The Pennsylvania State University
The Graduate School
Eberly College of Science

NEW STATISTICAL PROCEDURES FOR ANALYSIS OF HIV DATA AND HIGH DIMENSIONAL DATA

A Dissertation in
Statistics
by
Jingyi Ye

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December, 2017

The dissertation of Jingyi Ye was reviewed and approved* by the following:

Runze Li
Verne M. Willaman Professor of Statistics
Dissertation Co-Advisor, Co-Chair of Committee

Le Bao
Associate Professor of Statistics
Dissertation Co-Advisor, Co-Chair of Committee

Zhibiao Zhao
Associate Professor of Statistics

Lan Kong
Professor of Public Health Sciences

Aleksandra Slavkovic
Professor of Statistics
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

This dissertation consists of two parts. In the first part, we develop a new statistical procedure for analysing HIV data to improve efficiency of parameter estimates by incorporating extra available information. Also, we use the procedure to study the impact of this additional information. In the second part, we develop a new error variance function estimation procedure for ultrahigh dimensional varying coefficient models.

The human immunodeficiency virus (HIV) is a lentivirus that causes HIV infection and acquired immunodeficiency syndrome (AIDS). Accurate estimation and prediction of HIV epidemics can help people have a better understanding on HIV epidemics, and also help government make laws and formulate policies. Two key indicators, prevalence and incidence, are widely used to estimate HIV epidemics. HIV prevalence is the proportion of HIV positive population among the general population. HIV incidence is the proportion of new HIV infections among the general population.The new treatment, Antiretroviral treatment (ART), reduces the AIDS-related deaths, and changes the AIDS-related mortality rate substantially. In the UNAIDS 2014 Gap report, the number of people who are newly infected with HIV is continuing to decline in most countries and regions in the world, which suggests a slow-down of HIV epidemics. Traditionally, the increase of HIV prevalence rate mostly is due to the increase of new infections. However, reduction of AIDS-related deaths becomes another important reason of increasing HIV prevalence rate. In this case, knowing incidence helps people fuller understand the HIV epidemic. One of our goals is to utilize the newly available incidence assays in the process of estimation and projection of HIV epidemics, and to understand the contribution of such data in the presence of historical HIV prevalence data, which has been the main data source for estimating HIV epidemics. The Susceptible-Infectious-Recovered (SIR) system is widely used in the epidemiology. Under Bayesian framework, Incremental Mixture Importance Sampling (IMIS) can

be used to draw the posterior samples. The current method to incorporate new incidence assays with SIR model is to fit the historical prevalence data and new incidence data all over again even if the fitted results to the prevalence data are available. We propose a new method, Sequential IMIS, to estimate prevalence and incidence with assay data. Our method reduces the computing time in most scenarios, and enables the study the impact of incidence assay data in multiple scenarios. Also, we improve the stopping rule for IMIS to avoid the algorithm stops in the local maximum. Incidence assay data are impact by four parameters: prevalence, incidence, the false recent rate (FRR), and mean duration of recent infection (MDRI). We use the proposed method to study the impact to prevalence and incidence rates and impact to the changes of prevalence and incidence rates over time when incorporating the new incidence assay data. This impact takes both one time data and time series data into consideration. Our research shows that in most countries, incidence assay data can significantly improve the accuracy of the incidence estimate.

In the second part, we propose a new estimation procedure for error variance function estimation for ultrahigh dimensional varying coefficient models (VCM). Low dimensional VCM was systematically introduced in Hastie and Tibshirani (1993), and is one of the most commonly used nonparametric regression models in statistics. Error variance function estimation plays important roles in estimation of confidence interval and hypothesis testing for VCM, and is very challenging in the present of ultrahigh dimensional covariates. A naive way is to select variables first, and refit the model with low dimensional selected models. We first show both theoretically and empirically that this naive estimator significantly underestimates the error variance and may lead to an inconsistent estimate. We further propose a new estimation procedure for error variance function by using group least absolute shrinkage and selection operator (LASSO) and refitted cross-validation (RCV) techniques. We study the asymptotic property of the RCV estimate, and compare it with the naive estimate. Our findings include that the RCV estimator is consistent estimator and follows an asymptotic normal distribution with smallest variance. It significantly improves the naive estimator. We further conduct simulation studies to examine the finite sample performance of the RCV estimate.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisors Dr. Runze Li and Dr. Le Bao. They are great mentors for me, and are very supportive of both my research and career development. Throughout my Ph.D. study, they gave me great suggestions for my research topics, and guided me with their patience and enthusiasm. This dissertation would not have been possible without their guidance.

I would also like to thank my other committee members: Dr. Zhibiao Zhao, and Dr. Lan Kong for their insightful comments and kind support. They help me to fuller understand my research from other related academic fields for both theoretically perspective and applied perspective.

I am really grateful to all of those with whom I have had the pleasure to work during this and other related projects. They are my collaborators: Dr. Timothy B. Hallett, Dr. Zhao Chen, Dr. Jingyuan Liu, and those who gave valuable advice and help: Dr. Alex Welte, Dr. Xiaoyue Niu, Dr.Tim Brown and Dr. Kimberly Marsh.

Last but not least, I would like to thank my family, especially my parents: Dr. Xiangdong Ye, and Dr. Mengping Zhang. They have been great support for me during the five years.

The HIV research is supported by the Joint United Nations Programme on HIV/AIDS, the Bill & Melinda Gates Foundation, and National Institute of Allergy and Infectious Diseases, NIH, R56 AI120812-01A1. And the RCV project is funded by National Science Foundation, DMS 1512422 and National Institute on Drug Abuse, NIH, P50 DA039838.

Chapter 1
Introduction

This dissertation consists of two parts. The first part is devoted to developing more efficient statistical estimation procedures, Sequential Incremental Mixture Importance Sampling (S-IMIS), to incorporate incidence assays. Also, we use this new procedure to study the impact of assay data. In the second part, We propose a new error variance estimation procedure for varying coefficient models in the presence of ultrahigh dimensional covariates. This introduction will consist of two parts respectively, one for S-IMIS and the other one for error variance estimation in ultrahigh dimensional regression models.

1.1 Introduction for S-IMIS

1.1.1 Motivation

The human immunodeficiency virus (HIV) is a lentivirus that causes HIV infection and acquired immunodeficiency syndrome (AIDS). It ranks sixth in the causes of death among all people, and ranks second in the low-income group, according to World Health Organization (WHO) 2012 report. Two key indicators, prevalence and incidence, are widely used to estimate HIV epidemics. Prevalence is

the proportion of people living with HIV infection at a given time, such as at the end of a given year. Incidence is the proportion of new HIV infections that occur during a given year. Prevalence is the commonly used indicator, because incidence is hard to be measured due to lack of accurate testing method. HIV/AIDS epidemics are defined by the HIV prevalence in the general population. Based on the prevalence level in the population, we describe the HIV epidemics as concentrated epidemics and generalized epidemics. Generalized epidemics are defined as the epidemics with HIV prevalence among pregnant women above 1% on a national basis, which are considered as high-prevalence. generalized or low level epidemics are defined as HIV prevalence below 1% in the general population, but exceeds 5% in specific at-risk population (UNAIDS, 2000).

Since mid-1970s, HIV starts an explosive epidemic around the world. It becomes critical to estimate the HIV trend accurately. Fortunately, HIV prevalence did not increase infinitely, it saturated at some level as most other epidemics (UNAIDS, 1999). At that time, because of technology restriction, our focus is to estimate prevalence. After two decades, the prevalence of HIV reached the peak in most countries, which already resulted in more than a quarter of young and middle-aged adults are infected with HIV in many cities in sub-Saharan African countries (MAP, 2001). In most countries with available and adequate HIV surveillance data, there was an observed fall in HIV prevalence around 1990s (UNAIDS, 1999). Recent years, the number of people who are newly infected with HIV is continuing to decline in most parts of the world, which suggests a slow-down of HIV epidemics (UNAIDS, 2014b).

Numerous work has been done by governments and organizations all over the world, and we have made a great achievement by significantly reducing the new HIV infection during the past two decades. One most important achievement is availability of antiretroviral treatment (ART), which has substantially reduced the AIDS-specific mortality. According to UNAIDS Gap report 2014, *since 1995,*

ART has averted 7.6 million deaths globally, and have gained approximately 40.2 million life-years since the epidemic started. However, the situation is still critical especially in some low-income countries. Globally, nearly 75% of all people with HIV live in fifteen countries. In every region of the world, the majority of people with HIV live in three to four countries (UNAIDS, 2014b). Therefore, controlling new infections in those high risk countries becomes the critical part to fulfill the Fast-Track strategy to end the AIDS epidemic by 2030 (UNAIDS, 2014a). To achieve the goal, a 90% decline compared 2030 to 2010 of the number of new HIV infections and AIDS-related deaths is required.

Since 1980s, a long-time series of surveillance data has been collected, which allows us to estimate HIV prevalence and incidence. Due to technology restriction, in the past, only HIV prevalence data is available. HIV prevalence data records HIV status, and classifies population into HIV positive and negative groups. There are several kinds of prevalence data, such as the national population-based survey (NPBS) data, antenatal clinical (ANC) data, the prevention of mother-to-child transmission (PMTCT) data, and voluntary counseling and testing (VCT) data.

NPBS is a door-to-door survey with HIV related questions and HIV testing. It collects the most accurate prevalence data in most generalized epidemic countries. Countries with NPBS conducted give a better performance on the estimates, such as smaller confidence intervals than the countries without (UNAIDS, 2014c). NPBS provides estimates of HIV prevalence in the national level, as well as for different subgroups, such as different location groups, different age groups, and different sex groups. It is a major advantage compared to other HIV prevalence data. Also, it can be linked to other information, such as social, behavioral and other biomedical information. Those combined information gives the scientists the chance to study the epidemics in details (Gouws et al., 2008). However, due to the high cost of this survey, NPBS data is only available for very few years, especially in the low-income countries. Moreover, a selection bias might be exists. One of the reason is the

intention to reject the test of people with acknowledge of their infection. Another reason is the absence of survey due to the limited survey time (WHO/UNAIDS, 2003). A systematic selection model can be applied here to correct the bias, when HIV positive individuals are systematically opting out of HIV testing (Marra et al., 2015).

Another important data source is antenatal clinical (ANC) data. It records the HIV status of pregnant women attending antenatal clinics. ANC data is the major data source to assess the HIV trends (Ghys et al., 2006). Obviously, we have a clinical bias and a selection bias since pregnant women usually have more sex activities. Moreover, for those countries with limited participating clinics, the concern about the ANC data limited coverage arises(WHO/UNAIDS, 2003). But still, it is the most feasible data with acceptable cost, especially in those countries cannot afford frequent national surveys. In more than 115 countries worldwide, ANC data is available (UNAIDS/WHO/CDC, 2003).

There are also some other prevalence data such as the prevention of mother-to-child transmission (PMTCT), and voluntary counseling and testing (VCT). However, they are less available, representative of a small section of the population (sex workers, occupational groups), or subject to additional bias (for example, VCT). There is a much better data, case-report data, which records all the HIV infections in the countries. However, it is only available in very few high-income countries. In this dissertation, our focus in in Africa countries. Therefore, NPBS data and ANC data are the main data sources.

From the discussion above, we can see that it is not easy to estimate HIV prevalence and incidence with those surveillance data. First, there is a serious sampling bias, for example, only pregnant women are tested in ANC data. Since data from reviewed scientific publications are not available for many localities, it is very hard to estimate the bias accurately based on the historical data. Second, sample

size might be very small considering the scale of the clinics and the availability of the national survey in those countries with war zone. Moreover, in some low-income countries, the reliability of diagnostic tests is a big concern. In the last, there are some countries without any national survey and the estimates may have bad national representative (UNAIDS, 1999). All those reasons lead to a difficulty to estimate HIV prevalence accurately, not mention the HIV incidence.

In the past, due to the high epidemics, our main focus is on HIV prevalence. However, the estimation methods were pretty rough. As long-time surveillance data becomes available, more complicated models have been proposed. For concentrated epidemics, the workbook method is currently recommended. For generalized epidemics, UNAIDS has developed a new tool, Estimation and Projection Package (EPP), to construct the epidemic curves. EPP gives national and sub-national epidemic curves. It allows us to study the levels and trends easily. Since our main focus is in Africa, where the HIV prevalence is relative large, EPP model is our primary model is this dissertation. This model incorporates different data sources, which makes the model available even for new data sources. For countries with unclear epidemic level, both methods can be applied (Ghys et al., 2004).

1.1.2 Challenges for Estimating the Impact of Incidence Assays Data

Traditionally, the increase of HIV prevalence rates mostly due to the increase of new infections. However, reduction of AIDS-related deaths becomes another important reason of increasing HIV prevalence rates. In this case, knowing prevalence is not enough, we also need to know incidence to help people fully understand HIV epidemics. One of the United Nation (UN) Millennium Sustainable Development Goals is that “by 2030, end the epidemics of AIDS". The goal requires strong control on the new infection population, which is monitored by incidence rates. In

this dissertation, one of our major goals is to study the impact of newly available incidence assay data to the current incidence estimate, and to improve related computing algorithm.

Considering the limited budget in most generalized epidemic countries, accurate estimation and prediction of both prevalence and incidence rates become an important part to control the HIV epidemics. It can help people to have a better understanding on HIV epidemics, and also help government make laws and formulate policies. Multiple billion dollars has been used to stop HIV epidemics each year. Accuracy of the estimation and prediction may determine the usage of billion dollars, which affects millions of people's life.

Traditionally, due to lack of availability of incidence data, incidence is hard to estimate. One easy way to roughly estimate HIV incidence is to use the difference of prevalence estimates. For example, the incidence rate in 1980 is approximately the difference between prevalence estimates in 1979 and 1980 after considering the immigration, mortality and other social information. This method is acceptable in the early years since the difference was mostly due to the new infection. However, after the epidemic peak around 1990s, HIV prevalence rates started to fall in most countries. In some area, high HIV prevalence is due the longer years of HIV positive population. It is obviously not appropriate to use the estimates of prevalence rate to cover the usage of estimates of incidence rates. Estimating incidence becomes a new focus, which makes the availability on incidence data become an urgent work.

Recently, new technology makes it possible to test for new infection efficiently, which allows the collection of incidence data. Incidence data classifies the population into three groups, HIV negative groups, HIV non-recent infected groups and HIV recent infected groups. The main incidence data we discuss in this dissertation is incidence assay data. The data estimates the infectious window period based on antibody assays (Busch et al., 2010). Now, incidence assay data is collected

under the framework of national population based survey. The new NPBS further classifies the HIV positive groups into recent and non-recent infected groups. To avoid confusion, NPBS data in this dissertation still refers to the prevalence data we talked before. The new incidence data is called incidence assay data. Naturally, the new incidence assay data directly gives us information on the HIV incidence, comparing to "borrow" information from prevalence data. So, theoretically, we can have a better estimates on incidence based on incidence data. Since prevalence and incidence are highly related, it is necessary to use both data sources to estimate prevalence and incidence. Therefore, how to incorporate the incidence data to the old models and the impact to the old estimates become new problems.

Incidence data came to reality only about five years ago, and there are still argument about its real value. For governments, without a significant improve on estimates, a huge money used for incidence tests can be saved for other fields, especially for those countries with very limited budgets. In this dissertation, one of our goals is to estimate the impact of new incidence assays to the current prevalence and incidence estimates, and also improve related computing algorithm.

The major challenge to estimate impact of assay data, is the small sample size. Assay data came into reality only for a few years, and the data is inadequate in most Africa countries. Considering the small sample size of the original NPBS data, the final number of HIV positive individuals in the sample is very small. So, the different incidence rate does not change the likelihood as much as the different prevalence. For example, in Kenya rural area, the sample size of NPBS data in year 2012 is 7501, among which the HIV positive number is 383. According to the estimates from without assay data, the number of HIV recent-infected is around 15. Even the incidence estimates directly from assay data has a big difference from the estimates from estimates without assays, it can affect the final estimates very little considering the very small sample size. Another challenge is about the quality of incidence assays. Considering the small sample size, there is possible

randomness which can affects the quality of assay data. Also, since this is related new technology, there is also doubt of the reliably. This situation will become better after several years of incidence assay data been collected in the future.

1.1.3 Contributions

As we discussed above, multiple data sources are used to estimate the prevalence and incidence. For the new data source, incidence assay data, we first introduce the method to incorporate it to the existing model (Bao, Ye, and Hallett, 2014). Our goal is to study the impact of the assay data, and justify its real value. In that case, we cannot afford to wait for enough data and then study the impact. So, simulation study becomes a good strategy. This can consider all the possible scenarios of new assays. One nature way to incorporate the assay data, is to add a new data structure in EPP and fit the historical prevalence data and new incidence data all over again. However, we have to consider a large quantity of scenarios in order to give an overall suggestion to the impact of incidence data. There are too many countries and variables affecting the incidence estimates. It leads to an unaffordable computing cost. It is very time-consuming and even a waste to re-run the whole EPP again, especially when we already have the estimates without assays. In this dissertation, we propose sequential-IMIS to reduce the computing time by using the existing EPP results from prevalence data without assays. Our method reduces the computing time in most scenarios, and enables the study of the impact of incidence assay data in multiple scenarios. Also, we improve the stopping rule for IMIS to avoid the algorithm stops in the incorrect local maximum.

In order to study the impact, two impacts are taken into consideration, one is the impact to the estimator of prevalence and incidence, and another one is the changes of prevalence and incidence over time. We simulate all kinds of scenarios of incidence data based on different FRR and MDRI. We first consider

the simplest case, in which we will only have one-year incidence assay data. Based on the method, a time series incidence data is naturally the next study goal. We impose a simple time series structure on the infection rate to ensure the simulation covering both increasing and decreasing incidence trend. Our research shows that in most countries, incidence assay data can significantly improve the accuracy of the incidence estimate.

1.2 Introduction for Error Variance Estimation

High-dimensional data have frequently been collected in various scientific research areas such as tumor classifications, biomedical imaging, genomics, tomography and finance. Analysis of high-dimensional data poses many challenges for statisticians. As demonstrated in Donoho et al. (2000), there is urgent need for developing new statistical methodologies and theories for high-dimensional data analysis. A comprehensive overview of statistical challenges with high-dimensionality in various statistical problems can be found in Fan and Li (2006). Various challenges in analysis of big data may be found in Fan, Han, and Liu (2014). In the last two decades, statisticians were devoted to developing new variable selection and feature screening procedures, which are fundamental for high-dimensional data analysis. There have been a huge number of research works on high dimensional data analysis in the literature. It is impossible for us to give a comprehensive review here. Readers are referred to Fan and Lv (2008), Bühlmann and Van De Geer (2011) and references therein. Statistical inference on high-dimensional data calls for new statistical methodologies and theories.

1.2.1 Error Variance Estimation

Error variance estimation plays a critical role in statistical inference for high dimensional regression models. Confidence/Prediction interval construction and testing hypotheses on regression coefficients all require an accurate estimate of the error variance. For linear regression with finite dimensional predictors, the adjusted mean squared error provides an unbiased estimate of the error variance, and it performs well when the sample size is much larger than the number of predictors.

In the presence of ultrahigh dimensional covariates, error variance estimation indeed is a challenging task. It has been empirically observed that the mean squared error estimator leads to an underestimation of the error variance when model is significantly over-fitted. Specifically, it is typical to impose sparsity assumption in the presence of ultrahigh dimensional covariates. Feature screening and variable selection procedures will be implemented to reduce dimensionality. This leads to spurious correlation. Inclusion of spuriously correlated variables leads to significantly over-fitted model. As a result, this leads to an underestimate of error variance. On the other hand, if one does not impose sparsity assumption, the full model is saturated and all residuals equal 0. As a result, residual sums of squares equals 0, and the mean squared errors does not perform well in this case.

Sun and Zhang (2012) developed the scaled LASSO methods for estimating the regression coefficients and error variance jointly. They further demonstrated the scaled LASSO works reasonably well for error variance estimation in high dimensional linear models. Clearly, it is challenging to extend the scaled LASSO for other models. Fan, Guo, and Hao (2012) demonstrated the challenges of error variance estimation in the high-dimensional linear regression analysis. They firstly confirmed that the ordinal adjusted mean squared errors is an underestimate of the error variance, and it is not a consistent under certain conditions. They further developed an accurate error variance estimator by introducing refitted cross-

validation (RCV) techniques. Reid, Tibshirani, and Friedman (2013) compared eleven methods for error variance estimation methods. Readers are referred to their paper for details.

Due to the complex structure of high dimensional data, the high dimensional linear regression analysis may be a good start, but it may not be powerful to explore nonlinear features inherent into data. Nonparametric regression modeling provides valuable analysis for high dimensional data (Negahban et al., 2009; Hall and Miller, 2009). This is particularly the case for error variance estimation, as nonparametric modeling reduces modeling biases in the estimate. This paper aims to study issues of error variance estimation in ultrahigh dimensional nonparametric regression settings. Chen, Fan, and Li (2016) further extended the RCV method to ultrahigh dimensional additive models. The second part of this dissertation aims to develop new error variance estimation for ultrahigh dimensional varying coefficient models, which was systematically studied by Hastie and Tibshirani (1993).

1.2.2 Contribution

Let Y be a response variable, U be a continuous covariate and $\mathbf{X}$ be a p-dimensional covariate vector. Varying coefficient model to be studied in this dissertation has the following form

$$Y = \mathbf{X}^T \boldsymbol{\alpha}(U) + \sigma(U)\varepsilon,$$

where $\boldsymbol{\alpha}(u) = (\alpha_1(u), \cdots, \alpha_p(u))^T$ consists of the nonparametric coefficient functions, ε is a random error with mean 0 and variance 1. The second part of this dissertation aims to construct a root n consistent estimator for $\sigma^2(u)$ in the presence of ultrahigh dimensional covariate (i.e, $\log(p) = O(n^\zeta)$, for some $\zeta > 0$, where n is the sample size). We call $\sigma^2(u)$ to be the varying error variance function to

emphasize it is not a constant error variance. Existing work on error variance estimation is limited to estimate constant error variance. Thus, this dissertation is the first one to consider estimation of error variance function under ultrahigh dimensional setting.

In Chapter 5, we first study the behavior of naive error variance function estimator. The naive estimator is a three-step procedure: (a) we employ local linear grouped LASSO to select important variables; (b) we use local linear regression to estimate the coefficient functions in the selected model in order to avoid estimation bias inherent in the LASSO procedure, and (c) we applied kernel regression on the squared residuals to estimate the error variance function. We prove that this naive estimator has non-ignorable bias, and is not root nh_n convergent, where h_n is the bandwidth used in the kernel regression in Step (c).

To avoid the non-ignorable bias of the naive estimator, we propose a new estimator for the error variance function based on the RCV techniques (Fan, Guo, and Hao, 2012; Chen, Fan, and Li, 2016). Both Fan, Guo, and Hao (2012) and Chen, Fan, and Li (2016) proposed the RCV technique to estimate constant variance based on (global) least squares method, in which the projection matrix can be easily obtained. There are several challenging in using RCV techniques to estimate the error variance function using local model techniques. The first hurdle is the numerical optimization related to local linear group LASSO. To deal with these issues, we propose coordinate block descent (BCD) algorithm for the local linear group LASSO. The second hurdle is to develop the theoretical properties of the RCV estimator. The challenge is the difference between the observed responses and the fitted values cannot be represented as the same way as that for constant error variance estimator. This requires to develop entirely new theory for the newly proposed RCV estimator. We systematically study the asymptotical properties of the proposed RCV estimator, and establish the asymptotical normality of the resulting RCV estimator. The asymptotical normality implies that the proposed

RCV estimator shares the same asymptotic variance as the oracle estimator in which one knew the true values of regression coefficient functions in advance. Our Monte Carlo simulation study confirms our theoretical findings.

1.3 Organization of This Dissertation

Chapter 2 gives a literature review on the model and statistical procedures related to this dissertation research. Section 2.1 describes the history, structure, mechanics of EPP model and the within Bayesian framework. Section 2.2 describes IMIS algorithm. Section 2.3 presents the estimation methods for low dimensional varying coefficient model, and Section 2.4 summarizes a brief review on variable selection and feature screening. Section 2.5 discusses the challenge of analysing ultrahigh dimensional data, and introduces refitted cross-validation (RCV) method to estimate error variance along with some other common methods.

In Chapter 3, we propose sequential IMIS algorithm. Section 3.1 describes the proposed method and algorithm thoroughly. Section 3.2 is the simulation study based on EPP model to show the advantages of our proposed method.

In Chapter 4, we further study the impact of the new incidence assay data. Section 4.1 is the general methods and goals for the numerical study in this chapter. Section 4.2 studies the impact of single year incidence assays, and Section 4.3 studies the impact of time series incidence assays.

In Chapter 5, we devote to developing the error variance function estimation procedure for the ultrahigh dimensional varying coefficient models. Section 5.1 gives a brief introduction to the error variance estimating problem for high dimensional data. Section 5.2 introduces the model for ultrahigh dimension VCM with heteroscedastic error. Section 5.3 proposes the three-stage naive estimator for the error variance, along with the theoretical results. Section 5.4 proposes an RCV

estimator for the error variance. Theoretical property of RCV estimator is also given. Section 5.5 develops the algorithm to implement group LASSO, which is the variable selection method in this chapter. Then the simulation study and real data example are put in Section 5.6 and 5.7 respectively. Section 5.8 is the detailed proofs of the theory along with the regularity conditions.

Chapter 2
Literature Review

In the first part of this chapter, we introduce the Estimation and Projection Package (EPP) framework and the within computing method, Incremental Mixture Importance Sampling (IMIS), in Section 2.1 and 2.2 respectively. And in the second part, Section 2.3 presents the estimation methods for low dimensional varying coefficient model, and Section 2.4 summarizes a brief review on variable selection and feature screening. Section 2.5 discusses the challenge of analysing ultra-dimensional data, and introduces refitted cross-validation (RCV) method to estimate error variance along with some other common methods.

2.1 Estimation and Projection Package (EPP)

We have briefly talked about EPP software in previous section. In this section, we further introduce EPP model, including the history and model details.

2.1.1 History

Since 1980s, HIV surveillance data becomes available, which allows us to study the trend of HIV prevalence and incidence. The method to derive the HIV prevalence

and incidence estimates in early 1990s are pretty naive, which can only derive the point estimation. The method to derive the WHO end-1994 estimates is simply to use the HIV survey prevalence estimate of certain group times the group size proportion in the whole population, and take the summation based on the group. Estimated number of HIV positive individual $= \sum_i R_i p_i / n_i$, where p_i is the HIV positive in the sample group i, n_i is the group sample size, and R_i is the estimated population size of group i (Burton and Mertens, 1998). Then, it was extended to a two-step method to derive the WHO end-1997 country-specific estimates. The first step is to produce the point estimates from 1994 to 1997, then the second step is to fit an epidemic curve which could describe the spread of HIV in each country. The AIDS epidemic software, EPIMODEL, is used to calculate estimates of incidence and mortality from this epidemic curve. This is the first time that the HIV prevalence trend is estimated properly (Schwartländer et al., 1999). Then in 1999, Joint United Nations Programme on HIV/AIDS (UNAIDS) Reference Group on Estimates, Modelling and Projections is created to provide guidance on the procedures and assumptions used in preparing estimates of HIV/AIDS and its impact (Walker et al., 2003).

Estimation and Projection Package (EPP) is a software developed by UNAIDS to estimate national or sub-national time-series prevalence and incidence. Besides incidence and prevalence projection, EPP also produces deaths, and AIDS impacts, which can be the input to Spectrum (Brown et al., 2006). The Spectrum Projection Package is a tool to study the consequence of AIDS, such as AIDS-related mortality, AIDS group effect and other epidemiology information (Stover, 2004). EPP is first introduced in the UNAIDS/WHO end-2001 estimates for the generalized epidemics. There are four parameters in the model in the original EPP model: infection rate r, starting date $t(0)$, the peak prevalence at starting time with respect to each at-risk group $f(0)$, and final endemic prevalence parameter Φ. The method is to find the best curve with least square estimation of the four parameters, which minimizing

the square difference between the curve and the data points (Walker et al., 2003; Zaba et al., 2002). In this version, only prevalence projection is available. UNAIDS also gives a detailed introduction of EPP (Ghys et al., 2004).

Since then, EPP software and model have been updated every few years. In EPP 2005, the major updates of the model include: incidence trends are available; the infection rate $\mathtt{r}$ is allowed to change over time; EPP can adjust HIV prevalence considering the general population information (Brown et al., 2006). In EPP 2007, the Bayesian melding approach is added (Brown et al., 2008). It is part of the current method in EPP. In EPP 2009, ART information is available, which includes the CD4 eligibility criterion, HIV progression rate and numbers on first- and second-line ART. Also, IMIS is first introduced to use in the EPP model, which helps to draw posterior samples under the Bayesian framework (Brown et al., 2010). And r-stochastic model was first introduced into EPP. In EPP 2011, two new infection models, r-spline model and r-trend model are added, which allow the infection rate to change over time much more smooth with limited parameters (Hogan and Salomon, 2012). Also, the integration of EPP and Spectrum came into reality (Stover et al., 2012). The basic model structure has not changed since then (Bao et al., 2012).

The most recent update is in 2013. In this version, the range of selection of models is wider (Brown et al., 2014). The main model in Chapter 3 and Chapter 4, which is described in EPP 2013 (Bao et al., 2012). It is consisted of three parts: epidemiological dynamic models, an infection rate model, and data models.

2.1.2 Epidemiological Dynamic Models

The UNAIDS EPP 2013 is based on a simple susceptible-infected-removed(SIR) epidemiological model. We only consider the population with age between 15 to 50. At any given time $\mathtt{t}$, the targeted population is divided into two groups: $\mathsf{Z}(\mathtt{t})$ is the

number of uninfected individuals, and $Y(t)$ is the number of infected individuals. $N(t) = Z(t) + Y(t)$ is naturally the whole population size. We use the following differential equations to describe the rates at which the sizes of the groups change:

$$\begin{cases} \frac{dZ(t)}{dt} = E(t) - \frac{r(t)Y(t)Z(t)}{N(t)} - \mu Z(t) - \frac{a_{50}(t)Z(t)}{N(t)} + \frac{M(t)Z(t)}{N(t)}, \\ \frac{dY(t)}{dt} = \frac{r(t)Y(t)Z(t)}{N(t)} - HIVdeath - \frac{a_{50}(t)Y(t)}{N(t)} + \frac{M(t)Y(t)}{N(t)}. \end{cases} \tag{2.1.1}$$

At any given time t,

- $E(t)$ is the number of new adults entering the targeted population (age between 15-50), which depends on the population size of 15 years ago, the birth rate and the survival rate from birth to age 15,
- $r(t)$ is the average infection risk,
- μ is the non-AIDS death rate,
- HIV death is the AIDS related death rate,
- $a_{50}(t)$ is the number of adults exit the model after attaining age 50,
- $M(t)$ is the number of net migration into the population.

In the model, $E(t), \mu, a_{50}(t)$ and $M(t)$ are given either directly from the government statistic or from Spectrum software. They are considered fixed parameters. $r(t)$ can be derived from the infection model introduced following. Only $Z(t)$ and $Y(t)$ are the parameters need to be estimated.

2.1.3 Infection Models

The infection rate, $r(t)$, is the rate of infection, which defines as the average number of new infections caused by one HIV positive person at year t. In the earlier time, we often assume a constant infection as in EPP 2001, which is acceptable due to the high infection. However, as epidemic slows down, especially after ART comes to the reality, there are significant evidences of changing infection rate.

Constant infection rate is no longer a good choice. Instead of assuming a constant infection rate, several refined models were proposed to allow $r(t)$ change: r-jump model, r-spline model, r-stochastic model, and r-trend model. The r-jump model assumes that there is only one change in the infection rate. It adds the variation to the infection rate, however also leaves the problem that when to performance the "jump". It is hard to explain why the infection rate has a such sudden change at one point (Brown et al., 2008). The other three are more flexible models which allow the infection rate to change smoothly over the time of the epidemic. The stochastic model assumes the infection rate follows a Gaussian random walk with mean zero (Bao and Raftery, 2010). This model provides more flexibility than the r-jump model, but the parameter estimation is challenging when we have a long history of the epidemic since the infection rate at each year is treated as a parameter. The r-spline model assumes infection rate depend on time, and can be modeled by penalized B-spline (Hogan et al., 2010; Hogan and Salomon, 2012; Brown et al., 2014). R-trend model assumes the current infection rate depends on the past prevalence, the past incidence, and a stabilization condition. In this dissertation, we mainly use r-trend model according to UNAIDS recommendation (Bao et al., 2012).

R-trend Model

With epidemic starting time t_0 and initial infection rate r_0, the r-trend model assumes that the average infection risk $r(t)$ can be described as (Bao, 2012):

$$\log(r_{t+1}) - \log(r_t) = \beta_1(\beta_0 - r_t) + \beta_2\rho_t + \beta_3\gamma_t, \tag{2.1.2}$$

where

- β_0 is equilibrium condition which lead to no shift of $\log(r_t)$ at $r(t) = \beta_0$,
- β_1 describes the change of $\log(r_t)$, when it is not at equilibrium value,

- β_2 is the expected change of $\log(r_t)$ given a unit increase of the prevalence,
- β_3 describes the related change of $\log(r_t)$ and stabilization status,
- ρ_t is the prevalence rate at time t,
- $\gamma_t = (\rho_{t+1} - \rho_t)(t - t_0 - t_1)^+/\rho_t$ is the relative change of prevalence times the positive part of $t - t_0 - t_1$,
- t_0 is the starting year of epidemic,
- t_1 is the number of years that epidemic takes to stabilize.

We can see that r-trend model considers more factors which can affect the infection rate. Taken seven parameters, β_1, β_2, β_3, r_0, t_0, t_1, into consideration, r-trend model explores the shape of HIV prevalence trend in a more restricted parameter space than the r-spline model.

R-spline Model

B-spline is the most common tool in the functional data analysis to approximate a target function. In the r-spline model (Hogan et al., 2010; Hogan and Salomon, 2012; Brown et al., 2014), seven equal space functions were used as the basis function with coefficients $\beta_i, i = 1, \ldots 7$. Also, we apply a second-degree difference penalty on the coefficients, which can be described as:

$$\beta = \beta i - 1 + (\beta_{i-1} - \beta_{i-2}) + u_i, \tag{2.1.3}$$

where $u_i \sim N(0, \tau^2)$. Then we have 9 variables in r-spline model, coefficients $\beta_i, i = 1, \ldots, 7$, τ, along with initial pulse of infection to seed the epidemic Y_0. The model is more flexible, however with a big problem: It is risky to predict beyond the data period using spline models because the prediction is driven by a relatively small number of observations near the last year of data. To stabilize the prediction, after last year of data t_1, instead of B-spline, there are two ways to model the trend. One way, we assume $\log(r(t))$ follows a random work with mean $\log(r(t))$

and σ_t^2. Also, the variance σ_t^2 has the formula: $\sigma_t^2 = \sigma_{t1}^2(t - t_1)$. Another way is to use "equilibrium prior". We further assume that after the data year, prevalence will approach equilibrium, and $r(t)$ follows a normal distribution. It leads to a step-by-step trend of $r(t)$.

2.1.4 Data Models

It is clear that it is very hard to estimate all 8 or 10 parameters using only the epidemiology models and infection model. In this case, Bayesian framework is used to find posterior distribution of our target HIV epidemic indicators, prevalence and incidence. Under Bayesian framework, we need to find out the likelihood of each data source. Here we described the data models for commonly used prevalence data, antenatal clinical (ANC) data and the national population-based survey (NPBS) data, along with the newly available incidence assay data.

ANC data

ANC data is the major source in Africa countries to monitor the HIV epidemic trend. For clinic s in year t, the data records two numbers: the number of infected pregnant women, Y_{st}, and the number of pregnant women tested, N_{st}. In EPP 2013, ANC data model is described by a hierarchical model with a clinic random effect b_s accounting for the repeated measurement within clinic (Brown et al., 2014):

$$\Phi^{-1}(X_{st}) = \Phi^{-1}(\rho_t) + c + b_s + \epsilon_{st}, \tag{2.1.4}$$

where

- Φ is the standard normal cumulative distribution function,
- $X_{st} = (Y_{st} + 0.5)/(N_{st} + 1)$,
- ρ_t is the overall population prevalence in year t,

- c is the calibration constant for the ANC data, which determined based on the historical data:

$$\begin{cases} c = 0.11, & \text{for urban area,} \\ c = 0.17, & \text{for rural area.} \end{cases} \tag{2.1.5}$$

Now, we need to estimate likelihood of ANC data. However, there is no close form of this likelihood considering the complicated structure. In Alkema et al. (2007), they proposed a method to calculate the likelihood by integrating out σ^2 and b_s.

Assume γ_{st} is the prevalence at clinic s in year t, and Y_{st} then follows a binomial distribution $Y_{st} \sim \text{binom}(N_{st}, \gamma_{st})$. Then X_{st} is the posterior mean of γ_{st} with a non informative Jeffery's prior Beta$(\frac{1}{2}, \frac{1}{2})$. Then we have:

$$\begin{cases} b_s \sim N(0, \sigma^2), \\ \epsilon_{st} \sim N(0, \nu_{st}), \\ \nu_{st} = 2\pi \exp(\Phi^{-1}(\gamma_{st})^2)\gamma_{st}(1-\gamma_{st})/N_{st}. \end{cases} \tag{2.1.6}$$

We also assign a prior to σ^2 during the calculating:

$$\sigma^2 \sim \mathit{InverseGamma}(0.58, 93). \tag{2.1.7}$$

Then the likelihood becomes:

$$L(ANC|\rho) = \int\int L(ANC|\rho, b)p(b|\sigma^2)dbp(\sigma^2)d\sigma^2. \tag{2.1.8}$$

where $p(\cdot)$ stands for the prior distributions. We can numerically estimate the likelihood by using existing software functions.

NPBS data and assay data

The national population-based survey data give the information of HIV epidemic at the national level. NPBS data records two numbers for each country in year t: the number of uninfected people N_S and the number of infected people N_I. If the HIV positive individuals in NPBS data were further tested for whether they were recently infected by using incidence assays, then incidence assay data would separate N_I into two groups: non-recent infected people with number N_{NR}, and recent infected people with number N_R, where naturally $N_{NR} + N_R = N_I$.

Following the work of Welte et al. (2010), we assume (N_S, N_R, N_{NR}) follow a trinomial distribution with the following probabilities (Kassanjee et al., 2012):

$$\begin{cases} P(S) = 1 - \rho_t, \\ P(R) = (1 - \rho_t) I_t (\Omega_t - \beta_t T)/365 + \beta_t \rho_t, \\ P(NR) = \rho_t - (1 - \rho_t) I_t (\Omega_t - \beta_t T)/365 + \beta_t \rho_t, \end{cases} \tag{2.1.9}$$

where

- ρ_t and I_t are EPP output prevalence and incidence (calibrated to the national level) at time t,
- β_t is the false recent rate (FRR) at time t,
- Ω_t is the mean duration of recent infection (MDRI) at time t, and
- $T = 450$ is the cut-off length of recent infection.

So, for NPBS data, the number of infected people follows a binomial distribution:

$$N_I \sim (N_S + N_I, P(I)).$$

For assay data, the number of recent infected people further follows a binomial distribution:

$$N_R \sim \left(N_R + N_{NR}, \frac{P(R)}{P(R) + P(NR)} \right).$$

Then the log-likelihoods are defined as:

$$\begin{cases} l_{NPBS} = N_I P(I)(1 - P(I)) + (N_R + N_{NR}))(P(R) + P(NR)), \\ l_{assays} = N_R \dfrac{P(R)}{P(R) + P(NR)} + N_{NR} \dfrac{P(NR)}{P(R) + P(NR)}. \end{cases} \tag{2.1.10}$$

2.1.5 Prior Assumptions

For Bayesian framework, we first define the prior function of all parameters. In the ANC data structure, there is a parameter of the random clinical effect b_s. In this dissertation we define the prior as:

$$\begin{cases} b_s \sim N(0, 0.04^2), \quad c = 0.11, \quad \text{for urban area}, \\ b_s \sim N(0, 0.05^2), \quad c = 0.17, \quad \text{for rural area}. \end{cases} \tag{2.1.11}$$

In the r-trend model (2.1.2), there are seven parameters: β_0, β_1, β_2, β_3, t_0, t_1 and r_0. Prior distributions of those seven parameters are defined as:

$$\begin{cases} t_0 \sim \text{Uniform}[1970, 1990], \\ t_1 \sim \text{Uniform}[10, 30], \\ r_0 \sim \text{LogUniform}[\dfrac{1}{11.5}, 10], \\ \beta_0 \sim N(0, 0.2), \\ \beta_1 \sim N(0, 0.2), \\ \beta_2 \sim N(0, 0.2), \\ \beta_3 \sim N(0, 0.2). \end{cases} \tag{2.1.12}$$

Therefore, we have eight parameters in total for r-trend model.

In the r-spline model (2.1.3), there are nine parameters: $\beta_i, i = 1, \ldots, 7$, τ,and

Y_0. Prior distributions of those nine parameters are defined as:

$$\begin{cases} \tau^2 \sim \text{inverse-gamma}(0.001, 0.001), \\ Y_0 \sim \text{Uniform}[10^{-13}, 0.0025], \\ \beta_1 \sim N(1.5, 1), \\ \beta_i \sim N(0, \tau^2), \quad i = 2, \ldots, 7. \end{cases} \tag{2.1.13}$$

Therefore, we have ten parameters in total for r-spline model.

2.1.6 How EPP Works

Under Bayesian framework, the basic formula is:

$$f(\boldsymbol{\theta}|x) = \frac{f(x|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(x)} \propto L(\boldsymbol{\theta}|x)p(\boldsymbol{\theta}) \tag{2.1.14}$$

where $L(\boldsymbol{\theta}|x)$ is the data likelihood and $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. In Section 2.1.4, we already present all the data structures, and log-likelihoods of three data sources are l_{ANC}, l_{NPBS}, and l_{assays}. Then, The targeted log-likelihood is the summation of all the data source log-likelihood with the following format:

$$\begin{cases} \text{EPP without assays: } l_{EPP} = l_{ANC} + l_{NPBS}, \\ \text{EPP with assays: } l_{EPP} = l_{ANC} + l_{NPBS} + l_{assays}. \end{cases} \tag{2.1.15}$$

One of the advantage of EPP model is that it produces all key epidemic indicators together through the dynamic systems as described in Equation (2.1.1). However, we need to input $r(t)$ at each time step to make the equations work. Either r-trend or r-spline model involves with parameters more than we can directly estimate. Therefore, Bayesian framework is used. We can use the prior functions defined in Section 2.1.5, along with the epidemiology models and infection model to derive

the posterior samples of our targeted indicators.

One of the problems to study posterior samples of HIV prevalence and incidence, $\boldsymbol{\eta}$ = (prevalence, incidence) = (ρ, I), is that we have no idea of their prior functions ($p(\boldsymbol{\eta})$). What we have is the prior functions of the parameters $\boldsymbol{\theta}$ (8 in r-trend, 10 in r-spline). However, our data log likelihood functions l_{ANC}, l_{NPBS}, and l_{assays}, are based on HIV prevalence and incidence $\boldsymbol{\eta}$. The linkage between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is EPP model, which is defined as:

$$\boldsymbol{\eta} = g(\boldsymbol{\theta}), \tag{2.1.16}$$

where $g(\cdot)$ is the mapping from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ by EPP model. However, due to the complication of EPP, there is no close form of $g(\cdot)$.

Algorithm 2.1.1. *Take r-trend model as an example to illustrate the algorithm:*

algorithm for EPP model
1. Draw n prior samples of original parameters $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$.
2. For any sample $\boldsymbol{\theta}^{(i)} = (\beta_0, \beta_1, \beta_2, \beta_3, t_0, t_1, r_0, b_s)$,
(2.1) at time 1, using all the parameters, $r(1)$ is derived (by 2.1.2),
(2.2) using $r(1)$, $Y(1)$ and $Z(1)$ can be derived by (2.1.1),
(2.3) then, $\rho_1 = Y(1)/N(1)$, and $I_1 = r(1)Y(1)/N(1)$,
(2.4) using $\beta_0, \beta_1, \beta_2, \beta_3, t_0, t_1$ and ρ_1, $r(2)$ is derived.
3. Move the time forward, following the procedure in 2, until we have all the prevalence and incidence estima
$\boldsymbol{\eta}^{(i)}$ in all time steps.
4. Repeat 2-3 for all the samples $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$, and get $\boldsymbol{\eta} = \boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(n)}$.
For any time t,
5. Draw the posterior samples of $\boldsymbol{\eta}_t$ by certain method.
6. Take the median of posterior samples as the estimator at time t.

Then, we have the estimators of the prevalence and incidence at all time steps. One naturally problem arises is the method to draw the posterior samples. In some early versions of EPP, simple Sampling Importance Resampling (SIR) algorithm (Rubin, 1987) under Bayesian framework is used to draw the posterior samples of $\boldsymbol{\eta}$ (Alkema et al., 2007). The procedures are following:

Algorithm 2.1.2.

SIR algorithm for EPP model to draw posterior distribution
Once we have $\boldsymbol{\eta} = \boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(n)}$,
1. calculate the target likelihood function of $\boldsymbol{\eta}_t$, $L(\cdot\|\boldsymbol{\eta}_t^{(1)}), \ldots, L(\cdot\|\boldsymbol{\eta}_t^{(n)})$.
2. Calculate the weight of each sample, $w_t^{(i)} = \frac{L(\cdot\|\boldsymbol{\eta}_t^{(i)})}{\sum_{i=1}^{n} L(\cdot\|\boldsymbol{\eta}_t^{(i)})}, \quad i = 1, \ldots n.$
3. Draw the posterior samples of $\boldsymbol{\eta}_t$ with replacement with probabilities $w_t^{(1)}, \ldots, w_t^{(n)}$.

In the early version of EPP, since we only have 4 parameters, SIR was the primary algorithm (Alkema et al., 2007). As more data are collected, and the model flexibility is extended (from 4 parameters to 8 or 10 parameters), the computing cost has increased significantly. We need to simulate a large number of samples in order to cover the parameter space. As a response, UNAIDS updated the primary algorithm to IMIS to generate the posterior samples more efficient. The basic procedures are very similar to the procedures above, except we repeat "reweight-resample" multiple times until the algorithm converges. More details of IMIS will be given in the following section.

2.2 Incremental Mixture Importance Sampling (IMIS)

When we fit a Bayesian model, one of our major goals is to find posterior distribution. Sometimes, we can achieve that easily by using conjugate priors and Gibbs sampler. When we have analytic form of the exact posterior density function or the density function with an unknown normalizing constant, some simple sampling algorithm, such as inverse cumulative density function method (Devroye, 1986), rejection sampler (Gilks and Wild, 1992), and ratio-uniform sampler (Luengo and Martino, 2012) can be used to simulate the target distribution. However, in reality, the close form of posterior distribution is often not available. Therefore, we need to describe the unknown posterior distribution by drawing posterior samples with numerical method. Alternatively, we will use equation (2.1.14) to derive the

posterior samples:

$$f(\boldsymbol{\theta}|x) \propto L(\boldsymbol{\theta}|x)p(\boldsymbol{\theta}).$$

Then, the problem moves on to calculating likelihood and the normalization constant.

2.2.1 Methods to Generate Posterior Distribution

When direct sampling from target distribution is difficult, there are two common choices: Markov Chain Monte-Carlo (MCMC) and importance sampling. MCMC method is a class of algorithms to sample a distribution based on Markov chain, including Metropolis-Hastings (M-H) algorithm, Gibbs sampler, etc. The primary application of MCMC method is to calculate numerical approximations of multi-dimensional integrals. One of the examples is expectation. Assume X is a random variable with density function $f(\cdot)$, the goal is to calculate the expectation of function of X, $E[h(X)]$ with respect to f.

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \int h(x)f(x)dx = E[h(X)]. \tag{2.2.1}$$

This technique can be easily used to generate the normalization constant.

Importance sampling is the main focus in Chapter 3 and Chapter 4. *Importance sampling is a general technique for estimating properties of a particular distribution, while only having samples generated from a different distribution than the distribution of interest* (Chakraverty, 2014). For importance sampling, we have:

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i)\frac{f(X_i)}{g(X_i)}g(X_i) \to \int h(x)\frac{f(x}{g(x)}g(x)dx = E[h(X)], \tag{2.2.2}$$

where $g(\cdot)$ is an importance sampling function, called envelope, which is usually a common distribution, such as normal or uniform distribution.

One application of importance sampling, is Sample-Importance-Resample(SIR) algorithm (Givens and Hoeting, 2012). SIR is an approximation simulation method, which is an approach not exact simulation of target density function $f(\cdot)$. Still, we need to define an envelope $g(\cdot)$ first.

Algorithm 2.2.1.

General SIR algorithm
1. Draw n samples $\mathbf{x_1}, \ldots, \mathbf{x_n}$ from $g(\cdot)$.
2. Calculate the standardized importance weights,
$w_i = \frac{f(\mathbf{x_i})/g(\mathbf{x_i})}{\sum_{i=1}^{n} f(\mathbf{x_i})/g(\mathbf{x_i})}, \quad i = 1, \ldots n.$
3. Resample $\mathbf{y_1}, \ldots, \mathbf{y_m}$ with replacement with probabilities $w_1, \ldots, w_n$.
Then $\mathbf{y_1}, \ldots, \mathbf{y_m}$ form the shape of target density function.

The procedure is usually described as Sample-Importance-Resample. Under Bayesian framework, we consider the prior distribution density as the envelope. Then we have the weights as in Algorithm 2.1.1, which are simply proportion to log likelihoods. Due to the convenient property, SIR is commonly used to draw Bayesian posterior samples.

SIR is very easy and straightforward, and performances well for smooth and uni-modal distribution. However, it remains some problems. First, under Bayesian framework, we will choose a simple prior function, such as normal or uniform distribution, and simulate a large quantity of samples to cover the range of $\boldsymbol{\theta}$ as much as possible. As the number of parameters increases, we need to increase the initial sample size rapidly to ensure sufficient coverage of the high density region of the target distribution. For example, if $\theta \in [0, 100]$, we sample 1000 to cover the range, then for $\boldsymbol{\theta} \in [0, 100] \times [0, 100]$, we will need 1000^2 samples. When the dimension of $\boldsymbol{\theta}$ comes to 8 as r-trend model, we will need 1000^8 samples, which is not feasible in our research considering the complication of EPP model. Secondly, SIR has a bad performance when the posterior distribution has nonlinear ridges or be multi-modal. Therefore, Raftery and Bao (2010) derived Incremental Mixture Importance Sampling (IMIS) to fix the those problems.

2.2.2 Introduction to IMIS

IMIS was originally developed for calculating normalizing constants for finite mixture models, which is the integrated likelihood (Steele et al., 2006). The basic idea of IMIS is that instead of a large quantity of initial samples, a median size of initial samples are generated. Then, after calculating the importance weights of these samples, in each iteration, we will mainly sample around the points with high importance weight in last iteration, instead of sampling evenly in the range. Therefore, we "save" a lot of samples in the low-weight area, and can describe the high-weight area more clearly. When applying IMIS to EPP model, the algorithm becomes (cited from Raftery and Bao (2010)):

Algorithm 2.2.2. (IMIS)

1. *Initial Stage:*

   - *Sample* N_0 *inputs* $\theta_1, \theta_2, \ldots, \theta_{N_0}$ *from the prior distribution* $p(\theta)$.
   - *For each* θ_i*, calculate the likelihood* L_i*, and form the importance weights:*

$$w_i^{(0)} = \frac{L_i}{\sum_{j=1}^{N_0} L_j}.$$

2. *Importance Sampling Stage: For* $k = 1, 2 \ldots,$ *repeat the following steps:*

   - *Choose the current maximum weight input as the center* $\theta^{(k)}$*. Estimate* $\Sigma^{(k)}$ *as the weighted covariance of the* B *inputs with the smallest Mahalanobis distances to* $\theta^{(k)}$*, where the distances are calculated with respect to the covariance of the prior distribution and the weights are taken to be proportional to the average of the importance weights and* $\frac{1}{N_k}$.
   - *Sample* B *new inputs from a multivariate Gaussian distribution* H_k *with covariance matrix* $\Sigma^{(k)}$.

- *Calculate the likelihood of the new inputs and combine the new inputs with the previous ones. Form the importance weights:*

$$w_i^{(k)} = cL_i \times \frac{p(\theta_i)}{q^{(k)}(\theta_i)},$$

where c *is chosen so that the weights add to 1,* $q^{(k)}$ *is the mixture sampling distribution* $q^{(k)} = \frac{N_0}{N_k} p + \frac{B}{N_k} \sum_{s=1}^{k} H_s$, H_s *is the* s*-th multivariate normal distribution, and* $N_k = N_0 + Bk$ *is the total number of inputs up to iteration* k.

3. *3. Resample Stage: Once the stopping criterion is satisfied, resample* J *inputs with replacement from* $\theta_1, \ldots, \theta_{N_k}$ *with weights* $w_1, \ldots, w_{N_K}$*, where* K *is the number of iterations at the importance sampling stage.*

The algorithm has several control parameters to be set by the user: the number of initial samples N_0, the sample size at each importance sampling iteration B, and the number of resamples J. The algorithm is unbiased for any choice of control parameters, because it is an importance sampling algorithm, but the control parameters can affect its efficiency. We have found good results with the choices $N_0 = 1000d$, $B = 100d$ and $J = 3000$, where d is the dimension of the integrand.

A great advantage of importance sampling is that it is effectively self-monitoring, in that poor coverage of the target distribution by the importance sampling is immediately seen by the presence of large importance weights. We use the following specific criteria to assess the performance of the various importance sampling algorithms considered here:

- The maximum importance weight among the N_k inputs.
- The variance of the rescaled importance weights,

$$\widehat{V}(w) = \frac{1}{N_k} \sum_{i=1}^{N_k} (N_k w_i - 1)^2$$

- The entropy of the importance weights relative to uniformity,

$$\widehat{U}(w) = -\sum_{i=1}^{N_k} w_i \frac{\log(w_i)}{\log(N_k)}$$

- The expected number of unique points after re-sampling,

$$\widehat{Q}(w) = \sum_{i=1}^{N_k} (1 - (1 - w_i)^M)$$

- The effective sample size, $ESS(w) = \frac{N_k}{1+CV}$, where the coefficient of variation CV is defined as $CV = \frac{Var_q[L(\theta)p(\theta)/q(\theta)]}{E_q^2[L(\theta)p(\theta)/q(\theta)]} \doteq \widehat{V}(w)$ (Kong et al., 1994). We can also write

$$ESS(w) = \frac{N_k}{1 + \widehat{V}(w)} = \frac{N_k}{1 + N_k \sum_{i=1}^{N_k} w_i^2} - 2\sum_{i=1}^{N_k} w_i + 1 = \frac{1}{\sum_{i=1}^{N_k} w_i^2}$$

By using IMIS, the total sample we need to draw is $N_k = N_0 + Bk$, which is normally smaller than SIR sample size. The iteration stops when the number of unique points is larger than $(1 - e^{-1}) \times B$. This is the expected fraction when the importance sampling weights are all equal, which is the case when the importance sampling function is the same as the target distribution.There are four numbers calculated for each iteration: the raw marginal likelihood, the expected number unique points, the maximum weight, and the effective sample size. We intend to stop the algorithm, as the change of marginal likelihood goes to 0, the expected number of unique points and the effective sample size goes to B, and the maximum weight approaches $1/N_k$.

2.3 Varying Coefficient Models

In this section, we review statistical procedures for varying coefficient models (VCM). Let Y be the response variable, U and $\mathbf{X} = (X_1, \ldots, X_p)^T$ be its associated covariates. The VCM assumes that

$$Y = a_0(u) + a_1(u)X_1 + \cdots + a_p(u)X_p + \varepsilon, \tag{2.3.1}$$

where α_k's are unknown regression coefficients, and ε is a random error with $E(\varepsilon|\mathbf{X}) = 0$. This model was systematically studied in Hastie and Tibshirani (1993), in which the authors observed that

$$a_j(u) = \frac{E\left[X_j\{Y - \sum_{k\neq j} X_k a_k(u)\}|u\right]}{E\left(X_j^2|u\right)} = \frac{E\left[X_j^2\{Y - \sum_{k\neq j} X_k a_k(u)\}/X_j|u\right]}{E\left(X_j^2|u\right)}. \tag{2.3.2}$$

This offers a natural interpretation of the regression coefficient function in the VCM.

The VCM is a nonparametric regression model. Hastie and Tibshirani (1993) introduced two estimation methods of the coefficients. The first one is based on spline smoothing method, and is to approximate the regression coefficient functions α_k's by a linear combination of natural cubic spline bases, and using the following penalized least squares(Wahba, 1990):

$$J(\mathbf{a}(\cdot)) = \sum_{i=1}^{n} \left\{Y_i - \sum_{j=1}^{p} X_{ij} a_j(u_i)\right\}^2 + \sum_{j=1}^{p} \lambda_j \int a_j''(u)^2 du, \tag{2.3.3}$$

based on a random sample $\{u_i, \mathbf{x}_i, Y_i\}$ from model (2.3.1).

Hastie and Tibshirani (1993) also provide Bayesian interpretation of time-varying coefficient models. When u is a time variable t which is sampled over an equally-spaced grid points, and $\{\mathbf{x}_t, Y_t\}$ were collected over the time variable t, the

VCM can be written as

$$Y_t = \mathbf{x}_t^T \boldsymbol{\alpha}_t + \varepsilon_t, \tag{2.3.4}$$

where $\boldsymbol{\alpha}_t = (a_0(t), \cdots, a_p(t))^T$ and $\mathbf{x}_t = (1, X_{t1}, \cdots, X_{tp})^T$. This can be viewed as a dynamic linear models (West et al., 1985), and Bayesian method can be used to carry out the estimation of model parameter. Specifically, one may consider

$$\begin{aligned} Y_t &= \mathbf{x}_t^T \boldsymbol{\alpha}_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, V_t), \\ \boldsymbol{\alpha}_t &= G_t \boldsymbol{\alpha}_{t-1} + \omega_t, \qquad \omega_t \sim \mathbf{N}(\mathbf{0}, \mathbf{W}_t). \end{aligned}$$

When $p = 1$, and $G_t = 1$, and the variance V_t is constant, the mean of $\alpha(\cdot)$ can be approximately approached by an exponentially weighted moving average of $Y_j/X_j, 0 \le j \le t-1$. δ/X_j^2 is the weight assigned to Y_{t-j}/X_{t-j}, where δ is *discount factor* between 0 and 1. It also follows the idea from (2.3.2). The Bayesian procedures for dynamic linear model can be directly implement for the time-varying coefficient model to obtain the estimate of the regression coefficients and their posterior confidence interval.

To avoid solving large linear systems in spline smoothing methods, kernel regression and local polynomial regression have been proposed for the VCM in Hoover et al. (1998); Fan and Zhang (1999). For given point u_0 and u in the neighborhood of u_0, the local polynomial regression is to approximate $a_j(u)$ by

$$a_j(u) \approx \sum_{l=0}^{q} \frac{1}{l!} a_j^{(l)}(u_0)(u - u_0)^l, \tag{2.3.5}$$

where $a_j^{(l)}$ is the l-th derivative of $a_j(\cdot)$. The kernel regression is corresponding to $q = 0$. The local polynomial regression is to minimize the local least-squares:

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{p} \sum_{k=0}^{q} c_{j,k}(U_i - u_0)^k X_{ij} \right\}^2 K_h(U_i - u_0), \tag{2.3.6}$$

where $K(\cdot) = h^{-1}K(\cdot/h)$ with a kernel function $K(\cdot)$ and a bandwidth $h = h_n > 0$. Here $c_{j,k}, j = 1, \ldots, p, k = 0, \ldots, q$ is the estimated coefficient of k-th derivative of $a_j(u)$ around u_0. The local least squares estimator has a close form solution, which can be expressed by using matrix notation. Denote $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{W} = \text{diag}(K_h(U_1 - u), \ldots, K_h(U_n - u))$, and

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & X_{11}(U_1 - u)^q & \cdots & x_{1p} & \cdots & x_{1p}(U_1 - u)^q \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & X_{n1}(U_n - u)^q & \cdots & x_{np} & \cdots & x_{np}(U_n - u)^q \end{bmatrix}. \tag{2.3.7}$$

The local polynomial estimator of $a_j(u)$ is

$$\widehat{a}_j(u) = e_{j,\kappa}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{y}, \tag{2.3.8}$$

where $e_{j,\kappa}$ is a vector with length j, and 1 as κ-th element, 0 in the other position, and $\kappa = p(q+1)$. When $q = 1$ in (2.3.5), the resulting estimate is referred to as the local linear estimate, which is the estimation method in this dissertation.

Fan and Zhang (1999) derived the asymptotic bias and variance for $\widehat{a}_j(u)$, and further established the asymptotic normality of the local polynomial estimates. The asymptotic normality enables us to construct pointwise conference interval for $a_j(\cdot)$ at a specific point u. In practice, it is of great interest to construct simultaneously confidence band for $a_j(\cdot)$ for u over a certain interval.

Fan and Zhang (2000) developed a simultaneously confidence band for local polynomial regression. For local polynomial estimate in (2.3.8), the error variance can be estimated by smoothing the corresponding residuals. A normalized weighted residual sum of squares is used to estimate error variance:

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{2.3.9}$$

$$\widehat{\sigma}^2(u) = \frac{1}{\mathrm{tr}(\mathbf{W} - (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X})} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 K_h(U_i - u) \tag{2.3.10}$$

In order to construct confidence bands, one needs the estimation of $\mathrm{bias}(\widehat{a}_j(u)|\mathcal{D})$ and $\mathrm{var}(\widehat{a}_j(u)|\mathcal{D})$. Naturally, based on (2.3.8), bias of $\widehat{a}_j(u)$ can be calculated as:

$$\mathrm{bias}(\widehat{a}_j(u)|\mathcal{D}) = e_{j,\kappa}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\beta} \tag{2.3.11}$$

where $\mathcal{D} = (U_1, \ldots, U_n, X_{11}, \ldots, X_{1n}, \ldots, X_{p1}, \ldots, X_{pn})^T$, and

$$\beta_i = \sum_{j=1}^{p} \{a_j(U_i) - \sum_{k=0}^{q} \frac{1}{k!} a_j^{(k)}(u)(U_i - u)^k\} X_{ij}. \tag{2.3.12}$$

In order to estimate this bias, $q+2$ order Taylor expansion is used to approach $\boldsymbol{\beta}$ in (2.3.12), and we call this approach $\boldsymbol{\tau}$.

Then, the bias of $a_j(u)$ can be estimated by

$$\widehat{\mathrm{bias}}(\widehat{a}_j(u)|\mathcal{D}) = e_{j,\kappa}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\widehat{\boldsymbol{\tau}}, \tag{2.3.13}$$

where the i-th element of vector $\widehat{\boldsymbol{\tau}}$ is:

$$\sum_{j=1}^{p} \{\frac{1}{(q+1)!}\widehat{a}_j^{(q+1)}(u)(U_i - u)^{q+1} + \frac{1}{(q+2)!}\widehat{a}_j^{(q+2)}(u)(U_i - u)^{q+2})\} X_{ij}. \tag{2.3.14}$$

Also, since the estimated error variance can be derived from (2.3.10),

$$\widehat{\mathrm{var}}(\widehat{a}_j(u)|\mathcal{D}) = e_{j,\kappa}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}e_{j,\kappa}^T\widehat{\sigma}^2(u). \tag{2.3.15}$$

Notice here, the pilot bandwidth h^* used to calculate $\widehat{\sigma}^2(u)$ is of order $n^{-1/(2q+5)}$ according to the theorem conditions. The theoretical result of the asymptotic property of $\widehat{\mathrm{bias}}(\widehat{a}_j(u)|\mathcal{D})$ is given in Fan and Zhang (2000).

Denote $\|b(u)\|_\infty = \sup_{u\in\mathcal{I}} |b(u)|$ for a function $b(u)$. Fan and Zhang (2000) further derived the asymptotic distribution of the following statistic:

$$(-2\log h)^{1/2}\left(\left\|\widehat{\text{var}}(\widehat{a}_j(u)|\mathcal{D})(\widehat{a}_j - a_0 - \widehat{\text{bias}}(\widehat{a}_j(u)|\mathcal{D}))\right\|_\infty - d_{\nu,n}\right), \qquad (2.3.16)$$

where $d_{\nu,n}$ is calculable, and the detailed definition can be found in Theorem 1 (Fan and Zhang, 2000). This enables one to construct confidence band for $a_j(\cdot)$.

Practical implementation issues for local polynomial regression under the VCM framework have been studied. Bandwidth selection for local polynomial regression was studied in Zhang and Lee (2000); Wu and Chiang (2000); Cai et al. (2000); Fan et al. (2005). It is easy to understand the importance of choosing a right bandwidth. If the bandwidth is too small, there are too little samples in each interval, which leads to severe inaccuracy of the estimation. On the other hand, if the bandwidth is too large, the approximation that approaches the coefficients in each interval by Taylor expansion might be wrong and may lead to significant bias.

Since its introduction in Hastie and Tibshirani (1993), the VCM has been a very popular in longitudinal data analysis (Hoover et al., 1998; Chiang et al., 2001; Eubank et al., 2004; Qu and Li, 2006). The VCM has been applied to various scientific research. For example, the VCM has been used for analysis of data collected in ecological studies (Yi et al., 2004; Kürüm et al., 2014) and analysis of neuroimaging data (Zhu et al., 2011) . It has become a popular nonparametric models for analysis of data collected in social and behavior researches (Tan et al., 2012; Liu et al., 2013; Lanza et al., 2013; Vasilenko et al., 2014; Shiyko et al., 2013; Trail et al., 2014; Dziak et al., 2015; Yang et al., 2015).

Cai, Fan, and Li (2000) consider the VCM under generalized linear model framework. Denote $m(\mathbf{u},\mathbf{x})$ to be the regression function of the response variable Y on the covariate U and $\mathbf{X}$. That is, $m(u,\mathbf{x}) = E(Y|U=u,\mathbf{X}=\mathbf{x})$. Generalized

VCM (GVCM) assumes that

$$\eta(\mathbf{u}, \mathbf{x}) = g(m(\mathbf{u}, \mathbf{x})) = \sum_{j=1}^{p} a_j(u)x_j, \tag{2.3.17}$$

for a known link function $g(\cdot)$, where $a_j(u)$'s are unknown regression coefficient functions defined on a bounded compact set $\mathcal{U}$.

Cai, Fan, and Li (2000) give detailed estimation and hypothesis testing procedure for GVCM based on local likelihood technique. For u in the neighbourhood of given point u_0, approximate $a_j(u)$ by a local function $a_j + b_j(u - u_0)$, and then estimate $a_j(u)$ by estimating a_j and b_j. Suppose that $\{U_i, \mathbf{x}_i, Y_i\}$ is a random sample from model (2.3.17). Local likelihood approach is to maximize the following local likelihood function:

$$l_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} l(g^{-1}[\sum_{j=1}^{p} \{a_j + b_j(U_i - u_0)\}X_{ij}, Y_i])K_h(U_i - u_0), \tag{2.3.18}$$

where $\mathbf{a} = (a_1, a_2, \cdots, a_p)^T$, $\mathbf{b} = (b_1, b_2, \cdots, b_p)^T$, $K(\cdot) = h^{-1}K(\cdot/h)$ with a kernel function $K(\cdot)$ and a bandwidth $h = h_n > 0$. In general, there is no closed form solution for the local likelihood estimate. Thus, numerical optimization algorithm such as Newton-Raphson algorithm is used to find the maximizer of the local likelihood function (2.3.18). However, one has to run hundreds of the numerical optimization algorithm in order to estimate the regression coefficient functions over a set of hundreds of grid points. This leads to great computational burden. To save computational cost, Cai, Fan, and Li (2000) proposed a one-step Newton-Raphson estimator to maximize the likelihood with only one iteration for each grid point. Specifically, denote $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1, \cdots, a_p, b_1, \cdots, b_p)^T$, and $l_n'(\boldsymbol{\beta})$, $l_n''(\boldsymbol{\beta})$ be the gradient and the Hessian matrix of $l_n(\boldsymbol{\beta})$. With a given initial estimator

$\widehat{\boldsymbol{\beta}}_0 = (\widehat{\mathbf{a}}^{\mathsf{T}}(u_0), \widehat{\mathbf{b}}^{\mathsf{T}}(u_0))^{\mathsf{T}}$, **one-step** local likelihood estimator is defined to be,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} = \widehat{\boldsymbol{\beta}}_0 - (l_n''(\widehat{\boldsymbol{\beta}}_0))^{-1} l_n'(\boldsymbol{\beta}). \tag{2.3.19}$$

The estimation of the regression coefficient functions is carried out by evaluating the one-step estimator over a set of grid points over the support of U.

In practice, there are several practical implementation issues to carry out the one-step estimator. The first one is how to select the bandwidth. The empirical bias method (Carroll et al., 1998) is used. The second issue is how to set the initial values. It is crucial to set good initial value in the one-step estimator. An initial point far away from the true value may lead to a bad local maximum likelihood estimator (MLE). In the univariate setting, a natural choice of initial value is the least squares estimate. However, in the multivariate setting, there is no easy way to construct the initial values. Cai, Fan, and Li (2000) proposed a systematic scheme to construct the initial values. The third problem is that Hessian matrix of local likelihood function $l_n''(\boldsymbol{\beta})$ may be singular. This leads to the failure of the one-step estimator. Cai, Fan, and Li (2000) advocated the use of ridge regression for solving issues related to Hessian matrix singularity. and further studied the theoretical properties of the one-step estimator.

Condition 2.3.1. *The regularity conditions are:*

C1. The function $q_2(s, y) < 0$ *for* $s \in \mathcal{R}$ *and* y *in the range of the response variable.*

C2. The functions $f_U(u)$, $\Gamma(u)$, $V(m(u, x))$, $V'(m(u, x)$, *and* $g'''(m(u, x))$ *are continuous at the point* $u = u_0$. *Further, assume that* $f_U(u_0) > 0$ *and* $\Gamma(u_0) > 0$.

C3. $K(\cdot)$ *has a bounded support.*

C4. $a_j''(\cdot)$ *is continuous in a neighbourhood of* u_0 *for* $j = 1, \ldots, p$.

C5. $E(|\mathbf{X}|^3 | U = u)$ *is continuous at the point* $u = u_0$.

C6. $E(Y^4 | U = u, \mathbf{X} = \mathbf{x})$ *is bounded in a neighborhood of* $U = U_0$.

where $q_j(s,y) = (\partial^j/\partial s^j)l\{g^{-1}(s), y\}$.

The local likelihood function (2.3.18) is concave guaranteed by Condition Cl. It is satisfied for the canonical exponential family with a canonical link. Note that condition C2 implies that $q_1(\cdot,\cdot), q_2(\cdot,\cdot), q_3(\cdot,\cdot), \rho'(\cdot,\cdot)$, and $m'(\cdot,\cdot)$ are continuous.

In order to justify the one-step estimator $\widehat{\boldsymbol{\beta}}_{OS}$, one need to prove that it has an asymptotic distribution. In theorem 2.3.1, it is shown that $\widehat{\boldsymbol{\beta}}_{MLE}$ has an asymptotic distribution. Define $\mu_k = \int u^k K(u)du$, $\nu_k = \int u^k K^2(u)du$, $\mathbf{H} = \text{diag}(1, h) \otimes \mathbf{I}_p$, and $\otimes$ as the Kronecker product. The theorem in Cai, Fan, and Li (2000) is:

Theorem 2.3.1. *Suppose the conditions are hold, and that* $h = h_n \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$. *Then*

$$\sqrt{nh}\big[\mathbf{H}\{\widehat{\boldsymbol{\beta}}_{MLE}(u_0) - \beta(u_0)\} - \frac{h^2}{2(\mu_2 - \mu_1^2)} \times \begin{pmatrix} (\mu_2^2 - \mu_1\mu_3)a''(u_0) \\ (\mu_3^2 - \mu_1\mu_2)a''(u_0) \end{pmatrix} + o_p(h^2)\big] \xrightarrow{D} \mathbf{N}(0, \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^{-1}), \tag{2.3.20}$$

where

$$\Delta = f_U(u_0) \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes \Gamma(u_0), \qquad \Lambda = f_U(u_0) \begin{pmatrix} 1 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \Gamma(u_0).$$

The detail definition of $\Gamma(\cdot)$ *can be found in the Section 2.3 of Cai, Fan, and Li (2000).*

Theorem 2.3.2. *Under the assumptions in Theorem 2.3.1 ,* $\widehat{\boldsymbol{\beta}}_{OS}$ *has the same asymptotic distribution as* $\widehat{\boldsymbol{\beta}}_{MLE}$, *provided that the initial estimator* $\widehat{\boldsymbol{\beta}}_0$ *satisfies* $\mathbf{H}(\widehat{\boldsymbol{\beta}}_0 - \beta) = O_p(h^2 + (nh)^{-\frac{1}{2}})$.

In practice, it is of great interest to test whether the varying coefficients really vary over u. In other words, it is of interest to test whether the data can be fit well

by a linear regression model. This problem is equivalent to the hypothesis testing

$$H_0 : a_1(u) = a_1, \cdots, a_p(u) = a_p. \tag{2.3.21}$$

Cai, Fan, and Li (2000) proposed a likelihood ratio type statistic for testing this hypothesis:

$$T = 2[l(H_1) - l(H_0)], \tag{2.3.22}$$

where $l(H_1)$ and $l(H_0)$ are the likelihood under H_1 and H_0, respectively. In order to conduct the test, we need to obtain the null distribution of T. For parametric models, the corresponding likelihood ratio test follows a χ^2 distribution under null hypotheses. For varying coefficient models, the degrees of freedom tends to infinite under alternative hypothesis. Cai, Fan, and Li (2000) proposed *conditional bootstrap* to construct the null distribution of test statistics T. Denote $\{\widehat{a}_j\}$ as the MLE under the null hypothesis. For any given covariates $(U_i, \mathbf{X}_i)$, bootstrap sample Y_i^* can be generated from the distribution of Y decided by linear predictor $\widehat{\eta} = \sum_{j=1}^{p} \widehat{a}_j X_{ij}$, and the test statistic T^* can be calculated the same way as T. Therefore, a distribution of T^* is defined here, which can be used to approximate the distribution of T. This becomes a valid method since the asymptotic null distribution does not rely on the values of $\{\widehat{a}_j\}$ (Fan, Zhang, and Zhang, 2001). Huang et al. (2002); Qu and Li (2006) further generalize this strategy for other statistical settings. The confidence band proposed in Fan and Zhang (2000) may serve as an alternative method for testing hypothesis

$$H_0 : a_j(u) = a_0(u) \leftrightarrow H_1 : a_j(u) \neq a_0(u). \tag{2.3.23}$$

Readers are referred to these papers for details.

2.4 Variable Selection and Feature Screening

Suppose that $\{\mathbf{x}_i, y_i\}$, $i = 1, \cdots, n$ is a random sample from a linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where ε_i is a random error with mean 0 and variance σ^2.

At the initial stage of modelling, it is common to consider a large model by including many potential useful variables into the linear regression model in order to reduce modelling bias. To enhance model predictability and interpretation, it is always desirable to exclude unimportant variables from the final model. Thus, variable selection plays an important role in statistical modelling. Variable selection has been an active research topic since 1970s. Traditional variable selection criteria include Akaike's information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz et al., 1978). To carry out variable selection with the traditional variable selection criteria, one has to conduct best subset selection by exhaustively searching over all possible sets. This becomes infeasible when the dimensional of covariate vector is large due to computational cost. Thus, traditional variable selection typically is carried out by using backward elimination, forward addition or stepwise regression. Although the select model may not optimal, it provides a parsimonious model with good interpretation and prediction performance.

Penalized least squares method has been proposed for variable selection in the literature. The penalized least squares function is defined by

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.4.1}$$

where p is the dimension of $\boldsymbol{\beta}$, and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. Traditional information criteria including the AIC and BIC can be

written as an equivalent form of the penalized least squares (2.4.1) with L_0 penalty, defined by

$$p_\lambda(|\beta|) = \lambda I\{|\beta| \neq 0\}$$

with a properly chosen λ (Fan and Li, 2006). Bridge regression, proposed by Frank and Friedman (1993), is the penalized least squares with L_q penalty

$$p_\lambda(|\beta|) = \lambda|\beta|^q$$

for $0 \leq q \leq 2$. The LASSO, proposed by Tibshirani (1996), corresponds to the penalized least squares with L_1 penalty. Fan and Li (2001) provided insights into how to select a penalty function and advocates using nonconvex penalty function. They further proposed the smoothly clipped absolute deviation (SCAD) penalty, defined by

$$p_\lambda(\theta) = \int_0^\theta \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} dt, \tag{2.4.2}$$

in which $a = 3.7$ suggested by Fan and Li (2001), and $I(\cdot)$ is the indicator function, and $(b)_+$ denotes the positive part of b. Another popular nonconvex penalty is the MCP penalty proposed by Zhang (2010), and is defined by

$$p_\lambda(\theta) = \int_0^\theta \frac{(a\lambda - t)_+}{a} dt \tag{2.4.3}$$

for $a > 2$.

2.4.1 Grouped Variable Selection

Since grouped variable selection will be used to select important variables in varying coefficient models, this section is devoted to review this statistical procedure. Partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_J)$. Yuan and Lin (2006) proposed group

LASSO by minimizing the following penalized least squares

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2, \tag{2.4.4}$$

where $\|\cdot\|_2$ is the L_2-norm of a vector. When all dimensions of $\boldsymbol{\beta}_j$'s equal one, the group LASSO becomes the LASSO. Thus, the group LASSO is a natural extension of the LASSO to select grouped variables. Grouped variable selection is particularly useful under the setting analysis of variance in the presence of multiple levels for factors.

Although the objective function of the original LASSO is convex, the objective function of group LASSO is not convex. Thus, optimization problem associated the group LASSO is much more complicated than the original LASSO. Qin, Scheinberg, and Goldfarb (2013) proposed two algorithms to conduct the group LASSO method based on different sample size. For moderate sample size, a general version of Block Coordinate Descent (BCD) algorithm is proposed. This is the algorithm used in Chapter 5 of this dissertation. The detailed procedure is introduced following. For large size, an extension of (Fast) Iterative shrinkage thresholding (ISTA/FISTA) is proposed.

Let $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ be the response vector and the design matrix, respectively. Then the penalized least squares function of the group LASSO can be written as

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{J}\left\|\boldsymbol{\beta}_j\right\|. \tag{2.4.5}$$

BCD optimizes this objective function by updating the coefficients group by group. Within each sub-iteration, it finds the coefficients for this group, while the coefficients in other groups staying fixed. Define $M_j = \mathbf{X}_j^T\mathbf{X}_j$, and $p_j = \mathbf{X}_j^T(\sum_{i\neq j}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i))$.

The objective function can be re-written for j-th sub-iteration as:

$$\min_{\boldsymbol{\beta}_j} \frac{1}{2}\boldsymbol{\beta}_j^T M_j \boldsymbol{\beta}_j + p_j^T \boldsymbol{\beta}_j + \lambda \left\|\boldsymbol{\beta}_j\right\|. \tag{2.4.6}$$

When $\mathbf{X}_j$ is orthogonal, the coefficients have close form. Qin, Scheinberg, and Goldfarb (2013) drops the condition, and give a generalized algorithm to solve the problem by using Newton method.

If $\|p_j\| \leq \lambda$, $\boldsymbol{\beta}_j = 0$ is the optimal solution. If $\boldsymbol{\beta}_j \neq 0$, under optimality conditions

$$(M_j + \frac{\lambda}{\left\|\boldsymbol{\beta}_j\right\| I_n})\boldsymbol{\beta}_j = -p_j, \tag{2.4.7}$$

there exist $\Delta > 0$, so that (2.4.6) becomes

$$\min \frac{1}{2}\boldsymbol{\beta}_j^T M_j \boldsymbol{\beta}_j + p_j^T \boldsymbol{\beta}_j, \qquad \text{s.t.} \left\|\boldsymbol{\beta}_j\right\| \leq \Delta. \tag{2.4.8}$$

Denote x_j^* as the unique solution for (2.4.8) we are looking for, then $\left\|x_j^*\right\| = \Delta$. (2.4.7) leads to

$$\boldsymbol{\beta}_j^* = -(M_j + \frac{\lambda}{\Delta} I)p_j = \Delta z_j(\Delta), \tag{2.4.9}$$

where $z_j(\Delta) = -(\Delta M_j + \lambda I)^{-1} p_j$ with norm equals to 1. Our problem further changes to find the solution of Δ which satisfying $\|z_j(\Delta)\| = 1$.

Assume γ_i and q_i are the i-th eigenvalue and eigenvector of M_j, by decomposing M_j, the following function can be derived:

$$\|z_j(\Delta)\|^2 = \sum_i \frac{(q_i^T p_j)^2}{(\gamma_i \Delta + \lambda)^2}. \tag{2.4.10}$$

Newton's method is applied to find Δ iteratively with:

$$\phi(\Delta) = 1 - \frac{1}{\|z_j(\Delta)\|}. \tag{2.4.11}$$

The algorithm (denoted as BCD-GL) is summarized as:

Algorithm 2.4.1. *Block Coordinate Descent Group LASSO algorithm:*

BCD-GL algorithm
Given $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^m$, compute $M_j = \mathbf{X}_j^T \mathbf{X}_j$, for $j = 1, \ldots, J$. For $k = k^*$, $k^* \geq 1$:
1. Let $\boldsymbol{\beta}^{(k)}$ is from last iteration, for $j = j^*$:
(a). $p_j = \mathbf{X}_j^T (\sum_{i \neq j} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i))$.
(b). If $\|p_j\| \leq \lambda$, then $\boldsymbol{\beta}_j^{(k+1)} = 0$.
(c). If not, calculate $\|z_j(\Delta)\|$ by (2.4.10).
(d). then find Δ by applying Newton's root-finding method to (2.4.11).
(e). Then $z_j(\Delta)$ can be derived from its definition. $\boldsymbol{\beta}_j^{(k+1)} = \Delta z_j(\Delta)$.
2. After exploring all $j = 1, \ldots, J$, $\boldsymbol{\beta}^{(k+1)}$ is derived.
End the iteration k until it meets the stopping criteria, and the estimate in final step is $\boldsymbol{\beta}^{opt}$.

In the article, Iterative shrinkage thresholding (ISTA) algorithm is also introduced. Basically, it separate the targeted function (2.4.5) by setting $g(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and $h(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{J} \left\|\boldsymbol{\beta}_j\right\|$. After using quadratic approximation on $g(\boldsymbol{\beta})$, the target function can be transferred to an additive form. Then, the target function can be optimized by each coordinate. Qin, Scheinberg, and Goldfarb (2013) improve the algorithm by using block coordinate step when optimizing each coordinate (ISTA-BC). A hybrid version of BCD-GL and ISTA-BC is proposed, and global convergence of this hybrid version is proved.

2.4.2 Feature Screening

When the dimension p is greater than the sample size n, the least squares estimator of $\boldsymbol{\beta}$ is not well defined due to the singularity of $\mathbf{X}^T\mathbf{X}$. For a data, if $\log(p) = O(n^k), k > 0$, then we call the data ultrahigh dimensional data. This type of data widely appears within the field of genomic, economics, etc. In these area, the predictors may have influence on response could be millions while the number of subjects could be as few as tens or hundreds. In this case, traditional variable selection method may lead to a unacceptable computing burden, due to

the huge p. For one of the solution, Fan and Lv (2008) proposes a new idea based one the ranking of Pearson correlation.

The ridge regression is a useful technique to deal with singularity of the design matrix $\mathbf{X}$ and is defined by

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathrm{I}_p)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

where λ is a ridge parameter. The ridge regression estimator $\widehat{\boldsymbol{\beta}}_\lambda$ tends to the least squares estimator if it is well-defined when $\lambda \to 0$. On the other hand, $\lambda\widehat{\boldsymbol{\beta}}_\lambda$ tends to $\mathbf{X}^\mathsf{T}\mathbf{y}$ if $\lambda \to \infty$. This implies that $\widehat{\boldsymbol{\beta}}_\lambda \propto \mathbf{X}^\mathsf{T}\mathbf{y}$. In practice, all covariates and the response are marginally standardized so that their means and variances equals 0 and 1, respectively. Then $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{y}$ becomes the vector consists of the sample version of Pearson correlations between the response and individual covariate. This is the motivation of using Pearson correlation as a marginal utility for feature screening. Specifically, denote

$$\omega_j = \frac{1}{n}\mathbf{X}_j^\mathsf{T}\mathbf{y}, \quad \text{for } j = 1, 2, \ldots, p. \tag{2.4.12}$$

Here it is assumed that both $\mathbf{X}_j$ and $\mathbf{y}$ are marginally standardized. Thus, ω_j indeed is the sample correlation between the $j-$th predictor and the response variable.

Fan and Lv (2008) suggested ranking all predictors according to $|\omega_j|$ and select the top predictors which are relatively strongly correlated with the response. To be specific, for any given $\gamma \in (0, 1)$, the $[\gamma n]$ top ranked predictors are selected to obtain the submodel

$$\widehat{\mathcal{M}}_\gamma = \{1 \le j \le p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.4.13}$$

where $[\gamma n]$ denotes the integer part of γn. It reduces the ultrahigh dimensionality

down to a relatively moderate scale $[\gamma n]$, i.e. the size of $\widehat{\mathcal{M}}_\gamma$, and then the well-established penalized variable selection methods is applied to the submodel $\widehat{\mathcal{M}}_\gamma$. This screening procedure defined in (2.4.13) is referred to as sure independence screening (SIS) in the literature.

Several methods are developed since then for all kinds of different models. Zhu et al. (2011) develops a novel feature screening procedure, and it can be used for a lot of parametric and semiparametric models. Li, Zhong, and Zhu (2012) proposes independence screening procedure based on the distance correlation (DC-SIS). It has no model specification, and can be directly used for multivariate response variables. Also, Li, Peng, Zhang, and Zhu (2012) proposes a robust rank correlation based screening. This method can be used to deal with a data with large quantitative of outliers and influence points.

Also, the idea of SIS has been extended for varying coefficient model with ultrahigh dimensional covariates:

$$y = \beta_0(u) + \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon, \tag{2.4.14}$$

where $E(\varepsilon|\mathbf{x}, \mathbf{u}) = 0$, $\beta_0(u)$, $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_p(u))^T$ are nonparametric smooth functions over u. Fixing u, the varying coefficient model is a linear regression model.

Liu, Li, and Wu (2014) proposed a conditional correlation sure independence screening (CC-SIS) to measure the importance of predictor variable. Specifically, define

$$\rho(x_j, y|u) = \frac{\mathrm{cov}(x_j, y|u)}{\sqrt{\mathrm{cov}(x_j, x_j|u)\mathrm{cov}(y, y|u)}}, \tag{2.4.15}$$

which is a function of u. To avoid directly comparing two functions, Liu, Li and

Wu (2014) define the following marginal utility

$$\rho_{j0} = \mathrm{E}\,[\rho^2(x_j, y|u)]. \tag{2.4.16}$$

In order to estimate $\mathrm{cov}(x_j, y|u), \mathrm{cov}(x_j, x_j|u)$, and $\mathrm{cov}(y, y|u)$ in (2.4.15), we need to estimate $\mathrm{E}\,(y|u)$, $\mathrm{E}\,(y^2|u)$, $\mathrm{E}\,(x_j|u)$, $\mathrm{E}\,(x_j^2|u)$, and $\mathrm{E}\,(x_j y|u)$. Kernel smoothing method is used to estimate the five conditional expectation. Take $\mathrm{E}\,(y|u)$ as an example:

$$\widehat{\mathrm{E}}\,(y|u) = \sum_{i=1}^{n} \frac{K_h(u_i - u)y_i}{\sum_{i=1}^{n} K_h(u_i - u)}, \tag{2.4.17}$$

where $K(\cdot)$ is a kernel function, and $K_h(t) = h^{-1}K(t/h)$. Follow the idea, $\widehat{\mathrm{E}}\,(y|u)$, $\widehat{\mathrm{E}}\,(y^2|u)$, $\widehat{\mathrm{E}}\,(x_j|u)$, $\widehat{\mathrm{E}}\,(x_j^2|u)$, and $\widehat{\mathrm{E}}\,(x_j y|u)$ can be calculated. One thing should be noticed is that the bandwidth h should be the same for all the five conditional means in order to guarantee $|\widehat{\rho}(x_j, y|u)| \leq 1$.

Based on the definition of covariance, $\mathrm{cov}(x_j, y|u)$, $\mathrm{cov}(x_j, x_j|u)$, and $\mathrm{cov}(y, y|u)$ are easily estimated from the five conditional means. Furthermore, $\rho(x_j, y|u)$ can be estimated by plugging-in the estimate of the conditional mean. Thus, ρ_{j0} can be estimated by

$$\widehat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2(x_j, y|u_i). \tag{2.4.18}$$

Intuitively, the larger $\widehat{\rho}_j^*$ is, the more important x_j is. Thus, for a given d, the screening procedure is to select the d variables with largest $\widehat{\rho}_j^*$. That is,

$$\widehat{\mathcal{M}} = \{j : \widehat{\rho}_j^* \geq \widehat{\rho}_{(d)}, 1 \leq j \leq p\}, \tag{2.4.19}$$

where $\widehat{\rho}_{(1)} \geq \widehat{\rho}_{(2)} \geq \cdots \geq \widehat{\rho}_{(p)}$. Extending the strategy in Fan and Lv (2008) to varying coefficient models, the authors suggested setting $d = [n^{4/5}/\log(n^{4/5})]$, where $[a]$ stands for the integer part of a. By feature screening, we may reduce the dimensionality from p to d.

Define the true model index as:

$$\mathcal{M}_* = \{j : 1 \le j \le p, \beta_j(u) \neq 0 \text{ for some } u \in \mathcal{U}\}. \tag{2.4.20}$$

Liu, Li, and Wu (2014) further studied the theoretical properties of their proposed feature screening procedure. First, all the regularity conditions are listed here for clear understanding.

Condition 2.4.1. *The regularity conditions for CC-SIS:*
(B1): The following inequality holds uniformly in n*:*

$$\min_{j \in \mathcal{M}_*} \rho^*_{j0} > E\left\{ \frac{\lambda_{\max}\{cov(\mathbf{x}_{\mathcal{M}_*}, \mathbf{x}^T_{\mathcal{M}_*^c}|u)\, cov(\mathbf{x}_{\mathcal{M}_*^c}, \mathbf{x}^T_{\mathcal{M}_*}|u)\} \times \lambda_{\max}\{\boldsymbol{\rho}_{\mathcal{M}_*}(u)\boldsymbol{\rho}^T_{\mathcal{M}_*}(u)\}}{\lambda^2_{\min}\{cov(\mathbf{x}_{\mathcal{M}_*}|u)\}} \right\}. \tag{2.4.21}$$

(B2): Assume that conditioning on $\mathbf{x}^T_{\mathcal{M}_*}\boldsymbol{\beta}_{\mathcal{M}_*}(u)$ *and* $u, \mathbf{x}$ *and* ε *are independent. Further assume that the following linearity condition is valid:*

$$E\{\mathbf{x}|\mathbf{x}^T_{\mathcal{M}_*}\boldsymbol{\beta}_{\mathcal{M}_*}(u), u\} = cov(\mathbf{x}, \mathbf{x}^T_{\mathcal{M}_*}|u)\boldsymbol{\beta}_{\mathcal{M}_*}(u)\{cov(\mathbf{x}^T_{\mathcal{M}_*}\boldsymbol{\beta}_{\mathcal{M}_*}(u)|u)\}^{-1} \times \boldsymbol{\beta}^T_{\mathcal{M}_*}(u)\mathbf{x}_{\mathcal{M}_*}. \tag{2.4.22}$$

(C1) Denote the density function of u *by* $f(u)$*. Assume that* $f(u)$ *has continuous second-order derivative on* $\mathbb{U}$*.*
(C2) The kernel $K(\cdot)$ *is a symmetric density function with finite support and is bounded uniformly over its support.*
(C3) The random variables x_j *and* y *satisfy the sub exponential tail probability uniformly in* p*. That is , there exists* $x_0 > 0$*, such that for* $0 \le s \le S_0$*,*

$$\sup_{u \in \mathbb{U}} \max_{1 \le j \le p} E\{\exp(sx_j^2|u)\} < \infty, \tag{2.4.23}$$

$$\sup_{u \in \mathbb{U}} E\{\exp(sy^2|u)\} < \infty, \tag{2.4.24}$$

$$\sup_{u \in \mathbb{U}} \max_{1 \le j \le p} E\{\exp(sx_j y|u)\} < \infty. \tag{2.4.25}$$

(C4) All conditional means $E(y|u)$, $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$, *and* $E(x_j y|u)$, *their first and second order derivatives are finite uniformly in* $u \in \mathbb{U}$. *Further assume that*

$$\inf_{u\in\mathbb{U}} \min_{1\le j\le p} var(x_j|u) > 0, \tag{2.4.26}$$

$$\inf_{u\in\mathbb{U}} var(y|u) > 0. \tag{2.4.27}$$

Theorem 2.4.1. *Under Condition(B1) and (B2),*

$$\liminf_{n\to\infty} \Big\{ \min_{j\in\mathcal{M}_*} \rho_j^* - \max_{j\in\mathcal{M}_*^c} \rho_j^* \Big\} > 0. \tag{2.4.28}$$

This theorem suggests a clear separation between ρ_j^* for important and unimportant variables. This theorem enables the authors to further establish the ranking consistency property of the proposed procedure.

Theorem 2.4.2. *Under more conditions (B1),(B2), (C1)-(C4), suppose that bandwidth* $h \to 0$ *but* $nh^3 \to \infty$ *as* $n \to \infty$. *Then for* $p = o\{\exp(an)\}$ *with some* $a > 0$, *we have*

$$\liminf_{n\to\infty} \Big\{ \min_{j\in\mathcal{M}_*} \widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}_j^* \Big\} > 0 \text{ in probability.} \tag{2.4.29}$$

Also, the sure screening property is described as:

Theorem 2.4.3. *Under conditions(C1)-(C4) , suppose the bandwidth* $h = O(n^{-\gamma})$, *where* $0 < \gamma < 1/3$, *then we have*

$$P\Big(\max_{1\le j\le p} \big|\widehat{\rho}_j^* - \rho_{j0}^*\big| > c_3 \cdot n^{-\kappa} \Big) \le O\big\{ np \exp(-n^{1/3-\kappa}/\xi) \big\}. \tag{2.4.30}$$

And if we further assume that there exist some $c_3 > 0$ *and* $0 \le \kappa < \gamma$, *such that*

$$\min_{j\in\mathcal{M}_*} \rho_{j0}^* \ge 2c_3 n^{-\kappa}, \tag{2.4.31}$$

then

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}\right) \geq 1 - O\{ns_n \exp(-n^{1/3-\kappa}/\xi)\}, \tag{2.4.32}$$

where ξ *is some positive constant determined by* c_3*, and* s_n *is the cardinality of* $\mathcal{M}_*$*, which is sparse and may vary with* n*.*

This theorem shows that the selected variables contain the important variable with not too weak signal.

Besides CC-SIS method mentioned above, Fan, Ma, and Dai (2014) also propose a nonparametric independence screening(NIS) to select variables for ultrahigh dimensional sparse varying coefficient models. It is a follow-up work of Fan, Feng, and Song (2011). They use an iterative-NIS(INIS) approach to avoid the false negative and false positive.

2.5 Error Variance Estimation in Ultrahigh Dimensional Linear Regression Models

When facing a large quantity of unimportant variables, there might exist unimportant variables with high correlation with realized noises. Due to this spurious correlation, error variance estimation is challenging in the presence of ultrahigh dimensional covariates. The sparsity principle is a typical assumption in the analysis ultrahigh dimensional data. Thus, a common strategy to analyze ultrahigh dimensional data is two stage procedure: first conduct feature screening, and then follow a variable selection procedure to further clean up unimportant variable. In the presence of spurious correlation, spurious correlated variables will be ranked highly in the screening stage, and will be further selected by variable selection stage. The selected spurious correlated variables will explain the variation of the random error. This leads the mean squared errors significantly underestimate

the error variance. This phenomenon has confirmed by Fan, Guo, and Hao (2012).

Fan, Guo, and Hao (2012) gave the evidence of the influence of spurious correlation. Take null linear regression $\mathbf{y} = \boldsymbol{\varepsilon}$ as an example.

$$\widehat{\sigma}_n^2 = (1 - \widehat{\gamma}_n^2)\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{2.5.1}$$

and γ_n is the sample correlation of the spurious variables and the response, or noise in this example. Most variable selection methods tend to select variables with high correlation with response. Assume $\gamma_n = \max_{1\le j\le p} |\widehat{\mathrm{corr}}_n(X_j, Y)|$. Since all variables are not important, this γ_n leads to bias to the $\widehat{\sigma}_n^2$. And this γ_n could be very large considering the large p, thus $\widehat{\sigma}_n^2$ could be significantly underestimate. A simulation study gives further view of this problem.

Still the null linear regression model, take $n = 100$, and $p = 100$, or 10000. Simulate data with $\mathbf{x}_j \overset{\text{i.i.d.}}{\sim} N(0,1), j = 1,\ldots,p$, $\mathbf{y} \overset{\text{i.i.d.}}{\sim} N(0,1)$. γ_n and $\widehat{\sigma}_n^2$ can be computed from the data. Repeat for 500 random generated sample, and the figures shows the density of γ_n and $\widehat{\sigma}_n^2$. It is easy to see from the figure, as p goes larger, γ_n becomes larger which leads to a larger biased $\widehat{\sigma}_n^2$.

Let $\mathbf{y} = (Y_1,\ldots,Y_n)^T$ be the response, and $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ be the predictors, consider the linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.5.2}$$

where $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_p)^T$ be the coefficient, and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$ consists of the identical independent distributed random errors.

Denote $M_0 = \{j : \beta_j \neq 0\}$, and $s = |M_0|$, the cardinality of M_0. Further assume s is fixed or diverging at a mild rate. Denote M as a sub-index of $\{1,2,\ldots,p\}$. Then $\mathbf{X}_M$ is the predictors with respect to the index M, β_M are the coefficients with respect to $\mathbf{X}_M$, and $\mathbf{P}_M = \mathbf{X}_M(\mathbf{X}_M^T\mathbf{X}_M)^{-1}\mathbf{X}_M^T$.

The naive way to estimate error variance is to screen and select variables first

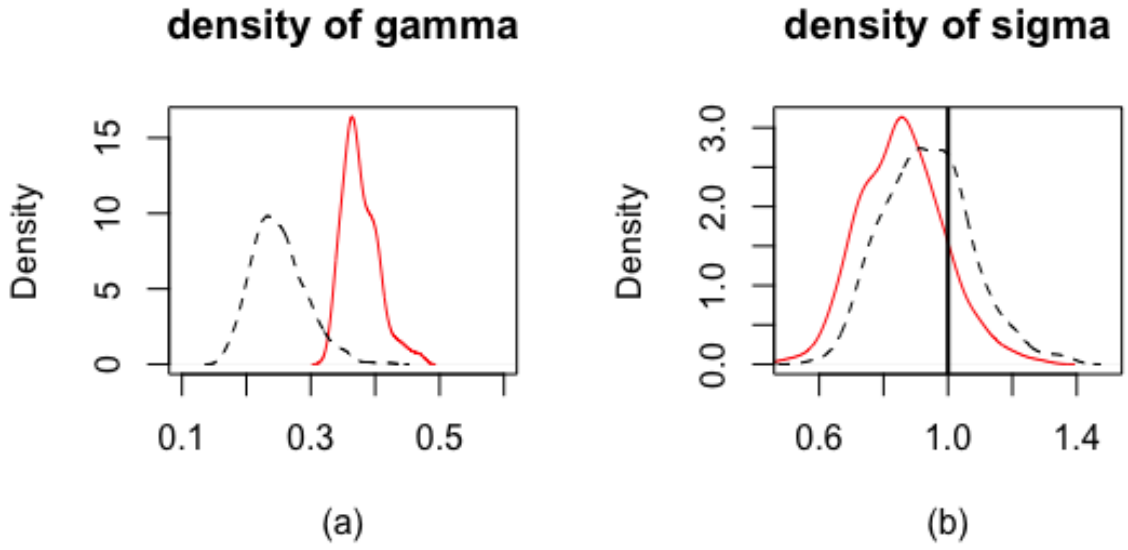


Figure 2.1: Spurious correlation.

(a) is the density of γ_n. Dash line refers to $p = 100$, and solid red line refers to $p = 10,000$. (b) is the density of $\widehat{\sigma}$. Dash line refers to $p = 100$, solid red line refers to $p = 10,000$, and the bold line is the true error variance. Adapted from Fan, Guo, and Hao (2012).

and use the mean squares error estimator (MSE) to estimate σ^2. To be more specific, define $\widehat{M}$ as the index of selected variables. Then,

$$\widehat{\sigma}^2_{\widehat{M}} = \frac{\mathbf{y}^T(\mathbf{I}_n - P_{\widehat{M}})\mathbf{y}}{n - \widehat{s}} = \frac{\varepsilon^T(\mathbf{I}_n - P_{\widehat{M}})\varepsilon}{n - \widehat{s}}. \tag{2.5.3}$$

Rewrite $\widehat{\sigma}^2_{\widehat{M}}$ as

$$\widehat{\sigma}^2_n = \frac{1}{n - \widehat{s}}(1 - \widehat{\gamma}^2_n)\varepsilon^T\varepsilon, \tag{2.5.4}$$

where $\widehat{\gamma}^2_n = (\varepsilon^T\mathbf{P}_{\widehat{M}}\varepsilon)/(\varepsilon^T\varepsilon)$.

For ease of theoretical analysis, assume that

$$P(\widehat{M} \supset M_0) \to 1, \qquad \text{as } n \to \infty. \tag{2.5.5}$$

This implies that all important variables have been retained in the selected model. The following assumptions have been imposed in Fan, Guo, and Hao (2012) to facilitate the asymptotic analysis of the naive estimator.

Define

$$\phi_{\min}(m) = \max_{M:|M|\le m} \{\lambda_{\min}(\frac{1}{n}\mathbf{X}_M^T\mathbf{X}_M)\},$$

and

$$\phi_{\max}(m) = \max_{M:|M|\le m} \{\lambda_{\max}(\frac{1}{n}\mathbf{X}_M^T\mathbf{X}_M)\}.$$

where $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of a matrix $\mathbf{A}$ respectively.

Condition 2.5.1. *The regularity conditions are following:*

Assumption 1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are IID with zero mean and finite variance σ^2 and independent of the design matrix $\mathbf{X}$.

Assumption 2. There is a constant $\lambda_0 > 0$ and b_n such that $b_n/n \to 0$ such that $P\{\phi_{\min}(b_n) \ge \lambda_0\} = 1$ for all n.

Assumption 3. There is a constant L such that $\max_{i,j} |X_{i,j}| \le L$, where X_{ij} is the (i,j) element of the design matrix $\mathbf{X}$.

Assumption 4. $E[\exp(|\varepsilon_1|/a)] \le b$ for some finite constant $a, b > 0$.

These assumptions may not be the weakest ones. The aim of Assumption 3,4 is to guarantee that $\widehat{\gamma}_n$ in theorem is of the order $O[\sqrt{\hat{s}\log(p)/n}]$. Some of the alternative conditions are:

Assumption 5. The random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. and there is a constant α such that $E[\exp\{(|X_{ij}|/\rho)^\alpha\}] \le L$ for all i and j and some constant $\alpha > 1$, and $\rho, L > 0$, where X_{ij} is the (i,j)th element of $\mathbf{X}$.

Assumption 6. ε_1 satisfies the condition that $E[\exp\{(|\varepsilon_1|/a)^\theta\}] \le b$ for some finite positive constants $a, b, \theta > 0$ and $1/\alpha + 1/\theta \le 1$, where α is defined by assumption 5.

Theorem 2.5.1. *Under Assumptions 1-4 or 1,2,5,6, the following results:*

(a) It the procedure satisfies the sure screening property with $\widehat{s} \leq b_n$ *and* $b_n = o(n)$*, then*

$$\sigma^2/(1-\widehat{\gamma}_n^2) \xrightarrow{p} \sigma^2, \qquad as\ n \to \infty \tag{2.5.6}$$

$$\sqrt{n}(\sigma^2/(1-\widehat{\gamma}_n^2) - \sigma^2) \xrightarrow{\mathcal{D}} N(0, E[\varepsilon_1^4] - \sigma^2) \qquad as\ n \to \infty \tag{2.5.7}$$

(b) In addition, $\log(p)/n = O(1)$*, then* $\widehat{\gamma}_n = O[\sqrt{\widehat{s}\log(p)/n}]$.

This means the consistency of the naive estimator depends on $\widehat{\gamma}$. It is a fraction of bias of $\widehat{\sigma}_n^2$. More specifically, if $\widehat{s}\log(p)/\sqrt{n} \to \infty$, the estimator is no longer root n consistent. This affirms the challenge of error variance estimation.

In order to provide a more reliable estimator, Fan, Guo, and Hao (2012) proposed a novel error variance estimator by refitted cross-validation (RCV) method. The RCV procedures can be described as follows:

- Split data evenly and get $(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})$, $(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})$.
- Select a model separately for each of two datasets. The selected indices for two datasets are $M^{(1)}, M^{(2)}$.
- Estimate error variance same as (2.5.3) with the variables selected from the other data set:

$$\widehat{\sigma}_i^2 = \frac{\mathbf{y}^{(i)T}(\mathbf{I}_{n/2} - P_{\widehat{M}_{2-i}}^{(i)})\mathbf{y}^{(i)}}{n - |\widehat{M}_{2-i}|}, \qquad i = 1, 2. \tag{2.5.8}$$

  Then, the final estimator is

$$\widehat{\sigma}_{RCV}^2 = (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)/2. \tag{2.5.9}$$

Fan, Guo, and Hao (2012) also established the asymptotic normality of the RCV estimator.

Theorem 2.5.2. *Assume that regularity conditions 1 and 2 hold and* $E[\varepsilon^4] \leq \infty$.

If a procedure satisfies the sure screening property with $\widehat{s}_1 \leq b_n$ *and* $\widehat{s}_2 \leq b_n$*, then*

$$\sqrt{n}(\widehat{\sigma}^2_{RCV} - \sigma^2) \xrightarrow{\mathcal{D}} N(0, E[\varepsilon_1^4] - \sigma^2), \qquad as\ n \to \infty. \tag{2.5.10}$$

Denote the oracle estimator $\widehat{\sigma}^2_O = n^{-1}\sum_{i=1}^n (Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$, where $\boldsymbol{\beta}$ is the true value of the regression coefficients. This theorem indicates that there is no $\widehat{\gamma}_n$ involving in the asymptomatic distribution, and the RCV estimator shares the same asymptotic variance of the oracle estimator. In other words, the theorem shows that RCV estimator has oracle property. Thus, it is expected that the RCV estimator has similar good performance as oracle estimator. This was confirmed by simulation studies in Fan, Guo, and Hao (2012).

Also, there are other methods to estimate the error variance or the noise level. Sun and Zhang (2012) propose an algorithm for scaled sparse linear regression to estimates the regression coefficients and error variance. The penalized loss function is

$$L_\lambda(\beta) = \frac{|y - X\beta|_2^2}{2n} + \lambda^2 \sum_{j=1}^p \rho(|\beta|/\lambda), \tag{2.5.11}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is coefficient, and $\rho(\cdot)$ is a penalty function. $\rho(t)$ is standardized to $\rho'(0+) = 1$. In order to find the optimization point of (2.5.11), $\widehat{\beta}$ need to satisfy:

$$\begin{cases} x_j'(y - X\widehat{\beta})/n = \lambda \mathrm{sgn}(\widehat{\beta}_j)\rho'(\left|\widehat{\beta}\right|/\lambda), \widehat{\beta} \neq 0, \\ x_j'(y - X\widehat{\beta})/n \in \lambda[0,1], \widehat{\beta} = 0. \end{cases} \tag{2.5.12}$$

Also, a proper λ is required. Sun and Zhang (2012) propose an iterative algorithm which can simultaneously update $\widehat{\sigma}, \lambda, \widehat{\beta}$:

$$\widehat{\sigma} \leftarrow \left|Y - X\widehat{\beta}^{old}\right|_2 / ((1-a)n)^{1/2}, \tag{2.5.13}$$

$$\lambda \leftarrow \widehat{\sigma}\lambda_0, \tag{2.5.14}$$

$$\widehat{\beta} \leftarrow \widehat{\beta}^{new}, L_\lambda(\widehat{\beta}_{new}) \leq L\hat{I}\dot{z}(\widehat{\beta}_{old}). \tag{2.5.15}$$

Set constant $(a, \lambda_0) = (p/n, 0)$, and initial value $\widehat{\beta}^{lse}$ as least squares estimator of β. So, the algorithm procedure is following. For (2.5.12), starting from $\widehat{\beta}(\lambda) = 0$, and $\lambda = |X'y/n|_\infty$, the solution paths of $\widehat{\beta}(\lambda)$ can be found for fixed λ. Then, use the algorithm in (2.5.13), and $\widehat{\beta}_{new}$ is updated by each step of solution paths of $\widehat{\beta}(\lambda)$. By this algorithm, error variance can be estimated.

Chapter 3
Incorporating Additional Data

In this chapter we first introduce the method to incorporate incidence assay data to the existing EPP model in Section 3.1. We further propose a sequential-IMIS to improve the computing efficiency for all the situations that incorporating additional data into a known model in Section 3.2. And a simulation study on EPP model is conducted in Section 3.3 to justify our method.

3.1 Incorporating Incidence Assays within EPP Framework

The Estimation and Projection Package(EPP) is a widely used software for estimating the trends of HIV epidemics. The use of prevalence data has been described in 1.1.1. As the HIV incidence data becomes available, we seek approaches for including this information in the model fitting precess, so that we can have more accurate estimates, especially for HIV incidence. The statistical model that simultaneously fit the prevalence data and incidence data is described inside Bao, Ye, and Hallett (2014). EPP model is consisted of three parts: epidemiological dynamic models(2.1.1), an infection rate model (2.1.3 and 2.1.2), and data models (2.1.4 and 2.1.6). As described in section 2.1.4, we have both prevalence data,

antenatal clinical (ANC) data and the national population-based survey (NPBS) data, and the newly available incidence assay data.

To incorporate the new incidence assay data, we propose combining the likelihood of the incidence assay data with the likelihood of other data, in a manner that is consistent with the biomarker-based incidence estimator using incidence assay data (Bao, Ye, and Hallett, 2014). Also, after the combining, new data likelihood is generated, and it allows EPP to produce new prevalence and incidence estimates. Our proposed method enables us to study the impact of assay data. We point out that by adding into incidence assay data, incidence rate has more accurate estimates, which suggests that incidence assay data provides additional information. However, one thing need to notice is that, due to the small sample size, the improvement could be too small to make a difference. And also, the simulation study only shows the results with one-time assay data.

Based on Bao, Ye, and Hallett (2014), we need to further analyze the impact of the assays for multiple scenarios or with time-series assay data. However, hundreds of simulation scenarios will be taken into consideration due to the complexity of the incidence assays. The computing burden is unacceptable if we still use the method proposed in Bao, Ye, and Hallett (2014). A more efficient computing method becomes necessary. However, it does not mean the original method is useless. It is still accurate method, which we treat as the benchmark.

Also, the current convergence rule stops the algorithm after the expected number of unique points is larger than $(1 - e^{-1}) \times B$, where B is the re-sampling sample size in IMIS. When the distribution has nonlinear ridge and multi-modal, it may cause a false converge, and leads to neglect of other modes. We propose adding another convergence criteria to avoid false converge.

3.2 Sequential IMIS

In some circumstances, the epidemics trends have been estimated by only using prevalence data without considering incidence data, because those two data sources might not be available to modellers at the same time. Instead of re-fitting the EPP package to fit the prevalence and incidence data simultaneously, we suggest taking the use of existing inference based on prevalence data, and gradually evolve it to approximate the inference based on both prevalence data and incidence data. We also expand this idea to the more general models beyond EPP.

3.2.1 Framework

Suppose that we have the old data source $\mathbf{x_A}$ and new data source $\mathbf{x_B}$, which are independent given $\boldsymbol{\theta}$. In the EPP application, prevalence data is $\mathbf{x_A}$, and incidence data is $\mathbf{x_B}$. They both can be used to estimate the same parameters $\boldsymbol{\theta}$. We have an original prior distribution $\mathrm{p}(\boldsymbol{\theta})$, and for each data sources, we have sampling distribution $\mathrm{f}(\mathbf{x_A}|\boldsymbol{\theta})$, and $\mathrm{f}(\mathbf{x_B}|\boldsymbol{\theta})$. Also, we have the posterior distribution without new data source $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A})$. Our goal is to find the posterior distribution $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$.

$$\begin{aligned}
\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B}) &= \frac{\mathrm{f}(\mathbf{x_A}, \mathbf{x_B}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{f}(\mathbf{x_A}, \mathbf{x_B}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&= \frac{\mathrm{f}(\mathbf{x_A}|\boldsymbol{\theta})\mathrm{f}(\mathbf{x_B}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{f}(\mathbf{x_A}, \mathbf{x_B}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&= \mathrm{f}(\mathbf{x_B}|\boldsymbol{\theta}) \times \frac{\mathrm{f}(\mathbf{x_A}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{f}(\mathbf{x_A}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \times \frac{\int \mathrm{f}(\mathbf{x_A}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int \mathrm{f}(\mathbf{x_A}, \mathbf{x_B}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\
&\propto \mathrm{f}(\mathbf{x_B}|\boldsymbol{\theta}) \times \mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A})
\end{aligned} \tag{3.2.1}$$

Intuitively, $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$ would be more similar to $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A})$, the posterior of partial data, than to $\mathrm{p}(\boldsymbol{\theta})$, the prior. From 3.2.1, we find out that the posterior distribution $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$ is proportion to the old result $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A})$ multiplied by new data density

$f(\mathbf{x_B}|\boldsymbol{\theta})$. It shows that instead of starting with prior distributions $p(\boldsymbol{\theta})$, we can use $f(\boldsymbol{\theta}|\mathbf{x_A})$ as initial sampling distribution. The proposed method significantly reduces computing time, especially when likelihood of old data $L(\boldsymbol{\theta}|\mathbf{x_A}) = f(\mathbf{x_A}|\boldsymbol{\theta})$ is hard to calculate.

3.2.2 Algorithm

Since we already used IMIS to draw the posterior distribution $f(\boldsymbol{\theta}|\mathbf{x_A})$, the following variables have been obtained in IMIS. We inherit the notation from Algorithm 2.2.2.

- N_0 is initial sample size,
- B is sample size drawn in each iteration,
- K is the total iterations,
- $N_K = N_0 + BK$ is the total sample size,
- $p(\cdot)$ is the prior function of $\boldsymbol{\theta}$,
- L_{old} is the likelihood function,
- $\theta^{(k)}$ is the center, and $\Sigma^{(k)}$ is weighted covariance in each iteration k,
- H_k is multivariate Gaussian distribution in iteration step k,
- $q^{(k)}$ is the mixture sampling distribution in each step k,

$$q^{(k)} = \frac{N_0}{N_k} p + \frac{B}{N_k} \sum_{s=1}^{k} H_s, \tag{3.2.2}$$

- $\boldsymbol{\theta}_i, \ldots \boldsymbol{\theta}_{N_K}$ are the samples, with likelihood $L_1, \ldots L_{N_K}$,
- $w_1, \ldots w_{N_K}$ are the weights,

$$w_i = cL_{old}(\boldsymbol{\theta}_i) \times \frac{p(\boldsymbol{\theta}_i)}{q^{(K)}(\boldsymbol{\theta}_i)}, \quad i = 1, \ldots N_K. \tag{3.2.3}$$

The algorithm details are provided below:

Algorithm 3.2.1. *(**S-IMIS**)*

1. Updating step:

The updated likelihood function becomes: $L_{old} \times L_{new}$, *where* L_{new} *is the likelihood of new data source. For each sample* $\boldsymbol{\theta}_i, \ldots \boldsymbol{\theta}_{N_K}$, *the new weights of these samples become:*

$$w_i^{(update)} \propto L_{old}(\boldsymbol{\theta}_i) \times L_{new}(\boldsymbol{\theta}_i) \times \frac{p(\boldsymbol{\theta}_i)}{q^{(m)}(\boldsymbol{\theta}_i)} \propto L_{new}(\boldsymbol{\theta}_i) \times w_i, i = 1, \ldots, N_K$$

2. New step:

Then, similar as IMIS, algorithm still goes on as the Importance Sampling Stage in Algorithm 2.2.2 from iteration $m = (k+1)$. *We still record multivariate Gaussian distribution* H_m *along with its mean and weighted covariance in each iteration step* m. $q^{(m)}$ *is calculated the same as 3.2.2. The new weights become:*

$$w_i^* \propto L_{new}(\boldsymbol{\theta}_i) \times \frac{p(\boldsymbol{\theta}_i)}{q^{(m)}(\boldsymbol{\theta}_i)}, \quad i = 1, \ldots, N_K$$

3. Resample step:

Resample J *inputs with replacement from* $\theta_1, \ldots, \theta_{N_M}$ *with weights* $w_1, \ldots, w_{N_M}$, *where* M *is the number of iterations at the importance sampling stage.*

3.2.3 Stopping Rule

The original stopping rule for IMIS is based on the expect number of unique points. One major application of sequential IMIS is to incorporate different data sources. One drawback of sequential IMIS is that it starts from a more concentrated distribution $f(\boldsymbol{\theta}|\mathbf{x_A})$ which may miss some high density regions of the target distribution $f(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$ if those two distributions differ a lot. To avoid the false convergence — the algorithm provides sufficient unique points but does not cover all high density

regions, we revise the stopping criterion by further requiring the marginal likelihood has been stable. To be specific, we require $|\mathtt{MarLike_k}/\mathtt{MarLike_{k-1}} - 1| < 0.001$ stands for consecutive 3 steps, where $\mathtt{k}$ is the step indicator.

3.2.4 Sequential IMIS in EPP

Under EPP framework, IMIS starts with drawing parameters from the prior distribution, and its proposal distribution are a mixture of multivariate normal distributions and the prior distribution, In EPP model, the incidence and prevalence are independent given the 8 r-trend parameters, $\boldsymbol{\theta}$. Therefore we can read the posterior $\mathtt{f}(\boldsymbol{\theta}|\mathbf{x_A})$ as the prior, and $\mathbf{x_B}$ as the only data, where $\mathbf{x_A}$ is the prevalence data and $\mathbf{x_B}$ is the incidence data. The resulting posterior would be $\mathtt{f}(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$, the target distribution of interest.

Then starting from the posterior $\mathtt{f}(\boldsymbol{\theta}|\mathbf{x_A})$, our focus is on different scenarios of incidence assays data $\mathbf{x_B}$. Simulation study allows us to change the parameters which have influence on incidence assays data based on our research requirements. This is a more efficient and flexible way to analysis the impact of incidence assays, compared to real data. Theoretically, a big difference between the estimation from $\mathbf{x_A}$ and $\mathbf{x_B}$ would cause inefficient or even wrong estimates from the algorithm. In this case, the posterior $\mathtt{f}(\boldsymbol{\theta}|\mathbf{x_A})$ might be a worse prior than the flat normal or uniform distributions. However, in practice, since assays data is conducted under survey data framework, they are highly unlikely to have that much inconsistency. We notice that, when the difference between estimation from two data sets $\mathbf{x_A}$ and $\mathbf{x_B}$ is large, using old posterior as prior might not a be good choice. Here, we also propose that once the point estimates based on the new data do not fall in 95% credible interval of the old posterior distribution, the sequential IMIS should not be used because it would be hard to move from the sharp distribution $\mathtt{f}(\boldsymbol{\theta}|\mathbf{x_A})$ to

the target distribution $f(\boldsymbol{\theta}|\mathbf{x_A}, \mathbf{x_B})$.

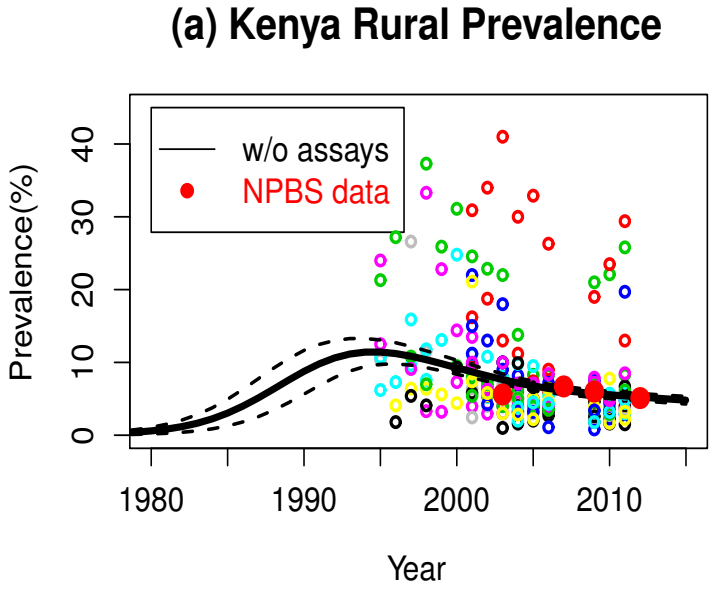
3.3 Numerical Studies

In this section, we presents both the method described in Bao, Ye, and Hallett (2014) (**Method 1**) and our newly proposed method S-IMIS (**Method 2**) to incorporate the incidence assay data. We consider **Method 1** as a benchmark. Also, a comparison of the new and original stopping rule is performed here.

The data we use is from Kenya rural area, and has Both ANC data and NPBS data up to year 2012. The NPBS data here refers to the prevalence data without and simulate new incidence assays data. Besides, we also have the posterior distribution of prevalence and incidence without assay data. In order to avoid future confusion, we call this posterior distribution **result 1**.

The basic framework of our numerical study is described below. Since we already have (**result 1**), from algorithm 2.2.2, the following parameters are recorded during the procedure: total iteration number K, prior functions $p(\cdot)$, the mixture sampling distribution $q^{(k)}$, all the samples $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_{N_K}$ and their weights $w_1, \ldots w_{N_K}$. Figure 3.1 plots the prevalence and incidence trends in **result 1**, along with the real NPBS and ANC data. The colorful ANC data varies due to the location and capability of different clinics. NPBS data point estimation are very close to the median of posterior estimates. The trends are calculated by taking median of posterior distribution in each year.

Now, assume in 2012, we can further collect incidence assay data. Here, we simulate incidence data based on the 2012 posterior median of prevalence and incidence estimates in **result 1** : 0.0524 as prevalence estimate and 0.002384 as incidence estimate. The ratio r between incidence and prevalence estimates is 0.0456.



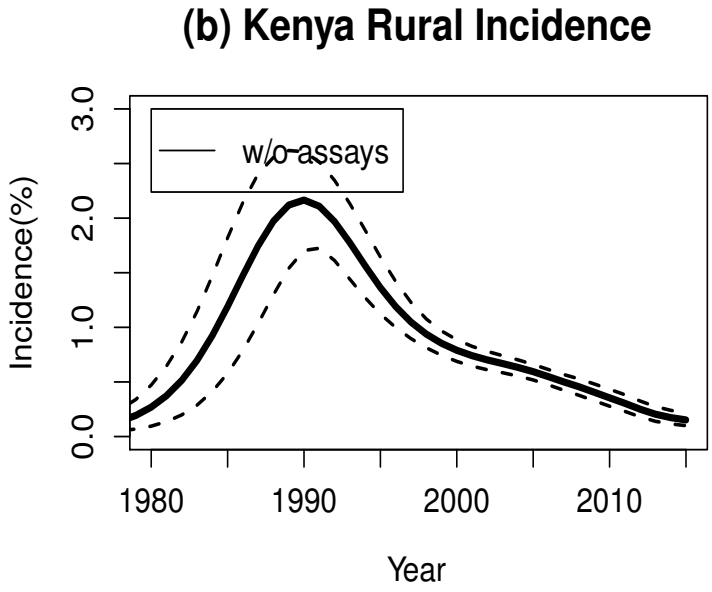


Figure 3.1: Kenya Rural prevalence and incidence trend.

(a) It shows the estimated prevalence estimates time trend derived by EPP without using incidence assays. Black line is the prevalence estimates trend and the dash line is the 95% credible interval of the posterior distribution. The large red dots are the NPBS data, and the colorful dots are ANC data. Each color represents one clinic. (b) This is the incidence estimates time trend plots derived by *result 1*. Black line is the incidence estimates trend and the dash line is the 95% credible interval of the posterior distribution.

As point out in (2.1.9), in order to simulate incidence assays data, four parameters should be taken into consideration: prevalence rate ρ, incidence rate I, false recent rate (FRR) β, and mean duration of recent infection (MDRI) Ω.

Setting 3.3.1. *Since our goal is to explore all the scenarios, we allow the parameters to change from following settings:*

$\cdot\rho$ *is the same as the prevalence estimate in* **result 1**,

$\cdot$I *is is formulated as the ratio with* ρ*, and detailed setting differs in different scenarios.*

$\cdot\beta = 0.025$,

$\cdot\Omega = 150$.

Once we have the simulated data, we can generate the new posterior distribution of prevalence incidence using the methods in Section 3.1 and Section 3.2. Still, in order to simplify the notation, we can these two new psoterior distributions as **result 2** and **result 3**, with respect to **Method 1** and **Method 2**. We can do comparison and further analysis based on those results.

3.3.1 Stopping Rule

First, we show the advantage of the new stopping rule. We believe combining this new additional rule, $|\mathtt{MarLike}_k/\mathtt{MarLike}_{k-1} - 1| < 0.001$ stands for consecutive 3 steps, and the original stopping rule could more efficiently prevent the false converge problem.

Simulation Design 1:

We set incidence rate as 0.1 times prevalence estimates, and the other parameters are the same as in Setting 3.3.1. We compare three stopping rules: the expected number of unique points criterion(old stopping rule); the expected number of unique points criterion and the additional marginal likelihood criterion (new stopping rule),

and an arbitrary large number, saying 100 iterations as the "correct" stopping rule. Our goal is to compare the performance under the old stooping rule v.s. the new stopping rule. And the results under 100 iterations is the benchmark here.

The outputs in Simulation Design 1 is in Appendix A. When we only apply the old rule, the algorithm ends after 3 iterations. From the outputs, we can find our the unique points have already achieved 718, which is larger than the critical value $1000 \times (1 - e^{-1}) = 632$. However, when we apply the new stopping rule, the ratio of marginal likelihood becomes $-30.836/(-21.596) - 1 = 0.428$. Under the new stopping rule, it arrives convergence in around 56 iterations.

To ensure the convergence, we continue iteration the IMIS algorithm until reaching 100 iterations. We compare the three scenarios in these figures, and show that estimates stopping by new rule has the almost exact the same result as the 100 iterations, however, it has a great difference from the estimates stopped by the old rule.

Figure 3.2 presents the time trend at different number of iterations, old rule in black with 3 iterations, new rule in blue with 56 iterations, and 100 iterations in green. The curves produced under the new stopping tule largely overlap with 100 iterations but clearly distinct from the ones produced under the old stopping rule.

Figure 3.3 shows the expected number of unique points and the ratio of likelihood as the IMIS algorithm continues. From Figure 3.3(b), we can see there is a obvious peak in the iteration 3, which leads the IMIS stop at a false converge. The peak is due to the existence of local maximum. However, the ratio of likelihood has not been stabilized, so that the new stopping rule forces IMIS run a few more iterations and move away from the local maximum to the true value.

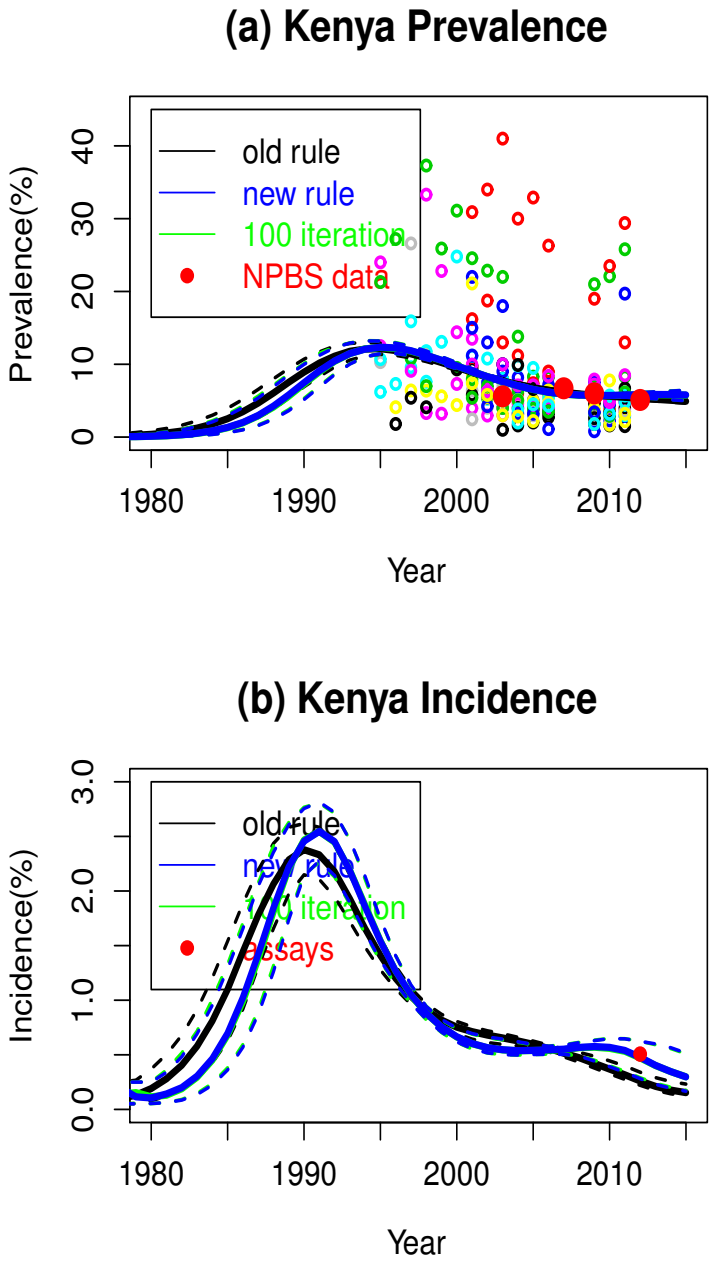


Figure 3.2: Stopping criteria plots.

(a) This is the prevalence estimates time trend at different number of iterations, old rule in black with 3 iterations, new rule in blue with 56 iterations, and 100 iterations in green. (b) This is the incidence estimates time trend at different number of iterations, old rule in black with 3 iterations, new rule in blue with 56 iterations, and 100 iterations in green.

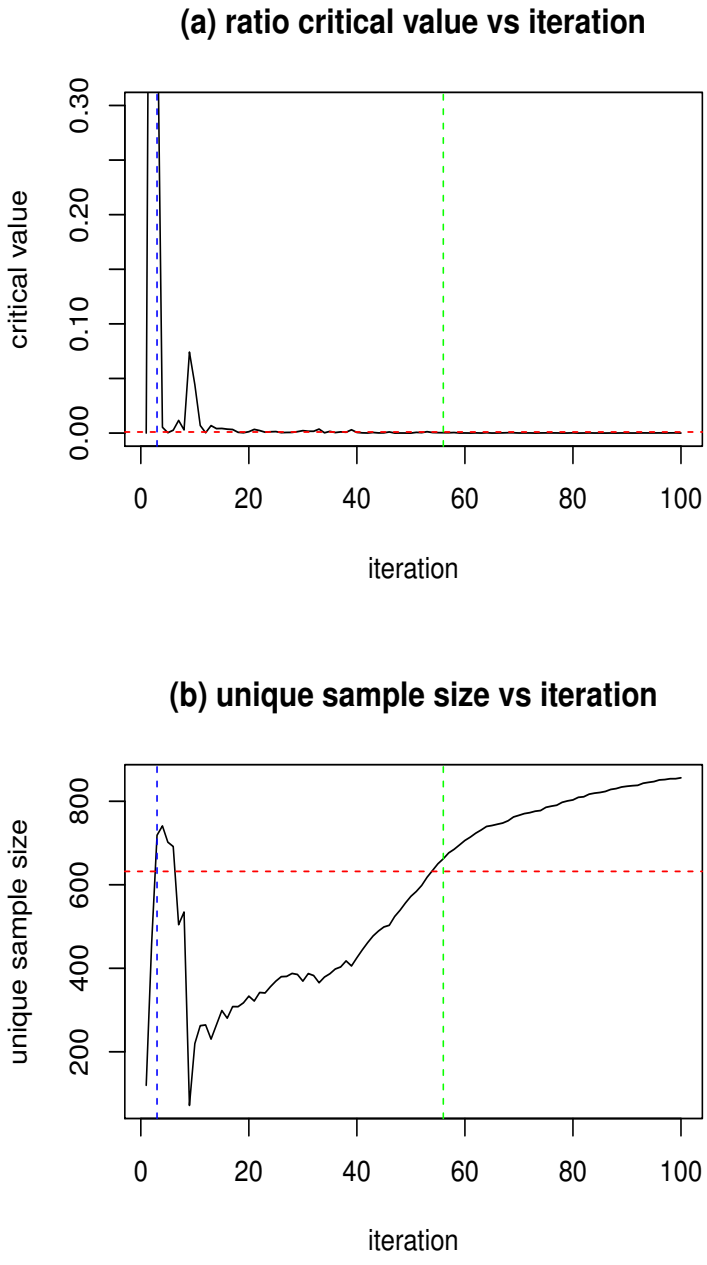


Figure 3.3: Stopping rule comparison.

The targeted values in both rule for the three scenarios. (a)Targeted values for new rule, the blue line is the old rule stop place, the green line is the new rule stop place, and the red line is the critical value (0.001). (b)Targeted values for old rule, the blue line is the old rule stop place, the green line is the new rule stop place, and the red line is the critical value (632).

3.3.2 Sequential IMIS

After establishing the stopping rule, we further compare **Method 1** and **Method 2**, in order to justify the proposed S-IMIS.

Simulation Design 2:

We set incidence rate from 0.01 to 0.15 times prevalence estimates, and the other parameters are the same as the settings above. We use both **Method 2**, that utilize the posterior $\mathrm{f}(\boldsymbol{\theta}|\mathbf{x_A})$, and **Method 1**. We apply the newly proposed stopping rule in Section 3.2.3 to the algorithm. Our goal is to validate sequential IMIS.

The proposed sequential IMIS (**Method 2**) will be considering as successful with two conditions. First, the resulting posterior samples are close enough to the posterior samples generated by **Method 1**. Second, the computing time is reduced comparing with the computing time of (**Method 1**).

Figure 3.4 compares the sequential IMIS results that fit both the prevalence data and the incidence data (black) under different simulations with ratio from 0.01 to 0.15. The ratio suggested by prevalence data is 0.045, which indicates a consistency of point estimates from both prevalence and incidence data.

Compare all the plots, and we find out that the difference between estimates from sequential method and the estimates from full-run EPP is insignificant until the ratio is 0.08. Since the ratio suggested by prevalence data is 0.045, we can roughly conclude that once the point estimates from both methods are close, sequential IMIS provides estimates close to the estimates from full-run EPP.

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15
prev	0.56	0.51	0.55	0.56	0.56	0.58	0.47	0.11	0.10	0.10	0.10	0.07	0.04	0.04	0.04
inc	0.58	0.51	0.56	0.56	0.54	0.59	0.47	0.10	0.09	0.11	0.08	0.08	0.07	0.04	0.05

Table 3.1: Comparison of distribution of **Result 2** and **Result 3**

In order to quantify the similarity between two posterior, we calculate the probability of **Result 3** posterior greater than **Result 2** posterior on the 15 ratios different simulation settings, which is the Wilcoxon test statistics divide by sample size of **Result 2** times sample size of **Result 3**. Table 3.1 shows the test statistic of each ratio. Since the statistics measure the probability of prevalence or incidence from S-IMIS and the full-run EPP, 0.5 suggests similar posterior distribution.

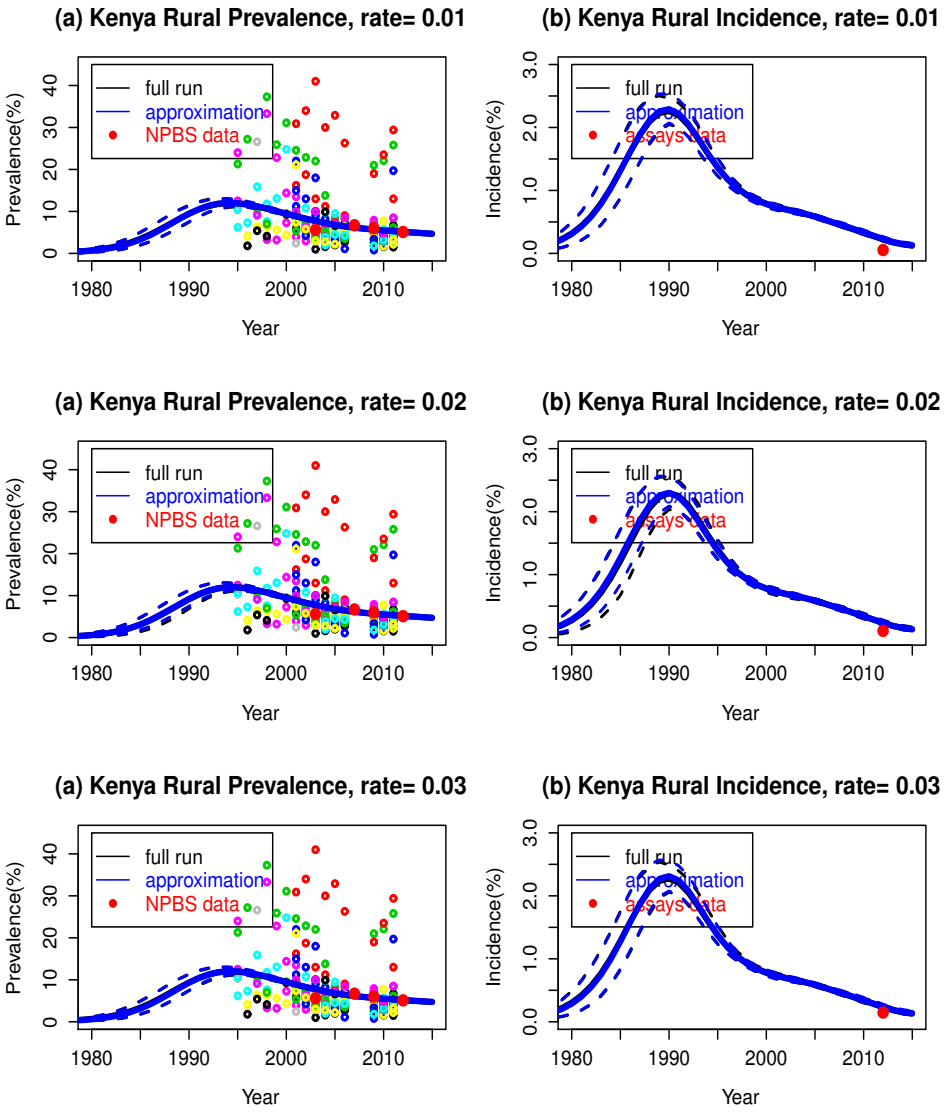


Figure 3.4: Kenya Rural prevalence and incidence for two methods comparison.

The results from simulation settings ratio = 0.01, 0.02, 0.03. (a) This is the prevalence estimates time trend plots derived by results from full-run EPP and Sequential IMIS. Blue line is the prevalence estimates trend and the dash line is the 95% credible interval of the posterior distribution. Black line is the prevalence estimates trend of full-run results and the dash line is the 95% credible interval of the posterior distribution. The large red dots are the NPBS data, and the colorful dots are ANC data. Each color represents one clinic. (b) This is the incidence estimates time trend plots for both methods. Blue line is the incidence estimates trend and the dash line is the 95% credible interval of the posterior distribution. Black line is the incidence estimates trend of full-run results and the dash line is the 95% credible interval of the posterior distribution. The large red dot is the incidence rate used in the simulation settings.

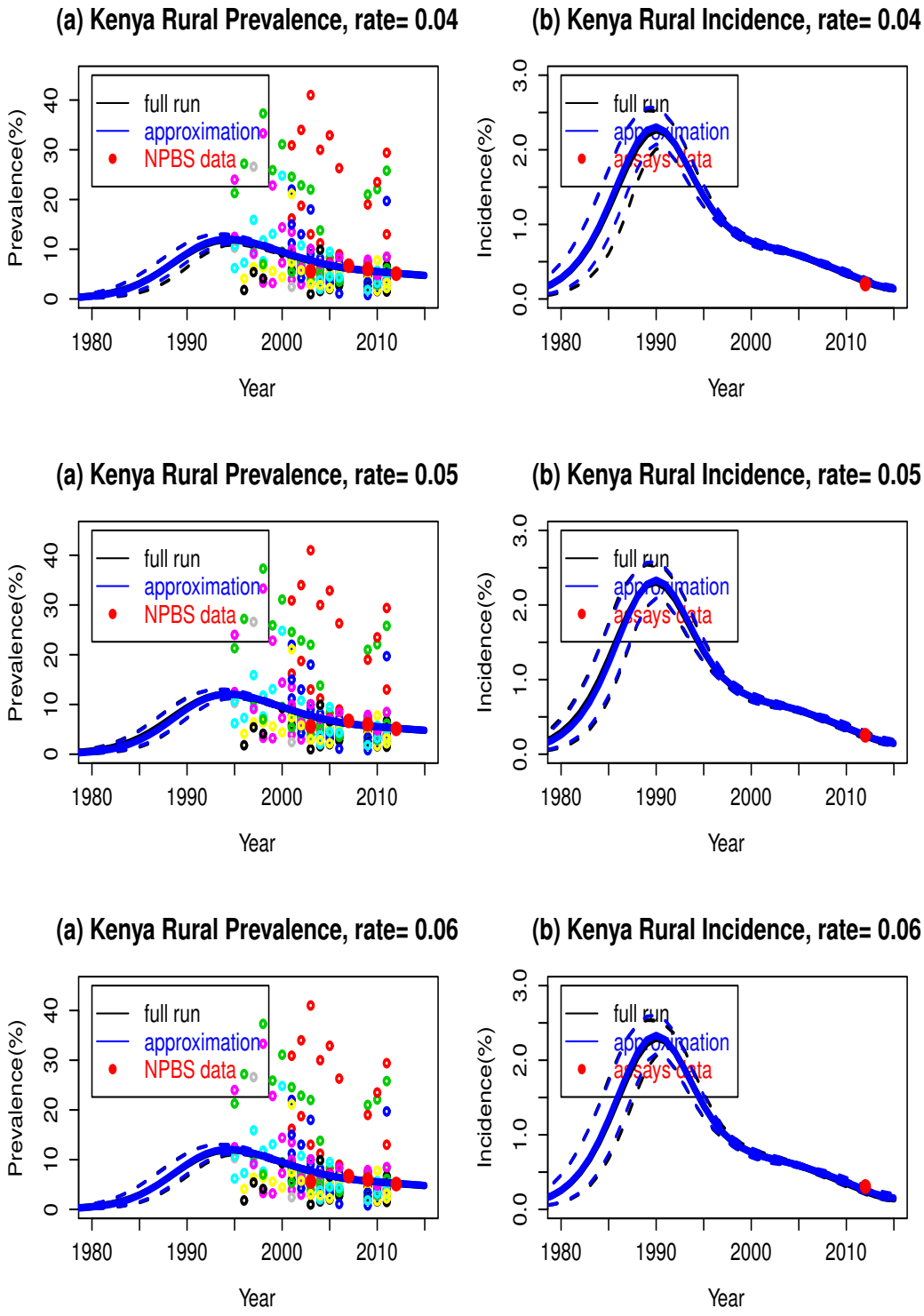


Figure 3.4(continue): The results from simulation settings ratio = 0.04, 0.05, 0.06.

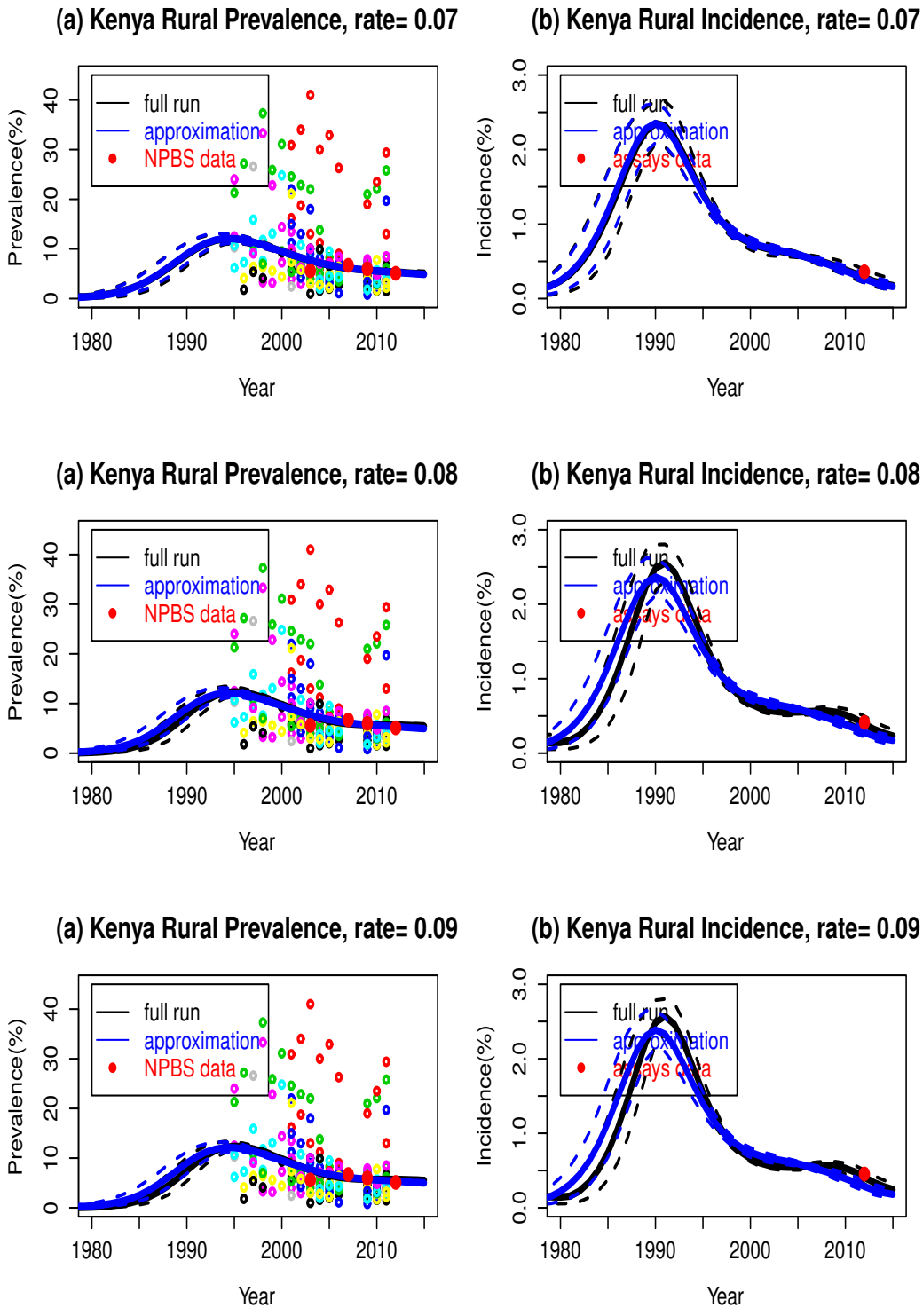


Figure 3.4(continue):The results from simulation settings ratio = 0.07, 0.08, 0.09.

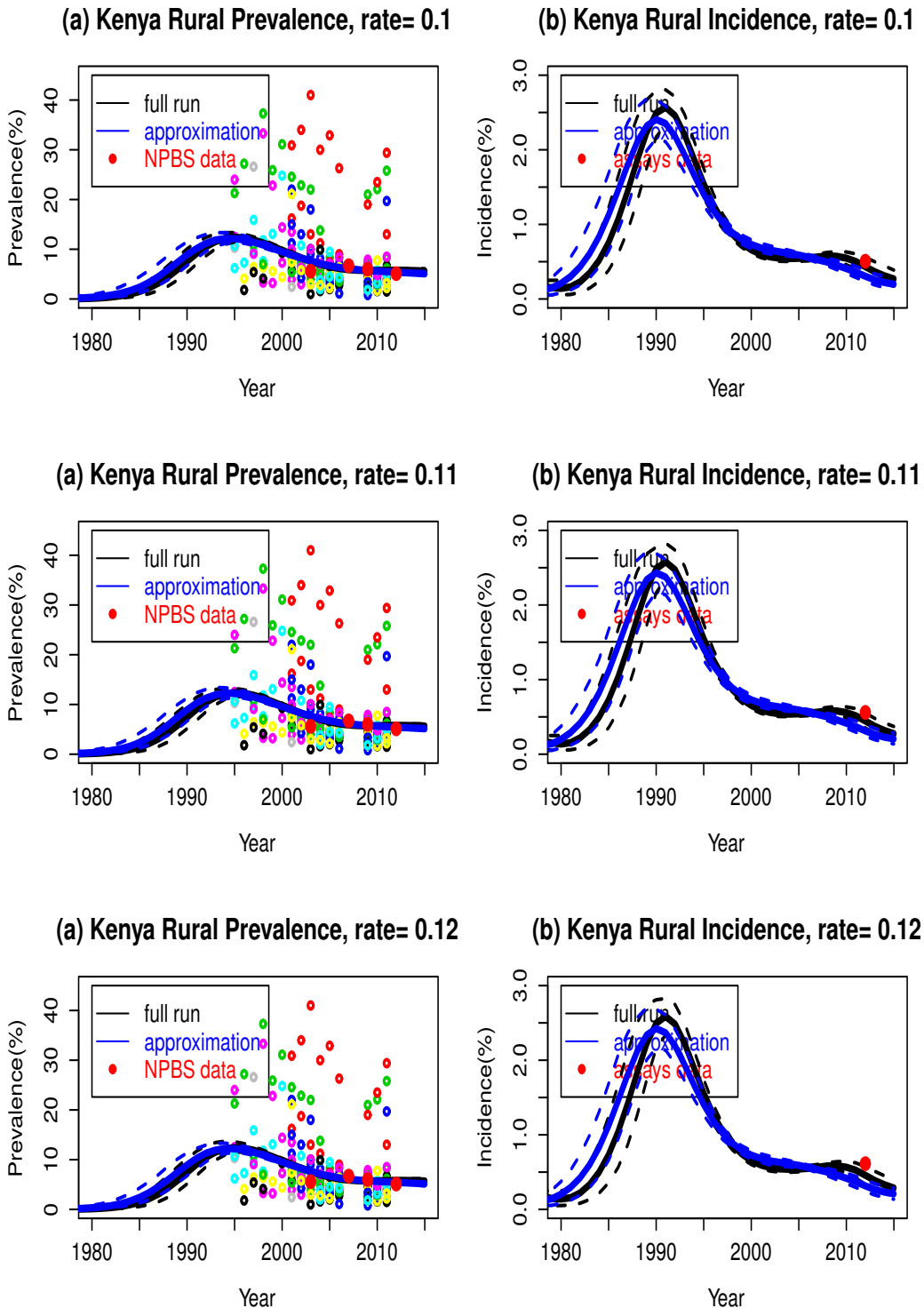


Figure 3.4(continue):The results from simulation settings ratio = 0.10, 0.11, 0.12.

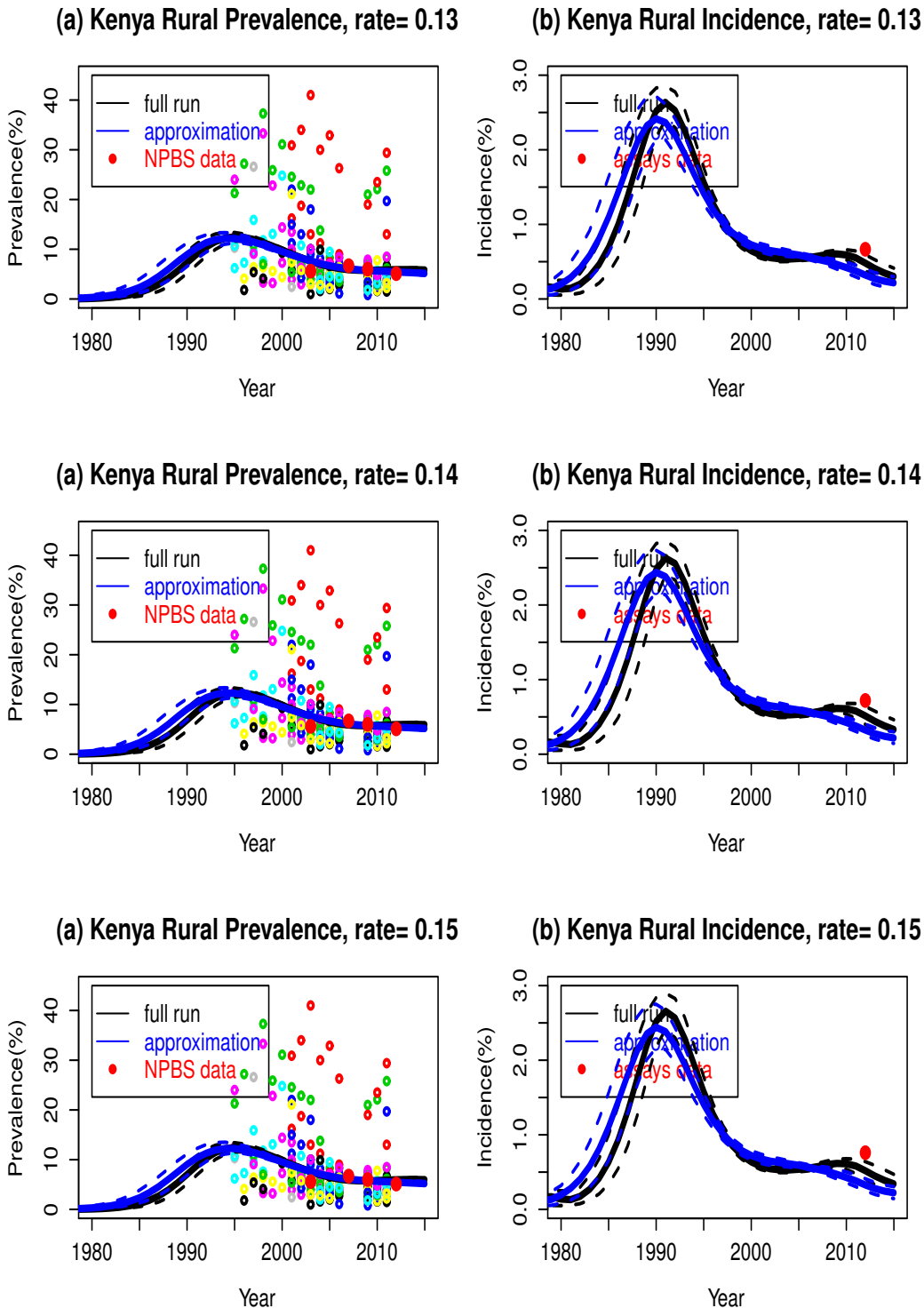


Figure 3.4(continue): The results from simulation settings ratio = 0.13, 0.14, 0.15.

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15
approx	834	826	843	861	783	785	960	1013	1148	858	1864	3736	5492	9204	12974
full-run	5213	6263	6834	6944	9718	7193	8530	9399	7833	9949	10288	7292	9224	9709	7810

Table 3.2: Computing time comparison

Table 3.2 shows the computing time of both methods of each ratio setting. We can see our method constantly has less computing time than (**Method 1**). Especially when the two data sets have consistent estimates, our method has a significant time reduce. When the ratio equals to 0.05, our methods spend less than 1/10 time compared to **Method 1**.

As a summary, if the two data sources has similar estimates, S-IMIS produce estimates with much less computing time, and the difference between **result 2** and **result 3** has not significant difference. However, once the difference between estimates from two data sets is signifiant, the original IMIS (**Method 1**) is suggested. For instance, in the EPP application, ratio larger than 0.08.

In reality, considering both data measure the same parameter $\mathsf{f}(\boldsymbol{\theta})$, they should yield to the same estimates especially with large sample size. But sometimes, when sample size is small and some other random effect, the two data could have really big difference. Then, we first need to check the reliability of the data for missing bias or use different model to fit the data. Once we confirm the collecting procedure and choice of model is correct, we can use the following criteria to determine whether to use **method 1** or **method 2**.

If the estimates from only new data source(incidence assay here) falls into the 95% credible interval of the old posterior distribution(**result 1**), S-IMIS is suggested. Otherwise, a full-run of the EPP is a better way. For example, in Figure 3.5(b), when ratio is We also propose a criteria to choose which method to use. illustrates the criteria to choose the methods. When ratio is 0.05, it falls in between of the black dash lines(95% credible interval), and S-IMIS is suggested. When ratios is 0.1, it is outside of black lines, and **method 1** is suggested. Figure 3.4

justifies our choice.

3.4 Outputs in Chapter 3

```
Output for iteration 100
[1] "5248 non-zero likelihoods, time used= 11.61 minutes"
[1] "Stage   MargLike   UniquePoint   MaxWeight   ESS"
[1]   1.000  -6.315 119.705   0.105  24.404
[1]   2.000 -21.596 458.063   0.027 267.749
[1]   3.000 -30.836 718.602   0.028 602.313
----------------------------------------------------old rule stops
[1]    4.000  -31.001  741.018    0.011 1168.251
[1]   5.000 -31.011 702.011   0.053 261.477
[1]   6.000 -31.092 691.626   0.071 165.041
[1]   7.000 -30.731 504.429   0.304  10.455
[1]   8.000 -30.821 534.685   0.185  23.181
[1]   9.000 -28.535  71.633   0.887   1.270
[1]  10.000 -29.801 219.687   0.585   2.853
[1]  11.000 -30.008 262.534   0.347   5.820
[1]  12.000 -30.018 264.268   0.334   6.782
[1]  13.000 -29.812 230.362   0.251   9.852
[1]  14.000 -29.934 264.618   0.141  20.488
[1]  15.000 -30.061 298.849   0.090  29.316
[1]  16.000 -29.952 280.488   0.129  25.147
[1]  17.000 -30.050 308.258   0.079  36.134
[1]  18.000 -30.029 308.098   0.090  35.491
[1]  19.000 -30.039 317.320   0.064  45.611
[1]  20.000 -30.076 333.308   0.059  54.033
```

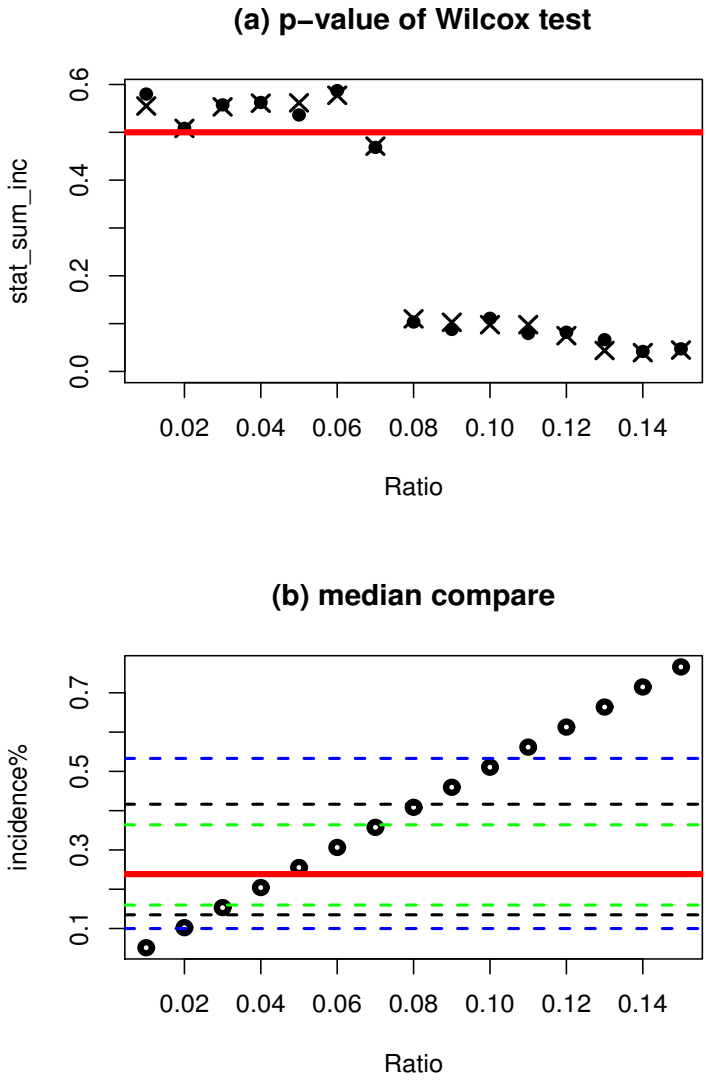


Figure 3.5: Criteria for the intervals.

(a) The dots are the two sample Wilcox test statistics between incidence posterior distributions of full-run and S-IMIS methods. The crosses are the p-value of two sample Wilcox test between prevalence posterior distributions of full-run and S-IMIS methods. The red line is 0.05 line. (b)The dots are the simulation settings of incidence rate. The red line is the posterior mean of incidence estimates of old result. the blue dash lines are the 98% credible interval of incidence estimates, the black dash lines are the 95% credible interval of incidence estimates, and the green dash lines are the 90% credible interval of incidence estimates.

```
[1]  21.000 -29.975 321.637   0.053  57.538
[1]  22.000 -30.044 342.039   0.044  63.097
[1]  23.000 -30.019 340.932   0.076  56.623
[1]  24.000 -30.055 356.013   0.048  70.901
[1]  25.000 -30.097 369.231   0.038  79.573
[1]  26.000 -30.109 379.874   0.035  88.377
[1]  27.000 -30.092 380.584   0.031  94.310
[1]  28.000 -30.110 387.671   0.031 101.423
[1]  29.000 -30.071 385.217   0.029 107.654
[1]  30.000 -30.004 369.242   0.094  62.584
[1]  31.000 -30.055 387.408   0.076  75.780
[1]  32.000 -30.005 382.624   0.058  72.828
[1]  33.000 -29.895 365.362   0.085  61.785
[1]  34.000 -29.887 379.018   0.060  87.057
[1]  35.000 -29.840 386.925   0.043 107.167
[1]  36.000 -29.826 398.287   0.040 115.750
[1]  37.000 -29.793 403.220   0.030 127.045
[1]  38.000 -29.816 417.667   0.028 142.495
[1]  39.000 -29.725 405.562   0.052 116.976
[1]  40.000 -29.743 425.286   0.043 143.728
[1]  41.000 -29.749 444.032   0.036 169.013
[1]  42.000 -29.747 461.510   0.030 200.673
[1]  43.000 -29.763 477.399   0.026 227.062
[1]  44.000 -29.767 489.319   0.023 252.849
[1]  45.000 -29.760 498.971   0.021 274.883
[1]  46.000 -29.729 503.016   0.026 260.210
[1]  47.000 -29.729 524.070   0.022 315.287
[1]  48.000 -29.730 539.137   0.019 336.400
```

```
 [1]  49.000 -29.732 556.620   0.016 370.904
 [1]  50.000 -29.732 572.317   0.014 414.458
 [1]  51.000 -29.713 583.981   0.012 451.223
 [1]  52.000 -29.724 598.408   0.011 494.694
 [1]  53.000 -29.762 618.051   0.010 549.297
 [1]  54.000 -29.776 634.161   0.009 605.264
 [1]  55.000 -29.787 650.282   0.008 667.427
 [1]  56.000 -29.797 662.256   0.008 717.341
---------------------------------------------- new rule stops
 [1]  57.000 -29.806 676.346   0.007 784.143
 [1]  58.000 -29.791 684.865   0.006 836.189
 [1]  59.000 -29.789 695.390   0.006 892.202
 [1]  60.000 -29.792 706.116   0.005 956.517
 [1]   61.000  -29.792  714.153    0.005 1007.579
 [1]   62.000  -29.786  723.499    0.005 1068.246
 [1]   63.000  -29.785  731.126    0.004 1121.385
 [1]   64.000  -29.794  739.661    0.004 1177.939
 [1]   65.000  -29.795  741.861    0.004 1197.924
 [1]   66.000  -29.798  744.911    0.004 1220.302
 [1]   67.000  -29.803  747.822    0.004 1241.028
 [1]   68.000  -29.793  753.302    0.004 1292.066
 [1]   69.000  -29.801  762.359    0.004 1366.748
 [1]   70.000  -29.799  766.289    0.003 1406.213
 [1]   71.000  -29.802  770.437    0.003 1445.499
 [1]   72.000  -29.799  772.604    0.003 1469.464
 [1]   73.000  -29.800  776.148    0.003 1504.058
 [1]   74.000  -29.796  777.777    0.003 1523.180
 [1]   75.000  -29.792  785.447    0.003 1610.189
```

```
[1]   76.000  -29.796  788.220    0.003 1638.571
[1]   77.000  -29.799  790.478    0.003 1663.605
[1]   78.000  -29.800  797.457    0.003 1753.247
[1]   79.000  -29.801  800.812    0.003 1794.805
[1]   80.000  -29.795  803.194    0.003 1828.138
[1]   81.000  -29.795  809.396    0.002 1918.866
[1]   82.000  -29.797  810.749    0.002 1937.280
[1]   83.000  -29.801  816.954    0.002 2030.865
[1]   84.000  -29.802  819.425    0.002 2068.248
[1]   85.000  -29.804  820.974    0.002 2091.062
[1]   86.000  -29.804  823.452    0.002 2129.940
[1]   87.000  -29.802  828.278    0.002 2216.591
[1]   88.000  -29.801  830.301    0.002 2248.874
[1]   89.000  -29.798  834.133    0.002 2318.954
[1]   90.000  -29.795  835.888    0.002 2350.428
[1]   91.000  -29.794  837.239    0.002 2377.634
[1]   92.000  -29.792  838.162    0.002 2396.648
[1]   93.000  -29.796  843.140    0.002 2497.283
[1]   94.000  -29.800  845.314    0.002 2542.739
[1]   95.000  -29.803  847.147    0.002 2581.025
[1]   96.000  -29.807  851.064    0.002 2672.996
[1]   97.000  -29.809  851.918    0.002 2692.195
[1]   98.000  -29.804  853.920    0.002 2741.951
[1]   99.000  -29.799  853.964    0.002 2731.405
[1]  100.000  -29.800  856.156    0.002 2783.588
```

Chapter 4

Impact of Incidence Assay Data

This chapter is a continuous study of Chapter 3. In Chapter 3, we justify our newly proposed sampling method, and a new stopping rule. In this Chapter, we further analyse the impact of adding in incidence assay data to the estimation of prevalence, incidence and the change of incidence over time. For example, we can compare the incidence estimate at 2015 before and after adding in incidence assay data to study the impact to prevalence and incidence. And we can compare the ratio of 2015 incidence and 2010 incidence before and after adding in incidence assay data to study the impact to incidence change over time. In this chapter, we start from the simplest case, which we only have one year assay data, and extend to the case that we have a time series assay data.

4.1 Methods and Goals

Our numeric study is similar to the one in Chapter 3. We will first repeat some importance concepts here. Our strategy is still to simulate incidence assay data, and compare the results before and after incorporating incidence assay data. Four parameters related to incidence assays:

· The prevalence (ρ) is the proportion of HIV positive population.

· The incidence (I) is the proportion of HIV recent infection population.
· The false recent rate (FRR, β), is the proportion of non-recent HIV infections in the population incorrectly classified as being recent. It describes the propensity for individuals with long-standing infection to return recent results.
· The mean duration of recent infection (MDRI, Ω), is the average length of time that people with newly acquired infection in the population are to be classified as having recently acquired infection.

After introducing new incidence assays data, the estimates of incidence is anticipated to be more accurate. And due to the high correlation between prevalence and incidence, it will also improve the accuracy of prevalence estimates. The two quality parameters, FRR and MDRI, describe the reliability of the assays data. We hope that FRR to be suitably low and MDRI to be sufficiently large. However, in reality, FRR cannot be zero and MDRI cannot be life-long. Also, in order to improve the accuracy of the current assay data, more expensive equipments and testing tools are required.

Our goal is to study the impact of assays data to the estimates in EPP model. To be more specific, there are two impacts we are interested in. One is to accurately estimate the impact to prevalence and incidence rates, and another one is to accurately estimate the impact to the changes of prevalence and incidence rates over time. The first one is easy to understand, since the absolute number of HIV infection and new infection is always a focus to our study. The changes of rates over time are also important because of Millennium Sustainable Development Goals: "by 2030, end the epidemics of AIDS" (UNAIDS, 2014a). As we have explained in Chapter 1, in order to achieve this goal, by 2030, a 90% decline compared to 2010 of the number of new HIV infections and AIDS-related deaths is required.

In order to better understand the contribution of incidence assays in the model fitting process, we experiment different hypothetical values of the quality parameters,

FRR and MDRI. However, there are four parameters, which we have no knowledge of the true values are, can impact the performance of assays data. To consider them together, then we might have a large quantity of scenarios in one country or region. And the data are not available or only one year in most countries, it is hard to just use real data to give a full evaluation of assay data. We will rely on the simulation study.

Since incidence assays add information mostly to incidence estimates, we will keep prevalence generating data the same as prevalence posterior median in **result 1** . For the remaining 3 parameters, incidence, FRR, MDRI, we change one parameter at a time. The details are in the next sections.

Some of the notations used in Chapter 3 are still used in this chapter. The posterior distribution generated from only prevalence data (without incidence assay data) is called **result 1**, and the posterior distribution generated from S-IMIS incorporating incidence assay data is called **result 3**.

4.2 Impact of Incorporating Single Year Incidence Assay Data

The settings are similar with the settings in Section 3.3. We still use synthetic data, based on prevalence data obtained from ANC, and in some cases NPBS, from Kenya Rural, to examine the impact of incorporating incidence assay data. We assume the incidence assays data are collected in 2012 as the last year of NPBS data. The following three key parameters are allowed to vary in the simulated scenarios: FRR, MDRI, and incidence rate. Here, we still use R, which is the ratio between the incidence rate using to generate data, and the incidence rate estimated in **result 1** to represent incidence. And we keep the same prevalence rate from which we simulate incidence assays as the prevalence rate in **result 1**. Our goal is

to compare **result 1**, which using only prevalence data and **result 3**, which using both prevalence and incidence assay data.

4.2.1 Impact to Parameter Estimates

The goal is to examine the impact of incidence assays data to absolute prevalence and incidence estimates as parameter changes. The settings is:

Setting 4.2.1. *We define their default values as follows:*
$\mathfrak{n} = 20,000, \beta = 0.025, \Omega = 150, R = 2$. *To simulate the incidence assay data, two parameters will be fixed at the default values, and the remaining one varies.*
· *For* R*, we try* $0.5, 1, 1.5, \ldots, 5$ *(10 levels);*
· *For* β*, we try* $0.005, 0.010, 0.015, \ldots, 0.05$ *(10 levels);*
· *For* Ω*, we try* $100, 110, 120, \ldots, 190$ *(10 levels) .*

We are interested in the estimates of prevalence and incidence in 2015. We compare posterior mean of prevalence, incidence respectively, and the probability of estimates in **result 3** smaller than estimates in **result 1**. The following are the simulation results. In all the figures in this section, the upper plot shows the 95% credible intervals in each scenario. And the lower plot is the percentage that the new estimate from **result 3** is smaller than old estimate from **result 1**.

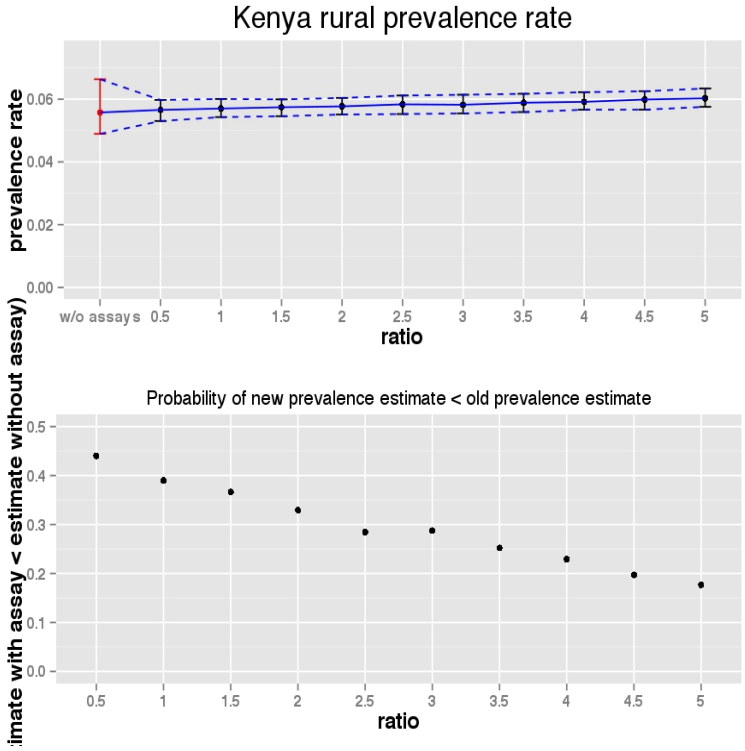


Figure 4.1: Impact of prevalence rate one time assay data with ratio changing.

(a) This is the prevalence posterior mean estimates with simulation ratio from 0.5 to 5, and the estimates from without assays data. The dots in the middle are the estimated prevalence means. The bars are the 95% credible interval of posterior prevalence estimates. (b)This is the percentage that the new prevalence estimates smaller than old posterior.

We find that when introducing the incidence assays data, even though prevalence of incidence assays data is generated the same as the prevalence of prevalence data, due to the effect of strong relationship between prevalence and incidence, prevalence estimates increase slightly. At the meantime, the credible interval has a significant decrease, which indicates an improvement in the estimates accuracy. The probability is expected to decrease considering the increase of ratio.

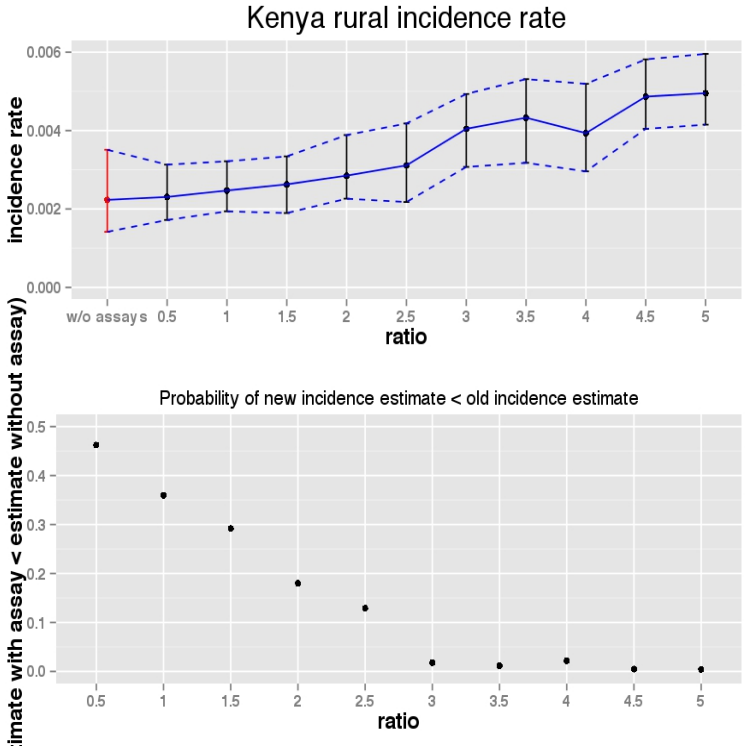


Figure 4.2: Impact of incidence rate for one time assay data with ratio changing.

(a) This is the incidence posterior mean estimates with simulation ratio from 0.5 to 5, and the estimates from without assays data. The dots in the middle are the estimated incidence means. The bars are the 95% credible interval of posterior incidence estimates. (b)This is the percentage that the new incidence estimates smaller than old posterior.

When introducing the incidence assays data, the incidence estimates increase significantly due to the increase of ratio. The credible interval decreases when the ratio is around 1, which suggests the incidence from new incidence assay data is similar with the incidence from prevalence data estimates. However, the interval enlarges as the difference between the incidence from two data estimates become larger. When the ratio is greater than 3, the estimates change dramatically, and leads to the probability goes to 0.

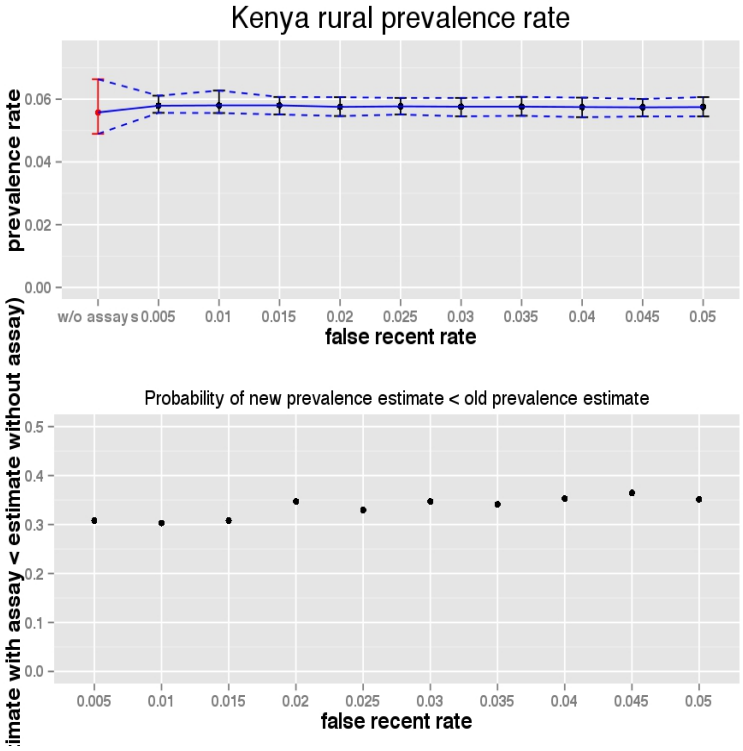


Figure 4.3: Impact of prevalence rate for one time assay data with β changing.

(a) This is the prevalence posterior mean estimates with simulation β from 0.005 to 0.05, and the estimates from without assays data. The dots in the middle are the estimated prevalence means. The bars are the 95% credible interval of posterior prevalence estimates. (b)This is the percentage that the new prevalence estimates smaller than old posterior.

When introducing the incidence assay data, prevalence estimates do not change over the increase of false recent rate β. However, the credible interval has a significant decrease, which indicates an improvement in the estimates accuracy. The probability deceases as false recent rate β increases.

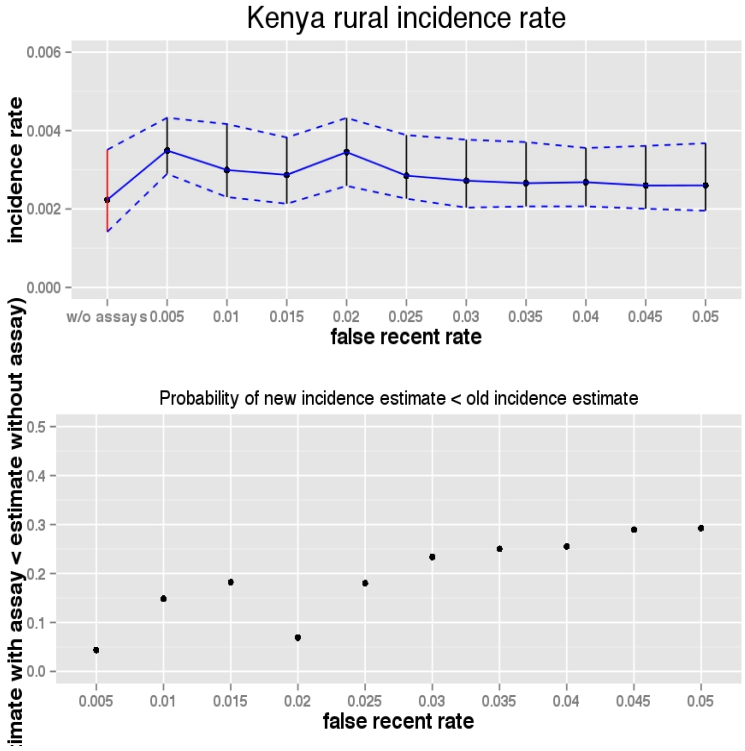


Figure 4.4: Impact of incidence rate for one time assay data with β changing.

(a) This is the incidence posterior mean estimates with simulation β from 0.005 to 0.05, and the estimates from without assays data. The black dots in the middle are the estimated incidence means. The bars are the 95% credible interval of posterior incidence estimates. (b)This is the percentage that the new incidence estimates smaller than old posterior.

As β increases, the estimates decreases. The probability increases as β increases.

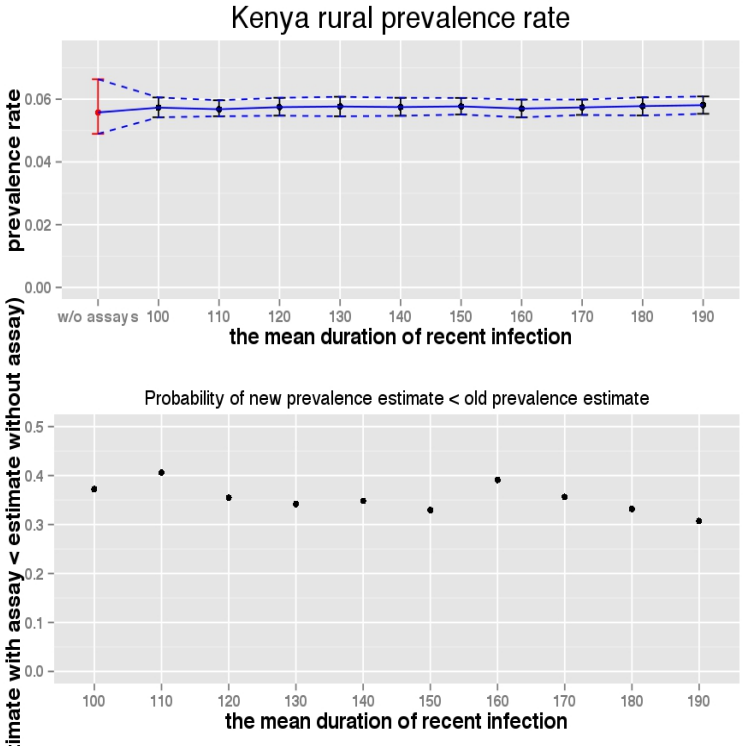


Figure 4.5: Impact of prevalence rate for one time assay data with Ω changing.

(a) This is the prevalence posterior mean estimates with simulation Ω from 100 to 190, and the estimates from without assays data. The dots in the middle are the estimated prevalence means. The bars are the 95% credible interval of posterior prevalence estimates. (b)This is the percentage that the new prevalence estimates smaller than old posterior.

The prevalence estimates do not change with the mean duration of recent infect Ω. However, the credible interval has a significant decrease, which indicates an improvement in the estimates accuracy. The probability slightly deceases as the mean duration of recent infect Ω increases.

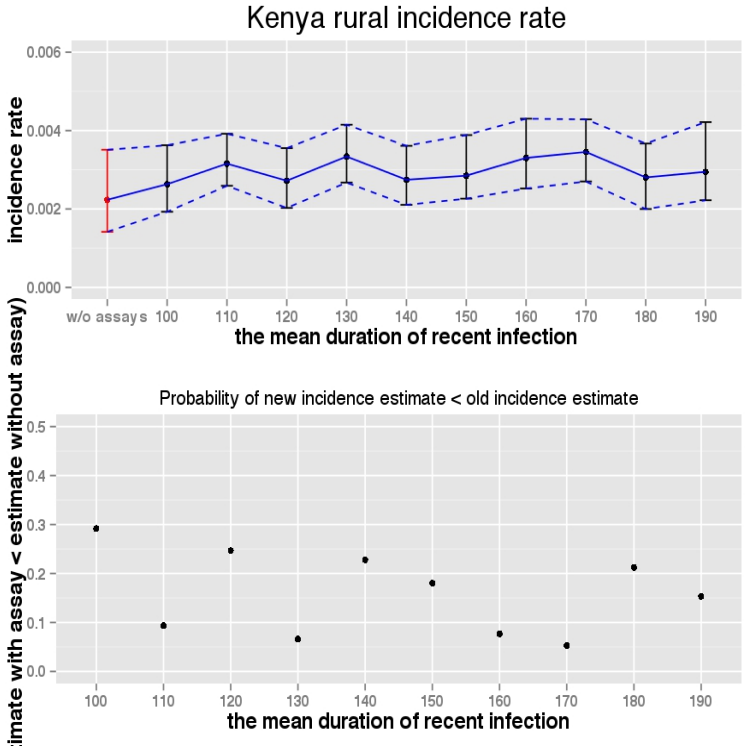


Figure 4.6: Impact of incidence rate for one time assay data with Ω changing.

(a) This is the incidence posterior mean estimates with simulation Ω from 100 to 190, and the estimates from without assays data. The black dots in the middle are the estimated incidence means. The bars are the 95% credible interval of posterior incidence estimates. (b)This is the percentage that the new incidence estimates smaller than old posterior.

Both the estimates and the probability do not have a significant change as Ω increases.

4.2.2 Impact to Change over Time

The goal is to examine the impact of incidence assays data to changes of prevalence and incidence estimates over time as parameter changes. Suppose We are interested in the change of estimated prevalence and incidence rate between 2007 and 2012. So, we check the probability of prevalence and incidence rates in

2007 less than 2012 for both **result 1** and **result 3**.

Setting 4.2.2. *We only consider ratio here.*
· *For* R*, we try* $0.5, 1, 1.5, \ldots, 5$ *(10 levels).*

We are interested in the ratio between prevalence, incidence estimates of 2007 and 2012. We show an extreme case that ratio is 5 to illustrate the impact of incidence assays data.

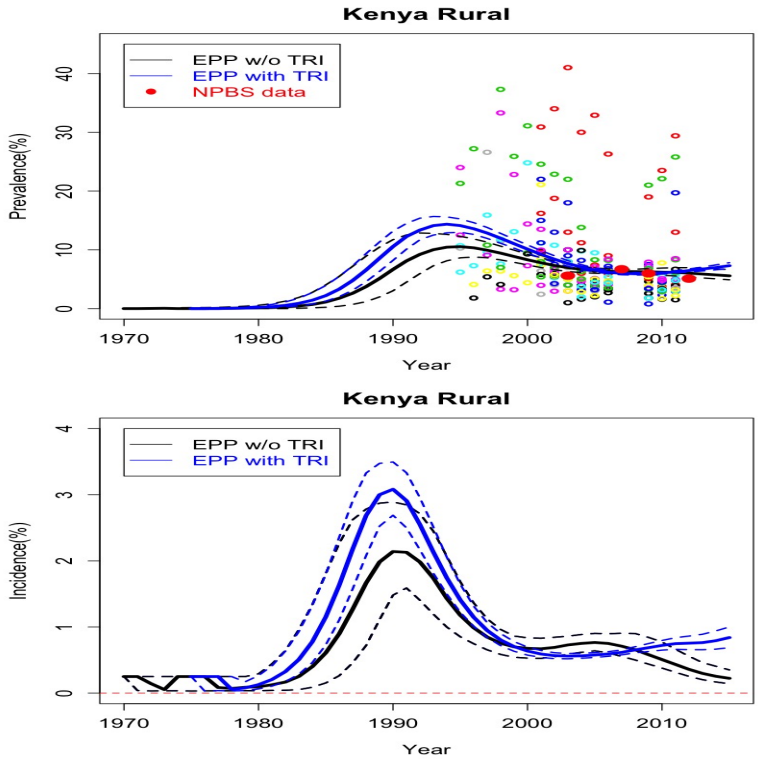


Figure 4.7: Kenya Rural, impact for single year assay data.

Kenya Rural. (a) This is the prevalence time trend posterior mean estimates with simulation ratio 5. Blue line is the prevalence estimates trend after introducing assays data and the dash line is the 95% credible interval of the posterior distribution. Black line is the prevalence estimates without assays data the dash line is the 95% credible interval of the posterior distribution. The large red dots are the NPBS data, and the colorful dots are ANC data. Each color represents one clinic. (b)This is the Kenya Rural incidence time trend posterior mean estimates with simulation ratio 5. Blue line is the prevalence estimates trend after introducing assays data and the dash line is the 95% credible interval of the posterior distribution. Black line is the prevalence estimates without assays data the dash line is the 95% credible interval of the posterior distribution. The large red dots is the simulation incidence rate.

There is a big difference between estimates with and without incidence assays data. After introducing assay data with large incidence, both prevalence and incidence have a significant increase in 2012.

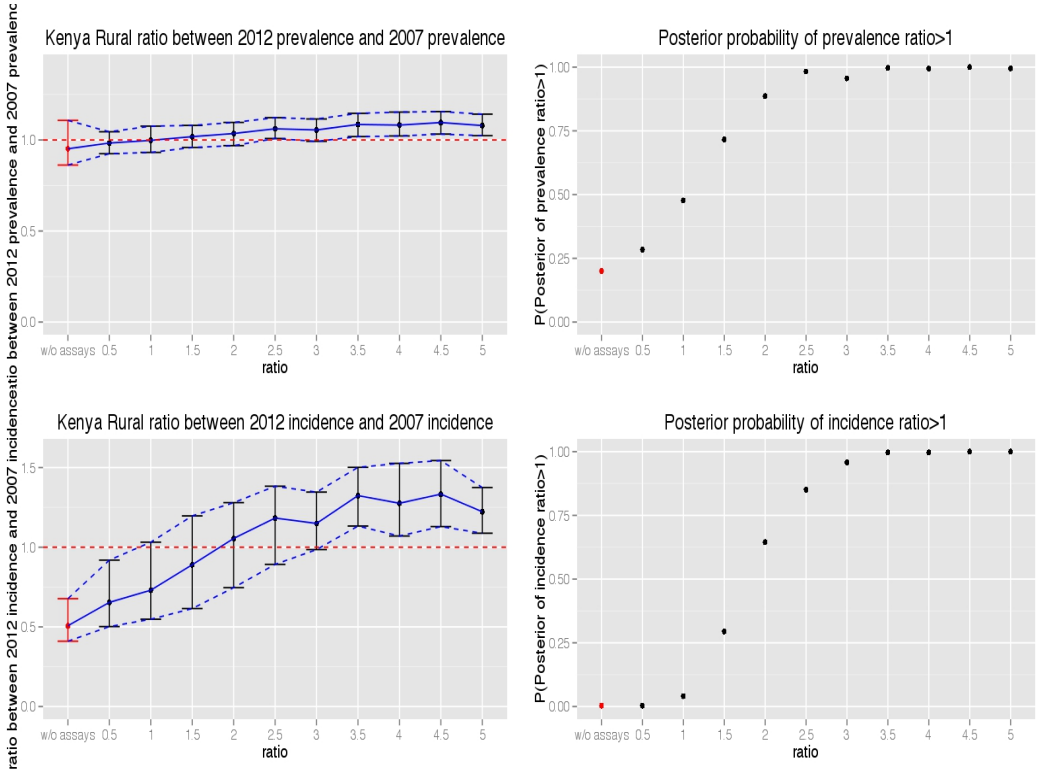


Figure 4.8: Kenya Rural, impact for change over time with single year assay data.

Kenya Rural. The left columns are the ratio between 2012 prevalence, incidence and 2017 as incidence increases. The right columns are the probability of 2012 prevalence, incidence larger than 2007.

Originally, both prevalence and incidence in 2007 is larger than 2012. Since the ratio increases significantly, which indicates the increase of simulation incidence, both probabilities increase as ratio increase.

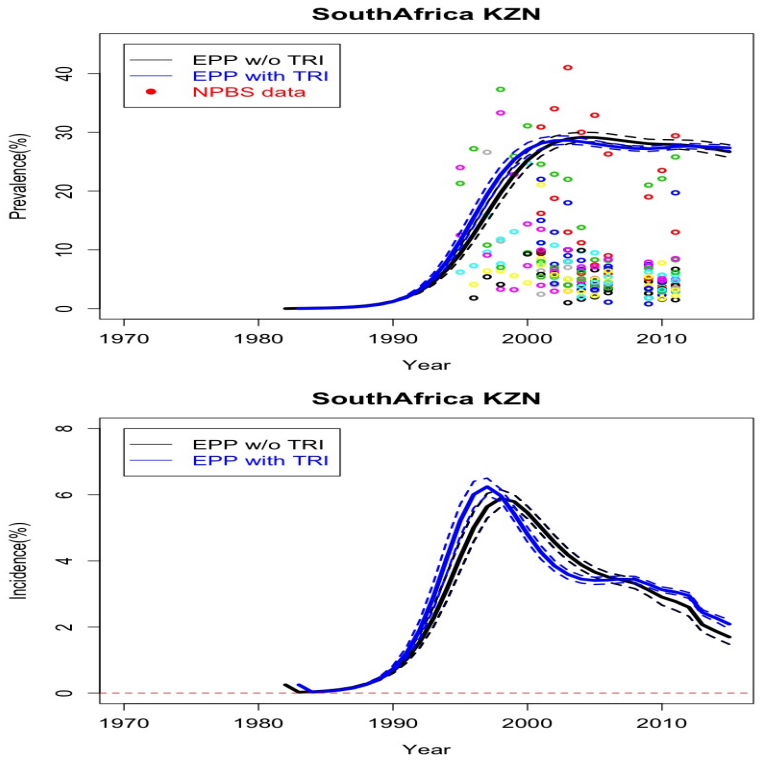


Figure 4.9: South Africa, impact for single year assay data.

South Africa KZN. (a) This is the prevalence time trend posterior mean estimates with simulation ratio 5. Blue line is the prevalence estimates trend after introducing assays data and the dash line is the 95% credible interval of the posterior distribution. Black line is the prevalence estimates without assays data the dash line is the 95% credible interval of the posterior distribution. The large red dots are the NPBS data, and the colorful dots are ANC data. Each color represents one clinic. (b)This is the Kenya Rural incidence time trend posterior mean estimates with simulation ratio 5. Blue line is the prevalence estimates trend after introducing assays data and the dash line is the 95% credible interval of the posterior distribution. Black line is the prevalence estimates without assays data the dash line is the 95% credible interval of the posterior distribution. The large red dots is the simulation incidence rate.

For this region, we can see the original prevalence estimates are stable at high level, and incidence estimates decrease a lot. So, when introducing the incidence assays data, it increases the incidence estimates however not enough to be larger than 2007.

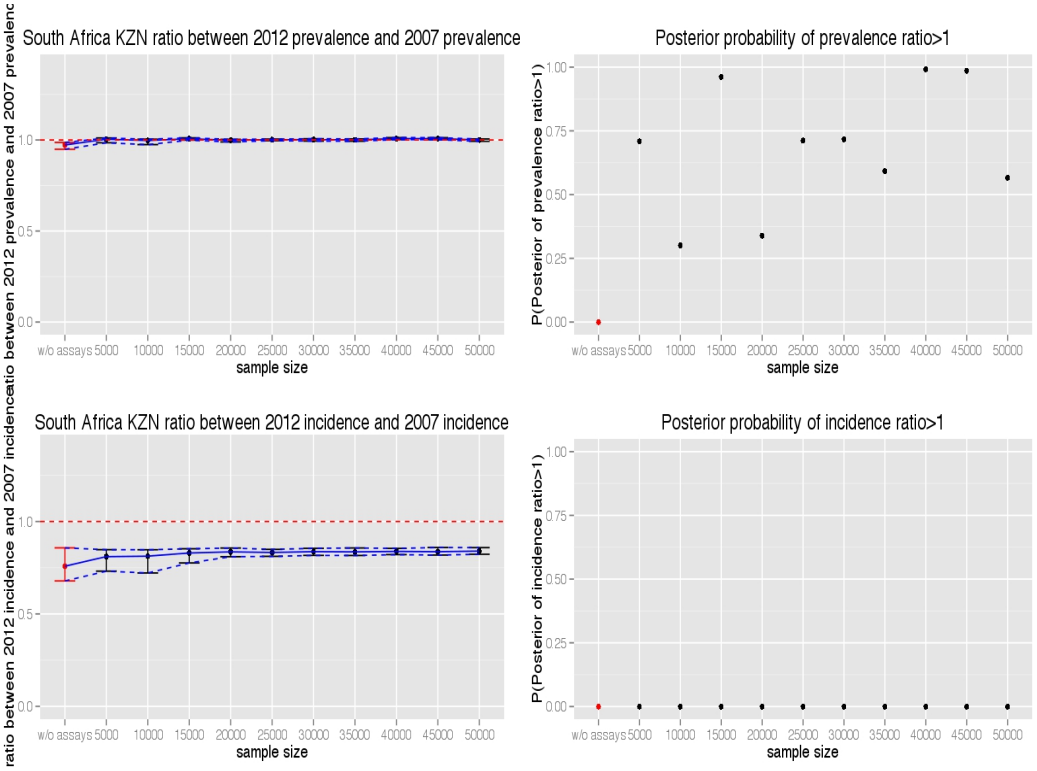


Figure 4.10: South Africa, impact for change over time with single year assay data.

South Africa KZN. The left columns are the ratio between 2012 prevalence, incidence and 2017 as incidence increases. The right columns are the probability of 2012 prevalence, incidence larger than 2007.

For prevalence estimates, since estimates without assays data is very flat after 2003, a little increase can result in a large change of the ratio between prevalence of 2007 and 2012. For incidence estimates, the estimates without assays data decrease significantly. So, even after introducing incidence data with large incidence, it is still not enough for incidence in 2012 to be higher than 2007.

We can draw the conclusion that the effect of incidence assay data to the change over time relies on the original trend in **result 1**. If original trend is flat, then it could significant change the trend. If the original trend is significantly, then a single year incidence data could do little to this change, especially when sample size is small.

4.3 Impact of Time Series Incidence Assays Data

In previous study, we examine the impact of single-year incidence assays with the presence of ANC prevalence, NPBS prevalence and pre-assumed structure of infection rate $r(t)$ within EPP. However, a time series assays data is more realistic. Then, a natural problem arises during the numerical study is that we cannot simulate incidence data only based on those parameters as in Section 3.3 and Section 4.2 since there should be a time effect relationship among prevalence and incidence for each year. Also, the ANC prevalence are biased and $r(t)$ trend assumptions might be incorrect. The r-trend model has a lot restriction on how the epidemic evolves in the future. For some countries, if the epidemic trend is not stable, $r-$trend model will be no longer suitable.

Furthermore, since we have no knowledge on not only future assays not but also prevalence data. It is hard to tell the impact of assays itself by simulate the datasets we use above. In that case, to unpack the contribution of time series incidence assays, we start with simple scenario without complications from ANC data and EPP assumptions. In this part of simulation study, assume that the key quantities summarizing the epidemic before 2012 are known, e.g. HIV prevalence, incidence and morality. It comes from applying EPP to a real dataset, e.g. Kenya rural with ANC data up to 2011 and NPBS data up to 2012. We also assume that the epidemic after 2012 is driven by a free parameter θ,

$$\log r(t) = \log r(t-1) + \theta, \quad \text{for } t > 2012 \tag{4.3.1}$$

We first examine the impact of θ to $r(t)$ and prevalence, incidence rate. Naturally, first take $\theta = 0$, which indicates the infection rate stays same after 2012. For Figure 4.11, it shows the prevalence and incidence still has a clear decreasing trend. If θ taking a negative number, it is easier to predict a faster decreasing trend for

prevalence and incidence. So, we take a positive theta $\theta = 0.01$. Even if this slope is very small, we can still see from Figure 4.12, that the incidence begins on increase after some time, though prevalence still decreasing. The two plots tell that prevalence and incidence are very sensitive to this θ. We use two settings to represent the scenarios that epic slows and epic grows.

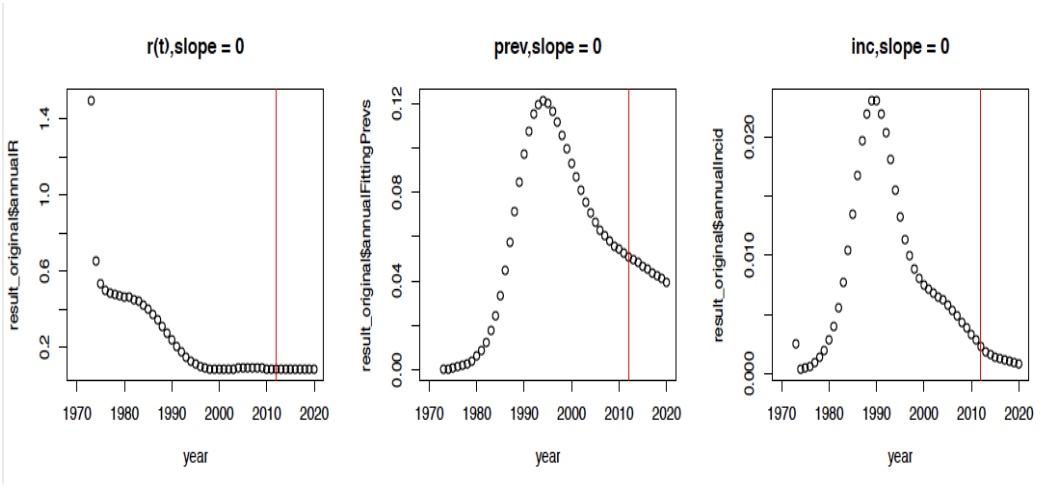


Figure 4.11: Kenya Rural. Data is up to 2012. Assume constant infection rate over time after 2012. Three plots are infection rate, prevalence rate, and incidence rate over time.

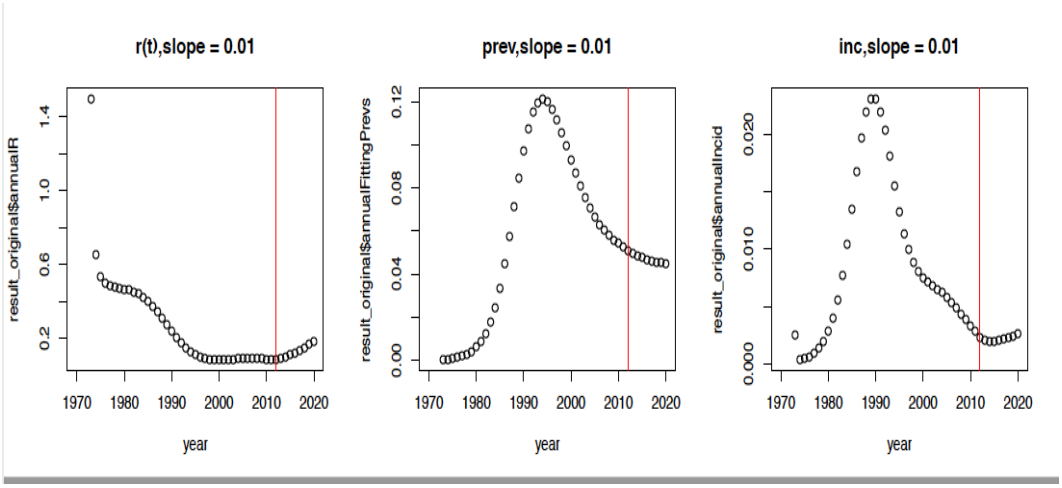


Figure 4.12: Kenya Rural. Data is up to 2012. Assume infection rate changes with a negative slope 0.01 over time after 2012. Three plots are infection rate, prevalence rate, and incidence rate over time.

We simulate NPBS data with incidence assays in 2014, 2016, 2018 and 2020, given different settings of θ, β and Ω. Very similar to the simulation study in Section 4.2, we consider the following scenarios:

Setting 4.3.1. *1. The true incidence (used for simulating surveys) is declining v.s. raising ($\theta = 0, 0.01$ respectively). 2. The false recent rate varies, $\beta = 0.005, 0.01, \ldots, 0.05$. 3.The mean duration varies, $\Omega = 110, 120, \ldots, 200$.*

Still incorporate incidence assay data using S-IMIS. Two new parameters are taken into consideration to evaluate the estimation. First, we compare the percentile of the true value among posterior samples. It means after we draw the posterior sample of incidence in 2020, it is easy to compute the quantile of the true value within these posterior samples. Naturally, the estimation is considering more accurate if this percentile is closer to 50%. Secondly, we compare estimated 2020 incidence rate and true value, which is known the simulation. We already have the simulated 2020 incidence as true value, and also we can estimate this incidence from these time series data. The estimation is considering more accurate if the posterior mean is closer to the true value. The results are summarized in the following plots:

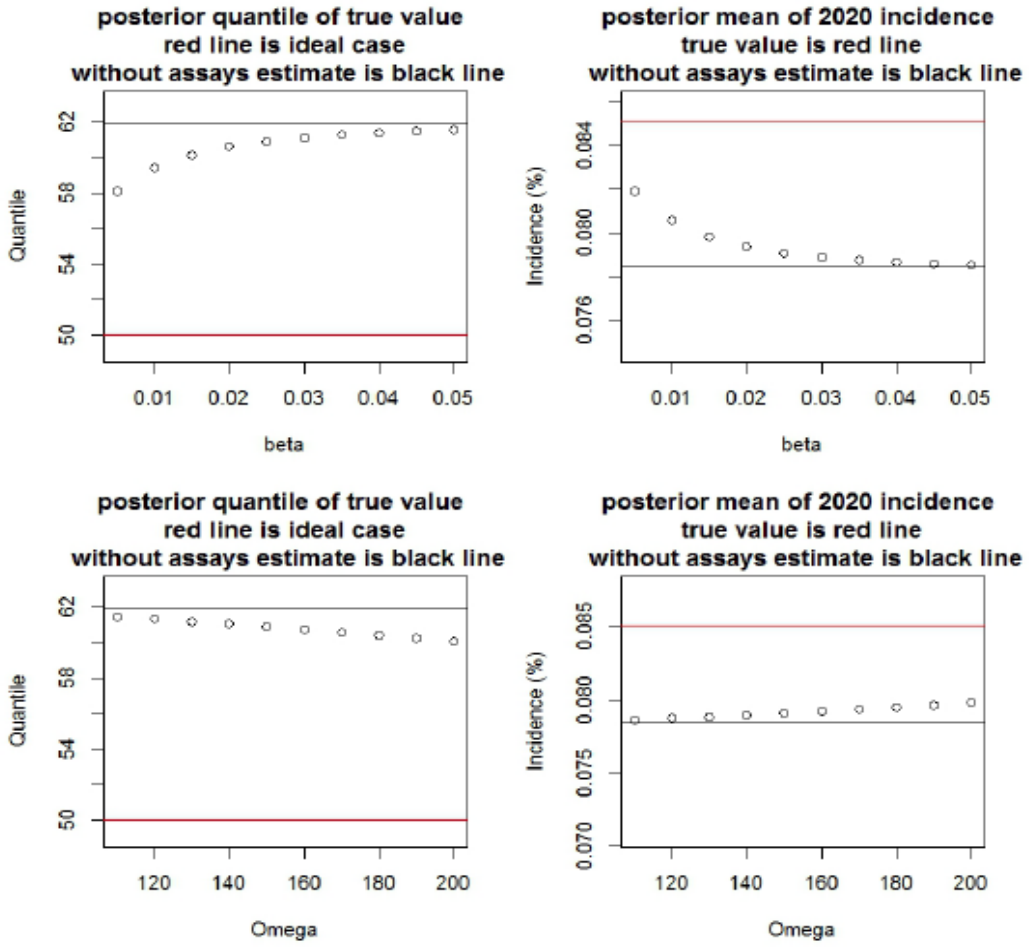


Figure 4.13: Kenya rural, impact to incidence change over time with β changes, for decreasing incidence simulation setting.

Kenya Rural. $\theta = 0$, $n = 7501$, the incidence declining after 2012. The upper two fix $\Omega = 150$, and allow β to change from 0.005 to 0.05. The lower two fix $\beta = 0.025$, and allow Ω to change from 110 to 150. The left two are the quantile of the true value within these posterior samples, and the right two are the posterior mean of estimated 2020 incidence and true value.

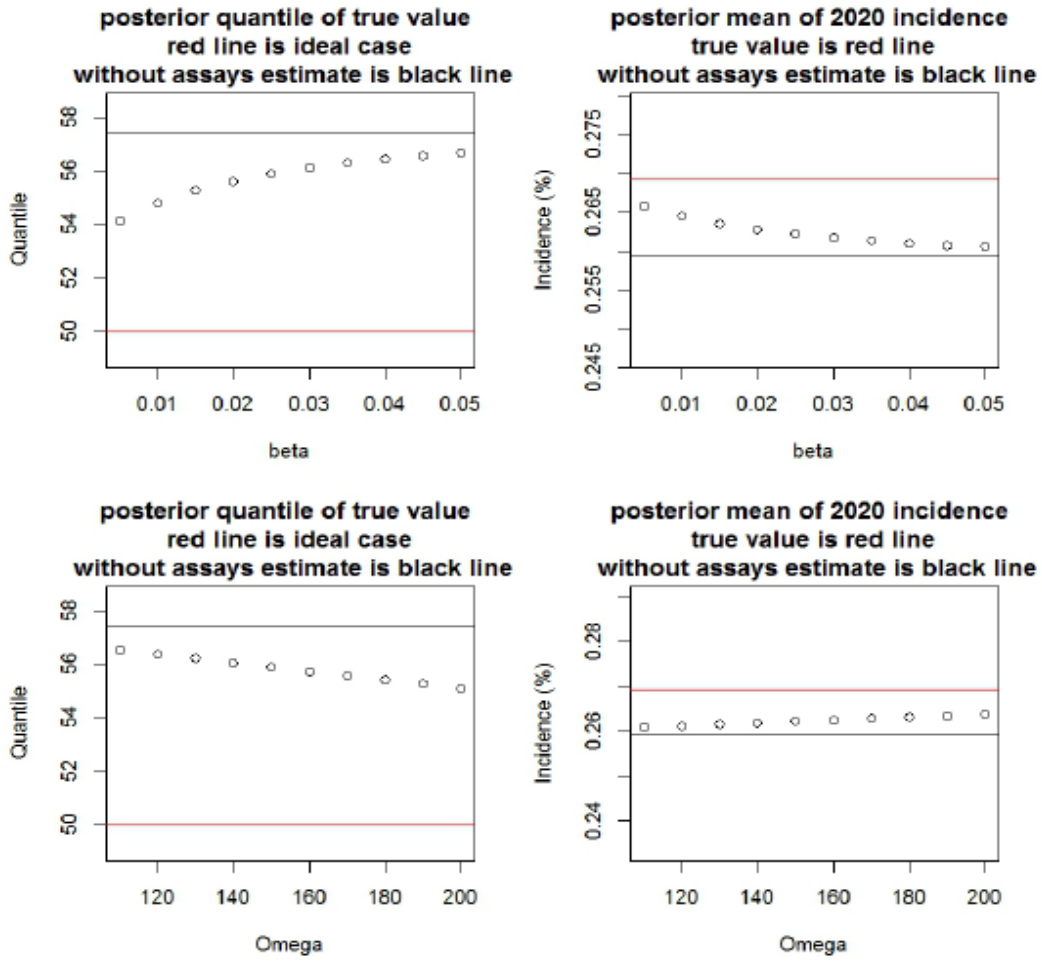


Figure 4.14: Kenya rural, impact to incidence change over time with β changes, for increasing incidence simulation setting.

Kenya Rural. $\theta = 0.01$, $n = 7501$, the incidence raising after 2012. The upper two fix $\Omega = 150$, and allow β to change from 0.005 to 0.05. The lower two fix $\beta = 0.025$, and allow Ω to change from 110 to 150. The left two are the quantile of the true value within these posterior samples, and the right two are the posterior mean of estimated 2020 incidence and true value.

From the Figure 4.13, we can see that as β increases, the quantile is far from the red line, which means more inaccurate estimation. Also, as Ω increases, the estimation tend to be more accurate, which is consistent with the result in previous section. Figure 4.14 shows similar results.

Chapter 5

Estimation of Ultrahigh Dimensional Varying-coefficient Model with Heteroscedastic Error

5.1 Introduction

Let Y be the response variable, U and $\mathbf{X} = (X_1, \ldots, X_p)^T$ be its associated covariates. The varying coefficient model (VCM) assumes that

$$Y = \alpha_1(U)X_1 + \cdots + \alpha_p(U)X_p + \varepsilon, \tag{5.1.1}$$

where α_k's are unknown regression coefficients, and ε is a random error with $E(\varepsilon|\mathbf{X}) = 0$. We set $X_1 \equiv 1$ to include an intercept in the model. This model was systematically studied in Hastie and Tibshirani (1993). Given U, the VCM becomes a linear regression model. And the regression coefficients can be interpreted in a similar way to those in linear regression model. Thus, it has become a very popular nonparametric regression models. A brief review has been given in Section 2.3. This chapter aims to develop an estimation procedure for error variance function.

Error variance estimation plays a critical role in statistical inference such as confidence interval estimation for high dimensional VCM. Due to spurious correlation inherent from model selection in the presence of high dimensional covariates, the traditional mean squared errors (i.e. naive estimator) leads to significant underestimation of the error variance. In this chapter, we utilize the RCV techniques (Fan, Guo, and Hao, 2012) and group LASSO technique to develop a new error variance function estimator. We study the asymptotic property of both naive estimator and the newly proposed estimator.

This chapter is organized as follows. Section 5.2 introduces the model for ultrahigh dimension VCM with heteroscedastic error. Section 5.3 described the three stage naive estimator for the error variance function, along with the theoretical results. Section 5.4 introduces the main method- RCV estimator for the error variance function. A detail results of asymptomatic consistency of RCV estimator is also given. Section 5.5 shows the algorithm to implement group LASSO, which is the variable selection method in this chapter. Then the simulation study and real data example are put in Section 5.6 and 5.7. Section 5.8 is the detailed proofs of the theory along with the regularity conditions.

5.2 Ultrahigh Dimensional VCM with Heteroscedastic Error

For ease of presentation, let us use vector notation. Suppose that $\{U_i, \mathbf{x}_i, y_i\}$, $i = 1, \cdots, n$ is a random sample from the VCM in (5.1.1).

$$y_i = \mathbf{x}_i^T \boldsymbol{\alpha}(U_i) + \varepsilon_i, \tag{5.2.1}$$

where $\boldsymbol{\alpha}(u) = (\alpha_1(u), \cdots, \alpha_p(u))^T$ is the unknown functional coefficient vector. We assume that U has a bounded compact set $\mathcal{U}$ on $\mathbb{R}^1$. Without loss of generality, we set $\mathcal{U} = [s_u, S^U] \subset \mathbb{R}^1$.

In this chapter, we consider the VCM with ultrahigh dimensional covariates. That is, $\log(p) = O(n^\zeta)$, for some $\zeta > 0$. Considering the heteroscedastic random error, this chapter assumes that

$$\mathrm{E}\,(\varepsilon|\mathbf{x}, U) = 0, \quad \text{and} \quad \sigma^2(U) = \mathrm{var}(\varepsilon|\mathbf{x}, U) < \infty. \tag{5.2.2}$$

For simplicity, we rewrite the VCM model as

$$y = \mathbf{x}^T\boldsymbol{\alpha}(U) + \sigma(U)\,\varepsilon, \tag{5.2.3}$$

where, with slightly abuse notation, thenrandom error ε with mean 0 and variance 1 is independent of $\mathbf{x}$ and U. The function $\sigma^2(u)$ defined on $\mathcal{U}$ is called the error variance function. This chapter aims to develop an estimation procedure for $\sigma^2(u)$ in the presence of ultrahigh dimensional covariates.

Sparsity is a commonly-used principle in the analysis of high dimensional data. By sparsity in VCM, many coefficient functions $\alpha_j(\cdot)$ equal 0. That is, $\alpha_j(u) \equiv 0$, for most $j \in \{1, \cdots, p\}$ and any $u \in \mathcal{U}$. Denote $\|f(u)\|_{\ell_2} = (\int_{\mathcal{U}} f^2(u)\,du)^{1/2}$. It is equivalent to, when the k-th derivative of$\alpha_j(u)$ exists for $k \geq 1$,

$$\|\alpha_j(u)\|_{\ell_2} = 0, \quad \text{and} \quad \frac{d^k\alpha_j(u)}{du^k} = 0, \quad \text{for any } u \in \mathcal{U}.$$

Let $\mathcal{S}$ be the active index set, and the cardinality of $\mathcal{S}$ is denoted by $s = |\mathcal{S}|$. Then $\mathcal{S}^c = \{j : \alpha_j(u) \equiv 0, \forall u \in \mathcal{U}, j = 1, \cdots, p\}$, the complement set of $\mathcal{S}$.

Identifying the index set $\mathcal{S}$ is the pivot of the high dimensional statistical analysis. The related high dimensional statistical procedures and statistical computation

attract much attention and are extensively studied in the literature over the last decades. Since the random error ε is unobservable, we have to get the residuals in order to estimate its variance function. This requires us to first identify the index $\mathcal{S}$, and then fit the data with the selected model. In the following sections, we propose multistage statistical procedures to estimate the error variance function.

5.3 Naive error variance function estimator

In this section, we first show the naive estimator, which is defined to be the traditional mean squared errors, leads to nonignorable bias. The naive estimator can be described as a three-stage estimator.

5.3.1 Stage 1: Estimate the Local Active Index Set $\mathcal{S}_u$

We shall use the local linear regression method, although general local polynomial fits are also applicable. The local linear regression has nice properties, such as high statistical efficiency, low computational burden and good boundary behavior (Fan and Gijbels, 1996).

To deal with high dimensionality, we have to select the important independent variables among p candidates in the first stage. In the same spirit of regularization methods for high/ultrahigh linear models, we consider minimizing the constrained least squares which can be written as:

$$\begin{aligned} &\min_{\substack{\alpha_j \in \mathcal{F} \\ j=1,\cdots,p}} \mathrm{E}\,(y - \mathbf{x}^{\mathsf{T}}\boldsymbol{\alpha}(U))^2 \\ &\text{subject to } \sum_{j=1}^{p} \|\alpha_j(u)\|_{\ell_2} \le \eta,\ u \in \mathcal{U}, \end{aligned} \tag{5.3.1}$$

where $\mathcal{F}$ is the functional space in which each function has k, $k > 1$, continuous

derivatives and is squared integrable with respect to the induced probability measure by $\{U, \mathbf{x}, y\}$. The corresponding Lagrangian is

$$\min_{\substack{\alpha_j \in \mathcal{F} \\ j=1,\cdots,p}} \mathrm{E}\,(y - \mathbf{x}^{\mathsf{T}}\boldsymbol{\alpha}(U))^2 + \lambda \sum_{j=1}^{p} \|\alpha_j(u)\|_{\ell_2}\,,\ u \in \mathcal{U}. \tag{5.3.2}$$

Using the empirical measure generated by samples replace the population measure will get the sparse estimators of the functional coefficients. As the functional space $\mathcal{F}$ is an infinite-dimensional space, it is hard to solve the optimization problem (5.3.2). By using Taylor's expansion, $\alpha_j(u) \approx \alpha(u_0) + \alpha_j'(u_0)(u - u_0)$ for u in the small neighbor of u_0. Using the local linear approximation, we can locally parameterize the functional coefficients in (5.3.2), and obtain the local nonparametric estimators as

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_{\mathrm{Loc}}(u)^{\mathsf{T}}, \widehat{\boldsymbol{\alpha}}'_{\mathrm{Loc}}(u)^{\mathsf{T}}h)^{\mathsf{T}} = \underset{a,b\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n} K_h(U_i - u)(y_i - \mathbf{x}_i^{\mathsf{T}}\mathbf{a} - \mathbf{x}_i^{\mathsf{T}}(\frac{U_i - u}{h})\mathbf{b})^2 \\ +\lambda \sum_{j=1}^{p} \sqrt{a_j^2 + b_j^2}, \quad \text{for fixed } u \in \mathcal{U},\end{aligned} \tag{5.3.3}$$

where $K_h(\cdot) = K(\cdot/h)/h$ with a bandwidth h, and $K(\cdot)$ is the kernel function. Equation (5.3.3) is a weighted group ℓ_2 regularization regression (also known as the group LASSO) with the tuning parameter λ that $\lambda \to 0$, as $n \to \infty$ (Yuan and Lin, 2006). One may use other convex or nonconvex penalties to replace the ℓ_2 norm in equation (5.3.3), such as

$$\begin{aligned}(\tilde{\boldsymbol{\alpha}}_{\mathrm{Loc}}(u)^{\mathsf{T}}, \tilde{\boldsymbol{\alpha}}'_{\mathrm{Loc}}(u)^{\mathsf{T}}h)^{\mathsf{T}} = \underset{a,b\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)(y_i - \mathbf{x}_i^{\mathsf{T}}\mathbf{a} - \mathbf{x}_i^{\mathsf{T}}(\frac{U_i - u}{h})\mathbf{b})^2 \\ + \sum_{j=1}^{p} P_\lambda(\sqrt{a_j^2 + b_j^2}), \quad \text{for fixed } u \in \mathcal{U},\end{aligned} \tag{5.3.4}$$

where $P_\lambda(\cdot)$ could be SCAD (Fan and Li, 2001) or MCP (Zhang, 2010) penalty with the tuning parameter λ.

Denote the local active index set at u by $\mathcal{S}_u$, such that $\mathcal{S}_u = \{j : \widehat{\alpha}_{\text{Loc},j}(u) \neq 0, \widehat{\alpha}'_{\text{Loc},j}(u) \neq 0, \exists u \in \mathcal{U}, j \in (1, \cdots, p)\}$. By the properties of the group ℓ_2 regularization regression, the estimators of paired coefficients $\widehat{\alpha}_{\text{Loc},j}(u)$ and $\widehat{\alpha}'_{\text{Loc},j}(u)$ are zero or not zero at the same time. We can fit the selected model only with the independent variables belong to $\mathcal{S}_u$ in the same way as the low dimensional local smoothing technique.

5.3.2 Stage 2: Locally Estimate the Functional Coefficients

In the stage 1, we use group ℓ_2 regularization regression obtain the local active index set $\mathcal{S}_u$. To avoid bias inherited from the group LASSO method, we propose to fit the following local linear regression to estimate the functional coefficients corresponding to $\widehat{\mathcal{S}}_u$ without penalty:

$$(\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_u}(u)^T, \widehat{\boldsymbol{\alpha}}'_{\mathcal{S}_u}(u)^T h)^T = \underset{a,b\in\mathbb{R}^{s_u}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(U_i - u)\Big(y_i - \mathbf{x}_{u,i}^T(\mathbf{a} + \mathbf{b}(\frac{U_i - u}{h}))\Big)^2. \tag{5.3.5}$$

And $s_u = |\mathcal{S}_u|$ is the cardinality of $\mathcal{S}_u$.

The solution of equation (5.3.5) can be considered as a weighted least squares estimator, which has the closed form

$$\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_u,j}(u) = \mathbf{e}_j^T(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u\mathbf{y} \tag{5.3.6}$$

where $\mathbf{W}_u$ is the $n \times n$ diagonal weights matrix with the j-th diagonal element being $K_h(U_j - u)$, $\mathbf{e}_j$ is a $2s_u$-dimensional unit vector with j-th element being 1,

and the design matrix

$$\mathbf{D}_u = \begin{pmatrix} \mathbf{x}_{u,1}^T & \frac{U_1-u}{h}\mathbf{x}_{u,1}^T \\ \vdots & \vdots \\ \mathbf{x}_{u,n}^T & \frac{U_n-u}{h}\mathbf{x}_{u,n}^T \end{pmatrix}. \tag{5.3.7}$$

The residuals are obtained by

$$\widehat{\varepsilon}_i(u) = y_i - \widehat{y}_i = y_i - \mathbf{x}_{u,i}^T \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_u}(u),\ i = 1, \cdots, n. \tag{5.3.8}$$

By localizing the residuals around any fixed $u \in \mathcal{U}$, $\sigma^2(u)$ is approximately a constant and the mean squared errors is a well-performed estimator of $\sigma^2(u)$.

5.3.3 Stage 3: Local Estimator of Varying-error-variance Function

In the stage 2, we obtain the residuals $\{\widehat{\varepsilon}_1(U_1), \cdots, \widehat{\varepsilon}_n(U_n)\}$ by (5.3.8). The varying-error-variance function $\sigma^2(u)$, $u \in \mathcal{U}$, can be locally estimated by the weighted mean square errors. The kernel estimator or local constant estimator is defined by

$$\widehat{\sigma}^2(u) = \frac{1}{c(u)} \sum_{i=1}^{n} K_{h^*}(U_i - u)\, \widehat{\varepsilon}_i^2(U_i). \tag{5.3.9}$$

where $c(u)$ is used to obtain the nearly unbiased estimator of $\sigma^2(u)$. We will detailed discuss it later. The bandwidth h^* may be different from the bandwidth h used in the previous stages to achieve a better estimate. Furthermore, the local linear regression also can be applied to estimate $\sigma^2(u)$ as

$$(\widehat{\sigma}^2(u), \widehat{\sigma}^{2'}(u)) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} K_{h^*}(U_i - u)\Big(\widehat{\varepsilon}_i^2(U_i) - \alpha - \beta(U_i - u)\Big)^2. \tag{5.3.10}$$

This method is proposed by Fan and Yao (1998). In this paper, we just study the estimator (5.3.9) for simplicity. The extension to local linear regression involves no fundamentally new ideas.

We call estimator (5.3.9) three-stage local variance estimator. The three stages include the first stage that reducing the dimensionality by group ℓ_2 regularization regression, the second stage that locally fitting the selected model and obtaining the residuals, and the third stage that locally estimate the variance function.

However, due to existence of spurious correlation in high dimensional data, lots of redundant variables may be selected to fit the model. It will lead to an underestimate of variance by the mean squared errors. In the next section, we will discuss this problem and propose new statistical procedure to calibrate the the high/ultrahigh dimensional multistage local variance estimator.

5.3.4 Theoretical properties of Naive Estimator

In this subsection, we study the theoretical properties of the multistage variance estimator $\widehat{\sigma}^2(u)$ for $u \in \mathcal{U}$. All regularity conditions are provided in Section 5.8.1.

In the classic parametric linear model, the unbiased variance estimator is

$$\widehat{\sigma}^2_{\text{LS}} = \frac{1}{n-s}\mathbf{y}^{\mathsf{T}}(\mathbf{I}-\mathbf{P})\mathbf{y}, \tag{5.3.11}$$

where $\mathbf{P}$ is the corresponding $n \times n$ projection matrix. The estimator is calibrated by replacing n with $n-s$, where s is the number of predictors used to fit the model. By Lemma 5 in Section 5.8.2, the weighted mean squared errors should be divided by $c_j(u) \triangleq \text{tr}((\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u(\mathbf{I}-\mathbf{S}))$ instead of n to make the variance estimator nearly unbiased. The following proposition shows the magnitude of the correction.

Proposition 1 Denote by $c(u) = \text{tr}((\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u(\mathbf{I}-\mathbf{S}))$, for any given $u \in \mathcal{U}$.

Under regularity conditions (C1)-(C10) in Section 5.8.1, it follows that, for any $u \in \mathcal{U}$,

$$f_U(u)\, n - 2\frac{K(0)}{h}\, s_u \le c(u) \le f_U(u)\, n - 2\frac{K(0)}{h}\, s_u + 2\frac{K(0)}{h}\, s_u^2 \tag{5.3.12}$$

Remark The leading term of upper and lower bound of $c(u)$ is the same as the effective number of local polynomial regression in Fan and Chen (1999). Because of the local fitting, the effective number is decide by the location $f_U(u)$. However, under the high-dimensional settings, the bounds are also related with the diverging dimension.

For the special case,

$$\begin{aligned} &\operatorname{tr}\left(\mathbf{W}_u - (\mathbf{D}_u^T \mathbf{W}_u \mathbf{D}_u)^{-1} \mathbf{D}_u^T \mathbf{W}^2 \mathbf{D}_u\right) \\ &= \operatorname{tr}(\mathbf{W}_u(\mathbf{I} - \mathbf{W}_u^{1/2}\mathbf{D}_u(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u^{1/2})). \end{aligned} \tag{5.3.13}$$

Denote by $\mathbf{P}_u = \{\mathbf{P}_{ij}(u)\}_{i,j} = \mathbf{W}_u^{1/2}\mathbf{D}_u(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u^{1/2}$. Notice that $\mathbf{P}_u$ is a projection matrix satisfying that $\operatorname{rank}(\mathbf{P}_u) = \operatorname{tr}(\mathbf{P}_u)$. For $\mathbf{W}_u = \operatorname{diag}(w_{n,1}, \cdots, w_{n,n})$, the i-th diagonal element is $w_{u,i}(1 - P_{ii})$,

$$\begin{aligned} f_U(u)n - \frac{K(0)}{h}s_u \le \sum_{i=1}^n w_{u,i} - \sum_{i=1}^n \mathbf{P}_{ii}(u) &\le \sum_{i=1}^n w_{u,i}(1 - \mathbf{P}_{ii}(u)) \\ &\le \frac{K(0)}{h}\sum_{i=1}^n (1 - \mathbf{P}_{ii}(u)) = \frac{K(0)}{h}\,(n - s_u). \end{aligned} \tag{5.3.14}$$

The result directly shows the difference of corrections between the ordinary least squares and the weighted least squares.

We construct the large sample properties of traditional two-stage local estimate under high-dimensional settings. Lots of redundant predictors may lead an biased estimator. Using the aforementioned notations, we have the following theorem.

Theorem 1 Suppose Condition (C1)-(C10) presented in Section 5.8.1 holds, and $\mathcal{S}_u^* \subset \mathcal{S}_u, s_u \leq s_n = o(n),\ i = 1, \cdots, n$, the bias term $\text{bias}_{\text{dim}} = \boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon}$ due to the diverging dimension follows that

$$\widehat{\gamma}_n^2 = \frac{\boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon}} = O_p\left(\frac{s_u \log(p)}{nh^2 h^*}\right), \tag{5.3.15}$$

where $\mathbf{W}_u^*$ respect to the bandwidth h^* in the third stage. Specially, when $\log(p)/n \to C_0 > 0$, term $\boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon}$ can not be ignored.

Theorem 2 Suppose Condition (C1)-(C10) presented in Section 5.8.1 holds, and $\mathcal{S}_u^* \subset \mathcal{S}_u, s_u \leq s_n = o(n),\ i = 1, \ldots, n$, where $\mathcal{S}_u^*$ is the true active index set, the nonparametric multistage local variance estimator defined in (5.3.9)

$$\widehat{\sigma}^2(u) = \frac{\widehat{\boldsymbol{\varepsilon}}^T \mathbf{W}_u^* \widehat{\boldsymbol{\varepsilon}}}{c(u)} = \frac{\mathbf{Y}^T (\mathbf{I} - \mathbf{S})^T \mathbf{W}_u^* (\mathbf{I} - \mathbf{S})\mathbf{Y}}{c(u)}, \text{ for } u \in \mathcal{U}, \tag{5.3.16}$$

follows the asymptotic normality

$$\sqrt{nh}\left(\widehat{\sigma}^2(u) - \sigma^2(u) - \text{bias}_{\text{n-para}} - \text{bias}_{\text{dim}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa(u)\nu_0}{f_U(u)}\right), \tag{5.3.17}$$

where the term

$$\text{bias}_{\text{n-para}} = \frac{h^2 \mu}{f_U(u)}\left((\sigma^2(u))' f_U'(u)) + \frac{1}{2}(\sigma^2(u))''\right)$$

stands for the bias due to the the nonparametric fitting and the term bias_{dim} defined in the theorem 1 stands for the bias from redundant predictors. Also, $\nu_0 = \int K^2(u)du$, $\kappa(u) = E(\varepsilon^4|u)$.

Theorem 1 and 2 show that many redundant predictors and diverging dimension can lead a non-negligible bias to variance estimator at each $u \in \mathcal{U}$.

5.4 RCV Variance Function Estimator

Spurious correlation is a common phenomenon in high/ultrahigh dimensional data analysis (Fan, Guo, and Hao, 2012). It severely influences the performance of variable selection procedures or independent sure screening procedures and leads lots of redundant variables into the fitted model. The intuitive interpretation of geometry is roughly that when scattering p points on n-dimensional space, there exist at most n orthogonal vectors in n-dimensional vector space. If $p \gg n$, it is very likely that there is a small angle between two vectors. In addition, to avoid missing any important variables, the tuning parameter λ usually is set very small, which also results in far more variables being selected into the fitted model than expected. In this case, using the mean squared errors as the estimator will underestimate the error variances. To overcome the defect and calibrate the variance estimator, we propose combining the refitted cross-validation (RCV) procedure (Fan, Guo, and Hao, 2012) with the high/ultrahigh dimensional multistage local variance estimator defined in previous section.

The refitted cross-validation procedure is to randomly split the samples into to two parts with equal sizes. Denote by $(y_i^{(j)}, \mathbf{x}_i^{(j)}, U_i^{(j)})$, $i \in \mathcal{I}_j, \mathcal{I}_j = \{i : i \in (1, \cdots, n)\}, |\mathcal{I}_j| = n/2, j = 1, 2$. For j-th group and any fixed u, we get the local active index set $\mathcal{S}_u^{(j)}$ via the group ℓ_2 regularization regression (5.3.3). Denote by $s_{u,j} = \left|\mathcal{S}_u^{(j)}\right|$ and $\mathbf{X}_{\mathcal{S}_u^{(j)}}^{(3-j)} = \{X_{i,j}\}$, $i \in \mathcal{I}_{3-j}, j \in \mathcal{S}_u^{(j)}$. Then we locally estimate the functional coefficients in the $\mathcal{S}_u^{(j)}$ by samples $(\mathbf{y}^{(3-j)}, \mathbf{X}_{\mathcal{S}_u^{(j)}}^{(3-j)}, U^{(3-j)})$. The local linear estimators of functional coefficients has the closed form

$$\widehat{\alpha}_k^{(3-j)}(u) = \mathbf{e}_k^T (\mathbf{D}_u^{(3-j)^T} \mathbf{W}_u^{(3-j)} \mathbf{D}_u^{(3-j)})^{-1} \mathbf{D}_u^{(3-j)^T} \mathbf{W}_u^{(3-j)} \mathbf{y}^{(3-j)}, \ j = 1, 2. \tag{5.4.1}$$

where the design matrix is $\mathbf{D}_u^{(3-j)} = (\mathbf{X}_{\mathcal{S}_u^{(j)}}^{(3-j)}, \mathrm{diag}((U_1^{(3-j)} - u)/h, \cdots, (U_{n/2}^{(3-j)} -$

$u)/h)\mathbf{X}^{(3-j)}_{\mathcal{S}_u^{(j)}})$. By (5.3.8), the residuals are

$$\widehat{\varepsilon}_i^{(3-j)}(U_i^{(3-j)}) = y_i^{(3-j)} - \mathbf{x}_{i,\mathcal{S}_u^{(j)}}^{(3-j)^T}\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_u^{(j)}}^{(3-j)}(U_i^{(3-j)}),\ j = 1, 2.$$

Define smoothing matrix

$$\mathbf{S}^{(3-j)} = \begin{pmatrix} (\mathbf{x}_{1,\mathcal{S}_u^{(j)}}^{(3-j)^T}, 0)(\mathbf{D}_{u_1}^{(3-j)^T}\mathbf{W}_{u_1}^{(3-j)}\mathbf{D}_{u_1}^{(3-j)})^{-1}\mathbf{D}_{u_1}^{(3-j)^T}\mathbf{W}_{u_1}^{(3-j)} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{X}_{n/2,\mathcal{S}_u^{(j)}}^{(3-j)^T}, 0)(\mathbf{D}_{u_{n/2}}^{(3-j)^T}\mathbf{W}_{u_{n/2}}^{(3-j)}\mathbf{D}_{u_{n/2}}^{(3-j)})^{-1}\mathbf{D}_{u_{n/2}}^{(3-j)^T}\mathbf{W}_{u_{n/2}}^{(3-j)} \end{pmatrix}. \tag{5.4.2}$$

The residuals can be expressed as a matrix form: $\widehat{\boldsymbol{\varepsilon}}^{(3-j)} = (\mathbf{I} - \mathbf{S}^{(3-j)})\mathbf{y}^{(3-j)}$. For any given u, the local error variance estimator is defined by

$$\widehat{\sigma}_j^2(u) = \frac{\widehat{\boldsymbol{\varepsilon}}^{(j)^T}\mathbf{W}_u^{(j)}\widehat{\boldsymbol{\varepsilon}}^{(j)}}{c(u)^{(j)}} = \frac{\mathbf{y}^{(j)^T}(\mathbf{I} - \mathbf{S}^{(j)})^T\mathbf{W}_u^{(j)}(\mathbf{I} - \mathbf{S}^{(j)})\mathbf{y}^{(j)}}{c(u)^{(j)}},\ j = 1, 2, \tag{5.4.3}$$

where $c(u)^{(j)} = \mathrm{tr}((\mathbf{I} - \mathbf{S}^{(j)})^T\mathbf{W}_u^{(j)}(\mathbf{I} - \mathbf{S}^{(j)}))$ to make the estimator nearly unbiased. Recall that the bandwidth used in $\mathbf{W}_u^{(j)}$ is different from that used in the fitting process (mentioned in Section 5.3.3). Finally, we obtain a new estimator

$$\widehat{\sigma}^2_{\mathrm{RCV}}(u) = \frac{\widehat{\sigma}_1^2(u) + \widehat{\sigma}_2^2(u)}{2},\ \text{for any } u \in \mathcal{U}. \tag{5.4.4}$$

This estimator is very similar to the other two respectively proposed in Fan, Guo, and Hao (2012) for linear regression models and Chen, Fan, and Li (2016) for additive models. However, it is a nonparametric estimator for varying-error-variance rather than the other two parametric estimators. Consequently, it is much more challenge in computation and establishing the asymptotic properties. In the following sections, we will further discuss these two problems. The procedure of refitted cross-validation (RCV) is illustrated schematically in Figure (5.1).

In this section, we study the theoretical properties of the RCV varaince estimator

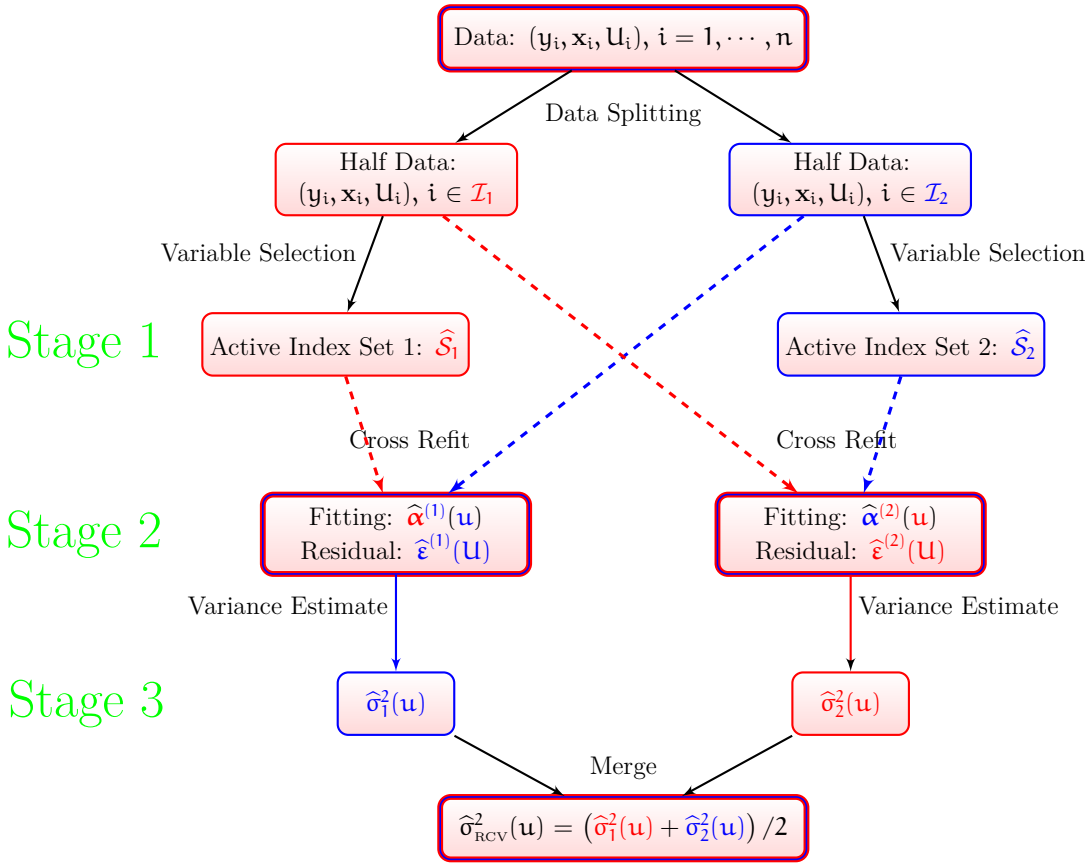


Figure 5.1: The flowchart of the multistage RCV variance estimate

$\widehat{\sigma}^2_{\text{RCV}}(u)$, for $u \in \mathcal{U}$. All regularity conditions are provided in Section 5.8.1.

Theorem 3 Suppose Conditions (C1)-(C10) in Section 5.8.1, and $\mathcal{S}_u^* \subset \mathcal{S}_u^{(j)}$, $s_{u,j} \leq s_{n,j} = o(n)$, $i = 1, \cdots, n$, $j = 1, 2$, where $\mathcal{S}_u^*$ is the true active index set, the RCV nonparametric variance estimator defined in (5.4.4) follows the asymptotic normality

$$\sqrt{nh}\left(\widehat{\sigma}^2_{\text{RCV}}(u) - \sigma^2(u) - \text{bias}_{\text{n-para}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa(u)\nu_0}{f_U(u)}\right). \tag{5.4.5}$$

Theorem 3 shows that the RCV nonparametric variance estimator can completely eliminate the negative effects of many redundant variables in terms of asymptotic consistency.

5.5 Statistical computation and implementation issues

Statistical computation is always a huge challenge for high/ultrahigh dimensional data analysis. Since the objective functions of regularization regression are usually not smooth and the dimension of data is very high, the computational instability and burden are much larger than the common cases. The fast and efficient algorithms are the goal that everyone pursues. Coordinate descent (CD) algorithm have been considered as an efficient and fast algorithm for the ℓ_1 regularization regression (LASSO) applied to high dimensional linear models (Wu and Lange, 2008; Friedman, Hastie, and Tibshirani, 2010). Naturally, we introduce the blocked coordinate descent (BCD) algorithm to deal with the newly proposed ℓ_2 group regularization regression defined by (5.3.3) in section 5.3.1.

For existence of the explicit solution of 1-dimensional ℓ_1 regularization regression (LASSO), the coordinate descent algorithm can be applied to solve the multiple or high-dimensional ℓ_1 regularization regression. However, because the design matrix is not orthogonal, there is no explicit solution for each sub-iteration of ℓ_2 regularization regression (Yuan and Lin, 2006). To overcome such problem, Qin, Scheinberg, and Goldfarb (2013) suggest using the Newton's method for marginal regression and construct the blocked coordinate descent algorithm. We employ the blocked coordinate descent algorithm (Qin, Scheinberg, and Goldfarb, 2013) to solve (5.3.3) that is a weighted least squares with ℓ_2 penalty. At each sub-iteration, we need to solve that

$$\min_{a_l, b_l \in \mathbb{R}^1} \frac{1}{2} \sum_{i=1}^{n} w_i \left(r_{il} - a_l X_{il} - b_l X_{il}(U_i - u_0)\right)^2 + \lambda \sqrt{a_l^2 + b_l^2}, \tag{5.5.1}$$

where $r_{il} = y_i - \sum_{j \neq l} \left(a_j X_{ij} + b_j X_{ij}(U_i - u_0)\right)$ is the residual of the previous iteration.

Denote by $\mathbf{r}_l = (r_{1l}, \cdots, r_{nl})^T$, $\mathbf{A}_{il} = (X_{il}, X_{il}(u_0 - U_i))^T$, $\mathbf{A}_l = (\mathbf{A}_{1l}, \cdots, \mathbf{A}_{nl})^T$, $\mathbf{W} = \mathrm{diag}(w_1, \cdots, w_n)$, and $\boldsymbol{\alpha}_l = (a_l, b_l)^T$, respectively. We rewrite the problem in matrix form

$$\begin{aligned} &\min_{\alpha_l \in \mathbb{R}^2} \frac{1}{2} \boldsymbol{\alpha}_l^T \mathbf{A}_l^T \mathbf{W} \mathbf{A}_l \boldsymbol{\alpha}_l - \mathbf{r}_l^T \mathbf{A}_l \boldsymbol{\alpha}_l + \lambda \left\| \boldsymbol{\alpha}_l \right\|_2 \\ \triangleq &\min_{\alpha_l \in \mathbb{R}^2} \frac{1}{2} \boldsymbol{\alpha}_l^T \mathbf{M}_l \boldsymbol{\alpha}_l - \mathbf{p}_l^T \boldsymbol{\alpha}_l + \lambda \left\| \boldsymbol{\alpha}_l \right\|_2 . \end{aligned} \tag{5.5.2}$$

Hence, $\boldsymbol{\alpha}_l = 0$ is the optimal solution of (5.5.2) if and only if $\left\| \mathbf{p}_l \right\|_2 \leq \lambda$. For the case that the solution $\boldsymbol{\alpha}_l^* \neq 0$, there exists a constant $\Delta > 0$ such that (5.5.2) is equivalent to

$$\begin{aligned} &\min_{\alpha_l} \frac{1}{2} \boldsymbol{\alpha}_l^T \mathbf{M}_l \boldsymbol{\alpha}_l - \mathbf{p}_l^T \boldsymbol{\alpha}_l, \text{ subject to } \left\| \boldsymbol{\alpha}_l \right\|_2 \leq \Delta, \\ &(\mathbf{M}_l + (\lambda / \left\| \boldsymbol{\alpha}_l \right\|_2) \mathbf{I}) \, \boldsymbol{\alpha}_l = \mathbf{p}_l. \end{aligned} \tag{5.5.3}$$

Thus, the unique solution $\boldsymbol{\alpha}_l^*$ (by the convexity) satisfies that

$$\left\| \boldsymbol{\alpha}_l^* \right\|_2 = \Delta, \ \boldsymbol{\alpha}_l^* = (\mathbf{M}_l + (\lambda/\Delta)\mathbf{I})^{-1} \mathbf{p}_l,$$

which can be represented as a function of Δ that $\boldsymbol{\alpha}_l^* = \Delta z_l(\Delta)$ with

$$z_l(\Delta) = (\Delta \mathbf{M}_l + \lambda \mathbf{I})^{-1} \mathbf{p}_l. \tag{5.5.4}$$

Obviously, $\left\| z_l(\Delta) \right\|_2 = 1$. Using the eigenvalue decomposition of $\mathbf{M}_l$,

$$\mathbf{M}_l = \gamma_1 \mathbf{q}_1 \mathbf{q}_1^T + \gamma_2 \mathbf{q}_2 \mathbf{q}_2^T, \tag{5.5.5}$$

where γ_i's and $\mathbf{q}_i$'s are the eigenvalues and the corresponding eigenvectors, it follows that

$$\left\| z_l(\Delta) \right\|_2^2 = (\frac{\mathbf{q}_1^T \mathbf{p}_l}{\gamma_1 \Delta + \lambda})^2 + (\frac{\mathbf{q}_2^T \mathbf{p}_l}{\gamma_2 \Delta + \lambda})^2.$$

Then applying the Newton's root finding method to solve

$$f(\Delta) = 1 - 1/\left\|z_l(\Delta)\right\|_2 . \tag{5.5.6}$$

The derivative is obtained by

$$\frac{d}{d\Delta}\frac{1}{\|z_l(\Delta)\|_2} = -\frac{1}{2}(\|z_l(\Delta)\|_2^2)^{-3/2}\frac{d}{d\Delta}\|z_l(\Delta)\|_2^2, \tag{5.5.7}$$

$$\frac{d}{d\Delta}\|z_l(\Delta)\|_2^2 = 2\sum_{i=1}^{2}(\frac{\mathbf{q}_i^T\mathbf{p}_l}{\gamma_i\Delta+\lambda})^2\frac{\gamma_i}{\gamma_i\Delta+\lambda}. \tag{5.5.8}$$

The pseudo code is listed in the following table:

Algorithm 5.5.1. *Extension of BCD-GL to local linear regression group LASSO*

Algorithm 1 BCD for (5.3.4)

Given initial value $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{2p}$ and λ. Set $k = 1$, compute $\mathbf{M}_l = \mathbf{A}_l^T\mathbf{W}\mathbf{A}_l$, and eigenvalue decomposition (5.5.5), for $l = 1, \cdots, p$.
repeat
  $\boldsymbol{\alpha} \Leftarrow \boldsymbol{\alpha}^{(k-1)}$
  for $l = 1, \cdots, p$ **do**
    compute $\mathbf{p}_l$ by (5.5.2)
    if $\|\mathbf{p}_l\|_2 \le \lambda$ **then**
      $\boldsymbol{\alpha}_l = (0, 0)^T$
    else
      compute the derivative by (5.5.7) and (5.5.8).
      find the root Δ of $f(\Delta) = 1 - 1/\|z_l(\Delta)\|_2$, using Newton's method.
      compute $z_l(\Delta)$ by (5.5.4).
      $\boldsymbol{\alpha}_l \Leftarrow \Delta z_l(\Delta)$.
    end if
  end for
  $\boldsymbol{\alpha}^{(k)} \Leftarrow \boldsymbol{\alpha}$, $k \Leftarrow k + 1$
until $k =$ maxIter or $\boldsymbol{\alpha}^{(k)}$ satisfies the stopping rule

---

5.6 Numerical studies

In the simulation studies, we examine the finite sample performance of the newly proposed multistage variance estimate and the RCV variance estimate by Monte Carlo simulation. All numerical studies were conducted by MATLAB code.

In the simulation studies, we use the oracle estimator as the benchmark. The oracle estimator is obtained by directly using the local linear regression method to fit the true model. In the stage 1, the variable selection stage, we use two different strategies. First, we locally apply the group ℓ_2 regularization regression at each given u. Second, we use CC-SIS procedure to select the significant functional coefficients over the entire interval $\mathcal{U}$. Therefore, for two different u_1 and u_2, the associated active index sets obtained by the first strategy may be different. However, the active index sets obtained by the second strategy are totally the same. In the stage 2, we locally estimate the selected functional coefficients and get the associated residuals at each U_i, $i = 1, \cdots, n$. In the stage 3, we carry out the pilot study and obtain the curve of variance function. To avoid the boundary issues of the local linear regression, the grid points are restricted by $0.05 \leq u \leq 0.95$.

We compare four different methods: Oracle(oracle), naive group LASSO(naive), group LASSO refitted cross-validation (GL-RCV), and conditional correlation sure independence screening refitted cross-validation(CC-SIS-RCV). The acronym in the parentheses are used in the tables of results and stand for the different methods. The oracle estimator is used as the benchmark. GL-RCV applies the local strategy in the variable selection stage and CC-SIS-RCV applies the global strategy. In our simulation, we use generalized cross-validation (GCV) to choose the tuning parameter λ is the group LASSO.

5.6.1 Bandwidth

Bandwidth selection is a big issue in nonparametric statistical procedures. Considering the use of multistage nonparametric estimate, it is more complicated for the proposed procedures. Theoretically, the three bandwidths using in the three stages (variable selection, refitting, error estimation) are not necessarily to be the same. The bandwidths in different methods (orale, naive, GL-RCV, CC-SIS-RCV) are also not necessarily to be the same. Therefore, there are 12 bandwidth combinations to be considered. We use the same batch of bandwidths for the oracle and the naive methods, and same for two RCV methods in order to simply this problem. Meanwhile, we use the same bandwidths at both variable selection stage and refitting stage for each procedure. In the simulation studies, we use mean squared errors of functional coefficients to choose the optimal the bandwidth h. It is defined as

$$\text{MSE}(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(\widehat{\alpha}_j(U_i) - \alpha_j(U_i))^2. \tag{5.6.1}$$

And we conduct a pilot study to select the optimal bandwidth,

$$h_{opt} = \text{argmin}_h \text{MSE}(h). \tag{5.6.2}$$

The following plot shows the pilot study results for the oracle estimate. From Figure ??, the minimum MSE is approached near bandwidth 0.2. So, we choose 0.2 as optimal bandwidth. Also, we can see the MSE curve decreases rapidly when bandwidth smaller than 0.2, and increases very slowly when bandwidth larger than 0.2, So, we also consider $h = 0.15$ for undersmoothing and $h = 0.3$ for oversmoothing for variable selection and refit steps.

For the RCV methods, since the sample size reduces to $n/2$, by using the bandwidth calculator $h_2/h_1 = (n_1/n_2)^{1/5}$, we obtain the new bandwidths batch $(0.18, 0.23, 0.35)$.

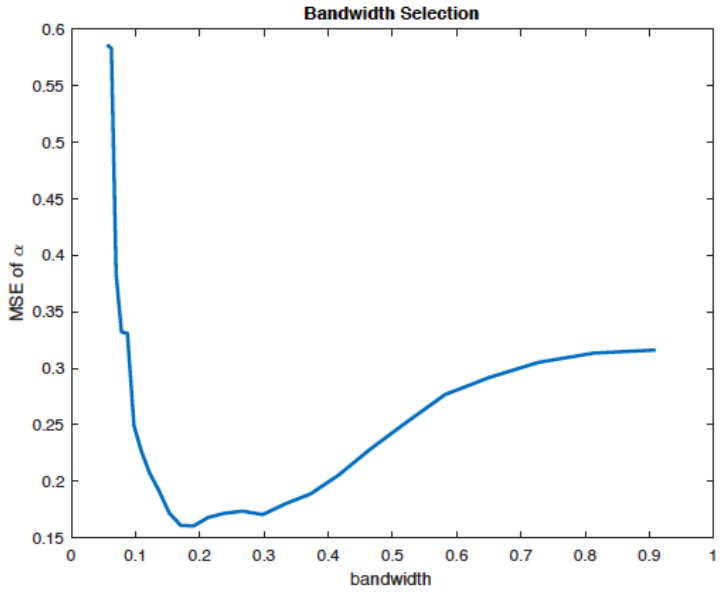


Figure 5.2: The MSE plot for varying error variance, and SNR = 2.

Next we choose the bandwidths for kernel method used to estimate $\widehat{\sigma}^2(u)$. According to Fan and Zhang (2000), it should be slightly small than the bandwidth used in the model fitting. Thus, we choose $(0.1, 0.15, 0.2)$ for the oracle and the naive methods, and $(0.15, 0.2, 0.3)$ for RCV methods.

method	bandwidth for variable selection and refitting	bandwidth for calculating $\widehat{\sigma}^2(u)$
oracle	$(0.15, 0.2, 0.3)$	$(0.1, 0.15, 0.25)$
naive	$(0.15, 0.2, 0.3)$	$(0.1, 0.15, 0.25)$
GL-RCV	$(0.18, 0.23, 0.35)$	$(0.15, 0.20, 0.25)$
CC-SIS-RCV	$(0.18, 0.23, 0.35)$	$(0.15, 0.20, 0.25)$

Table 5.1: RCV Simulation Settings

5.6.2 Simulation settings

The random samples $\{U_i, y_i, x_i\}$ of size $n = 400$, is generated from

$$y_i = \beta \left(\sum_{j=1}^{p} \alpha_j(U_i) X_{ij} \right) + \varepsilon_i, \quad i = 1, \cdots n.$$

where the covariates $\{\mathbf{x}_i\}$ follow multinormal distribution with mean vector zero and covariance matrix $\Sigma = \{\sigma_{ij} = 0.2^{|i-j|}\}$. The dimension of $\mathbf{x}$ is set to $p = 200$. Random errors $\{\varepsilon_i\}$ follow normal distribution with mean 0 and two variance settings: (1) the constant error variance $\sigma^2 = 1$; (2) the varying error variance $\sigma^2(u) = 1 + 0.125\sin(2\pi(u + 0.125))$. The random variable $\{U_i\}$ in functional coefficients $\alpha_j(u)$ comes from uniform distribution Unif($[0, 1]$). The nonzero functional coefficients are respectively

$$\alpha_1(u) = -2u^2 + 2u,\ \ \alpha_2(u) = 0.5\sin(2\pi u),\ \ \alpha_5(u) = 0.5 - 0.5u. \tag{5.6.3}$$

They has the similar degree of smoothness and associate with the covariates X_1, X_2 and X_5. β is used to control the signal-to-noise ratio (SNR). Three different SNR settings $0, 1, 2$ represent white noise, weak signals, and strong signals, respectively. The number of replications for each case is 600.

5.6.3 Simulation Results

From the settings above, we have six scenarios in total with respect to three kinds of SNR settings and two kinds of error variance settings. Also, within each scenario, we compare the undersmoothing, optimal, oversmoothing bandwidth. The detailed bandwidth settings are shown in Table 5.1. For small bandwidths, the effective number would be relatively small, so the it often appears singular problem. We choose to use ridge regression to avoid singularity.

We start with the easiest case, which is constant error variance. Figure 5.3 shows the estimates when SNR$= 0$, which represents the null model. The solid black line is the simulation settings or error variance, which we consider as the true value. The solid and dash colorful lines are mean estimates for each method along with their 95% credible intervals (within 600 replications). All four estimates lines

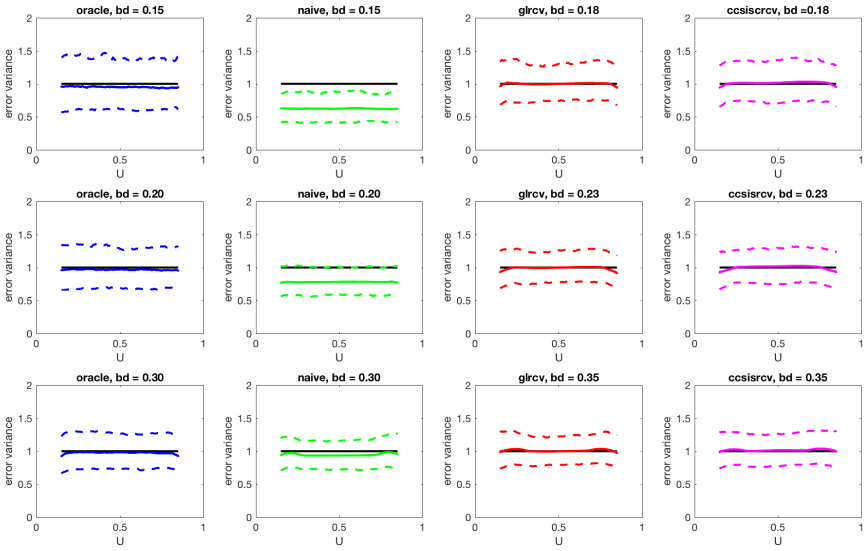


Figure 5.3: SNR = 0, constant error variance.

Error variance estimates for the compared four methods. The black lines are true error variance. The bold lines are mean estimates for each method, and the dash lines show 95% credible intervals.

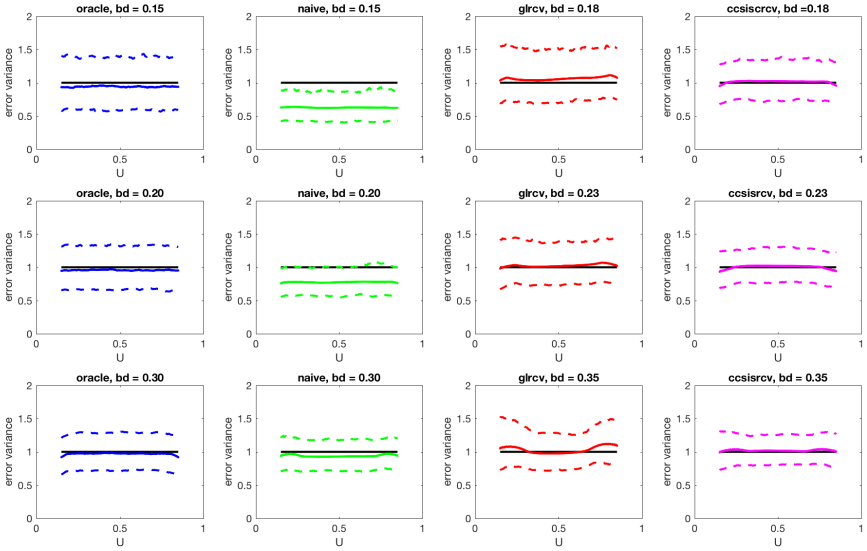


Figure 5.4: SNR = 1, constant error variance.

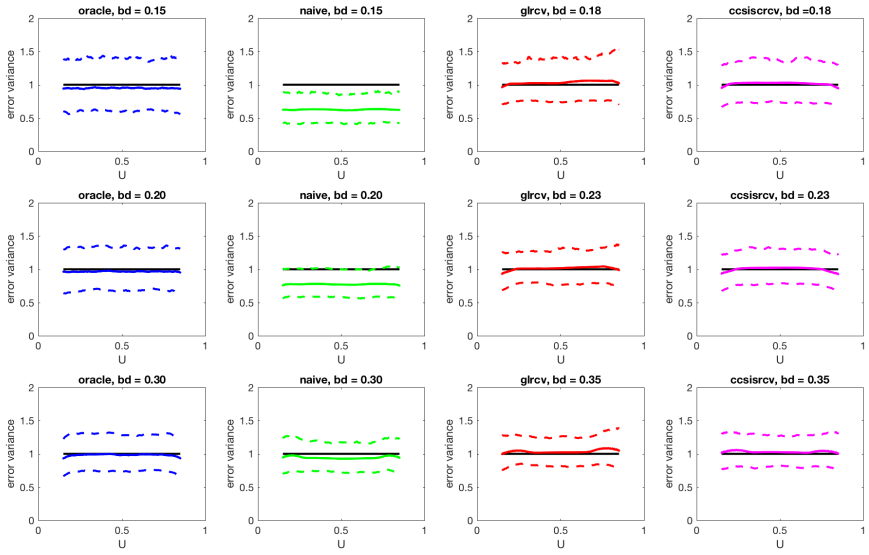


Figure 5.5: SNR = 2, constant error variance.

have a relative close shape compared to true value. However, for naive method, It has a clear underestimates due to the spurious correlation. Especially with small and regular bandwidths, the credible intervals can hardly cover the true lines. RCV method clearly has a better performance. GL-RCV and CC-SIS-RCV, they both fits the line well even compared to oracle method. Figure 5.4 shows the estimates when SNR= 1. It is the "hardest" scenario for all the methods, since it represents the situation that signal is similar as noise. The pattern is similar as Figure 5.3. However, GL-RCV performances not as well as the scenario when SNR= 0. Even so, RCV method still much better than naive. Figure 5.5 shows the estimates when SNR= 2, which represents the situation that signal is strong. The pattern shown in this plot is similar as previous two figures. Also, for oracle and two RCV methods, the estimates did not change too much as bandwidth change. However, for naive method, there is a clear increasing trend of estimates as bandwidth increases. Also, the boundary problem gets more clear as the bandwidth grows.

Next we explore the performance of four methods on varying error variance

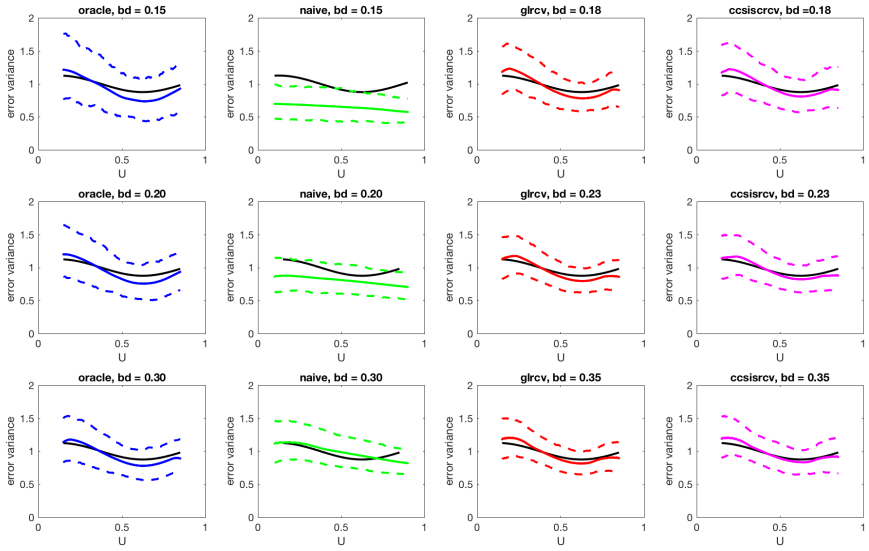


Figure 5.6: SNR = 0, varying error variance.

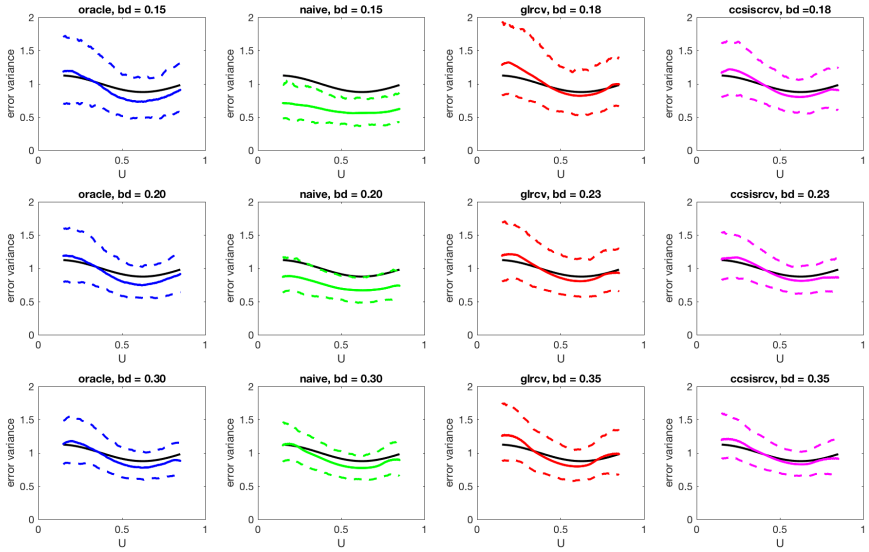


Figure 5.7: SNR = 1, varying error variance.

model. The patterns are similar to constant error variance case, however, the Figure 5.6 shows that for naive method, the three mean estimates lines cannot capture the pattern of true values. It has a clear underestimates. RCV method clear has a

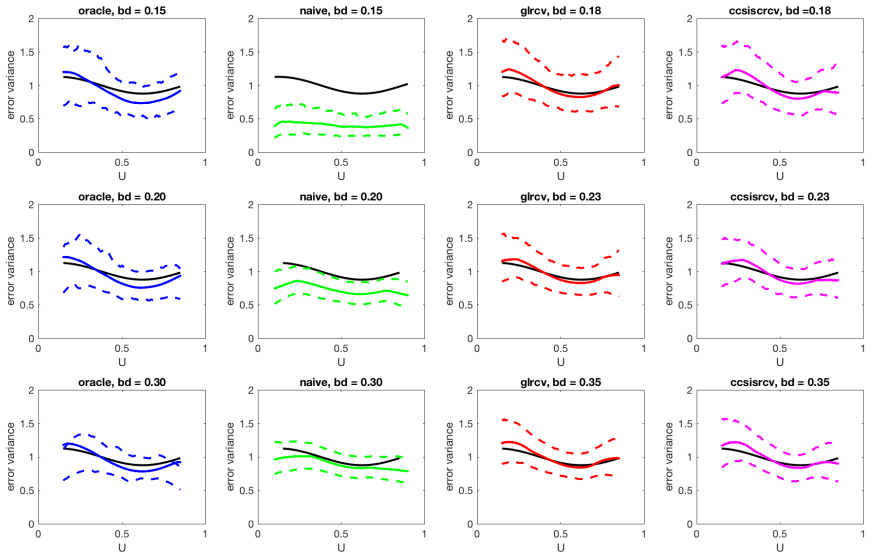


Figure 5.8: SNR = 2, varying error variance.

better performance. GL-RCV and CC-SIS-RCV, they both fits the line well. For varying error variance, the advantage of RCV method is obvious.

We also construct the 95% confidence intervals based on the standard deviation and the true value. So, the confidence intervals becomes:

$$[\sigma^2_{\text{method}} - 1.96\text{sd}(\widehat{\sigma}_{\text{method}}), \sigma^2_{\text{method}} + 1.96\text{sd}(\widehat{\sigma}_{\text{method}})] \tag{5.6.4}$$

And we calculate the probability of the 600 replication falls into this interval. The average probability are summarized in Table 5.2 following:

It is obviously that the estimates with probability closer to 95% has a better performance. From the table, we can see that RCV method has a similar performance with oracle method. However, naive method performances much worse especially when bandwidth is small and regular. Also, generally, probabilities of constant error variance are closer to 0.95 compared to probabilities of varying error variance. It is easy to understand since we ignore heteroscedasticity locally, however, it could

		constant			varying		
		SNR = 0	SNR = 1	SNR = 2	SNR = 0	SNR = 1	SNR = 2
oracle	small $\mathtt{h}$	0.9527	0.9515	0.9502	0.9161	0.9196	0.9307
	optimal $\mathtt{h}$	0.9499	0.9482	0.9494	0.9157	0.9158	0.9272
	large $\mathtt{h}$	0.9475	0.9498	0.9499	0.9228	0.9181	0.9251
naive	small $\mathtt{h}$	0.1013	0.1172	0.1175	0.0941	0.1164	0.0022
	optimal $\mathtt{h}$	0.4623	0.4741	0.4448	0.4322	0.4380	0.2830
	large $\mathtt{h}$	0.9264	0.9223	0.9224	0.8835	0.8784	0.8655
glrcv	small $\mathtt{h}$	0.9517	0.9347	0.9475	0.9149	0.9368	0.9426
	optimal $\mathtt{h}$	0.9494	0.9430	0.9483	0.9155	0.9412	0.9432
	large $\mathtt{h}$	0.9484	0.9335	0.9377	0.9229	0.9275	0.9358
ccsisrcv	small $\mathtt{h}$	0.9506	0.9512	0.9479	0.9375	0.9341	0.9278
	regular $\mathtt{h}$	0.9492	0.9430	0.9449	0.9372	0.9286	0.9282
	large $\mathtt{h}$	0.9505	0.9459	0.9412	0.9341	0.9350	0.9264

Table 5.2: Average probability of falling into 95% confidence interval.

cause problems especially when bandwidth is large and error variance changes rapidly. We also need to point out that group LASSO has a better performance when signal is strong. The other methods are not sensitive to signal.

To further shown the performance of the methods pointwise, we also choose a few points to present our results. Table 5.2-5.8 shows the bias and standard deviation at point 0.3, 0.4, 0.5, 0.6, 0.7 for 6 scenarios. From the tables, we can see the patterns more clear. Compared across methods, the bias and standard deviation of RCV methods is similar to oracle methods, sometimes even smaller. Furthermore, in all the scenarios, naive method produces estimates with much larger bias than RCV method compared to naive method. It shows that RCV method produces good estimates for varying coefficient model. Compared across bandwidth, the standard deviation do not change with bandwidth for all 4 methods. However, bias has a clear decreasing trend as bandwidth increases. It could due to the number of points to estimate is larger as bandwidth increases. Compared across points, there is no clear pattern that both bias and standard deviation changes with different points. Compared across error variance settings, the bias of varying error variance is larger than constant variance under then same SNR settings. Compared

		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oralce	small $\mathfrak{h}$	-0.058	0.209	-0.048	0.209	-0.05	0.189	-0.047	0.199	-0.053	0.203
	optimal $\mathfrak{h}$	-0.022	0.169	-0.032	0.166	-0.032	0.155	-0.033	0.157	-0.032	0.163
	large $\mathfrak{h}$	-0.013	0.145	-0.019	0.139	-0.017	0.134	-0.019	0.135	-0.021	0.137
naive	small $\mathfrak{h}$	-0.376	0.117	-0.374	0.112	-0.376	0.117	-0.368	0.12	-0.377	0.105
	optimal $\mathfrak{h}$	-0.222	0.109	-0.219	0.111	-0.217	0.109	-0.216	0.111	-0.218	0.107
	large $\mathfrak{h}$	-0.064	0.112	-0.064	0.113	-0.064	0.113	-0.063	0.118	-0.059	0.116
glrcv	small $\mathfrak{h}$	-0.002	0.156	-0.009	0.142	-0.001	0.144	0.006	0.144	0.012	0.15
	optimal $\mathfrak{h}$	-0.002	0.133	-0.007	0.122	-0.001	0.122	0.006	0.123	0.01	0.124
	large $\mathfrak{h}$	0.015	0.12	-0.005	0.114	-0.003	0.11	0.004	0.112	0.021	0.111
ccsisrcv	small $\mathfrak{h}$	0.013	0.16	0.011	0.158	0.016	0.169	0.028	0.168	0.03	0.158
	optimal $\mathfrak{h}$	0.01	0.131	0.014	0.139	0.019	0.142	0.024	0.142	0.019	0.136
	large $\mathfrak{h}$	0.012	0.125	0.007	0.124	0.014	0.122	0.014	0.123	0.026	0.125

Table 5.3: Bias and standard deviation for constant error variance, SNR = 0 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oracle	small $\mathfrak{h}$	-0.059	0.215	-0.04	0.211	-0.063	0.205	-0.058	0.201	-0.065	0.215
	optimal $\mathfrak{h}$	-0.046	0.172	-0.036	0.169	-0.047	0.17	-0.041	0.166	-0.05	0.165
	large $\mathfrak{h}$	-0.025	0.149	-0.018	0.147	-0.025	0.147	-0.026	0.142	-0.03	0.14
naive	small $\mathfrak{h}$	-0.37	0.112	-0.378	0.118	-0.378	0.119	-0.372	0.11	-0.375	0.123
	optimal $\mathfrak{h}$	-0.224	0.11	-0.232	0.113	-0.224	0.112	-0.219	0.11	-0.219	0.123
	large $\mathfrak{h}$	-0.07	0.118	-0.073	0.118	-0.07	0.118	-0.067	0.119	-0.066	0.119
glrcv	small $\mathfrak{h}$	0.04	0.209	0.039	0.213	0.053	0.207	0.069	0.194	0.075	0.199
	optimal $\mathfrak{h}$	0.014	0.174	0.009	0.17	0.016	0.162	0.027	0.16	0.044	0.158
	large $\mathfrak{h}$	0.023	0.155	-0.023	0.141	-0.023	0.136	-0.009	0.136	0.055	0.141
ccsisrcv	small $\mathfrak{h}$	0.026	0.157	0.023	0.154	0.023	0.165	0.019	0.161	0.017	0.165
	optimal $\mathfrak{h}$	0.019	0.127	0.023	0.129	0.02	0.135	0.02	0.134	0.013	0.128
	large $\mathfrak{h}$	0.023	0.12	0.018	0.116	0.018	0.118	0.012	0.119	0.023	0.117

Table 5.4: Bias and standard deviation for constant error variance, SNR = 1 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

across signal strength, biases of GL-RCV seems to be sensitive with signal strength. Generally, I would say that it performance better with strong signal. The estimates of the other methods do not change with signal strength.

		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oracle	small $\mathfrak{h}$	-0.052	0.202	-0.052	0.204	-0.046	0.202	-0.043	0.207	-0.054	0.196
	optimal $\mathfrak{h}$	-0.033	0.163	-0.025	0.167	-0.035	0.165	-0.029	0.162	-0.032	0.16
	large $\mathfrak{h}$	-0.009	0.14	-0.005	0.142	-0.018	0.143	-0.011	0.137	-0.012	0.14
naive	small $\mathfrak{h}$	-0.37	0.12	-0.37	0.117	-0.382	0.113	-0.378	0.118	-0.367	0.121
	optimal $\mathfrak{h}$	-0.222	0.11	-0.222	0.112	-0.23	0.106	-0.23	0.109	-0.219	0.11
	large $\mathfrak{h}$	-0.062	0.117	-0.065	0.115	-0.073	0.112	-0.07	0.111	-0.064	0.114
glrcv	small $\mathfrak{h}$	0.021	0.155	0.024	0.167	0.026	0.173	0.053	0.167	0.061	0.162
	optimal $\mathfrak{h}$	0.011	0.121	0.011	0.125	0.016	0.136	0.027	0.13	0.035	0.128
	large $\mathfrak{h}$	0.034	0.109	0.018	0.111	0.021	0.114	0.023	0.111	0.05	0.112
ccsisrcv	small $\mathfrak{h}$	0.028	0.166	0.027	0.168	0.029	0.159	0.027	0.166	0.013	0.164
	optimal $\mathfrak{h}$	0.018	0.138	0.021	0.137	0.024	0.13	0.023	0.135	0.013	0.137
	large $\mathfrak{h}$	0.038	0.123	0.026	0.122	0.025	0.12	0.02	0.121	0.03	0.124

Table 5.5: Bias and standard deviation for constant error variance, SNR = 2 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oracle	small $\mathfrak{h}$	0.01	0.217	-0.04	0.201	-0.112	0.167	-0.134	0.16	-0.136	0.164
	optimal $\mathfrak{h}$	0.024	0.177	-0.033	0.161	-0.086	0.139	-0.117	0.131	-0.113	0.136
	large $\mathfrak{h}$	0.032	0.154	-0.021	0.137	-0.066	0.122	-0.094	0.114	-0.088	0.119
naive	small $\mathfrak{h}$	-0.397	0.123	-0.354	0.111	-0.338	0.097	-0.322	0.092	-0.321	0.099
	optimal $\mathfrak{h}$	-0.233	0.116	-0.213	0.109	-0.207	0.098	-0.205	0.088	-0.204	0.089
	large $\mathfrak{h}$	-0.046	0.128	-0.07	0.114	-0.093	0.096	-0.106	0.092	-0.095	0.098
glrcv	small $\mathfrak{h}$	0.059	0.166	-0.008	0.142	-0.063	0.128	-0.092	0.109	-0.089	0.113
	optimal $\mathfrak{h}$	0.05	0.137	-0.007	0.12	-0.053	0.107	-0.077	0.097	-0.072	0.099
	large $\mathfrak{h}$	0.058	0.125	-0.001	0.108	-0.04	0.096	-0.06	0.091	-0.041	0.092
ccsisrcv	small $\mathfrak{h}$	0.075	0.18	0.014	0.149	-0.035	0.133	-0.066	0.127	-0.056	0.131
	optimal $\mathfrak{h}$	0.062	0.149	0.011	0.127	-0.03	0.114	-0.052	0.105	-0.049	0.113
	large $\mathfrak{h}$	0.055	0.127	0.01	0.113	-0.02	0.102	-0.038	0.096	-0.029	0.103

Table 5.6: Bias and standard deviation for varying error variance, SNR = 0 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oracle	small $\mathtt{h}$	0.014	0.231	-0.058	0.194	-0.129	0.166	-0.145	0.153	-0.129	0.158
	optimal $\mathtt{h}$	0.022	0.187	-0.042	0.154	-0.109	0.138	-0.126	0.122	-0.106	0.126
	large $\mathtt{h}$	0.027	0.158	-0.023	0.134	-0.078	0.118	-0.091	0.104	-0.084	0.108
naive	small $\mathtt{h}$	-0.389	0.129	-0.368	0.114	-0.339	0.1	-0.317	0.105	-0.326	0.102
	optimal $\mathtt{h}$	-0.228	0.119	-0.225	0.102	-0.219	0.096	-0.207	0.098	-0.211	0.098
	large $\mathtt{h}$	-0.051	0.123	-0.081	0.11	-0.101	0.099	-0.103	0.096	-0.099	0.099
glrcv	small $\mathtt{h}$	0.109	0.252	0.031	0.228	-0.037	0.175	-0.057	0.165	-0.044	0.18
	optimal $\mathtt{h}$	0.054	0.195	-0.015	0.173	-0.056	0.151	-0.07	0.136	-0.051	0.146
	large $\mathtt{h}$	0.055	0.173	-0.041	0.145	-0.078	0.128	-0.081	0.12	-0.023	0.127
ccsisrcv	small $\mathtt{h}$	0.074	0.183	0.008	0.155	-0.056	0.14	-0.075	0.13	-0.061	0.135
	optimal $\mathtt{h}$	0.059	0.153	0.01	0.131	-0.039	0.116	-0.062	0.109	-0.06	0.109
	large $\mathtt{h}$	0.055	0.136	0.003	0.118	-0.031	0.101	-0.047	0.097	-0.032	0.103

Table 5.7: Bias and standard deviation for varying error variance, SNR = 1 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

point		0.3		0.4		0.5		0.6		0.7	
		bias	std	bias	std	bias	std	bias	std	bias	std
oracle	small $\mathtt{h}$	0.0085	0.2169	-0.056	0.182	-0.1181	0.1568	-0.1419	0.1426	-0.129	0.154
	optimal $\mathtt{h}$	0.0326	0.1753	-0.033	0.1457	-0.0941	0.1244	-0.1209	0.1162	-0.1086	0.1209
	large $\mathtt{h}$	0.0439	0.1502	-0.0182	0.1281	-0.0699	0.1102	-0.0915	0.1028	-0.0827	0.1053
naive	small $\mathtt{h}$	-0.601	0.103	-0.55	0.096	-0.529	0.085	-0.498	0.08	-0.5	0.083
	optimal $\mathtt{h}$	-0.212	0.103	-0.216	0.097	-0.218	0.088	-0.219	0.081	-0.203	0.082
	large $\mathtt{h}$	-0.029	0.098	-0.029	0.091	-0.044	0.086	-0.045	0.08	-0.048	0.081
glrcv	small $\mathtt{h}$	0.0691	0.1779	0.0012	0.1496	-0.0437	0.1347	-0.0528	0.1371	-0.0335	0.1387
	optimal $\mathtt{h}$	0.0586	0.1443	0.0023	0.1216	-0.04	0.1108	-0.0501	0.1079	-0.0364	0.1091
	large $\mathtt{h}$	0.0713	0.1298	0.0157	0.1098	-0.0197	0.0989	-0.0349	0.0944	-0.0016	0.101
ccsisrcv	small $\mathtt{h}$	0.1194	0.1774	0.0366	0.1661	-0.045	0.1401	-0.0775	0.1247	-0.0427	0.1299
	optimal $\mathtt{h}$	0.0932	0.1448	0.0337	0.1335	-0.0397	0.1168	-0.0619	0.1058	-0.0358	0.1105
	large $\mathtt{h}$	0.1152	0.1355	0.0381	0.1152	-0.0206	0.1044	-0.0398	0.0961	-0.0006	0.104

Table 5.8: Bias and standard deviation for varying error variance, SNR = 2 at local point $\mathfrak{u} = 0.3, 0.4, 0.5, 0.6, 0.7$.

5.7 Real Data Application

To illustrate the methodology, we apply the newly proposed procedures to the Framingham Heart Study (FHS) data. It is a cardiovascular study that began in 1948 under the guidance of the National Heart, Lung and Blood Institute (Dawber, Meadors, and Moore 1951; Jaquish 2007). There are $n = 977$ samples (subjects) and $p = 349,985$ variables (nonrare SNPs). Our major goal is to spot the SNPs that are most associated with body mass index (BMI). Since each SNP has the dominant effect (D) and the additive effect (A), the total variables are $2p = 699,970$. This is typical ultrahigh dimensional data. According to existing experience and some previous research results, the effect of SNPs changes with age. Therefore, the varying coefficient model naturally becomes a reasonable choice to this data. In this model, $\mathbf{y}$ is the BMI, $\mathbf{U}$ is the age, and $\mathbf{x}$ is the SNPs. We first use conditional correlation sure independent screening (CC-SIS) to reduce the number of predictors from $2p = 699,970$ to a moderate size (132). The 132 SNPs with the largest 132 conditional correlations are listed in Table 5.10. According to the recommendation in Liu, Li, and Wu (2014), the variables with highest $d = n^{4/5}/(\log(n^{4/5})) \approx 44$ should be chosen. In this dissertation, we compare the estimates from different sample size $34, 44, 88, 132$. 34 is the number of variables Liu, Li, and Wu (2014) chose, and we use it as the benchmark for comparison.

For data $(\mathbf{y}, \mathbf{X}, \mathbf{u})$, we use naive method to approach the coefficients estimation $\widehat{\alpha}_j(u)$. The detailed procedure is described in Section 5.3. Generally speaking, we use group LASSO to further select variables first and refit the selected variables with low dimension varying-coefficient model method (local liner regression) for the estimation. Also, we calculate the $\widehat{\sigma}^2_{naive}$ for the comparison purpose. Meanwhile, we use the proposed RCV method to calculate $\widehat{\sigma}^2_{rcv}$. Also, two kinds of confidence interval of coefficients can be constructed based on the two kinds of $\widehat{\sigma}^2$. Since, the naive method is expected to significantly underestimate the error variance, the

method	rcv					naive				
d	$U = 34$	37	41	43	48	34	37	41	43	48
34	18.481	15.815	18.585	14.631	15.682	16.549	15.503	12.577	12.180	14.604
44	16.704	15.115	15.824	12.578	14.500	14.625	14.671	12.932	11.871	11.245
88	19.143	15.284	13.324	14.108	13.224	12.299	10.015	8.273	9.131	9.022
132	20.241	14.525	17.432	11.461	14.466	8.258	6.725	6.619	8.086	9.880

Table 5.9: The error variance estimates at 5 points of covariate U.

confidence bands based on naive method may have more significant variables than ones from RCV method.

After apply the two methods onto the data, we first compute $\widehat{\sigma}^2_{naive}$ and $\widehat{\sigma}^2_{rcv}$ for sample size $34, 44, 88, 132$. Since they are both function, we show the values at $(0.15, 0.25, 0.5, 0.75, 0.85)$ quantile of U.

In Table 5.9, it is easy to see that $\widehat{\sigma}^2_{naive}$ has a clear decreasing trend as the number of variables chosen getting larger. Meanwhile, $\widehat{\sigma}^2_{rcv}$ stays similar. And, for any particular covariate point, $\widehat{\sigma}^2_{rcv}$ always larger than $\widehat{\sigma}^2_{naive}$.

Since the optimize sample size is 44, it is used to conduct further study. We first use naive method, and the coefficient function $\{\widehat{\alpha}_j(U)\}, j = 1, \ldots, p$ can be estimated. We further compare $\int |\widehat{\alpha}_j(U)|^2 \, du$ for each j in $1 \ldots p$, can choose the predictors with largest nine L_2 norm. Along with the $\widehat{\sigma}^2$ estimated from the both naive methods, and RCV method, we can further construct confidence bands for these 9 variables.

The RCV confidence bands in Figure 5.9 are clearly larger than the naive confidence bands, which is consistent with our theoretical proof and simulation results.

Table 5.10: List of the selected 132 SNP's

	SNP	ρ^2		SNP	ρ^2
1	ss66511535(A)	0.0261	67	ss66209054(A)	0.0179
2	ss66517429(D)	0.0258	68	ss66526138(D)	0.0179
3	ss66112931(D)	0.0243	69	ss66102245(D)	0.0179
4	ss66088050(D)	0.0243	70	ss66265096(A)	0.0178
5	ss66155306(A)	0.0233	71	ss66056760(A)	0.0178
6	ss66070305(A)	0.0233	72	ss66265548(D)	0.0178
7	ss66398253(A)	0.0232	73	ss66103158(A)	0.0178
8	ss66319388(A)	0.0228	74	ss66409420(D)	0.0177
9	ss66489729(A)	0.0223	75	ss66039508(D)	0.0177
10	ss66258748(D)	0.0219	76	ss66183857(D)	0.0177
11	ss66398882(A)	0.0215	77	ss66356701(D)	0.0177
12	ss66353634(D)	0.0213	78	ss66272727(D)	0.0177
13	ss66144265(A)	0.0213	79	ss66097729(D)	0.0177
14	ss66346559(A)	0.0211	80	ss66040305(D)	0.0176
15	ss66085516(D)	0.0211	81	ss66536393(D)	0.0176
16	ss66461698(D)	0.021	82	ss66141492(A)	0.0176
17	ss66159949(A)	0.0209	83	ss66269504(D)	0.0176
18	ss66363198(A)	0.0208	84	ss66438421(D)	0.0176
19	ss66485483(A)	0.0207	85	ss66116314(D)	0.0176
20	ss66084087(D)	0.0205	86	ss66302804(A)	0.0176
21	ss66393072(A)	0.0204	87	ss66507772(D)	0.0175
22	ss66058021(D)	0.0204	88	ss66411959(A)	0.0175
23	ss66264934(A)	0.0203	89	ss66164865(D)	0.0175
24	ss66516012(A)	0.0202	90	ss66416614(A)	0.0174
25	ss66058506(A)	0.02	91	ss66305798(D)	0.0174
26	ss66102722(A)	0.0199	92	92(D)	0.0174
27	ss66404926(A)	0.0198	93	ss66336837(D)	0.0174
28	ss66089086(A)	0.0198	94	ss66052897(D)	0.0174
29	ss66491317(D)	0.0197	95	ss66143305(A)	0.0174
30	ss66236850(A)	0.0196	96	ss66320873(D)	0.0173
31	ss66153510(A)	0.0196	97	ss66222883(D)	0.0173
32	ss66282476(A)	0.0196	98	ss66139687(D)	0.0173
33	ss66188749(A)	0.0195	99	ss66430303(D)	0.0172
34	ss66306173(A)	0.0194	100	ss66346937(A)	0.0172
35	ss66435333(A)	0.0193	101	ss66079410(D)	0.0171
36	ss66101395(D)	0.0193	102	ss66404926(D)	0.0171
37	ss66305798(A)	0.0192	103	ss66362972(A)	0.0171
38	ss66137328(A)	0.0192	104	ss66082721(D)	0.017
39	ss66469808(D)	0.0191	105	ss66316243(A)	0.017
40	ss66509394(D)	0.0191	106	ss66220359(D)	0.017
41	ss66080432(D)	0.0189	107	ss66047081(A)	0.017
42	ss66173508(D)	0.0189	108	ss66071513(A)	0.017
43	ss66302110(A)	0.0189	109	ss66367461(D)	0.017
44	ss66041456(D)	0.0188	110	ss66370736(D)	0.017
45	ss66501923(A)	0.0187	111	ss66379476(A)	0.0169
46	ss66354801(D)	0.0186	112	ss66358965(A)	0.0169
47	ss66052226(A)	0.0186	113	ss66281927(A)	0.0169
48	ss66176990(D)	0.0186	114	ss66321055(D)	0.0168
49	ss66205766(A)	0.0185	115	ss66282595(A)	0.0168
50	ss66158449(A)	0.0185	116	ss66221481(D)	0.0168
51	ss66184897(A)	0.0185	117	ss66358314(D)	0.0168
52	ss66451201(A)	0.0185	118	ss66467683(D)	0.0168
53	ss66272727(A)	0.0185	119	ss66332935(D)	0.0168
54	ss66190611(A)	0.0184	120	ss66485672(A)	0.0168
55	ss66522817(A)	0.0183	121	ss66119591(D)	0.0167
56	ss66141715(D)	0.0183	122	ss66262246(A)	0.0167
57	ss66323107(A)	0.0183	123	ss66426981(A)	0.0167
58	ss66532146(A)	0.0182	124	ss66534314(D)	0.0167
59	ss66422022(A)	0.0182	125	ss66126299(A)	0.0167
60	ss66383596(A)	0.0182	126	ss66268589(A)	0.0167
61	ss66374124(D)	0.0182	127	ss66445258(D)	0.0167
62	ss66115664(D)	0.0181	128	ss66333299(D)	0.0167
63	ss66470378(A)	0.018	129	ss66340691(D)	0.0166
64	ss66219875(D)	0.018	130	ss66274749(D)	0.0166
65	ss66168969(A)	0.0179	131	ss66367982(A)	0.0166
66	ss66174853(A)	0.0179	132	ss66518380(A)	0.0166

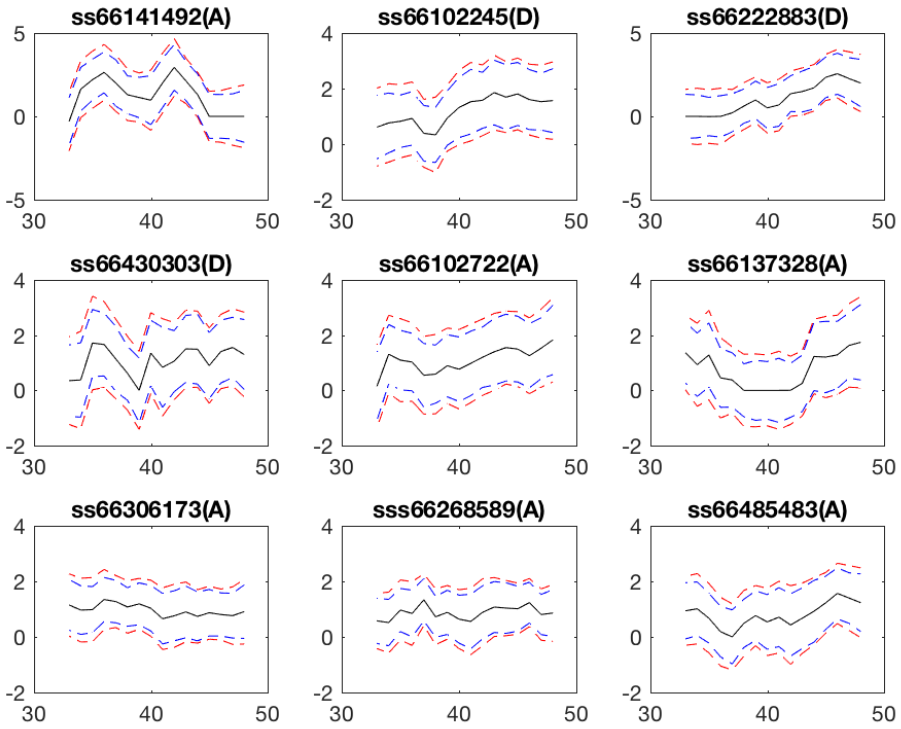


Figure 5.9: The coefficient function of 9 variables with largest L_2 norm of coefficient functions.

The red dash lines are the confidence bands respected to RCV method, and the blue dash lines are the confidence bands respected to naive method.

5.8 Regularity Conditions and Technical Proofs

5.8.1 Regularity Conditions

The following conditions are imposed to facilitate the proofs. They may not be the weakest conditions. Suppose $\mathcal{S}$ is a subset of $\{1, 2, \cdots, p\}$ and denote the cardinality of $\mathcal{S}$ by $|\mathcal{S}|$. Let $s = o(n)$. Define the conditional constrained

eigenvalues as

$$\begin{aligned}\phi_{\min}(s\,|u) &= \min_{|\mathcal{S}|<s} \lambda_{\min}\left(\mathrm{E}\left(\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathsf{T}}|U=u\right)\right)\\ \phi_{\max}(s\,|u) &= \max_{|\mathcal{S}|<s} \lambda_{\max}\left(\mathrm{E}\left(\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathsf{T}}|U=u\right)\right)\end{aligned} \tag{5.8.1}$$

(**C1**) For $j = 1, \cdots, p$ and any $u \in \mathcal{U}$, $\alpha_j''(u)$ exists and is continuous.

(**C2**) For an $s > 2$ and $j = 1, \cdots, p$, $\mathrm{E}\,|\varepsilon|^{2s} < \infty$, and $\mathrm{E}\,|X_j|^{2s} < \infty$.

(**C3**) The random variable U_i has a bound support $\mathcal{U} \subset \mathbb{R}^1$. Its density function $f_U(u)$ is Lipschitz continuous with order $\gamma \geq 2$ that is ,

$$|f_U(u_1) - f_U(u_2)| \leq C\,|u_1 - u_2|^{\gamma}, \quad \text{for some } C > 0,$$

and bounded away from 0 on its support.

(**C4**) For $j = 1, \cdots, p$, the joint distribution $f(u, X_j, \varepsilon)$ has a bounded support and uniformly continuous on its support.

(**C5**) $nh^4 \to 0$ and $n^2h^2/(\log(1/h))^2 \to \infty$.

(**C6**) Define $r(u) = \mathrm{E}\,(X^2\varepsilon^2\big|U = u)$. Assume that $r(u)$ is bounded away from 0 for $u \in \mathcal{U}$, and has a bounded first derivative on $\mathcal{U}$.

(**C7**) Let $\mathcal{S} \subset (1, \cdots, p)$ and $s_n = |\mathcal{S}|$. Assume that for $s_n = o(n)$, $\mathrm{E}\,(\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathsf{T}}\big|U = u)$ and $\mathrm{E}\,((\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{\mathsf{T}})^{-1}\big|U = u)$ are both Lipschitz continuous with order $\gamma \geq 2$. Assume that $\kappa(U_i) = \mathrm{E}\,(\varepsilon_i^4|U_i)$ exists.

(**C8**) The kernel $K(u)$ is a symmetric density function with finite support $\mathcal{U}$ and satisfies that

$$\sup_{u\in\mathcal{U}} |K(u)| < \infty.$$

Define

$$\mu_k = \int_{\mathcal{U}} u^k K(u)\,du, \quad \nu_k = \int_{\mathcal{U}} u^k K^2(u)\,du,$$

(**C9**) There exist two constants $\lambda_0 > 0$ and $\lambda_1 > 0$ such that

$$\begin{aligned} &P\left(\inf_u \phi_{\min}(s_u \,|U = u) \geq \lambda_0\right) = 1 \\ &P\left(\sup_u \phi_{\max}(s_u \,|U = u) \leq \lambda_1\right) = 1 \end{aligned} \tag{5.8.2}$$

5.8.2 Technical Proofs

Lemma 1. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d. random vectors, where the Y_i's are scalar random variables. Further assume that $E\,|Y|^s < \infty$ and $\sup_x \int |y|^s\, f(x, y)dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Given that $n^{2\varepsilon-1}h \to \infty$ for some $\varepsilon < 1 - s^{-1}$, then

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)Y_i - E\left(K_h(X_i - x)Y_i\right) \right| = O_p\left(\left(\frac{\log(1/h)}{nh} \right)^{1/2} \right). \tag{5.8.3}$$

Lemma 2. [Smirnov, 1944] Let $X_1, X_2, \cdots, X_n$ be independent identically distributed random variables with distribution function $F(t)$. $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t)$ is the corresponding empirical distribution function. For $\lambda \geq 0$, we have:

$$P\left(\sup_t (F_n(t) - F(t)) < \lambda\, n^{-\frac{1}{2}} \right) \to 1 - \exp\{-2\lambda^2\}, \quad n \to \infty. \tag{5.8.4}$$

For the two-dimensional independent identically distributed random variables (X_1, Y_1), ..., (X_n, Y_n) with distribution function $F(u, v)$ and empirical distribution function $F_n(u, v)$, we also have, for $\lambda \geq 0$,

$$P\left(\sup_{u,v} (F_n(u, v) - F(u, v)) < \lambda n^{-\frac{1}{2}} \right) \to 1 - \exp(-c\lambda^2), \quad n \to \infty, \tag{5.8.5}$$

where constant $0 < c \leq 2$.

Lemma 3. Let $G(v)$ and $K(u)$ be bounded differentiable functions. $F(u, v)$ is a bivariate joint cumulative distribution function. We have

$$\begin{aligned}
&\int_{s_u}^{S^u}\int_{s_v}^{S^v} G(v)K(u)\,dF(u,v) \\
=&-\int_{L_1\bigcup L_3} G(v_0)K(u)\,dF(u,v_0) - \int_{L_2\bigcup L_4} F(u_0,v)K(u)\,dG(v) \\
&+\int_{s_u}^{S^u}\int_{s_v}^{S^v} F(u,v)\,dK(u)\,dG(v),
\end{aligned}$$

where

$$L_1 : s_u \leq u \leq S^u, v_0 = s_v,$$

$$L_2 : u_0 = S^u, s_v \leq v \leq S^v,$$

$$L_3 : S^u \geq u \geq s_u, v_0 = S^v,$$

$$L_4 : u_0 = s_u, S^v \geq v \geq s_v.$$

.

Proof of Lemma 3.

Let

$$\begin{aligned}
a(u,v) &= G(v)K(u)\frac{\partial F(u,v)}{\partial u}, \\
b(u,v) &= G'(v)K(u)F(u,v).
\end{aligned}$$

Using the Green's identity, we obtain

$$\int_{\bigcup_i L_i} a(u,v)\,du + b(u,v)\,dv = \int_{s_u}^{S^u}\int_{s_v}^{S^v}\left(-\frac{\partial a(u,v)}{\partial v} + \frac{\partial b(u,v)}{\partial u}\right)\,du\,dv. \quad (5.8.6)$$

We calculate each term of the left hand side of Equation (5.8.6) as follows.

$$\int_{L_1} a(u,v)\,du = \int_{L_1} G(s_v)K(u)\frac{\partial F(u,s_v)}{\partial u}\,du = \int_{L_1} G(v_0)K(u)\frac{\partial F(u,v_0)}{\partial u}\,du,$$

where $v_0 = s_v$ on L_1. Similarly, we have

$$\int_{L_3} a(u,v)\,du = \int_{L_3} G(v_0)K(u)\frac{\partial F(u,v_0)}{\partial u}\,du,\ v_0 = S^v.$$

We notice that $\int_{L_2\cup L_4} a(u,v)du = 0$ due to $du = 0$ on $L_2 \cup L_4$. Thus, the left hand side equals

$$\int_{L_1\cup L_3} G(v_0)K(u)\frac{\partial F(u,v_0)}{\partial u}\,du + \int_{L_2\cup L_4} K(u_0)F(u_0,v)\,dG(v).$$

We calculate the right hand side of Equation (5.8.6). By denoting $\Omega = [s_u, S^u] \times [s_v, S^v]$, we get

$$\begin{aligned} & -\int_\Omega \frac{\partial a(u,v)}{\partial v}\,du\,dv \\ = & -\int_\Omega \frac{\partial}{\partial v}\left(G(v)K(u)\frac{\partial F(u,v)}{\partial u}\right)\,du\,dv \\ = & -\int_\Omega K(u)\left(G'(v)\frac{\partial F(u,v)}{\partial u} + G(v)\frac{\partial^2 F(u,v)}{\partial u\partial v}\right)\,du\,dv \\ = & -\int_\Omega K(u)\frac{\partial F(u,v)}{\partial u}\,du\,dG(v) - \int_\Omega K(u)G(v)\,dF(u,v) \end{aligned} \tag{5.8.7}$$

Analogously, we also have

$$\int_\Omega \frac{\partial b(u,v)}{\partial u}\,du\,dv = \int_\Omega K(u)\frac{\partial F(u,v)}{\partial u}\,du\,dG(v) + \int_\Omega F(u,v)\,dK(u)\,dG(v). \tag{5.8.8}$$

Combining Equations (5.8.7) and (5.8.8), the right hand side of equation (5.8.6) equals

$$\int_\Omega F(u,v)\,dK(u)\,dG(v) - \int_\Omega K(u)G(v)\,dF(u,v).$$

Consequently, we prove the result of Lemma 3.

Lemma 4. Denote the kernel estimator by

$$\hat{\beta}_n^{(j)}(u) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - U_i)X_{ij}\varepsilon_i = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{u - U_i}{h})X_{ij}\varepsilon_i. \tag{5.8.9}$$

Under the regularity conditions presented in the section 8.1, it follows that

$$P\left(\sup_{s_u \le u \le S^u} \left|\hat{\beta}_n^{(j)}(u) - E\,\hat{\beta}_n^{(j)}(u)\right| > M_1\right) \le \ 4\exp(-2c_1 M_1^2 nh^2), \tag{5.8.10}$$

where $c_1 > 0$ and $M_1 > 0$ are positive constants.

Proof of Lemma 4.

Denote the conditional mean function by $\beta^{(j)}(u) = E\,(X_{ij}\varepsilon_i | U_i = u)$, $i = 1, \cdots, n$, and the corresponding kernel estimator by

$$\hat{\beta}_n^{(j)}(u) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - U_i)X_{ij}\varepsilon_i = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{u - U_i}{h})X_{ij}\varepsilon_i, \tag{5.8.11}$$

where $K(\cdot)$ is symmetric kernel function with bounded support.

$$E\,\hat{\beta}_n^{(j)}(u) = \frac{1}{h}\int K(\frac{u - u_i}{h})\,\beta^{(j)}(u_i)\,du_i = \int K(u_i)\,\beta^{(j)}(u - u_i h)\,du_i, \tag{5.8.12}$$

then,

$$\begin{aligned} E\,\hat{\beta}_n^{(j)}(u) - \beta^{(j)}(u) &= \int K(u_i)\,\beta^{(j)}(u - u_i h)\,du_i - \int K(u_i)\,\beta^{(j)}(u)\,du_i \\ &= \int K(u_i)\left(-(\beta^{(j)}(u))' u_i h + \frac{1}{2}(\beta^{(j)}(u - \xi u_i h))'' u_i^2 h^2\right)\,du_i \\ &= O(h^2). \end{aligned} \tag{5.8.13}$$

The last equation holds by $(\beta^{(j)}(u))''$ exists and is uniformly bounded. For $\tau <$

$1/2, n^{2\tau}h^4 < nh^4 \to 0$. There exists a constant N_1, for any $n > N_1$, $n^{\tau}\left|E_n^{(j)}(u) - {}^{(j)}(u)\right| < c_1 n^{\tau}h^2 < c_2$. Therefore, we have

$$P\Big(\sup_{s_u \le u \le S^u}\left|E_n^{(j)}(u) - {}^{(j)}(u)\right| > c_2 n^{-\tau}\Big) < \epsilon. \tag{5.8.14}$$

Denote by $V_{ij} = X_{ij}\varepsilon_i$. Suppose V_{ij} is integrable. Then by dominated convergence theorem, we have

$$\begin{aligned} E_n^{(j)}(u) &= \frac{1}{h}\int_{-\infty}^{+\infty}\int_{s_{v_j}}^{S^{v_j}} v_{ij}K(\frac{u-u_i}{h})\,dF(u_i, v_{ij}) \\ &= \lim_{M\to\infty}\int_{-M}^{M}\int_{s_{v_j}}^{S^{v_j}} v_{ij}K(\frac{u-u_i}{h})\,dF(u_i, v_{ij}). \end{aligned} \tag{5.8.15}$$

By Lemma 3, the integral equals

$$\begin{aligned} \frac{1}{h}\lim_{M\to\infty}\Big[&-\int_{-M}^{M} s_{v_j}K(\frac{u-u_i}{h})\,dF(u_i, s_{v_j}) - \int_{-M}^{M} S^{v_j}K(\frac{u-u_i}{h})\,dF(u_i, S^{v_j}) \\ &-\int_{s_{v_j}}^{S^{v_j}} F(M, v_{ij})K(\frac{u-M}{h})\,dv_{ij} - \int_{s_{v_j}}^{S^{v_j}} F(-M, v_{ij})K(\frac{u+M}{h})\,dv_{ij}\Big] \\ &+ \frac{1}{h}\lim_{M\to\infty}\int_{-M}^{M}\int_{s_{v_j}}^{S^{v_j}} F(u_i, v_{ij})\,dK(\frac{u-u_i)}{h}\,dv_{ij} \end{aligned} \tag{5.8.16}$$

For $K(\cdot)$ is a kernel function with bounded support,

$$\lim_{M\to\infty}\int_{s_{v_j}}^{S^{v_j}} F(M, v_{ij})K(\frac{u-M}{h})\,dv_{ij} = \lim_{M\to\infty}\int_{s_{v_j}}^{S^{v_j}} F(-M, v_{ij})K(\frac{u+M}{h})\,dv_{ij} = 0. \tag{5.8.17}$$

Notice that $F(u_i, s_{v_j}) = 0, F(u_i, S^{v_j}) = F_U(u_i)$, the marginal cumulative distribution

function of random variable U. By integration by parts, we have

$$\begin{aligned}
&\int_{-M}^{M} s_{\nu_j} K(\frac{u-u_i}{h})\, dF(u_i, s_{\nu_j}) = 0 \\
&\int_{-M}^{M} S^{\nu_j} K(\frac{u-u_i}{h})\, dF(u_i, S^{\nu_j}) = \int_{-M}^{M} S^{\nu_j} F_U(u_i)\, dK(\frac{u-u_i}{h}).
\end{aligned} \tag{5.8.18}$$

Therefore, it follows

$$E_n^{(j)}(u) = -\frac{1}{h}\left[\int_{-\infty}^{+\infty} S^{\nu_j} F_U(u_i)\, dK(\frac{u-u_i}{h}) - \int_{-\infty}^{+\infty}\int_{s_{\nu_j}}^{S^{\nu_j}} F(u_i, \nu_{ij})\, dK(\frac{u-u_i}{h})\, d\nu_{ij}\right]. \tag{5.8.19}$$

Replacing the probability measure by the empirical measure, we can have

$$\begin{aligned}
{}_n^{(j)}(u) &= \frac{1}{h}\int_{-\infty}^{+\infty}\int_{s_{\nu_j}}^{S^{\nu_j}} \nu_{ij} K(\frac{u-u_i}{h})\, dF_n(u_i, \nu_{ij}) \\
&= -\frac{1}{h}\left[\int_{-\infty}^{+\infty} S^{\nu_j} F_{U,n}(u_i)\, dK(\frac{u-u_i}{h}) - \int_{-\infty}^{+\infty}\int_{s_{\nu_j}}^{S^{\nu_j}} F_n(u_i, \nu_{ij})\, dK(\frac{u-u_i}{h})\, d\nu_{ij}\right],
\end{aligned} \tag{5.8.20}$$

where $F_{U,n}(\cdot)$ and $F_n(\cdot)$ are the empirical distribution function corresponding to the sample $(V_{ij}, U_i), U_i\ i = 1, \cdots, n$, respectively.

$$\begin{aligned}
&\sup_{s_u \le u \le S^u} \left|{}_n^{(j)}(u) - E_n^{(j)}(u)\right| \\
\le\ &\sup_{s_u \le u \le S^u} \Big| \frac{S^{\nu_j}}{h}\int_{-\infty}^{+\infty} (F_{U,n}(u_i) - F_U(u_i))\, dK(\frac{u-u_i}{h}) \\
&+ \frac{1}{h}\int_{-\infty}^{+\infty}\int_{s_{\nu_j}}^{S^{\nu_j}} (F_n(u_i, \nu_{ij}) - F(u_i, \nu_{ij}))\, dK(\frac{u-u_i}{h})\, d\nu_{ij}\Big| \\
\le\ &\frac{S^{\nu_j}}{h}\sup_{-\infty \le u_i \le \infty} |F_{U,n}(u_i) - F_U(u_i)|\, M^* + \frac{1}{h}\sup_{\substack{-\infty \le u_i \le \infty \\ s_{\nu_j} \le \nu_{ij} \le S^{\nu_j}}} |F_n(u_i, \nu_{ij}) - F(u_i, \nu_{ij})|\, M^* S_j^*.
\end{aligned} \tag{5.8.21}$$

Using Lemma 1, it follows that

$$\begin{aligned}
&P\left(\sup_{s_u\le u\le S^u}\left|\ell_n^{(j)}(u)-E\,\ell_n^{(j)}(u)\right|>M_1\right)\\
&\le P\Big(\sup_{-\infty\le u_i\le\infty}|F_{U,n}(u_i)-F_U(u_i)|>M_2h\Big)\\
&\quad+\ P\Big(\sup_{\substack{-\infty\le u_i\le\infty\\ s_{v_j}\le v_{ij}\le S^{v_j}}}|F_n(u_i,v_{ij})-F(u_i,v_{ij})|>M_2h\Big) \qquad (5.8.22)\\
&\le 2\exp(-2M_2\,nh^2)+2\exp(-2M_2\,nh^2)\\
&=4\exp(-2M_2\,nh^2)
\end{aligned}$$

Lemma 5 Let $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)^T$ be an i.i.d. random vector with $E\,\varepsilon_1=0, \mathrm{Var}\varepsilon_1=1, E\,\varepsilon_1^4=\kappa$. If $\Sigma\ge 0$, then it follows

$$E\left(\frac{1}{n}\varepsilon^T\Sigma\varepsilon-\frac{1}{n}\mathrm{tr}(\Sigma)\right)^2=\frac{2}{n}\mathrm{tr}(\Sigma^2)+\frac{\kappa-3}{n^2}\sum_{i=1}^n\sigma_{ii}^2 \qquad (5.8.23)$$

Proof of Proposition 1. (Order of $c(u)$)

Recall that $c(u)=\mathrm{tr}((\mathbf{I}-\mathbf{S}_u)^T\mathbf{W}_u^*(\mathbf{I}-\mathbf{S}_u))=\mathrm{tr}(\mathbf{W}_u^*)-\mathrm{tr}(\mathbf{W}_u^*\mathbf{S}_u)-\mathrm{tr}(\mathbf{S}^T\mathbf{W}_u^*)+\mathrm{tr}(\mathbf{S}_u^T\mathbf{W}_u^*\mathbf{S}_u)$. Here the superscript $*$ means using the different bandwidth h^*. First, we have

$$\frac{1}{n}\mathrm{tr}(\mathbf{W}_u^*)=\frac{1}{n}\sum_{i=1}^n w_{u,i}^*=\frac{1}{n}\sum_{i=1}^n\frac{1}{h^*}K(\frac{U_i-u}{h^*})=E\,\frac{1}{h^*}K(\frac{U_i-u}{h^*})+O_p(\sqrt{\frac{\log(1/h^*)}{nh^*}}), \qquad (5.8.24)$$

where

$$E\,\frac{1}{h^*}K(\frac{U_i-u}{h^*})=\int\frac{1}{h^*}K(\frac{u_i-u}{h^*})f_U(u_i)\,du_i=\ f_U(u)+O_p(h^{*2}) \qquad (5.8.25)$$

Thus, $\mathrm{tr}(\mathbf{W}_u^*)=nf_U(u)+O_p(nh^{*2})$. Denote by $w_j^{(i)}=\mathbf{W}_{u_i}(j,j)$, $w_{u,i}^*=\mathbf{W}_u^*(i,i)$

and $\mathbf{P}^{(i)} = \mathbf{D}_{u_i}(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T$. Because $\mathbf{W}_{u_i}^{1/2} \mathbf{P}^{(i)} \mathbf{W}_{u_i}^{1/2}$ is a projection matrix, it follows that

$$\operatorname{tr}(\mathbf{W}_{u_i}^{1/2} \mathbf{P}^{(i)} \mathbf{W}_{u_i}^{1/2}) = \sum_{j=1}^{n} w_j^{(i)} \mathbf{P}_{jj}^{(i)} = 2 s_{u_i}. \tag{5.8.26}$$

Therefore, we have

$$\begin{aligned}
&\operatorname{tr}(\mathbf{W}_u^* \mathbf{S}) = \operatorname{tr}(\mathbf{S}^T \mathbf{W}_u^*) \\
&= \sum_{i=1}^{n} w_{u,i}^* \mathbf{e}_{n,i}^T \mathbf{D}_{u_i} (\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i} \mathbf{W}_{u_i} \mathbf{e}_{n,i} \\
&= \sum_{i=1}^{n} w_{u,i}^* w_i^{(i)} \mathbf{P}_{ii}^{(i)} \\
&= \frac{K(0)}{h} \sum_{i=1}^{n} w_{u,i}^* \mathbf{P}_{ii}^{(i)} \\
&= \frac{K(0)}{h} s_{u_i} + o_p(1).
\end{aligned} \tag{5.8.27}$$

By using equation (5.8.47) and Lemma 1, the last equation holds for

$$\begin{aligned}
&\sum_{i=1}^{n} w_{u,i}^* \mathbf{P}_{ii}^{(i)} \\
&= \sum_{i=1}^{n} w_{u,i}^* \mathbf{e}_{n,i}^T \mathbf{D}_{u_i} (\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{e}_{n,i} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^*} K\left(\frac{U_i - u}{h^*}\right) f_U^{-1}(U_i) \left[\mathbf{x}_i^T E\left((\mathbf{X}_{\mathcal{S}_{u_i}} \mathbf{X}_{\mathcal{S}_{u_i}}^T)^{-1} | U_i\right) \mathbf{x}_i\right] \left(1 + O_p\left(h^2 + \sqrt{\frac{\log(1/h)}{nh}}\right)\right) \\
&= E\left\{\frac{1}{h^*} K\left(\frac{U_i - u}{h^*}\right) f_U^{-1}(U_i) \mathbf{x}_i^T E\left[(\mathbf{X}_{\mathcal{S}_{u_i}} \mathbf{X}_{\mathcal{S}_{u_i}}^T)^{-1} | U_i\right] \mathbf{x}_i\right\} \left(1 + O_p\left(\sqrt{\frac{\log(1/h^*)}{nh^*}}\right)\right) \left(1 + O_p\left(h^2 + \sqrt{\frac{\log(}{n}\right.\right.
\end{aligned} \tag{5.8.28}$$

and

$$
\begin{aligned}
&\mathrm{E}\left\{\frac{1}{h^*}K(\frac{U_i-u}{h^*})f_U^{-1}(U_i)\mathbf{x}_i^T\mathrm{E}\left[(\mathbf{X}_{\mathcal{S}_{u_i}}\mathbf{X}_{\mathcal{S}_{u_i}}^T)^{-1}\,|U_i\right]\mathbf{x}_i\right\}\\
=&\,\mathrm{E}\left\{\frac{1}{h^*}K(\frac{U_i-u}{h^*})f_U^{-1}(U_i)\,\mathrm{tr}\big(\mathrm{E}\left[(\mathbf{X}_{\mathcal{S}_{u_i}}\mathbf{X}_{\mathcal{S}_{u_i}}^T)^{-1}\,|U_i\right]\mathbf{x}_i\mathbf{x}_i^T\big)\right\}\\
=&\,\mathrm{tr}\left(\mathrm{E}_U\left\{\frac{1}{h^*}K(\frac{U_i-u}{h^*})f_U^{-1}(U_i)\,\mathrm{E}\left[(\mathbf{X}_{\mathcal{S}_{u_i}}\mathbf{X}_{\mathcal{S}_{u_i}}^T)^{-1}\,|U_i\right]\mathrm{E}\left(\mathbf{x}_i\mathbf{x}_i^T\,|U_i\right)\right\}\right)\\
=&\,\mathrm{E}_U\left(\frac{1}{h^*}K(\frac{U_i-u}{h^*})f_U^{-1}(U_i)\,s_{u_i}\right).
\end{aligned}
\tag{5.8.29}
$$

Notice that the sample $\mathbf{x}_i$ consists of variables belonging to the active set $\mathcal{S}_{u_i}$. Together with equation (5.8.28) and equation (5.8.29), we have

$$
\sum_{i=1}^{n} w_{u,i}^*\mathbf{P}_{ii}^{(i)} = \; s_{u_i}\,(1+O_p(\sqrt{\frac{\log(1/h^*)}{nh^*}}))(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}})) = \; s_{u_i}+o_p(1).
\tag{5.8.30}
$$

Similarly, we can calculate the order of term

$$
\begin{aligned}
&\mathrm{tr}(\mathbf{S}^T\mathbf{W}_u^*\mathbf{S})\\
=&\;\sum_{i=1}^{n} w_{u,i}^*\mathbf{e}_{n,i}^T\mathbf{P}^{(i)}\mathbf{W}_{u_i}^2\mathbf{P}^{(i)}\mathbf{e}_{n,i}\\
=&\;\sum_{i=1}^{n} w_{u,i}^*(\mathbf{P}_{i1}^{(i)},\cdots,\mathbf{P}_{in}^{(i)})\mathbf{W}_{u_i}^2(\mathbf{P}_{i1}^{(i)},\cdots,\mathbf{P}_{in}^{(i)})^T\\
=&\;\sum_{i=1}^{n}\sum_{k=1}^{n} w_{u,i}^*\,(w_k^{(i)})^2\,(\mathbf{P}_{ik}^{(i)})^2\\
\le&\;\sum_{i=1}^{n}\sum_{k=1}^{n} w_{u,i}^*\,(w_k^{(i)})^2\,\mathbf{P}_{kk}^{(i)}\,\mathbf{P}_{ii}^{(i)}\\
\le&\;\sum_{i=1}^{n} w_{u,i}^*\,\mathbf{P}_{ii}^{(i)}\,\frac{K(0)}{h}\sum_{k=1}^{n} w_k^{(i)}\,\mathbf{P}_{kk}^{(i)}\\
=&\;2\frac{K(0)}{h}\,s_u\sum_{i=1}^{n} w_{u,i}^*\,\mathbf{P}_{ii}^{(i)}\\
=&\;2\frac{K(0)}{h}\,s_u^2.
\end{aligned}
\tag{5.8.31}
$$

As a result, it follows that

$$f_U(u)\,n - 2\frac{K(0)}{h}\,s_u \le c(u) \le f_U(u)\,n - 2\frac{K(0)}{h}\,s_u + 2\frac{K(0)}{h}\,s_u^2 \tag{5.8.32}$$

In the other hand, as an simplified version that $h^* = h$, we have

$$\begin{aligned} &\operatorname{tr}\left(\mathbf{W}_u - (\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u^2\mathbf{D}_u\right) \\ &= \operatorname{tr}(\mathbf{W}_u - \mathbf{W}_u\mathbf{D}_u(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u) \\ &= \operatorname{tr}(\mathbf{W}_u(\mathbf{I} - \mathbf{W}_u^{1/2}\mathbf{D}_u(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u^{1/2})) \end{aligned} \tag{5.8.33}$$

Denote by $\mathbf{P}_u = \{\mathbf{P}_{ij}(u)\}_{i,j} = \mathbf{W}_u^{1/2}\mathbf{D}_u(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u^{1/2}$. Notice that $\mathbf{I} - \mathbf{P}_u$ is a projection matrix satisfying that $\operatorname{rank}(\mathbf{I} - \mathbf{P}_u) = \operatorname{tr}(\mathbf{I} - \mathbf{P}_u)$. For $\mathbf{W}_u = \operatorname{diag}(w_{n,1}, \cdots, w_{n,n})$, equation (5.8.33) equals $\sum_i w_{u,i}(1 - \mathbf{P}_{ii}(u))$ and

$$\begin{aligned} f_U(u)n - \frac{K(0)}{h}s_u \le \sum_{i=1}^n w_{u,i} - \sum_{i=1}^n \mathbf{P}_{ii}(u) &\le \sum_{i=1}^n w_{u,i}(1 - \mathbf{P}_{ii}(u)) \\ &\le \frac{K(0)}{h}\sum_{i=1}^n (1 - \mathbf{P}_{ii}(u)) = \frac{K(0)}{h}\,(n - s_u). \end{aligned} \tag{5.8.34}$$

The proof of Proposition 1 is completed.

Proof of Theorem 1. and Theorem 2. Recall the varying-coefficient model

$$y_i = \sum_{j=1}^p \alpha_j(U_i)X_{ij} + \varepsilon_i, \tag{5.8.35}$$

where $\alpha_j(U), j = 1, \cdots, p$, are unknown functional coefficients. For each function $\alpha(u)$, we apply locally linear approximation in the neighbor of fixed point u_0 that

$$\alpha(u) \approx \alpha(u_0) + \alpha'(u_0)(u - u_0), \quad u \in \mathcal{N}_\varepsilon(u_0). \tag{5.8.36}$$

For local linear regression, the design matrix is defined by

$$\mathbf{D}_u = \begin{pmatrix} \mathbf{x}_1^T & \frac{U_1-u}{h}\mathbf{x}_1^T \\ \vdots & \vdots \\ \mathbf{x}_n^T & \frac{U_n-u}{h}\mathbf{x}_n^T \end{pmatrix}. \tag{5.8.37}$$

The local linear regression estimator of $\left(\boldsymbol{\alpha}^T(u), (\boldsymbol{\alpha}'(u))^T\right)^T$ is a weighted least squares estimator defined by

$$(\alpha_1(u), \cdots, \alpha_{s_u}(u), \alpha_1'(u)h, \cdots, \alpha_{s_u}'(u)h)^T = (\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u\mathbf{Y}, \tag{5.8.38}$$

where $\mathbf{W}_u$ is the weights matrix based on kernel function. Let

$$\mathbf{M} = \begin{pmatrix} \boldsymbol{\alpha}^T(U_1)\mathbf{x}_1 \\ \vdots \\ \boldsymbol{\alpha}^T(U_n)\mathbf{x}_n \end{pmatrix}, \quad \widetilde{\mathbf{M}}_u = \begin{pmatrix} \boldsymbol{\alpha}^T(u)\mathbf{x}_1 + (\boldsymbol{\alpha}'(u))^T(U_1-u)\mathbf{x}_1 \\ \vdots \\ \boldsymbol{\alpha}^T(u)\mathbf{x}_n + (\boldsymbol{\alpha}'^T(u))^T(U_n-u)\mathbf{x}_n \end{pmatrix}, \tag{5.8.39}$$

and

$$\widehat{\mathbf{M}} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}^T(U_1)\mathbf{x}_1 \\ \vdots \\ \widehat{\boldsymbol{\alpha}}^T(U_n)\mathbf{x}_n \end{pmatrix} = \begin{pmatrix} (\mathbf{x}_1^T, 0)(\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^T\mathbf{W}_{u_1} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{x}_n^T, 0)(\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^T\mathbf{W}_{u_n} \end{pmatrix}\mathbf{Y} \triangleq \mathbf{S}\mathbf{Y}, \tag{5.8.40}$$

where $\mathbf{S}$ is called the smooth matrix. The corresponding estimators of errors are $\widehat{\boldsymbol{\varepsilon}} = (\mathbf{I}-\mathbf{S})\mathbf{Y}$. The local error variance estimator is defined by

$$\widehat{\sigma}^2(u) = \frac{\widehat{\boldsymbol{\varepsilon}}^T\mathbf{W}_u^*\widehat{\boldsymbol{\varepsilon}}}{c(u)} = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\mathbf{Y}}{c(u)}, \tag{5.8.41}$$

where $c(u)$ is used to obtain the nearly unbiased estimator. The denominator could

be divided into three parts.

$$\begin{aligned}
&(\mathbf{M}^{\mathsf{T}}+\boldsymbol{\varepsilon}^{\mathsf{T}})(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})(\mathbf{M}+\boldsymbol{\varepsilon})\\
&=\mathbf{M}^{\mathsf{T}}(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\mathbf{M}+2\mathbf{M}^{\mathsf{T}}(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\boldsymbol{\varepsilon}\\
&+\boldsymbol{\varepsilon}(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\boldsymbol{\varepsilon}\\
&\triangleq \boldsymbol{\Delta}_1+\boldsymbol{\Delta}_2+\boldsymbol{\Delta}_3
\end{aligned} \tag{5.8.42}$$

First, we study

$$\begin{aligned}
&\mathbf{W}_u^{*1/2}\mathbf{S}\boldsymbol{\varepsilon}\\
&=\mathbf{W}_u^{*1/2}\begin{pmatrix}\mathbf{e}_{1,n}^{\mathsf{T}}\mathbf{D}_{u_1}(\mathbf{D}_{u_1}^{\mathsf{T}}\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^{\mathsf{T}}\mathbf{W}_{u_1}\\ \cdots\quad\cdots\quad\cdots\\ \mathbf{e}_{n,n}^{\mathsf{T}}\mathbf{D}_{u_n}(\mathbf{D}_{u_n}^{\mathsf{T}}\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^{\mathsf{T}}\mathbf{W}_{u_n}\end{pmatrix}\boldsymbol{\varepsilon}\\
&=\begin{pmatrix}w_{u,1}^{*1/2}\mathbf{e}_{1,n}^{\mathsf{T}}\mathbf{D}_{u_1}(\mathbf{D}_{u_1}^{\mathsf{T}}\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^{\mathsf{T}}\mathbf{W}_{u_1}\boldsymbol{\varepsilon}\\ \cdots\quad\cdots\quad\cdots\\ w_{u,n}^{*1/2}\mathbf{e}_{n,n}^{\mathsf{T}}\mathbf{D}_{u_n}(\mathbf{D}_{u_n}^{\mathsf{T}}\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^{\mathsf{T}}\mathbf{W}_{u_n}\boldsymbol{\varepsilon}\end{pmatrix}\triangleq\mathbf{H}
\end{aligned} \tag{5.8.43}$$

$$\left\|\mathbf{W}_u^{*1/2}\mathbf{S}\boldsymbol{\varepsilon}\right\|^2=\mathbf{H}^{\mathsf{T}}\mathbf{H}=(h_1,\cdots,h_n)(h_1,\cdots,h_n)^{\mathsf{T}}=\sum_{j=1}^{n}h_j^2. \tag{5.8.44}$$

For each $h_j,\ j=1,\cdots,n$, we have

$$\begin{aligned}
h_j^2=\ &\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j}(\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j})^{-1}\mathbf{D}_{u_j}^{\mathsf{T}}\left(w_{u,j}^{*1/2}\mathbf{e}_{n,j}\mathbf{e}_{n,j}^{\mathsf{T}}w_{u,j}^{*1/2}\right)\mathbf{D}_{u_j}(\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j})^{-1}\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\boldsymbol{\varepsilon}\\
\le\ &\lambda_{\max}\left[(\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j})^{-1}\mathbf{D}_{u_j}^{\mathsf{T}}\left(w_{u,j}^{*1/2}\mathbf{e}_{n,j}\mathbf{e}_{n,j}^{\mathsf{T}}w_{u,j}^{*1/2}\right)\mathbf{D}_{u_j}(\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j})^{-1}\right]\left\|\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\boldsymbol{\varepsilon}\right\|_\infty^2
\end{aligned}$$

Notice that face

$$\lambda_{\max}(\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})\le\lambda_{\max}(\mathbf{B})\,\lambda_{\max}(\mathbf{A}^{-2})=\lambda_{\max}(\mathbf{B})\,\lambda_{\min}^{-2}(\mathbf{A}). \tag{5.8.45}$$

Thus, the last equation is bounded by

$$\lambda_{\max}\left[\mathbf{D}_{u_j}^T\left(w_{u,j}^{*1/2}\mathbf{e}_{n,j}\mathbf{e}_{n,j}^T w_{u,j}^{*1/2}\right)\mathbf{D}_{u_j}\right]\ (\lambda_{\min}(\mathbf{D}_{u_j}^T\mathbf{W}_{u_j}\mathbf{D}_{u_j}))^{-2}. \tag{5.8.46}$$

With Lemma 1 and the symmetry of kernel function, it holds uniformly in u that

$$\begin{aligned}
&\frac{1}{n}\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u \\
&= \begin{pmatrix} n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T K_h(U_i-u) & n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\left(\frac{U_i-u}{h}\right)K_h(U_i-u) \\ n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\left(\frac{U_i-u}{h}\right)K_h(U_i-u) & n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\left(\frac{U_i-u}{h}\right)^2 K_h(U_i-u)\end{pmatrix} \\
&= \begin{pmatrix} f_U(u)E\left(\mathbf{x}\mathbf{x}^T|u\right) & 0 \\ 0 & \mu_2 f_U(u)E\left(\mathbf{x}\mathbf{x}^T|u\right)\end{pmatrix}(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}}))
\end{aligned} \tag{5.8.47}$$

Notice the fact that positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, then

$$\lambda_{\min}\begin{pmatrix}\mathbf{A} & 0\\ 0 & \mathbf{B}\end{pmatrix} = \min(\lambda_{\min}(\mathbf{A}),\lambda_{\min}(\mathbf{B})). \tag{5.8.48}$$

We also have

$$\begin{aligned}
&\lambda_{\min}(f_U(u)E\left(\mathbf{x}\mathbf{x}^T|u\right)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}})))\\
&=\lambda_{\min}(E\left(\mathbf{x}\mathbf{x}^T|u\right))f_U(u)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}}))\\
&\geq \lambda_0 f_U(u)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}}))\\
&\lambda_{\min}(\mu_2 f_U(u)E\left(\mathbf{x}\mathbf{x}^T|u\right)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}})))\\
&=\lambda_{\min}(E\left(\mathbf{x}\mathbf{x}^T|u\right))\mu_2 f_U(u)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}}))\\
&\geq \lambda_0\mu_2 f_U(u)(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}}))
\end{aligned} \tag{5.8.49}$$

As the result, it follows

$$\lambda_{\min}\left(\frac{1}{n}\mathbf{D}_{u_j}^{\mathsf{T}}\mathbf{W}_{u_j}\mathbf{D}_{u_j}\right) \geq \min\{1,\mu_2\}\lambda_0 f_U(u)\,(1+O_p(h^2+\sqrt{\frac{\log(1/h)}{nh}})) \quad (5.8.50)$$

Next we study the first term in equation (5.8.46). We have

$$\begin{aligned}
&\mathbf{D}_{u_j}^{\mathsf{T}}\left(w_{u,j}^{*1/2}\mathbf{e}_{n,j}\mathbf{e}_{n,j}^{\mathsf{T}}w_{u,j}^{*1/2}\right)\mathbf{D}_{u_j}\\
&=\begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_j & \cdots & \mathbf{x}_n \\ \mathbf{x}_1\frac{u_1-u_j}{h} & \cdots & 0 & \cdots & \mathbf{x}_n\frac{u_n-u_j}{h}\end{pmatrix}\begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & w_{u,j}^* & & \\ & & & \ddots & \\ & & & & 0\end{pmatrix}\begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} & \frac{u_1-u_j}{h}\mathbf{x}_1^{\mathsf{T}} \\ \vdots & \vdots \\ \mathbf{x}_j^{\mathsf{T}} & 0 \\ \vdots & \vdots \\ \mathbf{x}_n^{\mathsf{T}} & \frac{u_n-u_j}{h}\mathbf{x}_n^{\mathsf{T}}\end{pmatrix}\\
&=\begin{pmatrix} 0 & \cdots & w_{u,j}^{*1/2}\mathbf{x}_j & \cdots & 0\end{pmatrix}\begin{pmatrix} 0 & \cdots & w_{u,j}^{*1/2}\mathbf{x}_j & \cdots & 0\end{pmatrix}^{\mathsf{T}} = w_{n,j}^*\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}.
\end{aligned} \quad (5.8.51)$$

Thus, for $X_i\, i=1,\cdots,p$, has bounded support, it follows

$$\begin{aligned}
&\lambda_{\max}\left[\mathbf{D}_{u_j}^{\mathsf{T}}\left(w_{u,j}^{*1/2}\mathbf{e}_{n,j}\mathbf{e}_{n,j}^{\mathsf{T}}w_{u,j}^{*1/2}\right)\mathbf{D}_{u_j}\right]\\
&= w_{n,j}^*\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j \qquad (5.8.52)\\
&\leq \frac{K(0)}{h^*}\|\mathbf{x}_j\|^2 = \frac{K(0)}{h^*}O(n).
\end{aligned}$$

To find the order of $\mathbf{D}_u^T \mathbf{W}_u \boldsymbol{\varepsilon}$, we notice that

$$\begin{aligned}
&\mathbf{D}_u^T \mathbf{W}_u \boldsymbol{\varepsilon} \\
&= \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ \mathbf{x}_1 \frac{U_1-u}{h} & \cdots & \mathbf{x}_n \frac{U_n-u}{h} \end{pmatrix} \begin{pmatrix} w_{u,1}\varepsilon_1 \\ \vdots \\ w_{u,n}\varepsilon_n \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n w_{u,i}\mathbf{x}_i\varepsilon_i \\ \sum_{i=1}^n w_{u,i}\frac{U_i-u}{h}\mathbf{x}_i\varepsilon_i \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n w_{u,i}X_{i1}\varepsilon_i \\ \vdots \\ \sum_{i=1}^n w_{u,i}X_{is_u}\varepsilon_i \\ \sum_{i=1}^n w_{u,i}\frac{U_i-u}{h}X_{i1}\varepsilon_i \\ \vdots \\ \sum_{i=1}^n w_{u,i}\frac{U_i-u}{h}X_{is_u}\varepsilon_i \end{pmatrix}
\end{aligned} \tag{5.8.53}$$

is a $2s_u$-by-1 vector. For the symmetry of kernel function, $\int uK(u)du = 0$ and the last s_u terms have the same expectation 0. Using Lemma 4, we can uniformly construct the probabilistic bound of $\mathbf{D}_u^T \mathbf{W}_u \boldsymbol{\varepsilon}$.

$$\begin{aligned}
&P\Big(\max_{j=1,\cdots,p}\sup_u \left|\frac{1}{n}\sum_{i=1} K_h(U_i-u)X_{ij}\varepsilon_i\right| \geq M\Big) \\
\leq\ &\sum_{j=1}^p P\Big(\sup_u \left|\frac{1}{n}\sum_{i=1} K_h(U_i-u)X_{ij}\varepsilon_i\right| \geq M\Big) \\
\leq\ &\sum_{j=1}^p 4\exp\{-2c_1M^2nh^2\} \\
=\ &4\exp\{\log(p)-2c_1M^2nh^2\} \\
=\ &4\exp\{\log(p)\big(1-\frac{2c_1M^2nh^2}{\log(p)}\big)\}.
\end{aligned} \tag{5.8.54}$$

Let $M = c_2\sqrt{\frac{\log(p)}{2nh^2}}$, when c_2 is sufficiently large, the power is negative. Together with equation (5.8.44), (5.8.50) and (5.8.52), we have

$$\widehat{\gamma}_n^2 = \frac{\varepsilon^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\varepsilon}{\varepsilon^T \mathbf{W}_u^* \varepsilon} = O_p(\frac{s_u \log(p)}{nh^2 h^*}). \tag{5.8.55}$$

In here the order of $\varepsilon^T \mathbf{W}_u^* \varepsilon$ is obtain by LLN. The proof of Theorem 1 is completed.

Next, we study the limiting distribution of the leading term $\frac{1}{n}\varepsilon^T \mathbf{W}_u^* \varepsilon^T$. The expectation is

$$\begin{aligned}
&E\left(n^{-1}\varepsilon^T \mathbf{W}_u^* \varepsilon^T\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(K_{h^*}(U_i - u)\,\varepsilon_i^2\right) \\
&= E\left[E\left(\frac{1}{h^*}K(\frac{U_i - u}{h^*})\varepsilon_i^2 \,\middle|\, U_i\right)\right] \\
&= E\left[\frac{1}{h^*}K(\frac{U_i - u}{h^*})\,E\left(\varepsilon_i^2 \,\middle|\, U_i\right)\right] \\
&= \int \sigma^2(u_i)\frac{1}{h^*}K(\frac{u_i - u}{h^*}) f_U(u_i)\,du_i,
\end{aligned} \tag{5.8.56}$$

Denote by $\widehat{f}_U(u) = n^{-1}\sum_{i=1}^{n} K_h^*(U_i - u)$ and $\widehat{f}_U(u)$ converges to $f_U(u)$ in probability for any given $u \in \mathcal{U}$.

$$\begin{aligned}
&E\left(n^{-1}\varepsilon^T \mathbf{W}_u^* \varepsilon^T - \sigma^2(u)\widehat{f}_U(u)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(K_{h^*}(U_i - u)\,(\varepsilon_i^2 - \sigma^2(u)\right) \\
&= E\left[E\left(\frac{1}{h^*}K(\frac{U_i - u}{h^*})(\varepsilon_i^2 - \sigma^2(u)) \,\middle|\, U_i\right)\right] \\
&= E\left[\frac{1}{h^*}K(\frac{U_i - u}{h^*})\left(E\left(\varepsilon_i^2 \,\middle|\, U_i\right) - \sigma^2(u)\right)\right] \\
&= \int \left(\sigma^2(u_i) - \sigma^2(u)\right)\frac{1}{h^*}K(\frac{u_i - u}{h^*}) f_U(u_i)\,du_i,
\end{aligned} \tag{5.8.57}$$

Denote by $\tilde{u} = (U_i - u)/h^*$. Using Taylor series approximation, it follows

$$\begin{aligned}\sigma^2(u + \tilde{u}h^*) - \sigma^2(u) &= (\sigma^2(u))'\tilde{u}h^* + O(h^{*2}) \\ f_U(u + \tilde{u}h^*) &= f_U(u) + f'_U(u)\tilde{u}h^* + O(h^{*2})\end{aligned}$$

By changing variable in the integral, we got the nonparametric bias

$$\begin{aligned}&\int (\sigma^2(u + \tilde{u}h^*) - \sigma^2(u))K(\tilde{u})f_U(u + \tilde{u}h^*)\, d\tilde{u} \\ &= \int \left((\sigma^2(u))'\tilde{u}h^* + \frac{1}{2}(\sigma^2(u))''\tilde{u}^2h^{*2} + o(h^{*2})\right) \\ &\qquad \times \left(f_U(u) + f'_U(u)\tilde{u}h^* + O(h^{*2})\right) K(\tilde{u})\, d\tilde{u} \\ &= ((\sigma^2(u))'f'_U(u))\mu_2h^{*2} + \frac{1}{2}f_U(u)(\sigma^2(u))''\mu_2h^{*2} \\ &= h^{*2}\mu_2\big((\sigma^2(u))'f'_U(u)) + \frac{1}{2}(\sigma^2(u))''f_U(u)\big).\end{aligned} \tag{5.8.58}$$

Similarly, we have the variance

$$\begin{aligned}&\mathrm{Var}\left(n^{-1}\varepsilon^T\mathbf{W}^*_u\varepsilon^T\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(K_{h^*}(U_i - u)\,\varepsilon_i^2) \\ &= \mathrm{E}\left(\varepsilon_i^2\frac{1}{h^*}K(\frac{U_i - u}{h^*})\right)^2 - \left[\mathrm{E}\left(\varepsilon_i\frac{1}{h^*}K(\frac{U_i - u}{h^*})\right)\right]^2 \\ &= \mathrm{E}\left(\varepsilon_i^4\frac{1}{h^{*2}}K^2(\frac{U_i - u}{h^*})\right) - \left[((\sigma^2(u))^2f_U^2(u)(1 + O(h^{*2}) + O(h^{*4}))\right]\end{aligned} \tag{5.8.59}$$

Denote by $\kappa(U_i) = \mathrm{E}\,(\varepsilon_i^4|U_i)$. For the first term, we apply the similar strategy for

the expectation and obtain

$$\begin{aligned}
&\mathrm{E}\left(\varepsilon_i^4 \frac{1}{h^{*2}} K^2\left(\frac{U_i - u}{h^*}\right)\right) \\
&= \frac{1}{h^*}\int \left(\kappa(u) + \kappa'(u)\tilde{u}h^* + O(h^{*2})\right)\left(f_U(u) + f_U'(u)\tilde{u}h^* + O(h^{*2})\right) K^2(\tilde{u})\, d\tilde{u} \\
&= \frac{1}{h^*}\kappa(u) f_U(u)\nu_0 + h^*\kappa'(u) f_U'(u)\nu_2 + O(h^*) \\
&= \frac{1}{h^*}\kappa(u) f_U(u)\nu_0 (1 + O(h^{*2})).
\end{aligned} \tag{5.8.60}$$

Thus, the variance equals to

$$\mathrm{Var}\left(n^{-1}\boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon}^T\right) = \left(\frac{1}{h^*}\kappa(u) f_U(u)\nu_0 - (\sigma^2(u))^2 f_U^2(u)\right)(1 + O(h^{*2})). \tag{5.8.61}$$

By CLT, we have that

$$\sqrt{\frac{h^*}{n}}\left(\boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon}^T - \sigma^2(u) f_U(u) - \mathrm{bias}_{\text{n-para}} f_U(u)\right) \xrightarrow{d} \mathcal{N}(0, \kappa(u) f_U(u)\nu_0), \tag{5.8.62}$$

where the bias term equals to $h^{*2}\mu_2\big((\sigma^2(u))' f_U'(u)) + \frac{1}{2}(\sigma^2(u))'' f_U(u)\big)$. We also have the fact that

$$\begin{aligned}
&\boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{S})^T \mathbf{W}_u^* (\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon} \\
&\geq \boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon} - 2\left\|\mathbf{W}_u^{*1/2}\mathbf{S}\boldsymbol{\varepsilon}\right\| \left\|\mathbf{W}_u^{*1/2}\boldsymbol{\varepsilon}\right\| + \boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon}\left(1 - \sqrt{\frac{\boldsymbol{\varepsilon}^T \mathbf{S}^T \mathbf{W}_u^* \mathbf{S}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T \mathbf{W}_u^* \boldsymbol{\varepsilon}}}\right)^2.
\end{aligned} \tag{5.8.63}$$

Term $\mathbf{M}^T(\mathbf{I} - \mathbf{S})^T \mathbf{W}_u^* (\mathbf{I} - \mathbf{S})\mathbf{M}$. Define

$$\boldsymbol{\beta}_u = (\beta_{u,1}, \cdots, \beta_{u,s_u})^T, \quad \beta_{u,i} = \sum_{j=1}^{s_u}\left(\alpha_j(U_i) - \alpha_j(u) - \alpha_j'(u)(U_i - u)\right) X_{ij}.$$

By Taylor expansion, it can be bounded by

$$\beta_{u,i} = \sum_{j=1}^{s_u} \left(\frac{1}{2}\alpha_j''(u)(U_i - u)^2 + o((U_i - u)^2) \right) X_{ij}. \tag{5.8.64}$$

Consider the term

$$\mathbf{SM} = \begin{pmatrix} (\mathbf{x}_1^T, 0)(\mathbf{D}_{u_1}^T \mathbf{W}_{u_1} \mathbf{D}_{u_1})^{-1} \mathbf{D}_{u_1}^T \mathbf{W}_{u_1} \mathbf{M} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{x}_n^T, 0)(\mathbf{D}_{u_n}^T \mathbf{W}_{u_n} \mathbf{D}_{u_n})^{-1} \mathbf{D}_{u_n}^T \mathbf{W}_{u_n} \mathbf{M} \end{pmatrix}, \tag{5.8.65}$$

for i-th row, we have

$$\begin{aligned} & (\mathbf{x}_i^T, 0)(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{M} \\ = & (\mathbf{x}_i^T, 0)(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} (\mathbf{M} - \widetilde{\mathbf{M}} + \widetilde{\mathbf{M}}) \\ = & (\mathbf{x}_i^T, 0)(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \boldsymbol{\beta}_{u_i} + (\mathbf{x}_i^T, 0)(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \widetilde{\mathbf{M}}. \end{aligned} \tag{5.8.66}$$

Notice that

$$\widetilde{\mathbf{M}} = \begin{pmatrix} \boldsymbol{\alpha}^T(U_i)\mathbf{x}_1 + (\boldsymbol{\alpha}'(U_i))^T h \frac{U_1 - U_i}{h} \mathbf{x}_1 \\ \vdots \\ \boldsymbol{\alpha}^T(U_i)\mathbf{x}_n + (\boldsymbol{\alpha}'(U_i))^T h \frac{U_n - U_i}{h} \mathbf{x}_n \end{pmatrix} = \mathbf{D}_{u_i} \begin{pmatrix} \boldsymbol{\alpha}(U_i) \\ \boldsymbol{\alpha}'(U_i) h \end{pmatrix}, \tag{5.8.67}$$

then

$$(\mathbf{x}_i^T, 0)(\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \widetilde{\mathbf{M}} = \boldsymbol{\alpha}^T(U_i)\mathbf{x}_i. \tag{5.8.68}$$

Using equation (5.8.47) and the similar argument again, we get

$$(\frac{1}{n}\mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \mathbf{D}_{u_i})^{-1} \frac{1}{n} \mathbf{D}_{u_i}^T \mathbf{W}_{u_i} \begin{pmatrix} (U_1 - U_i)\mathbf{x}_1 \\ \vdots \\ (U_n - U_i)\mathbf{x}_n \end{pmatrix} \begin{pmatrix} \frac{1}{2}\alpha_1''(U_i) \\ \vdots \\ \frac{1}{2}\alpha_{s_u}''(U_i) \end{pmatrix} = O_p(h^2). \tag{5.8.69}$$

According to the above results, we have

$$
\begin{aligned}
&\mathbf{M}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\mathbf{M} \\
=& \begin{pmatrix} (\mathbf{x}_1^T,0)(\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\boldsymbol{\beta}_{u_1} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{x}_n^T,0)(\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\boldsymbol{\beta}_{u_n} \end{pmatrix}^T \mathbf{W}_u^* \begin{pmatrix} (\mathbf{x}_1^T,0)(\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\boldsymbol{\beta}_{u_1} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{x}_n^T,0)(\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\boldsymbol{\beta}_{u_n} \end{pmatrix} \\
=& \sum_{i=1}^n w_{u,i}^*\boldsymbol{\beta}_{u_i}^T\mathbf{W}_{u_i}\mathbf{D}_{u_i}(\mathbf{D}_{u_i}^T\mathbf{W}_{u_i}\mathbf{D}_{u_i})^{-1}(\mathbf{x}_i^T,0)^T(\mathbf{x}_i^T,0)(\mathbf{D}_{u_i}^T\mathbf{W}_{u_i}\mathbf{D}_{u_i})^{-1}\mathbf{D}_{u_i}^T\mathbf{W}_{u_i}\boldsymbol{\beta}_{u_i} \\
=& \sum_{i=1}^n w_{u,i}^*\left[\mathbf{1}^T(\mathbf{x}_i^T,0)^T(\mathbf{x}_i^T,0)\mathbf{1}\right]O_p(h^4)
\end{aligned}
$$

By the similar arguments before, we have

$$
\sup_u \frac{\mathbf{M}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\mathbf{M}}{c(u)} = O_p(s_n^2h^4) \tag{5.8.70}
$$

For any n-dimensional non-negative definite matrix $\Sigma \geq 0$, define the Σ-inner product and Σ-norm as, $\mathbf{a},\mathbf{b}$ are n-vectors,

$$
\langle \mathbf{a},\mathbf{b}\rangle_\Sigma = \mathbf{a}^T\Sigma\mathbf{b}, \quad \|\mathbf{a}\|_\Sigma = (\mathbf{a}^T\Sigma\mathbf{a})^{1/2}. \tag{5.8.71}
$$

It also follows the Cauchy-Schwarz inequality,

$$
\langle \mathbf{a},\mathbf{b}\rangle_\Sigma \leq \|\mathbf{a}\|_\Sigma \|\mathbf{b}\|_\Sigma . \tag{5.8.72}
$$

The crossing term is bounded by

$$
\begin{aligned}
&\mathbf{M}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\boldsymbol{\varepsilon} \\
=& \sum_{i=1}^n w_{u,i}^*(\varepsilon_i-(\mathbf{S}\boldsymbol{\varepsilon})_i)(\mathbf{x}_i^T,0)(\mathbf{D}_{u_i}^T\mathbf{W}_{u_i}\mathbf{D}_{u_i})^{-1}\mathbf{D}_{u_i}^T\mathbf{W}_{u_i}\boldsymbol{\beta}_{u_i} \\
=& O_p(s_nnh^2\sqrt{\frac{-\log(h^*)}{nh^*}}).
\end{aligned} \tag{5.8.73}
$$

Thus,

$$\sup_{u} \frac{\mathbf{M}^{\mathsf{T}}(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u^*(\mathbf{I}-\mathbf{S})\boldsymbol{\varepsilon}}{c(u)} = O_p(s_n h^2 \sqrt{\frac{-\log(h^*)}{nh^*}}). \tag{5.8.74}$$

Together with equation (5.8.62), (**??**, (5.8.70) and (5.8.74), the proof of Theorem 2 is completed.

Proof of Theorem 3.

We split whole samples $(\mathbf{Y}_i, \mathbf{x}_i, U_i),\ i = 1, \cdots, n$, into two parts $(\mathbf{Y}_i^{(j)}, \mathbf{x}_i^{(j)}, U_i^{(j)}),\ i \in \mathcal{I}_j,\ j = 1, 2; |\mathcal{I}_j| = |\mathcal{I}_{3-j}| = n/2$. Though group LASSO procedure, we select two active index sets $\mathcal{S}_1$ and $\mathcal{S}_2$ for different data sets, respectively. And $s_j = |\mathcal{S}_j|\ ,\ j = 1, 2.$ The indices of true covariates is $\mathcal{S}$. Suppose the selected active index sets include the true one, that is $\mathcal{S} \subset \mathcal{S}_j,\ j = 1, 2.$ Denote by $\mathcal{X}_{\mathcal{I}_j}$ the σ-field generated by $\mathbf{X}_i,\ i \in \mathcal{I}_j,\ j = 1, 2.$

First we study the leading term $\boldsymbol{\varepsilon}^{\mathsf{T}}(\mathbf{I}-\mathbf{S})^{\mathsf{T}}\mathbf{W}_u(\mathbf{I}-\mathbf{S})\boldsymbol{\varepsilon}$. We observe that

$$\begin{aligned}
&\boldsymbol{\varepsilon}^{(2)^{\mathsf{T}}}(\mathbf{I}-\mathbf{S}^{(2)})^{\mathsf{T}}\mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)} \\
&= \boldsymbol{\varepsilon}^{(2)^{\mathsf{T}}}(\mathbf{I} - \begin{pmatrix} (\mathbf{X}_1^{(2)^{\mathsf{T}}}, 0)(\mathbf{D}_{u_1}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_1}^{(2)}\mathbf{D}_{u_1}^{(2)})^{-1}\mathbf{D}_{u_1}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_1}^{(2)} \\ \cdots \quad \cdots \quad \cdots \\ (\mathbf{X}_{n/2}^{(2)^{\mathsf{T}}}, 0)(\mathbf{D}_{u_{n/2}}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_{n/2}}^{(2)}\mathbf{D}_{u_{n/2}}^{(2)})^{-1}\mathbf{D}_{u_{n/2}}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_{n/2}}^{(2)} \end{pmatrix}^{\mathsf{T}} \mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)}.
\end{aligned} \tag{5.8.75}$$

We notice that

$$\begin{aligned}
&\mathrm{E}\,(\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^{\mathsf{T}}}K_h(U_i^{(2)} - u)\,|\mathcal{X}_{\mathcal{I}_2}) \\
&\triangleq \mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\frac{1}{h}K(\frac{U_i - u}{h})\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^{\mathsf{T}}}) \\
&= \mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}\left(\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^{\mathsf{T}}}K_h(U_i^{(2)} - u))\,\middle|U_i\right) \\
&= f_U(u)\,\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\mathbf{X}^{(2)}\mathbf{X}^{(2)^{\mathsf{T}}}|u)
\end{aligned} \tag{5.8.76}$$

by Lemma 1, it follows that

$$\frac{2}{n}\sum_{i\in\mathcal{I}_1}\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^T}K_h(U_i^{(2)}-u)=f_U(u)\,\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}|u)(1+O_p(h^2+\sqrt{\frac{-\log(h)}{nh}})). \tag{5.8.77}$$

Similarly,

$$\frac{2}{n}\sum_{i\in\mathcal{I}_1}\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^T}K_h(U_i^{(2)}-u)\left(\frac{U_i^{(2)}-u}{h}\right)=o_p(1), \tag{5.8.78}$$

$$\frac{2}{n}\sum_{i\in\mathcal{I}_1}\mathbf{X}_i^{(2)}\mathbf{X}_i^{(2)^T}K_h(U_i^{(2)}-u)\left(\frac{U_i^{(2)}-u}{h}\right)^2=\mu_2 f_U(u)\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}|u)(1+O_p(h^2+\sqrt{\frac{-\log(h)}{nh}})). \tag{5.8.79}$$

It follows that

$$\begin{aligned}&\left(\frac{2}{n}\mathbf{D}_u^{(2)^T}\mathbf{W}_u^{(2)}\mathbf{D}_u^{(2)}\right)^{-1}\\ =&\begin{pmatrix}f_U^{-1}(u)\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}[(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T})^{-1}|u] & 0\\ 0 & \mu_2^{-1}f_U^{-1}(u)\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}[(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T})^{-1}|u]\end{pmatrix}(1+O_p(h^2+\sqrt{\frac{-\log(h)}{nh}}))\end{aligned} \tag{5.8.80}$$

By the definition of constrained eigenvalue and the condition, we have

$$\inf_u\lambda_{\min}\left(\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}|u)\right)\geq\lambda_0^*. \tag{5.8.81}$$

Using the similar strategy to the term $\mathbf{D}_{u_i}^{(2)^T}\mathbf{W}_{u_i}^{(2)}\boldsymbol{\varepsilon}^{(2)}$, we obtain

$$\frac{2}{n}\mathbf{D}_{u_i}^{(2)^T}\mathbf{W}_{u_i}^{(2)}\boldsymbol{\varepsilon}^{(2)}=\begin{pmatrix}\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}\left[\mathbf{X}_i^{(2)}\boldsymbol{\varepsilon}^{(2)}K_h(U_i^{(2)}-u)\right]\\ \mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}\left[\mathbf{X}_i^{(2)}\boldsymbol{\varepsilon}^{(2)}K_h(U_i^{(2)}-u)\left(\frac{U_i^{(2)}-u}{h}\right)\right]\end{pmatrix}+O_p(\sqrt{\frac{-\log(h)}{nh}}). \tag{5.8.82}$$

Consequently,

$$\frac{2}{n}\mathbf{D}_{u_i}^{(2)^T}\mathbf{W}_{u_i}^{(2)}\boldsymbol{\varepsilon}^{(2)}=\begin{pmatrix}O_p(\sqrt{\frac{-\log(h)}{nh}})\\ O_p(\sqrt{\frac{-\log(h)}{nh}})\end{pmatrix}. \tag{5.8.83}$$

Substituting the results into the smoothing matrix $\mathbf{S}^{(2)}$, we have

$$\begin{aligned}&(\mathbf{X}_i^{(2)^T}, 0)(\mathbf{D}_{u_i}^{(2)^T}\mathbf{W}_{u_i}^{(2)}\mathbf{D}_{u_i}^{(2)})^{-1}\mathbf{D}_{u_i}^{(2)^T}\mathbf{W}_{u_i}^{(2)}\\ &= f_U^{-1}(u)\mathbf{X}_i^{(2)^T}\mathrm{E}_{\mathcal{X}_{\mathcal{I}_2}}[(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T})^{-1}\,|u]O_p(\sqrt{\frac{-\log(h)}{nh}})(1+o_p(1))\end{aligned} \tag{5.8.84}$$

Thus,

$$(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)} = \boldsymbol{\varepsilon}^{(2)}(1+O_p(\sqrt{\frac{-\log(h)}{nh}})), \tag{5.8.85}$$

and

$$\begin{aligned}&df^{-1}\boldsymbol{\varepsilon}^{(2)^T}(\mathbf{I}-\mathbf{S}^{(2)})^T\mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)}\\ &= df^{-1}\boldsymbol{\varepsilon}^{(2)^T}\mathbf{W}_u^{(2)}\boldsymbol{\varepsilon}^{(2)}(1+O_p(\sqrt{\frac{-\log(h)}{nh}}))\\ &= df^{-1}\sum_{i\in\mathcal{I}_1} w_{u,i}\varepsilon_i^2(1+O_p(\sqrt{\frac{-\log(h)}{nh}}))\end{aligned} \tag{5.8.86}$$

By CLT and the obtained results, it follows that

$$\sqrt{\frac{2}{nh}}\left(\sum_{i\in\mathcal{I}_1} w_{u,i}\varepsilon_i^2 - \sigma^2(u)f_U(u)\right) \xrightarrow{d} \mathcal{N}(0, \kappa(u)f_U(u)\nu_0), \tag{5.8.87}$$

Moreover by the order of degree of freedom,

$$\sqrt{nh}\left(\frac{1}{df}\sum_{i\in\mathcal{I}_1} w_{u,i}\varepsilon_i^2 - \sigma^2(u)\right) \xrightarrow{d} \mathcal{N}\big(0, \frac{\kappa(u)\nu_0}{f_U(u)}\big). \tag{5.8.88}$$

Next we deal with the the rest terms. We consider

$$\sup_u \frac{\mathbf{M}^{(2)^T}(\mathbf{I}-\mathbf{S}^{(2)})^T\mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\mathbf{M}^{(2)}}{df} = O_p(h^6) \tag{5.8.89}$$

and

$$\begin{aligned}
&\mathbf{M}^{(2)^{\mathsf{T}}}(\mathbf{I}-\mathbf{S}^{(2)})^{\mathsf{T}}\mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)} \\
&= \sum_{i\in\mathcal{I}_1} w_{u,i}(\varepsilon_i^{(2)} - (\mathbf{S}^{(2)}\boldsymbol{\varepsilon}^{(2)})_i)\left(\mathbf{X}_i^{(2)^{\mathsf{T}}}\boldsymbol{\alpha}_{\mathcal{S}_2}(U_i) - (\mathbf{X}_i^{(2)^{\mathsf{T}}},0)(\mathbf{D}_{u_i}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_i}^{(2)}\mathbf{D}_{u_i}^{(2)})^{-1}\mathbf{D}_{u_i}^{(2)^{\mathsf{T}}}\mathbf{W}_{u_i}^{(2)}\mathbf{M}^{(2)}\right)
\end{aligned}$$

Using the similar argument above, we show that

$$(\mathbf{X}^{(2)^{\mathsf{T}}},0)(\mathbf{D}_u^{(2)^{\mathsf{T}}}\mathbf{W}_u^{(2)}\mathbf{D}_u^{(2)})^{-1}\mathbf{D}_u^{(2)^{\mathsf{T}}}\mathbf{W}_u^{(2)}\mathbf{M}^{(2)} = \mathbf{X}^{(2)^{\mathsf{T}}}\boldsymbol{\alpha}_{\mathcal{S}_2}(u)(1+O_p(h^2+\sqrt{\frac{-\log(h)}{nh}})), \tag{5.8.90}$$

and

$$\varepsilon_i^{(2)} - (\mathbf{S}^{(2)}\boldsymbol{\varepsilon}^{(2)})_i = \varepsilon_i^{(2)}(1 + O_p(\sqrt{\frac{-\log(h)}{nh}})). \tag{5.8.91}$$

Plugging the results into, it follows that

$$\begin{aligned}
&\mathbf{M}^{(2)^{\mathsf{T}}}(\mathbf{I}-\mathbf{S}^{(2)})^{\mathsf{T}}\mathbf{W}_u^{(2)}(\mathbf{I}-\mathbf{S}^{(2)})\boldsymbol{\varepsilon}^{(2)} \\
&= \sum_{i\in\mathcal{I}_1} w_{u,i}\varepsilon_i^{(2)}\mathbf{X}_i^{(2)^{\mathsf{T}}}\boldsymbol{\alpha}_{\mathcal{S}_2}(U_i)\, O_p(h^2 + \sqrt{\frac{-\log(h)}{nh}})(1 + O_p(\sqrt{\frac{-\log(h)}{nh}}))
\end{aligned} \tag{5.8.92}$$

using the Lemma 2 again, we have

$$\frac{2}{n}\sum_{i\in\mathcal{I}_1} w_{u,i}\varepsilon_i^{(2)}\mathbf{X}_i^{(2)^{\mathsf{T}}}\boldsymbol{\alpha}_{\mathcal{S}_2}(U_i) = O_p(\sqrt{\frac{-\log(h)}{nh}}). \tag{5.8.93}$$

The proof is completed.

Chapter 6

Contribution Remark and Future Work

In this section, we summarize the conclusions and important contributions in this dissertation first, and open a discussion of extending our proposed RCV method to partial varying-coefficient model.

6.1 Contribution Remark

This dissertation consists of two parts. In the first HIV related research, our goal is to study the impact of HIV prevalence and incidence estimates when adding into new incidence assay data. To fulfil that goal, in Chapter 3, we first introduce our proposed method to incorporate assay data (Bao, Ye, and Hallett, 2014). This method is accurate and we consider it as the benchmark. However, the method has too large computing cost, which makes large simulation studies impossible. Therefore, we propose a new method to draw posterior distribution, sequential IMIS. It can significantly reduce the computing time, and it has very close results with the benchmark method. We also propose a new stopping criteria to avoid

false converge. The idea of S-IMIS is also extended to a more general framework. In Chapter 4, by using the proposed S-IMIS in Chapter 3, we further study the impact of new incidence assay data. We consider two impacts, one is the impact to incidence estimate, and another one is to impact to the change of incidence estimate over time. Also, we consider both one-time incidence assay data and time series assay data. The results would show that the impact is highly related to the original prevalence and incidence trend. We discuss the different kinds of trends and the impact to each of the trend.

In the second RCV research, our goal is to estimate the error variance function for ultrahigh dimension varying-coefficient model. We first propose a three-stage naive estimator to estimate the coefficients functions and error variance function. We show from both theoretical proofs and simulation results that this naive estimator is biased under ultrahigh dimension setting due to spurious correlation. Then, we further proposed RCV estimator for error variance function estimation. We also show from both theoretical proofs and simulation results that RCV estimator has good asymptotic properties.

6.2 Future Work

For ultrahigh dimension data, error variance function estimation is always a big problem. Let Y_i be the response and z_i be the predictors. Considering an extension of varying-coefficient model: semi-parametric varying coefficient model with the form:

$$Y_i = \nu_i' \gamma(z_i) + \delta(z_i) + u_i, i = 1, \ldots, n, \tag{6.2.1}$$

where $\gamma(z)$ is a vector of unknown smooth function of z, $\delta(\cdot)$ is an unknown function. Compared to our varying coefficient model (2.3.1), it allows more flexibility which adds a new term function of z_i, and, it avoids much of the "curse of dimensionality"

problem, since the nonparametric functions are restricted only to part of the variable z.

Under this model frame, when restrict one or some of the coefficients β to constants, we can generalize a more specific model,

$$Y_i = \omega_i'\gamma + x_i'\beta(z_i) + u_i, i = 1, \ldots, n, \tag{6.2.2}$$

where ω_i is a vector of variables whose coefficient γ is a vector of constant. This model is called partially linear varying-coefficient model. Compared to VCM and semiparametric VCM, it allows some variables having changing coefficients and some variables having constant coefficients. RCV method can be extend to both of these two models to estimate the varying error variance function.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Alkema, L., A. E. Raftery, and S. J. Clark (2007). Probabilistic projections of hiv prevalence using bayesian melding. *The Annals of Applied Statistics*, 229–248.

Bao, L. (2012). A new infectious disease model for estimating and projecting hiv/aids epidemics. *Sexually transmitted infections 88*(Suppl 2), i58–i64.

Bao, L. and A. E. Raftery (2010). A stochastic infection rate model for estimating and projecting national hiv prevalence rates. *Sexually transmitted infections 86*(Suppl 2), ii93–ii99.

Bao, L., J. A. Salomon, T. Brown, A. E. Raftery, and D. R. Hogan (2012). Modelling national hiv/aids epidemics: revised approach in the unaids estimation and projection package 2011. *Sexually transmitted infections*, sextrans–2012.

Bao, L., J. Ye, and T. B. Hallett (2014). Incorporating incidence information within the unaids estimation and projection package framework: a study based on simulated incidence assay data. *AIDS (London, England) 28*(4), S515.

Brown, T., L. Bao, J. W. Eaton, D. R. Hogan, M. Mahy, K. Marsh, B. M. Mathers, and R. Puckett (2014). Improvements in prevalence trend fitting and incidence estimation in epp 2013. *AIDS 28*, S415–S425.

Brown, T., L. Bao, A. E. Raftery, J. A. Salomon, R. F. Baggaley, J. Stover, and P. Gerland (2010). Modelling hiv epidemics in the antiretroviral era: the unaids estimation and projection package 2009. *Sexually transmitted infections*, sti–2010.

Brown, T., N. Grassly, G. Garnett, and K. Stanecki (2006). Improving projections at the country level: the unaids estimation and projection package 2005. *Sexually Transmitted Infections 82*(suppl 3), iii34–iii40.

Brown, T., J. Salomon, L. Alkema, A. Raftery, and E. Gouws (2008). Progress and challenges in modelling country-level hiv/aids epidemics: the unaids estimation and projection package 2007. *Sexually Transmitted Infections 84*(Suppl 1), i5–i10.

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

Burton, A. H. and T. E. Mertens (1998). Provisional country estimates of prevalent adult human immunodeficiency virus infections as of end 1994: a description of the methods. *International Journal of Epidemiology 27*(1), 101–107.

Busch, M. P., C. D. Pilcher, T. D. Mastro, J. Kaldor, G. Vercauteren, W. Rodriguez, C. Rousseau, T. M. Rehle, A. Welte, M. D. Averill, et al. (2010). Beyond detuning: 10 years of progress and new challenges in the development and application of assays for hiv incidence estimation. *Aids 24*(18), 2763–2771.

Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association 95*(451), 888–902.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*(451), 941–956.

Carroll, R. J., D. Ruppert, and A. H. Welsh (1998). Local estimating equations. *Journal of the American Statistical Association 93*(441), 214–227.

Chakraverty, S. (2014). *Mathematics of Uncertainty Modeling in the Analysis of Engineering and Science Problems*. IGI Global.

Chen, Z., J. Fan, and R. Li (2016). Error variance estimation in ultrahigh dimensional additive models. *Journal of the American Statistical Association* (just-accepted).

Chiang, C.-T., J. A. Rice, and C. O. Wu (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association 96*(454), 605–619.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pp. 260–265. ACM.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture 1*, 32.

Dziak, J. J., R. Li, X. Tan, S. Shiffman, and M. P. Shiyko (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological methods 20*(4), 444.

Eubank, R., C. Huang, Y. M. Maldonado, N. Wang, S. Wang, and R. Buchanan (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), 653–667.

Fan, J. and J. Chen (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(4), 927–943.

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association 106*(494), 544–557.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(1), 37–65.

Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. *National science review 1*(2), 293–314.

Fan, J., T. Huang, et al. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*(6), 1031–1057.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, J. and R. Li (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, J., Y. Ma, and W. Dai (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics*, 153–193.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 1491–1518.

Fan, J. and W. Zhang (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics 27*(4), 715–731.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Ghys, P. D., T. Brown, N. Grassly, G. Garnett, K. Stanecki, J. Stover, and N. Walker (2004). The unaids estimation and projection package: a software package to estimate and project national hiv epidemics. *Sexually transmitted infections 80*(suppl 1), i5–i9.

Ghys, P. D., E. Kufa, and M. George (2006). Measuring trends in prevalence and incidence of hiv infection in countries with generalised epidemics. *Sexually Transmitted Infections 82*(suppl 1), i52–i56.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 337–348.

Givens, G. H. and J. A. Hoeting (2012). *Computational statistics*, Volume 710. John Wiley & Sons.

Gouws, E., V. Mishra, and T. Fowler (2008). Comparison of adult hiv prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. *Sexually transmitted infections 84*(Suppl 1), i17–i23.

Hall, P. and H. Miller (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*(3), 533–550.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hogan, D. R. and J. A. Salomon (2012). Spline-based modelling of trends in the force of hiv infection, with application to the unaids estimation and projection package. *Sexually transmitted infections 88*(Suppl 2), i52–i57.

Hogan, D. R., A. M. Zaslavsky, J. K. Hammitt, and J. A. Salomon (2010). Flexible epidemiological model for estimates and short-term projections in generalised hiv/aids epidemics. *Sexually transmitted infections 86*(Suppl 2), ii84–ii92.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Kassanjee, R., T. A. McWalter, T. Bärnighausen, and A. Welte (2012). A new general biomarker-based incidence estimator. *Epidemiology (Cambridge, Mass.) 23*(5), 721.

Kong, A., J. S. Liu, and W. H. Wong (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association 89*(425), 278–288.

Kürüm, E., R. Li, Y. Wang, and D. S (2014). Nonlinear varying-coefficient models with applications to a photosynthesis study. *Journal of agricultural, biological, and environmental statistics 19*(1), 57–81.

Lanza, S. T., S. A. Vasilenko, X. Liu, R. Li, and M. E. Piper (2013). Advancing the understanding of craving during smoking cessation attempts: A demonstration of the time-varying effect model. *nicotine & tobacco research 16*(Suppl_2), S127–S134.

Li, G., H. Peng, J. Zhang, and L. Zhu (2012). Robust rank correlation based screening. *The Annals of Statistics*, 1846–1877.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

Liu, X., R. Li, S. T. Lanza, S. A. Vasilenko, and M. Piper (2013). Understanding the role of cessation fatigue in the smoking cessation process. *Drug and alcohol dependence 133*(2), 548–555.

Luengo, D. and L. Martino (2012). Efficient random variable generation: ratio of uniforms and polar rejection sampling. *Electronics letters 48*(6), 326–327.

MAP (2001). The Status and Trends of HIV/AIDS/STI Epidemics in Asia and the Pacific. *MAP Provisional Report*.

Marra, G., R. Radice, T. Bärnighausen, S. N. Wood, M. E. McGovern, et al. (2015). A unified modeling approach to estimating hiv prevalence in sub-saharan african countries. Technical report.

Negahban, S., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009). A unified framework for high-dimensional analysis of $\mathfrak{m}$-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pp. 1348–1356.

Qin, Z., K. Scheinberg, and D. Goldfarb (2013). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation 5*(2), 143–169.

Qu, A. and R. Li (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics 62*(2), 379–391.

Raftery, A. E. and L. Bao (2010). Estimating and projecting trends in hiv/aids generalized epidemics using incremental mixture importance sampling. *Biometrics 66*(4), 1162–1173.

Reid, S., R. Tibshirani, and J. Friedman (2013). A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association 82*(398), 543–546.

Schwartländer, B., K. A. Stanecki, T. Brown, P. O. Way, R. Monasch, J. Chin, D. Tarantola, and N. Walker (1999). Country-specific estimates and models of hiv and aids: methods and limitations. *AIDs 13*(17), 2445–2458.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Shiyko, M., P. Naab, S. Shiffman, and R. Li (2013). Modeling complexity of ema data: Time-varying lagged effects of negative affect on smoking urges for subgroups of nicotine addiction. *nicotine & tobacco research 16*(Suppl_2), S144–S150.

Steele, R. J., A. E. Raftery, and M. J. Emond (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (imis). *Journal of Computational and Graphical Statistics 15*(3), 712–734.

Stover, J. (2004). Projecting the demographic consequences of adult hiv prevalence trends: the spectrum projection package. *Sexually transmitted infections 80*(suppl 1), i14–i18.

Stover, J., T. Brown, and M. Marston (2012). Updates to the spectrum/estimation and projection package (epp) model to estimate hiv trends for adults and children. *Sexually transmitted infections 88*(Suppl 2), i11–i16.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Tan, X., M. Shiyko, R. Li, Y. Li, and L. Dierker (2012). Intensive longitudinal data and model with varying effects. *Psychological Methods 17*, 61–77.

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

Trail, J. B., L. M. Collins, D. E. Rivera, R. Li, M. E. Piper, and T. B. Baker (2014). Functional data analysis for dynamical system identification of behavioral processes. *Psychological methods 19*(2), 175.

UNAIDS (1999). *Trends in HIV incidence and prevalence: natural course of the epidemic or results of behavioural change?* Joint United Nations Programme on HIV/AIDS (UNAIDS).

UNAIDS (2000). Guidelines for second generation hiv surveillance. *Geneva: WHO*.

UNAIDS (2014a). Fast-track: ending the aids epidemic by 2030. *Geneva: UNAIDS*.

UNAIDS (2014b). The gap report 2014. *UNAIDS / WHO*.

UNAIDS (2014c). Methodology: Understanding the hiv estimates. *Geneva: UNAIDS*.

UNAIDS/WHO/CDC (2003). Guidelines for conducting hiv sentinel serosurveys among pregnant women and other groups. *Geneva: WHO and UNAIDS*.

Vasilenko, S. A., M. E. Piper, S. T. Lanza, X. Liu, J. Yang, and R. Li (2014). Time-varying processes involved in smoking lapse in a randomized trial of smoking cessation therapies. *nicotine & tobacco research 16*(Suppl_2), S135–S143.

Wahba, G. (1990). Spline models for observing data. *Philadelphia: Society for Industrial and Applied Mathematics*.

Walker, N., K. A. Stanecki, T. Brown, J. Stover, S. Lazzari, J. M. Garcia-Calleja, B. Schwartländer, and P. D. Ghys (2003). Methods and procedures for estimating hiv/aids and its impact: the unaids/who estimates for the end of 2001. *Aids 17*(15), 2215–2225.

Welte, A., T. McWalter, O. Laeyendecker, and T. Hallett (2010). Using tests for recent infection to estimate incidence: problems and prospects for hiv. *Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable disease bulletin 15*(24).

West, M., P. J. Harrison, and H. S. Migon (1985). Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association 80*(389), 73–83.

WHO/UNAIDS (2003). Reconciling antenatal clinic-based surveillance and population-based survey estimates of hiv prevalence in sub-saharan africa. *Geneva: WHO and UNAIDS*.

Wu, C. O. and C.-T. Chiang (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 433–456.

Wu, T. T. and K. Lange (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 224–244.

Yang, H., J. A. Cranford, R. Li, and A. Buu (2015). Two-stage model for time-varying effects of discrete longitudinal covariates with applications in analysis of daily process data. *Statistics in medicine 34*(4), 571–581.

Yi, C., R. Li, P. S. Bakwin, A. Desai, D. M. Ricciuto, S. P. Burns, A. A. Turnipseed, S. C. Wofsy, J. W. Munger, K. Wilson, et al. (2004). A nonparametric method for separating photosynthesis and respiration components in co2 flux measurements. *Geophysical Research Letters 31*(17).

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zaba, B. et al. (2002). Improved methods and assumptions for estimation of the hiv/aids epidemic and its impact: Recommendations of the unaids reference group on estimates, modelling and projections. *AIDS (London, England) 16*(9), W1–14.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, W. and S.-Y. Lee (2000). Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis 74*(1), 116–134.

Zhu, H., L. Kong, R. Li, M. Styner, G. Gerig, W. Lin, and J. H. Gilmore (2011). Fadtts: functional analysis of diffusion tensor tract statistics. *NeuroImage 56*(3), 1412–1425.

Zhu, L.-P., L. Li, R. Li, and L.-X. Zhu (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 106*(496), 1464–1475.

Vita

Jingyi Ye

1. SKILLS

   - Proficient with: regression models analysis, experimental design, analysis of stochastic process, time series analysis, categorical data analysis, multivariate analysis, functional data analysis, data mining.
   - Proficient with software: R, Microsoft Office, Latex, SAS, Matlab.
   - Knowledge in software: SPSS, Python, C++, SQL.
   - Passed Society of Actuaries (SOA) Exams: Probability (2012), Financial Math (2012)

2. EDUCATION

   - Ph.D. in Statistics, Penn State University (2012 - 2017)
   - B.S. in Mathematics, Shandong University (2008-2012)
   - Exchange student in Statistics, University of Science and Technology of China (USTC) (2010-2012)

3. PROJECTS

   - Comparison between horizon model and transition model under non-Markovian framework (internship in Wells Fargo)
   - Estimation of Varying Error Variance for High Dimensional Data
   - Evaluating impact of additional data and sensitivity analysis
   - Consulting projects