STATISTICAL INFERENCE FOR VARYING COEFFICIENT MODELS

by

YIXIN CHEN

B.S., Henan University, China, 2008

M.S., Pittsburg State University, USA, 2010

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Abstract

This dissertation contains two projects that are related to varying coefficient models. The traditional least squares based kernel estimates of the varying coefficient model will lose some efficiency when the error distribution is not normal. In the first project, we propose a novel adaptive estimation method that can adapt to different error distributions and provide an efficient EM algorithm to implement the proposed estimation. The asymptotic properties of the resulting estimator is established. Both simulation studies and real data examples are used to illustrate the finite sample performance of the new estimation procedure. The numerical results show that the gain of the adaptive procedure over the least squares estimation can be quite substantial for non-Gaussian errors.

In the second project, we propose a unified inference for sparse and dense longitudinal data in time-varying coefficient models. The time-varying coefficient model is a special case of the varying coefficient model and is very useful in longitudinal/panel data analysis. A mixed-effects time-varying coefficient model is considered to account for the within subject correlation for longitudinal data. We show that when the kernel smoothing method is used to estimate the smooth functions in the time-varying coefficient model for sparse or dense longitudinal data, the asymptotic results of these two situations are essentially different. Therefore, a subjective choice between the sparse and dense cases may lead to wrong conclusions for statistical inference. In order to solve this problem, we establish a unified self-normalized central limit theorem, based on which a unified inference is proposed without deciding whether the data are sparse or dense. The effectiveness of the proposed unified inference is demonstrated through a simulation study and a real data application.

Key words: Varying coefficient models; adaptive estimation; local maximum likelihood; kernel smoothing; EM algorithm; dense longitudinal data; sparse longitudinal data; time-varying coefficient models; self-normalization.

STATISTICAL INFERENCE FOR VARYING COEFFICIENT MODELS

by

Yixin Chen

B.S., Henan University, China, 2008

M.S., Pittsburg State University, USA, 2010

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Approved by:

Major Professor
Dr. Weixin Yao

Copyright

Abstract

This dissertation contains two projects that are related to varying coefficient models. The traditional least squares based kernel estimates of the varying coefficient model will lose some efficiency when the error distribution is not normal. In the first project, we propose a novel adaptive estimation method that can adapt to different error distributions and provide an efficient EM algorithm to implement the proposed estimation. The asymptotic properties of the resulting estimator is established. Both simulation studies and real data examples are used to illustrate the finite sample performance of the new estimation procedure. The numerical results show that the gain of the adaptive procedure over the least squares estimation can be quite substantial for non-Gaussian errors.

In the second project, we propose a unified inference for sparse and dense longitudinal data in time-varying coefficient models. The time-varying coefficient model is a special case of the varying coefficient model and is very useful in longitudinal/panel data analysis. A mixed-effects time-varying coefficient model is considered to account for the within subject correlation for longitudinal data. We show that when the kernel smoothing method is used to estimate the smooth functions in the time-varying coefficient model for sparse or dense longitudinal data, the asymptotic results of these two situations are essentially different. Therefore, a subjective choice between the sparse and dense cases may lead to wrong conclusions for statistical inference. In order to solve this problem, we establish a unified self-normalized central limit theorem, based on which a unified inference is proposed without deciding whether the data are sparse or dense. The effectiveness of the proposed unified inference is demonstrated through a simulation study and a real data application.

Key words: Varying coefficient models; adaptive estimation; local maximum likelihood; kernel smoothing; EM algorithm; dense longitudinal data; sparse longitudinal data; time-varying coefficient models; self-normalization.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my appreciation to my major professor, Dr. Weixin Yao, for all his encouragement, guidance and suggestions.

I would also like to thank Dr. Gary Gadbury, Dr. Juan Du and Dr. Xinming Ou for their willingness to serve on my committee and for their valuable insight.

In addition, I would like to thank Dr. Marianne Korten for her willingness to be the chairperson of the examining committee for my doctoral degree.

My gratefulness extends to everyone who supported me in any respect during the completion of this dissertation.

Chapter 1

Adaptive Estimation for Varying Coefficient Models

1.1 Introduction

Since the introduction in Cleveland, et al. (1992) and Hastie and Tibshirani (1993), varying coefficient models have gained considerable attention due to their flexibility and good interpretability. They are useful extensions of the classical linear models and have been widely used to explore the dynamic pattern in many scientific areas, such as finance, economics, epidemiology, ecology, etc. By allowing coefficients to vary over the so-called index variable, the modeling bias can be significantly reduced and the 'curse of dimensionality' can be avoided (Fan and Zhang, 2008). In recent years, varying coefficient models have experienced rapid developments in both theory and methodology, see, for example, Wu, et al. (1998), Hoover, et al. (1998), Fan and Zhang (1999, 2000), Cai, et al. (2000), Fan and Huang (2005), Wang, et al. (2009), Wang and Xia (2009), etc. We refer readers to Fan and Zhang (2008) for a nice and comprehensive survey.

Let $y \in \mathcal{R}^1$ be the response, $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathcal{R}^d$ be the covariate vector, and $u \in \mathcal{R}^1$

is the index variable. The varying coefficient model is defined as

$$y = \sum_{j=1}^{d} g_j(u)x_j + \epsilon, \tag{1.1}$$

where $\{g_1(u), \ldots, g_d(u)\}^T$ are unknown smooth coefficient functions. Throughout this chapter, we assume the random error ϵ to be independent of $(u, \mathbf{x})$, with mean 0 and a finite second-order moment σ^2. By setting $x_1 \equiv 1$, it allows a varying intercept in the model.

Hastie and Tibshirani (1993), Hoover, et al. (1998), Chiang, et al. (2001), and Eubank, et al. (2004) proposed using smoothing spline to estimate coefficient functions. Polynomial spline was used in Huang, et al. (2002, 2004) and Huang and Shen (2004). Wu, et al. (1998), Hoover, et al. (1998), Fan and Zhang (1999), and Kauermann and Tutz (1999) adopted kernel smoothing to estimate coefficient functions. Fan and Zhang (2000) further studied a two-step estimation procedure to deal with the situation where the coefficient functions admit different degrees of smoothness. Recently, Wang and Xia (2009) proposed a shrinkage estimation procedure to select important nonparametric components. Wang, et al. (2009) developed a highly robust and efficient procedure based on local ranks.

Nevertheless, most of existing methods used least squares type criteria in estimation, which corresponds to the local likelihood when the error ϵ is normal. However, in the absence of normality, the traditional least squares based estimators will lose some efficiency.

In this chapter, we propose a novel adaptive kernel estimation procedure for varying coefficient models. The new adaptive method combines the kernel density estimation and the local maximum likelihood estimation such that the new estimator can adapt to different error distributions. The new adaptive estimator is shown to enjoy the asymptotic oracle property, i.e., it is asymptotically as efficient as if the error density were known. An efficient EM algorithm is proposed to implement the adaptive estimation method. We demonstrate through a simulation study that the new estimate is more efficient than the existing least squares based kernel estimate when the error distribution deviates from normal. In addition,

when the error is exactly normal, the new method is broadly comparable to the existing kernel approach. We further illustrate the effectiveness of the proposed adaptive estimation method with two real data examples.

The rest of this chapter is organized as follows. In Section 1.2, we introduce the new adaptive estimation method for the varying coefficient models and the EM algorithm. In Section 1.3, we compare our proposed adaptive estimation with the traditional least squares based estimation for five different error densities through a simulation study and then apply the new method to two real data examples. We conclude this chapter with a brief discussion in Section 1.4. All technical conditions and proofs are relegated to Section 1.5.

1.2 New Adaptive Estimation

1.2.1 Introduction to The New Method

Suppose that $\{\mathbf{x}_i, u_i, y_i, i = 1, \ldots, n\}$ is a random sample from model (1.1). For u in a neighborhood of u_0, we can approximate varying coefficient functions locally as

$$g_j(u) \approx g_j(u_0) + g_j'(u_0)(u - u_0) \equiv b_j + c_j(u - u_0), \text{ for } j = 1, \ldots, d. \tag{1.2}$$

The traditional local linear estimation of (1.1) is to minimize

$$\sum_{i=1}^{n} K_h(u_i - u_0) \left[y_i - \sum_{j=1}^{d} \{b_j + c_j(u_i - u_0)\} x_{ij} \right]^2, \tag{1.3}$$

for a given kernel density $K(\cdot)$ and a bandwidth h, where $K_h(t) = h^{-1}K(t/h)$. It is well known that the choice of kernel function is not critical in terms of estimation efficiency. Throughout this chapter, a Gaussian kernel will be used for $K(\cdot)$. Due to the least squares in (1.3), the resulting estimate may lose some efficiency when the error distribution is not normal. Therefore, it is desirable to develop an estimation procedure which can adapt to

different error distributions.

Let $f(\epsilon)$ be the density function of ϵ. If $f(\epsilon)$ were known, it would be natural to estimate the local parameters in (1.2) by maximizing the following local log-likelihood function

$$\sum_{i=1}^{n} K_h(u_i - u_0) \log f \left[y_i - \sum_{j=1}^{d} \{b_j + c_j(u_i - u_0)\} x_{ij} \right]. \tag{1.4}$$

However, in practice, $f(\epsilon)$ is generally unknown but can be replaced by a kernel density estimate based on the initial estimated residual $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$

$$\tilde{f}(\epsilon_i) = \frac{1}{n} \sum_{j \neq i}^{n} K_{h_0} (\epsilon_i - \tilde{\epsilon}_j), \quad \text{for } i, j = 1, 2, ..., n \tag{1.5}$$

where $\tilde{\epsilon}_i = y_i - \sum_{j=1}^{d} \tilde{g}_j(u_i) x_{ij}$ and $\tilde{g}_j(\cdot)$ can be estimated by least squares (or L_1 norm, i.e., median regression) based local linear estimate (1.3). Here we use leave-one-out kernel density estimate for $f(\epsilon_i)$ to remove the estimation bias. Let $\boldsymbol{\theta} = (b_1, \ldots, b_d, c_1, \ldots, c_d)^T$. Then our proposed adaptive local linear estimate for the local parameter $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}), \tag{1.6}$$

where

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} K_h(u_i - u_0) \log \left(\frac{1}{n} \sum_{j \neq i} K_{h_0} \left[y_i - \sum_{l=1}^{d} \{b_l + c_l(u_i - u_0)\} x_{il} - \tilde{\epsilon}_j \right] \right). \tag{1.7}$$

The idea of adaptiveness can be traced back to Beran (1974) and Stone (1975), where the adaptive estimation was proposed for location models. Later, this idea was extended to regression, time series and other models, see Bickel (1982), Manski (1984), Steigerwald (1992), Schick (1993), Drost and Klaassen (1997), Hodgson (1998), Yuan and De Gooijer (2007), and Yuan (2009). Linton and Xiao (2007) proposed an elegant adaptive nonparametric regression estimator by maximizing the local likelihood function. In fact, the adaptive

method proposed in Linton and Xiao (2007) can be seen as a special case of ours when $d = 1$ in (1.1). Wang and Yao (2012) extended the idea of adaptive estimation to sufficient dimension reduction.

1.2.2 Computation: An EM Algorithm

Unlike least squares criterion, (1.6) does not have an explicit solution. In this section, we propose an EM algorithm to maximize it by extending the generalized modal EM algorithm proposed in Yao (2013).

Let $\boldsymbol{\theta}^{(0)}$ be an initial parameter estimate, such as the least squares (or L_1 norm, i.e., median regression) based local linear estimate. We can update the parameter estimate according to the following algorithm.

Algorithm 1.2.1. *At $(k+1)$th step, we calculate the following E and M steps:*

E-Step*: Calculate the classification probabilities,*

$$\begin{aligned} p_{ij}^{(k+1)} =& \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]}{\sum_{j\neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]} \\ \propto& K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right],\ 1 \leq j \neq i \leq n. \end{aligned} \tag{1.8}$$

M-Step*: Update $\boldsymbol{\theta}^{(k+1)}$,*

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} =& \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)} K_h(u_i - u_0)\log\left(K_{h_0}\left[y_i - \sum_{l=1}^{d}\{b_l + c_l(u_i - u_0)\}x_{il} - \tilde{\epsilon}_j\right]\right)\right\} \\ =& \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)} K_h(u_i - u_0)\left[y_i - \tilde{\epsilon}_j - \boldsymbol{z}_i^T\boldsymbol{\theta}\right]^2\right\}, \\ =& \left(\sum_{i=1}^{n}\sum_{j\neq i} p_{ij}^{(k+1)} K_h(u_i - u_0)\boldsymbol{z}_i\boldsymbol{z}_i^T\right)^{-1} \sum_{i=1}^{n}\sum_{j\neq i} p_{ij}^{(k+1)} K_h(u_i - u_0)(y_i - \tilde{\epsilon}_j)\boldsymbol{z}_i \end{aligned} \tag{1.9}$$

where $\boldsymbol{z}_i = \{\boldsymbol{x}_i^T, \boldsymbol{x}_i^T(u_i - u_0)\}^T$ *and the second equation follows the use of Gaussian kernel for density estimation.*

The above EM algorithm monotonically increases the estimated local log-likelihood (1.7) after each iteration, as shown in the following theorem.

Theorem 1.2.1. *Each iteration of the above E and M steps will monotonically increase the local log-likelihood (1.7), i.e.,*

$$Q(\boldsymbol{\theta}^{(k+1)}) \geqslant Q(\boldsymbol{\theta}^{(k)}),$$

for all k, where $Q(\cdot)$ *is defined as in (1.7).*

1.2.3 Asymptotic Result

We now derive the asymptotic distribution of the proposed adaptive local linear estimator of $\boldsymbol{\theta}$. Define $\mu_k = \int u^k K(u)\mathrm{d}u$ and $\nu_k = \int u^k K^2(u)\mathrm{d}u$. Let $\mathbf{H} = \mathrm{diag}(1, h) \otimes \mathbf{I}_d$ with $\otimes$ denoting the Kronecker product and $\mathbf{I}_d$ being the $d \times d$ identity matrix. Let $q(\cdot)$ denote the marginal density of u, and

$$\Gamma_{jk}(u_i) = \mathrm{E}(x_{ij}x_{ik}|u_i) \text{ for } 1 \leq j, k \leq d,\ i = 1, ..., n, \tag{1.10}$$

$$\boldsymbol{\Gamma}(u_0) = \{\Gamma_{jk}(u_0)\}_{1 \leqslant j,k \leqslant d}\,. \tag{1.11}$$

Theorem 1.2.2. *Suppose that the regularity conditions in Section 1.5 hold. Then, with probability approaching 1, there exists a consistent local maximizer* $\hat{\boldsymbol{\theta}} = (\hat{b}_1, \ldots, \hat{b}_d, \hat{c}_1, \ldots, \hat{c}_d)^T$ *of (1.7) such that*

$$\sqrt{nh}\left\{\boldsymbol{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \boldsymbol{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j''(u_0)\boldsymbol{\psi}_j(1 + o_p(1))\right\} \xrightarrow{D} N(\boldsymbol{0}_{2d}, \delta_1^{-2}\delta_2 q(u_0)^{-1}\boldsymbol{S}^{-1}\boldsymbol{\Lambda}\boldsymbol{S}^{-1}),$$

where $\mathbf{0}_{2d}$ *is a* $2d\times 1$ *vector with each entry being 0,* $\rho(\cdot)=\log f(\cdot)$, $\delta_1 = E\{\rho''(\epsilon_i)\}$, $\delta_2 = E\{\rho'(\epsilon_i)^2\}$, $\boldsymbol{S} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \otimes \mathbf{\Gamma}(u_0)$, $\mathbf{\Lambda} = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \mathbf{\Gamma}(u_0)$, $\boldsymbol{\psi}_j = \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes (\Gamma_{jk}(u_0))_{1\le k\le d}^T$, *and* $\mathbf{\Gamma}(u_0)$ *is given by (1.11).*

A sketch of the proof of the above theorem is provided in Section 1.5. As shown in Linton and Xiao (2007), one important property of the proposed adaptive estimate is its asymptotic oracle property, i.e., it achieves the same asymptotic efficiency as if the error density were known. Therefore, the effect of estimating f by kernel density estimate will not affect the asymptotic distribution of the resulting estimator of $\boldsymbol{\theta}$.

1.3 Examples

1.3.1 Simulation Study

In this section, we conduct a simulation study to compare the proposed adaptive estimation (Adapt) with the traditional least squares based kernel estimation (LS) for varying coefficient models. The following five error distributions of ϵ were considered in our numerical experiment:

1. $N(0,1)$;
2. t_3;
3. $0.5N(-1,0.5^2)+0.5N(1,0.5^2)$;
4. $0.3N(-1.4,1)+0.7N(0.6,0.4^2)$;
5. $0.9N(0,1)+0.1N(0,10^2)$.

The standard normal distribution serves as a baseline in our comparison. The second one is a t-distribution with 3 degrees of freedom. The third density is bimodal and the fourth

one is left skewed. The last one is a contaminated normal mixture distribution, where 10% of the data from $N(0, 10^2)$ are most likely to be outliers.

For each of the above error distributions, we consider the following two models:

Model 1: $y = g_1(u) + g_2(u)x_1 + g_3(u)x_2 + \epsilon$, where $g_1(u) = \exp(2u-1), g_2(u) = 8u(1-u)$, and $g_3(u) = 2\sin^2(2\pi u)$.

Model 2: $y = g_1(u) + g_2(u)x_1 + g_3(u)x_2 + \epsilon$, where $g_1(u) = \sin(2\pi u), g_2(u) = (2u-1)^2 + 0.5$, and $g_3(u) = \exp(2u-1) - 1$.

In both models, x_1 and x_2 follow a standard normal distribution with correlation coefficient $\gamma = 1/\sqrt{2}$. The index variable u is a uniform random variable on [0, 1], and is independent of (x_1, x_2). We conduct two simulations with sample size n=200 and 400 respectively, each with 200 data replications. There are two bandwidths in the estimation, h in the local log-likelihood and h_0 in the kernel density estimation. The bandwidth h is chosen by cross-validation with more details in Fan and Zhang (1999), and $h_0 = h/\log(n)$ following Linton and Xiao (2007). The performance of estimator $\hat{g}(\cdot)$ is assessed via the square root of the average squared errors (RASE; Cai, et al., 2000; Wang, et al., 2009),

$$\text{RASE}^2 = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{3}[\hat{g}_j(u_k) - g_j(u_k)]^2, \tag{1.12}$$

where u_k, $k = 1, \ldots, N$, are the equally spaced grid points at which the functions $g_j(\cdot)$ were evaluated. We used N=200 in the numerical studies.

The simulation results are summarized in Tables 1.1 and 1.2. We can clearly see that the proposed adaptive estimation outperforms the least squares method when the error is non-normal. The gain in estimation efficiency can be quite substantial even for moderate sample sizes. When the error follows exactly normal distribution, our approach is still broadly comparable with the least squares based method.

Figures 1.1 and 1.2 plot the estimated coefficient functions and the 95% pointwise confidence intervals based on a typical sample when n=200 and the error distribution is the

contaminated normal mixture (Case 5). It is clear that the adaptive estimation method provides narrower confidence intervals than the least squares based method, as expected.

Table 1.1: *Model 1 estimation accuracy comparison–RASE and its standard error in brackets.*

ϵ	$n = 200$		$n = 400$	
	LS	Adapt	LS	Adapt
1	0.483(0.079)	0.439(0.081)	0.366(0.053)	0.324(0.053)
2	0.671(0.167)	0.601(0.139)	0.493(0.111)	0.422(0.086)
3	0.500(0.083)	0.401(0.077)	0.379(0.061)	0.277(0.048)
4	0.508(0.088)	0.376(0.082)	0.383(0.062)	0.262(0.045)
5	1.188(0.411)	0.720(0.220)	0.871(0.227)	0.459(0.098)

Table 1.2: *Model 2 estimation accuracy comparison–RASE and its standard error in brackets.*

ϵ	$n = 200$		$n = 400$	
	LS	Adapt	LS	Adapt
1	0.362(0.077)	0.380(0.074)	0.263(0.051)	0.275(0.049)
2	0.618(0.301)	0.566(0.201)	0.431(0.129)	0.384(0.076)
3	0.412(0.091)	0.351(0.080)	0.290(0.059)	0.215(0.041)
4	0.407(0.102)	0.319(0.089)	0.291(0.061)	0.207(0.051)
5	1.133(0.397)	0.669(0.224)	0.828(0.224)	0.436(0.101)

1.3.2 Real-Data Applications

Example 1 (Hong Kong environmental data). We now illustrate the adaptive estimation method via an application to an environmental data set. The data were collected daily in Hong Kong from January 1, 1994, to December 31, 1995 and have been analyzed by Fan and Zhang (1999), Cai, et al. (2000), Xia, et al. (2002) and Fan and Zhang (2008). In this data

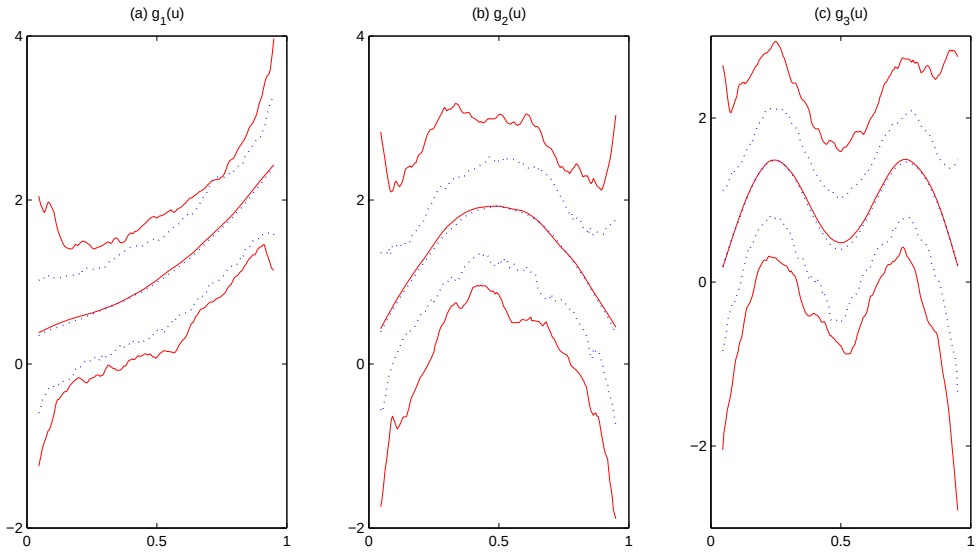

Figure 1.1: *Estimated coefficient functions with 95% pointwise confidence intervals (blue dotted line for Adapt and red solid line for LS) for model 1.*

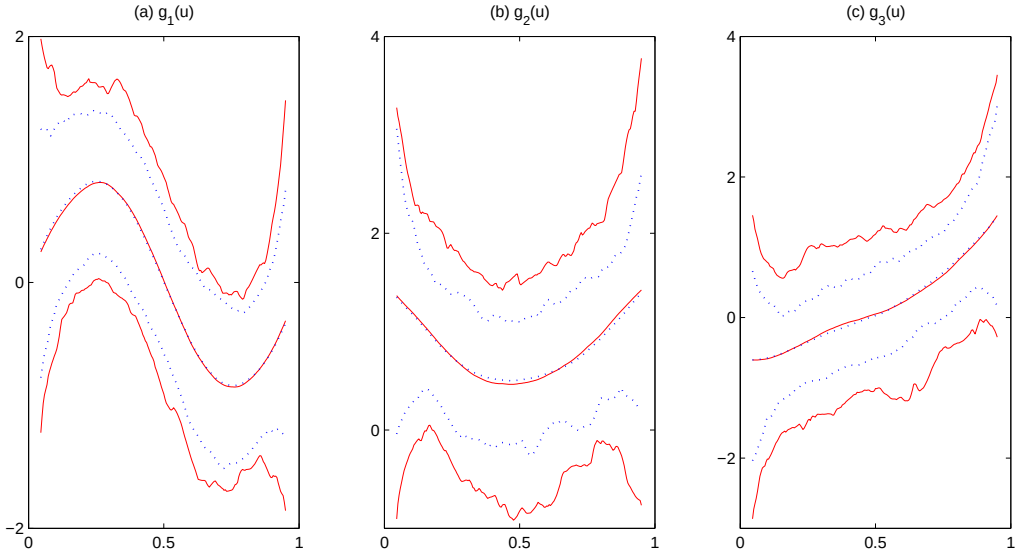

Figure 1.2: *Estimated coefficient functions with 95% pointwise confidence intervals (blue dotted line for Adapt and red solid line for LS) for model 2.*

set, a collection of daily measurements of pollutants and other environmental factors are included. Following Fan and Zhang (1999), we consider three pollutants: sulphur dioxide

x_2 (in $\mu g/m^3$), nitrogen dioxide x_3 (in $\mu g/m^3$), and respirable suspended particulates x_4 (in $\mu g/m^3$) (this variable is named as 'dust' in Fan and Zhang (1999), Fan and Zhang (2008), and Cai, et al. (2000)). The response variable is the logarithm of the number of daily hospital admissions y. We set $x_1 = 1$ as the intercept term and let u denote time which is scaled to the interval [0, 1]. As in the previous analyses, all three predictors are centered. The following varying coefficient model is considered to investigate the relationship between y and the levels of pollutants x_2, x_3, and x_4.

$$y = g_1(u) + g_2(u)x_2 + g_3(u)x_3 + g_4(u)x_4 + \epsilon.$$

We set aside 50 observations as testing set. The bandwidth h, selected by leave-one-out cross-validation, is around 0.146. The estimated coefficient functions together with 95% pointwise confidence intervals are depicted in Figure 1.3. We also compare the median squared prediction errors, MSPE $=$ Median$\{(y_j - \hat{y}_j)^2, j = 1, \ldots, k\}$, from our adaptive approach and the traditional least squares estimation, where $k = 50$ and $\hat{y}_j = \hat{g}_1(u_j) + \hat{g}_2(u_j)x_{j2} + \hat{g}_3(u_j)x_{j3} + \hat{g}_4(u_j)x_{j4}$. The MSPE from our adaptive approach is 0.0183, compared to 0.0178 from the LS estimation.

In Figure 1.5 (a), we give the residual QQ-plot for Hong Kong environmental data. From the plot, we can see that the residual is very close to normal, which explains why the MSPE of the adaptive approach is close to the MSPE of the LS estimation.

Example 2 (Boston housing data). The Boston Housing Data (corrected version (Gilley and Pace, 1996)), which has been analyzed by Fan and Huang (2005) and Wang and Xia (2009), is publicly available in the R package *mlbench*, (*http://cran.r-project.org/*). In this data set, the median value of owner-occupied homes in 506 U.S. census tracts in the Boston area in 1970 and some variables that might explain the variation of housing value are included. Following Fan and Huang (2005) and Wang and Xia (2009), we consider seven independent variables: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per $\$10,000$), NOX (nitric oxides concen-

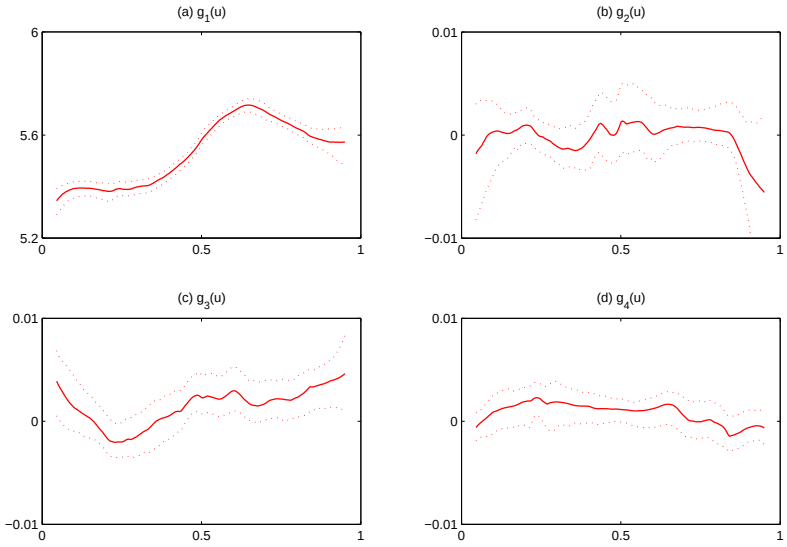

Figure 1.3: *Estimated coefficient functions (solid curves) with 95% pointwise confidence intervals (dotted curves) for Hong Kong environmental data.*

tration parts per 10 million), PTRATIO (pupil-teacher ratio by town), AGE (proportion of owner-occupied units built prior to 1940), and LSTAT (lower status of the population). The response variable is CMEDV (corrected median value of owner-occupied homes in USD 1000's). We denote the covariates CRIM, RM, TAX, NOX, PTRATIO, and AGE to be $x_2, x_3, \ldots, x_7$, respectively. We take $x_1 = 1$ as the intercept term and $u = \sqrt{\text{LSTAT}}$. By doing so, we can fit different regression models at different lower status population percentage (Fan and Huang, 2005). Following Fan and Huang (2005) we use the square root transformation on the index variable LSTAT to make the data symmetrically distributed. We construct the following varying coefficient model

$$y_i = g_1(u_i) + \sum_{j=2}^{7} g_j(u_i)x_{ij} + \epsilon_i.$$

Similar to the analysis in the previous example, we set aside 50 observations for checking prediction errors. The bandwidth h was selected by leave-one-out cross-validation, which is around 0.294. The estimated coefficient functions are depicted in Figure 1.4. From the

plot, we can see that the coefficient functions of x_2 (CRIM) and x_3 (RM) vary over time. The coefficient functions of x_4 (TAX), x_5 (NOX), and x_7 (AGE) are very close to zero and the coefficient function of x_6 (PTRATIO) shows no significant trend. These discoveries are consistent with those from Fan and Huang (2005) and Wang and Xia (2009). In terms of the median squared prediction error (MSPE), the MSPE from our adaptive approach is 0.0484, compared to 0.0604 from the LS estimation.

In Figure 1.5 (b), we give the residual QQ-plot for Boston housing data. Based on the tails of the QQ-plot, there is a clear deviation of the residuals from normal, which explains why the MSPE of the adaptive approach is much smaller than the MSPE of the LS estimation.

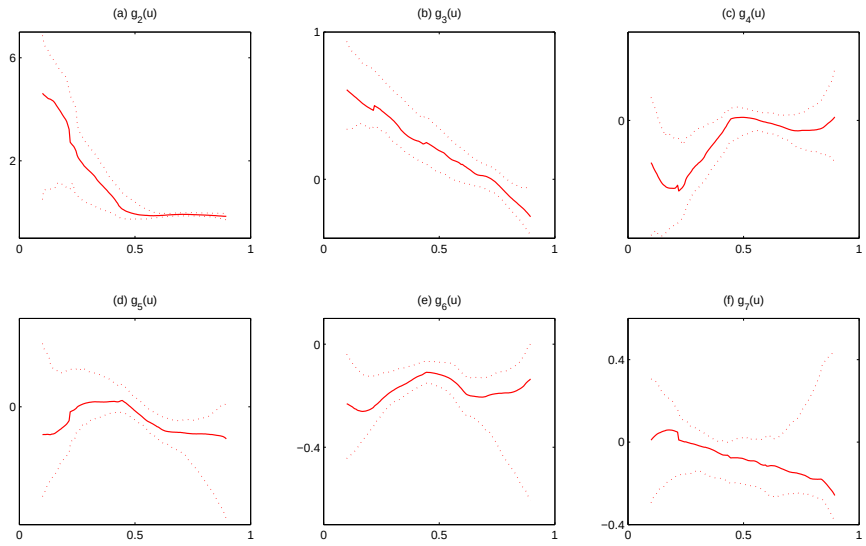

Figure 1.4: *Estimated coefficient functions (solid curves) with 95% pointwise confidence intervals (dotted curves) for Boston housing data.*

1.4 Discussion

In this chapter, we proposed an adaptive estimation for varying coefficient models. The new estimation procedure can adapt to different errors and thus provide a more efficient

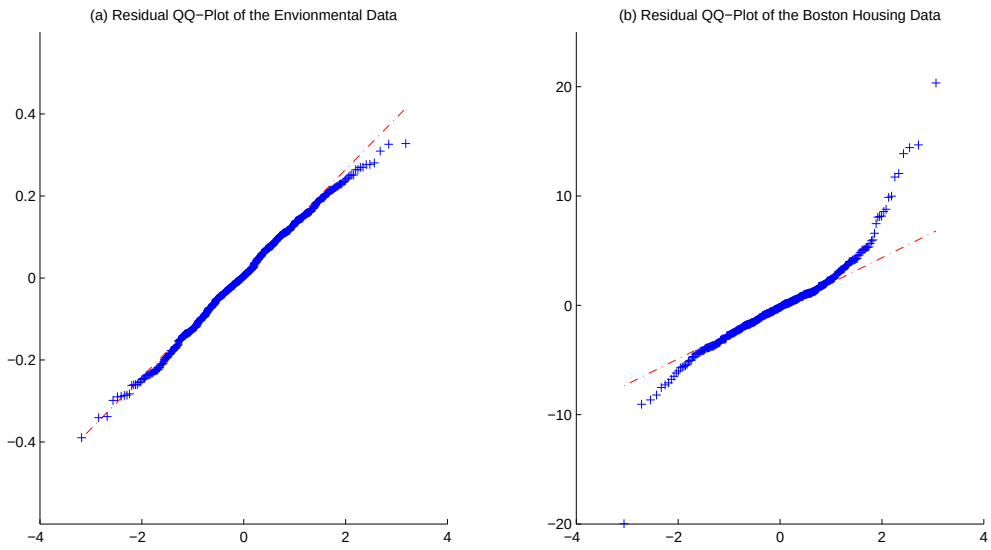

Figure 1.5: *Residual QQ-plot for two data examples: (a) Hong Kong environmental data; (b) Boston housing data.*

estimate than the traditional least squares based estimate. Simulation studies and two real data applications confirmed our theoretical findings.

It will be interesting to know whether we can also perform some adaptive hypothesis tests for the coefficient functions using the estimated error density. For example, we might be interested in testing some parametric assumptions, such as constant or zero, for the coefficient functions. It requires more research about whether the Wilks phenomenon for generalized likelihood ratio statistic proposed by Fan, et al. (2001) still holds for the proposed adaptive varying coefficient models.

The idea of the proposed adaptive estimator might also be generalized to many other models, such as varying coefficient partial linear models and nonparametric additive models. In addition, by combining this adaptive idea with shrinkage estimation, we can develop adaptive variable selection procedures. Such study is under way.

1.5 Proofs

We first impose some regularity conditions.

Conditions:

1. $K(\cdot)$ is bounded, symmetric, and has bounded support and bounded derivative;

2. $\{x_i\}_i$, $\{u_i\}_i$, $\{\epsilon_i\}_i$ are independent and identically distributed and $\{\epsilon_i\}_i$ is independent of $\{x_i\}_i$ and $\{u_i\}_i$. Additionally, the predictor $\mathbf{x}$ has a bounded support;

3. The probability distribution function $f(\cdot)$ of ϵ has bounded continuous derivatives up to order 4. Let $\rho(\epsilon) = \log f(\epsilon)$. Assume $\mathrm{E}[\rho^{'}(\epsilon_i)] = 0$, $\mathrm{E}[\rho^{''}(\epsilon_i)] < \infty$, $\mathrm{E}[\rho^{'}(\epsilon_i)^2] < \infty$ and $\rho^{'''}(\cdot)$ is bounded;

4. The marginal density of u has a continuous second derivative in some neighborhood of u_0 and $q(u_0) \neq 0$;

5. $h \to 0$, $nh \to \infty$ as $n \to \infty$ and $h_0 = h/\log(n)$;

6. $g_j(\cdot)$ has bounded, continuous 3^{rd} derivatives for $1 \leq j \leq d$.

These conditions are adopted from Fan and Zhang (1999) and Linton and Xiao (2007). They are not the weakest possible conditions. For instance, the independence of $\{\mathbf{x}_i\}_i$ and $\{\epsilon_i\}_i$ can be relaxed based on the discussion of Section 4 of Linton and Xiao (2007).

Proof of Theorem 1.2.1

Note that

$$
\begin{aligned}
&Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \\
&= \sum_{i=1}^{n} K_h(u_i - u_0) \log \left\{ \frac{\sum_{j \neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j \neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\} \\
&= \sum_{i=1}^{n} K_h(u_i - u_0) \log \sum_{j \neq i} \left(\frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j \neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right) \\
&\times \left(\frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right) \\
&= \sum_{i=1}^{n} K_h(u_i - u_0) \log \left\{ \sum_{j \neq i} p_{ij}^{(k+1)} \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\},
\end{aligned}
$$

where

$$
p_{ij}^{(k+1)} = \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \} x_{il} - \tilde{\epsilon}_j \right]}{\sum_{j \neq i} K_{h_0}\left[y_i - \sum_{l=1}^{d} \{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \} x_{il} - \tilde{\epsilon}_j \right]}.
$$

From the Jensen's inequality, we have

$$
\begin{aligned}
&Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \\
&\geqslant \sum_{i=1}^{n} K_h(u_i - u_0) \sum_{j \neq i} p_{ij}^{(k+1)} \log \left\{ \frac{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k+1)} + c_l^{(k+1)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]}{K_{h_0}\left[y_i - \sum_{l=1}^{d} \left\{ b_l^{(k)} + c_l^{(k)}(u_i - u_0) \right\} x_{il} - \tilde{\epsilon}_j \right]} \right\}.
\end{aligned}
$$

Based on the property of M-step of (1.9), we have $Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geq 0$. $\square$

Proof of Theorem 1.2.2

Note that the estimator $\hat{\boldsymbol{\theta}}$ is the maximizer of the following objective function

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} K_h(u_i - u_0) \log \tilde{f}\left[y_i - \sum_{l=1}^{d} \{b_l + c_l(u_i - u_0)\} x_{il} \right], \tag{1.13}$$

where

$$\tilde{f}(\epsilon_i) = \frac{1}{n} \sum_{j \neq i} K_{h_0}(\epsilon_i - \tilde{\epsilon}_j)$$

is the kernel density estimate of $f(\cdot)$, and $\tilde{\epsilon}_i$ is the residual based on the least squares local linear estimate. By the adaptive nonparametric regression result of Linton and Xiao (2007), the asymptotic result of $\hat{\boldsymbol{\theta}}$ in (1.13) is the same whether the true density $f(\cdot)$ is used or not. Therefore, we will mainly proof the existence and asymptotic distribution of $\hat{\boldsymbol{\theta}}$ assuming $f(\cdot)$ is known.

We will first prove that with probability approaching 1, there exists a consistent local maximizer $\hat{\boldsymbol{\theta}} = (\hat{b}_1, \ldots, \hat{b}_d, \hat{c}_1, \ldots, \hat{c}_d)^T$ of (1.7) such that

$$\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p\{(nh)^{-1/2} + h^2\}.$$

Then we establish the asymptotic distributions for such consistent estimate.

Denote $\boldsymbol{\theta}^* = \mathbf{H}\boldsymbol{\theta}$, $\mathbf{x}_i^* = (x_{i1}, x_{i2}, ..., x_{id}, (\frac{u_i - u_0}{h})x_{i1}, ..., (\frac{u_i - u_0}{h})x_{id})^T$, $K_i = K_h(u_i - u_0)$, $R(u_i, \mathbf{x}_i) = \sum_{j=1}^{d} g_j(u_i)x_{ij} - \sum_{j=1}^{d}[b_j + c_j(u_i - u_0)]x_{ij}$, and $a_n = (nh)^{-1/2} + h^2$. Let $\rho(\cdot) = \log f(\cdot)$, we have the objective function

$$L(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K_i \rho(y_i - \boldsymbol{\theta}^{*T} \mathbf{x}_i^*) = L(\boldsymbol{\theta}^*).$$

It is sufficient to show that for any given $\eta > 0$, there exists a large constant c such that

$$P\left\{ sup_{\|\mu\|=c} L(\boldsymbol{\theta}^* + a_n\mu) < L(\boldsymbol{\theta}^*) \right\} \geq 1 - \eta,$$

where μ has the same dimension as $\boldsymbol{\theta}$, a_n is the convergence rate. By using Taylor expansion, it follows that

$$\begin{aligned} L(\boldsymbol{\theta}^* + a_n\mu) - L(\boldsymbol{\theta}^*) &= \frac{1}{n}\sum_{i=1}^{n} K_i\{\rho(\epsilon_i + R(u_i,\mathbf{x}_i) - a_n\mu^T\mathbf{x}_i^*) - \rho(\epsilon_i + R(u_i,\mathbf{x}_i))\} \\ &= -\frac{1}{n}\sum_{i=1}^{n} K_i\rho^{'}(\epsilon_i + R(u_i,\mathbf{x}_i))a_n\mu^T\mathbf{x}_i^* + \frac{1}{2n}\sum_{i=1}^{n} K_i\rho^{''}(\epsilon_i + R(u_i,\mathbf{x}_i))a_n^2(\mu^T\mathbf{x}_i^*)^2 \\ &\quad -\frac{1}{6n}\sum_{i=1}^{n} K_i\rho^{'''}(z_i)a_n^3(\mu^T\mathbf{x}_i^*)^3 \\ &\triangleq I_1 + I_2 + I_3, \end{aligned}$$

where z_i is a value between $\epsilon_i + R(u_i,\mathbf{x}_i) - a_n\mu^T\mathbf{x}_i^*$ and $\epsilon_i + R(u_i,\mathbf{x}_i)$. For $I_1 = -\frac{1}{n}\sum_{i=1}^{n} K_i\rho^{'}(\epsilon_i + R(u_i,\mathbf{x}_i))a_n\mu^T\mathbf{x}_i^*$, $\mathrm{E}(I_1) = -\mathrm{E}\left[K_i\rho^{'}(\epsilon_i + R(u_i,\mathbf{x}_i))a_n\mu^T\mathbf{x}_i^*\right]$. By using Taylor expansion,

$$\rho^{'}(\epsilon_i + R(u_i,\mathbf{x}_i)) \approx \rho^{'}(\epsilon_i) + \rho^{''}(\epsilon_i)R(u_i,\mathbf{x}_i) + \frac{1}{2}\rho^{'''}(\epsilon_i)R^2(u_i,\mathbf{x}_i).$$

Based on the assumption that ϵ is independent of u and $\mathbf{x}$, and $\mathrm{E}[\rho^{'}(\epsilon_i)] = 0$, we have

$$\mathrm{E}(I_1) \approx -a_n\mathrm{E}\left\{K_i\left[\rho^{''}(\epsilon_i)R(u_i,\mathbf{x}_i) + \frac{1}{2}\rho^{'''}(\epsilon_i)R^2(u_i,\mathbf{x}_i)\right]\mu^T\mathbf{x}_i^*\right\}.$$

Since

$$\begin{aligned} R(u_i,\mathbf{x}_i) &= \textstyle\sum_{j=1}^{d} g_j(u_i)x_{ij} - \sum_{j=1}^{d}[b_j + c_j(u_i - u_0)]x_{ij} \\ &= \sum_{j=1}^{d}[\sum_{m=2}^{\infty}\frac{1}{m!}g_j^{(m)}(u_0)(u_i - u_0)^m]x_{ij} \\ &= O_p(h^2), \end{aligned}$$

then $\frac{1}{2}\rho^{'''}(\epsilon_i)R^2(u_i,\mathbf{x}_i) = [O_p(h^2)]^2 = O_p(h^4)$, which is a smaller order than $\rho^{''}(\epsilon_i)R(u_i,\mathbf{x}_i)$.

Thus,

$$\mathrm{E}(I_1) \approx -a_n \mathrm{E}\left\{K_i \rho^{''}(\epsilon_i) R(u_i, \mathbf{x}_i) \mu^T \mathbf{x}_i^*\right\} = -a_n \mathrm{E}\left[\rho^{''}(\epsilon_i)\right] \mathrm{E}\left[K_i R(u_i, \mathbf{x}_i) \mu^T \mathbf{x}_i^*\right].$$

Since $\delta_1 = \mathrm{E}\left\{\rho^{''}(\epsilon_i)\right\}$, then

$$\mathrm{E}(I_1) \approx -a_n \delta_1 \mathrm{E}\left[K_i R(u_i, \mathbf{x}_i) \mu^T \mathbf{x}_i^*\right] = -a_n \delta_1 \mathrm{E}\left\{\mathrm{E}\left\{R(u_i, \mathbf{x}_i) \mu^T \mathbf{x}_i^* | u_i\right\} K_i\right\}.$$

By $\mu^T \mathbf{x}_i^* \leq \|\mu\| \cdot \|\mathbf{x}_i^*\| = c \|\mathbf{x}_i^*\|$, we have $\mathrm{E}(I_1) = O(a_n c h^2)$.

$$\mathrm{var}(I_1) = \frac{1}{n} \mathrm{var}\left\{K_i \rho^{'}(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n \mu^T \mathbf{x}_i^*\right\} = \frac{1}{n}\{\mathrm{E}(A^2) - [\mathrm{E}(A)]^2\},$$

where $A = K_i \rho^{'}(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n \mu^T \mathbf{x}_i^*$. Since $\delta_2 = \mathrm{E}\left\{\rho^{'}(\epsilon_i)^2\right\}$, then

$$\begin{aligned}
\mathrm{E}(A^2) &= \mathrm{E}\left\{K_i^2 \rho^{'}(\epsilon_i + R(u_i, \mathbf{x}_i))^2 a_n^2 (\mu^T \mathbf{x}_i^*)^2\right\} \\
&\approx a_n^2 \mathrm{E}\left\{K_i^2 \rho^{'}(\epsilon_i)^2 (\mu^T \mathbf{x}_i^*)^2\right\} \\
&= a_n^2 \delta_2 \mathrm{E}\left\{\mathrm{E}\left\{(\mu^T \mathbf{x}_i^*)^2 | u_i\right\} K_i^2\right\} \\
&= O\left(a_n^2 c^2 \frac{1}{h}\right).
\end{aligned}$$

Note that $[\mathrm{E}(A)]^2 = [O(a_n c h^2)]^2 \ll \mathrm{E}(A^2)$, then $\mathrm{var}(I_1) \approx \frac{1}{n}\mathrm{E}(A^2) = O\left(a_n^2 c^2 \frac{1}{nh}\right)$. Hence, $I_1 = \mathrm{E}(I_1) + O_p(\sqrt{\mathrm{var}(I_1)}) = O_p(a_n c h^2) + O_p\left(\sqrt{a_n^2 c^2 \frac{1}{nh}}\right) = O_p(c a_n^2)$. For

$$I_2 = \frac{1}{2n} \sum_{i=1}^{n} K_i \rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i)) a_n^2 (\mu^T \mathbf{x}_i^*)^2,$$

we have

$$\begin{aligned}
\mathrm{E}(I_2) &= \frac{1}{2}a_n^2\mathrm{E}\left\{K_i\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))(\mu^T\mathbf{x}_i^*)^2\right\} \\
&= \frac{1}{2}a_n^2\mathrm{E}\left\{\rho^{''}(\epsilon_i)K_i(\mu^T\mathbf{x}_i^*)^2\right\}(1+o(1)) \\
&= \frac{1}{2}a_n^2\delta_1\mathrm{E}\left\{\mathrm{E}\left\{\mu^T\mathbf{x}_i^*\mathbf{x}_i^{*T}\mu|u_i\right\}K_i\right\}(1+o(1)) \\
&= \frac{1}{2}a_n^2\delta_1\mu^T\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i\right\}\mu(1+o(1)).
\end{aligned}$$

Note that $\mathbf{x}_i^*\mathbf{x}_i^{*T} = \left(x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l\right)_{1\le j,k\le d, l=0,1,2}$ and $\Gamma_{jk}(u_i) = \mathrm{E}(x_{ij}x_{ik}|u_i)$ for $1 \le j,k \le d$, then

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l|u_i\right\}K_i\right\} &= \mathrm{E}\left\{\mathrm{E}(x_{ij}x_{ik}|u_i)\left(\frac{u_i-u_0}{h}\right)^l K_i\right\} \\
&= \mathrm{E}\left\{\Gamma_{jk}(u_i)\left(\frac{u_i-u_0}{h}\right)^l K_i\right\}.
\end{aligned}$$

By using Taylor expansion, we obtain

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l|u_i\right\}K_i\right\} &= \frac{1}{h}\int\Gamma_{jk}(u_i)\left(\frac{u_i-u_0}{h}\right)^l K(\frac{u_i-u_0}{h})q(u_i)\mathrm{d}u_i \\
&= q(u_0)\Gamma_{jk}(u_0)\int t^l K(t)\mathrm{d}t(1+o(1)).
\end{aligned}$$

So we have

$$\mathrm{E}(I_2) = \frac{1}{2}a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu(1+o(1)),$$

where $\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \otimes \mathbf{\Gamma}(u_0)$ is a $2d \times 2d$ matrix. Thus,

$$\mathrm{E}(I_2) = O(a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu)$$

and

$$\begin{aligned}\text{var}(I_2) &= \frac{a_n^4}{4n}\text{var}\left[\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))K_i(\mu^T\mathbf{x}_i^*)^2\right] \\ &= \frac{a_n^4}{4n}\left\{\text{E}(B^2) - [\text{E}(B)]^2\right\},\end{aligned}$$

where $B = \rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))K_i(\mu^T\mathbf{x}_i^*)^2$. Let $\delta_3 = \text{E}\left(\rho^{''}(\epsilon_i)^2\right)$, then

$$\begin{aligned}\text{E}(B^2) &= \text{E}\left\{\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))^2K_i^2(\mu^T\mathbf{x}_i^*)^4\right\} \\ &\approx \text{E}\left\{\rho^{''}(\epsilon_i)^2K_i^2(\mu^T\mathbf{x}_i^*)^4\right\} \\ &= \delta_3\text{E}\left\{K_i^2(\mu^T\mathbf{x}_i^*)^4\right\} \\ &= O\left(\frac{1}{h}\right).\end{aligned}$$

Note that $[\text{E}(B)]^2 = [O(1)]^2 = O(1) \ll \text{E}(B^2)$, so $\text{var}(I_2) = O\left(\frac{a_n^4}{nh}\right)$. Based on the result $I_2 = \text{E}(I_2) + O_p(\sqrt{\text{var}(I_2)})$ and the assumption $nh \to \infty$, it follows that

$$I_2 = a_n^2\delta_1 q(u_0)\mu^T\mathbf{S}\mu(1 + o_p(1)).$$

Similarly, $I_3 = -\frac{1}{6n}\sum_{i=1}^n K_i\rho^{'''}(z_i)a_n^3(\mu^T\mathbf{x}_i^*)^3 = O_p(a_n^3)$.

Assume $\delta_1 < 0$. Noticing that $\mathbf{S}$ is a positive matrix, $\|\mu\| = c$, we can choose c large enough such that I_2 dominates both I_1 and I_3 with probability at least $1 - \eta$. Thus $P\left\{sup_{\|\mu\|=c}L(\boldsymbol{\theta}^* + a_n\mu) < L(\boldsymbol{\theta}^*)\right\} \geq 1 - \eta$. Hence with probability approaching 1, there exists a local maximizer $\hat{\boldsymbol{\theta}}^*$ such that $\left\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\right\| \leq a_n c$, where $a_n = (nh)^{-1/2} + h^2$. Based on the definition of $\boldsymbol{\theta}^*$, we can get, with probability approaching 1, $H(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_p((nh)^{-1/2} + h^2)$.

Next, we provide the asymptotic distribution for such consistent estimate. Since $\hat{\boldsymbol{\theta}}$

maximizes $L(\boldsymbol{\theta})$, then $L^{'}(\hat{\boldsymbol{\theta}}) = 0$. By Taylor expansion,

$$0 = L^{'}(\hat{\boldsymbol{\theta}}) = L^{'}(\boldsymbol{\theta}_0) + L^{''}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \frac{1}{2}L^{'''}(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2,$$

where $\tilde{\boldsymbol{\theta}}$ is a value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = -[L^{''}(\boldsymbol{\theta}_0)]^{-1}L^{'}(\boldsymbol{\theta}_0)(1 + o_p(1))$. Since $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n} K_i\rho(y_i - \boldsymbol{\theta}^{*T}\mathbf{x}_i^*)$ and $y_i - \boldsymbol{\theta}^{*T}\mathbf{x}_i^* = \epsilon_i + R(u_i, \mathbf{x}_i)$, then $L^{''}(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n} K_i\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))\mathbf{x}_i^*\mathbf{x}_i^{*T}$. We have the following expectation,

$$\begin{aligned}
\mathrm{E}[L^{''}(\boldsymbol{\theta}^*)] &= \mathrm{E}\left\{\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))K_i\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} \\
&\approx \mathrm{E}\left\{\rho^{''}(\epsilon_i)K_i\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} \\
&= \delta_1\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i\right\} \\
&= \delta_1 q(u_0)\mathbf{S}(1 + o(1)).
\end{aligned}$$

Throughout this chapter, we consider the element-wise variance of a matrix,

$$\mathrm{var}[L^{''}(\boldsymbol{\theta}^*)] = \frac{1}{n}\mathrm{var}\left\{K_i\rho^{''}(\epsilon_i + R(u_i, \mathbf{x}_i))\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} = O\left(\frac{1}{nh}\right).$$

Based on the result $L^{''}(\boldsymbol{\theta}^*) = \mathrm{E}[L^{''}(\boldsymbol{\theta}^*)] + O_p(\sqrt{\mathrm{var}[L^{''}(\boldsymbol{\theta}^*)]})$ and the assumption $nh \to \infty$, it follows that

$$L^{''}(\boldsymbol{\theta}^*) = \delta_1 q(u_0)\mathbf{S}(1 + o_p(1)).$$

For $L^{'}(\boldsymbol{\theta}^*)$, we can divide it into two parts.

$$\begin{aligned}
L^{'}(\boldsymbol{\theta}^*) &\approx -\frac{1}{n}\sum_{i=1}^{n} K_i\rho^{'}(\epsilon_i)\mathbf{x}_i^* - \frac{1}{n}\sum_{i=1}^{n} K_i\rho^{''}(\epsilon_i)R(u_i, \mathbf{x}_i)\mathbf{x}_i^* \\
&\triangleq -\mathbf{w}_n - \boldsymbol{\nu}_n.
\end{aligned}$$

The asymptotic result is determined by $\mathbf{w}_n$. In order to find the order of $\boldsymbol{\nu}_n$, we compute the following things.

$$\mathrm{E}(\boldsymbol{\nu}_n) = \mathrm{E}\left[K_i \rho^{''}(\epsilon_i) R(u_i, \mathbf{x}_i)\mathbf{x}_i^*\right] = \delta_1 \mathrm{E}\left\{\mathrm{E}\left\{R(u_i, \mathbf{x}_i)\mathbf{x}_i^* | u_i\right\} K_i\right\}.$$

Since $g_j^{'''}(\cdot)$ is bounded, then we have

$$R(u_i, \mathbf{x}_i) = \sum_{j=1}^{d}\left\{\sum_{m=2}^{\infty}\frac{1}{m!}g_j^{(m)}(u_0)(u_i-u_0)^m\right\}x_{ij} = \sum_{j=1}^{d}\frac{1}{2}g_j^{''}(u_0)(u_i-u_0)^2 x_{ij}(1+o_p(1)).$$

By $\mathbf{x}_i^* = (x_{i1}, ..., x_{id}, (\frac{u_i-u_0}{h})x_{i1}, ..., (\frac{u_i-u_0}{h})x_{id})^T$,

$$R(u_i, \mathbf{x}_i)\mathbf{x}_i^* \approx \left[\left(\frac{(u_i-u_0)^2}{2}\left\{\sum_{j=1}^{d} g_j^{''}(u_0)x_{ij}\right\}x_{ik}\right)_{1\le k\le d}, \left(\frac{(u_i-u_0)^3}{2h}\left\{\sum_{j=1}^{d} g_j^{''}(u_0)x_{ij}\right\}x_{ik}\right)_{1\le k\le d}\right]^T_{2d\times 1}.$$

Since

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{E}\left\{\left[\sum_{j=1}^{d} g_j^{''}(u_0)x_{ij}\right]x_{ik}|u_i\right\}\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \mathrm{E}\left\{\sum_{j=1}^{d} g_j^{''}(u_0)\mathrm{E}(x_{ij}x_{ik}|u_i)\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \sum_{j=1}^{d} g_j^{''}(u_0)\mathrm{E}\left\{\Gamma_{jk}(u_i)\frac{(u_i-u_0)^2}{2}K_i\right\}\\
&= \frac{h^2}{2}q(u_0)\sum_{j=1}^{d} g_j^{''}(u_0)\Gamma_{jk}(u_0)\int t^2 K(t)\mathrm{d}t(1+o(1))
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{E}\left\{\left[\sum_{j=1}^{d} g_j^{''}(u_0)x_{ij}\right]x_{ik}|u_i\right\}\frac{(u_i-u_0)^3}{2h}K_i\right\}\\
&= \mathrm{E}\left\{\sum_{j=1}^{d} g_j^{''}(u_0)\Gamma_{jk}(u_i)\frac{(u_i-u_0)^3}{2h}K_i\right\}\\
&= \frac{h^2}{2}q(u_0)\sum_{j=1}^{d} g_j^{''}(u_0)\Gamma_{jk}(u_0)\int t^3 K(t)\mathrm{d}t(1+o(1)),
\end{aligned}$$

then

$$\mathrm{E}(\boldsymbol{\nu}_n) = \delta_1 q(u_0)\frac{h^2}{2}\sum_{j=1}^{d} g_j^{''}(u_0)\boldsymbol{\psi}_j(1+o(1)),$$

where $\boldsymbol{\psi}_j = \begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix} \otimes (\Gamma_{jk}(u_0))_{1\le k\le d}^{T}$ is a $2d\times 1$ vector for $j=1,...,d$. Since $\mathrm{var}(\boldsymbol{\nu}_n) = \mathrm{var}\left\{K_i\rho^{''}(\epsilon_i)R(u_i,\mathbf{x}_i)\mathbf{x}_i^*\right\}/n = O(h^3/n)$, then based on the result $\boldsymbol{\nu}_n = \mathrm{E}(\boldsymbol{\nu}_n)+O_p(\sqrt{\mathrm{var}(\boldsymbol{\nu}_n)})$ and the assumption $nh\to\infty$, it follows that

$$\boldsymbol{\nu}_n = \delta_1 q(u_0)\frac{h^2}{2}\sum_{j=1}^{d} g_j^{''}(u_0)\boldsymbol{\psi}_j(1+o_p(1)).$$

Then

$$\begin{aligned}\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* =& -\left[L^{''}(\boldsymbol{\theta}^*)\right]^{-1}L^{'}(\boldsymbol{\theta}^*)(1+o_p(1))\\ =& -\left[\delta_1 q(u_0)\mathbf{S}\right]^{-1}(-\mathbf{w}_n-\boldsymbol{\nu}_n)(1+o_p(1))\\ =& \frac{\mathbf{S}^{-1}\mathbf{w}_n}{\delta_1 q(u_0)}(1+o_p(1)) + \mathbf{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j^{''}(u_0)\boldsymbol{\psi}_j(1+o_p(1)). \qquad (1.14)\end{aligned}$$

Based on the assumption $\mathrm{E}[\rho^{'}(\epsilon_i)]=0$, we can easily get $\mathrm{E}(\mathbf{w}_n)=0$.

$$\mathrm{var}(\mathbf{w}_n) = \frac{1}{n}\mathrm{var}\left\{K_i\rho^{'}(\epsilon_i)\mathbf{x}_i^*\right\} = \frac{1}{n}\mathrm{E}\left\{K_i^2\rho^{'}(\epsilon_i)^2\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\} = \frac{1}{n}\delta_2\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i^2\right\}.$$

Since $\mathbf{x}_i^*\mathbf{x}_i^{*T} = \left(x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l\right)_{1\le j,k\le d, l=0,1,2}$ and

$$\begin{aligned}\mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}\left(\frac{u_i-u_0}{h}\right)^l|u_i\right\}K_i^2\right\} &= \mathrm{E}\left\{\mathrm{E}\left\{x_{ij}x_{ik}|u_i\right\}\left(\frac{u_i-u_0}{h}\right)^l K_i^2\right\}\\ &= \mathrm{E}\left\{\Gamma_{jk}(u_i)\left(\frac{u_i-u_0}{h}\right)^l K_i^2\right\}\\ &= \frac{1}{h}q(u_0)\Gamma_{jk}(u_0)\int t^l K^2(t)\mathrm{d}t(1+o(1)),\end{aligned}$$

then

$$\mathrm{E}\left\{\mathrm{E}\left\{\mathbf{x}_i^*\mathbf{x}_i^{*T}|u_i\right\}K_i^2\right\} = \frac{1}{h}q(u_0)\mathbf{\Lambda}(1+o(1)),$$

where $\mathbf{\Lambda} = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \mathbf{\Gamma}(u_0)$ is a $2d \times 2d$ matrix. So

$$\mathrm{var}(\mathbf{w}_n) = \frac{1}{nh}\delta_2 q(u_0)\mathbf{\Lambda}(1+o(1)).$$

We next use the Lyapunov central limit theorem to obtain the asymptotic distribution of $\mathbf{w}_n$. The Lyapunov conditions are checked as follows. For any unit vector $\mathbf{d} \in \mathbb{R}^{2d}$, let $\mathbf{d}^T\mathbf{w}_n = \sum_{i=1}^n \xi_i$, where $\xi_i = \frac{1}{n}K_i\rho'(\epsilon_i)\mathbf{d}^T\mathbf{x}_i^*$. Since

$$\mathrm{E}(\xi_i^2) = \mathrm{E}\left\{\frac{1}{n^2}K_i^2\rho'(\epsilon_i)^2\mathbf{d}^T\mathbf{x}_i^*\mathbf{x}_i^{*T}\mathbf{d}\right\} = \frac{1}{n^2}\delta_2\mathbf{d}^T\mathrm{E}\left\{K_i^2\mathbf{x}_i^*\mathbf{x}_i^{*T}\right\}\mathbf{d} = \frac{1}{n^2h}\delta_2 q(u_0)\mathbf{d}^T\mathbf{\Lambda}\mathbf{d}(1+o(1)),$$

then $\left(\sum_{i=1}^n \mathrm{E}\left|\xi_i\right|^2\right)^3 = O\left(\left(\frac{1}{nh}\right)^3\right)$. Let $\delta_4 = \mathrm{E}\left\{\rho'(\epsilon_i)^3\right\}$, then

$$\mathrm{E}(\xi_i^3) = \mathrm{E}\left\{\frac{1}{n^3}K_i^3\rho'(\epsilon_i)^3(\mathbf{d}^T\mathbf{x}_i^*)^3\right\} = \frac{1}{n^3}\delta_3\mathrm{E}\left\{K_i^3(\mathbf{d}^T\mathbf{x}_i^*)^3\right\} = O(\frac{1}{n^3h^2}).$$

So $\left(\sum_{i=1}^n \mathrm{E}\left|\xi_i\right|^3\right)^2 = O\left(\left(\frac{1}{n^2h^2}\right)^2\right)$. Since $\left(\frac{1}{n^2h^2}\right)^2(nh)^3 = \frac{1}{nh} \to 0$, then $\left(\frac{1}{n^2h^2}\right)^2 = o\left(\left(\frac{1}{nh}\right)^3\right)$, which is equivalent to

$$\left(\sum_{i=1}^n \mathrm{E}\left|\xi_i\right|^3\right)^2 = o\left(\left(\sum_{i=1}^n \mathrm{E}\left|\xi_i\right|^2\right)^3\right).$$

Based on Lyapunov Central Limit Theorem,

$$\frac{\mathbf{w}_n}{\sqrt{\mathrm{var}(\mathbf{w}_n)}} \xrightarrow{D} N(\mathbf{0}_{2d}, \mathbf{I}_{2d}),$$

where $\mathbf{0}_{2d}$ is a $2d \times 1$ vector with each entry being 0; $\mathbf{I}_{2d}$ is a $2d \times 2d$ identity matrix. Pre-

viously, we already computed that $\text{var}(\mathbf{w}_n) = \frac{1}{nh}\delta_2 q(u_0)\mathbf{\Lambda}(1+o(1))$, by Slutsky's Theorem,

$$\sqrt{nh}\mathbf{w}_n \xrightarrow{D} N(\mathbf{0}_{2d}, \delta_2 q(u_0)\mathbf{\Lambda}).$$

Based on (1.14), we have the following result

$$\sqrt{nh}\left\{\mathbf{H}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) - \mathbf{S}^{-1}\frac{h^2}{2}\sum_{j=1}^{d} g_j^{''}(u_0)\boldsymbol{\psi}_j(1+o_p(1))\right\} \xrightarrow{D} N(\mathbf{0}_{2d}, \delta_1^{-2}\delta_2 q(u_0)^{-1}\mathbf{S}^{-1}\mathbf{\Lambda}\mathbf{S}^{-1}).$$

Chapter 2

Unified Inference for Sparse and Dense Longitudinal Data in Time-Varying Coefficient Models

2.1 Introduction

Longitudinal data sets arise in biostatistics and life-time testing problems when the responses of the individuals are recorded repeatedly over a period of time. Examples can be found in clinical trials, follow-up studies for monitoring disease progression, and observational cohort studies. In many longitudinal studies, repeated measurements of the response variable are collected at irregular and possibly subject-specific time points. Therefore, the measurements within each subject are possibly correlated with each other and data are often highly unbalanced, but different subjects can be assumed to be independent. Typically, the scientific interest is either in the pattern of change over time of the outcome measures or more simply in the dependence of the outcome on the covariates.

A useful nonparametric model to quantify the influence of covariates other than time is the time-varying coefficient model, in which coefficients are allowed to change smoothly

over time. Let $\{(y_{ij}, \mathbf{x}_i(t_{ij}), t_{ij}); i = 1, 2, ..., n; j = 1, 2, ..., n_i\}$ be a longitudinal sample from n randomly selected subjects, where t_{ij} is the time when the jth measurement of the ith subject is made, n_i is the number of repeated measurements of the ith subject, y_{ij} is the response, and $\mathbf{x}_i(t_{ij}) = \mathbf{x}_{ij} = (x_i^0(t_{ij}), x_i^1(t_{ij}), ..., x_i^k(t_{ij}))^T$ are the $(k+1)$-dimensional covariates for the ith subject at time t_{ij}. The total number of observations in this sample is $N = \sum_{i=1}^{n} n_i$. The time-varying coefficient model can be written as

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}), \tag{2.1}$$

where $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), ..., \beta_k(t))^T$ for all $t \geqslant 0$ are smooth functions of t, $\epsilon_i(t)$ is a realization of a zero-mean stochastic process $\epsilon(t)$, and $\mathbf{x}_{ij}$ and ϵ_i are independent. It allows the time-varying intercept to exist when $x^0(t) \equiv 1$.

To better account for the local correlation structure of the longitudinal data, similar to the nonparametric mixed-effects model used by Wu and Zhang (2002) and Kim and Zhao (2013), we add a subject-specific random trajectory $v_i(\cdot)$ to model (2.1) and consider the following *mixed-effects time-varying coefficient model*

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) + v_i(t_{ij}) + \sigma(t_{ij})\epsilon_{ij}, \tag{2.2}$$

where $v_i(t)$ is considered realizations of a mean 0 process with a covariance function $\gamma(t, t') = \text{cov}\{v_i(t), v_i(t')\} = \text{E}[v_i(t)v_i(t')]$, ϵ_{ij} are errors with $\text{E}(\epsilon_{ij}) = 0$ and $\text{E}(\epsilon_{ij}^2) = 1$, and $v_i(t)$ and ϵ_{ij} are assumed to be independent. Our primary goal in this chapter is to estimate the varying coefficients $\boldsymbol{\beta}(t)$ and construct confidence intervals for them.

Longitudinal data can be identified as sparse or dense according to the number of measurements within each subject. Statistical analyses for sparse or dense longitudinal data have been a subject of intense investigation in the recent ten years. Please see, for example, Yao, et al. (2005) and Ma, et al. (2012) for the studies of the sparse longitudinal data when n_i is assumed to be bounded or follow a given distribution with $\text{E}(n_i) < \infty$; and see, for

example, Fan and Zhang (2000), Zhang and Chen (2007), Degras (2011), and Cao, et al. (2012) for the studies of the dense longitudinal data when $n_i \to \infty$.

It is known that the boundary between sparse and dense cases is not always clear in practice. Researchers may classify the same data set differently and therefore, a subjective choice between the sparse and dense cases might pose challenges for statistical inference. Hoover, et al. (1998), Wu and Chiang (2000), Chiang, et al. (2001), and Huang, et al. (2002) established some asymptotic bias and variance of their proposed estimates under some general conditions. However, the established limiting variances contain some unknown functions, which are not easy to estimate. Therefore, the bootstrap procedures were used to evaluate the variability of their proposed estimates. Li and Hsing (2010) established a uniform convergence rate for weighted local linear estimation of mean and variance functions for functional/longitudinal data. Nevertheless, Kim and Zhao (2013) showed that the convergence rates and limiting variances under sparse and dense assumptions are different. This motivated them to develop some unified nonparametric approaches that can be used to conduct longitudinal data analysis without deciding whether the data are dense or sparse. However, Kim and Zhao (2013) only considered estimating the mean response curve without the presence of covariates effect.

In this chapter, we use the mixed-effects time-varying coefficient model (2.2) to take the covariates other than time into account. The model considered by Kim and Zhao (2013) is a special case of ours if $\mathbf{x}_{ij} = 1$. We show that when using kernel smoothing method to estimate the smoothing functions for sparse or dense longitudinal data, the asymptotic results of these two situations are essentially different. Therefore, a subjective choice between the sparse and dense cases might lead to wrong conclusions for statistical inference. In order to solve this problem, motivated by Kim and Zhao (2013), we establish a unified self-normalized central limit theorem, based on which a unified inference is proposed that can adapt to both sparse and dense cases. The resulting unified confidence interval does not depend on any unknown quantity other than the point estimator $\boldsymbol{\beta}(t)$ and thus is simple to

use in practice. The effectiveness of the proposed unified inference is demonstrated through a simulation study and an analysis of an acquired immune deficiency syndrome (AIDS) data set.

This chapter is organized as follows. In Section 2.2, we first introduce a sample-size weighted local constant estimator of the smoothing functions $\boldsymbol{\beta}(t)$ and provide the asymptotic properties for both sparse and dense longitudinal data. Under the mixed-effects time-varying coefficient model setting, we then propose a unified convergence theory based on a self-normalization technique. In Section 2.3, we provide numerical results from a simulation study and use the AIDS data to demonstrate the performance of the proposed unified approach. Section 2.4 contains some discussion. Regularity conditions and proofs are assembled in Section 2.5.

2.2 A Unified Approach for Longitudinal Data

2.2.1 Estimation Method

Hoover, et al. (1998) proposed a local constant fit for the time-varying coefficient model. However, they did not consider the effect of repeated measurements for each subject. Similar to Li and Hsing (2010), we consider a sample-size weighted local constant estimation method for the model (2.2). Let $f(\cdot)$ be the density function of t_{ij} and let t be an interior point of the support of $f(\cdot)$. The weighted local constant estimator we consider is

$$\hat{\boldsymbol{\beta}}(t) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}(t) \right]^2 K\left(\frac{t_{ij} - t}{h} \right) = \mathbf{H}_n^{-1} \mathbf{g}_n, \tag{2.3}$$

where $K(\cdot)$ is a kernel function which is symmetric about 0 and satisfies $\int_{\mathbb{R}} K(u) \mathrm{d}u = 1$ and

$h > 0$ is a bandwidth, with

$$\mathbf{H}_n = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}^T K(\frac{t_{ij} - t}{h}), \quad \mathbf{g}_n = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} y_{ij} K(\frac{t_{ij} - t}{h}). \tag{2.4}$$

2.2.2 Asymptotic Properties for Sparse and Dense Longitudinal Data

Kim and Zhao (2013) specified the sparse and dense cases clearly. Here we adopt their assumptions for the number of repeated measurements of each subject under these two scenarios:

- Sparse longitudinal data: $n_1, n_2, ..., n_n$ are independent and identically distributed positive-integer-valued random variables with $\mathrm{E}(n_i) < \infty$;
- Dense longitudinal data: $n_i \geqslant M_n$ for some $M_n \to \infty$ as $n \to \infty$.

Next, we show that the convergence rates and limiting variances of $\hat{\boldsymbol{\beta}}(t)$ are different for sparse and dense longitudinal data. To gain intuition about this, we decompose the difference between the estimated value $\hat{\boldsymbol{\beta}}(t)$ and the true value $\boldsymbol{\beta}(t)$ in the following way:

$$\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t) - \mathbf{H}_n^{-1} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \left[\mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) - \mathbf{x}_{ij}^T \boldsymbol{\beta}(t) \right] K(\frac{t_{ij} - t}{h}) = \mathbf{H}_n^{-1} \sum_{i=1}^{n} \boldsymbol{\xi}_i, \tag{2.5}$$

where the asymptotic distribution of $\hat{\boldsymbol{\beta}}(t)$ is determined by the right hand side, with

$$\boldsymbol{\xi}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{\xi}_{ij}, \quad \boldsymbol{\xi}_{ij} = \mathbf{x}_{ij} \left[v_i(t_{ij}) + \sigma(t_{ij}) \epsilon_{ij} \right] K(\frac{t_{ij} - t}{h}). \tag{2.6}$$

Based on the previous definition $\gamma(t, t') = \mathrm{cov}\left\{ v_i(t), v_i(t') \right\} = \mathrm{E}\left[v_i(t) v_i(t') \right]$, and $\mathrm{E}(\boldsymbol{\xi}_{ij} \boldsymbol{\xi}_{ij'}^T) = \mathrm{E}\left\{ \mathrm{E}\left(\boldsymbol{\xi}_{ij} \boldsymbol{\xi}_{ij'}^T \mid t_{ij}, t_{ij'} \right) \right\}$, we have, for $j \neq j'$,

$$\mathrm{E}(\boldsymbol{\xi}_{ij} \boldsymbol{\xi}_{ij'}^T) = \mathrm{E}\left\{ \mathbf{G}(t_{ij}, t_{ij'}) \gamma(t_{ij}, t_{ij'}) K(\frac{t_{ij} - t}{h}) K(\frac{t_{ij'} - t}{h}) \right\} \approx h^2 \mathbf{G}(t, t) f^2(t) \gamma(t, t), \tag{2.7}$$

where $\mathbf{G}(t_{ij}, t_{ij'}) = \mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij'}^T \mid t_{ij}, t_{ij'})$ and $\mathbf{G}(t,t) = \lim_{t' \to t} \mathbf{G}(t, t')$. Throughout this chapter, $a_n \approx b_n$ means that $a_n/b_n \to 1$. For the same subject and same time point,

$$\mathrm{E}(\boldsymbol{\xi}_{ij}\boldsymbol{\xi}_{ij}^T) = \mathrm{E}\left\{\mathbf{\Gamma}(t_{ij})\left[\gamma(t_{ij}, t_{ij}) + \sigma^2(t_{ij})\right] K^2(\frac{t_{ij}-t}{h})\right\} \approx \mathbf{\Gamma}(t) h f(t) \psi_k \left[\gamma(t,t) + \sigma^2(t)\right], \tag{2.8}$$

where $\mathbf{\Gamma}(t_{ij}) = \mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij}^T | t_{ij})$ and $\psi_K = \int_{\mathbb{R}} K^2(u)\mathrm{d}u$. Since

$$\mathrm{var}(\boldsymbol{\xi}_i | n_i) = n_i^{-2}\left\{\sum_{j=1}^{n_i} \mathrm{E}(\boldsymbol{\xi}_{ij}\boldsymbol{\xi}_{ij}^T) + \sum_{1 \leqslant j \neq j' \leqslant n_i} \mathrm{E}(\boldsymbol{\xi}_{ij}\boldsymbol{\xi}_{ij'}^T)\right\},$$

then by (2.7) and (2.8), we have the following result,

$$\mathrm{var}(\boldsymbol{\xi}_i | n_i) \approx \frac{1}{n_i}\mathbf{\Gamma}(t) h f(t) \psi_K \left[\gamma(t,t) + \sigma^2(t)\right] + (1 - \frac{1}{n_i})\mathbf{G}(t,t) h^2 f^2(t) \gamma(t,t). \tag{2.9}$$

Under the sparse assumption with $h \to 0$, $\mathrm{var}(\boldsymbol{\xi}_i | n_i) \approx \mathbf{\Gamma}(t) h f(t) \psi_K \left[\gamma(t,t) + \sigma^2(t)\right] / n_i$; under the dense assumption with $n_i \geq M_n$ and $M_n h \to \infty$, $\mathrm{var}(\boldsymbol{\xi}_i | n_i) \approx \mathbf{G}(t,t) h^2 f^2(t) \gamma(t,t)$. Therefore, the limiting variances for sparse and dense cases are substantially different. We state the asymptotic properties for these two scenarios in the following theorem.

Theorem 2.2.1. *Let*

$$\boldsymbol{\rho}(t) = \left[\frac{\boldsymbol{\beta}'(t) f'(t)}{f(t)} + \frac{\boldsymbol{\beta}''(t)}{2} + \mathbf{\Gamma}^{-1}(t)\mathbf{\Gamma}'(t)\boldsymbol{\beta}'(t)\right] \int_{\mathbb{R}} u^2 K(u)\mathrm{d}u.$$

Based on the regularity conditions in Section 2.5, we have the following asymptotic results.

- *Sparse data: Assume $nh \to \infty$ and $\sup_n nh^5 < \infty$. Then*

$$\sqrt{nh}\left[\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t)\right] \to N\left(\mathbf{0}_{k+1}, \mathbf{\Sigma}_{sparse}(t)\right), \tag{2.10}$$

where $\boldsymbol{0}_{k+1}$ *is a* $(k+1) \times 1$ *vector with each entry being 0,* $\tau = E(1/n_1)$*, and* $\boldsymbol{\Sigma}_{sparse}(t) = \boldsymbol{\Gamma}^{-1}(t)\psi_K \left[\gamma(t,t) + \sigma^2(t)\right] \tau / f(t)$.

- *Dense data: Assume* $n_i \geq M_n$*,* $M_n h \to \infty$*,* $nh \to \infty$ *and* $sup_n nh^4 < \infty$*. Then*

$$\sqrt{n}\left[\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t)\right] \to N\left(\boldsymbol{0}_{k+1}, \boldsymbol{\Sigma}_{dense}(t)\right), \tag{2.11}$$

where $\boldsymbol{\Sigma}_{dense}(t) = \boldsymbol{\Gamma}^{-1}(t)\,\boldsymbol{G}(t,t)\gamma(t,t)\boldsymbol{\Gamma}^{-1}(t)$.

Based on Theorem 2.2.1, the $\hat{\boldsymbol{\beta}}(t)$ has the traditional nonparametric convergence rate if the data are sparse but has the root n convergence rate if the data are dense. In addition, note that if $\mathbf{x} = 1$, then Theorem 2.2.1 simplifies to the asymptotic results provided by Kim and Zhao (2013).

Based on the asymptotic normalities in Theorem 2.2.1, the confidence intervals for $\boldsymbol{\beta}(t)$ are different under sparse and dense assumptions. Let $z_{1-\alpha/2}$ be the $1-\alpha/2$ standard normal quantile. Then an asymptotic $1-\alpha$ confidence interval for the smooth function $\beta_l(t)$, $l = 0, \ldots, k$ is

$$\hat{\beta}_l(t) - h^2\hat{\rho}_l(t) \pm z_{1-\alpha/2}(nh)^{-1/2}\left\{\left[\hat{\boldsymbol{\Gamma}}^{-1}(t)\psi_K\left[\hat{\gamma}(t,t) + \hat{\sigma}^2(t)\right]\hat{\tau}/\hat{f}(t)\right]^{1/2}\right\}_{l,l} \tag{2.12}$$

for sparse data, or

$$\hat{\beta}_l(t) - h^2\hat{\rho}_l(t) \pm z_{1-\alpha/2}n^{-1/2}\left\{\left[\hat{\boldsymbol{\Gamma}}^{-1}(t)\hat{\mathbf{G}}(t,t)\hat{\gamma}(t,t)\hat{\boldsymbol{\Gamma}}^{-1}(t)\right]^{1/2}\right\}_{l,l} \tag{2.13}$$

for dense data, where $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_k(t))^T$, $\hat{\beta}_l(t)$ is the $(l+1)$th element of $\hat{\boldsymbol{\beta}}(t)$, $\hat{\rho}_l(t)$ is the $(l+1)$th element of $\hat{\boldsymbol{\rho}}(t)$ and the subscript (l,l) refers to the $(l+1)$th diagonal element of a matrix. In the above formulas, $\hat{\tau} = n^{-1}\sum_{i=1}^{n} n_i^{-1}$, $\hat{\gamma}(t,t)$, $\hat{\sigma}^2(t)$, $\hat{f}(t)$, $\hat{\rho}_l(t)$, $\hat{\boldsymbol{\Gamma}}^{-1}(t)$, and $\hat{\mathbf{G}}(t,t)$ are consistent estimates of τ, $\gamma(t,t)$, $\sigma^2(t)$, $f(t)$, $\rho_l(t)$, $\boldsymbol{\Gamma}^{-1}(t)$, and $\mathbf{G}(t,t)$.

2.2.3 Proposed Unified Approach

From Section 2.2.2, the asymptotic results for sparse and dense longitudinal data are essentially different and thus a subjective choice between these two situations might pose challenges for statistical inference, which motivates us to find a unified approach.

In this section, we propose a unified self-normalized central limit theorem which can adapt to both sparse and dense cases for the mixed-effects time-varying coefficient model (2.2). Let

$$\mathbf{U}_n(t) = \mathbf{H}_n^{-1}\mathbf{W}_n\mathbf{H}_n^{-1},$$

where $\mathbf{H}_n$ has the same definition in (2.4), and

$$\mathbf{W}_n = \sum_{i=1}^{n}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left[y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(t_{ij})\right]K(\frac{t_{ij}-t}{h})\right\}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^T\left[y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(t_{ij})\right]K(\frac{t_{ij}-t}{h})\right\}.$$

We have the following unified central limit theorem.

Theorem 2.2.2. *Assume $nh/\log n \to \infty$ and $\sup_n nh^5 < \infty$ for sparse data, or $n_i \geq M_n$, $M_n h \to \infty$, $nh^2/\log n \to \infty$ and $sup_n nh^4 < \infty$ for dense data. Under the regularity conditions in Section 2.5,*

$$\boldsymbol{U}_n(t)^{-1/2}\left[\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t)\right] \to N(\boldsymbol{0}_{k+1}, \boldsymbol{I}_{k+1})$$

in both the sparse and the dense settings, where $\boldsymbol{I}_{k+1}$ is the $(k+1)\times(k+1)$ identity matrix.

Note that the central limit theorem proposed in Kim and Zhao (2013) is a special case of Theorem 2.2.2 if $\mathbf{x} = 1$ is assumed in model (2.2). Based on Theorem 2.2.2, a unified asymptotic pointwise $1-\alpha$ confidence interval for $\beta_l(t)$, $l = 0, \ldots, k$ can be written as follows:

$$\hat{\beta}_l(t) - h^2\hat{\rho}_l(t) \pm z_{1-\alpha/2}\left[\mathbf{U}_n(t)^{1/2}\right]_{l,l}. \tag{2.14}$$

The confidence intervals (2.12) and (2.13) in Section 2.2.2 require to estimate the within-subject covariance function $\gamma(t,t)$, the overall noise variance function $\sigma^2(t)$, and the conditional expectation $\mathbf{G}(t,t)$, which need extra smoothing procedures; but (2.14) does not need those estimations and can be used for both sparse and dense cases through the self-normalizer $\mathbf{U}_n(t)^{1/2}$.

For kernel regression, the selection of bandwidth is generally more important than the selection of kernel functions. As stated in Wu and Chiang (2000), under-smoothing or over-smoothing is mainly caused by inappropriate bandwidth choices in practice, but is rarely influenced by the kernel shapes. Since it is difficult to estimate the bias $h^2\boldsymbol{\rho}(t)$ in practice due to the unknown derivatives $f^{'}$, $\boldsymbol{\beta}^{'}$, $\boldsymbol{\beta}^{''}$ and $\boldsymbol{\Gamma}^{'}$, we use the same kernel function as in Kim and Zhao (2013), $K(u) = 2G(u) - G(u/\sqrt{2})/\sqrt{2}$, where $G(u)$ is the standard normal density. Then $\int_{\mathbb{R}} u^2 K(u)\mathrm{d}u = 0$ and therefore $\boldsymbol{\rho}(t) = \mathbf{0}_{k+1}$. This obviously does not solve the bias problem. For instance, if f, $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are four times differentiable, then we have the higher order bias term $O(h^4)$. As Kim and Zhao (2013) stated, the bias problem is an inherently difficult problem and no good solutions so far.

To select the bandwidth for $\hat{\boldsymbol{\beta}}$, we use the "leave-one-subject-out" cross-validation procedure suggested by Rice and Silverman (1991). Let $\hat{\boldsymbol{\beta}}_{-i}(t)$ be a kernel estimator of $\boldsymbol{\beta}(t)$ computed using the data with all the repeated measurements of the ith subject left out, and define

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left\{ y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{-i}(t_{ij}) \right\}^2 \tag{2.15}$$

to be the subject-based cross-validation. The optimal bandwidth is then defined to be the unique minimizer of $\mathrm{CV}(h)$.

2.3 Simulation and Real Data Application

2.3.1 Simulation Study

We follow Kim and Zhao (2013) to construct the subject-specific random trajectory $v_i(\cdot)$. Consider the model

$$y_{ij} = \sum_{l=0}^{2} \beta_l(t_{ij}) x_{ijl}(t_{ij}) + \sum_{m=1}^{3} \alpha_{im} \Phi_m(t_{ij}) + \sigma \epsilon_{ij} \quad (i = 1, ..., n;\ j = 1, ..., n_i),$$

where $\alpha_{im} \sim N(0, \omega_m)$ and $\epsilon_{ij} \sim N(0, 1)$. Let $\beta_0(t) = 5(t - 0.6)^2$, $\beta_1(t) = cos(3\pi t)$, $\beta_2(t) = sin(2\pi t)$, $\Phi_1(t) = 1$, $\Phi_2(t) = \sqrt{2} sin(2\pi t)$, $\Phi_3(t) = \sqrt{2} cos(2\pi t)$, $(\omega_1, \omega_2, \omega_3) = (0.6, 0.3, 0.1)$, $\sigma = 1$, and $n = 200$. Then the variance function $\gamma(t, t) = 0.6 + 0.6 sin^2(2\pi t) + 0.2 cos^2(2\pi t)$. The time points t_{ij} are uniformly distributed on $[0, 1]$. To generate covariates, let $b_{i1} \sim N(0, 0.3)$, $b_{i2} \sim N(0, 0.3)$, $\eta_{ij} \sim N(0, 1)$, $\delta_{ij} \sim N(0, 1)$ and $\varphi(t) = \sqrt{2}(t + 1)$, then set $x_{ij0} = 1$, $x_{ij1} = b_{i1}\varphi(t_{ij}) + \eta_{ij}$ and $x_{ij2} = b_{i2}\varphi(t_{ij}) + \delta_{ij}$ for $i = 1, ..., n$ and $j = 1, ..., n_i$. Under this setting, we have the following conditional expectations:

$$\mathbf{\Gamma}(t_{ij}) = \mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij}^T \mid t_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.6(t_{ij} + 1)^2 + 1 & 0 \\ 0 & 0 & 0.6(t_{ij} + 1)^2 + 1 \end{pmatrix},$$

$$\mathbf{G}(t_{ij}, t_{ij}) = \lim_{t_{ij'} \to t_{ij}} \mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij'}^T \mid t_{ij}, t_{ij'}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.6(t_{ij} + 1)^2 & 0 \\ 0 & 0 & 0.6(t_{ij} + 1)^2 \end{pmatrix}.$$

For the vector $N = (n_1, n_2, ..., n_n)$ of the number of repeated measurements on each

subject, we consider four cases

$$N_1: \quad n_i \sim U[\{5, 6, ..., 15\}]; \qquad N_2: \quad n_i \sim U[\{15, 16, ..., 35\}]; \tag{2.16}$$

$$N_3: \quad n_i \sim U[\{80, 81, ..., 120\}]; \qquad N_4: \quad n_i \sim U[\{150, 151, ..., 250\}]. \tag{2.17}$$

Here $U[D]$ represents the discrete uniform distribution on a finite set D. Five confidence intervals are compared in our simulation study:

1. the self-normalization based confidence interval in (2.14) (SN);
2. the asymptotic normality based confidence interval (2.12) for sparse data (NS);
3. the asymptotic normality based confidence intervals (2.13) for dense data (ND);
4. the bootstrap confidence interval with 200 bootstrap replications from sampling subjects with replacement (BS);
5. the infeasible confidence interval (NSD)

$$\hat{\beta}_l(t) - h^2\hat{\rho}_l(t) \pm z_{1-\alpha/2}n^{-1/2}\mathbf{S}_{l,l}, \tag{2.18}$$

where $\mathbf{S} = \left\{\mathbf{\Gamma}^{-1}(t)\mathbf{G}(t,t)\mathbf{\Gamma}^{-1}(t)(1-\hat{\tau})\gamma(t,t) + \mathbf{\Gamma}^{-1}(t)\hat{\tau}\psi_K\left[\gamma(t,t)+\sigma^2(t)\right]/\left[hf(t)\right]\right\}^{1/2}$.

The confidence interval NSD is used as a benchmark to compare the performance of the other confidence intervals, since NSD uses the true theoretical limiting variance function (2.9). Note, however, that NSD is practically infeasible, since it depends on many unknown functions. Similar to Kim and Zhao (2013), we use the true functions $\gamma(t,t)$, $\sigma^2(t)$, $f(t)$, $\mathbf{\Gamma}(t)$, and $\mathbf{G}(t,t)$ for NS, ND, and NSD, which gives an advantage to the above three methods. Note that the proposed self-normalization based confidence interval only requires a point estimate of $\boldsymbol{\beta}(t)$ and thus is very easy to implement.

To measure the performance of different confidence intervals, we use the following two criteria: empirical coverage probabilities and lengths of confidence intervals. Let $t_1 < \cdots <$

t_{20} be 20 grid points evenly spaced on $[0.1, 0.9]$. For each grid point t_j $(j = 1, ..20)$ and a given confidence level, we construct confidence intervals for smooth functions $\beta_0(t_j)$, $\beta_1(t_j)$, and $\beta_2(t_j)$, and compute the empirical coverage probabilities based on 500 replications. For each of the five confidence intervals, the empirical coverage probabilities and lengths are averaged at 20 grid points. The bandwidth used for each replicate is the average of 20 optimal bandwidths in (2.15) based on 20 replications (Kim and Zhao, 2013).

The results are showed in Tables 2.1, 2.2, and 2.3. It can be easily seen that the performance of the confidence intervals NS and ND for all $\beta_0(t)$, $\beta_1(t)$, and $\beta_2(t)$ strongly depends on the spareness or denseness of the data. When the number of repeated measurements on each subject is increased from the sparse setting N_1 to the dense setting N_4, the performance of the confidence interval NS assuming the sparse data becomes worse, while the confidence interval ND assuming the dense data becomes better. These two confidence intervals only perform well under their corresponding sparse or dense setting, which further confirms the theoretical results in Theorem 2.2.1.

Note that the confidence interval ND assuming dense data gives same widths for each simulation setting at a certain nominal level. This is because the asymptotic variances at 20 grid points assuming dense data are the same for each simulation setting. In addition, since we use the same way to generate two covariates x_{ij1} and x_{ij2}, the diagonal elements in $\mathbf{\Gamma}(t)$ and $\mathbf{G}(t, t)$ corresponding to $\beta_1(t)$ and $\beta_2(t)$ in (2.12), (2.13), and (2.18) are the same at a given grid point. Hence the widths of the confidence intervals of $\beta_1(t)$ and $\beta_2(t)$ are the same for NSD, NS, and ND .

Compared to NS and ND, the proposed self-normalization based confidence interval SN provides much robust and better performance. Firstly, it has similar widths and coverage probabilities as the bootstrap confidence interval (BS) and both of them perform closely to the infeasible confidence interval NSD; secondly, its computing time is much faster than the bootstrap confidence interval; finally, the asymptotic properties of the self-normalization method have been established in this chapter, whereas the theoretical properties of the

bootstrap procedure for longitudinal data have not been developed as far as we know.

Table 2.1: *Average empirical coverage percentages and lengths, in brackets, for $\beta_0(t)$ of five confidence intervals.*

$1-\alpha$	N	SN	NS	ND	NSD	BS
90%	N_1	88.0(0.367)	80.3(0.303)	68.6(0.236)	88.9(0.375)	89.0(0.380)
	N_2	88.0(0.301)	70.8(0.201)	78.1(0.236)	88.9(0.306)	88.7(0.307)
	N_3	90.1(0.258)	53.5(0.112)	87.1(0.236)	90.5(0.260)	90.1(0.258)
	N_4	89.1(0.248)	44.6(0.087)	87.4(0.236)	89.5(0.251)	89.3(0.249)
95%	N_1	92.8(0.437)	86.7(0.361)	75.5(0.281)	93.7(0.447)	93.5(0.451)
	N_2	93.7(0.359)	78.4(0.240)	85.2(0.281)	94.1(0.365)	94.0(0.365)
	N_3	94.2(0.307)	60.1(0.134)	92.1(0.281)	94.8(0.310)	94.2(0.308)
	N_4	93.7(0.296)	51.0(0.104)	92.4(0.281)	94.1(0.299)	93.6(0.297)

SN, the self-normalized confidence interval in (2.14); NS and ND, the asymptotic normality based confidence intervals (2.12) and (2.13) assuming sparse and dense data, respectively; NSD, the infeasible confidence interval in (2.18); BS, the bootstrap confidence interval; $N_1 - N_4$, the number of measurements on individual subject in (2.16) and (2.17).

2.3.2 Application to AIDS Data

In this section, we apply the self-normalization based confidence interval to the AIDS data (Qu and Li, 2006), which came from the Multi-Center AIDS Cohort Study. CD4 cells can be destroyed by human immune-deficiency virus(HIV) and thus the percentage of the CD4 cells in the blood of a human body will change after HIV infection. Because of this, CD4 cell count and the percentage in the blood are the most popular used markers for doctors to monitor the progression of the disease.

The HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991 was included in this data set. All individuals were scheduled to have their measurements made twice a year. Since many patients missed some of their

Table 2.2: *Average empirical coverage percentages and lengths, in brackets, for $\beta_1(t)$ of five confidence intervals.*

$1-\alpha$	N	SN	NS	ND	NSD	BS
90%	N_1	85.7(0.218)	82.1(0.198)	56.4(0.115)	87.4(0.226)	88.8(0.238)
	N_2	87.1(0.169)	76.6(0.132)	70.4(0.115)	88.5(0.174)	88.7(0.177)
	N_3	88.6(0.133)	61.2(0.074)	82.6(0.115)	89.8(0.136)	89.0(0.135)
	N_4	89.2(0.126)	54.5(0.057)	86.3(0.115)	90.1(0.128)	89.2(0.126)
95%	N_1	91.4(0.261)	88.4(0.236)	64.1(0.137)	92.6(0.270)	93.7(0.283)
	N_2	92.7(0.201)	84.1(0.157)	78.1(0.137)	93.4(0.207)	93.7(0.210)
	N_3	93.2(0.159)	68.8(0.088)	88.9(0.137)	94.1(0.163)	93.5(0.161)
	N_4	93.7(0.150)	61.4(0.068)	91.9(0.137)	94.6(0.153)	93.7(0.151)

SN, the self-normalized confidence interval in (2.14); NS and ND, the asymptotic normality based confidence intervals (2.12) and (2.13) assuming sparse and dense data, respectively; NSD, the infeasible confidence interval in (2.18); BS, the bootstrap confidence interval; $N_1 - N_4$, the number of measurements on individual subject in (2.16) and (2.17).

scheduled visits and all the HIV infections happened randomly during the study, the numbers of repeated measurements for each patient are not equal and their measurement times are different. Further details about the design, methods, and medical implications of the study can be found in Kaslow, et al. (1987).

The response variable is the CD4 percentage over time. Three covariates are: patient's age, smoking status with 1 as smoker and 0 as nonsmoker, and the CD4 cell percentage before their infection. The aim of our statistical analysis is to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage, and age at HIV infection on the mean CD4 percentage after the infection. Define t_{ij} to be the time (in years) of the jth measurement of the ith individual after HIV infection. In this data set, the patients have minimum 1 and maximum 14 measurements. Let Y_{ij} be the ith individual's CD4 percentage at time t_{ij} and X_{1i} be the smoking status for the ith individual (equal to 1 for smoker and 0 for nonsmoker). In order to have clear biological interpretations, we use centered age, obtained

Table 2.3: *Average empirical coverage percentages and lengths, in brackets, for $\beta_2(t)$ of five confidence intervals.*

$1-\alpha$	N	SN	NS	ND	NSD	BS
90%	N_1	86.6(0.219)	83.0(0.198)	57.1(0.115)	88.2(0.226)	88.9(0.232)
	N_2	86.9(0.169)	77.1(0.132)	70.6(0.115)	88.2(0.174)	88.1(0.174)
	N_3	88.5(0.134)	61.6(0.074)	82.8(0.115)	89.6(0.136)	88.7(0.135)
	N_4	88.9(0.126)	54.0(0.057)	85.8(0.115)	90.1(0.128)	89.0(0.127)
95%	N_1	92.0(0.260)	89.3(0.236)	65.2(0.137)	93.5(0.270)	93.8(0.276)
	N_2	93.0(0.201)	84.4(0.157)	78.6(0.137)	94.0(0.207)	93.7(0.208)
	N_3	93.5(0.160)	69.8(0.088)	89.3(0.137)	94.1(0.163)	93.8(0.161)
	N_4	93.7(0.150)	60.0(0.068)	91.3(0.137)	94.2(0.153)	93.6(0.150)

SN, the self-normalized confidence interval in (2.14); NS and ND, the asymptotic normality based confidence intervals (2.12) and (2.13) assuming sparse and dense data, respectively; NSD, the infeasible confidence interval in (2.18); BS, the bootstrap confidence interval; $N_1 - N_4$, the number of measurements on individual subject in (2.16) and (2.17).

by subtracting the sample average age at infection from the ith individual's age at infection and denoted by X_{2i}, and centered pre-infection CD4 percentage, obtained by subtracting the average pre-infection CD4 percentage of the sample from the ith patient's actual pre-infection CD4 percentage, which is denoted by X_{3i}. Then we construct the time-varying coefficient model for the AIDS data as follows:

$$Y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})X_{1i} + \beta_2(t_{ij})X_{2i} + \beta_3(t_{ij})X_{3i} + \epsilon_{ij},$$

where $\beta_0(t)$ represents the baseline CD4 percentage and can be interpreted as the mean CD4 percentage at time t for a nonsmoker with average pre-infection CD4 percentage and average age at HIV infection. Therefore, $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$ represent the time-varying effects for cigarette smoking, age at HIV infection, and pre-infection CD4 percentage, respectively, on the post-infection CD4 percentage at time t.

We use the kernel smoothing method stated in (2.3) to estimate the smoothing functions $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$. The bandwidth was chosen by using the leave-one-subject-out cross-validation method. The self-normalization based 95% confidence intervals were constructed for $\beta_0(t), \ldots, \beta_3(t)$ at 100 equally spaced time points between 0.1 and 5.9 years. We also constructed the bootstrap 95% confidence intervals at the same 100 time points, based on 1000 bootstrap replications. Figure 2.1 depicts the fitted coefficient functions (solid curves) with 95% self-normalization based confidence intervals (dashed curves) and bootstrap confidence intervals (dotted curves). It can be easily seen that the self-normalization based confidence intervals are very close to bootstrap confidence intervals. Indeed, they almost overlap with each other. However, the computing time for the self-normalization based confidence interval is much faster than the bootstrap confidence interval. The former one only takes approximately 5 seconds, whereas the latter one needs almost 50 minutes based on a personal computer with Intel(R) Core(TM) i5 CPU, 4GB installed memory, and 32-bit operating system.

Based on the constructed confidence intervals, the mean baseline CD4 percentage of the population decreases with time, but at a rate that appears to be slowing down at four years after the infection. Since the confidence intervals for cigarette smoking and age of HIV infection cover 0 most of the time, these two covariates do not significantly affect the post-infection CD4 percentage. The pre-infection CD4 percentage appears to be positively associated with higher post-infection CD4 percentage. Our findings basically agree with Wu and Chiang (2000), Fan and Zhang (2000), Huang, et al. (2002), and Qu and Li (2006).

2.4 Discussion

In this chapter, we proposed a unified inference for the time-varying coefficient model (2.2) for the longitudinal data based on the new established unified self-normalized central limit theorem. The new inference tool allows us to do inference for the longitudinal data without

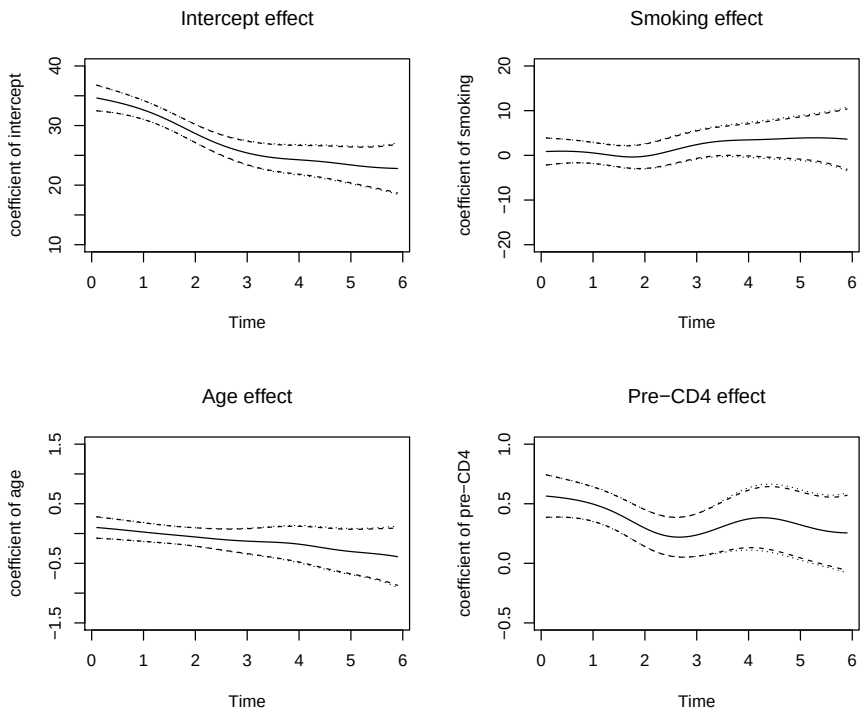

Figure 2.1: *Application to AIDS data. Estimated coefficient curves for the baseline CD4 percentage and the effects of smoking, age and pre-infection CD4 percentage on the percentage of CD4 cells. Solid curves, estimated effects; dashed curves,* 95% *self-normalization based confidence intervals; dotted curves,* 95% *bootstrap confidence intervals.*

subjectively deciding whether the data are sparse or dense. The effectiveness of the proposed unified inference is demonstrated through a simulation study and an analysis of an AIDS data set.

The weighted local constant estimators that we considered in this chapter only use one smoothing parameter, which may not be able to provide adequate smoothing for all the coefficient curves at the same time. Wu and Chiang (2000) proposed the componentwise local least squares criteria to estimate the time-varying coefficients using different amounts of smoothing. The reason that we use one smoothing parameter is for the simplicity of computation and our proposed unified inference can be extended to the case of different smoothing parameters as well.

For time-varying coefficient models, the commonly asked questions are whether the co-

efficient functions $\boldsymbol{\beta}(\cdot)$ are varying over time and whether certain covariates are significant. Therefore, we may wish to test whether a certain component of $\boldsymbol{\beta}(\cdot)$ is identically zero or constant. The generalized likelihood ratio statistics for the nonparametric testing problems proposed in Fan, et al. (2001) might be considered, but the theoretical and practical aspects for longitudinal data would require substantial development.

2.5 Proofs

The following conditions are imposed to facilitate the proof and are adopted from Wu and Chiang (2000), Huang, et al. (2002) and Kim and Zhao (2013). They are not the weakest possible conditions.

Regularity conditions:

1. The observation time points follow a random design in the sense that t_{ij}, for $j = 1, ..., n_i$ and $i = 1, ..., n$, are chosen independently from an unknown distribution with a density $f(\cdot)$ on a finite interval. The density function $f(\cdot)$ is continuously differentiable in a neighborhood of t and is uniformly bounded away from 0 and infinity.

2. In a neighborhood of t, $\boldsymbol{\beta}(\cdot)$ is twice continuously differentiable, $\sigma^2(\cdot)$ is continuously differentiable. In a neighborhood of (t, t), $\gamma(t, t') = \text{cov}\{v_i(t), v_i(t')\}$ is continuously differentiable and $\gamma(t, t) = \lim_{t' \to t} \text{cov}\{v_i(t), v_i(t')\}$. Furthermore, $\sigma^2(t) < \infty$ and $\gamma(t, t) < \infty$.

3. $\{v_i(\cdot)\}_i$, $\{t_{ij}\}_{ij}$, $\{\epsilon_{ij}\}_{ij}$ are independent and identically distributed and mutually indepndnet.

4. $\{\mathbf{x}_{ij}\}_{ij}$, $\{v_i(\cdot)\}_i$, $\{\epsilon_{ij}\}_{ij}$ are mutually independent. $\{\mathbf{x}_{ij}\}_i$ are independent and identically distributed. For the same i, $\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i}$ have the identical distribution and can be correlated. $\text{E}\left[\|\mathbf{x}_{ij}\| \cdot \left\|\mathbf{x}_{ij'}\right\| \cdot \left\|\mathbf{x}_{ij''}\right\| | t_{ij}, t_{ij'}, t_{ij''}\right] < \infty$ for $1 \leqslant j \neq j' \neq j'' \leqslant n_i$.

5. $\boldsymbol{\Gamma}(t)$ is invertible and differentiable.

6. $\mathrm{E}\{|v_i(\cdot)+\sigma(\cdot)\epsilon_{ij}|^3\}$ is continuous in a neighborhood of t and $\mathrm{E}\{|v_i(\cdot)+\sigma(\cdot)\epsilon_{ij}|^3\}<\infty$.

7. $K(\cdot)$ is bounded, symmetric, and has bounded support and bounded derivative.

Since $\sigma^2(t)$ and $\gamma(t,t)$ are unknown in most applications and the unified approach that we proposed does not need the specific structures of $\sigma^2(t)$ and $\gamma(t,t)$, therefore, we do not require further specific structures for $\sigma^2(t)$ and $\gamma(t,t)$, except for their continuity in the above condition 2.

Proof of Theorem 2.2.1. Based on (2.5), the asymptotic results for sparse or dense longitudinal data depend on the limiting distribution of $\boldsymbol{\xi}_i$ which is defined in (2.6). In order to obtain the limiting distribution of $\boldsymbol{\xi}_i$, we define the following notations.

$$\mathbf{H}_n=\sum_{i=1}^{n}\mathbf{V}_i,\qquad \mathbf{V}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{V}_{ij},\qquad \mathbf{V}_{ij}=\mathbf{x}_{ij}\mathbf{x}_{ij}^T K(\frac{t_{ij}-t}{h}),$$

$$\mathbf{b}_n=\sum_{i=1}^{n}\boldsymbol{\zeta}_i,\qquad \boldsymbol{\zeta}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{\zeta}_{ij},\qquad \boldsymbol{\zeta}_{ij}=\mathbf{x}_{ij}\left[\mathbf{x}_{ij}^T\boldsymbol{\beta}(t_{ij})-\mathbf{x}_{ij}^T\boldsymbol{\beta}(t)\right]K(\frac{t_{ij}-t}{h}).$$

$$\boldsymbol{\Gamma}(t_{ij})=\mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij}^T|t_{ij}),\quad \Gamma_1(t_{ij})=\mathrm{E}(x_{ijl}^2x_{ijr}^2|t_{ij}),\quad \boldsymbol{\Gamma}_2(t_{ij})=\mathrm{E}(X_{ijm}^2\mathbf{x}_{ij}\mathbf{x}_{ij}^T|t_{ij}),$$

where $l,r,m=0,...,(k+1)$. We first want to find the order of $\mathbf{H}_n$ and $\mathbf{b}_n$. Their orders are determined by $\mathbf{V}_{ij}$ and $\boldsymbol{\zeta}_{ij}$. Throughout this chapter, we consider the element-wise variance of a matrix. Based on Taylor's expansion and the symmetry of the kernel function $K(\cdot)$, we have the following results,

$$\begin{aligned}
\mathrm{E}(\mathbf{V}_{ij}) &= \mathrm{E}\left\{\mathrm{E}(\mathbf{V}_{ij}|t_{ij})\right\}\\
&= \mathrm{E}\left\{\mathrm{E}(\mathbf{x}_{ij}\mathbf{x}_{ij}^T|t_{ij})K(\frac{t_{ij}-t}{h})\right\}\\
&= h\int\left[\boldsymbol{\Gamma}(t)+\boldsymbol{\Gamma}'(t)ht_0+o(h)\right]K(t_0)\left[f(t)+f'(t)ht_0+o(h)\right]\mathrm{d}t_0\\
&= \boldsymbol{\Gamma}(t)hf(t)\left[1+O(h^2)\right],
\end{aligned}$$

and

$$\begin{aligned}
\text{var}(\mathbf{V}_{ij}(l,r)) &= \text{var}\left(x_{ijl}x_{ijr}K(\frac{t_{ij}-t}{h})\right) \\
&= \text{E}\left\{\left[x_{ijl}x_{ijr}K(\frac{t_{ij}-t}{h})\right]^2\right\} - \left\{\text{E}\left[x_{ijl}x_{ijr}K(\frac{t_{ij}-t}{h})\right]\right\}^2 \\
&= \text{E}\left[\Gamma_1(t_{ij})K^2(\frac{t_{ij}-t}{h})\right] - O(h^2) \\
&= h\Gamma_1(t)f(t)\psi_K + o(h) - O(h^2) \\
&= O(h),
\end{aligned}$$

where (l,r) refers to the element of $\mathbf{V}_{ij}$ in the lth row and rth column. Therefore, $\text{var}(\mathbf{V}_{ij}) = O(h)$. Similarly, we have the following results for $\boldsymbol{\zeta}_{ij}$,

$$\begin{aligned}
\text{E}(\boldsymbol{\zeta}_{ij}) &= \text{E}\left\{\text{E}\left\{\mathbf{x}_{ij}\left[\mathbf{x}_{ij}^T\boldsymbol{\beta}(t_{ij}) - \mathbf{x}_{ij}^T\boldsymbol{\beta}(t)\right]K(\frac{t_{ij}-t}{h})|t_{ij}\right\}\right\} \\
&= \text{E}\left\{\boldsymbol{\Gamma}(t_{ij})\left[\boldsymbol{\beta}(t_{ij}) - \boldsymbol{\beta}(t)\right]K(\frac{t_{ij}-t}{h})\right\} \\
&= h^3 f(t)\boldsymbol{\Gamma}(t)\left[\frac{\boldsymbol{\beta}^{'}(t)f^{'}(t)}{f(t)} + \frac{\boldsymbol{\beta}^{''}(t)}{2} + \boldsymbol{\Gamma}^{-1}(t)\boldsymbol{\Gamma}^{'}(t)\boldsymbol{\beta}^{'}(t)\right]\int t_0^2 K(t_0)\text{d}t_0 + o(h^3) \\
&= \boldsymbol{\Gamma}(t)h^3 f(t)\boldsymbol{\rho}(t) + o(h^3),
\end{aligned}$$

$$\begin{aligned}
\text{var}(\zeta_{ijm}) &= \text{var}\left\{x_{ijm}\left[\mathbf{x}_{ij}^T\boldsymbol{\beta}(t_{ij}) - \mathbf{x}_{ij}^T\boldsymbol{\beta}(t)\right]K(\frac{t_{ij}-t}{h})\right\} \\
&= \text{E}\left\{\text{E}\left\{\left[x_{ijm}\mathbf{x}_{ij}^T\left[\boldsymbol{\beta}(t_{ij}) - \boldsymbol{\beta}(t)\right]K(\frac{t_{ij}-t}{h})\right]^2|t_{ij}\right\}\right\} - \left[O(h^3)\right]^2 \\
&= \int\left[\boldsymbol{\beta}(t_{ij}) - \boldsymbol{\beta}(t)\right]^T\boldsymbol{\Gamma}_2(t_{ij})\left[\boldsymbol{\beta}(t_{ij}) - \boldsymbol{\beta}(t)\right]K^2(\frac{t_{ij}-t}{h})f(t_{ij})\text{d}t_{ij} - O(h^6) \\
&= O(h^3),
\end{aligned}$$

where $\boldsymbol{\rho}(t) = \left[\frac{\boldsymbol{\beta}^{'}(t)f^{'}(t)}{f(t)} + \frac{\boldsymbol{\beta}^{''}(t)}{2} + \boldsymbol{\Gamma}^{-1}(t)\boldsymbol{\Gamma}^{'}(t)\boldsymbol{\beta}^{'}(t)\right] \int_{\mathbb{R}} u^2 K(u)\mathrm{d}u$, ζ_{ijm} and x_{ijm} are the m^{th} elements of $\boldsymbol{\zeta}_{ij}$ and $\mathbf{x}_{ij}$, respectively. Therefore, $\mathrm{var}(\boldsymbol{\zeta}_{ij}) = O(h^3)$. In order to find the order of $\mathbf{H}_n$, we consider that in either the sparse or the dense case,

$$\mathrm{E}(\mathbf{V}_i|n_i) = \mathrm{E}(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{V}_{ij}|n_i) = \mathrm{E}(\mathbf{V}_{ij}) = \boldsymbol{\Gamma}(t)hf(t)\left[1 + O(h^2)\right]$$

is not random. Then we have

$$\begin{aligned}
\mathrm{var}(\mathbf{V}_i) &= \mathrm{E}\left\{\mathrm{var}(\mathbf{V}_i|n_i)\right\} + \mathrm{var}\left\{\mathrm{E}(\mathbf{V}_i|n_i)\right\} = \mathrm{E}\left\{\mathrm{var}(\mathbf{V}_i|n_i)\right\} \\
&= \mathrm{E}\left\{\mathrm{var}(\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{V}_{ij}|n_i)\right\} \\
&= \mathrm{E}\left\{\frac{1}{n_i^2}\left[\sum_{j=1}^{n_i}\mathrm{var}(\mathbf{V}_{ij}) + \sum_{1\leqslant j\neq j^{'}\leqslant n_i}\mathrm{cov}(\mathbf{V}_{ij},\mathbf{V}_{ij^{'}})\right]\right\}.
\end{aligned}$$

Since $\mathrm{cov}(\mathbf{V}_{ij},\mathbf{V}_{ij^{'}}) \leqslant \sqrt{\mathrm{var}(\mathbf{V}_{ij})\mathrm{var}(\mathbf{V}_{ij^{'}})}$, and $\mathbf{V}_{ij}$ has the same distribution as $\mathbf{V}_{ij^{'}}$, then

$$\mathrm{var}(\mathbf{V}_i) \leqslant \mathrm{E}\left\{\frac{1}{n_i^2}\left[n_i\mathrm{var}(\mathbf{V}_{ij}) + n_i(n_i-1)\mathrm{var}(\mathbf{V}_{ij})\right]\right\} = \mathrm{E}[\mathrm{var}(\mathbf{V}_{ij})] = \mathrm{var}(\mathbf{V}_{ij}).$$

So we get $\mathrm{var}(\mathbf{H}_n) = \sum_{i=1}^{n}\mathrm{var}(\mathbf{V}_i) \leqslant n\mathrm{var}(\mathbf{V}_{ij}) = O(nh)$, which means that $\mathrm{var}(\mathbf{H}_n) = O(nh)$. Based on the above results, we obtain the order of $\mathbf{H}_n$ as follows,

$$\begin{aligned}
\mathbf{H}_n &= \mathrm{E}(\mathbf{H}_n) + O_p\left(\sqrt{\mathrm{var}(\mathbf{H}_n)}\right) \\
&= n\mathrm{E}\left[\mathrm{E}(\mathbf{V}_i|n_i)\right] + O_p(\sqrt{nh}) \\
&= n\boldsymbol{\Gamma}(t)hf(t)\left[1 + O_p(h^2)\right] + O_p(\sqrt{nh}) \\
&= \left[1 + O_p\left\{h^2 + \frac{1}{\sqrt{nh}}\right\}\right] n\boldsymbol{\Gamma}(t)hf(t).
\end{aligned}$$

Similarly, $\mathbf{b}_n = n\boldsymbol{\Gamma}(t)h^3 f(t)\boldsymbol{\rho}(t) + o_p(nh^3) + O_p(\sqrt{nh^3})$. Hence,

$$\begin{aligned}
\mathbf{H}_n^{-1}\mathbf{b}_n &= \frac{\boldsymbol{\Gamma}^{-1}(t)\left[n\boldsymbol{\Gamma}(t)h^3 f(t)\boldsymbol{\rho}(t) + o_p(nh^3) + O_p(\sqrt{nh^3})\right]}{\left[1 + O_p(h^2 + \sqrt{\tfrac{1}{nh}})\right] nhf(t)} \\
&= \frac{nh^3 f(t)\boldsymbol{\rho}(t) + o_p(nh^3) + O_p(\sqrt{nh^3})}{\left[1 + O_p(h^2 + \sqrt{\tfrac{1}{nh}})\right] nhf(t)} \\
&= h^2\boldsymbol{\rho}(t) + \boldsymbol{\delta}_n,
\end{aligned}$$

where $\boldsymbol{\delta}_n = o_p(h^2) + O_p(\sqrt{\frac{h}{n}})$.

For dense longitudinal data, $\sqrt{n}\boldsymbol{\delta}_n = o_p(\sqrt{n}h^2) + O_p(\sqrt{h})$, $n_i \geqslant M_n$, $M_n h \to \infty$, $nh \to \infty$, and $sup_n nh^4 < \infty$, then we have $\boldsymbol{\delta}_n = o_p(1/\sqrt{n})$. Since $\text{var}(\sum_{i=1}^n \boldsymbol{\xi}_i) = n\text{var}(\boldsymbol{\xi}_i) = n\text{E}[\text{var}(\boldsymbol{\xi}_i|n_i)] = n\text{var}(\boldsymbol{\xi}_i|n_i) \approx n\mathbf{G}(t,t)h^2 f^2(t)\gamma(t,t)$, then

$$\left[nh^2 f^2(t)\right]^{-1} \text{var}(\sum_{i=1}^n \boldsymbol{\xi}_i) \approx \mathbf{G}(t,t)\gamma(t,t).$$

We next use the Lyapunov central limit theorem to obtain the asymptotic distribution of $\sum_{i=1}^n \boldsymbol{\xi}_i$. The Lyapunov conditions are checked as follows. For any unit vector $\mathbf{d} \in \mathbb{R}^{k+1}$, let $\mathbf{d}^T \sum_{i=1}^n \boldsymbol{\xi}_i = \sum_{i=1}^n \mathbf{d}^T\boldsymbol{\xi}_i = \sum_{i=1}^n \boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_i = \mathbf{d}^T\boldsymbol{\xi}_i$. Then

$$\text{E}(\boldsymbol{\theta}_i^2) = \text{E}(\mathbf{d}^T\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\mathbf{d}) = \mathbf{d}^T\text{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\mathbf{d} = \mathbf{d}^T\text{var}(\boldsymbol{\xi}_i)\mathbf{d}.$$

Since $\text{var}(\boldsymbol{\xi}_i) = \text{E}\{\text{var}(\boldsymbol{\xi}_i|n_i)\} + \text{var}\{\text{E}(\boldsymbol{\xi}_i|n_i)\} = \text{E}\{\text{var}(\boldsymbol{\xi}_i|n_i)\} = \text{var}(\boldsymbol{\xi}_i|n_i) = O(h^2)$, then $\text{E}(\boldsymbol{\theta}_i^2) = O(h^2)$ and thus $\left(\sum_{i=1}^n \text{E}|\boldsymbol{\theta}_i|^2\right)^3 = O\{(nh^2)^3\} = O(n^3h^6)$. Based on $\mathbf{d}^T\boldsymbol{\xi}_i \leqslant \|\mathbf{d}\| \cdot \|\boldsymbol{\xi}_i\| = \|\boldsymbol{\xi}_i\|$, we have

$$\text{E}(\boldsymbol{\theta}_i^3) \leqslant \text{E}(\|\boldsymbol{\xi}_i\|^3) \leqslant \text{E}\left\{\frac{1}{n_i^3}\left(\sum_{j=1}^{n_i}\left\|\boldsymbol{\xi}_{ij}\right\|\right)^3\right\} \approx \text{E}\left[\left\|\boldsymbol{\xi}_{ij}\right\| \cdot \left\|\boldsymbol{\xi}_{ij'}\right\| \cdot \left\|\boldsymbol{\xi}_{ij''}\right\|\right] = O(h^3),$$

which implies that $\left(\sum_{i=1}^{n} \mathrm{E}|\boldsymbol{\theta}_i|^3\right)^2 = O((nh^3)^2) = O(n^2h^6)$. Since

$$\frac{n^2h^6}{n^3h^6} = \frac{1}{n} \to 0 \quad as \quad n \to \infty,$$

thus $\left(\sum_{i=1}^{n} \mathrm{E}|\boldsymbol{\theta}_i|^3\right)^2 = o\left\{\left(\sum_{i=1}^{n} \mathrm{E}|\boldsymbol{\theta}_i|^2\right)^3\right\}$. The Lyapunov conditions are satisfied and hence the following result is obtained based on the Lyapunov central limit theorem.

$$\frac{\sum_{i=1}^{n} \boldsymbol{\xi}_i}{h\sqrt{n}f(t)} \to N(\mathbf{0}_{k+1}, \mathbf{G}(t,t)\gamma(t,t)).$$

Since $\sum_{i=1}^{n} \boldsymbol{\xi}_i = \mathbf{H}_n \left[\hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}(t) - \mathbf{H}_n^{-1}\mathbf{b}_n\right] = \mathbf{H}_n \left[\hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t) - \boldsymbol{\delta}_n\right]$, $\mathbf{H}_n = \left[1 + O_p\left\{h^2 + \frac{1}{\sqrt{nh}}\right\}\right] n\boldsymbol{\Gamma}(t)hf(t)$ and $\boldsymbol{\delta}_n = o_p(1/\sqrt{n})$, then

$$\frac{\sum_{i=1}^{n} \boldsymbol{\xi}_i}{h\sqrt{n}f(t)} \approx \sqrt{n}\boldsymbol{\Gamma}(t)\left[\hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t)\right].$$

By Slutsky's theorem, $\sqrt{n}\left[\hat{\boldsymbol{\beta}}_n(t) - \boldsymbol{\beta}(t) - h^2\boldsymbol{\rho}(t)\right] \to N(\mathbf{0}_{k+1}, \boldsymbol{\Gamma}^{-1}(t)\mathbf{G}(t,t)\gamma(t,t)\boldsymbol{\Gamma}^{-1}(t))$.

Similarly, for sparse longitudinal data, since $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n$ are independent and identically distributed, then the result follows from $\boldsymbol{\delta}_n = o_p(1/\sqrt{nh})$ and $\mathrm{var}(\sum_{i=1}^{n} \boldsymbol{\xi}_i) \approx nh\tau\psi_K f(t)[\gamma(t,t) + \sigma^2(t)]\boldsymbol{\Gamma}(t)$. $\square$

Proof of Theorem 2.2.2. Based on Theorem 2.2.1, if we can show $n\mathbf{U}_n(t) \to \boldsymbol{\Sigma}_{dense}(t)$ and $nh\mathbf{U}_n(t) \to \boldsymbol{\Sigma}_{sparse}(t)$, then the Theorem 2.2.2 can be proved.

Denote $K_{ij} = K(\frac{t_{ij}-t}{h})$. Let

$$\begin{aligned}
\mathbf{W}_n &= \sum_{i=1}^{n} \left\{\frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}\left[y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(t_{ij})\right]K_{ij}\right\}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}^T\left[y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(t_{ij})\right]K_{ij}\right\} \\
&= \sum_{i=1}^{n} \left(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \boldsymbol{\xi}_i\boldsymbol{\alpha}_i^T + \boldsymbol{\alpha}_i\boldsymbol{\xi}_i^T + \boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T\right),
\end{aligned}$$

where $\boldsymbol{\xi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}$, and $\boldsymbol{\alpha}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left[\mathbf{x}_{ij}^T\boldsymbol{\beta}(t_{ij})-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}(t_{ij})\right]K_{ij}$. Similarly as Kim and Zhao (2013), by Theorem 3.1 in Li and Hsing (2010), $\left|\hat{\boldsymbol{\beta}}(z)-\boldsymbol{\beta}(z)\right| = O_p(l_n)\mathbf{1}_{k+1}$ uniformly for z in the neighborhood of t, where $l_n = h^2+\sqrt{\frac{\log n}{n}}$ for dense data, $l_n = h^2+\sqrt{\frac{\log n}{nh}}$ for sparse data, $\mathbf{1}_{k+1}$ is a $(k+1)\times 1$ vector with all elements equal to 1. Then

$$\boldsymbol{\alpha}_i = O_p(|\boldsymbol{\alpha}_i|) = O_p(l_n)\frac{1}{n_i}\sum_{j=1}^{n_i}\left|\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{1}_{k+1}K_{ij}\right|.$$

Since $\boldsymbol{\xi}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{\xi}_{ij}$ which is defined in (2.6), we can get

$$\begin{aligned}\sum_{i=1}^{n}\left|\boldsymbol{\xi}_i\boldsymbol{\alpha}_i^T+\boldsymbol{\alpha}_i\boldsymbol{\xi}_i^T+\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T\right| &= O_p(l_n)\sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\left|\boldsymbol{\xi}_{ij}\right|\sum_{j=1}^{n_i}\left|\mathbf{x}_{ij}^T(\mathbf{x}_{ij}^T\mathbf{1}_{k+1})K_{ij}\right| \\ &+O_p(l_n)\sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\left|\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{1}_{k+1}K_{ij}\right|\sum_{j=1}^{n_i}\left|\boldsymbol{\xi}_{ij}^T\right| \\ &+O_p(l_n^2)\sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\left|\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{1}_{k+1}K_{ij}\right|\sum_{j=1}^{n_i}\left|\mathbf{x}_{ij}^T(\mathbf{x}_{ij}^T\mathbf{1}_{k+1})K_{ij}\right|.\end{aligned}$$

Based on the proof of Theorem 2.2.1,

$$\boldsymbol{\xi}_{ij} = \mathrm{E}(\boldsymbol{\xi}_{ij})+O_p(\sqrt{\mathrm{var}(\boldsymbol{\xi}_{ij})}) = O_p(\sqrt{\mathrm{E}(\boldsymbol{\xi}_{ij}\boldsymbol{\xi}_{ij}^T)}) = O_p(\sqrt{h}),$$

$$\mathbf{x}_{ij}\mathbf{x}_{ij}^TK_{ij} = \mathbf{V}_{ij} = \mathrm{E}(\mathbf{V}_{ij})+O_p(\sqrt{\mathrm{var}(\mathbf{V}_{ij})}) = O_p(h)+O_p(\sqrt{h}) = O_p(\sqrt{h}).$$

Since $\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{1}_{k+1}K_{ij} = \mathbf{x}_{ij}\mathbf{x}_{ij}^TK_{ij}\mathbf{1}_{k+1}$, $\mathbf{x}_{ij}^T(\mathbf{x}_{ij}^T\mathbf{1}_{k+1})K_{ij} = \mathbf{1}_{k+1}^T\mathbf{x}_{ij}\mathbf{x}_{ij}^TK_{ij}$ and $l_n^2 = o(l_n)$, then $\sum_{i=1}^{n}\left|\boldsymbol{\xi}_i\boldsymbol{\alpha}_i^T+\boldsymbol{\alpha}_i\boldsymbol{\xi}_i^T+\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T\right| = O_p(nhl_n)$. Recall that $\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_n$ are independent, then

$$\begin{aligned}\mathbf{W}_n &= \mathrm{E}\left(\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right)+O_p\left(\sqrt{\mathrm{var}(\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)}\right)+O_p(nhl_n) \\ &= \sum_{i=1}^{n}\mathrm{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)+O_p(x_n),\end{aligned}$$

where $x_n = \sqrt{\sum_{i=1}^n \text{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)} + nhl_n$. By Theorem 2.2.1, we have the following results for dense and sparse cases.

$$\begin{aligned} n\mathbf{H}_n^{-1}\sum_{i=1}^n \text{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\mathbf{H}_n^{-1} &= n\{[1+o_p(1)]n\boldsymbol{\Gamma}(t)hf(t)\}^{-1}\sum_{i=1}^n \text{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\{[1+o_p(1)]n\boldsymbol{\Gamma}(t)hf(t)\}^{-1} \\ &\approx \frac{\boldsymbol{\Gamma}^{-1}(t)\text{var}(\sum_{i=1}^n \boldsymbol{\xi}_i)\boldsymbol{\Gamma}^{-1}(t)}{nh^2f^2(t)} \\ &\to \boldsymbol{\Gamma}^{-1}(t)\mathbf{G}(t,t)\gamma(t,t)\boldsymbol{\Gamma}^{-1}(t) = \boldsymbol{\Sigma}_{dense}(t), \end{aligned}$$

$$\begin{aligned} nh\mathbf{H}_n^{-1}\sum_{i=1}^n \text{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\mathbf{H}_n^{-1} &= nh\{[1+o_p(1)]n\boldsymbol{\Gamma}(t)hf(t)\}^{-1}\sum_{i=1}^n \text{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\{[1+o_p(1)]n\boldsymbol{\Gamma}(t)hf(t)\}^{-1} \\ &\approx \frac{\boldsymbol{\Gamma}^{-1}(t)\text{var}(\sum_{i=1}^n \boldsymbol{\xi}_i)\boldsymbol{\Gamma}^{-1}(t)}{nhf^2(t)} \\ &\to \boldsymbol{\Gamma}^{-1}(t)\psi_K[\gamma(t,t)+\sigma^2(t)]\tau/f(t) = \boldsymbol{\Sigma}_{sparse}(t). \end{aligned}$$

Therefore, it remains to show $x_n = o(nh^2)$ for dense data and $x_n = o(nh)$ for sparse data.

Dense case: Since $n_i \geqslant M_n$ for some $M_n \to \infty$ as $n \to \infty$, then

$$\begin{aligned} \text{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T(l,r)) &= \text{var}\left\{\frac{1}{n_i^2}\sum_{j=1}^{n_i} x_{ijl}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\sum_{j=1}^{n_i} x_{ijr}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\right\} \\ &\leqslant \text{E}\left\{\left[\frac{1}{n_i^2}\sum_{j=1}^{n_i} x_{ijl}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\sum_{j=1}^{n_i} x_{ijr}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\right]^2\right\} \\ &= O(h^4), \end{aligned}$$

which implies $\text{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T) = O(h^4)$ and thus $\sum_{i=1}^n \text{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T) = O(nh^4)$. Hence

$$x_n = \sqrt{\sum_{i=1}^n \text{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)} + nhl_n = O(\sqrt{n}h^2 + nh^3 + h\sqrt{n\log n}) = o(nh^2).$$

Then we have

$$\begin{aligned} n\mathbf{U}_n(t) &= n\mathbf{H}_n^{-1}\mathbf{W}_n\mathbf{H}_n^{-1} \\ &= n\mathbf{H}_n^{-1}\left(\sum_{i=1}^{n}\mathrm{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)+O_p(x_n)\right)\mathbf{H}_n^{-1} \\ &\approx n\mathbf{H}_n^{-1}\left(\sum_{i=1}^{n}\mathrm{E}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)\right)\mathbf{H}_n^{-1}\to\boldsymbol{\Sigma}_{dense}(t). \end{aligned}$$

Therefore, $\mathbf{U}_n(t)^{-1/2}\left[\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)-h^2\boldsymbol{\rho}(t)\right]\to N(\mathbf{0}_{k+1},\mathbf{I}_{k+1})$.

Sparse case: Since $\mathrm{E}(n_i)<\infty$, then we have

$$\begin{aligned} \mathrm{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T(l,r)) &\leqslant \mathrm{E}\left\{\left[\frac{1}{n_i^2}\sum_{j=1}^{n_i}x_{ijl}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\sum_{j=1}^{n_i}x_{ijr}\left[v_i(t_{ij})+\sigma(t_{ij})\epsilon_{ij}\right]K_{ij}\right]^2\right\} \\ &= O(h), \end{aligned}$$

which means that $\sum_{i=1}^{n}\mathrm{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)=O(nh)$. Thus the asymptotic result for sparse case follows from $x_n=\sqrt{\sum_{i=1}^{n}\mathrm{var}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T)}+nhl_n=O(\sqrt{nh}+nh^3+\sqrt{nh\log n})=o(nh)$.

Chapter 3

Future Work: Mixture of Varying Coefficient Models

3.1 Motivation

The varying coefficient models have the following structure:

$$Y = \sum_{j=1}^{p} \beta_j(t) X_j + \sigma^2(t)\epsilon = \mathbf{X}^T \boldsymbol{\beta}(t) + \sigma^2(t)\epsilon, \tag{3.1}$$

where $Y \in \mathcal{R}^1$, $\mathbf{X} = (X_1, \cdots, X_p)^T \in \mathcal{R}^p$, $t \in \mathcal{R}^1$ and $\epsilon \sim N(0,1)$. By allowing $\beta_j(\cdot)$ to be functions of covariate t, we can study how the coefficients change over different values of t, e.g., time and temperature. The varying coefficient model is useful when all subjects obey the same relationship between the response and covariates. However, in some applications, the subjects might come from an heterogeneous population which consists of several homogeneous subpopulations/clusters. For this type of application, a single varying coefficient model of (3.1) is no longer adequate. To analyze such heterogeneous data, we consider the mixture framework.

Our motivation to consider the mixture of varying coefficient models is from the CO_2–GDP

dataset. The CO_2–GDP dataset contains two related variables for 175 countries for the years 1980-2005, the Carbon dioxide (CO_2) emission per capita and the economy size (GDP) per capita. Huang and Yao (2012) studied the relationship between a country's CO_2 emission from its industrial activities and GDP per capita in year 2005 using a 2-component mixture of regression models. The identity of the two components indicates a country's development path, either in a high or a low CO_2 emission way, as compared with its GDP per capita. This is a cross-sectional analysis, and it is of greater interest to ask whether and how the development paths evolve over time, since we have data of the same structure over many years (1980-2005). Finite mixture of varying coefficient models allow us to overcome the challenge of incorporating both functional and heterogeneity features of the data.

Existing literature of mixture of varying coefficient models focus on estimation and applications, and there lacks of comprehensive studies on the asymptotic properties, and theories on testing hypothesis. For example, Lu and Song (2012) proposed a mixture of varying coefficient models to study heterogeneous longitudinal data in medical research.

In the future work, we will systematically investigate mixture of varying coefficient models, where each varying coefficient model component follows the definition of Fan and Zhang (1999). In addition to the varying coefficients and variance functions, the proportions are also nonparametric functions of a covariate. Therefore, the proposed model is fully nonparametric. In Section 3.2, we provide the estimation procedure and an efficient EM algorithm for mixture of varying coefficient models and establish the asymptotic property in Section 3.3. The technical conditions and proofs are relegated to Section 3.4. In addition, we plan to study the hypothesis tests for the varying coefficients in mixture of varying coefficient models in the future.

3.2 Mixture of Varying Coefficient Models

Suppose that observation $\{(\mathbf{x}_i, y_i, t_i), i = 1, \cdots, n\}$ are i.i.d random samples from $\{(\mathbf{X}, Y, T)\}$. Let $\mathcal{C}$ be a latent class variable, and assume that conditioning on $T = t$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = c|T = t) = \pi_c(t)$ for $c = 1, 2, \cdots, C$, where for any t, $\pi_1(t) + \cdots + \pi_C(t) = 1$. Conditioning on $\mathcal{C} = c$ and $T = t$, Y follows a varying coefficient model:

$$y = \mathbf{x}^T \boldsymbol{\beta}_c(t) + \sigma_c(t)\epsilon.$$

where $\mathbf{x} = (x_1, \cdots, x_p)^T$, $\boldsymbol{\beta}_c(\cdot) = (\beta_{c1}(\cdot), \cdots, \beta_{cp}(\cdot))^T$ and $\sigma_c(\cdot)$ are unknown smooth functions, and ϵ follows a standard normal distribution. As the latent class variable $\mathcal{C}$ is not observed, the conditional distribution of Y given $\mathbf{X} = \mathbf{x}, T = t$ can be written as

$$Y|\mathbf{X} = \mathbf{x}, T = t \sim \sum_{c=1}^{C} \pi_c(t) N(\mathbf{x}^T \boldsymbol{\beta}_c(t), \sigma_c^2(t)). \tag{3.2}$$

By considering C a positive integer, we refer to model (3.2) as a *finite mixture of varying coefficient models*. This model can be viewed as a generalization of semiparametric mixture of regression models with the varying mixing proportions (Huang and Yao, 2012), by allowing coefficients and variances to depend on covariate t. It is also a generalization of nonparametric mixture of regression (Huang, et al., 2013), where the one dimensional nonparametric regression function in each component is replaced by a varying coefficient model.

3.3 Preliminary Results

3.3.1 Estimation Procedure

The log-likelihood function of model (3.2) is

$$\sum_{i=1}^{n} \log \left[\sum_{c=1}^{C} \pi_c(t_i) \phi\{Y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c(t_i), \sigma_c^2(t_i)\} \right], \tag{3.3}$$

where $\phi(\cdot|\mu, \sigma^2)$ is the normal density function. In this chapter, we apply kernel regression to estimate the unknown smooth functions $\boldsymbol{\beta}_c(\cdot)$, $\sigma_c(\cdot)$, and $\pi_c(\cdot)$. For any fixed t, we use local constants $\boldsymbol{\beta}_c, \sigma_c^2$, and π_c to approximate $\boldsymbol{\beta}_c(t), \sigma_c^2(t)$, and $\pi_c(t)$, $c = 1, \cdots, \mathcal{C}$. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a rescaled kernel for a kernel function $K(\cdot)$ with a bandwidth $h > 0$. Then, the corresponding local log-likelihood function for data $\{(\mathbf{x}_i, y_i, t_i), i = 1, \cdots, n\}$ is

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi(y_i | \mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\} K_h(t_i - t), \tag{3.4}$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_C^2)^T$, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_{C-1})^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_C^T)^T$, and $\boldsymbol{\beta}_c = (\beta_{c1}, \cdots, \beta_{cp})^T$, $c = 1, \cdots, C$. By using local constant approximation, we have a closed-form solution in the M-step of the proposed EM algorithm, while local linear or higher order do not provide a closed-form solution for σ_c^2 and π_c. The extension from local constant to local linear or higher order is trival and of minor interest. For convenience in theoretical development and computation, we use local constant approximation to estimate all the nonparametric functions.

Computing Algorithms

For a given t, (3.4) is a weighted likelihood of finite mixture model, and thus an EM algorithm is a natural choice to solve (3.4). However, such a pointwise implementation will suffer the mislabel problem, see (Huang, et al., 2013). We use a modified EM algorithm for model estimation. The key of the algorithm is to estimate the common labels in E-step

which do not depend on the choice of t. In the M-step, we update the estimated curves simultaneously at a set of grid points. The modified EM algorithm is described as follows:

E-step: For $i = 1, \cdots, n$ and $c = 1, \cdots, C$, calculate

$$r_{ic} = \frac{\pi_c(t_i)\phi\{y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c(t_i), \sigma_c^2(t_i)\}}{\sum_{c=1}^{C}\pi_c(t_i)\phi\{y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c(t_i), \sigma_c^2(t_i)\}}, \tag{3.5}$$

M-step: For $c = 1, \cdots, C$, and t in a set of grid points, calculate

$$\hat{\pi}_c = \frac{\sum_{i=1}^{n} r_{ic}K_h(t_i - t)}{\sum_{i=1}^{n} K_h(t_i - t)}, \tag{3.6}$$

$$\hat{\boldsymbol{\beta}}_c = (S^T W_c S)^{-1} S^T W_c \mathbf{y}, \tag{3.7}$$

$$\hat{\sigma}_c^2 = (\mathbf{y} - X\hat{\boldsymbol{\beta}}_c)^T W_c (\mathbf{y} - X\hat{\boldsymbol{\beta}}_c)/tr(W_c), \tag{3.8}$$

where $S = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, $w_{ci} = r_{ic}K_h(t_i - t)$, $W_c = \text{diag}\{w_{c1}, \cdots, w_{cn}\}$, a $n \times n$ diagonal matrix and $tr(W_c)$ is the trace of W_c.

The modified EM algorithm is essentially similar to the EM type algorithm (section 2.3.1) in Huang and Yao (2012). Hence, it poses the ascent property in an asymptotic sense; see Theorem 4(a) of Huang and Yao (2012).

3.3.2 Asymptotic Property

Let $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\pi}}\}$ be the solution of maximizing the local likelihood function (3.4). Then we can estimate $\pi_c(t)$, $\boldsymbol{\beta}_c(t)$, and $\sigma_c^2(t)$ using $\hat{\pi}_c$, $\hat{\boldsymbol{\beta}}_c$, and $\hat{\sigma}_c^2$, respectively. In this section we study the asymptotic properties of these estimates. Let $\boldsymbol{\theta}(t) = (\boldsymbol{\beta}(t)^T, (\boldsymbol{\sigma}^2(t))^T, \boldsymbol{\pi}(t)^T)^T$, where $\boldsymbol{\beta}(t) = (\boldsymbol{\beta}_1(t)^T, \cdots, \boldsymbol{\beta}_c(t)^T)^T$, $\boldsymbol{\sigma}^2(t) = (\sigma_1^2(t), \cdots, \sigma_c^2(t))^T$, and $\boldsymbol{\pi}(t) = (\pi_1(t), \cdots, \pi_{C-1}(t))^T$,

and let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, (\hat{\boldsymbol{\sigma}}^2)^T, \hat{\boldsymbol{\pi}}^T)^T$. Denote

$$\rho(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \phi\left\{y|\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2\right\}, \qquad \ell(\boldsymbol{\theta}, \mathbf{x}, y) = \log \rho(y|\mathbf{x}, \boldsymbol{\theta});$$

$$q_\theta(\boldsymbol{\theta}, \mathbf{x}, y) = \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x}, y)}{\partial \boldsymbol{\theta}}, \qquad q_{\theta\theta}(\boldsymbol{\theta}, \mathbf{x}, y) = \frac{\partial^2 \ell(\boldsymbol{\theta}, \mathbf{x}, y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T};$$

$$\boldsymbol{\Lambda}(t|v) = E\{q_\theta(\boldsymbol{\theta}(v), \mathbf{x}, y)|T = t\}, \qquad \mathcal{I}(t) = -E\left[q_{\theta\theta}\{\boldsymbol{\theta}(t), \mathbf{x}, y\}|T = t\right].$$

Theorem 3.3.1. *Suppose that the regularity conditions (A)—(H) in Section 3.4 hold. Then, with probability approaching 1, there exists a consistent local maximizer* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, (\hat{\boldsymbol{\sigma}}^2)^T, \hat{\boldsymbol{\pi}}^T)^T$ *of (3.4) such that*

$$\sqrt{nh}\{\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \mathcal{I}^{-1}(t)[\frac{f'(t)\boldsymbol{\Lambda}'(t|t)}{f(t)} + \frac{1}{2}\boldsymbol{\Lambda}''(t|t)]v_2 h^2 + o_p(h^2)\} \xrightarrow{D} N(\mathbf{0}_m, \tau f^{-1}(t)\mathcal{I}(t)),$$

where $f(\cdot)$ *is the marginal density function of* T, $\tau = \int K^2(t)\,dt$, *and* $v_2 = \int t^2 K(t)\,dt$.

3.4 Proofs

Regularity Conditions

A, The sample $\{(\mathbf{X}_i, Y_i, T_i), i = 1, \cdots, n\}$ is independent and identically distributed from model (3.2).

B, The unknown functions $\boldsymbol{\theta}(t)$ has continuous second derivatives. Furthermore, $\sigma_c^2(t) > 0$, $\pi_c(t) > 0$ and $\pi_1(t) + \cdots + \pi_{\mathcal{C}}(t) = 1$ hold for $c = 1, \cdots, \mathcal{C}$ and all $u \in \mathcal{U}$.

C, The support for T, denoted by $\mathcal{U}$, is closed and bounded of $\mathbb{R}^1$. The marginal density function of T, $f(t)$, is twice continuously differentiable and positive for $t \in \mathcal{U}$.

D, The third derivative $|\partial^3 l(\boldsymbol{\theta}, \mathbf{x}, y, t)/\partial\theta_j\partial\theta_k\partial\theta_l| \leqslant M_{jkl}(\mathbf{x}, y, t)$, where $\mathrm{E}\{M_{jkl}(\mathbf{X}, Y, T)\}$ is bounded for all j, k, l.

E, The following conditions hold for all i and j,

$$\mathrm{E}(|\frac{\partial \ell(\boldsymbol{\theta}, \mathbf{X}, Y)}{\partial \theta_j}|^3) < \infty, \ \mathrm{E}(|\frac{\partial^2 \ell(\boldsymbol{\theta}, \mathbf{X}, Y)}{\partial \theta_i \partial \theta_j}|^2) < \infty.$$

Furthermore, $\mathrm{E}(q_{\theta\theta}(\boldsymbol{\theta}, X, Y)|T = t)$ is continuous in t, and the second derivative matrix, $\mathcal{I}(t)$, is positive definite for $t \in \mathcal{U}$.

F, $\mathrm{E}\{q_\theta(\boldsymbol{\theta}, X, Y) q_\theta^T(\boldsymbol{\theta}, X, Y)|T = t\}$ is continuous in t.

G, The kernel function $K(\cdot)$ has a bounded support, and satisfies that

$$K(t) > 0, \ K(-t) = K(t), \ \int K(t)du = 1.$$

H, $h \to 0$, $nh \to \infty$ as $n \to \infty$.

All these conditions are mild conditions and have been used in the literature of local likelihood estimation. Conditions A - C are basic assumptions in our model. Conditions D - F are similar to the regularity conditions to prove the asymptotic normality of MLEs. Condition G is the definition of kernel and Condition H is a standard assumption in the nonparametric regression.

Proof of Theorem 3.3.1 We will first prove that with probability approaching 1, there exists a consistent local maximizer $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, (\hat{\boldsymbol{\sigma}}^2)^T, \hat{\boldsymbol{\pi}}^T)^T$ of (3.4) such that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p\{(nh)^{-1/2} + h^2\}.$$

Then we establish the asymptotic distributions for such consistent estimate.

Denote $\gamma_n = (nh)^{-1/2} + h^2$, $K_i = K_h(t_i - t)$, and $q_{\theta_j\theta_k\theta_l}(\boldsymbol{\theta}, \mathbf{x}, y) = \frac{\partial^3 \ell(\boldsymbol{\theta}, \mathbf{x}, y)}{\partial\theta_j \partial\theta_k \partial\theta_l}$, where $j, k, l = 1, 2, ..., (pC + 2C - 1)$. We have the following objective function

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c, \sigma_c^2)\right\} K_h(t_i - t) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)K_i.$$

It is sufficient to show that for any given $\eta > 0$, there exists a large constant, a, such that

$$P\left\{sup_{\|\boldsymbol{\mu}\|=a} L(\boldsymbol{\theta} + \gamma_n\boldsymbol{\mu}) < L(\boldsymbol{\theta})\right\} \geq 1 - \eta,$$

where $\boldsymbol{\mu}$ has the same dimension as $\boldsymbol{\theta}$, γ_n is the convergence rate. By using Taylor expansion, it follows that

$$\begin{aligned}
L(\boldsymbol{\theta} + \gamma_n\boldsymbol{\mu}) - L(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} K_i\{\ell(\boldsymbol{\theta} + \gamma_n\boldsymbol{\mu}, \mathbf{x}_i, y_i) - \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)\} \\
&= \frac{1}{n}\sum_{i=1}^{n} K_i\{\gamma_n q_\theta^T(\boldsymbol{\theta}, \mathbf{x}_i, y_i)\boldsymbol{\mu} + \frac{1}{2}\gamma_n^2\boldsymbol{\mu}^T q_{\theta\theta}(\boldsymbol{\theta}, \mathbf{x}_i, y_i)\boldsymbol{\mu} \\
&\quad + \frac{1}{6}\gamma_n^3 \sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} \mu_j\mu_k\mu_l q_{\theta_j\theta_k\theta_l}(\boldsymbol{\xi}, \mathbf{x}_i, y_i)\} \\
&= I_1 + I_2 + I_3,
\end{aligned}$$

where $m = pC + 2C - 1$, $\boldsymbol{\xi}$ is a value between $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \gamma_n\boldsymbol{\mu}$ such that $||\boldsymbol{\xi} - \boldsymbol{\theta}|| \leq \gamma_n a$.

Let $f(\cdot)$ be the marginal density function of T, and $\boldsymbol{\Lambda}(t|v) = \mathrm{E}\{q_\theta(\boldsymbol{\theta}(v), \mathbf{x}, y)|T = t\}$. Note that $\boldsymbol{\Lambda}(t|t) = \mathrm{E}\{q_\theta(\boldsymbol{\theta}(t), \mathbf{x}, y)|T = t\} = 0$. Then for $I_1 = \frac{1}{n}\sum_{i=1}^{n} \gamma_n q_\theta^T(\boldsymbol{\theta}, \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i$, we have the following results.

$$\mathrm{E}(I_1) = \mathrm{E}[\gamma_n q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i] = \frac{\gamma_n}{h}\int \boldsymbol{\Lambda}^T(t_i|t)\boldsymbol{\mu}K(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i = O(\gamma_n a h^2),$$

and

$$\mathrm{Var}(I_1) = \frac{1}{n}\mathrm{Var}[\gamma_n q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i] = \frac{1}{n}\{\mathrm{E}(A^2) - [\mathrm{E}(A)^2]\},$$

where $A = \gamma_n q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i$. Let $\boldsymbol{\Gamma}(t|t_0) = \mathrm{E}[q_\theta(\boldsymbol{\theta}(t_0), \mathbf{x}, y)q_\theta^T(\boldsymbol{\theta}(t_0), \mathbf{x}, y)|t]$. Then

$$\begin{aligned}
\mathrm{E}(A^2) &= \gamma_n^2 \mathrm{E}[\boldsymbol{\mu}^T q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t_0), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i^2] \\
&= \gamma_n^2 \boldsymbol{\mu}^T \mathrm{E}\{\mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t_0), \mathbf{x}_i, y_i)|t_i]K_i^2\}\boldsymbol{\mu} \\
&= \gamma_n^2 \boldsymbol{\mu}^T \mathrm{E}[\boldsymbol{\Gamma}(t_i|t)K_i^2]\boldsymbol{\mu} \\
&= \gamma_n^2 \boldsymbol{\mu}^T \frac{1}{h^2}\{\int \boldsymbol{\Gamma}(t_i|t)K_i^2 f(t_i)\mathrm{d}t_i\}\boldsymbol{\mu} \\
&= O(\frac{\gamma_n^2 a^2}{h}).
\end{aligned}$$

Note that $[\mathrm{E}(A)]^2 = [O(\gamma_n a h^2)]^2 = O(a^2h^4\gamma_n^2) \ll \mathrm{E}(A^2)$, then $\mathrm{Var}(I_1) \approx \frac{1}{n}\mathrm{E}(A^2) = O(\frac{a^2\gamma_n^2}{nh})$. Hence, $I_1 = \mathrm{E}(I_1) + O_p(\sqrt{\mathrm{Var}(I_1)}) = O_p(\gamma_n a h^2) + O_p(\frac{a\gamma_n}{\sqrt{nh}}) = O_p(a\gamma_n^2)$.

For $I_2 = \frac{1}{2n}\sum_{i=1}^n \gamma_n^2 \boldsymbol{\mu}^T q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i$, we have

$$\begin{aligned}
\mathrm{E}(I_2) &= \frac{\gamma_n^2}{2}\mathrm{E}[\boldsymbol{\mu}^T q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\boldsymbol{\mu}K_i] \\
&= \frac{\gamma_n^2}{2}\boldsymbol{\mu}^T \mathrm{E}\{\mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)|t_i]K_i\}\boldsymbol{\mu}.
\end{aligned}$$

Let $\mathbf{S}(t|t_0) = \mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t_0), \mathbf{x}, y)|t]$ and $\mathcal{I}(t) = -\mathbf{S}(t|t) = -\mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}, y)|t]$. Then

$$\begin{aligned}
\mathrm{E}(I_2) &= \frac{\gamma_n^2}{2}\boldsymbol{\mu}^T \mathrm{E}[\mathbf{S}(t_i|t)K_i]\boldsymbol{\mu} \\
&= \frac{\gamma_n^2}{2}\boldsymbol{\mu}^T [\frac{1}{h}\int \mathbf{S}(t_i|t)K(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i]\boldsymbol{\mu} \\
&= -\frac{\gamma_n^2}{2}\boldsymbol{\mu}^T \mathcal{I}(t)f(t)\boldsymbol{\mu}(1 + o(1)).
\end{aligned}$$

Let $\mathbf{B} = \frac{1}{2n}\sum_{i=1}^n q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i$ and denote $\mathbf{B}(j,k)$ be the element in the j^{th} row and k^{th} column of the matrix $\mathbf{B}$. Then $q_{\theta_j\theta_k}(\boldsymbol{\theta}(t), \mathbf{x}, y)$ is the element in the j^{th} row and k^{th} column

of the matrix $q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}, y)$. Let $\delta(t|t_0) = \mathrm{E}[q^2_{\theta_j\theta_k}(\boldsymbol{\theta}(t_0), \mathbf{x}, y)|t]$. It can be shown that

$$\begin{aligned}
\mathrm{Var}(\mathbf{B}(j,k)) &= \frac{1}{4n}\mathrm{Var}[q_{\theta_j\theta_k}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i] \\
&< \frac{1}{4n}\mathrm{E}[q^2_{\theta_j\theta_k}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i^2] \\
&= \frac{1}{4n}\mathrm{E}\{\mathrm{E}[q^2_{\theta_j\theta_k}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)|t_i]K_i^2\} \\
&= \frac{1}{4n}\mathrm{E}[\delta(t_i|t)K_i^2] \\
&= \frac{1}{4nh^2}\int \delta(t_i|t)K^2(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i \\
&= O(\frac{1}{nh}).
\end{aligned}$$

Throughout this chapter, we consider the element-wise variance of a matrix. So, $\mathrm{Var}(\mathbf{B}) = O(\frac{1}{nh})$. Hence, $\mathrm{Var}(I_2) = O(\frac{\gamma_n^4}{nh})$. Based on the result $I_2 = \mathrm{E}(I_2) + O_p(\sqrt{\mathrm{Var}(I_2)})$ and the assumption $nh \to \infty$, it follows that

$$I_2 = -\frac{\gamma_n^2}{2}\boldsymbol{\mu}^T\mathcal{I}(t)f(t)\boldsymbol{\mu}(1 + o(1)) = O_p(\gamma_n^2).$$

Similarly, $I_3 = \frac{\gamma_n^3}{6n}\sum_{i=1}^n\{\sum_{j=1}^m\sum_{k=1}^m\sum_{l=1}^m \mu_j\mu_k\mu_l q_{\theta_j\theta_k\theta_l}(\boldsymbol{\xi}, \mathbf{x}_i, y_i)\}K_i = O_p(\gamma_n^3)$.

Noticing that $\mathcal{I}(t) = -\mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}, y)|t] = -\mathrm{E}\{\frac{\partial^2\ell(\boldsymbol{\theta}(t), \mathbf{x}, y)}{\partial\boldsymbol{\theta}(t)\partial\boldsymbol{\theta}(t)^T}\}$ is a positive matrix, $||\boldsymbol{\mu}|| = a$, we can choose a large enough such that I_2 dominates both I_1 and I_3 with probability at least $1 - \eta$. Thus, $P\left\{sup_{\|\boldsymbol{\mu}\|=a}L(\boldsymbol{\theta} + \gamma_n\boldsymbol{\mu}) < L(\boldsymbol{\theta})\right\} \geq 1 - \eta$. Hence with probability approaching 1, there exists a local maximizer $\hat{\boldsymbol{\theta}}$ such that $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| \leq \gamma_n a$, where $\gamma_n = (nh)^{-1/2} + h^2$. Therefore, with probability approaching 1, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p((nh)^{-1/2} + h^2)$.

Next, we provide the asymptotic distribution for such consistent estimate. Since $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$, then $L'(\hat{\boldsymbol{\theta}}) = 0$. By Taylor expansion,

$$0 = L'(\hat{\boldsymbol{\theta}}) = L'(\boldsymbol{\theta}) + L''(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}L'''(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2, \tag{3.9}$$

where $\tilde{\boldsymbol{\theta}}$ is a value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Then $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = -[L^{''}(\boldsymbol{\theta})]^{-1}L^{'}(\boldsymbol{\theta})(1 + o_p(1))$. Since $L^{''}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} K_i = \frac{1}{n}\sum_{i=1}^{n} q_{\theta\theta}(\boldsymbol{\theta}, \mathbf{x}_i, y_i)K_i$, then we have

$$\begin{aligned}
\mathrm{E}[L^{''}(\boldsymbol{\theta})] &= \mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i] \\
&= \mathrm{E}\{\mathrm{E}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)|t_i]K_i\} \\
&= \mathrm{E}[\mathbf{S}(t_i|t)K_i] \\
&= \frac{1}{h}\int \mathbf{S}(t_i|t)K(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i \\
&= -\mathcal{I}(t)f(t)(1 + o(1)),
\end{aligned}$$

and $\mathrm{Var}[L^{''}(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{Var}[q_{\theta\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i] = O(\frac{1}{nh})$. Based on the result $L^{''}(\boldsymbol{\theta}) = \mathrm{E}[L^{''}(\boldsymbol{\theta})] + O_p\{\sqrt{\mathrm{Var}[L^{''}(\boldsymbol{\theta})]}\}$ and the assumption $nh \to \infty$, it follows that

$$L^{''}(\boldsymbol{\theta}) = -\mathcal{I}(t)f(t)(1 + o(1)).$$

The asymptotic result is determined by $L^{'}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)}{\partial \boldsymbol{\theta}} K_i = \frac{1}{n}\sum_{i=1}^{n} q_{\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i$. We have

$$\begin{aligned}
\mathrm{E}[L^{'}(\boldsymbol{\theta})] &= \mathrm{E}[q_{\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i] \\
&= \mathrm{E}\{\mathrm{E}[q_{\theta}(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)|t_i]K_i\} \\
&= \mathrm{E}[\boldsymbol{\Lambda}(t_i|t)K_i] \\
&= \frac{1}{h}\int \boldsymbol{\Lambda}(t_i|t)K(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i \\
&= h^2 f(t)[\frac{f^{'}(t)\boldsymbol{\Lambda}^{'}(t|t)}{f(t)} + \frac{1}{2}\boldsymbol{\Lambda}^{''}(t|t)]v_2(1 + o(1)),
\end{aligned}$$

where $v_2 = \int t^2 K(t)\mathrm{d}t$.

$$\begin{aligned}
\mathrm{Var}[L^{'}(\boldsymbol{\theta})] &= \frac{1}{n}\mathrm{Var}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i] \\
&= \frac{1}{n}\{\mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i^2] - \mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i]\mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i]^T\} \\
&= \frac{1}{n}\{\mathrm{E}\{\mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)|t_i]K_i^2\} - O(h^4)\} \\
&= \frac{1}{n}\{\mathrm{E}[\boldsymbol{\Gamma}(t_i|t)K_i^2] - O(h^4)\} \\
&= \frac{1}{n}\{\frac{1}{h^2}\int \boldsymbol{\Gamma}(t_i|t)K^2(\frac{t_i - t}{h})f(t_i)\mathrm{d}t_i - O(h^4)\} \\
&= \frac{1}{n}\{\frac{1}{h}\boldsymbol{\Gamma}(t|t)f(t)\tau(1 + o(1)) - O(h^4)\} \\
&= \frac{1}{nh}\boldsymbol{\Gamma}(t|t)f(t)\tau(1 + o(1)),
\end{aligned}$$

where $\tau = \int K^2(t)\mathrm{d}t$. We next use the Lyapunov central limit theorem to obtain the asymptotic distribution of $L^{'}(\boldsymbol{\theta})$. The Lyapunov conditions are checked as follows.

For any unit vector $\mathbf{d} \in \mathbb{R}^m$, where $m = pC + 2C - 1$, let $\mathbf{d}^T L^{'}(\boldsymbol{\theta}) = \sum_{i=1}^n \zeta_i$, where $\zeta_i = \frac{1}{n}\mathbf{d}^T q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i$. Since

$$\mathrm{E}(\zeta_i^2) = \frac{1}{n^2}\mathbf{d}^T\mathrm{E}[q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i^2]\mathbf{d} = O(\frac{1}{n^2 h}),$$

and

$$\mathrm{E}(\zeta_i^3) = \mathrm{E}\{\frac{1}{n^3}\mathbf{d}^T q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)q_\theta^T(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)\mathbf{d}\mathbf{d}^T q_\theta(\boldsymbol{\theta}(t), \mathbf{x}_i, y_i)K_i^3\} = O(\frac{1}{n^3 h^2}),$$

then $(\sum_{i=1}^n \mathrm{E}|\zeta_i|^2)^3 = O(\frac{1}{n^3 h^3})$ and $(\sum_{i=1}^n \mathrm{E}|\zeta_i|^3)^2 = O(\frac{1}{n^4 h^4})$. Note that $\frac{1}{n^4 h^4}(nh)^3 = \frac{1}{nh} \to 0$, so $\frac{1}{n^4 h^4} = o(\frac{1}{n^3 h^3})$, which is equivalent to say $(\sum_{i=1}^n \mathrm{E}|\zeta_i|^3)^2 = o((\sum_{i=1}^n \mathrm{E}|\zeta_i|^2)^3)$. Based on

Lyapunov central limit theorem,

$$\frac{L^{'}(\boldsymbol{\theta}) - \mathrm{E}[L^{'}(\boldsymbol{\theta})]}{\sqrt{\mathrm{Var}[L^{'}(\boldsymbol{\theta})]}} \xrightarrow{D} N(\mathbf{0}_m, \mathbf{I}_m),$$

where $\mathbf{0}_m$ is a $m \times 1$ vector with each entry being 0 and $\mathbf{I}_m$ is a $m \times m$ identity matrix. Previously, we already computed that $\mathrm{Var}[L^{'}(\boldsymbol{\theta})] = \frac{1}{nh}\boldsymbol{\Gamma}(t|t)f(t)\tau(1 + o(1))$, by Slutsky's Theorem,

$$\sqrt{nh}\{L^{'}(\boldsymbol{\theta}) - \mathrm{E}[L^{'}(\boldsymbol{\theta})]\} \xrightarrow{D} N(\mathbf{0}_m, \boldsymbol{\Gamma}(t|t)f(t)\tau).$$

By the Condition F we have $\mathcal{I}(t) = \boldsymbol{\Gamma}(t|t)$. Hence, based on (3.9), we have the following result

$$\sqrt{nh}\{\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) - \mathcal{I}^{-1}(t)[\frac{f^{'}(t)\boldsymbol{\Lambda}^{'}(t|t)}{f(t)} + \frac{1}{2}\boldsymbol{\Lambda}^{''}(t|t)]v_2h^2 + o_p(h^2)\} \xrightarrow{D} N(\mathbf{0}_m, \tau f^{-1}(t)\mathcal{I}(t)).$$

Bibliography

Beran, R. (1974). Asymptotic efficient adaptive rank estimates in location models. *The Annals of Statistics*, 2, 63-74.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, 10, 647-671.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95, 888-902.

Cao, G., Yang, L. and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, 24, 359-377.

Chiang, C-T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96, 605-619.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (J.M. Chambers and T.J. Hastie, eds.), pp. 309-376. Wadsworth/Brooks-Cole, Pacific Grove, CA.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, 21, 1735-1765.

Drost, F. C. and Klaassen, C. A. J. (1997). Efficient estimation in semiparametric GRACH models. *Journal of Econometrics*, 81, 193-221.

Eubank, R. L., Huang, C. F., Maldonado, Y. M., Wang, N., Wang, S. and Buchanan, R. J. (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society*, Ser. B, 66, 653-667.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11, 1031-1057.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, 29, 153-193.

Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society*, Ser. B, 62, 303-322.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27, 1491-1518.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1), 179-195.

Gilley, O. W. and Pace, R. K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31, 403-405.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society*, Ser. B, 55, 757-796.

Hodgson, D. J. (1998). Adaptive estimation of cointegrating regressions with ARMA errors. *Journal of Econometrics*, 85, 231-267.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 809-822.

Huang, M., Li, R., and Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108, 929-941.

Huang, J. Z. and Shen H. (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics*, 31, 515-534.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111-128.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14, 763-788.

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107(498), 711-724.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. (1987). The multicenter AIDS cohort study - rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, 126, 310-318.

Kauermann G. and Tutz G. (1999). On model diagnostics using varying coefficient models. *Biometrika*, 86, 119-128.

Kim, S. and Zhao, Z. (2013). Unified inference for sparse and dense longitudinal models. *Biometrika*, 100, 1, 203-212.

Linton, O. and Xiao, Z. (2007). A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory*, 23, 371-413.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38, 3321-3351.

Lu, Z. and Song, X. (2012). Finite mixture varying coefficient models for analyzing longitudinal heterogeneous data. *Statistics in Medicine*, 31(6), 544-560.

Manski, C. F. (1984). Adaptive estimation of non-linear regression models. *Econometric Reviews*, 3, 145-194.

Ma, S., Yang, L. and Carroll, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica*, 22, 95-122.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62, 379-391.

Raykar, V. C. and Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. In proceedings of the sixth *SIAM International Conference on Data Mining*, Bethesda, April 2006, 524-528.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society*, Ser. B, 53, 233-243.

Schick, A. (1993). On efficient estimation in regression models. *The Annals of Statistics*, 21, 1486-1521.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistical Society*, B, 53, 683-690.

Steigerwald, D. G. (1992). Adaptive estimation in time series regression models. *Journal of Econometrics*, 54, 251-276.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3, 267-284.

Wang, L., Kai, B. and Li, R. (2009). Local rank inference for varying coefficient models. *Journal of the American Statistical Association*, 104, 1631-1645.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104, 747-757.

Wang, Q. and Yao, W. (2012). An adaptive estimation of MAVE. *Journal of Multivariate Analysis*, 104, 88-100.

Wu, C. O. and Chiang, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 10, 433-456.

Wu, C. O., Chiang, C-T. and Hoover, D. R. (1998). Asymptotoc Confidence Regions for Kernel Smoothing of a Varying Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, 93, 1388-1402.

Wu, H. and Zhang, J. T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97, 883-897.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society*, Ser. B, 64, 363-388.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577-590.

Yao, W. (2013). A note on EM algorithm for mixture models. *Statistics and Probability Letters*, 83, 519-526.

Yuan, A. and De Gooijer, J. G. (2007). Semiparametric regression with kernel error model. *Scandinavian Journal of Statistics*, 34, 841-869.

Yuan, A. (2009). Semiparametric inference with kernel likelihood. *Journal of Nonparametric Statistics*, 21, 207-228.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50, 689-699.

Zhang, J. T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35, 1052-1079.