DOCTORAL THESIS

Application of partial consistency for the semi-parametric models

Zhao, Jingxin

Date of Award:
2017

HONG KONG BAPTIST UNIVERSITY

Doctor of Philosophy

THESIS ACCEPTANCE

DATE: August 30, 2017

STUDENT'S NAME: ZHAO Jingxin

THESIS TITLE: Application of Partial Consistency for the Semi-parametric Models

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairman:	Dr. Tam Hon Wah Associate Professor, Department of Computer Science, HKBU (Designated by Dean of Faculty of Science)
Internal Members:	Prof. Liao Lizhi Professor, Department of Mathematics, HKBU (Designated by Head of Department of Mathematics)
	Dr. Yang Can Assistant Professor, Department of Mathematics, HKBU
External Members:	Prof. Song Xinyuan Associate Professor Department of Statistics The Chinese University of Hong Kong
	Prof. Zhang Wenyang Chair Professor of Statistics Department of Mathematics The University of York
Proxy:	Dr. Zeng Tieyong Associate Professor, Department of Mathematics, HKBU (as proxy for Prof. Zhang Wenyang)
In-attendance:	Dr. Peng Heng Associate Professor, Department of Mathematics, HKBU

Issued by Graduate School, HKBU

Application of Partial Consistency for the Semi-parametric Models

ZHAO Jingxin

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Principal Supervisor:

Dr. PENG Heng (Hong Kong Baptist University)

August 2017

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis, dissertation submitted to this or other institution for a degree, diploma or other qualification.

I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 

Date: August 2017

ABSTRACT

The semi-parametric model enjoys a relatively flexible structure and keeps some of the simplicity in the statistical analysis. Hence, there are abundance discussions on semi-parametric models in the literature. The concept of partial consistency was firstly brought up in Neyman and Scott(1948). It was said the in cases where infinite parameters are involved, consistent estimators are always attainable for those "structural" parameters. The "structural' parameters are finite and govern infinite samples. Since the nonparametric model can be regarded as a parametric model with infinite parameters, then the semi-parametric model can be easily transformed into a infinite-parametric model with some "structural" parameters. Therefore, based on this idea, we develop several new methods for the estimating and model checking problems in semi-parametric models.

The implementation of applying partial consistency is through the method "local average". We consider the nonparametric part as piecewise constant so that infinite parameters are created. The "structural" parameters shall be the parametric part, the model residual variance and so on. Due to the partial consistency phenomena, classical statistic tools can then be applied to obtain consistent estimators for those "structural" parameters. Further more, we can take advantage of the rest of parameters to estimate the nonparametric part. In this thesis, we take the varying coefficient model as the example. The estimation of the functional coefficient is discussed and relative model checking methods are presented.

The proposed new methods, no matter for the estimation or the test, have remarkably lessened the computation complexity. At the same time, the estimators and the tests get satisfactory asymptotic statistical properties. The simulations we conducted for the new methods also support the asymptotic results, giving a relatively efficient and accurate performance. What's more, the local average method is easy to understand and can be flexibly applied to other type of models. Further developments could be done on this potential method.

In Chapter 2, we introduce a local average method to estimate the functional coefficients in the varying coefficient model. As a typical semi-parametric model, the

varying coefficient model is widely applied in many areas. The varying coefficient model could be seen as a more flexible version of classical linear model, while it explains well when the regression coefficients do not stay constant. In addition, we extend this local average method to the semi-varying coefficient model, which consists of a linear part and a varying coefficient part. The procedures of the estimations are developed, and their statistical properties are investigated. Plenty of simulations and a real data application are conducted to study the performance of the proposed method.

Chapter 3 is about the local average method in variance estimation. Variance estimation is a fundamental problem in statistical modelling and plays an important role in the inferences in model selection and estimation. In this chapter, we have discussed the problem in several nonparametric and semi-parametric models. The proposed method has the advantages of avoiding the estimation of the nonparametric function and reducing the computational cost, and can be easily extended to more complex settings. Asymptotic normality is established for the proposed local average estimators. Numerical simulations and a real data analysis are presented to illustrate the finite sample performance of the proposed method.

Naturally, we move to the model checking problem in Chapter 4, still taking varying coefficient models as an example. One important and frequently asked question is whether an estimated coefficient is significant or really "varying". In the literature, the relative hypothesis tests usually require fitting the whole model, including the nuisance coefficients. Consequently, the estimation procedure could be very compute-intensive and time-consuming. Thus, we bring up several tests which can avoid unnecessary functions estimation. The proposed tests are very easy to implement and their asymptotic distributions under null hypothesis have been deduced. Simulations are also studied to show the properties of the tests.

Keywords: Partial consistency; Local average; Varying coefficient model; Variance estimation; Model checking; Model fitting.

ACKNOWLEDGEMENTS

I am so glad that I can take this chance to express my deep gratitude to my supervisor Dr. PENG Heng. It is he that guides me to the academic world. He taught me right from the first step, how to go through the literature, how to start a research, how to write a paper... Sometimes he is strict like a police officer, and sometimes he is funny like a friend. The experience he shared and the opportunity he supplied help me grow up quickly. This work can no longer exist without his direction and help. I am so thankful for those I have learnt from him, not only the statistical knowledge but also the thoughts about life. At the meantime, I would also like to thank my co-supervisor Dr. Tong Tiejun, for his useful advice and comments.

In addition, I wish to thank Dr. HUANG Tao. He carefully helped me revise most part of Chapter 3. He has given me a lot of invaluable suggestions and improved the quality of this work a lot. I also want to thank Prof. ZHU Lixing and Dr. YANG Can, for their excellent lectures and fruitful discussions.

As well, I would like to thank the lab-mates and my lovely friends. We have studied, discussed and played together in the past four years. It is a very valuable memory in my life. I am grateful to Dr. DAI Wenlin, Dr. DONG Kai, Dr. ZHU Xuehu, Dr. LIU Peng, Dr. GUO Xu, Mr. ZHOU Min, Ms. LUO Dehui, Ms. LI Lingzhu, Ms. MING Jingsi and Mr. TAN Falong, for their selfless help and kindly care.

At last, I give my great thank to my husband, XIE Chuanlong and my parents. They provide me fully support and endless encouragement in the past and in the future.

Table of Contents

Declaration i

Abstract ii

Acknowledgements iv

Table of Contents v

List of Tables viii

List of Figures ix

Chapter 1 Introduction 1

1.1 Semi-parametric model 2

1.2 Statistical inferences for the semi-parametric model 3

1.2.1 Model fitting 3

1.2.2 Variance estimation 4

1.2.3 Model checking 6

1.3 Partial consistency and local average 7

1.4 Outline of the thesis 8

Chapter 2 Local Average Fitting in Varying Coefficient Model 10

2.1 Introduction 10

2.2 Methodology 13

2.2.1 Varying coefficient model 13

2.2.2 Semi-varying coefficient model 15

2.2.3 local average method 17

2.3 Theorem 20
2.4 Application and Simulation 23
2.4.1 A simple application 23
2.4.2 Simulation for varying coefficient model 26
2.4.3 Simulation for semi-varying coefficient model 30
2.5 Concluding remarks 34
2.6 Appendix 35

Chapter 3 Variance Estimation for Semi-parametric Regression Models 40
3.1 Introduction 40
3.2 Variance estimation by local average 43
3.2.1 Review of classical methods 43
3.2.2 Local average method 45
3.2.3 Theoretical properties 46
3.3 Extensions of the estimator 47
3.3.1 Partially linear models 47
3.3.2 Varying coefficient models 48
3.3.3 Refined local average variance estimator 50
3.3.4 More extensions 52
3.4 Applications of local average variance estimation 54
3.4.1 Confidence interval of variance estimation 54
3.4.2 Nonparametric hypothesis testing 55
3.4.3 Variance function estimation 56
3.5 Numerical studies 57
3.5.1 Simulations for variance estimation 57
3.5.2 Applications of variance estimation 67
3.5.3 Real data analysis 70
3.6 Conclusion and discussion 72
3.7 Appendix 73

Chapter 4 Hypothesis Testing in Varying Coefficient Models 79
4.1 Introduction 79

4.2 Methodology 82
4.3 Theorem 89
4.4 Simulation 92
4.5 Concluding remarks 97
4.6 Appendix 99

Bibliography 112

Curriculum Vitae 118

List of Tables

2.1 Typical time (in seconds) used by different estimators, I 30
2.2 Simulation results of the constant coefficients 31
2.3 Simulation results of different estimators 33
2.4 Typical time (in seconds) used by different estimators, II 34

3.1 RMSE for Example 1 58
3.2 MSE for Example 1 with Small Sample $n = 15$ 60
3.3 Simulation Results for Example 2 61
3.4 Simulation Results for Example 3 63
3.5 Simulation Results for Example 4 64
3.6 Simulation Results for Example 5 65
3.7 Simulation results for uniformly distributed covariates 66
3.8 Simulation results for normally distributed covariates 67
3.9 Coverage rate of the confidence interval constructed by (9) and (10) . 68
3.10 Residual Variance Estimates for the Bank Data, I 71
3.11 Residual Variance Estimates for Bank Data, II 72

4.1 Proportion of rejections for null model with T_1 94
4.2 Proportion of rejections for null model with T_2 95
4.3 Proportion of rejections for null model with T_3 96

List of Figures

2.1 Scatter of daily hospital admissions and expected curve when pollutant levels are set at averages. Solid line: full model. Dashed: semi-varying coefficient model. 24

2.2 The estimated coefficient functions with pointwise 95% confidence intervals for the full model. 25

2.3 The estimated coefficient functions for semi-varying coefficient model. 26

2.4 Some typical results. Solid curve: true value; dashed curve: estimated value. 28

2.5 MISE as a function of bandwidth. Solid curve: I=4; dashed curve: I=5; dotted curve: I=10. 29

2.6 Functional coefficient estimation in semi-varying coefficient model. Solid curves:true functions; dashed curve:estimation functions. 31

3.1 Example 8, Left: Power function of the test statistic T in Section 4.2 with $n = 200$. Solid line: $\sigma = 0.25$ with $\varepsilon_i \sim N(0,1)$, dashed line: $\sigma = 0.5$ with $\varepsilon_i \sim N(0,1)$, dot-dash line:$\sigma = 0.25$ with $\varepsilon \sim \sqrt{3/5} \cdot t_5$, two-dash line: $\sigma = 0.5$ with $\varepsilon \sim \sqrt{3/5} \cdot t_5$. Right: Hypothesis function: Solid line: $B = 1$, dashed line: $B = 0.5$, dot-dash line: $B = 0.25$, two-dash line: $B = 0$. 69

3.2 Example 9, Left: Boxplots of the mean absolute deviation curve based on 400 simulations for the proposed variance function estimation with $I = 4, 8$, the refined method, FY1998, as proposed by Fan and Yao (1998), and the ideal estimator, from left to right. Right: The sample residuals, the estimated variance function (solid line) by the refined method, and the true variance function (dashed line). 70

4.1 Null distributions of test statistics T_1, T_2 and T_3. Solid curve: standard normal; dotted curve: n=400; dash-dot curve: n=800; dashed curve: n=1600 96

4.2 Example 1: Left: True function when $a = 0$(solid), $a = 0.2$(dashed), $a = 0.5$(dotted), $a = 0.8$(dash-dotted), $a = 1$(dotted-solid). Right: Power functions for the proposed tests under different alternatives. Solid curve: T_1; dotted curve: T_2; dashed curve: T_3. 97

4.3 Example 2: Left: True function when $a = 0$(solid), $a = 0.2$(dashed), $a = 0.5$(dotted), $a = 0.8$(dash-dotted), $a = 1$(dotted-solid). Right: Power functions for the proposed tests under different alternatives. Solid curve: T_1; dotted curve: T_2; dashed curve: T_3. 98

Chapter 1

Introduction

Investigating the relationship of responding variables and predictors is one of the main purposes in statistical inference. Usually researchers will assume a possible statistic model for the association and parametric models always get the first priority. Under widely discussion for decades, the parametric models have been well recognized for their efficiency and interpretable ability, for example, the classical linear model. It has been a long time since the linear model was first brought up, but its popularity never seems to fade.

Certainly the parametric model has it own weakness. Once the model is established, its structure is fixed. If the specific assumption is incorrect, all the relative statistical inferences will be in vain. The non-parametric model then appears to avoid the plausible wrong results. Without a specific assumption for the model, it gives a more flexible frame of the relation of variables. However, the non-parametric model is usually hard to interpret and it is born with the "curse of dimensionality".

Along with the coming of the age of "big data", we are faced with large volume of various information. The fixed structure of parametric models can no longer meet demand and the non-parametric model can hardly tackle the task with high dimensional data. Therefore, the semi-parametric model seems to be a good choice.

1.1 Semi-parametric model

In order to combine the advantages of parametric models and non-parametric models, the semi-parametric model usually keeps certain structure while contains some non-parametric parts. Thus, the restrictive conditions on the parametric model have been relaxed a little and we still have some structure to make analysis. The semi-parametric model is more flexible than the parametric one and is easier to control than the non-parametric one.

The most institutive example of the non-parametric model is the partially linear model:

$$Y = f(X_1) + \beta^T X_2 + \epsilon,$$

where X_1, X_2 are the covariates, Y is the respond variable, β is the coefficient and ϵ is the error term. Notice that for the independent variable X_1, a non-parametric function is put on. At the meantime, a linear assumption is imposed on variable X_2. This model can well explain the cases that we are sure of the linear relationship between X_2 and Y but don't have enough information of X_1. Hence, a nonspecific function on X_1 makes it a safe and flexible selection.

There are many other powerful approaches in the semi-parametric model, such as low-dimensional interaction models, multiple-index models, additive models, varying coefficient models and so on. The semi-parametric model either put some non-parametric components in the parametric model, or add some structural assumptions to the non-parametric model. Thus, the semi-parametric model shares the advantages of both parametric and non-parametric models.

Among those semi-parametric models, the varying coefficient model has drawn general attention. Hastie and Tibshirani(1993) firstly well defined the varying coefficient model. Based on the linear models, the constant coefficients are replaced by smooth non-parametric functions in the varying coefficient model. Thus the regression coefficients could change over some factors, as it is called — "varying coefficient". Commonly, it is defined as

$$y = \sum_{l=1}^{p} a_l(u) x_l + \epsilon$$

where u is the index variable to monitor the smooth parameters functions $a_l(\cdot)(l =$

$1, ..., p$). Since the coefficients of this model are allowed to vary, the bias can then be significantly reduced. In addition, the "curse of dimensionality" can be avoided as well. What's more, by setting conditions on the covariates, the varying coefficient model can be easily transformed into other semi-parametric models. For example, if we let $x_1 = 1$, $a_l(\cdot)(l = 2, ..., p)$ be constant, then the varying coefficient model turns to be a partially linear model.

The wide applications of the varying coefficient model have been well recognized. The changing coefficient property is quite appealing for the analysis of nonlinear time series data, longitudinal data and survival data. Hence, the varying coefficient model has been successfully applied in economics, finance, epidemiology, medical science and many other areas.

In this thesis, we take the varying coefficient model as a typical example of semi-parametric models. The following content is mostly focused on the varying coefficient model. Nevertheless, as can be seen, the proposed methods are easy to adapt to other type of semi-parametric models.

1.2 Statistical inferences for the semi-parametric model

In the literature, there are plenty of statistical inference methods designed for the semi-parametric model . Here in this section, we will briefly introduce some of the representative ones for model fitting, variance estimation and model checking. More details will be discussed in the relevant chapters.

1.2.1 Model fitting

For estimating the functional coefficients in the varying coefficient model, there are mainly two kinds of methods. One is the basis functions estimation and the other one is based on local polynomial.

In the basis functions estimation, the coefficient functions will be approximated by a basis expansion, then the least squares is employed to get the estimators. This is the method that Hastie and Tibshirani (1993) used when they proposed the varying

coefficient model, where the natural cubic splines are chosen. There are also other smoothing splines estimators, which have been studied by Hoover et al.(1998), Wu and Chiang(2000) and Chiang et al.(2001). Huang, et al.(2002) studied the statistical properties of the general basis functions estimators. The basis functions estimators usually have a specific expression of the estimated function, so that they are intuitional. However, they need to deal with the problems of how to choose the basis functions and related parameters from numerous options.

Since the varying coefficient model can be considered as a local linear model , the local polynomial estimator seems more appropriate. The local polynomial estimators have been popularly discussed through years. In the local polynomial estimation method, one applies the Taylor expansions to the functional coefficients and calculates the estimators by least squares with kernel. A weighted local polynomial estimator was used in Hoover et al.(1998) to estimate $a_l(\cdot)$. Conventionally, we call it one step local polynomial estimator. However, the one step local polynomial estimator is proposed on the assumption that all the functional coefficients share the same degree of smoothness. Fan and Zhang(1999) have proved that the one step local polynomial estimator is not efficient if the coefficient functions have different degrees of smoothness. Therefore they presented a two step estimator. In the first step, the initial estimators are obtained by local linear estimation. Then in the second step, after substituting the initial estimators into the model, a higher order local polynomial with suitable bandwidth is used for the objective coefficient function. Thus the needs for different degrees of smoothness can be satisfied.

1.2.2 Variance estimation

The variance estimation of the error term is always playing a vital part in model inferences. There are mainly two kinds of estimators for the residual variance: residual sum of squares estimators and difference-based estimators.

Firstly we consider a simple non-parametric model,

$$y = f(x) + \epsilon$$

where y is the response variable, f is some unknown mean function and the error term ϵ is distributed with zero mean and constant variance σ^2.

To estimate the residual variance σ^2, a preliminary thought is to find the mean function f first, then calculate σ^2 from the residual sum of squares. That is also the basic idea of the first kind of method. Usually, a parametric model is assumed to estimate the unknown mean function, then use the fitted value to calculate the estimated residuals. However, since the information of the mean function is usually incomplete or even missing, one can hardly construct an accurate mean function. What's worse, a misspecified form for the function may lead to large bias.

To avoid a particular assumption for the mean function, non-parametric methods are brought out, such as Nadaraya-Watson kernel regression(1964). This method still aims to find the fitted mean function values then get the residuals. Instead of making an assumption for the mean function, the kernel regression takes advantage of the data themselves. In this way, the forms of the mean function are based on the observed data and become more flexible. The kernel regression makes the mean function estimations more accurate so that the estimation of the residual variance becomes more reliable too. However, this kind of variance estimators faces other problems. The bandwidth choice of the kernel regression is always an issue. Besides that, the estimators depend critically on the smoother matrix and are messy to analyze.

In 1984, Rice proposed the first-order difference-based estimator when discussing the bandwidth choice for the nonparametric regression. The difference-based estimators, i.e. the second kind, use differences to get rid of the mean function trend. By taking differences, the mean function value $f(x_i)$ of the adjacent x_i could be removed if the mean function is smooth enough. Thus we can skip the step of estimating the mean function f and estimate the residual variance directly.

Later on, Gasser et al. (1986) proposed a second-order difference-based estimator. More difference-based estimator with different orders were introduced afterwards, for example, Hall et al.(1990). However, none of the above estimators achieves the asymptotic optimal rate $o_p(n^{-1})$ for the mean squared error (Dette et al., 1998), until M*ü*ller et al.(2003) studied the class of difference-based estimators. Under certain assumptions for the difference weights ,the asymptotic optimal rate could be achieved. Tong and Wang(2005) suggested a new estimator, taking regression on the lag-k Rice's

estimator. In equally spaced design, this estimator could reduce the asymptotic rate to $O_p(n^{-3/2})$.

Compared to the residual sum of squares estimators, the difference-based estimators are much more direct, saving a lot of procedure on mean function construction. Thus it makes the residual variance estimation become independent of the mean function construction. However, most difference-based estimators require the design points to be univariate and ordered, which makes it complicated to extend to high-dimensional domains. What's more, the over difference may lead to large estimation variance.

The variance estimation of semi-parametric model is in the same situation. One has either to fit the whole model, both the parametric and the non-parametric components, or to come up with a strategy to eliminate the non-parametric part.

1.2.3 Model checking

As we have mentioned before, parametric models are always welcome if the assumptions hold. Then it is natural to ask the question whether possible structure is feasible in the model, or whether a non-parametric component is necessary. Therefore, the model checking procedure is of great importance in the statistical inference.

Take the varying coefficient model again as the example. Researchers have investigated many kinds of difference between the null and the alternative hypothesis to get the test statistics and the corresponding critical values.

Usually, the log-likelihood or the residual squares under the null and the alternative hypothesis will be compared to see the discrepancy. The test statistics is set to be the difference or the ratio. Due to the complexity of varying coefficient, the reject rules are usually obtained by bootstrap. Cai, et al. (2000a) , Cai, et al(2000b) and Huang, et al.(2002) have studied a lot on this kind of method for different estimators and data types.

Differently, with the local polynomial estimator, Fan and Zhang(2000) investigated the deviations of the estimated coefficient function and the true coefficient function. The asymptotic distribution for the maximum of the normalized deviations has been deduced. Hence, hypothesis tests can be conducted. However, this

test statistic involves many unknown quantities. The estimation of the unknown quantities needs high order polynomials so that it is complicated to compute.

In Fan, et al.(2001), the generalized likelihood ratio (GLR) tests has been brought up and they illustrated this test with varying coefficient model in detail. Using the difference of the log-likelihood under alternative and null hypothesis as the test statistics, the GLR test is proved to be optimal and to follow the Wilk's phenomena. Instead of the maximum likelihood estimator in classical maximum likelihood ratio test, any reasonable nonparametric regression estimators can be used in the alternative log-likelihood in GLR test. Thus the GLR tests is widely applicable to many model, providing many possibilities for model checking problems. Also, one can notice that the tests in Cai, et al. (2000a) , Cai, et al(2000b) and Huang, et al.(2002) are actually some specific scenarios of GLR tests.

1.3 Partial consistency and local average

Neyman and Scott(1948) firstly proposed the concept of partial consistency. Consider a sequence of independent variable $\{x_i\}$ and this sequence of variable follows a distribution depending on two set of parameters, Θ_1 and Θ_2. Θ_1 contains only finite number of parameters while Θ_2 has infinite ones. What's more, the parameters in Θ_1 governs infinite number of variables x_i while the ones in Θ_2 are only related to finite number of variables. Then the sequence $\{x_i\}$ is consistent in respect to the parameters in Θ_1. In Neyman and Scott(1948), they called the parameters in Θ_1 "structural" and the parameters in Θ_2 "incidental".

For the semi-parametric model, we can regard the parameters in the parametric part as the "structural" parameters. As for the non-parametric part, we can transform it into a parameter one with infinite "incidental" parameters. Thus, based on the partial consistency phenomena, we come up with several new ideas for estimating parameters and conducting tests in semi-parametric models.

To create infinite "incidental" parameters, we implement the "local average" method. The non-parametric part could be considered as piecewise constant, with the "piece" being very small. Then in each "piece", we take average to get the true

function value. This true function value is only valid in its own piece and responding to finite variables. Thus, for the non-parametric part we will get infinite constant, which are also the "incidental" parameters. Since the parametric part is shared by all observations, the relevant parameters then can be treated as "structural". Similarly, residual variance is a structural parameter as it influences all the samples. Next, we can apply classical statistic tools to obtain consistent estimators for those "structural" parameters. Moreover, the "incidental" parameters are still of great use. In this thesis, we have proved that those "incidental" parameters are unbiased estimators for the true function values. Then it becomes a simple non-parametric model. Now classical non-parametric smoothing and model checking methods can be used to investigate the non-parametric part.

1.4 Outline of the thesis

The rest of this thesis is arranged as follow.

In Chapter 2, we discussed the estimation problems in varying coefficient models and in partial linear varying coefficient models. The local average method is developed for estimating the functional coefficients and the parameters in linear part. Asymptotic properties are deduced and the simulations are consistent with the theoretical results. We also apply the method on a real data set about the air pollutants in Kong Hong.

Chapter 3 is focus on the estimation of residual variance. Starting with a simple non-parametric model, the local average estimators is brought up. Then, we extend the model to the partially linear model and the varying coefficient model. The local average estimators adapt well to these semi-parametric models. Further more, we propose a refined local average estimator to improve the performance. We also point out several applications of the local average estimator. Then the numerical simulations are conducted and a real data set about the bank employees is used to illustrate the method.

At last, the model checking problem is discussed in Chapter 4. We proposed three test statistics for the varying coefficient model based on the local average method.

Actually the first two test statistics are an extended application of Chapter 2 and the third test is founded on the results of Chapter 3. We have obtained the asymptotic distributions of the three test statistics under null hypothesis and have done a lot of simulations. Both the size study and the power study give satisfactory results.

Chapter 2

Local Average Fitting in Varying Coefficient Model

2.1 Introduction

As we all know, a parametric model is very convenient to establish and its statistical properties have been well studied, especially for the classical linear model. So they are widely used in industry, for example, the linear regression appears everywhere. However, with the volume and the varieties of data increasing rapidly, the classical parametric models can no longer fit the complicated data relationship. The fixed data structures of parametric models are not very realistic in applications, and a mis-specification could lead to a large bias.

To add more flexibility to the model and to keep some useful properties of the parametric models, many semi-parametric models are proposed. Examples include but not limit to the partially linear model(Engle, Granger, Rice and Weiss 1986; Robinson 1988), the additive model(Friedman and Stuetzle 1981; Buja, Hastie and Tibshirani 1989),the multiple-index model(Härdle and Stoker 1989) and the varying coefficient model(Hastie and Tibshirani 1993). All these semi-parametic models are intended to relax some parts of the conditions imposed on the parametric models.

Among the above semi-parametric models, the varying coefficient model arouses interest of many researchers. It could be regarded as an extension of the classical linear model. With dynamic coefficients, the varying coefficient model can express

more accurately about the relationship of the response and the covariates. For instance, in this age of big data, the E-business would collect many information from the consumers and make use of these information to do target promotion. It will become more convincing if the association is allowed to change over time (or age). Similarly, the varying coefficient model is successfully applied in economics, finance, epidemiology, medical science and many other areas. The property of changing coefficient is quite appealing for analysis of nonlinear time series data, longitudinal data and survival data.

The varying coefficient model is firstly well described in Hastie and Tibshirani (1993). Usually, it will take a form as

$$y = \sum_{l=1}^{p} a_l(u)x_l + \epsilon \tag{2.1}$$

where u is the index variable to monitor the smooth parameters functions $a_l(\cdot)(l = 1, ..., p)$. For given covariates u and $(x_1, x_2, ..., x_p)$

$$\mathrm{E}[\epsilon|u, x_1, ..., x_p] = 0, \qquad \mathrm{Var}[\epsilon|u, x_1, ..., x_p] = \sigma^2.$$

With a same degree smoothness assumption on the coefficient functions $a_l(\cdot)$, Hastie and Tibshirani (1993) proposed an estimation method with smoothing spline. In their estimation, the functional coefficients are expressed by the natural cubic spline basis functions. Another global smoothing method is based on polynomial splines, brought out in Huang et al.(2002,2004). Nevertheless, Huang et al's polynomial spline method can choose multiple smoothing parameters so that it still works well when the coefficient functions have different smoothness.

On the other hand, since the natural character of the varying coefficient model is locally linear, the kernel-local polynomial smoothing is also popular in the literature. Hoover et al.(1998) used a weighted local polynomial estimator to estimate $a_l(\cdot)$ and the asymptotic distributions of this estimator are derived by Wu et al.(1998). However, when different degrees of smoothness exist in the coefficients, this one-step local polynomial method may not be adequate. To adjust to different smoothness, Fan and Zhang(1999) raised a two-step estimator. In the first step, an initial estimators for the lower smoothness coefficients are obtained by kernel-local linear method. Then in the second step, the initial estimators are substituted back to extract the higher

smoothness coefficient. Thus, with a tunable bandwidth in the second step, the two-step estimator performs better than the one-step method when estimating the smoother coefficient. Even if there is no higher smoothness in coefficients, the two-step estimator can work equally well as the one-step one.

Sometimes, in practice, some coefficients are known to be constant. To fully utilize this information, we can write a common model as

$$y = \sum_{l=1}^{p} a_l(u)x_l + \mathbf{Z}^T\boldsymbol{\beta} + \epsilon \tag{2.2}$$

Thus we get a semivarying coefficient model, where $\boldsymbol{\beta}$ is a q-dimensional unknown vector and $a_l(\cdot)(l = 1, ..., p)$, as in model (2.1), are unknown functions to be estimated. In most of the existing estimating methods, they would treat the estimation of the constant parameter $\boldsymbol{\beta}$ as primary, since a good estimator of $\boldsymbol{\beta}$ will transform the model to a standard varying coefficient one.

Zhang et al.(2002) studied this semivarying coefficient model. They suggested to first consider the constant coefficients as functional too and then take average to get the final estimate of $\boldsymbol{\beta}$. The resulting estimator has bias of order $O_p(h^2)$ but does not reach the lowest asymptotic variance for the semiparametric model. Then in Fan and Huang(2005), a profile least-square estimator is put forward. The technique is as follow. First, get the estimated functional coefficients $\widehat{a_l}$ expressed in a function of $\boldsymbol{\beta}$, using the one-step kernel-local polynomial method for the model (2.1). Then substitute $\widehat{a_l}$ back to model (2.2) to solve $\boldsymbol{\beta}$ via ordinary least squares. It is shown that the covariance matrix of profile least-square estimator reaches the lowest asymptotic variance for the semiparametric model. To further reduce the estimation bias of $\boldsymbol{\beta}$, Xia et al.(2004) presented a semi-local least squares estimator. Based on the local kernel estimation, a summation of the residual squares is made to acquire a global estimator of $\boldsymbol{\beta}$. In this way, the bias of the constant coefficient $\boldsymbol{\beta}$ becomes $O(h^3)$. Alternatively, general series method can also be applied to semivarying coefficient model, see Ahmad et al.(2005).

In this chapter, we come up with another idea for the varying coefficient model. The new estimator keeps the local linearity of the model and simplifies the estimation procedure. The main structure of this estimating procedure is as follows: Regard the varying coefficient model as locally piecewise linear so that we can get a series of

points estimatior for the functional coefficient $a_l(\cdot)$. The next step is to use some classical smoothing method to get the whole curve estimator from the estimated points. This local average estimator has three advantages. First, the computation burden is sharply lightened meanwhile the accuracy is ensured. Second, the second step provides necessary opportunity to control the bandwidth for different smoothness. Third, this estimator can easily adapt to the semivarying coefficient model, resulting a global estimator of constant coefficient $\boldsymbol{\beta}$. Moreover, we do not need to substitute back the estimated $\hat{\boldsymbol{\beta}}$ to get the varying coefficients.

The remainder of this chapter contains a more detailed discussion of how local average method being executed and a theoretical explanation about why it is feasible. In Section 2, we will illustrate the implementation of our new estimator concretely, in both the varying coefficient model and the semivarying coefficient model. Section 3 studies the asymptotic properties of the local average estimator. Section 4 gives a simple application and a lot of simulations and comparisons of the local average estimator and other estimators. Section 5 is the summary, covering the conclusion and discussion.

2.2 Methodology

Since the local average estimator also belongs to the local smoothing methods, we will focus our discussions on the local kernel estimations. In this section, detailed descriptions, especially the formulas, of the existing local kernel methods are provided. Thus we can have a clear comparison among the local smoothing methods for the varying-coefficient model and the semi-varying coefficient model.

2.2.1 Varying coefficient model

Rewrite the simple varying coefficient model (2.1) in matrix form

$$y = X^T \mathbf{a}(U) + \epsilon$$

where X is a p-dimension vector $(x_1, ..., x_p)^T$ and the unknown functional coefficient $\mathbf{a}(U) = (a_1(U), ..., a_p(U))^T$. Without losing generality, we suppose we are

aiming to estimate the last functional coefficient, i.e., $a_p(\cdot)$, with a random sample $(U_i, X_i, y_i), i = 1, ..., n$.

The first attempt is the one-step estimation, which uses a kernel local linear smoothing. For each given u_0, the coefficient function $a_l(u)$ is approximated locally as

$$a_l(u) \approx a_l + b_l(u - u_0), l = 1, ..., p$$

in a small neighbourhood of u_0. Then this leads to minimizing

$$\sum_{i=1}^{n} \{y_i - X_i^T \mathbf{a} - X_i^T \mathbf{b}(U_i - u)\}^2 K_h(U_i - u)$$

with respect to $(\mathbf{a}, \mathbf{b})$, for a given kernel function K and a bandwidth h. Let

$$\begin{aligned}
&\mathbf{X} = (X_1, ..., X_n)^T, \qquad \mathbf{U}_u = \text{diag}(U_1 - u, ..., U_n - u), \\
&\Lambda_u = (\mathbf{X}, \mathbf{U}_u \mathbf{X}), \qquad Y = (y_1, ..., y_n)^T, \\
&W_{h,u} = \text{diag}(K_h(U_1 - u), ..., K_h(U_n - u)).
\end{aligned}$$

Then we can have the one-step estimator

$$\hat{\mathbf{a}}_p(u) = e_{p,2p}^T (\Lambda_u^T W_{h,u} \Lambda_u)^{-1} \Lambda_u^T W_{h,u} Y$$

Here and hereafter, we will always use notation $e_{k,m}$ to denote the unit vector of length m with 1 at the kth position.

This one-step estimator intuitively demonstrates the structure of the varying coefficient model. It achieves a bias of $O(h^2)$ and a variance of $O((nh)^{-1})$ under the assumption that all the coefficient functions possess the same smoothness. However, when the same smoothness assumption could not hold, Fan and Zhang(1999) pointed out that one single bandwidth is not enough to reach the optimal rate. To realize different bandwidth for smoother coefficient, two step estimator is proposed.

The first step is the same as the one-step estimator, then we can get the initial estimators for all the coefficients with an initial bandwidth h_0

$$\hat{\mathbf{a}}_l(u) = e_{l,2p}^T (\Lambda_u^T W_{h_0,u} \Lambda_u)^{-1} \Lambda_u^T W_{h_0,u} Y, \qquad l = 1, ..., p.$$

For $l = 1, ..., p-1$, substitute $\hat{\mathbf{a}}_l(u)$ back to the model 2.1, then we get a new responsor

$$\tilde{y}_i = y_i - \sum_{i=1}^{p-1} \hat{\mathbf{a}}_l(U_i) x_{il}.$$

Assume the smoother coefficient $\mathbf{a}_p(u)$ has a fourth derivative. With a different bandwidth h_2, let

$$X_p = (x_{1p}, x_{2p}, ..., x_{np})^T, \qquad \tilde{Y} = (\tilde{y}_1, ..., \tilde{y}_n)^T,$$
$$\Gamma_u = (X_p, \mathbf{U}_u X_p, \mathbf{U}_u^2 X_p, \mathbf{U}_u^3 X_p), \qquad .$$

Then for any given u, the two step estimator is

$$\hat{\mathbf{a}}_p(u) = e_{1,4}^T (\Gamma_u^T W_{h_2,u} \Gamma_u)^{-1} \Gamma_u^T W_{h_2,u} \tilde{Y}$$

With a tunable bandwidth h_2, the bias of the two step estimator is of $O(h_2^4)$ and the variance is of $O((nh)^{-1})$. So the two step estimator can achieve the optimal rate of convergence $O_p(n^{-8/9})$. Nevertheless, the optimal rate is at the cost of the computation burden. Since the substitution process needs all the estimators of the nuisance coefficients at every U_i, the two step estimation is very time-consuming.

2.2.2 Semi-varying coefficient model

The most important part in semi-varying coefficient model estimation is the constant coefficient estimation. A good estimator of the constant coefficient will turn the problem into a simple varying coefficient one. Then the remains can be solved by the method we mentioned in the above subsection. In the existing local kernel estimation, the main idea is similar, that is, to use a local polynomial to replace the function coefficients.

Based on the simple varying coefficient model, we add a linear part $Z^T\mathbf{b}$ Thus we can consider this semi-varying coefficient model

$$y = X^T \mathbf{a}(U) + Z^T \mathbf{b} + \epsilon$$

where Z is a q-dimension vector $(z_1, ..., z_q)^T$ and the unknown constant coefficient $\mathbf{b} = (b_1, ..., b_q)^T$.

Zhang et al.(2002) proposed a clever idea. First treat $\mathbf{b}$ as functional, then take average over samples to get the final estimator. By the one step estimator we discussed in last section, for each b_j and U_i we have

$$\hat{b}_j(U_i) = e_{p+j,2(p+q)}^T (\Lambda_{U_i}^T W_{h,u_i} \Lambda_{U_i})^{-1} \Lambda_{U_i}^T W_{h,U_i} Y$$

Then we take average of $\hat{b}_j(U_i)$ over $i = 1, ..., n$ to get the final estimator

$$\hat{b}_j = \frac{1}{n}\sum_{i=1}^{n} e_{p+j,2(p+q)}^T(\Lambda_{U_i}^T W_{h,U_i}\Lambda_{U_i})^{-1}\Lambda_{u_i}^T W_{h,U_i}Y$$

The bias of Zhang et al.'s estimator is of order $O_p(h^2)$ and the covariance matrix is of order $O_p(n^{-1})$. We can notice that this estimator is developed from a local estimator. The global property of the constant coefficient $\mathbf{b}$ is not fully utilized.

Fan and Huang(2005) proposed another estimator. Based on the sample, rewrite the model as

$$y_i - Z_i^T\mathbf{b} = X_i^T\mathbf{a}(U_i) + \epsilon_i, \quad i = 1, ..., n,$$

Then still by the one step estimator in last section, the functional coefficients could be expressed as

$$\hat{\mathbf{a}}(U_i) = (\mathbf{I}_p, \mathbf{0}_p)(\Lambda_{U_i}^T W_{h,U_i}\Lambda_{U_i})^{-1}\Lambda_{U_i}^T W_{h,U_i}(Y - \mathbf{Z}\mathbf{b}),$$

where $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)^T$, $\mathbf{I}_p$ is a size p identity matrix and $\mathbf{0}_p$ is a size $p \times p$ zero matrix. Substitute $\hat{\mathbf{a}}(U_i)$ back to the model, we obtain

$$y_i - Z_i^T\mathbf{b} = X_i^T(\mathbf{I}_p, \mathbf{0}_p)(\Lambda_{U_i}^T W_{h,U_i}\Lambda_{U_i})^{-1}\Lambda_{U_i}^T W_{h,U_i}(Y - \mathbf{Z}\mathbf{b}) + \epsilon_i, \quad i = 1, ..., n.$$

Denote

$$\mathbf{H} = \begin{pmatrix} X_1^T(\mathbf{I}_p, \mathbf{0}_p)(\Lambda_{U_1}^T W_{h,U_1}\Lambda_{U_1})^{-1}\Lambda_{U_1}^T W_{h,U_1} \\ \vdots \\ X_n^T(\mathbf{I}_p, \mathbf{0}_p)(\Lambda_{U_n}^T W_{h,U_n}\Lambda_{U_n})^{-1}\Lambda_{U_n}^T W_{h,U_n} \end{pmatrix},$$

then we can have

$$(\mathbf{I}_n - \mathbf{H})Y = (\mathbf{I}_n - \mathbf{H})\mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}.$$

Finally the Fan and Huang's estimator is gained by least squares:

$$\widehat{\mathbf{b}} = \{\mathbf{Z}^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})Y.$$

Same as Zhang et al.'s estimator, this one also has a bias of $O_p(h^2)$ and a variance of order $O_p(n^{-1})$. Besides, Fan and Huang(2005) have showed that unlike Zhang et al.'s estimator, theirs is semiparametrically efficient. But the cumbersome process of computing the nuisance coefficients is obvious a shortcoming.

In 2004, Xia et al. presented a semi-local least squares estimator. The constant coefficient $\mathbf{b}$ is estimated globally while the functional ones are estimated locally. Minimise the function

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - X_i^T\{\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j(U_i - U_j)\} - Z_i^T\mathbf{b}]^2 K_h(U_i - U_j)$$

respect to $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_j (j = 1, ..., n)$ and $\mathbf{b}$. Then the estimator of $\mathbf{b}$ is

$$\widehat{\mathbf{b}} = (\mathbf{0}_{q\times 2pn}, \mathbf{I}_q)(\Omega^T W_h \Omega)^{-1}\Omega^T W_h \breve{Y}$$

where

$$\Omega = (A, \mathbf{Z}^*), \quad A = \mathbf{I}_n \otimes (\mathbf{X}, h^{-1}\mathbf{U}\mathbf{X}) - h^{-1}\mathbf{U} \otimes (\mathbf{0}_{n\times p}, \mathbf{X}), \quad \mathbf{Z}^* = \mathbf{1} \otimes \mathbf{Z},$$
$$\mathbf{U} = diag\{U_1, U_2, ..., U_n\}, \quad W_h = diag\{W_{h,U_1}, ..., W_{h,U_n}\}, \quad \breve{Y} = \mathbf{1} \otimes Y,$$

$\mathbf{0}_{k\times l}$ is a zero matrix with size $k \times l$ and $\mathbf{1}$ is an n-dimensional vector with all entries being 1.

Xia et al.(2004) have showed that this semi-local least squares estimator has bias of $O(h^3)$ and the variance is $O(n^{-1})$. Since the bias has been reduced, the undersmoothing is avoid. At the meantime we notice that the computation burden is still heavy, since the size of the design matrix Ω is increasing with n^2.

2.2.3 local average method

To simplify the estimation process and facilitate different bandwidths, we propose a local average estimator. The primary concept is that we treat the functional coefficient as local constant within a small neighbourhood. Thus we can use the average of the function values in this small neighbourhood as the estimator of the function value of the midpoint in this interval. After collecting all the midpoint estimators, we can use some classic smoothing method to rebuild the function. For the first step, a sufficiently small interval is used to ensure the unbiasedness. Then in the second step, different bandwidths could be selected for different smoothness acquirements.

The details of the implement is as following. Let's consider the simple varying-coefficient model

$$y = X^T\mathbf{a}(U) + \epsilon.$$

In the beginning, we sort the samples $(U_i, X_i, y_i), i = 1, ..., n$ according to U_i in an ascending order. $U_{(1)} \leq U_{(2)} \leq ... \leq U_{(n)}$. Then divide them into k groups with I samples in each group. Since we should make sure the interval is sufficiently small, we usually take small fixed I and k may increase with the sample size n. Thus, $n = Ik$. (In practise, the possible remainders are removed out. As I is small enough, the number of the removed samples is negligible.)

Denote the jth observation in ith group as $(U_{ij}, X_{ij}, y_{ij}), i = 1, ..., k, j = 1, ..., I$. Hence, we have

$$y_{ij} = X_{ij}^T \mathbf{a}(U_{ij}) + \epsilon_{ij}, \quad i = 1, ..., k, \quad j = 1, ..., I.$$

Since we treat the functional coefficient $\mathbf{a}(\cdot)$ as constant in a small neighbourhood, we assume in every group $\mathbf{a}(U_{i1}) = \mathbf{a}(U_{i2}) = \cdots = \mathbf{a}(U_{iI}) = \mathbf{a}_i = \mathbf{a}(\bar{U}_{i\cdot})$. Thus, for the ith group, we have

$$\mathbf{Y}_i = \mathbf{X}_i \mathbf{a}_i + \boldsymbol{\epsilon}_i^*$$

where

$$\mathbf{Y}_i = (y_{i1}, y_{i2}, ..., y_{iI})^T \in \mathbb{R}^{I\times 1}, \quad \mathbf{a}_i = (a_1(\bar{U}_{i\cdot}), a_2(\bar{U}_{i\cdot}), ..., a_p(\bar{U}_{i\cdot}))^T \in \mathbb{R}^{p\times 1},$$
$$\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{iI})^T \in \mathbb{R}^{I\times p} \quad \text{and} \quad \boldsymbol{\epsilon}_i = (\epsilon_{i1}^*, \epsilon_{i2}^*, ..., \epsilon_{iI}^*)^T \in \mathbb{R}^{I\times 1}.$$

Combine all the k groups, we get

$$\mathbf{Y} = \mathbb{X}\mathbf{a} + \boldsymbol{\epsilon}^*$$

where

$$\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, ..., \mathbf{Y}_k^T)^T \in \mathbb{R}^{n\times 1}, \quad \mathbf{a} = (\mathbf{a}_1^T, \mathbf{a}_2^T, ..., \mathbf{a}_k^T)^T \in \mathbb{R}^{kp\times 1},$$
$$\mathbb{X} = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \cdots \oplus \mathbf{X}_k \in \mathbb{R}^{n\times kp} \quad \text{and} \quad \boldsymbol{\epsilon}^* = (\boldsymbol{\epsilon}_1^{*T}, \boldsymbol{\epsilon}_2^{*T}, ..., \boldsymbol{\epsilon}_k^{*T})^T \in \mathbb{R}^{n\times 1}.$$

Now we can get our primary estimator

$$\begin{aligned}\hat{\mathbf{a}} &= (\hat{a}_1(\bar{U}_{1\cdot}), ..., \hat{a}_p(\bar{U}_{1\cdot}), \hat{a}_1(\bar{U}_{2\cdot}), ..., \hat{a}_p(\bar{U}_{2\cdot}), ..., \hat{a}_1(\bar{U}_{k\cdot}), ..., \hat{a}_p(\bar{U}_{k\cdot}))^T \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}\end{aligned}$$

Still take the last functional coefficient $a_p(\cdot)$ as example, from the primary estimator $\hat{\mathbf{a}}$, k relevant estimators $(\hat{a}_p(\bar{U}_{1\cdot}), \hat{a}_p(\bar{U}_{2\cdot}), ..., \hat{a}_p(\bar{U}_{k\cdot}))$ are acquired. We can prove that for any $i = 1, ..., k$,

$$\mathrm{E}(\hat{a}_p(\bar{U}_{i\cdot})) = a_p(\bar{U}_{i\cdot}) + O_p(\frac{\log n}{n}), \ \ \mathrm{Var}(\hat{a}_p(\bar{U}_{i\cdot})) = e_{p,p}^T \mathrm{E}[(\mathbf{X}_i^T\mathbf{X}_i)^{-1}|\bar{U}_{i\cdot}]e_{p,p}\sigma^2$$

The details of the proofs are in the Appendix.

Then the problem becomes a nonparametric model

$$\hat{a}_p(\bar{U}_{i\cdot}) = a_p(\bar{U}_{i\cdot}) + v_p(\bar{U}_{i\cdot})\epsilon, \quad i = 1, ..., k$$

where $v_p^2(\bar{U}_{i\cdot}) = e_{p,p}^T \mathrm{E}[(\mathbf{X}_i^T\mathbf{X}_i)^{-1}|\bar{U}_{i\cdot}]e_{p,p}$.

Now we can use some classical smoothing method to get the final estimator. In this step, we can choose bandwidth h and other estimating parameters freely for different smoothness. In this chapter, we adapt the local polynomial smoothing. The theoretical study in next section is also based on this adaption. Assume the last functional coefficient $a_p(\cdot)$ has bounded forth derivative. For given u, denote

$$\bar{\mathbf{U}} = \begin{pmatrix} 1 & (\bar{U}_{1\cdot} - u) & \cdots & (\bar{U}_{1\cdot} - u)^3 \\ \vdots & \vdots & & \vdots \\ 1 & (\bar{U}_{k\cdot} - u) & \cdots & (\bar{U}_{k\cdot} - u)^3 \end{pmatrix},$$

and put

$$\hat{\mathbf{a}}_p = (\hat{a}_p(\bar{U}_{1\cdot}), \hat{a}_p(\bar{U}_{2\cdot}), ..., \hat{a}_p(\bar{U}_{k\cdot}))^T, \quad \bar{\mathbf{W}} = diag(K_h(\bar{U}_{1\cdot} - u), ..., K_h(\bar{U}_{k\cdot} - u))$$

Here we can choose bandwidth h freely for different smoothness. Then the final local-average estimator can be obtained by weighted least squares

$$\hat{a}_p(u) = e_{1,4}^T(\bar{\mathbf{U}}^T\bar{\mathbf{W}}\bar{\mathbf{U}})^{-1}\bar{\mathbf{U}}^T\bar{\mathbf{W}}\hat{\mathbf{a}}_p$$

An extension to the semi-varying coefficient model is natural. Opposite to Zhang et al.(2002)'s idea, we keep the constant property of the linear part but treat the functional coefficients as piecewise constant. In this way, a global estimator of the constant coefficient is feasible. For the varying coefficient part, either a back substitution or continuation with classical smoothing rebuild is available.

Consider the semi-varying coefficient model

$$y = X^T\mathbf{a}(U) + Z^T\mathbf{b} + \epsilon.$$

After reordering and grouping the samples (U_i, X_i, Z_i, y_i) according to U_i, we reindex the jth observation in ith group as $(U_{ij}, X_{ij}, Z_{ij}, y_{ij})$. Let $\mathbf{\Phi} = (\mathbb{X}, \mathbb{Z})$ and $\boldsymbol{\theta} = (\mathbf{a}^T, \mathbf{b}^T)^T$, where $\mathbb{Z} = (\mathbf{Z}_1^T, \mathbf{Z}_2^T, ..., \mathbf{Z}_k^T)^T \in \mathbb{R}^{n\times q}$, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{iI})^T \in \mathbb{R}^{I\times q}$,

$\mathbb{X} = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \cdots \oplus \mathbf{X}_k \in \mathbb{R}^{n \times kp}$ and $\mathbf{a} = (\mathbf{a}_1^T, \mathbf{a}_2^T, ..., \mathbf{a}_k^T)^T \in \mathbb{R}^{kp \times 1}$ as we just denoted above,. Then we can write the model as

$$\mathbf{Y} = \mathbf{\Phi}\boldsymbol{\theta} + \boldsymbol{\epsilon}^*.$$

Therefore the local average estimator of the constant coefficient $\mathbf{b}$ can be easily calculated by ordinary least squares

$$\widehat{\mathbf{b}} = (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{Y}.$$

2.3 Theorem

The following technical conditions are needed for investigating the theoretical properties of the proposed estimators:

1. $a_l'(\cdot)$ and $a_l''(\cdot)$ are continuous and bounded for $l = 1, ..., p$.
2. The function a_p has a continuous fourth derivative in a neighborhood of u_0.
3. All predictors are bounded, i.e., $\|\mathbf{X}\|^2 < \infty$, $\|\mathbf{Z}\|^2 < \infty$.
4. The residuals ϵ_i are standard normally distributed.
5. The density function f of U has a continuous second derivative and there exists $\gamma > 0$ such that $f(u) > \delta$ for any u.
6. The function $K(t)$ is a symmetric density function with a compact support.
7. The group size I is a fixed small integer such that $I/n \to 0$.

We employed the following notations throughout this chapter:

$$\xi_i = \int t^i K(t)dt, \quad \nu_i = \int t^i K^2(t)dt$$
$$\Psi(I) = \mathrm{E}[(\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T)^{-1}|U]$$

Then we can have the following asymptotic properties of the local average estimator with the adaption of local cubic polynomial smoothing.

Theorem 2.1. *Under conditions 1-7, if $h \to 0$ and $nh \to \infty$, then the asymptotic conditional bias of $\hat{a}_p(u_0)$ in simple varying-coefficient model is given by*

$$bias(\hat{a}_p(u_0)|U,X) = \frac{1}{4!}\frac{\xi_4^2 - \xi_2\xi_6}{\xi_4 - \xi_2^2}a_p^{(4)}(u_0)h^4 + o_p(h^4)$$

and the asymptotic conditional variance of $\hat{a}_p(u_0)$ is given by

$$var(\hat{a}_p(u_0)|U,X) = \frac{(\xi_4^2\nu_0 - 2\xi_4\xi_2\nu_2 + \xi_2^2\nu_4)\sigma^2}{nhf(u_0)(\xi_4 - \xi_2^2)^2}e_{p,p}^T\Psi(I)e_{p,p} + o_p(\frac{1}{nh})$$

where $\Psi(I) = E[(\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T)^{-1}|U]$.

The proofs of Theorem 2.1 and other theorems are given in the Appendix. By Theorem 2.1, the asymptotic bias is independent of the group size I as long as it is small enough. The group size I is involved in the asymptotic variance in the factor $\Psi(I)$, i.e., $\mathrm{E}[(\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T)^{-1}|U]$. As a matrix generalization of the random variable case, $\Psi(I_1) - \Psi(I_2)$ is positive definite if $I_1 < I_2$. Thus the asymptotic variance of the local average estimator will decrease as I gets larger. This can be easily understood. We can consider $\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T$ as a covariance matrix estimator if X is centralized. Then the larger I will give a better estimator with smaller variance, which naturally leads to less variation in our final local average estimator.

Compared with Fan and Zhang(1999)'s two step estimator in the theoretical aspect , the local average estimator has the same asymptotic variance of $O((nh)^{-1})$. For the asymptotic bias, our local average estimator is of $O(h^4)$ as well, but the formula is more concise since we don't have the term dominated by the initial bandwidth. Also, the conditional MSE of the local average estimator can achieves the optimal rage of convergence $O_p(n^{-8/9})$ when h is taken of order $n^{-1/9}$. Other theoretical advantages of Fan and Zhang(1999)'s two step estimator also hold in the local average estimator. For example, the estimators has the same optimal convergent rate as in the ideal situation where $a_1, ..., a_{p-1}$ are known.

The following Theorem 2.2 provides the asymptotic properties of the average estimator in the case that we consider the objective coefficient a_p shares the same smoothness with others. That is to say, a_p has only continuous and bounded second derivative. So in the local polynomial smoothing step, we applied a linear fit.

Theorem 2.2. *Under conditions 1,3-7, if $h \to 0$ and $nh \to \infty$, then the asymptotic conditional bias of $\hat{a}_p(u_0)$ in simple varying-coefficient model is given by*

$$bias(\hat{a}_p(u_0)|U, X) = \frac{1}{2}\xi_2 a_p''(u_0)h^2 + o_p(h^2)$$

and the asymptotic conditional variance of $\hat{a}_p(u_0)$ is given by

$$var(\hat{a}_p(u_0)|U, X) = \frac{\nu_0\sigma^2}{nhf(u_0)} e_{p,p}^T \Psi(I) e_{p,p} + o_p(\frac{1}{nh})$$

where $\Psi(I) = E[(\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T)^{-1}|U]$.

Now the asymptotic bias is of $O(h^2)$ and the asymptotic variance is of $O((nh)^{-1})$. The statistical efficiency is still kept. What's more, the asymptotic result is the same as that of the one-step estimator, and the bias is one term less compared with the two-step estimator. In other words, the local average estimator performs as well as the one-step estimator when there is no smoothness difference among the coefficients.

Notice that we apply local polynomial smoothing in the second step and the above asymptotic properties are all based on this setting. Obviously, the asymptotic results will change if different smoothing method is chosen. However, the primary estimators are asymptotic unbiased and their variances have explicit forms. What's more, those estimators are independent. Therefore, most classical regression models are feasible for the smoothing step and their asymptotic properties will not be skewed. The final asymptotic results will be a simple plug-in merely. In this way, our local average estimator gains more flexibility. Prior information about the objective functional coefficients could be fully utilized with various smoothing methods.

Theorem 2.3. *Under conditions 1, 3-7, the local average estimator of* $\mathbf{b}$ *in semi-varying coefficient model is asymptotic normal, i.e.*

$$\sqrt{n}(\hat{\mathbf{b}} - \mathbf{b}) \to N(0, \sigma^2\Sigma^{-1})$$

where

$$\Sigma = E(ZZ^T) - E\{E[(\frac{1}{I}\sum_{j=1}^{I} Z_j X_j^T)(\frac{1}{I}\sum_{j=1}^{I} X_j X_j^T)^{-1}(\frac{1}{I}\sum_{j=1}^{I} X_j Z_j^T)|U_1, ..., U_I]\}.$$

Theorem 2.3 states that the local average estimator of the constant coefficient is consistent. Actually it can be shown that the estimator is also asymptotically unbiased. The group size I effects the asymptotic variance only as long as it is small enough. If we consider the case when $p = 1$ and $X = 1$, then the model will turn into a partially linear model:

$$y = a(U) + Z^T\mathbf{b} + \epsilon.$$

By Theorem 2.3, the asymptotic variance will become $\frac{I}{I-1}\sigma^2\widetilde{\Sigma}^{-1}$, with

$$\widetilde{\Sigma} = \mathrm{E}[\{X - \mathrm{E}(X|Z)\}\{X - \mathrm{E}(X|Z)\}^T].$$

This is consistent with the result of Cui *et al.*(2014). However, notice that Bickel *et al.*(1993) have shown that $\sigma^2\widetilde{\Sigma}^{-1}$ is the semiparametric information bound. This implies that our local estimator doesn't reach the semiparametric efficient bound for general varying-coefficient partially linear model. We consider this inefficiency as the expense for the asymptotically unbiased and the computation simplicity.

2.4 Application and Simulation

2.4.1 A simple application

In this section, we apply our local average method to an environmental data set. The data set records daily measurements of air pollutants and other environmental factors in Hong Kong from January 1, 1994 to December 31, 1995; see Fan and Zhang(1999). Here we want to study the association between the air pollutants level and the number of hospital admissions for circulation and respiration problem. The air pollutants we considered are Sulphur Diocide, Nitrogen Dioxide and respirable suspended particulate, denoted as X_2, X_3 and X_4. All are measured in $\mu g/m^3$. The respond variable Y represents the number of daily hospital admissions and $U = t =$time. Also we will include an intercept term $X_1 = 1$.

First we scatter the daily number of hospital admissions for circulation and respiration with time t in Figure 2.1. We could see a clear increasing trend and some possible seasonal circular waves. Thus we believe there must be some intersecting

relationship between time and air pollutants levels. The following varying coefficient model is proposed to fit the data

$$Y = a_1(t) + a_2(t)X_2 + a_3(t)X_3 + a_4(t)X_4 + \epsilon.$$

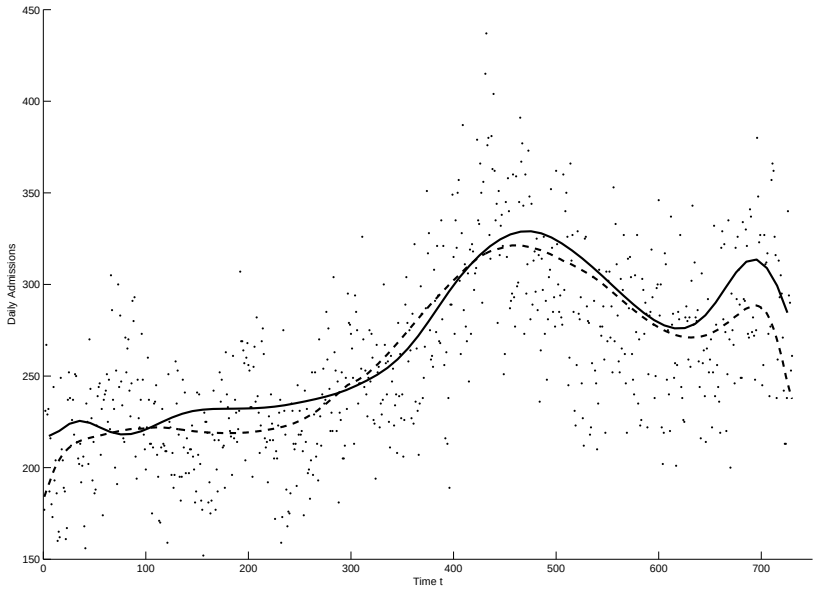

Figure 2.1: Scatter of daily hospital admissions and expected curve when pollutant levels are set at averages. Solid line: full model. Dashed: semi-varying coefficient model.

In this application, we choose $I = 10$ to be our group size. The bandwidths were chosen to be 30% of the interval length in the smoothing step. The estimated coefficient functions were depicted in Figure 2.2. Together with the estimated functions, we also plot the pointwise 95% confidence bonds in dashed. The confidence bonds are calculated directly from the Theorem 2.1 with residual variance estimated by local average method(See next chapter for more details). From the figure we can conclude there is time effect on at least one coefficient. In addition, if we set the pollutants levels at their averages, we will get how the expected number of hospital admissions change over time. The solid line in Figure 2.1 shows this results. Now the increase in the Year 1995 and the seasonal effect are more obvious.

Of course we want to know whether the coefficients are statistically significantly time varying but this problem is beyond the discussion of this chapter. According

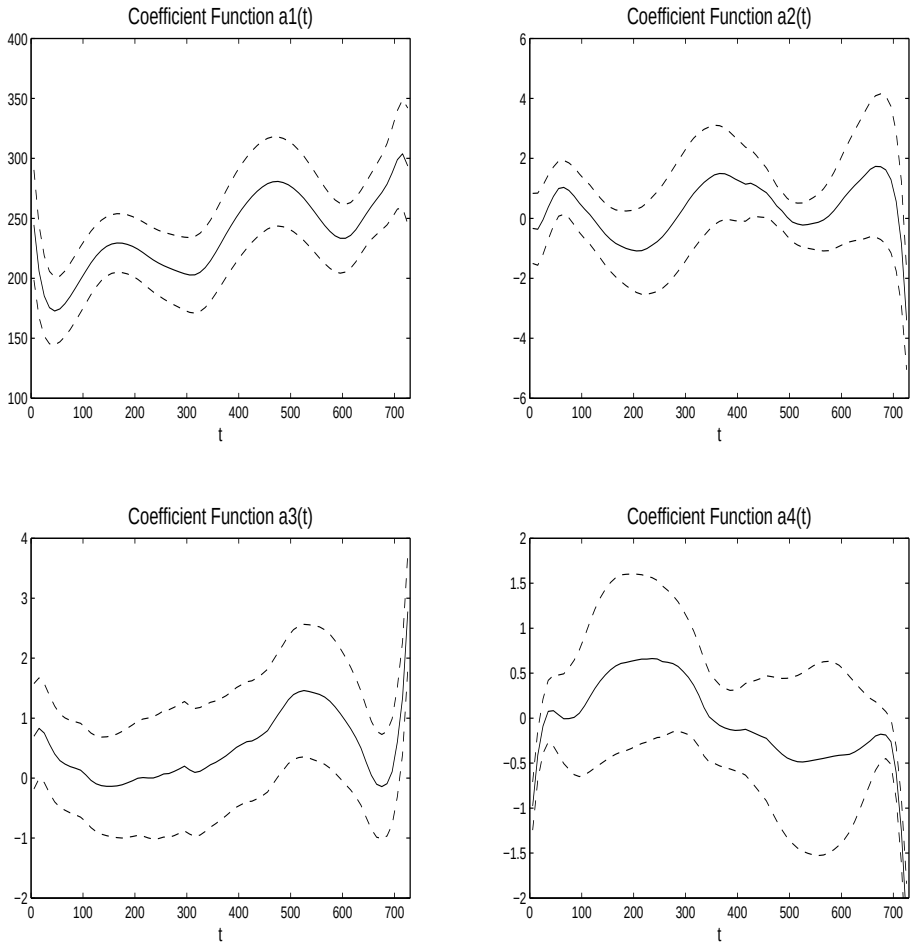

Figure 2.2: The estimated coefficient functions with pointwise 95% confidence intervals for the full model.

to the analysis of Fan and Zhang(2000), we can not reject the hypothesis that the fourth coefficient a_4 is a constant. So we let the coefficient of X_4 be a constant and proposed a semi-varying coefficient model

$$Y = a_1(t) + a_2(t)X_2 + a_3(t)X_4 + a_4X_4 + \epsilon$$

For this model, we still let $I = 10$ and use the 30% of the interval length to be the bandwidths. Then we get the estimator of the constant coefficient a_4 is -0.0581. For the nonparametric part, we just smooth the primary results. Figure 2.3 plots the estimated coefficient functions. They describe how the coefficients vary with time. Compared with the coefficient functions in Figure 2.2, the varying extent of the

coefficients in semi-varying model is more strong. We also plot the expected number of hospital admissions under this semi-varying model. It is shown in dashed in Figure 2.1. The overall trends of the two expected curves are alike and main differences appear at boundaries. In the point view of fitness, the dashed semi-varying curve seems to perform better. In all, the daily hospital admissions for respiratory and circulatory shows an overall increasing trend and some seasonal patterns.

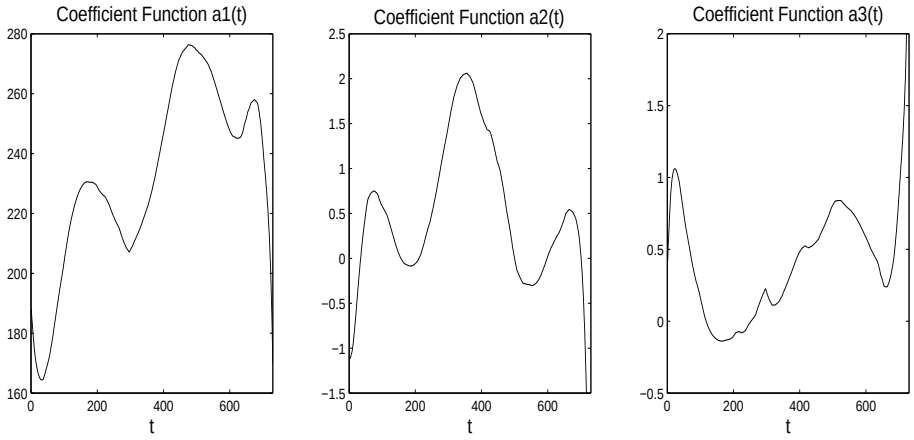

Figure 2.3: The estimated coefficient functions for semi-varying coefficient model.

2.4.2 Simulation for varying coefficient model

To investigate the performance of our method, we select following three examples.

Example 1. $Y = \sin(60U)X_1 + 4U(1-U)X_2 + \epsilon$.

Example 2. $Y = \sin(6\pi U)X_1 + \sin(2\pi U)X_2 + \epsilon$.

Example 3. $Y = \sin(8\pi(U-0.5))X_1$
$$+ \{3.5[\exp(-(4U-1)^2) + \exp(-(4U-3)^2)] - 1.5\}X_2 + \epsilon$$

where U is uniformly distributed on $[0,1]$, X_1, X_2 are bivariate normally distributed with mean vector $\vec{0}$ and covariance matrix $\begin{pmatrix} 1 & 2^{-1/2} \\ 2^{-1/2} & 1 \end{pmatrix}$. Moreover, ϵ, U and (X_1, X_2) are independent. The error term ϵ is normally distributed with mean zero and variance σ^2. To make signal-to-noise ratio be about 5:1, the variance σ^2 is chosen as

$$\sigma^2 = 0.2\mathrm{Var}[m(U, X_1, X_2)] \text{ with } m(U, X_1, X_2) = \mathrm{E}[Y|U, X_1, X_2].$$

These examples were proposed in Fan and Zhang(1999). In their paper, they used these three examples to study the performances of one-step estimator and two-step estimator. Thus we use these examples as well so that it will be convenient to make comparisons.

For each example, the objective functional coefficient is a_2 and 100 replications are conducted with sample size 500 and 1000. Mean integrated squared errors (MISE) are recorded to evaluate the performance of the estimators. We plot the MISE curve against bandwidth h to find the optimal h for each example. As in Fan and Zhang(1999), among the 100 replications, we choose the one such that the local-average estimator with $I = 10$ has the median performance. When the sample is selected, the one-step estimator and two-step estimator are also computed. Here, the optimal h for one-step and two-step estimator are both referred to the results presented in Fan and Zhang(1999). The Figure 2.4 shows the results of three different estimators for each example when sample size $n = 500$.

Since the two coefficients in each example have different smoothness, the one-step estimator can't meet the requirement of different bandwidths . When a_2 needs a large bandwidth, a linear fit with such large bandwidth is inefficient for a_1. If we adjust the bandwidth to a small one, a_2 will be undersmoothed. Therefore we can find the bias of one-step estimator in Figure 2.4 is quite large, especially for example 2 and example 3. Local-average estimator and two-step estimator are similar in example 1 and example 2. Both of them perform very well, due to the flexible choice of bandwidth. In example 3, the two-step estimator works well, though the local-average estimator gives a better estimation.

The simulation results are consistent with the theoretical results. When there exist different degrees of smoothness, the local-average estimator and the two-step estimator perform better than the one-step estimator since their asymptotic biases are of $O(h^4)$ while that of one-step estimator is $O(h^2)$. In this way we can explain why the one-step estimator has more bias in Figure 2.4. Besides, the asymptotic bias of local-average estimator is one term less than that of the two-step estimator. So the performance of local-average estimator is sometimes better than two-step estimator.

Figure 2.5 plots the MISE of local average estimator as a function of bandwidth.

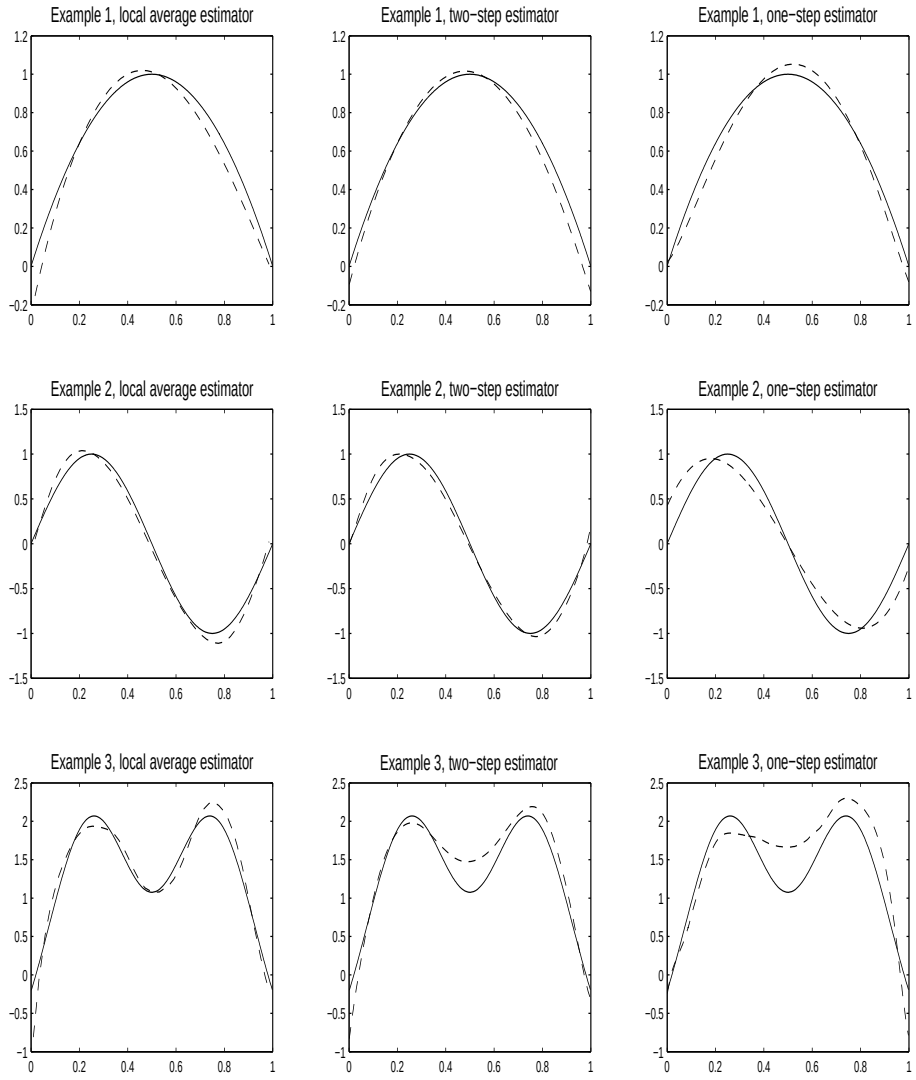

Figure 2.4: Some typical results. Solid curve: true value; dashed curve: estimated value.

We can find that as the sample size grows, the estimation results become better. The change with group size I is in accord with our theoretical conclusion. A larger I leads to a smaller asymptotic variance and has no influence on bias, so that the MISE becomes smaller. However, one can notice that the improvement of $I = 5$ from $I = 4$ is almost the same with that of $I = 10$ from $I = 5$. The marginal effect is decreasing quickly. Therefore $I = 10$ can already give a good estimation, though theoretically

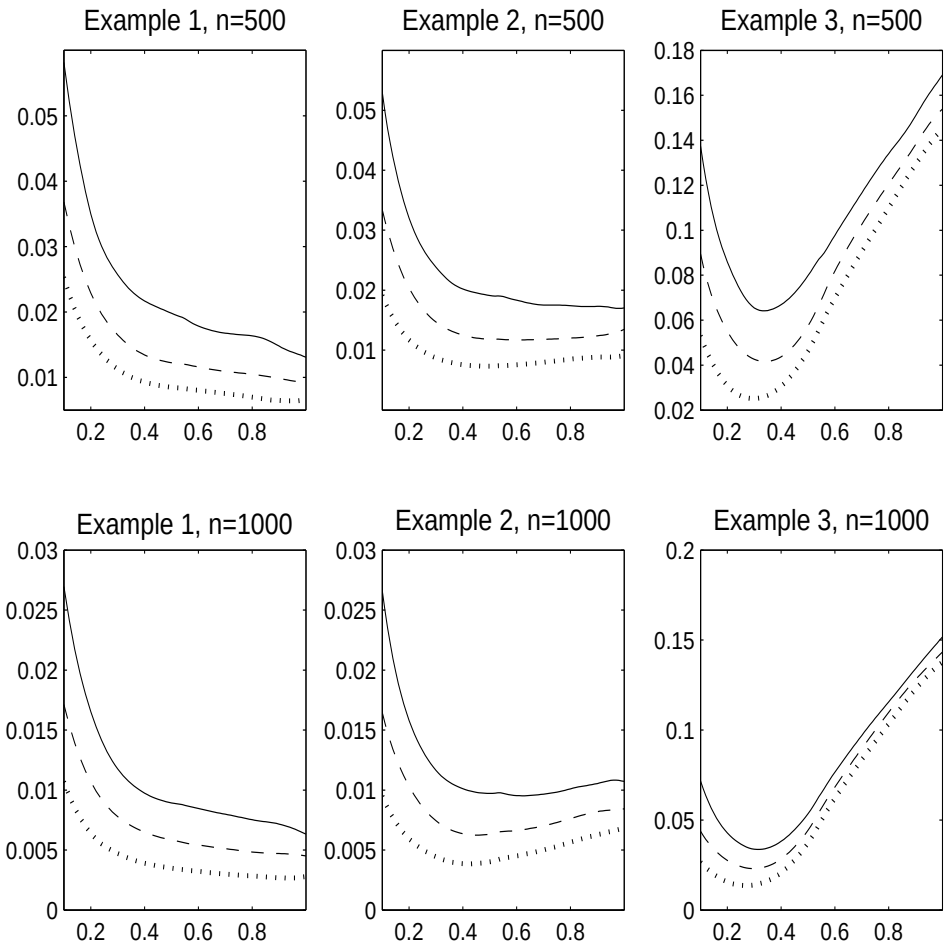

Figure 2.5: MISE as a function of bandwidth. Solid curve: I=4; dashed curve: I=5; dotted curve: I=10.

a large I may be preferred. One can also notice that the trends for different I are similar. This indicates that group size I and the bandwidth h in the smoothing step are independent. Thus it should not bother a lot to choose the group size I.

Another outstanding advantage of local average estimator is the computation simplicity. Table 2.1 shows the typical time spent by local average estimator, two step estimator and one step estimator. The sample data used in each example is the same as in Figure 2.4 and all the required bandwidths are selected in advance. The time listed in the table are obtained by the functions "tic" and "toc" in MATLAB. It is not a standard measurement of the computation time, but it can still give us a general understanding of the situations. We can find the significant advantage of

the local average estimator, as shown in the Table 2.1. It is not difficult to find the reason. In each estimator, most of the computations are involved in the weighted least squares process. For two step estimator and one step estimator, the weighted least squares process has to deal with a $n \times n$ matrix. However, for local average estimator, the largest matrix size in weighted least squares process is $k \times k$. Since $n = kI$ and I is a positive integer, the matrix size of local average estimator is much smaller than that of the other two estimators in weighted least squares process. In this way, the local average estimator saves a lot of computations. It can be thought that in the "average" step of our estimator, we have done some data mining to get a more corrected, ordered and simplified data set. The "average" step not only concentrates the information but also makes the disturbance abate.

Table 2.1: Typical time (in seconds) used by different estimators, I

	Example 1	Example 2	Example 3
local average	0.21	0.20	0.20
two step	1.37	1.32	1.22
one step	1.41	1.39	1.48

2.4.3 Simulation for semi-varying coefficient model

For the simulation of semi-varying coefficient model, we use the following examples.

$$\text{Example 4. } Y = \sin(2\pi U)X_1 + \cos(2\pi U)X_2 + X_3 + \epsilon.$$

$$\text{Example 5. } Y = \sin(2\pi U)X_1 + \{3.5[\exp(-(4U-1)^2) + \exp(-(4U-3)^2)] - 1.5\}X_2 + X_3 + \epsilon.$$

$$\text{Example 6. } Y = \sin(6\pi(U-0.5))X_1 + \sin(2\pi U)X_2 + X_3 + \epsilon.$$

where U is uniformly distributed on $[0, 1]$, X_1, X_2 and X_3 follow a standard normal distribution. ϵ is normally distributed with mean 0 and variance σ^2. The σ^2 in each example is selected so that the signal-to-noise ratio is 5:1. Further, U, X_1, X_2, X_3 and ϵ are all independent.

1000 replications with sample size $n = 250$ and $n = 500$ are conducted for each of the above examples. For the constant coefficients, the mean, the standard error and

the mean squared error(mse) of the estimators are recorded in the Table 2.2. For the functional coefficients, we only estimate the second coefficient a_2 in each example for a simple illustration.

Table 2.2: Simulation results of the constant coefficients

n	I	Example 1			Example 2			Example 3		
		mean	std	mse	mean	std	mse	mean	std	mse
250	4	1.0006	0.0561	0.0031	1.0073	0.0802	0.0065	1.0022	0.0588	0.0035
	5	0.9971	0.0505	0.0026	0.9958	0.0685	0.0047	1.0010	0.0532	0.0028
	10	0.9993	0.0456	0.0021	0.9980	0.0631	0.0040	1.0021	0.0472	0.0022
500	4	0.9992	0.0396	0.0016	0.9997	0.0558	0.0031	1.0001	0.0409	0.0017
	5	1.0012	0.0364	0.0013	1.0019	0.0504	0.0025	1.0003	0.0350	0.0012
	10	1.0000	0.0315	0.0010	1.0025	0.0448	0.0020	1.0011	0.0319	0.0010

Form Table 2.2, we can find that our estimators in all the example are very close to the true value 1. The estimator becomes better when the sample size grows larger. For the group size I, we can find that it makes no particular difference on the mean while a larger I gives a smaller standard deviation. This phenomena is consistent with our theoretical results, since I only appears in the asymptotic variance.

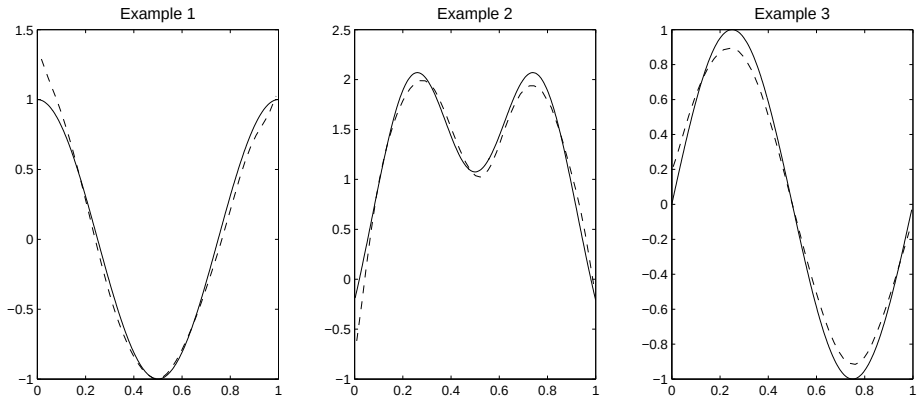

Figure 2.6: Functional coefficient estimation in semi-varying coefficient model. Solid curves:true functions; dashed curve:estimation functions.

Figure 2.6 is a simple illustration of the estimation of the functional coefficient for the semi-varying coefficient model. When we get the estimator $\widehat{\mathbf{b}}$ for the constant coefficient, we also get the primary estimators of the functional coefficient $\widehat{\mathbf{a}}$. Of course we can substitute $\widehat{\mathbf{b}}$ back to the original model and get a simple varying

coefficient model to solve the problem. Here in this simulation, we just make use of the primary estimators $\widehat{\mathbf{a}}$ to continue the estimation. The same as in simple varying-coefficient model, we adapt the local cubic polynomial method. 1000 random samples with sample size $n = 500$ are drawn from each example model. Integrated squared errors are recorded for all the 1000 estimators based on the group size $I = 10$. The bandwidth is selected as 0.4, 0.27, 0.5 for example 4, 5, 6 respectively. Figure 2.6 plots the estimator that gains the median performance in each example. We can see that the local average estimator also works well for the functional coefficients. Compare with the simple varying-coefficient case, which depicted in Figure 2.4, it seems that the linear part doesn't screw the efficiency. Besides, notice that in example 6, the two functional coefficients possess different smoothness. Obviously, our local average estimator can well handle this problem.

Now we want to compare the local average estimator with other estimators. Another 1000 replicates with sample size $n = 500$ of each example are generated and we use different methods to estimate the constant coefficient $b = 1$. Table 2.3 lists the mean, the standard deviation and the mse of the local average estimator($\hat{b}_{\mathrm{LA}}$) with group size $I = 10$, Zhang et al.(2002)'s estimator($\hat{b}_{\mathrm{Z}}$), Fan and Huang(2005)'s estimator($\hat{b}_{\mathrm{F}}$) and Xia et al.(2004)'s estimator($\hat{b}_{\mathrm{X}}$). First notice that all the means are very close to the true value. The difference is less than 0.001, which is a quite small error. The local average estimator seems to have less bias, but this advantage is insignificant. For the standard deviation, the one of the local average estimator is the largest. So the mse of the local average estimator is the largest, too. We should have expected this result since Theorem 2.3 has already implied the inefficiency of the local average estimator.

However, we can get rid out this inefficiency easily. Here we introduce an updated estimator $\hat{b}_{\mathrm{update}}$. We will estimate the functional coefficients first, then substitute them back to get a linear regression model to evaluate the constant coefficients. The detailed process is as follows. Just as what we did in Figure 2.6, we make use of the primary estimators of the functional coefficient $\widehat{\mathbf{a}}$, which comes along with the estimator. With a local cubic polynomial fit, we will obtain the functional coefficients estimator $\widehat{\mathbf{a}}(U_i), i = 1, ..., n$. Substitute $\widehat{\mathbf{a}}(U_i), i = 1, ..., n$ back and rearrange the

Table 2.3: Simulation results of different estimators

Example 1					
	$\hat{b}_{\mathrm{LA}}$	$\hat{b}_{\mathrm{Z}}$	$\hat{b}_{\mathrm{F}}$	$\hat{b}_{\mathrm{X}}$	$\hat{b}_{\mathrm{update}}$
mean	0.9997	0.9993	0.9995	0.9993	0.9997
std	0.0313	0.0292	0.0287	0.0292	0.0286
mse	0.0010	0.0009	0.0008	0.0009	0.0008
Example 2					
	$\hat{b}_{\mathrm{LA}}$	$\hat{b}_{\mathrm{Z}}$	$\hat{b}_{\mathrm{F}}$	$\hat{b}_{\mathrm{X}}$	$\hat{b}_{\mathrm{update}}$
mean	0.9997	0.9990	0.9993	0.9990	0.9996
std	0.0429	0.0400	0.0392	0.0398	0.0392
mse	0.0018	0.0016	0.0015	0.0016	0.0015
Example 3					
	$\hat{b}_{\mathrm{LA}}$	$\hat{b}_{\mathrm{Z}}$	$\hat{b}_{\mathrm{F}}$	$\hat{b}_{\mathrm{X}}$	$\hat{b}_{\mathrm{update}}$
mean	0.9998	0.9992	0.9995	0.9992	0.9994
std	0.0316	0.0294	0.0289	0.0293	0.0291
mse	0.0010	0.0009	0.0008	0.0009	0.0008

model, we can have

$$\tilde{y}_i \equiv y_i - X_i^T \widehat{\mathbf{a}}(U_i) = Z_i^T \mathbf{b} + \epsilon_i^*, \quad i = 1, ..., n.$$

Treat this as a linear regression model, we can get the updated estimator easily by least squares.

The last column of Table 2.3 shows the simulation results of the updated estimator. It can be found that the standard deviation is remarkably decreased compared with the original local average estimator. What's more, it is more or less the same as the standard deviation of Fan and Huang(2005)'s estimator $\hat{b}_{\mathrm{F}}$. In Fan and Huang(2005), they have claimed that their estimator $\hat{b}_{\mathrm{F}}$ is semiparametrically efficient. So is the updated estimator. The efficiency of the updated estimator can be shown by the Theorem 5.45 (One-step estimation) in Van der Vaart(2000). Since the functional coefficients in updated estimator can be written in the form of the local average estimator, the updated estimator is just a one-step estimator given the local average estimator. That's also why we call it "updated". From the table we can find that the updated estimator always gets the best mse.

Then we still want to discuss the computation simplicity. Table 2.4 shows the

typical time used by the above 5 different estimators. The three example samples are the same ones in Figure 2.6, in which the local average estimator gains median performance. The same with Table 2.1, we use the functions 'tic' and 'toc' in MATLAB to do the timing. Here we can see a huge advantage of the local average estimator. The time spent by other estimators are tens of that spent by local average estimator. $\hat{b}_{\mathrm{X}}$ has to deal with a $n^2 \times n^2$ matrix so it needs more time. For the updated estimator, it is close to Fan and Huang(2005)'s estimator.

Table 2.4: Typical time (in seconds) used by different estimators, II

	Example 1	Example 2	Example 3
$\hat{b}_{\mathrm{LA}}$	0.01	0.01	0.01
$\hat{b}_{\mathrm{Z}}$	0.34	0.32	0.29
$\hat{b}_{\mathrm{F}}$	0.45	0.47	0.39
$\hat{b}_{\mathrm{X}}$	14.47	13.12	13.83
$\hat{b}_{\mathrm{update}}$	0.55	0.40	0.40

From all these simulations, we can conclude that the local average estimator $\hat{b}_{\mathrm{LA}}$ can give a good estimation and dramatically reduce the computation burden. Though it is not asymptotically efficient, local average estimator can be a good primary estimator or pilot estimator. Furthermore, the updated estimator can make up the inefficiency while doesn't increase too much computation complexity.

2.5 Concluding remarks

In this chapter, we introduce a new estimator for the varying coefficient model — local average estimator. Naturally, we apply it to the semi-varying coefficient model. Both of the theoretical and simulation results show that the local average method gives out a satisfying estimation. For the simple varying coefficient model, our estimator can easily deal with the different smoothness problem and reach an optimal rate of convergence $O_p(n^{-8/9})$. For the semi-varying coefficient model, the local average estimator for the constant part is asymptotically unbiased and asymptotically normal.

The most impressive contribution of our estimator is the computation simplicity. With a "taking average" step, we concentrate the information and decrease the sample

size. Though a smaller sample size seems to lose some information, the theoretical result indicates that the local average estimator has the same order of variance as we have known all the nuisance estimators. This is because we also decrease the dispersion of the error term in the "average" step. Therefore, we ease the computation burden but still keep the efficiency. From the simulation part we can find that the time spent by local average estimator is significantly less than that of other estimators. This sharply reduced computation time gives the local average estimator a great credit.

This chapter only discussed some most basic applications of the local average method. Much more researches can be conducted in the future. For the semi-varying coefficient model, the updated estimator of the constant coefficients $\mathbf{b}$ has been proposed in the simulation section. Clearly, more theoretical and computational results of this updated estimator need to be studied. Besides, all the estimation in this chapter is based on that we have already known which coefficients are constant and which are functional. Lack of this prior information, we are faced with a model selection problem. Here we can still solve this problem with local average method: after getting the primary estimator, we can decide whether this coefficient has a constant trend through some classical model checking procedures. What's more, the local average method can also adapt to the additive model.

2.6 Appendix

The proof of Theorem 2.2 is almost the same as that of Theorem 2.1. The only difference is that we use a polynomial of order 3 for smoothing in Theorem 2.1 while order 1 is used in Theorem 2.2. Thus we will prove Theorem 2.1 and Theorem 2.3 only.

Proof of Theorem 2.1. We only need to show the mean and variance of the primary estimator are well defined, so that we can transform it into a simple smoothing problem. The final asymptotic conditional bias and variance are just a simple application of the smoothing method.

Note that the primary estimator $\mathbf{a} = (\mathbf{a}_1^T, \mathbf{a}_2^T, ..., \mathbf{a}_k^T)^T$, and that the components

$\mathbf{a}_i,\ i = 1, ..., k$ are actually calculated separately. So without losing generality, we will discuss $\mathbf{a}_i$ only.

$$\begin{aligned}
\hat{\mathbf{a}}_i &= (\hat{a}_1(\bar{U}_{i\cdot}), \hat{a}_2(\bar{U}_{i\cdot}), ..., \hat{a}_p(\bar{U}_{i\cdot}))^T \\
&= (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Y}_i \\
&= (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T \begin{pmatrix} X_{i1}^T\mathbf{a}(U_{i1}) + X_{i1}^T\mathbf{a}(\bar{U}_{i\cdot}) - X_{i1}^T\mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{i1} \\ X_{i2}^T\mathbf{a}(U_{i2}) + X_{i2}^T\mathbf{a}(\bar{U}_{i\cdot}) - X_{i2}^T\mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{i2} \\ \cdots \\ X_{iI}^T\mathbf{a}(U_{iI}) + X_{iI}^T\mathbf{a}(\bar{U}_{i\cdot}) - X_{iI}^T\mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{iI} \end{pmatrix} \\
&= \mathbf{a}(\bar{U}_{i\cdot}) + (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\boldsymbol{\epsilon}_i + (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T \begin{pmatrix} X_{i1}^T(\mathbf{a}(U_{i1}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ X_{i2}^T(\mathbf{a}(U_{i2}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ \cdots \\ X_{iI}^T(\mathbf{a}(U_{iI}) - \mathbf{a}(\bar{U}_{i\cdot})) \end{pmatrix}
\end{aligned}$$

Now we need to prove that for each $a_l(\cdot), l = 1, ..., p$ and any i, j, $|a_l(U_{ij}) - a_l(\bar{U}_{i\cdot})| = O_p(\frac{\ln n}{n})$. Let $F(\cdot)$ be the cumulative distribution of U, i.e., $F'(u) = f(u)$. Besides, let $\tau = F(U)$, so we can regard τ as a uniformly distributed variable in the interval [0,1]. We denote two consecutive order statistics by $U_{(i+1)}, U_{(i)}$, and $\tau_{(i+1)}, \tau_{(i)}$ are the corresponding uniformly distributed variables. Then we have

$$\begin{aligned}
|a_l(U_{ij}) - a_l(\bar{U}_{i\cdot})| &= |a_l'(\xi_{ij})||U_{ij} - \bar{U}_{i\cdot}| \\
&\leqslant |a_l'(\xi_{ij})| \cdot \frac{I-1}{2} \max |U_{(i+1)} - U_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max |F^{-1}(\tau_{(i+1)}) - F^{-1}(\tau_{(i)})| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max (F^{-1})'(\eta)|\tau_{(i+1)} - \tau_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max \frac{1}{f(u_\eta)}|\tau_{(i+1)} - \tau_{(i)}| \\
&\leqslant \frac{(I-1)|a_l'(\xi_{ij})|}{2\delta} \max |\tau_{(i+1)} - \tau_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2\delta} O_p(\frac{\ln n}{n})
\end{aligned}$$

where ξ_{ij} is between U_{ij} and $\bar{U}_{i\cdot}$, η is between $\tau_{(i+1)}$ and $\tau_{(i)}$, $u_\eta = F^{-1}(\eta)$. The last equation holds by the Theorem 3.1 of Lars Holst(1980). Therefore,

$$\hat{\mathbf{a}}_i = \mathbf{a}(\bar{U}_{i\cdot}) + (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\boldsymbol{\epsilon}_i + \mathbf{O}_p(\frac{\ln n}{n})$$

and

$$\mathrm{E}(\hat{\mathbf{a}}_i) = \mathbf{a}(\bar{U}_{i\cdot}) + O_p(\frac{\ln n}{n}),$$
$$\mathrm{var}(\hat{\mathbf{a}}_i) = \mathrm{E}[(\mathbf{X}_i^T\mathbf{X}_i)^{-1}|\mathbf{U}]\sigma^2 = \mathrm{E}[(\sum_{j=1}^{I} X_j X_j^T)^{-1}|\mathbf{U}]\sigma^2.$$

What's more, since the ordering is no longer needed in the following smoothing step, we can naively consider the primary estimators $(\bar{U}_{i\cdot}, \hat{\mathbf{a}}_i), 1 = 1, ..., k$ are independent and identically distributed.

In our setting, we use the local polynomial regression to do the smoothing. Then we can simply apply the Theorem 3.1 in Fan and Gijbels(1996) to get the final results. □

Proof of Theorem 2.3. Similarly as in proof of Theorem 2.1, for each group, we can write

$$\begin{aligned}
\mathbf{Y}_i &= (y_{i1}, y_{i2}, ..., y_{iI})^T \\
&= \begin{pmatrix} X_{i1}^T(\mathbf{a}(U_{i1}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ X_{i2}^T(\mathbf{a}(U_{i2}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ \cdots \\ X_{iI}^T(\mathbf{a}(U_{iI}) - \mathbf{a}(\bar{U}_{i\cdot})) \end{pmatrix} + \mathbf{X}_i^T\mathbf{a}(\bar{U}_{i\cdot}) + \mathbf{Z}_i^T\mathbf{b} + \boldsymbol{\epsilon}_i \\
&= \mathbf{O}_p(\frac{\ln n}{n}) + \mathbf{X}_i^T\mathbf{a}(\bar{U}_{i\cdot}) + \mathbf{Z}_i^T\mathbf{b} + \boldsymbol{\epsilon}_i
\end{aligned}$$

Then

$$\begin{aligned}
\widehat{\mathbf{b}} &= (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{Y} \\
&= (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{O}_p(\frac{\ln n}{n}) + \mathbf{b} + (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\epsilon} \qquad (2.3)
\end{aligned}$$

Here we can notice that the local average estimator is asymptotically unbiased. Now we will continue to prove the asymptotic normality.

By equation (2.3), we can have

$$\begin{aligned}
\sqrt{n}(\widehat{\mathbf{b}} - \mathbf{b}) &= (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{O}_p(\frac{\ln n}{\sqrt{n}}) + \sqrt{n}(\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\epsilon} \\
&= \sqrt{n}(\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\epsilon} + o_p(1)
\end{aligned}$$

Since

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_1 \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} & \mathbf{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_k & \mathbf{Z}_k \end{pmatrix},$$

then

$$\mathbf{\Phi}^T\mathbf{\Phi} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}_1^T\mathbf{Z}_1 \\ \mathbf{0} & \mathbf{X}_2^T\mathbf{X}_2 & \cdots & \mathbf{0} & \mathbf{X}_2^T\mathbf{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_k^T\mathbf{X}_k & \mathbf{X}_k^T\mathbf{Z}_k \\ \mathbf{Z}_1^T\mathbf{X}_1 & \mathbf{Z}_2^T\mathbf{X}_2 & \cdots & \mathbf{Z}_k^T\mathbf{X}_k & \sum_{i=1}^k \mathbf{Z}_i^T\mathbf{Z}_i \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix}$$

where $\mathbf{A} = diag(\mathbf{X}_1^T\mathbf{X}_1, ..., \mathbf{X}_k^T\mathbf{X}_k)$, $\mathbf{B} = (\mathbf{Z}_1^T\mathbf{X}_1, ..., \mathbf{Z}_k^T\mathbf{X}_k)^T$ and $\mathbf{C} = \sum_{i=1}^k \mathbf{Z}_i^T\mathbf{Z}_i$.

Then it can be inverted blockwise as following:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1} & (\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}$$

So

$$\begin{aligned} & (\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\mathbf{\Phi}^T\mathbf{\Phi})^{-1} \\ =& \begin{pmatrix} -(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1} & (\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix} \\ =& \{\sum_{i=1}^k \mathbf{Z}_i^T\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Z}_i\}^{-1} \begin{pmatrix} -\sum_{i=1}^k \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1} & \mathbf{I} \end{pmatrix} \end{aligned}$$

Then we can have

$$\begin{aligned} \sqrt{n}(\mathbf{0}_{1\times kp}, \mathbf{1}_{1\times q})(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\boldsymbol{\epsilon} &= \{\frac{1}{n}\sum_{i=1}^k \mathbf{Z}_i^T\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Z}_i\}^{-1} \\ &\quad \times \{\frac{1}{\sqrt{n}}\sum_{i=1}^k (\mathbf{Z}_i^T - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1})\boldsymbol{\epsilon}_i\} \\ &= R_1^{-1}R_2 \end{aligned}$$

Since

$$R_1 = \frac{1}{k}\sum_{i=1}^k [\frac{1}{I}\sum_{j=I}^I Z_{ij}Z_{ij}^T - (\frac{1}{I}\sum_{j=I}^I Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^I X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^I X_{ij}Z_{ij}^T)]$$

By law of large numbers, as $k \to \infty$

$$
\begin{aligned}
R_1 \to & \mathrm{E}[\frac{1}{I}\sum_{j=I}^{I} Z_{ij}Z_{ij}^T - (\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)] \\
= & \mathrm{E}[\mathrm{E}[\frac{1}{I}\sum_{j=I}^{I} Z_{ij}Z_{ij}^T - (\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)|U_{i1},...U_{iI}]] \\
= & \mathrm{E}[\frac{1}{I}\sum_{j=I}^{I} \mathrm{E}[Z_{ij}Z_{ij}^T|U_{ij}]] \\
& - \mathrm{E}[\mathrm{E}[(\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)|U_{i1},...U_{iI}]] \\
= & \mathrm{E}[ZZ^T] - \mathrm{E}[\mathrm{E}[(\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)|U_{i1},...U_{iI}]] \\
= & \Sigma
\end{aligned}
$$

Next we deal with the term R_2. Regardless of the ordering and given $\{(U_i, X_i, Z_i)\}$, $i = 1, ..., n$, $\boldsymbol{\epsilon}_i$ is independent of each other and has mean zero. Therefore, R_2 is asymptotically normal with mean zero. Then we only need to show the limiting variance of R_2.

$$
\begin{aligned}
& \mathrm{Var}(R_2|U, X, Z) \\
= & \frac{1}{n}\sum_{i=1}^{k}(\mathbf{Z}_i^T - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1})(\mathbf{Z}_i^T - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1})^T \times \mathrm{E}[\boldsymbol{\epsilon}_i^2|U, X, Z] \\
= & \sigma^2\frac{1}{n}\sum_{i=1}^{k}(\mathbf{Z}_i^T\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Z}_i) \\
\to & \sigma^2\mathrm{E}[\frac{1}{I}\sum_{j=I}^{I} Z_{ij}Z_{ij}^T - (\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)] \\
= & \mathrm{E}[ZZ^T] - \mathrm{E}[\mathrm{E}[(\frac{1}{I}\sum_{j=I}^{I} Z_{ij}X_{ij}^T)(\frac{1}{I}\sum_{j=I}^{I} X_{ij}X_{ij}^T)^{-1}(\frac{1}{I}\sum_{j=I}^{I} X_{ij}Z_{ij}^T)|U_{i1},...U_{iI}]] \\
= & \sigma^2\Sigma
\end{aligned}
$$

Therefore, by the Slutsky theorem, we have

$$\sqrt{n}(\widehat{\mathbf{b}} - \mathbf{b}) \to N(\mathbf{0}, \sigma^2\Sigma^{-1})$$

□

Chapter 3

Variance Estimation for Semi-parametric Regression Models

3.1 Introduction

In regression analysis, a resident part of the model is the error term, which is not only a representation of the randomness of the model, but also contains useful information such as the signal-to-noise ratio and goodness-of-fit. Variance estimation is a fundamental problem in statistical modelling and plays an important role in the inferences after model selection and estimation. Moreover, it provides a benchmark of prediction error when a given model is compared to the "oracle" model. In practice, variance estimation has wide application in many areas, such as quality control (Box and Ramirez, 1987), confidence interval construction (Carroll, 1987), and immunoassay (Butt, 1984), among others (Carroll and Ruppert, 1988).

Many methods of variance estimation have been brought up throughout the decades. For parametric regression models, the variance can be estimated by the least squares method or the maximum likelihood estimation method. For nonparametric and semiparametric regression models, it is a bit more challenging to estimate the variance accurately. For example, consider a simple nonparametric regression

model,

$$y_i = f(x_i) + \epsilon_i, i = 1, ..., n, \tag{3.1}$$

where $f(\cdot)$ is an unknown smooth function, and the error terms ϵ_i's are i.i.d. with mean zero and constant variance σ^2. Two main estimation approaches are reported in the literature. One approach is first to estimate the nonparametric function $f(\cdot)$ and then to calculate the estimate of residual variance based on the residual sum of squares (RSS) and its degrees of freedom. In theory, to estimate the nonparametric function $f(\cdot)$ accurately and efficiently, smoothness conditions are imposed on $f(\cdot)$ that may increase the possibility of model misspecification. In practice, smoothing techniques such as kernel methods or spline methods are implemented to estimate $f(\cdot)$. The selection of the bandwidth or the number and positions of basis functions is challenging and computationally intensive, which can greatly affect the estimation of $f(\cdot)$ and thus the estimations of RSS and its degrees of freedom.

Another approach is the difference-based method. This approach can directly estimate σ^2 without knowing much information about the nonparametric function $f(\cdot)$. Rice (1984) proposed a first-order difference-based method to estimate the residual variance. The basic idea is to avoid the estimation of unknown function $f(x_i)$ and remove its effect by taking the differences for adjacent x_i's and calculating the residual variance directly. More difference-based methods were introduced later, for example, by Gasser et al. (1986), Hall and Marron (1990), and Klipple and Eubank (2007). None of the above estimators achieves the asymptotic optimal efficiency for the mean squared error (Dette, Munk and Wagner, 1998). Müller et al. (2003) studied a class of difference-based estimators, and under certain assumptions for the difference weights, the asymptotic optimal efficiency can be achieved. Tong and Wang (2005) suggested a regression type of difference-based estimator for an equally spaced design. Furthermore, motivated by Tong and Wang (2005), Park, Kim and Lee (2012) explored the difference-based variance estimator for small sample nonparametric regression modeling problems. Brown and Levine (2007) considered the variance function estimation in nonparametric regression by the difference-based approach. The optimal convergence rates over broad function classes and bandwidths are fully characterized. Wang *et al.* (2008) further investigated the minimax rate of the convergence of the variance

function estimation in nonparametric regression and particularly studied the effect of the unknown mean function on the estimation of the variance function. Cai *et al.* (2009) extended Wang *et al.* (2008)'s results to the variance function estimation in multivariate nonparametric regression with a fixed design.

The difference-based methods are much more robust and computationally simple, but they often assume the nonparametric regression model has certain simple structure and the design point is univariate and equally spaced. It is unclear how to generalize this to more complicated nonparametric and semi-parametric models. Wang *et al.* (2011) and Brown *et al.* (2016) tried to use the difference-based approach to estimate the semi-parametric partially linear and multivariate partially linear model and obtained an optimal efficient estimator of the linear components in the model. Furthermore, testing methods were also developed for the statistical inference problems. It is appropriate to try to extend the difference-based approach to more complex nonparametric or semi-parametric regression models and not to focus solely on variance or variance function estimation.

In this chapter, we propose a local average method to estimate the residual variance for nonparametric and semi-parametric models based on the idea of partial consistency. First, we approximate the nonparametric function by a step constant function locally and then reparameterize nonparametric and semi-parametric models into high-dimensional linear models. The ordinary least squares method can then be implemented to calculate the residual and estimate the residual variance. The reparameterized linear model is high-dimensional because the number of nuisance parameters increases with the sample size. These nuisance parameters cannot be estimated consistently. However, the residual variance can be estimated consistently and with near efficiency because of the partial consistency. Compared to the difference-based method, our proposed method can be easily generalized to more complicated models and does not require equally spaced assumptions for the design point.

The remainder of this chapter is organized as follows. In Section 2, we first review the classical estimation methods for residual variance and then propose a local average variance estimation method and investigate its asymptotic properties. In Section 3, we extend the proposed method to semi-parametric partially linear

models and varying coefficient models and further propose a refined local average variance estimator and discuss numerical implementation issues. Furthermore, for the simple nonparametric regression, based on the basic motivation of our local average method, and combined with the ideas of kernel function and moving average, we suggest other two methods to improve the efficiency and stability of our original local average method. In Section 4, we discuss some applications of the proposed local average variance estimation method. Numerical studies and real data analysis are presented to illustrate the finite performance of the proposed method in Section 5. Some discussion and conclusions are given in Section 6. All of the proofs are relegated to the Appendix.

3.2 Variance estimation by local average

3.2.1 Review of classical methods

Consider the linear model,

$$y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, i = 1, ..., n, \quad \text{or} \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (y_1, ..., y_n)^T$ is an n-vector of responses, $\mathbf{X} = (X_1, ..., X_n)^T$ is an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a p-vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ is an n-vector of i.i.d. random errors with mean zero and variance σ^2. The ordinary least squares estimator $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}}{\text{tr}(\mathbf{I} - \mathbf{H})},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection operator onto the linear space generated by the column vectors of $\mathbf{X}$, $\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$ is the residual sum of squares (RSS), and $\text{tr}(\mathbf{I} - \mathbf{H})$ is the corresponding degrees of freedom. This estimator $\hat{\sigma}^2$ is the best linear unbiased estimator with asymptotic normality.

Consider the nonparametric regression model 3.1: in general, the residual vector and the RSS can be calculated, respectively, as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \hat{\mathbf{f}} = (\mathbf{I} - \mathbf{S})\mathbf{Y}, \quad \text{RSS} = \mathbf{Y}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{Y},$$

where $\mathbf{S}$ is a smoothing matrix and, in practice, can be constructed by various methods such as the kernel method and the spline method. Although $f(\cdot)$ can be estimated

efficiently, the bias would seriously affect the estimation of the residual variance. Moreover, it is a bit challenging to estimate the degrees of freedom of this RSS accurately because the matrix $(\mathbf{I}-\mathbf{S})$ is not a projection matrix as $\mathbf{H}$. In most situations, $\text{tr}(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})$ is used to approximate the degrees of freedom of this RSS, whose calculation depends on the choice of the bandwidth h or the number and locations of basis functions. Furthermore, the optimal bandwidth h or basis functions for the efficient variance estimator would not be the same as the optimal bandwidth h or basis functions for the estimator of the mean function. Because selection of the optimal bandwidth or basis functions is involved in many complex computational procedures, it may lead to an unstable or even a biased estimator of the residual variance.

Another approach is the difference-based method proposed by Rice (1984). Consider model 3.1 and, without loss of generality, assume $0 \leq x_1 \leq \cdots \leq x_n \leq 1$, then the difference-based estimator is defined as

$$\hat{\sigma}^2_{\text{R}}(k) = \frac{1}{2(n-k)} \sum_{i=k+1}^{n} (y_i - y_{i-k})^2.$$

Note that $y_i - y_{i-k} = f(x_i) - f(x_{i-k}) + \epsilon_i - \epsilon_{i-k}$. When x_i's are equally spaced, the first term of the difference of the mean functions $f(x_i)$ and $f(x_{i-k})$ should be of the order $O(n^{-1})$ given k is constant. It is much smaller than the second term, the difference of errors ϵ_i and ϵ_{i-k}, in terms of orders of magnitude, and thus can be ignored in calculating $\hat{\sigma}^2_{\text{R}}(k)$. Gasser et al. (1986) proposed a second-order difference-based estimator to improve the estimation efficiency,

$$\hat{\sigma}^2_{\text{GSJ}} = \frac{1}{n-2} \sum_{i=2}^{n-1} c_i^2 \hat{\epsilon}_i^2,$$

where $\hat{\epsilon}_i$ is the difference between y_i and the value at x_i of the line joining (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}), and c_i's are chosen such that $E[c_i^2 \hat{\epsilon}_i^2] = \sigma^2$ for all i when $f(\cdot)$ is linear. Müller et al. (2003) considered a class of difference-based estimators,

$$\hat{\sigma}^2_{\text{MSW}} = \frac{1}{2\sum_{i\neq j}^{n} W_{ij}} \sum_{i\neq j}^{n} W_{ij}(y_i - y_j)^2,$$

where W_{ij}'s are weights only depending on x_i's. They showed under certain assumptions that the asymptotic optimal rate of the mean squared errors can be achieved by $\hat{\sigma}^2_{\text{MSW}}$. For the difference-based method, notice that when x_i's are equally spaced

in $[0,1]$, i.e. $x_i = i/n$,

$$E[\hat{\sigma}^2_{\mathrm{R}}(k)] \asymp \sigma^2 + Jd_k, \quad 1 \le k \le m, \quad m = o(n),$$

where $d_k = k^2/n^2$ and $J = \int_0^1 \{f'(x)\}^2 dx/2$. Based on this fact, Tong and Wang (2005) considered a linear model $s_k = \alpha + \beta d_k + e_k$, where $s_k = \sum_{i=k+1}^{n} (y_i - y_{i-k})^2/\{2(n-k)\}$, $1 \le k \le m$, and proposed a variance estimator,

$$\hat{\sigma}^2_{\mathrm{T}} = \widehat{\alpha},$$

where $\widehat{\alpha}$ is the estimated intercept. This method reduces the asymptotic rate of mean squared errors to $O_p(n^{-1})$ with an optimal bandwidth m.

The difference-based methods have the advantages of avoiding the estimation of the nonparametric function and reducing the computational cost, but their extensions to more general settings are somewhat limited. For example, some methods such as Tong and Wang's method require the x_i's to be equally spaced, and it is unclear how the variation of x_i's affects the estimation of the residual variance. Moreover, the difference-based operation increases the complexity of the model variance, and it is unclear how to generalize it to more complicated nonparametric and semiparametric models.

3.2.2 Local average method

For the sake of simplicity, assume that x_i's are ordered in an ascending order $x_1 \le x_2 \le \ldots \le x_n$, denote $x_{ij} = x_{I*(i-1)+j}$, and then split x_i into $k = n/I$ groups with I observations in each group:

$$\underbrace{x_{11}, x_{12}, ..., x_{1I}}_{1}, \underbrace{x_{21}, x_{21}, ..., x_{2I}}_{2}, \cdots\cdots, \underbrace{x_{k1}, x_{k2}, ..., x_{kI}}_{k}.$$

The number of samples I in each group is a fixed constant and independent of the sample size, and the number of group k is proportionate to the sample size. When the interval is small enough, those x_i's falling in the same interval are expected to be close, so do the nonparametric function values $f(x_i)$'s. Therefore, for the ith group, we take the local average of y_{ij}'s as the estimated function values of $f(x_{ij})$'s:

$$\hat{f}(x_{i1}) = \hat{f}(x_{i2}) = \cdots = \hat{f}(x_{iI}) = \frac{1}{I}\sum_{j=1}^{I} y_{ij} \triangleq y_{i\cdot}$$

and then propose a local average variance estimator

$$\hat{\sigma}_L^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - \hat{f}(x_{ij}))^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - y_{i\cdot}))^2. \tag{3.2}$$

In fact, by the step function approximation, the nonparametric regression model 3.1 can be reparameterized and rewritten as a high-dimensional linear model:

$$y_{ij} = X_{ij}^T\boldsymbol{\alpha} + \varepsilon_{ij}^*, i = 1, ..., k, j = 1, ..., I, \tag{3.3}$$

where $y_{ij} = y_{I(i-1)+j}$, X_{ij} is a k-vector with zero elements except one for the ith element, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)^T$ is a k-vector of unknown parameters, and $\varepsilon_{ij}^* = \varepsilon_{ij} + f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij})$. The ordinary least squares method can then be used to estimate the model, residuals, and the residual variance, where k can be regarded as the degrees of freedom of RSS.

Remark: Model (3.3) is quite similar to an ANOVA model in form, and one may think to apply the variance estimation approach to obtain an efficient estimator of the variance. However, model (3.3) is different from an ANOVA model in two significant ways. First, the parameters $\boldsymbol{\alpha}$ in model (3.3) are treated as fixed effects since they depend on the unknown function $f(\cdot)$, and it would be inappropriate to regard them as random effects following some covariance matrix structure. Second, the error in model (3.3) combines both a real mean zero random error and an approximation error, and the effect of such an approximated error on the final variance estimator is unclear.

3.2.3 Theoretical properties

We need the following two regularity conditions to investigate the theoretical properties of the proposed local average variance estimator.

(A) The residuals ϵ_i's are i.i.d. with $E[\epsilon_i] = 0, E[\epsilon_i^2] = \sigma^2$, and $E[\epsilon_i^4] = \mu_4 < \infty$.

(B) The function $f(\cdot)$ and its corresponding first and second derivatives are all bounded.

Theorem 3.1. *Under Conditions (A) and (B), the local average variance estimator $\hat{\sigma}_L^2$ in (3.2) for the nonparametric model (3.1) is asymptotically normal,*

$$\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) \xrightarrow{D} N(0, \mu_4 - \frac{I-3}{I-1}\sigma^4).$$

Remark: The reparameterized linear model is high-dimensional because the number of parameters $\boldsymbol{\alpha}$ increases in proportion to the sample size, and thus the estimation of $\boldsymbol{\alpha}$ or $f(x_{ij})$ cannot be consistent. However, our interest is the estimation of the residual variance, and Theorem 3.1 shows that the proposed local average estimator $\hat{\sigma}_L^2$ in (3.2) is a consistent estimation of σ^2. This is the so-called partial consistency phenomenon (Neyman and Scott, 1948; Fan, Huang and Peng, 2005). Here, σ^2 is called the structural parameter of the model and can be estimated consistently, while $\boldsymbol{\alpha}$ is called the incidental parameter and cannot get consistent estimator.

3.3 Extensions of the estimator

3.3.1 Partially linear models

Consider a partially linear regression model,

$$y_i = Z_i^T \boldsymbol{\beta} + f(x_i) + \epsilon_i, i = 1, ..., n, \tag{3.4}$$

where $Z_i = (z_{i1}, ..., z_{ip})^T$ is a p-vector of covariates, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a p-vector of parameters, $f(\cdot)$ is an unknown smooth function, and the error terms ϵ_i's are i.i.d. with mean zero and constant variance σ^2. Denote $\mathbf{Y} = (y_1, ..., y_n)^T$ and $\mathbf{Z} = (Z_1, ..., Z_n)^T$. Klipple and Eubank (2007) proposed a difference-based variance estimator for (3.4),

$$\hat{\sigma}_K^2 = \frac{\mathbf{Y}^T \mathbf{D}^T (\mathbf{I} - \mathbf{P}) \mathbf{D} \mathbf{Y}}{\mathrm{tr}\{\mathbf{D}^T (\mathbf{I} - \mathbf{P}) \mathbf{D}\}},$$

where $\mathbf{D}$ is the so-called differencing matrix, and $\mathbf{P} = \mathbf{D}\mathbf{Z}(\mathbf{Z}^T \mathbf{D}^T \mathbf{D} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{D}^T$ is the projection matrix. This method is quite complicated because the difference weights have to be chosen carefully to balance the bias and the variance.

By the local average method, we assume x_i's are ordered in ascending order $x_1 < x_2 < \cdots < x_n$, and then split the data into $k = n/I$ groups by x_i's. Denote the jth data in the ith group as $(x_{ij}, Z_{ij}, y_{ij}, \epsilon_{ij}) = (x_t, Z_t, y_t, \epsilon_t)$, $t = (i-1) \times I + j$. For the ith group, we approximate the function values $f(x_{ij})$'s by their average $\frac{1}{I} \sum_{j=1}^{I} f(x_{ij})$. When I is small, the approximation error $f(x_{ij}) - \frac{1}{I} \sum_{j=1}^{I} f(x_{ij})$ is of the order $O(n^{-1})$, much smaller than the order of the difference of ϵ_{ij}'s. Similarly,

(3.4) can be reparameterized and rewritten as

$$y_{ij} = \alpha_i + Z_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}^*, i = 1, ..., k, j = 1, ..., I, \tag{3.5}$$

where $\alpha_i = \frac{1}{I}\sum_{j=1}^{I} f(x_{ij})$ and $\epsilon_{ij}^* = \epsilon_{ij} + f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij})$. Thus, by the ordinary least squares method, we have

$$\hat{\boldsymbol{\beta}} = \{\textstyle\sum_{i,j}(Z_{ij} - \overline{Z}_{i\cdot})^T(Z_{ij} - \overline{Z}_{i\cdot})\}^{-1}\{\sum_{i,j}(Z_{ij} - \overline{Z}_{i\cdot})^T(y_{ij} - \overline{y}_{i\cdot})\},$$
$$\hat{\alpha}_i = \tfrac{1}{I}\textstyle\sum_{j=1}^{I}\{y_{ij} - Z_{ij}^T\hat{\boldsymbol{\beta}}\}, \quad i = 1, ..., k.$$

Therefore, we propose a local average variance estimator

$$\hat{\sigma}_L^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - Z_{ij}^T\hat{\boldsymbol{\beta}} - \hat{\alpha}_i)^2. \tag{3.6}$$

Theorem 3.2. *Under Conditions (A) and (B), the local average variance estimator $\hat{\sigma}_L^2$ in (3.6) for the partial linear model (3.4) is asymptotically normal,*

$$\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) \xrightarrow{D} N(0, \mu_4 - \frac{I-3}{I-1}\sigma^4).$$

Remark: The asymptotic variance of the estimator $\hat{\sigma}_L^2$ in (3.6) for the partial linear model (3.4) is the same as that of $\hat{\sigma}_L^2$ in (3.2) for the nonparametric model (3.1). This means that the estimation of $\boldsymbol{\beta}$ has little effect on the variance estimation for the local average method. As shown by Fan, Huang and Peng (2005), for such a high-dimensional model (3.6), the structural parameter $\boldsymbol{\beta}$ can be estimated consistently and almost efficiently.

3.3.2 Varying coefficient models

Consider a varying coefficient model

$$y_i = \sum_{l=1}^{p} a_l(U_i)x_{il} + \epsilon_i = X_i^T\boldsymbol{a}(U_i) + \epsilon_i, i = 1, ..., n, \tag{3.7}$$

where U_i is the index variable, $X_i = (x_{i1}, ..., x_{ip})^T$ is a p-vector of covariates, and $\boldsymbol{a}(\cdot) = (a_1(\cdot), ..., a_p(\cdot))^T$ is a p-vector of unknown varying coefficient functions. By the local polynomial fitting, Zhang and Lee (2000) proposed a variance estimator

$$\hat{\sigma}_{\text{poly}}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_{\text{poly}}^2(U_i).$$

For each given point u_0,

$$\hat{\sigma}^2_{\text{poly}}(u_0) = \frac{\mathbf{Y}^T\{\mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\}\mathbf{Y}}{\text{tr}\{\mathbf{W} - (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\}},$$

where $\mathbf{W} = \text{diag}(K_h(U_1 - u_0), ..., K_h(U_n - u_0))$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is the bandwidth, and $\mathbf{X} = (X_1, ..., X_n)^T \otimes (1, (U_1 - u_0), ..., (U_1 - u_0)^q)$, with $\otimes$ denoting the Kronecker product. As discussed above, it is not an easy task to select an appropriate bandwidth simultaneously for all varying coefficient functions. Moreover, an optimal bandwidth selection can be quite arduous and time-consuming, although the primary interest is only to estimate the residual variance.

For our proposed local average method, without loss of generality, we assume the index variable U_i's are ordered in ascending order $U_1 < U_2 < \cdots < U_n$ and then split the data into $k = n/I$ groups by u_i's. Denote the jth data in the ith group as $(U_{ij}, X_{ij}, y_{ij}, \epsilon_{ij}) = (U_t, X_t, y_t, \epsilon_t)$, $t = (i-1) * I + j$. If, for the ith group, we treat the coefficient functions as piecewise constant $\boldsymbol{a}(U_{i1}) = \boldsymbol{a}(U_{i2}) = \cdots = \boldsymbol{a}(U_{iI}) = \boldsymbol{a}_i$, then (3.7) can be rewritten as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{a}_i + \boldsymbol{\epsilon}^*_i \quad \text{or} \quad \mathbf{Y} = \mathbb{X}\boldsymbol{a} + \boldsymbol{\epsilon}^*,$$

where $\mathbf{Y}_i = (y_{i1}, ..., y_{iI})^T$ is a I-vector of responses, $\mathbf{X}_i = (X_{i1}, ..., X_{iI})^T$ is an $I \times p$ matrix of covariates, $\boldsymbol{a}_i = (a_{i1}, ..., a_{ip})^T$ is a p-vector of unknown parameters, $\boldsymbol{\epsilon}^*_i = (\epsilon^*_{i1}, ..., \epsilon^*_{iI})^T$ is an I-vector of random errors, and $\mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_k)^T$, $\mathbb{X} = \text{diag}(\mathbf{X}_1, ..., \mathbf{X}_k)$, $\boldsymbol{a} = (\boldsymbol{a}_1, ..., \boldsymbol{a}_k)^T$, $\boldsymbol{\epsilon}^* = (\boldsymbol{\epsilon}^*_1, \boldsymbol{\epsilon}^*_2, ..., \boldsymbol{\epsilon}^*_k)^T$. Thus, by the ordinary least squares method, we propose a local average variance estimator

$$\hat{\sigma}^2_L = \frac{\mathbf{Y}^T\mathbf{P}\mathbf{Y}}{\text{tr}(\mathbf{P})}, \tag{3.8}$$

where $\mathbf{P} = \mathbf{I} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

We need the following regularity conditions to investigate the theoretical properties of the proposed local average variance estimator.

(A') The residuals ϵ_i's are normally distributed with mean zero and variance σ^2.

(B') The coefficient functions $a_l(\cdot), l = 1, ..., p$, and their corresponding first and second derivatives are all bounded.

(C') All predictors are bounded, $\|X\| < \infty$.

Theorem 3.3. *Under Conditions (A') - (C'), the local average variance estimator (3.8) for the varying coefficient model (3.7) is asymptotically normal,*

$$\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) \xrightarrow{D} N(0, \frac{2I\sigma^4}{I-p}).$$

Remark: The normality assumption (A') is assumed only to facilitate the proof. In general, the asymptotic variance is $\frac{I(\mu_4-3\sigma^4)}{k(I-p)^2}\mathbf{p}^T\mathbf{p}+\frac{2I\sigma^4}{I-p}$, where $\mathbf{p}$ is the diagonal vector of $\mathbf{P}$. The first term disappears under the normality assumption because $\mu_4-3\sigma^4=0$. Without the normality assumption, this asymptotic variance is still bounded. As the projection matrix $\mathbf{P}$ is idempotent and symmetric, we have

$$\frac{(I-p)^2k}{I} = \frac{k^2(I-p)^2}{n} = \frac{tr(\mathbf{P})^2}{n} \leq \mathbf{p}^T\mathbf{p} \leq tr(\mathbf{P}^T\mathbf{P}) = tr(\mathbf{P}) = k(I-p).$$

Therefore, the asymptotic variance is bounded by $\frac{\mu_4-3\sigma^4}{n} + \frac{2I\sigma^4}{n(I-p)}$ and $\frac{1}{n}\frac{I}{I-p}(\mu_4-\sigma^4)$.
Remark: For the nonparametric model, the partially linear model and the varying coefficient model, our proposed local average variance estimators all achieve the asymptotic optimal rate for the mean squared errors (Dette, Munk, and Wagner, 1998), i.e., $\text{MSE}(\hat{\sigma}^2) = n^{-1}\text{var}(\epsilon^2) + o(n^{-1})$. The difference-based estimator proposed by Tong and Wang (2005) also achieves the asymptotic optimal rate for the mean squared errors but only under an equally spaced design setting. Moreover, our method is straightforward with light computational cost and can be easily extended to more complicated settings.

3.3.3 Refined local average variance estimator

Theorems 3.1-3.3 show that the asymptotic variance of our proposed local average estimators decreases as I increases. In next section, numerical studies show that our method is robust to the selection of interval length I. Even $I = 2$ can lead to a consistent variance estimator, although $I > p$ for (3.8) are required to make the method valid. Selecting an optimal I may, in practice, improve the performance of our proposed method but require additional computational cost. Here, we propose a refined local average variance estimator by aggregating estimators over different I's, which is expected to be more stable.

First, consider the local average variance estimator $\hat{\sigma}_L^2$ in (3.2) for the nonparametric model 3.1. By some calculations, we can show that

$$\mathrm{E}[\hat{\sigma}_L^2(I)] = \beta_1 + \beta_2 \frac{I(I+1)}{n^2} + o(\frac{1}{n^2}), I = 2, ..., m,$$

where $m = o(n)$ is a predefined constant, $\beta_1 = \sigma^2$, and $\beta_2 = J = \frac{1}{12}\int f'(x)^2 dx$. We naturally consider a linear model

$$\hat{\sigma}_L^2(I) = \beta_1 + \beta_2 \frac{I(I+1)}{n^2} + e_I,\ I = 2, ..., m.$$

Under the normality assumption for ϵ and by Theorem 3.1, it is easy to show that $\mathrm{var}(e_I) \approx \frac{I}{I-1}\frac{2\sigma^4}{n}$. Hence, we instead consider a linear model

$$s_I = \beta_1 t_{I1} + \beta_2 t_{I2} + e_I^*,\ I = 2, ..., m,$$

where $s_I = w_I \hat{\sigma}^2(I), t_{I1} = w_I, t_{I2} = w_I \frac{I(I+1)}{n^2}$, and $w_I = \sqrt{n(I-1)/I}$. The variance of e_I^* is now approximately a constant $2\sigma^4$. By the ordinary least squares, we propose a refined local average variance estimator

$$\hat{\sigma}_*^2 = \hat{\beta}_1 = \frac{(\sum s_I t_{I1})(\sum t_{I2}^2) - (\sum t_{I1} t_{I2})(\sum\ t_{I2} s_I)}{(\sum t_{I1}^2)(\sum t_{I2}^2) - (\sum t_{I1} t_{I2})^2}.$$

In this way, we use the regression technique to reduce the bias and variance of multiple local average variance estimators, which is expected to improve the stability of this estimator.

Similar techniques can be extended to the partially linear model and the varying coefficient model. We have, for the partially linear model,

$$\mathrm{E}[\hat{\sigma}^2(I)] = (1 + \frac{I}{n(I-1)})\sigma^2 + \frac{I(I+1)}{n^2} J + o(\frac{1}{n}),$$

where J is a constant depending on the nonparametric function $f(\cdot)$, and for the varying coefficient model,

$$\mathrm{E}[\hat{\sigma}^2(I)] = \sigma^2 + C\frac{(I-1)(I+1)}{n^2} + o(1),$$

where C is a constant depending on the varying coefficient functions $\boldsymbol{a}(\cdot)$ and the covariance matrix Σ of X.

3.3.4 More extensions

The basic idea of the proposed local average error variance estimation method is trying to use the local average to reduce the bias of the nonparametric function estimators as much as possible, even though the variance of the nonparametric function estimators would become larger and they could not be consistent and efficient estimators of the nonparametric function. However the error variance is a global model parameter. The inflated variance of the nonparametric function estimators can be integrated by the sum of quantitative values of all observations, and the effect of slight bias of the nonparametric function estimators could be ignored. Hence it could be expected that the global parameter estimators, such as the estimator of residual variance, would still be nearly efficient.

Our proposed local average method for the estimation of the error variance is trying to use a simple way to illustrate its flexibility and efficiency when estimating the global parameter in the nonparametric and the semiparametric regression model, from both the theoretical and practical insights. It is obvious that there are many ways to improve our proposed local average method.

3.3.4.1 Moving Average Method

Though the local average estimator for the nonparametric function remarkably reduces the bias of the nonparametric function estimation, when I, the number of observations in the every group is somehow large, the biases of the estimators for the samples near the boundary of each group are much larger than those of the estimators for the samples close to the center of every group. To overcome this inefficiency, inspired by the idea of the moving average and K-mean, for each sample, we replace its nonparametric function estimator by the mean of its nearest I observations. For convenient illustration, we only consider the simple nonparametric regression model 3.1. Define the so called projection matrix $\mathbf{S}$ with a very small constant odd I as

$$\mathbf{S}(i,j) = \frac{1}{I}\mathbf{1}(|i-j| < (I+1)/2)$$

where $\mathbf{1}(\cdot)$ is the indicator function, and then the error variance is estimated as

$$\hat{\sigma}^2_M(I) = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{Y}}{\mathrm{tr}\{(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\}}.$$

3.3.4.2 Kernel Based Method

As most of the difference-based approaches, our local average method and the moving average method require near equally spaced sampling for the observations. If not, the bias of the nonparametric function estimators cannot be considerably reduced. Then the efficiency of the variance estimator will be seriously affected. To minimize the bias of the nonparametric function estimation when the observations are not equal-spaced sampled, we employ the idea of kernel regression. Define the bandwidth as

$$h = \frac{I}{2n}(\max(x_i) - \min(x_i))$$

where I is a small constant and does not depend on the sample size n. Hence the length of every window is in an order of $O(1/n)$. The average number of observations falling in each window is near I. Similar as above, for the simple nonparametric regression model 3.1, define the projection matrix $\mathbf{S}$ as

$$\mathbf{S}(i,j) = \mathbf{1}\left(\frac{|x_i - x_j|}{h} < 1\right) \Bigg/ \sum_{j=1}^{n} \mathbf{1}\left(\frac{|x_i - x_j|}{h} < 1\right)$$

where $\mathbf{1}$ is the indicator function. Then the variance of the error term is estimated as

$$\hat{\sigma}_K^2(I) = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{Y}}{\mathrm{tr}\{(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\}}.$$

It is obvious that the indicator function can be replaced by any kernel function. For convenience, we just use the indicator function as the kernel function. Compared to the local average estimator, the number of observations to calculate the average in each group is now varying, not a constant. It depend on the sampling distribution of the observations and the bandwidth h.

No matter in the moving average method or in the kernel based method, the number of the samples for estimating the "average" is a very small constant and those observations are quite close to each others, even though the number could be variant with different windows. Hence the bias of the nonparametric function estimation based on the moving average or the kernel based method should be very small, even though its variance could be somehow large. It is similar as the original local average method. Furthermore, for a series of variance estimators $\sigma^2(I)$ with different I, similar as the refined local average variance estimator, we could also construct a regression model and use the weighted least squares to get a refined moving average estimator or

kernel based estimator for the error variance. However, these two extensions are not easy to adapt to the nonparametric regression model or semi-parametric regression model, and hence cannot be directly applied to estimate the error variance for more complex nonparametric regression models or semi-parametric regression models. It would need further investigations.

3.4 Applications of local average variance estimation

In this section, we focus on the simple nonparametric regression model 3.1 and illustrate some potential applications of the proposed local average variance estimation method. Similar procedures can be implemented for more complicated inference problems and for more complicated nonparametric and semiparametric models.

3.4.1 Confidence interval of variance estimation

Theorem 3.1 shows that the asymptotic variance of the proposed local average variance estimator is $\mu_4 - \frac{I-3}{I-1}\sigma^4$, which depends on the unknown parameter σ^2 and the fourth moment of the error distribution. Under the normality assumption for the error, it can be simplified as $\frac{2I}{I-1}\sigma^4$ because $\mu_4 = 3\sigma^2$. By the idea of the variance-stabilizing transformation (van der Vaart, 1998), we can construct the confidence interval for σ^2 based on the proposed local average variance estimator.

First, if the error follows a normal distribution, then by some calculations, the variance-stable transformation is

$$\phi(\sigma^2) = \sqrt{\frac{I-1}{2I}}\log(\sigma^2).$$

By Theorem 3.1, we then have

$$\sqrt{n}(\phi(\hat{\sigma}_L^2) - \phi(\sigma^2)) \xrightarrow{d} N\left(0, \phi'(\sigma^2)^2 \cdot \frac{2I}{I-1}\sigma^4\right) \xrightarrow{d} N(0,1).$$

This yields an asymptotic $1-\alpha$ level confidence interval for the variance σ^2,

$$\left(\exp\left(\log\hat{\sigma}_L^2 - \sqrt{\frac{2I}{I-1}}\frac{1}{\sqrt{n}}z_{\alpha/2}\right), \exp\left(\log\hat{\sigma}_L^2 + \sqrt{\frac{2I}{I-1}}\frac{1}{\sqrt{n}}z_{\alpha/2}\right)\right), \qquad (3.9)$$

where z_α is the $1-\alpha$ quantile of the standard normal distribution.

If the normality assumption is not satisfied by the error distribution, we can still use the variance-stable transformation

$$\phi(\sigma^2) = \sqrt{\frac{I-1}{I-3}}\arcsin\left(\sqrt{\frac{I-3}{I-1}}\frac{\sigma^2}{\sqrt{\mu_4}}\right)$$

to stabilize the variance of the estimator. By using the proposed local average technique and some straightforward calculations, we can derive a consistent estimator of μ_4,

$$\hat{\mu}_4 = \frac{I^4}{((I-1)^3+1)(I-1)}\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}-y_{i\cdot})^4 - \frac{6((I-1)^3+(I-1)(I-2)/2)}{I^4}\hat{\sigma}_L^4\right),$$

where y_{ij} and $y_{i\cdot}$ are defined in Section 2. Thus, an asymptotic $1-\alpha$ level confidence interval for the variance σ^2 is given by

$$\left(\sqrt{\hat{\mu}_4}\sqrt{\frac{I-1}{I-3}}\sin\left(\sqrt{\frac{I-3}{I-1}}\left(\sqrt{\frac{I-1}{I-3}}\arcsin\left(\sqrt{\frac{I-3}{I-1}}\frac{\hat{\sigma}_L^2}{\sqrt{\hat{\mu}_4}}\right)-\frac{1}{\sqrt{n}}z_{\alpha/2}\right)\right),\right.$$
$$\left.\sqrt{\hat{\mu}_4}\sqrt{\frac{I-1}{I-3}}\sin\left(\sqrt{\frac{I-3}{I-1}}\left(\sqrt{\frac{I-1}{I-3}}\arcsin\left(\sqrt{\frac{I-3}{I-1}}\frac{\hat{\sigma}_L^2}{\sqrt{\hat{\mu}_4}}\right)+\frac{1}{\sqrt{n}}z_{\alpha/2}\right)\right)\right) \quad (3.10)$$

3.4.2 Nonparametric hypothesis testing

The proposed variance estimator and the idea of local average can also be used for the nonparametric testing. For example, for the nonparametric regression model 3.1, we consider the following hypothesis problem:

$$H_0 : f(x) = a + bx \quad \text{vs} \quad H_1 : f(x) \neq a + bx \quad \text{for any } a \text{ and } b.$$

By the idea of the generalized likelihood ratio of Fan, Zhang and Zhang (2001), we can construct the following test statistic,

$$T = \frac{n(I-1)}{I}\frac{RSS_0 - RSS_1}{RSS_1},$$

where RSS_0 is the sum of the squares of the residuals estimated by the least squares method under the null hypothesis H_0, and $RSS_1 = (n-k)\hat{\sigma}_L^2$.

As shown by Fan, Zhang and Zhang (2001), the null distribution of such test statistic T would be expected to hold the Wilks' phenomenon, that is, it is model

free and follows asymptotically χ^2 distribution where the degree of freedom may tend to infinity with the sample size. Hence, it is reasonable to suggest the following bootstrap procedure to approximate the null distribution of the test statistic T under the null hypothesis.

Step 1. Based on the samples $(X_i, Y_i), i = 1, ..., n$, first calculate the least squares estimator $\hat{a}$ and $\hat{b}$ under the null hypothesis and $\hat{\sigma}_L^2$ by the local average method, then compute the test statistic T.

Step 2. Construct a bootstrap example $(X_i, Y_i^*), i = 1, ..., n$ by

$$Y_i^* = \hat{a} + \hat{b}X_i + \varepsilon_i^*,$$

where ε_i^* is sampled from the normal distribution with mean zero and variance $\hat{\sigma}_L^2$. Then, based on the bootstrap sample, calculate the test statistic T^*.

Step 3. Repeat Step 2 for B times to obtain the bootstrapped test statistics and sort them in increasing order: $T_1^* \leq \cdots \leq T_B^*$.

Step 4. If $T \leq T_{B(1-\alpha)}^*$, we accept the null hypothesis; otherwise, we reject the null hypothesis.

There is no need to estimate the nonparametric regression function for the proposed testing procedure, which avoids the use of a complex algorithm to select optimal tuning parameter, thus reducing the computational cost and increasing the stability of the testing results. Our numerical simulation results in next section show the proposed testing procedure performs reasonably well. Further theoretical investigations are needed, especially for more complicated testing problems.

3.4.3 Variance function estimation

Consider the following nonparametric regression model,

$$y_i = f(x_i) + \sigma(x_i)\varepsilon_i,$$

where $\sigma(x_i)$ is a smoothing variance function, and ε_i is a random variable with mean zero and variance one.

By the idea of the local average, define

$$\varepsilon_{Iij}^{*2} \hat{=} \frac{I}{I-1}(y_{ij} - y_{i\cdot})^2, i = 1, ..., k, j = 1, ..., I,$$

and then by some calculations, we can show that

$$\mathrm{E}\varepsilon_{Iij}^{*2} = \sigma^2(x_{ij}) + O(\frac{1}{n}).$$

Hence, given $(x_{ij}, \varepsilon_{Iij}^{*2}), i = 1, ..., k, j = 1, ..., I$, a local linear regression estimator $\hat{\sigma}_I^2(x_{ij})$ for the variance function $\sigma^2(\cdot)$ can be obtained. Similar to the idea of the refined local average variance estimator, we can derive a refined estimator of the variance function. Consider m different $I_1, ..., I_m$, define a weighted refined variance estimator

$$\varepsilon_t^{*2} = \sum_{l=1}^{m} I_l \varepsilon_{I_l i_l j_l}^{*2} \Big/ \sum_{l=1}^{m} I_l,$$

where $t = (i_l - 1)I_l + j_l$, $l = 1, ..., m$, $j_l = 1, ..., I_l$. Then, given $(x_t, \varepsilon_t^{*2}), t = 1, ..., n$, a local linear regression estimator $\hat{\sigma}^2(x_t)$ can be obtained. Such an estimator does not depend on the value of I_l and is expected to be more stable.

3.5 Numerical studies

3.5.1 Simulations for variance estimation

Example 1. Consider a nonparametric model

$$y_i = 5\sin(\omega\pi x_i) + \epsilon_i, i = 1, ..., n,$$

where $x_i = i/n$, and the random errors ϵ_i's follow an i.i.d. normal distribution with mean zero and variance σ^2. Let $n = (100, 500)$, $\omega = (1, 2, 4)$, and $\sigma = (0.5, 1.5, 4)$.

In Table 3.1, following the model setting above, the simulation results of the local average variance estimator $\hat{\sigma}_L^2$ with $I = 2$ and 5, the refined local average variance estimator $\hat{\sigma}_*^2$, which is calculated based on the estimators of $\hat{\sigma}_L^2$ with $I = 2, 3, ..., 11$, the moving averaging based estimator $\hat{\sigma}_M^2$ with $I = 5$, and the kernel based estimator $\hat{\sigma}_L$ with $I = 5$ are listed. If x_i is equally spaced sampled, the moving averaging based variance estimator $\hat{\sigma}_M^2$ and the kernel based variance estimator $\hat{\sigma}_K^2$ are equivalent. Hence, to make the comparison more reasonable, we let $x_i, i = 1, ..., n$ sampled from

the $[0,1]$ uniform distribution for the simulation of the kernel based estimator $\hat{\sigma}_K^2$ with $I = 5$ in Table 3.1.

For each combination of (n, ω, σ), the simulations are repeated 10000 times and the relative mean squared errors, RMSE $= n\text{MSE}/(2\sigma^4) = \frac{n}{2\sigma^4}(\text{bias}^2 + \text{var})$, is calculated for each variance estimator. The closer the RMSE is to 1, the better is the estimator. The proposed variance estimators are compared with $\hat{\sigma}_{\text{R}}^2, \hat{\sigma}_{\text{GSJ}}^2$ and $\hat{\sigma}_{\text{T}}^2(m_s)$. For $\hat{\sigma}_{\text{T}}^2(m_s)$, the data in Table 1 in Tong and Wang (2005) is referenced with $m_s = n^{\frac{1}{2}}$.

Table 3.1: RMSE for Example 1

n	ω	σ	$\hat{\sigma}_{\text{R}}^2$	$\hat{\sigma}_{\text{GSJ}}^2$	$\hat{\sigma}_{\text{T}}^2(m_s)$	$\hat{\sigma}_L^2(2)$	$\hat{\sigma}_L^2(5)$	$\hat{\sigma}_*^2$	$\hat{\sigma}_M^2(5)$	$\hat{\sigma}_K^2(5)$
100	1	0.5	1.56	1.98	1.19	2.11	2.41	1.46	1.49	1.38
		1.5	1.51	2.03	1.12	2.02	1.27	1.33	1.46	1.35
		4	1.53	2.00	1.14	1.98	1.24	1.28	1.49	1.36
	2	0.5	2.00	2.01	1.81	2.89	14.71	1.79	1.49	1.72
		1.5	1.52	2.02	1.14	2.09	1.56	1.37	1.45	1.36
		4	1.45	1.95	1.15	2.07	1.27	1.32	1.45	1.35
	4	0.5	9.06	1.96	26.77	11.19	195.97	3.04	1.69	6.77
		1.5	1.54	2.01	1.46	2.34	4.14	1.49	1.45	1.39
		4	1.58	1.98	1.15	2.01	1.39	1.38	1.43	1.35
500	1	0.5	1.48	1.94	1.06	1.98	1.26	1.30	1.48	1.34
		1.5	1.48	1.94	1.05	2.01	1.28	1.31	1.46	1.35
		4	1.52	1.96	1.05	1.99	1.28	1.31	1.43	1.33
	2	0.5	1.49	1.93	1.06	2.04	1.38	1.34	1.46	1.37
		1.5	1.53	1.98	1.05	1.98	1.26	1.30	1.41	1.35
		4	1.48	1.94	1.06	1.96	1.24	1.30	1.42	1.37
	4	0.5	1.58	1.97	1.51	2.09	3.00	1.40	1.46	1.41
		1.5	1.50	1.95	1.07	2.03	1.27	1.33	1.44	1.34
		4	1.51	1.96	1.04	1.97	1.25	1.28	1.43	1.34

Table 3.1 depicts the RMSE of all estimators. We can see that in general, the order exists as $\text{RMSE}_{\hat{\sigma}_{\text{T}}^2(m_s)} < \text{RMSE}_{\hat{\sigma}_L^2(5)} < \text{RMSE}_{\hat{\sigma}_{\text{R}}^2} < \text{RMSE}_{\hat{\sigma}_L^2(2)} \doteq \text{RMSE}_{\hat{\sigma}_{\text{GSJ}}^2}$. As the sample size n tends to infinity, the RMSE of $\hat{\sigma}^2(I)$ tends to $\frac{I}{I-1}$ as shown in Theorem 3.1. Moreover, as expected, the performance of our local average estimator depends on the smoothness of the nonparametric function $f(\cdot)$, the sample size, and the signal-to-noise ratio. When f is rough and σ is small, for instance, $(n, \omega, \sigma) = (100, 4, 0.5)$,

the RMSE of $\hat{\sigma}_L^2(5)$ is quite large because the bias as shown above is about $\frac{I(I+1)}{n^2}J$. When the sample size n is much larger than the group size I, the bias is negligible, and the RMSE then converges to $\frac{n}{2\sigma^4}$var, which is $1+\frac{1}{I-1}$ for this example. For the refined local average variance estimator $\hat{\sigma}_*^2$, in general, it not only performs better or at least as well as $\hat{\sigma}_L^2(2)$ and $\hat{\sigma}_L^2(5)$, but is also much more robust. The results of $\hat{\sigma}_M^2(5)$ are very stable, and it is better than $\hat{\sigma}_{GSJ}^2$. The performance of $\hat{\sigma}_K(5)$ is similar as the refined local average variance estimate $\hat{\sigma}_*^2$ though a little worse, but it only depends one particular I and does not need use further least squares to improve the estimation. Furthermore, when the variation of samples are very large, such kernel based variance estimator would be more stable and adaptive.

To assess the performance of the proposed variance estimators for data with small sample size, we consider the above example but with $n = 15$. We compared our proposed local average methods and moving averaging based method with $\hat{\sigma}_{GSL}^2$, $\hat{\sigma}_T^2(m_s)$ and $\hat{\sigma}_{ols}^2(d_k, m_1)$ which is proposed by Park, Kim and Lee (2012) specially for small sample nonparametric regression. For convenience, we follow the setting of Park, Kim and Lee (2012) for the above Example 1. The refined local average estimator is calculated by $I = 2$ to 5. The numerical simulation results about the mean square error (MSE) of the estimates are shown in Table 3.2. When the signal-noise ratio is some large, or $\sigma = 0.01$, $\hat{\sigma}_{GSL}^2$ and $\hat{\sigma}_{ols}^2(d_k, m_1)$ would be much better than other methods. When the variation of noise is increased or the signal-noise ration is decreased, the performances of our methods, especially the moving averaging based method with $I = 2$ are not worse than other methods. Hence those methods have their own advantages for the error variance estimation when the sample size is very small.

Example 2. Consider a bivariate nonparametric model

$$y_i = f(x_i, u_i) + \epsilon_i, \quad i = 1, ..., n,$$

where $x_i = i/n$, u_i is a categorical variable with two levels, the bivariate function $f(x_i, u_i) = 5\sin(\pi x_i)$ when $u_i = 0$ and $f(x_i, u_i) = 5\sin(2\pi x_i)$ when $u_i = 1$, and the random errors ϵ_i's follow an i.i.d. normal distribution with mean zero and variance σ^2. Our local average method continues to work well. The data can be naturally split into two subgroups according to the categorical variable. Within each subgroup, the

Table 3.2: MSE for Example 1 with Small Sample $n = 15$

ω	σ	$\hat{\sigma}^2_{\mathrm{GSJ}}$	$\hat{\sigma}^2_{\mathrm{T}}(m_s)$	$\hat{\sigma}^2_{ols}(d_k, m_1)$	$\hat{\sigma}^2_L(2)$	$\hat{\sigma}^2_L(5)$	$\hat{\sigma}^2_*$	$\hat{\sigma}^2_M(2)$	$\hat{\sigma}^2_M(5)$
1	0.01	2.09E-05	3.28E-02	4.91E-03	9.83E-02	2.67E+00	1.49E-03	4.36E-02	7.94E-02
	0.5	2.79E-02	6.22E-02	5.06E-02	1.57E-01	2.80E+00	9.06E-02	2.34E-02	1.01E-01
	2	8.13E+00	6.11E+00	5.67E+00	5.40E+00	7.63E+00	5.46E+00	2.78E+00	3.41E+00
	4	1.66E+02	1.05E+02	1.39E+02	7.93E+01	5.55E+01	7.61E+01	4.97E+01	5.05E+01
2	0.01	5.05E-03	5.16E-01	1.05E-01	1.53E+00	3.88E+01	1.45E-03	6.81E-01	1.61E+00
	0.5	3.13E-02	6.12E-01	3.40E-01	1.74E+00	3.93E+00	2.25E-01	5.60E-01	1.67E+00
	2	8.13E+00	6.67E+00	7.03E+00	8.76E+00	4.90E+01	7.56E+00	1.89E+00	5.51E+00
	4	1.66E+02	1.05E+02	9.57E+01	8.36E+01	1.14E+02	8.09E+01	4.47E+01	5.51E+01
4	0.01	1.01E+00	2.52E+01	5.51E+00	2.22E+01	1.44E+02	2.52E+01	9.84E+00	4.56E+01
	0.5	1.04E+00	2.56E+01	5.73E+00	2.23E+01	1.45E+02	2.60E+01	9.42E+00	4.61E+01
	2	9.62E+00	3.60E+01	1.48E+01	3.68E+01	1.63E+02	4.16E+01	5.99E+00	5.50E+01
	4	1.65E+02	1.27E+02	1.31E+02	1.39E+02	2.53E+02	1.43E+02	3.32E+01	1.23E+02

above bivariate nonparametric model reduces to a univariate nonparametric model and we totally get two different nonparametric models. The local average method can be applied to obtain a variance estimator for each subgroup, then a single variance estimator is calculated by taking a weighted average of these within-subgroup variance estimates. We consider $n = (200, 400, 800)$, $\sigma = (0.5, 1.5, 4)$, and for each combination of (n, σ), run the simulation for 10000 times. The results of local average estimators of $\hat{\sigma}_L^2(2)$,$\hat{\sigma}_L^2(5)$ and $\hat{\sigma}_L^2(10)$ are summarized in Table 3.3.

Table 3.3: Simulation Results for Example 2

n	σ	$\hat{\sigma}_L^2(2)$			$\hat{\sigma}_L^2(5)$			$\hat{\sigma}_L^2(10)$		
		mean	std	rmse	mean	std	rmse	mean	std	rmse
200	0.5	0.27	0.04	2.99	0.33	0.04	13.06	0.51	0.06	117.31
	1.5	2.27	0.32	2.04	2.33	0.26	1.51	2.52	0.27	2.91
	4	16.02	2.28	2.02	16.08	1.81	1.29	16.26	1.76	1.23
400	0.5	0.26	0.03	2.22	0.27	0.02	2.85	0.32	0.02	17.28
	1.5	2.26	0.23	2.02	2.27	0.18	1.31	2.32	0.18	1.41
	4	15.97	1.59	1.97	16.03	1.28	1.28	16.08	1.21	1.15
800	0.5	0.25	0.02	2.04	0.26	0.01	1.50	0.27	0.01	3.29
	1.5	2.25	0.16	2.05	2.25	0.13	1.29	2.27	0.12	1.18
	4	16.00	1.13	2.01	16.02	0.90	1.25	16.02	0.84	1.11

Example 3. Consider a partially linear model

$$y_i = Z_i^T \boldsymbol{\beta} + f(x_i) + \epsilon_i, i = 1, ..., n,$$

where $\boldsymbol{\beta} = (1, 3, 0, 0, 0)^T$, $x_i = i/n$, $f(x_i) = 5sin(\omega \pi x_i)$, and the random errors ϵ_i's follow an i.i.d. normal distribution with mean zero and variance σ^2. Z_i follows a multivariate normal distribution with mean zero and covariance matrix with 1 on the diagonal and 0.5 on the off-diagonal. Similar to Example 1, we consider $\omega = (1, 2, 4)$, $\sigma = (0.5, 1.5, 4)$, and $n = (100, 500)$, and for each combination of (ω, σ, n), run the simulation for 10000 times. We calculate $\hat{\sigma}_L^2(I)$ for $I = 2$ and 5, respectively, and then compare them to the difference-based estimator $\hat{\sigma}_{\mathrm{K}}^2$ proposed by Klipple and Eubank (2007) with $m = 2$ and the GSJS weights. It is worth mentioning that under this setting, the estimator $\hat{\sigma}_{\mathrm{K}}^2$ is the same as the estimator proposed by Wang, Brown, and Cai (2011). Table 3.4 depicts the RMSE of $\hat{\sigma}_L^2(2), \hat{\sigma}_L^2(5)$, $\hat{\sigma}_{\mathrm{K}}^2$, and $\hat{\sigma}_*^2$ with $m = 11$

(I is from 2 to 11). It shows that in general, $\text{RMSE}_{\hat{\sigma}^2_{L(2)}} \doteq \text{RMSE}_{\hat{\sigma}^2_K} > \text{RMSE}_{\hat{\sigma}^2_*} > \text{RMSE}_{\hat{\sigma}^2_{L(5)}}$, except for some unstable results for $\hat{\sigma}^2_L(5)$ with $\sigma = 0.5$.

Example 4. Consider a semiparametric additive model

$$y_i = Z_i^T\boldsymbol{\beta} + f_1(x_{i,1}) + f_2(x_{i,2}) + f_3(x_{i,3}) + \epsilon_i, \quad i = 1, ..., n,$$

where $\boldsymbol{\beta} = (1, 3, 0, 0, 0)^T$, $f_1(x) = -\sin(2x)$, $f_2(x) = x^2 - 25/12$, $f_3(x) = \exp(-x) - 2\sinh(5/2)/5$, and Z_i is the same as in Example 3. $X_i = (x_{i,1}, x_{i,2}, x_{i,3})$ is a 3-dimensional random vector, each marginal distribution is a uniform distribution on $[0, 1]$, and the correlation matrix is compound symmetric with ρ on the off-diagonal. Thus, $\mathbf{X}$ is ensured to be bounded. We consider $\rho = (0.25, 0.75)$, $n = (200, 400)$, and $\sigma = (0.5, 1.5, 4)$ and run the simulation for 10000 times.

Rather than take $X_i = (x_{i,1}, x_{i,2}, x_{i,3})$ as a whole and cut it on a 3-dimensional space, we group $x_{i,1}, x_{i,2}$, and $x_{i,3}$ separately. For each $p, p = 1, 2, 3$, we order $x_{i,p}$ in an ascending order and split $x_{i,p}$ into $k = n/I$ groups with I observations in each group. Denote $x_{ij,p} = x_{I*(i-1)+j,p}$, then for the ith group, we use the average to estimate the function values within this group, i.e. $\hat{f}_p(x_{ij,p}) = \frac{1}{I}\sum_{j=1}^{I} f_p(x_{ij,p}) \triangleq \alpha_{i,p}$. Thus, the semiparametric additive model can be written as

$$y_i = Z_i^T\boldsymbol{\beta} + \alpha_{i_1,1} + \alpha_{i_2,2} + \alpha_{i_3,3} + \epsilon_i, \quad i = 1, ..., n, \quad i_1, i_2, i_3 = 1, ..., k.$$

The residual variance can then be estimated by the ordinary least squares method via a high-dimensional linear regression model

$$\mathbf{Y} = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \alpha_{1,1}, ..., \alpha_{k,1}, \alpha_{1,2}, ..., \alpha_{k,2}, \alpha_{1,3}, ..., \alpha_{k,3})^T$, $\mathbf{D}$ is the design matrix with the first five columns equal to $\mathbf{Z}$, and the rest is a zero-one matrix indicating the presence of $\alpha_{i,p}$. Table 3.5 depicts the results of $\hat{\sigma}^2_L(4)$, $\hat{\sigma}^2_L(5)$, and $\hat{\sigma}^2_L(10)$. It shows that the estimation results are better as the sample size increases and are quite robust to the correlation coefficient ρ of the predictor $\mathbf{X}$.

Example 5. Consider a varying coefficient model

$$y_i = \sin(2\pi u_i)x_{i1} + 4u_i(1 - u_i)x_{i2} + \epsilon_i, i = 1, ..., n,$$

where u_i's follow a uniform distribution on $[0, 1]$, (x_{i1}, x_{i2})'s follow an i.i.d. bivariate normal distribution with means 0, variances 1, and correlation coefficient $1/\sqrt{2}$, and

Table 3.4: Simulation Results for Example 3

n	ω	σ	$\hat{\sigma}^2_L(2)$			$\hat{\sigma}^2_L(5)$			$\hat{\sigma}^2_K$			$\hat{\sigma}^2_*$		
			mean	std	rmse	mean	std	rmse	mean	std	rmse	mean	std	rmse
100	1	0.5	0.23	0.05	2.21	0.26	0.04	1.58	0.25	0.05	2.13	0.23	0.04	1.72
		1.5	2.03	0.42	2.24	2.13	0.35	1.32	2.26	0.47	2.15	2.05	0.35	1.60
		4	14.45	3.03	2.27	15.01	2.45	1.36	16.01	3.32	2.16	14.56	2.44	1.57
	2	0.5	0.25	0.05	2.20	0.35	0.05	10.18	0.25	0.05	2.14	0.22	0.04	2.14
		1.5	2.04	0.43	2.22	2.22	0.36	1.32	2.25	0.47	2.14	2.05	0.35	1.60
		4	14.45	3.04	2.28	15.17	2.49	1.35	16.03	3.31	2.14	14.59	2.45	1.56
	4	0.5	0.31	0.06	6.46	0.69	0.09	162.15	0.25	0.05	2.15	0.23	0.06	3.25
		1.5	2.12	0.45	2.18	2.56	0.41	2.64	2.25	0.47	2.16	2.04	0.36	1.74
		4	14.49	3.08	2.30	15.49	2.53	1.30	15.94	3.34	2.18	14.60	2.49	1.59
500	1	0.5	0.25	0.02	2.06	0.25	0.02	1.25	0.25	0.02	1.95	0.25	0.02	1.34
		1.5	2.21	0.20	2.06	2.22	0.16	1.28	2.25	0.20	1.93	2.21	0.16	1.37
		4	15.67	1.42	2.07	15.80	1.13	1.28	16.03	1.43	2.00	15.73	1.14	1.34
	2	0.5	0.25	0.02	2.03	0.25	0.02	1.27	0.25	0.02	1.96	0.25	0.02	1.42
		1.5	2.21	0.20	2.05	2.23	0.16	1.27	2.25	0.20	1.97	2.21	0.16	1.37
		4	15.67	1.41	2.06	15.82	1.13	1.28	16.00	1.43	2.01	15.72	1.14	1.35
	4	0.5	0.25	0.02	2.00	0.27	0.02	2.54	0.25	0.02	1.97	0.25	0.02	1.42
		1.5	2.21	0.20	2.07	2.24	0.16	1.25	2.25	0.20	2.02	2.21	0.16	1.37
		4	15.69	1.41	2.04	15.83	1.13	1.29	16.00	1.43	1.99	15.72	1.16	1.38

Table 3.5: Simulation Results for Example 4

ρ	n	σ	$\hat{\sigma}^2_L(4)$			$\hat{\sigma}^2_L(5)$			$\hat{\sigma}^2_L(10)$		
			mean	std	rmse	mean	std	rmse	mean	std	rmse
0.25	200	0.5	0.22	0.05	4.65	0.23	0.04	2.67	0.24	0.03	1.50
		1.5	2.02	0.43	4.64	2.11	0.35	2.83	2.17	0.26	1.52
		4	14.38	3.01	4.57	14.99	2.45	2.74	15.41	1.87	1.50
	400	0.5	0.24	0.03	4.25	0.24	0.03	2.73	0.25	0.02	1.46
		1.5	2.14	0.31	4.25	2.18	0.25	2.65	2.21	0.19	1.48
		4	15.22	2.21	4.29	15.48	1.73	2.54	15.71	1.35	1.50
0.75	200	0.5	0.23	0.05	4.55	0.23	0.04	2.73	0.24	0.03	1.50
		1.5	2.02	0.43	4.68	2.11	0.35	2.75	2.17	0.27	1.53
		4	14.43	3.04	4.56	15.01	2.44	2.70	15.42	1.89	1.53
	400	0.5	0.24	0.03	4.33	0.24	0.03	2.62	0.25	0.02	1.47
		1.5	2.14	0.31	4.30	2.18	0.25	2.63	2.21	0.19	1.45
		4	15.23	2.23	4.35	15.52	1.76	2.60	15.73	1.34	1.47

the errors ϵ_i's follow an i.i.d. normal distribution with mean zero and variance σ^2. We consider $n = (100, 500)$ and $\sigma = (0.5, 1.5, 4)$ and run the simulation for 10000 times. We calculate $\hat{\sigma}^2_L(I)$ for $I = 5, 10$, respectively, and $\hat{\sigma}^2_*$ with $m = 11$. We also calculate the variance estimator $\hat{\sigma}^2_{\text{poly}}$ proposed by Zhang and Lee (2000) with the tricube kernel function, $q = 3$ and $h = 0.3$ at $u_i = 0.5$.

Table 3.6 lists the sample average and the sample standard deviation of all estimators. It shows that the sample average of each estimator becomes closer to the true value and the standard deviation decreases as the sample size increases. The sample standard deviation of local average variance estimators is smaller than that of Zhang and Lee's variance estimator. Especially, the sample standard deviation of $\hat{\sigma}^2_L(5)$ is about half of that of $\hat{\sigma}^2_{\text{poly}}$.

Example 6. Examples 1-3 all assume the covariate x's are equally spaced, which may not be realistic in practice. Here, we focus on the case in which the covariate x's follow some distributions. We first consider

$$y_i = 5\sin(2\pi x_i) + \epsilon_i, i = 1, ..., n,$$

where the covariates x_i's follow an uniform distribution on $[0, 1]$. This example is similar to Example 1 with $\omega = 2$ but with non-equally spaced x's. We consider

Table 3.6: Simulation Results for Example 5

n	σ	$\hat{\sigma}^2(5)$			$\hat{\sigma}^2(10)$			$\hat{\sigma}^2_{\text{poly}}$			$\hat{\sigma}^2_*$		
		mean	std	rmse	mean	std	rmse	mean	std	rmse	mean	std	rmse
100.0	0.5	0.2565	0.0475	1.8377	0.2711	0.0437	1.8814	0.2490	0.0914	6.6897	0.2513	0.0451	1.6284
	1.5	2.2588	0.4106	1.6660	2.2728	0.3593	1.2799	2.2534	0.8173	6.5975	2.2523	0.4002	1.5819
	4.0	15.9687	2.9274	1.6739	16.0016	2.5175	1.2379	16.0827	5.8288	6.6371	15.9754	2.8279	1.5620
500.0	0.5	0.2502	0.0203	1.6411	0.2507	0.0177	1.2614	0.2497	0.0358	5.1169	0.2501	0.0198	1.5626
	1.5	2.2502	0.1845	1.6806	2.2503	0.1571	1.2190	2.2509	0.3197	5.0480	2.2522	0.1806	1.6112
	4.0	15.9971	1.2933	1.6334	15.9983	1.1344	1.2567	16.0043	2.2397	4.8985	15.9710	1.2873	1.6191

$n = (100, 200, 400, 800)$ and $\sigma = (0.5, 1.5, 4)$ and run the simulation for 10000 times. We calculate the local average variance estimator $\hat{\sigma}_L^2(I)$ for $I = 2$ and 5, respectively, and $\hat{\sigma}_*^2$ with $m = 11$. The results are summarized in Table 3.7, which shows that the performance of our estimators is still very good. The RMSEs in Table 3.7 are comparable to those in Table 3.1.

Table 3.7: Simulation results for uniformly distributed covariates

n	σ	$\hat{\sigma}_L^2(2)$			$\hat{\sigma}_L^2(5)$			$\hat{\sigma}_*^2$		
		mean	std	rmse	mean	std	rmse	mean	std	rmse
100	0.5	0.26	0.05	2.35	0.29	0.05	3.03	0.26	0.04	1.56
	1.5	2.27	0.45	2.01	2.29	0.36	1.29	2.26	0.37	1.34
	4	16.00	3.18	1.98	16.03	2.53	1.25	16.04	2.65	1.37
200	0.5	0.25	0.04	2.00	0.26	0.03	1.54	0.25	0.03	1.37
	1.5	2.25	0.32	2.00	2.26	0.25	1.27	2.25	0.26	1.33
	4	15.95	2.28	2.02	16.00	1.78	1.24	16.02	1.83	1.31
400	0.5	0.25	0.02	1.96	0.25	0.02	1.26	0.25	0.02	1.29
	1.5	2.25	0.22	2.00	2.25	0.18	1.24	2.25	0.18	1.29
	4	16.01	1.61	2.03	16.01	1.29	1.29	15.99	1.29	1.29
800	0.5	0.25	0.02	1.96	0.25	0.01	1.24	0.25	0.01	1.32
	1.5	2.25	0.16	2.06	2.25	0.13	1.27	2.25	0.13	1.36
	4	15.98	1.14	2.05	16.00	0.89	1.23	15.99	0.90	1.28

We also consider

$$y_i = 5\sin(\frac{2}{3}\pi x_i) + \epsilon_i, i = 1, ..., n,$$

where the covariates x's follow a standard normal distribution. This example is similar to Example 1 but with $\omega = 2/3$ and non-equally spaced x's. To ensure that our data are bounded, we discard the data with $|x_i| > 1.5$, leaving about 90% of the data. We consider $n = (100, 200, 400, 800)$, $\sigma = (0.5, 1.5, 4)$, and run the simulation for 10000 times. We again calculate the local average variance estimator $\hat{\sigma}_L^2(I)$ for $I = 2$ and 5, respectively, and $\hat{\sigma}_*^2$ with $m = 11$. The results are summarized in Table 3.8, which shows that the performance of our estimator is still satisfactory, although the convergency rate is a bit slow. The reason is that, for a normal distribution, most data are concentrated around the mean and fewer data exist at the boundary. Hence the bias at the boundary negatively affects the result when the sample size is small. Compared to those in Tables 3.1 and 3.7, the RMSEs in Table 3.8 are generally

larger. This is a consequence of the effective sample size n being smaller since we have removed some data. If we multiply the RMSE by 0.9, which is the utilization rate of the data, the RMSEs become the same as those in Tables 3.1 and 3.7.

Table 3.8: Simulation results for normally distributed covariates

n	σ	$\hat{\sigma}^2_L(2)$			$\hat{\sigma}^2_L(5)$			$\hat{\sigma}^2_*$		
		mean	std	rmse	mean	std	rmse	mean	std	rmse
100	0.5	0.31	0.07	7.36	0.46	0.09	40.13	0.33	0.07	9.17
	1.5	2.33	0.50	2.54	2.45	0.43	2.23	2.33	0.41	1.73
	4	16.07	3.46	2.33	16.24	2.81	1.55	16.07	2.83	1.56
200	0.5	0.27	0.04	3.18	0.31	0.04	8.01	0.27	0.04	2.51
	1.5	2.27	0.34	2.32	2.32	0.28	1.63	2.27	0.28	1.56
	4	16.04	2.44	2.32	16.07	1.97	1.51	16.02	1.97	1.52
400	0.5	0.26	0.03	2.48	0.27	0.02	2.47	0.25	0.02	1.65
	1.5	2.25	0.24	2.30	2.27	0.19	1.48	2.25	0.20	1.52
	4	16.03	1.71	2.29	15.99	1.36	1.44	16.00	1.41	1.55
800	0.5	0.25	0.02	2.40	0.25	0.02	1.64	0.25	0.02	1.55
	1.5	2.25	0.17	2.31	2.25	0.14	1.47	2.25	0.14	1.51
	4	16.02	1.22	2.33	16.00	0.95	1.41	16.00	1.00	1.55

3.5.2 Applications of variance estimation

Example 7. Consider the nonparametric regression model

$$y_i = x_i + 2\exp(-16x_i^2) + \sigma\varepsilon_i, i = 1, ..., n,$$

where $x_i \sim \text{Unif}[-2, 2]$, and $\varepsilon_i \sim N(0, 1)$ or $\sqrt{3/5}t_5$ independent of x_i. We simulate 1000 random samples of size $n = 200, 400$ with $\sigma = 0.25, 0.5$, and then construct the 95% confidence interval for σ^2 with $I = 4, 5$ by equations 3.9 and 3.10.

The coverage rates of the confidence interval constructed by 3.9 and 3.10 under different model settings are shown in Table 3.9. From the table, we can see that the confidence intervals constructed by 3.9 appear sensible to the normality assumption. When the error does not follow a normal distribution, the coverage rate tends to be much lower. The coverage rates of the confidence intervals constructed by 3.10 are relatively stable, no matter the error follows a normal distribution or a t-distribution.

Table 3.9: Coverage rate of the confidence interval constructed by (9) and (10)

Error distribution	n	Equation	$\sigma = 0.25$		$\sigma = 0.5$	
			$I = 4$	$I = 5$	$I = 4$	$I = 5$
$N(0,1)$	200	(9)	0.906	0.841	0.947	0.959
		(10)	0.960	0.930	0.932	0.931
	400	(9)	0.940	0.925	0.952	0.942
		(10)	0.957	0.954	0.920	0.937
t_5	200	(9)	0.792	0.748	0.804	0.799
		(10)	0.929	0.946	0.882	0.907
	400	(9)	0.793	0.771	0.784	0.800
		(10)	0.919	0.932	0.900	0.908

The effect of I on the coverage rate is negligible, even though the result of $I = 5$ seems better than that of $I = 4$ when the sample size increases.

Example 8. For the nonparametric testing problem, we consider the following model

$$y_i = x_i + B\exp(-16x_i^2) + \sigma\varepsilon_i, i = 1, ..., n,$$

where $\varepsilon_i \sim N(0,1)$ or $\sqrt{3/5}t_5$, $x_i \sim$ Unif$[-2,2]$, and x_i and ε_i are independent. We investigate the performance of the proposed testing procedure under the local alternative hypotheses with B changing from 0 to 1. We simulate 1000 random samples of size $n = 200$ with $\sigma = 0.25, 0.5$ and $I = 5$.

The power functions of the proposed test statistics under different error distributions are shown in Figure 3.1. It is obvious that the error distribution has little effect on the power function. When the signal-to-noise ratio is relative large, our proposed testing method has good power. Nevertheless, further theoretical investigation is required to understand the proposed nonparametric test statistic.

Example 9. Similar to Example 2 in Fan and Yao (1998), we simulate 400 random samples of size $n = 200$ from the model

$$y_i = x_i + 2\exp(-16x_i^2) + \sigma(x_i)\varepsilon_i, i = 1, ..., n,$$

with $\sigma(x) = 0.4\exp(-2x^2) + 0.2$ and x_i and ε_i are independent, $x_i \sim Unif[-2,2]$ and $\varepsilon_i \sim N(0,1)$. For each simulated example, the proposed variance function estimator

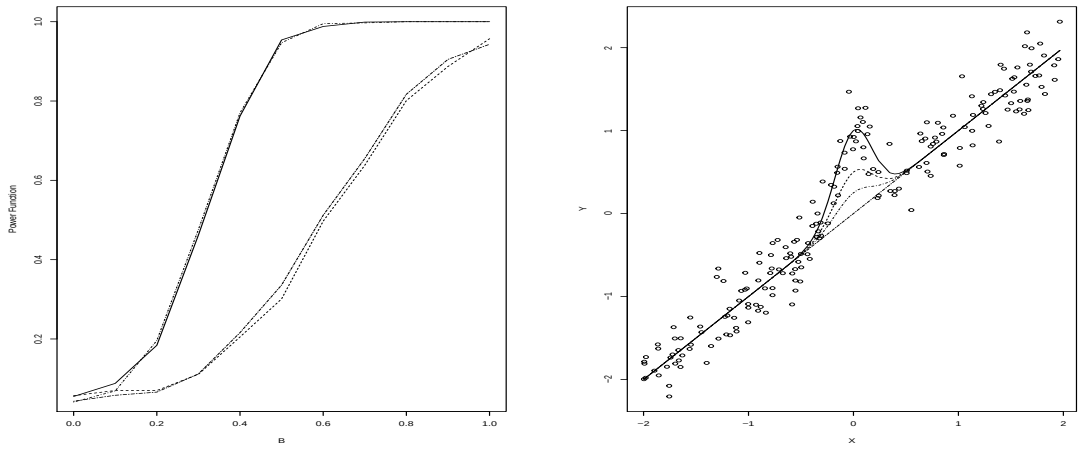

Figure 3.1: Example 8, Left: Power function of the test statistic T in Section 4.2 with $n = 200$. Solid line: $\sigma = 0.25$ with $\varepsilon_i \sim N(0,1)$, dashed line: $\sigma = 0.5$ with $\varepsilon_i \sim N(0,1)$, dot-dash line:$\sigma = 0.25$ with $\varepsilon \sim \sqrt{3/5} \cdot t_5$, two-dash line: $\sigma = 0.5$ with $\varepsilon \sim \sqrt{3/5} \cdot t_5$. Right: Hypothesis function: Solid line: $B = 1$, dashed line: $B = 0.5$, dot-dash line: $B = 0.25$, two-dash line: $B = 0$.

is evaluated by the mean absolute deviation error,

$$\text{MAD} = n_{\text{grid}}^{-1} \sum_{i=1}^{n_{\text{grid}}} |\hat{\sigma}(x_i) - \sigma(x_i)|,$$

where $\{x_i, i = 1, ..., n_{\text{grid}}\}$ are equally-spaced grid points on $[-1.8, 1.8]$ with $n_{\text{grid}} = 101$. The efficient estimation method (FY1998, Fan and Yao 1998) running by the C program code downloaded from Fan's personal website is used for comparison. To fit $\sigma(x)$ by the proposed method, we use the "locpol" function in R package *locpol* with the Epanechnikov kernel function and the bandwidth selected by the "thumbBw" function.

As shown by the boxplots in Figure 3.2, our proposed variance function method is comparable to the method proposed by Fan and Yao (1998); particularly, the refined method seems to perform slightly better than theirs in terms of both the mean and the standard deviation of MAD. Although without any theoretical justification, we think this may due to the stability of the proposed variance function estimation method, which avoids the selection of the optimal tuning parameter and the estimation of the unknown function.

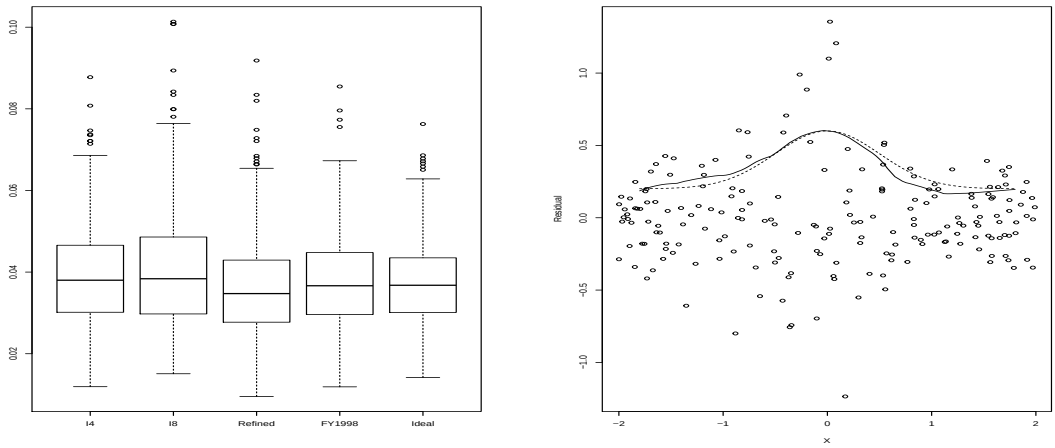

Figure 3.2: Example 9, Left: Boxplots of the mean absolute deviation curve based on 400 simulations for the proposed variance function estimation with $I = 4, 8$, the refined method, FY1998, as proposed by Fan and Yao (1998), and the ideal estimator, from left to right. Right: The sample residuals, the estimated variance function (solid line) by the refined method, and the true variance function (dashed line).

3.5.3 Real data analysis

Here, we apply our method to a real data set (Albright, Winston, and Zappe, 1999). In 1995, a bank was accused of gender discrimination. It was pointed out that the salaries of the female employees were significantly lower than those of the male employees. The data contains the information of 208 employees. Fan and Peng (2004) processed the data by deleting outliers and integrating the information. Thus 199 samples remained in the processed data set with seven variables: *EduLev*, a categorical variable indicating the employees' education level from 1 to 5 with 5 being the highest; *JobGrade*, a categorical variable indicating the job level from 1 to 6 with 6 being the highest; *Gender*, 0 for male and 1 for female; *Age*, the employee's age in 1995; *PCJob*, 1 for computer-related job and 0 for otherwise; *YrsExp*, the work experience of the employees by 1995, including their previous work experience in other banks; and *Salary*, annual salary in thousands of dollars in 1995. Fan and Peng(2004) proposed two models for investigating the gender influence on the salary:

the linear model

$$\begin{aligned}\text{Salary} =&\beta_0 + \beta_1\text{Female} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i \\ &+ \sum_{i=1}^{5}\beta_{6+i}\text{JobGrd}_i + \beta_{12}\text{YrsExp} + \beta_{13}\text{Age} + \epsilon\end{aligned}$$

and the partially linear additive model

$$\begin{aligned}\text{Salary} =&\beta_0 + \beta_1\text{Female} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i \\ &+ \sum_{i=1}^{5}\beta_{6+i}\text{JobGrd}_i + f_1(\text{YrsExp}) + f_2(\text{Age}) + \epsilon\end{aligned}$$

For the linear model, both the ordinary least squares method and the SCAD penalized likelihood method are applied to estimate the residual variance. For the partially linear additive model, similar to Fan, Guo, and Hao (2012), a 5-knot quadratic spline model with SCAD penalty is applied for the nonparametric functions, and the degrees of freedom are estimated by sample size n minus the number of non-zero parameters. Our local average estimators are also calculated based on the partially linear additive model. Table 3.10 summarizes the results. It shows that three residual sum of squares estimators and $\hat{\sigma}^2(5)$ are within the interval [12.5,12.9], but $\hat{\sigma}^2(4)$ and $\hat{\sigma}^2_*$ are much smaller and within the interval [11.45, 11.80].

Table 3.10: Residual Variance Estimates for the Bank Data, I

Method	OLS	SCAD_1	SCAD_2	$\hat{\sigma}^2(4)$	$\hat{\sigma}^2(5)$	$\hat{\sigma}^2_*$
Estimated Residual Variance	12.53	12.64	12.87	11.49	12.88	11.78

In fact, the results in Fan and Peng (2004) indicate that neither Age nor Gender is statistically significant in both models. Thus, we removed Age as it is closely correlated with YrsExp, for which the correlation coefficient is 0.69, and considered the interaction of YrsExp and Gender via a bivariate nonparametric function,

$$\text{Salary} = \beta_1 + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i + \sum_{i=1}^{5}\beta_{6+i}\text{JobGrd}_i + f(\text{YrsExp}, \text{Gender}) + \epsilon.$$

The local average estimators with different interval length I are computed based on both the partially linear additive model and the bivariate nonparametric model. The

results are listed in Table 3.11, which shows that the bivariate nonparametric model produces a more stable result, and the estimated residual variances are all around 12. It is also consistent with the refined local average estimator in Table 3.10. For the partially linear additive model, however the estimators fluctuate with different I. We conclude that the bivariate nonparametric model describes the data better, and the noise of the data set is around 12.

Table 3.11: Residual Variance Estimates for Bank Data, II

Model	$\hat{\sigma}^2(3)$	$\hat{\sigma}^2(4)$	$\hat{\sigma}^2(5)$	$\hat{\sigma}^2(6)$	$\hat{\sigma}^2(7)$
Partial Additive Model	10.50	11.49	12.88	12.38	11.11
Bivariate Nonparametric Model	11.91	12.20	11.81	12.22	12.18

3.6 Conclusion and discussion

The residual sum of squares method and difference-based method are two classical approaches in variance estimation. The residual sum of squares method is natural but needs to estimate the unknown function accurately and efficiently, whereas the difference-based method is robust and requires light computational cost but is limited to the univariate case and may not be optimal.

Our proposed local average method has the advantages of avoiding efficiently estimating the nonparametric functions and reducing much computational cost. What's more, it can be easily implemented for various nonparametric and semi-parametric models. The basic assumption is that the unknown nonparametric function is smooth enough and can be approximated well by a constant step function locally. Thus, we can reparameterize the nonparametric and semi-parametric models into a high-dimensional linear model and estimate the residual variance directly. Under some regularity conditions, we have proved that the local average variance estimator is asymptotically normal and achieves the asymptotic optimal rate $O_p(n^{-1})$. Simulation studies show that under certain circumstances the local average estimator involves bias. As the bias is closely related to the group size, we propose a refined local average variance estimator by aggregating variance estimators of different group sizes.

Furthermore, the basic idea of our proposed local average estimator for the er-

ror variance is trying to reduce the bias of the local estimates of the nonparametric function as much as possible even though it would inflates the variance of the estimators. In fact, when further using those local estimators to estimate some global parameters in the model, those inflated variance of the local parameter estimators or the nonparametric function estimators can be integrated, and the reduced biases can be ignored. With small sacrifice, we could make a trade-off between the computation and the efficiency of the parameter estimator. In this chapter, we try to use our proposed local average estimator for the residual variance estimation as one of the simple examples to illustrate such idea clearly.

3.7 Appendix

Proof of Theorem 3.1. Note that

$$\begin{aligned}
& \sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) \\
= \; & \sqrt{n}\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - \frac{1}{I}\sum_{j=1}^{I} y_{ij})^2 - \sigma^2\right] \\
= \; & \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\left(f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij}) + \epsilon_{ij} - \frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}\right)^2 - \sqrt{n}\sigma^2 \\
\hat{=} \; & S_1 + S_2 + S_3,
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij}))^2, \\
S_2 &= \frac{2I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij}))(\epsilon_{ij} - \frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}), \\
S_3 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(\epsilon_{ij} - \frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})^2 - \sqrt{n}\sigma^2.
\end{aligned}$$

Under Condition (b), we have

$$\begin{aligned} S_1 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\frac{1}{I^2}\sum_{i=1}^{k}\sum_{j=1}^{I}\left(\sum_{l=1}^{I}(f(x_{ij})-f(x_{il}))\right)^2 \\ &\le \frac{I}{I-1}\frac{1}{\sqrt{n}}\frac{1}{I}\sum_{i=1}^{k}\sum_{j=1}^{I}\sum_{l=1}^{I}(f(x_{ij})-f(x_{il}))^2 \\ &= \frac{2}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{1\le j<l\le I}(f(x_{ij})-f(x_{il}))^2 \\ &\le \frac{I}{\sqrt{n}}\sum_{i=1}^{n-1}(f(x_{i+1})-f(x_i))^2 = o_p(n^{-3/2}). \end{aligned}$$

For S_2, we have

$$\begin{aligned} S_2 &\le \frac{2I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\sum_{l=1}^{I}\sum_{m=1}^{I}|f(x_{ij})-f(x_{il})||\epsilon_{(n)}-\epsilon_{(1)}| \\ &\le \frac{2I}{I-1}\frac{1}{\sqrt{n}}(\epsilon_{(n)}-\epsilon_{(1)})\sum_{i=1}^{k}\sum_{1\le j<l\le I}|f(x_{ij})-f(x_{il})|, \end{aligned}$$

where $\epsilon_{(n)}$ and $\epsilon_{(1)}$ are the largest and smallest of ϵ_i's, respectively. By applying Lemma A.1 of Hsing and Carroll (1992), we have $n^{-1/4}|\epsilon_{(n)}-\epsilon_{(1)}| \xrightarrow{p} 0$, and then

$$S_2 \le \frac{I^3}{\sqrt{n}}(\epsilon_{(n)}-\epsilon_{(1)})\sum_{i=1}^{n-1}|f(x_{i+1})-f(x_i)| = o_p(n^{-1/4}).$$

Therefore,

$$\begin{aligned} \sqrt{n}(\hat{\sigma}_L^2-\sigma^2) &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})^2-\sqrt{n}\sigma^2+o_p(1) \\ &= \sqrt{I}\sqrt{k}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{I-1}\sum_{j=1}^{I}(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})^2\right)-\sigma^2\right]+o_p(1). \end{aligned}$$

Denote $U_i = \frac{1}{I-1}\sum_{j=1}^{I}(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})^2, 1\le i\le k$. It is easy to show that U_i's are independently and identically distributed with

$$\mathrm{E}[U_i]=\sigma^2 \text{ and } \mathrm{Var}[U_i]=\frac{\mu_4}{I}-\frac{I-3}{I(I-1)}\sigma^4.$$

Therefore, by the Central Limit Theorem, we have

$$\sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k}U_i-\sigma^2\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{\mu_4}{I}-\frac{I-3}{I(I-1)}\sigma^4\right),$$

or, equivalently,

$$\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) \xrightarrow{D} \mathcal{N}\left(0, \mu_4 - \frac{I-3}{I-1}\sigma^4\right).$$

□

Proof of Theorem 3.2. Denote $y_{ij} = f(x_{ij}) + Z_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$,$y_{ij}^* = \alpha_i + Z_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$, and $\hat{y}_{ij} = \hat{\alpha}_i + Z_{ij}^T\hat{\boldsymbol{\beta}}$. We then have

$$\begin{aligned}
\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) &= \sqrt{n}\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - \hat{y}_{ij})^2 - \sigma^2\right] \\
&= \sqrt{n}\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - y_{ij}^* + y_{ij}^* - \hat{y}_{ij})^2 - \sigma^2\right] \\
&\hat{=} S_1 + S_2 + S_3,
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - y_{ij}^*)^2, \\
S_2 &= \frac{2I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij} - y_{ij}^*)(y_{ij}^* - \hat{y}_{ij}), \\
S_3 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}^* - \hat{y}_{ij})^2 - \sqrt{n}\sigma^2.
\end{aligned}$$

Because $\alpha_i = \frac{1}{I}\sum_{j=1}^{I} f(x_{ij})$, the estimation error $y_{ij} - y_{ij}^* = f(x_{ij}) - \frac{1}{I}\sum_{j=1}^{I} f(x_{ij})$ is of the order $O(n^{-1})$. We can then show that $S_1 = O_p(n^{-3/2})$ and $S_2 = O_p(n^{-1/2})$. Thus,

$$\begin{aligned}
\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}^* - \hat{y}_{ij})^2 - \sqrt{n}\sigma^2 + o_p(1) \\
&= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}[(Z_{ij} - \frac{1}{I}\sum_{j=1}^{I} Z_{ij})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\epsilon_{ij} - \frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})]^2 \\
&\quad -\sqrt{n}\sigma^2 + o_p(1) \\
&\hat{=} S_4 + S_5 + S_6,
\end{aligned}$$

where

$$
\begin{aligned}
S_4 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\left[\left(Z_{ij}-\frac{1}{I}\sum_{j=1}^{I}Z_{ij}\right)^T\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^2 \\
S_5 &= \frac{2I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\left(Z_{ij}-\frac{1}{I}\sum_{j=1}^{I}Z_{ij}\right)^T\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\left(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}\right) \\
S_6 &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\left(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}\right)^2-\sqrt{n}\sigma^2+o_p(1).
\end{aligned}
$$

Applying Theorem 1 of Cui, Lu and Peng (2014), the estimation error for the coefficients $\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}$ is of the order $O_p(n^{-1/2})$. We can then show that $S_4 = O_p(n^{-1/2})$ and $S_5 = o_p(1)$. Thus,

$$
\begin{aligned}
\sqrt{n}(\hat{\sigma}_L^2-\sigma^2) &= \frac{I}{I-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\sum_{j=1}^{I}\left(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}\right)^2-\sqrt{n}\sigma^2+o_p(1) \\
&= \sqrt{I}\sqrt{k}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{I}{I-1}\left(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij}\right)^2-\sigma^2\right]+o_p(1).
\end{aligned}
$$

Similar to the proof of Theorem 3.1, denote $U_i = \frac{1}{I-1}\sum_{j=1}^{I}(\epsilon_{ij}-\frac{1}{I}\sum_{j=1}^{I}\epsilon_{ij})^2, 1\le i\le k$, and U_i's are independently and identically distributed with

$$
\mathrm{E}[U_i]=\sigma^2 \text{ and } \mathrm{Var}[U_i]=\frac{\mu_4}{I}-\frac{I-3}{I(I-1)}\sigma^4.
$$

Therefore, by the Central Limit Theorem, we have

$$
\sqrt{k}(\frac{1}{k}\sum_{i=1}^{k}U_i-\sigma^2)\xrightarrow{d}N(0,\frac{\mu_4}{I}-\frac{I-3}{I(I-1)}\sigma^4),
$$

or, equivalently,

$$
\sqrt{n}(\hat{\sigma}_L^2-\sigma^2)\xrightarrow{d}N(0,\mu_4-\frac{I-3}{I-1}\sigma^4).
$$

□

Proof of Theorem 3.3. Denote $y_{ij}=X_{ij}^T\mathbf{a}(U_{ij})+\epsilon_{ij}$, $y_{ij}^*=X_{ij}^T\mathbf{a}_i+\epsilon_{ij}$, and $\hat{y}_{ij}=X_{ij}^T\hat{\mathbf{a}}_i$. We then have

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\mathbf{Y}^T\mathbf{P}\mathbf{Y}}{tr(\mathbf{P})}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}-\hat{y}_{ij})^2}{n-pk} \\
&= \frac{\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}-y_{ij}^*+y_{ij}^*-\hat{y}_{ij})^2}{n-pk} \\
&\hat{=} S_1+S_2+S_3,
\end{aligned}
$$

where

$$\begin{aligned} S_1 &= \frac{1}{n-pk}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}-y_{ij}^*)^2, \\ S_2 &= \frac{2}{n-pk}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}-y_{ij}^*)(y_{ij}^*-\hat{y}_{ij}), \\ S_3 &= \frac{1}{n-pk}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}^*-\hat{y}_{ij})^2. \end{aligned}$$

Under Condition (C'), the estimation error $y_{ij}-y_{ij}^* = X_{ij}^T(\mathbf{a}(U_{ij})-\mathbf{a}_i)$ is of the order $O(n^{-1})$. We can then show that $S_1 = O_p(n^{-2})$ and $S_2 = O_p(n^{-1})$. Thus,

$$\begin{aligned} \hat{\sigma}_L^2 &= \frac{1}{n-pk}\sum_{i=1}^{k}\sum_{j=1}^{I}(y_{ij}^*-\hat{y}_{ij})^2 + o_p(n^{-1}) \\ &= \frac{1}{k}\sum_{i=1}^{k}\frac{\boldsymbol{\epsilon}_i^T\mathbf{P}_i\boldsymbol{\epsilon}_i}{I-p} + o_p(n^{-1}), \end{aligned}$$

where $\mathbf{P}_i = I - \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T$. Denote $Z_i = \frac{\boldsymbol{\epsilon}_i^T\mathbf{P}_i\boldsymbol{\epsilon}_i}{I-p}$, and then it is easy to show that

$$\mathrm{E}[Z_i] = \sigma^2 \quad \text{and} \quad \mathrm{Var}[Z_i] = \frac{1}{(I-p)^2}\mathrm{Var}[\boldsymbol{\epsilon}_i^T\mathbf{P}_i\boldsymbol{\epsilon}_i] = \frac{\mu_4-3\sigma^4}{(I-p)^2}\mathbf{p}_i^T\mathbf{p}_i + \frac{2\sigma^4}{I-p},$$

where $\mathbf{p}_i$ is the diagonal vector of $\mathbf{P}_i$. Denote $V_i = Z_i - \sigma^2$, and V_i's are independent with

$$\mathrm{E}[V_i] = 0 \text{ and } \mathrm{E}[V_i^2] \equiv \sigma_i^2 = \frac{\mu_4-3\sigma^4}{(I-p)^2}\mathbf{p}_i^T\mathbf{p}_i + \frac{2\sigma^4}{I-p}.$$

Denote

$$s_k^2 = \sum_{i=1}^{k}\sigma_i^2 = \frac{\mu_4-3\sigma^4}{(I-p)^2}\sum_{i=1}^{k}\mathbf{p}_i^T\mathbf{p}_i + \frac{2k\sigma^4}{I-p} = \frac{\mu_4-3\sigma^4}{(I-p)^2}\mathbf{p}^T\mathbf{p} + \frac{2k\sigma^4}{I-p},$$

where $\mathbf{p}$ is the diagonal vector of $\mathbf{P}$.

Because $\mathbf{p}_i^T\mathbf{p}_i \le \mathrm{tr}(\mathbf{P}_i^T\mathbf{P}_i) = \mathrm{tr}(\mathbf{P}_i) = I-p$, then

$$\int V_i^2 dP = \mathrm{E}[V_i^2] = \frac{\mu_4-3\sigma^4}{(I-p)^2}\mathbf{p}_i^T\mathbf{p}_i + \frac{2\sigma^4}{I-p} \le \frac{\mu_4-\sigma^4}{I-p} < \infty.$$

Thus, for $\delta > 0$, we have

$$\lim_{k\to\infty}\int_{|V_i|>\delta\sqrt{k}} V_i^2 dP = 0 \quad \text{and} \quad \lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\int_{|V_i|>\delta\sqrt{k}} V_i^2 dP = 0.$$

Because $\mathbf{p}^T\mathbf{p} \geq \frac{\text{tr}(\mathbf{P})^2}{n} = \frac{k^2(I-p)^2}{n} = \frac{(I-p)^2k}{I}$, then $s_k^2 > \frac{\mu_4-\sigma^4}{I}k > 0$. Thus, for $\delta > 0$ and $m = \frac{\mu_4-\sigma^4}{I}$, we have

$$\lim_{k\to\infty} \sum_{i=1}^{k} \frac{1}{s_k^2} \int_{|V_i|>\delta\sqrt{s_k^2}} V_i^2 dP < \lim_{k\to\infty} \frac{1}{mk} \sum_{i=1}^{k} \int_{|V_i|>\delta\sqrt{mk}} V_i^2 dP = 0,$$

and the Lindeberg conditions of the Central Limit Theorem are satisfied by $V_i, i = 1, ..., k$.

When the error ϵ_i's follow a normal distribution, $\mu_4 = 3\sigma^4$ and $s_k^2 = \frac{2k\sigma^4}{I-p}$. Therefore, we have

$$\frac{\sum_{i=1}^{k} V_i}{s_k} = \frac{k\left(\frac{\sum_{i=1}^{k} Z_i}{k} - \sigma^2\right)}{\sqrt{\frac{2k\sigma^4}{I-p}}} \xrightarrow{d} N(0,1),$$

and

$$\sqrt{n}(\hat{\sigma}_L^2 - \sigma^2) = \frac{\sqrt{I}k}{\sqrt{k}}\left(\frac{\sum_{i=1}^{k} Z_i}{k} - \sigma^2\right) + \sqrt{n}O_p(n^{-1}) \xrightarrow{D} \mathcal{N}(0, \frac{2I\sigma^4}{I-p}).$$

□

Chapter 4

Hypothesis Testing in Varying Coefficient Models

4.1 Introduction

In statistics literature, the scholars have discussed a lot about the regression analysis. The relationship of the dependent and independent variables always arouses our interest. Beginning with the simple linear model, the parametric models have shown their remarkable virtues—convenient setup, well studied statistical properties and mature applications. However, all of these will be in vain if the parametric assumption is wrong. For the nonparametric model, the model flexility has been granted but the statistical inference would be difficult without any structural assumptions. What's more, the nonparametric models suffer severely from the "curse of dimensionality". Therefore, people intend to combine these two kinds of models and semi-parametric models have been heatedly discussed. With some nonparametric components, the semi-parametric model is much more flexible than the ordinary parametric models. At the meantime, the semi-parametric models keep certain structures so that the aim of the statistical inference will be more clear and the adverse influence of dimensionality can be reduced. Popular examples include partially liner models(Engle, Granger, Rice and Weiss 1986; Robinson 1988), additive models(Friedman and Stuetzle 1981; Buja, Hastie and Tibshirani 1989) and varying coefficient models(Hastie and Tibshirani 1993).

The varying coefficient model was firstly introduced by Hastie and Tibshirani (1993). With the predictors $(u, x_1, ..., x_p)$ and the response variable y, it usually takes a form as

$$y = \sum_{l=1}^{p} a_l(u)x_l + \epsilon, \tag{4.1}$$

where u is the covariate that controls the coefficient functions $a_l(\cdot)$, $l = 1, ..., p$. It is also assumed that $\mathrm{E}[\epsilon|u, x_1, ..., x_p] = 0$ and $\mathrm{Var}[\epsilon|u, x_1, ..., x_p] = \sigma^2$. In all, the varying coefficient model possesses a linear structure, which is favorable in the statistical analysis. Different form the classical linear model, the coefficients in varying coefficient model are replaced by functions. Thus, the association of the variables will not be restricted to a fixed relationship. The relations could vary with different values of covariate u. This property is very appealing in practical applications since we can allow the coefficients change over time, i.e. let $u = t$. In this way, the varying coefficient model is a natural choice when we want to analyse nonlinear time series data, longitudinal data and survival data. Hence, the varying coefficient model obtains great success in areas of economics, epidemiology, medical science and so on.

As we have introduced in Chapter 2, there are mainly two kinds of methods for estimating the functional coefficients. One is the basis functions estimation. The coefficient functions will be approximated by a basis expansion, then the least squares is employed. When Hastie and Tibshirani (1993) proposed the varying coefficient model, the natural cubic splines are used for estimation. Other smoothing splines estimations have been studied by Hoover et al.(1998), Wu and Chiang(2000) and Chiang et al.(2001). Huang, et al.(2002) studied the statistical properties of the general basis functions estimators. Though the basis functions estimators are intuitional and easy to understand, they involve arbitrary choices of the basis functions and related parameters. The other popularly used estimating method is the local polynomial estimator. Since the varying coefficient model is essentially a local linear model, the local polynomial estimators seem more suitable. Taylor expansions are applied to the functional coefficients and the estimators will be calculated by the least squares with kernel. Hoover et al.(1998) used a weighted local polynomial estimator to estimate $a_l(\cdot)$ but this one step local polynomial estimator is not efficient if the coefficient functions have different degrees of smoothness. Fan and Zhang(1999) presented a

two step estimator. With the initial estimators substituted into the model, a higher order local polynomial estimator with suitable bandwidth is calculated for the objective coefficient function. Then the needs for different degrees of smoothness can be satisfied in the second step.

Naturally the question will arise that whether certain coefficient $a_l(\cdot)$, $l = 1, ..., p$ is really varying. More generally, we want to know if the coefficient function admits any structures. The researchers have investigated many kinds of difference between the null and the alternative hypothesis to get the test statistics and the corresponding rejection rules. With the local polynomial estimators, Fan and Zhang(2000) studied the deviations of the estimated coefficient function and the true coefficient function. The asymptotic distribution for the maximum of the normalized deviations has been deduced so that hypothesis tests can be conducted. However, this test statistic involves many unknown quantities. The estimation for these unknown parameters is complicated. Another approach is that we compare the log-likelihood or the residual squares under the null and the alternative hypothesis. Take the difference or the ratios as the test statistics and use bootstrap to the get the rejection rules. See Cai, et al. (2000a) , Cai, et al(2000b) and Huang, et al.(2002) for different estimators and data types. Fan, et al.(2001) proposed generalized likelihood ratio (GLR) tests and illustrated the idea with varying coefficient model in detail. The GLR test uses the difference of the log-likelihood under the alternative hypothesis and the null hypothesis as the test statistics. Different from the classical maximum likelihood ratio test, the maximum likelihood estimator in log-likelihood under the alternative is replaced by any reasonable nonparametric regression estimators. Thus the GLR tests can be widely applied to many models. Also, Fan, et al. have proved that the GLR tests are optimal and follow the Wilk's phenomena. The GLR tests provide many possibilities for model checking problems. In fact, the tests in Cai, et al. (2000a) , Cai, et al(2000b) and Huang, et al.(2002) are some specific scenarios of GLR tests.

Notice that for all the commonly used test methods, the varying coefficient model has to be fully fitted first. The estimation procedure, no matter using the basis functions method or the local polynomial method, will need loads of computational work. If the bootstrap is also implemented, the computation burden will be even

heavier. In this chapter, we present three tests to deal with this problem. The new tests avoid unnecessary estimation for the nuisance parameters and functions. The basic idea of the new tests is as follow. Firstly, we use local average method to estimate the function coefficient roughly and get some primary point estimators about the objective coefficient function. Secondly, some classical nonparametric test can be applied to the primary estimator to check the model assumptions. Moreover, we can get the estimator of the residual variance via the local average method as we did in Chapter 3. Then we can make use of the residual variance estimators to conduct the hypothesis test.

The remainder of this chapter is organized as follow. In Section 2, classical nonparametric tests for the varying coefficient model are briefly introduced first. Then the three new tests are discussed in detail. The testing problem is stated and the test statistics are presented. In Section 3, the theorems about the asymptotic distribution under null hypothesis and the Wilk's type of results are developed. Several simulations are conducted in Section 4 to demonstrate the performance of the proposed tests. Section 5 is the summary.

4.2 Methodology

Generally, we want to test whether the functional coefficients in model 4.1 possess any structures. Ranging form a simple constant to a given family of functions, the test problems can be various. For simplicity, we suppose that we only want to check whether the last functional coefficient, i.e., $a_p(\cdot)$, is constant. The extensions to general cases are similar. Write the varying coefficient model in matrix form

$$y = X^T \mathbf{a}(u) + \epsilon.$$

With the random sample $(U_i, X_i, y_i), i = 1, ..., n$, the hypothesis testing problem is:

$$H_0 : a_p(\cdot) = c, \leftrightarrow H_1 : a_p(\cdot) \neq c \tag{4.2}$$

where c is an unknown constant.

First we discuss the tests based on the local polynomial estimator. If we consider all the functional coefficients share the same degree of smoothness, it is sufficient to

use the one step local linear estimator. In a small neighbourhood of any given point u_0, the coefficient function $a_l(u)$ can be approximated locally as

$$a_l(u) \approx a_l + b_l(u - u_0), l = 1, ..., p$$

Then minimizing

$$\sum_{i=1}^{n} \{y_i - X_i^T \mathbf{a} - X_i^T \mathbf{b}(U_i - u)\}^2 K_h(U_i - u)$$

with respect to $(\mathbf{a}, \mathbf{b})$, for a given kernel function K and a bandwidth h, to get the estimator

$$\hat{\mathbf{a}}_p(u) = e_{p,2p}^T (\Lambda_u^T W_{h,u} \Lambda_u)^{-1} \Lambda_u^T W_{h,u} Y$$

where

$$\begin{aligned}
&\mathbf{X} = (X_1, ..., X_n)^T, \qquad \mathbf{U}_u = \text{diag}(U_1 - u, ..., U_n - u), \\
&\Lambda_u = (\mathbf{X}, \mathbf{U}_u \mathbf{X}), \qquad Y = (y_1, ..., y_n)^T, \\
&W_{h,u} = \text{diag}(K_h(U_1 - u), ..., K_h(U_n - u))
\end{aligned}$$

and $e_{k,m}$ denotes the unit vector of length m with 1 at the kth position. For convenience, we also denote the observed covariates vector by $\mathcal{D}$, i.e,

$$\mathcal{D} = (U_1, ..., U_n, x_{11}, ..., x_{1n}, ..., x_{p1}, ..., x_{pn})^T.$$

With local linear estimators, Fan and Zhang(2000) has shown that under some regularity conditions, the asymptotic distribution of the maximum discrepancy between the estimated functional coefficient and true functional coefficient can be reduced. Let the test statistic

$$T = (-2 \log h)^{1/2} (\sup |\{\widehat{\text{var}}(\hat{a}_p|\mathcal{D})\}^{1/2} (\hat{a}_p - \hat{c} - \widehat{\text{bias}}(\hat{a}_p|\mathcal{D}))| - d_{v,n}),$$

then

$$P(T < x) \to \exp\{-2e^{-x}\}.$$

So the null hypothesis H_0 should be rejected if large value of T is observed. However, this test statistic involves many unknown components. The estimation for these unknown parameters is quite complicated. Higher order derivatives of $a_p(\cdot)$ is needed for the bias term, and the whole model need fitting to get the variance estimator.

Fan at el.(2001) came up with the method of GLR tests. The main idea of the GLR tests is that the estimators under alternative hypothesis can be flexible. What's more, since the GLR tests enjoy the Wilk's type of property, the nuisance parameters in null hypothesis can also be easily estimated. As long as the estimators are reasonable, the test statistics will remain well performed. Still consider the above testing problem 4.2, the GLR test statistic is give by

$$\lambda_n = \frac{n}{2}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1} \approx \frac{n}{2}\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}$$

with

$$\mathrm{RSS}_0 = \sum_{i=1}^{n}(y_i - \hat{a}_1(U_i)x_{i1} - \cdots - \hat{a}_{p-1}(U_i)x_{i,p-1} - \hat{c}x_{ip})^2,$$
$$\mathrm{RSS}_1 = \sum_{i=1}^{n}(y_i - \tilde{a}_1(U_i)x_{i1} - \cdots - \tilde{a}_{p-1}(U_i)x_{i,p-1} - \tilde{a}_p(U_i)x_{ip})^2.$$

Here, we use the notations $\hat{a}_1(U_i), ..., \hat{a}_{p-1}(U_i)$ to represent the estimators under the null hypothesis while $\tilde{a}_1(U_i), ..., \tilde{a}_p(U_i)$ are for those under the alternative. In Fan at el(2001), the local liner estimators are applied in both cases. Under the conditions in Fan at el(2001), it is proved that the asymptotic null distribution of the GLR test is χ^2 with large degrees of freedom, i.e

$$r\lambda_n \overset{a}{\sim} \chi^2_{b_n},$$

where r and b_n are some value related to the kernel function and some basic properties of the covariates. In addition, due to the Wilk's phenomena, r and b_n can be simply gained by bootstrap simulation, though the explicit formulas are also given.

Analogously, the tests with basis function estimators can be conducted in the same way. Each functional coefficient $a_l(\dot{})$ is expanded by basis functions

$$a_l(u) \approx \sum_{s=0}^{K_l}\gamma_{ls}B_{ls}(u), l = 1, ..., p$$

where $B_{ls}(s = 1, ..., K_l)$ is a set of basis functions. Then the estimators $\hat{\gamma}_{ls}$ can be estimated by minimising

$$\sum_{i=1}^{n} w_i\{y_i - \sum_{l=1}^{p}\sum_{s=1}^{K_l}x_{il}B_{ls}(U_i)\gamma_{ls}\}^2$$

where w_i's are some reasonable weights. For the testing problem 4.2, Huang at el.(2002) suggested a test statistic $T = (\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$ with

$$\mathrm{RSS}_0 = \sum_{i=1}^{n} w_i \{y_i - \sum_{l=1}^{p-1} \sum_{s=1}^{K_l} x_{il} B_{ls}(U_i)\hat{\gamma}_{ls} - \hat{c}x_{ip}\}^2,$$

$$\mathrm{RSS}_1 = \sum_{i=1}^{n} w_i \{y_i - \sum_{l=1}^{p} \sum_{s=1}^{K_l} x_{il} B_{ls}(U_i)\tilde{\gamma}_{ls}\}^2.$$

$\hat{\gamma}_{ls}$ and $\tilde{\gamma}_{ls}$ are the estimators under null and alternative hypothesis respectively. Next step is using bootstrap to calculate the rejection region. We can find that this is actually a specific incidence of the GLR test.

As we have mentioned, the existing tests require intensive computation. Even if the bootstrap can be avoided in the GLR test, the estimation of the nuisance coefficients can still be time-consuming. What's worse, the one step local polynomial estimator is not efficient if the coefficient functions have different degrees of smoothness. So the estimation process may consist of several two step local polynomial estimations. Though the basis function estimator can adapt to different smoothness for different coefficients, the selection of the basis functions and corresponding smoothing parameters is not easy. Since only the last coefficient $a_p(\cdot)$ is of interest in the hypothesis testing, such big effort to estimate the other coefficients is not preferable. Therefore, we want to find some methods that can get rid of the unnecessary estimation. Thus the computation burden can be reduced.

In this chapter, we propose three tests based on the local average method. Instead of getting the fully smoothed function estimators, we only calculate some primary point estimators. The hypothesis tests then can be conducted based on these primary point estimators. Since the functional coefficient can be treated as locally constant, thus we can get a ordinary least squares estimator in a certain small neighbourhood of u. Latter in this chapter we will prove that this ordinary least squares estimator is an asymptotically unbiased estimator for the function value of "average-point" in this neighbourhood. After collecting all the "average-point" estimators, we can use some classic nonparametric tests to study the function structure.

The details of the implementation is demonstrated as following. Firstly, we sort the samples $(U_i, X_i, y_i), i = 1, ..., n$ according to U_i in an ascending order, i.e., $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$. Then divide the sorted samples into k groups with I samples

in each group. Since we should make sure that the neighbourhood is sufficiently narrow, small and fixed I is selected. Then k will increase with the sample size n. Thus, $n = Ik$. (In practise, the remainders will be removed out if n is not evenly divisible by I. As I is small enough, the number of the removed samples is also small. So this part is negligible.)

Denote the jth observation in the ith group as $(U_{ij}, X_{ij}, y_{ij}), i = 1, ..., k, j = 1, ..., I$. Hence, we have

$$y_{ij} = X_{ij}^T \mathbf{a}(U_{ij}) + \epsilon_{ij}, \quad i = 1, ..., k, \quad j = 1, ..., I.$$

Since we treat the functional coefficient $\mathbf{a}(\cdot)$ as constant in the small neighbourhood, we assume in every group $\mathbf{a}(U_{i1}) \approx \mathbf{a}(U_{i2}) \approx \cdots \approx \mathbf{a}(U_{iI}) \approx \mathbf{a}_i \approx \mathbf{a}(\bar{U}_{i\cdot})$. Thus, for the ith group, we have

$$\mathbf{Y}_i = \mathbf{X}_i \mathbf{a}_i + \boldsymbol{\epsilon}_i^*$$

where

$$\begin{aligned} &\mathbf{Y}_i = (y_{i1}, y_{i2}, ..., y_{iI})^T \in \mathbb{R}^{I\times 1}, \quad \mathbf{a}_i = (a_1(\bar{U}_{i\cdot}), a_2(\bar{U}_{i\cdot}), ..., a_p(\bar{U}_{i\cdot}))^T \in \mathbb{R}^{p\times 1}, \\ &\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{iI})^T \in \mathbb{R}^{I\times p} \quad \text{and} \quad \boldsymbol{\epsilon}_i = (\epsilon_{i1}^*, \epsilon_{i2}^*, ..., \epsilon_{iI}^*)^T \in \mathbb{R}^{I\times 1}. \end{aligned}$$

Combine all the k groups, we get

$$\mathbf{Y} = \mathbb{X}\mathbf{a} + \boldsymbol{\epsilon}^*$$

where

$$\begin{aligned} &\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, ..., \mathbf{Y}_k^T)^T \in \mathbb{R}^{n\times 1}, \quad \mathbf{a} = (\mathbf{a}_1^T, \mathbf{a}_2^T, ..., \mathbf{a}_k^T)^T \in \mathbb{R}^{kp\times 1}, \\ &\mathbb{X} = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \cdots \oplus \mathbf{X}_k \in \mathbb{R}^{n\times kp} \quad \text{and} \quad \boldsymbol{\epsilon}^* = (\boldsymbol{\epsilon}_1^{*T}, \boldsymbol{\epsilon}_2^{*T}, ..., \boldsymbol{\epsilon}_k^{*T})^T \in \mathbb{R}^{n\times 1}. \end{aligned}$$

Now we can get the primary point estimators

$$\begin{aligned} \hat{\mathbf{a}} &= (\hat{a}_1(\bar{U}_{1\cdot}), ..., \hat{a}_p(\bar{U}_{1\cdot}), \hat{a}_1(\bar{U}_{2\cdot}), ..., \hat{a}_p(\bar{U}_{2\cdot}), ..., \hat{a}_1(\bar{U}_{k\cdot}), ..., \hat{a}_p(\bar{U}_{k\cdot}))^T \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \end{aligned} \tag{4.3}$$

From the estimators $\hat{\mathbf{a}}$, we can get k estimators related to $a_p(\cdot)$, namely, $\hat{a}_p(\bar{U}_{1\cdot})$, $\hat{a}_p(\bar{U}_{2\cdot})$, ..., $\hat{a}_p(\bar{U}_{k\cdot})$. In the next section, we will see that $\mathrm{E}(\hat{a}_p(\bar{U}_{i\cdot})) = a_p(\bar{U}_{i\cdot}) + O_p(\frac{\log n}{n})$ and $\mathrm{Var}(\hat{a}_p(\bar{U}_{i\cdot})) = e_{p,p}^T \mathrm{E}[(\sum_{i=1}^I X_i X_i^T)^{-1} | \bar{U}_{i\cdot}] e_{p,p} \sigma^2$ for $i = 1, ..., k$.

Then the testing problem can be regarded as based on a simple nonparametric model

$$\hat{a}_p(\bar{U}_{i\cdot}) = a_p(\bar{U}_{i\cdot}) + v_p(\bar{U}_{i\cdot})\epsilon, \quad i = 1, ..., k$$

with $v_p^2(\bar{U}_{i\cdot}) = e_{p,p}^T \mathrm{E}[(\sum_{i=1}^I X_i X_i^T)^{-1} | \bar{U}_{i\cdot}] e_{p,p}$.

By now, we have transformed the varying-coefficient model into a simple univariate nonparametric model. The problem becomes easier no matter in estimation or in model checking. We can apply some classical methods to conduct the hypothesis test. Notice that the model is heteroscedastic, we first consider the test proposed by Zheng(1996). Under the null hypothesis, the coefficient $a_p(\cdot)$ is constant, then we can have a simple estimator for the constant c

$$\hat{c} = \frac{1}{k} \sum_{i=1}^{k} \hat{a}_p(\bar{U}_{i\cdot})$$

and the test statistics is

$$T_1 = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} \frac{1}{h} K(\frac{\bar{U}_{i\cdot} - \bar{U}_{j\cdot}}{h}) \hat{e}_i \hat{e}_j.$$

where $\hat{e}_i = \hat{a}_p(\bar{U}_{i\cdot}) - \hat{c}$.

If the function on the conditional variance $v_p^2(\bar{U}_{i\cdot}) = e_{p,p}^T \mathrm{E}[(\sum_{i=1}^I X_i X_i^T)^{-1} | \bar{U}_{i\cdot}] e_{p,p}$ is known or can be estimated efficiently, we can also apply the GLR test to this problem. Then test statistics is

$$T_2 = \frac{n}{2I} \log \frac{\sum_{i=1}^k (\hat{a}_p(\bar{U}_{i\cdot}) - \frac{1}{k} \sum_{i=1}^k \hat{a}_p(\bar{U}_{i\cdot}))^2}{\sum_{i=1}^k (\hat{a}_p(\bar{U}_{i\cdot}) - \tilde{m}_h(\bar{U}_{i\cdot}))^2}$$

where $\tilde{m}_h(\bar{U}_{i\cdot})$ is a reasonable estimator for $a_p(\bar{U}_{i\cdot})$ under the alternative hypothesis, for example, the local linear estimator.

Notice that in T_2, we only need to estimate one functional coefficient $a_p(\cdot)$, i.e. $\tilde{m}_h(\cdot)$. If we apply the GLR test directly to the original varying coefficient model, we have to estimate other $2(p-1)$ uninterested functional coefficients. The computation burden has been sharply lessened after we employing the local average method.

Nevertheless, the function $v_p^2(\bar{U}_{i\cdot})$ is usually unknown. Notice that the GLR test is based on the residual squares, so we may ignore all the function estimation as long as we can find a way to estimate the residual variance efficiently. Remind that in the first step, we get the primary point estimators for all the functional coefficients

of the varying coefficient model. We can just substitute these point estimators back to the varying coefficient model to get the estimated residual squares. That is, the fitted value $\hat{\mathbf{Y}}$ can be written as,

$$\hat{\mathbf{Y}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$$

Then the residual errors are estimated

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Therefore, the estimation for the sum of residual squares under the alternative hypothesis is

$$\widehat{\mathrm{RSS}}_1 = \mathbf{Y}^T\mathbf{P}\mathbf{Y} \tag{4.4}$$

where $\mathbf{P} = \mathbf{I} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

The estimation of RSS_0 under null hypothesis is similar. Since we have known that the last coefficient is constant , then the model becomes

$$y = \ddot{X}^T\ddot{\mathbf{a}}(u) + cx_p + \epsilon.$$

where $\ddot{X} = (x_1, ..., x_{p-1})^T$ and $\ddot{\mathbf{a}}(u) = (a_1(u), a_2(u), ..., a_{p-1}(u))^T$. Divide the samples into k groups according to the variate U_i as above. Let $\boldsymbol{\Phi} = (\ddot{\mathbb{X}}, X_p)$ and $\boldsymbol{\theta} = (\ddot{\mathbf{a}}^T, c)^T$, where $\ddot{\mathbb{X}} = \ddot{\mathbf{X}}_1 \oplus \ddot{\mathbf{X}}_2 \oplus \cdots \oplus \ddot{\mathbf{X}}_k \in \mathbb{R}^{n\times k(p-1)}$ and $\ddot{\mathbf{a}} = (\ddot{\mathbf{a}}_1^T, \ddot{\mathbf{a}}_2^T, ..., \ddot{\mathbf{a}}_k^T)^T \in \mathbb{R}^{k(p-1)\times 1}$ as we just denoted above, with all the p-dimension covariate X replaced by the $(p-1)$-dimension covariate $\ddot{X}$. Then we can write the model as

$$\mathbf{Y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\epsilon}^*.$$

Therefore the sum of residual squares under the null hypothesis can be estimated as

$$\widehat{\mathrm{RSS}}_0 = \mathbf{Y}^T\ddot{\mathbf{P}}\mathbf{Y}$$

where $\ddot{\mathbf{P}} = \mathbf{I} - \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$.

Then the final test statistic is

$$T_3 = \frac{n}{2}\frac{\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1}{\widehat{\mathrm{RSS}}_1} = \frac{n}{2}\frac{\mathbf{Y}^T\ddot{\mathbf{P}}\mathbf{Y} - \mathbf{Y}^T\mathbf{P}\mathbf{Y}}{\mathbf{Y}^T\mathbf{P}\mathbf{Y}}$$

For the first two tests T_1 and T_2, the wanted functional coefficient has been taken out so that the problem is simplified to a simple nonparametric regression model.

This merit makes the proposed tests very attractive when the testing problem is focused on individual coefficient. The uninterested coefficients are left behind so that no extra estimation process is needed. The last test T_3 based on the estimated RSS is conducted on the original varying coefficient model, so a simultaneous result can be obtained if the testing problem is about several coefficients. In the next section, we will present the asymptotic distributions of the proposed tests and their statistical properties.

4.3 Theorem

To deduce the asymptotic properties of the proposed tests, the following assumptions are required

1. The marginal density function $f(u)$ of U is bounded away form $\delta > 0$ and from ∞. Its first-order derivatives are bounded.

2. $a(u)$ has the continuous second derivative.

3. X is bounded. $\Gamma(U) = \mathrm{E}[(\sum_{i=1}^{I} X_i X_i^T)^{-1}|U]$ exists for any given U in its support and is continuously differentiable.

4. $\mathrm{E}[\epsilon^4] = \mu_4 < \infty$.

5. The function $K(t)$ is a symmetric density function with a compact support.

In addition, we introduce some notations here for further use in the theorems. Let

$$\kappa_1 = K(0) - \frac{1}{2}\int K^2(t)dt \qquad \kappa_2 = \int \{K(t) - \frac{1}{2}K * K(t)\}^2 dt$$

$$\Psi_n = \sum_{i=1}^{k} M_i^{-2} \sum_{j=1}^{I} m_{ij}^4 \qquad M_i = B_i - C_i^T A_i^{-1} C_i \qquad m_{ij} = C_i^T A_i^{-1} \ddot{X}_{ij} - x_{ijp}$$

$$A_i = \sum_{j=1}^{I} \ddot{X}_{ij} \ddot{X}_{ij}^T \qquad B_i = \sum_{j=1}^{I} x_{ijp}^2 \qquad C_t = \sum_{j=1}^{I} \ddot{X}_{ij} x_{ijp}$$

$$D_{i1} = \sum_{j=1}^{I} \ddot{X}_{ij} \epsilon_{ij} \qquad D_{i2} = \sum_{j=1}^{I} x_{ijp} \epsilon_{ij}$$

where $K * K$ denotes the convolution of kernel function K.

As we have mentioned in Chapter 2, the primary point estimators are asymptotically unbiased and we also have the explicit forms of their variances. Here, we use a lemma to state the statistical properties of these point estimators obtained by local average method.

Lemma 4.1. *Given assumptions 1- 4, with small and fixed I, when $n \to \infty$, then for the primary local average estimator (4.3), we have*

$$E(\hat{a}_l(\bar{U}_{i\cdot})) \to a_l(\bar{U}_{i\cdot}), \quad Var(\hat{a}_l(\bar{U}_{i\cdot})) \equiv \sigma_l^2(\bar{U}_{i\cdot}) = (e_{l,p}^T \Gamma(\bar{U}_{i\cdot}) e_{l,p})\sigma^2, \ \ l = 1, ...p, i = 1, .., k.$$

Note that when $n \to \infty$, the points in the same group will be narrowed to one single point since the support of U is bounded. Hence, we consider the distribution of the average point $\bar{U}_{i\cdot}$ is the same as the distribution of U when $n \to \infty$. In fact, follow the proof of Lemma 4.1, we could choose any point in the group as the "average-point" and the asymptotic properties remain. The foundation of the Lemma 4.1 is that the change of the function values in a small neighbourhood could be neglected. Under the assumption 1-4, the difference of the function values in the same group is of $O(\frac{\log n}{n})$.

Combining the Lemma 4.1 and Lemma 3.3 in Zheng(1996), we can get the asymptotic distribution of T_1 under the null hypothesis.

Theorem 4.1. *Given assumptions 1-5, if $h \to 0$ and $nh \to \infty$, then under the null hypothesis (4.2),*

$$nh^{1/2} T_1 \xrightarrow{d} N(0, \sigma_1^2)$$

where σ_1^2 is the asymptotic variance of $nh^{1/2}T_1$,

$$\sigma_1^2 = 2I^2 \int K^2(s)ds \cdot \int \sigma_p^4(u) f(u) d(u)$$

and σ_1^2 can be consistently estimated by $\widehat{\sigma_1^2}$,

$$\widehat{\sigma_1^2} = \frac{2I^2}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} \frac{1}{h} K^2(\frac{\bar{U}_{i\cdot} - \bar{U}_{j\cdot}}{h}) \hat{e}_i^2 \hat{e}_j^2.$$

We can also standardize the test statistic as

$$\begin{aligned} V_1 &\equiv \sqrt{\frac{n-I}{n}} \frac{nh^{1/2} T_1}{\hat{\sigma}_1} \\ &= \frac{\sum_{i=1}^{k} \sum_{j \neq i}^{k} K(\frac{\bar{U}_{i\cdot} - \bar{U}_{j\cdot}}{h}) \hat{e}_i \hat{e}_j}{\{\sum_{i=1}^{k} \sum_{j \neq i}^{k} 2K^2(\frac{\bar{U}_{i\cdot} - \bar{U}_{j\cdot}}{h}) \hat{e}_i^2 \hat{e}_j^2\}^{1/2}}. \end{aligned}$$

Then the following corollary arises naturally.

Corollary 4.1. *Given assumptions 1-5, if $h \to 0$ and $nh \to \infty$, then under the null hypothesis (4.2),*

$$V_1 \xrightarrow{d} N(0,1)$$

Thus we can use Corollary 4.1 to calculate the critical value for the test.

Theorem 4.2. *Given assumptions 1-5, then under H_0, as $h \to 0$, $nh^{3/2} \to \infty$*

$$r_n T_2 \overset{a}{\sim} \chi^2_{a_n}$$

where

$$r_n = \frac{\kappa_1}{\kappa_2}[\int \sigma_p^2(u)du][\int \sigma_p^2(u)f(u)du][\int \sigma_p^4(u)du]^{-1},$$
$$a_n = \frac{\kappa_1^2}{\kappa_2}h^{-1}[\int \sigma_p^2(u)du]^2[\int \sigma_p^4(u)du]^{-1}$$

One may find that the the asymptotic results of T_2 is actually the same as the Remark 4.2 in Fan et al. (2001), with the weight function $w(x) = 1$. As a result, we could also use a weighted residual sum of squares in the test to offset the heteroscedastic influence. Let

$$\mathrm{RSS}_0' = \sum_{i=1}^{k}(\hat{a}_p(\bar{U}_{i\cdot}) - \frac{1}{k}\sum_{i=1}^{k}\hat{a}_p(\bar{U}_{i\cdot}))w(\bar{U}_{i\cdot}), \quad \mathrm{RSS}_1' = \sum_{i=1}^{k}(\hat{a}_p(\bar{U}_{i\cdot}) - \tilde{m}_h(\bar{U}_{i\cdot}))w(\bar{U}_{i\cdot})$$

where $w(u) = [\sigma_p^2(u)]^{-1}$, then

$$T_2' = \frac{n}{2I}\log\frac{\mathrm{RSS}_0'}{\mathrm{RSS}_1'}$$

and by Remark 4.2 in Fan et al. (2001), we have

$$r_n' T_2' \overset{a}{\sim} \chi^2_{a_n'}$$

with $r_n' = \frac{\kappa_1}{\kappa_2}$ and $a_n' = \frac{\kappa_1^2}{\kappa_2}h^{-1}|\Omega|$. Ω is the support of U. In this way, the Wilks type of result holds. When weighted residual sum of squares are used, the asymptotic result is the same with that of GLR test directly applied on the original varying coefficient model. The difference is that our proposal saves a lot of computation. If we directly use GLR test for the original varying coefficient model, we have to estimate other $p-1$ functional coefficients both under the null hypothesis and the alternative hypothesis.

In our method, we only need to estimate one coefficient function and one variance function $\sigma_p^2(\cdot)$.

Next we consider the asymptotic distribution of T_3.

Theorem 4.3. *Given assumptions 1-5, for a given I, as $n \to \infty$, then under H_0,*

$$\frac{2(I-p)}{I}\sigma_3^{-1}(T_3 - \frac{n}{2(I-p)}) \to N(0,1),$$

where $\sigma_3^2 = \Psi_n(\frac{\mu_4}{\sigma^4}-3)+2n/I$. Furthermore, if ϵ follows a mesokurtic distribution(say, normal distribution), then

$$\frac{2(I-p)}{I}T_3 \overset{a}{\sim} \chi^2_{n/I}.$$

When ϵ is distributed with a normal distribution, the null distribution of T_3 is quite simple. The underlying χ^2 distribution is only related to the group size I, the covariates dimension p and the sample size n. The Wilks phenomenon is valid. Unlike T_1 and T_2, the estimation for the asymptotic mean and variance is not needed. This is a great merit of T_3. Even if the assumption of mesokurtic distribution is not satisfied, the calculation of Ψ_n is not complex. What's more, by CauchySchwarz inequality we have $n/2I \leq \Psi_n \leq n/I$.

4.4 Simulation

In this section, we investigate the finite-sample performance of the proposed tests. Consider the model

$$Y = a_1(U)X_1 + a_2(U)X_2 + \epsilon,$$

where U is uniformly distributed on $[0,1]$, the covariates (X_1, X_2) follows a bivariate normal distribution with zero mean and covariance $\begin{pmatrix} 1 & 2^{-1/2} \\ 2^{-1/2} & 1 \end{pmatrix}$, and $\epsilon \sim N(0, \sigma^2)$. We set $\sigma^2 = 0.2\text{Var}\{\text{E}[Y|U, X_1, X_2]\}$ so that the signal to noise ration is about $5:1$. What's more, U, ϵ and (X_1, X_2) are independent. We will conduct tests on whether the second coefficient $a_2(\cdot)$ is constant, i.e.

$$H_0 : a_2(\cdot) = c, \leftrightarrow H_1 : a_2(\cdot) \neq c$$

First we consider the null model with

$$a_1(U) = \sin(60U), \qquad a_2(U) = 1.$$

The test statistics T_1, T_2 and T_3 are simulated 1000 times for different sample size n and group size I. Then we calculated the proportion of rejections as the empirical size for the three tests under $\alpha = 0.05$ significant level. For test statistic T_1, the bandwidth h is chosen to be $\gamma \cdot n^{-2/5}$ as in Zheng(1996). For test statistic T_2, the local linear smoothing is adopted under the alternative hypothesis and the bandwidth is taken as $\gamma \cdot n^{-1/5}$ according to Fan and Gijbels(1996). What's more, since we let the covariates U and (X_1, X_2) are independent, the variance of our primary local average estimator then becomes a constant. Therefore, in T_1 and T_2, Zheng's method and the GLR method are both applied to a simple homoscedastic model. The parameter γ is a constant to control the bandwidth, so that we can get a general idea about the influence of different bandwidths on the test statistics. In both tests, the Epanechnikov kernel is employed and so it is in the later simulations.

We summarise the results of size study of T_1, T_2 and T_3 in Table 4.1, Table 4.2 and Table 4.3 respectively.

As can be seen,in most cases the test T_1 has size close to 0.05. When sample size n becomes larger, the sizes tend to the asymptotic value. What's more, when the sample size is large enough, the influence caused by the choice of group size I and bandwidth seems slight.

Table 4.2 shows the rejection rates of T_2 in the finite-sample setting. In general, the sizes get closer to 0.05 as n increases. Based on this table, we still can not tell the difference among different bandwidths and the group sizes. Notice that the sizes in the table are all larger than 0.05, though the convergent trend exits.

From Table 4.3, we can see that the simulation results is satisfactory. The sizes of T_3 converge to 0.05 rapidly as n grows. The test statistics with $I = 10$ outperform those of the other two cases. This is consistent with the results in Chapter 3. When sample size is large enough , larger group size I will give a better estimation of the residual variance.

In order to have a more intuitional understanding about the asymptotic distribution of the test statistics under the null hypothesis, we plot the density functions of the three proposed test statistics for the null model. For each test, 1000 replicates are conducted from the null model and their sample distributions is obtained by a

Table 4.1: Proportion of rejections for null model with T_1

		T_1		
Parameter γ	Sample size n	I = 4	I = 5	I = 10
0.5	200	0.035	0.039	0.026
	400	0.042	0.042	0.037
	600	0.041	0.038	0.037
	800	0.045	0.045	0.038
	1000	0.047	0.045	0.040
	1200	0.042	0.043	0.041
	1400	0.041	0.042	0.041
	1600	0.042	0.044	0.038
1	200	0.035	0.032	0.032
	400	0.037	0.036	0.039
	600	0.038	0.036	0.038
	800	0.040	0.036	0.038
	1000	0.043	0.040	0.042
	1200	0.041	0.041	0.039
	1400	0.038	0.043	0.044
	1600	0.042	0.042	0.036
2	200	0.025	0.024	0.026
	400	0.026	0.030	0.032
	600	0.031	0.028	0.034
	800	0.033	0.028	0.030
	1000	0.032	0.032	0.030
	1200	0.033	0.033	0.038
	1400	0.034	0.034	0.034
	1600	0.033	0.037	0.035

kernel density estimate.(Here we employ the function "ksdensity" in MATLAB.) In addition, all the test statistics are standardized so that we can compare the sample distributions with the standard normal distribution. The bandwidths are chosen to be $\frac{1}{2}n^{-2/5}$ and $n^{-1/5}$ for T_1 and T_2 respectively. Group size I are selected as 10 in all the three test statistics. Sample size n is set to be 400, 800, 1600 so that we can get a general idea about the convergent trend of the sample distributions.

It can be seen from Figure 4.1 that the sample distributions of standardized T_1 and T_3 have a similar bell shape as the standard normal distribution. What's more,

Table 4.2: Proportion of rejections for null model with T_2

		T_2		
Parameter γ	Sample size n	I = 4	I = 5	I = 10
0.5	200	0.093	0.105	0.138
	400	0.071	0.085	0.107
	600	0.068	0.076	0.090
	800	0.070	0.072	0.080
	1000	0.069	0.066	0.082
	1200	0.065	0.070	0.072
	1400	0.065	0.070	0.075
	1600	0.068	0.063	0.069
1	200	0.083	0.088	0.094
	400	0.075	0.079	0.089
	600	0.070	0.077	0.085
	800	0.069	0.070	0.074
	1000	0.062	0.066	0.078
	1200	0.066	0.070	0.071
	1400	0.066	0.067	0.069
	1600	0.067	0.058	0.066
2	200	0.089	0.084	0.092
	400	0.079	0.075	0.091
	600	0.074	0.078	0.079
	800	0.075	0.077	0.073
	1000	0.066	0.075	0.077
	1200	0.074	0.067	0.073
	1400	0.073	0.075	0.073
	1600	0.076	0.064	0.070

the sample distributions behave like the standard normal more as sample size n gets larger. For the test statistics T_2, there seems to present some discrepancy between the sample distribution and the standard normal distribution. Though the sample distribution is quite close to the normal standard, we can still find a long right tail. This may explain some of the facts that the size of test T_2 is usually larger than the significance level.

Next we conduct the power study of the proposed tests. Take the following two

Table 4.3: Proportion of rejections for null model with T_3

	T_3		
Sample size n	I = 4	I = 5	I = 10
200	0.105	0.097	0.067
400	0.118	0.078	0.057
600	0.094	0.069	0.059
800	0.084	0.075	0.049
1000	0.092	0.087	0.054
1200	0.078	0.086	0.053
1400	0.077	0.07	0.056
1600	0.081	0.053	0.052

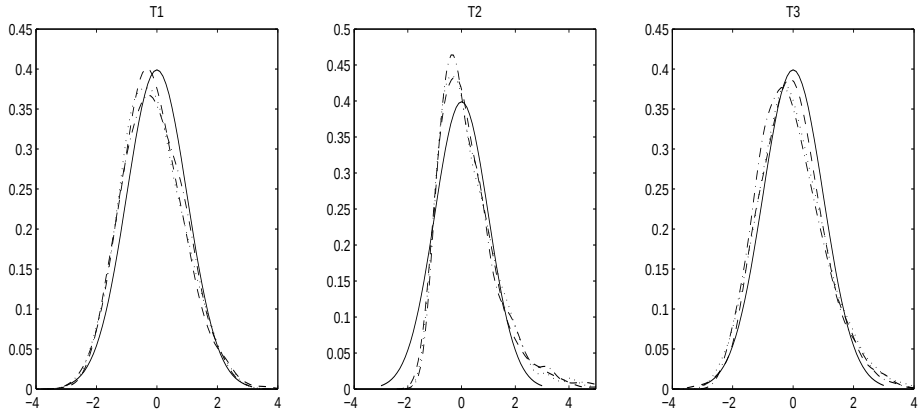

Figure 4.1: Null distributions of test statistics T_1, T_2 and T_3. Solid curve: standard normal; dotted curve: n=400; dash-dot curve: n=800; dashed curve: n=1600 .

families of alternative models as examples:

$$\text{Example1.} \quad a_1(U) = \sin(60U), \quad a_2(U) = a \cdot 4U(1-U) + (1-a).$$

$$\text{Example2.} \quad a_1(U) = \sin(6\pi U), \quad a_2(U) = a \cdot \sin(2\pi U) + (1-a).$$

with the index parameter $a = 0, 0.1, ..., 1$. Obviously, the null hypothesis holds when $a = 0$. Then the functional coefficient $a_2(\cdot)$ gradually departs from the constant as the index parameter a arises to 1.

Under these two families of alternative models, we compute the power functions of the three proposed tests. Here we let the sample size $n = 800$ and the group size $I = 10$. As in the above examples, 1000 simulations are conducted for each test under every alternative model. In T_1, we use a bandwidth of $\frac{1}{2}n^{-2/5}$ and $n^{-1/5}$ is chosen in

T_2.

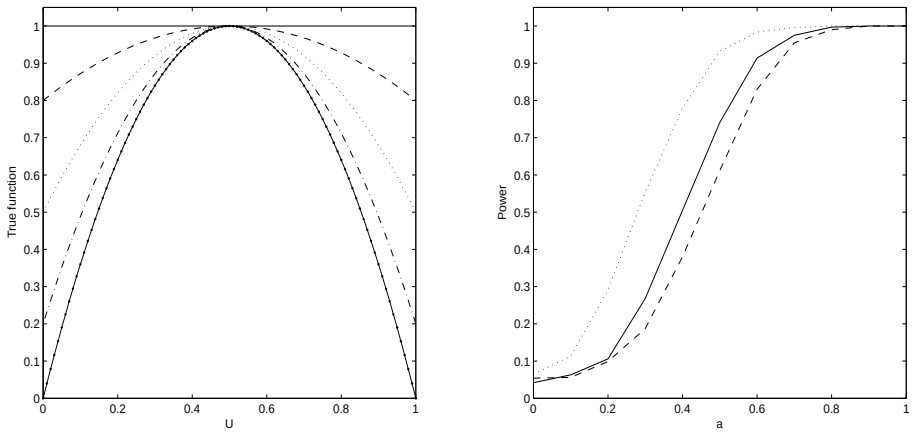

Figure 4.2: Example 1: Left: True function when $a = 0$(solid), $a = 0.2$(dashed), $a = 0.5$(dotted), $a = 0.8$(dash-dotted), $a = 1$(dotted-solid). Right: Power functions for the proposed tests under different alternatives. Solid curve: T_1; dotted curve: T_2; dashed curve: T_3.

The left panel of Figure 4.2 plots the true curves of the functional coefficient $a_2(\cdot)$ in Example 1, ranging from the null hypothesis to the alternatives. The right panel depicts the power functions at 0.05 significance level. It can be seen that in most cases, the powers of the proposed tests present a order of $T_2 > T_1 > T_3$. Moreover, all the three power functions increase to 1 rapidly, indicating the sensitivity for the alternatives of the proposed tests.

Figure 4.3 also shows the true functions of $a_2(\cdot)$ and the power functions of the tests at 0.05 significance level. As expected, the results reveal the proposed test statistics are useful. Still we find the power of T_2 is higher than those of T_1 and T_3. Combining the facts gotten from the size study, it seems that the test T_2 is inclined to reject the null hypothesis.

4.5 Concluding remarks

We have brought up three test statistics for the varying coefficient models based on the local average method. In T_1 and T_2, the problem of the varying coefficient model is transformed into a simple nonparametric one. Hence, classical test approaches

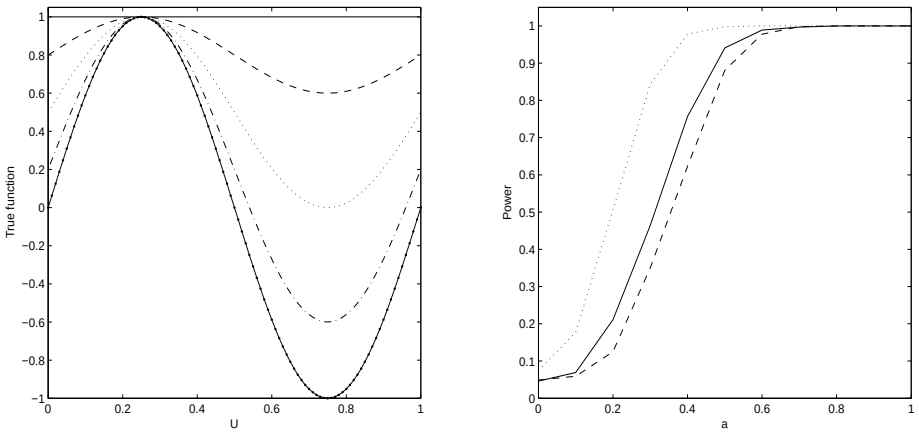

Figure 4.3: Example 2: Left: True function when $a = 0$(solid), $a = 0.2$(dashed), $a = 0.5$(dotted), $a = 0.8$(dash-dotted), $a = 1$(dotted-solid). Right: Power functions for the proposed tests under different alternatives. Solid curve: T_1; dotted curve: T_2; dashed curve: T_3.

can be applied on to this transformed model. Since this transformed model is heteroscedastic, then we have two different test statistics. For T_1, we use Zheng(1996)'s method, which has no requirement on the residual variance. T_2 takes the way of the GLR tests proposed by Fan et al(2001), where we need to know the underlying variance function. Interestingly, the finite-sample simulation results indicate that, T_1 tends to be conservative while T_2 seems to prefer to reject the null hypothesis. The test statistic T_3 is inspired by the GLR test. Instead of estimating the coefficients of the model, we directly estimate the residual variance under the null hypothesis and the alternative hypothesis. Both of the asymptotic results and the simulations of T_3 are quite satisfactory. Moreover, all the three test statistics have proved themselves with large values of power in detecting the alternatives.

The most remarkable contribution of the proposed tests is the lessening of the computation. In most of the existing tests, one need to estimate all the functional coefficients of the varying coefficient model. Even the test problem is only about one coefficient, large effort has to be put into the estimation of other nuisance coefficients. For the proposed tests, T_1 and T_2 only need to deal with one functional coefficient, and T_3 leaves out all the smoothing procedures. Thus, we dramatically improve the computation efficiency. T_1 and T_2 are very suitable for the problems concentrated on

one coefficient. For those tests on several coefficients, it is recommended to use T_3, or to apply T_1 or T_2 multiple times.

Further studies can be conducted on the local average method. In this chapter, we only take the test problem $a_2(\cdot) = c$ as an example. Extensions to the general case $a_2(\cdot) = m(u, \theta)$, where $m(u, \theta)$ is some parametric model, are direct and of the same frame. In addition, the application to other models is also feasible, for example, the additive model. In all, the local average method is a very potential tool and researchers could develop more on this topic.

4.6 Appendix

Proof of Lemma 4.1. Note that the primary point estimators $\mathbf{a} = (\mathbf{a}_1^T, \mathbf{a}_2^T, ..., \mathbf{a}_k^T)^T$ is a result of ordinary least squares from k independent linear regressions. The components $\mathbf{a}_i,\ i = 1, ..., k$ are actually calculated separately. So without losing generality, we will discuss $\mathbf{a}_i$ only.

$$
\begin{aligned}
\hat{\mathbf{a}}_i &= (\hat{a}_1(\bar{U}_{i\cdot}), \hat{a}_2(\bar{U}_{i\cdot}), ..., \hat{a}_p(\bar{U}_{i\cdot}))^T \\
&= (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{Y}_i \\
&= (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \begin{pmatrix} X_{i1}^T \mathbf{a}(U_{i1}) + X_{i1}^T \mathbf{a}(\bar{U}_{i\cdot}) - X_{i1}^T \mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{i1} \\ X_{i2}^T \mathbf{a}(U_{i2}) + X_{i2}^T \mathbf{a}(\bar{U}_{i\cdot}) - X_{i2}^T \mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{i2} \\ \cdots \\ X_{iI}^T \mathbf{a}(U_{iI}) + X_{iI}^T \mathbf{a}(\bar{U}_{i\cdot}) - X_{iI}^T \mathbf{a}(\bar{U}_{i\cdot}) + \epsilon_{iI} \end{pmatrix} \\
&= \mathbf{a}(\bar{U}_{i\cdot}) + (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \boldsymbol{\epsilon}_i + (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \begin{pmatrix} X_{i1}^T (\mathbf{a}(U_{i1}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ X_{i2}^T (\mathbf{a}(U_{i2}) - \mathbf{a}(\bar{U}_{i\cdot})) \\ \cdots \\ X_{iI}^T (\mathbf{a}(U_{iI}) - \mathbf{a}(\bar{U}_{i\cdot})) \end{pmatrix}
\end{aligned}
$$

Now we need to prove that for each $a_l(\cdot), l = 1, ..., p$ and any i, j, $|a_l(U_{ij}) - a_l(\bar{U}_{i\cdot})| = O_p(\frac{\ln n}{n})$. Let $F(\cdot)$ be the cumulative distribution of U, i.e., $F'(u) = f(u)$. Besides, let $\tau = F(U)$, so we can regard τ as a uniformly distributed variable in the interval [0,1]. We denote two consecutive order statistics by $U_{(i+1)}, U_{(i)}$, and $\tau_{(i+1)}, \tau_{(i)}$

are the corresponding uniformly distributed variables. Then we have

$$\begin{aligned}
|a_l(U_{ij}) - a_l(\bar{U}_{i\cdot})| &= |a_l'(\xi_{ij})||U_{ij} - \bar{U}_{i\cdot}| \\
&\leqslant |a_l'(\xi_{ij})| \cdot \frac{I-1}{2} \max |U_{(i+1)} - U_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max |F^{-1}(\tau_{(i+1)}) - F^{-1}(\tau_{(i)})| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max (F^{-1})'(\eta)|\tau_{(i+1)} - \tau_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2} \max \frac{1}{f(u_\eta)}|\tau_{(i+1)} - \tau_{(i)}| \\
&\leqslant \frac{(I-1)|a_l'(\xi_{ij})|}{2\delta} \max |\tau_{(i+1)} - \tau_{(i)}| \\
&= \frac{(I-1)|a_l'(\xi_{ij})|}{2\delta} O_p(\frac{\ln n}{n})
\end{aligned}$$

where ξ_{ij} is between U_{ij} and $\bar{U}_{i\cdot}$, η is between $\tau_{(i+1)}$ and $\tau_{(i)}$, $u_\eta = F^{-1}(\eta)$. The last equation holds by the Theorem 3.1 of Lars Holst(1980). Therefore,

$$\hat{\mathbf{a}}_i = \mathbf{a}(\bar{U}_{i\cdot}) + (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\boldsymbol{\epsilon}_i + \mathbf{O}_p(\frac{\ln n}{n})$$

and

$$\begin{aligned}
&\mathrm{E}(\hat{\mathbf{a}}_i) \to \mathbf{a}(\bar{U}_{i\cdot}), \\
&\mathrm{Var}(\hat{\mathbf{a}}_i) = \mathrm{E}[(\mathbf{X}_i^T\mathbf{X}_i)^{-1}|\bar{U}_{i\cdot}]\sigma^2 = \Gamma(\bar{U}_{i\cdot})\sigma^2,
\end{aligned}$$

as $n \to \infty$.

What's more, since the ordering is no longer needed in the following model checking step, we can naively consider the primary local estimators $(\bar{U}_{i\cdot}, \hat{\mathbf{a}}_i), 1 = 1, ..., k$ are independent and identically distributed. □

Proof of Theorem 4.1. By Lemma 4.1, the test problem can be transformed into

$$\hat{a}_p(\bar{U}_{i\cdot}) = a_p(\bar{U}_{i\cdot}) + \eta_i, \quad i = 1, ..., k$$

where the new error terms $\eta_i, i = 1, ..., k$ are independent and have zero mean. Under null hypothesis, $\hat{a}_p(\bar{U}_{i\cdot}) = c + \eta_i$. Thus $\hat{c} = c + \frac{1}{k}\sum_{i=1}^k \eta_i$ and $\hat{a}_p(\bar{U}_{i\cdot}) - \hat{c} = -\sum_{j\neq i}^k \frac{1}{k}\eta_j +$

$\frac{k-1}{k}\eta_i$. Then

$$\begin{aligned}
T_1 &= \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})(\hat{a}_p(\bar{U}_{i\cdot})-\hat{c})(\hat{a}_p(\bar{U}_{j\cdot})-\hat{c}) \\
&= \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\{-\sum_{s\neq i}^{k}\frac{1}{k}\eta_s+\frac{k-1}{k}\eta_i\}\{-\sum_{t\neq j}^{k}\frac{1}{k}\eta_t+\frac{k-1}{k}\eta_j\} \\
&= S_1+S_2+S_3,
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\sum_{s\neq i}^{k}\frac{1}{k}\eta_s\sum_{t\neq i}^{k}\frac{1}{k}\eta_t \\
S_2 &= \frac{-2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\sum_{s\neq i}^{k}\frac{1}{k}\eta_s\eta_j\frac{k-1}{k} \\
S_3 &= \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\eta_i\eta_j(\frac{k-1}{k})^2.
\end{aligned}$$

It is easy to see that $\mathrm{E}[S_1]=\mathrm{E}[S_2]=O(\frac{1}{k})$, and $\mathrm{E}[S_1^2]=O(\frac{1}{k^2})+O(\frac{1}{k^4h})$, $\mathrm{E}[S_2^2]=O(\frac{1}{k^2})+O(\frac{1}{k^3h})$. Then by Chebyshev inequality, we can have $kh^{1/2}S_1=o_p(1)$ and $kh^{1/2}S_2=o_p(1)$.

Denote $S_3^*=\frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\eta_i\eta_j$, then $S_3=(\frac{k-1}{k})^2S_3^*$. By Lemma 3.3 in Zheng(1996)

$$kh^{1/2}S_3^*\xrightarrow{d}N(0,\Sigma)$$

where $\Sigma=2\int K^2(s)ds\cdot\int\{\mathrm{E}[\eta^2|u]\}^2f(u)d(u)$. Then by Slusky theorem,

$$nh^{1/2}T_1\xrightarrow{d}N(0,\Sigma_1).$$

Next we will show that Σ_1 can be consistently estimated by $\hat{\Sigma}_1$.

$$\begin{aligned}
\hat{\Sigma}_1 =& \frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\hat{e}_i^2\hat{e}_j^2 \\
=& \frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\{-\sum_{s\neq i}^{k}\frac{1}{k}\eta_s+\frac{k-1}{k}\eta_i\}^2\{-\sum_{t\neq j}^{k}\frac{1}{k}\eta_t+\frac{k-1}{k}\eta_j\}^2 \\
=& \frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\{(\sum_{s\neq i}^{k}\frac{1}{k}\eta_s)^2-2\frac{k-1}{k}\eta_i(\sum_{s\neq i}^{k}\frac{1}{k}\eta_s)+\frac{(k-1)^2}{k^2}\eta_i^2\} \\
&\times\{(\sum_{t\neq j}^{k}\frac{1}{k}\eta_t)^2-2\frac{k-1}{k}\eta_j(\sum_{t\neq j}^{k}\frac{1}{k}\eta_t)+\frac{(k-1)^2}{k^2}\eta_j^2\}
\end{aligned}$$

It is obvious that every term that contains $-2\frac{k-1}{k}\eta_i(\sum_{s\neq i}^{k}\frac{1}{k}\eta_s)$ or $-2\frac{k-1}{k}\eta_j(\sum_{t\neq j}^{k}\frac{1}{k}\eta_t)$ has zero mean since η's are independent. Thus, the expansion of the expectation of $\hat{\Sigma}_1$ remains three parts, i.e.,

$$\mathrm{E}[\hat{\Sigma}_1] = \mathrm{E}[S_4 + S_5 + S_6],$$

where

$$\begin{aligned}
S_4 &= \frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})(\sum_{s\neq i}^{k}\frac{1}{k}\eta_s)^2(\sum_{t\neq j}^{k}\frac{1}{k}\eta_t)^2,\\
S_5 &= \frac{4I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})(\sum_{s\neq i}^{k}\frac{1}{k}\eta_s)^2\frac{(k-1)^2}{k^2}\eta_j^2,\\
S_6 &= \frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\frac{(k-1)^4}{k^4}\eta_i^2\eta_j^2.
\end{aligned}$$

It is easy to show that $S_4 = O(\frac{1}{k^2})$, $S_5 = O(\frac{1}{k})$, then

$$\begin{aligned}
\mathrm{E}[\hat{\Sigma}_1] &= \mathrm{E}[S_6] + O(\frac{1}{k})\\
&= \frac{(k-1)^4}{k^4}\mathrm{E}[\frac{2I^2}{k(k-1)}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\frac{1}{h}K^2(\frac{\bar{U}_{i\cdot}-\bar{U}_{j\cdot}}{h})\eta_i^2\eta_j^2] + O(\frac{1}{k})\\
&= \frac{(k-1)^4}{k^4}\Sigma_1 + O(\frac{1}{k})
\end{aligned}$$

So, as $n\to\infty$, $\mathrm{E}[\hat{\Sigma}_1]\to\Sigma_1$. □

Proof of Theorem 4.2. By Lemma 4.3, the test problem is transformed into

$$\hat{a}_p(\bar{U}_{i\cdot}) = a_p(\bar{U}_{i\cdot}) + \eta_i, \quad i = 1, ..., k$$

where the new error terms $\eta_i, i = 1, ..., k$ are independent, $\mathrm{E}[\eta_i] = 0$, $\mathrm{Var}[\eta_i] = \sigma_p^2(x|u)$. Then we apply the GLR test for the problem without unifying the variance. The proof is similar to that of Theorem 5 in Fan et al(2001), except that some variance terms are no longer constant. In addition, we use a simple kernel estimator for the alternative while in Fan et al(2001), local linear estimator is used. However, these two kinds of estimators share the same asymptotic properties in our case. So they are equivalent.

$$T_2 = \frac{n}{2I}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_0} \approx \frac{n}{2I}\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}$$

where $\text{RSS}_0 = \sum_{i=1}^k (\hat{a}_p(\bar{U}_{i\cdot}) - \hat{c})^2$, $\text{RSS}_1 = \sum_{i=1}^k (\hat{a}_p(\bar{U}_{i\cdot}) - \tilde{m}_h(\bar{U}_{i\cdot}))^2$. $\tilde{m}_h(\bar{U}_{i\cdot})$ is the estimator under the alternative hypothesis. Here we use a kernel estimator with $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$. Thus $\tilde{m}_h(\bar{U}_{i\cdot}) = \frac{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\hat{a}_p(\bar{U}_{j\cdot})}{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})}$. Then under null the hypothesis, we have

$$\begin{aligned}
\frac{1}{k}\text{RSS}_1 &= \frac{1}{k}\sum_{i=1}^k (c + \eta_i - \frac{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})(c+\eta_j)}{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})})^2 \\
&= \frac{1}{k}\sum_{i=1}^k (\eta_i - \frac{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_j}{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})})^2 \\
&= \frac{1}{k}\sum_{i=1}^k \eta_i^2 + \frac{1}{k}\sum_{i=1}^k (\frac{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_j}{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})})^2 - \frac{2}{k}\sum_{i=1}^k \eta_i \frac{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_j}{\sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})} \\
&\to \int \sigma_p^2(x) f(u) du + W_1 - 2W_2
\end{aligned}$$

Since

$$\begin{aligned}
W_1 &= \frac{1}{k^3}\sum_{i=1}^k \sum_{j=1}^k \sum_{j'=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot}) K_h(\bar{U}_{i\cdot} - \bar{U}_{j'\cdot})\eta_j \eta_{j'} \frac{1}{f_u^2(\bar{U}_{i\cdot})} + o_p(1) \\
&= \frac{1}{k^2}\sum_{j=1}^k \sum_{j'=1}^k \int K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot}) K_h(\bar{U}_{i\cdot} - \bar{U}_{j'\cdot}) d\bar{U}_{i\cdot} \eta_j \eta_{j'} + o_p(1) \\
&= O_p(\frac{1}{kh}) + O_p(\frac{1}{kh^{1/2}})
\end{aligned}$$

and

$$\begin{aligned}
W_2 &= \frac{1}{k^2}\sum_{i=1}^k \sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_i \eta_j \frac{1}{f_u(\bar{U}_{i\cdot})} + o_p(1) \\
&= \frac{1}{2}\frac{1}{k^2}\sum_{i=1}^k \sum_{j=1}^k K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_i \eta_j (\frac{1}{f_u(\bar{U}_{i\cdot})} + \frac{1}{f_u(\bar{U}_{j\cdot})}) + o_p(1) \\
&= O_p(\frac{1}{kh}) + O_p(\frac{1}{kh^{1/2}})
\end{aligned}$$

So we have

$$\frac{1}{k}\text{RSS}_1 = \int \sigma_p^2(u) f(u) du + o_p(1)$$

On the other hand,

$$\begin{aligned}\frac{1}{k}(\mathrm{RRS}_0 - \mathrm{RRS}_1) &= \frac{1}{k}\sum_{i=1}^{k}(\hat{a}_p(\bar{U}_{i\cdot}) - \hat{c})^2 - \frac{1}{k}\sum_{i=1}^{k}(\hat{a}_p(\bar{U}_{i\cdot}) - \tilde{m}_h(\bar{U}_{i\cdot}))^2 \\ &= \frac{1}{k}\sum_{i=1}^{k}(\eta_i - \bar{\eta})^2 - \frac{1}{k}\sum_{i=1}^{k}(\eta_i - (\tilde{m}_h(\bar{U}_{i\cdot}) - c))^2 \\ &= 2\frac{1}{k}\sum_{i=1}^{k}\eta_i(\tilde{m}_h(\bar{U}_{i\cdot}) - c) - \frac{1}{k}\sum_{i=1}^{k}(\tilde{m}_h(\bar{U}_{i\cdot}) - c)^2 - \frac{1}{k}\sum_{i=1}^{k}\bar{\eta}^2 \\ &= W_3 - W_4 + O_p(\frac{1}{n})\end{aligned}$$

Since

$$\begin{aligned}W_3 &= 2\frac{1}{k}\sum_{i=1}^{k}\eta_i\frac{\sum_{j=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_j}{\sum_{j=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})} \\ &= \frac{2}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_i\eta_j\frac{1}{f_u(\bar{U}_{i\cdot})}(1 + o_p(1)) \\ &= \frac{1}{k^2}\sum_{i=1}^{k}\frac{1}{h}K(0)\eta_i^2\frac{2}{f_u(\bar{U}_{i\cdot})} + \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j\neq i}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_i\eta_j\frac{2}{f_u(\bar{U}_{i\cdot})} + o_p(1) \\ &= W_{31} + W_{32} + o_p(1)\end{aligned}$$

and

$$\begin{aligned}W_4 =& \frac{1}{k}\sum_{i=1}^{k}(\frac{\sum_{j=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})\eta_j}{\sum_{j=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})})^2 \\ =& \frac{1}{k^3}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{j'=1}^{k}K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})K_h(\bar{U}_{i\cdot} - \bar{U}_{j'\cdot})\eta_j\eta_{j'}\frac{1}{f_u^2(\bar{U}_{i\cdot})}(1 + o_p(1)) \\ =& \frac{1}{k^2}\sum_{j=1}^{k}\sum_{j'=1}^{k}\{\mathrm{E}[K_h(\bar{U}_{i\cdot} - \bar{U}_{j\cdot})K_h(\bar{U}_{i\cdot} - \bar{U}_{j'\cdot})\frac{1}{f_u^2(\bar{U}_{i\cdot})}]\}\eta_j\eta_{j'}(1 + o_p(1)) \\ =& \frac{1}{k^2}\sum_{j=1}^{k}\sum_{j'=1}^{k}\frac{1}{h}\int K(v)K(v + \frac{\bar{U}_{j\cdot} - \bar{U}_{j'\cdot}}{h})dv f_u^{-1}(\bar{U}_{j\cdot})\eta_j\eta_{j'}(1 + o_p(1)) \\ =& \frac{1}{k^2}\sum_{j=1}^{k}\frac{1}{h}\int K^2(v)dv f_u^{-1}(\bar{U}_{j\cdot})\eta_j^2 \\ &+ \frac{1}{k^2}\sum_{j=1}^{k}\sum_{j'\neq j}^{k}\frac{1}{h}\int K(v)K(v + \frac{\bar{U}_{j\cdot} - \bar{U}_{j'\cdot}}{h})dv f_u^{-1}(\bar{U}_{j\cdot})\eta_j\eta_{j'} + o_p(1) \\ =& W_{41} + W_{42} + o_p(1)\end{aligned}$$

then,

$$\frac{1}{k}(\mathrm{RRS}_0 - \mathrm{RRS}_1) = (W_{31} - W_{41}) + (W_{32} - W_{42}) + o_p(1).$$

$$W_{31} - W_{41} = \frac{1}{k^2}\sum_{j=1}^{k}\frac{1}{h}\{2K(0) - \int K^2(v)dv\}f_u^{-1}(\bar{U}_{j\cdot})\eta_j^2$$
$$\to \frac{1}{kh}\{2K(0) - \int K^2(v)dv\}\int \sigma_p^2(u)du$$

$$W_{32} - W_{42} = \frac{1}{k^2 h}\sum_{j'\neq j}\eta_j\eta_{j'}\{2K(\frac{\bar{U}_{j\cdot} - \bar{U}_{j'\cdot}}{h}) - \int K(v)K(v + \frac{\bar{U}_{j\cdot} - \bar{U}_{j'\cdot}}{h})dv\}f_u^{-1}(\bar{U}_{j\cdot})$$

Denote $W_5 = W_{32} - W_{42}$, then by Proposition 3.2 in Peter de Jong(1987), we have $W_5 \to N(0, \mathrm{Var}[W_5])$, with

$$\mathrm{Var}[W_5] = \frac{2}{k^2 h}\int\{2K(s) - \int K(v)K(v+s)dv\}^2 ds \int[\sigma_p^2(u)]^2 du$$

The conditions checking for Proposition 3.2 in Peter de Jong(1987) is almost the same with that in the proof of theorem 5 in Fan et al(2001), so we omit the process here. □

Proof of Theorem 4.3. By Theorem 3.3 in Chapter 3,

$$\frac{\widehat{\mathrm{RSS}}_1}{k(I-p)} - \sigma^2 = O_p(\frac{1}{\sqrt{n}}).$$

Then from the definition of the test, we have

$$T_3 = \frac{n}{2}\frac{\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1}{\widehat{\mathrm{RSS}}_1} = \frac{n}{2k(I-p)}\cdot\frac{1}{\sigma^2(1+o_p(1))}(\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1)$$

As in the computation, $U_i's$ are ranked in ascending order, denoted as $\{U_{(i)}\}$. Let S_t be the index set of $\{U_{(I(t-1)+1)}, ..., U_{(tI)}\}$, and $\mathbf{1}_{ti}$ is $\mathbf{1}\{i \in S_t\}, t = 1, ..., k$. In addition, for convenience, we denote $\widehat{\boldsymbol{\beta}}$ as the local average estimator for the first $(p-1)$ functional coefficients under the null hypothesis, i.e. $\widehat{\boldsymbol{\beta}}_t = (\hat{a}_1(\bar{U}_{t\cdot}), \hat{a}_2(\bar{U}_{t\cdot}), ..., \hat{a}_{p-1}(\bar{U}_{t\cdot}))^T, t = 1, ..., k$, and $\widetilde{\boldsymbol{\beta}}$ is for the alternative. $\hat{c}$ is the constant estimator under null the hypothesis and $\tilde{\gamma}_t = \tilde{a}_p(\bar{U}_{t\cdot}), t = 1..., k$ is the functional estimators under the alternative

hypothesis. Then $\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1$ can be expanded as

$$\begin{aligned}
\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1 =& \sum_{t=1}^{k}\sum_{i=1}^{n}(y_i - \widehat{\boldsymbol{\beta}}_t^T \ddot{X}_i - \hat{c}x_{ip})^2 \mathbf{1}_{ti} - \sum_{t=1}^{k}\sum_{i=1}^{n}(y_i - \widetilde{\boldsymbol{\beta}}_t^T \ddot{X}_i - \tilde{\gamma}_t x_{ip})^2 \mathbf{1}_{ti} \\
=& \sum_{t=1}^{k}\sum_{i=1}^{n}\{(y_i - \widehat{\boldsymbol{\beta}}_t^T \ddot{X}_i - \hat{c}x_{ip})^2 - (y_i - \widehat{\boldsymbol{\beta}}_{t0}^T \ddot{X}_i - cx_{ip})^2\}\mathbf{1}_{ti} \\
&+ \sum_{t=1}^{k}\sum_{i=1}^{n}\{(y_i - \widehat{\boldsymbol{\beta}}_{t0}^T \ddot{X}_i - cx_{ip})^2 - (y_i - \widetilde{\boldsymbol{\beta}}_t^T \ddot{X}_i - \tilde{\gamma}_t x_{ip})^2\}\mathbf{1}_{ti} \\
\equiv& \mathrm{DRSS}_1 + \mathrm{DRSS}_2
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{t0}$ is the estimator for $(a_1(\bar{U}_{t\cdot}), ..., a_{p-1}(\bar{U}_{t\cdot}))$ if the constant coefficient c is known.

$$\begin{aligned}
\mathrm{DRSS}_1 =& \sum_{t=1}^{k}\sum_{i=1}^{n}\{\ddot{X}_i^T(\widehat{\boldsymbol{\beta}}_{t0} - \widehat{\boldsymbol{\beta}}_t) + x_{ip}(c - \hat{c})\}\{2y_i - \widehat{\boldsymbol{\beta}}_t^T \ddot{X}_i - \widehat{\boldsymbol{\beta}}_{t0}^T \ddot{X}_i - \hat{c}x_{ip} - cx_{ip}\}\mathbf{1}_{ti} \\
=& \sum_{t=1}^{k}\sum_{i=1}^{n}\{\ddot{X}_i^T(\widehat{\boldsymbol{\beta}}_{t0} - \widehat{\boldsymbol{\beta}}_t) + x_{ip}(c - \hat{c})\} \\
&\times \{2\ddot{\mathbf{a}}^T(U_i)\ddot{X} + cx_{ip} - \widehat{\boldsymbol{\beta}}_t^T \ddot{X}_i - \widehat{\boldsymbol{\beta}}_{t0}^T \ddot{X}_i - \hat{c}x_{ip}\}\mathbf{1}_{ti} \\
=& \sum_{t=1}^{k}\sum_{i=1}^{n}\{\ddot{X}_i^T(\widehat{\boldsymbol{\beta}}_{t0} - \widehat{\boldsymbol{\beta}}_t) + x_{ip}(c - \hat{c})\} \\
&\times \{\ddot{X}_i^T(\ddot{a}(U_i) - \widehat{\boldsymbol{\beta}}_t) + \ddot{X}_i^T(\ddot{a}(U_i) - \widehat{\boldsymbol{\beta}}_{t0}) + x_{ip}(c - \hat{c})\}\mathbf{1}_{ti}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathrm{DRSS}_2 =& \sum_{t=1}^{k}\sum_{i=1}^{n}\{\ddot{X}_i^T(\widetilde{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}}_{t0}) + x_{ip}(\tilde{\gamma}_t - c)\} \\
&\times \{\ddot{X}_i^T(\ddot{a}(U_i) - \widetilde{\boldsymbol{\beta}}_t) + \ddot{X}_i^T(\ddot{\mathbf{a}}(U_i) - \widehat{\boldsymbol{\beta}}_{t0}) + x_{ip}(c - \tilde{\gamma}_t)\}\mathbf{1}_{ti}
\end{aligned}$$

Since

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_t =& \{\sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T\}^{-1}\{\sum_{i\in S_t}\ddot{X}_i y_i - \hat{c}\sum_{i\in S_t}\ddot{X}_i x_{ip}\} \\
=& \{\sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T\}^{-1}\{\sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T\ddot{\mathbf{a}}(U_i) + (c - \hat{c})\sum_{i\in S_t}\ddot{X}_i x_{ip} + \sum_{i\in S_t}\ddot{X}_i\epsilon_i\} \\
=& \{\sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T\}^{-1} \\
&\times \{\sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T(\ddot{\mathbf{a}}(U_i) - \ddot{\mathbf{a}}(\bar{U}_{t\cdot})) + \sum_{i\in S_t}\ddot{X}_i\ddot{X}_i^T\ddot{\mathbf{a}}(\bar{U}_{t\cdot}) + (c - \hat{c})\sum_{i\in S_t}\ddot{X}_i x_{ip} + \sum_{i\in S_t}\ddot{X}_i\epsilon_i\}
\end{aligned}$$

Remind of the notations

$$A_t = \sum_{i \in S_t} \ddot{X}_i \ddot{X}_i^T \qquad B_t = \sum_{i \in S_t} x_{ip}^2 \qquad C_t = \sum_{i \in S_t} \ddot{X}_i x_{ip}$$
$$D_{t1} = \sum_{i \in S_t} \ddot{X}_i \epsilon_i \qquad D_{t2} = \sum_{i \in S_t} x_{ip} \epsilon_i$$
$$M_t = B_t - C_t^T A_t^{-1} C_t \qquad m_{ti} = C_t^T A_t^{-1} \ddot{X}_i - x_{ip}$$

Then together with the Lemma 4.1, the estimator $\widehat{\boldsymbol{\beta}}_t$ can be written as

$$\widehat{\boldsymbol{\beta}}_t = \ddot{\mathbf{a}}(\bar{U}_{t\cdot}) + A_t^{-1} D_{t1} + A_t^{-1} C_t (c - \hat{c}) + O_p(\frac{\log n}{n}).$$

Likewise, we have

$$\widetilde{\boldsymbol{\beta}}_t = \ddot{\mathbf{a}}(\bar{U}_{t\cdot}) + A_t^{-1} D_{t1} + A_t^{-1} C_t (c - \tilde{\gamma}_t) + O_p(\frac{\log n}{n})$$
$$\widehat{\boldsymbol{\beta}}_{t0} = \ddot{\mathbf{a}}(\bar{U}_{t\cdot}) + A_t^{-1} D_{t1} + O_p(\frac{\log n}{n})$$

In addition, from Lemma 4.1 we obtain

$$\begin{aligned}
\tilde{\gamma}_t &= c + e_{p,p}^T \{\sum_{i \in S_t} X_i X_i^T\}^{-1} \{\sum_{i \in S_t} X_i \epsilon_i\} + O_p(\frac{\log n}{n}) \\
&= c + e_{p,p}^T \begin{bmatrix} A_t & C_t \\ C_t^T & B_t \end{bmatrix}^{-1} \begin{bmatrix} D_{t1} \\ D_{t2} \end{bmatrix} + O_p(\frac{\log n}{n}) \\
&= c + (B_t - C_t^T A_t^{-1} C_t)^{-1} \begin{bmatrix} -C_t^T A_t^{-1} & 1 \end{bmatrix} \begin{bmatrix} D_{t1} \\ D_{t2} \end{bmatrix} + O_p(\frac{\log n}{n}) \\
&= c - M_t^{-1} \sum_{i \in S_t} m_{ti} \epsilon_i + O_p(\frac{\log n}{n})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{DRSS}_1 =& \sum_{t=1}^{k}\sum_{i=1}^{n} m_{ti}(\hat{c}-c)\times\{m_{ti}(\hat{c}-c) - 2\ddot{X}_i^T A_t^{-1} D_{t1} + 2\epsilon_i\}(1+o_p(1))\mathbf{1}_{ti} \\
=&(\hat{c}-c)^2 \sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}^2 - 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\ddot{X}_i^T A_t^{-1} D_{t1} \\
&+ 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\epsilon_i + o_p(1) \\
=&(\hat{c}-c)^2 \sum_{t=1}^{k} M_t - 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t}(C_t^T A_t^{-1}\ddot{X}_i - x_{ip})\ddot{X}_i^T A_t^{-1} D_{t1} \\
&+ 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\epsilon_i + o_p(1) \\
=&(\hat{c}-c)^2 \sum_{t=1}^{k} M_t - 2(\hat{c}-c)\sum_{t=1}^{k}(C_t^T - C_t^T)A_t^{-1} D_{t1} \\
&+ 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\epsilon_i + o_p(1) \\
=&(\hat{c}-c)^2 \sum_{t=1}^{k} M_t + 2(\hat{c}-c)\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\epsilon_i + o_p(1)
\end{aligned}$$

Since $\hat{c}-c = O_p(\frac{1}{\sqrt{n}})$, $\sum_{t=1}^{k} M_t = O_p(n)$ and by Law of Large Numbers,

$$\frac{1}{n}\sum_{t=1}^{k}\sum_{i\in S_t} m_{ti}\epsilon_i = O_p(\frac{1}{\sqrt{n}}),$$

thus $\text{DRSS}_1 = O_p(1)$.

Next we consider DRSS_2

$$\begin{aligned}
\mathrm{DRSS}_2 =& \sum_{t=1}^{k}\sum_{i=1}^{n} -m_{ti}(\tilde{\gamma}_t - c) \times \{m_{ti}(\tilde{\gamma}_t - c) - 2\ddot{X}_i^T A_t^{-1} D_{t1} + 2\epsilon_i\}(1 + o_p(1))\mathbf{1}_{ti} \\
=& -\sum_{t=1}^{k}\sum_{i\in S_t}(\tilde{\gamma}_t - c)^2 m_{ti}^2 + 2\sum_{t=1}^{k}(\tilde{\gamma}_t - c)\sum_{i\in S_t} m_{ti}\ddot{X}_i^T A_t^{-1} D_{t1} \\
& - 2\sum_{t=1}^{k}\sum_{i\in S_t}(\tilde{\gamma}_t - c)m_{ti}\epsilon_i + o_p(1) \\
=& -\sum_{t=1}^{k} M_t(\tilde{\gamma}_t - c)^2 - 2\sum_{t=1}^{k}\sum_{i\in S_t}(\tilde{\gamma}_t - c)m_{ti}\epsilon_i + o_p(1) \\
=& -\sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t}\sum_{j\in S_t} m_{ti}m_{tj}\epsilon_i\epsilon_j + 2\sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t}\sum_{j\in S_t} m_{ti}m_{tj}\epsilon_i\epsilon_j + o_p(1) \\
=& \sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t}\sum_{j\in S_t} m_{ti}m_{tj}\epsilon_i\epsilon_j + o_p(1) \\
=& \sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t} m_{ti}^2\epsilon_i^2 + \sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t}\sum_{\substack{j\in S_t \\ j\neq i}} m_{ti}m_{tj}\epsilon_i\epsilon_j + o_p(1) \\
\equiv& P_1 + P_2 + o_p(1)
\end{aligned}$$

Now we will show that

$$P_1 - k\sigma^2 \to N(0, v_1), \qquad P_2 \to N(0, v_2),$$

where $v_1 = (\mu_4 - \sigma^4)\sum_{t=1}^{k} M_t^{-2}\sum_{i\in S_t} m_{ti}^4$, $v_2 = 2k\sigma^4 - 2\sigma^4\sum_{t=1}^{k} M_t^{-2}\sum_{i\in S_t} m_{ti}^4$.

Since

$$P_1 - k\sigma^2 = \sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t} m_{ti}^2\epsilon_i^2 - k\sigma^2 = \sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t} m_{ti}^2(\epsilon_i^2 - \sigma^2),$$

let $Z_t = M_t^{-1}\sum_{i\in S_t} m_{ti}^2(\epsilon_i^2 - \sigma^2)$, then Z_t is independent random variable and

$$\mathrm{E}[Z_t] = 0, \qquad \mathrm{Var}[Z_t] = (\mu_4 - \sigma^4)M_t^{-2}\sum_{i\in S_t} m_{ti}^4$$

Because $\frac{1}{2} \le M_t^{-2}\sum_{i\in S_t} m_{ti}^4 \le 1$ for any t, and $v_1 = \sum_{t=1}^{k}\mathrm{Var}[Z_t] = O(k)$, it is easy to prove that the Lindeberg's condition is satisfied for Z_t. Therefore, by central limit theorem, $\frac{\sum_{t=1}^{k} Z_t}{v_1} \to N(0,1)$, i.e.,

$$P_1 - k\sigma^2 \to N(0, v_1)$$

with $v_1 = (\mu_4 - \sigma^4)\sum_{t=1}^{k} M_t^{-2}\sum_{i\in S_t} m_{ti}^4$.

The proof of the normality of P_2 is an application of Proposition 3.2 in Peter de Jong (1987). Denote

$$\Pi_{ij} = \sum_{t=1}^{k} \mathbf{1}_{ti}\mathbf{1}_{tj}, i.e., \Pi_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are in the same group} \\ 0 & \text{otherwise} \end{cases}$$

Define $W_{ij} = 2M_t^{-1}m_{ti}m_{tj}\Pi_{ij}\epsilon_i\epsilon_j$, then $P_2 = \sum_{i<j} W_{ij}$. Then by the Proposition 3.2 in Peter de Jong(1987), we need to check the following conditions

1. $\mathrm{E}[W_{ij}|\epsilon_i] = 0$ a.s. for all $i, j \le n$.
2. $\mathrm{Var}[P_2] \to v_2$.
3. $G_{\mathrm{I}}, G_{\mathrm{II}}, G_{\mathrm{IV}}$ is of smaller order than v_2^2.

where

$$G_{\mathrm{I}} = \sum_{i<j} \mathrm{E}[W_{ij}^4]$$

$$G_{\mathrm{II}} = \sum_{i<j<m} \mathrm{E}[W_{ij}^2W_{im}^2 + W_{ji}^2W_{jm}^2 + W_{mi}^2W_{mj}^2]$$

$$G_{\mathrm{IV}} = \sum_{i<j<m<l} \mathrm{E}[W_{ij}W_{im}W_{lj}W_{lm} + W_{ij}W_{il}W_{mj}W_{ml} + W_{im}W_{il}W_{jm}W_{jl}]$$

Condition 1 is obvious by the definition. To prove condition 2, note that $\mathrm{E}[P_2] = 0$, then

$$\begin{aligned}
\mathrm{Var}[P_2] &= \mathrm{E}[P_2^2] \\
&= \mathrm{E}[(\sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t}\sum_{\substack{j\in S_t \\ j\neq i}} m_{ti}m_{tj}\epsilon_i\epsilon_j)^2] \\
&= \sum_{t=1}^{k} M_t^{-1}\mathrm{E}[(\sum_{i\in S_t}\sum_{\substack{j\in S_t \\ j\neq i}} m_{ti}m_{tj}\epsilon_i\epsilon_j)^2] \\
&= 2\sigma^4\sum_{t=1}^{k} M_t^{-1}\sum_{i\in S_t} m_{ti}^2\sum_{\substack{j\in S_t \\ j\neq i}} m_{tj}^2 \\
&= 2\sigma^4\sum_{t=1}^{k} M_t^{-2}(\sum_{i\in S_t} m_{ti}^2)^2 - 2\sigma^4\sum_{t=1}^{k} M_t^{-2}\sum_{i\in S_t} m_{ti}^4 \\
&= 2k\sigma^4 - 2\sigma^4\sum_{t=1}^{k} M_t^{-2}\sum_{i\in S_t} m_{ti}^4
\end{aligned}$$

So condition 2 is satisfied and we obtain $v_2^2 = O(k^2)$. For condition 3,

$$\begin{aligned} G_{\mathrm{I}} &= \sum_{i<j} \mathrm{E}[W_{ij}^4] \\ &= \sum_{i<j} \mathrm{E}[(2M_t^{-1} m_{ti} m_{tj} \Pi_{ij} \epsilon_i \epsilon_j)^4] \\ &= 8\mu_4^2 \sum_{t=1}^{k} M_t^{-4} \sum_{i \in S_t} m_{ti}^4 \sum_{\substack{j \in S_t \\ j \neq i}} m_{tj}^4 \\ &= O(k) \end{aligned}$$

Similarly, we can prove that $G_{\mathrm{II}} = O(k)$, $G_{\mathrm{IV}} = O(k)$.

Therefore, combining the asymptotic results of P_1 and P_2, we have

$$\mathrm{DRSS}_2 - k\sigma^2 \to N(0, v_3)$$

where $v_3 = v_1 + v_2 + 2\mathrm{Cov}(P_1, P_2)$. It is easy to prove that $\mathrm{Cov}(P_1, P_2) = 0$. Then

$$v_3 = (\mu_4 - 3\sigma^4) \sum_{t=1}^{k} \mathrm{E}[M_t^{-2} \sum_{i \in S_t} m_{ti}^4] + 2k\sigma^4 = (\mu_4 - 3\sigma^4)\Psi_n + 2k\sigma^4$$

Since v_3 is of order $O(k)$, then we have

$$v_3^{-1/2}(\widehat{\mathrm{RSS}}_0 - \widehat{\mathrm{RSS}}_1 - k\sigma^2) \to N(0, 1)$$

Let $\sigma_3^2 = (\frac{\mu_4}{\sigma^4} - 3)\Psi_n + 2k$, then

$$\frac{2(I-p)}{I} \sigma_3^{-1} (T_3 - \frac{n}{2(I-p)}) \to N(0, 1)$$

□

Bibliography

[1] Ahmad, I., Leelahanon, S., and Li, Q. Efficient estimation of a semiparametric partially linear varying coefficient model. *Annals of Statistics* (2005), 258–283.

[2] Albert, S., Wayne, W., and Zappe, C. Data analysis and decision making with microsoft excel. *Hampshire, United Kingdom: South-Western College Publishing Co. Ltd* (2008).

[3] Bickel, P. J., Klaassen, C., Ritov, Y., and Wellner, J. Efficient and adaptive inference in semiparametric models, 1993.

[4] Brown, L. D., Levine, M., et al. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics 35*, 5 (2007), 2219–2232.

[5] Brown, L. D., Levine, M., and Wang, L. A semiparametric multivariate partially linear model: a difference approach. *Journal of Statistical Planning and Inference 178* (2016), 99–111.

[6] Buckley, M., Eagleson, G., and Silverman, B. The estimation of residual variance in nonparametric regression. *Biometrika* (1988), 189–199.

[7] Buja, A., Hastie, T., and Tibshirani, R. Linear smoothers and additive models. *The Annals of Statistics* (1989), 453–510.

[8] Cai, T. T., Levine, M., and Wang, L. Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis 100*, 1 (2009), 126–136.

[9] Cai, Z., Fan, J., and Li, R. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association 95*, 451 (2000), 888–902.

[10] Cai, Z., Fan, J., and Yao, Q. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*, 451 (2000), 941–956.

[11] Cui, X., Lu, Y., and Peng, H. Estimation of partially linear regression model under partial consistency property. *arXiv preprint arXiv:1401.2163* (2014).

[12] Dette, H., Munk, A., and Wagner, T. Estimating the variance in nonparametric regressionwhat is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*, 4 (1998), 751–764.

[13] Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association 81*, 394 (1986), 310–320.

[14] Fan, J., and Gijbels, I. *Local polynomial modelling and its applications*, vol. 66. Chapman and Hall, London, 1996.

[15] Fan, J., Guo, S., and Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 1 (2012), 37–65.

[16] Fan, J., Huang, T., et al. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*, 6 (2005), 1031–1057.

[17] Fan, J., and Jiang, J. Nonparametric inferences for additive models. *Journal of the American Statistical Association 100*, 471 (2005), 890–907.

[18] Fan, J., and Jiang, J. Nonparametric inference with generalized likelihood ratio tests. *Test 16*, 3 (2007), 409–444.

[19] Fan, J., Peng, H., et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*, 3 (2004), 928–961.

[20] Fan, J., Peng, H., and Huang, T. Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *Journal of the American Statistical Association 100*, 471 (2005), 781–796.

[21] Fan, J., and Yao, Q. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* (1998), 645–660.

[22] Fan, J., Zhang, C., and Zhang, J. Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics* (2001), 153–193.

[23] Fan, J., and Zhang, J.-T. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*, 2 (2000), 303–322.

[24] Fan, J., and Zhang, W. Statistical estimation in varying coefficient models. *Annals of Statistics* (1999), 1491–1518.

[25] Fan, J., and Zhang, W. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics 27*, 4 (2000), 715–731.

[26] Friedman, J. H., and Stuetzle, W. Projection pursuit regression. *Journal of the American statistical Association 76*, 376 (1981), 817–823.

[27] Gasser, T., Sroka, L., and Jennen-Steinmetz, C. Residual variance and residual pattern in nonlinear regression. *Biometrika* (1986), 625–633.

[28] Groves, T., and Rothenberg, T. A note on the expected value of an inverse matrix. *Biometrika 56*, 3 (1969), 690–691.

[29] Hall, P., and Marron, J. S. On variance estimation in nonparametric regression. *Biometrika* (1990), 415–419.

[30] Hardle, W., and Mammen, E. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* (1993), 1926–1947.

[31] Härdle, W., and Stoker, T. M. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association 84*, 408 (1989), 986–995.

[32] Hastie, T., and Tibshirani, R. Generalized additive models. *Statistical science* (1986), 297–310.

[33] Hastie, T., and Tibshirani, R. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* (1993), 757–796.

[34] Holst, L. On the lengths of the pieces of a stick broken at random. *Journal of Applied Probability* (1980), 623–634.

[35] Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*, 4 (1998), 809–822.

[36] Huang, J. Z., and Shen, H. Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian journal of statistics 31*, 4 (2004), 515–534.

[37] Huang, J. Z., Wu, C. O., and Zhou, L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*, 1 (2002), 111–128.

[38] Huang, J. Z., Wu, C. O., and Zhou, L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* (2004), 763–788.

[39] Jong, P. A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields 75*, 2 (1987), 261–277.

[40] Klipple, K., and Eubank, R. Difference-based variance estimators for partially linear models. *Festschrift in honor of Distinguished Professor Mir Masoom Ali on the occasion of his retirement* (2007), 313–323.

[41] Müller, U. U., Schick, A., and Wefelmeyer, W. Estimating the error variance in nonparametric regression by a covariate-matched u-statistic. *Statistics: A Journal of Theoretical and Applied Statistics 37*, 3 (2003), 179–188.

[42] Neyman, J., and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society* (1948), 1–32.

[43] Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. Varying coefficient regression models: a review and new developments. *International Statistical Review 83*, 1 (2015), 36–64.

[44] Park, C. G., Kim, I., and Lee, Y.-S. Error variance estimation via least squares for small sample nonparametric regression. *Journal of Statistical Planning and Inference 142*, 8 (2012), 2369–2385.

[45] Rice, J. Bandwidth choice for nonparametric regression. *The Annals of Statistics* (1984), 1215–1230.

[46] Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* (1988), 931–954.

[47] Tong, T., and Wang, Y. Estimating residual variance in nonparametric regression using least squares. *Biometrika* (2005), 821–830.

[48] Van der Vaart, A. W. *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[49] Wang, L., Brown, L. D., Cai, T. T., et al. A difference based approach to the semiparametric partial linear model. *Electronic Journal of Statistics 5* (2011), 619–641.

[50] Wang, L., Brown, L. D., Cai, T. T., and Levine, M. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics* (2008), 646–664.

[51] Wu, C. O., Chiang, C.-T., and Hoover, D. R. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association 93*, 444 (1998), 1388–1402.

[52] Xia, Y., Zhang, W., and Tong, H. Efficient estimation for semivarying-coefficient models. *Biometrika 91*, 3 (2004), 661–681.

[53] Zhang, T., Wu, W. B., et al. Inference of time-varying regression models. *The Annals of Statistics 40*, 3 (2012), 1376–1402.

[54] Zhang, W., and Lee, S.-Y. Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis 74*, 1 (2000), 116–134.

[55] Zhang, W., Lee, S.-Y., and Song, X. Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis 82*, 1 (2002), 166–188.

[56] Zheng, J. X. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics 75*, 2 (1996), 263–289.

CURRICULUM VITAE

Academic qualifications of the thesis author, Miss ZHAO Jingxin:

- Received the degree of Bachelor of Science(Honors) in Statistics and Operation Research from Hong Kong Baptist University, 2013.

August 2017