Functional Data Analysis for Sparse Functional Data

Yin Zhang

M.S., Zhejiang University, P.R.China, 2011
B.S., Zhejiang University, P.R.China, 2009

A Dissertation Presented to the Graduate Faculty
of University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May, 2018

_____

_____

_____

_____

Abstract

With the development of science and modern technology, more and more data are being collected continuously over a time interval in various disciplines, such as public health, biology, medicine and finance. Such data can be viewed as "functional data". Functional data analysis (FDA), which deals with the analysis and theory of functional data, has been receiving increasing popularity over the past decades. In this dissertation, we propose several functional data analysis methods and apply them to NIH cohort study, which is a study in the field of growth modeling.

It is well known that early year catch-down growth is highly prevalent in developing countries for the reason of malnutrition (Black et al. [2008]). Children who suffers from malnutrition in the first 5 years of life will be at increasing risk for the development in cognitive and physical growth. Therefore, characterizing the catch-down growth and identifying the associate important risk factors is one of the most popular topics. In our study, we aim to investigate the relationship between height-for-age Z score (HAZ) at year 3 and a collection of predictors. However, we meet two problems. First, all functional predictors are sparsely and irregularly observed, that is, the measurement time varies from individual to individual. Functional predictors over the entire time interval must be estimated in order to perform the regression. In addition, some predictors, such as height, should be monotone over time, and a non-monotone estimation of height would make no sense. Secondly, the relationship between the response and functional predictors is not usually linear. Furthermore, here exists outliers in the response.

To address the first problem, we propose a new method based on a monotone transformation, functional principal component (FPC) analysis and a penalized regression to estimate monotone functions for sparse growth data. We also prove

the asymptotic properties for this proposed estimator. Extensive numerical studies show that our proposed method outperforms the existing methods in terms of model fitting and monotonicity of the estimation. In addition, the proposed method can also be utilized as a data preprocessing procedure for other methods, such as functional clustering and classification, where the functional predictors are required to be completely known.

To address the second problem, we build a functional single index model for the non-linear relationship between response and functional predictors. The functional single index model is not only flexible but also interpretable. To deal with outliers, we propose a local modal regression (LMR) (Yao et al. [2012]) based estimation method. We show that by using the optimal bandwidth, the LMR estimator is not only robust when there are outliers or the error distribution is heavy tailed, but also asymptotically as efficient as the ordinary least squares based estimator when the error distribution is a Gaussian distribution. In addition, we conduct extensive simulation studies to demonstrate the robustness and efficiency of the resulting estimator by comparing it with least squares estimator and Huber estimator across different error distributions.

Acknowledgements

Most importantly, I want to thank my parents for their love and support. Without them, I would end up nowhere.

I want to give my deepest thanks to Prof. Jianhui Zhou, my advisor, for his enduring guidance, understanding, patience, and most importantly, his friendship during my graduate studies in University of Virginia (UVa).

Prof. Zhou was always very nice to help me out when I was stuck by tough problems in my research. He was usually able to identify where the problem was, and more importantly, motivate me to dive into the logic and critical thinking behind the problem, thereby to explore some new directions that I could work on, which is especially helpful for me as a beginner. He also advised me to learn how to do the time management when I was occupied by overwhelming tasks. Moreover, Prof. Zhou encouraged me to communicate with people with different background to know the world well. All of these are very helpful not only to the PhD life in UVa but also the future.

Besides my advisor, I would like to thank two of my thesis committee: Prof. Dan Spitzner, Prof. Tingting Zhang. I took many of their courses, from which I learned and benefited a lot. I still remember they could explain the most abstract topics in plain language, and to make even an entry-level student to think deeply as an expert. They were also very nice to provide general advice as well as specific solutions when I consulted to them about the courses and research.

I am especially grateful to Dr. William Petri and Prof. Jennie Ma. They were my supervisors when I was a research assistant at UVa medicine school. We have been working on several projects in the field of public health for more than 2 years. They were very kind to provide me with possible directions that I can work on in

the research. They are also very supportive to my thesis, and my thesis topic is derived from the projects I worked with them. Without them, I would not have the opportunity to do the statistical research in the field of public health.

I also want to thank Prof. Karen Kafadar, Prof. Daniel Keenan and Prof. Chao Du. Their expertise in statistics, as well as their kindness in helping students, have made my studies much more efficient and easier. Thanks to my fellow students in the department. My days would be less enjoyable without their vitality and enthusiasm.

My gratitude also goes to Mrs. Karen Dalton. She has been like a mother, being there when I need her most, reminding me what I have to take care of, and trying every effort to make the department more like a home.

Finally, I would like to thank all people who have helped me during my PhD study at the University of Virginia.

Contents

Abstract ii

Acknowledgements iv

1 Introduction **1**

1.1 Overview of Methods of Estimating Monotone Functions 1

1.2 Overview of Functional Regression Models, Variable Selection and Robust Techniques 3

1.2.1 Overview of Functional Regression Models 3

1.2.2 Overview of Variable Selection 5

1.2.3 Introduction to Local Modal Regression 11

1.2.4 Introduction to B-spline Basis Function 15

1.3 Motivations, Contributions and Organization of the Dissertation . . . 16

2 Estimating Monotone Functions for Sparse Growth Data **20**

2.1 Introduction 20

2.2 Model 22

2.2.1 Framework 22

2.2.2 Model Estimation 23

2.3 Asymptotic Properties 31

2.4 Simulation 33

2.5 Application 36

3 Robust and Efficient Variable Selection for Functional Single Index Model with a Scalar Response 39

3.1 Introduction 39

3.2 Functional Single Index Model 41

3.3 Robust and Efficient Variable Selection for FSIM 42

3.3.1 Local Modal Estimation 42

3.3.2 Algorithm 44

3.4 Asymptotic Properties 47

3.5 Relative Efficiency, Tuning Parameters and Influence Function 51

3.5.1 Bandwidth selection and relative efficiency 51

3.5.2 Tuning parameter in practice 54

3.5.3 Influence function 56

3.6 Simulation Studies 59

3.7 An Application in Growth Modeling 61

4 Proofs 68

4.1 Proof of Theorems in Chapter 2 68

4.2 Notations of Chapter 3 75

4.3 Proof of Theorems in Chapter 3 76

5 Conclusions 91

Chapter 1

Introduction

1.1 Overview of Methods of Estimating Monotone Functions

Methods that estimating monotone functions have played an increasingly important role in data analysis. One reason is that some functions such as people's height should be assumed monotone over time. Another reason is that monotonicity is a useful way of regularizing or stabilizing estimated functions since it can remove the small scale wiggles that are common near the boundaries (Ramsay [1998]).

Various methods for estimating smooth monotone functions have been developed over past decades. Ramsay [1988] estimated monotone functions by using linear combinations of monotone regression splines. They imposed a constraint that the coefficients must be non-negative, which indicates that the estimation procedure involves linear inequality constraints. Moreover, it is sufficient to constrain coefficients to be non negative to guarantee monotonicity, but not necessary. Therefore, there is always room for improving the estimation by allowing negative coefficients

while still preserving monotonicity. Kelly and Rice [1990] developed a related approach involving constrained optimization with respect to linear combinations of regression splines. Bloch and Silverman [1997] applied a monotone transformation technique to the estimation of discriminant functions.

However, in many applications, functional data are collected at sparse and irregular time points. The measurement times are usually different across individuals, and it is demanding to have the individual function estimated monotonely over the entire time interval. This problem can not be handled by the preceding methods because those methods can only estimate the function within individual time interval, that is, from individual starting time to ending time. To address this problem, another line of work have been established, which leverages the pooled information to estimate the entire functions, such as Yao et al. [2005] and Goldsmith et al. [2011]. In general, these methods assume the observed functions are iid realizations of a random process, each function can borrow the information from the properties of this process. A major downside of these methods is that they can not guarantee to produce the monotone functions, so they can not be directly used in application where variables are assumed to be monotone. For example, it does not make any sense to have a non monotone estimated function for people's height.

Motivated by these issues, we develop a procedure of estimating smoothing monotone functions based on a monotone transformation and the functional principal (FPC) analysis, in which FPC scores are obtained by a penalized regression. We demonstrate that this extends applicability of existing monotone estimation methods to situations where data are sparsely and irregularly observed. Numerical studies show that our proposed method outperforms the existing methods in terms of model fitting and monotonicity of the estimation. In addition, the proposed procedure can also be utilized as a data preprocessing step for other methodology such as func-

tional regression, clustering and classification, where the functional predictors are required to be completely known. An illustrative comparison of proposed method and existing methods can be found in Fig 2.2.

1.2 Overview of Functional Regression Models, Variable Selection and Robust Techniques

1.2.1 Overview of Functional Regression Models

With the development of science and modern technology, more and more data are being collected continuously over a time interval or intermittently at dense discrete time points in various disciplines, such as biology, medicine and finance. Such data can be viewed as "functional data", assuming that they are realizations from an underlying continuous process, although only being observed at discrete time points. Functional data are intrinsically infinite dimensional, which poses challenges for both theoretical development and numerical computation. On the other hand, the high or infinite dimensional structure of the data is a rich source of information, which brings many opportunities for retrieving information from the data.

In practice, researchers are often concerned with the relationship between a response and multiple predictors. Consequently, regression models receive great popularity and are broadly used. When the predictors and/or the response are in the form of functions, the analysis is referred to as functional regression. Based on data types of the response and predictors, functional regression models can be divided into three categories: i) scalar-on-function regression; ii) function-on-scalar regression and iii) function-on-function regression. In this dissertation, we focus on the scalar-on-function regression, and the developed methods can be generalized to the

other two cases. One popular functional regression model is the functional linear model (FLM) (Ramsay [2006]), where the conditional mean of response y, given the functional predictors $\boldsymbol{X}(\cdot) = (X_1(\cdot), ..., X_p(\cdot))^T$, is an integral of a linear combination of $\boldsymbol{X}(\cdot)$. Without loss of generality, we assume that both $\boldsymbol{X}(\cdot)$ and y have zero mean. The model is expressed as

$$E(y|\boldsymbol{X}) = \int \boldsymbol{\beta}^T(t)\boldsymbol{X}(t)dt, \tag{1.1}$$

where $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), ..., \beta_p(\cdot))^T$ is the p-dimensional coefficient function. The linear relationship assumed in model (1.1) is too restricted to hold in many applications. Müller and Stadtmüller [2005] proposed a more flexible generalized functional linear model (GFLM) to accommodate nonlinear relationship

$$E(y|\boldsymbol{X}) = g\left(\int \boldsymbol{\beta}^T(t)\boldsymbol{X}(t)dt\right), \tag{1.2}$$

where the relationship between the scalar response and the functional predictors is characterized by a known link function $g(\cdot)$. The GFLM is useful when the linearity assumption is violated, and can be applied to logistic or Poission regression, where the response is binary or integers. Li et al. [2012] proposed GFLM with semi-parametric single-index interactions, to handle the case where there are multiple covariates, a finite number of latent features in the functional predictor, and their interactions. However, those methods are at risk of misspecifying the model with a pre-selected link function. To gain more flexibility, Chen et al. [2011] extended the GFLM to the functional single index model (FSIM) by using an unknown link function g while keeping the effects of covariate functions the same as in (1.2), and estimating the link function g by nonparametric kernel smoothing. In addition, they

further generalized the single index model to multiple indice model, which makes the model even more flexible. In a recent work of Ma [2016], FSIM was also studied by employing B-spline basis functions to approximate the coefficient functions and the link function under the quadratic loss function. They provided uniform convergence rates of the proposed spline estimators, and constructed asymptotic simultaneous confidence bands for the coefficient functions.

When g is unknown, the single index model can be further extended to multiple indices such as Chen et al. [2011], where the number of indices could be unknown. Such "multiple functional index models" typically relax the additive error structure and are expressed as

$$y = g\left(\int \boldsymbol{\beta}_1^T(t)\boldsymbol{X}(t)dt, ..., \int \boldsymbol{\beta}_m^T(t)\boldsymbol{X}(t)dt, \epsilon\right) \tag{1.3}$$

with an unknown multivariate function g on $\mathbb{R}^{p+1}$. This line of research follows the paradigm of sufficient dimension reduction, which was first proposed in Li [1991], known as sliced inverse regression (SIR) for multivariate data. More recently this line of approaches has been extended to longitudinal data in Jiang et al. [2014] and to functional data in Ferré and Yao [2003]; Ferré and Yao [2005]; Cook et al. [2010].

1.2.2 Overview of Variable Selection

When the dimension of predictors is high, the least square estimator and maximum likelihood estimator suffer from high collinearity among predictors, which would further cause unstable estimation and statistical accuracy. In addition, dense structure in the resulting estimator will raise difficulty to distinguish the important variables from unimportant ones, which will further prohibit model interpretation. Moreover, higher dimensionality usually requires more computation cost. Therefore, high di-

mensionality simultaneously brings the challenges of the statistical accuracy, model interpretability, and computational complexity, which are three important pillars of any statistical procedures (Fan and Lv [2010]).

Variable selection is born to overcome these challenges by producing a sparse model, which is able to achieve statistical accuracy, model interpretability, and low computational complexity simultaneously. Suppose we have n independent and identically distributed (iid) observations $(\boldsymbol{x_i}, \boldsymbol{y_i})_{i=1}^n$ from population $(\boldsymbol{X}, y)$ and it is usually assumed that $E(y|X) = \boldsymbol{\beta}^T \boldsymbol{X}$, with regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. In variable selection literatures, the p-dimensional regression coefficients are assumed to be sparse, i.e, some β_j's are zero. The nonzero coefficient indicates that the associated predictor is important. Variable selection aims to identify all important variables whose coefficients do not vanish and to provide effective estimates of those coefficients.

Classical Variable Selection

Earlier variable selection methods focus on comparing models with different subsets of predictors and choosing the best one under some criterion. One important criterion is the Akaike information criterion (AIC) proposed in Akaike [1973] when considering to choose a model that minimizes the Kullback-Leibler (KL) divergence of the fitted model from the true model. Using the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$, Akaike [1973] shows that, the estimated KL divergence is asymptotically equivalent to

$$-l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda I(\boldsymbol{\beta}_j \neq 0) \tag{1.4}$$

with $\lambda = 1$. Another important criterion is the Bayesian information criterion (BIC) proposed in Schwarz et al. [1978], which is equivalent to minimizing (1.4) with $\lambda = (\log n)/2$. Indeed, the approaches of variable selection by AIC and BIC can be unified as minimizing

$$-l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda ||\boldsymbol{\beta}||_0 \tag{1.5}$$

where $||\boldsymbol{\beta}||_0$ denotes the L_0 norm of $\boldsymbol{\beta}$, i.e., the number of the nonzero components of $\boldsymbol{\beta}$.

If the normality joint distribution assumption is imposed, the maximum likelihood estimator coincides with the least square estimator. Then the criterion Mallows's $C_p = RSS_d/s^2 + 2d - n$, in Mallows [1973] corresponds to (1.5) with $\lambda = 1$, where s^2 is the mean squared error of the full model and RSS_d is the residual sum of squares of the best subset with d variables. Other criteria include the generalized cross-validation (GCV) given by $RSS_d/(1 - d/n)^2$, cross-validation (CV), and the risk inflation criterion (RIC) in Foster and George [1994]. For more discussions on regularization, see Bickel et al. [2006] and Fan and Lv [2010]. The classical variable selection methods suffer from its NP computational complexity, especially when the dimension is high.

Variable Selection by Penalization

Recent variable selection methods focus on penalized methods, which aims to minimizing

$$-l_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(\boldsymbol{\beta}_j) \tag{1.6}$$

which substitutes the L_0 penalty in (1.5) with a general penalty functions $p_\lambda(\cdot)$. By choosing an appropriate $p_\lambda(\cdot)$, one is able to produce sparse estimators with desirable properties. A natural generalization of the penalized L_0 regression is the penalized L_q regression, called bridge regression in Frank and Friedman [1993], in which $p_\lambda(\beta) = \lambda|\beta|^q$, for $0 < q \leqslant 2$. The bridge regression bridges the best subset section (penalized L_0) and ridge regression (penalized L_2), and includes the well known penalized L_1 regression as a special case. The penalized L_1 regression is also called Least Absolute Shrinkage and Selection Operator (LASSO) in Tibshirani [1996]. LASSO has received great popularity because it can conduct model estimation and automatic variable selection simultaneously. However, it is also noticed for some limitations of neither being able to select group variables nor to handle the cases $n < p$, as described in Zou and Hastie [2005]. To address this issue, Zou and Hastie [2005] proposed a L_1 and L_2 hybrid penalty, known as elastic net, which is able to do automatic variable selection and continuous shrinkage. It also can select groups of correlated variables, even if $n < p$. In addition, Zou [2006] proposed adaptive LASSO, which reduces the penalty when the initial estimate is large, which is able to reduce the estimation bias for large coefficients. Additionally, adaptive LASSO possesses the Oracle property (Fan and Li [2001]).

There are also many nonconvex penalty functions in the existing variable selection literatures. Such penalty function is usually defined via its first derivative. For example, Fan and Li [2001] introduce the smoothly clipped absolute deviation (SCAD) by considering the following penalized least square problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \tag{1.7}$$

They pointed out that a good penalty function should result in an estimator with

three nice properties

- *Unbiasedness:* The estimator is nearly unbiased when the true unknown parameter is large;
- *Sparsity:* The estimator is a thresholding rule, which automatically sets small estimated coefficients to zero;
- *Continuity:* The estimator is continuous to avoid instability in model prediction.

As argued in Fan and Li [2001], the sufficient conditions for the penalty function $p_{\lambda_n}(|\beta_j|)$ to achieve the aforementioned three properties are *1)* $p'_{\lambda_n}(|\beta_j|) = 0$ *for large* $|\beta_j|$, *2) the minimum of the function* $|\beta_j| + p'_{\lambda_n}(|\beta_j|)$ *is attained at 0*, and *3)* $p_{\lambda_n}(|\beta_j|)$ *is singular at the origin.* Accordingly they proposed the following SCAD penalty function

$$p_{\lambda_n}(|\beta_j|) = \begin{cases} \lambda_n|\beta_j| & |\beta_j| \leqslant \lambda_n \\ -(|\beta_j|^2 - 2a\lambda_n|\beta_j| + \lambda_n^2)/(2(a-1)) & \lambda_n < |\beta_j| \leqslant a\lambda_n \\ (a+1)\lambda_n^2/2 & |\beta_j| > a\lambda_n \end{cases} \tag{1.8}$$

with its first derivative

$$p'_{\lambda_n}(|\beta_j|) = \begin{cases} \lambda_n & |\beta_j| \leqslant \lambda_n \\ (a\lambda_n - |\beta_j|)/(a-1) & \lambda_n < |\beta_j| \leqslant a\lambda_n \\ 0 & |\beta_j| > a\lambda_n \end{cases} \tag{1.9}$$

The plots of the penalty function with its first derivative are displayed in Figure 1.1. It can be seen that $p_{\lambda_n}(|\beta_j|)$ is not differentiable at 0 with respect to β_j, thus it is not easy to minimize the penalized least squares functions due to its singularity.

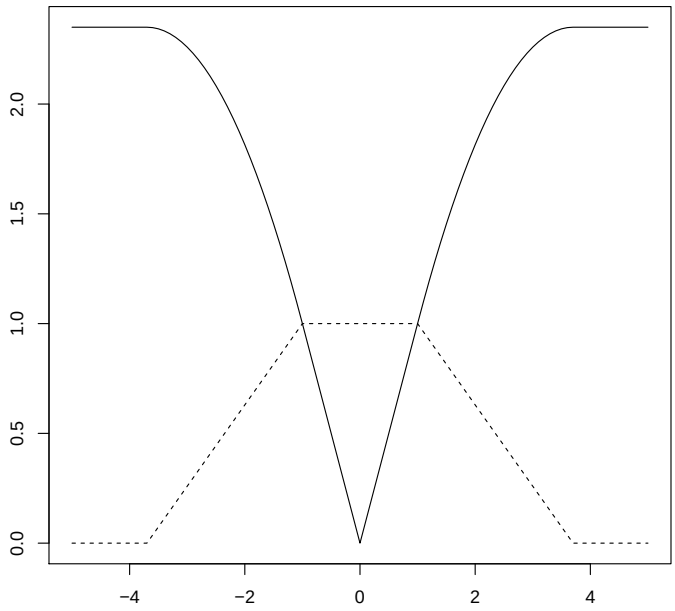

Figure 1.1: SCAD penalty function (solid line) and its derivative (dotted line) $a = 3.7$ and $\lambda = 1$

Fan and Li [2001] suggested to approximate the penalty function by a quadratic function as

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\beta_{j0}|) + \frac{p'_{\lambda_n}(|\beta_{j0}|)}{2|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2) \qquad \text{for } \beta_j \approx \beta_{j0} \tag{1.10}$$

Hence, given a good initial value $\boldsymbol{\beta}^0 = (\beta_{10}, ..., \beta_{p0})^T$, such as the unpenalized least square estimator, the penalized estimator can be obtained via the following iterative equation until convergence.

$$\boldsymbol{\beta}^{t+1} = (\boldsymbol{X}^T\boldsymbol{X} + 2n\boldsymbol{\Sigma}_n^t)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{1.11}$$

where $\boldsymbol{\Sigma}_n^t = \text{diag}\left(p'_{\lambda_n}(|\beta_{1t}|), ..., p'_{\lambda_n}(|\beta_{pt}|)\right)$, $t = 0, 1, 2, ...$

The SCAD method can be generalized to group variable selection by borrowing

the ideas of group LASSO (Yuan and Lin [2006]). If we replace the scalar parameter in (1.7) with vector parameter, we can get the objective function with the group SCAD penalty

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}||\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{X}_j\boldsymbol{\beta}_j||_2^2 + \sum_{j=1}^{p} p_{\lambda_n}(||\boldsymbol{\beta}_j||_2) \tag{1.12}$$

where $||\cdot||_2$ is the L_2 norm. The hybrid structure of L_1 norm and L_2 norm in the group SCAD penalty will result in either shrinking all the variables in a group to zero or keeping all of them nonzero, i.e., group variable selection.

Another similar work is the minimax concave penalty (MCP) in Zhang [2010], whose derivative is given by

$$p'_{\lambda}(t) = (a\lambda - t)_+/a \tag{1.13}$$

It is easy to see that MCP translates the flat part of the derivative of SCAD to the origin, while SCAD takes off at the origin and then gradually levels off.

1.2.3 Introduction to Local Modal Regression

In many fields such as biology, finance, economics and others, it is common that the data distributions have outliers or are heavily tailed or/and skewed. In these cases, the least square estimator may not perform well due to its sensitivity to outliers or deviation from normality. To address this issue, the M-estimator has been investigated in literatures, see Hardle and Gasser [1984], Tsybakov [1986], Fan et al. [1994] and Fan and Jiang [2000] among others. However, these methods lose efficiency when there are no outliers and the error distribution is normal. Motivated by this, Yao et al. [2012] proposed a local modal regression (LMR) procedure for

nonparametric regression model to achieve both robustness and efficiency.

Specifically, Yao et al. [2012] considered a univariate nonparametric regression problem

$$Y = m(X) + \epsilon, \tag{1.14}$$

where $\mathrm{E}(\epsilon|X = x) = 0$, $\mathrm{Var}(\epsilon|X = x) = \sigma^2(x)$, and $m(\cdot)$ is an unknown smooth function to be estimated. According to the local polynomial regression, $m(x) = \mathrm{E}(Y|X = x)$ can be locally approximated by a polynomial regression, that is, for x in a neighborhood of x_0,

$$m(x) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x_0)}{j!}(x - x_0)^j \equiv \sum_{j=0}^{p} \beta_j (x - x_0)^j \tag{1.15}$$

where $\beta_j = m^{(j)}(x_0)/j!$. Supposing (x_i, y_i), $j = 1, ..., n$ are independent and identically distributed observations of (X, Y) from the regression model (1.14), the LMR estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ can be obtained by maximizing the following function

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\phi_{h_2}\left(y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j\right) \tag{1.16}$$

with respect to $\boldsymbol{\beta}$, where $K_h(t) = K(t/h)/h$, a scaled kernel function of $K(t)$ with a bandwidth h, and $\phi_{h_2}(t)$ is a scaled kernel density function of $\phi(t)$ with another bandwidth h_2. As they mentioned, the choice of ϕ is not very crucial, and they used standard normal density. The choice of $K(\cdot)$ is not very important either.

Specifically, when $p = 0$, (1.16) reduces to

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\phi_{h_2}(y_i - \beta_0) \tag{1.17}$$

which is a kernel density estimate of (X, Y) at (x_0, y_0) with $y_0 = \beta_0$. Hence, the resulting estimator $\hat{\beta}_0$, by maximizing (1.17), is indeed the mode of the kernel density estimate in the y direction given $X = x_0$ (Weiss [1994] 8.3.2). This is the reason that they call their method local modal regression.

Note that if we treat $-\phi_{h_2}(t)$ as the loss function in (1.16), the robustness of LMR procedure can be further interpreted from the view of M-estimation (Avella Medina and Ronchetti [2015]). The bandwidth h_2 controls the robustness of the procedure. Figure 1.2 (top) shows that a smaller h_2 are more resistant with outliers than a larger h_2, for the reason that the process with a smaller h_2 considers observations with smaller residuals, thereby to reduce the impact of outliers. Another way to understand the robustness is by investigating the process of estimation. Actually, the LMR estimation is a weighted least square estimation, and the weight of each observation is $\frac{-\phi'_{h_2}(r)}{r}$, where r is the residual for that observation. Therefore, by 1.2(bottom) we can see that outliers will be placed a lower weight because its residual is typically larger. The smaller h_2 is, the lower weight is.

The advantages of the LMR, as demonstrated in Yao et al. [2012], are (1) the resulting estimator is more efficient than the ordinary local polynomial regression estimator in the presence of outliers or heavy tail error distribution and (2) the proposed procedure is as asymptotically efficient as the local polynomial regression estimator when there are no outliers and the error distribution is a Gaussian distribution.

However, as commented in Yao et al. [2012], the LMR procedure suffers from the

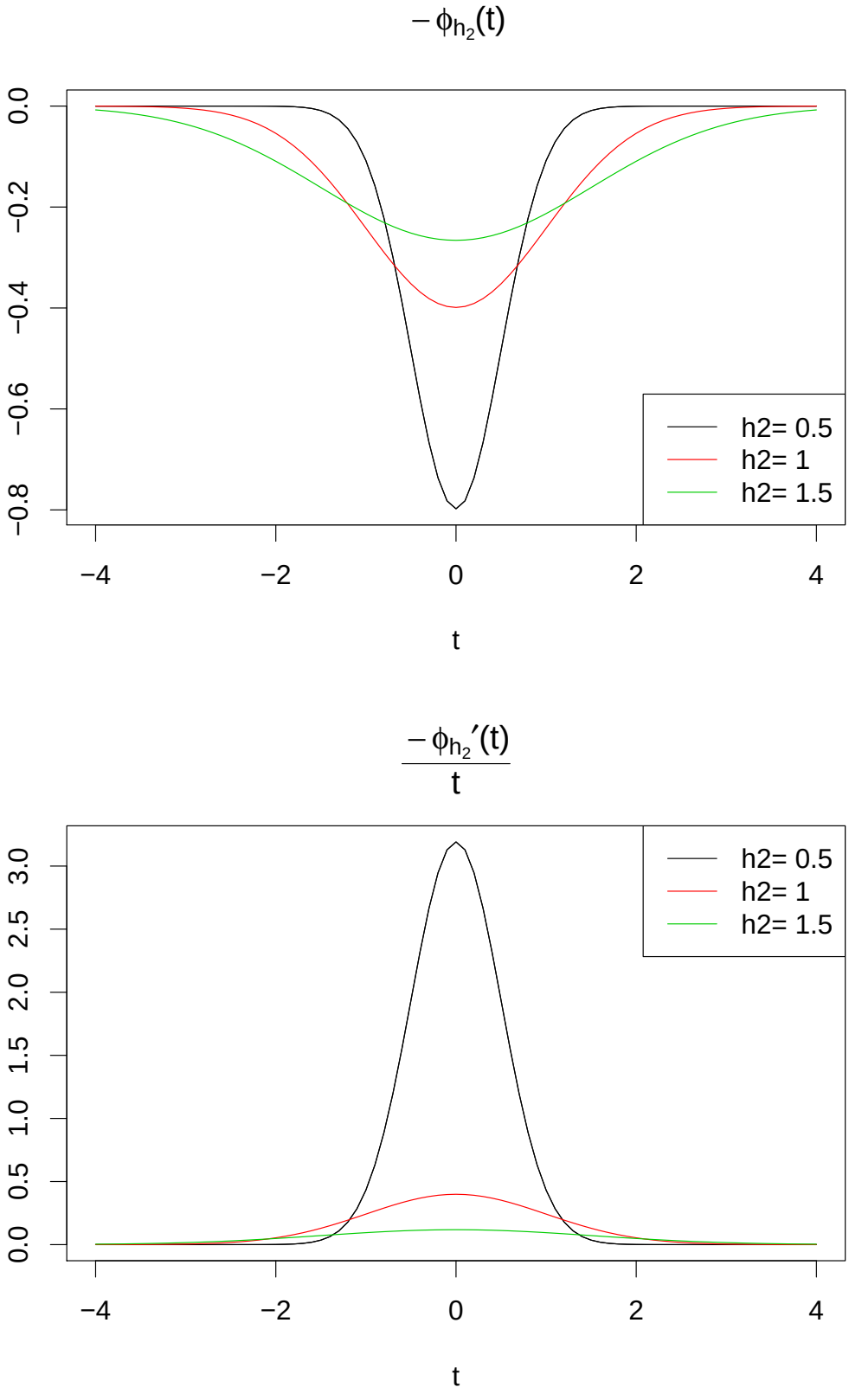

Figure 1.2: Function $-\phi_{h_2}(t)$ and $\frac{-\phi'_{h_2}(t)}{t}$ with different values of h_2

"curse of dimensionality" in the multivariate case. Hence, in order to apply LMR procedure to the multivariate cases, Liu et al. [2013] proposed a single index model to avoid the curse of dimensionality while keeping the model flexibility.

1.2.4 Introduction to B-spline Basis Function

B-spline basis function is one of the most widely used functional basis to approximate smooth functions in the literature of functional data analysis. In this subsection, we introduce the definition of B-spline basis functions of order-d defined on a compact interval $[0, 1]$. Given a sequence of knots $0 = \tau_0 < \tau_1 < ... < \tau_K < \tau_{K+1} = 1$, we define the augmented knot sequence $\{\boldsymbol{\xi}\}$ such that

- $\xi_1 \leqslant \xi_2 \leqslant ... \leqslant \xi_d \leqslant \tau_0$
- $\xi_{j+d} = \tau_j,\ j = 1, ..., K$
- $\tau_{K+1} \leqslant \xi_{K+d+1} \leqslant \xi_{K+d+2} \leqslant ... \leqslant \xi_{K+2d}$

The actual values of these additional knots beyond the boundary are arbitrary, and it is customary to make them all the same as τ_0 and τ_{K+1}, respectively.

Denote, by $B_{i,m}(x)$, the ith B-spline basis function of order m for the knot-sequence $\{\boldsymbol{\xi}\}$, $1 \leqslant m \leqslant d$. The B-spline basis functions are defined recursively in terms of divided differences as follows,

$$B_{i,m}(x) = \frac{x - \xi_i}{\xi_{i+m-1} - \xi_i} B_{i,m-1}(x) + \frac{\xi_{i+m} - x}{\xi_{i+m} - \xi_{i+1}} B_{i+1,m-1}(x), \tag{1.18}$$

for $i = 1, ..., K + 2d - 1$, with

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \xi_i \leqslant x < \xi_{i+1}; \\ 0 & \text{otherwise,} \end{cases} \tag{1.19}$$

which is also known as Haar basis functions.

Figure 1.3 shows the B-spline basis of different orders with 9 interior knots evenly distributed on $[0, 1]$. For more details on B-spline basis functions, see Hastie et al. [2011] and Schumaker [1981].

1.3 Motivations, Contributions and Organization of the Dissertation

In this dissertation, we apply the proposed methods to an application in the field of growth modeling. It's well known that early year catch-down growth is highly prevalent in developing countries for the reason of malnutrition (Black et al. [2008]). Children who suffers from malnutrition in the first 5 years of life will be at increasing risk for the development in cognitive and physical growth. Therefore, characterizing the catch-down growth and identifying the most important risk factors that resulting in it is one of the most important topics in the field of growth modeling. Undoubtably a functional regression model is the best choice for this problem.

However, the following challenges in our study keep us away from the existing methods. First of all, almost all functional predictors, such as height, weight, are sparsely and irregularly observed, especially many individuals are enrolled later or dropped earlier. However, we require the estimation over the entire time interval. Some functional predictors, such as height, should be reasonably assumed monotone over the entire time interval, a non monotone increasing estimation would make no sense. To our best knowledge, existing method can not handle with this case. Secondly, although functional regression has been widely used in the longitudinal data and functional data analysis (Müller [2008]), most of them used a pre-specified

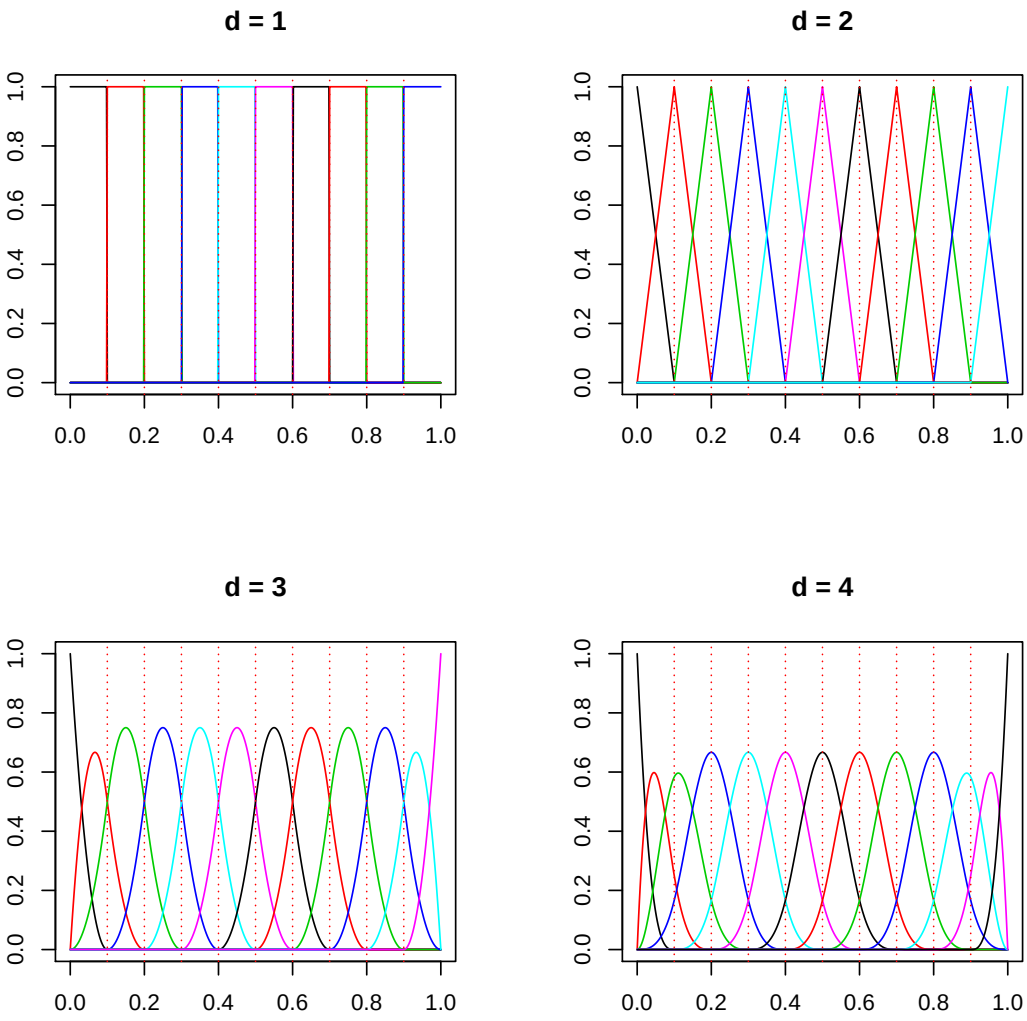

Figure 1.3: B-splines basis functions of different orders

link function. This may increase the risk of model misspecification, leading to a poor model fitting as a result. In addition, the existence of outliers in the response is also a crucial concern in our study.

In the first part of this dissertation, we develop a procedure of estimating monotone functions based on a monotone transformation, the functional principal component (FPC) analysis and a penalized regression to handle the situations where data are sparsely and irregularly observed. In the second part of this dissertation, we construct a functional single index regression model and propose a local model regression based robust variable selection method to identify the most important predictors, estimate their effects as well as the link function.

The contributions of this dissertation are as follows. First, we provide a new method to estimate the monotone functions for sparse functional data, which means the data is sparsely and irregularly observed for each individual, while the pooled data is sufficiently dense. We also prove the asymptotic consistency property for this proposed estimator. In addition, this method can be used as data preprocessing step for other methodology such as functional regression, clustering and classification. Secondly, we establish a robust variable selection procedure for the functional single index regression, where the predictors are functions and the response is a scalar. The asymptotic consistency, sparsity and normality properties have been derived for the resulting penalized estimator. We propose a method to select the optimal bandwidth based on the asymptotic relative efficiency. We show that by using the optimal bandwidth, the LMR estimator is either robust when there are outliers or the error distribution is heavy tailed, or as asymptotically efficient as the ordinary least square based estimator when the data include no outliers and the error distribution is a Gaussian distribution. In addition, we investigate the influence function of proposed estimator.

The remainder of this dissertation is organized as follows. In Chapter 2 we propose a new method to estimate the monotone functions for sparse functional data. Asymptotic consistency for the proposed estimator is derived. Extensive simulation studies are conducted to show that the proposed method outperforms the existing ones. We apply the developed methods to the NIH study in the field of growth modeling. In Chapter 3 we introduce the functional single index model and the local modal regression based estimation method. We employ the group SCAD to do the variable selection. The asymptotic properties of the robust estimator are studied and extensive simulation studies are conducted to demonstrate the efficiency and robustness of proposed estimator.A method to select the best bandwidth is provided and the influence function of the estimator is derived. All proofs of asymptotic properties are deferred to Chapter 4. This dissertation is concluded in Chapter 5 with conclusions and discussions.

Chapter 2

Estimating Monotone Functions for Sparse Growth Data

2.1 Introduction

Unarguably, advancements in technology and computation have lead to a rapidly increasing number of applications where measurements are functions. These developments have been accompanied by intense methodological development where the predictors are functions, such as functional regression, functional data clustering and functional data classification (Cardot et al. [2003], Cardot and Sarda [2005], Crainiceanu et al. [2009], James [2002], Jacques and Preda [2014], Rossi and Villa [2006]).

It is very common that many functional predictors are sparsely and irregularly observed, that is, the measurement time varies from individual to individual. In addition, some individuals may have different starting measurement time and/or ending measurement time. However, in many applications such as functional regression, the complete functions over the entire time interval are usually required.

Existing methods such as PACE (Yao et al. [2005]) and funeigen (Goldsmith et al. [2011]) which utilize information from the pooled data provide a potential solution to such problems.

On the other hand, some functional predictors, such as subjects' height, should be reasonably assumed monotone over the entire time interval. However, existing methods such as PACE and funeigen which can not guarantee the monotonicity may face the risk of interpretation if the resulting estimators are not monotone. Indeed, monotone estimating methods in (Ramsay [1998], Ramsay [2006]) are able to provide monotone estimators. However, they are not applicable in the situation where the time of first measurement and/or last measurement vary across individuals.

Therefore, it will be necessary to obtain the estimation of the monotone functional predictors over the entire time interval in the situation where the functional predictors are sparsely and irregularly observed. The limitations and constraints aforementioned keep us away from using existing methods. To address this problem, in this chapter, we propose a new method based on a monotone transformation and the functional principal component (FPC) analysis, where the FPC scores are obtained by a penalized regression. Numerical studies show that our proposed method outperforms the existing methods in terms of model fitting and monotonicity of the estimation. In addition, the proposed method can also be utilized as a data preprocessing procedure of other methods, such as functional regression, clustering and classification, where the functional predictors are required to be completely known. Four typical examples in NIH cohort study are demonstrated to illustrative the comparison between the proposed method and existing methods, shown in Fig 2.2.

This chapter is organized as follows. In Section 2.2.1 we introduce the model setup. The procedure of model estimation is presented in Section 2.2.2. The asymp-

totic properties of the proposed estimator are shown in Section 2.3. Extensive simulation studies are carried out in Section 2.4 to illustrate the efficiency of the proposed method. A data example from the field of growth modeling is analyzed in Section 2.5. The proofs of asymptotic properties are deferred to Chapter 4.

2.2 Model

2.2.1 Framework

Suppose we have n trajectories $X_i(t)$, $i = 1, ..., n$, which are independent realizations of a smooth random function $X(t)$. The domain of $X(t)$ is a bounded and closed time interval $\mathcal{T}$. For simplicity, we assume $\mathcal{T} = [0, 1]$. We denote sparse functional data as $Y_{ij} = X_i(T_{ij})$, that is, the jth observation of the trajectory $X_i(\cdot)$, made at a random time T_{ij}, $i = 1, ..., n$, $j = 1, ..., m_i$, so we have $0 \leqslant T_{i1} \leqslant T_{im_i} \leqslant 1$.

We further assume $X_i(t)$ is a smooth monotone nondecreasing random function over the time interval $\mathcal{T}$, and can be expressed in the form of

$$X_i(t) = \beta_i + \int_0^t \exp[W_i(s)]ds \tag{2.1}$$

where β_i is equal to $X_i(0)$ and $W_i(t) = \log(X_i'(t))$ (Ramsay [1998]). So the $W_i(t)$ can be viewed as the log rate of $X_i(t)$. The mean function $\mathrm{E}W_i(t) = \mu(t)$ and covariance function $\mathrm{cov}(W_i(s), W_i(t)) = G(s, t)$ are both unknown. We assume that the covariance surface function G can be directly represented through the Karhunen-Loève expansion in terms of a collection of eigenfunctions ϕ_k and nonincreasing eigenvalues λ_k: $G(s, t) = \sum_k \lambda_k \phi_k(t)\phi_k(s)$,$t, s \in \mathcal{T}$. Thereby we can have the ith trajectory $W_i(t)$ expressed as $W_i(t) = \mu(t) + \sum_k \xi_{ik}\phi_k(t)$, where ξ_{ik} are uncorrelated random variables with mean 0 and variance $\mathrm{E}\xi_{ik}^2 = \lambda_k$, where $\sum_k \lambda_k < \infty$, $\lambda_1 \geqslant \lambda_2 \geqslant$

....

Based on these assumptions, the model we consider can be written as

$$\begin{aligned} Y_{ij} &= X_i(T_{ij}) + \epsilon_{ij}, \quad j = 1, ..., m_i, && (2.2) \\ X_i(t) &= \beta_i + \int_0^t \exp[W_i(s)]ds, \quad t \in \mathcal{T}, && (2.3) \\ W_i(t) &= \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \quad t \in \mathcal{T} && (2.4) \end{aligned}$$

for $i = 1, ..., n$, where ϵ_{ij} is the additional iid measurement errors, with $\mathrm{E}(\epsilon_{ij}) = 0$, $\mathrm{Var}(\epsilon_{ij}) = \sigma^2$.

2.2.2 Model Estimation

The procedure of model estimation can be divided into 3 steps. First, for each subject i, we obtain an initial least square estimator of $W_i(t)$, say $\tilde{W}_i(t)$, over the time interval $[T_{i1}, T_{im_i}]$ by (2.2) and (2.3); β_i cannot estimated because no data is observed between $[0, T_{i1}]$, instead we estimate $\beta_i^* = \beta_i + \int_0^{T_{i1}} \exp[W_i(s)]ds$ in this step. Second, we use local linear smoothing method (Fan and Gijbels [1996]) to non-parametrically estimate the mean function $\mu(t)$ and covariance surface function $G(s,t)$ of $W(t)$, denoted by $\hat{\mu}(t)$ and $\hat{G}(s,t)$ respectively, based on the pooled data from $\tilde{W}_i(t)$. The subsequent eigenvalues and eigenfunctions in (2.4) can naturally be obtained by decomposing $\hat{G}(s,t)$ into Karhunen-Loève expansion. Third, under some mild assumptions, we propose a penalized regression method to predict the FPC scores, and the proposed predictor includes the best linear unbiased predictor (BLUP) as a special case. Once the FPC scores are available, we can easily obtain the final estimations of $W_i(t)$ and $X_i(t)$ over the entire time interval $\mathcal{T}$.

Initial estimation of $W_i(t)$ and β_i^*

We attempt to obtain the least square estimation of β_i and $W_i(t)$, $t \in [T_{i1}, T_{im_i}]$ by minimizing the following quadratic loss function

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(Y_{ij} - \beta_i - \int_0^{T_{ij}} \exp[W_i(s)]ds\right)^2. \tag{2.5}$$

One observation can be made from (2.5) is that the estimation of β_i and $W_i(t)$ does not depend on information from subject i', if $i' \neq i$. Another is that it is very difficult to estimate β_i because no data is observed between $[0, T_{i1}]$, instead we estimate $\beta_i^* = \beta_i + \int_0^{T_{i1}} \exp[W_i(s)]ds$ in this step. Therefore, (2.5) can be reduced to

$$\sum_{j=1}^{m_i}\left(Y_{ij} - \beta_i^* - \int_{T_{i1}}^{T_{ij}} \exp[W_i(s)]ds\right)^2. \tag{2.6}$$

Due to the intrinsic nonlinear structure of (2.6), a closed form solution is difficult to obtain. Instead, we use an iterative algorithm which updates the estimation of β_i^* and $W_i(t)$ alternately until a convergence is achieved. Additionally, because the function $W_i(t)$ is unconstrained, we may assume that $W_i(t)$ can be well represented as a linear combination of a series of basis functions, that is appropriate to the problem at hand. In this dissertation, we use B-spline basis. Specifically, let $\boldsymbol{B}_i(t) = (B_{i1}(t), ..., B_{ik}(t))^T$ be a series of B-spline basis with number of basis functions K_i, then we have

$$W_i(t) = \boldsymbol{c}_i^T \boldsymbol{B}_i(t)$$

where $\boldsymbol{c}_i$ is the B-spline coefficients. Then (2.6) can be re-written as

$$\sum_{j=1}^{m_i} \left(Y_{ij} - \beta_i^* - \int_{T_{i1}}^{T_{ij}} \exp[\boldsymbol{c}_i^T \boldsymbol{B}_i(s)] ds \right)^2, \tag{2.7}$$

and the least square estimators $\tilde{\boldsymbol{c}}_i$ and $\tilde{\beta}_i^*$ can be obtained by minimizing (2.7).

The following algorithm is presented for the initial estimation of β_i^* and $W_i(t)$, $i = 1, ..., n$. Let $m_i(t) = \int_{T_{i1}}^{t} \exp[\boldsymbol{c}_i^T \boldsymbol{B}_i(s)] ds$ and $\boldsymbol{z}_i(t) = \frac{\partial m_i(t)}{\partial \boldsymbol{c}_i} = \int_{T_{i1}}^{t} \boldsymbol{B}_i(s) \exp\{\boldsymbol{c}_i^T \boldsymbol{B}_i(s)\} ds$. The iterative algorithm can be stated as follows,

Step 0: given an initial estimate $\boldsymbol{c}_i^{(0)}$, which may be a vector of 0s.

Step 1: given $\boldsymbol{c}_i^{(\nu-1)}$ from the $(\nu-1)$th iteration, estimate $\beta_i^{*(\nu-1)}$ by linear regression.

Step 2: given $\boldsymbol{c}_i^{(\nu-1)}$ and $\beta_i^{*(\nu-1)}$ from the $(\nu-1)$th iteration, obtain $\boldsymbol{c}_i^{(\nu)}$ by Newton Raphson method. The update vector satisfies

$$\boldsymbol{Z}_i^T \boldsymbol{Z}_i (\boldsymbol{c}_i^{(\nu)} - \boldsymbol{c}_i^{(\nu-1)}) = \boldsymbol{Z}_i^T \boldsymbol{r}_i \tag{2.8}$$

where $\boldsymbol{Z}_i = \boldsymbol{Z}_i^{(\nu-1)}$ is $m_i \times K_i$ and has rows $\boldsymbol{z}_i(T_{ij})$ and $\boldsymbol{r}_i$ is the vector of length m_i containing the residuals $r_{ij} = Y_{ij} - \hat{\beta}_i^* - \hat{m}_i(T_{ij})$.

Step 3: Repeat **Step 1** and **Step 2** until convergence.

Note that for each subject i, $\tilde{\beta}_i^*$ and $\tilde{W}_i(\cdot)$ is obtained individually, that is, without communication between the individuals. This results in that we can only have estimation of $\tilde{W}_i(t)$ over the domain $[T_{i1}, T_{im_i}]$, instead of $[0, 1]$. The extension of $\tilde{W}_i(t)$ to entire domain $[0, 1]$ requires borrowing information from pooled data, and this will be stated in the second and third step.

For this step, we can further prove that the least square estimator $\tilde{\beta}_i^*$ and $\tilde{W}_i(\cdot)$ are both consistent estimators over its individual domain, as shown in Section (2.3).

Estimation of $\mu(t)$ and $G(s,t)$

We assume that the mean function $\mu(t)$, covariance surface function $G(s,t)$, and the eigenfunctions $\phi_k(t)$ are all smooth. We use local linear smoothing method (Fan and Gijbels [1996]) for mean and covariance surface estimation, fitting local lines in one dimension and local planes in two dimensions by weighted least squares.

Denote $\tilde{W}_{ij} = \tilde{W}_i(T_{ij})$, where $\tilde{W}_i(T_{ij})$ is the evaluation of initial estimator $\tilde{W}_i(t)$ at time T_{ij}. We define the local linear smoothing estimator of $\mu(t)$ by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} k_1\left(\frac{T_{ij}-t}{h_\mu}\right)\{\tilde{W}_{ij} - b_0 - b_1(t-T_{ij})\}^2$$

with respect to b_0 and b_1. $k_1(\cdot)$ is a kernel function and h_μ is the bandwidth. The estimate of $\mu(t)$ is then $\hat{\mu}(t) = \hat{b}_0(t)$. The local linear surface smoother for $G(s,t)$ is defined by minimizing

$$\sum_{i=1}^{n}\sum_{1\leqslant j\neq l\leqslant m_i} k_2(\frac{T_{ij}-s}{h_G}, \frac{T_{il}-t}{h_G}) \times \{G_i(T_{ij},T_{il}) - b_0 - b_{11}(s-T_{ij}) - b_{12}(t-T_{il})\}^2,$$

with respect to (b_0, b_{11}, b_{12}). The estimate of $G(s,t)$ is then $\hat{G}(s,t) = \hat{b}_0(s,t)$. The estimates of eigenfunctions and eigenvalues correspond to the solutions $\hat{\phi}_k$ and $\hat{\lambda}_k$ of the eigenequations,

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k\hat{\phi}_k(t)$$

where the $\hat{\phi}_k$ are subject to $\int_{\mathcal{T}} \hat{\phi}_k(t)^2dt = 1$ and $\int_{\mathcal{T}} \hat{\phi}_k(t)\hat{\phi}_m(t)dt = 0$ for $m < k$. We estimate the eigenfunctions by discretizing the smoothed covariance, as discribed by Rice and Silverman [1991] and Capra and Müller [1997].

Penalized estimation of FPC score ξ_{ik}

In the situations where the density of the grid of measurements for each subject is sufficiently large, the scores $\xi_{ik} = \int_0^1 (W_i(t) - \mu(t))\phi_k(t)dt$ can be well estimated by numerical integration. Since data are only available only at discrete random times T_{ij} in the framework, reflecting the sparseness of the data, the integrals will not provide reasonable approximations to ξ_{ik}. In this dissertation, we proposed a penalized least square method to estimate the scores ξ_{ik}. We impose a constraint on the model that, every estimated trajectory should get close to the mean function to some degree, which depends on the density of its measurements. We think that the area with fewer number of measurements should be subject to a stronger constraint. This constraint is extremely useful when extending the trajectory $\tilde{W}_i(t)$ from $[T_{i1}, T_{im_i}]$ to the entire time interval $[0, 1]$, because we want to push the estimated trajectory to the mean trajectory in area $[0, T_{i1}] \cup [T_{im_i}, 1]$ for the absence of data availability. Without imposing this constraint may lead to a bad performance in the extension. Therefore, a data-driven penalty term is designed to incorporate this constraint. Our proposed method includes the best unbiased linear prediction (BLUP) in Yao et al. [2005] as a special case.

Write $\boldsymbol{W}_i = (W_i(T_{i1}), ..., W_i(T_{im_i}))^T$, $\boldsymbol{\mu}_i = (\mu(T_{i1}), ..., \mu(T_{im_i}))^T$, and $\boldsymbol{\phi}_{ik} = (\phi_k(T_{i1}), ..., \phi_k(T_{im_i}))^T$. The BLUP of ξ_{ik} is

$$\tilde{\xi}_{ik} = E(\xi_{ik}|\boldsymbol{W}_i) = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{W_i}^{-1}(\boldsymbol{W}_i - \boldsymbol{\mu}_i) \tag{2.9}$$

where $\boldsymbol{\Sigma}_{W_i} = \text{cov}(\boldsymbol{W}_i, \boldsymbol{W}_i)$.

Estimates for the scores ξ_{ik} are obtained from (2.9) by substituting estimates for

$\boldsymbol{\mu}_i$, λ_k, $\boldsymbol{\phi}_{ik}$ and $\boldsymbol{\Sigma}_{W_i}$, which are obtained from the entire data ensemble, leading to

$$\hat{\xi}_{ik} = \hat{E}[\xi_{ik}|\tilde{\boldsymbol{W}}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Sigma}}_{W_i}^{-1}(\tilde{\boldsymbol{W}}_i - \hat{\boldsymbol{\mu}}_i) \tag{2.10}$$

where $\tilde{\boldsymbol{W}}_i = (\tilde{W}_i(T_{i1}), ..., \tilde{W}_i(T_{im_i}))^T$. As we can see from (2.9), the BLUP is obtained from the mean of conditional distribution given $\tilde{\boldsymbol{W}}_i$. We discover that this BLUP can also be obtained from a least square problem, which is stated in following theorem.

Theorem 2.2.1. *Let $\tilde{\boldsymbol{\alpha}}_{ik}$ be the minimizer of*

$$E(\xi_{ik} - \boldsymbol{\alpha}_{ik}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i))^2 \tag{2.11}$$

with respect to $\boldsymbol{\alpha}_{ik}$, then

$$\tilde{\boldsymbol{\alpha}}_{ik} = \boldsymbol{\Sigma}_{W_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k,$$

and $\tilde{\boldsymbol{\alpha}}_{ik}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i)$ is the BLUP of $\tilde{\xi}_{ik}$.

The same estimates for the scores ξ_{ik} can be obtained by replacing $\tilde{\boldsymbol{\alpha}}_{ik}$ with its estimate $\hat{\boldsymbol{\alpha}}_{ik} = \hat{\boldsymbol{\Sigma}}_{W_i}^{-1} \hat{\boldsymbol{\phi}}_{ik} \hat{\lambda}_k$ and the estimation of $W_i(t)$ and $X_i(t)$ can be subsequently obtained

$$\begin{aligned} \hat{W}_i^K(t) &= \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(t) \\ \hat{X}_i^K(t) &= \hat{\beta}_0 + \int_{T_{i1}}^{t} \exp \hat{W}_i^K(s) ds \end{aligned}$$

where K is the number of eigenfunctions used, which is determined by the accumulative proportion of explained variance $\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{\infty} \lambda_i}$.

From (2.11) we can see that the estimator is a pure least square estimator without any constraint, which leaves space for improvement. In this chapter, we impose a data driven constraint by adding a penalty term onto the objective function (2.11). Let $a_{ik,pen} = \boldsymbol{\alpha}_{ik,pen}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i)$, the new objective function is shown as follows:

$$\sum_{k=1}^{K} \mathrm{E}(\xi_{ik} - a_{ik,pen})^2 + \gamma_i \mathrm{E} \int_0^1 \left(\sum_{k=1}^{K} a_{ik,pen}\phi_k(t) \right)^2 v_i(t)dt \tag{2.12}$$

with respect to $\boldsymbol{\alpha}_{ik,pen}$, where $v_i(t)$ is a weight function subject to $\int_0^1 v_i(t)dt = 1$, $\gamma_i \geqslant 0$ is the regularization parameter, K is the total number of eigenfunctions used. We denote the minimizer of (2.12) as $\tilde{\xi}_{ik,pen} = \tilde{a}_{ik,pen} = \tilde{\boldsymbol{\alpha}}_{ik,pen}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i)$ and the associate estimator as $\hat{\xi}_{ik,pen} = \hat{\boldsymbol{\alpha}}_{ik,pen}^T(\tilde{\boldsymbol{W}}_i - \hat{\boldsymbol{\mu}}_i)$. Therefore, the penalized estimator of $W_i(t)$ is

$$\hat{W}_{i,pen}^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik,pen}\hat{\phi}_k(t). \tag{2.13}$$

From (2.12) and (2.13) we can see that the penalty term in (2.12) would constrain the estimated trajectory deviating from the mean trajectory to some degree, which is determined by γ_i and weight function $v_i(t)$. In addition, the proposed penalized estimator is equivalent to BLUP estimator if $\gamma_i = 0$ and will get close to mean trajectory as $\gamma_i \to \infty$. Thus, this penalty term provides a solution between BLUP estimator and the mean trajectory.

By using similar arguments in the proof of Theorem 2.2.1, it is easy for us to have

$$\hat{\boldsymbol{\alpha}}_{ik,pen} = \hat{\boldsymbol{\Sigma}}_{W_i}^{-1}\hat{\boldsymbol{\phi}}_{ik}\hat{\lambda}_k(1 + \gamma_i\hat{D}_{ik})^{-1}$$

where $\hat{D}_{ik} = \int_0^1 \hat{\phi}_k^2(t) v_i(t) dt$. Therefore,

$$\begin{aligned} \hat{\xi}_{ik,pen} &= \hat{\boldsymbol{\alpha}}_{ik,pen}^T(\tilde{\boldsymbol{W}}_i - \hat{\boldsymbol{\mu}}_i) = (1+\gamma_i \hat{D}_{ik})^{-1} \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Sigma}}_{W_i}^{-1} (\tilde{\boldsymbol{W}}_i - \hat{\boldsymbol{\mu}}_i) \\ &= (1+\gamma_i \hat{D}_{ik})^{-1} \hat{\xi}_{ik} \end{aligned} \tag{2.14}$$

and finally we have

$$\hat{X}_{i,pen}^K(t) = \tilde{\beta}_i^* + \int_{T_{i1}}^t \exp \hat{W}_{i,pen}^K(s) ds \tag{2.15}$$

Further we can see from (2.14) that the penalized estimator is a multiplication of BLUP estimator and a constant $(1+\gamma_i \hat{D}_{ik})^{-1}$. Unlike γ_i, $\hat{D}_{ik}$ is private to the kth principal component score of subject i. It is determined by individual weight function and kth eigenfunctions, and only affects its kth principal component score.

Therefore, by tuning γ_i and choosing appropriate weight function $v_i(t)$, our proposed method provides potential ability to improve the estimation.

Selection of weight function $v(\cdot)$ and regularization parameter γ_i

The selection of weight function relies on the following principle: for every subject, its dense area (the area where the data is densely distributed) should receive lighter weight than the sparse area. This is because dense area with more data should have a good estimation based on the existing data. However, sparse area should borrow more information from the pooled data. Therefore, we propose a weight function as follows,

$$v_i(t) \propto \sum_{j=1}^{m_i} \exp\left(\frac{(t-T_{ij})^2}{2 \times length(\mathcal{T})} \right),$$

where $length(\mathcal{T}) = 1$. Other choices of weight function can be $v_i(t) \propto 1$ if $t \in [T_{i1}, T_{im_i}]$ and 0 otherwise, which focus on the trajectory extension.

For the parameter γ_i, we use K-fold Cross Validation (CV) to select γ_i. Specifically, we randomly split the set $\{1, 2, ..., n\}$ into K subset $S_1, ..., S_K$. CV aims to minimize the following function

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{m_i} \left(\tilde{\boldsymbol{W}}_i(T_{ij}) - \hat{\boldsymbol{W}}_i^{(-k)}(T_{ij}) \right)^2$$

where $\hat{\boldsymbol{W}}_i^{(-k)}$ are obtained via minimizing (2.12) by excluding the observations in S_k, for $k = 1, ..., K$. In practice, common choices are $K = 5$ and $K = 10$. In our simulation studies and real data applications in Section 2.4 and Section 2.5, we set $K = 5$ and all data are equally split.

2.3 Asymptotic Properties

First, we derive the asymptotic consistency properties for the initial estimators. For subject i, $i = 1, ..., n$, let $W_{i0}(t)$ be its true trajectory and β_{i0} be the associate true scalar coefficient in model (2.3). Then we have the following consistency theorems. The proofs are deferred to Chapter 4.

Theorem 2.3.1. *Suppose the true trajectory $W_{i0}(t)$ can be well represented by a serial of B-spline basis $\boldsymbol{B}_i(t)$, and the number of basis is independent of m_i, then we have*

$$\sqrt{m_i}|\tilde{\beta}_i - \beta_{i0}| = O_p(1) \tag{2.16}$$

$$\sqrt{m_i} \left(\int_{T_{i1}}^{T_{i2}} (\tilde{W}_i(t) - W_{i0}(t))^2 dt \right)^{1/2} = O_p(1) \tag{2.17}$$

where $\tilde{\beta}_i$ and $\tilde{W}_i(\cdot)$ are initial estimator of β_i and $W_i(t)$.

We now show that the initial estimator possesses the normality property.

Theorem 2.3.2. *Suppose the true trajectory $W_{i0}(t)$ can be well represented by a serial of B-spline basis $\boldsymbol{B}_i(t)$, and the number of basis is independent of m_i, then we have*

$$\begin{aligned}\sqrt{m_i}(\tilde{\beta}_i - \beta_{i0}) &\rightarrow N(0, \sigma^2) \\ \sqrt{m_i}(\tilde{W}_i(t) - W_{i0}(t)) &\rightarrow N(0, \boldsymbol{B}_i^T(t)(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{B}_i(t)\sigma^2)\end{aligned}$$

where $\boldsymbol{Z}_i$ is defined in (2.8).

Next, we derive consistency results demonstrating the consistency of the estimated FPC scores $\hat{\xi}_{ik,pen}$ in (2.14) for the true conditional expectations $\tilde{\xi}_{ik}$ in (2.9). Proofs are deferred to Chapter 4.

The target trajectories we aim to predict are

$$\begin{aligned}\breve{W}_{i,pen}(t) &= \mu(t) + \sum_{k=1}^{\infty} \tilde{\xi}_{ik,pen}\phi_k(t) \\ \breve{X}_{i,pen}(t) &= \beta_0 + \int_{T_{i1}}^{t} \exp \breve{W}_{i,pen}(s)ds\end{aligned}$$

with $\tilde{\xi}_{ik,pen}$ as defined in (2.9). With some mild assumptions, we have the following consistency theorem.

Theorem 2.3.3. *Suppose conditions (A1.1)-(A5) and (B1.1)-B(2.2b) in Section (4.1) hold. Then, as $n \rightarrow \infty$, we have*

$$\lim_{n\rightarrow\infty} \hat{\xi}_{ik,pen} = \tilde{\xi}_{ik,pen} \quad \textit{in probability} \tag{2.18}$$

and for all $t \in \mathcal{T}$

$$\lim_{K\to\infty}\lim_{n\to\infty} \hat{W}^K_{i,pen}(t) = \breve{W}_{i,pen}(t) \quad \textit{in probability} \tag{2.19}$$

$$\lim_{K\to\infty}\lim_{n\to\infty} \hat{X}^K_{i,pen}(t) = \breve{X}_{i,pen}(t) \quad \textit{in probability} \tag{2.20}$$

where $\hat{W}^K_{i,pen}(t)$ *and* $\hat{X}^K_{i,pen}$ *are defined in (2.13) and (2.15).*

2.4 Simulation

In this section, a simulation study is conducted to assess the finite sample performance of the proposed method.

Consider an example consisting of n nondecreasing trajectories which are very similar to children's height. The ith trajectory $x_i(t)$, $i = 1, 2, ..., n$ has the following form,

$$x_i(t) = (170 + 10\eta_i)\frac{\exp(5t)}{1+\exp(5t)} + e_i,$$

where $t \in [0, 1]$, $\eta_i \sim N(0, 1)$ and $e_i \sim N(0, 6)$. Since the observed data is supposed to be sparse and irregular, we randomly pick m_i time points t_{ij}, $j = 1, ..., m_i$ for each trajectory, here m_i is an integer chosen between 5 to 15 with equal probability. Then the jth observation of trajectory i is

$$y_{ij} = x_i(t_{ij}) + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, 0.005x(t_{ij}))$ is the error term.

A total of 100 simulation replications are conducted for the model setup. In the simulations, we use the cubic B-splines with order $d = 4$ for the initial estimation of

β_i and $W_i(t)$, and the number of basis is chosen by cross validation. The bandwidth h_G and h_μ are selected via cross validation. The number of eigenfunctions K is determined by the cumulative proportion of explained variance, which is 0.99 in this simulation. The regularization parameter γ_i are tuned by cross validation, as described in Section 2.2.2.

We compare the performances of the following models

- The proposed model with penalty term (Mono_Pen)
- The proposed model without penalty term (Mono)
- PACE proposed by Yao et al. [2005] (PACE)
- funeigen proposed by Goldsmith et al. [2011] (Funeigen)
- Initial estimation proposed by Ramsay [1998] (IE)

The performance of the estimator $\hat{x}(\cdot)$ will be assessed using the total mean squared errors (MSE)

$$\text{MSE\_total} = n^{-1} \sum_{i=1}^{n} \int_0^1 (\hat{x}_i(t) - x_i(t)^2 dt, \tag{2.21}$$

and the proportion of monotone estimation

$$\text{mono\_prop} = \frac{\text{number of monotone estimated trajectories}}{n}$$

Further, as we all know that it is easier to estimate the curve in $[T_{i1}, T_{i2}]$ than $[0, T_{i1}] \cup [T_{i2}, 1]$, for each individual trajectory. We define the scaled MSE_in and

MSE_out as

$$\text{MSE\_in} = n^{-1} \sum_{i=1}^{n} \frac{\int_{T_{i1}}^{T_{i2}} (\hat{x}_i(t) - x_i(t)^2 dt}{T_{i2} - T_{i1}},$$

and

$$\text{MSE\_out} = n^{-1} \sum_{i=1}^{n} I(T_{i1} + 1 - T_{i2} > 0) \left(\frac{\int_0^{T_{i1}} (\hat{x}_i(t) - x_i(t)^2 dt + \int_{T_{i2}}^{1} (\hat{x}_i(t) - x_i(t)^2 dt}{T_{i1} + 1 - T_{i2}} \right)$$

to see the performance of the estimator in different intervals.

The simulation result with 100 repetitions is summarized in table (2.1), from where several observations can be made. First, existing methods PACE and funeigen can not guarantee monotone estimation, they are beaten by the proposed model in terms of total MSE. Second, if the monotonicity transformation (which is an exponential transformation) is added, we found that the MSE_in decreases, but the MSE_out increases a lot; this is because the variation will be magnified when transforming the data ($W(t)$) back to original data ($X(t)$), especially for those area where there is much fewer or even no data points. So that's why we need to impose a restriction (penalty term) to borrow more information from the mean trajectory. Third, after adding a penalty term, both the MSE_total and MSE_out decreases a lot, while the MSE_in is still comparable with that without penalty. Fourth, if we compare the proposed method with initial estimation in terms of MSE_in, we found that the MSE_in is slightly improved, perhaps for the reason that initial estimation suffers from the boundary effect, when doing the extension, the initial boundary is not the boundary of the extended trajectory. Finally, although funeigen has worst performance generally, but it performs best in MSE_out, indicating that funeigen has the best ability of prediction beyond the individual domain.

Therefore, we can conclude that generally the proposed method outperforms the existing methods in terms of MSE and monotonicity.

Table 2.1: Table: simulation

Method	**MSE_total**	**MSE_in**	**MSE_out**	**mono_prop**
Mono_Pen	1.363	1.041	3.592	1.000
Mono	1.765	1.022	6.480	1.000
PACE	1.755	1.396	3.841	0.184
Funeigen	2.124	2.013	2.422	0.784
IE	-	1.104	-	1.000

2.5 Application

In this section, we apply our proposed method to the NIH cohort study. This study consists of 626 newborn infants living in an urban slum of Mirpur Thana in Dhaka, Bangladesh. Each study subject was visited for collecting information related to child morbidity about every 3 months.

Our interest is to investigate the relationship between height and other demographic factors by using a functional regression model. The fact is that the measurements of height of children are not temporally aligned, some get enrolled later and some get dropped earlier. Therefore, the first step is to estimate the monotone height trajectories over the entire time interval by the proposed method. Figure 2.1 shows the original data and the estimated trajectories.

We pick four typical subjects from the cohort to show the difference between the existing methods and proposed method, as presented in Figure 2.2. All 4 examples show that the estimated trajectory may not be monotone by existing methods such as PACE and funeigen, which need to be improved by the monotone technique. The top 2 examples show that without the penalty, the our monotone technique may

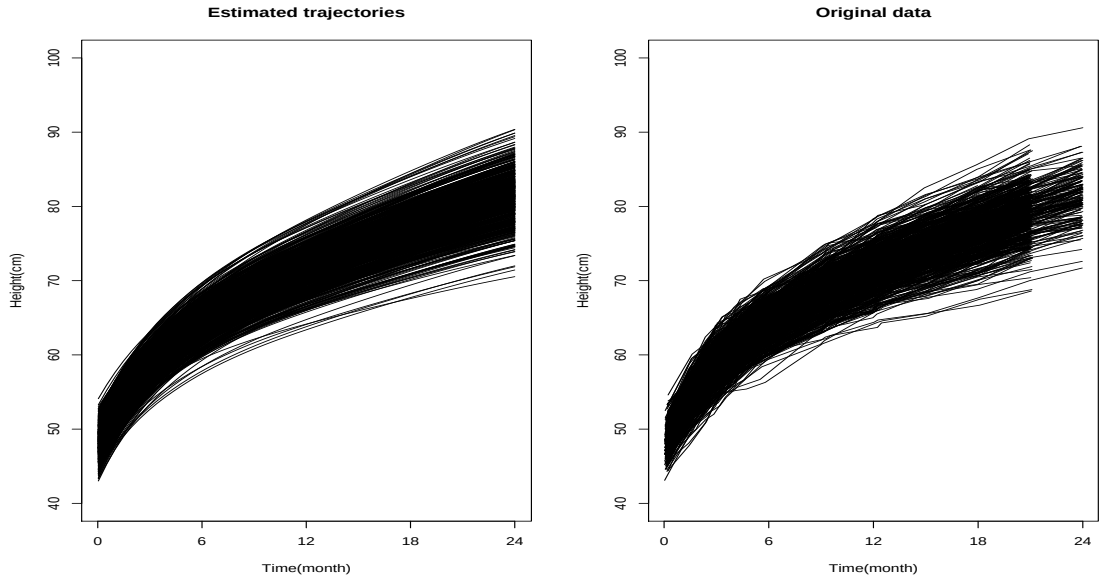

Figure 2.1: Estimated trajectories and original data

not level off in the tail area, especially when it is sparse. This situation could be improved by the proposed penalty. In addition, the initial estimation can only give us monotone estimation over individual domain, instead of the entire interval.

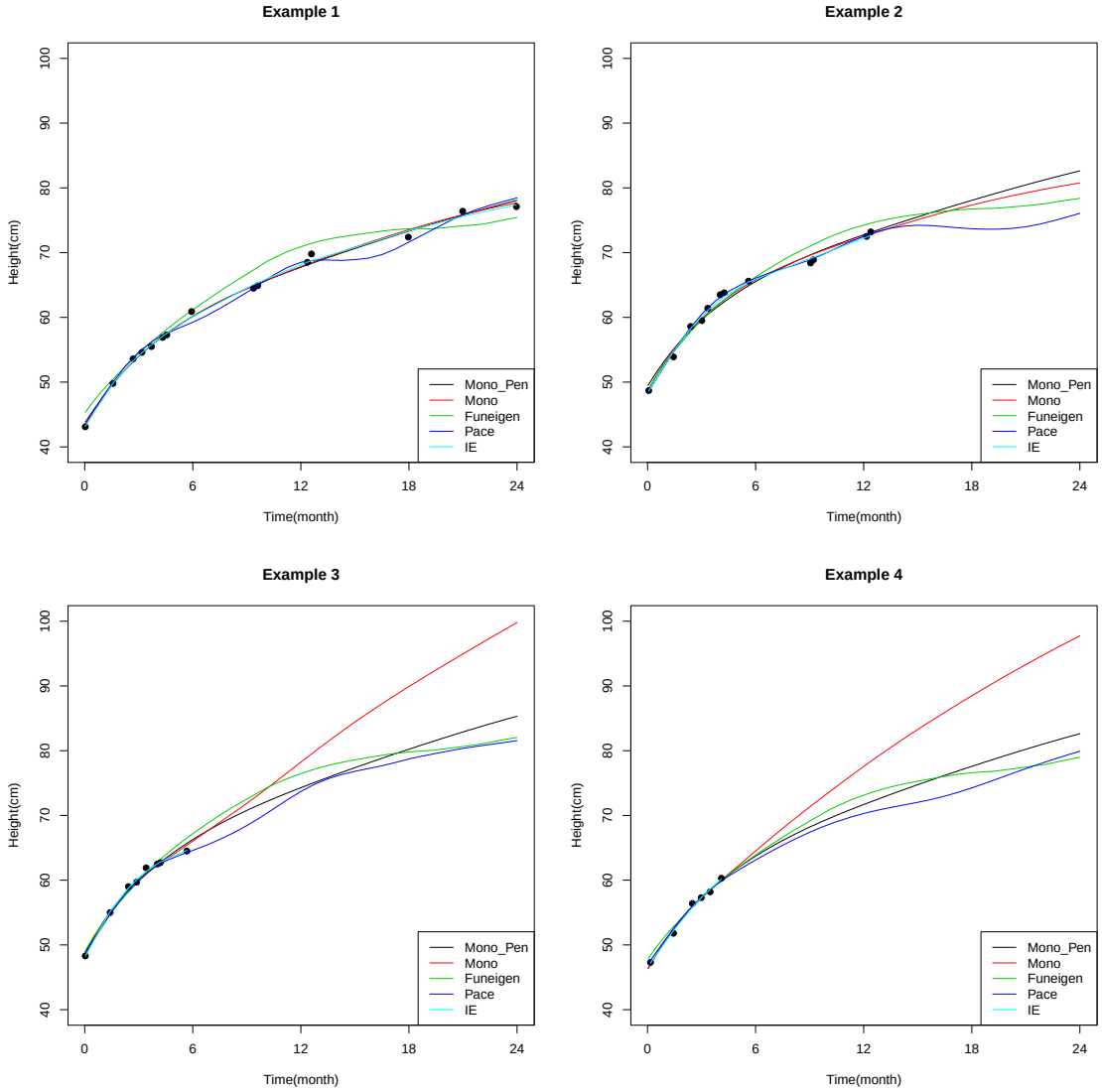

Figure 2.2: Four typical examples in NIH cohort study to illustrate the difference between proposed method and existing methods: Given that the pooled data are densely distributed over the entire time interval $[0, 24]$, existing methods 1 such as PACE (Yao et al. [2005]) and funeigen (Goldsmith et al. [2011]) may have decreasing estimation in the sparse area, that is, monotonicity is not guaranteed, and existing methods 2 (labeled with IE) (Ramsay [1998]) cannot have estimation over the entire time interval. On the other hand, our proposed method is able to produce the monotone estimation over the entire time interval for every individual curve.

Chapter 3

Robust and Efficient Variable Selection for Functional Single Index Model with a Scalar Response

3.1 Introduction

Functional single index model (FSIM) is widely used in many applications to investigate the relationship between a collection of functional predictors and a scalar response. However, with the increasing number of functional predictors being collected, it is desirable to select the most important ones for the purpose of statistical accuracy, model interpretability, and computational complexity.

Outliers are usually accompanied with the ever growing data. It is well known that the least squares estimation is inefficient and biased in the presence of outliers. In fact, least squares estimation is highly sensitive to outliers because it will

be dragged towards the outliers, and the variance of the estimates is artificially inflated. To address this problem, robust techniques are usually considered. A nonparametric M-type of regression (Huber [2011]) will place a lower weight on the outliers and thereby to reduce their impact, so usually it will be more efficient than the least-square based nonparametric regression when there are outliers or the error distribution has a heavy tail. However, these methods lose some efficiency when the error is normal distributed. On the other hand, local modal regression (LMR) proposed by Yao et al. [2012] possesses the advantages of both the least-square based estimator and M-type estimators, that is, LMR estimator is as efficient as least-square based estimator when there are no outliers and the error is normal distributed and robust to outliers otherwise, for the univariate regression, as shown in Yao et al. [2012].

In this chapter, we take the inspirations from the LMR to propose a robust and efficient variable selection procedure for the functional single index model to produce a sparse estimator with partial zero components. The main contribution is that the sampling properties of the proposed estimator are systemically studied. First, we show that the resulting estimator has the Oracle property. Furthermore, we prove that the proposed estimator is as efficient as least-square based estimator when the error is normal distributed and more efficient than least-square based estimator when the error distribution is heavy-tailed.

This chapter is organized as follows. In Section 3.2 we introduce the functional single index model. In Section 3.3 we propose an LMR based robust estimation and variable selection method for FSIM. The asymptotic properties of proposed estimator are presented in Section 3.4. The relative efficiency, tuning parameter and influence function problem is placed in Section 3.5. Extensive simulation studies are carried out in Section 3.6 to illustrate the efficiency and robustness of proposed

estimator, comparing with the least-square based and M-type estimator. A data example in the field of growth modeling is analyzed in Section 3.7. The proofs of asymptotic properties are deferred to Chapter 4.

The following notations are adopted in this chapter. We denote $||\cdot||_2$ as the L_2 norm for both vectors and functions. For example, if $\boldsymbol{x} = (x_1, ..., x_p)^T$ is a $p \times 1$ vector, $||\boldsymbol{x}||_2^2 = \sum_{j=1}^p x_j^2$; if $\alpha(t)$, $t \in [0,1]$ is a functional predictor, $||\alpha||_2^2 = \int_0^1 \alpha^2(t)dt$. If $\boldsymbol{\alpha}(t) = (\alpha_1(t), ..., \alpha_p(t))^T$, $t \in [0,1]$ is a $p \times 1$ vector of functional predictor, then $||\boldsymbol{\alpha}||_2^2 = ||\alpha_1||_2^2 + ... + ||\alpha_p||_2^2$. Given a $M \times M$ positive definite matrix $\boldsymbol{H}$, for any $M \times 1$ vector $\boldsymbol{\beta}$, we define $||\boldsymbol{\beta}||_H^2 = \boldsymbol{\beta}^T \boldsymbol{H} \boldsymbol{\beta}$. Let $\otimes$ denote the Kronecker product.

3.2 Functional Single Index Model

Suppose we have a scalar response y and a collection of functional predictors $\boldsymbol{X}(t) = (X_1(t), ..., X_p(t))^T$, defined on the interval $[0,1]$. Let $(\boldsymbol{X}_i(t), y_i)$, $i = 1, ..., n$, $t \in [0,1]$ be independent and identically distributed observations from $(\boldsymbol{X}(t), y)$. The functional single index model can be expressed as

$$y_i = g\left(\int_0^1 \boldsymbol{\alpha}(t)^T \boldsymbol{X}_i(t)dt\right) + \epsilon_i \tag{3.1}$$

where $g(\cdot)$ is a smooth unknown link function, ϵ_i is a random error with unknown variance σ^2 and $\boldsymbol{\alpha}(t) = (\alpha_1(t), ..., \alpha_p(t))^T$ are the coefficient functions.

For model identifiability, we assume $||\boldsymbol{\alpha}||_2 = 1$ and $\alpha_1(0) > 0$. To study this identifiability issue, we use the argument similar as Ma [2016]. Let

$$g\left(\sum_{j=1}^p \int_0^1 \alpha_j(t)^T X_{ij}(t)dt\right) = g^*\left(\sum_{j=1}^p \int_0^1 \alpha_j^*(t)^T X_{ij}(t)dt\right) \tag{3.2}$$

where $\alpha_j(t)$ and $\alpha_j^*(t)$ are functional coefficients satisfying conditions (C1) and (C2) in Section 3.4 and $g(\cdot)$ and $g^*(\cdot)$ are nonconstant link functions. There exist some $\boldsymbol{\delta}_j = \{\delta_{jl}\}$, $\boldsymbol{\delta}_j^* = \{\delta_{jl}^*\}$ and $\boldsymbol{\xi}_{ij} = \{\xi_{ijl}\}$ such that $\alpha_j(t) = \sum_{l=1}^{\infty} \delta_{jl}\varphi_l(t)$, $\alpha_j^*(t) = \sum_{l=1}^{\infty} \delta_{jl}^*\varphi_l(t)$ and $X_{ij}(t) = \sum_{l=1}^{\infty} \xi_{ijl}\varphi_l(t)$, for each $i = 1, ..., n$, $j = 1, ..., p$, and $t \in [0, 1]$, where $\varphi_l(t)$, $l = 1, 2, ...,$ are an orthogonal basis. Thus we have $g(\sum_{j=1}^{p} \boldsymbol{\delta}_j^T \boldsymbol{\xi}_{ij}) = g^*(\sum_{j=1}^{p} \boldsymbol{\delta}_j^{*T} \boldsymbol{\xi}_{ij})$. By taking partial derivative with respect to each element in $\boldsymbol{\xi}_{ij}$, we have

$$\frac{\partial g/\partial \xi_{ijl}}{\partial g/\partial \xi_{i11}} = \frac{\delta_{jl}}{\delta_{11}} = \frac{\delta_{jl}^*}{\delta_{11}^*} = \frac{\partial g^*/\partial \xi_{ijl}}{\partial g^*/\partial \xi_{i11}} \tag{3.3}$$

Hence, we have $\delta_{jl} = \delta_{jl}^*$ or $\delta_{jl} = -\delta_{jl}^*$ if $\sum_{j=1}^{p}\sum_{l=1}^{\infty} \delta_{jl}^2 = 1$ and $\sum_{j=1}^{p}\sum_{l=1}^{\infty} \delta_{jl}^{*2} = 1$, for all l and j, which implies that $\alpha_j(t) = \alpha_j^*(t)$ or $\alpha_j(t) = -\alpha_j^*(t)$ for all $t \in [0, 1]$ and $j = 1, ..., p$. Thus, we only need to impose one linear constraint on the deterministic coefficient $\boldsymbol{\delta_j}$ for model identifiability. A straightforward one is let $\alpha_1(0) > 0$, which is equivalent to $\sum_{l=1}^{\infty} \delta_{1l}\varphi_l(0) > 0$.

REMARK 3.2.1. *Without the loss of generality, we assume $\alpha_1(0) > 0$. Otherwise, we can assume there exist some integer j and $t_0 \in (0, 1)$, such that $\alpha_k(t) = 0$, $k < j$ and $\alpha_j(t) = 0$, $t <= t_0$, but $\alpha_j(t + \epsilon) > 0$, for any small positive ϵ.*

3.3 Robust and Efficient Variable Selection for FSIM

3.3.1 Local Modal Estimation

The parameters to be estimated are the coefficient functions $\boldsymbol{\alpha}(\cdot)$ and the link function $g(\cdot)$. In practice, it is common to use B-spline functions to represent

the coefficient functions $\boldsymbol{\alpha}(t) = (\alpha_1(t), ..., \alpha_p(t))^T$. Assume the coefficient function $\alpha_j(t)$ can be well approximated by a series of normalized B-spline basis functions $\boldsymbol{B}(t) = (B_1(t), ..., B_K(t))^T$ with order d and number of basis K, and K is independent of n. By omitting the approximation error, we have

$$\alpha_j(t) = \sum_{k=1}^{K} \beta_{j,k} B_k(t) = \boldsymbol{B}(t)^T \boldsymbol{\beta}_j$$

where $\boldsymbol{\beta}_j = (\beta_{j,1}, ..., \beta_{j,K})^T$, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_p^T)^T$ is a $Kp \times 1$ vector. Hence, model (3.1) can be approximated in the matrix form

$$y_i = g\left(\int_0^1 \boldsymbol{\alpha}(t)^T \boldsymbol{X}_i(t) dt\right) + \epsilon_i = g(\boldsymbol{C}_i^T \boldsymbol{\beta}) + \epsilon_i \tag{3.4}$$

where $\boldsymbol{C}_i = (\boldsymbol{C}_{i1}^T, ..., \boldsymbol{C}_{ip}^T)^T$ is a $Kp \times 1$ vector, with $\boldsymbol{C}_{ij} = \int_0^1 \boldsymbol{B}(t) X_{ij}(t) dt$.

In practice, it can be assumed that the predictors $\boldsymbol{X}_i(t)$ are completely known, although they are densely or sparsely observed at discrete time points $t_{i1}, ..., t_{i,m_i}$. Otherwise, we can use a pre-smoothing technique from Chapter 2, Yao et al. [2005] or Goldsmith et al. [2011] before modeling. Thus $\boldsymbol{C}_i$ is known, since it can be well approximated by the Riemann integration.

It is easy to see that once we get an estimator for $\boldsymbol{\beta}$ from (3.4), the estimation of coefficient functions $\boldsymbol{\alpha}(t)$ can be obtained via the equation $\hat{\boldsymbol{\alpha}}(t) = (I_p \otimes \boldsymbol{B}(t)^T)\hat{\boldsymbol{\beta}}$. Thus we only need to estimate $\boldsymbol{\beta}$ and g in (3.4).

In this section, we aim to propose a local modal estimator for $\boldsymbol{\beta}$ and g, due to the possible presence of outliers. For ease of computation, we assume that the scaled density function $\phi(\cdot)$ in (1.16) is standard normal distribution. To do the variable selection, we consider the group SCAD as the penalty term. Therefore, the

penalized local modal estimator $\hat{\boldsymbol{\beta}}$ and $\hat{g}$ can be obtained by minimizing

$$-\sum_{i=1}^{n}\phi_{h_2}\left(y_i - g(\boldsymbol{C}_i^T\boldsymbol{\beta})\right) + n\sum_{j=1}^{p} p_{\lambda_n}(||\boldsymbol{\beta_j}||_H) \tag{3.5}$$

where $\phi_{h_2}(t) = \phi(t/h)/h$ is a scaled kernel density function of $\phi(t)$. h_2 is a bandwidth and is assumed only depending on the conditional error distribution of ϵ give U, where $U = \int_0^1 \boldsymbol{\alpha}(t)^T\boldsymbol{X}(t)dt$. $p_{\lambda_n}(\cdot)$ is the SCAD penalty defined as (1.8) and λ_n is the regularization parameter.

3.3.2 Algorithm

However, simultaneously estimating $\boldsymbol{\beta}$ and g from (3.5) is difficult. We propose an iterative algorithm to update the estimations of g and $\boldsymbol{\beta}$ alternately until their change do not exceed a pre-specified threshold value. The algorithm is presented as below.

Step 0 (Initialization step) Obtain an initial value $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\boldsymbol{\beta}}_1^{(0)T}, ..., \hat{\boldsymbol{\beta}}_p^{(0)T})^T$ of $\boldsymbol{\beta}$ such that $\sum_j ||\hat{\boldsymbol{\beta}}_j^{(0)}||_H^2 = 1$;

Step I: Let $\hat{\boldsymbol{\beta}}^{(k)}$ be the estimate after the kth iteration. Note that the link function $g(\cdot)$ can be locally approximated by a linear function, that is, for v in a neighborhood of u,

$$g(v) \approx g(u) + g'(u)(v-u) \equiv a + b(v-u),$$

then in the $(k+1)$th iteration, one can find $(\hat{a}^{(k+1)}(u), \hat{b}^{(k+1)}(u))$ by minimizing

$$-\sum_{i=1}^{n}\phi_{h_2}\left(y_i - a - b(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)} - u)\right) K_h(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)} - u) \tag{3.6}$$

with a symmetric kernel function $K_h = K(t/h)/h$, where h is the bandwidth. However, we can not have a closed form for ($\hat{a}^{(k+1)}(u)$ and $\hat{b}^{(k+1)}(u)$), instead, we need another iterative algorithm. For ease of presentation, we suppress the superscript of $\hat{\boldsymbol{\beta}}^{(k)}$ here. Let $\hat{\boldsymbol{\theta}}_{(0)} = (\hat{a}_{(0)}, \hat{b}_{(0)})^T$ be the initial value. Start from $l = 0$, the algorithm for solving (3.6) can be stated as follows,

- Given the $\hat{a}_{(l)}$ and $\hat{b}_{(l)}$, we define $w_{i(l)}$ as

$$w_{i(l)} = -K_h(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}} - u)\frac{\phi'_{h_2}(\hat{e}_{i(l)})}{\hat{e}_{i(l)}} \tag{3.7}$$

  where $\hat{e}_{i(l)} = r_{i(l)}/\text{MAD}(r_{i(l)})$ with $r_{i(l)} = y_i - \hat{a}_{(l)} - \hat{b}_{(l)}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}} - u)$ the residuals in the lth iteration and MAD is the median absolute deviation function.

- Update $\hat{a}_{(l+1)}, \hat{b}_{(l+1)}$ by minimizing

$$\sum_{i=1}^{n}\Big(y_i - a - b(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}} - u)\Big)^2 w_{i(l)} \tag{3.8}$$

  w.r.t a and b, where $w_{i(l)}$ is from (3.7).

- repeat (3.7) and (3.8) until convergence.

Then we define $\hat{g}^{(k+1)}(u) = \hat{a}^{(k+1)}(u)$.

Step II: Given the estimate of $g(\cdot)$, say, $\hat{g}^{(k+1)}(\cdot)$ in the $(k+1)$th iteration, update $\hat{\boldsymbol{\beta}}^{(k+1)}$ by solving the following minimization problem

$$\hat{\boldsymbol{\beta}}^{(k+1)} := \underset{\boldsymbol{\beta}, \sum_j ||\boldsymbol{\beta_j}||_H^2 = 1}{\text{argmin}} \sum_{i=1}^{n} -\phi_{h_2}\left(y_i - \hat{g}^{(k+1)}(\boldsymbol{C}_i^T\boldsymbol{\beta})\right) + n\sum_{j=1}^{p} p_{\lambda_n}(||\boldsymbol{\beta_j}||_H) \tag{3.9}$$

In this step we suppress the superscript of $\hat{g}^{(k+1)}$. However, since $\hat{g}(\cdot)$ may not be a linear function, solving (3.9) is a nonlinear optimization problem, and the

computation can be challenging. To address this problem, we first approximate the term $\hat{g}(\boldsymbol{C}_i^T\boldsymbol{\beta})$ by its first order Taylor expansion at $\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)}$. Then the first term of (3.9) is approximated as

$$-\sum_{i=1}^{n}\phi_{h_2}\left(y_i-\hat{g}^{(k+1)}(\boldsymbol{C}_i^T\boldsymbol{\beta})\right) \approx -\sum_{i=1}^{n}\phi_{h_2}\left(y_i-\hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})-\hat{g}'(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})(\boldsymbol{C}_i^T\boldsymbol{\beta}-\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})\right)$$

For the second term, we use the quadric approximation as Fan and Li [2001], if $\hat{\boldsymbol{\beta}}^{(k)}$ is a neighborhood of $\boldsymbol{\beta}$

$$\begin{aligned} p_{\lambda_n}(||\boldsymbol{\beta_j}||_H) &\approx p_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{k})}||_H)+p'_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{k})}||_H)(||\boldsymbol{\beta_j}||_H-||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{k})}||_H) \\ &\approx \frac{p'_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{k})}||_H)}{2||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(\boldsymbol{k})}||_H}\boldsymbol{\beta}_j^T\boldsymbol{H}\boldsymbol{\beta}_j+\text{Constant} \\ &= \boldsymbol{\beta}^T(\boldsymbol{D}_\lambda\otimes\boldsymbol{H})\boldsymbol{\beta}+\text{Constant} \end{aligned}$$

with $\boldsymbol{D}_\lambda=\text{diag}(\frac{p'_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}^{(\boldsymbol{k})}||_H)}{2||\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}^{(\boldsymbol{k})}||_H},...,\frac{p'_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(\boldsymbol{k})}||_H)}{2||\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(\boldsymbol{k})}||_H})$. Hence $\hat{\boldsymbol{\beta}}^{(k+1)}$ can be obtained by minimizing

$$-\sum_{i=1}^{n}\phi_{h_2}\left(y_i-\hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})-\hat{g}'(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})(\boldsymbol{C}_i^T\boldsymbol{\beta}-\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})\right)+n\boldsymbol{\beta}^T(\boldsymbol{D}_\lambda\otimes\boldsymbol{H})\boldsymbol{\beta}$$

Therefore, $\hat{\boldsymbol{\beta}}^{(k+1)}$ is obtained by a similar iterative algorithm as that solving (3.6). Specifically, given an initial value $\hat{\boldsymbol{\beta}}_{(0)}$ and $l=0$, repeat the following 3 steps until convergence

- Given the $\hat{\boldsymbol{\beta}}_{(l)}$, we define $w_{i(l)}$ as

$$w_{i(l)}=-\frac{\phi'_{h_2}(\hat{e}_{i(l)})}{\hat{e}_{i(l)}}$$

where $\hat{e}_{i(l)} = r_{i(l)}/\text{MAD}(r_{i(l)})$ with $r_{i(l)} = y_i - \hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)}) - \hat{g}'(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}_{(l)} - \boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})$ the residuals in the lth iteration and MAD is the median absolute deviation function.

- Obtain $\hat{\boldsymbol{\beta}}_{(l+1)}$ by minimizing

$$\sum_{i=1}^{n}\left(y_i - \hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)}) - \hat{g}'(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})(\boldsymbol{C}_i^T\boldsymbol{\beta} - \boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)})\right)^2 w_{i(l)} + n\boldsymbol{\beta}^T(\boldsymbol{D}_\lambda \otimes \boldsymbol{H})\boldsymbol{\beta}$$

then we have

$$\hat{\boldsymbol{\beta}}_{(l+1)} = (\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U} + n\boldsymbol{D}_\lambda \otimes \boldsymbol{H})^{-1}\boldsymbol{U}^T\boldsymbol{V}$$

where $\boldsymbol{U} = \left(\hat{g}'(\boldsymbol{C}_1^T\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{C}_1, ..., \hat{g}'(\boldsymbol{C}_n^T\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{C}_n\right)^T$, $\boldsymbol{V} = (v_1, ..., v_n)^T$ with $v_i = y_i - \hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)}) + \hat{g}'(\boldsymbol{C}_1^T\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}^{(k)}$, $\boldsymbol{W} = \text{diag}(w_{1(l)}, ..., w_{n(l)})$

- Normalization: $\hat{\boldsymbol{\beta}}_{(l+1)} = \hat{\boldsymbol{\beta}}_{(l+1)}/||\hat{\boldsymbol{\beta}}_{(l+1)}||_{\boldsymbol{I}_p\otimes\boldsymbol{H}}$

Step III: Repeat **Step I** and **Step II** until convergence.

REMARK 3.3.1. *The initial value $\hat{\boldsymbol{\beta}}^{(0)}$ can be randomly initialized or the least square estimator. The complexity of this algorithm is $O(NK^2 + K^3)\times$ (average number of iterations in step I + average number of iterations in step II) $\times$ average number of iterations in step III, where N is number of subjects and K is the number of basis spline functions.*

3.4 Asymptotic Properties

For the ease of presentation, we denote the true coefficient functions in model (3.1) by $\boldsymbol{\alpha}_0(t) = (\alpha_{01}(t), ..., \alpha_{0p}(t))^T$, with $t \in [0, 1]$ and the true link function by $g_0(\cdot)$.

Let $F(u, h_2) = E(\phi''_{h_2}(\epsilon)|U = u)$ and $G(u, h_2) = E(\phi'_{h_2}(\epsilon)^2|U = u)$, where $\phi'_{h_2}(t)$ and $\phi''_{h_2}(t)$ are first and second derivative of $\phi_{h_2}(t)$. Denote $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K^2(u)du$ where $K(\cdot)$ is the kernel in (3.6). To establish the asymptotic properties of the proposed estimators, the following regularity conditions are imposed.

(C1) For all $1 \leqslant j \leqslant p$, $X_j(t)$ are random square-integrable functions. In addition, we assume $\max_i ||\boldsymbol{X}_i||_2/\sqrt{n} = o_p(1)$.

(C2) The coefficient functions $\alpha_{01}(t), ..., \alpha_{0p}(t)$ are continuously differentiable on $[0, 1]$.

(C3) $||\boldsymbol{\alpha_0}||_2 = 1$ and $\alpha_{01}(0) > 0$.

(C4) The marginal density of $U = \int_0^1 \boldsymbol{\alpha}_0(t)^T \boldsymbol{X}(t)dt$, denoted by f, is positive and has a continuous second derivative.

(C5) The link function $g_0(\cdot)$ has a continuous second derivative.

(C6) $K(\cdot)$ is a symmetric density function with bounded support and satisfies the Lipschitz condition.

(C7) Let $\tau_1, ..., \tau_K$ be the interior knots of $[0, 1]$ for a B-spline basis $\boldsymbol{B}(t), t \in [0, 1]$. Moreover, let $\tau_0 = 0$ and $\tau_{K+1} = 1$, $\Delta_i = \tau_i - \tau_{i-1}$. Then there exists a constant C_0 such that

$$\frac{\max\{\Delta_i\}}{\min\{\Delta_i\}} \leqslant C_0$$

(C8) $\liminf_{n\to\infty} \liminf_{||\beta_j||_H \to 0^+} \lambda_n^{-1} p_{\lambda_n}(||\boldsymbol{\beta_j}||_H) > 0, j = s + 1, ..., p$

(C9) $F(u, h_2)$ and $G(u, h_2)$ are continuous with respect to u

(C10) $F(u, h_2) < 0$ for any h_2

(C11) $E(\phi'_h(\epsilon|U = u)) = 0$,$E(\phi''_h(\epsilon)^2|U = u))$,$E(\phi'_h(\epsilon)^3|U = u))$ and $E(\phi'''_h(\epsilon)|U = u))$ are continuous with respect to u.

REMARK 3.4.1. *The conditions (C1)-(C2) are common conditions in functional regression, such as in Fan et al. [2005], Li and Liang [2008]. The condition (C3) is used for model identifiability. Conditions (C4)-(C6) are very similar as those for functional single index model in Liu et al. [2013]. Condition (C7) implies that* $\tau_0, ..., \tau_{K+1}$ *is a* C_0*-quasi-uniform sequence of partitions of* $[0, 1]$. *Condition (C8) is the assumption about the penalty function, which is similar to Fan and Li [2001] and Li and Liang [2008]. Conditions (C9)-(C11) are used in modal nonparametric regression in Yao et al. [2012].*

Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, ..., \boldsymbol{\beta}_{p0})^T$ be minimizer of $||\boldsymbol{\alpha}_0(t) - (I_p \otimes \boldsymbol{B}(t)^T)\boldsymbol{\beta}||_2$ with respect to $\boldsymbol{\beta}$, given the basis function $\boldsymbol{B}(t)$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{g}$ be the solution by minimizing (3.5). Denote

$$a_n = \max_j \{p'_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H) \ \ \boldsymbol{\beta_{j0}} \neq \mathbf{0}\} \tag{3.10}$$

and

$$b_n = \max_j \{p''_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H) \ \ \boldsymbol{\beta_{j0}} \neq \mathbf{0}\}. \tag{3.11}$$

We have the following Theorem 3.4.1 which gives the consistency of the proposed penalized estimator.

Theorem 3.4.1. (Consistency) *Suppose the conditions (C1)-(C11) in the appendix.*

hold. If $h \to 0$, $nh^3 \to \infty$, $a_n = O_p(n^{-1/2})$ *and* $b_n \to 0$ *as* $n \to \infty$, *we have*

$$||\hat{\alpha}_j(\cdot) - \alpha_{j0}(\cdot)|| = O_p(n^{-1/2}) \quad j = 1, ..., p \tag{3.12}$$

It is clear to see that $a_n \to 0$ if $\lambda_n \to 0$, by the property of SCAD. Therefore, by appropriately choosing the λ_n, we are able to obtain a $O(n^{1/2})$ consistent estimator. Without the loss the generality, we assume that the true coefficient function $\alpha_{j0}(t) = 0$ for $j = s+1, ..., p$. We now show that this estimator possesses the sparsity property, $\hat{\alpha}_j(\cdot) = 0$ for $j = s+1, ..., p$, which is stated as follows.

Theorem 3.4.2. (Sparsity) *Suppose the conditions (C1)-(C11) in the appendix hold. If* $h \to 0$, $nh^3 \to \infty$, $nh^4 \to 0$, $\lambda_n \to 0$ *and* $n^{1/2}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to* 1, *we have*

$$\hat{\alpha}_j(\cdot) = 0 \quad j = s+1, ..., p \tag{3.13}$$

Without loss of generality, we assume $\hat{\boldsymbol{\alpha}}(\cdot) = (\hat{\boldsymbol{\alpha}}_1(\cdot), \mathbf{0})^T$ be the consistent estimator. Thus given the function basis $\boldsymbol{B}(t)$, we have $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0})^T$ and $\boldsymbol{C} = (\boldsymbol{C}_1^T, \boldsymbol{C}_0^T)^T$. Denote

$$\begin{aligned}
\boldsymbol{b} &= (\boldsymbol{b}_1^T, ..., \boldsymbol{b}_s^T)^T \text{ with } \boldsymbol{b}_j = p'_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)\frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}}{||\boldsymbol{\beta_{j0}}||_H}, \\
\boldsymbol{\Sigma} &= \text{diag}(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_s) \\
&\quad \boldsymbol{\Sigma}_j = p''_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)\frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}\boldsymbol{\beta}_{j0}^T\boldsymbol{H}}{||\boldsymbol{\beta_{j0}}||_H^2} + p'_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)(\frac{\boldsymbol{H}}{||\boldsymbol{\beta_{j0}}||_H} - \frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}\boldsymbol{\beta}_{j0}^T\boldsymbol{H}}{||\boldsymbol{\beta_{j0}}||_H^3}), \\
\boldsymbol{W} &= \text{E}[\phi''_{h_2}(\epsilon)(g'_0(U))^2\boldsymbol{C}_1\boldsymbol{C}_1^T] - \text{E}[(g'_0(U))^2\phi''_{h_2}(\epsilon)\boldsymbol{C}_1\text{E}(\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2)}\boldsymbol{C}_1^T|U)], \\
\boldsymbol{Q} &= \text{E}\left(\phi'_{h_2}(\epsilon)^2 g'_0(U)^2(\boldsymbol{C}_1 - \text{E}(\boldsymbol{C}_1\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2)}|U)(\boldsymbol{C}_1 - \text{E}(\boldsymbol{C}_1\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2)}|U)^T\right)
\end{aligned}$$

Let $\boldsymbol{A}^{-}$ denote the generalized inverse of any matrix $\boldsymbol{A}$, then we have the following theorem.

Theorem 3.4.3. *Suppose the conditions (C1)-(C11) in the appendix hold. If* $h \to 0$, $nh^3 \to \infty$, $nh^4 \to 0$, $\lambda_n \to 0$ *and* $n^{1/2}\lambda_n \to \infty$ *as* $n \to \infty$, *we have*

$$Var(\hat{\boldsymbol{\alpha}}_1(t))^{-1/2}\{\hat{\boldsymbol{\alpha}}_1(t) - \boldsymbol{\alpha}_1(t) + (\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T(\boldsymbol{W} + \boldsymbol{\Sigma})^{-}\boldsymbol{b})\} \to N(\boldsymbol{0}, \boldsymbol{I}_s)$$

where $Var(\hat{\boldsymbol{\alpha}}_1(t)) = (\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T\, Var(\hat{\boldsymbol{\beta}}_1)(\boldsymbol{I}_s \otimes \boldsymbol{B}(t))$ *and* $Var(\hat{\boldsymbol{\beta}}_1) = \frac{1}{n}(\boldsymbol{W}+\boldsymbol{\Sigma})^{-}\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^{-}$.

Theorem 3.4.4. *When* $\boldsymbol{\alpha}(\cdot)$ *is a known constant, and* h_2 *is a constant and does not depend on the* n. *Under the regularity conditions (C1)-(C11), if* $h \to 0$, $nh \to \infty$ *as* $n \to \infty$, *then*

$$\sqrt{nh}\left(\hat{g}(u;\boldsymbol{\alpha}) - g_0(u) - \frac{1}{2}h^2\mu_2 g_0''(u)\right) \xrightarrow{d} N\left(0, \frac{\nu_0 G(u,h_2)}{f(u)F(u,h_2)^2}\right) \tag{3.14}$$

3.5 Relative Efficiency, Tuning Parameters and Influence Function

3.5.1 Bandwidth selection and relative efficiency

The distinguishing feature of our proposed method is that it introduces an additional tuning parameters h_2, which does not depend on the sample size. In this section we show that this parameter can be automatically selected by the observed data in order to achieve both robustness and efficiency of the resulting estimators.

Based on the theorem 3.4.4 and the asymptotic variance of the estimator (Fan and Gijbels [1996]) from local linear regression (LLR) for $g(\cdot)$, we can show that the

ratio of asymptotic variance of the LMR estimator to that of the LLR estimator for $g(\cdot)$ is given by

$$R(u, h_2) = \frac{G(u, h_2)F^{-2}(u, h_2)}{\mathrm{Var}(\epsilon|U = u)}. \tag{3.15}$$

Therefore, the optimal choice of h_2 is obtained by minimizing (3.15),

$$h_{2,opt} = \underset{h_2}{\mathrm{argmin}}\, R(u, h_2) = \underset{h_2}{\mathrm{argmin}}\, G(u, h_2)F^{-2}(u, h_2) \tag{3.16}$$

from where we also can see the optimal bandwidth only depends on the conditional distribution of ϵ given $U = u$.

To obtain the optimal bandwidth h_{opt}, we can minimize the mean square error (MSE) of $\hat{g}(u)$, based on theorem 3.4.4 and we have

$$h_{opt} = R(u, h_{2,opt})^{1/5} h_{LLR}, \tag{3.17}$$

where

$$h_{LLR} = \left(\frac{\nu_0}{\mu_2^2} \frac{\mathrm{Var}(\epsilon|U = u)}{g_0''(u)^2 f(u)} \right)^{1/5} \times n^{-1/5}$$

is the asymptotic optimal bandwidth for LLR (Fan and Gijbels [1996]).

The asymptotic relative efficiency (ARE) between LMR estimator with h_{opt} and $h_{2,opt}$ and the LLR estimator with h_{LLR} can be easily obtained,

$$ARE = \frac{MSE(LLR)}{MSE(LMR)} = R(u, h_{2,opt})^{-4/5},$$

from where we can see that $R(u, h_2)$ completely determines the ARE, so we only

need to study $R(u, h_2)$. The properties of $R(u, h_2)$ are presented as follows,

Theorem 3.5.1. *Let $g_{\epsilon|u}(\cdot)$ be the conditional density of ϵ given $U = u$. For $R(u, h_2)$, we have the following results*

(a) If $g_{\epsilon|u}(\cdot)$ is a normal density, $R(u, h_2) > 1$ for any finite h_2 and $\lim_{h_2 \to \infty} R(u, h_2) = 1$;

(b) When $g_{\epsilon|u}(\cdot)$ is a mixed normal distribution $\alpha N(0, \sigma^2(u)) + (1-\alpha)N(0, \delta^2(u))$ with $\alpha \in (0.5, 1)$ and $\delta^2(u) > \sigma^2(u)$, if $h_2 > \frac{(\alpha\sigma^2 + (1-\alpha)\delta^2)^2}{2\alpha(1-\alpha)\sigma^2\delta^2}$, then we have $R(u, h_2) < 1$.

From (a) one can see that when ϵ follows a normal distribution, the optimal LMR (with $h_2 \to \infty$) is the same as LLR. This is the reason why LMR will not lose efficiency under normal distribution. From (b) one can see that when the normal distributed is contaminated with another normal distribution with larger variance, we can find a h_2 such that the LMR estimator is better than LLR estimator in terms of efficiency.

Using a similar argument as Theorem 3.4.3, we obtain the asymptotic variance of the least square based penalized estimator for $\boldsymbol{\alpha}(t)$, which is

$$\frac{1}{n}(\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T(\boldsymbol{A}_0 + \boldsymbol{\Sigma})^- \boldsymbol{A}_2(\boldsymbol{A}_0 + \boldsymbol{\Sigma})^-(\boldsymbol{I}_s \otimes \boldsymbol{B}(t))$$

where $\boldsymbol{A}_k = \mathrm{E}\{|\epsilon|^k g_0'(U)^2(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)^T)\}$ with $k = 0, 2$, and $\boldsymbol{\Sigma}$ is defined in (3.14). When ϵ is independent of U, $\boldsymbol{Q}$ and $\boldsymbol{W}$ in Theorem (3.4.3) is equivalent to $\boldsymbol{Q} \equiv \boldsymbol{Q}(h_2) = \mathrm{E}\left(\phi_{h_2}'(\epsilon)^2\right) \mathrm{E}\left(g_0'(U)^2(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)^T\right)$ and $\boldsymbol{W} \equiv \boldsymbol{W}(h_2) = \mathrm{E}\left(\phi_{h_2}''(\epsilon)\right) \mathrm{E}\left(g_0'(U)^2(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1|U)^T\right)$

We can define the ratio of the asymptotic variance of the LMR estimator to that

of the LLR estimator as follows

$$R_\alpha(h_2,t)=\frac{tr\{(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))^T((\boldsymbol{W}+\boldsymbol{\Sigma})^-\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^-)(\boldsymbol{I}_s\otimes\boldsymbol{B}(t)))\}}{tr\{(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))^T((\boldsymbol{A}_0+\boldsymbol{\Sigma})^-\boldsymbol{A}_2(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-)(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))\}}$$

it is not hard to see that $R_\alpha(h_2,t)\to G(h_2)F(h_2)^{-2}$ as $n\to\infty$. Therefore, the $h_{2,opt}$ from (3.16) is also the minimizer of $R_\alpha(h_2)$ if ϵ is independent of U. Thus the asymptotic relative efficiency between the estimator based on LMR method and the LLR estimator is

$$ARE=R_\alpha(h_2,t)^{-1}\to G(h_2)^{-1}F(h_2)^2$$

Further we have the following results.

Theorem 3.5.2. *Under the conditions given in Theorem 3.4.3, we have*

$$\inf_{h_2}(\boldsymbol{W}+\boldsymbol{\Sigma})^-\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^-\leqslant(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-\boldsymbol{A}_2(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-$$

and the equality holds true when ϵ is normal and independent of U. That is, the LMR estimator is at least efficient as the least square estimator.

3.5.2 Tuning parameter in practice

In practice, we need to get the estimate $F(u_i,h_2)$ and $G(u_i,h_2)$ for every u_i, then define

$$\hat{F}(h_2)=\frac{1}{n}\sum_{i=1}^n\phi''_{h_2}(\hat{\epsilon}_i)\text{ and }\hat{G}(h_2)=\frac{1}{n}\sum_{i=1}^n\{\phi'_{h_2}(\hat{\epsilon}_i)\}^2.$$

Then $R(h_2)$ can be estimated by $\hat{R}(h_2) = \hat{G}(h_2)\hat{F}^{-2}(h_2)/\sigma^2$, and the optimal $\hat{h}_{2,opt}$ is obtained by minimizing $\hat{R}(h_2)$. Based on our experience, h_2 is usually comparable to σ, so we use grid search method for finding h_2, and the grid points are $0.5\hat{\sigma} \times 1.02^j$, $j = 0, ..., 90$, as suggested by Yao et al. [2005].

The optimal h is much easier to estimate if $\hat{h}_{2,opt}$ is found. By (3.17), we can have

$$\hat{h}_{opt} = \hat{R}(\hat{h}_{2,opt})^{1/5}\hat{h}_{LLR}$$

where $\hat{h}_{LLR} = \left(\frac{\hat{\nu}_0}{\hat{\mu}_2^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\sigma}_i^2}{\hat{g}''(u_i)^2\hat{f}(u_i)}\right)^{1/5}$ can be obtained by plug-in method (Fan and Gijbels [1996]).

However, u_i is not available because $\boldsymbol{\beta}$ is not known. To address this problem, (1) we get an initial estimate of $\boldsymbol{\beta}$ and $g(u)$ by using linear regression and LOWESS, respectively. (2), we can estimate (or re-estimate) the $\hat{h}_{2,opt}$ and $\hat{h}_{opt}$. (3), we can update the $\boldsymbol{\beta}$ and $\hat{g}$ for 5 iterations. (4) Repeat (2) and (3) until all are converged.

Due to the heavily computational burden of our algorithm, tuning K and λ_n simultaneously will be unrealistic. For simplicity, we set $K = 0.5n^{0.5}$ where n is the sample size in the simulation and application part. The regularization parameter λ_n in our proposed method controls the amount of shrinkage of the coefficient functions. By SCAD penalty function, parameter greater than $a\lambda_n$ will not be penalized, and parameter less than λ_n will shrink to exact zero. Thus, the size of model is determined by λ_n and appropriately choosing λ_n is the most important for model selection. Since we need to select the optimal λ_n from a large range of values by using grid search, cross validation is too computationally expensive to be used here. Instead, we consider the BIC-type criterion, for the reason that BIC results in a consistent estimator, as mentioned in Peng and Huang [2011].

The BIC-type criterion is defined as follows

$$\text{BIC-type}_\lambda = n\log\left(\text{RSS}_\lambda/n\right) + m \times \log(n) \tag{3.18}$$

where $\text{RSS}_\lambda = \sum_{i=1}^{n}\left(y_i - \hat{g}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}})\right)^2$ is the residual sum of squares, and m denotes the number of nonzero estimates in $\hat{\boldsymbol{\beta}}$.

3.5.3 Influence function

According to Huber [2011], the influence function of parameter θ at distribution F is defined

$$\psi_{\hat{\theta},F} = \lim_{\epsilon\to 0}\frac{\hat{\theta}(F_\epsilon(x)) - \hat{\theta}(F)}{\epsilon}$$

where $F_\epsilon(x)$ is called the ϵ-contaminated distribution function which is defined as

$$F_\epsilon(x) = (1-\epsilon)F + \epsilon\delta_x$$

where δ_x is the probability measure which assigns probability 1 to x and 0 to all other elements. By the definition, it can be seen that the influence function is the marginal effect of oversampling x on a particular estimator for an uncontaminated distribution.

In the FSIM, suppose that $(\boldsymbol{C}, y)$ is a generic observation from the distribution F. Recall that given $\boldsymbol{\beta}$, for any fixed u, we estimate $g(u)$ by maximizing the following objective function

$$\rho(\boldsymbol{C}, y; a) = \text{E}\phi_{h_2}\left(y - a - b(\boldsymbol{C}^T\boldsymbol{\beta} - u)\right)K_h(\boldsymbol{C}^T\boldsymbol{\beta} - u),$$

with $\hat{g}(u) = \hat{a}$. According to Huber [2011], one can have the influence function of $\hat{g}(u)$ at $(\tilde{\boldsymbol{C}}, \tilde{y})$ as follows

$$\begin{aligned} IF(\tilde{\boldsymbol{C}}, \tilde{y}; a) &= -\mathrm{E}[\frac{\partial}{\partial a}\psi(\boldsymbol{C}, y; a)]^{-1}\psi(\tilde{\boldsymbol{C}}, \tilde{y}; a) \\ &= [\mathrm{E}(\phi''_{h_2}(r)K_h(\boldsymbol{C}^T\boldsymbol{\beta} - u)]^{-1}\phi'_{h_2}(\tilde{r})K_h(\tilde{\boldsymbol{C}}^T\boldsymbol{\beta} - u) \end{aligned}$$

where $r = y - a - b(\boldsymbol{C}^T\boldsymbol{\beta} - u)$ and $\tilde{r} = \tilde{y} - a - b(\tilde{\boldsymbol{C}}^T\boldsymbol{\beta} - u)$.

With observations $(\boldsymbol{C}_i, y_i)$, $i = 1, ..., n$, the associated empirical influence function of $g(u)$ can be obtained

$$\hat{IF}(\tilde{\boldsymbol{C}}, \tilde{y}; \hat{a}) = [\frac{1}{n}\sum_{i=1}^{n}\phi''_{h_2}(r_i)K_h(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}} - u)]^{-1}\phi'_{h_2}(\tilde{r})K_h(\tilde{\boldsymbol{C}}^T\hat{\boldsymbol{\beta}} - u).$$

Note that $\hat{IF}(\tilde{\boldsymbol{C}}, \tilde{y}; \hat{a})$ measures the marginal effect of a new observation $(\tilde{\boldsymbol{C}}, \tilde{y})$ to the existing estimator $\hat{g}$ at u. We can see that $\hat{IF}(\tilde{\boldsymbol{C}}, \tilde{y}; \hat{a})$ is also a function of h_2 and h. It is difficult to derive a specific relationship between $\hat{IF}$ and h or h_2. However, one might expect that a smaller h would lead to a smaller value of $\hat{IF}$ because $K_h(\tilde{\boldsymbol{C}}^T\hat{\boldsymbol{\beta}} - u)$ will decrease significantly as h decreases, on the other side, the denominator will keep positive because of the neighboring points of u. For the h_2, one might expect a larger h_2 will lead to a larger value of $\hat{IF}$.

If proper h and h_2 are given, for the LMR estimator, the value of influence function of an outlier with larger absolute value of $\tilde{r}$ will be very small, because $\phi'_{h_2}(\tilde{r})$ goes to 0 as $|\tilde{r}| \to \infty$. However, it is easy to see that this value will be unbounded for LLR estimator (by simply replacing the loss $-\phi_{h_2}(t)$ with t^2). This implies that marginal effect of an outlier is relatively smaller for LMR estimator than that of the LLR estimator, that is, LMR is more robust than LLR.

The influence function of estimator of $\boldsymbol{\beta}$ can be similarly derived. Given $g(\cdot)$ and

a generic observation $(\boldsymbol{C}, y)$, the objective function for estimating the $\boldsymbol{\beta}$ is

$$\rho(\boldsymbol{C}, y; \boldsymbol{\beta}) = \mathrm{E}\left[\phi_{h_2}(y - g(\boldsymbol{C}^T\boldsymbol{\beta})\right] - \sum_{j=1}^{p} p_\lambda(||\boldsymbol{\beta_j}||_H),$$

by taking derivative with respect to $\boldsymbol{\beta}$ twice, we have

$$\begin{aligned}
\psi(\boldsymbol{C}, y; \boldsymbol{\beta}) &= \partial\rho(\boldsymbol{C}, y; \boldsymbol{\beta})/\partial\boldsymbol{\beta} \\
&= -\phi'_{h_2}(r)g'(\boldsymbol{C}^T\boldsymbol{\beta})\boldsymbol{C} - \sum_{j=1}^{p} p'_\lambda(||\boldsymbol{\beta_j}||_H)\frac{\partial||\boldsymbol{\beta_j}||_H}{\partial\boldsymbol{\beta}}
\end{aligned}$$

and

$$\begin{aligned}
\psi'(\boldsymbol{C}, y; \boldsymbol{\beta}) &= \partial\psi(\boldsymbol{C}, y; \boldsymbol{\beta})/\partial\boldsymbol{\beta} \\
&= [\phi''_{h_2}(r)g'(\boldsymbol{C}^T\boldsymbol{\beta})^2 - \phi'_{h_2}(r)g''(\boldsymbol{C}^T\boldsymbol{\beta})]\boldsymbol{C}\boldsymbol{C}^T - \sum_{j=1}^{p} p'_\lambda(||\boldsymbol{\beta_j}||_H)\frac{\partial||\boldsymbol{\beta_j}||_H}{\partial\boldsymbol{\beta}} \\
&\quad - \sum_{j=1}^{p}\left(p'_\lambda(||\boldsymbol{\beta_j}||_H)\frac{\partial^2||\boldsymbol{\beta_j}||_H}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} + p''_\lambda(||\boldsymbol{\beta_j}||_H)\frac{\partial||\boldsymbol{\beta_j}||_H}{\partial\boldsymbol{\beta}}\frac{\partial||\boldsymbol{\beta_j}||_H}{\partial\boldsymbol{\beta}^T}\right)
\end{aligned}$$

Therefore, the influence function for $\hat{\boldsymbol{\beta}}$ at at $(\tilde{\boldsymbol{C}}, \tilde{y})$ is

$$IF(\boldsymbol{C}, y; \boldsymbol{\beta}) = -[E\psi'(\boldsymbol{C}, y; \boldsymbol{\beta})]^{-1} \times \psi(\tilde{\boldsymbol{C}}, \tilde{y}; \boldsymbol{\beta})$$

and the associated empirical influence function with observations $(\boldsymbol{C}_i, y_i)$, $i = 1, ..., n$ is

$$\hat{IF}(\boldsymbol{C}, y; \hat{\boldsymbol{\beta}}) = -[E\psi'(\boldsymbol{C}, y; \hat{\boldsymbol{\beta}})]^{-1} \times \psi(\tilde{\boldsymbol{C}}, \tilde{y}; \hat{\boldsymbol{\beta}})$$

3.6 Simulation Studies

In this section, a simulation study is conducted to assess the finite sample performance of the proposed method.

We use the following data generator to generate the functional predictors. For $i = 1, 2, ..., n$ and $j = 1, ..., 6$,

$$x_{ij}(t) = \begin{cases} (10 + \eta_{ij})\frac{\exp(5t)}{1+\exp(5t)} + e_{ij} & j = 1, 4 \\ (10 + \eta_{ij})\sin(5t) + e_{ij} & j = 2, 5 \\ (10 + \eta_{ij})\cos(5t) + e_{ij} & j = 3, 6 \end{cases}$$

where $t \in [0, 1]$, $\eta_{ij} \sim N(0, 3)$ and $e_{ij} \sim N(0, 6)$. Each predictor function is sampled through 21 equally spaced measurements in $[0, 1]$.

The true coefficient functions $\alpha_j(t)$ are given by

$$\alpha_1(t) = C_\alpha \left(\frac{1 + \sqrt{t}}{2}\right), \quad \alpha_2(t) = C_\alpha \sin(\pi t),$$
$$\alpha_k(t) = 0, \quad k = 3, ..., 6$$

where $C_\alpha > 0$ is a constant such that $||\boldsymbol{\alpha}|| = 1$, for the sake of model identifiability.

Letting $u_i = \sum_{j=1}^{6} \int_0^1 x_{ij}(t)\alpha_j(t)dt$, the scalar response y is generated as follows:

$$y_i = g(u_i) + \sigma(g(u_i))\epsilon_i, \tag{3.19}$$

where $g(\cdot)$ is the link function, ϵ_i is the error term, and σ controls the Signal-to-Noise Ratio (SNR). In this example, we use the following link function $g(u_i) = \sqrt{\frac{u_i - \min(u_i)}{\max(u_i) - \min(u_i)}}$ and $\sigma(g(u_i)) = 0.1sd(g(u_i))$, where sd standards for standard deviation.

In this example, we investigate 3 error distributions of ϵ with different sample sizes $n = 400$,

1. $N(0, 1)$. This density serves as a bench mark with no outliers;
2. $0.95N(0, 1) + 0.05N(0, 50)$, the normal errors contaminated by 5% outliers from another normal distribution $N(0, 50)$;
3. $t_3/\sqrt{3}$, the scaled heavy-tailed $t-$distribution with 3 degrees of freedom.

and other 2 estimators besides the proposed one: local linear regression (LLR) estimator and Huber estimator (The parameter of Huber loss is fixed at 1.345, leading to 95% efficiency when error is normal distributed and there are no outliers Huber [2011]).

For the model performance, the performance of the estimator $\hat{g}(\cdot)$ will be assessed using the square root of mean squared errors (RMSE(g))

$$\text{RMSE}(g) = \left(n^{-1} \sum_{i=1}^{n} ||\hat{g}(u_i) - g(u_i)||^2 \right)^{1/2}, \tag{3.20}$$

and the errors of the estimated coefficient functions are quantified by square root of squared error RSE($\hat{\boldsymbol{\alpha}}$) given as

$$\text{RSE}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{k=1}^{p} \int_0^1 \left(\hat{\alpha}_k(t) - \alpha_k(t)\right)^2 dt \right)^{1/2} \tag{3.21}$$

The performance of variable selection is measured by variable selection score (VS-score),

$$\text{VS-score} = \frac{\#\text{nonzero components identified} \times 2 + \#\text{zero components identified}}{\text{Full score}}$$

A total of 100 simulation replications are conducted for each model setup. In all simulations, we use the cubic B-splines with order $d = 4$ to approximate the coefficient functions, the bandwidth h, h_2, number of B-spline basis K, regularization parameter λ_n are tuned by methods described in Section 3.5.

Table 3.1 reports the RMSE, RSE and VS-score for the proposed LMR estimator, Huber estimator and LLR estimator for the different types of error distribution. The latter two are used for comparison to show the efficiency and robustness of the proposed estimator.

Several observations can be made from Table 3.1. First, when the error distribution is normal (Distribution 1), LMR and LLR outperform the Huber estimator in terms of the estimation of link function and coefficient functions, because Huber loses some efficiency. The comparable results between LMR and LLR show that LMR does not loss efficiency. Second, when the error distribution is contaminated with a proportion of outliers (Distribution 2), LLR performs much worse than LMR and Huber because its sensitivity to outliers; LMR performs better than Huber in this case. Finally, when the error distribution has a heavy tail (Distribution 3), LLR does not behave as well as LMR and Huber, and LMR and Huber has almost the same performance. Therefore, we can conclude that our proposed method perform best for different error distributions.

3.7 An Application in Growth Modeling

In this section, we apply our proposed method to the NIH cohort study. This study consists of 626 newborn infants living in an urban slum of Mirpur Thana in Dhaka, Bangladesh. Each study subject was visited for collecting information related to child morbidity about every 3 months.

Table 3.1: *Simulation results for different error distributions*

Error Distribution	Estimator	RASE(g)	RSE($\boldsymbol{\alpha}$)	VS-score
1	LMR	0.0103	0.267	0.994
	Huber	0.0125	0.271	1.000
	LLR	0.0105	0.269	0.988
2	LMR	0.0152	0.283	0.964
	Huber	0.0159	0.291	0.950
	LLR	0.0243	0.297	0.950
3	LMR	0.0102	0.267	0.999
	Huber	0.0102	0.267	0.994
	LLR	0.0143	0.271	0.993

Our interest is to investigate the relationship between HAZ at month 36 and 5 predictors (weight(kg), height(cm), family income, maternal education and water quality), where weight(kg) and height(cm) are functional predictors over the time interval $[0, 24]$ (month) and others are time invariant variables.

Some data preprocessing is conducted on functional predictors before modeling since the these measurements are not perfectly temporally aligned. First, to improve the model estimation, we delete the subjects with fewer than 5 height measurements. Second, for the height, we use the monotone estimation method proposed in Chapter 2 since height should be monotone over the month $0 - 24$; while weight is not necessary to be increasing over the time, so we use existing method PACE. Next, we interpolate equally distant 12 points on $[0, 24]$ as the input of the proposed FSIM model.

After data preprocessing, we use the remaining 355 children in our analysis. A functional single index model is built to investigate the relationship between HAZ at month 36 and these 5 predictors. The group SCAD penalty is employed for model

selection. The functional single index model can be written as follows:

$$y_i = g\left(\int_0^{24} \alpha_1(t)X_{i1}(t)dt + \int_0^{24} \alpha_2(t)X_{i2}(t)dt\right) + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

where y_i is the HAZ at month 36, $X_{ij}(\cdot), j = 1, 2$ are weight and height respectively and $X_{ij}, j = 3, 4, 5$ are family income, maternal education and water quality, $i = 1, ..., 355$. Figure 3.1 shows the distribution of the predictors (after pre-smoothing) and the response. The tuning parameters are selected by the method in Section 3.5. Table 3.2 shows the summary statistics of and the estimation the scalar predictors. All the data are standardized before put into the model.

Table 3.2: *Summary statistics and estimation of scalar predictors*

Predictors	Mean (Proportion)	SD	Estimation
Income (Taka)	6977	3194	0.0275
Maternal Education	0.627	–	0.1071
Water	0.955	–	0.0000

The estimations in Table 3.2 indicates that an increase of a unit standard deviation of income results in an increase of 0.0275 in HAZ at year 3 and maternal education is able to raise the HAZ by 0.1071.

Figure 3.2 shows the estimation of nonzero coefficient functions and the link function with their 95% confidence intervals. Link function is basically monotone, like a piecewise linear function of the index u. The effect of height is positive after Month 18. Therefore, we might expect that a larger value of height after Month 18 and a larger value of index is more likely to have a higher HAZ at year 3.

Figure 3.3 shows two examples in the study with different index values. We can see that the black line have much larger value of height after Month 18 than red one, and their index are 0.56 and -5.09. Their HAZ at year 3 are 1.11 and -3.28.

This observation is consistent with our expectation.

We use Test RMSE as metric to evaluate the model prediction. We split the data into training data and test data. For the training data, we use cross validation to choose the optimal parameters. Specifically, we split the training data into 5 partitions, at k-th round ($k = 1, ..., 5$), the k-th partition will be used for validation and others for training model. The best model is chosen based on the best average RMSE of validation data over all 5 rounds. Then the Test RMSE is evaluated on the test data. We also evaluate prediction performance of the random forest, which is a very powerful predictive model. Table 3.3 shows the results.

Table 3.3: *Test RMSE of proposed method and Random Forest*

	Test RMSE
FSIM	0.80
RF	0.58

From Table 3.3 we can see that our proposed model is not as good as random forest in terms of prediction. However, random forest is not interpretable.

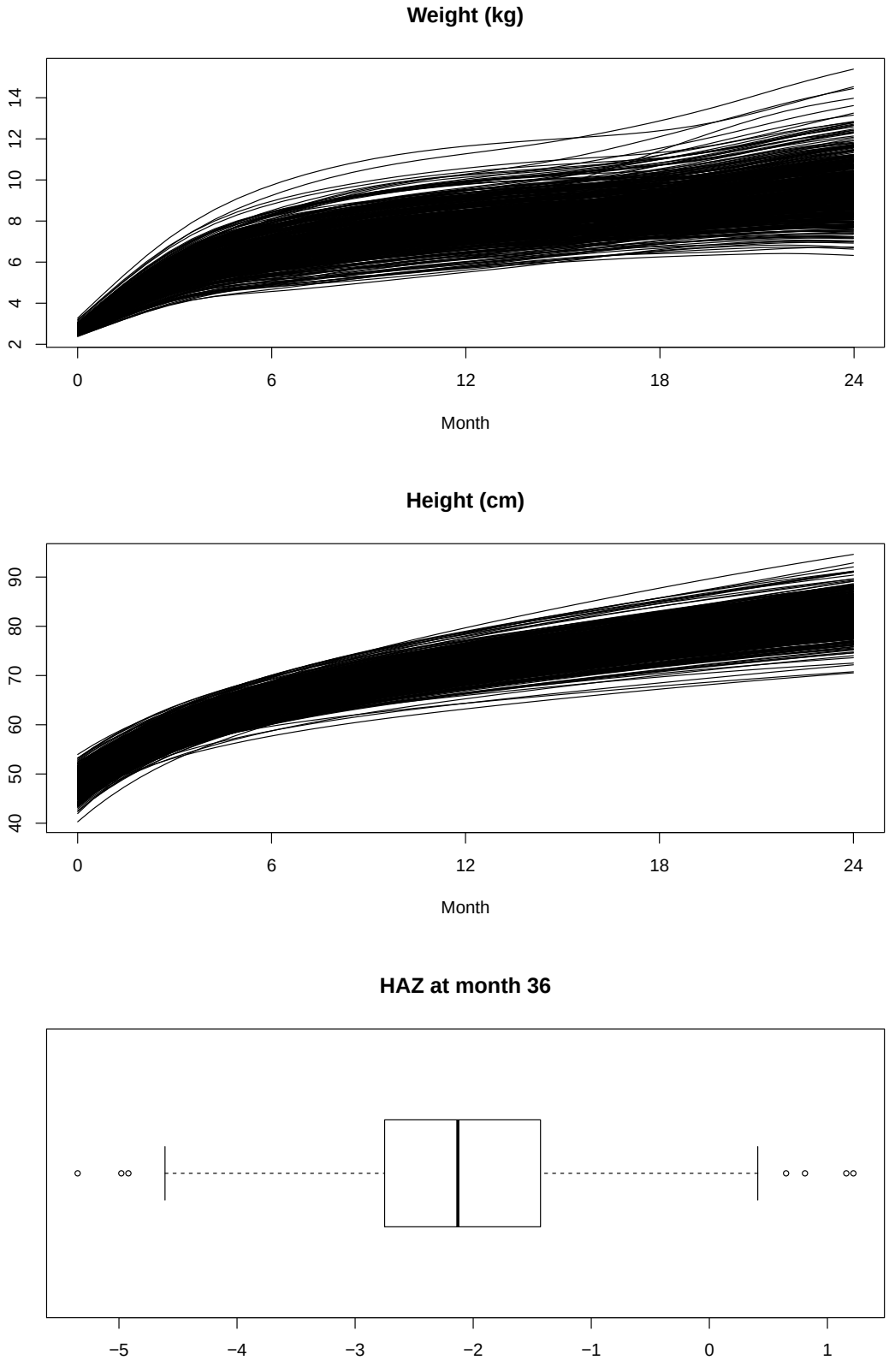

Figure 3.1: Distribution of the predictors and response

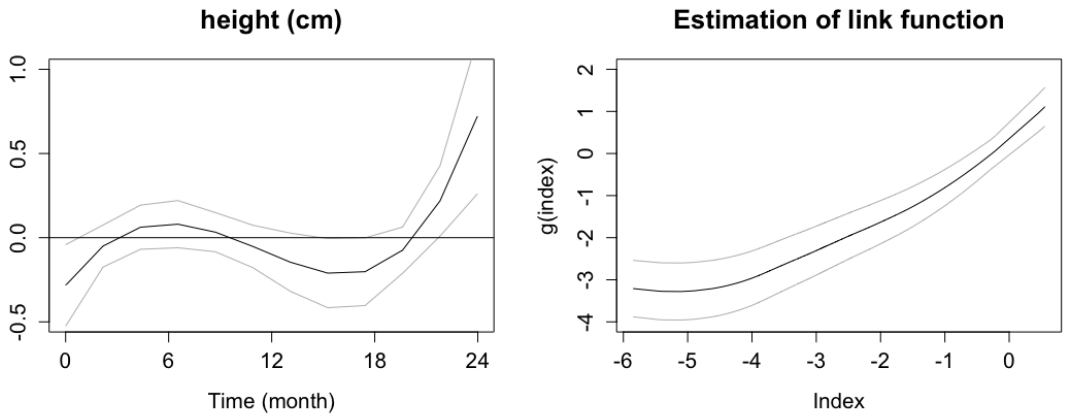

Figure 3.2: Estimation of nonzero functional coefficient and link function with 95% pointwise confidence interval

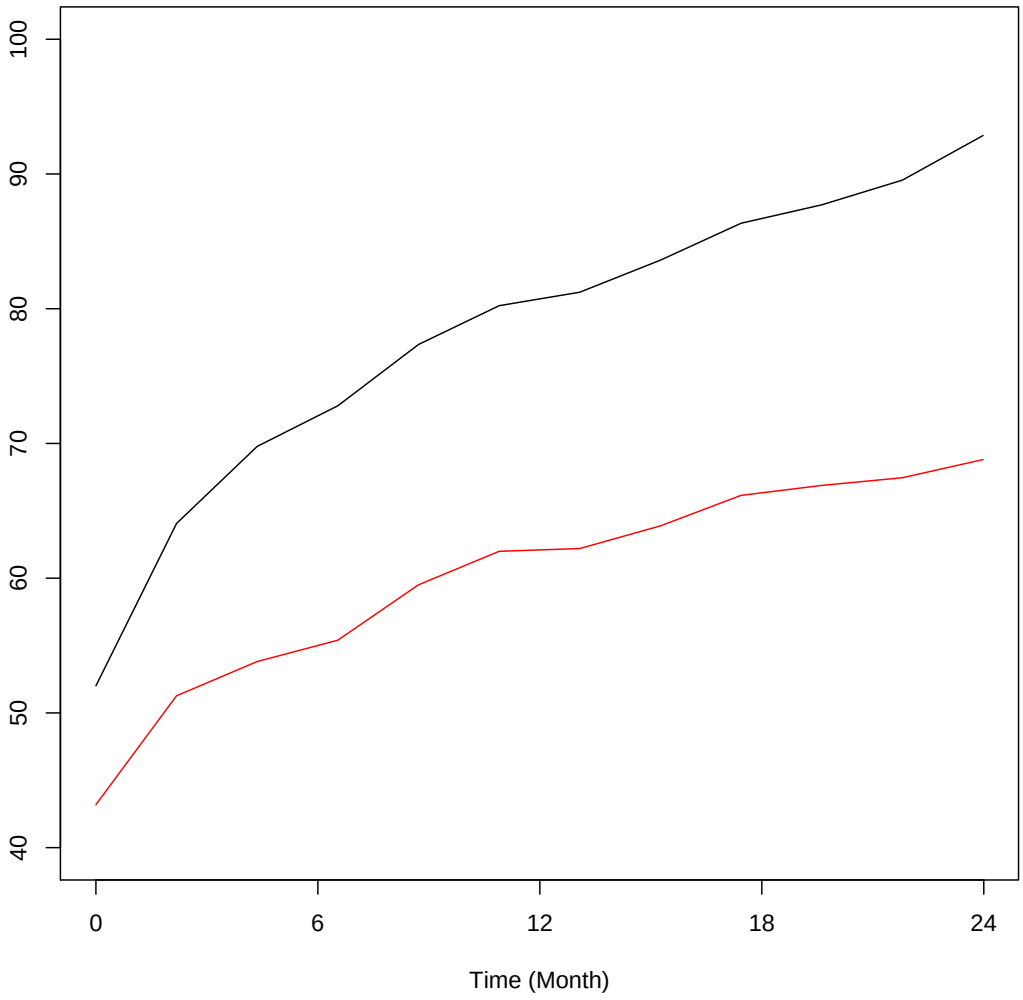

Figure 3.3: Estimation of nonzero functional coefficient and link function

Chapter 4

Proofs

4.1 Proof of Theorems in Chapter 2

Proof of Theorem 2.2.1. Since

$$\mathrm{E}(\xi_{ik} - \boldsymbol{\alpha}_{ik}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i))^2 \quad = \quad \lambda_k - 2\lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\alpha}_{ik} + \boldsymbol{\alpha}_{ik}^T \boldsymbol{\Sigma}_{W_i} \boldsymbol{\alpha}_{ik},$$

by taking the partial derivative with respect to $\boldsymbol{\alpha}_{ik}$ and setting it to zero, we have

$$\tilde{\boldsymbol{\alpha}}_{ik} = \boldsymbol{\Sigma}_{W_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k,$$

then $\tilde{\boldsymbol{\alpha}}_{ik}^T(\boldsymbol{W}_i - \boldsymbol{\mu}_i) = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{W_i}^{-1}(\tilde{\boldsymbol{W}}_i - \tilde{\boldsymbol{\mu}}_i)$ is the BLUP of ξ_{ik}. □

Proof of Theorem 2.3.1. We suppress i in the proof. Let $\boldsymbol{c}_0$ is the B-spline coefficient of $W_0(\cdot)$ given the B-spline basis functions $\boldsymbol{B}(t)$. Denote $\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{c}_0^T)^T$ be the true parameter. Let $\boldsymbol{v}$ be a vector. Define $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \delta \boldsymbol{v}$. Define $n = EN$. Let $Q(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n (y_i - \beta - \int_{T_1}^t \exp(\boldsymbol{c}^T \boldsymbol{B}(s))ds)^2$. We first show that, for any given

$\epsilon > 0$, there exists a large $C > 0$ such that

$$P\{\sup_{||\boldsymbol{v}||=C} Q(\tilde{\theta}) - Q(\theta) > 0\} \geqslant 1 - \epsilon$$

This implies with probability at least $1 - \epsilon$ that there exists a local minimum in the ball $\{\theta_0 + \delta\boldsymbol{v} : ||\boldsymbol{v}|| \leqslant C\}$. Hence, there exists a local maximizer such that $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}|| = O_p(\delta)$. It is easy to derive that

$$\begin{aligned} Q(\tilde{\theta}) - Q(\theta) &= \frac{1}{n}\sum_i (y_i - \beta - m(t_i))^2 - \frac{1}{n}\sum_i (y_i - \beta_0 - m_0(t_i))^2 \\ &= \frac{2}{n}\sum_i (y_i - \beta_0 - m_0(t_i))[(\beta_0 - \beta)(m_0(t_i) - m(t_i))] \\ &\quad + \frac{1}{n}\sum_i ((\beta_0 - \beta)(m_0(t_i) - m(t_i)))^2 \\ &= 2I_1 + I_2 \end{aligned}$$

We first investigate

$$\begin{aligned} m(t_i) - m_0(t_i) &= \int_{T_1}^{t_i} \exp(\boldsymbol{c}^T\boldsymbol{B}(s))ds - \int_{T_1}^{t_i} \exp(\boldsymbol{c}_0^T\boldsymbol{B}(s))ds \\ &= \left(\int_{T_1}^{t_i} \exp(\boldsymbol{c}_0^T\boldsymbol{B}(s))\boldsymbol{B}^T(s)ds\right)(\boldsymbol{c} - \boldsymbol{c}_0 + o_p(1))^T \\ &= \boldsymbol{M}^T(\boldsymbol{c} - \boldsymbol{c}_0 + o_p(\mathbf{1})) \end{aligned}$$

where $\boldsymbol{M} = \int_{T_1}^{t_i} \exp(\boldsymbol{c}_0^T\boldsymbol{B}(s))\boldsymbol{B}(s)ds$, then

$$\begin{aligned} I_2 &= \frac{1}{n}\sum_i ((\beta_0 - \beta)(m_0(t_i) - m(t_i)))^2 \\ &= \frac{1}{n}\sum_i ((\beta_0 - \beta)(\boldsymbol{M}^T(\boldsymbol{c} - \boldsymbol{c}_0 + o_p(\mathbf{1})))^2 \\ &\geqslant C_0(\frac{\delta^2||\boldsymbol{v}||^2}{n}) \end{aligned}$$

where $C_0 > 0$ is a constant depending on $\boldsymbol{v}$. By similar arguments,

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_i (y_i - \beta_0 - m_0(t_i))[(\beta_0 - \beta)(m_0(t_i) - m(t_i))] \\ &= \frac{1}{n}\sum_i \epsilon_i[(\beta_0 - \beta)(m_0(t_i) - m(t_i))] \\ &= O_p(\frac{\delta^2||\boldsymbol{v}||}{n}) \end{aligned}$$

It can be seen that by choosing an appropriate $\boldsymbol{v}$, we always can find a positive C_0 to make sure the above inequality hold. Hence, I_2 dominates I_1 if $||\boldsymbol{v}||$ is large enough. Therefore, there exists local minimizer $\tilde{\beta}$ and $\tilde{\boldsymbol{c}}$ such that $\tilde{\beta} - \beta = O_p(\delta)$ and $||\tilde{\boldsymbol{c}} - \boldsymbol{c}|| = O_p(\delta)$. Hence,

$$\begin{aligned} \left(\int_{T_1}^{T_2} (\tilde{W}(t) - W(t))^2 dt\right) &= \int_{T_1}^{T_2} (\boldsymbol{B}^T(t)(\tilde{\boldsymbol{c}} - \boldsymbol{c}))dt \\ &= (\tilde{\boldsymbol{c}} - \boldsymbol{c})^T \int \boldsymbol{B}\boldsymbol{B}^T (\tilde{\boldsymbol{c}} - \boldsymbol{c}) \\ &= O_p(\delta^2) \end{aligned}$$

since $\int \boldsymbol{B}\boldsymbol{B}^T = O(1)$, this completes the proof. □

Proof of Theorem 2.3.2. Denote $\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{c}_0^T)^T$ be the true parameter. Denote

$\tilde{\boldsymbol{\theta}}$ as the minimizer of 2.3, then we have

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{N_i} (y_j - \beta - m(t_j))^2 = \sum_{j=1}^{N_i} \left((y_j - \beta - \int \exp(\boldsymbol{c}^T \boldsymbol{B}(s)) ds \right)^2 \\
&= \sum_{j=1}^{N_i} \left(y_j - \beta - \int \exp(\boldsymbol{c}^T \boldsymbol{B}(s)) ds - \boldsymbol{z}_j^T (\boldsymbol{c} - \boldsymbol{c}_0 + o(1)) \right)^2 \\
&= \sum_{j=1}^{N_i} \left(\epsilon_j - (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{z}_j^T (\boldsymbol{c} - \boldsymbol{c}_0 + o(1)) \right) \\
&= \sum_{j=1}^{N_i} \left(\epsilon_j \boldsymbol{q}_j^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0 + o(1)) \right)
\end{aligned}$$

where $\boldsymbol{q}_j = (1, \boldsymbol{z}_j^T)^T$. Therefore, we have

$$\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \rightarrow N(0, (\boldsymbol{Q}^T \boldsymbol{Q})^{-1} \sigma^2)$$

and

$$\begin{aligned}
\sqrt{n}(\tilde{\beta} - \beta_0) &\rightarrow N(0, \sigma^2) \\
\sqrt{n}(\tilde{\boldsymbol{c}} - \boldsymbol{c}_0) &\rightarrow N(0, (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \sigma^2)
\end{aligned}$$

Since $\tilde{W}(t) = \boldsymbol{B}^T(t) \tilde{\boldsymbol{c}}$, it is easy to derive that

$$\sqrt{n}(\tilde{W}(t) - W(t)) \rightarrow N(0, \boldsymbol{B}^T(t) (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{B}(t) \sigma^2)$$

□

Let $\tilde{\boldsymbol{T}}_i = (T_{i1}, ..., T_{iN_i})^T$. We assume that $W_i(t)$ can be well approximated by $\tilde{W}_i(t)$ for $t \in [T_{i1}, T_{i2}]$. Then $\tilde{\boldsymbol{W}}_i = (\tilde{W}_i(T_{i1}), ..., \tilde{W}_i(T_{iN_i}))^T$, corresponding to the values of $\tilde{W}_i(t)$ at time $\tilde{\boldsymbol{T}}_i$, can be viewed as observed data. We assume the data

$(T_{ij}, \tilde{W}_{ij})$, from model (ref), have the same distribution as $(T, \tilde{W})$, with joint distribution $g(t, w)$. Assume that the observation times T_{ij} are iid with marginal density $f(t)$, but that dependence is allowed between observations $\tilde{W}_{ij}$ and $\tilde{W}_{ik}$, coming from the same subject. We have following assumptions,

(A1.1) The number of observations N_i made for the ith subject or cluster is a random variable with $N_i \overset{iid}{\sim} N$, where $N > 0$ is a positive discrete random variable, with $EN < \infty$ and $P(N > 1) > 0$.

(A1.2) $(T_{ij}, \tilde{W}_{ij})$ is independent of N_i

Assume that $(T_{ij}, T_{il}, \tilde{W}_{ij}, \tilde{W}_{il})$ is distributed as $(T_1, T_2, \tilde{W}_1, \tilde{W}_2)$ with joint density function $g_2(t_1, t_2, w_1, w_2)$. We assume regularity conditions for the marginal and joint densities, $f(t), g(t, w)$ and $g_2(t_1, t_2, w_1, w_2)$.

(B1.1) $(d^2/dt^2)f(t)$ exists and is continuous on $\mathcal{T}$ and $f(t) > 0$ on $\mathcal{T}$.

(B1.2) $(d^2/dt^2)g(t, y)$ exists and is uniformly continuous on $\mathcal{T} \times R$.

(B1.3) $(d^2/dt_1^{l_1} dt_2^{l_2})g_2(t_1, t_2, y_1, y_2)$ exists and is uniformly continuous on $\mathcal{T}^2 \times R^2$, for $l_1 + l_2 = l$,$0 \leqslant l_1, l_2 \leqslant 2$.

Let $k_1(\cdot)$ and $k_2(\cdot, \cdot)$ be nonnegative univariate and bivariate kernel functions used in the smoothing steps for the mean μ and covariance G in Section (2.2.2). We assume that k_1 and k_2 are compactly supported densities with following properities

(B2.1a) k_1 is compactly supported, $||k_1||^2 = \int k_1^2(u)du < \infty$.

(B2.2a) k_1 is a kernel function of order $(0, 2)$.

(B2.1b) k_2 is compactly supported, $||k_2||^2 = \int k_2^2(u, v)dudv < \infty$.

(B2.2b) k_2 is a kernel function of order $((0, 0), 2)$.

Here we say that a bivariate kernel function k_2 is of order (ν, l), where ν is multi-index $\nu = (\nu_1, \nu_2)$, if

$$\int\int u^{l_1} v^{l_2} k_2(u,v)dudv$$
$$= \begin{cases} 0 & 0 \leqslant l_1 + l_2 < l, l_1 \neq nu_1, l_2 \neq nu_2 \\ (-1)^{|\nu|}|\nu|! & l_1 = nu_1, l_2 = nu_2 \\ \neq 0 & l_1 + l_2 = l \end{cases}$$

where $|\nu| = \nu_1 + \nu_2$. A univariate kernel k_1 is of order (ν, l) for a univariate $\nu = \nu_1$ and $l_2 = 0$ on the right side, integrating only over the argument u on the left side.

Let h_μ and h_G be the bandwidths for estimating $\hat{\mu}$ and $\hat{G}$. As the number of subjects $n \to \infty$, the following conditions are required.

(A2.1) $h_\mu \to 0$, $nh_\mu^4 \to \infty$, and $nh_\mu^6 < \infty$.

(A2.2) $h_G \to 0$, $nh_G^6 \to \infty$, and $nh_G^8 < \infty$.

Define the Fourier transforms of $k_1(u)$ and $k_2(u,v)$ by $\zeta_1(t) = \int e^{-iut} k_1(u)du$ and $\zeta_2(t,s) = \int e^{-(iut+ivs)} k_2(u,v)dudv$. They satisfy the following:

(A3.1) $\zeta_1(t)$ is absolutely integrable, that is, $\int |\zeta_1(t)|dt < \infty$.

(A3.2) $\zeta_2(t,s)$ is absolutely integrable, that is, $\int\int |\zeta_2(t,s)|dtds < \infty$.

Assume that the fourth moment of Y centered at $\mu(T)$ is finite, that is,

(A4) $E|Y - \mu(T)|^4 < \infty$.

Then we have the following lemmas about the uniform convergence rates for local linear estimators $\hat{\mu}(t)$ and $\hat{G}(s,t)$ on compact sets $\mathcal{T}$ and $\mathcal{T}^2$,

Lemma 4.1.1. *Suppose conditions (A1.1)-(A4) and (B1.1)-B(2.2b) hold, we have*

$$\sup_{t\in\mathcal{T}} |\hat{\mu}(t) - \mu(t)| = O_p(\frac{1}{\sqrt{N}h_\mu})$$

and

$$\sup_{t,s\in\mathcal{T}} |\hat{G}(s,t) - G(s,t)| = O_p(\frac{1}{\sqrt{N}h_G^2})$$

Next, consider

Lemma 4.1.2. *Suppose conditions (A1.1)-(A4) and (B1.1)-B(2.2b) hold, we have*

$$|\hat{\lambda}_k - \lambda_k| = O_p(\frac{1}{\sqrt{N}h_G^2})$$
$$||\hat{\phi}_k - \phi_k||_H = O_p(\frac{1}{\sqrt{N}h_G^2})$$

and

$$\sup_{t\in\mathcal{T}} |\hat{\phi}_k(t) - \phi_k(t)| = O_p(\frac{1}{\sqrt{N}h_G^2})$$

The next lemma shows that the target trajectory is well defined.

Lemma 4.1.3. *For the positive definite covariance operator G generated by the continuous symmetric function $G(s,t)$ on $\mathcal{T}^2$, as $K \to \infty$,*

$$\sup_{t\in[0,1]} E[\widetilde{W}_i^K(t) - \widetilde{W}_i(t)]^2 \to 0.$$

For the consistency theorem, we also assume that the data asymptotically follow a linear scheme

(A5) The number, location, and values of measurements for a given subject or cluster remain unaltered as $N \to \infty$.

Proof of Theorem 2.3.3. Recall that $\hat{\xi}_{ik,pen} = (1+\gamma\hat{D}_{ik})^{-1}\hat{\lambda}_k\hat{\boldsymbol{\phi}}_{ik}^T\hat{\boldsymbol{\Sigma}}_{W_i}^{-1}(\tilde{\boldsymbol{W}}_i - \hat{\boldsymbol{\mu}}_i)$. The (j,l)th entry of the matrix $\hat{\boldsymbol{\Sigma}}_{W_i}$ is $\hat{\boldsymbol{\Sigma}}_{W_i}(j,l) = \hat{G}(T_{ij}, T_{il})$. Applying lemma (4.1.1) and lemma (4.1.2) and Slutsky's theorem, (2.18) follows.

We next prove (2.19) and (2.20) for each fixed $t \in \mathcal{T}$. Let $\breve{W}_{i,pen}^K(t) = \mu(t) + \sum_{k=1}^K \tilde{\xi}_{ik,pen}\phi_k(t)$. Note that

$$|\hat{W}_{i,pen}^K(t) - \breve{W}_{i,pen}(t)| \leqslant |\hat{W}_{i,pen}^K(t) - \breve{W}_{i,pen}^K(t)| + |\breve{W}_{i,pen}^K(t) - \breve{W}_{i,pen}(t)|$$

Lemma (4.1.3) shows that $\breve{W}_{i,pen}^K(t) \xrightarrow{p} \breve{W}_{i,pen}(t)$, as $K \to \infty$. For fixed K, observing that $\hat{\xi}_{ik,pen} \xrightarrow{p} \tilde{\xi}_{ik,pen}$ as $N \to \infty$, $\sup_{t\in[0,1]} |\hat{W}_{i,pen}^K(t) - \breve{W}_{i,pen}^K(t)| \xrightarrow{p} 0$. This implies that given $\epsilon, \delta > 0$, there exist K_0 such that for $K \geqslant K_0$, $P(|\breve{W}_{i,pen}^K(t) - \breve{W}_{i,pen}(t)| > \epsilon/2) \leqslant \delta/2$. For each K, there exists $N_0(K)$ such that for $N \geqslant N_0(K)$, $P(|\hat{W}_{i,pen}^K(t) - \breve{W}_{i,pen}^K(t)| > \epsilon/2) \leqslant \delta/2$. Thus for $K \geqslant K_0$ and $N \geqslant N_0(K)$, $P(|\hat{W}_{i,pen}^K(t) - \breve{W}_{i,pen}(t)| \geqslant \epsilon) \leqslant \delta$, which leads to (2.19). Note that $\hat{X}_{i,pen}^K(t) = \hat{\beta}_{i0} + \int_{T_{i1}}^t \exp \hat{W}_{i,pen}^K(s)ds$ for each fixed $t \in \mathcal{T}$. Thus (2.20) is followed by $\hat{\beta}_i \xrightarrow{p} \beta_{i0}$ (theorem (2.3.1)) and (2.19). □

4.2 Notations of Chapter 3

For ease of presentation of the proof, we use the following notations. Let $U = \int_0^1 \boldsymbol{\alpha}_0(t)^T\boldsymbol{X}(t)dt$, $U_i = \int_0^1 \boldsymbol{\alpha}_0(t)^T\boldsymbol{X}_i(t)dt$. Let $\boldsymbol{B}(t) = (B_1(t), ..., B_K(t))^T$ be a basis function, and $\tilde{\boldsymbol{\alpha}}_0(t) = (\tilde{\boldsymbol{\alpha}}_{01}(t), ..., \tilde{\boldsymbol{\alpha}}_{0p}(t))^T =$
$(\boldsymbol{B}(t)^T\boldsymbol{\beta}_{01}, ..., \boldsymbol{B}(t)^T\boldsymbol{\beta}_{0p})^T = (I_p \otimes \boldsymbol{B}(t)^T)\boldsymbol{\beta}_0$ be the best approximation of $\boldsymbol{\alpha}_0(t)$ in the space spanned $\boldsymbol{B}(t)$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, ..., \boldsymbol{\beta}_{0p}^T)^T$ is the associate coefficient.

Let $\hat{\boldsymbol{\alpha}}_0(t) = (\hat{\boldsymbol{\alpha}}_1(t), ..., \hat{\boldsymbol{\alpha}}_p(t))^T = (\boldsymbol{B}(t)^T\hat{\boldsymbol{\beta}}_1, ..., \boldsymbol{B}(t)^T\hat{\boldsymbol{\beta}}_p)^T = (I_p \otimes \boldsymbol{B}(t)^T)\hat{\boldsymbol{\beta}}$ be the associated estimator. Let $\hat{U}_i = \int_0^1 \hat{\boldsymbol{\alpha}}_0(t)^T \boldsymbol{X}_i(t)dt$. Denote

$$\theta_0 = \begin{pmatrix} g_0(u) \\ hg_0'(u) \end{pmatrix} \quad \hat{\theta} = \begin{pmatrix} \hat{a}(u;\hat{\boldsymbol{\beta}}) \\ h\hat{b}(u;\hat{\boldsymbol{\beta}}) \end{pmatrix}$$

and

$$\hat{U}_i^* = \begin{pmatrix} 1 \\ (\hat{U}_i - u)/h \end{pmatrix} \quad U_i^* = \begin{pmatrix} 1 \\ (U_i - u)/h \end{pmatrix}$$

Denote $\mu_j = \int t^j K(t)dt$ and $\nu_j = \int t^j K^2(t)dt$ the jth moment of the kernel function and square kernel function.

4.3 Proof of Theorems in Chapter 3

Lemma 4.3.1. *Assume that conditions (C1)-(C7), (C9)-(C11) hold, if $h \to 0$, $nh \to \infty$ as $n \to \infty$, we have*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^j \phi''_{h_2}(\epsilon_i) &= f(u)\mu_j F(u, h_2) + o_p(1) \\ \frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^j r_i\phi''_{h_2}(\epsilon_i) &= \frac{g_0''(u)}{2}h^2 f(u)\mu_{j+2}F(u, h_2) + o_p(h^2) \end{aligned}$$

where ϵ_i is iid random variable with zero mean, and $r_i = g_0(U_i) - g_0(u) - g_0'(u)(U_i - u)$ is the remaining term of the first Taylor expansion of $g_0(U_i)$ at u.

Lemma 4.3.2. *Assume that conditions (C1)-(C7), (C9)-(C11) hold, if $h \to 0$,*

$nh^3 \to \infty$ as $n \to \infty$, for a fixed u and any $\hat{\boldsymbol{\beta}}$ subject to $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0} = O_p(\delta)$, we have

$$\begin{aligned}
\hat{a}(u;\hat{\boldsymbol{\beta}}) - g_0(u) &= \frac{1}{nf(u)F(u,h_2)}\sum_{i=1}^{n} K_h(U_i-u)\phi'_{h_2}(\epsilon_i) - \frac{g'_0(u)E(\boldsymbol{C}^T\phi''_{h_2}(\epsilon)|U=u)}{F(u,h_2)}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}) \\
&\quad +O_p(h^2)+o_p(\delta) \\
\hat{b}(u;\hat{\boldsymbol{\beta}}) - g'_0(u) &= O_p(\sqrt{1/(nh^3)}+h^2)+o_p(\delta/h)
\end{aligned}$$

Proof of Lemma 4.3.1. Let $T_{1n}^j = \frac{1}{n}\sum_{i=1}^n K_h(U_i-u)\left(\frac{U_i-u}{h}\right)^j$ and $t=(u_i-u)/h$, then

$$\begin{aligned}
\mathrm{E}(T_{1n}^j) &= \mathrm{E}\left[K_h(U_i-u)\left(\frac{U_i-u}{h}\right)^j \phi''_{h_2}(\epsilon_i)\right] \\
&= \mathrm{E}\left[K_h(U_i-u)\left(\frac{U_i-u}{h}\right)^j \mathrm{E}(\phi''_{h_2}(\epsilon_i)|U_i)\right] \\
&= \int \frac{1}{h}K\left(\frac{u_i-u}{h}\right)\left(\frac{u_i-u}{h}\right)^j f(u_i)F(u_i,h_2)du_i \\
&= \int t^j K(t)f(u+th)F(u+th,h_2)du \\
&= f(u)\mu_j F(u,h_2)+O(h)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(T_{1n}^j) &= \frac{1}{n}\mathrm{Var}\left(K_h(U_i-u)\left(\frac{U_i-u}{h}\right)^j \phi''_{h_2}(\epsilon_i)\right) \\
&\leqslant \frac{1}{n}\mathrm{E}\left[K_h^2(U_i-u)\left(\frac{U_i-u}{h}\right)^{2j} \mathrm{E}(\phi''_{h_2}(\epsilon_i)^2|U_i)\right] \\
&= \frac{1}{nh}\int \frac{1}{h}K^2\left(\frac{u_i-u}{h}\right)\left(\frac{u_i-u}{h}\right)^{2j} f(u_i)\mathrm{E}(\phi''_{h_2}(\epsilon_i)^2|U_i=u_i)du_i \\
&= O\left(\frac{1}{nh}\right)
\end{aligned}$$

Based on the result $T_{1n}^j = \mathrm{E}(T_{1n}^j) + \sqrt{\mathrm{Var}(T_{1n}^j)}$, we have

$$T_{1n}^j = f(u)\mu_j F(u, h_2) + o_p(1)$$

Let $T_{2n}^j = \frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i-u}{h}\right)^j r_i \phi''_{h_2}(\epsilon_i)$, since $r_i = \sum_{k=2}^{\infty} \frac{g_0^k(u)}{k!}(U_i - u)^k$, then

$$\begin{aligned} T_{2n}^j &= \frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^j r_i \\ &= \frac{g_0''(u)}{2} h^2 T_{1n}^{j+2}(1 + O_p(h)) \\ &= \frac{g_0''(u)}{2} h^2 f(u)\mu_{j+2} F(u, h_2) + o_p(1) \end{aligned}$$

□

Proof of Lemma 4.3.2. $\hat{\boldsymbol{\theta}} = (\hat{a}(u; \hat{\boldsymbol{\beta}}), \hat{b}(u; \hat{\boldsymbol{\beta}}))^T$ is obtained through minimizing (3.6), which implies

$$\frac{1}{n}\sum_{i=1}^n \hat{U}_i^* K_h(\hat{U}_i - u)\phi'_{h_2}(Y_i - \hat{\boldsymbol{\theta}}^T \hat{U}_i^*) = \mathbf{0} \tag{4.1}$$

By Taylor expansion and simple calculation, (4.1) equals

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n [U_i^* + (\hat{U}_i^* - U_i^*)] \times [K_h(U_i - u) + L(\hat{U}_i - U_i)] \\ & \times \phi'_{h_2}((Y_i - \boldsymbol{\theta}_{\mathbf{0}}^T U_i^*) + \boldsymbol{\theta}_{\mathbf{0}}^T(U_i^* - \hat{U}_i^*) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathbf{0}})^T(U_i^* + (\hat{U}_i^* - U_i^*))) \\ = & \frac{1}{n}\sum_{i=1}^n [U_i^* + (\hat{U}_i^* - U_i^*)] \times [K_h(U_i - u) + L(\hat{U}_i - U_i)] \\ & \times \{\phi'_{h_2}(Y_i - \boldsymbol{\theta}_{\mathbf{0}}^T U_i^*) + \phi''_{h_2}(Y_i - \boldsymbol{\theta}_{\mathbf{0}}^T U_i^*)[\boldsymbol{\theta}_{\mathbf{0}}^T(U_i^* - \hat{U}_i^*) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathbf{0}})^T(U_i^* + (\hat{U}_i^* - U_i^*))]\} \end{aligned}$$

where L is a Lipschitz constant. By eliminating the higher order term it yields that

$$
\begin{aligned}
= \quad & \frac{1}{n}\sum_{i=1}^{n} U_i^* K_h(U_i-u)\phi'_{h_2}(Y_i-\boldsymbol{\theta}_{\mathbf{0}}^T U_i^*) - \frac{1}{n}\sum_{i=1}^{n} U_i^* K_h(U_i-u)\boldsymbol{\theta}_{\mathbf{0}}^T\phi''_{h_2}(Y_i-\boldsymbol{\theta}_{\mathbf{0}}^T U_i^*)(\hat{U}_i^*-U_i^*) \\
& -\frac{1}{n}\sum_{i=1}^{n} U_i^* K_h(U_i-u)\phi''_{h_2}(Y_i-\boldsymbol{\theta}_{\mathbf{0}}^T U_i^*)(U_i^*+(\hat{U}_i^*-U_i^*))^T(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathbf{0}}) \\
& -\frac{1}{n}\sum_{i=1}^{n}(\hat{U}_i^*-U_i^*)K_h(U_i-u)\phi'_{h_2}(Y_i-\boldsymbol{\theta}_{\mathbf{0}}^T U_i^*) + \begin{pmatrix} 0 \\ o_p(\delta^2/h) \end{pmatrix} \\
& +\frac{1}{n}\sum_{i=1}^{n} U_i^* L(\hat{U}_i-U_i)\phi'_{h_2}(Y_i-\boldsymbol{\theta}_{\mathbf{0}}^T U_i^*) + o_p(\delta) \\
= \quad & I_1 - I_2 - I_3 + I_4 + I_5
\end{aligned}
$$

By applying Lemma 4.3.1, it is easy to show that

$$
\begin{aligned}
I_1 &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\phi'_{h_2}(\epsilon_i+r_i) \\ \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\left(\frac{U_i-u}{h}\right)\phi'_{h_2}(\epsilon_i+r_i) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\phi'_{h_2}(\epsilon_i)+O_p(h^2) \\ \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\left(\frac{U_i-u}{h}\right)\phi'_{h_2}(\epsilon_i)+O_p(h^3) \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
I_2 &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)g'_0(u)(\hat{U}_i-U_i)\phi''_{h_2}(\epsilon_i+r_i) \\ \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\left(\frac{U_i-u}{h}\right)g'_0(u)(\hat{U}_i-U_i)\phi''_{h_2}(\epsilon_i+r_i) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)g'_0(u)(\boldsymbol{C}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))\phi''_{h_2}(\epsilon_i+r_i) \\ \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\left(\frac{U_i-u}{h}\right)g'_0(u)(\boldsymbol{C}_i^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))\phi''_{h_2}(\epsilon_i+r_i) \end{pmatrix} \\
&= \begin{pmatrix} F(u,h_2)f(u)g'_0(u)\{\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}+O_p(h^2) \\ F(u,h_2)hg'_0(u)\{f(u)\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}'+O_p(h^3) \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
I_3 &= \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)U_i^*(U_i^*)^T(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})\phi''_{h_2}(\epsilon_i+r_i) \\
&\quad +\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)U_i^*(\hat{U}_i^*-U_i^*)^T(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})\phi''_{h_2}(\epsilon_i+r_i) \\
&= S_1+S_2
\end{aligned}
$$

where

$$
\begin{aligned}
S_1 &= F(u,h_2)\begin{pmatrix} f(u) & hf(u)\mu_2 \\ hf(u)\mu_2 & f(u)\mu_2 \end{pmatrix}(1+O_p\left(\sqrt{\frac{1}{nh}}+h^2\right))\begin{pmatrix} \hat{a}-g_0(u) \\ h(\hat{b}-g_0'(u)) \end{pmatrix} \\
&= F(u,h_2)f(u)\begin{pmatrix} \hat{a}-g_0(u) \\ h(\hat{a}-g_0(u))+h(\hat{b}-g_0'(u)) \end{pmatrix}(1+O_p\left(\sqrt{\frac{1}{nh}}+h^2\right)) \\
S_2 &= \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\begin{pmatrix} 0 & (\hat{U}_i-U_i)/h \\ 0 & (U_i-u)(\hat{U}_i-U_i)/h^2 \end{pmatrix}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta_0})\phi''_{h_2}(\epsilon_i+r_i) \\
&= F(u,h_2)\begin{pmatrix} f(u)\{\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}(\hat{b}-g_0'(u)) \\ hg_0'(u)\{f(u)\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}'(\hat{b}-g_0'(u)) \end{pmatrix} \\
&\quad \times(1+O_p\left(\sqrt{\frac{1}{nh}}+h^2\right))
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
I_3 &= F(u,h_2)f(u)\begin{pmatrix} \hat{a}-g_0(u)+\{\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}(\hat{b}-g_0'(u)) \\ h(\hat{a}-g_0(u))+h(\hat{b}-g_0'(u))(1+O_p(\delta)) \end{pmatrix} \\
&\quad \times(1+O_p\left(\sqrt{\frac{1}{nh}}+h^2\right)) \\
I_4 &= \begin{pmatrix} 0 \\ \frac{1}{n}\sum_{i=1}^n K_h(U_i-u)\left(\frac{U_i-u}{h}\right)\phi_{h_2}'(\epsilon_i+r_i)+o_p(\delta^2/h) \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ O_p(\sqrt{\frac{1}{nh}})+o_p(\delta^2/h) \end{pmatrix} \\
I_5 &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n L(\hat{U}_i-U_i)\phi_{h_2}'(\epsilon_i+r_i)+o_p(\delta) \\ \frac{1}{n}\sum_{i=1}^n L(\hat{U}_i-U_i)(\frac{U_i-u}{h})\phi_{h_2}'(\epsilon_i+r_i)+o_p(\delta) \end{pmatrix} \\
&= o_p(\delta)
\end{aligned}
$$

By combining I_1-I_5, it is easy to show that

$$
\begin{aligned}
&[\hat{a}-g_0(u)+\{\mathrm{E}(\boldsymbol{C}^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}(\hat{b}-g_0'(u))] \\
&= \frac{1}{nF(u,h_2)f(u)}\sum_{i=1}^n K_h(U_i-u)\phi_{h_2}'(\epsilon_i)-g_0'(u)\mathrm{E}(C^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}) \\
&\quad +O_p(h^2)+o_p(\delta)
\end{aligned}
$$

and

$$
\begin{aligned}
[\hat{a}-g_0(u)+(\hat{b}-g_0'(u))(1+O_p(\delta))] &= \frac{1}{nhf(u)}\sum_{i=1}^n K_h(U_i-u)(\frac{U_i-u}{h})\epsilon_i- \\
&\frac{g_0'(u)}{F(u,h_2)f(u)}[F(u,h_2)f(u)\mathrm{E}(C^T|U=u)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0})]'+o_p(\delta^2/h^2+\delta/h)
\end{aligned}
$$

The lemma is completed by solving the above two equations and following the

assumptions $h \to 0$, $nh^3 \to \infty$. □

REMARK 4.3.1. *In the following theorems, we will primarily use the following results which can be easily obtained by applying Lemma 4.3.2,*

$$\begin{aligned}
\hat{b}(u;\hat{\boldsymbol{\beta}}) - g_0'(u) &= o_p(1) \\
\hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}) - g_0(U_i) &= \frac{1}{nf(U_i)}\sum_{j=1}^{n}\frac{K_h(U_j-U_i)\phi_{h_2}'(\epsilon_j)}{F(u,h_2)} + g_0'(U_i)[\boldsymbol{C}_i - E(\boldsymbol{C}_i\frac{\phi_{h_2}''(\epsilon_i)}{F(u,h_2)}|U_i)]^T \\
&\quad (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + O_p(h^2) + o_p(\delta) \\
\hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}}) - \hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0) &= -\frac{g_0'(U_i)E(\phi_{h_2}''(\epsilon_i)\boldsymbol{C}_i^T|U_i)](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{F(U_i,h_2)} + O_p(h^2) + o_p(\delta)
\end{aligned}$$

Proof of Theorem 3.4.1. Let $Q(\hat{g}_{\boldsymbol{\beta}},\boldsymbol{\beta}) = \sum_{i=1}^{n}\phi_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\hat{\boldsymbol{\beta}})) - np_{\lambda_n}(||\boldsymbol{\beta_j}||_H)$ and let $\delta = n^{-1/2} + a_n$ and $\boldsymbol{v}$ be a vector. Define $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \delta\boldsymbol{v}$, where $\boldsymbol{\beta}_0$ is the best approximation of $\boldsymbol{\alpha}_0(\cdot)$ in the B-spline space. We first show that, for any given $\epsilon > 0$, there exists a large $C > 0$ such that

$$P\{\sup_{||\boldsymbol{v}||=C} Q(\hat{g}_{\boldsymbol{\beta}},\boldsymbol{\beta}) - Q(\hat{g}_{\boldsymbol{\beta}_0},\boldsymbol{\beta}_0) < 0\} \geqslant 1-\epsilon \tag{4.2}$$

Let $L(\hat{g}_{\boldsymbol{\beta}},\boldsymbol{\beta}) = \frac{1}{n}(Q(\hat{g}_{\boldsymbol{\beta}},\boldsymbol{\beta}) - Q(\hat{g}_{\boldsymbol{\beta}_0},\boldsymbol{\beta}_0))$, then

$$\begin{aligned}
L(\hat{g}_{\boldsymbol{\beta}},\boldsymbol{\beta}) &\geqslant \frac{1}{n}\sum_{i=1}^{n}[\phi_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta})) - \phi_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0))] - \sum_{j=1}^{s}[p_{\lambda_n}(||\boldsymbol{\beta_j}||_H - p_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)] \\
&= \frac{1}{n}\sum_{i=1}^{n}[\phi_{h_2}'(y_i - \hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0))(\hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0) - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta}))] \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\phi_{h_2}''(y_i - \hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0))(\hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0) - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta}))^2 \\
&\quad - \sum_{j=1}^{s}[p_{\lambda_n}(||\boldsymbol{\beta_j}||_H - p_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)] \\
&= I_1 + I_2 - I_3
\end{aligned}$$

By applying Lemma 4.3.2, it is easy to show that

$$\begin{aligned}
I_1 &= \frac{1}{n}\sum_{i=1}^{n}[\phi'_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0))(\hat{g}_{\boldsymbol{\beta}_0}(\boldsymbol{C}_i^T\boldsymbol{\beta}_0) - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta}))] \\
&= \frac{1}{n}\sum_{i=1}^{n}\phi'_{h_2}(\epsilon_i - \frac{1}{nf(U_i)}\sum_{j=1}^{n}K_h(U_j - U_i)\frac{\phi'_{h_2}(\epsilon_j)}{F(U_i,h_2)} + O_p(h^2) + o_p(\delta)) \\
&\quad \times[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T g'_0(U_i)\mathrm{E}(\boldsymbol{C_i}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i) + O_p(h^2) + o_p(\delta)] \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T g'_0(U_i)\mathrm{E}(\boldsymbol{C_i}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\phi'_{h_2}(\epsilon_i) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\phi''_{h_2}(\epsilon_i)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T(g'_0(U_i))\mathrm{E}(\boldsymbol{C_i}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\frac{1}{nf(U_i)}\sum_{j=1}^{n}K_h(U_j - U_i)\frac{\phi'_{h_2}(\epsilon_j)}{F(U_i,h_2)} \\
&\quad +O_p(h^2) + o_p(\delta) \\
&= S_1 - S_2 + O_p(h^2) + o_p(\delta)
\end{aligned}$$

We first investigate S_2, by simple calculation and applying Lemma 4.3.1,

$$\begin{aligned}
S_2 &= \frac{1}{n}\sum_{j=1}^{n}\phi'_{h_2}(\epsilon_j)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}\frac{g'_0(U_i)\mathrm{E}(\boldsymbol{C_i}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)}{f(U_i)F(U_j,h_2)}K_h(U_i - U_j)\phi''_{h_2}(\epsilon_i)\right) \\
&= \frac{1}{n}\sum_{j=1}^{n}\phi'_{h_2}(\epsilon_j)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T g'_0(U_j)\mathrm{E}(\boldsymbol{C_j}\frac{\phi''_{h_2}(\epsilon_j)}{F(U_j,h_2)}|U_j)(1 + o_p(h))
\end{aligned}$$

thus it is easy to derive that

$$S_1 - S_2 = o_p(h)$$

Therefore, $I_1 = O_p(\delta^2||\boldsymbol{v}||)$. By similar arguments,

$$\begin{aligned} I_2 &= \frac{1}{n}\sum_{i=1}^{n}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T(g_0'(U_i))\mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)+O_p(h^2)+o_p(\delta)]^2\phi''_{h_2}(\epsilon_i) \\ &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(g_0'(U_i))^2\mathrm{E}(\boldsymbol{C}_i^T\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\phi''_{h_2}(\epsilon_i)\right)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\ &= \delta^2\boldsymbol{v}^T\left(\frac{1}{n}\sum_{i=1}^{n}(g_0'(U_i))^2\mathrm{E}(\boldsymbol{C}_i^T\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\phi''_{h_2}(\epsilon_i)\right)\boldsymbol{v} \\ &\leqslant -\delta^2||\boldsymbol{v}||^2M_0 \end{aligned}$$

since $F(u,h_2)=\mathrm{E}(\phi''_{h_2}(\epsilon_i)|U_i=u)<0$ for any u, there must exist a $M_0>0$, depending on $\boldsymbol{v}$, because $\frac{1}{n}\sum_{i=1}^{n}(g_0'(U_i))^2\mathrm{E}(\boldsymbol{C}_i^T\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)\mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)$ is a nonnegative definite matrix, by choosing an appropriate $\boldsymbol{v}$, we always can find a positive M_0 to make sure the above inequality hold. Hence, I_2 dominates I_1 if $||\boldsymbol{v}||$ is large enough.

$$\begin{aligned} I_3 &= \sum_{j=1}^{s}\left[p'_{\lambda_n}(||\boldsymbol{\beta_{j0}}||_H)(||\boldsymbol{\beta_j}||_H-||\boldsymbol{\beta_{j0}}||_H)+(p''_{\lambda_n}+o_p(1))(||\boldsymbol{\beta_{j0}}||_H)(||\boldsymbol{\beta_j}||_H-||\boldsymbol{\beta_{j0}}||_H)^2\right] \\ &= O_p(a_n\delta||\boldsymbol{v}||+\delta^2||\boldsymbol{v}||^2b_n) \end{aligned}$$

Note that $b_n\to 0$, thus by choosing a large $||\boldsymbol{v}||$, I_3 is domiated by I_2. Therefore, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=O_p(\delta)$. Hence,

$$\begin{aligned} ||\hat{\alpha}_j(\cdot)-\alpha_{j0}(\cdot)||^2 &= \int_0^1(\hat{\alpha}_j(t)-\alpha_{j0}(t))^2dt \\ &= \int_0^1(\boldsymbol{B}(t)^T(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j))^2dt \\ &\leqslant 2(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)^T\boldsymbol{H}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j) \end{aligned}$$

Since $||\boldsymbol{H}||=O(1)$, $a_n=O_p(n^{-1/2})$, we have $||\hat{\alpha}_j(\cdot)-\alpha_{j0}(\cdot)||=O_p(n^{-1/2})$

□

Proof of Theorem 3.4.2. By the property of SCAD function, we can see that $a_n = 0$ as $\lambda_n \to 0$. Note that $\hat{\alpha}_j(\cdot) = \boldsymbol{B}(\cdot)^T\hat{\boldsymbol{\beta}}_j$, and by Theorem 3.4.1, it is sufficient to show that, when $n \to \infty$, for any $\boldsymbol{\beta}$ that satisfies $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| = O(n^{-1/2})$ and given small $\eta = C_0 n^{-1/2}$, with probability tending to 1, we have

$$\frac{\partial Q(\hat{g}_{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\partial \beta_{jk}} > 0 \quad \text{for } 0 < \beta_{jk} < \eta \tag{4.3}$$

$$\frac{\partial Q(\hat{g}_{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\partial \beta_{jk}} < 0 \quad \text{for } -\eta < \beta_{jk} < 0 \tag{4.4}$$

Consequently, (4.3) and (4.4) imply that the minimizer of $Q(\hat{g}_{\boldsymbol{\beta}}, \boldsymbol{\beta})$ attains at $\beta_{jk} = 0$, for $j = s+1, ..., p$ and $k = 1, ...K$.

$$\begin{aligned}\frac{\partial Q(\hat{g}_{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\partial \beta_{jk}} &= \sum_{i=1}^{n} \phi'_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta})) \cdot (-\hat{g}'_{\boldsymbol{\beta}}(\boldsymbol{C}_i^T\boldsymbol{\beta}))C_{ijk} + np'_{\lambda_n}(||\boldsymbol{\beta_j}||_H)\frac{\boldsymbol{e}_k^T\boldsymbol{H}\boldsymbol{\beta}_j}{||\boldsymbol{\beta_j}||_H} \\ &= I_1 + I_2\end{aligned}$$

where $\boldsymbol{e}_k$ is a $K \times 1$ vector with kth element 1 and 0 otherwise. By applying Lemma 4.3.1, it is straightforward to show that

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_{i=1}^{n} \phi'_{h_2}(\epsilon_i - \frac{1}{nf(U_i)}\sum_{l=1}^{n} K_h(U_l - U_i)\frac{\phi'_{h_2}(\epsilon_l)}{F(U_i, h_2)} \\ &\quad + O_p(h^2) + o_p(\delta)) \cdot (-g'_0(U_i) + o_p(1)) \cdot C_{ijk} \\ &= -\sum_{i=1}^{n} \phi'_{h_2}(\epsilon_i)(g'_0(U_i)C_{ijk} + \sum_{i=1}^{n} \phi''_{h_2}(\epsilon_i)\frac{g'_0(U_i)}{nf(U_i)}\sum_{l=1}^{n} K_h(U_l - U_i)\frac{\phi'_{h_2}(\epsilon_l)}{F(U_i, h_2)}C_{ijk} \\ &\quad + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\sum_{i=1}^{n} \phi''_{h_2}(\epsilon_i)(g'_0(U_i))^2[\boldsymbol{C}_i - \mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i, h_2)}|U_i)]C_{ijk})(1 + o_p(1)) \\ &= (-S_1 + S_2 + S_3)(1 + o_p(1)) \end{aligned}$$

We first investigate S_2, it is easy to see that

$$\begin{aligned} S_2 &= \sum_{l=1}^{n} \phi'_{h_2}(\epsilon_l) \sum_{i=1}^{n} \left(\frac{g'_0(U_i) C_{ijk}}{nf(U_i)F(U_i, h_2)} K_h(U_i - U_l) \phi''_{h_2}(\epsilon_i) \right) \\ &= \sum_{l=1}^{n} \phi'_{h_2}(\epsilon_l) g'_0(U_l) \mathrm{E}(C_{ljk} \frac{\phi''_{h_2}(\epsilon_i)}{F(U_i, h_2)} | U_l)(1 + O_p(\sqrt{1/(nh)} + h^2)) \end{aligned}$$

Hence, $S_2 - S_1 = O_p(\sqrt{n})$, and with

$$S_3 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \sum_{i=1}^{n} \phi''_{h_2}(\epsilon_i)(g'_0(U_i))^2 [\boldsymbol{C}_i - \mathrm{E}(\boldsymbol{C}_i \frac{\phi''_{h_2}(\epsilon_i)}{F(U_i, h_2)} | U_i)] C_{ijk} = O_p(n\delta ||\boldsymbol{v}||)$$

We can obtain that $I_1 = O_p(n\delta||\boldsymbol{v}||)$. Let h_{kt} be the (k,t)th element of $\boldsymbol{H}$, we have

$$I_2 = np'_{\lambda_n}(||\boldsymbol{\beta_j}||_H) \frac{h_{kk}\beta_{jk} + \sum_{t \neq k} h_{kt}\beta_{jt}}{||\boldsymbol{\beta_j}||_H}$$

By the property of B spline and the fact $h_{kk} > 0$, we can conclude that the sign of $(h_{kk}\beta_{jk} + \sum_{t \neq k} h_{kt}\beta_{jt})$ is determined by that of β_{jk}. Furthermore, we can see that

$$I_1 + I_2 = n\lambda_n \left(O_p \left(\frac{||\boldsymbol{v}||}{n^{1/2}\lambda_n} \right) + \frac{p'_{\lambda_n}(||\boldsymbol{\beta_j}||_H)}{\lambda_n} \cdot \frac{h_{kk}\beta_{jk} + \sum_{t \neq k} h_{kt}\beta_{jt}}{||\boldsymbol{\beta_j}||_H} \right)$$

Under the condition (C8) and assumptions $\lambda_n \to 0$ and $n^{1/2}\lambda_n \to \infty$ as $n \to \infty$, the sign of derivative is completely determined by that of β_{jk}. This completes the proof. □

Proof of Theorem 3.4.3. Let $\boldsymbol{C}_i = (\boldsymbol{C}_{i1}, \boldsymbol{C}_{i0})^T$. With γ being the Lagrange multiplier, we know that $(\hat{\boldsymbol{\beta}}_1^T, 0)^T$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} -\phi'_{h_2}(y_i - \hat{g}_{\boldsymbol{\beta}}(\boldsymbol{C}_{i1}^T \hat{\boldsymbol{\beta}}_1))(-\hat{g}'_{\boldsymbol{\beta}}(\boldsymbol{C}_{i1}^T \hat{\boldsymbol{\beta}}_1)) \boldsymbol{C}_{i1} + \boldsymbol{\kappa} + \gamma(\boldsymbol{I}_s \otimes \boldsymbol{H}) \hat{\boldsymbol{\beta}}_1 = \boldsymbol{0} \qquad (4.5)$$

Where $\boldsymbol{\kappa}$ is a $K_n \times s$ vector with jth block subvector $\boldsymbol{\kappa}_j = p'_{\lambda_n}(||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||_H) \cdot \frac{\boldsymbol{H}\hat{\boldsymbol{\beta}}_j}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||_H}$. By applying Lemma 4.3.2, we have

$$\frac{1}{n}\sum_{i=1}^{n} -\phi'_{h_2}(\epsilon_i - \frac{1}{nf(U_i)}\sum_{j=1}^{n}\frac{K_h(U_j-U_i)\phi'_{h_2}(\epsilon_j)}{F(U_i,h_2)} + g'_0(U_i)[\boldsymbol{C}_i - \mathrm{E}(\boldsymbol{C}_i\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i)]^T$$
$$\times(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O_p(h^2) + o_p(\delta)) \times (-g'_0(U_i))\boldsymbol{C}_{i1}(1+o_p(1)) + \boldsymbol{\kappa} + \gamma(\boldsymbol{I}_s \otimes \boldsymbol{H})\hat{\boldsymbol{\beta}}_1 = \boldsymbol{0}$$

By Taylor expansion, we will find $\boldsymbol{\kappa}_j = \boldsymbol{b}_j + (\boldsymbol{\Sigma}_j + o_p(1))(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{j0})$, with $\boldsymbol{b}_j = p'_{\lambda_n}(||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H)\cdot\frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}}{||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H}$ and $\boldsymbol{\Sigma}_j = p''_{\lambda_n}(||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H)\frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}\boldsymbol{\beta}_{j0}^T\boldsymbol{H}}{||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H^2} + p'_{\lambda_n}(||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H)(\frac{\boldsymbol{H}}{||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H} - \frac{\boldsymbol{H}\boldsymbol{\beta}_{j0}\boldsymbol{\beta}_{j0}^T\boldsymbol{H}}{||\boldsymbol{\beta}_{\boldsymbol{j0}}||_H^3})$. Let $\boldsymbol{b} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_s)^T$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_s)$. By omitting the high order term, we have

$$\begin{aligned}
\boldsymbol{0} &= \frac{1}{n}\sum_{i=1}^{n} g'_0(U_i)\boldsymbol{C}_{i1}\phi'_{h_2}(\epsilon_i) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{g'_0(U_i)\boldsymbol{C}_{i1}}{nf(U_i)F(U_i,h_2)}K_h(U_j-U_i)\phi'_{h_2}(\epsilon_j) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\phi''_{h_2}(\epsilon_i)(g'_0(U_i))^2\boldsymbol{C}_{i1}(\boldsymbol{C}_{i1} - \mathrm{E}(\boldsymbol{C}_{i1}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i))^T(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{\boldsymbol{10}}) \\
&\quad +\boldsymbol{b} + \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{\boldsymbol{10}}) + \gamma(\boldsymbol{I}_s \otimes \boldsymbol{H})\hat{\boldsymbol{\beta}}_1 \\
&= -I_1 + I_2 + I_3 + \boldsymbol{b} + \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{\boldsymbol{10}}) + \gamma(\boldsymbol{I}_s \otimes \boldsymbol{H})\hat{\boldsymbol{\beta}}_1
\end{aligned}$$

By interchanging the summation in I_2, we have

$$I_1 - I_2 = -\frac{1}{n}\sum_{i=1}^{n} g'_0(U_i)(\boldsymbol{C}_{i1} - \mathrm{E}(\boldsymbol{C}_{i1}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i))\phi'_{h_2}(\epsilon_i)$$

By mulitplying $P_{\beta_0} = I - \boldsymbol{\beta}_{\boldsymbol{10}}\boldsymbol{\beta}_{\boldsymbol{10}}^T = I - \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1^T + o_p(1)$ on both sides, we have

$$P_{\beta_0}[\boldsymbol{W} + \boldsymbol{\Sigma}](\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{\boldsymbol{10}}) + \boldsymbol{b} = -P_{\beta_0}\frac{1}{n}\sum_{i=1}^{n} g'_0(U_i)(\boldsymbol{C}_{i1} - \mathrm{E}(\boldsymbol{C}_{i1}\frac{\phi''_{h_2}(\epsilon_i)}{F(U_i,h_2)}|U_i))\phi'_{h_2}(\epsilon_i)$$

where $W = \mathrm{E}[\phi''_{h_2}(\epsilon)(g'_0(U))^2\boldsymbol{C}_1\boldsymbol{C}_1^T] - \mathrm{E}[(g'_0(U))^2\phi''_{h_2}(\epsilon)\boldsymbol{C}_1\mathrm{E}(\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2}\boldsymbol{C}_1^T|U)]$. By fol-

lowing the Central Limit Theorem and Slutsky's Theorem, we can have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0 + (\boldsymbol{W}+\boldsymbol{\Sigma})^{-}\boldsymbol{b}) \to N(\mathbf{0}, (\boldsymbol{W}+\boldsymbol{\Sigma})^{-}\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^{-})$$

where $\boldsymbol{Q} = \mathrm{E}\left(\phi'_{h_2}(\epsilon)^2 g'_0(U)^2(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2)}|U)(\boldsymbol{C}_1 - \mathrm{E}(\boldsymbol{C}_1\frac{\phi''_{h_2}(\epsilon)}{F(U,h_2)}|U)^T\right)$. Then we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_1) = \frac{1}{n}(\boldsymbol{W}+\boldsymbol{\Sigma})^{-}\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^{-}$$

and by the equation that $\hat{\boldsymbol{\alpha}}_1(t) = (\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T\hat{\boldsymbol{\beta}}_1$, we are able to have

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}_1(t))^{-1/2}(\hat{\boldsymbol{\alpha}}_1(t) - \tilde{\boldsymbol{\alpha}}_1(t) + (\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T(\boldsymbol{W}+\boldsymbol{\Sigma})^{-}\boldsymbol{b})) \to N(\mathbf{0}, \boldsymbol{I}_s)$$

where $\mathrm{Var}(\hat{\boldsymbol{\alpha}}_1(t)) = (\boldsymbol{I}_s \otimes \boldsymbol{B}(t))^T\mathrm{Var}(\hat{\boldsymbol{\beta}}_1)(\boldsymbol{I}_s \otimes \boldsymbol{B}(t))$, this completes the proof.

□

Proof of Theorem 3.4.4.

$$\sqrt{nh}\,(\hat{g}(u;\alpha) - g_0(u)) = \sqrt{nh}\,(\hat{g}(u;\alpha) - \hat{g}(u;\alpha_0)) + \sqrt{nh}\,(\hat{g}(u;\alpha_0) - g_0(u))$$

where $\hat{g}(u;\alpha_0)$ is a local linear estimator of $g_0(\cdot)$ if α_0 is known. According to Theorems 2.1, 2,2 in Yao et al. [2005], we obtain

$$\sqrt{nh}\left(\hat{g}(u;\alpha_0) - g_0(u) - \frac{1}{2}h^2\mu_2 g''_0(u)\right) \xrightarrow{d} N(0, \frac{\nu_0 G(u,h_2)}{f(u)F(u,h_2)^2}).$$

We also can show that

$$\sqrt{nh}\,(\hat{g}(u;\alpha) - \hat{g}(u;\alpha_0)) = o_p(1),$$

where α is assumed to be known. Thus, we can have

$$\sqrt{nh}\left(\hat{g}(u;\boldsymbol{\alpha}) - g_0(u) - \frac{1}{2}h^2\mu_2 g_0''(u)\right) \xrightarrow{d} N\left(0, \frac{\nu_0 G(u,h_2)}{f(u)F(u,h_2)^2}\right)$$

this completes the proof. □

Proof of Theorem 3.5.1. (a) Since $\phi(\cdot)$ is a normal density, we have

$$\begin{aligned} F(u,h_2) &= \int \phi''(t/h_2)g_{\epsilon|u}(t)dt = -\frac{1}{\sqrt{2\pi}}(\sigma^2(u)+h_2^2)^{-1} \\ G(u,h_2) &= \int (\phi'(t/h_2))^2 g_{\epsilon|u}(t)dt = \frac{\sigma^2(u)}{2\pi h_2^2}(2\sigma^2(u)+h_2^2)^{-1} \end{aligned}$$

Hence

$$R(u,h_2) = \frac{h_2^4 + 2\sigma^2(u)h_2^2 + 2\sigma^4(u)}{h_2^4 + 2\sigma^2(u)h_2^2} > 1$$

and when $h_2 \to \infty$, $R(u,h_2) \to 1$.

(b) Suppose $g_{\epsilon|u} = \alpha N(0,\sigma^2(u)) + (1-\alpha)N(0,\delta^2(u))$, with $\delta^2(u) > \sigma^2(u)$. Let $\sigma^2 = \sigma^2(u)$ and $\delta^2 = \delta^2(u)$. It is easy to derive that

$$G(u,h_2)/F^2(u,h_2) = \frac{(\alpha\sigma^2+(1-\alpha)\delta^2)h^4 + 2(\alpha\sigma^4+(1-\alpha)\delta^4)h^2 + 2(\alpha\sigma^2+(1-\alpha)\delta^2)^2}{h^4 + 2(\alpha\sigma^2+(1-\alpha)\delta^2)h^2}$$

and

$$E(\epsilon^2) = \alpha\sigma^2 + (1-\alpha)\delta^2$$

So we can see that when $h_2 \to \infty$, $\frac{G(u,h_2)}{E(\epsilon^2)F^2(u,h_2)} \to 1$ and if $\alpha = 1$, $\frac{G(u,h_2)}{E(\epsilon^2)F^2(u,h_2)} \geqslant 1$. Let $f(h_2) = G(u,h_2)/F^2(u,h_2) - E(\epsilon^2)$. When $\alpha \in (0.5,1)$, by solving $f(h_2) < 0$, it is easy to derive that $h_2 > \frac{(\alpha\sigma^2+(1-\alpha)\delta^2)^2}{2\alpha(1-\alpha)\sigma^2\delta^2}$. This completes the proof. □

Proof of Theorem 3.5.2.

$$R_\alpha(h_2,t)=\frac{tr\{(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))^T((\boldsymbol{W}+\boldsymbol{\Sigma})^-\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^-)(\boldsymbol{I}_s\otimes\boldsymbol{B}(t)))\}}{tr\{(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))^T((\boldsymbol{A}_0+\boldsymbol{\Sigma})^-\boldsymbol{A}_2(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-)(\boldsymbol{I}_s\otimes\boldsymbol{B}(t))\}}$$

By using similar arguments as proof of Theorem 3.5.1

$$\boldsymbol{Q}\equiv\boldsymbol{Q}(h_2)\to -h_2^{-6}\phi^2(0)\mathrm{E}\{\epsilon^2 g_0'(U)^2(\boldsymbol{C}_1-\mathrm{E}(\boldsymbol{C}_1|U)(\boldsymbol{C}_1-\mathrm{E}(\boldsymbol{C}_1|U)^T\}$$

and

$$\boldsymbol{W}\equiv\boldsymbol{W}(h_2)\to -h_2^{-3}\phi(0)\mathrm{E}\{g_0'(U)^2(\boldsymbol{C}_1-\mathrm{E}(\boldsymbol{C}_1|U)(\boldsymbol{C}_1-\mathrm{E}(\boldsymbol{C}_1|U)^T\}$$

as $h_2\to\infty$. Thus, we have

$$\lim_{h_2\to\infty}(\boldsymbol{W}+\boldsymbol{\Sigma})^-\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^-=(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-\boldsymbol{A}_2(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-.$$

If ϵ is normal and independent of $\boldsymbol{X}$, we can have

$$\lim_{h_2\to\infty}G(u,h_2)F^{-2}(u,h_2)/\sigma^2(u)=1.$$

Thus $\inf_{h_2}(\boldsymbol{W}+\boldsymbol{\Sigma})^-\boldsymbol{Q}(\boldsymbol{W}+\boldsymbol{\Sigma})^-\leqslant(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-\boldsymbol{A}_2(\boldsymbol{A}_0+\boldsymbol{\Sigma})^-$ holds.

□

Chapter 5

Conclusions

This dissertation is motivated by two challenges from NIH study. First, all functional predictors are sparsely and irregularly observed, that is, the measurement time varies from individual to individual. Functional predictors over the entire time interval must be estimated in order to do the regression. In addition, some predictors, such as height, should be monotone over time, and a non monotone estimation of height would make no sense. Second, the relationship between the response and functional predictors are not usually linear and there exists outliers in the response. We solve these issues by proposing two methods as presented in Chapter 2 and Chapter 3.

In Chapter 2, we develop a procedure of estimating monotone functions based on a monotone transformation, the functional principal component (FPC) analysis and a penalized regression to handle the situations where data are sparsely and irregularly observed. We also prove the asymptotic consistency property for this proposed estimator. Extensive simulation studies show the proposed method outperforms the existing methods.

In Chapter 3, to deal with the outliers in the response, we propose a robust procedure of variable selection for functional single index model. The penalty term is

also the group SCAD. The asymptotic consistency, sparsity and normality properties have been derived for the resulting penalized estimator. We propose a method to select the optimal bandwidth based on the asymptotic relative efficiency. We show that by using the optimal bandwidth, the LMR estimator is either robust when the error distribution is heavy tailed, or as asymptotically efficient as the ordinary least square based estimator when the error distribution is a Gaussian distribution. In addition, we investigate the influence function of proposed estimator. Extensive simulation studies show the proposed method outperforms the existing methods.

Bibliography

H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

M. Avella Medina and E. Ronchetti. Robust statistics: a selective overview and new directions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(6): 372–393, 2015.

P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart. Regularization in statistics. *Test*, 15(2):271–344, 2006.

R. E. Black, L. H. Allen, Z. A. Bhutta, L. E. Caulfield, M. De Onis, M. Ezzati, C. Mathers, J. Rivera, Maternal, C. U. S. Group, et al. Maternal and child undernutrition: global and regional exposures and health consequences. *The lancet*, 371(9608):243–260, 2008.

D. A. Bloch and B. W. Silverman. Monotone discriminant functions and their applications in rheumatology. *Journal of the American Statistical Association*, 92 (437):144–153, 1997.

W. B. Capra and H.-G. Müller. An accelerated-time model for response curves. *Journal of the American Statistical Association*, 92(437):72–83, 1997.

H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41, 2005.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003.

D. Chen, P. Hall, H.-G. Müller, et al. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747, 2011.

R. Cook, L. Forzani, and A. Yao. Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression. *Statistica Sinica*, pages 235–238, 2010.

C. M. Crainiceanu, A.-M. Staicu, and C.-Z. Di. Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561, 2009.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

J. Fan and J. Jiang. Variable bandwidth and one-step local m-estimator. *Science in China Series A: Mathematics*, 43(1):65–81, 2000.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

J. Fan, T.-C. Hu, and Y. K. Truong. Robust non-parametric function estimation. *Scandinavian journal of statistics*, pages 433–446, 1994.

J. Fan, T. Huang, et al. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057, 2005.

L. Ferré and A.-F. Yao. Functional sliced inverse regression analysis. *Statistics*, 37 (6):475–488, 2003.

L. Ferré and A.-F. Yao. Smoothed functional inverse regression. *Statistica Sinica*, pages 665–683, 2005.

D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4): 830–851, 2011.

W. Hardle and T. Gasser. Robust non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 42–51, 1984.

T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2011.

P. J. Huber. *Robust statistics*. Springer, 2011.

J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

C.-R. Jiang, W. Yu, J.-L. Wang, et al. Inverse regression for longitudinal data. *The Annals of Statistics*, 42(2):563–591, 2014.

C. Kelly and J. Rice. Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, pages 1071–1085, 1990.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

R. Li and H. Liang. Variable selection in semiparametric regression modeling. *Annals of statistics*, 36(1):261, 2008.

Y. Li, N. Wang, and R. J. Carroll. Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 2012.

J. Liu, R. Zhang, W. Zhao, and Y. Lv. A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, 122:226–238, 2013.

S. Ma. Estimation and inference in functional single-index models. *Annals of the Institute of Statistical Mathematics*, 68(1):181–208, 2016.

C. L. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

H.-G. Müller. Functional modeling of longitudinal data. *Longitudinal Data Analysis*, 1:223–252, 2008.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Annals of Statistics*, pages 774–805, 2005.

H. Peng and T. Huang. Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4):1362–1379, 2011.

J. Ramsay. Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375, 1998.

J. O. Ramsay. Monotone regression splines in action. *Statistical science*, pages 425–441, 1988.

J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

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

F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7):730–742, 2006.

L. Schumaker. Spline functions: basic theory. 1981. *John Wiley&Sons, New York*, 1981.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. B. Tsybakov. Robust reconstruction of functions by the local-approximation method. *Problemy Peredachi Informatsii*, 22(2):69–84, 1986.

R. Weiss. Multivariate density estimation: theory, practice, and visualization. *Journal of the American Statistical Association*, 89(425):359–361, 1994.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

W. Yao, B. G. Lindsay, and R. Li. Local modal regression. *Journal of nonparametric statistics*, 24(3):647–663, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.