pollution data in parts of China.

	WI		Banded diagonal		Lag-1 banded cov		Lag-2 banded cov	
	Estimate	sd	Estimate	sd	Estimate	sd	Estimate	sd
Case 1								
$\hat{\beta}_1$	0.459	0.097	0.440	0.079	0.436	0.075	0.437	0.074
$\hat{\beta}_2$	0.518	0.096	0.545	0.071	0.548	0.068	0.545	0.067
$\hat{\beta}_3$	-0.174	0.038	-0.188	0.034	-0.188	0.033	-0.186	0.032
$\hat{\beta}_4$	3.375	0.524	3.294	0.395	3.293	0.379	3.293	0.375
$\hat{\theta}(s)$	—	0.167	—	0.117	—	0.116	—	0.116
Case 2								
$\hat{\beta}_1$	0.452	0.095	0.439	0.075	0.435	0.071	0.436	0.067
$\hat{\beta}_2$	0.505	0.096	0.542	0.070	0.544	0.064	0.546	0.064
$\hat{\beta}_3$	-0.167	0.036	-0.180	0.030	-0.180	0.027	-0.184	0.026
$\hat{\beta}_4$	3.378	0.520	3.293	0.388	3.298	0.371	3.299	0.368
$\hat{\theta}(s)$	—	0.161	—	0.112	—	0.110	—	0.110

From the estimation results in Table 3.2, a significant decrease is shown after interpolating the estimated spatio-temporal covariance matrix. Based on the results under proportional bandwidth selection, the estimation standard error of  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  and  $\beta_4$  experience a decline of 23.9%, 29.4%, 14.7% and 29.1% under the condition of Lag-1 banded covariance matrix. Estimation results show that  $\text{SO}_2$ ,  $\text{NO}_2$ ,  $\text{CO}$  are all positively correlated with  $\text{PM}_{2.5}$  except  $\text{O}_3$ .

Furthermore, the nonparametric spatial main effect  $\theta(s)$  attains a 32.8% decrease in standard error. By comparing the four estimation result in the table above, we learn that estimation efficiency attains significant decrease so long as including the estimated covariance matrix (even in the Diagonal banded covariance matrix case). Also, we see few efficiency gain by comparing the case of Lag-2 to Lag-1 banded covariance matrix. Therefore, Lag-1 banded covariance is the best choice since it can provide us with highly

efficient estimation results with fewer computation time.

The results indicate that except for  $O_3$ , all the other 3 air pollutants ( $SO_2$ ,  $NO_2$  and  $CO$ ) are positively correlated with  $PM_{2.5}$  with high significance. However,  $O_3$  has a unique pattern on its own which is different from the other 4 pollutants. This gives us new guidance on air quality management in the sense that  $O_3$  should also be taken into account besides  $PM_{2.5}$ . We should develop new ways on controlling air pollution caused by  $O_3$ .

### 3.7.2 COVID-19 in Hubei Province

We also apply the methodology on newly-developed Coronavirus disease (COVID-19) in parts of Hubei Province, which also show some improvements over the estimation of  $\beta$ . COVID-19 is an infectious disease caused by a new virus. The disease causes respiratory illness with symptoms such as cough, fever, and in more severe cases, difficulty breathing. It spreads primarily through contact with an infected person when they cough or sneeze. From January to March 2020, China experiences massive area of COVID-19 infection. In this section, we will study the cure rate in Hubei Province.

We include all 17 cities of Hubei Province in our analysis: Wuhan, Huanggang, Ezhou, Huangshi, Xianning, Xiaogan, Tianmen, Xiantao, Qianjiang, Jingzhou, Jingmen, Suizhou, Xiangfan, Yichang, Shennongjia, Shiyan, Enshi. Wuhan experiences the highest infection number among all 17 cities. We consider two temporal lines in our analysis: the first time span is from Jan 28 to Feb 23, 2020, and the second one from Jan 28 to Mar 17, 2020. The first temporal line is chosen due to the reason that flow of people is strictly controlled after Feb 23, 2020. And the last date of the second temporal basis represents the inflection point. The daily renewed data is collected from the Health Commission of Hubei Province (<http://wjw.hubei.gov.cn/>). The variables we consider in the analysis are Confirm number and Heal number. We will explore the relationship between Heal number and Confirm number within the spatial region of Hubei Province.

Before moving on to the estimation part, we first simply discuss the cure rate (calculated by Heal number divided by Confirm number) in five cities of Hubei Province on the temporal line: Huanggang, Suizhou, Wuhan, Xiangyang and Xiaogan. Figure 3.3 characterizes the trend of cure rate for these five cities.

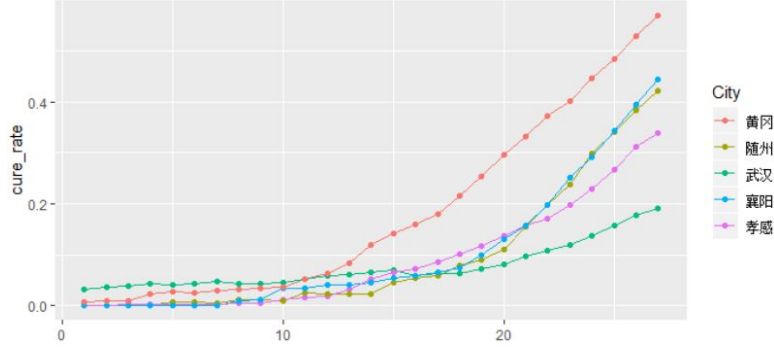


图 3.3: Changing trend of cure rate in five cities of Hubei Province.

From Figure 3.3, we see that the trend shows an exponential increase. Therefore, we take a box-cox transform of the covariates (log-transform) before data analysis. The model in this setting is described as:

$$Heal_t(s) = Confirm_t^T(s)\beta + f(s) + \epsilon_t(s); s \in S, t \in T.$$

Below are the estimation results for  $\beta$  and  $f(s)$  in Hubei Province using proportional bandwidth. The two cases in the following table corresponds to estimation under two time spans (Case 1: Jan 28 to Feb 23, 2020; Case 2: Jan 28 to Mar 17, 2020), respectively.

表 3.3: Estimation of cure rate and spatial effect for COVID-19 under four types of covariance structures (Proportional bandwidth selection).

	WI		Banded diagonal		Lag-1 banded cov		Lag-2 banded cov	
	Estimate	sd	Estimate	sd	Estimate	sd	Estimate	sd
Case 1								
$\hat{\alpha}$	0.0013	0.00032	0.0009	0.00025	0.0009	0.00024	0.0010	0.00024
$\hat{\theta}_1(s)$	—	0.00429	—	0.00357	—	0.00333	—	0.00326
Case 2								
$\hat{\alpha}$	0.0084	0.00191	0.0093	0.00159	0.0089	0.00149	0.0090	0.00143
$\hat{\theta}_1(s)$	—	0.01170	—	0.00904	—	0.00877	—	0.00862

From Table 3.3, we can see that the estimation efficiency of  $\beta$  and  $\theta_1(s)$  experiences a significant decline by taking the estimated spatio-temporal covariance matrix into account. The standard error of estimated cure rate  $\hat{\beta}$  attains a 25.0% and 22.0% decrease under the case of interpolating estimated lag-1 covariance regarding two temporal basis, respectively. We also notice less estimation efficiency gain after interpolating the estimated lag-2 covariance compared to using lag-1 covariance. Furthermore, the color change in Figure 3.4 below shows that including the estimated covariance has the capacity to count down the spatial correlation between neighboring sites.

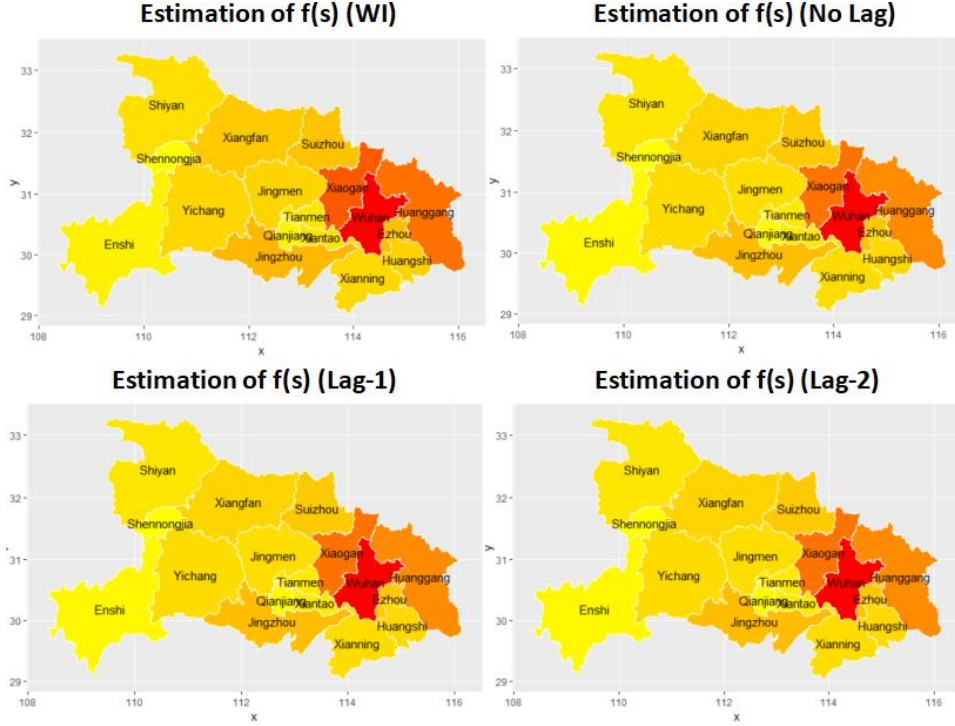


图 3.4: Plot of estimated spatial effect  $\hat{\theta}(s)$  based on proportional bandwidth selection under four types of covariance structures for COVID-19 data in Hubei Province from Jan 28th to Feb 23th, 2020.

We learn from the results presented that cure rate for COVID-19 in the long run gets higher than in a relatively short period. This is due to the fact that with the development of medical care for COVID-19 and less people get caught in this disease in March, the total number of cured patients is on the increasing path while the cumulative population infected with COVID-19 tends to be stable. Figure 5 also shows the severity of COVID-19 infection in Wuhan and the surrounded cities, including Xiaogan and Huanggang.



## 第四章 Causal and mediation effect analysis

### 4.1 My contributions on this research:

I did all the methodology and data analysis work under the supervision of Prof. Yehua Li in UCR.

### 4.2 Overview

In this chapter, we will propose proper statistical methods in analyzing the risk of cardiovascular disease (CVD) using data from the Aerobics Center Longitudinal Study (ACLS). There were in total 12591 participants considered in our analysis. Liu and her collaborators explored the association between RE and risk of CVD morbidity, all-cause mortality and total CVD events through a traditional joint modeling approach by utilizing the same dataset [43]. However, the underlying causal relationship between RE and CVD risk was not discussed under their model framework. Also, BMI may not be the only mediator affecting CVD risk. Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Glucose and Total Cholesterol (TC) are all possible mediation effects which link physical activity (PA) to CVD. Therefore, we need to develop novel methods in comprehensively understanding both the causal mechanism of doing exercise on CVD risk and the mediation effects based on more probable and reasonable variables.

Here, the causal effect of RE on CVD risk will be investigated through marginal structural models (MSM). MSM is primarily used in deriving the natural direct and indirect effects of a specified exposure on the target outcome with one or more mediation pathways considered [22, 30, 31]. Hong developed a nonparametric statistical method which can relax the assumption of no treatment-mediator interaction in simultaneous estimation of the natural direct and indirect effects [22]. The main methodology behind is the estimation of the marginal mean regarding each counterfactual outcome through weighting each experimental unit. The weight applied here represents the ratio of the conditional probability w.r.t a mediator value under the control condition to the conditional probability w.r.t the same mediator value under the experimental circumstance. The outcome is characterized as a function of the direct and indirect effects. This novel approach is applicable for tackling with large number of covariates.

On the basis of [22], Lange and his collaborators proposed a unified estimation procedure which can parameterize the natural direct and indirect effects directly, and can be implemented in standard softwares [31]. The model framework here can be generalized to a wide class of outcomes and mediators. Accurate specification of the models regarding the distribution of exposures and mediators is necessary in achieving estimation with higher efficiency. Consider a binary exposure case and a generalized linear model framework, denote  $A$  as the exposure,  $X$  as the confounding variables,  $M$  as the possible mediators and  $Y$  as the target outcome. Specifically, the estimation procedure can be divided into the following 5 steps:

**Step 1.** Derive an appropriate model of  $A$  on  $X$  with the given dataset.

**Step 2.** Derive an appropriate model of  $M$  on  $A$  and  $X$  with the given dataset.

**Step 3.** Construct an artificial duplicated dataset through replicating each observation in the original data set twice and incorporating a new variable  $A^*$ , which is set as the original value regarding  $A$  in the first replication and the opposite value of  $A$  in the second replication.

**Step 4.** Calculate the weights  $W_i$  through applying the fitted models from Step 1 and Step 2, which is commonly of the form below:

$$W_i = \frac{1}{P(A=A_i|X=X_i)} \frac{P(M=M_i|A=A_i^*, X=X_i)}{P(M=M_i|A=A_i, X=X_i)}.$$

**Step 5.** Fit an appropriate model for the outcome including  $A$ ,  $A^*$  and sometimes  $X$  (depending on what we concern about in estimation) as covariates and weighted by  $W_i$  in Step 4.

The validity and rationality of using this 5-step estimation procedure is briefly discussed in the Appendix 4 of [31] under the case of a generalized linear MSM.

To make a step further, the methodology in [31] was extended to the case of multiple causal pathways by including more than one possible mediators [30]. Also, ranking on different causal pathways is conducted in this work. Before conduct Step 3 in the estimation procedure of [31], we need to test if each mediator considered is independent with the other ones conditional on the  $A$  and  $X$  in order to avoid possible interventions among different mediators. The estimation can be continued only when the no significant intervention exists. Furthermore, Step 4 is refined by replicating the original dataset for  $2^K$  times with proper assigned value for the auxiliary variables  $A_1, A_2, \dots, A_K$  when the number of possible mediators is  $K$ .



However, the above researches mainly focus on modeling binary and continuous outcomes by using generalized linear models in the field of causal inference. The use of MSM on a survival response is rarely studied. Also, application of causal effect analysis on kinesiology is a significant problem since issues related to human health are more and more important in recent years. We intend to learn about what the potential causal mechanism is under one or multiple pathways. Here, we will extend this method to the case of risk regarding total CVD events for the baseline part of the data.

Furthermore, we explore mediation effects and the relationship between RE and CVD risk under multiple longitudinal pathways through a multivariate joint modeling method based on both baseline and follow-up data [48]. Since data for the follow-up clinical trial is also incorporated, it makes a step further on the basis of the previous causal effect analysis. The joint modeling estimation procedure is divided into 2 stages. In stage 1, a multivariate mixed model is fitted regarding the longitudinal trajectories. Then the outputs from stage 1 is utilized to fit a survival model. This estimation procedure is applicable on various types of longitudinal trajectories, including left, right and interval censored data. Parameters in the joint model are estimated by using a Bayesian approach.

The main idea of our novel approach lies in jointly modeling the latent shared random effects behind the trajectories regarding the risk factors causing CVD. Bayesian estimation is applied by accommodating the covariance structure among all the longitudinal trajectories considered. We will derive a multiple joint model suited for our data by also including several confounding variables. The relative importance of these mediation effects we attained in causing CVD will be explicitly discussed based on the corresponded hazard ratios in the estimation results. In addition, the effect of RE on CVD risk is further investigated comprehensively by taking multiple longitudinal trajectories into account.

The rest of this chapter is organized as follows. In Section 3, we give a description and exploratory analysis of the ACLS data. In Section 4, we propose a causal and mediation effect analysis using MSM under the condition of a binary exposure and a survival outcome. In Section 5, we develop a multivariate joint modeling approach to further explore the mediation effects and the relationship between RE and CVD risk more comprehensively. In Section 6 and 7, we present the numerical results, conclusions and discussions.

### 4.3 Data description

ACLS is an observational study based on subjects receiving medical examination at the Cooper Clinic in Dallas, Texas. The 13722 participants are required to do clinical tests regularly. However, 1131 individuals are excluded from the analysis due to pre-existing conditions such as CVD, cancer, diabetes, abnormal resting, or exercise electrocardiogram which may cause inevitable bias in the estimation results. Therefore, only the remaining 12591 participants are included in our analysis. Ages of the participants range from 18 to 89, in which male individuals account for 79% of the total population. These people received at least two medical examinations from 1987 to 2006. The study is examined and approved by the Cooper Institute Institutional Review Board. All the participants provided informed consent for the baseline and follow-up examinations.

Detailed information regarding the study is illustrated precisely in US Department of Health and Human Services (2008) and [43]. RE can be quantified by weekly exercise frequency (times/week) and weekly average exercise time (minutes/week). Exploratory analysis on the potential risk factors is conducted in this research. Since DBP is significantly correlated with SBP with pearson correlation coefficient 0.643 ( $p\text{-value} < 0.001$ ) for baseline data, we only include SBP in our analysis as most literature do. The correlation plot for the remaining four risk factors is illustrated in Figure 4.1 below:

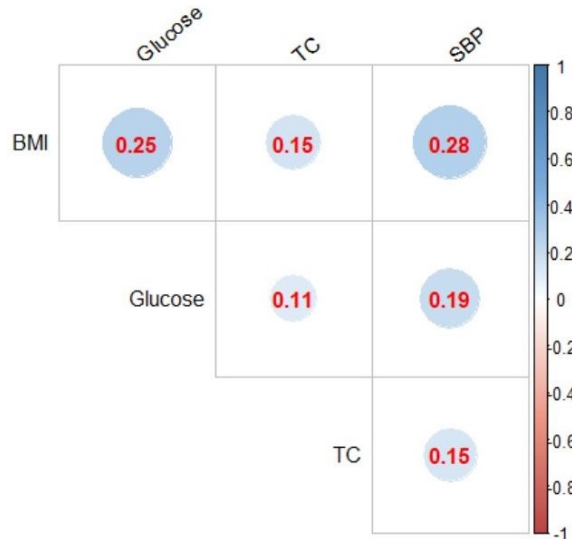


图 4.1: Correlation plot of BMI, Glucose, TC and SBP under baseline data (Note: The sizes of the circles indicate the intensity of correlation. All correlations in the plot are statistically significant).

Furthermore, the basic statistics (mean and standard deviation) regarding the four mediators are shown in Table 1 below:

表 4.1: Mean and standard deviation of the four mediators by RE frequency.

Variable	RE=0 (n=9153) <sup>2</sup>		RE=1 (n=258)		RE=2 (n=902)		RE=3 (n=1534)		RE $\geq$ 4 (n=744)	
	mean	sd <sup>3</sup>	mean	sd	mean	sd	mean	sd	mean	sd
BMI <sup>1</sup>	26.3	4.2	25.7	3.4	25.2	3.7	25.5	3.6	25.2	3.6
Glucose <sup>1</sup>	99.1	16.9	97.0	9.7	96.3	11.2	97.6	14.6	97.3	12.5
TC <sup>1</sup>	210.3	40.3	199.2	36.7	198.9	36.7	199.8	36.3	195.1	36.3
SBP <sup>1</sup>	120.7	14.2	119.2	13.1	118.9	13.0	120.1	13.4	120.5	13.9

<sup>1</sup> The units for BMI, Glucose, TC and SBP are  $kg/m^2$ ,  $mg/dl$ ,  $mg/dl$  and  $mmHg$ , respectively.

<sup>2</sup> Number of participants included in the corresponded group with specific weekly RE frequency.

<sup>3</sup> sd denotes standard deviation.

By applying a two-sample test on the population with and without resistance exercise, we acknowledge that the mean value is significantly lower in the case of doing RE for all the four risk factors compared to no RE at all. This implies that RE is greatly beneficial on lowering the risk regarding these variables. Furthermore, the overall pattern shows that RE for twice a week seems to achieve the lowest level under these four circumstances except for TC. Therefore, it seems that this RE frequency is considered to be the best routine on human health.

However, the causal effect of RE on the risk regarding total CVD events remains unknown. Since Glucose, TC and SBP are all possible risk factors which mediates the effect of RE on CVD risk, we intend to know which ones are the important mediation pathways besides BMI. Furthermore, the problem on whether the natural direct effects show different patterns or performances among different age groups worth investigation. In order to achieve a relatively comprehensive understanding on these problems, we will develop new methods on better interpreting the underlying mechanism behind our data. A marginal structural model under the case of survival outcome will be proposed to determine the direct causal relationship between RE and CVD risk as well as the causal mediation effects. Then, we will present a multivariate joint modeling approach to find out the potential mediation effects based on both baseline and follow-up clinical observations.

#### 4.4 Causal effect analysis using MSM

Marginal Structural Model (MSM) is widely used in the field of causal inference. In practical problems, we concern about not only correlations among specific variables, but also causal relationships in a deeper sense. Under one or multiple pathways, the total effect of exposure on the target outcome can be decomposed into two parts: natural direct effect, and natural indirect effect. Natural direct effect corresponds to direct influence of exposure on the outcome, which is considered to be a crucial criteria on determining whether causal relationship exists in between. On the contrast, natural indirect effect refers to the impact of exposure on outcome interfered by other risk factors. We illustrate these two concepts using the variables in the ACLS data through a graph below, where  $X$  denotes the confounding variables (current smoking status, heavy alcohol drinking, etc.):

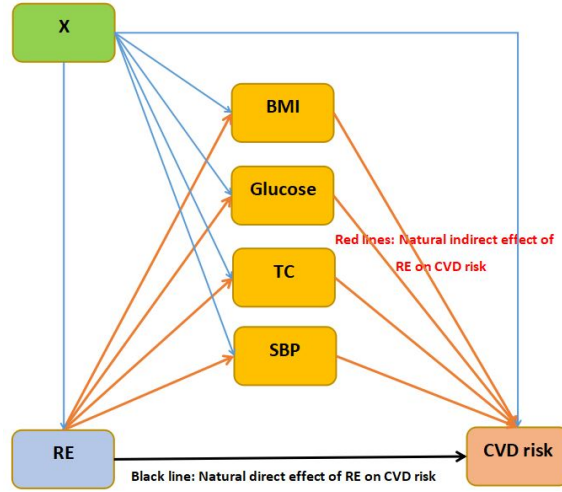


图 4.2: Natural direct and indirect effect of RE on CVD risk.

In the ACLS data, we denote exposure (RE) as  $A$ , the survival time (event time of total CVD) as  $Y$ , and possible mediator (BMI, Glucose, TC, or SBP) as  $M$ . Consider the problem in a binary-exposure case in which  $A$  is regarded as a binary variable ( $A = 0$  means RE absent and  $A = 1$  means RE present) and denote  $Y_{a^*M_a}$  as a nested counterfactual meaning the time to event which would be observed if  $A$  was set to  $a^*$  and mediator  $M$  was set to the value it would have taken if  $A$  was set to  $a$ , then the natural direct effect of  $A$  on  $Y$  is expressed as  $\mathbb{E}\left[\frac{h(Y_{1M_0})}{h(Y_{0M_0})}\right]$  and natural indirect effect as  $\mathbb{E}\left[\frac{h(Y_{1M_1})}{h(Y_{1M_0})}\right]$  under the framework of a survival setting, where  $h(\cdot)$  represents the hazard

ratio function. Thus, the total effect of RE on CVD risk is the multiplication of natural direct and indirect effect:  $\mathbb{E}[\frac{h(Y_{1M1})}{h(Y_{0M0})}]$ . Hong (2010) introduced the concept: Ratio of Mediator Probability Weighting with technical details [22], which is a foundation work for following researches on estimation in MSM (see [30, 31]). The assumption of treatment-by-mediator interaction is relaxed by using this weighting method so as to estimate the natural direct and indirect effects at the same time. Any parametric relationship among the outcome, mediators, exposure and confounding variables can be omitted under this framework. The final outcome is characterized as a function of the natural direct, indirect effect and other confounding variables.

For estimation, we first construct a linear model for  $M$  conditional on both  $A$  and  $X$ . Then an artificial dataset is created by repeating each observation in the original dataset twice and including a new variable  $A^*$ , which is set as the original exposure ( $A$ ) in the first replication and equals to the opposite value of the exposure ( $1 - A$ ) in the second replication [31]. Under the condition of a continuous risk factor  $M$ , we calculate the ratio of mediator probability weight  $W$  by applying the fitted models from the linear model for  $M$  to the newly-generated dataset:  $W = \frac{g^{(a)}(M_a=m|A=a,X)}{g^{(a^*)}(M_{a^*}=m|A=a^*,X)}$ , where  $g^{(a)}(.)$  and  $g^{(a^*)}(.)$  represent the conditional distribution function w.r.t  $A$  and  $A^*$ , respectively. This weight formula is used to adjust the difference between the direct and indirect pathways caused by the counterfactual variable  $M_{a^*}$ , which is considered to achieve more stable estimates when including confounding variables  $X$  in the model [31]. Under Assumption 1, 2, 4, 6, 7, 8 in [22] and the Bayes theorem, we know that the rationale on using weight  $W$  of this form is illustrated below:

$$\begin{aligned}
 & \mathbb{E}(h(Y_{a^*M_a})|X) \\
 &= \int \int_{m,y} h(y) \cdot f(h(Y_{a^*m}) = y|A = a^*, M_a = m, X) \cdot g^{(a^*)}(M_a = m|A = a^*, X) dy dm \\
 &= \int \int_{m,y} h(y) \cdot f(h(Y_{a^*m}) = y|A = a^*, M_{a^*} = m, X) \cdot g^{(a)}(M_a = m|A = a, X) dy dm \\
 &= \int \int_{m,y} h(y) \cdot f(h(Y_{a^*m}) = y|A = a^*, M_{a^*} = m, X) \cdot g^{(a^*)}(M_{a^*} = m|A = a^*, X) \\
 &\quad \cdot \frac{g^{(a)}(M_a=m|A=a,X)}{g^{(a^*)}(M_{a^*}=m|A=a^*,X)} dy dm \\
 &= \mathbb{E}[h(Y) \frac{g^{(a)}(M_a=m|A=a,X)}{g^{(a^*)}(M_{a^*}=m|A=a^*,X)} | A = a^*, X] \\
 &= \mathbb{E}[h(Y)W | A = a^*, X].
 \end{aligned}$$

Finally, we fit a Cox proportional hazard model to event time  $Y$  including only  $A$ ,  $A^*$  and the confounders  $X$  as covariates and weighted by weight  $W$  in the previous step.

The exact form of the model in the last step is shown below:

$$h_i(Y|A, A^*, X) = h_0(Y) \exp(\alpha_1 A_i + \alpha_2 A_i^* + \gamma X_i),$$

where  $h_0(t)$  denotes the baseline hazard ratio and  $\alpha_1, \alpha_2, \gamma$  are the coefficients for parameters  $A, A^*, X$ , respectively. Baseline hazard function  $h_0(Y)$  can be estimated by B-splines. The parameters in this model are estimated through maximizing the partial log likelihoods with respect to  $\alpha_1$  and  $\alpha_2$  multiplied by weight  $W$  in the likelihoods regarding the weighted Cox regression function. Applicability and validity of this estimation methodology under the setting of a survival outcome has been demonstrated in [55] with practical implementations on cancer patients, and is clearly discussed in [30, 31], which claim that the marginal structural model is also applicable with a survival response.

For the Cox proportional hazard model proposed, we can infer that the estimated  $\hat{\alpha}_1$  and  $\hat{\alpha}_2$  represent the log hazard ratios of natural direct and indirect effects. From this, we can also derive the proportion of effects mediated by the potential risk factors:  $\hat{\alpha}_2/(\hat{\alpha}_1 + \hat{\alpha}_2)$ . By doing exponential transformation, we attain the corresponding estimated hazard ratios  $\exp(\hat{\alpha}_1)$  and  $\exp(\hat{\alpha}_2)$ . Since  $\hat{\alpha}_1 + \hat{\alpha}_2$  denotes the total effect, the term  $\exp(\hat{\alpha}_1 + \hat{\alpha}_2)$  means the total hazard ratio of exposure  $A$  on the survival outcome. The total effect can also be obtained by a cox proportional hazard model without including any potential risk factors using the original dataset. Furthermore, confidence intervals regarding the natural direct effect, natural indirect effect and total effect can be derived through bootstrap resampling [31].

With multiple mediators in our model, we need to test whether the mediation trajectories forms nonintertwined causal pathways [30]. Only when each risk factor included is independent of the others can we proceed to the next estimation step by incorporating them together in analysis. Otherwise, we can only include one of them at a time in the following steps.

## 4.5 Construction of multivariate joint modeling

The exercise amount of each participant is defined by weekly RE frequency and working time. In order to determine the effect of RE frequency on the risk of total CVD events with multiple mediators, we propose a multivariate joint model by connecting Cox proportional hazard model and multivariate linear mixed model together for analysis.

We take average of the weekly RE and AE frequency in their follow-up observations to reduce measurement errors. Important confounders in the data are fully considered in the model, including baseline examination year, age, sex, baseline smoking status, heavy alcohol drinking, BMI, Glucose, TC, SBP, parental history of CVD, and meeting the AE guidelines.

In the modeling part, we first fit a multivariate linear mixed effect model on the longitudinal trajectories of the mediators, and then develop a Cox proportional hazard model for the risk of total CVD events. The linear mixed model is constructed as below:

$$Y_{li}(t) = M_{li}(t) + \epsilon_{li}(t),$$

$$M_{li}(t) = (\beta_{l0} + b_{l0i}) + (\beta_{l1} + b_{l1i})t + \beta_{l2}RE_i(t) + \beta_{l3}Z_i(t), l = 1, \dots, k,$$

where  $Y_{li}(t)$  and  $M_{li}(t)$  are the observed and underlying true value of the  $l$ -th mediator for the  $i$ -th subject at time  $t$ , respectively. The error term  $\epsilon_{li}(t)$  follows a normal distribution with mean 0 and covariance  $\Sigma_0$ . Population intercepts and slopes are denoted as  $(\beta_{l0}, \beta_{l1}), l = 1, \dots, k$ . The subject-specific intercepts and slopes  $(b_{l0i}, b_{l1i})$  are random effects following a multivariate normal distribution with mean  $\mathbf{0}$  and an unknown variance-covariance matrix  $\Sigma_1$ , which characterize the temporal trend and the baseline mediation effects and is determined by a correlation matrix  $\mathbf{R}_1$  and variance parameter  $\sigma_1$ . The confounding covariates  $Z_i(t)$  includes examination year, age, sex, current smoking status, alcohol drinking and meeting the AE guidelines in the following analysis.

The risk of total CVD events is modelled by a Cox proportional hazard model:

$$h_i(t|M_{1i}(t), \dots, M_{ki}(t), \mathbf{RE}_i, \mathbf{X}_i) = h_0(t) \exp(\gamma_1 \mathbf{RE}_i + \gamma_2 \mathbf{RE}_i^2 + \gamma_3 \mathbf{X}_i + \sum_{j=1}^k \alpha_j M_{ji}(t)),$$

where  $\mathbf{X}_i$  incorporates all the confounding variables. Function  $h_0(t)$  is the baseline risk which is modelled through B-splines (Rizopoulos, 2012). Parameter  $\alpha_j, j = 1, \dots, k$  denotes the effect of the  $l$ -th mediator on the risk of total CVD events. The coefficients  $\gamma_1$  and  $\gamma_2$  quantify the direct effect of RE on total CVD events, whereas  $\sum_{j=1}^k \alpha_j \beta_{j2}$  represents the indirect effect of RE on CVD risk. The use of a quadratic form regarding RE in the model is demonstrated in [27], which indicates that RE attains a quadratic U-shape form with CVD risk.

Figure 4.2 characterizes the modeling framework with important estimation results of the ACLS data by a flow chart:

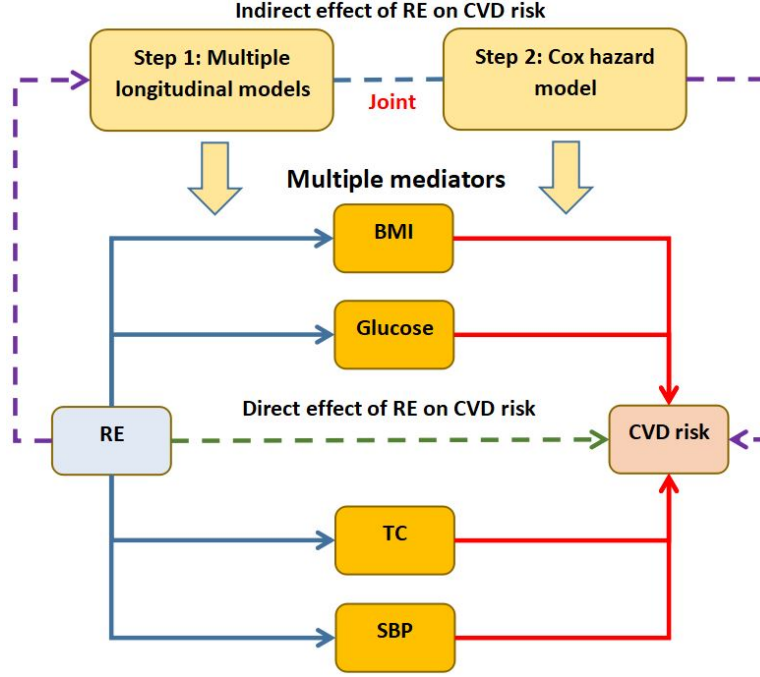


图 4.3: Flow chart of the multivariate joint modeling approach.

The parameters in the multivariate joint model are estimated by a Bayesian method. Denote  $\boldsymbol{\theta}$  to be the whole parameter set. By assuming independence among the longitudinal trajectories, the event time of total CVD events and longitudinal trajectories, and the longitudinal outcomes within each subject, we infer the posterior distribution  $p(\boldsymbol{\theta}, \mathbf{b}_1, \mathbf{b}_2)$  is of the following form:

$$p(\boldsymbol{\theta}, \mathbf{b}_1, \mathbf{b}_2) \propto \prod_{i=1}^n \prod_{l=1}^k \prod_{t=1}^{T_i} p(Y_{li}(t)|b_{l0i}, b_{l1i}, \boldsymbol{\theta})p(T_i|b_{l0i}, b_{l1i}, \boldsymbol{\theta})p(b_{l0i}, b_{l1i}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

where  $p(Y_{li}(t)|b_{l0i}, b_{l1i}, \boldsymbol{\theta})$  is a Gaussian likelihood of  $Y_{li}(t)$  conditional on the random effects  $(b_{l0i}, b_{l1i})$ . For correlation matrix  $\mathbf{R}_1$  regarding random effects in the model, the LKJ-Correlation prior (named after Daniel Lewandowski, Dorota Kurowicka and Harry Joe) is used for Bayesian estimation on  $\mathbf{R}_1$  (see [35]). When the density of the prior for a correlation matrix is proportional to the determinant of which raised to the power of a positive regularization parameter minus one, it is called the LKJ-Correlation prior. It is an extension work to [26], which presents a more efficient method to generate correlation matrix uniformly over the space of all positive definite correlation matrix. The mathematical form of a LKJ distribution is shown below:

$$f(R|\xi) = [2^{\sum_{k=1}^{K-1} (2(\xi-1) + K-k)(K-k)} \prod_{k=1}^{K-1} \{B(\xi + (K-k-1)/2, \xi + (K-k-1)/2)\}^{K-k}] |R|^{\xi-1},$$



where  $K$  is the dimension of correlation matrix  $R$  and  $\xi$  denotes the positive scalar parameter. It is a weakly informative approach for producing priors which outperforms the inverse-Wishart priors in the sense that sampling methods are slow when nonsingular sample is required in the inverse-Wishart case. The LKJ-Correlation prior is more useful and acceptable in Bayesian computing approaches and has widely been used by many research works, see [4] for an example. Furthermore, a half-Student- $t$  prior is utilized to estimate the variance parameter  $\sigma_1$ , which is especially suitable for our data. The half-Student  $t$  prior belongs to the half- $t$  family, which is often used as a weakly informative prior distribution. It is the absolute value of the Student- $t$  distribution which is centered at zero. An expression for the half-Student- $t$  prior of  $\sigma_1$  is given as below:

$$p(\sigma_1) \propto (1 + \frac{1}{v}(\frac{\sigma_1}{u})^2)^{-(v+1)/2},$$

where  $u$  is the scale parameter and  $v$  denotes the degree of freedom. This prior performs more flexible near zero value in comparison to the inverse-gamma family and has the capacity to restrict  $\sigma_1$  from large values. We choose scale parameter  $\xi = 1.5$  for the correlation matrix prior and  $v = 3$  degrees of freedom for the prior of the variance parameters. The other parameters in the longitudinal part of the joint model are all estimated through standard default priors with mean zero. For the survival part, we consider using independent normal priors with mean zero for the coefficients  $\gamma_1, \gamma_2, \gamma_3$  and  $\alpha_j, j = 1, \dots, k$ . Furthermore, the baseline hazard function  $h_0(t)$  is approximated by B-splines which is estimated through setting a hierarchical prior. Detailed information for this is referred to [29]. Finally, an importance sampling method is used to correct for the estimation bias in this multivariate joint modeling procedure (see [48]).

## 4.6 Numerical results

We apply the marginal structural model and multivariate joint modeling approach on our ACLS data in order to determine the mediation effects linking PA to CVD. The causal mechanism between RE/AE and the risk of total CVD events is examined for the baseline data with the existence of multiple mediation effects. Since each pair of the four mediators are intertwined with each other significantly through chi-square test under a generalized linear model, we only incorporate one mediator in the causal pathway at a time ([30]). Therefore, we derive four estimation results corresponded to each of the four mediators. The results on natural direct and indirect effects between RE/AE and

CVD risk are illustrated in the following table (denote natural direct and indirect effect as NDE and NIE, respectively):

表 4.2: Estimated natural direct and indirect effect of RE and AE on the risk of total CVD events.

Risk factor selected	HR of CVD (95% CI) by RE		HR of CVD (95% CI) by AE	
	NDE	NIE	NDE	NIE
BMI	0.441 (0.356,0.547)	0.968 (0.797,1.175)	0.781 (0.628,0.972)	0.943 (0.776,1.144)
Glucose	0.424 (0.342,0.526)	0.999 (0.823,1.212)	0.744 (0.599,0.925)	0.995 (0.820,1.207)
TC	0.430 (0.346,0.533)	0.982 (0.809,1.191)	0.752 (0.605,0.934)	0.982 (0.810,1.191)
SBP	0.430 (0.347,0.533)	0.980 (0.808,1.189)	0.741 (0.595,0.921)	1.001 (0.825,1.215)

The estimation results show that whatever mediator is selected in the model, the natural direct hazard ratio of transforming from no RE to doing RE is significantly lower than 0.6 (lower than 1). This indicates that RE significantly lowers the risk of total CVD events in a causal sense. The four mediators contribute little since the natural indirect effects are all close to 1. By comparing only the values of indirect effects regarding the four mediators, we know that BMI obtains the largest mediation effect, while Glucose being the least one. The results for TC and SBP are very close to each other with SBP attains slightly lower value, so we regard them as approximately equal importance on mediation. Similarly, the results imply that the causal effect of AE on lowering CVD risk is also significant with less power compared to those of RE. BMI also contributes the highest among these four mediators in this case.

Furthermore, we explore the difference of natural direct effect of RE and AE on CVD risk among different age groups. The participants are divided into two groups, with one less than 50 years old, and the other more than 50 years old. Below are the corresponded comparison results:

表 4.3: Comparison for natural direct effect of RE and AE on CVD risk between two age groups.

Mediator selected	HR of CVD (95% CI) by RE		HR of CVD (95% CI) by AE	
	Age<50 years	Age≥50 years	Age<50 years	Age≥50 years
BMI	0.457 (0.321,0.651)	0.426 (0.323,0.560)	0.854 (0.585,1.247)	0.766 (0.585,1.003)
Glucose	0.448 (0.315,0.638)	0.397 (0.301,0.523)	0.823 (0.563,1.205)	0.716 (0.548,0.934)
TC	0.453 (0.318,0.645)	0.405 (0.308,0.533)	0.841 (0.574,1.232)	0.721 (0.553,0.941)
SBP	0.455 (0.319,0.648)	0.402 (0.306,0.529)	0.829 (0.566,1.215)	0.703 (0.538,0.919)

The results for different age groups indicates that the hazard ratios of CVD in the older age group are all relatively lower than those in the younger group whatever RE or AE is set as the exposure. So older people benefits more from resistance and aerobic exercise in prevention of CVD. Furthermore, the decrease rate of natural direct effect is relatively higher in the case of AE compared to RE.

By also incorporating the follow-up clinical trials into the model, we obtain more findings by using the multivariate joint modeling approach. Consider the hazard ratios in a standardized scale (per standard deviation), we conclude that TC is the most important mediator causing the rise of CVD risk, which exceeds the other 3 mediators to a great extent. The corresponded hazard ratio and 95% confidence interval (CI) is 1.343 (1.234,1.536). BMI rates right after TC, which attains a hazard ratio of 1.173 (1.113,1.257). The two least important mediators are SBP and Glucose, with hazard ratios of 1.133 (1.062,1.247) and 1.131 (1.052,1.195), respectively. So the relative importance of these 4 mediators in terms of hazard ratios can be expressed as TC>BMI>SBP>Glucose. This is an important finding since we get to know another mediator more important than BMI in causing CVD: total cholesterol. Table 4 below summarizes the estimation of the regression coefficients and hazard ratios in the survival part of multivariate joint modeling:

**表 4.4: Estimation of regression coefficients and hazard ratios regarding RE and the four mediators in the survival part of multivariate joint modeling.**

Variable	Regression coefficient		Hazard ratio	
	estimate	95% CI	estimate	95% CI
Linear term of weekly RE frequency	-0.675	(-0.775,-0.591)	0.509	(0.461,0.554)
Quadratic term of weekly RE frequency	0.176	(0.153,0.198)	1.192	(1.165,1.219)
BMI (Mediator)	0.159	(0.107,0.228)	1.173	(1.113,1.257)
Glucose (Mediator)	0.124	(0.051,0.178)	1.131	(1.052,1.195)
TC (Mediator)	0.295	(0.210,0.430)	1.343	(1.234,1.536)
SBP (Mediator)	0.125	(0.060,0.221)	1.133	(1.062,1.247)

All the estimates in the survival part of the model are statistically significant. From Table 4, we also notice that RE has a quadratic U-shape form association with the risk of total CVD events. The lowest hazard ratio is attained with RE frequency of 2 times/week. This result is consistent with [43]. Traditional Cox proportional hazard models are also implemented for the ACLS data, which show that resistance exercise

of 1 to 3 times/week or shorter than 120 minutes in time helps lowering CVD risk significantly. However, doing RE with higher frequency or longer time will have counter effects in causing CVD. Stratified analysis is also done to investigate the hazard ratio of total CVD events by weekly RE frequency and RE working time meeting and not meeting AE guidelines, which also shows that moderate RE is beneficial to reducing CVD risk whether meeting AE guidelines or not.

The results for the multiple longitudinal model part implies that RE has significant effects on lowering BMI, Glucose and TC. The regression coefficients and 95% CIs are -0.042 (-0.052,-0.031), -0.083 (-0.159,-0.008) and -0.841 (-1.042,-0.641) corresponded to these three mediators, respectively. By simply multiplying the coefficients with those in the survival part of the model, we learn that BMI, Glucose and TC are all potential mediators which link RE to CVD. The mediation effects corresponded to the three mediator pathways are -0.0067, -0.0103 and -0.2481, respectively. This result further justifies that RE lowers the risk of total CVD events under each potential mediator pathway. However, a slight positive relationship is revealed between RE and SBP with no statistical significance. This is mainly due to the reason that although RE helps lowering blood pressure in the long term, heavy lifting, holding breath during exertion or other inappropriate RE working methods may have side effects on SBP, which prompts it to rise in a short period. Therefore, moderate and scientific RE is required in order to reduce SBP.

## 4.7 Conclusion and Discussion

The novelty of our study lies in the work of incorporating multiple risk trajectories in exploring the mediation effects linking resistance exercise to CVD risk. Besides BMI, we also focus on some other mediators which links RE with CVD risk, including Glucose, TC and SBP. While the MSM model under the case of a survival outcome helps us learn the underlying causal mechanism between RE and CVD with baseline data, the multivariate joint modeling approach gives us more insight into the mediation effects by combining baseline and follow-up data together for analysis. Several important confounding variables are also included in our joint model.

The significant result regarding causal effect of RE on lowering CVD risk marks a great progress in the field of kinesiology. This result shows more significance than high correlation relationship in between. Furthermore, we know that although vascular

responses to RE training were the same for younger and older women [56], the peak forearm blood flow response was greater in the older age group. This provides evidence for our result indicating older people benefits more from resistance and aerobic exercise in the prevention of CVD.

The relative importance rating of the four probable mediators is derived by utilizing the multivariate joint modeling method, which helps us learn more on the mediation effects regarding these longitudinal trajectories besides BMI only (e.g., [12, 43]). Total Cholesterol is the most crucial factor among all four mediators considered and has also been extensively studied by many researchers. In a cohort study of Chinese steelworkers in 2011, the effect of TC is explored and examined on the risk of deadly CVD events. Furthermore, research on examining the correlation of blood pressure and CVD risk is referred to [61]. In their study, the relationship between blood pressure and risk of heart disease is investigated for population from different parts of the world. In practice of clinical trials, efforts have been made on lowering blood pressure and cholesterol in order to reduce CVD risk (See [24, 46, 58]. The combined effect of total cholesterol and systolic blood pressure is also tested on CVD risks in [36]. The study implied that people with high levels of systolic blood pressure and cholesterol attains 7 times higher risk of CVD compared to populations with the lowest levels regarding these 2 combinations. If the highest level of combination is attained, CVD risk can grow up to 12 times higher. In addition, strong evidence suggests that high continuous associations exist between TC and Cardiovascular Heart Disease in both Western and Asian countries. This applies to both middle-aged and old-aged populations [33].

The joint modeling results also show that moderate amount of resistance exercise with approximately twice a week can significantly reduce CVD risk. In fact, engaging in regular RE is beneficial on one's health and body building (See [50, 52, 63]). This is greatly helpful in prevention of CVD [9]. However, RE of no less than 4 times/week does not provide additional benefit and will instead increase the chance of CVD. Large amount of resistance exercise may cause high muscle loads which can bring great stress. This may also lead to Cardiovascular related diseases [44].

However, due to limited variables included in our analysis, only a few mediators is tested to examine whether they are essential in connecting resistance exercise and CVD risks. If more variables are involved in the model, we will have a more comprehensive understanding on this problem and gain more insight in analyzing CVD risks. Maybe

new mediators can be discovered to enrich our cognition in the future.

## 参考文献

- [1] Brooke A Fischer Aggarwal, Ming Liao, and Lori Mosca. Physical activity as a potential mechanism through which social support may reduce cardiovascular disease risk. *The Journal of Cardiovascular Nursing*, 23(2):90, 2008.
- [2] Andrea A Baccarelli, Yinan Zheng, Xiao Zhang, Dou Chang, Lei Liu, Katherine Wolf, Zhou Zhang, John P Mccracken, Anaité Díaz, and Pier Bertazzi. Air pollution exposure and lung function in highly exposed subjects in Beijing, China: a repeated-measure study. *Particle & Fibre Toxicology*, 11(1):51, 2014.
- [3] Randy W Braith and Kerry J Stewart. Resistance exercise training: its role in the prevention of cardiovascular disease. *Circulation*, 113(22):2642–2650, 2006.
- [4] Eike C Brechmann and Harry Joe. Parsimonious parameterization of correlation matrices using truncated vines and factor analysis. *Computational Statistics & Data Analysis*, 77:233–251, 2014.
- [5] Bo Cai, Andrew B Lawson, Md Monir Hossain, Jungsoon Choi, Russell S Kirby, and Jihong Liu. Bayesian semiparametric model with spatially–temporally varying coefficients selection. *Statistics in Medicine*, 32(21):3670–3685, 2013.
- [6] Huaihou Chen, Guanqun Cao, and Ronald A Cohen. Multivariate semiparametric spatial methods for imaging data. *Biostatistics*, 18(2):386–401, 2017.
- [7] Song Xi Chen and Cheng Yong Tang. Nonparametric inference of value-at-risk for dependent financial returns. *Journal of Financial Econometrics*, 3(2):227–255, 2005.
- [8] Zhao-Yue Chen, Tian-Hao Zhang, Rong Zhang, Zhong-Min Zhu, Chun-Quan Ou, and Yuming Guo. Estimating PM<sub>2.5</sub> concentrations based on non-linear exposure-lag-response associations with aerosol optical depth and meteorological measures. *Atmospheric Environment*, 173:30–37, 2018.
- [9] Véronique A Cornelissen, Robert H Fagard, Ellen Coeckelberghs, and Luc Vanhees. Impact of resistance training on blood pressure and other cardiovascular risk factors: a meta-analysis of randomized, controlled trials. *Hypertension*, 58(5):950–958, 2011.

- [10] Jon O Ebbert, Muhamad Y Elrashidi, and Michael D Jensen. Managing overweight and obesity in adults to reduce cardiovascular disease risk. *Current Atherosclerosis Reports*, 16(10):445, 2014.
- [11] Avraham Ebenstein, Maoyong Fan, Michael Greenstone, Guojun He, and Maigeng Zhou. Evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River policy. *Proceedings of the National Academy of Sciences*, 110(32):12936–12941, 2013.
- [12] Aharon Erez, Shaye Kivity, Anat Berkovitch, Assi Milwidsky, Robert Klempfner, Shlomo Segev, Ilan Goldenberg, Yechezkel Sidi, and Elad Maor. The association between cardiorespiratory fitness and cardiovascular risk may be modulated by known cardiovascular risk factors. *American Heart Journal*, 169(6):916–923, 2015.
- [13] Jianqing Fan, Tao Huang, and Runze Li. Analysis of longitudinal data with semi-parametric estimation of covariance function. *Journal of the American Statistical Association*, 102(478):632–641, 2007.
- [14] Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.
- [15] Mario Francisco-Fernandez and Jean D Opsomer. Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics*, 33(2):279–295, 2005.
- [16] Guannan Geng, Qiang Zhang, Randall V Martin, Aaron van Donkelaar, Hong Huo, Huizheng Che, Jintai Lin, and Kebin He. Estimating long-term PM<sub>2.5</sub> concentrations in China using satellite-based aerosol optical depth and a chemical transport model. *Remote Sensing of Environment*, 166:262–270, 2015.
- [17] Alexandros Gryparis, Brent A Coull, Joel Schwartz, and Helen H Suh. Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(2):183–209, 2007.
- [18] Alexandros Gryparis, Konstantina Dimakopoulou, Xanthi Pedeli, and Klea Katsouyanni. Spatio-temporal semiparametric models for NO<sub>2</sub> and PM<sub>10</sub> concentration levels in Athens, Greece. *Science of the Total Environment*, 479:21–30, 2014.



- 
- [19] Yuming Guo, Hongmei Zeng, Rongshou Zheng, Shanshan Li, Adrian G. Barnett, Siwei Zhang, Xiaonong Zou, Rachel Huxley, Wanqing Chen, and Gail Williams. The association between lung cancer incidence and ambient air pollution in China: A spatiotemporal analysis. *Environmental Research*, 144(JAN.PT.A):60–65, 2016.
- [20] Peter Hall, Hans-Georg Müller, and Jane-Ling Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006.
- [21] Wolfgang Härdle. *Applied nonparametric regression*. Number 19. Cambridge University Press, 1990.
- [22] Guanglei Hong. Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, Biometrics Section*, pages 2401–2415. American Statistical Association Alexandria, VA, 2010.
- [23] Siegfried Hörmann, Piotr Kokoszka, et al. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.
- [24] Rod Jackson, Carlene MM Lawes, Derrick A Bennett, Richard J Milne, and Anthony Rodgers. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual’s absolute cardiovascular risk. *The Lancet*, 365(9457):434–441, 2005.
- [25] Wolfgang Jank and Galit Shmueli. Modelling concurrency of events in on-line auctions via spatiotemporal semiparametric models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(1):1–27, 2007.
- [26] Harry Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189, 2006.
- [27] Masamitsu Kamada, Eric J Shiroma, Julie E Buring, Motohiko Miyachi, and I-Min Lee. Strength training and all-cause, cardiovascular disease, and cancer mortality in older women: A cohort study. *Journal of the American Heart Association*, 6(11):e007677, 2017.

- [28] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367, 2008.
- [29] Stefan Lang and Andreas Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.
- [30] Theis Lange, Mette Rasmussen, and Lau Caspar Thygesen. Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, 179(4):513–518, 2014.
- [31] Theis Lange, Stijn Vansteelandt, and Maarten Bekaert. A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, 176(3):190–195, 2012.
- [32] Carl J Lavie, Ross Arena, Damon L Swift, Neil M Johannsen, Xuemei Sui, Duck-chul Lee, Conrad P Earnest, Timothy S Church, James H O’ Keefe, Richard V Milani, et al. Exercise and the cardiovascular system: clinical science and cardiovascular outcomes. *Circulation Research*, 117(2):207–219, 2015.
- [33] Malcolm R Law, Nicholas J Wald, and SG Thompson. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ*, 308(6925):367–372, 1994.
- [34] Duck Chul Lee, Russell R. Pate, Carl J. Lavie, Xuemei Sui, Timothy S. Church, and Steven N. Blair. Leisure-time running reduces all-cause and cardiovascular mortality risk. *Journal of the American College of Cardiology*, 64(5):472–481, 2014.
- [35] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009.
- [36] Sarah Lewington and Robert Clarke. Combined effects of systolic blood pressure and total cholesterol on cardiovascular disease risk. *Circulation*, 112(22):3373–3374, 2005.
- [37] Yehua Li. Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, 98(2):355–370, 2011.

- 
- [38] Xuan Liang, Shuo Li, Shuyi Zhang, Hui Huang, and Song Xi Chen. PM<sub>2.5</sub> data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121(17):10–220, 2016.
- [39] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing’s PM<sub>2.5</sub> pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.
- [40] Xihong Lin and Raymond J Carroll. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96(455):1045–1056, 2001.
- [41] Xihong Lin and Raymond J Carroll. Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):69–88, 2006.
- [42] Yang Liu, Christopher J Paciorek, and Petros Koutrakis. Estimating regional spatial and temporal variability of PM<sub>2.5</sub> concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, 117(6):886–892, 2009.
- [43] Yanghui Liu, Duck-Chul Lee, Yehua Li, Weicheng Zhu, Riquan Zhang, Xuemei Sui, Carl J Lavie, and Steven N Blair. Associations of resistance exercise with cardiovascular disease morbidity and mortality. *Medicine and Science in Sports and Exercise*, 51(3):499–508, 2019.
- [44] John C Longhurst, Charles L Stebbins, et al. The power athlete. *Cardiology Clinics*, 15(3):413–429, 1997.
- [45] Chunsheng Ma. Semiparametric spatio-temporal covariance models with the ARMA temporal margin. *Annals of the Institute of Statistical Mathematics*, 57(2):221–233, 2005.
- [46] Giuseppe Mancia. Defining blood pressure goals: is it enough to manage total cardiovascular risk? *Journal of Hypertension*, 27:S3–S8, 2009.
- [47] Randall V Martin. Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42(34):7823–7843, 2008.

- [48] Katya Mauff, Ewout Steyerberg, Isabella Kardys, Eric Boersma, and Dimitris Rizopoulos. Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 30:999–1014, 2020.
- [49] Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization, et al. *Global atlas on cardiovascular disease prevention and control*. Geneva: World Health Organization, 2011.
- [50] Miriam E Nelson, Maria A Fiatarone, Christina M Morganti, Isaiah Trice, Robert A Greenberg, and William J Evans. Effects of high-intensity strength training on multiple risk factors for osteoporotic fractures: a randomized controlled trial. *JAMA*, 272(24):1909–1914, 1994.
- [51] Ralph Paffenbarger. Physical exercise to reduce cardiovascular disease risk. *Proceedings of the Nutrition Society*, 59(3):421–422, 2000.
- [52] Michael L Pollock, Glenn A Gaesser, Janus D Butcher, Jean-Pierre Després, Rod K Dishman, Barry A Franklin, and Carol Ewing Garber. Acsm position stand: the recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness, and flexibility in healthy adults. *Medicine & Science in Sports & Exercise*, 30(6):975–991, 1998.
- [53] C Arden Pope Iii, Richard T Burnett, Michael J Thun, Eugenia E Calle, Daniel Krewski, Kazuhiko Ito, and George D Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9):1132–1141, 2002.
- [54] John A Rice and Bernard W Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1):233–243, 1991.
- [55] Justine Rochon, Andreas du Bois, and Theis Lange. Mediation analysis of the relationship between institutional research activity and patient survival. *BMC Medical Research Methodology*, 14(1):9, 2014.

- 
- [56] Linda Marie Rossow. Cardiovascular and muscular responses to eight weeks of resistance exercise training in young and older women. *Dissertations & Theses Gradworks*, 2013.
- [57] Ina Shaw and Brandon S Shaw. Resistance training’s role in the prevention of sports injuries. In *Sports Injuries*, pages 123–136, 2015.
- [58] Neil J Stone, Jennifer G Robinson, Alice H Lichtenstein, C Noel Bairey Merz, Conrad B Blum, Robert H Eckel, Anne C Goldberg, David Gordon, Daniel Levy, Donald M Lloyd-Jones, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2889–2934, 2014.
- [59] Minghui Tao, Liangfu Chen, Rong Li, Lili Wang, Jun Wang, Zifeng Wang, Guiqian Tang, and Jinhua Tao. Spatial oscillation of the particle pollution in eastern China during winter: Implications for regional air quality and climate. *Atmospheric Environment*, 144:100–110, 2016.
- [60] María Victoria Toro, Lázaro V Cremades, and Josep Calbó. Relationship between VOC and NO<sub>x</sub> emissions and chemical production of tropospheric ozone in the Aburra Valley (Colombia). *Chemosphere*, 65(5):881–888, 2006.
- [61] Peggy CW van den Hoogen, Edith JM Feskens, Nico JD Nagelkerke, Alessandro Menotti, Aulikki Nissinen, and Daan Kromhout. The relation between blood pressure and mortality due to coronary heart disease among men in different parts of the world. *New England Journal of Medicine*, 342(1):1–8, 2000.
- [62] Aaron Van Donkelaar, Randall V Martin, Michael Brauer, and Brian L Boys. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental Health Perspectives*, 123(2):135–143, 2015.
- [63] Kevin R Vincent, Randy W Braith, Ross A Feldman, Henrique E Kallas, and David T Lowenthal. Improved cardiorespiratory endurance following 6 months of resistance exercise in elderly men and women. *Archives of Internal Medicine*, 162(6):673–678, 2002.

- [64] Jun Wang, Xiaoguang Xu, Robert Spurr, Yuxuang Wang, and Easan Drury. Improved algorithm for MODIS satellite retrievals of aerosol optical thickness over land in dusty atmosphere: Implications for air quality monitoring in China. *Remote Sensing of Environment*, 114(11):2575–2583, 2010.
- [65] Lili Wang, Nan Zhang, Zirui Liu, Yang Sun, Dongsheng Ji, and Yuesi Wang. The influence of climate factors, meteorological conditions, and boundary-layer structure on severe haze pollution in the Beijing-Tianjin-Hebei region during january 2013. *Advances in Meteorology*, 2014(685971):1–14, 2014.
- [66] Naisyin Wang. Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90(1):43–52, 2003.
- [67] Naisyin Wang, Raymond J Carroll, and Xihong Lin. Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100(469):147–157, 2005.
- [68] Ying Wang, Guoshun Zhuang, Chang Xu, and Zhisheng An. The air pollution caused by the burning of fireworks during the lantern festival in Beijing. *Atmospheric Environment*, 41(2):417–431, 2007.
- [69] Sara Wilcox, Deborah Parra-Medina, Melva Thompson-Robinson, and Julie Will. Nutrition and physical activity interventions to reduce cardiovascular disease risk in health care settings: a quantitative review with a focus on women. *Nutrition Reviews*, 59(7):197–214, 2001.
- [70] Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.
- [71] Haipeng Xing and Zhiliang Ying. A semiparametric change-point regression model for longitudinal observations. *Journal of the American Statistical Association*, 107(500):1625–1637, 2012.
- [72] Peirong Xu, Jun Zhang, Xingfang Huang, and Tao Wang. Efficient estimation for marginal generalized partially linear single-index models with longitudinal data. *Test*, 25(3):413–431, 2016.

- [73] Fumo Yang, Jihua Tan, Qingquan Zhao, Z Du, K He, Y Ma, F Duan, and GJAC Chen. Characteristics of PM<sub>2.5</sub> speciation in representative megacities and across China. *Atmospheric Chemistry & Physics*, 11(11):1025–1051, 2011.
- [74] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.
- [75] Yue Zhang and Bin Zhang. Semiparametric spatial model for interval-censored data with time-varying covariate effects. *Computational Statistics & Data Analysis*, 123:146–156, 2018.





## 致谢

首先感谢我的导师姚方老师，姚方老师对我的要求很严格，体现在工作的严谨性，学术的精益求精，开阔思路视野，以及灵活变通能力，受益颇深，使我思考问题和解决问题的能力、主观能动性均有了很大的提高。非常感谢姚方老师。

感谢黄辉老师对我的帮助。黄老师是我博士期间第一任导师，对我的要求很严格，在课题讨论中对我思考问题的角度和灵活性有很大的启发和帮助，为我在学术方面的发展奠定了坚实的基础。谢谢黄辉老师。

感谢陈松蹊老师，读博期间很有幸加入了陈老师课题组，参与京津冀大气污染评估课题，在采集整理数据、分析数据上有很大帮助。在非参数统计领域中受益很大。谢谢陈松蹊老师。

同时感谢课题组梁萱、郭斌、张澍一、邹韬等师兄师姐和陈磊同学对我的帮助。

感谢李业华老师。李老师是我美国学访的指导老师。李老师在联合建模、生存分析给予了我很多指点，对我的论文逐句精雕细刻，拓宽思路，要求很高。我的收获很大。谢谢李业华老师。

感谢我的父母对我的帮助和支持。感谢在我成长道路上遇到的所有人。

最后，感谢各位评审老师的仔细阅读和宝贵意见。谢谢！



## 北京大学学位论文原创性声明和使用授权说明

### 原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：                    日期：      年      月      日

### 学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 ☐ 一年 / ☐ 两年 / ☐ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：                    导师签名：                    日期：      年      月      日