STATISTICAL ANALYSIS OF SPATIAL DYNAMIC PATTERN IN SPATIAL DATA ANALYSIS

HONGJIA YAN

(B.Sc. and M.Sc.)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF YORK

August 2013

Abstract

In this thesis, inspired by the Boston House Price data, we propose a semiparametric spatial dynamic model, that extends the ordinary spatial autoregressive models to accommodate the effects of some covariates associated with the House price. A profile likelihood-based estimation procedure is proposed and the asymptotic normality of the proposed estimators are derived. We also investigate the connection between $cross-validation$ method and AIC/BIC methods in the semiparametric family. In the proposed model, it is easier to apply the AIC/BIC method than the $cross-validation$ method. We illustrate how to identify the parametric/nonparametric components in the proposed semiparametric model. We also show how many unknown parameters an unknown bivariate function amounts to, and propose an AIC/BIC nonparametric model selection. Simulation studies are conducted to examine the performance of the proposed methods, and their results show that the methods work very well. Finally, we apply the proposed methods to analyze the Boston House Price data, which lead to some interesting findings.Although, the proposed model and methodology are stimulated by the Boston House Price data, they could be widely used in many other scientific problems.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to give my sincere thanks to my supervisor, Professor Wenyang Zhang, who has patiently supervised me over the last three years. He has provided me so much support and brilliant advice. I truely appreciate all the valuable time and effort he has spent on me.

I also want to thank the staff of the Department of Mathematics, especially my TAP panel (Dr Stephen Connor, Dr Degui Li, and Dr Heping He). They have provided so much valuable advice on my PhD thesis. My lovely friends Mr Yuan Ke, Mr John Box, and Mr Xiang Li, thank you for offering me support and advice.

Special thanks to my dearest mother, Ms Jiaqi Ye, who provides me endless love and support, and without whom, I could never have accomplish my PhD; my father, Mr Yude Yan, who encouraged me to pursue a PhD; and my devoted friend, Mr Shang Ma, who gives me endless support, love and understanding.

Author's Declaration

Chapters 2 is concerned with the fundamental method in nonparametric statistics. We reviewed the framework of the local polynomial modelling in Chapter 2, which is mostly from the Book: Fan, J and Gijbels, I (1996) Local Polynomial modeling and Its applications.

In Chapter 7, the results are mainly from my published paper: The Connection between cross-validation and AIC in a semiparametric family joint with Prof. Wenyang Zhang (Department of Mathematics The University of York, UK) and Dr. Heng,Peng (Department of Mathematics, Hong Kong Baptist University, Hong Kong). In Chapter 11.1, the basic information about Boston and the geographic map of Boston are from Wikipedia.

The rest of the chapters are mostly related to my submitted paper joint with Prof. Wenyang Zhang, Dr Yan Sun (School of Economics Shanghai University of Finance and Economics, P. R. China) and Dr Zudi, Liu (School of Mathematical Sciences,The University of Adelaide, Australia).

To the best of my knowledge and belief this thesis does not infringe the copyright of any other person.

1 Introduction

1.1 Background of the model and the project

The Boston House Price data are frequently used in the literature to illustrate new statistical methods. If we use y_i to denote the median value of owner-occupied homes at location s_i, a spatial autoregressive model for the data would be

$$y_i = \sum_{j \neq i} w_{ij} y_j + \epsilon_i, \quad i = 1, \cdots, n, \tag{1.1}$$

where w_{ij} is the impact of y_j on y_i. Apparently, (1.1) does not adequately address what affects the house price and how. It is better to incorporate the effects of some important covariates, such as crime rate and accessibility to radial highways, into the model. If X_i, a p dimensional vector, is the vector of the covariates associated with y_i, a reasonable model to fit the data would be

$$y_i = \sum_{j \neq i} w_{ij} y_j + X_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n. \tag{1.2}$$

where w_{ij} and $\boldsymbol{\beta}$ are unknown. However, there are two problems with model (1.2): first, there are too many unknown parameters; second, the model has not taken into account the location effects of the impacts of the covariates – the impacts of some covariates may vary with location. To control the number of unknown parameters and take the location effects into account, we propose the following model to fit the data

$$y_i = \alpha \sum_{j \neq i} w_{ij} y_j + X_i^{\mathrm{T}} \boldsymbol{\beta}(s_i) + \epsilon_i, \quad i = 1, \ \cdots, \ n, \tag{1.3}$$

where w_{ij} is a specified certain physical or economic distance, $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ \cdots, \ \beta_p(\cdot))^{\mathrm{T}}$, ϵ_i, $i = 1, \ \cdots, \ n$, are i.i.d., and follow $N(0, \ \sigma^2)$, $\{X_i, \ i = 1, \ \cdots, \ n\}$ is independent of $\{\epsilon_i, \ i = 1, \ \cdots, \ n\}$.The unknown parameters α, σ^2 and $\boldsymbol{\beta}(\cdot)$ are unknown and cannot be estimated. Model (1.3) is the model addressed in this thesis. Hereafter, y_i is of course not necessarily the house price, it is a generic response variable.

In model (1.3), the spatial neighbouring effect of y_j, $j \neq i$, on y_i is formulated through αw_{ij}, where w_{ij} is a specified certain physical or economic distance. Such method to define spatial neighbouring effect is common, see Ord (1975), Anselin (1988), Su and Jin (2010).

Model (1.3) is a useful extension of spatial autoregressive models (Gao *et al.*, 2006; Kelejian and Prucha, 2010; Ord, 1975; Su and Jin, 2010) and varying coefficient models (Cheng *et al.*, 2009; Fan and Zhang, 1999, 2000; Li and Zhang, 2011; Sun *et al.*, 2007; Zhang *et al.*, 2002; Zhang *et al.*, 2009; Wang and Xia, 2009; and Tao and Xia, 2011). One characteristic of model (1.3) is

$$E(\epsilon_i | y_1, \ \cdots, \ y_{i-1}, \ y_{i+1}, \ \cdots, \ y_n) \neq 0$$

although $E(\epsilon_i) = 0$, the standard least squares estimation will not work for (1.3). Given the local linear modelling and profile likelihood idea, we propose a local likelihood based estimation procedure for the unknown parameters and functions in (1.3) and derive the asymptotic properties of the obtained estimators.

Cross-validation and AIC/BIC are the most commonly used tools in model selection. Due to the structure of model (1.3), it is easier to apply the AIC/BIC method than the Cross-validation method in model selection. Inspired by this, we investigated the connection between these two methods in a semiparametric model, and find that they are equivalent to each other. Given the above reasons along with others, this question becomes a very interesting and important topic.

In reality, some of the components of $\boldsymbol{\beta}(\cdot)$ in model (1.3) may be constant, and we do not know which components are functional and which are constant. Methodologically speaking, if mistakenly treating a constant component as functional, we would pay a price on the variance side of the obtained estimator. However, if mistakenly treating a functional component as constant, we would pay a price on the bias side of the obtained estimator. The identification of constant/functional components in $\boldsymbol{\beta}(\cdot)$ is imperative. From a practical point of view, the identification of constant components is also important. For the data set we study in this paper, $\boldsymbol{\beta}(\cdot)$ can be interpreted as the vector of the impacts of the covariates concerned on the house price. The identification will reveal which covariates have location-varying impacts with the House Price, and which do not. This is apparently of great interest. Because it is easier to apply the AIC/BIC method than Cross-Validation method, we show how many unknown parameters an unknown bivariate function amounts to, and propose an AIC/BIC nonparametric version to identify the constant components of $\boldsymbol{\beta}(\cdot)$ in model (1.3).

Throughout this thesis, $\mathbf{0}_k$ is a k dimensional vector with each component being 0, I_k is an identity matrix of size k and $U[0,\ 1]^2$ is a two dimensional uniform distribution on $[0,\ 1] \times [0,\ 1]$.

1.2 Thesis organization

In Chapter 2, we review the fundamental methodology in nonparametric statistics-the local polynomial modelling method. In chapter 2.1, we introduce the framework of the local polynomial modelling method, the bivariate case. The method for multivariate data is presented in chapter 2.2. This chapter provides insights into nonparametric estimation and should offer a better understanding of the methods we derived in this thesis.

In Chapter 3, we describe the estimation procedure for the proposed model (1.3). The asymptotic properties of the proposed methods are presented in Chapter 4, followed by the proofs of the theorems and lemmas, in Chapters 5 and 6.

In model selection, there are three methods people would mostly use : AIC, BIC and cross-validation method. Due to the structure of the proposed model (1.3), it is not straightforward to apply the cross-validation method here, in contrast, AIC/BIC methods are easier to apply.As a result, in Chapter 7, we have showed the connection between cross-validation and AIC/BIC in the semiparametric family. We illustrated the equivalence of these two methods in this chapter. The simulation studies showing the connection are listed in chapter 7.4. The theoretical proofs are presented in chapter 7.5.Investigating the connection between these methods is also an very interesting and important research object.

In Chapter 8, the model selection methods are introduced. The thresholding K method is presented in chapter 8.1,followed by the Curvature-to-Average ratio (CTAR) based method, which is illustrated in chapter 8.2. In chapter 8.3, we show how many unknown parameters an unknown bivariate function amounts to, and propose an AIC/BIC of nonparametric version for model selection here.

The performances of the proposed estimation and model selection procedures are assessed by the simulation studies in Chapter 9 and Chapter 10. In Chapter 9, we estimate the unknown function $\beta(\cdot)$ and unknown parameter α under different situations. The Oracle properties of the estimation procedure are also presented in Chapter 9.

In Chapter 11, we explore how the covariates, which are commonly found to be associated with House Price, affect the median value of owner-occupied homes in Boston, and how the impacts of these covariates change with the location, based on the proposed model and estimation procedure.

The conclusions and a discussion of the future research are presented in Chapter 12.

2 Local Polynomial Modelling

2.1 Framework of local polynomial modelling

In statistics, regression analysis is one of the most useful and commonly used tools. In this chapter, we review the techniques applied in nonlinear regression, especially the local polynomial modelling, which is one of the most widely used techniques in nonparametric statistics. We begin by introducing this technique in the case of one-dimensional variables. We introduce the multivariate cases in the next sub-chapter.Consider that we generate n $i.i.d$ bivariate data samples, i.e, $\{(X_i, Y_i), \quad i = 1, \cdots, n\}$,the data are generated from the following model:

$$Y = l(X) + \sigma(X)\varepsilon \tag{2.1}$$

Here, we assume that $E(\varepsilon) = 0$, $Var(\varepsilon) = 1$, and that independent variable X and ε are independent. We also assume the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X, i.e, the design density, by $f(\cdot)$. We assume that the $(p+1)^{th}$ derivative of $l(x)$ at the point x_0 exists. The Taylor expansion for the point x , which is in the neighborhood of x_0 ,is:

$$l(x) \approx \sum_{i=0}^{p} \frac{l^{(i)}(x_0)}{i!}(x - x_0)^i \tag{2.2}$$

This polynomial is fitted locally by the weighted least squares regression problem. Treating $\frac{l^{(j)}(x_0)}{j!} = \beta_j$, for $j = 0, 1, \cdots, p$, we minimize the following weighted least squares regression:

$$\sum_{i=1}^{n}\{Y_i - \sum_{j=1}^{n}\beta_j(X_i - x_0)^j\}^2 K_h(X_i - x_0) \tag{2.3}$$

Here, K is the kernel function, and h is the bandwidth, which controls the size of the neighborhood, $K_h(\cdot) = K(\cdot/h)/h$. We minimize this problem with respect to β_j and the solution to the least squares problem is denoted by $\hat{\beta}_j$, $j = 0, 1, \cdots, p$. For a better understanding, we can view the weighted least squares in a matrix form. According to the notations in Fan and Gijbels (1996), denote the design matrix $\mathbf{X}$, $\mathbf{y}$,$\hat{\beta}$ as :

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_n \end{pmatrix},$$

let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights:

$$\mathbf{W} = diag\{K_h(X_1 - x_0), K_h(X_2 - x_0), \cdots, K_h(X_n - x_0)\} \tag{2.4}$$

The weighted least squares problems can be rewritten as:

$$min(\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}\mathbf{W}(\mathbf{y} - \mathbf{X}\beta) \tag{2.5}$$

with $\beta = (\beta_0, \cdots, \beta_p)^T$, and the solution in the matrix form can be viewed as :

$$\hat{\beta} = (\mathbf{X^T W X})^{-\mathbf{1}}\mathbf{X^T W y} \tag{2.6}$$

We know the conditional expectation of $\mathbf{y}$ given $\mathbf{X}$ is :

$$l(x_0) = E(Y \mid X = x_0) \tag{2.7}$$

By equation 2.6 , we can easily derive the conditional bias and variance of the estimator $\hat{\beta}$:

$$E(\hat{\beta} \mid X) = (\mathbf{X^T W X})^{\mathbf{-1}} \mathbf{X^T W l} = \beta + (\mathbf{X^T W X})^{\mathbf{-1}} \mathbf{X^T W r} \tag{2.8}$$

$$Var(\hat{\beta} \mid X) = (\mathbf{X^T W X})^{\mathbf{-1}} (\mathbf{X^T \Sigma X})(\mathbf{X^T W X})^{\mathbf{-1}} \tag{2.9}$$

where $\mathbf{l} = \{l(X_1), \cdots, l(X_n)\}^T$, $\beta = \{l(x_0), \cdots, l^p(x_0)/p!\}^T$;
and $\mathbf{r} = \mathbf{l} - \mathbf{X}\beta$ is the residual vector of the approximation,
$\Sigma = diag\{K_h^2(X_1 - x_0)\sigma^2(X_1), \cdots, K_h^2(X_n - x_0)\sigma^2(X_n)\}$

However, the exact bias and variance of $\hat{\beta}$ are not directly usable due to the unknown quantities: the residual $\mathbf{r}$ and the diagonal matrix Σ. There are two ways to solve the problem. One method is the "plug-in" method. We find the estimator of the unknown quantities, then we plug them into the equations. Another method is founded by Ruppert and Wand (1994). They found the approximation of the conditional bias and variance by their first order asymptotic expansions. Before illustrating the results, we would first introduce some notations we will use in the Theorem. Denote the moments of K and K^2 by $\mu_j = \int \mu^j K(\mu) d\mu$ and $\nu_j = \int \mu_j K^2(\mu) d\mu$ respectively. The unit vector $e_{v+1} = (0, 0, \cdots, 1, 0, \cdots, 0)^T$ and 1 is the $(v+1)_{th}$ component. There are also some matrices and vectors of moments that appear in the asymptotic expressions.

$$S = (\mu_{j+l})_{0\leq j,l\leq p} \quad c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T$$

$$\tilde{S} = (\mu_{j+l+1})_{0\leq j,l\leq p} \quad \tilde{c_p} = (\mu_{p+2}, \cdots, \mu_{2p+2})^T$$

$$S_* = (\nu_{j+l})_{0\leq j,l\leq p}$$

We have the following theorem:

Theorem 2.1 *Assume that $t(x_0) > 0$ and that $t(\cdot)$,$l^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0.Further, assume that $h \to 0$ and $nh \to \infty$ Then the asymptotic conditional variance of $\hat{l_v}(x_0)$ is given by*

$$Var(\hat{l_v}(x_0) \mid X) = e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) nh^{1+2\nu}} + Op(\frac{1}{nh^{1+2\nu}}) \tag{2.10}$$

the asymptotic conditional bias for $p - \nu$ odd is given by

$$Bias(\hat{l_v}(x_0) \mid X) = e_{v+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} l^{(p+1)} h^{p+1-\nu} + op(h^{p+1-\nu}) \tag{2.11}$$

the asymptotic conditional bias for $p - \nu$ even is given by

$$Bias(\hat{l_v}(x_0) \mid X) = e_{v+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+2)!} \{l^{(p+2)}(x_0)$$

$$+(p+2) m^{(p+1)}(x_0) \frac{f'(x_0)}{f(x_0)}\} h^{p+2-\nu} + op(h^{p+2-\nu})$$

There are several advantages of local polynomial fitting. One of the most important is that local polynomial fitting is nearly optimal in an asymptotic minimax sense. The computational costs for the local polynomial estimators are very low due to their simplicity. Fan and Marron (1994) showed that it was possible to do local polynomial fitting in $O(n)$ operations. This method is an effective and easily applied method in nonparametric statistics that adapts to various types of designs. In the next chapter, we discuss this method for multivariate data.

2.2 Local polynomial modelling for multivariate data

Given multivariate covariate $\mathbf{X}$ and dependent variable $\mathbf{Y}$, we want to estimate the mean regression function.For simplicity, we only introduce the local linear fitting here, i.e $p = 1$. However, the key idea behind and the methodology of using local polynomial fitting for higher p are the same. We still minimize the multivariate version of the weighted least squares regression:

$$\sum_{i=1}^{n}\{\mathbf{Y_i} - \beta_0 - \sum_{j=1}^{d}\beta_j(X_{ij} - x_j)\}^2 K_B(\mathbf{X_i} - \mathbf{x}), \tag{2.12}$$

with respect to $\beta = (\beta_0, \cdots, \beta_d)^T$. We define:

$$K_B(\mathbf{u}) = \frac{1}{\mid B \mid} K(B^{-1}\mathbf{u}) \tag{2.13}$$

The bandwidth matrix B is a nonsingular $d \times d$ matrix and $\mid B \mid$ is the determinant of the bandwidth matrix. K is defined as a d-variate nonnegative kernel function. The solution for the multivariate version of the weighted least squares regression problem is:

$$\hat{\beta} = (\mathbf{X_D}^\mathbf{T}\mathbf{W}\mathbf{X_D})^{-\mathbf{1}}\mathbf{X_D}^\mathbf{T}\mathbf{W}\mathbf{y}, \tag{2.14}$$

where,

$$\mathbf{X}_D = \begin{pmatrix} 1 & (X_{11}-x_1) & \cdots & (X_{1d}-x_d) \\ 1 & (X_{21}-x_1) & \cdots & (X_{2d}-x_d) \\ \vdots & \vdots & & \vdots \\ 1 & (X_{n1}-x_1) & \cdots & (X_{nd}-x_d) \end{pmatrix},$$

The weight matrix is $\mathbf{W} = diag\{K_B(\mathbf{X_1}-\mathbf{x}), \cdots, K_B(\mathbf{X_n}-\mathbf{x})\}$. According to Ruppert and Wand(1994), we can find the conditional bias and variance. The conditional bias of the estimator $\hat{l}(\mathbf{x})$ is given by:

$$E\{\hat{l}(\mathbf{x}) - l(\mathbf{x}) \mid X\} = \frac{1}{2}\mu_2(K)[tr\{H(\mathbf{x})BB^T\} + op\{tr(BB^T)\}] \quad (2.15)$$

And the conditional variance of the estimator is :

$$Var\{\hat{l}(\mathbf{x}) \mid X\} = \frac{1}{n \mid B \mid}\nu_0(K)\frac{\sigma^2(\mathbf{x})}{f(\mathbf{x})}\{1 + op(1)\} \quad (2.16)$$

where $\nu_0(K) = \int K^2(\mu)d\mu$, and $H(\mathbf{x})$ is defined as the Hesian matrix of m at $\mathbf{x}.f$ denote the d-variate marginal density function of $\mathbf{X} = (X_1, \cdots, X_d)^T$.

In the next chapter, we introduce the estimation procedure for the designed model. We use the methodology of the local polynomial modelling, which we have elaborated on this chapter.

3 Estimation Procedure

Let $w_{ii} = 0$, $W = (w_{ij})$, $Y = (y_1,\ \cdots,\ y_n)^{\mathrm{T}}$, $A = I - \alpha W$, and $\mathbf{m} = \left(X_1^{\mathrm{T}}\boldsymbol{\beta}(s_1),\ \cdots,\ X_n^{\mathrm{T}}\boldsymbol{\beta}(s_n)\right)^{\mathrm{T}}$. By simple calculation, we have that the conditional density function of Y given $\mathbf{m}$ is $N\left(A^{-1}\mathbf{m},\ (A^{\mathrm{T}}A)^{-1}\sigma^2\right)$, which leads to the following log likelihood function

$$-\frac{n}{2}\log(2\pi) - n\log(\sigma) + \log(|A|) - \frac{1}{2\sigma^2}(AY - \mathbf{m})^{\mathrm{T}}(AY - \mathbf{m}). \quad (3.1)$$

Our estimation is profile likelihood based. We first construct the estimator $\tilde{\boldsymbol{\beta}}(\cdot;\ \alpha)$ of $\boldsymbol{\beta}(\cdot)$ pretending α is known, then let $(\hat{\alpha},\ \hat{\sigma}^2)$ maximise (3.1) with $\boldsymbol{\beta}(\cdot)$ being replaced by $\tilde{\boldsymbol{\beta}}(\cdot;\ \alpha)$. $\hat{\alpha}$ and $\hat{\sigma}^2$ are our estimators of α and σ^2, respectively. After the estimator of α is obtained, the estimator of $\boldsymbol{\beta}(\cdot)$ is taken to be $\tilde{\boldsymbol{\beta}}(\cdot;\ \alpha)$ with α and the bandwidth used being replaced by $\hat{\alpha}$ and a slightly larger bandwidth, respectively. The details are as follows.

For any $s = (u,\ v)^{\mathrm{T}}$, we denote $(\partial\boldsymbol{\beta}(s)/\partial u,\ \partial\boldsymbol{\beta}(s)/\partial v)$ by $\dot{\boldsymbol{\beta}}(s)$, where $\partial\boldsymbol{\beta}(s)/\partial u = (\partial\beta_1(s)/\partial u,\ \cdots,\ \partial\beta_p(s)/\partial u)^{\mathrm{T}}$.

For any given s, by the Taylor's expansion, we have

$$\boldsymbol{\beta}(s_i) \approx \boldsymbol{\beta}(s) + \dot{\boldsymbol{\beta}}(s)(s_i - s)$$

when s_i is in a small neighbourhood of s, which leads to the following objective function for estimating $\boldsymbol{\beta}(s)$

$$\sum_{i=1}^{n}\left(y_i^* - X_i^{\mathrm{T}}\mathbf{a} - X_i^{\mathrm{T}}\mathbf{B}(s_i - s)\right)^2 K_h(\|s_i - s\|), \quad (3.2)$$

where y_i^* is the ith component of AY, $K_h(\cdot) = K(\cdot/h)/h^2$, $K(\cdot)$ is a kernel function, and h is a bandwidth. Let $(\hat{\mathbf{a}},\ \hat{\mathbf{B}})$ minimise (3.2), the 'estimator' $\tilde{\boldsymbol{\beta}}(s;\ \alpha)$ of $\boldsymbol{\beta}(s)$ is taken to be $\hat{\mathbf{a}}$. By simple calculations, we have

$$\tilde{\boldsymbol{\beta}}(s;\ \alpha) = \hat{\mathbf{a}} = (I_p,\ \mathbf{0}_{p\times 2p}) \left(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}\right)^{-1} \mathcal{X}^{\mathrm{T}}\mathcal{W}AY, \tag{3.3}$$

where $\mathbf{0}_{p\times q}$ is a matrix of size $p \times q$ with each entry being 0, and

$$\mathcal{X} = \begin{pmatrix} X_1 & \cdots & X_n \\ X_1 \otimes (s_1 - s) & \cdots & X_n \otimes (s_n - s) \end{pmatrix}^{\mathrm{T}},$$

$$\mathcal{W} = \operatorname{diag}\left(K_h(\|s_1 - s\|),\ \cdots,\ K_h(\|s_n - s\|)\right).$$

Replacing $\boldsymbol{\beta}(s_i)$ in (3.1) by $\tilde{\boldsymbol{\beta}}(s_i;\ \alpha)$ and ignoring the constant term, we have the objective function for estimating α and σ^2

$$-n\log(\sigma) + \log(|A|) - \frac{1}{2\sigma^2}(AY - \tilde{\mathbf{m}})^{\mathrm{T}}(AY - \tilde{\mathbf{m}}), \tag{3.4}$$

where $\tilde{\mathbf{m}}$ is $\mathbf{m}$ with $\boldsymbol{\beta}(s_i)$ being replaced by $\tilde{\boldsymbol{\beta}}(s_i;\ \alpha)$. Let α_i, $i = 1,\ \cdots,\ n$, be the eigenvalues of W,

$$\tilde{\sigma}^2 = \frac{1}{n}(AY - \tilde{\mathbf{m}})^{\mathrm{T}}(AY - \tilde{\mathbf{m}}),$$

and $(\hat{\alpha},\ \hat{\sigma}^2)$ maximise (3.4). Noticing that $|A| = \prod_{i=1}^{n}(1 - \alpha\alpha_i)$, by simple calculation, we have $\hat{\alpha}$ is the maximiser of

$$-n\log(\tilde{\sigma}) + \sum_{i=1}^{n}\log(|1 - \alpha\alpha_i|), \tag{3.5}$$

and $\hat{\sigma}^2$ is $\tilde{\sigma}^2$ with α being replaced by $\hat{\alpha}$.

Maximizing (3.5) is not difficult as it is an one dimensional optimization problem and we can use a grid point method to solve it.

The estimator $\hat{\boldsymbol{\beta}}(\cdot)$ $\left(= (\hat{\beta}_1(\cdot),\ \cdots,\ \hat{\beta}_p(\cdot))^{\mathrm{T}}\right)$ is $\tilde{\boldsymbol{\beta}}(\cdot;\ \alpha)$ with α being replaced by $\hat{\alpha}$ and the bandwidth h with a slightly larger bandwidth h_1. The reason to replace the bandwidth h by a slightly larger one is that h is for the estimation of constant parameters such as α and σ^2, and thus is usually smaller than the one for estimating functional parameters, because undersmooth is needed for the estimators of constant parameters to achieve the optimal convergence rate.

In reality, some components of $\boldsymbol{\beta}(\cdot)$ may be constant. If a component of $\boldsymbol{\beta}(\cdot)$ is a constant, say $\beta_1(\cdot) = \beta_1$, we use the average of $\hat{\beta}_1(s_i)$, $i = 1,\ \cdots,\ n$, to estimate the constant β_1, that is

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1(s_i).$$

How to identify the constant components of $\boldsymbol{\beta}(\cdot)$ is addressed in the following chapter.

4 Asymptotic Properties of the Proposed Estimators

In this chapter, we present the asymptotic properties of the proposed estimators. We only present the asymptotic results and leave the theoretical proofs for chapters 5 and 6.

Although we assume that ϵ_i in (1.3) follows normal distribution in our model assumption, we do not need this assumption when deriving the asymptotic properties of the proposed estimators. So, in this chapter, we do not assume that ϵ_i follows normal distribution unless otherwise stated.

In this chapter, for w_{ij} in (1.3), we assume that there exists a sequence $\rho_n > 0$ such that $w_{ij} = O(1/\rho_n)$ uniformly with respect to $i, and\ j$ and the matrices W and A^{-1} are uniformly bounded in both row and column sums.

We now introduce some notations needed in the presentation of the asymptotic properties of the proposed estimators: Let $\mu_j = E\epsilon_1^j$,

$$\kappa_0 = \int_{R^2} K(\|s\|)ds,\ \ \kappa_2 = \int_{R^2} [(1,0)s]^2 K(\|s\|)ds = \int_{R^2} [(0,1)s]^2 K(\|s\|)ds,$$

$$\nu_0 = \int_{R^2} K^2(\|s\|)ds,\ \ \nu_2 = \int_{R^2} [(1,0)s]^2 K^2(\|s\|)ds = \int_{R^2} [(0,1)s]^2 K^2(\|s\|)ds,$$

$$\Psi = E(X_1X_1^{\mathrm{T}}),\ \Gamma = EX_1,\ Z_1(s) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g_{ii}\boldsymbol{\beta}(s_i)K_h(\|s_i - s\|),$$

$$Z_2(s) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j\neq i}^{n} g_{ij}\boldsymbol{\beta}(s_j)K_h(\|s_i - s\|), \quad Z(s) = Z_1(s) + \Psi^{-1}\Gamma\Gamma^{\mathrm{T}} Z_2(s),$$

$$Z = \kappa_0^{-1}\Big(f^{-1}(s_1)X_1^{\mathrm{T}}Z(s_1), \cdots, f^{-1}(s_n)X_n^{\mathrm{T}}Z(s_n)\Big)^{\mathrm{T}},$$

$$A = I_n - \alpha W, \quad G = (g_{ij}) = WA^{-1},$$

$$\pi_1 = \lim_{n\to\infty} \frac{\mathrm{tr}((G+G^{\mathrm{T}})G)}{n}, \quad \pi_2 = \lim_{n\to\infty} \frac{\mathrm{tr}(G)}{n}, \quad \pi_3 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2,$$

$$\lambda_1 = \lim_{n\to\infty} \frac{1}{n} E[(G\mathbf{m} - Z)^{\mathrm{T}}(G\mathbf{m} - Z)],$$

$$\lambda_2 = \lim_{n\to\infty} \frac{1}{n} E[(G\mathbf{m} - Z)^{\mathrm{T}} G_c], \quad \lambda_3 = \lim_{n\to\infty} \frac{1}{n} E[(G\mathbf{m} - Z)^{\mathrm{T}} \mathbf{1}_n]$$

where $G_c = (g_{11}, \cdots, g_{nn})^{\mathrm{T}}$ and $\mathbf{1}_n$ is a n dimensional vector with each component being 1. Further, let

$$\Omega = \begin{pmatrix} \frac{1}{\sigma^2}\lambda_1 + \pi_1 & \frac{1}{\sigma^2}\pi_2 \\ \frac{1}{\sigma^2}\pi_2 & \frac{1}{2\sigma^4} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \frac{\mu_4 - 3\sigma^4}{\sigma^4}\pi_3 + \frac{2\mu_3}{\sigma^4}\lambda_2 & \frac{\mu_3}{2\sigma^6}\lambda_3 + \frac{\mu_4 - 3\sigma^4}{2\sigma^6}\pi_2 \\ \frac{\mu_3}{2\sigma^6}\lambda_3 + \frac{\mu_4 - 3\sigma^4}{2\sigma^6}\pi_2 & \frac{\mu_4 - 3\sigma^4}{4\sigma^8} \end{pmatrix},$$

$$s = (u, v)^{\mathrm{T}}, \quad \boldsymbol{\beta}_{uu}(s) = \left(\frac{\partial^2 \beta_1(s)}{\partial u^2}, \cdots, \frac{\partial^2 \beta_p(s)}{\partial u^2}\right)^{\mathrm{T}},$$

$$\boldsymbol{\beta}_{vv}(s) = \left(\frac{\partial^2 \beta_1(s)}{\partial v^2}, \cdots, \frac{\partial^2 \beta_p(s)}{\partial v^2}\right)^{\mathrm{T}}$$

and

$$S = \begin{pmatrix} (X_1^{\mathrm{T}}, \mathbf{0}_{1\times 2p})\left(\mathcal{X}_{(1)}^{\mathrm{T}} \mathcal{W}_{(1)} \mathcal{X}_{(1)}\right)^{-1} \mathcal{X}_{(1)}^{\mathrm{T}} \mathcal{W}_{(1)} \\ \vdots \\ (X_n^{\mathrm{T}}, \mathbf{0}_{1\times 2p})\left(\mathcal{X}_{(n)}^{\mathrm{T}} \mathcal{W}_{(n)} \mathcal{X}_{(n)}\right)^{-1} \mathcal{X}_{(n)}^{\mathrm{T}} \mathcal{W}_{(n)} \end{pmatrix}$$

where $\mathcal{X}_{(i)}$ and $\mathcal{W}_{(i)}$ are $\mathcal{X}$ and $\mathcal{W}$ respectively with s being replaced by s_i, $i = 1, \ \cdots, \ n$.

Using simple calculations, we can see the matrix Ω defined above is the limit of the Fisher information matrix of α and σ^2. As the singularity of matrix Ω may have serious implications for the convergence rate of the proposed estimators, we present the asymptotic properties for the case in which Ω is nonsingular and the case in which Ω is singular separately. We present the nonsingular case in Theorems 1 - 3, and the singular case in Theorems 4 - 7.

Theorem 1. *Under the Conditions (1)-(7) or Conditions (1)-(6), $(\tilde{7})$ and (8) in Chapter 5, α in model (1.3) is identifiable and Ω is nonsingular, and when $n^{1/2}h^2/\log^2 n \to \infty$ and $nh^8 \to 0$, $\hat{\alpha}$ and $\hat{\sigma}^2$ are consistent estimators of α and σ^2, respectively.*

Theorem 1 shows the conditions under which Ω is nonsingular and the consistency of $\hat{\alpha}$ and $\hat{\sigma}^2$ under such conditions. Based on Theorem 1, we can derive the asymptotic nomality of $\hat{\alpha}$ and $\hat{\sigma}^2$.

Theorem 2. *Under the assumptions of Theorem 1, if the second partial derivative of $\boldsymbol{\beta}(s)$ is Lipschitz continuous and $nh^6 \to 0$,*

$$\sqrt{n}\left(\hat{\alpha}-\alpha,\ \hat{\sigma}^2-\sigma^2\right)^T \xrightarrow{D} N(\mathbf{0},\ \Omega^{-1}+\Omega^{-1}\Sigma\Omega^{-1}).$$

Further, if ϵ_i is normally distributed,

$$\sqrt{n}\left(\hat{\alpha}-\alpha,\ \hat{\sigma}^2-\sigma^2\right)^T \xrightarrow{D} N(\mathbf{0},\ \Omega^{-1}).$$

Theorem 2 implies that the convergence rate of $\hat{\alpha}$ is of order $n^{-1/2}$ when Ω is nonsingular, which is the optimal rate for parametric estimation. We will see, in Theorem 5, this rate can not be achieved by $\hat{\alpha}$ when Ω is singular.

Theorem 3. *Under the assumptions of Theorem 1, if $nh_1^6 = O(1)$ and*

$h/h_1 \to 0$,

$$\sqrt{nh_1^2 f(s)}\Big(\hat{\boldsymbol{\beta}}(s)-\boldsymbol{\beta}(s)-2^{-1}\kappa_0^{-1}\kappa_2 h_1^2\{\boldsymbol{\beta}_{uu}(s)+\boldsymbol{\beta}_{vv}(s)\}\Big) \xrightarrow{D} N\Big(\mathbf{0},\ \kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\Big).$$

for any given s.

Theorem 3 shows $\hat{\boldsymbol{\beta}}(\cdot)$ is asymptotic normal and achieves the convergence rate of order $n^{-1/6}$, which is the optimal rate for bivariate nonparametric estimation.

We now turn to the case where Ω is singular.

Theorem 4. *Under the Conditions (1)-(6) and (9) in Chapter 5, α is identifiable and Ω is singular, and if $nh^8 \to 0$, $n^{1/2}h^2/\log^2 n \to \infty$, $\rho_n \to \infty$, $\rho_n h^4 \to 0$ and $nh^2/\rho_n \to \infty$, $\hat{\alpha}$ is a consistent estimator of α.*

Theorem 5. *Under the assumptions of Theorem 4, if the second partial derivative of $\boldsymbol{\beta}(s)$ is Lipschitz continuous and $nh^6 \to 0$,*

$$\sqrt{n/\rho_n}(\hat{\alpha}-\alpha) \xrightarrow{D} N(0,\ \sigma_\alpha^2),$$

where

$$\begin{aligned}\sigma_\alpha^2 &= \Big[\frac{1}{\sigma^2}\lambda_4 + \lim_{n\to\infty}\frac{\rho_n}{n}tr((G+G^T)G)\Big]^{-2} \times \\ &\quad \Big\{\frac{1}{\sigma^2}\lambda_4 + \lim_{n\to\infty}\frac{\rho_n}{n}tr((G+G^T)G) + \frac{2\mu_3}{\sigma^4}\lim_{n\to\infty}\frac{\rho_n}{n}E[(G\mathbf{m}-SG\mathbf{m})^T G_c]\Big\}\end{aligned}$$

and

$$\lambda_4 = \lim_{n\to\infty}\frac{\rho_n}{n}E[(G\mathbf{m}-SG\mathbf{m})^T(G\mathbf{m}-SG\mathbf{m})].$$

Theorem 5 shows the convergence rate of $\hat{\alpha}$ is of order $(\rho_n/n)^{-1/2}$ which is slower than $n^{-1/2}$ when $\rho_n \longrightarrow \infty$. However, we will see, from Theorem 7, this has no effect on the asymptotic properties of $\hat{\boldsymbol{\beta}}(\cdot)$.

Theorem 6. *Under the assumptions of Theorem 5,*

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

Theorem 6 shows that although the asymptotic variance of $\hat{\sigma}^2$ is different to that when Ω is nonsingular, $\hat{\sigma}^2$ still enjoys convergence rate of $n^{-1/2}$.

Theorem 7. *Under the assumptions of Theorem 4, if $nh_1^6 = O(1)$ and $h/h_1 \to 0$,*

$$\sqrt{nh_1^2 f(s)}\Big(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s) - 2^{-1}\kappa_0^{-1}\kappa_2 h_1^2\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\}\Big) \xrightarrow{D} N\Big(\mathbf{0}, \kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\Big)$$

for any given s.

From Theorem 3 and Theorem 7, we can see the singularity of Ω has no effect on the asymptotic distribution of $\hat{\boldsymbol{\beta}}(\cdot)$.

5 Proofs of Theorems

To avoid any confusion with notations, we use α_0 to denote the true value of α in this chapter. Further, we rewrite $A = I_n - \alpha W$ as $A(\alpha)$ to emphasis its dependence on α, and abbreviate $A(\alpha_0)$ as A.

The following regularity conditions are needed to establish the asymptotic properties of the estimators.

Conditions

(1) The kernel function $K(\cdot)$ is a bounded positive, symmetric and Lipshitz continuous function with a compact support on R. And $h \to 0$.

(2) $\{\beta_i(\cdot),\ i = 1, \cdots, p\}$ have continuous second partial derivatives.

(3) $\{X_i\}$ is a sequence of iid. random sample from the population and is independent of ϵ_i, $i = 1, \cdots, n$. Moreover, $E(X_1 X_1^{\mathrm{T}})$ is positive definite, $E\|X_1\|^{2q} < \infty$ and $E|\epsilon_1|^{2q} < \infty$ for some $q > 2$.

(4) $\{s_i\}$ is a sequence of fixed design points on a bounded support $\mathcal{S}$. Further, there exists a positive joint density function $f(\cdot)$ satisfying a Lipshitz condition such that

$$\sup_{s \in \mathcal{S}} \Big| \frac{1}{n} \sum_{i=1}^{n} [r(s_i) K_h(\|s_i - s\|) - \int r(t) K_h(\|t - s\|) f(t) dt \Big| = O(h)$$

for any bounded continuous function $r(\cdot)$ and $K_h(\cdot) = K(\cdot/h)/h^2$ where $K(\cdot)$ satisfies Condition (1). $f(\cdot)$ is bounded away from zero on $\mathcal{S}$.

(5) There exists a sequence $\rho_n > 0$ such that the elements w_{ij} of W are $O(1/\rho_n)$ uniformly in all $i,\ j$. As a normalization, $w_{ii} = 0$ for all i. Furthermore, the matrices W and A^{-1} are uniformly bounded in both row and column sums.

(6) $A^{-1}(\alpha)$ are uniformly bounded in either row or column sums, uniformly in α in a compact support Δ . The true α_0 is an interior point in Δ.

(7) $\lim_{n\to\infty} \frac{1}{n} E[(G\mathbf{m} - Z)^{\mathrm{T}}(G\mathbf{m} - Z)] = \lambda_1 > 0.$

($\tilde{7}$) $\lambda_1 = 0.$

(8) ρ_n is bounded and for any $\alpha \neq \alpha_0$,

$$\lim_{n\to\infty} \Big\{\frac{1}{n} \log \left|\sigma^2 A^{-1}(A^{-1})^{\mathrm{T}}\right| - \frac{1}{n} \log \left|\sigma_a^2(\alpha) A^{-1}(\alpha)(A^{-1}(\alpha))^{\mathrm{T}}\right|\Big\} = 0$$

where $\sigma_a^2(\alpha) = \frac{\sigma^2}{n} \mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}} A(\alpha)A^{-1}\}$.

(9) $\rho_n \to \infty$, the row sums of G have the uniform order $O(1/\sqrt{\rho_n})$, and

$$\lim_{n\to\infty} \frac{\rho_n}{n} E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}(G\mathbf{m} - SG\mathbf{m})] = \lambda_4 > 0.$$

Remark 1: Condition (1)-(3) are commonly seen in nonparametric estimation. They are not the weakest possible ones, but they are imposed to facilitate the technical proofs. Since the sampling units can be regarded as given, the fixed bounded design Condition (4) is made for technical convenience. Of course as in Linton(1995), Condition (4) does not preclude $\{s_i\}_{i=1}^n$ from being generated by some random mechanism. For example, if s_i's were iid. with joint density $f(\cdot)$, then Condition (4) holds with probability one which can be obtained similarly as Hansen (2008). In this case, we can obtain our results by first conditional on $\{s_i\}_{i=1}^n$ and then go on as usual.

Remark 2: Condition (5)-(8) parallel the corresponding conditions of Lee (2004) and Su and Jin (2010), in which Condition (5)-(6) concern the essential features of the weight matrix for the model. Condition (7) is a sufficient condition which ensures that the likelihood function of α has a unique maximizer. When Condition $(\tilde{7})$ holds and the elements of W are uniformly bounded, the uniqueness of the maximizer can be guaranteed by Condition (8). These two kinds of conditions ensure that Ω which is the limit of the information matrix of the finite dimensional parameters is nonsingular. So they are the crucial conditions for $\sqrt{n}-$ rate of convergence of the finite dimensional parameter estimators.

Remark 3: When $\rho_n \to \infty$, Ω can be nonsingular only if Condition (7) holds. For the situation under Condition $(\tilde{7})$, Ω will become singular. The singularity of the matrix may have implications on the rate of convergence of the estimators. Nevertheless, we follow Lee (2004) and Su and Jin (2010) to consider the situation in which

$$\lim_{n\to\infty} \frac{\rho_n}{n} E[(G\mathbf{m})^{\mathrm{T}}(I_n - S)^{\mathrm{T}}(I_n - S)G\mathbf{m}] = \lambda_4 \in (0, \infty).$$

In this case, it is natural to assume that the elements of $(I_n - S)G\mathbf{m}$ have the uniform order $O_P(1/\sqrt{\rho_n})$ which can be satisfied by the assumption that the row sums of G are of uniform order $O(1/\sqrt{\rho_n})$.

In the following, let H be a diagonal matrix of size $3p$ with its first p elements on the diagonal being 1 and the remaining elements being h, $P = (I_n - S)^{\mathrm{T}}(I_n - S)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^{\mathrm{T}}$. Moreover, like α_0, we use σ_0^2 to denote the true value of σ^2 to avoid confusion of notation. Since the following notations will be frequently used in the proofs, we list here for easy reference.

$l(\alpha, \sigma^2) = -\frac{n}{2}\log(\sigma^2) + \log(|A(\alpha)|) - \frac{1}{2\sigma^2}(A(\alpha)Y)^{\mathrm{T}}PA(\alpha)Y,$

$l_c(\alpha) = -\frac{n}{2}\log\tilde{\sigma}^2(\alpha) + \log|A(\alpha)|,$

$\tilde{\sigma}^2(\alpha) = \frac{1}{n}(A(\alpha)Y)^{\mathrm{T}}PA(\alpha)Y,$

$\bar{\sigma}^2(\alpha) = \frac{1}{n}E[(A(\alpha)Y)^{\mathrm{T}}PA(\alpha)Y],$

$\sigma_a^2(\alpha) = \frac{\sigma_0^2}{n}\mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}}A(\alpha)A^{-1}\}.$

To prove the theorems, the following lemmas are needed and their proofs can be founded in chapter 6.

Lemma 1. Let $\{Y_i\}$ be a sequence of independent random variables and $\{s_i\} \in R^2$ are nonrandom vectors. Suppose that for some $q > 2$, $\max_i E|Y_i|^q < \infty$. Then under Condition (1), we have

$$\sup_{s\in\mathcal{S}}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big[K_h(\|s_i - s\|)Y_i - E\{K_h(\|s_i - s\|)Y_i\}\Big]\Big| = O_p\Big(\Big\{\frac{\log n}{nh^2}\Big\}^{1/2}\Big),$$

provided that $n^{1-2/q}h^2/\log^2 n \to \infty$ and $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}K_h(\|s_i - s\|) < \infty$ for any $s \in \mathcal{S}$.

Lemma 2. Under the Conditions (1)-(4), then when $n^{1/2}h^2/\log^2 n \to \infty$,

(1) $n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}H^{-1} = \begin{pmatrix} \kappa_0 f(s)\Psi & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{2p\times p} & \kappa_2 f(s)\Psi\otimes I_2 \end{pmatrix} + O_P(c_n\mathbf{1}_{3p}\mathbf{1}_{3p}^{\mathrm{T}})$

holds uniformly in $s \in \mathcal{S}$ where $c_n = h + \{\frac{\log n}{nh^2}\}^{1/2}$,

(2) $\boldsymbol{\beta}(s) - (I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathbf{m} = -\frac{\kappa_2 h^2}{2\kappa_0}\Big\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\Big\} +$

$o_p(h^2\mathbf{1}_p)$ holds uniformly in $s \in \mathcal{S}$.

Lemma 3. Under the Conditions (1)-(5), then when $n^{1/2}h^2/\log^2 n \to \infty$,

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}G\mathbf{m} - n^{-1}E(H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}G\mathbf{m}) = o_P(1)$$

uniformly in $s \in \mathcal{S}$.

Lemma 4. Under the Conditions (1)(3)(4) and (5), when $n^{1/2}h^2/\log^2 n \to \infty$, we have (1) $\frac{1}{n}E[\mathrm{tr}(P)] = 1 + o(1)$, (2) $\frac{1}{n}E[\mathrm{tr}(G^{\mathrm{T}}P) - \mathrm{tr}(G)] = o(1)$, (3) $\frac{1}{n}E[\mathrm{tr}(G^{\mathrm{T}}PG) - \mathrm{tr}(G^{\mathrm{T}}G)] = o(1)$. Further, when $nh^2/\rho_n \to \infty$, then (4) $\frac{\rho_n}{n}E[\mathrm{tr}(P) - n] = o(1)$, (5) $\frac{\rho_n}{n}E[\mathrm{tr}(G^{\mathrm{T}}P) - \mathrm{tr}(G)] = o(1)$, (6) $\frac{\rho_n}{n}E[\mathrm{tr}(G^{\mathrm{T}}PG) - \mathrm{tr}(G^{\mathrm{T}}G)] = o(1)$.

Lemma 5. Under the Conditions (1)-(5), then when $n^{1/2}h^2/\log^2 n \to \infty$, (1) $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(nh^2)$. Moreover, under the assumption that the second partial derivatives of $\boldsymbol{\beta}(s)$ are all Lipschitz continuous, we have (2) $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = O_P(nh^3 + \{nh^2\log n\}^{1/2})$.

Lemma 6. Under the Conditions (1)-(5), when $n^{1/2}h^2/\log^2 n \to \infty$ and $nh^8 \to 0$, we have (1) $n^{-1/2}L^{\mathrm{T}}P\mathbf{m} = o_P(1)$ for $L = \mathbf{m}, \boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$, (2) $n^{-1}L^{\mathrm{T}}PG\mathbf{m} = o_P(1)$ for $L = \mathbf{m},\ \boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$.

Lemma 7. Under the Conditions (1)-(5), when $n^{1/2}h^2/\log^2 n \to \infty$, we have (1) $n^{-1}\Big\{(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m} - E[(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}]\Big\} = o_P(1)$, (2) $n^{-1}E[(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}] = n^{-1}E[(G\mathbf{m} - Z)^{\mathrm{T}}(G\mathbf{m} - Z)] + o(1)$.

Lemma 8. Under the Conditions (1)-(5), when $n^{1/2}h^2/\log^2 n \to \infty$, we have (1) $n^{-1/2}\{\boldsymbol{\epsilon}^{\mathrm{T}}P\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}\} = o_P(1)$, (2) $n^{-1/2}\{\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}P\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}\boldsymbol{\epsilon}\} =$

$o_P(1)$, (3) $n^{-1/2}\{\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}PG\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}G\boldsymbol{\epsilon}\}=o_P(1)$, (4) $n^{-1/2}\{(G\mathbf{m})^{\mathrm{T}}P\boldsymbol{\epsilon}-(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon}\}=o_P(1)$.

Lemma 9. Suppose that $B=(b_{ij})_{1\le i,j\le n}$ is a sequence of symmetric matrices with row and column sums uniformly bounded and its elements are also uniformly bounded. Let $\sigma^2_{Q_n}$ be the variance of Q_n where $Q_n=(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^{\mathrm{T}}B\boldsymbol{\epsilon}-\sigma_0^2\mathrm{tr}(B)$. Assume that the variance $\sigma^2_{Q_n}$ is $O(n)$ with $\{\frac{\sigma^2_{Q_n}}{n}\}$ bounded away from zero, then we have under Conditions (3)- (5) that $\frac{Q_n}{\sigma_{Q_n}}\xrightarrow{D}N(0,1)$.

Lemma 10. Under the Conditions (1)-(5), the row sums of matrix G having the uniform order $O(1/\sqrt{\rho_n})$ and $n^{1/2}h^2/\log^2 n\to\infty$, we have (1) $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m}=o_P(\rho_n^{-1/2}nh^2)$. Moreover, if the second partial derivatives of $\boldsymbol{\beta}(s)$ are all Lipschitz continuous, then (2) $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m}=O_P(\rho_n^{-1/2}nh^3+\{nh^2\log n/\rho_n\}^{1/2})$.

Lemma 11. Under the Conditions (1)-(5) and the row sums of matrix G having the uniform order $O(1/\sqrt{\rho_n})$, then when $n^{1/2}h^2/\log^2 n\to\infty$, $\rho_n\to\infty$, $\rho_n h^4\to 0$ and $nh^2/\rho_n\to\infty$, we have (1) $\frac{\rho_n}{n}\mathbf{m}^{\mathrm{T}}P\mathbf{m}=o_P(1)$, (2) $\frac{\rho_n}{n}L^{\mathrm{T}}PG\mathbf{m}=o_P(1)$ for $L=\mathbf{m},\boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$, (3) $\sqrt{\frac{\rho_n}{n}}(G\boldsymbol{\epsilon})^{\mathrm{T}}P\mathbf{m}=o_P(1)$, (4) $\frac{\rho_n}{n}\Big\{(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}-E[(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}]\Big\}=o_P(1)$, (5) $\sqrt{\frac{\rho_n}{n}}\{\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}P\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}\boldsymbol{\epsilon}\}=o_P(1)$, (6) $\sqrt{\frac{\rho_n}{n}}\{\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}PG\boldsymbol{\epsilon}-\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}G\boldsymbol{\epsilon}\}=o_P(1)$, (7) $\sqrt{\frac{\rho_n}{n}}\{(G\mathbf{m})^{\mathrm{T}}P\boldsymbol{\epsilon}-(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon}\}=o_P(1)$.

Lemma 12. Suppose that $B=(b_{ij})_{1\le i,j\le n}$ is a sequence of symmetric matrices with row and column sums uniformly bounded. Let $\sigma^2_{Q_n}$ be the variance of Q_n where $Q_n=(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^{\mathrm{T}}B\boldsymbol{\epsilon}-\sigma_0^2\mathrm{tr}(B)$. Assume that the variance $\sigma^2_{Q_n}$ is $O(n/\rho_n)$ with $\{\frac{\rho_n}{n}\sigma^2_{Q_n}\}$ bounded away from zero, the elements of B are of uniform order $O(1/\rho_n)$ and the row sums of G of uniform order $O(1/\sqrt{\rho_n})$. Then we have under $\rho_n \to \infty$ and Conditions (3)-(5) that $\frac{Q_n}{\sigma_{Q_n}} \xrightarrow{D} N(0,1)$.

In the proofs of the theorems, we will use the facts that for constant matrices B and D, $\mathrm{var}(\boldsymbol{\epsilon}^{\mathrm{T}} B \boldsymbol{\epsilon}) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^{n} b_{ii}^2 + \sigma_0^4[\mathrm{tr}(BB^{\mathrm{T}}) + \mathrm{tr}(B^2)]$ and

$E(\boldsymbol{\epsilon}^{\mathrm{T}} B \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} D \boldsymbol{\epsilon}) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^{n} b_{ii}d_{ii} + \sigma_0^4[\mathrm{tr}(B)\mathrm{tr}(D) + \mathrm{tr}(BD) + \mathrm{tr}(BD^{\mathrm{T}})]$.

Moreover, we will frequently use the following facts by Condition (5) (see Lee, 2004) without clearly pointed out:

(1) the elements of $G = WA^{-1}$ are $O(1/\rho_n)$ uniformly in all $i,\ j$.

(2) The matrix $G = WA^{-1}$ is uniformly bounded in both row and column sums.

Proof of Theorem 1: First we will show that Ω is nonsingular. Let $\mathbf{d} = (d_1, d_2)^{\mathrm{T}}$ be a constant vector such that $\Omega\mathbf{d} = \mathbf{0}_2$. Then it is sufficient to show that $\mathbf{d} = \mathbf{0}_2$. From the second equation of $\Omega\mathbf{d} = \mathbf{0}_2$ we have that $d_2 = -2\sigma_0^2 \lim_{n\to\infty} \frac{1}{n}\mathrm{tr}(G)d_1$. Plug d_2 into the first equation of $\Omega\mathbf{d} = \mathbf{0}_2$ and we get that

$$d_1\Big\{\frac{1}{\sigma_0^2}\lambda_1 + \lim_{n\to\infty}\Big[\frac{1}{n}\mathrm{tr}((G+G^{\mathrm{T}})G) - \frac{2}{n^2}\mathrm{tr}^2(G)\Big]\Big\} = 0.$$

It follows by Condition (7) that $\lambda_1 > 0$. Moreover, $\mathrm{tr}\{(G+G^{\mathrm{T}})G\} - \frac{2}{n}\mathrm{tr}^2(G) = \frac{1}{2}\mathrm{tr}\{(\tilde{G}^{\mathrm{T}} + \tilde{G})(\tilde{G}^{\mathrm{T}} + \tilde{G})^{\mathrm{T}}\} \geq 0$ where $\tilde{G} = G - \frac{1}{n}\mathrm{tr}(G)I_n$. As we have by Condition (5) that the elements of $\tilde{G}$ are uniformly $O(1/\rho_n)$ and its row and column sums are also uniformly bounded, then it can be easily shown that $\mathrm{tr}\{(\tilde{G}^{\mathrm{T}} + \tilde{G})(\tilde{G}^{\mathrm{T}} + \tilde{G})^{\mathrm{T}}\} = O(\frac{n}{\rho_n})$.

Therefore, if Condition ($\tilde{7}$) holds, Condition (8) implies that the limit of $\frac{1}{n}\text{tr}((G+G^{\text{T}})G)-\frac{2}{n^2}\text{tr}^2(G) = \frac{1}{2n}\text{tr}\{(\tilde{G}^{\text{T}}+\tilde{G})(\tilde{G}^{\text{T}}+\tilde{G})^{\text{T}}\} > 0$. Therefore, $d_1 = 0$ and $d_2 = 0$.

Next we will follow the idea of Lee (2004) to show the consistency of $\hat{\alpha}$. Define $Q(\alpha)$ to be $\max\limits_{\sigma^2} E\Big[l(\alpha, \sigma^2)\Big]$ by ignoring the constant term. The optimal solutions of this maximization problem are $\bar{\sigma}^2(\alpha) = \frac{1}{n}E[(A(\alpha)Y)^{\text{T}}PA(\alpha)Y]$. Consequently,

$$Q(\alpha) = -n/2 \cdot \log \bar{\sigma}^2(\alpha) + \log|A(\alpha)|.$$

According to White (1994, Theorem 3.4), it suffices to show the uniform convergence of $n^{-1}\{l_c(\alpha) - Q(\alpha)\}$ to zero in probability on Δ and the unique maximizer condition that

$$\limsup_{n\to\infty} \max_{\alpha\in N^c(\alpha_0,\delta)} n^{-1}|Q(\alpha) - Q(\alpha_0)| < 0 \ \text{ for any } \delta > 0 \qquad (5.1)$$

where $N^c(\alpha_0, \delta)$ is the complement of an open neighborhood of α_0 in Δ with diameter δ.

Note that $\frac{1}{n}l_c(\alpha) - \frac{1}{n}Q(\alpha) = -\frac{1}{2}\{\log\tilde{\sigma}^2(\alpha) - \log\bar{\sigma}^2(\alpha)\}$, then to show the uniform convergence, it is sufficient to show that $\tilde{\sigma}^2(\alpha) - \bar{\sigma}^2(\alpha) = o_P(1)$ uniformly on Δ and $\bar{\sigma}^2(\alpha)$ is uniformly bounded away from zero on Δ. Since

$$\begin{aligned}
& \tilde{\sigma}^2(\alpha) - \bar{\sigma}^2(\alpha) \\
= \; & n^{-1}\Big\{(A(\alpha)A^{-1}\mathbf{m})^{\text{T}}PA(\alpha)A^{-1}\mathbf{m} - E[(A(\alpha)A^{-1}\mathbf{m})^{\text{T}}PA(\alpha)A^{-1}\mathbf{m}]\Big\} \\
& +n^{-1}\Big\{(A(\alpha)A^{-1}\boldsymbol{\epsilon})^{\text{T}}PA(\alpha)A^{-1}\boldsymbol{\epsilon} - \sigma_0^2 E[\text{tr}\{(A(\alpha)A^{-1})^{\text{T}}PA(\alpha)A^{-1}\}]\Big\} \\
& +2n^{-1}(A(\alpha)A^{-1}\mathbf{m})^{\text{T}}PA(\alpha)A^{-1}\boldsymbol{\epsilon},
\end{aligned}$$

and $A(\alpha)A^{-1} = I_n + (\alpha_0 - \alpha)G$ by $WA^{-1} = G$, it follows from Lemma 6 and Lemma 7(1) that

$$n^{-1}\Big\{(A(\alpha)A^{-1}\mathbf{m})^{\mathrm{T}}PA(\alpha)A^{-1}\mathbf{m} - E[(A(\alpha)A^{-1}\mathbf{m})^{\mathrm{T}}PA(\alpha)A^{-1}\mathbf{m}]\Big\} = o_P(1)$$

and

$$n^{-1}(A(\alpha)A^{-1}\mathbf{m})^{\mathrm{T}}PA(\alpha)A^{-1}\boldsymbol{\epsilon} = o_P(1).$$

Next, we have by Lemma 4(1)-(3), Lemma 8(1)-(3) and Chebyshev inequality that

$$n^{-1}\Big\{(A(\alpha)A^{-1}\boldsymbol{\epsilon})^{\mathrm{T}}PA(\alpha)A^{-1}\boldsymbol{\epsilon} - \sigma_0^2 E[\mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}}PA(\alpha)A^{-1}\}]\Big\} = o_P(1).$$

Therefore, $\tilde{\sigma}^2(\alpha) - \bar{\sigma}^2(\alpha) = o_P(1)$ uniformly on Δ.

Now we will show that $\bar{\sigma}^2(\alpha)$ is bounded away from zero uniformly on Δ. As we know by simple calculation and Lemma 4(1)-(3) that

$$\begin{aligned} \bar{\sigma}^2(\alpha) &\geq \sigma_0^2 n^{-1} E\Big[\mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}}PA(\alpha)A^{-1}\}\Big] \\ &= \sigma_0^2 n^{-1}\mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}}A(\alpha)A^{-1}\} + o(1), \end{aligned} \qquad (5.2)$$

it suffices to show that $\sigma_a^2(\alpha) = \frac{\sigma_0^2}{n}\mathrm{tr}\{(A(\alpha)A^{-1})^{\mathrm{T}}A(\alpha)A^{-1}\}$ is uniformly bounded away from zero on Δ. To do so, we define an auxiliary spatial autoregressive (SAR) process: $Y = \alpha_0 WY + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2 I_n)$. Then its log likelihood function without the constant term is

$$l_a(\alpha, \sigma^2) = -\frac{n}{2}\log\sigma^2 + \log|A(\alpha)| - \frac{1}{2\sigma^2}(A(\alpha)Y)^{\mathrm{T}}A(\alpha)Y.$$

Set $Q_a(\alpha)$ to be $\max\limits_{\sigma^2} E_a[l_a(\alpha, \sigma^2)]$ by ignoring the constant term, where E_a is the expectation under this SAR process. It can be easily shown that

$$Q_a(\alpha) = -n/2 \cdot \log \sigma_a^2(\alpha) + \log |A(\alpha)|,$$

By Jensen inequality, for all $\alpha \in \Delta$, $\max\limits_{\sigma^2} E_a[l_a(\alpha, \sigma^2)] \leq E_a[l_a(\alpha_0, \sigma_0^2)]$, thus $Q_a(\alpha) \leq Q_a(\alpha_0)$. As

$$\frac{1}{n}[Q_a(\alpha) - Q_a(\alpha_0)] = -\frac{1}{2} \log \sigma_a^2(\alpha) + \frac{1}{2} \log \sigma_0^2 + \frac{1}{n}\Big(\log |A(\alpha)| - \log |A(\alpha_0)|\Big)$$

uniformly on Δ, then it follows that

$$-\frac{1}{2} \log \sigma_a^2(\alpha) \leq -\frac{1}{2} \log \sigma_0^2 + \frac{1}{n}\Big(\log |A(\alpha_0)| - \log |A(\alpha)|\Big).$$

If we can show that

$$n^{-1}\{\log |A(\alpha_2)| - \log |A(\alpha_1)|\} = O(1) \ \text{ uniformly in } \alpha_1 \text{ and } \alpha_2 \text{ on } \Delta \tag{5.3}$$

then $-\frac{1}{2} \log \sigma_a^2(\alpha)$ is bounded from above for any $\alpha \in \Delta$. Therefore, the statement that $\sigma_a^2(\alpha)$ is uniformly bounded away from zero on Δ can be established by a counter argument.

Now we will verify (5.3), it follows by the mean value theorem and Condition (5)-(6) that

$$\begin{aligned} n^{-1}\{\log |A(\alpha_2)| - \log |A(\alpha_1)|\} &= -n^{-1}\mathrm{tr}\{WA^{-1}(\tilde{\alpha})\}(\alpha_2 - \alpha_1) \\ &= O(\rho_n^{-1})(\alpha_2 - \alpha_1) \qquad (5.4) \end{aligned}$$

where $\tilde{\alpha}$ lies between α_1 and α_2. (5.3) is then established by Δ being a bounded set.

To show the uniqueness condition (5.1), write

$$\begin{aligned} n^{-1}[Q(\alpha)-Q(\alpha_0)] &= n^{-1}[Q_a(\alpha)-Q_a(\alpha_0)] \\ &\quad +2^{-1}[\log\sigma_a^2(\alpha)-\log\bar{\sigma}^2(\alpha)]+2^{-1}[\log\bar{\sigma}^2(\alpha_0)-\log\sigma_0^2], \end{aligned}$$

it follows by Lemma 4(1) and Lemma 6(1) that $\bar{\sigma}^2(\alpha_0)-\sigma_0^2=\frac{1}{n}E[\mathbf{m}^{\mathrm{T}}P\mathbf{m}]+\sigma_0^2\frac{1}{n}E[\mathrm{tr}(P)]-\sigma_0^2=o(1)$. Hence, $\log\bar{\sigma}^2(\alpha_0)-\log\sigma_0^2=o(1)$ as $\bar{\sigma}^2(\alpha_0)$ and σ_0^2 are both bounded away from zero. Moreover, we have already shown in (5.2) that $\lim\limits_{n\to\infty}[\sigma_a^2(\alpha)-\bar{\sigma}^2(\alpha)]\leq 0$, hence,

$$\limsup_{n\to\infty}\max_{\alpha\in N^c(\alpha_0,\delta)} n^{-1}[Q(\alpha)-Q(\alpha_0)]\leq 0 \text{ for any } \delta>0.$$

Now we will show that the above inequality holds strictly. Because $\bar{\sigma}^2(\alpha)$ is bounded away from zero and has a quadratic form of α with its coefficients bounded by Lemma 4(1)-(3), 6 and 7(2), this together with (5.4), we get that $n^{-1}Q(\alpha)$ is uniformly equicontinuous in α on Δ.

By the compactness of $N^c(\alpha_0,\delta)$, we suppose there would exist an $\delta>0$ and a sequence $\{\alpha_n\}$ in $N^c(\alpha_0,\delta)$ converging to a point $\alpha^*\neq\alpha_0$ such that $\lim\limits_{n\to\infty}n^{-1}[Q(\alpha_n)-Q(\alpha_0)]=0$. Next, as $\alpha_n\to\alpha^*$, we have $\lim\limits_{n\to\infty}n^{-1}[Q(\alpha_n)-Q(\alpha^*)]=0$. Hence, it follows that

$$\lim_{n\to\infty}n^{-1}[Q(\alpha^*)-Q(\alpha_0)]=0. \tag{5.5}$$

Since we have known that $Q_a(\alpha^*)-Q_a(\alpha_0)\leq 0$ and $\lim\limits_{n\to\infty}[\sigma_a^2(\alpha^*)-\bar{\sigma}^2(\alpha^*)]\leq 0$, (5.5) is possible only if (i) $\lim\limits_{n\to\infty}[\sigma_a^2(\alpha^*)-\bar{\sigma}^2(\alpha^*)]=0$

and (ii) $\lim_{n\to\infty} n^{-1}[Q_a(\alpha^*) - Q_a(\alpha_0)] = 0$ both hold. However, (i) is a contradiction when Condition (7) holds as

$$\lim_{n\to\infty}[\sigma_a^2(\alpha^*) - \bar{\sigma}^2(\alpha^*)] = -(\alpha_0 - \alpha^*)^2 \lim_{n\to\infty} n^{-1}E[(G\mathbf{m} - Z)^{\mathrm{T}}(G\mathbf{m} - Z)] = 0,$$

by Lemma 4(1)-(3), 6 and 7(2). If Condition ($\tilde{7}$) holds, the contradiction follows from (ii) by Condition (8).

For the consistency of $\hat{\sigma}^2$, as it follows by some calculation, $A(\hat{\alpha})A^{-1} = I_n + (\alpha_0 - \hat{\alpha})G$, Lemma 6, 7, 8(1)-(3), Chebyshev inequality and $\hat{\alpha} \xrightarrow{P} \alpha_0$ that

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}(A(\hat{\alpha})Y - SA(\hat{\alpha})Y)^{\mathrm{T}}(A(\hat{\alpha})Y - SA(\hat{\alpha})Y) \\
&= \frac{1}{n}(A(\hat{\alpha})A^{-1}\mathbf{m})^{\mathrm{T}}PA(\hat{\alpha})A^{-1}\mathbf{m} + \frac{2}{n}(A(\hat{\alpha})A^{-1}\mathbf{m})^{\mathrm{T}}PA(\hat{\alpha})A^{-1}\boldsymbol{\epsilon} \\
&\quad + \frac{1}{n}(A(\hat{\alpha})A^{-1}\boldsymbol{\epsilon})^{\mathrm{T}}PA(\hat{\alpha})A^{-1}\boldsymbol{\epsilon} \\
&= \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}P\boldsymbol{\epsilon} + o_P(1) = \sigma_0^2 + o_P(1).
\end{aligned}$$

Proof of Theorem 2: Denote $\boldsymbol{\theta} = (\alpha, \sigma^2)^{\mathrm{T}}$ and $\boldsymbol{\theta}_0 = (\alpha_0, \sigma_0^2)^{\mathrm{T}}$, we get by Taylor expansion that

$$0 = \frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\sigma}^2)^{\mathrm{T}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ and thus converges to $\boldsymbol{\theta}_0$ in probability by Theorem 1. Then the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ can be obtained by showing that

$$-\frac{1}{n}\frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \xrightarrow{P} \Omega \quad \text{and} \quad \frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} \xrightarrow{D} N(\mathbf{0}, \Sigma + \Omega)$$

where Ω is a nonsingular matrix by Theorem 1.

By straightforward calculation, it can be easily obtained that

$$\begin{aligned} \tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha^2} &= -\tfrac{1}{n}\mathrm{tr}([WA^{-1}(\alpha)]^2) - \tfrac{1}{\sigma^2 n}(WY)^{\mathrm{T}}PWY, \\ \tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma^2\partial\sigma^2} &= \tfrac{1}{2\sigma^4} - \tfrac{1}{\sigma^6 n}(A(\alpha)Y)^{\mathrm{T}}PA(\alpha)Y, \\ \tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha\partial\sigma^2} &= -\tfrac{1}{\sigma^4 n}(WY)^{\mathrm{T}}PA(\alpha)Y. \end{aligned} \tag{5.6}$$

As $A(\tilde{\alpha})A^{-1} = I_n + (\alpha_0 - \tilde{\alpha})G$ by $G = WA^{-1}$, we have

$$\frac{1}{n}\frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial\sigma^2\partial\sigma^2} - \frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\sigma^2\partial\sigma^2} = o_P(1) \quad \text{and} \quad \frac{1}{n}\frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial\alpha\partial\sigma^2} - \frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\alpha\partial\sigma^2} = o_P(1).$$

using Lemma 6, 7, 8(1)-(3), Chebyshev inequality and $\tilde{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$. Let $G(\alpha) = WA^{-1}(\alpha)$, then it follows by the mean value theorem that

$$\begin{aligned} & \frac{1}{n}\frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial\alpha^2} - \frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\alpha^2} \\ = \quad & -\frac{2}{n}\mathrm{tr}(G^3(\bar{\alpha}))(\tilde{\alpha} - \alpha_0) + \left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}^2}\right)\frac{1}{n}(G\mathbf{m} + G\boldsymbol{\epsilon})^{\mathrm{T}}P(G\mathbf{m} + G\boldsymbol{\epsilon}) \end{aligned}$$

for some $\bar{\alpha}$ between $\tilde{\alpha}$ and α_0. Note that $G(\alpha)$ is bounded in row and column sums uniformly in a neighborhood of α_0 by Condition (5)-(6). Therefore, $\frac{1}{n}\mathrm{tr}(G^3(\bar{\alpha})) = O(1/\rho_n)$. Since we have $\frac{1}{n}(G\mathbf{m} + G\boldsymbol{\epsilon})^{\mathrm{T}}P(G\mathbf{m} + G\boldsymbol{\epsilon}) = O_P(1)$ by Lemma 6(2), 7, 8(3) and Markov inequality, it follows that $\frac{1}{n}\frac{\partial^2 l(\tilde{\boldsymbol{\theta}})}{\partial\alpha^2} - \frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\alpha^2} = o_P(1)$ by $\tilde{\alpha} \xrightarrow{P} \alpha_0$ and $\tilde{\sigma}^2 \xrightarrow{P} \sigma_0^2$.

Next input $\boldsymbol{\theta}_0$ into (5.6) and we can get by Lemma 6 that

$$
\begin{aligned}
-\tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\alpha^2} &= \tfrac{1}{n}\mathrm{tr}(G^2) + \tfrac{1}{\sigma_0^2 n}(G\mathbf{m})^{\mathrm{T}}P(G\mathbf{m}) + \tfrac{1}{\sigma_0^2 n}\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}PG\boldsymbol{\epsilon} + o_P(1),\\
-\tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\sigma^2\partial\sigma^2} &= -\tfrac{1}{2\sigma_0^4} + \tfrac{1}{\sigma_0^6 n}\boldsymbol{\epsilon}^{\mathrm{T}}P\boldsymbol{\epsilon} + o_P(1),\\
-\tfrac{1}{n}\tfrac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\alpha\partial\sigma^2} &= \tfrac{1}{\sigma_0^4 n}\boldsymbol{\epsilon}^{\mathrm{T}}G^{\mathrm{T}}P\boldsymbol{\epsilon} + o_P(1).
\end{aligned}
$$

Thus, the result of $-\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \xrightarrow{P} \Omega$ can be obtained using Lemma 7, Lemma 8(1)-(3) and Chebyshev inequality.

In the following we will establish the asymptotic distribution of $\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}$. It follows by Lemma 5(2) that $\frac{1}{\sqrt{n}}(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = O_P(n^{1/2}h^3 + \{h^2\log n\}^{1/2}) = o_P(1)$ when $nh^6 \to 0$ and $h^2\log n \to 0$. Then we have by straightforward calculation, Lemma 6(1) and Lemma 8 that

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\alpha} &= -\frac{1}{\sqrt{n}}\mathrm{tr}(G) + \frac{1}{\sigma_0^2\sqrt{n}}(WY)^{\mathrm{T}}PAY\\
&= \frac{1}{\sigma_0^2\sqrt{n}}\Big[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \{\boldsymbol{\epsilon}^{\mathrm{T}}G\boldsymbol{\epsilon} - \sigma_0^2\mathrm{tr}(G)\}\Big] + o_P(1),
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\sigma^2} &= -\frac{\sqrt{n}}{2\sigma_0^2} + \frac{1}{2\sigma_0^4\sqrt{n}}(AY)^{\mathrm{T}}PAY\\
&= \frac{1}{2\sigma_0^4\sqrt{n}}\{\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - n\sigma_0^2\} + o_P(1).
\end{aligned}
$$

Next we have by straightforward calculation that

$$
\begin{aligned}
&\mathrm{var}((G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \{\boldsymbol{\epsilon}^{\mathrm{T}}G\boldsymbol{\epsilon} - \sigma_0^2\mathrm{tr}(G)\})\\
= \;&\sigma_0^2 E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}(G\mathbf{m} - SG\mathbf{m})] + (\mu_4 - 3\sigma_0^4)\sum_{i=1}^{n} g_{ii}^2 + \sigma_0^4[\mathrm{tr}(GG^{\mathrm{T}}) + \mathrm{tr}(G^2)]\\
&+2\mu_3 E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}G_c],
\end{aligned}
$$

$\text{var}(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - n\sigma_0^2) = n(\mu_4 - \sigma_0^4)$ and

$$\begin{aligned} & \text{cov}\{(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \{\boldsymbol{\epsilon}^{\mathrm{T}}G\boldsymbol{\epsilon} - \sigma_0^2\text{tr}(G)\}, \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - n\sigma_0^2\} \\ = \; & \mu_3 E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\mathbf{1}_n] + (\mu_4 - \sigma_0^4)\text{tr}(G). \end{aligned}$$

Hence, it follows by Lemma 7(2) and some calculation that

$$E\left(\frac{1}{n}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right) = \Sigma + \Omega + o(1).$$

As the components of $\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \left(\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \alpha}, \frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \sigma^2}\right)^{\mathrm{T}}$ are linear-quadratic forms of double arrays, using Lemma 9 we gain $\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{D} N(\mathbf{0}, \Sigma + \Omega)$.

Proof of Theorem 3: It can be easily shown that

$$\begin{aligned} \sqrt{nh_1^2 f(s)}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}(s)) &= \sqrt{nh_1^2 f(s)}(I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1)^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\boldsymbol{\epsilon} \\ &\quad + \sqrt{nh_1^2 f(s)}(\alpha_0 - \hat{\alpha})(I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1)^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1 WY \\ &\quad + \sqrt{nh_1^2 f(s)}(I_p, \mathbf{0}_{p\times 2p})\{(\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1)^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathbf{m} - \boldsymbol{\beta}(s)\} \\ &\equiv J_{n1} + J_{n2} + J_{n3} \end{aligned}$$

where $\mathcal{X}_1$ and $\mathcal{W}_1$ are $\mathcal{X}$ and $\mathcal{W}$ respectively with h replaced by h_1.

Let H_1 be H with h replaced by h_1. It follows by straightforward calculation that

$$\sqrt{n^{-1}h_1^2 f(s)}E\{H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\boldsymbol{\epsilon}\} = \mathbf{0}_{3p\times 1},$$

and

$$\begin{aligned} & n^{-1}h_1^2 f(s)\mathrm{cov}\{H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\boldsymbol{\epsilon}\} = \sigma_0^2 n^{-1}h_1^2 f(s)E\{H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1^2\mathcal{X}_1 H_1^{-1}\} \\ = \;& \sigma_0^2 f^2(s)\begin{pmatrix} \nu_0\Psi + o_P(\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}) & o_P(\mathbf{1}_p\mathbf{1}_{2p}^{\mathrm{T}}) \\ o_P(\mathbf{1}_{2p}\mathbf{1}_p^{\mathrm{T}}) & \nu_2\Psi\otimes I_2 + o_P(\mathbf{1}_{2p}\mathbf{1}_{2p}^{\mathrm{T}}) \end{pmatrix} \end{aligned}$$

then it follows by central limit theorem, Lemma 2(1) and Slutsky's Theorem that

$$J_{n1} \xrightarrow{D} N\Big(\mathbf{0}, \nu_0\kappa_0^{-2}\sigma_0^2\Psi^{-1}\Big).$$

Moreover, it follows immediately from Lemma 3 that

$$n^{-1}\{H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1 G(\mathbf{m}+\boldsymbol{\epsilon}) - E[H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1 G(\mathbf{m}+\boldsymbol{\epsilon})]\} = o_P(1)$$

This together with Lemma 2(1) and Condition (4) leads to

$$(I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1)^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1 G(\mathbf{m}+\boldsymbol{\epsilon}) = O_P(1).$$

Next when $nh_1^6 = O(1)$ and $h/h_1 \to 0$, we have $\sqrt{\frac{h_1^2}{n}}(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(n^{1/2}h_1h^2) = o_P(1)$ using Lemma 5(1). Hence it can be seen from the proof of Theorem 2 that $\sqrt{\frac{h_1^2}{n}}\frac{\partial l(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = o_P(1)$ and $\sqrt{nh_1^2}(\hat{\alpha}-\alpha_0) = o_P(1)$ under the assumptions of Theorem 3. Therefore,

$$J_{n2} = \sqrt{f(s)}\sqrt{nh_1^2}(\alpha_0-\hat{\alpha})(I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1)^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1(G\mathbf{m}+G\boldsymbol{\epsilon}) = o_P(1).$$

For J_{n3}, it can be obtained by Lemma 2(2) that

$$J_{n3} = \frac{\kappa_2 h_1^2}{2\kappa_0}\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\} + o_P(h_1^2\mathbf{1}_p).$$

Finally combing the results of J_{n1}, J_{n2} and J_{n3}, when $nh_1^6 = O(1)$ and $h/h_1 \to 0$ we get the theorem.

Proof of Theorem 4: It is obvious from the proof of nonsigularity of Ω in Theorem 1 that under Condition (9), Ω is singular.

Next like Lee (2004), to prove the consistency of $\hat{\alpha}$, it suffices to show that

$$\frac{\rho_n}{n}\{l_c(\alpha) - l_c(\alpha_0) - [Q(\alpha) - Q(\alpha_0)]\} = o_P(1) \text{ uniformly on } \Delta,$$

where $Q(\alpha) = -n/2 \cdot \log \bar{\sigma}^2(\alpha) + \log|A(\alpha)|$ defined as in the proof of Theorem 1 and α_0 is a unique maximizer.

It follows by the mean value theorem that

$$\begin{aligned} & \frac{\rho_n}{n}\{l_c(\alpha) - l_c(\alpha_0) - [Q(\alpha) - Q(\alpha_0)]\} \\ = \quad & -\frac{\rho_n}{2}\frac{\partial[\log \tilde{\sigma}^2(\tilde{\alpha}) - \log \bar{\sigma}^2(\tilde{\alpha})]}{\partial \alpha}(\alpha - \alpha_0) \\ = \quad & \frac{1}{\tilde{\sigma}^2(\tilde{\alpha})}\frac{\rho_n}{n}\left\{[(WY)^{\mathrm{T}}PA(\tilde{\alpha})Y - L_n(\tilde{\alpha})] - \frac{\tilde{\sigma}^2(\tilde{\alpha}) - \bar{\sigma}^2(\tilde{\alpha})}{\bar{\sigma}^2(\tilde{\alpha})}L_n(\tilde{\alpha})\right\}(\alpha - \alpha_0) \end{aligned}$$

where $\tilde{\alpha}$ lies between α and α_0, and $L_n(\tilde{\alpha}) = E[(WY)^{\mathrm{T}}PA(\tilde{\alpha})Y]$.

Note that $A(\tilde{\alpha})A^{-1} = I_n + (\alpha_0 - \tilde{\alpha})G$, by applying Lemma 4(5)(6), 11 and Chebyshev inequality we can get

$$\frac{\rho_n}{n}\{(WY)^{\mathrm{T}}PA(\tilde{\alpha})Y - L_n(\tilde{\alpha})\} = o_P(1) \text{ and } \frac{\rho_n}{n}L_n(\tilde{\alpha}) = O(1).$$

Moreover, using the same lines as in the proof of Theorem 1, we can establish that $\tilde{\sigma}^2(\tilde{\alpha}) - \bar{\sigma}^2(\tilde{\alpha}) = o_P(1)$ for any $\tilde{\alpha}$ on Δ with $\bar{\sigma}^2(\alpha)$ being uniformly bounded away from zero on Δ. Thus $\tilde{\sigma}^2(\alpha)$ is uniformly bounded away from zero in probability. Consequently,

$$\frac{\rho_n}{n}\{l_c(\alpha)-l_c(\alpha_0)-[Q(\alpha)-Q(\alpha_0)]\}=o_P(1) \text{ uniformly on } \Delta.$$

The uniqueness condition of α_0 can be obtained by the uniform equicontinuity of $\frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)]$ on Δ and $\lim_{n\to\infty}\frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)]<0$ when $\alpha\neq\alpha_0$ using a counter argument as in the proof of Theorem 1.

Write

$$\begin{aligned}\frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)] &= -\frac{\rho_n}{2}[\log\bar\sigma^2(\alpha)-\log\bar\sigma^2(\alpha_0)]+\frac{\rho_n}{n}[\log|A(\alpha)|-\log|A(\alpha_0)|] \\ &\equiv -\frac{1}{2}J_{n1}+J_{n2}.\end{aligned}$$

It follows by the mean value theorem

$$J_{n1}=\frac{\rho_n}{\bar\sigma^{*2}(\alpha)}(\bar\sigma^2(\alpha)-\bar\sigma^2(\alpha_0))$$

where $\bar\sigma^{*2}(\alpha)$ lies between $\bar\sigma^2(\alpha)$ and $\bar\sigma^2(\alpha_0)$. As $\bar\sigma^2(\alpha)$ is uniformly bounded away from zero on Δ, $\bar\sigma^{*2}(\alpha)$ is also uniformly bounded away from zero on Δ. Further, we can see by Lemma 4(5)(6) and Lemma 11 that $\rho_n(\bar\sigma^2(\alpha)-\bar\sigma^2(\alpha_0))$ is a quadratic form of α with its coefficients bounded. Therefore, J_{n1} is uniformly equicontinuious on Δ by the above results.

For J_{n2}, it can be seen by the mean value theorem that

$$J_{n2}=-\frac{\rho_n}{n}\text{tr}(WA^{-1}(\tilde\alpha))(\alpha-\alpha_0)$$

where $\tilde\alpha$ lies between α and α_0, and $\text{tr}(WA^{-1}(\tilde\alpha))=O\Big(n/\rho_n\Big)$ by

Condition (5)-(6). Therefore, J_{n2} is uniformly equicontinuous on Δ.

In conclusion, $\frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)]$ is uniformly equicontinuous on Δ.

Next we will show that when $\alpha \neq \alpha_0$, $\lim\limits_{n\to\infty} \frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)] < 0$. Using similar lines as in the proof of Theorem 1, let $Q_a(\alpha) = -\frac{n}{2}\log\sigma_a^2(\alpha) + \log|A(\alpha)|$, and write

$$\begin{aligned}\frac{\rho_n}{n}[Q(\alpha)-Q(\alpha_0)] &= \frac{\rho_n}{n}[Q_a(\alpha)-Q_a(\alpha_0)] - \frac{\rho_n}{2}[\log\bar{\sigma}^2(\alpha) - \log\sigma_a^2(\alpha)] \\ &\quad + \frac{\rho_n}{2}[\log\bar{\sigma}^2(\alpha_0) - \log\sigma_0^2]. \end{aligned} \tag{5.7}$$

As it follows by the mean value theorem, Lemma 4(4)-(6) and Lemma 11(1)(2) that

$$\begin{aligned}-\frac{\rho_n}{2}[\log\bar{\sigma}^2(\alpha) - \log\sigma_a^2(\alpha)] &= -\frac{\rho_n}{2\sigma^{*2}(\alpha)}[\bar{\sigma}^2(\alpha) - \sigma_a^2(\alpha)] \\ &= -\frac{1}{2\sigma^{*2}(\alpha)}(\alpha_0-\alpha)^2\frac{\rho_n}{n}E[(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}] + o(1)\end{aligned}$$

where $\sigma^{*2}(\alpha)$ lies between $\bar{\sigma}^2(\alpha)$ and $\sigma_a^2(\alpha)$ and it therefore uniformly bounded away from zero on Δ. Then for any $\alpha \neq \alpha_0$, when condition (9) holds, $-\frac{\rho_n}{2}[\log\bar{\sigma}^2(\alpha) - \log\sigma^2(\alpha)] < 0$ for sufficient large n.

For the third term on the right side in (5.7), it can be obtained by the mean value theorem, Lemma 4(4) and Lemma 11(1) that

$$\frac{\rho_n}{2}[\log\bar{\sigma}^2(\alpha_0) - \log\sigma_0^2] = \frac{\rho_n}{2\sigma^{*2}}\{\bar{\sigma}^2(\alpha_0) - \sigma_0^2\} = o(1)$$

where σ^{*2} lies between $\bar{\sigma}^2(\alpha_0)$ and σ_0^2, and is bounded away from zero.

In consequence, $\lim\limits_{n\to\infty} \frac{\rho_n}{n}\{Q(\alpha)-Q(\alpha_0)\} < 0$ when $\alpha \neq \alpha_0$, as we have shown $Q_a(\alpha) - Q_a(\alpha_0) \leq 0$ in the proof of Theorem 1.

Proof of Theorem 5: By Taylor expansion, we have that

$$0 = \frac{\partial l_c(\hat{\alpha})}{\partial \alpha} = \frac{\partial l_c(\alpha_0)}{\partial \alpha} + \frac{\partial^2 l_c(\tilde{\alpha})}{\partial \alpha^2}(\hat{\alpha} - \alpha_0)$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0, and thus converges to α_0 in probability by Theorem 4. Then the asymptotic distribution of $\hat{\alpha}$ can be obtained by proving that when $\rho_n \to \infty$,

$$-\frac{\rho_n}{n}\frac{\partial^2 l_c(\tilde{\alpha})}{\partial \alpha^2} \xrightarrow{P} \sigma_1^2 \text{ and } \sqrt{\frac{\rho_n}{n}}\frac{\partial l_c(\alpha_0)}{\partial \alpha} \xrightarrow{D} N(0, \sigma_2^2/\sigma_0^4),$$

where $\sigma_1^2 = \frac{1}{\sigma_0^2}\lim\limits_{n\to\infty}\frac{\rho_n}{n}E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}(G\mathbf{m} - SG\mathbf{m})]$ and $\sigma_2^2 = \sigma_0^4\sigma_1^2$.

As we have by $A(\alpha)A^{-1} = I_n + (\alpha_0 - \alpha)G$, Lemma 11 and Chebyshev inequality that $\frac{\rho_n}{n}(WY)^{\mathrm{T}}PWY = \frac{\rho_n}{n}(G\mathbf{m} + G\boldsymbol{\epsilon})^{\mathrm{T}}P(G\mathbf{m} + G\boldsymbol{\epsilon}) = O_P(1)$ and $\frac{\rho_n}{n}(WY)^{\mathrm{T}}PA(\alpha)Y = O_P(1)$, then when $\rho_n \to \infty$,

$$\begin{aligned}
&\frac{\rho_n}{n}\frac{\partial^2 l_c(\alpha)}{\partial \alpha^2} \\
= \;& \frac{\rho_n}{n}\Big\{\frac{2}{\tilde{\sigma}^4(\alpha)n}[(WY)^{\mathrm{T}}PA(\alpha)Y]^2 - \frac{1}{\tilde{\sigma}^2(\alpha)}(WY)^{\mathrm{T}}PWY - \mathrm{tr}([WA^{-1}(\alpha)]^2)\Big\} \\
= \;& -\frac{1}{\tilde{\sigma}^2(\alpha)}\cdot\frac{\rho_n}{n}(WY)^{\mathrm{T}}PWY - \frac{\rho_n}{n}\mathrm{tr}([WA^{-1}(\alpha)]^2) + o_P(1).
\end{aligned}$$

Further using Lemma 6(1), 8(1) and the above results, we can get when $\rho_n \to \infty$ that

$$\tilde{\sigma}^2(\alpha) = \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}P\boldsymbol{\epsilon} + o_P(1) = \sigma_0^2 + o_P(1)$$

for any $\alpha \in \Delta$. Therefore it follows by the mean value theorem that

$$\begin{aligned}
& \frac{\rho_n}{n}\left\{\frac{\partial^2 l_c(\tilde{\alpha})}{\partial\alpha^2} - \frac{\partial^2 l_c(\alpha_0)}{\partial\alpha^2}\right\} \\
= & \Big\{\frac{1}{\tilde{\sigma}^2(\alpha_0)} - \frac{1}{\tilde{\sigma}^2(\tilde{\alpha})}\Big\}\frac{\rho_n}{n}(WY)^{\mathrm{T}}PWY - \frac{\rho_n}{n}\{\mathrm{tr}(G^2(\tilde{\alpha})) - \mathrm{tr}(G^2(\alpha_0))\} + o_P(1) \\
= & -\frac{\rho_n}{n}\mathrm{tr}(G^3(\bar{\alpha}))(\tilde{\alpha} - \alpha_0) + o_P(1)
\end{aligned}$$

where $G(\alpha) = WA^{-1}(\alpha)$. As $\mathrm{tr}(G^3(\bar{\alpha})) = O(n/\rho_n)$ uniformly on Δ by Condition (5)-(6), we obtain that $\frac{\rho_n}{n}\left\{\frac{\partial^2 l_c(\tilde{\alpha})}{\partial\alpha^2} - \frac{\partial^2 l_c(\alpha_0)}{\partial\alpha^2}\right\} = o_P(1)$ using $\tilde{\alpha} \xrightarrow{P} \alpha_0$.

Next it follows from $\tilde{\sigma}^2(\alpha_0) \xrightarrow{P} \sigma_0^2$, Lemma 11 and Chebyshev inequality that

$$-\frac{\rho_n}{n}\frac{\partial^2 l_c(\alpha_0)}{\partial\alpha^2} = \frac{1}{\sigma_0^2}\frac{\rho_n}{n}E[(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}] + \frac{\rho_n}{n}[\mathrm{tr}(G^2) + \mathrm{tr}(GG^{\mathrm{T}})] + o_P(1).$$

Therefore, $-\frac{\rho_n}{n}\frac{\partial^2 l_c(\tilde{\alpha})}{\partial\alpha^2} \xrightarrow{P} \sigma_1^2$ by the row sums of G being uniform order $O(1/\sqrt{\rho_n})$.

In the following we will establish the asymptotic distribution of $\sqrt{\frac{\rho_n}{n}}\frac{\partial l_c(\alpha_0)}{\partial\alpha}$. As it follows that $\sqrt{\frac{\rho_n}{n}}(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(n^{1/2}h^3 + \{h^2\log n\}^{1/2}) = o_P(1)$ when $nh^6 \to 0$, $h^2\log n \to 0$ by Lemma 10(2) and $\sqrt{\frac{\rho_n}{n}}(G\boldsymbol{\epsilon})^{\mathrm{T}}P\mathbf{m} = o_P(1)$ by Lemma 11(3). Then we have by straightforward calculation and Lemma 6(1), 8(1), 11(5)(7) that the first order derivative of $\sqrt{\frac{\rho_n}{n}}l_c(\alpha)$ at α_0 is

$$\sqrt{\frac{\rho_n}{n}}\frac{\partial l_c(\alpha_0)}{\partial\alpha} = \frac{1}{\tilde{\sigma}^2(\alpha_0)}\sqrt{\frac{\rho_n}{n}}\big\{(WY)^{\mathrm{T}}PAY - \tilde{\sigma}^2(\alpha_0)\mathrm{tr}(G)\big\},$$

with

$$\begin{aligned} & \sqrt{\frac{\rho_n}{n}}\{(WY)^{\mathrm{T}}PAY - \tilde{\sigma}^2(\alpha_0)\mathrm{tr}(G)\} \\ = & \sqrt{\frac{\rho_n}{n}}\Big\{(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}[G - \frac{1}{n}\mathrm{tr}(G)I_n]\boldsymbol{\epsilon}\Big\} + o_P(1), \end{aligned}$$

and

$$\begin{aligned} \sigma_{qn}^2 & \equiv \mathrm{var}\Big\{(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}[G - \frac{1}{n}\mathrm{tr}(G)I_n]\boldsymbol{\epsilon}\Big\} \\ & = \sigma_0^2 E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}(G\mathbf{m} - SG\mathbf{m})] + (\mu_4 - 3\sigma_0^4)\sum_{i=1}^{n}\{g_{ii} - \frac{\mathrm{tr}(G)}{n}\}^2 \\ & \quad + \sigma_0^4[\mathrm{tr}((G + G^{\mathrm{T}})G) - \frac{2}{n}\mathrm{tr}^2(G)] + 2\mu_3 E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}(G_c - \frac{1}{n}\mathrm{tr}(G)\mathbf{1}_n)]. \end{aligned}$$

As we have by Lemma 12 that

$$\sigma_{qn}^{-1}\Big\{(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}[G^{\mathrm{T}} - \frac{1}{n}\mathrm{tr}(G)I_n]\boldsymbol{\epsilon}\Big\} \xrightarrow{D} N(0, 1),$$

it follows that

$$\sqrt{\frac{n}{\rho_n}}(\hat{\alpha} - \alpha_0) = \Big(-\frac{\rho_n}{n}\frac{\partial^2 l_c(\tilde{\alpha})}{\partial \alpha^2}\Big)^{-1} \cdot \sqrt{\frac{\rho_n}{n}}\frac{\partial l_c(\alpha_0)}{\partial \alpha} \xrightarrow{D} N\Big(0, \sigma_0^2\lambda_4^{-1}\Big).$$

by $\frac{\rho_n}{n}\sigma_{qn}^2 \to \sigma_2^2$ and $\tilde{\sigma}^2(\alpha_0) \xrightarrow{P} \sigma_0^2$.

Proof of Theorem 6: By straightforward calculation, Lemma 6(1), Lemma 8(1), 11, Chebyshev inequality and Theorem 5, we get when $\rho_n \to \infty$ that

$$\sqrt{n}(\hat{\sigma}^2 - \sigma_0^2) = \frac{1}{\sqrt{n}}(A(\hat{\alpha})Y - SA(\hat{\alpha})Y)^{\mathrm{T}}(A(\hat{\alpha})Y - SA(\hat{\alpha})Y) - \sqrt{n}\sigma_0^2$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}}(\mathbf{m}+\boldsymbol{\epsilon})^{\mathrm{T}}P(\mathbf{m}+\boldsymbol{\epsilon}) - \sqrt{n}\sigma_0^2 \\
&\quad + \frac{2}{\sqrt{\rho_n}}\sqrt{\frac{n}{\rho_n}}(\alpha_0-\hat{\alpha})\frac{\rho_n}{n}(G\mathbf{m}+G\boldsymbol{\epsilon})^{\mathrm{T}}P(\mathbf{m}+\boldsymbol{\epsilon}) \\
&\quad + \frac{1}{\sqrt{n}}\{\sqrt{\frac{n}{\rho_n}}(\alpha_0-\hat{\alpha})\}^2\frac{\rho_n}{n}(G\mathbf{m}+G\boldsymbol{\epsilon})^{\mathrm{T}}P(G\mathbf{m}+G\boldsymbol{\epsilon}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\epsilon_i^2-\sigma_0^2)+o_P(1)
\end{aligned}
$$

This together with central limit theorem for iid random variables leads to

$$\sqrt{n}(\hat{\sigma}^2-\sigma_0^2) \xrightarrow{D} N(0,\mu_4-\sigma_0^4).$$

Proof of Theorem 7: The result can be obtained using the same lines as the proof of Theorem 3, except that here $J_{n2} = \sqrt{f(s)}\sqrt{\frac{nh_1^2}{\rho_n}}(\alpha_0 - \hat{\alpha})(I_p, \mathbf{0}_{p\times 2p})(n^{-1}H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1H_1^{-1})^{-1}\frac{\sqrt{\rho_n}}{n}H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1G(\mathbf{m}+\boldsymbol{\epsilon})$. It follows by Lemma 2(1), Markov inequality, the row sums of the matrix G having uniform order $O(1/\sqrt{\rho_n})$ and Condition (4) that

$$(I_p, \mathbf{0}_{p\times 2p})(n^{-1}H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1\mathcal{X}_1H_1^{-1})^{-1}\frac{\sqrt{\rho_n}}{n}H_1^{-1}\mathcal{X}_1^{\mathrm{T}}\mathcal{W}_1G(\mathbf{m}+\boldsymbol{\epsilon}) = O_P(1).$$

Next, it can be seen from the proof of Theorem 5 and Lemma 10(1) that $\sqrt{\frac{\rho_n h_1^2}{n}}(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(n^{1/2}h_1h^2) = o_P(1)$ when $nh_1^6 = O(1)$ and $h/h_1 \to 0$. Hence, $\sqrt{\frac{nh_1^2}{\rho_n}}(\hat{\alpha}-\alpha) \xrightarrow{P} 0$ according to the arguments establishing Theorem 5. Consequently we have that $J_{n2} = o_P(1)$.

6 Proofs of Lemmas

In this chapter, we set $m_i = X_i^{\mathrm{T}} \boldsymbol{\beta}(s_i)$, x_{il} to be the lth $(l = 1, \cdots, p)$ element of X_i, $i = 1, \cdots, n$, $r_n = (\frac{\log n}{nh^2})^{1/2}$, $[D]_{ij}$ to be the (i, j)th elements of the matrix D, and c as a positive finite constant that may take different values at each appearance. Moreover, the operator Vec$(\cdot)$ creates a column vector from the matrix by simply stacking its column vectors below one another.

Frequently we will use the facts (see Lee, 2004) without clearly pointing out that the matrix G is uniformly bounded in both row and column sums, and the elements g_{ij} of G are $O(1/\rho_n)$ uniformly in all i, j.

Proof of Lemma 1: Let $\tau_n = n^{1/q}(\log n)^{1/2}$ and the following proof is organized as Hansen (2008). First, we deal with the truncation error in replacing Y_i with the truncated process $Y_i 1(|Y_i| \le \tau_n)$. Second, we replace the supremum with a maximization over a finite N-point grid. Third, we use Bernstein inequality to bound the remainder.

The first step is to truncate Y_i. Define $R(s) = \frac{1}{n}\sum\limits_{i=1}^{n} K_h(\|s_i - s\|)Y_i 1(|Y_i| > \tau_n)$. Since $P(|Y_n| > \tau_n) \le \tau_n^{-q} E|Y_n|^q$ and $\sum\limits_{n=1}^{\infty} \tau_n^{-q} = \sum\limits_{n=1}^{\infty} n^{-1}(\log n)^{-q/2} < \infty$ for $q > 2$. It follows that with probability one $|Y_n| \le \tau_n$ for all sufficient large n. Since τ_n is increasing, we have for all sufficient large n, $|Y_i| \le \tau_n$ for all $i \le n$. This implies that $\sup_s |R(s)|$ is eventually zero with probability one.

Next by a standard argument and Condition (4)

$$E[R(s)] \le \frac{1}{n}\sum_{i=1}^{n} K_h(\|s_i - s\|)E|Y_i|^q/\tau_n^{q-1} \le c\tau_n^{1-q},$$

it follows that with probability one $\sup_s E|R(s)| = O(\tau_n^{1-q}) = O(r_n)$.

Combing the above results, we have that with probability one

$$\sup_{s\in\mathcal{S}} |R(s) - ER(s)| = O(r_n).$$

For the second step we create a grid to cover the region $\mathcal{S}$. As $\mathcal{S}$ is a compact region, we can find a finite positive constant c_1 such that $\mathcal{S} \subseteq \{s : \|s\| \leq c_1\}$. Next we create a grid using regions of the form $N_l = \{s : \|s - s_l\| \leq r_n h\}$. By selecting s_l to lay on a grid, the region $\{s : \|s\| \leq c_1\}$ can be covered with $N \leq c_1^2 h^{-2} r_n^{-2}$ such regions N_l. Therefore the supremum can be replaced by a maximization over these N-point grid.

From the assumption of the kernel function, we know that there exists a finite positive constant L, when $\|s\| > L$, $K(\|s\|) = 0$, and there exists a finite positive constant c_2 such that for all $s, s' \in R^2$, $|K(\|s\|) - K(\|s'\|)| \leq c_2 \Big| \|s\| - \|s'\| \Big| \leq c_2 \|s - s'\|$. Define $W^*(\|s\|) = c_2 I(\|s\| \leq 2L)$, thus for $s \in N_l$, we have $\|\frac{s-s_l}{h}\| \leq r_n$ and

$$|K(\|\frac{s_i - s}{h}\|) - K(\|\frac{s_i - s_l}{h}\|)| \leq r_n W^*(\|\frac{s_i - s_l}{h}\|). \qquad (6.1)$$

Now define $R_1(s) = \frac{1}{n}\sum\limits_{i=1}^{n} K_h(\|s_i - s\|) Y_i 1(|Y_i| \leq \tau_n)$ and $\tilde{R}_1(s) = \frac{1}{n}\sum\limits_{i=1}^{n} W_h^*(\|s_i - s\|)|Y_i| 1(|Y_i| \leq \tau_n)$ where $W_h^*(\|s_i - s\|) = W^*(\|\frac{s_i - s}{h}\|)/h^2$. Note that $E|\tilde{R}_1(s)| \leq \frac{1}{n}\sum\limits_{i=1}^{n} W_h^*(\|s_i - s\|) E|Y_i| < c_3$ for some positive constant c_3 by Condition (4). Then we have by (6.1) that

$$\sup_{s\in N_l} |R_1(s) - ER_1(s)| \;\; \leq \;\; |R_1(s_l) - ER_1(s_l)| + r_n[|\tilde{R}_1(s_l)| + E|\tilde{R}_1(s_l)|]$$

$$\begin{aligned} &\leq |R_1(s_l) - ER_1(s_l)| + r_n|\tilde{R}_1(s_l) - E\tilde{R}_1(s_l)| + 2r_nE|\tilde{R}_1(s_l)| \\ &\leq |R_1(s_l) - ER_1(s_l)| + |\tilde{R}_1(s_l) - E\tilde{R}_1(s_l)| + 2c_3r_n \end{aligned}$$

with the final inequality because $r_n \leq 1$ for sufficient large n. Therefore, for sufficient large n

$$\begin{aligned} P(\sup_{s\in\mathcal{S}} |R_1(s) - ER_1(s)| > 4c_3r_n) &\leq N \max_{1\leq l\leq N} P(\sup_{s\in N_l} |R_1(s) - ER_1(s)| > 4c_3r_n) \\ &\leq N \max_{1\leq l\leq N} P(|R_1(s_l) - ER_1(s_l)| > c_3r_n) \\ &\quad + N \max_{1\leq l\leq N} P(|\tilde{R}_1(s_l) - E\tilde{R}_1(s_l)| > c_3r_n). \end{aligned}$$

Third, we will use Bernstein inequality to bound the above probabilites. Let $V_i(s) = Y_{i1}K(\|\frac{s_i-s}{h}\|) - E[Y_{i1}K(\|\frac{s_i-s}{h}\|)]$ where $Y_{i1} = Y_i1(|Y_i| \leq \tau_n)$. As $|Y_{i1}| \leq \tau_n$ and $K(\|\frac{s_i-s}{h}\|) \leq c_4$ for some positive constant c_4, it follows that $|V_i(s)| \leq 2c_4\tau_n$ and for any s, $\sum_{i=1}^n \text{var}(V_i(s)) = \sum_{i=1}^n K^2(\|\frac{s_i-s}{h}\|)D(Y_{i1}) \leq c_5nh^2$ by Condition (4) for some positive constant c_5. Then by Bernstein inequality for independent variables it follows that for any s and sufficient large n,

$$\begin{aligned} P(|R_1(s) - ER_1(s)| > c_3r_n) &= P(|\sum_{i=1}^n V_i(s)| > c_3r_nnh^2) \\ &\leq 2\exp\left\{\frac{-c_3^2r_n^2n^2h^4}{2\sum_{i=1}^n \text{var}(V_i(s)) + \frac{4}{3}c_3c_4\tau_nr_nnh^2}\right\} \\ &\leq 2\exp\left\{\frac{-c_3^2\log n}{2c_5 + 4c_4}\right\} \leq 2n^{-c_3} \end{aligned}$$

since $(c_3/3\tau_nr_n)^2 = c_3^2/9\log^2 n/(n^{1-2/q}h^2) \to 0$ and taking $c_3 > \max\{2c_5 + 4c_4, 1\}$.

Using the same arguments, we can get that for any s and sufficient large n $P(|\tilde{R}_1(s) - E\tilde{R}_1(s)| > c_3 r_n) \leq 2n^{-c_3}$. Therefore,

$$P(\sup_{s\in\mathcal{S}} |R_1(s) - ER_1(s)| > 4c_3 r_n) \leq ch^{-2} r_n^{-2} n^{-c_3} = o(1).$$

Proof of Lemma 2:

(1) Note that

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}H^{-1} = \begin{pmatrix} \frac{1}{n}\sum\limits_{i=1}^{n} X_i X_i^{\mathrm{T}} K_h(\|s_i - s\|) & \frac{1}{n}\sum\limits_{i=1}^{n} X_i X_i^{\mathrm{T}} \otimes (\frac{s_i - s}{h})^{\mathrm{T}} K_h(\|s_i - s\|) \\ \frac{1}{n}\sum\limits_{i=1}^{n} X_i X_i^{\mathrm{T}} \otimes \frac{s_i - s}{h} K_h(\|s_i - s\|) & \frac{1}{n}\sum\limits_{i=1}^{n} X_i X_i^{\mathrm{T}} \otimes \frac{s_i - s}{h}(\frac{s_i - s}{h})^{\mathrm{T}} K_h(\|s_i - s\|) \end{pmatrix}.$$

Then by Lemma 1, Lipschitz continuity of $f(\cdot)$, Condition (4) and symmetry of the kernel funtion we have that

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}H^{-1} = \begin{pmatrix} \kappa_0 f(s)\Psi + O_p(\{h + r_n\}\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}) & O_p(\{h + r_n\}\mathbf{1}_p\mathbf{1}_{2p}^{\mathrm{T}}) \\ O_p(\{h + r_n\}\mathbf{1}_{2p}\mathbf{1}_p^{\mathrm{T}}) & \kappa_2 f(s)\Psi \otimes I_2 + O_p(\{h + r_n\}\mathbf{1}_{2p}\mathbf{1}_{2p}^{\mathrm{T}}) \end{pmatrix}$$

holds uniformly in $s \in \mathcal{S}$.

(2) Note that

$$\boldsymbol{\beta}(s) - (I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathbf{m} =$$

$$(I_p, \mathbf{0}_{p\times 2p})H^{-1}(n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}H^{-1})^{-1} n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\Big\{\mathcal{X}\begin{pmatrix} \boldsymbol{\beta}(s) \\ \mathrm{Vec}(\dot{\boldsymbol{\beta}}^{\mathrm{T}}(s)) \end{pmatrix} - \mathbf{m}\Big\}.$$

As

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\Big\{\mathcal{X}\begin{pmatrix}\boldsymbol{\beta}(s)\\ \mathrm{Vec}(\dot{\boldsymbol{\beta}}^{\mathrm{T}}(s))\end{pmatrix}-\mathbf{m}\Big\}=$$

$$\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}X_iX_i^{\mathrm{T}}\{\boldsymbol{\beta}(s)+\dot{\boldsymbol{\beta}}(s)(s_i-s)-\boldsymbol{\beta}(s_i)\}K_h(\|s_i-s\|)\\ \frac{1}{n}\sum\limits_{i=1}^{n}X_i\otimes\frac{s_i-s}{h}X_i^{\mathrm{T}}\{\boldsymbol{\beta}(s)+\dot{\boldsymbol{\beta}}(s)(s_i-s)-\boldsymbol{\beta}(s_i)\}K_h(\|s_i-s\|)\end{pmatrix},$$

it follows by the second order Taylor expansion that for s_i in a small neighborhood of s,

$$\boldsymbol{\beta}(s_i)=\boldsymbol{\beta}(s)+\dot{\boldsymbol{\beta}}(s)(s_i-s)+\frac{1}{2}\begin{pmatrix}(s_i-s)^{\mathrm{T}}\ddot{\boldsymbol{\beta}}_1(s_i^*)(s_i-s)\\ \vdots\\ (s_i-s)^{\mathrm{T}}\ddot{\boldsymbol{\beta}}_p(s_i^*)(s_i-s)\end{pmatrix},$$

where $\ddot{\boldsymbol{\beta}}_l(s)=\frac{\partial^2\beta_l(s)}{\partial s\partial s^{\mathrm{T}}}$, $l=1,\cdots,p$, and $s_i^*=s+\theta(s_i-s)$ with $\theta\in(0,1)$. Then we can obtain that

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\Big\{\mathcal{X}\begin{pmatrix}\boldsymbol{\beta}(s)\\ \mathrm{Vec}(\dot{\boldsymbol{\beta}}^{\mathrm{T}}(s))\end{pmatrix}-\mathbf{m}\Big\}=$$

$$-\frac{h^2}{2}\sum_{l=1}^{p}\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}X_ix_{il}(\frac{s_i-s}{h})^{\mathrm{T}}\ddot{\boldsymbol{\beta}}_l(s_i^*)(\frac{s_i-s}{h})K_h(\|s_i-s\|)\\ \frac{1}{n}\sum\limits_{i=1}^{n}X_i\otimes(\frac{s_i-s}{h})x_{il}(\frac{s_i-s}{h})^{\mathrm{T}}\ddot{\boldsymbol{\beta}}_l(s_i^*)(\frac{s_i-s}{h})K_h(\|s_i-s\|)\end{pmatrix}.$$

Now using Lemma 1, Condition (4), symmetry of the kernel function and continuity of the second order partial derivatives of $\boldsymbol{\beta}(s)$, it is easy to show that

$$\frac{1}{n}\sum_{i=1}^{n}X_ix_{il}(\frac{s_i-s}{h})^{\mathrm{T}}\ddot{\boldsymbol{\beta}}_l(s_i^*)(\frac{s_i-s}{h})K_h(\|s_i-s\|)$$

$$= \quad \kappa_2 f(s) E(X_1 x_{1l})(1,0,0,1)\mathrm{Vec}(\ddot{\boldsymbol{\beta}}_l(s)) + o_P(\mathbf{1}_p),$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \otimes (\frac{s_i - s}{h}) x_{il} (\frac{s_i - s}{h})^{\mathrm{T}} \ddot{\boldsymbol{\beta}}_l(s_i^*)(\frac{s_i - s}{h}) K_h(\|s_i - s\|) = o_P(\mathbf{1}_{2p})$$

hold uniformly in $s \in \mathcal{S}$. Therefore,

$$\begin{aligned} & n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\Big\{\mathcal{X}\begin{pmatrix} \boldsymbol{\beta}(s) \\ \mathrm{Vec}(\dot{\boldsymbol{\beta}}^{\mathrm{T}}(s)) \end{pmatrix} - \mathbf{m}\Big\} \\ = & \begin{pmatrix} -\frac{1}{2}h^2\kappa_2 f(s)\Psi\Big\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\Big\} \\ \mathbf{0}_{2p\times 1} \end{pmatrix} + o_P(h^2\mathbf{1}_{3p}) \end{aligned}$$

holds uniformly in $s \in \mathcal{S}$.

Next, it follows from Lemma 2(1) that

$$\begin{aligned} & (n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X}H^{-1})^{-1} = \\ & \begin{pmatrix} \kappa_0^{-1} f^{-1}(s)\Psi^{-1} & \mathbf{0}_{p\times 2p} \\ \mathbf{0}_{2p\times p} & \kappa_2^{-1} f^{-1}(s)\Psi^{-1}\otimes I_2 \end{pmatrix} + O_p(\{h + r_n\}\mathbf{1}_{3p}\mathbf{1}_{3p}^{\mathrm{T}}) \end{aligned}$$

holds uniformly in s. Hence, by the above results, we have

$$\boldsymbol{\beta}(s) - (I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathbf{m} = -\frac{\kappa_2 h^2}{2\kappa_0}\Big\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\Big\} + o_p(h^2\mathbf{1}_p)$$

holds uniformly in $s \in \mathcal{S}$.

Proof of Lemma 3: Note that

$$n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}G\mathbf{m} = \left(\begin{array}{c} \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} g_{ij}m_jX_iK_h(\|s_i-s\|) \\ \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} g_{ij}m_jX_i\otimes\frac{s_i-s}{h}K_h(\|s_i-s\|) \end{array} \right).$$

In the following we will show that

$$\sup_{s\in\mathcal{S}}|\frac{1}{n}\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n}g_{ij}m_jX_iK_h(\|s_i-s\|)-E[\sum_{j=1}^{n}g_{ij}m_jX_iK_h(\|s_i-s\|)]\Big\}| = o_P(1),$$

and

$$\sup_{s\in\mathcal{S}}|\frac{1}{n}\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n}g_{ij}m_jX_i\otimes\frac{s_i-s}{h}K_h(\|s_i-s\|)-E[\sum_{j=1}^{n}g_{ij}m_jX_i\otimes\frac{s_i-s}{h}K_h(\|s_i-s\|)]\Big\}| = o_P(1).$$

It is obvious that these two results can be established by the same arguments, here we only show the first one. Note that

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}g_{ij}m_jX_iK_h(\|s_i-s\|) \\ = \;& \frac{1}{n}\sum_{i=1}^{n}g_{ii}m_iX_iK_h(\|s_i-s\|)+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}g_{ij}Em_jX_iK_h(\|s_i-s\|) \\ & +\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}g_{ij}(m_j-Em_j)X_iK_h(\|s_i-s\|). \end{aligned}$$

As g_{ii} and $\sum\limits_{j\neq i}^{n}g_{ij}Em_j$ are both bounded for any i, it follows by Lemma 1 that

$$\frac{1}{n}\sum_{i=1}^{n}g_{ii}m_iX_iK_h(\|s_i-s\|) \;=\; \frac{1}{n}\sum_{i=1}^{n}E[g_{ii}m_iX_iK_h(\|s_i-s\|)]+O_P(r_n)$$

$$= \quad \Psi\frac{1}{n}\sum_{i=1}^{n} g_{ii}\boldsymbol{\beta}(s_i)K_h(\|s_i-s\|)+o_P(1),$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} g_{ij}Em_jX_iK_h(\|s_i-s\|) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} E[X_ig_{ij}Em_jK_h(\|s_i-s\|)]+O_P(r_n)\\ &= \Gamma\Gamma^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} g_{ij}\boldsymbol{\beta}(s_j)K_h(\|s_i-s\|)+o_P(1)\end{aligned}$$

hold uniformly in $s\in\mathcal{S}$.

In the following we only need to show that for any d ($d=1,\cdots,p$)

$$\sup_{s\in\mathcal{S}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} g_{ij}(m_j-Em_j)x_{id}K_h(\|s_i-s\|)\Big| = o_P(1).$$

This result can be established using the second step in Lemma 1 where we take $r_n=(\log n)^{-1/2}$ and then Chebyshev inequality instead of Bernstein inequality.

Proof of Lemma 4: (1) It follows by Lemma 2(1) and some calculation that

$$\begin{aligned}&S=\kappa_0^{-1}n^{-1}(1+o_P(1))\\ &\cdot\begin{pmatrix} f^{-1}(s_1)X_1^{\mathrm{T}}\Psi^{-1}X_1K_h(\|s_1-s_1\|) & \cdots & f^{-1}(s_1)X_1^{\mathrm{T}}\Psi^{-1}X_nK_h(\|s_n-s_1\|)\\ \vdots & \vdots & \vdots\\ f^{-1}(s_n)X_n^{\mathrm{T}}\Psi^{-1}X_1K_h(\|s_1-s_n\|) & \cdots & f^{-1}(s_n)X_n^{\mathrm{T}}\Psi^{-1}X_nK_h(\|s_n-s_n\|)\end{pmatrix}\end{aligned} \tag{6.2}$$

As $P = (I_n - S)^{\mathrm{T}}(I_n - S) = I_n - S^{\mathrm{T}} - S + S^{\mathrm{T}}S$, we note that

$$\begin{aligned} E[\mathrm{tr}(S)] &= \frac{1}{\kappa_0 n}\sum_{i=1}^{n} E[f^{-1}(s_i)X_i^{\mathrm{T}}\Psi^{-1}X_iK_h(\|s_i - s_i\|)](1+o(1)) \\ &= \frac{K(0)}{\kappa_0 nh^2}\sum_{i=1}^{n} E[f^{-1}(s_i)X_i^{\mathrm{T}}\Psi^{-1}X_i](1+o(1)) \\ &= \frac{pK(0)}{\kappa_0 nh^2}\sum_{i=1}^{n} f^{-1}(s_i) = O(h^{-2}), \end{aligned}$$

hence it follows by $nh^2 \to \infty$ that $n^{-1}E[\mathrm{tr}(S)] = n^{-1}E[\mathrm{tr}(S^{\mathrm{T}})] = o(1)$.

Since by straightforward calculation we have that the (k,l)th $(k, l = 1, \cdots, n)$ element of matrix $S^{\mathrm{T}}S$ takes the form

$$\begin{aligned} [S^{\mathrm{T}}S]_{kl} &= \kappa_0^{-2}n^{-2}(1+o_P(1)) \\ &\quad \cdot X_k^{\mathrm{T}}\{\sum_{i=1}^{n} f^{-2}(s_i)\Psi^{-1}X_iX_i^{\mathrm{T}}\Psi^{-1}K_h(\|s_k - s_i\|)K_h(\|s_l - s_i\|)\}X_l, \end{aligned}$$

and it follows by Lemma 1, continuity of $f(\cdot)$ and Condition (4) that

$$\begin{aligned} & \frac{1}{nh^2}\sum_{i=1}^{n} f^{-2}(s_i)\Psi^{-1}X_iX_i^{\mathrm{T}}\Psi^{-1}K^2(\|\frac{s_i - s}{h}\|) \\ = & \frac{1}{nh^2}\sum_{i=1}^{n} f^{-2}(s_i)\Psi^{-1}K^2(\|\frac{s_i - s}{h}\|) + O_P(r_n) \\ = & \nu_0 f^{-1}(s)\Psi^{-1}(1+o_P(1)) \end{aligned}$$

holds uniformly in $s \in \mathcal{S}$. Thus

$$\begin{aligned} n^{-1}E[\mathrm{tr}(S^{\mathrm{T}}S)] &= \frac{\nu_0}{\kappa_0^2 n^2 h^2}E[\sum_{k=1}^{n} f^{-1}(s_k)X_k^{\mathrm{T}}\Psi^{-1}X_k](1+o(1)) \\ &= \frac{\nu_0 p}{\kappa_0^2 n^2 h^2}\sum_{k=1}^{n} f^{-1}(s_k)(1+o(1)) = o(1). \end{aligned}$$

Consequently, $n^{-1}\mathrm{tr}(P) = n^{-1}\mathrm{tr}(I_n) - 2n^{-1}\mathrm{tr}(S) + n^{-1}\mathrm{tr}(S^{\mathrm{T}}S) = 1 + o(1)$.

Results (2) and (3) can be established by the same arguments as in (1) and straightforward calculation.

Next it can be seen clearly from the above proof that when $nh^2/\rho_n \to \infty$, we can obtain results (4)-(6) by the fact that the elements of G having the uniform order $O(1/\rho_n)$.

Proof of Lemma 5: (1) It follows from Lemma 2(2) and some calculation that

$$\begin{aligned}(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} &= -\frac{\kappa_2 h^2}{2\kappa_0}(G\mathbf{m})^{\mathrm{T}}(I_n - S^{\mathrm{T}})\begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} \\ &\quad +(G\mathbf{m})^{\mathrm{T}}(I_n - S^{\mathrm{T}})(X_1, \cdots, X_n)^{\mathrm{T}}\mathbf{1}_p o_P(h^2).\end{aligned}$$

Next we use (6.2), Lemma 2(1), Lemma 1, Condition (4), continuity of $f(\cdot)$ and the second partial derivatives of $\boldsymbol{\beta}(\cdot)$ to get that

$$\begin{aligned}& S^{\mathrm{T}}\begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} \\ &= \begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} + (X_1, \cdots, X_n)^{\mathrm{T}}\mathbf{1}_p o_P(1),\end{aligned}$$

and

$$S^{\mathrm{T}}(X_1, \cdots, X_n)^{\mathrm{T}}\mathbf{1}_p = (X_1^{\mathrm{T}}\mathbf{1}_p, \cdots, X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}} O_P(1). \tag{6.3}$$

Consequently we have by Markov inequality that

$$(G\mathbf{m})^{\mathrm{T}} P\mathbf{m} = n^{-1}(G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots, X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}} o_P(nh^2) = o_P(nh^2)$$

(2) If $f(\cdot)$ and the second partial derivatives of $\boldsymbol{\beta}(s)$ are all Lipshitz continuous, then we can obtain by Lemma 1, Condition (4) and similar arguments as in Lemma 2(2) that

$$\begin{aligned} & n^{-1}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\Big\{\mathcal{X}\begin{pmatrix} \boldsymbol{\beta}(s) \\ \mathrm{Vec}(\dot{\boldsymbol{\beta}}^{\mathrm{T}}(s)) \end{pmatrix} - \mathbf{m}\Big\} \\ = & \begin{pmatrix} -\frac{1}{2}h^2\kappa_2 f(s)\Psi\Big\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\Big\} \\ \mathbf{0}_{2p\times 1} \end{pmatrix} + O_P(\{h^3 + h^2 r_n\}\mathbf{1}_{3p}). \end{aligned}$$

This together with Lemma 2(1) leads to

$$\boldsymbol{\beta}(s) - (I_p, \mathbf{0}_{p\times 2p})(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathbf{m} = -\frac{\kappa_2 h^2}{2\kappa_0}\Big\{\boldsymbol{\beta}_{uu}(s) + \boldsymbol{\beta}_{vv}(s)\Big\} + O_p(\{h^3 + h^2 r_n\}\mathbf{1}_p)$$

holding uniformly in $s \in \mathcal{S}$. Hence

$$\begin{aligned} (G\mathbf{m})^{\mathrm{T}} P\mathbf{m} &= -\frac{\kappa_2 h^2}{2\kappa_0}(G\mathbf{m})^{\mathrm{T}}(I_n - S^{\mathrm{T}})\begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} \\ & \quad + (G\mathbf{m})^{\mathrm{T}}(I_n - S^{\mathrm{T}})(X_1, \cdots, X_n)^{\mathrm{T}}\mathbf{1}_p O_P(\{h^3 + h^2 r_n\}) \end{aligned}$$

If $f(\cdot)$ and the second partial derivatives of $\boldsymbol{\beta}(\cdot)$ are all Lipschitz continuous, then

$$\begin{aligned} & S^{\mathrm{T}} \begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} \\ = & \begin{pmatrix} X_1^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_1) + \boldsymbol{\beta}_{vv}(s_1)] \\ \vdots \\ X_n^{\mathrm{T}}[\boldsymbol{\beta}_{uu}(s_n) + \boldsymbol{\beta}_{vv}(s_n)] \end{pmatrix} + (X_1, \cdots, X_n)^{\mathrm{T}} \mathbf{1}_p O_P(h + r_n). \end{aligned}$$

Therefore, we have by (6.3) and Markov inequality that

$$\begin{aligned} (G\mathbf{m})^{\mathrm{T}} P\mathbf{m} &= n^{-1}(G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots, X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}} O_P(n\{h^3 + h^2 r_n\}) \\ &= O_P(nh^3 + \{nh^2 \log n\}^{1/2}) \end{aligned}$$

Proof of Lemma 6: (1) In the following, we will show that $n^{-1/2}L^{\mathrm{T}}P\mathbf{m} = o_P(1)$ for $L = \mathbf{m},\ \boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$.

Note that $n^{-1/2}\mathbf{m}^{\mathrm{T}}P\mathbf{m} = n^{-1/2}(\mathbf{m} - S\mathbf{m})^{\mathrm{T}}(\mathbf{m} - S\mathbf{m})$, and it follows by Lemma 2(2) that

$$\mathbf{m} - S\mathbf{m} = (X_1, \cdots, X_n)^{\mathrm{T}}\mathbf{1}_p O_P(h^2). \tag{6.4}$$

Therefore,

$$n^{-1/2}\mathbf{m}^{\mathrm{T}}P\mathbf{m} = n^{-1}\sum_{i=1}^{n}(X_i^{\mathrm{T}}\mathbf{1}_p)^2 O_P(n^{1/2}h^4) = o_P(1)$$

using law of large numbers and $nh^8 \to 0$.

Since we have by (6.3), (6.4) and Chebyshev inequality that

$$\begin{aligned} n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}P\mathbf{m} &= n^{-1/2}(\boldsymbol{\epsilon}-S\boldsymbol{\epsilon})^{\mathrm{T}}(\mathbf{m}-S\mathbf{m}) \\ &= \{n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}(X_1,\cdots,X_n)^{\mathrm{T}}\mathbf{1}_p - n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}(X_1,\cdots,X_n)^{\mathrm{T}}\mathbf{1}_p\}O_P(h^2) \\ &= n^{-1/2}\sum_{i=1}^{n}X_i^{\mathrm{T}}\mathbf{1}_p\epsilon_i O_P(h^2) = O_P(h^2), \end{aligned}$$

Hence $n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}P\mathbf{m} = o_P(1)$.

Similarly, we can show that $n^{-1/2}(G\boldsymbol{\epsilon})^{\mathrm{T}}P\mathbf{m} = O_P(h^2) = o_P(1)$.

(2) Here, we will show that $n^{-1}L^{\mathrm{T}}PG\mathbf{m} = o_P(1)$ for $L = \mathbf{m}, \boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$.

Clearly, it follows by Lemma 5(1) that $n^{-1}\mathbf{m}^{\mathrm{T}}PG\mathbf{m} = o_P(h^2) = o_P(1)$.

For simplification, in the following we set $\tilde{X} = (X_1^{\mathrm{T}}\mathbf{1}_p,\cdots,X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}}$ and $V = (f^{-1}(s_1)X_1^{\mathrm{T}}\Psi^{-1}\mathbf{1}_p,\cdots,f^{-1}(s_n)X_n^{\mathrm{T}}\Psi^{-1}\mathbf{1}_p)^{\mathrm{T}}$.

Note that

$$\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}PG\mathbf{m} = \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}G\mathbf{m} - \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}G\mathbf{m} - \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}SG\mathbf{m} + \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}SG\mathbf{m}.$$

As $E(\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}G\mathbf{m}) = 0$, and

$$\mathrm{var}(\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}G\mathbf{m}) = \frac{\sigma_0^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}g_{ij}^2Em_j^2 + \frac{\sigma_0^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}g_{ij}g_{ik}Em_jEm_k = O(\frac{1}{n}),$$

we obtain by Chebyshev inequality that $n^{-1}\boldsymbol{\epsilon}^{\mathrm{T}}G\mathbf{m} = o_P(1)$.

It follows by (6.2), Lipschitz continuity of $f(\cdot)$, Lemma 3 and Condition (4) that

$$S^{\mathrm{T}}G\mathbf{m} = \{Z + V\cdot o_P(1) + \tilde{X}\cdot o_P(1)\}(1+o_P(1)). \tag{6.5}$$

Therefore,

$$\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}G\mathbf{m} = \{\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}Z + \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}V \cdot o_P(1) + \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}\tilde{X} \cdot o_P(1)\}(1 + o_P(1)) = o_P(1)$$

by law of large numbers.

Similarly,

$$SG\mathbf{m} = \{Z + V \cdot o_P(1)\}(1 + o_P(1)). \tag{6.6}$$

by Lemma 3 and Condition (4). Therefore, we have by law of large numbers that

$$\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}SG\mathbf{m} = \{\frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}Z + \frac{1}{n}\boldsymbol{\epsilon}^{\mathrm{T}}V \cdot o_P(1)\}(1 + o_P(1)) = o_P(1).$$

Next it follows by (6.2) and Lemma 1 that $S\boldsymbol{\epsilon} = V \cdot o_P(1)$. This together with (6.6), we obtain that

$$\frac{1}{n}(S\boldsymbol{\epsilon})^{\mathrm{T}}SG\mathbf{m} = \{\frac{1}{n}V^{\mathrm{T}}Z + \frac{1}{n}V^{\mathrm{T}}V\}o_P(1).$$

Therefore, $n^{-1}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}SG\mathbf{m} = o_P(1)$ by law of large numbers.

Similarly, we can show that $n^{-1}(G\boldsymbol{\epsilon})^{\mathrm{T}}PG\mathbf{m} = o_P(1)$.

Proof of Lemma 7: (1) It can be seen that

$$\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m} = \frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m} - \frac{2}{n}(G\mathbf{m})^{\mathrm{T}}SG\mathbf{m} + \frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m}.$$

For the first term we have that

$$\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m} = \frac{1}{n}\sum_{j=1}^{n}(\sum_{i=1}^{n} g_{ij}^2)m_j^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n} g_{ij}g_{ik}m_jm_k,$$

and

$$\mathrm{var}\{\frac{1}{n}\sum_{j=1}^{n}(\sum_{i=1}^{n} g_{ij}^2)m_j^2\} = \frac{1}{n^2}\sum_{j=1}^{n}(\sum_{i=1}^{n} g_{ij}^2)^2 D(m_j^2) = O(\frac{1}{n}),$$

it follows by Chebyshev inequality that $\frac{1}{n}\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n} g_{ij}^2 m_j^2 - E\{\frac{1}{n}\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n} g_{ij}^2 m_j^2\} = o_P(1)$.

Let $\bar{m}_i = m_i - Em_i,\ i = 1, \cdots, n$, then

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n} g_{ij}g_{ik}m_jm_k - E\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n} g_{ij}g_{ik}m_jm_k\} \\ = \ & \frac{1}{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\sum_{i=1}^{n} g_{ij}g_{ik}\bar{m}_j\bar{m}_k + \frac{1}{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\sum_{i=1}^{n} g_{ij}g_{ik}\bar{m}_jEm_j \\ & + \frac{1}{n}\sum_{k=1}^{n}\sum_{j\neq k}^{n}\sum_{i=1}^{n} g_{ij}g_{ik}\bar{m}_kEm_k. \end{aligned}$$

Define $J_{n1} = \frac{1}{n}\sum\limits_{j=1}^{n}\sum\limits_{k\neq j}^{n}\sum\limits_{i=1}^{n} g_{ij}g_{ik}\bar{m}_j\bar{m}_k$, $J_{n2} = \frac{1}{n}\sum\limits_{j=1}^{n}\sum\limits_{k\neq j}^{n}\sum\limits_{i=1}^{n} g_{ij}g_{ik}\bar{m}_jEm_j$ and $J_{n3} = \frac{1}{n}\sum\limits_{k=1}^{n}\sum\limits_{j\neq k}^{n}\sum\limits_{i=1}^{n} g_{ij}g_{ik}\bar{m}_kEm_k$, with

$$\begin{aligned} \mathrm{var}(J_{n1}) &= E(J_{n1}^2) \\ &= \frac{2}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}^{n}(\sum_{i=1}^{n} g_{ij}g_{ik})^2[\boldsymbol{\beta}^{\mathrm{T}}(s_j)D(X_1)\boldsymbol{\beta}(s_j)][\boldsymbol{\beta}^{\mathrm{T}}(s_k)D(X_1)\boldsymbol{\beta}(s_k)] \\ &\leq \max_{j,k}(\sum_{i=1}^{n}|g_{ij}g_{ik}|)\frac{2}{n^2}\sum_{j=1}^{n}\{\sum_{i=1}^{n}|g_{ij}|\boldsymbol{\beta}^{\mathrm{T}}(s_j)D(X_1)\boldsymbol{\beta}(s_j)\}^2 = O(\frac{1}{n}), \end{aligned}$$

$$\mathrm{var}(J_{n2}) = \frac{1}{n^2}\sum_{j=1}^{n}(\sum_{k\neq j}^{n}\sum_{i=1}^{n} g_{ij}g_{ik})^2 D(m_j)(Em_j)^2 = O(\frac{1}{n}),$$

and

$$\mathrm{var}(J_{n3}) = \frac{1}{n^2}\sum_{k=1}^{n}(\sum_{j\neq k}^{n}\sum_{i=1}^{n} g_{ij}g_{ik})^2 D(m_k)(Em_k)^2 = O(\frac{1}{n}).$$

Therefore, by Chebyshev inequality that $J_{n1} = o_P(1)$, $J_{n2} = o_P(1)$ and $J_{n3} = o_P(1)$. In conclusion, we obtain that $\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m} - E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m}\} = o_P(1)$.

It follows from (6.6) that

$$\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}SG\mathbf{m} = \{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}Z + \frac{1}{n}(G\mathbf{m})^{\mathrm{T}}V \cdot o_P(1)\}(1 + o_P(1)).$$

Using similar arguments as establishing $\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m} - E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}G\mathbf{m}\} = o_P(1)$, we have that $\frac{1}{n}(G\mathbf{m})^{T}L - E\{\frac{1}{n}(G\mathbf{m})^{T}L\} = o_P(1)$ for $L = Z$ and V. Moreover, $E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}SG\mathbf{m}\} = E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}Z\} + o(1)$. Therefore,

$$\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}SG\mathbf{m} - E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}SG\mathbf{m}\} = o_P(1).$$

For the term $\frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m}$, again by (6.6) we have that

$$\begin{aligned}\frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m} &= \Big\{\frac{1}{n}Z^{\mathrm{T}}Z + \frac{1}{n}V^{\mathrm{T}}V \cdot o_P(1) + \frac{2}{n}Z^{\mathrm{T}}V \cdot o_P(1)\Big\}(1 + o_P(1)) \\ &= \frac{1}{n}E(Z^{\mathrm{T}}Z) + o_P(1)\end{aligned}$$

by law of large numbers, and $E\{\frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m}\} = \frac{1}{n}E(Z^{\mathrm{T}}Z) + o(1)$. Thus

$$\frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m} - E\{\frac{1}{n}(SG\mathbf{m})^{\mathrm{T}}SG\mathbf{m}\} = o_P(1).$$

In conclusion, we obtain that $\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m} - E\{\frac{1}{n}(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}\} = o_P(1)$.

(2) We have seen from (6.6) that

$$\begin{aligned}\frac{1}{n}E\{(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m}\} &= \frac{1}{n}E\{(G\mathbf{m} - Z + V\cdot o_P(1))^{\mathrm{T}}(G\mathbf{m} - Z + V\cdot o_P(1))\}(1+o(1)) \\ &= \frac{1}{n}E[(G\mathbf{m} - Z)^{\mathrm{T}}(G\mathbf{m} - Z)] + o(1).\end{aligned}$$

Proof of Lemma 8: (1) Since

$$n^{-1/2}\{\boldsymbol{\epsilon}^{\mathrm{T}}P\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}\} = -2n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}S\boldsymbol{\epsilon} + n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}S^{\mathrm{T}}S\boldsymbol{\epsilon},$$

$E[n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}S\boldsymbol{\epsilon}] = \sigma_0^2 n^{-1/2}E[\mathrm{tr}(S)] = O(\{nh^4\}^{-1/2}) = o(1)$, and

$$\begin{aligned}\mathrm{var}(n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}}S\boldsymbol{\epsilon}) &\leq \frac{1}{n}E(\boldsymbol{\epsilon}^{\mathrm{T}}S\boldsymbol{\epsilon})^2 \\ &= \frac{1}{n}\Big[(\mu_4 - 3\sigma_0^4)\sum_{i=1}^{n}E[S]_{ii}^2 + \sigma_0^4 E\{[\mathrm{tr}(S)]^2 + \mathrm{tr}(SS^{\mathrm{T}}) + \mathrm{tr}(S^2)\}\Big]\end{aligned}$$

It can be seen from the proof of Lemma 4(1) that $n^{-1}E[\mathrm{tr}(SS^{\mathrm{T}})] = o(1)$,

$$\frac{1}{n}\sum_{i=1}^{n}E[S]_{ii}^2 = \frac{1}{\kappa_0^2 n^3 h^4}\sum_{i=1}^{n}E[f^{-2}(s_i)(X_i^{\mathrm{T}}\Psi^{-1}X_i)^2]K^2(0) = O(\frac{1}{n^2h^4}) = o(1),$$

and

$$\mathrm{tr}(S) = \frac{pK(0)}{\kappa_0 nh^2}\sum_{i=1}^{n}f^{-1}(s_i) + o_p(1).$$

It can be seen that $n^{-1/2}\mathrm{tr}(S) = O_p(\{nh^4\}^{-1/2}) = o_P(1)$. Hence, $n^{-1}[\mathrm{tr}(S)]^2 = o_P(1)$ and $n^{-1}E\{[\mathrm{tr}(S)]^2\} = o(1)$.

It follows by straightforward calculation, Lemma 1 and Condition

(4) that

$$\begin{aligned}
[S^2]_{ii} &= \frac{1+o_P(1)}{\kappa_0^2 n^2} f^{-1}(s_i) X_i^{\mathrm{T}} \Psi^{-1} \sum_{j=1}^{n} f^{-1}(s_j) X_j X_j^{\mathrm{T}} \Psi^{-1} K_h^2(\|s_j - s_i\|) X_i \\
&= \frac{\nu_0}{\kappa_0^2 n h^2} f^{-1}(s_i) X_i^{\mathrm{T}} \Psi^{-1} X_i (1 + o_P(1)).
\end{aligned}$$

Thus

$$\frac{1}{n} E[\mathrm{tr}(S^2)] = \frac{\nu_0(1+o(1))}{\kappa_0^2 n^2 h^2} \sum_{i=1}^{n} E[f^{-1}(s_i) X_i^{\mathrm{T}} \Psi^{-1} X_i] = O(\frac{1}{nh^2}) = o(1).$$

Consequently, we have by Chebyshev inequality that $n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}} S \boldsymbol{\epsilon} = o_P(1)$.

Similarly, it can be shown that $n^{-1/2}\boldsymbol{\epsilon}^{\mathrm{T}} S^{\mathrm{T}} S \boldsymbol{\epsilon} = o_P(1)$. Hence we have shown that $n^{-1/2}(\boldsymbol{\epsilon}^{\mathrm{T}} P \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}} \boldsymbol{\epsilon}) = o_P(1)$.

Results (2) and (3) can be obtained by the same arguments as in (1) and straightforward calculation.

(4) Note that

$$n^{-1/2}\{(G\mathbf{m})^{\mathrm{T}} P \boldsymbol{\epsilon} - (G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}} \boldsymbol{\epsilon}\} = -n^{-1/2}(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}} S \boldsymbol{\epsilon}.$$

Moreover, $E[n^{-1/2}(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}} S\boldsymbol{\epsilon}] = 0$ and $\mathrm{var}[n^{-1/2}(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}} S\boldsymbol{\epsilon}] = \sigma_0^2 n^{-1} E[(G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}} S S^{\mathrm{T}} (G\mathbf{m} - SG\mathbf{m})]$. As it follows by (6.6), Lemma 1 and Condition (4) that

$$\begin{aligned}
S^{\mathrm{T}} SG\mathbf{m} &= \{S^{\mathrm{T}} Z + S^{\mathrm{T}} V \cdot o_P(1)\}(1 + o_P(1)) \\
&= \{Z + \tilde{X} \cdot o_P(1) + V \cdot o_P(1)\}(1 + o_P(1))
\end{aligned}$$

where $V = (f^{-1}(s_1) X_1^{\mathrm{T}} \Psi^{-1} \mathbf{1}_p, \cdots, f^{-1}(s_n) X_n^{\mathrm{T}} \Psi^{-1} \mathbf{1}_p)^{\mathrm{T}}$ and $\tilde{X} = (X_1^{\mathrm{T}} \mathbf{1}_p, \cdots, X_n^{\mathrm{T}} \mathbf{1}_p)^{\mathrm{T}}$.

This together with (6.5) it can be seen that

$$S^{\mathrm{T}}G\mathbf{m} - S^{\mathrm{T}}SG\mathbf{m} = \{Z + \tilde{X} + V\}o_P(1).$$

Hence $n^{-1}E[(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}SS^{\mathrm{T}}(G\mathbf{m}-SG\mathbf{m})] = o(1)$. Consequently, it can be obtained by Chebyshev inequality that $n^{-1/2}(G\mathbf{m}-SG\mathbf{m})^{\mathrm{T}}S\boldsymbol{\epsilon} = o_P(1)$. Therefore, $n^{-1/2}\{(G\mathbf{m})^{\mathrm{T}}P\boldsymbol{\epsilon} - (G\mathbf{m} - SG\mathbf{m})^{\mathrm{T}}\boldsymbol{\epsilon}\} = o_P(1)$.

Proof of Lemma 9: The asymptotic distribution of the linear-quadratic random form Q_n can be established via the martingale central limit theorem. Our proof of this lemma follows closely the arguments in Kelejian and Prucha (2001) and Lee (2004).

Note that

$$Q_n = \sum_{i=1}^{n}(\sum_{j=1}^{n} g_{ij}m_j - g_i^s)\epsilon_i + \sum_{i=1}^{n} b_{ii}\epsilon_i^2 + 2\sum_{i=1}^{n}\sum_{k=1}^{i-1} b_{ik}\epsilon_i\epsilon_k - \sigma_0^2\mathrm{tr}(B) = \sum_{i=1}^{n} V_{ni}$$

where g_i^s is the ith element of $SG\mathbf{m}$ and $V_{ni} = (\sum_{j=1}^{n} g_{ij}m_j - g_i^s)\epsilon_i + b_{ii}(\epsilon_i^2 - \sigma_0^2) + 2\epsilon_i \sum_{k=1}^{i-1} b_{ik}\epsilon_k$.

Define $\sigma-$ fields $\mathcal{T}_i =< \epsilon_1, \cdots, \epsilon_i >$ generated by $\epsilon_1, \cdots, \epsilon_i$. Because $\{\epsilon_i\}_{i=1}^n$ are iid with zero mean, finite variance and independent with $\{X_j\}_{j=1}^n$,

$$E(V_{ni}|\mathcal{T}_{i-1}) = E(\sum_{j=1}^{n} g_{ij}m_j - g_i^s)E\epsilon_i + b_{ii}(E\epsilon_i^2 - \sigma_0^2) + 2E\epsilon_i \sum_{k=1}^{i-1} b_{ik}\epsilon_k = 0.$$

Hence, the $\{(V_{ni}, \mathcal{T}_i)|1 \leq i \leq n\}$ forms a martingale difference double array and $\sigma_{Q_n}^2 = \sum_{i=1}^{n} E(V_{ni}^2)$ with $\sigma_{Q_n}^2$ being bounded away from zero at n rate. Define the normalized variables $V_{ni}^* = V_{ni}/\sigma_{Q_n}$. Then

$\{(V_{ni}^*, \mathcal{T}_i) | 1 \leq i \leq n\}$ is a martingale difference double array and $\frac{Q_n}{\sigma_{Q_n}} = \sum_{i=1}^{n} V_{ni}^*$. In order for the martingale central limit theorem to be applicable we would show that there exists a $\delta > 0$ such that $\sum_{i=1}^{n} E|V_{ni}^*|^{2+\delta} = o(1)$ and $\sum_{i=1}^{n} E(V_{ni}^{*2} | \mathcal{T}_{i-1}) \xrightarrow{P} 1$.

For any positive constant p and q satisfying $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$\begin{aligned} |V_{ni}| &\leq |b_{ii}| \cdot |\epsilon_i^2 - \sigma_0^2| + |\epsilon_i| (|\sum_{j=1}^{n} g_{ij} m_j - g_i^s| + 2 \sum_{k=1}^{i-1} |b_{ik}| \cdot |\epsilon_k|) \\ &= |b_{ii}|^{\frac{1}{p}} (|b_{ii}|^{\frac{1}{q}} \cdot |\epsilon_i^2 - \sigma_0^2|) + |\sum_{j=1}^{n} g_{ij} m_j - g_i^s|^{\frac{1}{p}} (|\sum_{j=1}^{n} g_{ij} m_j - g_i^s|^{\frac{1}{q}} |\epsilon_i|) \\ &\quad + \sum_{k=1}^{i-1} |b_{ik}|^{\frac{1}{p}} (|b_{ik}|^{\frac{1}{q}} 2 |\epsilon_k| \cdot |\epsilon_i|). \end{aligned}$$

Applying Holder inequality we obtain that

$$\begin{aligned} |V_{ni}|^q &\leq \Big[|\sum_{j=1}^{n} g_{ij} m_j - g_i^s| + \sum_{k=1}^{i} |b_{ik}| \Big]^{\frac{q}{p}} \\ &\quad \cdot \Big[|\sum_{j=1}^{n} g_{ij} m_j - g_i^s| \cdot |\epsilon_i|^q + |b_{ii}| \cdot |\epsilon_i^2 - \sigma_0^2|^q + \sum_{k=1}^{i-1} |b_{ik}| 2^q |\epsilon_i|^q |\epsilon_k|^q \Big] \end{aligned}$$

Let $c_1 > 1$ be a finite constant such that $E(|\epsilon_1^2 - \sigma_0^2|) \leq c_1$, $E|\epsilon_1|^q \leq c_1$, and $(E|\epsilon_1|^q)^2 \leq c_1$. Set $\mathcal{D} = \{X_i\}_{i=1}^n$, we have

$$E[|V_{ni}|^q | \mathcal{D}] \leq 2^q c_1 \Big[|\sum_{j=1}^{n} g_{ij} m_j - g_i^s| + \sum_{k=1}^{i} |b_{ik}| \Big]^q$$

As the the matrix B are uniformly bounded in row sums, there exists a constant c_2 such that $\sum_{j=1}^{n} |b_{ij}| \leq c_2$ for all i. Take $q = 2 + \delta$, it follows by Cr inequality and (6.6) that

$$\begin{aligned}
&\sum_{i=1}^{n} E[|V_{ni}|^{2+\delta}] = \sum_{i=1}^{n} E\{E[|V_{ni}|^{2+\delta}|\mathcal{D}]\} \\
\leq\ & c_1 2^{3+2\delta} \sum_{i=1}^{n}\{E|\sum_{j=1}^{n} g_{ij}m_j - g_i^s|^{2+\delta} + (\sum_{k=1}^{i}|b_{ik}|)^{2+\delta}\} \\
\leq\ & c_1 2^{3+2\delta}\{2^{2+2\delta}\sum_{i=1}^{n}\Big[E|\sum_{j=1}^{n} g_{ij}(m_j - Em_j)|^{2+\delta} + |\sum_{j=1}^{n} g_{ij}Em_j|^{2+\delta}\Big] + cn\}.
\end{aligned}$$

Because $\{m_j\}$ are independent variables, we have that

$$\begin{aligned}
&E|\sum_{j=1}^{n} g_{ij}(m_j - Em_j)|^{2+\delta} \\
\leq\ & c\{\sum_{j=1}^{n} E|g_{ij}(m_j - Em_j)|^{2+\delta} + (\sum_{j=1}^{n} E[g_{ij}(m_j - Em_j)]^2)^{\frac{2+\delta}{2}}\} \leq c
\end{aligned}$$

by $\sum\limits_{j=1}^{n}|g_{ij}|$ being uniformly bounded for all i. Therefore, $\sum\limits_{i=1}^{n} E[|V_{ni}|^{2+\delta}] = O(n)$. Hence $\sum\limits_{i=1}^{n} E|V_{ni}^*|^{2+\delta} = \frac{1}{(\sigma_{Q_n}^2)^{\frac{2+\delta}{2}}}\sum\limits_{i=1}^{n} E|V_{ni}|^{2+\delta} = O(\frac{n}{n^{1+\delta/2}}) = o(1)$.

It remains to show that $\sum\limits_{i=1}^{n} E(V_{ni}^{*2}|\mathcal{T}_{i-1}) \xrightarrow{P} 1$. As $E(V_{ni}^2|\mathcal{D},\mathcal{T}_{i-1}) = (\mu_4 - \sigma_0^4)b_{ii}^2 + [(\sum\limits_{j=1}^{n} g_{ij}m_j - g_i^s) + 2\sum\limits_{k=1}^{i-1} b_{ik}\epsilon_k]^2\sigma_0^2 + 2\mu_3 b_{ii}[(\sum\limits_{j=1}^{n} g_{ij}m_j - g_i^s) + 2\sum\limits_{k=1}^{i-1} b_{ik}\epsilon_k]$, it follows that

$$\begin{aligned}
&E(V_{ni}^2|\mathcal{T}_{i-1}) - E(V_{ni}^2) \\
=\ & 4\sigma_0^2\{\sum_{k=1}^{i-1} b_{ik}^2(\epsilon_k^2 - \sigma_0^2) + \sum_{k=1}^{i-1}\sum_{l\neq k}^{i-1} b_{ik}b_{il}\epsilon_k\epsilon_l\} + 4[\sigma_0^2 E(\sum_{j=1}^{n} g_{ij}m_j - g_i^s) + \mu_3 b_{ii}]\sum_{k=1}^{i-1} b_{ik}\epsilon_k
\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{n} E(V_{ni}^{*2}|\mathcal{T}_{i-1}) - 1 &= \frac{1}{\sigma_{Q_n}^2}\sum_{i=1}^{n}[E(V_{ni}^2|\mathcal{T}_{i-1}) - E(V_{ni}^2)] \\ &= \frac{4\sigma_0^2}{\frac{1}{n}\sigma_{Q_n}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\{\sum_{k=1}^{i-1} b_{ik}^2(\epsilon_k^2-\sigma_0^2) + \sum_{k=1}^{i-1}\sum_{l\neq k}^{i-1} b_{ik}b_{il}\epsilon_k\epsilon_l\} \\ &\quad + \frac{4}{\frac{1}{n}\sigma_{Q_n}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}[\sigma_0^2 E(\sum_{j=1}^{n} g_{ij}m_j - g_i^s) + \mu_3 b_{ii}]\sum_{k=1}^{i-1} b_{ik}\epsilon_k \\ &= \frac{4\sigma_0^2}{\frac{1}{n}\sigma_{Q_n}^2}(J_{n1}+J_{n2}) + \frac{4}{\frac{1}{n}\sigma_{Q_n}^2}J_{n3}\end{aligned}$$

with $J_{n1} = \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{i-1} b_{ik}^2(\epsilon_k^2-\sigma_0^2)$, $J_{n2} = \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{i-1}\sum\limits_{l\neq k}^{i-1} b_{ik}b_{il}\epsilon_k\epsilon_l$, and $J_{n3} = \frac{1}{n}\sum\limits_{i=1}^{n}[\sigma_0^2 E(\sum\limits_{j=1}^{n} g_{ij}m_j - g_i^s) + \mu_3 b_{ii}]\sum\limits_{k=1}^{i-1} b_{ik}\epsilon_k$.

Clearly, $EJ_{nl} = 0, l = 1,2,3$. By Chebyshev inequality, to show $J_{nl} = o_P(1)$, it is only need to prove $EJ_{nl}^2 = o(1)$. It is obvious by uniform boundness of b_{ik} and uniform boundness of $\sum\limits_{i=1}^{n}|b_{ik}|$ that $E(J_{n1}^2) = \frac{1}{n^2}\sum\limits_{k=1}^{n-1}(\sum\limits_{i=k+1}^{n} b_{ik}^2)^2 D(\epsilon_1^2) \leq \frac{1}{n^2}D(\epsilon_1^2)\max_{i,k}|b_{ik}|^2\sum\limits_{k=1}^{n-1}(\sum\limits_{i=k+1}^{n}|b_{ik}|)^2 = O(\frac{1}{n})$.

Since $J_{n2} = \frac{1}{n}\sum\limits_{k=1}^{n-1}\sum\limits_{l\neq k}^{n-1}(\sum\limits_{i=\max\{k,l\}+1}^{n} b_{ik}b_{il})\epsilon_k\epsilon_l$, we have

$$\begin{aligned}E(J_{n2}^2) &= \frac{2\sigma_0^4}{n^2}\sum_{k=1}^{n-1}\sum_{l\neq k}^{n-1}(\sum_{i=\max\{k,l\}+1}^{n} b_{ik}b_{il})^2 \leq \frac{2\sigma_0^4}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}(\sum_{i=1}^{n}|b_{ik}b_{il}|)^2 \\ &\leq \frac{2\sigma_0^4}{n^2}\max_{i,l}|b_{il}|\max_{k}\sum_{i=1}^{n}|b_{ik}|\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}|b_{ik}b_{il}| = O(\frac{1}{n})\end{aligned}$$

As J_{n3} can be written as $J_{n3} = \frac{1}{n}\sum\limits_{k=1}^{n-1}[\sum\limits_{i=k+1}^{n}(\sigma_0^2 E[\sum\limits_{j=1}^{n} g_{ij}m_j - g_i^s] +$

$\mu_3 b_{ii})b_{ik}]\epsilon_k$, it follows that

$$\begin{aligned}
E(J_{n3}^2) &= \frac{\sigma_0^2}{n^2}\sum_{k=1}^{n-1}[\sum_{i=k+1}^{n}(\sigma_0^2 E[\sum_{j=1}^{n} g_{ij}m_j - g_i^s] + \mu_3 b_{ii})b_{ik}]^2 \\
&\leq \frac{\sigma_0^2}{n^2}\max_i\{\sigma_0^2|E[\sum_{j=1}^{n} g_{ij}m_j - g_i^s]| + \mu_3|b_{ii}|\}^2\sum_{k=1}^{n}(\sum_{i=1}^{n}|b_{ik}|)^2 = O(\frac{1}{n})
\end{aligned}$$

by $|E[\sum_{j=1}^{n} g_{ij}m_j - g_i^s]| \leq \sum_{j=1}^{n}|g_{ij}Em_j| + E|g_i^s| = O(1)$ for any i, where $E|g_i^s| = O(1)$ is obtained using (6.6).

Because $J_{nl} = o_P(1)$ for $l = 1, 2, 3$ and $\lim_{n\to\infty}\frac{\sigma_{Q_n}^2}{n} > 0$, $\sum_{i=1}^{n} E(V_{ni}^{*2}|\mathcal{T}_{i-1})$ converges in probability to 1. The central limit theorem for martingale difference double array is thus applicable to establish the result.

Proof of Lemma 10: (1) Here we will show that $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(\rho_n^{-1/2}nh^2)$.

It can be seen from the proof of Lemma 5(1) that

$$(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = (G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}}o_P(h^2).$$

As $\frac{\sqrt{\rho_n}}{n}(G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}} = \frac{\sqrt{\rho_n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}m_j X_i^{\mathrm{T}}\mathbf{1}_p$, and

$$\begin{aligned}
E|\frac{\sqrt{\rho_n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}m_j X_i^{\mathrm{T}}\mathbf{1}_p| &\leq \frac{\sqrt{\rho_n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E|g_{ij}m_j X_i^{\mathrm{T}}\mathbf{1}_p| \\
&\leq c\frac{\sqrt{\rho_n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}|g_{ij}| = O(1),
\end{aligned}$$

using that $\max_i \sum_{j=1}^{n}|g_{ij}| = O(1/\sqrt{\rho_n})$, then by Markov inequality we have

$$\frac{\sqrt{\rho_n}}{n}(G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}} = O_P(1). \tag{6.7}$$

Therefore, $(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = o_P(\rho_n^{-1/2}nh^2)$.

(2) If $f(\cdot)$ and the second partial derivatives of $\boldsymbol{\beta}(s)$ are all Lipschitz continuous, then it follows from the proof of Lemma 5(2) that

$$(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = (G\mathbf{m})^{\mathrm{T}}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}}O_P(h^3 + h^2 r_n).$$

Together with (6.7) we have

$$(G\mathbf{m})^{\mathrm{T}}P\mathbf{m} = O_P(\rho_n^{-1/2}nh^3 + \{nh^2 \log n/\rho_n\}^{1/2}).$$

Proof of Lemma 11: In the following proofs we will always use the facts that the elements of G having the uniform order $O(1/\rho_n)$ and the row sums of the matrix G having the uniform order $O(1/\sqrt{\rho_n})$.

First we will show that $\frac{\rho_n}{n}\mathbf{m}^{\mathrm{T}}P\mathbf{m} = o_P(1)$. It can be seen from (6.4) that

$$\begin{aligned}\frac{\rho_n}{n}\mathbf{m}^{\mathrm{T}}P\mathbf{m} &= \frac{\rho_n}{n}(\mathbf{m} - S\mathbf{m})^{\mathrm{T}}(\mathbf{m} - S\mathbf{m}) \\ &= \frac{1}{n}(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots, X_n^{\mathrm{T}}\mathbf{1}_p)(X_1^{\mathrm{T}}\mathbf{1}_p, \cdots, X_n^{\mathrm{T}}\mathbf{1}_p)^{\mathrm{T}}O_P(\rho_n h^4) = o_P(1)\end{aligned}$$

by law of large numbers.

Now we will show that $\frac{\rho_n}{n}L^{\mathrm{T}}PG\mathbf{m} = o_P(1)$ for $L = \mathbf{m}, \boldsymbol{\epsilon}$ and $G\boldsymbol{\epsilon}$.

It follows immediately from Lemma 10(1) and $\rho_n h^4 \to 0$ that $\frac{\rho_n}{n}\mathbf{m}^{\mathrm{T}}PG\mathbf{m} = o_P(\rho_n^{1/2}h^2) = o_P(1)$.

Next by the same lines as establishing Lemma 3 and Condition (4) that

$$\frac{\sqrt{\rho_n}}{n}H^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}G\mathbf{m} = \begin{pmatrix} \Gamma\Gamma^{\mathrm{T}}\frac{\sqrt{\rho_n}}{n}\sum\limits_{i=1}^{n}\sum\limits_{j\neq i}^{n} g_{ij}\boldsymbol{\beta}(s_j)K_h(\|s_i - s\|) \\ \mathbf{0}_{2p\times 1} \end{pmatrix} + o_P(\mathbf{1}_{3p}) \tag{6.8}$$

holds uniformly in $s \in \mathcal{S}$. Then using the same lines as establishing Lemma 6(2), the facts that the elements of G having the uniform order $O(1/\rho_n)$, the row sums of the matrix G having the uniform order $O(1/\sqrt{\rho_n})$ and $\rho_n/n \to 0$, we obtain that $\frac{\rho_n}{n}\boldsymbol{\epsilon}^{\mathrm{T}}PG\mathbf{m} = o_P(1)$ and $\frac{\rho_n}{n}(G\boldsymbol{\epsilon})^{\mathrm{T}}PG\mathbf{m} = o_P(1)$.

Next it follows the same lines as establishing $n^{-1/2}(G\boldsymbol{\epsilon})^{\mathrm{T}}P\mathbf{m} = o_P(1)$ in Lemma 6(1) that $\sqrt{\rho_n/n}(G\boldsymbol{\epsilon})^{\mathrm{T}}P\mathbf{m} = o_P(1)$ when $\rho_n h^4 \to 0$.

As we have by Lemma 2(1) and (6.8) that

$$\sqrt{\rho_n}SG\mathbf{m} = \begin{pmatrix} \kappa_0^{-1}f^{-1}(s_1)X_1^{\mathrm{T}}\Psi^{-1}\Gamma\Gamma^{\mathrm{T}}\tilde{Z}(s_1) \\ \vdots \\ \kappa_0^{-1}f^{-1}(s_n)X_n^{\mathrm{T}}\Psi^{-1}\Gamma\Gamma^{\mathrm{T}}\tilde{Z}(s_n) \end{pmatrix} + o_P(1),$$

where $\tilde{Z}(s) = \lim\limits_{n\to\infty}\frac{\sqrt{\rho_n}}{n}\sum\limits_{i=1}^{n}\sum\limits_{j\neq i}^{n} g_{ij}\boldsymbol{\beta}(s_j)K_h(\|s_i - s\|)$, and

$$\frac{\rho_n}{n}(G\mathbf{m})^{\mathrm{T}}PG\mathbf{m} = \frac{1}{n}(\sqrt{\rho_n}G\mathbf{m} - \sqrt{\rho_n}SG\mathbf{m})^{\mathrm{T}}(\sqrt{\rho_n}G\mathbf{m} - \sqrt{\rho_n}SG\mathbf{m}),$$

the results (4) can be obtained similarly using the same lines as showing Lemma 7 with $G\mathbf{m}$ and $SG\mathbf{m}$ replaced by $\sqrt{\rho_n}G\mathbf{m}$ and $\sqrt{\rho_n}SG\mathbf{m}$ respectively.

The results (5) and (6) can be obtained from the proof of Lemma 8(2) and 8(3) under the assumptions of Lemma 11.

Finally, the result (7) can be obtained as Lemma 8(4).

Proof of Lemma 12: The proof can be established using the same lines as Lemma 9 under the assumptions of Lemma 12.

7 The Connection between Cross-validation and AIC in a Semiparametric Family

7.1 Introduction

Cross-validation as an important tool in statistical analysis,is intuitively appealing and easy to implement.However, it is also computationally expensive. Although cross-validation tends to pick up a model which is unnecessarily complex, it is nevertheless frequently used in practice, see Xia and Li (2002).

The AIC is another important criterion used for model selection in a family of hierarchical models, see Akaike (1973, 1974). The equivalence of cross-validation and AIC has been established in Stone (1977). Some discussions about the frequently used criteria, which includes cross-validation and AIC, for model selection can be found in Allen (1974), Arlot and Celisse (2010), Davies *et al.*(2005), and Lv and Liu (2010).

The existing works about the connection between cross-validation and AIC are mainly for parametric models. The methodologies in those works can also be extended to accommodate some semiparametric models, if the orthogonal basis based decomposition smoothing method is used to deal with the unknown functions involved in the models concerned. This is because after decomposition of the unknown functions, the models would become parametric, though with some tuning parameters. However, for semiparametric models, if kernel smoothing is used, the situation would be different. This is because the parameterization of unknown functions in kernel smoothing is done locally. On the other hand, the model selection should be done globally. How to link them together is not trivial. As far as cross-validation and AIC are concerned, a natural question is whether cross-validation is still equivalent to AIC in semiparametric models when kernel smoothing is used. This question is answered here.

We have to make some assumptions on the semiparametric models addressed due to the 'curse of dimensionality'. Let y be a response variable, U and X be covariates. U and y are scalars, X is a p-dimensional vector. We assume the conditional log-density function of y given $(U,\ X^{\mathrm{T}})$ is

$$f(y;\ \boldsymbol{\theta},\ a_1(U),\ \cdots,\ a_\kappa(U),\ \ X), \tag{7.1}$$

where $f(\cdot;\ \cdots)$ is specified, $\boldsymbol{\theta}$ is a q-dimensional unknown constant parameter, $a_j(\cdot)$, $j = 1,\ \cdots,\ \kappa$, are unknown functions.

Model (7.1) represents a large family of semiparametric models, including generalised linear models, varying coefficient models (Fan and Zhang, 1999; Sun *et al.*, 2007; Wang *et al.*, 2009; Wang and Xia, 2009; Zhang *et al.*, 2009), multiparameter likelihood models (Cheng *et al.*, 2009), partially linear models (Liang and Li, 2009; Ma *et al.*, 2006), and semivarying coefficient models (Zhang *et al.*, 2002; Li and Palta, 2009; Kai *et al.*, 2010; Li and Zhang, 2011).

In this chapter, we are going to show the connection between cross-validation and AIC in the family of semiparematric models (7.1). We begin in chapter 7.2 with a description of a maximum likelihood based semiparametric estimation procedure for the unknown functions and constants in (7.1). In chapter 7.3, we give the definitions of the AIC and cross-validation for (7.1) and show the connection between them. The connection will also be demonstrated by simulation in chapter 7.4 followed by the technical proof of a theorem which establishes the connection.

7.2 Estimation procedure

Suppose $(U_i,\ X_i^{\mathrm{T}},\ y_i)^{\mathrm{T}}$, $i = 1,\ \cdots,\ n$, are i.i.d. from $(U,\ X^{\mathrm{T}},\ y)^{\mathrm{T}}$. In this chapter, we are going to present an estimation procedure for the unknown functions and constants in (7.1).

7.2.1 Estimation of the unknown constant parameter

The estimation of the constant parameter consists of two steps: we first use local maximum likelihood estimation to get some initial estimators of the constant parameter, then average the initial estimators to get the final estimator. The details are as follows.

For any given u, by Taylor's expansion, we have

$$a_j(U_i) \approx a_j(u) + \dot{a}_j(u)(U_i - u)$$

which leads to the local log-likelihood function

$$L(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b}) = \sum_{i=1}^{n} f(y_i;\ \boldsymbol{\theta},\ a_1{+}b_1(U_i{-}u),\ \cdots,\ a_\kappa{+}b_\kappa(U_i{-}u),\ X_i)K_h(U_i{-}u), \tag{7.2}$$

where $\mathbf{a} = (a_1,\ \cdots,\ a_\kappa)^{\mathrm{T}}$, $\mathbf{b} = (b_1,\ \cdots,\ b_\kappa)^{\mathrm{T}}$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, h is a bandwidth.

Let $(\tilde{\boldsymbol{\theta}}(u)^{\mathrm{T}},\ \tilde{\mathbf{a}}(u)^{\mathrm{T}},\ \tilde{\mathbf{b}}(u)^{\mathrm{T}})$ maximise (7.2). $\tilde{\boldsymbol{\theta}}(u)$ is an initial estimator of $\boldsymbol{\theta}$. Let u go over U_i, $i = 1, \cdots, n$. The final estimator of $\boldsymbol{\theta}$ is taken to be

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{\theta}}(U_i).$$

7.2.2 Estimation of unknown functions

The estimation of unknown functions is the standard local maximum likelihood estimation: replacing the $\boldsymbol{\theta}$ in (7.2) by its estimator and changing the bandwidth h to a slightly larger one h_1. We have the following objective function for the estimation of unknown functions

$$\sum_{i=1}^{n} f(y_i;\, \hat{\boldsymbol{\theta}},\, a_1+b_1(U_i-u),\, \cdots,\, a_\kappa+b_\kappa(U_i-u),\, X_i)K_{h_1}(U_i-u). \quad (7.3)$$

Let $(\hat{a}_1(u), \cdots, \hat{a}_\kappa(u), \hat{b}_1(u), \cdots, \hat{b}_\kappa(u))$ maximise (7.3). $\hat{a}_j(u)$ is the estimator of $a_j(u)$, $j = 1, \cdots, \kappa$.

The reason to change the bandwidth h to a slightly larger one h_1 is that the bandwidth used in estimation of the unknown constant parameter is usually smaller than the optimal bandwidth for estimation of unknown functions in order to achieve a better convergence rate.

The estimators of the unknown functions and unknown constants are needed in the computation of either AIC or CV. The estimation procedure presented in this chapter is to provide such estimators.

7.3 Equivalence of CV and AIC

In this chapter, we will first give the definitions of cross-validation and AIC, then show that they are equivalent to each other when the kernel function is taken to be the density function of the uniform distribution on $[-1,\ 1]$.

7.3.1 Cross-validation

For each i, $i = 1, \cdots, n$, we delete the ith observation and estimate $\boldsymbol{\theta}$ and $a_j(\cdot)$, $j = 1, \cdots, \kappa$, based on the other observations. The resulting estimators are denoted by $\hat{\boldsymbol{\theta}}^{\setminus i}$ and $\hat{a}_j^{\setminus i}(\cdot)$, respectively. The cross-validation sum is defined as

$$\mathrm{CV} = -\sum_{i=1}^{n} f(y_i;\ \hat{\boldsymbol{\theta}}^{\setminus i},\ \hat{a}_1^{\setminus i}(U_i),\ \cdots,\ \hat{a}_\kappa^{\setminus i}(U_i),\ \ X_i).$$

7.3.2 AIC

To define AIC for semiparametric models is not straightforward when kernel smoothing is used. The main problem is to find out how many unknown constant parameters an unknown function, in general, amounts to. Cheng *et al.*(2009) came up with an ad-hoc solution for this problem, and suggested that an unknown function amounts to $h^{-1}(\nu_0 + \nu_2/\mu_2)$ unknown parameters, where

$$\mu_i = \int u^i K(u)du, \quad \nu_i = \int u^i K^2(u)du.$$

While their ad-hoc solution did work well for their models, it is nevertheless worth revisiting this problem more thoroughly as their approach is based on the local residual sum of squares, and this problem, as it stands, should be in a global sense and should be solved globally.

To find out how many unknown constant parameters an unknown function, in general, amounts to, we only need to come down to the standard univariate nonparametric regression model

$$\eta_i = a(U_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{7.4}$$

where (U_i, η_i), $i = 1, \cdots, n$, are i.i.d. The degrees of freedom of the residual sum of squares of (7.4) can be reasonably viewed as 'n − *the number of unknown parameters in (7.4)*', which implies that unknown function $a(\cdot)$ amounts to

$$\mathcal{T} = n - \frac{1}{\mathrm{var}(\epsilon_1)} E\left\{ \sum_{i=1}^{n} (\eta_i - \hat{a}(U_i))^2 \, | \mathcal{D} \right\}$$

unknown constant parameters, where $\mathcal{D} = (U_1, \cdots, U_n)$, $\hat{a}(U_i)$ is the local linear estimator of $a(U_i)$. Using the standard argument in Fan and Gijbels (1996) and Lemma 1 in chapter 7.5, we have

$$\mathcal{T} = (2K(0) - \nu_0) h^{-1} (1 + o_P(1))$$

when $h = o_P(n^{-1/5})$ and $nh \longrightarrow \infty$, where $K(\cdot)$ is the kernel function to produce the local linear estimator $\hat{a}(U_i)$, h is the bandwidth. So, we conclude that an unknown function amounts to $(2K(0) - \nu_0)h^{-1}$ unknown constant parameters, and define the AIC for (7.1) as

$$\mathrm{AIC} = -\sum_{i=1}^{n} f(y_i; \ \hat{\boldsymbol{\theta}}, \ \hat{a}_1(U_i), \ \cdots, \ \hat{a}_\kappa(U_i), \ \ X_i) + \mathcal{K},$$

where $\mathcal{K}$ is the number of "unknown parameters" in the model, which is

$$\mathcal{K} = q + \kappa(2K(0) - \nu_0) h_1^{-1}.$$

Remark: *In general, bias is associated with an overly parsimonious model, whereas excessive variability is associated with an overly complex model. Clearly, the term,* $(2K(0) - \nu_0)h_1^{-1}$*, increases as the bandwidth decreases, reflecting an inflation in variability, and the term decreases as the bandwidth increases, reflecting an inflation in bias. Thus, complexity is inversely related to bandwidth.*

The connection between CV and AIC is established through the following theorem.

Theorem 1. *Under the conditions (1) - (6) stated in chapter 7.5, we have*

$$CV = AIC - \kappa(K(0) - \nu_0)h_1^{-1} + o_P(1).$$

Remark: When the kernel function is taken to be the density function of the uniform distribution on $[-1,\ 1]$, it is easy to see $K(0) = \nu_0$, which implies that the CV is asymptotically the same as AIC.

Theorem 1 provides not only the connection between CV and AIC but also a way to compute the CV, which would significantly reduce the computational burden in the computation of CV.

7.4 Simulation study

In this chapter, we are going to use a simulated example to demonstrate the connection between CV and AIC. We will also use either of these two criteria to do model selection, and compare their performances.

Example 1. We generated a sample $(y_i,\ X_i,\ U_i)$, $i = 1,\ \cdots,\ n$, from the logistic regression model

$$\log\left\{\frac{P(y=1|X,\ U)}{1-P(y=1|X,\ U)}\right\} = x_1a_1(U) + x_2a_2(U) + x_3a_3 + x_4a_4.$$

Here the $X_i = (x_{i1},\ x_{i2},\ x_{i3},\ x_{i4})^{\mathrm{T}}$ were independently generated from a normal distribution $N(\mathbf{0},\ I_4)$, and the U_i were independently generated from a uniform distribution $U(0,\ 1)$. We set

$$a_1(u) = \sin(2\pi u), \quad a_2(u) = \cos(2\pi u), \quad a_3 = 2, \quad a_4 = 1.$$

We use

$$d = (\mathrm{CV} - \mathrm{AIC} + \kappa(K(0) - \nu_0)h_1^{-1})/(q + \kappa K(0)h_1^{-1})$$

to measure the difference between CV and AIC$-\kappa(K(0)-\nu_0)h_1^{-1}$. The reason for us to use d rather than $(\mathrm{CV} - \mathrm{AIC} + \kappa(K(0) - \nu_0)h_1^{-1})/\mathrm{CV}$ to measure the difference is that CV is usually very large and the ratio would be very small. In fact, from the proof of Theorem 1, we can see

$$\mathrm{CV} = -\sum_{i=1}^{n} f(y_i;\ \hat{\boldsymbol{\theta}},\ \hat{a}_1(U_i),\ \cdots,\ \hat{a}_\kappa(U_i),\ \ X_i) + (q + \kappa K(0)h_1^{-1}).$$

So, it is $(q + \kappa K(0)h_1^{-1})$ that plays a key role in CV. So, we use d to measure the difference between CV and AIC $- \kappa(K(0) - \nu_0)h_1^{-1}$.

We set the sample size n in the range from 200 to 2000. For each given n, we do 100 simulations, and for each simulation, we compute the d. The mean and standard deviation (SD) of the ds obtained from the 100 simulations for each given n are presented in Figure 1. Figure

1 shows the CV is indeed very close to the AIC$-\kappa(K(0)-\nu_0)h_1^{-1}$ when sample size is larger than 1500, which is in line with our theoretical result presented in Theorem 1.

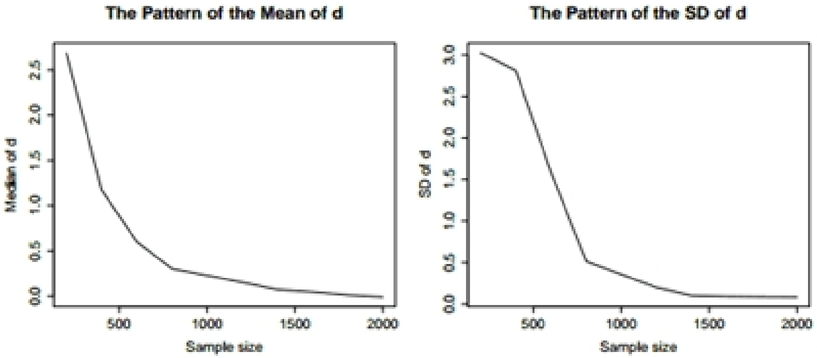

Figure 1: *The left figure is for the mean of d, in which the horizontal axis is sample size, the vertical axis is the mean of the d. The right figure is for the standard deviation of d, in which the horizontal axis is sample size, the vertical axis is the standard deviation of the d.*

We now pretend we don't know which coefficients in the model, from which the sample is generated, are functional, and apply either of the two criteria to identify the functional coefficients. The candidate family is

$$\bigcup_{k=1}^{4} \bigcup_{1\le i_1<\cdots<i_k\le 4} \{\mathcal{M}_{i_1,\cdots,i_k}\} \bigcup \{\mathcal{M}_0\}$$

where $\mathcal{M}_{i_1,\cdots,i_k}$ represents the model

$$\log\left\{\frac{P(y=1|X,\ U)}{1-P(y=1|X,\ U)}\right\} = \sum_{j\in\{i_1,\ \cdots,\ i_k\}} x_j a_j(U) + \sum_{l\in\{i_1,\ \cdots,\ i_k\}^c} x_l a_l$$

and $\mathcal{M}_0$ represents the model

$$\log\left\{\frac{P(y=1|X,\ U)}{1-P(y=1|X,\ U)}\right\} = \sum_{j=1}^{4} x_j a_j$$

where $\{i_1,\ \cdots,\ i_k\}^c$ is the complement of $\{i_1,\ \cdots,\ i_k\}$.

We set the sample size to be 1500. The reason to set such a large sample size is because binary data carries much less information, and there are two unknown functions and two unknown constants to estimate in the model. In order to have enough information to construct decent estimators of the unknowns, we have to set the sample size in the magnitude of thousands. Because the computation involved in the model selection is very expensive, we only carry out 100 simulations. In each simulation, we compute the CV and AIC for each potential candidate model, and select the one with the smallest CV for the CV based approach. Similarly we select the smallest AIC for the AIC based approach. We find, in the 100 simulations, the ratio of picking the right model, $\mathcal{M}_{1,2}$, is 95% for the CV based approach, and 94% for the AIC based approach. From these results, we can see both criteria perform reasonably well, and their performances are similar. We also set the sample size to be 1200 and 2000, the obtained results are similar, which are presented in Table 1 and 2.

Table 1: **The Ratio of AIC Picking Each Candidate Model**

Sample Size	$\mathcal{M}_{1,2}$	$\mathcal{M}_{1,2,3}$	$\mathcal{M}_{1,2,4}$	$\mathcal{M}_{1,2,3,4}$	others
$n = 1200$	93%	5%	2%	0%	0%
$n = 1500$	94%	5%	1%	0%	0%
$n = 2000$	97%	1%	1%	1%	0%

Table 2: **The Ratio of CV Picking Each Candidate Model**

Sample Size	$\mathcal{M}_{1,2}$	$\mathcal{M}_{1,2,3}$	$\mathcal{M}_{1,2,4}$	$\mathcal{M}_{1}$	$\mathcal{M}_{2,3}$	others
$n = 1200$	93%	3%	2%	1%	1%	0%
$n = 1500$	95%	5%	0%	0%	0%	0%
$n = 2000$	97%	3%	0%	0%	0%	0%

7.5 Proofs

The connection between CV and AIC is built on the following technical conditions

(1) $h = o(n^{-1/6})$, $h_1 = o(n^{-1/6})$, $nh^3 \longrightarrow \infty$, $nh_1^3 \longrightarrow \infty$.

(2) Let $\mathbf{0}_{p\times q}$ be a $p \times q$ matrix with each entry being 0. We assume $\dot{L}(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b}) = \mathbf{0}_{(2\kappa+q)\times 1}$ has an unique root.

(3) The density function of U, $g(u)$, is continuous and positive on its support $[0,\ 1]$. The second derivative of $\mathbf{a}(\cdot)$ is continuous.

(4) The kernel function $K(\cdot)$ is a symmetric and positive density function, and has bounded derivative on its support set $[-A,\ A]$.

(5) For some $s > 2$, $E(|X|^{2s}|U = u) < \infty$ is continuous and $E(y^{2s}|U = u,\ X = x) < \infty$.

(6) $f(y;\ \boldsymbol{\theta},\ a_1,\ \cdots,\ a_\kappa,\ X)$ has bounded second derivative with respect to $(\boldsymbol{\theta}^{\mathrm{T}},\ a_1,\ \cdots,\ a_\kappa)$.

The following Lemma in Fan and Zhang (1999) is needed in the proof of the Theorem.

Lemma 1. *Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d random vectors, where the Y_i's are scalar random variables. Assume further that $E|y|^s < \infty$ and $\sup_x \int |y|^s f(x, y)dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_{x \in D} |n^{-1} \sum_{i=1}^{n} \{K_h(X_i - x)Y_i - E[K_h(X_i - x)Y_i]\}| = O_P[\{nh/\log(1/h)\}^{-1/2}]$$

provided that $n^{2\varepsilon - 1}h \longrightarrow \infty$ for some $\varepsilon < 1 - s^{-1}$.

Proof of Theorem 1:

Let

$$\mathbf{a}(\cdot) = (a_1(\cdot),\ \cdots,\ a_\kappa(\cdot))^{\mathrm{T}}, \quad \hat{\mathbf{a}}(\cdot) = (\hat{a}_1(\cdot),\ \cdots,\ \hat{a}_\kappa(\cdot))^{\mathrm{T}},$$

$$\hat{\mathbf{a}}^{\backslash i}(\cdot) = (\hat{a}_1^{\backslash i}(\cdot),\ \cdots,\ \hat{a}_\kappa^{\backslash i}(\cdot))^{\mathrm{T}}.$$

The AIC and CV can be written as

$$\mathrm{CV} = -\sum_{i=1}^{n} f(y_i;\ \hat{\boldsymbol{\theta}}^{\backslash i},\ \hat{\mathbf{a}}^{\backslash i}(U_i),\ X_i), \quad \mathrm{AIC} = -\sum_{i=1}^{n} f(y_i;\ \hat{\boldsymbol{\theta}},\ \hat{\mathbf{a}}(U_i),\ X_i) + \mathcal{K}.$$

Let

$$\Delta = \sum_{i=1}^{n} \Big\{ f(y_i;\ \hat{\boldsymbol{\theta}},\ \hat{\mathbf{a}}(U_i),\ \ X_i) - f(y_i;\ \hat{\boldsymbol{\theta}}^{\setminus i},\ \hat{\mathbf{a}}^{\setminus i}(U_i),\ \ X_i) \Big\}$$

It is obvious that

$$\mathrm{CV} = -\sum_{i=1}^{n} f(y_i;\ \hat{\boldsymbol{\theta}},\ \hat{\mathbf{a}}(U_i),\ \ X_i) + \Delta,$$

and

$$\Delta = -\sum_{i=1}^{n} \Big((\hat{\boldsymbol{\theta}}^{\setminus i} - \hat{\boldsymbol{\theta}})^{\mathrm{T}},\ (\hat{\mathbf{a}}^{\setminus i}(U_i) - \hat{\mathbf{a}}(U_i))^{\mathrm{T}} \Big)\, \dot{f}(y_i;\ \hat{\boldsymbol{\theta}},\ \hat{\mathbf{a}}(U_i),\ X_i)(1{+}o_P(1)), \tag{7.5}$$

where

$$\dot{f}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \Big(\dot{f}_1(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)^{\mathrm{T}},\ \dot{f}_2(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)^{\mathrm{T}} \Big)^{\mathrm{T}}$$

$$\dot{f}_1(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \partial f(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)/\partial \boldsymbol{\theta}, \quad \dot{f}_2(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \partial f(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)/\partial \mathbf{a}.$$

Let I_k be an identity matrix of size k,

$$L^{\setminus i}(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b}) = \sum_{k=1,\ k\neq i}^{n} f(y_k;\ \boldsymbol{\theta},\ \mathbf{a} + \mathbf{b}(U_k - u),\ X_k) K_h(U_k - u),$$

$$\dot{L}(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b}) = \partial L / \partial (\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{a}^{\mathrm{T}},\ \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}, \quad H_i = \mathrm{diag}\left(I_{q+\kappa},\ (U_i - u) I_\kappa \right),$$

and

$$\mathbf{F}(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k) = \Big(\dot{f}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)^{\mathrm{T}},\ \dot{f}_2(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)^{\mathrm{T}} \Big)^{\mathrm{T}}.$$

By simple calculation, we have

$$\dot{L}(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b}) = \sum_{k=1}^{n} H_k \mathbf{F}(y_k;\ \boldsymbol{\theta},\ \mathbf{a} + \mathbf{b}(U_k - u),\ X_k) K_h(U_k - u).$$

We suppress the u in $\tilde{\boldsymbol{\theta}}(u)$, $\tilde{\mathbf{a}}(u)$, or $\tilde{\mathbf{b}}(u)$ to make the notations more simple. Let $\mathbf{a}_0 = \mathbf{a}(u)$, $\mathbf{b}_0 = \dot{\mathbf{a}}(u)$, and θ_0 be the true $\boldsymbol{\theta}$. By the Taylor's expansion and

$$\dot{L}(\tilde{\boldsymbol{\theta}},\ \tilde{\mathbf{a}},\ \tilde{\mathbf{b}}) = \mathbf{0}_{(2\kappa+q)\times 1},$$

we have

$$\mathbf{0}_{(2\kappa+q)\times 1} = H_0^{-1}\mathbf{J}_1 + H_0^{-1}(\mathbf{J}_2 + \mathbf{J}_3)(\boldsymbol{\xi} - \boldsymbol{\xi}_0) + O_P\left[n\left\{\|H_0(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\|^3\right\}\right], \tag{7.6}$$

where H_0 is the H_i with $(U_i - u)$ being replaced by h, and

$$\boldsymbol{\xi} = (\tilde{\boldsymbol{\theta}}^{\mathrm{T}},\ \tilde{\mathbf{a}}^{\mathrm{T}},\ \tilde{\mathbf{b}}^{\mathrm{T}})^{\mathrm{T}}, \quad \boldsymbol{\xi}_0 = (\boldsymbol{\theta}_0^{\mathrm{T}},\ \mathbf{a}_0^{\mathrm{T}},\ \mathbf{b}_0^{\mathrm{T}})^{\mathrm{T}},$$

$$\mathbf{J}_1 = \sum_{k=1}^{n} H_k \mathbf{F}(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0 + \mathbf{b}_0(U_k - u),\ X_k) K_h(U_k - u),$$

$$\mathbf{J}_2 = \sum_{k=1}^{n} H_k G(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0 + \mathbf{b}_0(U_k - u),\ X_k) H_k K_h(U_k - u),$$

$$G(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k) = \begin{pmatrix} \ddot{f}_{11}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{21}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{21}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) \\ \ddot{f}_{12}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{22}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{22}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) \\ \ddot{f}_{12}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{22}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) & \ddot{f}_{22}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) \end{pmatrix}$$

$$\ddot{f}_{11}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \partial^2 f(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}},$$

$$\ddot{f}_{12}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \partial^2 f(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)/\partial\boldsymbol{\theta}\partial\mathbf{a}^{\mathrm{T}},$$

$$\ddot{f}_{22}(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i) = \partial^2 f(y_i;\ \boldsymbol{\theta},\ \mathbf{a},\ X_i)/\partial\mathbf{a}\partial\mathbf{a}^{\mathrm{T}}, \quad \mathbf{J}_3 = (J_1,\ \cdots,\ J_{q+2\kappa})^{\mathrm{T}},$$

$$J_l = \sum_{k=1}^{n}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^{\mathrm{T}} H_k \ddot{F}_l(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0+\mathbf{b}_0(U_k-u),\ X_k) H_k K_h(U_k-u), \quad l \le q+\kappa,$$

when $l > q + \kappa$,

$$J_l = \sum_{k=1}^{n}(U_k-u)(\boldsymbol{\xi}-\boldsymbol{\xi}_0)^{\mathrm{T}} H_k \ddot{F}_l(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0+\mathbf{b}_0(U_k-u),\ X_k) H_k K_h(U_k-u),$$

where

$$(F_1(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k),\ \cdots,\ F_{q+2\kappa}(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k))^{\mathrm{T}} = \mathbf{F}(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k),$$

and

$$\ddot{F}_l(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k) = \partial^2 F_l(y_k;\ \boldsymbol{\theta},\ \mathbf{a},\ X_k)/\partial(\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{a}^{\mathrm{T}},\ \mathbf{a}^{\mathrm{T}})^{\mathrm{T}}\partial(\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{a}^{\mathrm{T}},\ \mathbf{a}^{\mathrm{T}}).$$

Let $\boldsymbol{\xi}^{\backslash i}$ and $\mathbf{J}_l^{\backslash i}$, $l = 1,\ 2,\ 3$, be the counterparts of $\boldsymbol{\xi}$ and $\mathbf{J}_l$ when the ith observation is deleted. Obviously,

$$\mathbf{0}_{(2\kappa+q)\times 1} = H_0^{-1}\mathbf{J}_1^{\backslash i} + H_0^{-1}(\mathbf{J}_2^{\backslash i} + \mathbf{J}_3^{\backslash i})(\boldsymbol{\xi}^{\backslash i} - \boldsymbol{\xi}_0) + O_P\left[n\left\{\|H_0(\boldsymbol{\xi}^{\backslash i} - \boldsymbol{\xi}_0)\|^3\right\}\right]. \tag{7.7}$$

Using exactly the same argument as that in the proof of Theorem 2 in Zhang and Peng (2010), we have

$$\|H_0(\boldsymbol{\xi}^{\backslash i} - \boldsymbol{\xi}_0)\| = O_P(\delta_n) \quad \text{and} \quad \|H_0(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\| = O_P(\delta_n) \tag{7.8}$$

uniformly in terms of u, where $\delta_n = h^2 + \{-\log(h)/(nh)\}^{1/2}$. (7.8) together with (7.6) and (7.7) lead to

$$H_0^{-1}\mathbf{J}_1 + H_0^{-1}(\mathbf{J}_2+\mathbf{J}_3)(\boldsymbol{\xi}-\boldsymbol{\xi}_0) = H_0^{-1}\mathbf{J}_1^{\backslash i} + H_0^{-1}(\mathbf{J}_2^{\backslash i}+\mathbf{J}_3^{\backslash i})(\boldsymbol{\xi}^{\backslash i}-\boldsymbol{\xi}_0) + O_P\left(n\delta_n^3\right)$$

which leads to

$$\begin{aligned}\boldsymbol{\xi}^{\backslash i} - \boldsymbol{\xi} \;=\;& \mathbf{J}_2^{-1}\Big\{H_i\mathbf{F}(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0 + \mathbf{b}_0(U_i-u),\ X_i)K_h(U_i-u) \\ & +H_iG(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0 + \mathbf{b}_0(U_i-u),\ X_i)H_iK_h(U_i-u)(\boldsymbol{\xi}^{\backslash i}-\boldsymbol{\xi}_0)\Big\} \\ & +O_P(\delta_n^3). \qquad (7.9)\end{aligned}$$

Using Lemma 1, by some tedious calculations, we have

$$\mathbf{J}_2 = ng(u)H_0\Omega(u)H_0(1+o_P(1))$$

uniformly, where

$$\Omega(u) = \mathrm{diag}\left(E\left\{\ddot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0,\ X)|U=u\right\},\ \mu_2E\left\{\ddot{f}_{22}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0,\ X)|U=u\right\}\right)$$

and

$$\ddot{f}(y;\ \boldsymbol{\theta},\ \mathbf{a},\ X) = \begin{pmatrix} \ddot{f}_{11}(y;\ \boldsymbol{\theta},\ \mathbf{a},\ X) & \ddot{f}_{21}(y;\ \boldsymbol{\theta},\ \mathbf{a},\ X) \\ \ddot{f}_{12}(y;\ \boldsymbol{\theta},\ \mathbf{a},\ X) & \ddot{f}_{22}(y;\ \boldsymbol{\theta},\ \mathbf{a},\ X)\end{pmatrix}.$$

So,

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\backslash i} - \hat{\boldsymbol{\theta}} \;=\;& \frac{1}{n^2}(I_q,\ \mathbf{0}_{q\times 2\kappa})\sum_{k=1}^{n}\Big\{g(U_k)^{-1}H_0^{-1}\Omega(U_k)^{-1}H_0^{-1}H_{i,k} \\ & \mathbf{F}(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k) + \dot{\mathbf{a}}(U_k)(U_i-U_k),\ X_i)K_h(U_i-U_k)\Big\}\times\end{aligned}$$

$$(1+o_P(1)), \tag{7.10}$$

where $H_{i,k}$ is the H_i with u being replaced by U_k.

Replacing the $\boldsymbol{\theta}$ in $L(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b})$ by $\hat{\boldsymbol{\theta}}$, in $L^{\setminus i}(\boldsymbol{\theta},\ \mathbf{a},\ \mathbf{b})$ by $\hat{\boldsymbol{\theta}}^{\setminus i}$, and using exactly the same argument as that for deriving (7.9), we have

$$\hat{\mathbf{a}}^{\setminus i}(u)-\hat{\mathbf{a}}(u) = \frac{1}{ng(u)}\left[E\left\{\ddot{f}_{22}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0,\ X)|U=u\right\}\right]^{-1}(\Gamma_1+\Gamma_2)(1+o_P(1)), \tag{7.11}$$

where

$$\Gamma_1 = \dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0+\mathbf{b}_0(U_i-u),\ X_i)K_{h_1}(U_i-u),$$

$$\Gamma_2 = -\sum_{k\neq i}\ddot{f}_{12}(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}_0+\mathbf{b}_0(U_k-u),\ X_k)(\hat{\boldsymbol{\theta}}^{\setminus i}-\hat{\boldsymbol{\theta}})K_{h_1}(U_k-u).$$

(7.5), (7.10) and (7.11) together with Lemma 1 leads to

$$\Delta = -(\Delta_1+\Delta_2)(1+o_P(1))$$

with

$$\Delta_1 = \sum_{i=1}^{n}\dot{f}_1(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\quad X_i)^{\mathrm{T}}(\hat{\boldsymbol{\theta}}^{\setminus i}-\hat{\boldsymbol{\theta}})$$

and

$$\Delta_2 = \sum_{i=1}^{n}\dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\quad X_i)^{\mathrm{T}}\left(\hat{\mathbf{a}}^{\setminus i}(U_i)-\hat{\mathbf{a}}(U_i)\right),$$

and

$$\begin{aligned}\Delta_1 \quad = \quad & \frac{1}{n^2}\sum_{k=1}^{n}g(U_k)^{-1}\sum_{i=1}^{n}\Big\{\dot{f}_1(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\quad X_i)^{\mathrm{T}}\times \\ & (I_q,\ \mathbf{0}_{q\times\kappa})\left[E\left\{\ddot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k),\ X)|U=U_k\right\}\right]^{-1}\times\end{aligned}$$

$$
\begin{aligned}
& \dot{f}(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k)+\dot{\mathbf{a}}(U_k)(U_i-U_k),\ X_i)K_h(U_i-U_k)\Big\}\times \\
& (1+o_P(1)) \\
= \quad & \frac{1}{n}\sum_{k=1}^{n}\mathrm{tr}\Big[(I_q,\ \mathbf{0}_{q\times\kappa})\left[E\left\{\ddot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k),\ X)|U=U_k\right\}\right]^{-1}\times \\
& E\left\{\dot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k),\ X)\dot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_k),\ X)^{\mathrm{T}}|U=U_k\right\}(I_q,\ \mathbf{0}_{q\times\kappa})^{\mathrm{T}}\Big]\times \\
& (1+o_P(1)) \\
= \quad & -q(1+o_P(1)),
\end{aligned}
$$

$$
\Delta_2=(\Delta_{2,1}+\Delta_{2,2})(1+o_P(1)),
$$

where

$$
\begin{aligned}
\Delta_{2,1} \quad = \quad & \frac{K_{h_1}(0)}{n}\sum_{i=1}^{n}\Big\{\frac{1}{g(U_i)}\dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ \ X_i)^{\mathrm{T}} \\
& \left[E\left\{\ddot{f}_{22}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ X)|U=U_i\right\}\right]^{-1}\dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ \ X_i)\Big\} \\
= \quad & -K_{h_1}(0)\kappa+O_P(n^{-1/2}h_1^{-1}),
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta_{2,2} \quad = \quad & -\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{g(U_i)}\dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ \ X_i)^{\mathrm{T}} \\
& \left[E\left\{\ddot{f}_{22}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ X)|U=U_i\right\}\right]^{-1} \\
& \sum_{k\neq i}\ddot{f}_{12}(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i)+\dot{\mathbf{a}}(U_i)(U_k-U_i),\ X_k)(\hat{\boldsymbol{\theta}}^{\setminus i}-\hat{\boldsymbol{\theta}})K_{h_1}(U_k-U_i)\Big\} \\
= \quad & -\frac{1}{n^3}\sum_{l=1}^{n}\sum_{i=1}^{n}\sum_{k\neq i}^{n}\Big\{\frac{K_h(U_i-U_l)K_{h_1}(U_k-U_i)}{g(U_i)g(U_l)}\dot{f}_2(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ \ X_i)^{\mathrm{T}} \\
& \left[E\left\{\ddot{f}_{22}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i),\ X)|U=U_i\right\}\right]^{-1} \\
& \ddot{f}_{12}(y_k;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_i)+\dot{\mathbf{a}}(U_i)(U_k-U_i),\ X_k)
\end{aligned}
$$

$$
\begin{aligned}
&(I_q,\ \mathbf{0}_{q\times\kappa})\left[E\left\{\ddot{f}(y;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_l),\ X)|U=U_l\right\}\right]^{-1}\\
&\dot{f}(y_i;\ \boldsymbol{\theta}_0,\ \mathbf{a}(U_l)+\dot{\mathbf{a}}(U_l)(U_i-U_l),\ X_i)\Big\}\left(1+o_P(1)\right)=o_P(1).
\end{aligned}
$$

So,

$$\Delta = q + h_1^{-1}K(0)\kappa + o_P(1)$$

which leads to

$$\text{CV} = \text{AIC} - \kappa(K(0)-\nu_0)h_1^{-1} + o_P(1).$$

8 Model selection

8.1 Thresholding K method

8.1.1 Introduction

As we mentioned in Chapter 3, some components of $\boldsymbol{\beta}(\cdot)$ in model (1.3) may be constant in reality, and it is important to identify such constant components.To identify the constant components is basically a model selection problem. In this chapter, we introduce several model selection methods. First, we present the *Thresholding K* method. We treat all components $\beta_1(\cdot), \cdots, \beta_p(\cdot)$ as functional and estimate them in the simulations.After we get the estimator of each component $\hat{\beta}_1(\cdot), \cdots, \hat{\beta}_p(\cdot)$, we calculate the discrepancy of the estimator from its average. For example, say the p^{th} component, we define its discrepancy as k_p, $k_p = \sum_{i=1}^{i=n}(\hat{\beta}_p(S_i) - \hat{\beta}_p)$, here $\hat{\beta}_p = \sum_{i=1}^{i=n} \hat{\beta}_p(S_i)/n$. We then sort these p discrepancies in descending order, suppose the order is $k_{(1)} < k_{(2)} < k_{(3)} \cdots < k_{(p)}$.We use this value range as the starting and ending point of the thresholding method. Here, we use the minimum discrepancy $k_{(1)}$ as the initial thresholding value K_1. We increase the starting thresholding value until it reaches the maximum discrepancy. We calculate the Mean Integrated Squared of Error each time under different thresholding values. The ones with the minimum MISE values are our optimal thresholding K_0. The details are illustrated below

8.1.2 Identify the constant component

(1) We take the initial K_1 as thresholding to identify the constant component. For the j^{th} simulation, we generate data set j, and es-

timate all the components as functional. Then we calculate the discrepancy of each component. If the discrepancy of the component is smaller than K_1, we treat this component as constant. Suppose, only $\beta_1(\cdot)$ has a discrepancy smaller than K_1, we treat $\beta_1(\cdot)$ as a constant component denoted as β_1. We then conduct the estimation procedure again,but this time, we treat $\beta_1(\cdot)$ as constant. Suppose the obtained estimated $\boldsymbol{\beta}$ is : $\boldsymbol{\beta} = (\hat{\beta}_1, \hat{\beta}_2(\cdot), \cdots, \hat{\beta}_p(\cdot))$. We calculate the ISE for the j^{th} simulation, which is defined as:

$$ISE_j = \frac{1}{n}\sum_{i=1}^{n}[(\hat{\alpha} - \alpha)\sum_{k \neq i} w_{ik} y_k + X_i^T(\hat{\beta}(s_i) - \beta(s_i))]^2 \tag{8.1}$$

(2)We continue the procedure for n times. We calculate ISE value every time. Then, we could get the mean integrated squared of error for the thresholding value K_1, we denote the value as $MISE(K_1) = \sum_{j=1}^{j=n} ISE_j/n$.

(3)As we said before, we use the minimum discrepancy as the initial K_1, which may not be the optimal one.Now, we will increase the thresholding value until it reaches the maximum discrepancy $k_{(p)}$. In our simulations, we increase 30% each time.This is due to the computational limitations. For different thresholding value K_j, we could calculate the $MISE(K_j)$ for each K. The one with minimum $MISE$ value is the optimal threshold . From the simulation results, we notice that the optimal threshold is not unique, in fact, it is an interval.

(4) After we find the optimal thresholding, we conduct the simulation n times again.We use the ratios to evaluate the performance of this

method. We calculate three ratios :The ratio of picking right model, denote as R ; The ratio of picking wrong model, treating constant as function; The ratio of picking wrong model, treating function as constant. From the simulation, we find the results are quite satisfied.The thresholding method works very well.

8.1.3 Difficulty of the thresholding method

The key idea of the thresholding method is to select an optimal thresholding to identify the constant component. The natural question is how to select the optimal threshold. In the previous chapter, we defined the thresholding values with minimum mean integrated squared error as the optimal threshold K_0. In simulations,we calculate the $MISEs$ based on what was known of the true model. So the key point of finding the optimal thresholding value is the same as finding the $MISEs(K)$ of each thresholding value K. If we find a good estimator of $MISE(K)$, denoted as $\hat{M}(K)$,then finding the optimal thresholding K would not be problematic. Unfortunately, we have not thus far come up with a method for finding the optimal thresholding value K_0. Finding an accurate estimator of $MISEs(K)$ will be complicated work, and we intend to pursue such research in future work.

8.1.4 The aim of thresholding method

The main aim of the thresholding method is a benchmark. We compare the results we get from the thresholding method and from AIC/BIC in simulations.We mentioned the challenge of applying the thresholding method in real data analysis. It is very difficult to find the optimal thresholding value K_0 in real data analysis.We could see the thresholding method do works very well in simulations. If the results we get from AIC/BIC can compare with the results using the thresholding method,we could consider AIC/BIC as a powerful tool for identifying the constant component. AIC/BIC can be easily applied to real data analysis, and the results are comparable. We applied the AIC method to the Boston House Price data set. The results are in the chapter covering real data analysis.

8.2 Curvature-to-Average ratio based method to identify the constant component

In the previous chapter, we derived the thresholding K method to help us identify the constant components. In the thresholding method, we calculate the discrepancy of each component, and then sort them in the increasing order. We use the smallest discrepancy as our initial thresholding value. In the simulation studies, we could tell the thresholding method do provide us fancy results. However, we came across the scale problem. For example, the component $\beta_i(s_t)$ is constant at every location $s_t(u_t, v_t)$. However, the value of β_i is large. The component $\beta_j(s_t)$ is functional. But the value of $\beta_i(\cdot)$ is extremely small at every location $s_t(u_t, v_t)$ .Under this situation, the discrepancy of the constant component is much larger than the functional component.We will easily identify the wrong component as constant due to the scale problem, which occurs in real data sets. Accordingly, we come up with a new method called Curvature-to-Average (CTAR) based method. This method remove the effects of the scales. The basic ideas are sim-

ilar to thresholding method, and the details are listed below.
(1) For each location $S_i(u_i, v_i)$, we find the estimator of the unknown function $\beta_j(S_i)$, denoted as $\hat{\beta}_j(S_i)$. We treat each component as functional component at the beginning.
(2) We calculate the average value of each component, we denote it as $\bar{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_j(S_i)$
(3) We then calculate the noise $s = \sqrt{\sum\limits_{i=1}^{n} \left(\hat{\beta}_j(S_i) - \bar{\beta}_j\right)^2 / n - 1}$ .We can get the ratio for the j_{th} component β_j, i.e, $r_j = s/\bar{\beta}_j$.
(4) We use the ratio r_j as a thresholding. We set an thresholding λ. If the ratio r_j is smaller than the thresholding λ, then we treat the component β_j as constant. Otherwise, the component is functional. We repeat the procedure for n times, then we can calculate the ratio of picking the right model.

The CTAR method well deal with the scale problem , and simulation results showed it works very well. However, we are still faced with the question of how to find the optimal λ.The difficulty of finding the optimal λ is similar to the difficulty of finding the optimal threshold K_0 as we mentioned previously.

8.3 Identification of constant components based on AIC and BIC

8.3.1 Criterion for identification

In this chapter, we appeal the AIC or BIC to identify the constant components. The AIC for (1.3), in which some components of $\boldsymbol{\beta}(\cdot)$

may be constant, is defined as follows

$$\text{AIC} = n\log(\hat{\sigma}) - \log(|\hat{A}|) + \frac{1}{2\hat{\sigma}^2}(\hat{A}Y - \hat{\mathbf{m}})^{\mathrm{T}}(\hat{A}Y - \hat{\mathbf{m}}) + \mathcal{K}, \quad (8.2)$$

where $\hat{A}$ and $\hat{\mathbf{m}}$ are A and $\mathbf{m}$ with the unknown parameters and functions being replaced by their estimators, $\mathcal{K}$ is the number of unknown parameters in model (1.3). The BIC can be defined in a similar way.

Because there are unknown functions in model (1.3), the first hurdle in the calculation of AIC of model (1.3) is to find how many unknown constants an unknown bivariate function amounts to. In the following, based on the residual sum of squares of standard bivariate nonparametric regression model, we propose an ad hoc way to solve this problem.

Suppose we have the following standard bivariate nonparametric regression model,

$$\eta_i = g(s_i) + e_i, \quad i = 1, \cdots, n, \quad (8.3)$$

where $E(e_i) = 0$ and $\text{var}(e_i) = \sigma_e^2$. The residual sum of squares of (8.3) is

$$\text{RSS} = \sum_{i=1}^{n} \{\eta_i - \hat{g}(s_i)\}^2$$

where $\hat{g}(\cdot)$ is the local linear estimator of $g(\cdot)$. On the other hand,

$$E(\text{RSS}/\sigma_e^2) = n - \text{ the number of unknown parameters in the regression function}$$

So, the number $\mathcal{T}$ of unknown constants the unknown function $g(\cdot)$

amounts to can be reasonably viewed as

$$\mathcal{T} = n - E(\mathrm{RSS}/\sigma_e^2) = n - \sigma_e^{-2} E\left[\sum_{i=1}^{n} \{\eta_i - \hat{g}(s_i)\}^2\right].$$

To make $\mathcal{T}$ more convenient to use, we derive the asymptotic form of $\mathcal{T}$. Let

$$\mathbf{S}_i = \begin{pmatrix} 1 & s_1^{\mathrm{T}} - s_i^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & s_n^{\mathrm{T}} - s_i^{\mathrm{T}} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix},$$

and

$$\mathcal{W}_i = \mathrm{diag}\left(K_h(u_1 - u_i)K_h(v_1 - v_i),\ \cdots,\ K_h(u_n - u_i)K_h(v_n - v_i)\right),$$

we have

$$\hat{g}(s_i) = (1,\ 0,\ 0)\left(\mathbf{S}_i^{\mathrm{T}} \mathcal{W}_i \mathbf{S}_i\right)^{-1} \mathbf{S}_i^{\mathrm{T}} \mathcal{W}_i \boldsymbol{\eta}$$

By the standard argument in Fan and Gijbels (1996) and the Lemma 1 in Fan and Zhang (1999), we have

$$\mathcal{T} = \left(2K^2(0) - \nu_0^2\right) h^{-2} + o(h^{-2})$$

when $h = o(n^{-1/6})$ and $nh^2 \longrightarrow \infty$, where $\nu_0 = \int K^2(t)dt$.

We conclude that an unknown bivariate function amounts to $(2K^2(0) - \nu_0^2)\, h^{-2}$ unknown constants. Based on this conclusion, if the number of constant components in $\boldsymbol{\beta}(\cdot)$ is q, the $\mathcal{K}$ in (8.2) will be $q + (p-q)\,(2K^2(0) - \nu_0^2)\, h^{-2}$.

To identify the constant components in $\boldsymbol{\beta}(\cdot)$ in (1.3) is basically a model selection problem. Theoretically speaking, we go for the model with the smallest AIC (or BIC). However, in practice, it is almost computationally impossible to compute the AICs for all possible models. We have to use some algorithm to reduce the computational burden. In the following, we are going to introduce two algorithms for the model selection.

8.3.2 Computational algorithms

In this chapter, we use AIC as an example to demonstrate the introduced algorithms. The model in which $\boldsymbol{\beta}(\cdot)$ has its i_1th, i_2th, $\cdots$, and i_kth components being constant is denoted by $\{i_1, \ \cdots, \ i_k\}$.

Backward elimination

The first algorithm we introduce is the backward elimination. Details are as follows.

(1) We start with the full model, $\{1, \ \cdots, \ p\}$, and compute its AIC by (8.2). Denote the full model by $\mathcal{M}_p$, its AIC by AIC_p.

(2) For any integer k, suppose the current model is $\mathcal{M}_k = \{i_1, \ \cdots, \ i_k\}$ with AIC given by AIC_k. Take $\mathcal{M}_{k-1}$ to be the model with the largest maximum of log likelihood function among the models $\{i_1, \ \cdots, \ i_{j-1}, \ i_{j+1}, \ \cdots, \ i_k\}$, $j = 1, \ \cdots, \ k$. If $\text{AIC}_k < \text{AIC}_{k-1}$, the chosen model is $\mathcal{M}_k$, and the model selection is ended; otherwise, continue to compute $\mathcal{M}_l$ and AIC_l until either $\text{AIC}_l < \text{AIC}_{l-1}$ or $l = 0$.

Curvature-to-Average ratio (CTAR) based method

A more aggressive way to reduce the computational burden involved in the model selection procedure is based on the ratio of the curvature of the estimated function to its average. Explicitly, we first treat all $\beta_j(\cdot)$, $j = 1, \cdots, p$, as functional. For each j, $j = 1, \cdots, p$, we compute the curvature-to-average ratio (CTAR) R_j of the estimated function $\hat{\beta}_j(\cdot)$:

$$R_j = \frac{1}{\bar{\beta}_j^2} \sum_{i=1}^{n} \left\{ \hat{\beta}_j(s_i) - \bar{\beta}_j \right\}^2, \quad \bar{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_j(s_i), \quad j = 1, \cdots, p.$$

We sort R_j, $j = 1, \cdots, p$, in an increasing order, say $R_{i_1} \leq \cdots \leq R_{i_p}$, then compute the AICs for the models $\{i_1, \cdots, i_k\}$ from $k = 0$ to the turning point k_0 where the AIC starts to increase. The chosen model is $\{i_1, \cdots, i_{k_0}\}$.

The algorithm based on the CTAR is much faster than the backward elimination based algorithm, however, from simulations, we find it less accurate although it still works reasonably well.

9 Performance of the Estimation Procedure

9.1 Different sample sizes

In this chapter, we will use simulated examples to examine the performances of the proposed estimation. In all simulated examples and the real data analysis later on, we set w_{ij} to be

$$w_{ij} = \exp(-\|s_i - s_j\|)/\sum_{k \neq i} \exp(-\|s_i - s_k\|), \quad \|s_i\| = (s_i^{\mathrm{T}} s_i)^{1/2}. \tag{8.1}$$

We first examine the performance of the proposed estimation procedure.

Example 2. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2 \pi), \quad \beta_2(s) = \cos(\|s\|^2 \pi),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We are going to apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, α and σ^2, and examine the accuracy of the proposed estimation procedure.

We use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ as the kernel function in the estimation procedure. The bandwidth used in the estimation is 0.45.

We use mean squared error (MSE) to assess the accuracy of an estimator of an unknown constant parameter, mean integrated squared error (MISE) to assess the accuracy of an estimator of an unknown function.

For each given sample size n, we do 200 simulations. We compute the MSEs of the estimators of the unknown constants and the MISEs of the estimators of the unknown functions for sample sizes $n = 400$, $n = 500$ and $n = 600$. The obtained results are presented in Table 3. Table 3 shows the proposed estimation procedure works very well. For a more visible illustration of the performance of the proposed estimation procedure, we set sample size $n = 500$ and do 200 simulations. We single out the one with median performance among the 200 simulations. The estimate of α coming from this simulation is 0.42, the estimate of σ^2 is 0.95. The estimated unknown functions from this simulation are presented in Figures 2 and 3, and are superimposed with the true functions. All these show our estimation procedure works very well.

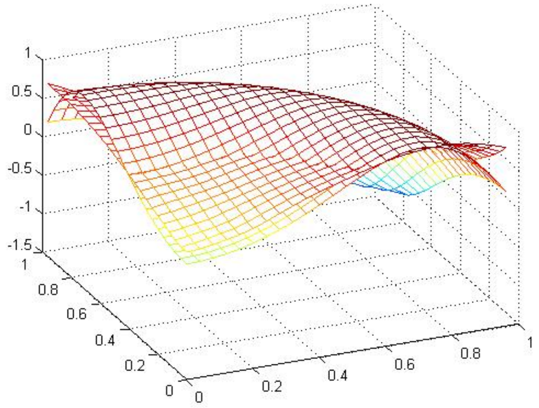

Figure 2: $\beta_1(s) = \sin(\|s\|^2\pi)$

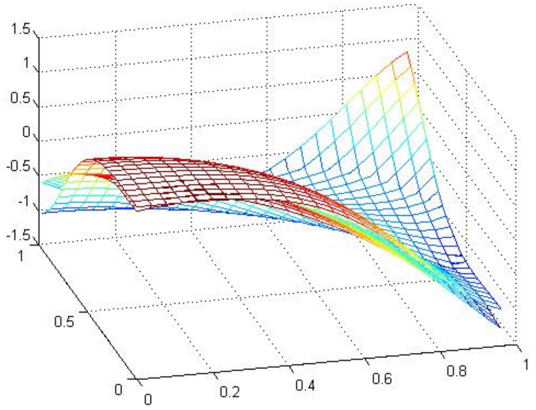

Figure 3: $\beta_2(s) = \cos(\|s\|^2\pi)$

Table 3: **Example2 :The MISEs and MSEs for different sample sizes**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\alpha}$	$\hat{\sigma}^2$
n=400	0.0512	0.0480	0.00788	0.0099
n=500	0.0432	0.0382	0.00429	0.0070
n=600	0.0380	0.0345	0.00325	0.0046

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 400$, $n = 500$ and $n = 600$, corresponding to the estimator of an unknown constant is the MSEs of the estimator.

Example 3. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, ϵ_i

from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We are going to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, α and σ^2, and examine the accuracy of the proposed estimation procedure.

We still use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ as the kernel function in the estimation procedure. The bandwidth used in the estimation is 0.31.

We use mean squared error (MSE) to assess the accuracy of an estimator of an unknown constant parameter, mean integrated squared error (MISE) to assess the accuracy of an estimator of an unknown function.

For each given sample size n, we do 200 simulations. We compute the MSEs of the estimators of the unknown constants and the MISEs of the estimators of the unknown functions for sample size $n = 500$, $n = 600$ and $n = 700$. The obtained results are presented in Table 4. Table 4 shows the proposed estimation procedure works very well. For a more visible illustration of the performance of the proposed estimation procedure, we set sample size $n = 500$ and do 200 simulations. We single out the one with median performance among the 200 simulations. The estimate of α coming from this simulation is 0.47, the estimate of σ^2 is 0.98. The estimated unknown functions from this simulation are presented in Figures 4 and 5, and are superimposed with the true functions. Both figures show our estimation procedure works very well.

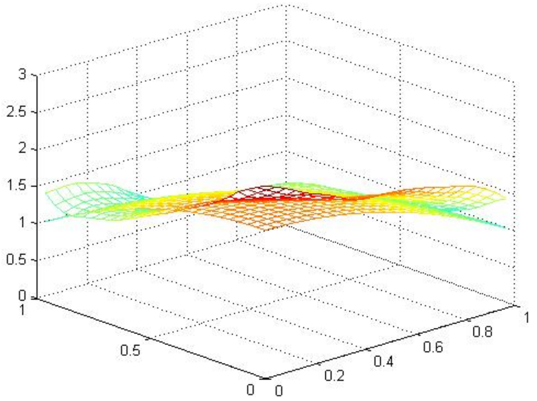

Figure 4: $\beta_1(s) = 2 - (\|s\|^2)$

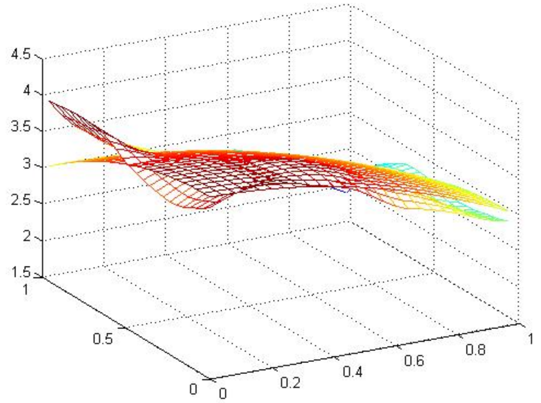

Figure 5: $\beta_2(s) = 4 - (\|s\|^2)$

9.2 Different bandwidths

A natural question here is how wide the local neighborhood should be in estimation procedure. Likewise, it is essential to know how to select the bandwidth so that our approximation will be valid. There is a trade-off between bias and variance during estimation. If we select

Table 4: **Example 3 :The MISEs and MSEs for different sample sizes**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\alpha}$	$\hat{\sigma}^2$
n=500	0.03159	0.03032	0.06892	0.0079
n=600	0.02432	0.02382	0.03791	0.0042
n=700	0.01862	0.01845	0.01255	0.0026

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 500$, $n = 600$ and $n = 700$, corresponding to the estimator of an unknown constant is the MSEs of the estimator.

the bandwidth too large, the variance will be small, however, we would pay a price on bias part. If the bandwidth is too small, the variance of the estimated local parameters will be large.In this chapter, we estimate beta function based on different bandwidths, and examine how bandwidths affect our estimation. Due to the computational limit, we only estimate the beta functions when sample size n equals 500.

Example 4. We set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi), \quad \beta_2(s) = \cos(\|s\|^2\pi),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, and ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We are going to apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, and examine the accuracy of the proposed estimation procedure.We use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ as the kernel function in the estimation procedure. We estimated the unknown functions beta based on different bandwidths. The obtained results are presented in Table 5. We can tell the minimum MISE for unknown functions are both achieved at the bandwidth $h = 0.4$.

We use MISE to assess the accuracy of an estimator of an unknown function.

For each given bandwidth h, we do 200 simulations. We compute the MISEs of the estimators of the unknown functions for sample size $n = 500$. The ways in which the selection of bandwidth affected the estimation accuracy, are illustrated are in Figures 6 and 7. Here, $\hat{\beta}(\cdot) = \hat{\beta}_1(\cdot) + \hat{\beta}_2(\cdot)$

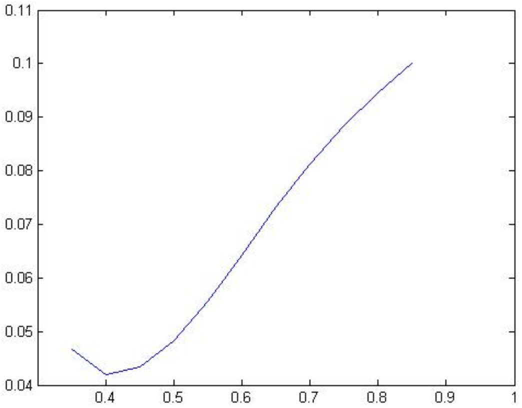

Figure 6: MISE for $\beta_1(s) = \sin(\|s\|^2\pi)$ based on different bandwidth

Example 5. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

Table 5: **Example 4: The MISEs for $\beta(\cdot)$ with different bandwidths**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\beta}(\cdot)$
0.35	0.04667	0.04279	0.08946
0.4	0.04194	0.03785	0.07980
0.45	0.04329	0.03813	0.08142
0.5	0.04838	0.04106	0.08942
0.55	0.05574	0.04544	0.10119
0.6	0.06432	0.05063	0.11496
0.65	0.07321	0.05602	0.12923
0.7	0.08134	0.06160	0.14294
0.75	0.08848	0.06774	0.15622
0.8	0.09462	0.07497	0.16960
0.85	0.10005	0.08365	0.18371

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 500$ based on different bandwidths

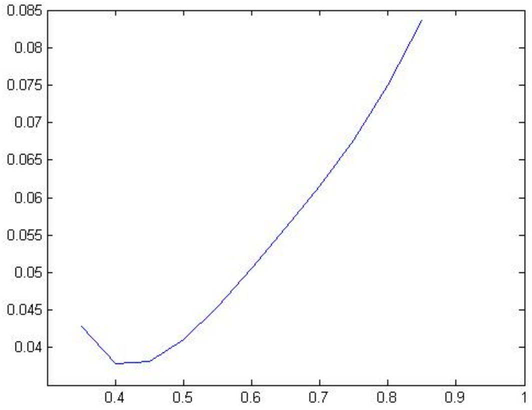

Figure 7: MISE for $\beta_2(s) = \cos(\|s\|^2\pi)$ based on different bandwidth

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, and ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We estimated the unknown functions beta based on different bandwidths. The results are presented in Table 6.

We use MISE to assess the accuracy of an estimator of an unknown function.

For each given bandwidth h, we do 200 simulations. We compute the MISEs of the estimators of the unknown functions for sample size $n = 500$. To illustrate how the selection of bandwidth affects the estimation accuracy, we graph the results in Figures 8 and 9.

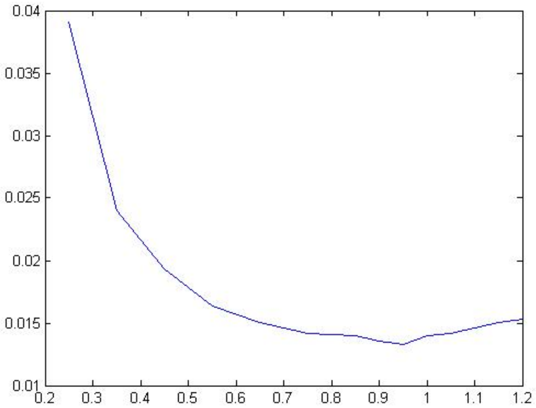

Figure 8: MISE for $\beta_1(s) = 2 - (\|s\|^2)$ based on different bandwidth

9.3 Performance of estimation procedure when alpha is known

In the previous chapter, we use simulated examples to examine the performances of the proposed estimation, the unknown functions and

Table 6: **Example 5: The MISEs for $\beta(\cdot)$ with different bandwidths**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\beta}(\cdot)$
0.25	0.03905	0.03713	0.07618
0.35	0.02396	0.02347	0.04743
0.45	0.01935	0.01879	0.03814
0.55	0.01638	0.01613	0.03251
0.65	0.01502	0.01437	0.02939
0.75	0.01413	0.01367	0.02780
0.85	0.01396	0.01302	0.02698
0.9	0.01356	0.01289	0.02645
0.95	0.01332	0.01276	0.02608
1.0	0.01395	0.01289	0.02684
1.05	0.01417	0.01325	0.02742
1.10	0.01463	0.01396	0.02859
1.15	0.01502	0.01412	0.02914
1.20	0.01535	0.01457	0.02992

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 500$ based on different bandwidth

the unknown parameters. Our estimation procedure is profile likelihood. We pretend that the beta function is known, and we get the estimator of α using grid method and the estimator of σ^2 . Then, we find the estimator of the unknown functions. A very obvious question would be how the estimation procedure works when we know the value of α. In this chapter, we estimate the unknown functions under the condition when the value of α is known.

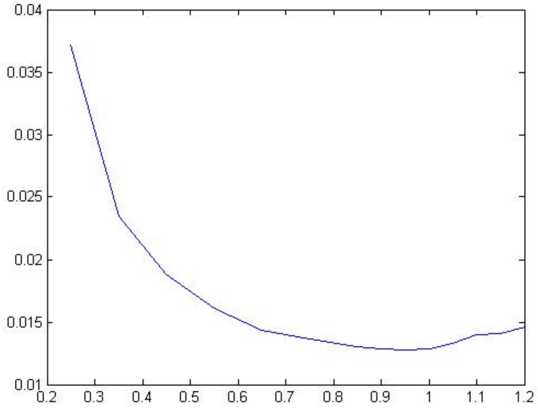

Figure 9: MISE for $\beta_2(s) = 4 - (\|s\|^2)$ based on different bandwidth

9.3.1 Different sample sizes

Example 6. In model (1.3), we set $p = 1$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi),$$

and independently generate X_i from $N(\mathbf{0},\ I)$, s_i from $U[0,\ 1]^2$, and ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We are going to apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, and examine the accuracy of the proposed estimation procedure.The bandwidth used in the estimation is 0.25.

We use MISE to assess the accuracy of an estimator of an unknown function.

For each given sample size n, we do 200 simulations. We compute MISEs of the estimators of the unknown functions for sample sizes

$n = 400$, $n = 500$ and $n = 600$. The results are presented in Table 7, which shows the proposed estimation procedure still works very well.

Table 7: **Example 6: The MISEs for $\beta(\cdot)$ with α known**

	$\hat{\beta}_1(\cdot)$
n=400	0.05319
n=500	0.04179
n=600	0.03578

The column corresponding to the estimator of an unknown function is the MISEs of the estimators for $n = 400$, $n = 500$ and $n = 600$.

Example 7. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi), \quad \beta_2(s) = \cos(\|s\|^2\pi),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, and ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, and examine the accuracy of the proposed estimation procedure.

Other settings are the same. The bandwidth used in the estimation is 0.45. We use MISE to assess the accuracy of an estimator of an unknown function.

For each given sample size n, we do 200 simulations. We compute MISEs of the estimators of the unknown functions for sample sizes $n = 400$, $n = 500$ and $n = 600$. The results are presented in Table 8, which are quite satisfying.

Table 8: **Example 7: The MISEs for $\beta(\cdot)$ with α known**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$
n=400	0.05086	0.04746
n=500	0.04311	0.03802
n=600	0.03792	0.03419

The column corresponding to the estimator of an unknown function is the MISEs of the estimators for $n = 400$, $n = 500$ and $n = 600$.

Example 8. In model (1.3), we set $p = 1$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2),$$

and independently generate X_i from $N(\mathbf{0},\ I)$, s_i from $U[0,\ 1]^2$, ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We estimate $\beta_1(\cdot)$, and examine the accuracy of the proposed estimation procedure. The bandwidth used in the estimation is 0.15.

We use MISE to assess the accuracy of an estimator of an unknown function. The results are presented in Table 9.

Table 9: **Example 8: The MISEs for $\beta(\cdot)$ with α known**

	$\hat{\beta}_1(\cdot)$
n=500	0.03019
n=600	0.02379
n=700	0.01773

The column corresponding to the estimator of an unknown function is the MISEs of the estimators for $n = 500$, $n = 600$ and $n = 700$

Example 9. We set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, and examine the accuracy of the proposed estimation procedure.We use the Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ as the kernel function in the estimation procedure, and estimate the unknown functions beta based bandwidth $h = 0.31$. MISE is used to assess the accuracy of an estimator of an unknown function.

For each given sample size n, we do 200 simulations. We compute the MISEs of the estimators of the unknown functions for sample size $n = 500$, $n = 600$ and $n = 700$. The results are presented in Table 10.

Table 10: **Example 9: The MISEs for $\beta(\cdot)$ with α known**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$
n=500	0.03059	0.02973
n=600	0.02395	0.02278
n=700	0.01793	0.01824

The column corresponding to the estimator of an unknown function is the MISEs of the estimators for $n = 500$, $n = 600$ and $n = 700$.

9.3.2 Different bandwidths

In this chapter, we estimate the unknown functions based on different bandwidth. To compare the results from the previous sections, we estimate beta using the same bandwidths we used before. Due to the same reasons given before, namely the computational limitations, we only calculate the MISEs when the sample size n equals 500.

Example 10. We set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi), \quad \beta_2(s) = \cos(\|s\|^2\pi),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, and examine the accuracy of the proposed estimation procedure. We estimated the unknown functions beta based on different bandwidths. The results are presented in Table 11.

We use MISE to assess the accuracy of an estimator of an unknown function.

For each given bandwidth h, we do 200 simulations. We compute the MISEs of the estimators of the unknown functions for sample size $n = 500$. The ways in which the selection of bandwidths affects the estimation accuracy, are presented in Figures 10 and 11.

Table 11: **Example 10: The MISEs for $\beta(\cdot)$ with α known under different bandwidths**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\beta}(\cdot)$
0.35	0.04664	0.04278	0.08943
0.4	0.04192	0.03765	0.07958
0.45	0.04311	0.03802	0.08114
0.5	0.04792	0.04104	0.08897
0.55	0.05573	0.04544	0.10116
0.6	0.06432	0.05062	0.11493
0.65	0.07320	0.05601	0.12922
0.7	0.08133	0.06159	0.14292
0.75	0.08842	0.06772	0.15614
0.8	0.09462	0.07466	0.16927
0.85	0.10004	0.08362	0.18365

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 500$ based on different bandwidths

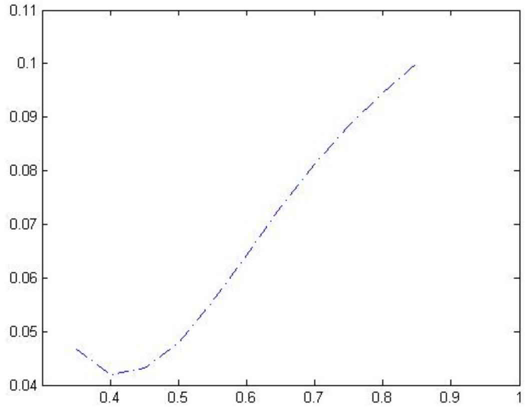

Figure 10: MISE for $\beta_1(s) = \sin(\|s\|^2\pi)$ based on different bandwidths with α known

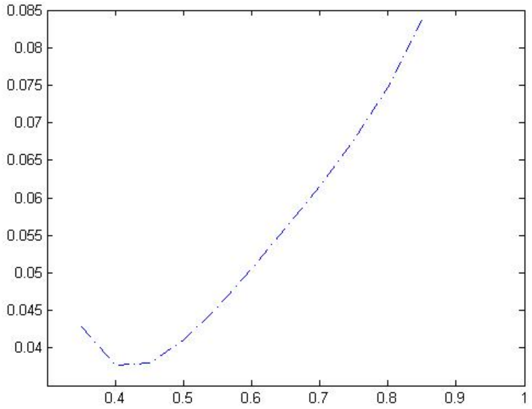

Figure 11: MISE for $\beta_2(s) = \cos(\|s\|^2\pi)$ based on different bandwidths with α known

Example 11. We set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

and independently generate X_i from $N(\mathbf{0}_2,\ I_2)$, s_i from $U[0,\ 1]^2$, ϵ_i from $N(0,\ \sigma^2)$, $i = 1,\ \cdots,\ n$. y_i, $i = 1,\ \cdots,\ n$, are generated through model (1.3). We are going to apply the proposed estimation method based on the generated $(s_i,\ X_i^{\mathrm{T}},\ y_i)$, $i = 1,\ \cdots,\ n$, to estimate $\beta_1(\cdot)$, $\beta_2(\cdot)$, and examine the accuracy of the proposed estimation procedure.We use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ as the kernel function in the estimation procedure. We estimated the unknown functions beta based on different bandwidth. The obtained results are presented in Table 12.

We use the mean integrated squared error (MISE) to assess the accuracy of an estimator of an unknown function.For each given bandwidth h, we do 200 simulations. We compute the MISEs of the estimators of the unknown functions for sample size $n = 500$. To have a more visible idea, we draw graphs about how the selection of bandwidth affected the estimation accuracy. They are presented in Figures 12 and 13.

Table 12: **Example 11: The MISEs for $\beta(\cdot)$ with α known under different bandwidths**

	$\hat{\beta}_1(\cdot)$	$\hat{\beta}_2(\cdot)$	$\hat{\beta}(\cdot)$
0.25	0.03814	0.03623	0.07438
0.35	0.02243	0.02232	0.04475
0.45	0.01723	0.01715	0.03439
0.55	0.01522	0.01502	0.03025
0.65	0.01444	0.01397	0.02843
0.75	0.01386	0.01325	0.02712
0.85	0.01337	0.01271	0.02608
0.90	0.01324	0.01257	0.02582
0.95	0.01326	0.01258	0.02585
1.00	0.01338	0.01275	0.02613
1.05	0.01356	0.01301	0.02657
1.10	0.01379	0.01332	0.02711
1.15	0.01404	0.01366	0.02771
1.20	0.01432	0.01426	0.02858

The column corresponding to the estimator of an unknown function is the MISEs of the estimator for $n = 500$ based on different bandwidths

9.4 Oracle property of the estimation

In the previous chapter, we estimate the unknown functions under the condition that either α is known or α is unknown, and find that the

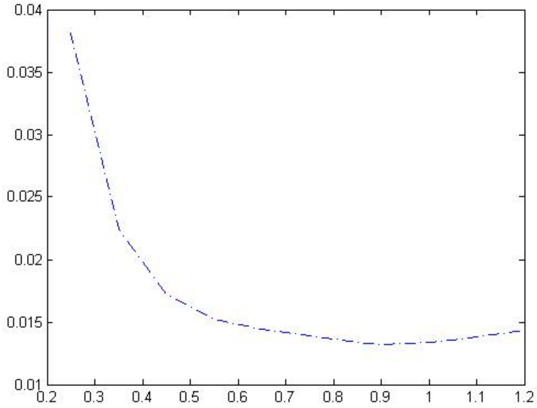

Figure 12: MISE for $\beta_1(s) = 2 - (\|s\|^2)$ based on different bandwidths with α known

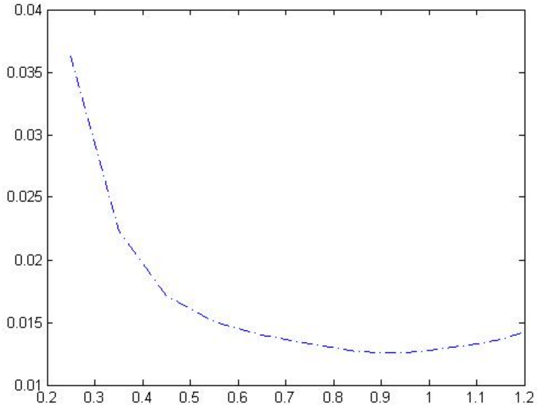

Figure 13: MISE for $\beta_2(s) = 4 - (\|s\|^2)$ based on different bandwidths with α known

mean integrated squared errors for the unknown functions are very similar for the two situations. Clearly, we get more accurate results when α is known. However, there is not a big difference. All of the above results shows that our estimator has an Oracle property.This is understandable from asymptotic point of view, because the convergence rate of an estimator of unknown constant is of order $n^{-1/2}$. It is much faster than the convergence rate of an estimator of unknown function which is of order $(\sqrt{n}h)^{-1}$. So the estimation for the unknown functions under the condition ,whether α is unknown or known,would not differ from each other.

Example 12. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

We have estimated the unknown function $\beta_1(\cdot)$ and $\beta_2(\cdot)$ under condition α are unknown and known respectively under different bandwidths. We used the MISE to measure the accuracy of the estimation procedure.In Table 13, we compare the MISE for the unknown function $\beta_1(\cdot)$ under the condition α is known and unknown. We also use the graphs to obtain an invisible view. Figure 14 compares the MISEs under different bandwidths with α unknown and α known. The two lines are clearly close together, which proves that our estimator has an oracle property.

In Table 14, we compare the MISE for the unknown function $\beta_2(s) = 4 - (\|s\|^2)$ under the conditions α is known and α is unknown. Figure 15 and compares the MISEs under different bandwidths with α unknown and α known.

Example 13. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi), \quad \beta_2(s) = \cos(\|s\|^2\pi),$$

Table 13: **Example 12: The MISEs of $\hat{\beta}_1(\cdot)$ with α known and unknown using different bandwidths**

	α known	α unknown
0.25	0.03814	0.03905
0.25	0.02243	0.02396
0.25	0.01723	0.01935
0.25	0.01522	0.01638
0.65	0.01444	0.01502
0.75	0.01386	0.01413
0.85	0.01337	0.01396
0.9	0.01324	0.01356
0.95	0.01326	0.01332
1.0	0.01338	0.01395
1.05	0.01356	0.01417
1.10	0.01379	0.01463
1.15	0.01404	0.01502
1.20	0.01432	0.01535

The column corresponding to the estimator of an unknown function $\beta_1(s) = 2 - (\|s\|^2)$ is the MISEs of the estimator for $n = 500$ based on different bandwidths with α known and unknown.

In Table 15, we also compare the MISEs for the unknown function $\beta_1(\cdot) = \sin(\|s\|^2\pi)$ when α is known and unknown. Figures 16 and 17 provide an invisible view, comparing the MISEs under different bandwidths with α unknown and α known.

In Table 16, we compare the MISEs for the unknown function $\beta_2(s) = \cos(\|s\|^2\pi)$ when α is known and unknown. Figures 18 and 19 compare the MISEs under different bandwidths with α unknown and known.

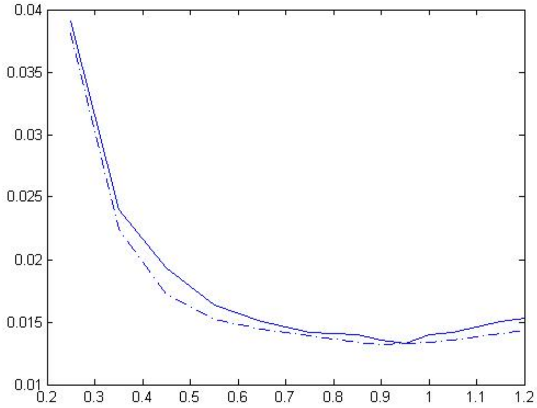

Figure 14: MISE for $\beta_1(s) = 2 - (\|s\|^2)$ based on different bandwidths

The two lines are the MISEs of the estimator for $n = 500$ based on different bandwidth with α known and unknown.The solid line is the MISEs with α unknown. The dashed line is the MISEs with α known. The MISEs with α known are smaller than the MISEs with α unknown. However, the two lines are very close.

All of the above examples proves our estimators of the unknown functions have an Oracle property.

9.5 Performance of estimation procedure for alpha

In the previous chapter, we estimate the unknown functions under the condition that the unknown parameter is known. In this chapter, we examine the performance of estimation procedure for the unknown parameter α when beta functions are known.The convergence rate of an estimator of unknown constant is of order $n^{-1/2}$. It is much faster than

Table 14: **Example 12: The MISEs of $\hat{\beta}_2(\cdot)$ with α known and unknown using different bandwidths**

	α known	α unknown
0.25	0.03623	0.03713
0.35	0.02232	0.02347
0.45	0.01715	0.01879
0.55	0.01502	0.01613
0.65	0.01397	0.01437
0.75	0.01325	0.01367
0.85	0.01271	0.01302
0.9	0.01257	0.01289
0.95	0.01258	0.01276
1.0	0.01275	0.01289
1.05	0.01301	0.01325
1.10	0.01332	0.01396
1.15	0.01366	0.01412
1.20	0.01426	0.01457

The column corresponding to the estimator of an unknown function $\beta_2(s) = 4 - (\|s\|^2)$ is the MISEs of the estimator for $n = 500$ based on different bandwidths with α known and unknown

the convergence rate of an estimator of unknown function which is of order $(\sqrt{n}h)^{-1}$. There is no significant difference between the situation that α is known and α is unknown in estimating. The situation totally changes when we estimate the unknown parameter under the condition that beta functions are known. The accuracy of estimation greatly increases. It is easy to understand, because α converges much faster than the unknown functions.We would use two examples to prove this.

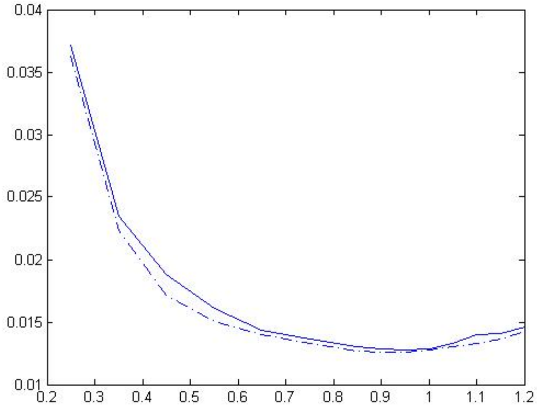

Figure 15: MISE for $\beta_2(s) = 4 - (\|s\|^2)$ based on different bandwidth

The two lines are the MISEs of the estimator for $n = 500$ based on different bandwidth with α known and unknown.The solid line is the MISEs with α unknown. The dashed line is the MISEs with α known.

Example 14. We set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.5, \quad \beta_1(s) = \sin(\|s\|^2\pi), \quad \beta_2(s) = \cos(\|s\|^2\pi),$$

We use mean squared error (MSE) to assess the accuracy of an estimator of an unknown constant parameter.For each given sample size n, we do 200 simulations. We compute the MSEs of the estimators of the unknown constants for sample sizes $n = 400$, $n = 500$ and $n = 600$. The obtained results are presented in Table 17, which shows the proposed estimation procedure works very well. The accuracy of estimating the unknown parameter α increases significantly.

Table 15: **Example 13: The MISEs of $\hat{\beta}_1(\cdot)$ with α known and unknown using different bandwidths**

	α known	α unknown
0.35	0.04664	0.04667
0.4	0.04192	0.04194
0.45	0.04311	0.04329
0.5	0.04792	0.04838
0.55	0.05573	0.05574
0.6	0.06432	0.06432
0.65	0.07320	0.07321
0.7	0.08133	0.08134
0.75	0.08842	0.08848
0.8	0.09462	0.09642
0.85	0.10004	0.10005

The column corresponding to the estimator of an unknown function $beta_1(\cdot) = \sin(\|s\|^2\pi)$ *is the MISEs of the estimator for* $n = 500$ *based on different bandwidths with* α *known and unknown*

Example 15. In model (1.3), we set $p = 2$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2),$$

The obtained results are presented in Table 18. Table 18 shows the proposed estimation procedure works very well. The accuracy of estimating the unknown parameter α does increase a lot.

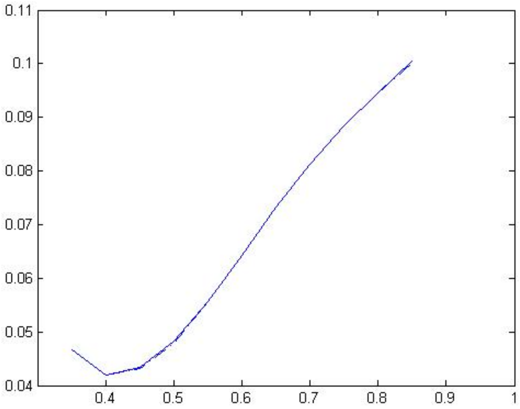

Figure 16: MISE for $\beta_1(\cdot) = \sin(\|s\|^2\pi)$ based on different bandwidths

The two lines are the MISEs of the estimator for $n = 500$ *based on different bandwidths with* α *known and unknown.*

10 Performance of the Model Selection Procedure

In this chapter, we use simulated examples to examine the performances of the model selection methods previously derived . We use the ratio of picking the right model to measure the performance of our model selection methods, and find that all the model selection methods perform quite well in simulation studies.

10.1 Thresholding K simulation

Example 16. In model (1.3), we set $p = 3$, $\beta_1(\cdot) = \sin(\|s\|^2\pi)$ and $\beta_2(\cdot) = \cos(\|s\|^2\pi)$, $\beta_3(\cdot) = \beta_3 = 1$. We generate X_i, s_i, ϵ_i,

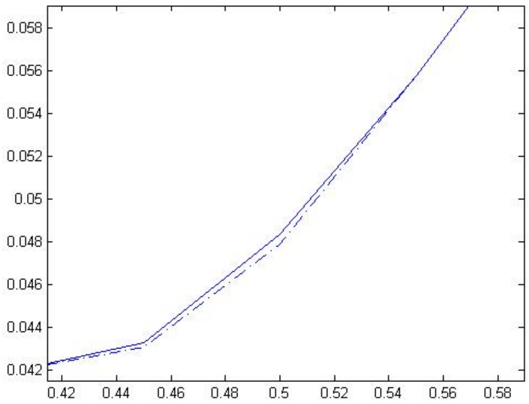

Figure 17: MISE for $\beta_1(\cdot) = \sin(\|s\|^2\pi)$ based on different bandwidths

We zoom Figure 17 in the bandwidth range [0.35, 0.55] to have a better view.The solid line are the MISEs of the estimator for $n = 500$ *based on different bandwidth with* α *unknown . The dashed line are the MISEs of the estimator for* $n = 500$ *based on different bandwidths with* α *known*

y_i $i = 1, \cdots, n$, in the same way as before, except that X_i is from $N(\mathbf{0}_3, I_3)$. Based on the generated data, we are going to apply the proposed thresholding K method to select the correct model, and examine the performances of the proposed method in identifying the constant components in model (1.3).

We still use the Epanechnikov kernel as the kernel function in the model selection.The bandwidth we used for estimation is $h = 0.45$. We set the sample size to equal 500.We first calculate the discrepancy of each component p.We repeat the procedure 200 times, and we take their average as our starting point. Table 19 represents the discrepancy of each component.

Table 16: **Example 13:The MISEs of $\hat{\beta}_2(\cdot)$ with α known and unknown using different bandwidths**

	α known	α unknown
0.35	0.04278	0.04279
0.4	0.03765	0.03785
0.45	0.03802	0.03813
0.5	0.04104	0.04106
0.55	0.04544	0.04544
0.6	0.05062	0.05063
0.65	0.05601	0.05602
0.7	0.06159	0.06160
0.75	0.06772	0.06774
0.8	0.07466	0.07497
0.85	0.08362	0.08365

The column corresponding to the estimator of an unknown function $\beta_2(s) = \cos(\|s\|^2\pi)$ is the MISEs of the estimator for $n = 500$ based on different bandwidth with α known and unknown

Table 17: **Example14 : The MSEs for α with $\beta(\cdot)$ known**

	$\hat{\alpha}$
n=400	0.004075
n=500	0.001156
n=600	0.000789

The column corresponding to the estimator of an unknown parameter is the MSEs of the estimators for $n = 400$, $n = 500$ and $n = 600$.

We use 10.8760 as the initial value K_1. Then,we increase our thresholding value K_i 30% each time until it reaches our maximum

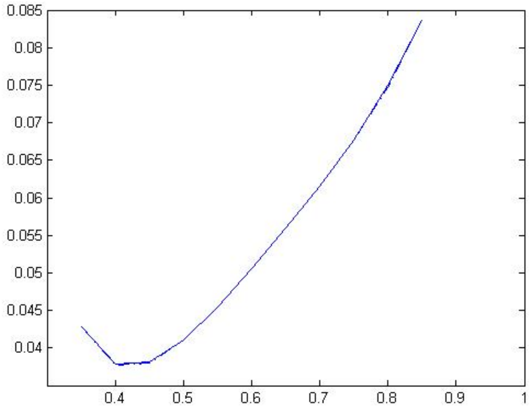

Figure 18: MISE for $\beta_2(s) = \cos(\|s\|^2\pi)$ based on different bandwidth

The two lines are the MISEs of the estimator for $n = 500$ based on different bandwidth with α known and unknown. They are vey close to each other

Table 18: **Example15 :The MSEs for α with $\beta(\cdot)$ known**

	$\hat{\alpha}$
n=500	0.001045
n=600	0.000566
n=700	0.000218

The column corresponding to the estimator of an unknown parameter is the MSEs of the estimators for $n = 500$, $n = 600$ and $n = 700$.

value 197.3109. We repeat the simulation 200 times and then calculate the MISE for this 12 thresholding value. Table 20 represents the MISE values .To illustrate, Figure 20 shows the MISE values for each thresholding K.We could find the thresholding K values with minimum

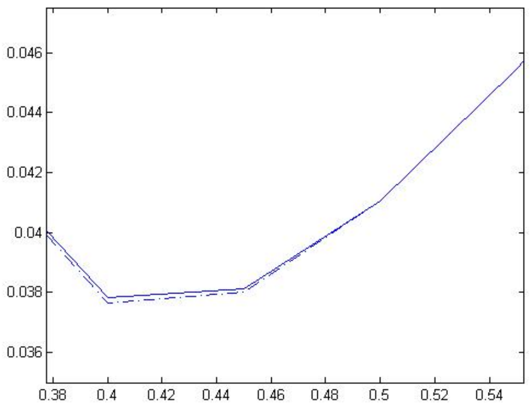

Figure 19: MISE for $\beta_2(s) = \cos(\|s\|^2\pi)$ based on different bandwidth

We zoom Firgure 19 in the bandwidth range [0.35, 0.55] to have a better view.The solid line are the MISEs of the estimator for $n = 500$ *based on different bandwidth with* α *unknown . The dashed line are the MISEs of the estimator for* $n = 500$ *based on different bandwidth with* α *known*

Table 19: **Example 16: The Discrepancies for different component**

	Discrepency
$\beta_1(\cdot)$	133.9557
$\beta_2(\cdot)$	197.3109
$\beta_3(\cdot)$	10.8760

The column corresponding to discrepancy of each component.The sample size $n = 500$

MISE, which would be our optimal thresholding value K_0.

We also find the ratios for each thresholding K_i. We calculate three

Table 20: **Example 16 :The MISEs of different Thresholding values**

thresholding value	MISEs
K=10.876	0.1162
K=14.1388	0.1102
K=18.3804	0.1032
K=23.8945	0.1008
K=31.0629	0.0991
K=40.3818	0.0984
K=52.4963	0.0975
K=68.2452	0.0975
K=88.7188	0.1057
K=115.2245	0.1624
K=149.9348	0.2777
K=194.9153	0.5457

The column corresponding to MISEs of each thresholding value K_i.The sample size $n = 500$. We repeat the simulation 200 times

ratios:The ratio of picking the right model, Right model; the ratio of treating constant component as function, constant as function; the ratio of treating function as constant, function as constant .Table 21 presents the results.

The thresholding value $K = 52.4963$ and $K = 68.2452$ achieved the minimum MISEs and the ratios of picking the right model are all 1 .This would be our optimal thresholding K value range, $[52.4963, 68.2452]$. Previously, we ran 200 times simulation. Theoretically, we should conduct 1000 simulations for at least 2 thresholding values within the optimal range. Due to the computational limitations, we only conduct 1000 simulations for thresholding value $K = 52.4963$. We calculate

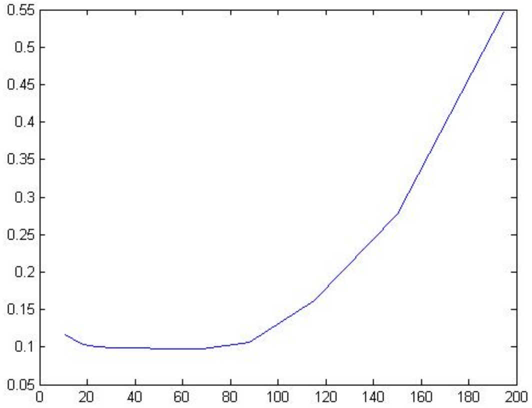

Figure 20: MISEs based on different thresholding value

The sample size is $n = 500$. *In the interval [52.4963, 68.2452] , we get the minimum MISEs. This would be our optimal thresholding K range.*

the ratios of picking the model, and the result is listed in Table 22.

Example 17. In model (1.3), we set $p = 3$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2), \beta_3(\cdot) = \beta_3 = 1$$

We generate $X_i,\ s_i,\ \epsilon_i,\ y_i\ i = 1,\ \cdots,\ n$, in the same way as before, except that X_i is from $N(\mathbf{0}_3,\ I_3)$.

We use the Epanechnikov kernel as the kernel function in the model selection, and the bandwidth used to estimate is $h = 0.31$. The sample size is set at 500.We first calculate the discrepancy of each component p. We repeat the procedure 200 times, and take their average as our starting point. Table 23 represents the discrepancy of each component.

Table 21: **Example 16 :The Ratios of picking the model under different thresholding K values**

thresholding value	Right model	constant as function	function as constant
K=10.876	0.43	0.57	0
K=14.1388	0.67	0.23	0
K=18.3804	0.89	0.11	0
K=23.8945	0.95	0.05	0
K=31.0629	0.98	0.02	0
K=40.3818	0.99	0.01	0
K=52.4963	1	0	0
K=68.2452	1	0	0
K=88.7188	0.99	0	0.01
K=115.2245	0.77	0	0.23
K=149.9348	0.42	0	0.58
K=194.9153	0.26	0	0.74

The ratios of picking model of each thresholding value K_i.The sample size $n = 500$. We repeat the simulation 200 times

Table 22: **Example 16: The Ratios of Optimal Thresholding Value K**

thresholding value	Right model	constant as function	function as constant
K=52.4963	0.996	0.003	0.001

The ratios of picking model of optimal thresholding value $K_0 = 52.4963$.The sample size $n = 500$. We repeat the simulation 1000 times

We use 2.8838 as the initial value K_1. Then we increase our thresh-

Table 23: **Example 17: The Discrepancies for different components**

	Discrepency
$\beta_1(\cdot)$	88.5951
$\beta_2(\cdot)$	88.4794
$\beta_3(\cdot)$	2.8838

The column corresponding to discrepancy of each component.The sample size $n = 500$

olding value K_i 30% each time until it reaches our maximum value 88.5951. We repeat the simulation 200 times and then calculate the MISE for the following 14 thresholding values. Table 24 represents these MISEs values . Figure 20 shows the MISE values for each thresholding K.We find the thresholding K values with minimum MISE. And this would be our optimal Thresholding value K_0.

We also find the ratios for each thresholding K_i. We calculate three ratios. The ratio of picking the right model, Right model; the ratio of treating constant component as function, constant as function; the ratio of treating function as constant, function as constant. Table 25 represents the results.

The thresholding values $K = 13.9195$,$K = 18.0954$,and $K = 23.5240$ all achieved the minimum MISEs and the ratios of picking the right model are all 1 .This would be our optimal thresholding K value range, $[13.9195, 23.5240]$. Previously, we performed 200 simulations. Theoretically, we should conduct 1000 simulations for at least 2 thresholding values within the optimal range. Due to the compu-

Table 24: **Example 17 :The MISEs of different Thresholding values**

thresholding value	MISEs
K=2.8838	0.036102
K=3.7489	0.036087
K=4.8736	0.035924
K=6.3367	0.034897
K=8.2364	0.034132
K=10.7073	0.033937
K=13.9195	0.033827
K=18.0954	0.033827
K=23.5240	0.033827
K=30.5812	0.035678
K=39.7556	0.037982
K=51.6823	0.043212
K=67.1871	0.059798
K=87.3431	0.078125

The column corresponding to MISEs of each thresholding value K_i.The sample size $n = 500$. We repeat the simulation 200 times

tational limitation, we only conduct 1000 simulations for thresholding value $K = 18.0954$. We calculate the ratios of picking the model, and the result is listed in Table 26.

The above examples, indicate that the MISEs of detecting the constant component as a functional component is approximately 10 times larger than the MISE of detecting the functional component as a constant component. As we explained , the error order of treating the functional component as constant is $O(1)$. In contrast, the error order of treating constant as function is $O(\frac{1}{\sqrt{nh}})$. As a result, we should be

Table 25: **Example 17 :The Ratios of picking the model under different thresholding K values**

thresholding value	Right model	constant as function	function as constant
K=2.8838	0.62	0.38	0
K=3.7489	0.67	0.33	0
K=4.8736	0.71	0.29	0
K=6.3367	0.74	0.26	0
K=8.2364	0.81	0.19	0
K=10.7073	0.94	0.06	0
K=13.9195	1	0	0
K=18.0954	1	0	0
K=23.5240	1	0	0
K=30.5812	0.92	0	0.08
K=39.7556	0.85	0	0.15
K=51.6823	0.77	0	0.23
K=67.1871	0.69	0	0.31
K=87.3431	0.54	0	0.46

The ratios of picking model of each thresholding value K_i.The sample size $n = 500$. We repeat the simulation for 200 times

Table 26: **Example 17: The Ratios of Optimal Thresholding Value K**

thresholding value	Right model	constant as function	function as constant
K=18.0954	0.997	0.001	0.002

The ratios of picking model of optimal thresholding value $K_0 = 18.0954$.The sample size $n = 500$. We repeat the simulation for 1000 times

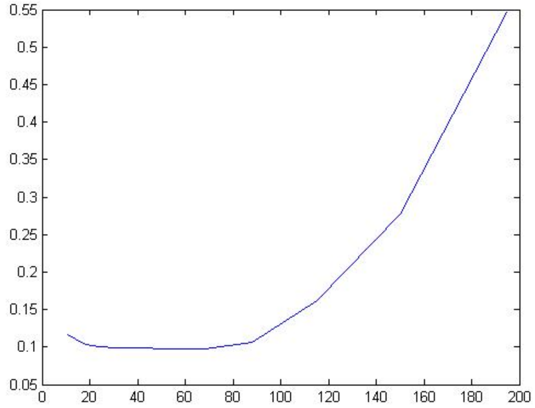

Figure 21: MISEs based on different thresholding value

The sample size is $n = 500$*. In the interval [13.9195, 23.5240] , we get the minimum MISEs. This would be our optimal thresholding K range.*

very careful when we select the Thresholding K. If the thresholding chosen is too large, then the ratio of picking the constant as function will increase. However, if it is too small, the ratio of picking the function as constant will increase.The second mistake is much worse than the first, and within the optimal range, the thresholding method will provide fantastic results for selecting models. We have not discovered how to get the estimator of MISEs as we discussed before. We can not deny the existence of the optimal thresholding range.Perhaps in the future, we could construct the way of identifying the optimal thresholding range.In Table 22 and Table 26, we repeat the simulation for 1000 times.The thresholding method works quite well in model selection.

10.2 CTAR method to identify constant component

In this chapter, we use the simulated example to examine the performance of the CTAR method derived above. As previously discussed, CTAR could deal with the scale problems well.

Example 18. In model (1.3), we set $p = 3$, $\beta_1(\cdot) = \sin(\|s\|^2\pi)$ and $\beta_2(\cdot) = \cos(\|s\|^2\pi)$, $\beta_3(\cdot) = \beta_3 = 1$. We generate X_i, s_i, ϵ_i, y_i $i = 1, \cdots, n$, in the same way as before, except that X_i is from $N(\mathbf{0}_3, I_3)$. Based on the generated data, we apply the proposed CTAR method to select the correct model, and examine the performances of the proposed method in identifying the constant components in model (1.3).The bandwidth we used for estimation is $h = 0.45$. We use different λs in the CTAR. We set the sample size $n = 500$. For each λ we did 200 times simulations. We also calculate the ratios of picking right models for each λ. Table 27 illustrated the results of the CTAR, and their performances are quite satisfying.

In Table 23, we notice that the ratio of picking the right model is 1 in the range $[0.25, 0.3]$. Regarding the thresholding method, we have an optimal thresholding K range. It is the same for CTAR method, for which we have an optimal λ range. In our example, the optimal λ range is $[0.25, 0.3]$. Within this range we could detect the constant component in our model efficiently and accurately. Previously, we performed 200 times simulations for each λ value. Theoretically, we should conduct 1000 times simulation for at least 2 λ values within the optimal range. Due to the computational limitations, we only conduct 1000 simulation for $\lambda = 0.28$. We calculate the ratios of picking the

Table 27: **Example 18: The ratios of picking model with different λ**

thresholding value	Right model	constant as function	function as constant
λ=0.05	0	1	0
λ=0.1	0.35	0.65	0
λ=0.15	0.53	0.47	0
λ=0.2	0.89	0.11	0
λ=0.25	1	0	0
λ=0.3	1	0	0
λ=0.35	0.92	0	0.08

The ratios of picking model of each value λ.The sample size $n = 500$. We repeat the simulation 200 times.

right model, and the result is listed in Table 28.

Table 28: **Example 18: The ratios of picking model using the optimal λ**

thresholding value	Right model	constant as function	function as constant
λ=0.28	0.995	0.004	0.001

The ratio of picking the model of optimal λ value $\lambda_0 = 0.28$.The sample size $n = 500$. We repeat the simulation 1000 times.

Example 19. We set $p = 3$, $\sigma^2 = 1$,

$$\alpha = 0.4, \quad \beta_1(s) = 2 - (\|s\|^2), \quad \beta_2(s) = 4 - (\|s\|^2), \beta_3(\cdot) = \beta_3 = 1$$

We generate X_i, s_i, ϵ_i, y_i $i = 1, \cdots, n$, in the same way as before, except that X_i is from $N(\mathbf{0}_3,\ I_3)$. The bandwidth used for estimation is $h = 0.31$, and use different λs in the CTAR. We set the sample size $n = 500$. For each λ we performed 200 times simulations. We also calculated the ratios of picking the right models for each λ. Table 29 illustrates the results of the CTAR, from which it can be seen that the performances are quite satisfying.

Table 29: **Example 19: The ratios of picking model with different λ**

thresholding value	Right model	constant as function	function as constant
λ=0.05	0	1	0
λ=0.1	0.59	0.41	0
λ=0.15	0.87	0.13	0
λ=0.2	1	0	0
λ=0.25	1	0	0
λ=0.3	0.97	0	0.03
λ=0.35	0.89	0	0.11

The ratios of picking model of each value λ.The sample size $n = 500$. We repeat the simulation for 200 times

In Table 29, we note that the ratio of picking the right model is 1 in the range $[0.2, 0.25]$. The optimal λ range is $[0.2, 0.25]$. Within this range we could detect the constant component in our model efficiently and accurately.We conduct 1000 times simulation for $\lambda = 0.23$. We calculate the ratios of picking the right model. The result is listed in Table 30. We could tell the CTAR do works very well in model selection.

The basic idea of the CTAR is the same as the Thresholding K

Table 30: **Example 19: The ratios of picking model using the optimal λ**

thresholding value	Right model	constant as function	function as constant
λ=0.23	0.996	0.002	0.002

The ratios of picking model of optimal λ value $\lambda_0 = 0.23$.The sample size $n = 500$. We repeat the simulation 1000 times

method, both of which works very well. Within the optimal range we detect the constant component in our model efficiently and accurately. However, we encounter the same problem of how to identify the optimal λ range.

10.3 AIC and BIC method

10.3.1 Bandwidth selection in AIC/BIC based model selection

Selecting bandwidth is always an essential problem. We first select the same bandwidth in estimating and model selection, and our results are quite dissatisfying.*AIC* and *BIC* can be described as the trade off between bias and variance in model construction, or loosely speaking between the accuracy and the complexity of the model. The formulas we used to calculate *AIC* and *BIC* in our model are

$$\text{AIC} = n\log(\hat{\sigma}) - \log(|\hat{A}|) + \frac{1}{2\hat{\sigma}^2}(\hat{A}Y - \hat{\mathbf{m}})^{\mathrm{T}}(\hat{A}Y - \hat{\mathbf{m}}) + \mathcal{K}, \quad (8.1)$$

$$\text{BIC} = 2(n\log(\hat{\sigma}) - \log(|\hat{A}|) + \frac{1}{2\hat{\sigma}^2}(\hat{A}Y - \hat{\mathbf{m}})^{\mathrm{T}}(\hat{A}Y - \hat{\mathbf{m}})) + \mathcal{K}\log(n), \tag{8.2}$$

We can not choose a bandwidth h, which is the optimal bandwidth for both estimating and model selection. Thus, we use different bandwidths for these two parts. As for the Thresholding K method and the CTAR method, there is an optimal value range. It is the same for AIC and BIC method, which has an optimal bandwidth range within which the ratio of picking the right model is close to 1. We illustrate the results later.

10.3.2 Simulation results

Example 20.. In model (1.3), we set $p = 3$, $\beta_1(\cdot)$ and $\beta_2(\cdot)$ the same as that in Example 16, $\beta_3(\cdot) = \beta_3 = 1$. We generate X_i, s_i, ϵ_i, y_i $i = 1, \cdots, n$, in the same way as that in Example 16, except that X_i is from $N(\mathbf{0}_3,\ I_3)$. Based on the generated data, we are going to apply the proposed AIC or BIC to select the correct model, and examine the performances of the proposed AIC, BIC and the two algorithms in identifying the constant components in model (1.3).

We still use the Epanechnikov kernel as the kernel function in the model selection, however, the bandwidth used is 0.25 for AIC and 0.35 for BIC, which is smaller than that for estimation. In general, the bandwidth used for model selection should be smaller than that for estimation. In fact, we have tried different bandwidths, it turned out any bandwidth in a reasonable range such as $[0.2,\ 0.3]$ for AIC, [0.3, 0.4] for BIC would do the job very well.

Due to the very expensive computation involved, for any given sample size n, we only do 200 simulations, and in each simulation, we apply either AIC or BIC coupled with either of the two proposed algorithms to select model. For each candidate model, the ratios of picking up this model in the 200 simulations are computed for different cases. The results are presented in Table 31. We can see, from Table 31, the proposed BIC with Backward elimination performs best, and the others are doing reasonably well too.

To make the case more convincing, for sample size 500, we do 1000 simulations for each method. The ratio of picking up each candidate model in the 1000 simulations are presented in Table 32 for each method. It is very clear, the results in Table 32 are consistent with that in Table 31. We conclude all of the proposed model selection methods work well, and the proposed BIC with Backward elimination performs best.

10.4 Optimal bandwidth range

At the beginning of the previous subsection, we discussed an optimal bandwidth range for AIC and BIC model selection methods.Due to the computational cost, we only conduct 200 simulations for bandwidth from $h = 0.15$ to $h = 0.45$ for sample size $n = 500$. We use the backward elimination method. We calculated the ratio of picking the right model. Table 33 shows the results. We could find that the ratio of picking the right model in the bandwidth range $[0.2, 0.3]$ for AIC and $[0.3, 0.4]$ for BIC are close to 1. These would be our optimal

Table 31: **Ratio of Picking Up Each Candidate Model using AIC/BIC**

	{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	{1, 2, 3}	{}
n=400	0	0	0.91	0	0.04	0.02	0.03	0
n=500	0	0	0.98	0	0.02	0	0	0
n=600	0	0	1	0	0	0	0	0
n=400	0	0	0.9	0	0.07	0	0.03	0
n=500	0	0	0.94	0	0.06	0	0	0
n=600	0	0	0.96	0	0.03	0	0.01	0
n=400	0	0	0.92	0	0.05	0	0.03	0
n=500	0	0	0.98	0	0.01	0	0.01	0
n=600	0	0	1	0	0	0	0	0
n=400	0	0	0.89	0	0.09	0	0.02	0
n=500	0	0	0.95	0	0.05	0	0	0
n=600	0	0	0.97	0	0.03	0	0	0

The ratios of picking up each candidate model in 200 *simulations for different sample sizes.* $\{i_1,\ \cdots,\ i_k\}$ *stands for the model in which* $\boldsymbol{\beta}(\cdot)$ *has its* i_1*th,* $\cdots$*,* i_k*th components being constant, and the column corresponding to which is the ratios of picking up this model among* 200 *simulations. Row 2 to row 4 are the ratios obtained based on AIC and Backward elimination when sample size* $n = 400$*,* $n = 500$ *and* $n = 600$*. Row 5 to row 7 are the ratios obtained based on AIC and the CTAR based algorithm, Row 8 to row 10 are the ratios obtained based on BIC and Backward elimination, and Row 11 to row 13 are the ratios obtained based on BIC and the CTAR based algorithm.*

bandwidth ranges. Within them, we could get a satisfying result in model selection.

Table 32: **Ratio of Picking Up Each Candidate Model with simulation time of 1000s**

{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	{1, 2, 3}	{}
0	0	0.989	0	0.01	0	0.001	0
0	0	0.959	0	0.033	0	0.008	0
0	0	0.992	0	0.005	0	0.003	0
0	0	0.963	0	0.031	0	0.006	0

The ratios of picking up each candidate model in 1000 *simulations for sample size* $n = 500$. $\{i_1, \cdots, i_k\}$ *stands for the model in which* $\boldsymbol{\beta}(\cdot)$ *has its* i_1*th,* $\cdots$, i_k*th components being constant, and the column corresponding to which is the ratios of picking up this model among* 1000 *simulations. Row 2 are ratios obtained based on AIC and Backward elimination when sample size* $n = 500$. *Row 3 are the ratios obtained based on AIC and the CTAR based algorithm, Row 4 are the ratios obtained based on BIC and Backward elimination, and Row 5 are the ratios obtained based on BIC and the CTAR based algorithm.*

11 Real Data Analysis

11.1 Introduction

Boston is the capital of the Commonwealth of Massachusetts. It is also the largest and one of the oldest cities in the United States. Boston has a population of around 600 thousand people. The city covers 125 square km. Greater Boston is the fifth-largest area in the United States. Many world-famous universities and research institutes are located in the city and surrounding areas, which makes Boston an international center for education and researching. The universities and research institutes in this area have an notable effect on the region's economy.The region's economic base includes research, finance, man-

Table 33: **Ratio of Picking Up Each Candidate Model using a different bandwidth**

	{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	{1, 2, 3}	{}
h=0.15	0	0	0.79	0	0.04	0.12	0.05	0
h=0.2	0	0	0.92	0	0	0.07	0.01	0
h=0.25	0	0	0.98	0	0.02	0	0	0
h=0.3	0	0	1	0	0	0	0	0
h=0.35	0	0	0.91	0	0.02	0	0.01	0.06
h=0.4	0	0	0.84	0	0.01	0	0	0.15
h=0.45	0	0	0.72	0	0	0	0	0.28
h=0.15	0	0	0.75	0	0.03	0.18	0.04	0
h=0.2	0	0	0.87	0	0	0.11	0.02	0
h=0.25	0	0	0.93	0	0.07	0	0	0
h=0.3	0	0	1	0	0	0	0	0
h=0.35	0	0	0.98	0	0.01	0	0.01	0
h=0.4	0	0	0.94	0	0	0	0	0.06
h=0.45	0	0	0.83	0	0	0	0	0.17

The ratios of picking up each candidate model in 200 *simulations for sample size n=500.* $\{i_1, \cdots, i_k\}$ *stands for the model in which* $\boldsymbol{\beta}(\cdot)$ *has its* $i_1th, \cdots, i_kth$ *components being constant, and the column corresponding to which is the ratios of picking up this model among* 200 *simulations. Row 2 to row 8 are the ratios obtained based on AIC and Backward elimination when sample size* $n = 500$*. Row 9 to row 15 are the ratios obtained based on BIC and Backward elimination.*

ufacturing, etc. and Boston has also received the highest amount of annual funding compared with all other cities in the United States. As a result, Boston is a supreme financial center, that ranks number twelve in the top twenty Global Financial Centers. All of these positive factors make Boston a city with the highest costs of living in the

United States, i.e. 3rd in the United States and 36th globally.

Figure 22: Map of Boston

11.2 Brief description of the data set

The aim of this chapter is to apply the model we developed in the previous chapters to a real data set, specifically the Boston House Price data. More precisely, we apply the proposed model (1.3) together with the associated model selection and estimation method in our analysis.The data set consists of 5 covariates and 1 response variable.The sample size n equals 506. The locations of the houses consists of longitudes and latitudes,converted into U(0,1).We first introduce the data and analyze the data set.

Response variable **MEDV**: Median value of owner-occupied homes in \$1000's

Covariate1 **CRIM**: per capita crime rate by town

Covariate2 **RM**: average number of rooms per dwelling

Covariate 3 **RAD**: index of accessibility to radial highways

Covariate 4 **TAX**: full-value property-tax rate per $10,000 dollar

Covariate 5 **LSTAT**: The Percent of the lower status of the population

11.2.1 Descriptions of covariates

As in the previous chapter, we know there are 5 covariates and 1 response variable in this data set. We calculate the mean and standard deviation of each covariate, and the order of the data set is the same.We could have some rough idea of our data set. Table 34 describes our statistics. Figure 23 are the histograms of the 5 covariates.

Table 34: **Descriptive Statistics**

	Mean	Std.Deviation	N
MEDV	22.5328	9.1971	506
CRIM	3.6135	8.6015	506
RM	6.284634	0.7026	506
RAD	9.55	8.707	506
TAX	408.24	168.537	506
LSTAT	12.6530	7.1410	506

We use the statistical software SAS-"Statistical Analysis System" to get the above results. The sample size is 506.

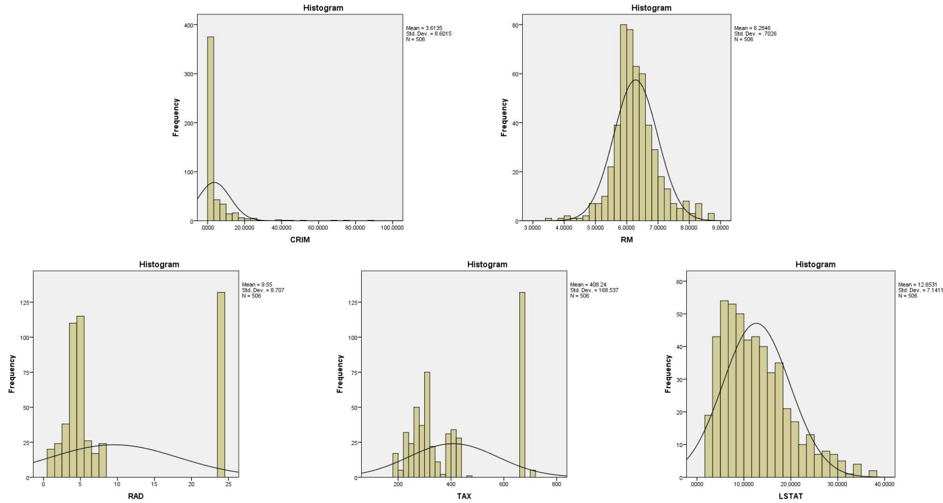

Figure 23: *The Histograms of the 5 Covariates.The left one on the upper panel is the histogram of* **CRIM***, the right one on the upper panel is the histogram of* **RM***. The left one in the lower panel is the histogram of* **RAD***, the middle on in the lower panel is the histogram of* **TAX***, the right one in the lower panel is the histogram of* **LSTAT**

11.3 Parametric way of analyzing data

In the previous chapter, we analyze the data set to have some basic ideas. In this chapter, we use the parametric way to analyze our data set, that is the Linear Regression Model. We ignore the spatial interaction term, and consider the effect on the covariate to be constant.Then our model would be as follows:

$$y_i = X_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad i = 1,\ \cdots,\ n, \tag{8.1}$$

where $\boldsymbol{\beta} = (\beta_0,\ \cdots, \beta_p)$,

In the linear regression model, we only need to estimate the coefficients $\boldsymbol{\beta} = (\beta_0,\ \cdots, \beta_p)$. And we use AIC backward elimination method to help us select the significant variables. Table 35 and Table 36 provide the results of AIC Backward elimination and estimation, which indicate that the 5 covariates are all significant.However, the linear regression model is not realistic nor is it adequate to analyze data. Spatial interaction is a real-world phenomenon that should be considered in the model. As a result, in the next chapter, we apply the new model we proposed to fit the data set.

Table 35: **Model Summary**

Model	R	R square	Adjusted R square	Std. Error of the estimate
1	.809	.655	.651	.5.4310

a. Predictors:(Constant), LSTAT,RM,CRIM,RAD,TAX
b. Dependent Variable: MEDV

Table 36: **Coefficients**

Model	B	Std.Error	Standardized Coefficients Beta	t	Sig
Constant	1.120	3.328		.336	.737
CRIM	-.087	.037	-.082	-2.368	.018
RM	5.090	.442	.389	11.509	.000
RAD	.173	.071	.164	2.447	.015
TAX	-.013	.004	-.230	-3.463	.001
LSTAT	-.537	.050	-.417	-10.733	.000

a. Predictors:(Constant), LSTAT,RM,CRIM,RAD,TAX
b. Dependent Variable: MEDV
c.Model: Linear Regression Model

11.4 Model selection and estimation

In this chapter, we are going to apply the proposed model (1.3) together with the proposed model selection and estimation method to analyse the Boston house price data. Specifically, we are going to explore how some factors such as CRIM, RM,RAD, TAX, and LSTAT affect the median value of owner-occupied homes in \$1000's (denoted by MEDV), and whether the effects of these factors vary over location.

We use model (1.3) to fit the data with y_i, x_{i1}, x_{i2}, x_{i3}, x_{i4} and x_{i5} being MEDV, CRIM, RM, RAD, TAX and LSTAT, respectively, and $X_i = (x_{i1}, \cdots, x_{i5})^{\mathrm{T}}$. The kernel function used in either estimation procedure or model selection is taken to be the Epanechnikov kernel.

We first try to find which factors have location varying effects on the house price, and which factors do not. This is equivalent to identifying the constant coefficients in the model used to fit the data. We apply the proposed BIC coupled with Backward elimination to do the model selection, and the bandwidth used is chosen to be 17% of the range of the locations. The obtained result shows the coefficients of x_{i3} and x_{i5} are constant, which means all factors, except RAD and LSTAT, have location varying effects on the house price.

We now apply the chosen model

$$y_i = \alpha \sum_{j \neq i} w_{ij} y_j + x_{i1}\beta_1(s_i) + x_{i2}\beta_2(s_i) + x_{i3}\beta_3 + x_{i4}\beta_4(s_i) + x_{i5}\beta_5 + \epsilon_i, \tag{8.2}$$

$i = 1, \cdots, n$, where w_{ij} is defined by (8.1), to fit the data. The proposed estimation procedure is used to estimate the unknown functions and constants, and the bandwidth used in the estimation procedure is taken to be 60% of the range of the locations. The estimates of the unknown constants are presented in Table 37, and the estimates of the unknown functions are presented in Fig 23.

As β_3 and β_5 can be interpreted as the impacts of RAD and LSTAT, respectively, Table 37 shows the index of accessibility to radial highways has positive impact on house price and the percentage of the lower status of the population has negative impact on house price. Apparently, this makes sense. Table 35 also shows that the estimate of α is 0.221, which is an unignorable effect, and indicates the house prices in a neighbourhood do affect each other. This is a true phenomenon in real world.

Table 37: **Estimates of The Unknown Constant Coefficients**

$\hat{\alpha}$	$\hat{\beta}_3$	$\hat{\beta}_5$
0.2210	0.3589	-0.4473

From Fig 24, we can see the impact $\beta_1(\cdot)$ of the per capita crime rate by town on house price is negative and is clearly varying over location. The impact $\beta_2(\cdot)$ of the average number of rooms per dwelling on house price is positive and is also varying over location. It is interesting to see that the impact of the average number of rooms per dwelling is lower in the area where the impact of crime rate is high than the area where the impact of crime rate is low. This implies that the crime rate is a dominate factor on the house price in the area where the impact of crime rate is high. Fig 24 also shows the association between the house price and the full-value property-tax rate is varying over location, and it is generally positive, however, there are some areas where this

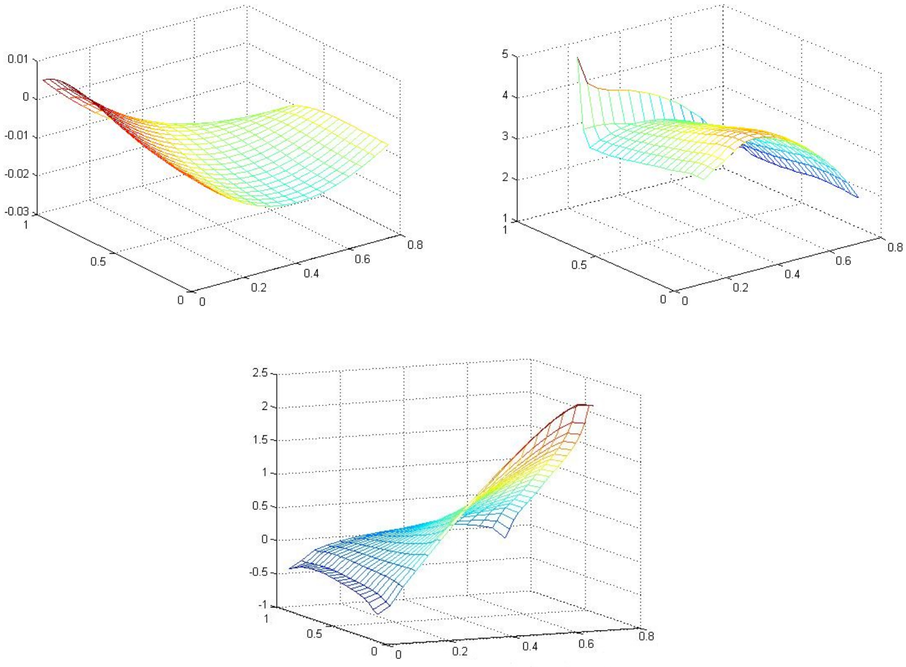

Figure 24: *The 3D plots of $\hat{\beta}_1(s)$, $\hat{\beta}_2(s)$ and $\hat{\beta}_4(s)$. The left one in the upper panel is $\hat{\beta}_1(s)$, right one in the upper panel is $\hat{\beta}_2(s)$, and the one in the lower panel is $\hat{\beta}_4(s)$.*

association is negative. We can also see that the impact of the average number of rooms per dwelling is lower in the area, where the association between the house price and the full-value property-tax rate is strong, than the area where the association is weak.

12 Conclusions and Future Work

In Chapter 2 of this thesis, we introduce the framework of the local polynomial modelling and the multivariate version of the method. In Chapter 3, we propose the estimation procedure for the designed model, in which we use kernel smoothing and local polynomial fitting. However, we never discuss how to select the bandwidth. The choice of the bandwidth plays an important role,and its selection is one of the most important tasks in estimation. This may require further work. In Chapters 4, 5 and 6, we provide the asymptotic properties of the proposed estimators followed by the proofs of the theorems and lemmas. As we mentioned above, due to the structure of our model, it is not straight-forward to apply the cross-validation method. In Chapter 7, we discuss the connection between these two important model selection methods. In Chapter 8, we provided several model selection methods.

We also designed the hypothesis testing for our model. Due to the computational limitations, we did not have enough time to do simulations and apply this method to real data.We will illustrate our idea here.

Bootstrapping method is a nonparametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results. In our model, we use the bootstrapping method to help us detect whether the component of $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot),\ \cdots,\ \beta_p(\cdot))^{\mathrm{T}}$, is a function, constant or zero. Our aim is to construct a hypothesis testing to test whether a specific covariate is significant. If it is, whether its coefficient is functional or constant.

Procedure

We construct the hypothesis testing as follows. Without lose of generality, we take the first component of $\boldsymbol{\beta}(\cdot)$, i.e. $\beta_1(\cdot)$ as an example. We want to test whether $\beta_1(\cdot)$ is constant. If $\beta_1(\cdot)$ is constant, then we do another hypothesis test to determine whether $\beta_1(\cdot)$ is zero. Our test statistic is $T_1 = \sum_{j=1}^{K} \mid \hat{\beta}_1(s_j) - \bar{\beta}_1 \mid \frac{1}{K}$. Our Hypothesis testing is reject H_0 when $T_1 > C$,otherwise we accept the H_0, where $P(T_1 > C) = \alpha$. To find C, we appeal Bootstrapping method as follows. We use the estimation method introduced in Chapter 3 to get the estimators of $\boldsymbol{\beta}(\cdot)$, where $\hat{\boldsymbol{\beta}}(\cdot) = (\hat{\beta}_1, \hat{\beta}_2(\cdot), \cdots, \hat{\beta}_p(\cdot))^{\mathrm{T}}$ and the estimator of $\hat{\alpha}$. In bootstrapping re-sampling, we fixed $\hat{\boldsymbol{\beta}}(\cdot)$ and $\hat{\alpha}$. We calculate the residuals of the response variable y_i, which is $\hat{\epsilon}_i = y_i - \hat{y}_i, i = 1, 2, ..., n$. We conduct the bootstrap re-sampling for the residuals $\{\ \hat{\epsilon}_1,\ \cdots,\ \hat{\epsilon}_n\}$ r times to get the bootstrap samples $\hat{\epsilon}_k^* = (\hat{\epsilon}_{1k}^*, \cdots, \hat{\epsilon}_{nk}^*)$, $k = 1, 2, \cdots, r$. According to the results presented by Efron and Tibshirani (1993,chap.19), which suggest that bootstrap confidence intervals on 1000 bootstrap samples generally provides accurate results, and using 2,000 bootstrap replications should be very safe. $\hat{\epsilon}_b^*$ are the re-sampled residuals for the b^{th} bootstrap sample.

After we get the b^{th} bootstrap sample for $\hat{\epsilon}$, where $\hat{\epsilon}_b^* = (\hat{\epsilon}_{1b}^*, \cdots, \hat{\epsilon}_{nb}^*)^{\mathrm{T}}$, We could get the b^{th} bootstrap sample for the response variable y_b^*, where $y_b^* = (y_{1b}^*,\ \cdots,\ y_{nb}^*)^{\mathrm{T}}$ The procedure is as follows:

$$y_{ib}^* = \hat{\alpha} \sum_{j \neq i} w_{ij} y_{jb}^* + X_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}(s_i) + \hat{\epsilon}_{ib}^*, \quad i = 1,\ \cdots,\ n, \quad \hat{\boldsymbol{\beta}}(s_i) = (\hat{\beta}_1, \hat{\beta}_2(s_i), \cdots, \hat{\beta}_p(\cdot)) \tag{8.1}$$

We could see it in the matrix form:

$$(I - \hat{\alpha} W)Y^* = X^T \hat{\boldsymbol{\beta}}(\cdot) + \epsilon^* \tag{8.2}$$

After we get the b^{th} bootstrap y_b^*, we could use the same estimation procedures we used before to get the b^{th} bootstrap sample of $\hat{\boldsymbol{\beta}}(\cdot)$, where $\hat{\boldsymbol{\beta}}_b^*(\cdot) = (\hat{\beta}_{b1}^*, \cdots, \hat{\beta}_{bn}^*(\cdot))^{\mathrm{T}}$. We calculate $T_1^* = \sum\limits_{j=1}^{N} \mid \hat{\beta}_{b1}^*(s_j) - \bar{\beta}_1{}^* \mid \frac{1}{N}$.We repeat the bootstrapping procedure r times, i.e. $T_1^*, T_2^*, \cdots, T_r^*$ .It is straight-forward to find the p-value for the hypothesis testing based on the bootstrap sampling. Then we could get the results of whether $\beta_1(\cdot)$ is a significant covariate or not.If we accept the null hypothesis, then we continue. We then test whether $\beta_1 = 0$. Hypothesis testing is good tool to help us detect the parametric/nonparametric components. Unfortunately, due to the huge computational cost, we do not have enough time to finish it.

The method of re-sample the residuals:

We randomly draw. Each bootstrap residual sample selects n values with the replacement among the n values of the original residual sample,

$$\hat{\epsilon}_{ib}^* = random - draw(\hat{\epsilon}_1, \ \cdots, \ \hat{\epsilon}_n)$$

There is still a great deal of interesting research left. Three years is too short a time to investigate all of them. However, We will continue our work along these lines in the future.

List of References

Akaike, H. (1973). 'Information theory and an extension of the maximum likelihood principle.' In: B. N. Petrov and F. Csáki, eds., *2nd International Symposium on Information Theory* (Akadémia Kiadó, Budapest), 267-281.

Anselin, L. (1988): *Spatial Econometrics: Methods and Models.* The Netherlands: Kluwer Academic Publishers.

Akaike, H. (1974). 'A new look at the statistical model identification'. *IEEE Transactions on Automatic Control* **AC - 19**, 716-723.

Allen, D. M. (1974). 'The relationship between variable selection and data augmentation and a method for prediction.' *Technometrics*, **16**, 125-127.

Arlot, S. and Celisse, A. (2010). 'A survey of cross-validation procedures for model selection'. *Statistics Surveys*, **4**, 40-79.

Cheng, M., Zhang, W. and Chen, L. (2009). 'Statistical estimation in generalized multiparameter likelihood models.' *Journal of the American Statistical Association*, **104**, 1179-1191.

Davies, S. L., Neath, A. A. and Cavanaugh J. E. (2005). 'Cross validation model selection criteria for linear regression based on the KullbackLeibler discrepancy'. *Statistical Methodology*, **2**, 249-266.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J. and Zhang, W. (1999). 'Statistical estimation in varying coefficient models.' *The Annals of Statistics*, **27**, 1491-1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715-731.

Gao, J., Lu, Z. and Tjostheim, D. (2006). Estimation in semiparametric spatial regression. *The Annals of Statistics*, **34**, 1395-1435

Hansen Bruce E.(2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24**, 726-748.

Kelejian, H. H. and Prucha, I. R. (2001). On the aymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* , **104**, 219-257.

Kelejian, H. H. and Prucha, I. R., 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, **157**, 53-67.

Kai, B., Li, R., and Zou, H. (2011). 'New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models.' *The Annals of Statistics*, **39**, 305-332.

Lee, L.-F. (2004). Asymptotic distribution of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, **72** , 1899-1925.

Li, J. and Palta, M. (2009). 'Bandwidth selection through cross validation for semi-parametric varying-coefficient partially linear models.' *Journal of Statistical Computation and Simulation.* **79**, 1277-1286.

Li, J. and Zhang, W. (2011). 'A semiparametric threshold model for censored longitudinal data analysis.' *Journal of the American Statistical Association*, **106**, 685-696.

Liang, H. and Li, R. (2009). 'Variable selection for partially linear models with measurement Errors.' *Journal of American Statistical Association.* **104**, 234-248.

Lv, J. and Liu, J. S. (2010). 'Model selection principles in misspecified models.' `http://arxiv.org/abs/1005.5483v1`

Ma, Y., Chiou, J.-M. and Wang, N. (2006). 'Efficient semiparametric estimator for heteroscedastic partially-linear models,' *Biometrika*, **93**, 75-84.

Linton O. (1995). Second order approximation in the partially linear regression model. *Econometrica*, **63**, 1079-1112.

Ord, J. K., 1975. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**, 120-126.

Su L. and Jin S. (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics* , **157**, 18-33.

Stone, M. (1977). 'An asymptotic equivalence of choice of model by cross-validation and Akaikes criterion.' *Journal of the Royal Statistical Society, Series B*, **39**, 44-47.

Sun, Y., Zhang, W. and Tong, H. (2007). 'Estimation of the covariance matrix of random effects in longitudinal studies'. *The Annals of Statistics*. **35**, 2795-2814.

Tao, H. and Xia, Y. (2011) Adaptive semi-varying coefficient model selection, *Statistica Sinica*, **22**, 575-599.

White, H. (1994). *Estimation, inference and specification analysis.* Cambridge University Press.

Wang, H. and Xia, Y. (2009). 'Shrinkage estimation of the varying coefficient model.' *Journal of the American Statistical Association.* **104**, 747-757.

Wang, L., Kai, B. and Li, R. (2009). 'Local rank inference for varying coefficient models.' *Journal of American Statistical Association*, **104**, 1631-1645.

Xia, Y. and Li, W. K. (2002). 'Asymptotic behavior of bandwidth selected by cross-validation method under dependence.' *Journal Multivariate Analysis*, **83**, 265-287.

Zhang, W., Fan, J. and Sun, Y. (2009). 'A semiparametric model for cluster data.' *The Annals of Statistics*, **37**, 2377-2408.

Zhang, W., Lee, S. Y. and Song, X. (2002). 'Local polynomial fitting in semivarying coefficient models.' *Journal of Multivariate Analysis*, **82**, 166-188.

Zhang, W. and Peng, H. (2010). ‘Simultaneous confidence band and hypothesis test in generalized varying-coefficient models.’ *Journal of Multivariate Analysis*, **101**, 1656-1680.

Wang, H. and Xia, Y. (2009) Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association.* **104**, 747-757.