# Functional Regression Models

by

**Boyi Hu**

M.Sc., University of British Columbia, 2018
M.Sc., Lakehead University, 2016
B.Sc., University of Science and Technology of China, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science


**SIMON FRASER UNIVERSITY**
**Fall 2023**

# Declaration of Committee

**Name:** **Boyi Hu**

**Degree:** **Doctor of Philosophy**

**Thesis title:** **Functional Regression Models**

**Committee:**

**Chair:** Owen Ward
Assistant Professor, Statistics and Actuarial Science

**Jiguo Cao**
Supervisor
Professor, Statistics and Actuarial Science

**Richard Lockhart**
Committee Member
Professor, Statistics and Actuarial Science

**Haolun Shi**
Examiner
Assistant Professor, Statistics and Actuarial Science

**Lily Wang**
External Examiner
Professor, Statistics
George Mason University

# Abstract

The conventional method for functional quantile regression is to fit the regression model for each quantile of interest separately. The slope function of the regression, as a bivariate function indexed by time and quantile, is actually estimated as a univariate function of time only by first fixing the quantile. This estimation strategy has two major limitations. The monotonicity of conditional quantiles can not be guaranteed, and the smoothness of the slope estimator as a bivariate function can not be controlled. We develop a new framework for functional quantile regression to overcome the two limitations. We propose to simultaneously fit the functional quantile regression model for multiple quantiles under some constraints so that the estimated quantiles satisfy the monotonicity conditions. Meanwhile, the smoothness of the slope estimator is controlled.

Motivated by an application of modeling the impact of daily temperature, annual precipitation and irrigation system on soybean yield, we propose two locally sparse estimation methods under a semi-parametric functional quantile regression model. In the target application, the daily temperature is a functional predictor, and the influence of daily temperature on soybean yield may not always exist during the whole growing season. We aim to identify the time regions where the influence exists. For this purpose, in two projects, we use two different penalized estimation methods, functional SCAD and modified group lasso, to obtain locally sparse estimations for the bivariate slope function associated with the functional predictor.

Focusing on the soybean yield application introduced above, we further propose a novel semi-parametric functional generalized linear model (FGLM) to analyze the relationship between the environmental factors and the soybean yield. In this project, we consider the data from different years as from different populations due to the fact that the climate conditions can be very different year by year. Based on the new assumption, the main challenge is that we only have limited number of observations for each year. To solve this issue, we combine a density ratio model with the proposed semi-parametric FGLM so that the new framework can be fitted using the pool data. We propose to use a combination of penalized B-spline and empirical likelihood method to fit the model. The proposed method is highly flexible and robust to model misspecification.

**Keywords:** Functional quantile regression; Locally sparse estimation; Bivariate splines; Functional generalized linear model; Empirical likelihood; Penalized B-splines

# Dedication

To my wife, Xixi, and my son, Hubi. You have made me stronger, better and more fulfilled than I could have ever imagined. To my dear parents, thanks for your unconditional love and endless support.

# Acknowledgements

At the finish line of my PhD study, I would like to first express my sincerest gratitude to my advisor Dr. Jiguo Cao for bringing me into the field of functional data analysis. His tremendous support and endless encouragement helped me through my PhD journey and allowed me to pursue my research dream. Without his insightful and creative suggestions, I would have never gotten the opportunities of working on so many interesting research topics. In addition to his academical advice, his support and patience when we welcomed our son also allowed me to become a more supportive partner and a better father for my family. I am truly fortunate to have him as my advisor.

I also wish to express my deepest appreciation to Dr. Richard Lockhart for being extremely patient with all my questions and for organizing the reading group on high-dimensional data analysis. He bore with me during so many long discussions on the problems I encountered in my research, and provided me with knowledgeable and brilliant insights.

I am deeply grateful to my thesis committee members for taking their time to read my thesis and help me improve it. Many thanks to Dr. Haolun Shi and Dr. Richard Lockhart for serving on my committee, and Dr. Owen Ward for chairing my defence. Special thanks to Dr. Lily Wang from George Mason University for taking time from her busy schedule to serve as the external examiner.

I am lucky to have the opportunities to work as a teaching assistant for courses instructed by Professors Richard Lockhart, the late Steve Thompson, Tim Swartz, and Liangliang Wang. Their supportive guidance helped improve my teaching skills.

I also want to thank all the staff and faculty members of the Department of Statistics and Actuarial Science for creating a delightful and inclusive environment.

My journey would be lonely without the company of my PhD friends in the department including but not limited to Charlie Zhou, Zhiyang Zhou, Sidi Wu, Alex Wang, William Ruth, Tianyu Guan, John Wang, Yuping Yang, Shijia Wang and Shufei Ge. Thank you for all the fun we had together!

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Nowadays, with modern technology, more and more data can be recorded continuously over a period of time such as daily temperature curves or daily humidity curves. They are both examples of functional data, which has become a commonly encountered type of data. Functional data analysis (FDA) focuses on the analysis and the theory of such data. Under the FDA framework, sample of functional data are considered as realizations of some random functions. From this point of view, functional data are intrinsically infinite dimensional objects. Then, in addition to the curves, analyzing images, surfaces or other infinite dimensional objects varying over a continuum also belongs to FDA. In FDA, functional quntile regression and functional generalized linear model are two popular and useful tools in real data applications. In this thesis, we propose new methodologies for these two models.

Functional quantile regression is an useful tool to analyze the relationship between functional predictors and scalar response. The simplest version of functional quantile regressions can be formulated as

$$Q_Y(u \mid X) = c(u) + \int_{\mathscr{T}} X(t)\beta(t,u)dt, \tag{1.1}$$

where $Y$ is scalar response, $X(t)$ is a functional predictor over the time domain $\mathscr{T}$, $Q_Y(u \mid X)$ is the $u$-th quantile of $Y$ conditioning on $X(t)$. The function $\beta(t,u)$ in model (1.1) is called the slope function, which is a bivariate function indexed by the time $t$ and quantile $u$. In real applications, $\beta(t,u)$ is of great interest because it describes the dynamic influence of the $X(t)$ on the conditional distribution of $Y$.

The commonly used estimation strategy for model (1.1) is to first fix a quantile $u$, and then estimate $\beta(t,u)$ as a univariate function of $t$ only, $\beta_u(t)$ [6, 41]. The advantage of this conventional method is that it can reduce the dimensionality of the parameter space but it also have two major limitations. Since the model (1.1) is fitted for different quantiles separately, the estimation of $\beta(t,u)$ obtained from the conventional method is usually very wiggle as a bivariate function. Moreover, the estimated conditional quantiles of $Y$ may cross each other, which violates the monotonicity condition of a valid quantile function. In practice, the non-monotone quantile estimations can lead to invalid interpretation. In Chapter 2, propose a novel estimation strategy to overcome these two

limitations. We use bivariate splines [64, 46] to approximate $\beta(t,u)$ and then fit the model (1.1) for multiple quantiles simultaneously. We impose constraints and regularization into our estimation procedure to control the smoothness of the estimator for $\beta(t,u)$ and to guarantee the monotonicity of the quantile estimations.

Motivated by an application of analyzing the impact of daily temperature, annual precipitation and irrigation system on soybean yield, in Chapter 3 and 4, we consider a more general semi-parametric functional quantile regression with both scalar predictors and functional predictors. The $u$-th quantile of a scalar response $Y$ conditioning on scalar predictors, $\boldsymbol{Z}$, and functional predictors $\boldsymbol{X}(t)$, $Q_\tau(Y|\boldsymbol{Z},\boldsymbol{X}(t))$ can be modeled as: For this purpose, we use the same soybean yield data set as in Chapter 2, and for this data set, we consider the following functional quantile regression model,

$$Q_Y(u|\boldsymbol{Z},X) = \boldsymbol{Z}^\tau \boldsymbol{\alpha}(u) + \sum_k \int_{\mathcal{T}} \beta_k(t,u) X_k(t) dt, \tag{1.2}$$

where $Y$ is a scalar response, $\boldsymbol{Z} = (1, Z_1, Z_2, \ldots, Z_p)$ is a vector of scalar predictors and $\{X_k(t)\}$ are functional predictors defined over the same domain $\mathcal{T}$. For the target application, $Y$ is the annual soybean yield, $X_k(t)$ are the daily temperature related measurements, and $Z_1$ and $Z_2$ are the annual precipitation and the ratio of irrigated area of each county in Kansas. In (1.2), each entry of $\boldsymbol{\alpha}(u)$ is a varying coefficient depending on the quantiles and $\beta_k(t,u)$ are a bivariate slope functions associated with different functional predictors.

Reviewing the literature [72, 79], we know that there exists a comfortable range of temperature for soybean growth and within the range, small fluctuations of the temperature has no influence on the soybean growth. In other words, under model (1.2), there should exist some sub-regions within the domain of $X_k(t)$ where $\beta_k(t,u) = 0$. We call this property as local sparsity. In Chapter 3 and 4, we focus on finding locally sparse estimators for $\beta(t,u)$ under the model (1.2). Specifically, in Chapter 3, we adopt the conventional estimation strategy to fit the model (1.2). That is, we first fix the quantile $u$, and then estimate $\boldsymbol{\alpha}(u)$ as a vector of scalar parameters and estimate $\beta_k(t,u)$ as univariate functions. We propose a method that combines smoothed quantile loss [32] and functional SCAD method [50] to obtain locally sparse estimation for $\beta_k(t,u)$. In Chapter 4, we follow the idea of Chapter 2 to approximate $\beta_k(t,u)$ using bivariate splines and fit the model (1.2) for multiple quantiles simultaneously. With the help of a modified group lasso [89] type penalty, we propose another locally sparse estimator for $beta_k(t,u)$ based on the simultaneous functional quantile regression.

Focusing on the soybean yield application, in Chapter 5, we further propose a novel semi-parametric functional generalized linear model (FGLM) to analyze the relationship between the environmental factors and the soybean yield. In the real data analysis of Chapter 3 and 4, we assume that the observations from different years, namely between 1991 and 2006, have the same distribution. In other words, we assume that model (1.1) and (1.2) hold for all the observations collected between 1991 and 2006. This assumption neglect the possibility that the climate conditions

of Kansas can be very different across years and we do not have enough environmental variables, such as daily humidity and sunshine, in the data set to account for this variation.

For this reason, in this project, we consider the data from different years as from different populations. Then a new challenge arises. We only have limited number of observations for each year. The small sample size could bring difficulty in the estimation of the unknown functions associated with functional predictors. To solve this issue, we combine a density ratio model with the proposed semi-parametric FGLM so that the new framework can be fitted using the pool data. The proposed method is very flexible and robust to model misspecification.

# Chapter 2

# Simultaneous Functional Quantile Regression

## 2.1. Introduction

The $u$-th quantile of a scalar response $Y$ conditioning on a functional covariate $X(t)$, $Q_Y(u \mid X)$ can be modeled as:

$$Q_Y(u \mid X) = c(u) + \int_{\mathscr{T}} X(t)\beta(t,u)dt, \tag{2.1}$$

where $X(t)$ is a stochastic process defined on a compact interval $\mathscr{T}$, and $\beta(t,u)$ is a bivariate slope function indexed by both time $t$ and quantile $u$. The model (2.1) is called the functional quantile regression model. The slope function $\beta(t,u)$ is of primary interest because it describes how the quantile of the response variable is related to the functional covariate.

In the literature, a commonly used strategy for estimating $\beta(t,u)$ is to treat it as a univariate function of $t$ by fixing the quantile $u$ first. This conventional strategy has two major limitations. First, the slope function $\beta(t,u)$ is usually assumed to be smooth over both $t$ and $u$, which is also favorable in real applications. However, fitting the regression models for different quantiles separately cannot guarantee that the resulting estimator for $\beta(t,u)$ is smooth over $u$. Second, for some observations, the estimation of $Q_Y(u \mid X)$ may not be monotonically increasing in $u$ as it should be. These crossing quantiles can further lead to invalid distribution estimation for the response variable.

In this chapter, we address the above two limitations. Different from the existing methods that $\beta(t,u)$ is estimated as the univariate function of $t$ for each fixed $u$, we propose to use bivariate spline basis functions to approximate $\beta(t,u)$ directly and then estimate the corresponding basis coefficients. Under our framework, the smoothness of the estimation is guaranteed by the smoothness of the bivariate spline approximation, which is ensured by adding some linear constraints on the spline coefficients. In addition, we further impose some extra linear constraints to mitigate the crossing-quantile problem. In this way, we can make sure that the estimated quantiles for each subject are monotone. The monotonicity issue, to some extent, can be fixed by using some monotonization techniques, such as [13]. But it can not improve the estimation for $\beta(t,u)$, because the monotonicity of the quantiles are not considered in the estimation procedure for $\beta(t,u)$ and the

monotonization is only applied to the estimated quantiles. For example, [41] proposed first to estimate $\beta(t,u)$ for the model (2.1), and then to estimate conditional quantile functions based on the estimated $\beta(t,u)$. For the quantile functions that are not monotone, he adjusted them to become monotone by using the technique of [13]. However, the estimation for $\beta(t,u)$ was left unchanged.

The model we consider is an extension of linear quantile regression (LQR) model, which describes the linear relationship between conditional quantiles of a scalar response and some predictor variables [43]. By estimating multiple conditional quantiles, LQR allows us to depict and then make the inference on the entire distribution of the response conditioning on the predictors. Linear quantile regression has been well studied and makes many contributions to real-world applications [45, 83].

Nowadays, functional variables becomes more and more common in real-world applications. FDA has become a comprehensive branch of statistics that provides a useful and convenient framework to analyze functional data with some high dimensional structures, such as curves, images, and surfaces, which are so-called functional data. Estimation for a quantile regression with a scalar response and some functional covariates is a fruitful research topic as related questions arise in many recent applications, such as [7], [11], [86], [81] and [92].

The model (2.1) was first formulated in [6] as a natural extension of classical linear quantile regression. In the paper, a penalized spline estimator for $\beta(t,u)$ was proposed for a fixed $u$ without any dimension reduction on the functional covariate. Later on, for the same model (2.1), [41] proposed to first use functional principal component analysis (FPCA) to truncate the functional covariate $X(t)$ for dimension reduction and then to estimate the slope function $\beta(t,u)$ for a fixed $u$ based on the conventional linear quantile regression framework. [41] also established an optimal convergence rate for the proposed estimator under the minimax sense.

## 2.2. Proposed Method

### 2.2.1 Estimation Procedure

Let $Y$ be a scalar random variable, and $X(t)$ be a random function with mean curve $\mu(t)$, where $t \in \mathscr{T}$, and $\mathscr{T} \subset \mathbb{R}$ is a compact set. Let $\mathbf{\Omega} = \mathscr{T} \times \mathscr{A}$, where $\mathscr{A} \subset (0,1)$ is an interval. For any $u \in \mathscr{A}$, the $u$-th quantile of $Y$ given the functional covariate $X(t)$ is modelled by the following functional quantile model,

$$Q_Y(u \mid X) = c(u) + \int_{\mathscr{T}} X(t)\beta(t,u)dt. \tag{2.1}$$

To estimate the slope function $\beta(t,u)$ in (2.1), We propose to first approximate $\beta(t,u)$ by bivariate splines, and then estimate the corresponding coefficients.

There are multiple types of bivariate splines that can be used for the approximation, such as tensor products of B-splines [75, 64, 91] and bivariate Bernstein polynomials over triangulations [46]. In this project, we choose the Bernstein polynomials over a triangulation to approximate the

bivariate slope function in (2.1). In comparison with the tensor products of B-splines, the bivariate Bernstein polynomials enjoy the advantage that the triangulation technique allows the local refinement, that is, we can flexibly adjust the number of bivariate basis functions with different resolutions in various local areas of the two-dimensional space $\mathscr{T} \times [0, 1]$, which is convenient in many applications. Of course, the Bernstein polynomials and the triangulation technique are not a must for the proposed method and other bivariate bases should also work.

Figure 2.1 shows the example of local refinement of a triangulation. The left panel of Figure 2.1 shows a triangulation over $[0, 1] \times [0, 1]$. The right panel of Figure 2.1 shows the triangulation after a local refinement by adding a new vertex D inside the triangle $\triangle$ABC. The triangle $\triangle$ABC is further split into three triangles: $\triangle$ABD, $\triangle$BCD and $\triangle$ACD.

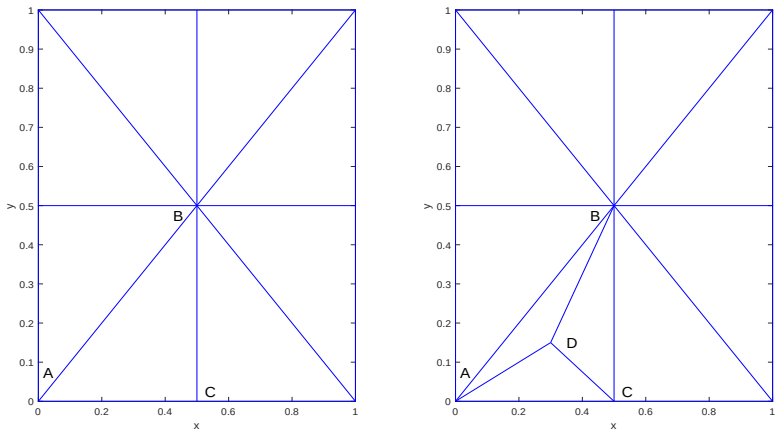



Figure 2.1: Example of local refinement of triangulation. The left panel shows a triangulation over $[0, 1] \times [0, 1]$. The right panel shows the triangulation after a local refinement by adding a new vertex D inside the triangle $\triangle$ABC.

Suppose that $\mathscr{A}$ is the interval containing multiple quantiles of interest. Our goal is to find a function $s(t, u) \in S_d^r(\Delta)$ that can well approximate the slope function $\beta(t, u)$ on the domain $\mathscr{T} \times \mathscr{A}$. To make our writing and proofs in the subsequent sections clearer, we use $\{b_j(t, u)\}_{j=1}^J$ to denote the Bernstein polynomials defined over the triangulation $\Delta = \{\Lambda_1, \ldots, \Lambda_M\}$, where $j = 1, \ldots, J$ is the index for the polynomials. The relationship between $J$ and $M$ is $J = (d + 2)(d + 1)M/2$, because there are $(d + 2)(d + 1)/2$ Bernstein polynomials associated with each triangle of $\Delta$. In addition, for each basis function $b_j(t, u)$, we denote its support by $\Delta_j$, which is a specific triangle of $\Delta$ with $\Delta_j =$ the triangle of $\Delta$ that is the support of $b_j(t, u)$. In other words, $b_j(t, u) \neq 0$ for $(t, u) \in \Delta_j$, and $b_j(t, u) = 0$ for $(t, u) \notin \Delta_j$. If two Bernstein polynomials $b_j(t, u)$ and $b_k(t, u)$ are associated with the same triangle, then $\Delta_j$ and $\Delta_k$ are identical.

The function $s(t,u) \in S_d^r(\mathscr{T} \times \mathscr{A})$ that approximates $\beta(t,u)$ can be written as a linear combination of Bernstein polynomials $\{b_j(t,u)\}_{j=1}^J$. Then on the domain $\mathscr{T} \times \mathscr{A}$, we have the approximation

$$\beta(t,u) \approx s(t,u) = \sum_{j=1}^{J} \gamma_j b_j(t,u) \in S_d^r(\Delta), \tag{2.2}$$

where $\{\gamma_j\}_{j=1}^J$ are the corresponding coefficients.

Under some conventional assumptions on $X(t)$, which are commonly assumed in the literature [84, 71, 60, 59, 73] and are usually satisfied in real applications, by the Mercer's theorem, $X(t)$ admits the decomposition

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \tag{2.3}$$

where $\phi_k(t)$, $k = 1, \ldots$, are called functional principal components (FPC) and $\xi_k$ are called functional principal component scores. By the decomposition (2.3) and the approximation (2.2), model (2.1) can be approximately re-expressed as

$$\begin{aligned} Q_Y(u \mid X) &\approx c(u) + \int_{\mathscr{T}} \mu(t)\beta(t,u)dt + \int_{\mathscr{T}} \sum_{k=1}^{\infty} \xi_k \phi_k(t) s(t,u) dt, \\ &= c_0(u) + \int_{\mathscr{T}} \sum_{k=1}^{\infty} \xi_k \phi_k(t) s(t,u) dt, \end{aligned}$$

where $c_0(u) = c(u) + \int_{\mathscr{T}} \mu(t)\beta(t,u)dt$. Let $\{b_{0,j}(u)\}_{j=1}^{J_0}$ denote the univariate B-spline basis functions defined over the interval $\mathscr{A}$. Then we further approximate $c_0(u)$ by $c_0(u) \approx \sum_{j=1}^{J_0} \gamma_{0,j} b_{0,j}(u) = \boldsymbol{b}_0^{\mathrm{T}}(u)\boldsymbol{\gamma}_0$, where $\boldsymbol{b}_0^{\mathrm{T}}(u) = (b_{0,1}(u), \ldots, b_{0,J_0}(u))$ and $\boldsymbol{\gamma}_0^{\mathrm{T}} = (\gamma_{0,1}(u), \ldots, \gamma_{0,J_0}(u))$.

Under the functional data context, functional observations as the infinite dimension subjects can not fit in the conventional framework of linear quantile regression. In addition, the observed functional data is not always smooth enough for using numerical integration to approximate the integral in (2.1). To tackle these problems and to extend the classical linear quantile regression to functional quantile regression, we usually need to truncate the functional observations $\{x_i(t)\}_{i=1}^n$ to reduce the dimensionality and to smooth them. A plausible approach for the dimensionality reduction is to truncate $X(t)$ by using its first $m$ functional principal components obtained from the decomposition (2.3).

As we have mentioned in the introduction, different from the conventional methods, we want to estimate $\beta(t,u)$ as a bivariate function directly. Therefore, all the quantiles of interest should be considered simultaneously in the estimation procedure. There are a lot of papers that have discussed the advantage of combining multiple quantile regression models, such as [96, 40, 93, 31]. One commonly used approach is to consider the sum of those models.

We know that for a real-valued random variable $Y$, the minimizer of $E\{\rho_u(Y-u)\}$ is the $u$-quantile of $Y$, where $\rho_u(x) = x\,(u - \mathbb{1}\{x < 0\})$ is called the check function [43]. Assume that we observe independent and identically distributed data pairs $\{y_i, x_i(t)\}_{i=1}^n$ as the realizations of

$\{Y, X(t)\}$. We use $A \in \mathscr{A}$ to denote a set of quantiles of interest, which are assumed to be uniformly distributed in $\mathscr{A}$, and use $n_A$ to denote the cardinality of $A$. We first apply FPCA on $\{y_i, x_i(t)\}_{i=1}^n$ to obtain the estimated FPCs $\{\hat{\phi}_k(t)\}$, and FPC scores $\{\hat{\xi}_{ik}\}_{k=1}^m$. Then, based on the approximation (2.2), a reasonable estimator for $\beta(t,u)$ should minimize the following loss function

$$\frac{1}{nn_A}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \int_{\mathscr{T}}\sum_{k=1}^{m}\hat{\xi}_{ik}\hat{\phi}_k(t)s(t,u_r)dt\right), \tag{2.4}$$

with respect to $s(t,u) \in S_d^r(\Delta)$ and $\boldsymbol{\gamma}_0$.

The conventional framework of linear quantile regression is designed for finite dimensional subjects, and the slope parameter to be estimated in classical linear quantile regression is also finite-dimensional. Although the functional observations can be truncated by FPCA into a finite dimension, the slope function $\beta(t,u)$ in the model (2.1) is still infinite-dimensional. As a consequence, the direct extension (2.4) of the conventional linear quantile regression framework to functional data can lead to invalid estimation for $\beta(t,u)$ and the uniqueness of the minimizer of (2.4) cannot be guaranteed.

To clarify this, let $\{\hat{s}(t,u), \hat{\boldsymbol{\gamma}}_0\}$ be a minimizer of (2.4) and fix the truncation level at $m$. Assume that there exists another function $s_1(t,u) \in S_d^r(\Delta)$ such that $s_1(t,u)$ is orthogonal to the first $m$ estimated FPCs of $X(t)$. Then $\{\hat{s}(t,u) + s_1(t,u), \hat{\boldsymbol{\gamma}}_0\}$ is another minimizer of (2.4). More specifically, by FPCA, we can obtain the $(m+1)$-th FPC denoted as $\hat{\phi}_{m+1}(t)$, which is orthogonal to the first $m$ estimated functional principal components, $\hat{\phi}_1(t), \ldots, \hat{\phi}_m(t)$. If there exists some measurable function $w(u)$ such that $\hat{s}(t,u) + w(u)\hat{\phi}_{m+1}(t)$ also belongs to the space $S_d^r(\Delta)$, then $\left\{\hat{s}(t,u) + w(u)\hat{\phi}_{m+1}(t), \hat{\boldsymbol{\gamma}}_0\right\}$ is also a minimizer of (2.4). This implies that $\hat{s}(t,u) + w(u)\hat{\phi}_{m+1}(t)$ is another estimator for $\beta(t,u)$. However, the information of $\hat{\phi}_k(t)$ for any $k \geq m+1$ is excluded from our estimation procedure when we choose the truncation level as $m$, and the estimator for $\beta(t,u)$ derived from the estimation procedure should not include any information about $\hat{\phi}_k(t)$ for any $k \geq m+1$. Therefore, the objective function (2.4) derived directly from the conventional linear quantile regression is problematic under the functional data context.

To overcome this problem, we propose to penalize the $L^2$-norm of the approximation $s(t,u)$ during the estimation procedure. In addition, the roughness of the slope function estimator $s(t,u)$ is also a concern under functional data context. Therefore, a roughness penalty for $s(t,u)$ is also used during the estimation procedure. Roughness penalty is a very useful tool to control the smoothness of functions through the estimation procedure. Detailed discussions on roughness penalty for functional data can be found in [66], [8], [68], [67] and [5]. In this project, we consider the following roughness penalty $R(s;\omega_0,\omega_1,\omega_2)$,

$$R(s;\omega_0,\omega_1,\omega_2) = \sum_{\Lambda\in\Delta}\int_{\Lambda}\sum_{d_1+d_2=2}\omega_{d_1}\binom{2}{d_1}\left[\nabla_t^{d_1}\nabla_u^{d_2}s(t,u)\right]^2 dtdu,$$

where $\omega_0$, $\omega_1$ and $\omega_2$ are three tuning parameters representing the weights corresponding to second derivatives in different directions. More specifically, $\omega_0$ is the weight corresponding to $\frac{\partial^2 s}{\partial t^2}$. $\omega_1$ is the weight corresponding to $\frac{\partial^2 s}{\partial t \partial u}$. $\omega_2$ is the weight corresponding to $\frac{\partial^2 s}{\partial u^2}$. Since in our estimation procedure, we include a tuning parameter $\lambda_{2,n}$ for the whole roughness penalty $R(s;\omega_0,\omega_1,\omega_2)$, then $\omega_0$ can be fixed as a constant $\omega_0 = 1$. In addition, if the smoothness of the target slope function along quantile index and functional index can be assumed to be identical or without much difference, then we can simply set $\omega_0 = \omega_1 = \omega_2 = 1$ to reduce the computational cost. Then, $R(s;\omega_0,\omega_1,\omega_2)$ becomes to $R(s)$,

$$R(s) = \sum_{\Lambda \in \Delta} \int_{\Lambda} \sum_{d_1+d_2=2} \binom{2}{d_1} \left( \nabla_t^{d_1} \nabla_u^{d_2} \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{\gamma} \right)^2 dtdu,$$

which is the most commonly used roughness penalty discussed in the literature listed above.

Let $\lambda_{1,n}$ and $\lambda_{2,n}$ be two nonnegative tuning parameters. Then, we estimate the slope function $\beta(t,u)$ in (2.1) by minimizing

$$\frac{1}{nn_A} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \rho_{u_r} \left( y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \int_{\mathscr{T}} \sum_{k=1}^{m} \hat{\xi}_{ik} \hat{\phi}_k(t) s(t,u) dt \right) + \lambda_{1,n} \|s\|^2_{L^2(\boldsymbol{\Omega})} + \lambda_{2,n} R(s;\omega_0,\omega_1,\omega_2), \tag{2.5}$$

with respect to $s(t,u) \in S_d^r(\Delta)$ and $\boldsymbol{\gamma}_0$, where the norm $\|s\|^2_{L^2(\boldsymbol{\Omega})}$ is defined as $\|s\|^2_{L^2(\boldsymbol{\Omega})} = \int_{\mathscr{T}\times\mathscr{A}} s^2(t,u)dtdu$.

For any $s(t,u) \in S_d^r(\Delta)$, we have the expression

$$s(t,u) = \sum_{j=1}^{J} \gamma_j b_j(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{\gamma}, \tag{2.6}$$

where $\boldsymbol{B}(t,u) = (b_1(t,u), \cdots, b_J(t,u))^{\mathrm{T}}$ and $\boldsymbol{\gamma}$ is the vector of coefficients satisfying some linear constraint

$$\boldsymbol{H}\boldsymbol{\gamma} = \boldsymbol{0}. \tag{2.7}$$

The constraint (2.7) ensures that $s(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{\gamma} \in C^r(\mathscr{T}\times\mathscr{A})$. The matrix $\boldsymbol{H}$ depends on the triangulation $\Delta$, the degree $d$, and the smoothness parameter $r$ of the spline space $S_d^r(\Delta)$ [46]. For example, when $r = 1$, $s(t,u)$ is assumed to have the continuous first partial derivatives over both $t$ and $u$. An useful technique to remove the constraint (2.7) is QR decomposition [82]. For a given $\boldsymbol{H}$, by QR decomposition, we have

$$\boldsymbol{H}^{\mathrm{T}} = (\boldsymbol{Q}^*, \boldsymbol{Q}) \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix}, \tag{2.8}$$

where $(\boldsymbol{Q}^*, \boldsymbol{Q})$ is a matrix with orthogonal columns and $\boldsymbol{R}$ is a upper triangle matrix with nonzero diagonal elements. With the decomposition (2.8), the constraint $\boldsymbol{H}\boldsymbol{\gamma} = \mathbf{0}$ can be removed by rewriting $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma} = \boldsymbol{Q}\boldsymbol{\theta}. \tag{2.9}$$

Suppose we observe $X_i(t)$ for $t \in T$ and we use $n_T$ to denote the cardinality of $T$. By (2.6), the penalty $\|s\|^2_{L^2(\boldsymbol{\Omega})}$ can be approximated by $\|s\|^2_{L^2(\boldsymbol{\Omega})} \approx \frac{1}{n_A n_T}\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{\gamma}$, where $\boldsymbol{B}_{A,T}$ is a $J$ by $n_A n_T$ matrix with the $j$th row of $\boldsymbol{B}_{A,T}$ being the evaluations of Bernstein polynomials $b_j(t,u)$ for all $t \in T$ and $u \in A$. The roughness penalty $R(s;\omega_0,\omega_1,\omega_2)$ or $R(s)$ can also be written as the matrix form $\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\gamma}$, where the matrix $\boldsymbol{D}$ is a $J$ by $J$ positive definite and block diagonal matrix with each block corresponding to one triangle of the triangulation $\Delta$, and the size of each block depends on the degree $d$.

Define $L_0(\boldsymbol{\theta},\boldsymbol{\gamma}_0) = (nn_A)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}\right)$, the whole quantile loss based on FPCA. Then by (2.6) and (2.9), the minimization problem (2.5) can be converted into the following,

$$\min_{\boldsymbol{\theta},\boldsymbol{\gamma}_0} \; L_0(\boldsymbol{\theta},\boldsymbol{\gamma}_0) + \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta}, \tag{2.10}$$

where $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1},\ldots,\xi_{im})^{\mathrm{T}}$, $\hat{\boldsymbol{P}}(u)$ is an $m \times J$ matrix with the $(k,j)$-entry being $\hat{p}_{k,j}(u) = \int_{(t,u)\in\Delta_j}\hat{\phi}_k(t)b_j(t,u)dt$. Note that, for the matrix $\hat{\boldsymbol{P}}(u)$, and a specific $u$, say $u = u_r \in A$, many entries of $\hat{\boldsymbol{P}}(u_r)$ are zeros because the integral $\int_{(t,u)\in\Delta_j}\hat{\phi}_k(t)b_j(t,u_r)dt$ is equal to zero if the triangle $\Delta_j$, which is the support of $b_j(t,u_r)$, does not intersect with the horizontal line $u = u_r$.

If we denote the minimizer of (2.10) by $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$, then our proposed estimator for $\beta(t,u)$ in (2.1) is

$$\hat{\beta}(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\hat{\boldsymbol{\theta}}. \tag{2.11}$$

In practice, to guarantee the estimated conditional quantile functions of all the subjects to be monotone, some extra linear constraints on $\boldsymbol{\theta}$ can be further imposed. Specifically, given $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$, the estimated $u$-quantile of the $i$th subject is $\hat{Q}_Y(u \mid X = x_i) = \boldsymbol{b}_0^{\mathrm{T}}(u)\hat{\boldsymbol{\gamma}}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u)\boldsymbol{Q}\hat{\boldsymbol{\theta}}$.

The monotonicity of $\hat{Q}_Y(u \mid X = x_i)$ can be approximately expressed as $\hat{Q}_Y(u_r \mid X = x_i) \le \hat{Q}_Y(u_r' \mid X = x_i)$ for any $u_r < u_r'$, $u_r, u_r' \in A$. Then a reasonable way to mimic the monotonicity of these quantile functions is to impose the following constraints into the optimization,

$$\{\boldsymbol{b}_0^{\mathrm{T}}(u_r) - \boldsymbol{b}_0^{\mathrm{T}}(u_r')\}\boldsymbol{\gamma}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\{\hat{\boldsymbol{P}}(u_r) - \hat{\boldsymbol{P}}(u_r')\}\boldsymbol{Q}\boldsymbol{\theta} \le \mathbf{0}$$

for any quantile $u_r < u_r'$ and any $i = 1,\ldots,n$, which guarantee that the estimated conditional quantiles of $Y \mid X_i(t)$ do not cross each other [4, 51]. Then, we can solve (2.10) under the constraints

$$\{\boldsymbol{b}_0^{\mathrm{T}}(u_r) - \boldsymbol{b}_0^{\mathrm{T}}(u_r')\}\boldsymbol{\gamma}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\{\hat{\boldsymbol{P}}(u_r) - \hat{\boldsymbol{P}}(u_r')\}\boldsymbol{Q}\boldsymbol{\theta} \le \mathbf{0} \tag{2.12}$$

for all $i = 1,\ldots,n$, and any $u_r < u_r', u_r, u_r' \in A$.

### 2.2.2 Computation

This subsection is concerned with the computational aspect of the minimization problem (2.10). As introduced in [43], for a specific quantile $u$, the minimization of loss function derived from the classical linear quantile regression model is equivalent to a constrained linear programming problem.

For the proposed method, we need to solve the minimization problem (2.10). Following the same idea in [43], (2.10) can also be formulated into the minimization of the following quadratic programming problem with respect to $\boldsymbol{\theta}$, $\boldsymbol{\gamma}_0$, and $\{w_{i,r}, v_{i,r}\}_{i=1,\ldots,n,r=1,\ldots,n_A}$,

$$\frac{1}{nn_A}\sum_{r=1}^{n_A}\left\{u_r\sum_{i=1}^{n}w_{i,r}+(1-u_r)\sum_{i=1}^{n}v_{i,r}\right\}+\lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta}, \tag{2.13}$$

subject to $y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta} = w_{i,r} - v_{i,r}$, $w_{i,r} \geq 0$, and $v_{i,r} \geq 0$, for all $i = 1, \ldots, n$, and $r = 1, \ldots, n_A$.

If we want to further impose the monotonicity constraints (2.12) on (2.13), then the constrained optimization can be similarly formulated as the following problem with respect to $\boldsymbol{\theta}$, $\boldsymbol{\gamma}_0$, and $\{w_{i,r}, v_{i,r}\}_{i=1,\ldots,n,r=1,\ldots,n_A}$,

$$\frac{1}{nn_A}\sum_{r=1}^{n_A}\left\{u_r\sum_{i=1}^{n}w_{i,r}+(1-u_r)\sum_{i=1}^{n}v_{i,r}\right\}+\lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta},$$

subject to $y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta} = w_{i,r} - v_{i,r}$, $\{\boldsymbol{b}_0^{\mathrm{T}}(u_r) - \boldsymbol{b}_0^{\mathrm{T}}(u_r')\}\boldsymbol{\gamma}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\{\hat{\boldsymbol{P}}(u_r) - \hat{\boldsymbol{P}}(u_r')\}\boldsymbol{Q}\boldsymbol{\theta} \leq \mathbf{0}$, $w_{i,r} \geq 0$, and $v_{i,r} \geq 0$, for all $i = 1, \ldots, n$, $r = 1, \ldots, n_A$ and any $u_r < u_r', u_r, u_r' \in A$.

In summary, the complete algorithm can be splitted into two parts:

- Derive the coefficients in (2.13) with or without the monotonicity constraints (2.12), such as $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^n$, $\{\hat{\boldsymbol{P}}(u)\}_{u\in A}$, $\boldsymbol{Q}$, etc. Specifically, we first derive the estimated FPCs $\{\hat{\phi}_k(t)\}_{k=1}^m$ and corresponding scores $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^n$. Next we compute the matrices related to the bivariate spline basis, $\boldsymbol{B}^{\mathrm{T}}(t,u)$, $\boldsymbol{Q}$, $\boldsymbol{B}_{A,T}$ and $\boldsymbol{D}$. Given $\{\hat{\phi}_k(t)\}_{k=1}^m$ and $\boldsymbol{B}^{\mathrm{T}}(t,u)$, $\{\hat{\boldsymbol{P}}(u)\}_{u\in A}$ are approximated by using numerical integration based on Simpson's rule.

- With all the preparations in the previous step, we can code the quadratic programming problem (2.13) with or without the constraints (2.12) in MATLAB and solve it in MATLAB as well.

### 2.2.3 Tuning Parameter Selection

In our proposed method, to obtain the estimation of $c(u)$ and $\beta(t,u)$ in model (2.1), we need to first decide the truncation level $m$, and the values of tuning parameters $\lambda_{1,n}$ and $\lambda_{2,n}$.

For the truncation level $m$, we suggest to use following BIC criterion to choose $m$,

$$BIC(m) = \log\left(n^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left\{y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 - \int_{\mathscr{T}}\sum_{k=1}^{m}\hat{\xi}_{ik}\hat{\phi}_k(t)\hat{\beta}(t,u)dt\right\}\right) + \frac{(m+1)\log n}{n}. \tag{2.14}$$

For the selection of penalty parameters $\lambda_{1,n}$ and $\lambda_{2,n}$, ideally, leave-one-out cross-validation should be the best way to do it. However, the computational cost for each fitting is expensive, therefore, in practice, we usually use five-fold or ten-fold cross-validation to select the parameter values. As we will show in the next section, the value of $\lambda_{2,n}$ depends on the value of $\lambda_{1,n}$. For this reason, we propose a sequential procedure to choose values for $\lambda_{1,n}$ and $\lambda_{2,n}$. and the specific cross-validation procedure is described as follows. We use ten-fold cross-validation as an example.

We first use the complete sample $\{x_i(t)\}_{i=1}^n$ to estimate FPCs $\{\hat{\phi}_k(t)\}_{k=1}^m$, and corresponding scores $\{\hat{\boldsymbol{\xi}}_i\}_{i=1,\ldots,n}$. Then for a fixed $m$, we use ten-fold cross validation to find the optimal value for tuning parameters $\lambda_{1,n}$ and $\lambda_{2,n}$. More specifically, we first apply the cross validation on the following objective function with only one penalty $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta}$,

$$L_{n,1}(\boldsymbol{\theta},\boldsymbol{\gamma}_0) = L_0(\boldsymbol{\theta},\boldsymbol{\gamma}_0) + \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta},$$

to decide the optimal value for $\lambda_{1,n}$ among all candidates, denoted as $\hat{\lambda}_{1,n}$.

Next, based on $\hat{\lambda}_{1,n}$, we apply the cross validation again on the full objective function with two penalties

$$L_{n,2}(\boldsymbol{\theta},\boldsymbol{\gamma}_0) = L_0(\boldsymbol{\theta},\boldsymbol{\gamma}_0) + \hat{\lambda}_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta},$$

to find the optimal value for $\lambda_{2,n}$ among all candidates, denoted as $\hat{\lambda}_{2,n}$. Then $(\hat{\lambda}_{1,n},\hat{\lambda}_{2,n})$ are the optimal values for $(\lambda_{1,n},\lambda_{2,n})$ for the current truncation level $m$. We will repeat this sequential selection procedure for multiple values of $m$, and then choose the optimal value for $m$ based on the criterion (2.14).

## 2.3. Theoretical Results

To investigate the asymptotic properties of the proposed slope function estimator $\hat{\beta}(t,u)$ defined in (2.11), we assume the following conditions on the distribution of the random function $X(t)$, the conditional distribution of $Y \mid X(t)$, and the slope function $\beta(t,u)$.

(A1) $\{Y_i, X_i(t)\}_{i=1}^n$ are independent and identically distributed.

(A2) $\int_{\mathscr{T}} E\left(X^4(t)\right) dt < \infty$, and $E\left(\xi_k^4\right) < C\kappa_k^2$ for all $k \geq 1$.

(A3) For some $\alpha > 1$ and for any $k \geq 1$, $C^{-1}k^{-\alpha} \leq \kappa_k \leq Ck^{-\alpha}$, $\kappa_k - \kappa_{k+1} \geq C^{-1}k^{-\alpha-1}$.

(A4) $\partial F_{Y|X}(y \mid X)/\partial y \vee |\partial^2 F_{Y|X}(y \mid X)/\partial y^2| \leq C$, and $\inf_{u\in\mathscr{A}} f_{Y|X}(Q_{Y|X}(u \mid X) \mid X) \geq C^{-1}$.

(A5) $\beta(t,u) \in W_q^{d+1}(\mathscr{T} \times \mathscr{A})$, and for some $\zeta > \alpha/2 + 1$, $\sup_{u\in\mathscr{A}} |\beta_k(u)| \leq Ck^{-\zeta}$, $k = 1, \ldots,$ where $W_q^{d+1}(\mathscr{T}\times\mathscr{A})$ is a Sobolev space defined over $\mathscr{T}\times\mathscr{A}$, and $\beta_k(u) = \int_{\mathscr{T}} \beta(t,u)\phi_k(t)dt$.

(A6) There exists a finite number $p_0$ such that $\kappa_k = 0$ for all $k \geq p_0$.

The i.i.d. assumption is conventional and the scenario of dependent data is not considered in this project. A2 are commonly assumed restrictions on the moments of $X(t)$ and $\xi_k$. There is no condition on the moment of $Y$ needed. A3 is adapted from (A3) of [41], which ensures the identifiability of $\phi_k(t)$ as well as the estimation accuracy of $\hat{\phi}_k(t)$. A4 are common conditions on the conditional distribution and density functions of $Y$ under quantile regression context. A5 determines the estimation accuracy of $\hat{\beta}(t,u)$ by using the truncated functional covariate, and the Sobolev space assumption ensures that bivariate splines can be used to approximate $\beta(t,u)$. A6 implies that the functional covariate $X_i(t)$ can be represented by a finite number of pairs of FPCs and corresponding FPC scores.

For a triangle $\Lambda$, let $|\Lambda|$ be length of its longest edge, and then for a triangulation $\Delta$, we define $|\Delta| := \max\{|\Lambda| : \Lambda \in \Delta\}$ ( i.e., the length of the longest edge of all triangles in the triangulation $\Delta$). Recall that $n_A$ and $n_T$ represent the cardinalities of $A$ and $T$ as previously defined. The following theorem gives the rate of convergence of the slope function estimator $\hat{\beta}(t,u)$ for a given truncation level $m$ when the FPCA is used to reduce the dimension of the functional covariate.

For any fixed $u \in (0,1)$, we use $\beta_u(t)$ to denote $\beta(t,u)$ and use $\hat{\beta}_u(t)$ to denote $\hat{\beta}(t,u)$. Define

$$A_1 = \{r \in (1,\ldots,n_A) : \left\|\hat{\beta}_{u_r}(t) - \beta_{u_r}(t)\right\|_{L^2} \geq M\kappa_m^{-1/2}m^{1/2}n^{-1/2}, \\ \text{for some constant } M > 0\},$$

where $\left\|\hat{\beta}_{u_r}(t) - \beta_{u_r}(t)\right\|_{L^2} = \left\{\int_{\mathscr{T}} \left(\hat{\beta}_{u_r}(t) - \beta_{u_r}(t)\right)^2 dt\right\}^{1/2}$. The set $A_1$ can be regarded as an index set of quantiles for which the estimation are not good enough.

**Theorem 1.** *Under the conditions A1-A5, and assume further that* $|\Delta| = o\left(m^{-(1+2\alpha)/(2d+2)}n^{-3/(2d+2)}\right)$*, and* $n_A^{-1}|\Delta|^{-1}m^{(\alpha-1)/3} = o(1)$*. Suppose the tuning parameters* $\lambda_{1,n}$ *and* $\lambda_{2,n}$ *satisfy* $\lambda_{1,n} \asymp n_A^{-1}n_T^{-1}m^{-1/2}n|\Delta|^{d+1}$*, and* $\lambda_{2,n} = o(\lambda_{1,n}n_A^{-1}n_T^{-1}|\Delta|^4)$*, then*

$$\left\|\hat{\beta}(t,u) - \beta(t,u)\right\|_{L^2(\mathbf{\Omega})} \approx O_p\left(\kappa_m^{-1/2}m^{1/2}n^{-1/2} \vee m^{-(2\zeta+1)/2}\right).$$

*In addition, for* $A_1$ *we have* $|A_1| = o_p(m^{-1-\alpha}n^{-1/2}n_A)$.

**Remark 1.** The first term of the stochastic order of $\left\|\hat{\beta}(t,u)-\beta(t,u)\right\|_{L^2(\boldsymbol{\Omega})}$ in Theorem 1 is decreasing as the sample size $n$ becomes larger, and is increasing with a larger the truncation level $m$ (i.e, adding more FPCs in the estimation). The second term represents the information loss if we include too few FPCs in the estimation procedure. Then based on condition A5, we can obtain a theoretically optimal truncation level $m \asymp n^{1/(\alpha+2\zeta)}$.

The following theorem presents the asymptotic distribution of the slope estimator $\hat{\beta}(t,u)$. We now assume that $p_0$ is known and finite as in [48]. Under A6 and by Lemma 1 and Lemma 3 in the supplementary material, there exist $\boldsymbol{\gamma}_0^*$ and $\boldsymbol{\theta}^*$ such that

$$\sup_{(t,u)\in\mathscr{T}\times\mathscr{A}}|\beta(t,u)-\boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\boldsymbol{\theta}^*|\le C_1|\Delta|^{d+1},\quad \sup_{u\in\mathscr{A}}|c(u)-\boldsymbol{b}_0^{\mathrm{T}}(u)\boldsymbol{\gamma}_0^*|\le C_2|\Delta|^{d+1},$$

for some constant $C_1$ and $C_2$. Let $\boldsymbol{\Gamma}^*=(\boldsymbol{\gamma}_0^*,\boldsymbol{\theta}^*)^{\mathrm{T}}$, $\boldsymbol{Z_i}(u)=\left[\boldsymbol{b}_0^{\mathrm{T}}(u),\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u)\boldsymbol{Q}\right]$, $\tilde{\boldsymbol{B}}(t,u)=(\boldsymbol{0}_{1\times n_B},\boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q})^{\mathrm{T}}$, and $\tilde{\boldsymbol{Z}}_i=(\boldsymbol{Z}_i^{\mathrm{T}}(u_1),\ldots,\boldsymbol{Z}_i^{\mathrm{T}}(u_{n_A}))$. Then define $\boldsymbol{\Sigma}_1=n_A^{-1}\sum_{r=1}^{n_A}E\left[f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*)\boldsymbol{Z}_i^{\mathrm{T}}(u_r)\boldsymbol{Z}_i(u_r\right.$ and

$$\boldsymbol{\Sigma}_2=\frac{1}{2n}\boldsymbol{\Sigma}_1+\lambda_{1,n}\begin{bmatrix}\boldsymbol{0}&\boldsymbol{0}\\ \boldsymbol{0}&\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\end{bmatrix}+\lambda_{2,n}\begin{bmatrix}\boldsymbol{0}&\boldsymbol{0}\\ \boldsymbol{0}&\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\end{bmatrix},$$

where $f_i$ is the conditional density of $Y_i \mid X_i(t)$. Let $\boldsymbol{U}_1$ be an $n_A$ by $n_A$ matrix with its $(r,r')$-entry being $u_r\wedge u_{r'}-u_ru_{r'}$ for any $r,r'=1,\ldots,n_A$. Define $\boldsymbol{U}_2=n_A^{-2}E\left[\tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\boldsymbol{U}_1\tilde{\boldsymbol{Z}}_i\right]$, and $\boldsymbol{\Sigma}=(2n\boldsymbol{\Sigma}_2)^{-1}\boldsymbol{U}_2/n\,(2n\boldsymbol{\Sigma}_2)^{-1}$.

**Theorem 2.** *Under the conditions of Theorem 1, A6 and $n_An|\Delta|^{d+2}=o(1)$, as $n\to\infty$ and $n_A\to\infty$, for fixed $(t,u)$, we have*

$$\sigma_\beta^{-1/2}(t,u)\left\{\hat{\beta}(t,u)-\beta(t,u)\right\}\to N(0,1)$$

*in distribution, where $\sigma_\beta(t,u)=\tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(t,u)$.*

**Remark 2.** Due to the fact that the number of quantile levels, $n_A$, is used to ensure a good estimate of the bivariate function in the quantile interval, $n_A$ shouldn't be too small. Meanwhile, larger $n_A$ will result in larger number of triangle basis functions, which will increase the variance of the estimator. So in our theorems, $n_A$ needs to satisfy $n_A^{-1}|\Delta|^{-1}m^{(\alpha-1)/3}=o(1)$ and $n_An|\Delta^{d+2}|=o(1)$.

The next theorem presents how to construct a simultaneous confidence region (SCR) for $\beta(t,u)$. Let $\Gamma_{min}(\cdot)$ and $\Gamma_{max}(\cdot)$ represent the minimum and maximum eigenvalues of a square matrix. Let $\boldsymbol{\Omega}_s$ denote the set of vertices of the triangulation $\Delta$ and $|\boldsymbol{\Omega}_s|$ denote the cardinality of the set $\boldsymbol{\Omega}_s$.

**Theorem 3.** *Under the conditions of Theorem 2, and further assume that $\Gamma_{min}(\boldsymbol{\Sigma})$ and $\Gamma_{max}(\boldsymbol{\Sigma})$ are bounded away from 0 and $\infty$ with probability tending to one as $n\to\infty$,*

*(1) As $n,n_A\to\infty$, we have*

$$\sigma_\beta^{-1/2}(t,u)\left\{\hat{\beta}(t,u)-\beta(t,u)\right\}\to\vartheta(t,u),\tag{2.1}$$

*in distribution, where $\vartheta(t,u)$ is a Gaussian random field with mean 0 defined on $\mathbf{\Omega}$ with the covariance function*

$$\begin{aligned} C(t,u,t',u') :&= Cov\left(\vartheta(t,u),\vartheta(t',u')\right) \\ &= \sigma_\beta^{-1/2}(t,u)\sigma_\beta^{-1/2}(t',u')\tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(t',u'). \end{aligned}$$

*Specifically, $C(t,u,t,u) = Var\left(\vartheta(t,u)\right) = 1$.*

*(2) For any $a \in (0,1)$,*

$$\lim_{n\to\infty} P\left\{\sup_{(t,u)\in\mathbf{\Omega}_s}\left|\sigma_\beta^{-1/2}(t,u)\left\{\hat{\beta}(t,u)-\beta(t,u)\right\}\right| \le Q_\beta(a)\right\} = 1-a, \tag{2.2}$$

*where $\mathbf{\Omega}_s$ as a subset of $\mathbf{\Omega}$ becomes denser as $n\to\infty$, and $Q_\beta(a) = (2\log|\mathbf{\Omega}_s|)^{1/2} - (2\log|\mathbf{\Omega}_s|)^{-1/2}\{\log(-0.5\log(1-a)) + 0.5\,[\log(\log|\mathbf{\Omega}_s|) + \log 4\pi]\}$. Then an asymptotic $100(1-a)\%$ simultaneous confidence region (SCR) for $\beta(t,u)$ over $\mathbf{\Omega}_s$ is given by $\hat{\beta}(t,u) \pm \sigma_\beta^{1/2}(t,u)Q_\beta(a)$.*

**Remark 3.** In Theorem 2, the condition $n_A n|\Delta|^{d+2} = o(1)$ is used for undersmoothing of the slope estimator, which is widely applied in the series approximating estimations [87, 88]. By consistently estimating the asymptotic variance $\sigma_\beta(t,u)$, the result in Theorem 2 can be used to establish the pointwise confidence interval of the slope function. Compared with the asymptotic $100(1-a)\%$ point confidence interval in Theorem 2, $\hat{\beta}(t,u) \pm \sigma_\beta^{1/2}(t,u)z_a$, the width of the simultaneous confidence region in Theorem 3 for any $(t,u) \in \mathbf{\Omega}_s$ is inflated by the rate $Q_\beta(a)/z_a$, where $z_a$ is the $a$-quantile of the standard normal distribution.

## 2.4. Simulation Studies

In this section, we use simulation studies to compare the performance of the proposed method and the conventional method in [41] based on the estimation of $\beta(t,u)$ under three scenarios. In Scenario I, the true slope function $\beta(t,u)$ is free from $u$, that is, $\beta(t,u)$ is just a univariate function of $t$, which is similar to the simulation study in [41]. In Scenario II, the true slope function $\beta(t,u)$ is a bivariate function of both $t$ and $u$, which is more complicated than the slope function in the first scenario and is not considered in [41]. The data generating models are introduced as follows.

### 2.4.1 Data Generating Models

**Scenario I.** In this scenario, we suppose that realizations of the $\{X(t), Y\}$ are generated from

$$X(t) = \sum_{i=1}^{10} i^{-1} r_i \phi_i(t) + \mu(t), \quad \mu(t) = \sqrt{3}\phi_1(t),$$
$$Y = \int_0^1 \rho_1(t) X(t) dt + \epsilon, \quad \rho_1(t) = \sum_{j=1}^{10} \tau_j \phi_j(t),$$

where $\phi_j(t) = 2^{1/2}\cos(j\pi t)$, $\tau_1 = 1$, $\tau_j = 4(-1)^{j+1}j^{-2}$ for $j \geq 2$, $r_i$ are i.i.d. Uniform$(-\sqrt{3}, \sqrt{3})$ random variables, and $\epsilon$ is an $N(0,1)$ random variables independent of other random variables. Under this setting, the underlying slope function $\beta(t,u)$ is $\beta(t,u) = \rho_1(t) = \sum_{j=1}^{10} \tau_j \phi_j(t)$.

**Scenario II.** In this scenario, we suppose that the realizations of the pair $\{X(t), Y\}$ are generated from

$$X(t) = \sum_{i=1}^{10} i^{-1} r_i \phi_i(t) + \mu(t), \quad \mu(t) = 3\sqrt{3}\phi_1(t),$$
$$Y = \int_0^1 \rho_1(t) X(t) dt + \sigma(X)\epsilon, \quad \sigma(X) = \int_0^1 \rho_2(t) X(t) dt,$$
$$\rho_1(t) = \sum_{j=1}^{10} \tau_j \phi_j(t), \quad \rho_2(t) = \phi_1(t),$$

where $\phi_j(t) = 2^{1/2}\cos(j\pi t)$, $\tau_1 = 1$, $\tau_j = 4(-1)^{j+1}j^{-2}$ for $j \geq 2$, $r_i$ are i.i.d. Uniform$(-\sqrt{3}, \sqrt{3})$ random variables, and $\epsilon$ is an Gamma$(1,2)$ random variable independent of other random variables. Under this setting, the underlying slope function $\beta(t,u)$ is $\beta(t,u) = \rho_1(t) + \rho_2(t)Q_\epsilon(u) = \sum_{j=1}^{10} \tau_j \phi_j(t) + \rho_2(t)Q_\epsilon(u)$, where $Q_\epsilon(u)$ is the $u$-quantile of the random variable $\epsilon$.

**Scenario III.** In this scenario, we suppose that the realizations of the pair $\{X(t), Y\}$ are generated from

$$X(t) = \sum_{i=1}^{10} i^{-1} r_i \phi_i(t) + \mu(t), \quad \mu(t) = 3\sqrt{3}\phi_1(t),$$
$$Y = \int_0^1 \rho_1(t) X(t) dt + \sigma(X)\epsilon, \quad \sigma(X) = \int_0^1 \rho_2(t) X(t) dt,$$
$$\rho_1(t) = \sum_{j=1}^{10} \tau_j \phi_j(t), \quad \rho_2(t) = \phi_1(t),$$

where $\phi_j(t) = 2^{1/2}\cos(j\pi t)$, $\tau_1 = 1$, $\tau_j = 4(-1)^{j+1}j^{-2}$ for $j \geq 2$, $r_i$ are i.i.d. Uniform$(-\sqrt{3}, \sqrt{3})$ random variables, and $\epsilon$ is an $Gamma(1,2)$ random variable independent of other random variables. Under this setting, the underlying slope function $\beta(t,u)$ is $\beta(t,u) = \rho_1(t) + \rho_2(t)Q_\epsilon(u) =$

Table 2.1: The average of the mean squared error (MSE) defined in (2.1) and the average of maximum absolute error (MAE) defined in (2.2) for the estimations of the slope function $\beta(t,u)$ by using the proposed simultaneously functional quantile regression (SFQR) method and the conventional functional quantile regression (FQR) method under three scenarios with the sample size $n = 300, 400, 500$ respectively in 100 simulation repetitions.

| | n | SFQR | | FQR | | SFQR/FQR | |
|---|---|---|---|---|---|---|---|
| | | MSE | MAE | MSE | MAE | MSE | MAE |
| Scenario I | 300 | 0.085 | 0.699 | 0.099 | 0.706 | 0.848 | 0.991 |
| | 400 | 0.068 | 0.630 | 0.080 | 0.648 | 0.849 | 0.97 |
| | 500 | 0.062 | 0.600 | 0.067 | 0.596 | 0.927 | 1.007 |
| Scenario II | 300 | 0.321 | 1.383 | 0.505 | 1.548 | 0.635 | 0.894 |
| | 400 | 0.253 | 1.241 | 0.381 | 1.380 | 0.664 | 0.900 |
| | 500 | 0.212 | 1.201 | 0.309 | 1.249 | 0.686 | 0.962 |
| Scenario III | 300 | 0.267 | 1.196 | 0.647 | 1.567 | 0.412 | 0.763 |
| | 400 | 0.193 | 1.009 | 0.459 | 1.337 | 0.422 | 0.755 |
| | 500 | 0.172 | 0.945 | 0.379 | 1.222 | 0.453 | 0.773 |

$\sum_{j=1}^{10} \tau_j \phi_j(t) + \rho_2(t) Q_\epsilon(u)$, where $Q_\epsilon(u)$ is the $u$-quantile of the random variable $\epsilon$. Note that in this scenario, the error term has the asymmetric distribution, which is different from the other scenarios.

### 2.4.2 Summary of Simulation Results

In the simulations of Scenario I and 2, the set $A$ consists of 30 quantiles uniformly distributed within the interval $[0.2, 0.8]$. In the simulations of Scenario III, the set $A$ consists of 17 quantiles uniformly distributed within the interval $[0.1, 0.9]$. The truncation level $m$ for the functional covariate $\{x_i(t)\}_{i=1}^n$ by using FPCA is chosen as $m = 3$. The candidate sets for the tuning parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ are chosen as $\{10^{-2}, 10^{-2.5}, 10^{-3}, 10^{-3.5}, 10^{-4}\}$ and $\{10^{-1}, 10^{-2}, 10^{-3}\}$. We consider two metrics, the mean squared error (MSE) and the maximum absolute error (MAE), to evaluate the performance of the two methods for the estimation of the slope function $\beta(t,u)$ for the $i$th repetition of simulation, $i = 1, \ldots, 100$,

$$\text{MSE} = \frac{1}{n_A n_T} \sum_{t \in T} \sum_{u_r \in A} \{\hat{\beta}^{(i)}(t, u_r) - \beta(t, u_r)\}^2, \tag{2.1}$$

$$\text{MAE} = \max_{t \in T, u_r \in A} |\hat{\beta}^{(i)}(t, u_r) - \beta(t, u_r)|. \tag{2.2}$$

MSE measures the average deviation between the estimator $\hat{\beta}(t,u)$ and the true $\beta(t,u)$, and MAE measures the maximum deviation between them. The simulation results are summarized in Table 2.1.

Under Scenario I, the true slope function $\beta(t,u)$ is a univariate function of time $t$, and it does not change with the quantile $u$. We regard Scenario I as the simple case. Under Scenario II, the true slope function $\beta(t,u)$ changes with both $t$ and $u$, and the error $\epsilon$ is Normally distributed. Under Scenario

III, the true slope function $\beta(t,u)$ also changes with both $t$ and $u$. But different from Scenario II, the error $\epsilon$ has a asymmetric distribution. We regard both Scenario II and 3 as the complex cases. Table 2.1 shows that for the simple case, the proposed method and the conventional method [41] have similar performance on the estimation of $\beta(t,u)$. For the complex cases, the performance of the proposed method is much better than the conventional method [41] . For instance, when the sample size $n = 300$, our method has reduced the average MSE by 36.5% in comparison with the conventional method [41]. The improvement in MAE is less significant than in MSE, but our method still brings down the average MAE by 10.6%. When the sample size $n$ increases to 500, the improvement gap of the proposed method is decreasing in comparison with the conventional method [41] . Nevertheless, the MSE of our estimate is still 31.6% less than the estimate obtained from the conventional method. This empirical result is supported by our intuition. For the simple case, the functional covariate $X(t)$ has the same effect on all the quantiles of the response variable $Y$, and therefore, incorporating multiple quantiles into the estimation procedure may not have any advantage compared with the conventional method [41]. While for the complex cases, the effect of $X(t)$ on different quantiles of $Y$ is different, and combining the strength of multiple quantiles should provide a better estimation.

Fig.2.2, Fig.2.3 and Fig.2.4 present the boxplots of the mean squared error (MSE) and the maximum absolute error (MAE) for the estimation of $\beta(t,u)$ under the setting that the true slope function $\beta(t,u)$ is a univariate function of $t$ and does not change with $u$ for sample size $n = 300, 400$ and 500. Fig.2.5, Fig.2.6, Fig.2.7 Fig.2.8, Fig.2.9, and Fig.2.10 present the boxplots of the mean squared error (MSE) and the maximum absolute error (MAE) for the estimation of $\beta(t,u)$ under the settings that the true slope function $\beta(t,u)$ changes with both $u$ and $t$ based on sample size $n = 300, 400$ and $500$.

Based on Fig.2.2-Fig.2.14, for MSE, the advantage of our method is significant compared with the conventional method. The proposed method in general has a smaller variance and a smaller median in comparison with the estimator derived from the conventional method. In addition, we can also see that for MSE, the advantage of the proposed method is more significant when the true $\beta(t,u)$ changes with both $t$ and $u$ compared with the situations when the true $\beta(t,u)$ only changes with $t$. But for MAE, the proposed estimator for $\beta(t,u)$ usually has a larger variation than the estimator obtained from the conventional method.

To compare the influence of truncation level $m$ on the estimation of $\beta(t,u)$, we carry out additional simulations based on 100 repetitions under Scenario II and 3 with the sample size $n = 500$. The estimations are obtained with different truncation levels, $m = 3$ and $m = 4$. The MSE of the estimations are calculated based on 17 quantiles uniformly distributed on $[0.1, 0.9]$. The results are summarized in Table 2.2.

Table 2.2: The average of the mean squared error (MSE) defined in (2.1) for the estimations of the slope function $\beta(t, u)$ by using the proposed simultaneously functional quantile regression (SFQR) method of two truncation levels $m = 3, 4$ under two scenarios with the sample size $n = 500$ in 100 simulation repetitions.

| Truncation Level m | 3 | 4 |
|---|---|---|
| Scenario II | 0.181 | 0.259 |
| Scenario III | 0.172 | 0.285 |

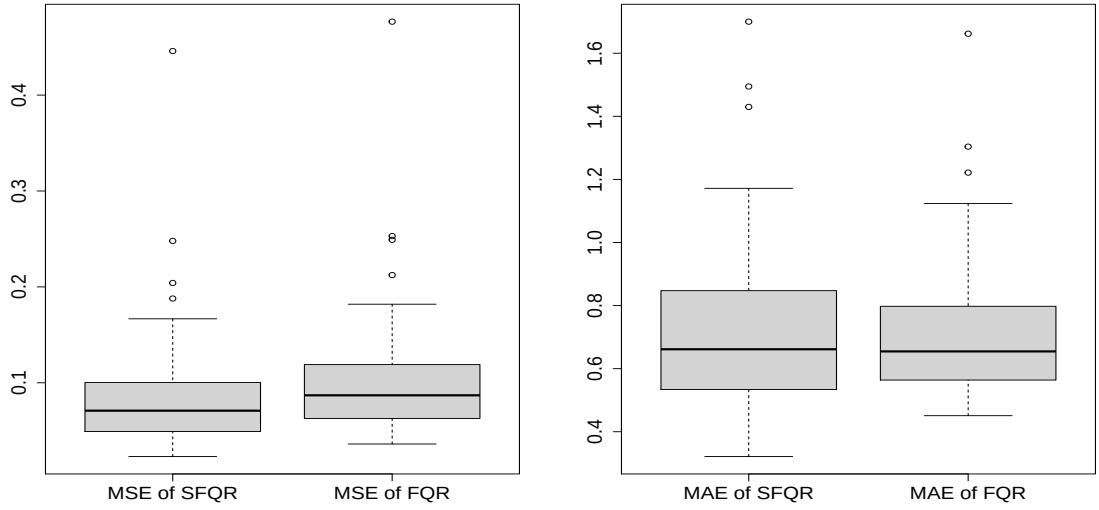



Figure 2.2: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario I with sample size 300.

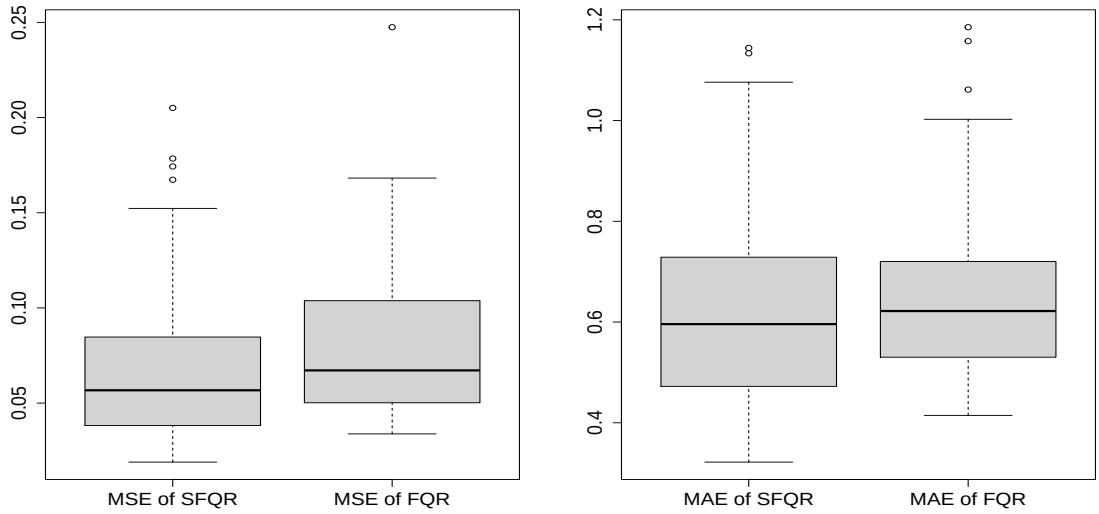



Figure 2.3: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario I with sample size 400.

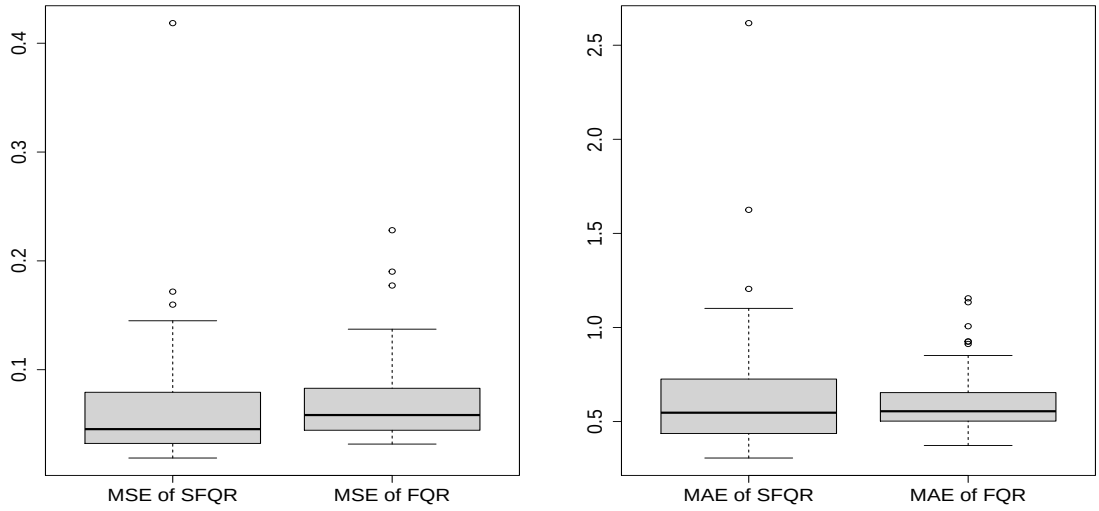



Figure 2.4: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario I with sample size 500.

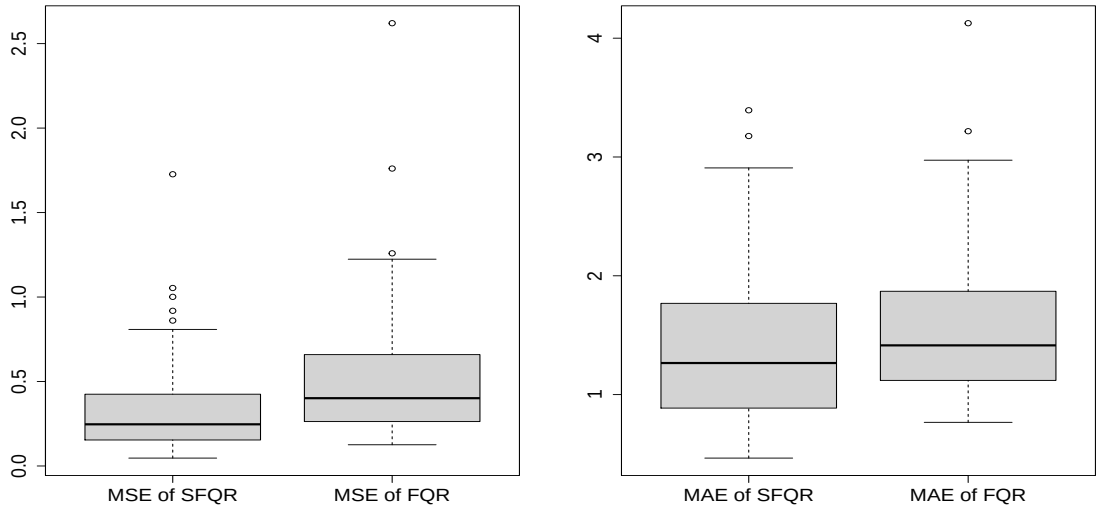



Figure 2.5: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario II with sample size 300.

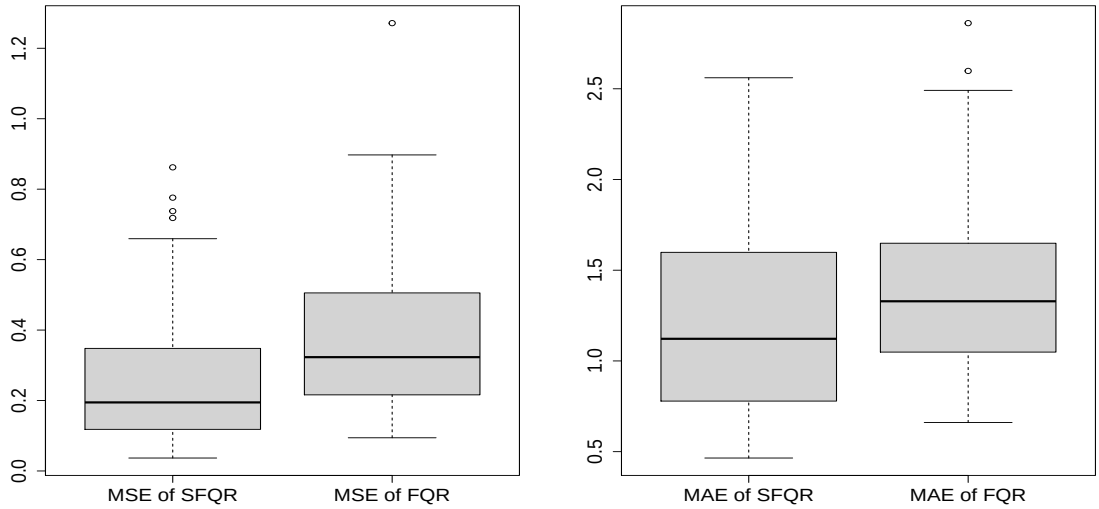



Figure 2.6: Boxplots of the mean squared error (MSE) defined in (2.1) and the maximum absolute error (MAE) defined in (2.2) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario II with sample size 400.

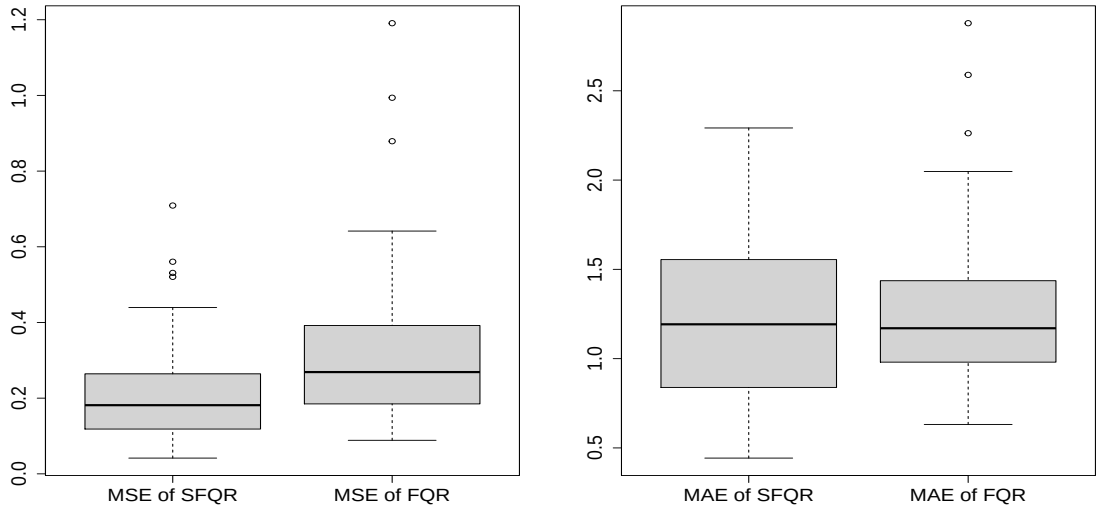



Figure 2.7: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario II with sample size 500.

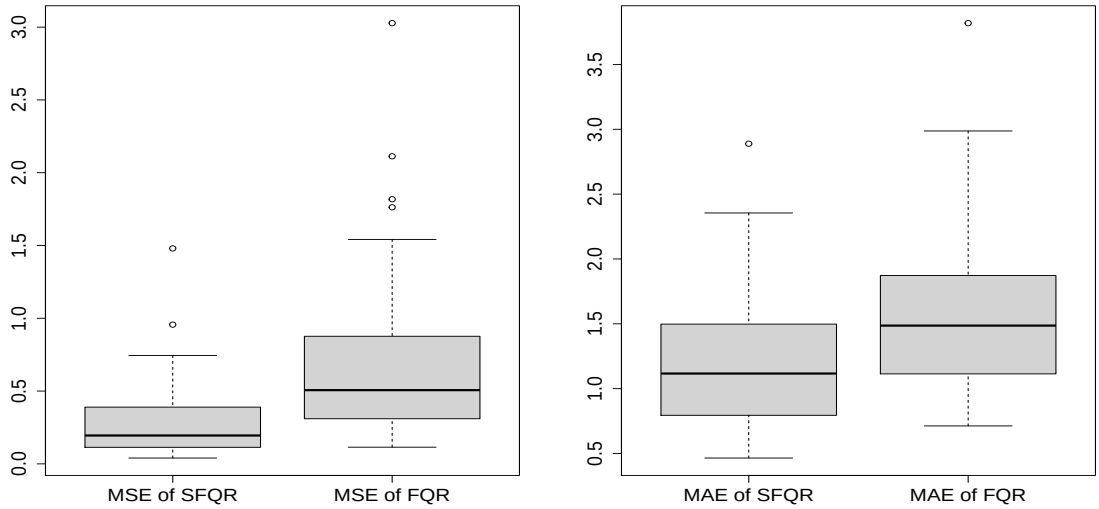



Figure 2.8: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario III with sample size 300.

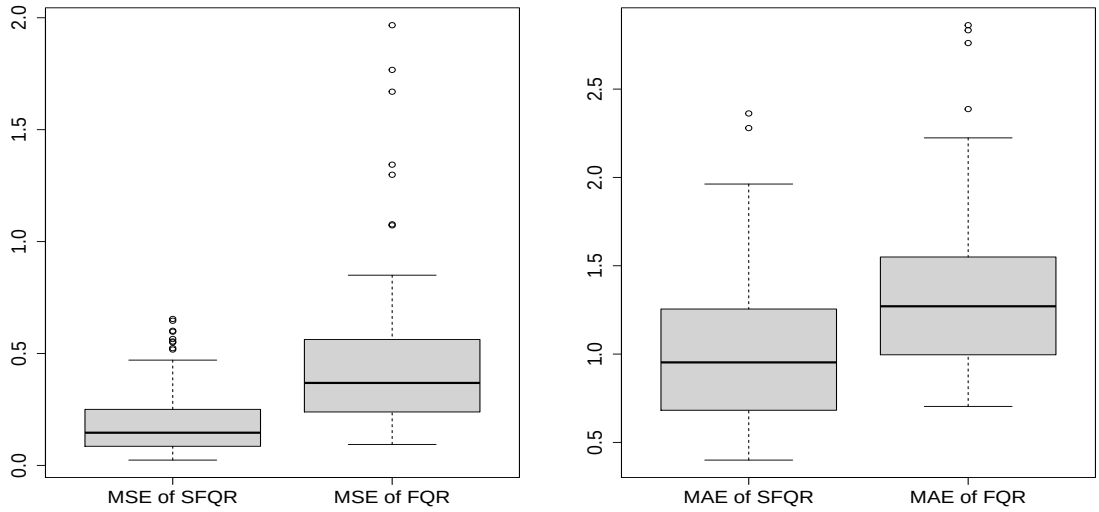



Figure 2.9: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario III with sample size 400.

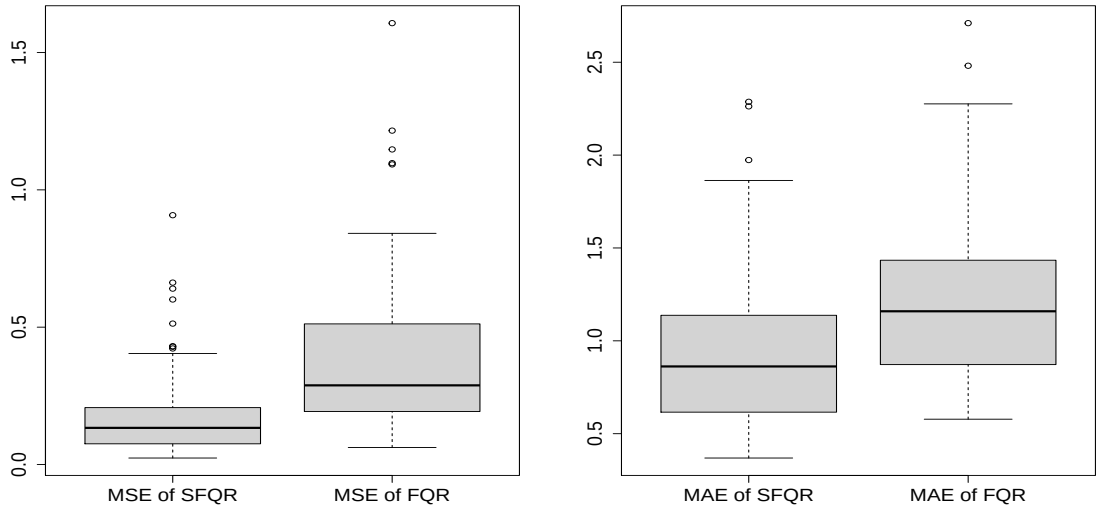



Figure 2.10: Boxplots of the average of mean squared error (MSE) and the average of maximum absolute error (MAE) for the estimation of $\beta(t, u)$ using the proposed simultaneous functional quantile regression (SFQR) and the conventional functional quantile regression (FQR) under Scenario III with sample size 500.

## 2.5. Real Data Analysis

### 2.5.1 The Capital Bike Share Program

The Urban population is growing rapidly in recent years. Meanwhile, air pollution, greenhouse gas emissions, and other environmental problems are getting worse and worse as an increasing number of people need to drive to work. As an alternative to driving to work, especially in big cities, which are facing traffic, environmental and health issues, biking is a healthy and eco-friendly way.

Instead of owing bikes, renting bikes is considered a more economical and environmental-friendly alternative. Nowadays, bike-sharing systems have become an essential part of urban mobility in many major cities, and the number of cities that are becoming bike-friendly is increasing day by day.

As an outdoor activity, customers' rental behaviors are affected by weather conditions. A successful bike business needs to have a good strategy to adjust the supply of available bikes to meet the demands based on the weather conditions. Therefore, it is of great interest to quantify the weather condition's effect on the bike rental. Weather conditions can be measured using a wide range of factors. In this article, we investigate the relationship between the total daily number of bike rentals and the hourly temperature.

The data set we use in this study is from the Capital Bike Share study [21], which contains rentals to cyclists without membership in the Capital Bike Share program in Washington D.C. from January 1st, 2011 to December 31th 2012. The hourly counts of casual bike rentals every day, the weather conditions, and the hourly temperature measurements are all recorded in the data set. The demands of bike rentals are quite different between weekdays and weekends. In the study, we restrict our analysis to the data observed on the weekends. Specifically, we only consider the temperature measurements and the counts of bike rentals obtained between 7:00 and 17:00 on Saturdays and Sundays without raining or snowing. The goal of our analysis is to investigate how the hourly temperature affects the lower, middle, and upper quantiles of the daily total bike rentals during the weekends.

Figure 2.11 shows the estimated slope function $\hat{\beta}(t, u)$ for $u = 10\%, 20\%, 50\%$, and $90\%$. In the top two panels of Figure 2.11, the slope function is negative in the early morning and becomes positive in the noon and afternoon, which are the peak demand periods for bike rentals. Since the temperature in the early morning is usually much cooler than the temperature in the noon and afternoon, the cumulative effect of temperature on the bike rental is positive. It indicates that as long as the overall temperature of that day is not too low, then the lower bounds of the bike rental will usually not be bad because the 10% and 20% quantiles measure the worst situations for the bike rental demand. The result on the 50% quantile displayed in the bottom left panel in Figure 2.11 represents the normal situation. It shares a similar pattern with the lower quantiles that the cool morning and the warm noon are preferred for a normal bike demand.

The bottom right panel in Figure 2.11 shows that when $u = 90\%$, the slope function is negative in the early morning before 9:00 and late afternoon after 15:30. This may be due to the fact that

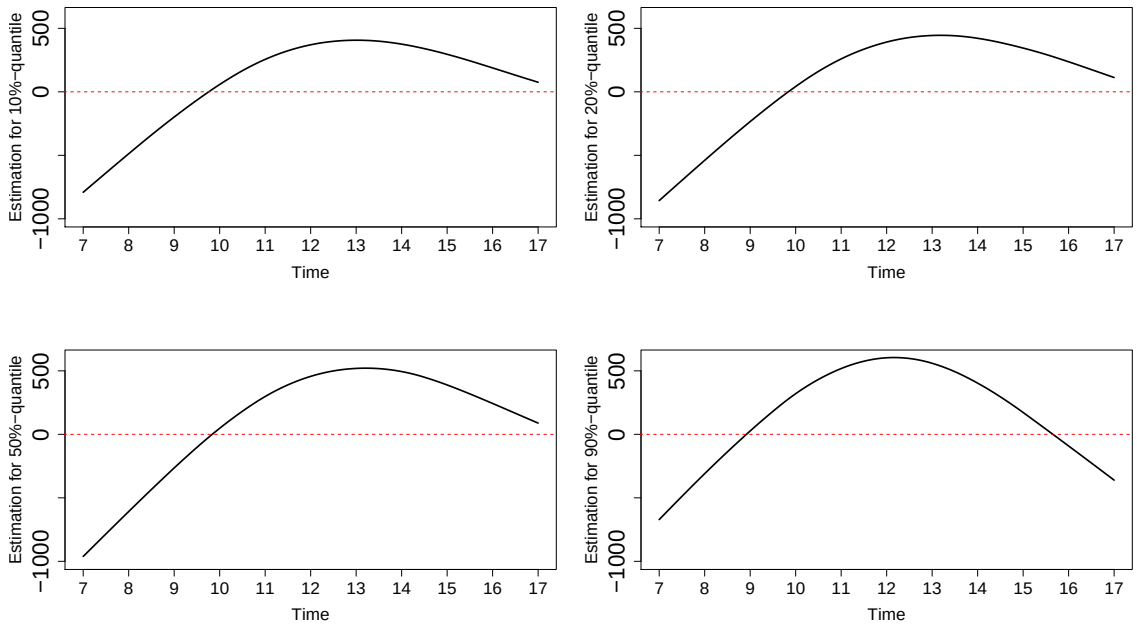



Figure 2.11: The estimated slope function $\hat{\beta}(t, u)$ for the regression model (2.1) at quantiles $u = 10\%, 20\%, 50\%, 90\%$ based on the data collected from Capital Bike Share program in Washington D.C. during 7:00 to 17:00 every weekend.

a high temperature in the morning deters unnecessary bike rental at noon and afternoon. If the temperature is high in the morning, then the temperature of the whole day is usually very high as well. In addition, the late afternoon is usually the hottest time of the day, and on some days, the late afternoon temperature can be too high for biking. On the other hand, a cool morning may indicate a comfortable biking temperature for the peak demand periods: the noon and the afternoon. Since 90% quantile almost measures the most ideal situation for the bike rental demand, this plot indicates that for a high bike rental demand, the weather needs to be cool in the morning, and comfortable or moderate in the afternoon.

To give an overall visualization of the estimated $\hat{\beta}(t, u)$, Figure 2.12(a) displays the heat map of $\hat{\beta}(t, u)$ estimated from the proposed method for the time $t$ from 7:00 to 17:00 and the quantile $u$ from 10% to 90%. The estimated slope function $\hat{\beta}(t, u)$ is positive after 9:00 for the quantiles $u$ from 10% to 60%, while it gradually becomes negative in the late afternoon for quantiles $u$ from 60% to 90%.

In comparison with the proposed method, Figure 2.12(b) shows the heat map of the estimation for $\beta(t, u)$ derived from the conventional method [41]. We can observe that this estimation is not smooth. In addition to that, the proposed method can overcome the issue of the monotonicity of the quantile estimates. Figure 2.13 shows the comparison of the estimated quantile functions of the $60^{\text{th}}$ and the $100^{\text{th}}$ subjects derived from the conventional method [41] and the proposed method. Let $Q^*_{60}(u)$ and $Q^*_{100}(u)$ be the estimated quantile functions of the $60^{\text{th}}$ and the $100^{\text{th}}$ subjects

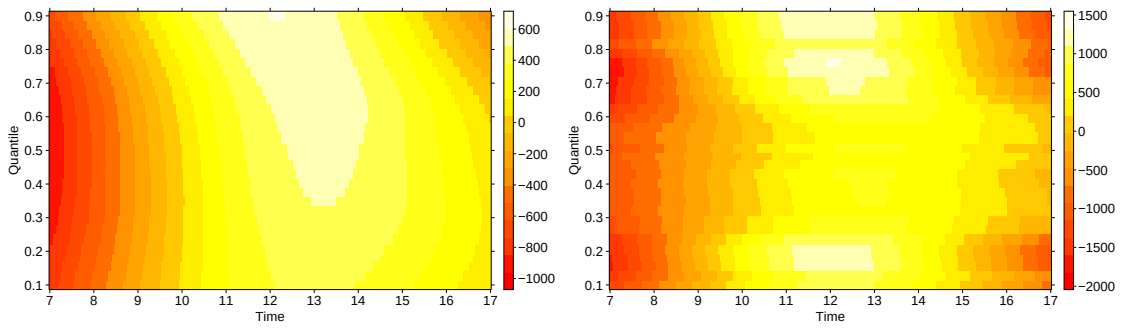



Figure 2.12: The heat maps of the estimated slope function $\hat{\beta}(t, u)$ for the regression model (2.1) derived from the proposed method (Panel (a)) and the conventional method [41] (Panel (b)) based on the data collected from Capital Bike Share program in Washington D.C. during 7:00 to 17:00 every weekend.

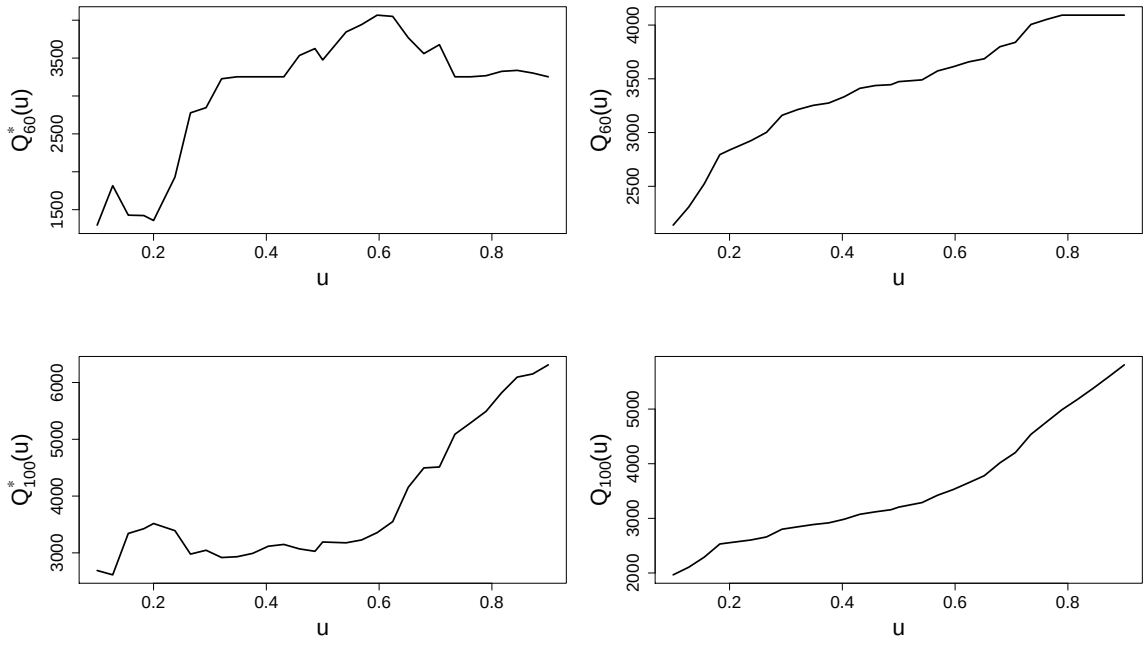



Figure 2.13: Estimated quantile functions of the 60th and the 100th subjects derived from the conventional method [41] (shown in the left two panels) and the proposed method (shown in the right two panels) based on the data collected from the Capital Bike Share program in Washington D.C. during 7:00 to 17:00 every weekend.

derived from the conventional method [41], and $Q_{60}(u)$ and $Q_{100}(u)$ be the corresponding estimated quantile functions derived from the proposed method. We can directly observe that $Q^*_{60}(u)$ and $Q^*_{100}(u)$ are not monotone over the interval $u \in [0.1, 0.9]$ as they should be, while $Q_{60}(u)$ and $Q_{100}(u)$ are both monotonically increasing in $u \in [0.1, 0.9]$.

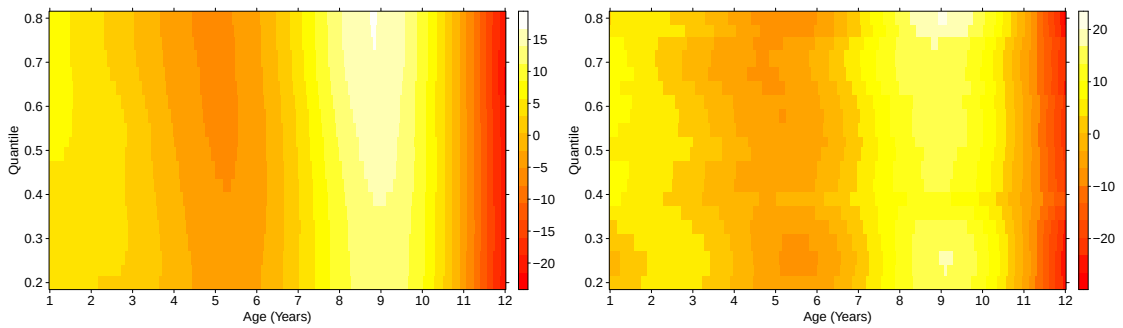



Figure 2.14: The heat maps of the estimated slope function $\hat{\beta}(t,u)$ for the regression model (2.1) derived from the proposed method (Panel (a)) and the conventional method [41] (Panel (b)) based on the Berkeley growth data for one-year-old to twelve-year-old.

### 2.5.2 Berkeley Growth Data

Child's height growth is an important health indicator, and abnormal growth usually implies an underlying health problem or growth disorder. It is thus helpful to understand the relationship between children's growth history and their adult height to evaluate the health and growth progress of children. If the predicted adult height of a child has an abnormally small lower quantile, then interventions should be considered during their teenage year to treat any potential health problem that affects height growth.

To investigate this relationship, we use the children's growth history between one-year-old and twelve-year-old as a functional covariate [11], and the conditional quantile of their eighteen-year-old heights as the response variable. We apply the proposed method to the Berkeley growth data [78] to estimate the slope function $\beta(t,u)$ from the model (2.1).

Figure 2.14(a) displays $\hat{\beta}(t,u)$ for $u \in [0.2, 0.8]$ and $t \in [1, 12]$. We can observe that the major variation of $\hat{\beta}(t,u)$ along the direction of $u$ (y-axis variable) occurs between one-year-old and six-year-old. For any fixed age $t \geq 7$, $\hat{\beta}(t,u)$ does not change too much as a function of $u$.

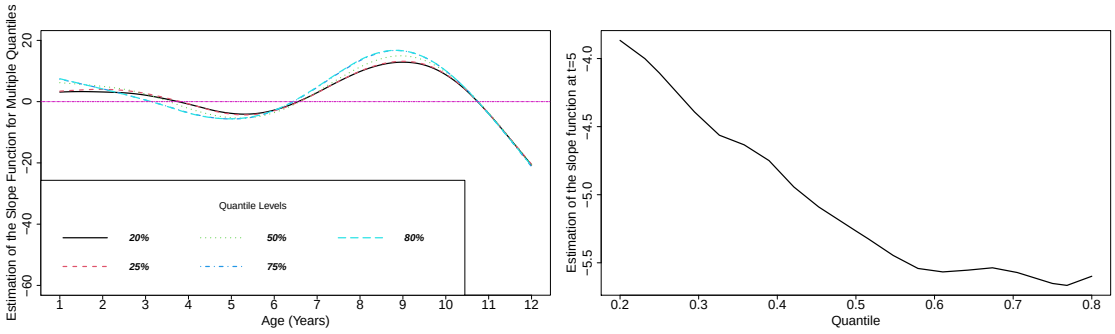



Figure 2.15: The estimated slope function $\hat{\beta}(t,u)$ for the regression model (2.1) at $u = 20\%, 25\%, 50\%, 75\%$, and $80\%$ over the age $t$ from one-year-old to twelve-year-old and the estimated slope function $\hat{\beta}(t,u)$ at age $t = 5$ for $u$ from 20% to 80% based on the Berkeley growth data.

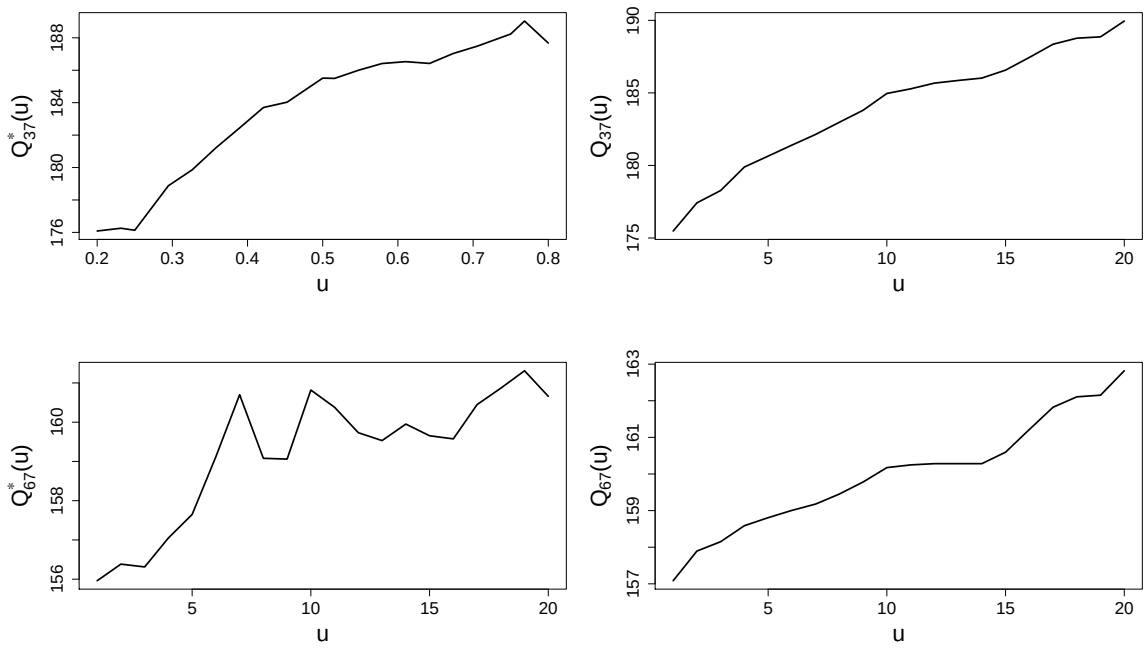



Figure 2.16: Estimated quantile functions of the 37th and the 67th subjects derived from the conventional method [41] (shown in the left two panels) and the proposed method (shown in the right two panels) based on the Berkeley growth data for one-year-old to twelve-year-old.

More specifically, Figure 2.15(a) displays $\beta(t, u)$ as a function of $t$ for $u = 20\%, 25\%, 50\%, 75\%$, and $80\%$. It shows that the children's growth history between age seven years and eleven years is always positively correlated with the quantiles of their adult height. This interval may be regarded as a growth spurt. If one child has a significantly lower height compared to the normal level during the growth spurt period, then some intervention should be considered.

Figure 2.15(b) shows the estimated slope function $\hat{\beta}(t, u)$ as a function of $u$ from $0.2$ to $0.8$ when $t = 5$, which is a negative function for any $u \in [0.2, 0.8]$. It indicates that the early growth spurt is not always a good sign for children's adult height. The early spurt may decrease children's potential to have a higher adult height due to the sex hormone levels in their bodies [74]. These children grow taller than other kids when they are young. However, their skeletons mature more rapidly. Consequently, they may stop growing at an early age, and eventually, end up having average or below average height as adults.

Similar to the previous application, Figure 2.14 and Figure 2.16 also show the comparison of the performance of the proposed method and the conventional method. In Figure 2.16, $Q^*_{37}(u)$ and $Q^*_{67}(u)$ are defined in the same way as the previous application. Clearly, the quantile estimations obtained from the conventional method [41] are not monotone over the interval $u \in [0.2, 0.8]$, while the proposed method can guarantee the desired monotonicity.

## 2.6. Conclusions and Discussion

We propose a novel framework for the simultaneous functional quantile regression to overcome the two major limitations of the conventional methods. When the true slope function is not a univariate function of time index, our framework can provide a better estimation for the slope function compared with the conventional estimation strategy that estimates the slope function as the univariate function by first fixing the quantile index. This advantage of the proposed method is examined by simulation studies in comparison with the method of [41]. In addition, the proposed framework can solve the two major limitations of the conventional methods. Within the proposed framework, the estimated conditional quantile functions are guaranteed to be monotone and their smoothness can be controlled.

In the current model (2.1), we only consider a single functional covariate. It may not be flexible enough to capture all of the information from data. In [11], they proposed a generalized version of the model (2.1) by using the composition of some link function and the linear functional of the functional covariate. In practice, it is common that several accompanying scalar covariates are observed along with the functional covariate. For this reason, in [77], a linear combination of scalar covariates is included in the model. Moreover, we often observe multiple functional covariates simultaneously in applications. To take multiple functional covariates into account, [54] further extended the model to incorporate a linear combination of multiple functional covariates with different slope functions.

Although we present our method based on the model (2.1), our proposed method can be further extended to different settings of the functional quantile regression model, such as sparse functional observations [84, 9]. Therefore, for the future work, we will extend our framework to the model that contains multivariate functional covariates and finite dimensional covariates. We will also investigate the properties and performance of our method in the scenario of sparse functional observations.

## 2.7. Proofs in Section 2.3

We use $\|\cdot\|_{l^2}$ to denote the $l^2$-norm of a vector. We assume that A1-A5 given in the previous chapter are satisfied.

**Lemma 1.** *[16] Let $l$ and $d$ be integers with $0 \leq l \leq d$. For a given set of knots $T$ on $[a, b]$. a B-spline basis of degree $d$ is defined based on the knots $T$, denoted as a vector $\boldsymbol{b}_0^{\mathrm{T}}(x)$. Then for any function $g(x) \in \boldsymbol{W}_q^{l+1}([a, b])$ and for any $1 \leq q \leq \infty$, any $0 \leq r \leq l$, there exists a coefficient vector $\boldsymbol{\gamma}_0$ such that*

$$\|D^r(\boldsymbol{b}_0^{\mathrm{T}}(x)\boldsymbol{\gamma}_0 - g(x))\|_{L^q([a,b])} \leq Ch^{l+1-r} \left\|D^{l+1}g\right\|_{L^q([a,b])}.$$

*where $C$ is a constant and $h$ is the largest length of a knot interval in $T$.*

**Lemma 2.** *[46] Let $\{B(t, u)_j\}_{j=1}^J$ be the Bernstein polynomials basis defined on the spline space $S_d^r(\Delta)$ over a $\pi$-quasi-uniform triangulation $\Delta$. Then there exist positive constants $c_1$ and $c_2$, which*

*only depend on $d$, $r$ and shape parameter $\pi$, such that*

$$c_1|\Delta|^2\|\boldsymbol{\gamma}\|_{l^2}^2 \le \|\boldsymbol{B}(t,u)^{\mathrm{T}}\boldsymbol{\gamma}\|_2^2 \le c_2|\Delta|^2\|\boldsymbol{\gamma}\|_{l^2}^2 .$$

Given any domain $\boldsymbol{\Omega} \in \mathbb{R}^2$, for any $1 \le q \le \infty$, we define the $q-$norm of any function $f \in \boldsymbol{\Omega}$ by

$$\|f\|_{q,\boldsymbol{\Omega}} = \begin{cases} \left(\int_{\boldsymbol{\Omega}} |f(u)|^q du\right)^{1/q}, & 1 \le q < \infty \\ \text{ess sup}_{u\in\boldsymbol{\Omega}}|f(u)|, & q = \infty, \end{cases}$$

**Lemma 3.** *[46] Suppose $\Delta$ is a quasi-uniform triangulation over a polygonal domain $\boldsymbol{\Omega}$, and $f(x,y) \in W_q^{d+1}(\boldsymbol{\Omega})$, where $W_q^{d+1}(\boldsymbol{\Omega})$ is a Sobolev space defined on $\boldsymbol{\Omega}$. Suppose $d \ge 3r+2$. Then, for any integers $a_1$, $a_2$, $0 \le a_1 + a_2 \le d$, there exists $f^*(x,y) \in S_d^r(\Delta)$ such that*

$$\left\|\nabla_x^{a_1}\nabla_y^{a_2}(f^* - f)\right\|_{q,\boldsymbol{\Omega}} \le C|\Delta|^{d+1-a_1-a_2}|g|_{d+1,q,\boldsymbol{\Omega}},$$

*for some constant $C$, which only depends on $d$, $r$ and $\pi$, the shape parameter of the triangulation $\Delta$.*

Define $L_n(\boldsymbol{\gamma}_0,\boldsymbol{\theta}) = \frac{1}{n_A}\sum_{r=1}^{n_A} E_n\left\{\rho_{u_r}\left(v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta})\right)\right\} + \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta}$, where $v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta}) = Y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}$. Suppose $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$ is a solution of

$$\min_{\boldsymbol{\gamma}_0,\boldsymbol{\theta}} L_n(\boldsymbol{\gamma}_0,\boldsymbol{\theta}), \tag{2.1}$$

and define $\hat{\boldsymbol{\gamma}} = \boldsymbol{Q}\hat{\boldsymbol{\theta}}$ and $\hat{\beta}(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\hat{\boldsymbol{\theta}}$.

For convenience, we first define several notations. Let $\beta_k(u) = \int \beta(t,u)\phi_k(t)dt$ for $k = 1,\ldots,m$. For $i = 1,\ldots,n$ and $k = 1,\ldots,m$, let $\eta_{i,k} = \kappa_k^{-1/2}\xi_{i,k}$, $\hat{\eta}_{i,k} = \kappa_k^{-1/2}\hat{\xi}_{i,k}$, $d_k(u) = \kappa_k^{1/2}\beta_k(u)$, $\hat{d}_k(u) = \kappa_k^{1/2}\sum_{j=1}^J \hat{\gamma}_j\hat{p}_{k,j}(u)$. where $\hat{\gamma}_j$ is the $j$th entry of $\hat{\boldsymbol{\gamma}}$. For $k = 0$, let $\hat{\eta}_{i,0} = \eta_{i,0} = 1$ and $d_0(u) = c_0(u)$ as well as $\hat{d}_0(u) = \boldsymbol{b}_0^{\mathrm{T}}(u)\hat{\boldsymbol{\gamma}}_0$.

Then for any $i = 1,\ldots,n$, define

$$\begin{aligned}
&\boldsymbol{d}^m(u) = (d_1(u),\ldots,d_m(u))^{\mathrm{T}},\\
&\boldsymbol{d}^{m+1}(u) = (d_0(u),d_1(u),\ldots,d_m(u))^{\mathrm{T}},\\
&\hat{\boldsymbol{d}}^m(u) = (\hat{d}_1(u),\ldots,\hat{d}_m(u))^{\mathrm{T}},\\
&\hat{\boldsymbol{d}}^{m+1}(u) = (\hat{d}_0(u),\hat{d}_1(u),\ldots,\hat{d}_m(u))^{\mathrm{T}},\\
&\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i,1},\ldots,\hat{\xi}_{i,m})^{\mathrm{T}},\\
&\hat{\boldsymbol{\eta}}_i^m = (\hat{\eta}_{i,1},\ldots,\hat{\eta}_{i,m})^{\mathrm{T}},\\
&\hat{\boldsymbol{\eta}}_i^{m+1} = (\hat{\eta}_{i,0},\hat{\eta}_{i,1},\ldots,\hat{\eta}_{i,m})^{\mathrm{T}} = (1,\hat{\eta}_{i,1},\ldots,\hat{\eta}_{i,m})^{\mathrm{T}}.
\end{aligned}$$

For a sequence $\{a_i\}_{i=1}^n$, define $E_n\{a_i\} = \frac{1}{n}\sum_{i=1}^n a_i$.

**Lemma 4.** *[41] Let* $\mathbb{S}^m = \{\boldsymbol{h} \in \mathbb{R}^{m+1} | \, \|\boldsymbol{h}\|_{l^2} = 1\}$. *Then,*

$$-E_n[\{u - \mathbb{1}(Y_i \leq \hat{\boldsymbol{\eta}}_i^{m+1} \cdot (\boldsymbol{d}^{m+1}(u) + M\sqrt{m/n}\boldsymbol{h}))\}(\hat{\boldsymbol{\eta}}_i^{m+1} \cdot \boldsymbol{h})] \\ \geq c_1 M(1 - o_p(1))\sqrt{m/n} - O_p(\sqrt{m/n}) - M^2 o_p(\sqrt{m/n}),$$

*where* $c_1$ *and* $M$ *are some positive constants, as well as the stochastic orders are evaluated uniformly for any* $u \in (0, 1)$ *and any* $\boldsymbol{h} \in \mathbb{S}^m$

**Lemma 5.** *Let* $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$ *be a solution of* (2.1)*,* $\lambda_{1,n} \asymp n_A^{-1} n_T^{-1} m^{-1/2} n |\Delta|^{d+1}$*, and* $\lambda_{2,n} = o(\lambda_{1,n} n_A n_T |\Delta|^4)$. *The estimator* $\hat{\beta}(t, u)$ *can be expressed as*

$$\begin{aligned}\hat{\beta}(t, u) &= \sum_{k=1}^{m} \kappa_k^{-1/2} \hat{d}_k(u) \hat{\phi}_k(t) + \sum_{k=m+1}^{\infty} \kappa_k^{-1/2} \hat{d}_k(u) \hat{\phi}_j(t) \\ &= \sum_{k=1}^{m} \kappa_k^{-1/2} \hat{d}_k(u) \hat{\phi}_k(t) + \hat{\beta}_{re}(t, u),\end{aligned}$$

*where* $\kappa_k^{-1/2} \hat{d}_k(u) = \int \hat{\beta}(t, u) \hat{\phi}_k(t) dt$ *and* $\hat{\beta}_{re}(t, u) = \sum_{k=m+1}^{\infty} \kappa_k^{-1/2} \hat{d}_k(u) \hat{\phi}_j(t)$. *Then, we have*

$$\left\| \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} \hat{\boldsymbol{\theta}} \right\|_{l^2} \leq \lambda_{1,n}^{-1} (1 - o(1)) O_p(mn^{-\frac{1}{2}} \vee m|\Delta|^{d+1}),$$

*and*

$$\hat{\beta}_{re}(t, u) \approx O_p(m^{1/2} n^{-1/2}).$$

*Proof*: Under the condition A5, and by Lemma 1 and Lemma 3, there exist $\boldsymbol{\gamma}_0^*$ and $\boldsymbol{\theta}^*$ such that

$$\sup_{(t,u) \in \mathscr{T} \times \mathscr{A}} |\beta(t, u) - \boldsymbol{B}^{\mathrm{T}}(t, u) \boldsymbol{Q} \boldsymbol{\theta}^*| \leq C_1 |\Delta|^{d+1}, \\ \sup_{u \in \mathscr{A}} |c(u) - \boldsymbol{b}_0^{\mathrm{T}}(u) \boldsymbol{\gamma}_0^*| \leq C_2 |\Delta|^{d+1},$$

for some constant $C_1$ and $C_2$. Since $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$ is the minimizer of (2.1), then

$$L_n(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}) \leq L_n(\hat{\boldsymbol{\gamma}}_0, \boldsymbol{\theta}^*). \tag{2.2}$$

For any $u_r \in A$, let

$$(\tilde{d}_0(u_r), \tilde{\boldsymbol{d}}^m(u_r)) = \mathrm{argmin}_{f_0, f_1, \ldots, f_m} E_n \left\{ \rho_{u_r} \left( Y_i - f_0 - \hat{\boldsymbol{\eta}}_i^m \cdot \boldsymbol{f}^m \right) \right\},$$

where $\boldsymbol{f}^m = (f_1, \ldots, f_m)^{\mathrm{T}}$. From [41], we know that

$$\sup_{u \in \mathscr{A}} \left\{ (d_0(u) - c_0(u))^2 + \left\| \tilde{\boldsymbol{d}}^m(u) - \boldsymbol{d}^m(u) \right\|_{l^2}^2 \right\} = O_p(mn^{-1}).$$

Therefore,

$$\begin{aligned}
&\left|n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\tilde{d}_0(u_r)-\hat{\boldsymbol{\eta}}_i^m\cdot\tilde{\boldsymbol{d}}^m\right)\right\}-n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}^*\right)\right\}\right| \\
&\quad\leq\{n_An\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\Bigg\{\left|\tilde{d}_0(u_r)+\hat{\boldsymbol{\eta}}_i^m\cdot\tilde{\boldsymbol{d}}^m(u_r)-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\boldsymbol{\theta}^*\right| \\
&\qquad+|2u_r-1|\left|\tilde{d}_0(u_r)+\hat{\boldsymbol{\eta}}_i^m\cdot\tilde{\boldsymbol{d}}^m(u_r)-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\boldsymbol{\theta}^*\right|\Bigg\}, \\
&\quad\leq 2n_A^{-1}n^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\left\{\left|\hat{\boldsymbol{\eta}}_i^m\cdot\tilde{\boldsymbol{d}}^m(u_r)-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\boldsymbol{\theta}^*\right|+\left|\tilde{d}_0(u_r)-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*\right|\right\}, \\
&\quad\leq 2n_A^{-1}n^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\left\{\|\hat{\boldsymbol{\eta}}_i^m\|_{l^2}\left\|\tilde{\boldsymbol{d}}^m(u_r)-\boldsymbol{\kappa}\hat{\boldsymbol{P}}(u_r)Q\boldsymbol{\theta}^*\right\|_{l^2}+\left|\tilde{d}_0(u_r)-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*\right|\right\}, \\
&\quad\leq O_p(m^{1/2})\times O_p(m^{1/2}n^{-1/2}\vee|\Delta|^{d+1})=O_p(mn^{-1/2}\vee m^{1/2}|\Delta|^{d+1}),
\end{aligned}\tag{2.3}$$

where $\boldsymbol{\kappa}$ is a $m$ by $m$ diagonal matrix with the diagonal entries being $(\kappa_1^{1/2},\ldots,\kappa_m^{1/2})$. In addition, by definition of $(\tilde{d}_0(u_r),\tilde{\boldsymbol{d}}^m(u_r))$,

$$\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\tilde{d}_0(u_r)-\hat{\boldsymbol{\eta}}_i^m\cdot\tilde{\boldsymbol{d}}^m(u_r)\right)\right\}\leq\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\hat{\boldsymbol{\theta}})\right)\right\}.\tag{2.4}$$

Since

$$\begin{aligned}
&n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\hat{\boldsymbol{\theta}})\right)\right\} \\
&\qquad<n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}^*)\right)\right\},
\end{aligned}$$

then by (2.3) and (2.4),

$$\begin{aligned}
&\left|n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\hat{\boldsymbol{\theta}})\right)\right\}\right. \\
&\quad\left.-n_A^{-1}\sum_{r=1}^{n_A}E_n\left\{\rho_{u_r}\left(Y_i-\boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0^*-\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}^*)\right)\right\}\right|\leq O_p(mn^{-1/2}\vee m^{1/2}|\Delta|^{d+1}).
\end{aligned}\tag{2.5}$$

By Lemma 2 and [82],

$$\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\leq C|\Delta|^{-2}\left\|\hat{\boldsymbol{\theta}}\right\|_{l^2}^2=C|\Delta|^{-4}\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2.$$

On the other hand,

$$\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\approx n_An_T\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2.$$

Since $\lambda_{2,n} = o(\lambda_{1,n} n_A n_T |\Delta|^4)$, when $n_A$ and $n_T$ are large enough,

$$\frac{\lambda_{1,n}^{-1}\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}} = o(1). \tag{2.6}$$

Combine (2.2), (2.5) and (2.6), we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}} &\le O_p(\lambda_{1,n}^{-1}mn^{-1/2} \vee \lambda_{1,n}^{-1}m^{1/2}|\Delta|^{d+1}) + \boldsymbol{\theta}^{*\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta}^*, \\ &= O_p(\lambda_{1,n}^{-1}mn^{-1/2} \vee \lambda_{1,n}^{-1}m^{1/2}|\Delta|^{d+1}) + O(n_A n_T),\end{aligned}$$

which implies $\left\|\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\right\|_{l^2} = O_p(\lambda_{1,n}^{-1/2}m^{1/2}n^{-1/4} \vee \lambda_{1,n}^{-1/2}m^{1/4}|\Delta|^{(d+1)/2} \vee n_A^{1/2}n_T^{1/2})$.

By Lemma 3, there exists a vector $\hat{\boldsymbol{\theta}}_m$ such that $\max_{t\in T,u\in A}|\hat{\beta}_m^*(t,u) - \sum_{k=1}^m \left(\int \hat{\beta}(t,u)\hat{\phi}_k(t)dt\right)\hat{\phi}_k(t)| \le C|\Delta|^{d+1}$, where $\hat{\beta}_m^*(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\hat{\boldsymbol{\theta}}_m$. Since $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}})$ is a minimizer of (2.1), then $L_n(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}) \le L_n(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\theta}}_m)$. That is,

$$\begin{aligned}&n^{-1}n_A^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}\right) + \lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}} \\ &+ \lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\hat{\boldsymbol{\theta}} \le n^{-1}n_A^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right) \\ &+ \lambda_{1,n}\hat{\boldsymbol{\theta}}_m^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}_m + \lambda_{2,n}\hat{\boldsymbol{\theta}}_m^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\hat{\boldsymbol{\theta}}_m.\end{aligned} \tag{2.7}$$

On the other hand,

$$\begin{aligned}&n^{-1}n_A^{-1}\Bigg|\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}\right) \\ &\quad - \sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right)\Bigg|, \\ &\quad = n^{-1}n_A^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\Bigg\{\left|\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right| \\ &\quad + (2u_r - 1)\left(\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right)\Bigg\}, \\ &\quad \le n^{-1}n_A^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\left\{2\left|\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right|\right\}, \\ &\quad \le n^{-1}n_A^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\left\{2\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}\left\|\hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{P}}(u_r)Q\hat{\boldsymbol{\theta}}_m\right\|_{l^2}\right\}, \\ &\quad = O_p(m^{1/2}|\Delta|^{d+1}),\end{aligned} \tag{2.8}$$

provided that $E_n\left\{\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}\right\} = O_p(m^{1/2})$ [41]. By Lemma 2, (2.7) and (2.8), we have

$$\begin{aligned}
&n_A^{-1}n_T^{-1}\left\{\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}_m\right\}\\
&\leq O_p(\lambda_{1,n}^{-1}n_A^{-1}n_T^{-1}m^{1/2}|\Delta|^{d+1}) + O_p(\lambda_{1,n}^{-1}\lambda_{2,n}n_A^{-1}n_T^{-1}|\Delta|^{-2}\left\|\hat{\boldsymbol{\theta}}_m\right\|_{l^2}^2),\\
&= O_p(\lambda_{1,n}^{-1}n_A^{-1}n_T^{-1}m^{1/2}|\Delta|^{d+1} \vee \lambda_{1,n}^{-1}\lambda_{2,n}n_A^{-1}n_T^{-1}|\Delta|^{-4}) = O_p(m/n). \qquad (2.9)
\end{aligned}$$

Then, by (2.9)

$$\begin{aligned}
&\left\|\hat{\beta}(t,u) - \sum_{k=1}^{m}\left(\int\hat{\beta}(t,u)\hat{\phi}_k(t)dt\right)\hat{\phi}_k(t)\right\|_{L^2(\boldsymbol{\Omega})}^2,\\
=&\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2 - \left\|\sum_{k=1}^{m}\left(\int\hat{\beta}(t,u)\hat{\phi}_k(t)dt\right)\hat{\phi}_k(t)\right\|_{L^2(\boldsymbol{\Omega})}^2\\
=&\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2 - \left\|\sum_{k=1}^{m}\kappa_k^{-1/2}\hat{d}_k(u)\hat{\phi}_k(t) - \hat{\beta}_m^*(t,u) + \hat{\beta}_m^*(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2,\\
\leq&\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2 - \left\|\hat{\beta}_m^*(t,u)\right\|_{L^2(\boldsymbol{\Omega})}^2 + O(|\Delta|^{d+1}),\\
\approx&n_A^{-1}n_T^{-1}\left\{\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}_m\right\} + O(|\Delta|^{d+1}),\\
=&O_p(m/n).
\end{aligned}$$

Therefore, we conclude that $\hat{\beta}_{re}(t,u) \approx O_p(m^{1/2}n^{-1/2})$.

**Lemma 6.** *Let ($\hat{\boldsymbol{\gamma}}_0$ ,$\hat{\boldsymbol{\theta}}$) be a solution of* (2.1)*. Then,*

$$\begin{aligned}
&\left\|n_A^{-1}n^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\left\{u_r - \mathbb{1}(Y_i < \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\hat{\boldsymbol{\theta}})\right\}\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\|_{l^2}\\
&\leq \frac{c\{m + (n_A \wedge C|\Delta|^{-1})\}}{n_A n}\max_{i\in 1,\ldots,n}\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}\max_{u_r\in A}\sigma_{\max}\{\hat{\boldsymbol{P}}(u_r)\}\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_{l^2}\\
&+ \left\|\left(2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q} + 2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\|_{l^2}, \qquad (2.10)
\end{aligned}$$

*where $\sigma_{\max}(x)$ is the largest singular value of a matrix $x$.*

*Proof*: The check function $\rho_u(x)$ is

$$\rho_u(x) = |x| + (2u-1)x,$$

then we can rewrite (2.1) as

$$\min_{\boldsymbol{\gamma}_0,\boldsymbol{\theta}} \frac{1}{n_A}\sum_{r=1}^{n_A} E_n\{|v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta})| + (2u_r-1)v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta})\} + \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta}. \tag{2.11}$$

To further remove the absolute value sign, the minimization problem (2.11) is equivalent to the following constrained optimization problem

$$\min_{\boldsymbol{\varphi},\boldsymbol{\gamma}_0,\boldsymbol{\theta}} g(\boldsymbol{\varphi},\boldsymbol{\gamma}_0,\boldsymbol{\theta}) = \frac{1}{n_A}\sum_{r=1}^{n_A} E_n\{\varphi_{i,r} + (2u_r-1)v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta})\} + \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta} \tag{2.12}$$

such that

$$\begin{aligned} -\varphi_{i,r} + v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta}) \le 0,\\ -\varphi_{i,r} - v_{i,r}(\boldsymbol{\gamma}_0,\boldsymbol{\theta}) \le 0. \end{aligned}$$

Suppose ($\hat{\boldsymbol{\varphi}}$, $\hat{\boldsymbol{\gamma}}_0$ ,$\hat{\boldsymbol{\theta}}$) is a solution of (2.12). Define three sets

$$\begin{aligned} I_1 &= \{(i,r) : v_{i,r}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}) > 0\},\\ I_2 &= \{(i,r) : v_{i,r}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}) < 0\},\\ I_3 &= \{(i,r) : v_{i,r}(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}}) = 0\}. \end{aligned}$$

The gradient of objective function $g$ is

$$\nabla g = \frac{1}{n}\begin{bmatrix} \mathbf{1} \\ \frac{1}{n_A}\sum_{r=1}^{n_A}\sum_{i=1}^{n}(2u_r-1)\nabla v_{i,r} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ 2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q} + 2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q} \end{bmatrix}$$

The active constraints are given by

$$\begin{bmatrix} -\boldsymbol{e}_{i,r} \\ \nabla v_{i,r} \end{bmatrix} \quad \forall (i,r) \in I_1,$$

$$\begin{bmatrix} -\boldsymbol{e}_{i,r} \\ -\nabla v_{i,r} \end{bmatrix} \quad \forall (i,r) \in I_2,$$

and

$$\begin{bmatrix} -\boldsymbol{e}_{i,r} \\ \nabla v_{i,r} \end{bmatrix}, \quad \begin{bmatrix} -\boldsymbol{e}_{i,r} \\ -\nabla v_{i,r} \end{bmatrix} \quad \forall (i,r) \in I_3,$$

where $\boldsymbol{e}_i$ is a $n_A n$-vector with a one in the position corresponding to the pair $(i, r)$ and zeros elsewhere.

By checking Kuhn-Tucker condition, we have the following necessary condition for the optimality [17],

$$\nabla g + \frac{1}{n} \sum_{(i,r)\in I_1} \delta_{i,r} \begin{bmatrix} -\boldsymbol{e}_{i,r} \\ \nabla v_{i,r} \end{bmatrix} + \frac{1}{n} \sum_{(i,r)\in I_2} \delta_{i,r} \begin{bmatrix} -\boldsymbol{e}_{i,r} \\ -\nabla v_{i,r} \end{bmatrix} + \frac{1}{n} \sum_{(i,r)\in I_3} \left\{ \delta_{i,r} \begin{bmatrix} -\boldsymbol{e}_{i,r} \\ \nabla v_{i,r} \end{bmatrix} + \nu_{i,r} \begin{bmatrix} -\boldsymbol{e}_{i,r} \\ -\nabla v_{i,r} \end{bmatrix} \right\} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{2.13}$$

where $\delta_{i,r} \geq 0$ and $\nu_{i,r} \geq 0$. Since the sets $I_1$, $I_2$ and $I_3$ are disjoint, then the top $nn_A$ equations of (2.13) imply that

$$\delta_{i,r} = 1 \quad \forall (i,r) \in I_1, \tag{2.14}$$

$$\delta_{i,r} = 1 \quad \forall (i,r) \in I_2, \tag{2.15}$$

$$\delta_{i,r} + \nu_{i,r} = 1 \quad \forall (i,r) \in I_3. \tag{2.16}$$

Now we consider the rest of equations in (2.13). By (2.14), (2.15) and (2.16), we can derive that

$$\frac{1}{n_A n} \sum_{(i,r)\in I_1 \cup I_2} \{\text{sign}(v_{i,r}) + 2u_r - 1\} \nabla v_{i,r} + \frac{1}{n_A n} \sum_{(i,r)\in I_3} \{\delta_{i,r} - \nu_{i,r} + 2u_r - 1\} \nabla v_{i,r} + \begin{bmatrix} \mathbf{0} \\ 2\lambda_{1,n} \hat{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{B}_{A,T} \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} + 2\lambda_{2,n} \hat{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{Q} \end{bmatrix} = \mathbf{0}, \tag{2.17}$$

where $\nabla v_{i,r}$ is defined as

$$\nabla v_{i,r} = \begin{bmatrix} -\boldsymbol{b}_0^{\mathrm{T}}(u_r) \\ -\hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q} \end{bmatrix}.$$

The equations (2.17) implies that

$$\frac{1}{n_A n} \sum_{(i,r)\in I_1 \cup I_2 \cup I_3} \{u_r - \mathbb{1}(Y_i < \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q} \hat{\boldsymbol{\theta}})\} \boldsymbol{b}_0^{\mathrm{T}}(u_r)$$
$$= \frac{1}{n_A n} \sum_{(i,r)\in I_3} \{1/2 - (\delta_{i,r} - \nu_{i,r})/2 - \mathbb{1}(Y_i < \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q} \hat{\boldsymbol{\theta}})\} \boldsymbol{b}_0^{\mathrm{T}}(u_r), \tag{2.18}$$

and

$$\frac{1}{n_A n} \sum_{(i,r)\in I_1 \cup I_2 \cup I_3} \{1 - u_r - \mathbb{1}(Y_i < \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q} \hat{\boldsymbol{\theta}})\} \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q}$$
$$= \frac{1}{n_A n} \sum_{(i,r)\in I_3} \{1/2 - (\delta_{i,r} - \nu_{i,r})/2 - \mathbb{1}(Y_i < \boldsymbol{b}_0^{\mathrm{T}}(u_r)\hat{\boldsymbol{\gamma}}_0 + \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q} \hat{\boldsymbol{\theta}})\} \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) \boldsymbol{Q}$$
$$- 2\lambda_{1,n} \hat{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{B}_{A,T} \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} - 2\lambda_{2,n} \hat{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{Q}. \tag{2.19}$$

Let $n_B$ denote the number of B-spline basis functions, and $n_{knots}$ denote the number of its interior knots. The relationship bewteen $n_B$, $n_{knots}$ and $d$ is

$$n_B = d + n_{knots}.$$

Since we only require the chosen B-spline basis has the same approximation power as the bivariate splines used for the approximation of $\beta(t,u)$, we can choose the interior knots uniformly distributed within the range of $u$ with $n_{knots} = C|\Delta|^{-1}$ for some constant $C$.

Define $\boldsymbol{Z}_i(u) = \left[\boldsymbol{b}_0^{\mathrm{T}}(u), \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u)\boldsymbol{Q}\right]$, and $\hat{\boldsymbol{\Gamma}} = \begin{bmatrix}\hat{\gamma}_0 \\ \hat{\boldsymbol{\theta}}\end{bmatrix}$. Recall that $I_3$ is defined as

$$I_3 = \{(i,r) : v_{i,r}(\hat{\gamma}_0, \hat{\boldsymbol{\theta}}) = 0\}.$$

Let

$$\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{Z}_1(u_1) \\ \boldsymbol{Z}_2(u_1) \\ \cdots \\ \boldsymbol{Z}_n(u_1) \\ \boldsymbol{Z}_1(u_2) \\ \cdots \\ \boldsymbol{Z}_n(u_r) \\ \cdots\end{bmatrix} = \begin{bmatrix}\boldsymbol{b}_0^{\mathrm{T}}(u_1), \hat{\boldsymbol{\xi}}_1^{\mathrm{T}}\hat{\boldsymbol{P}}(u_1)\boldsymbol{Q} \\ \boldsymbol{b}_0^{\mathrm{T}}(u_1), \hat{\boldsymbol{\xi}}_2^{\mathrm{T}}\hat{\boldsymbol{P}}(u_1)\boldsymbol{Q} \\ \cdots, \cdots \\ \boldsymbol{b}_0^{\mathrm{T}}(u_2), \hat{\boldsymbol{\xi}}_1^{\mathrm{T}}\hat{\boldsymbol{P}}(u_2)\boldsymbol{Q} \\ \cdots, \cdots \\ \boldsymbol{b}_0^{\mathrm{T}}(u_r), \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q} \\ \cdots, \cdots\end{bmatrix} = [\boldsymbol{M}_1, \boldsymbol{M}_2],$$

where $\boldsymbol{M}_1 = \begin{bmatrix}\boldsymbol{b}_0^{\mathrm{T}}(u_1) \\ \boldsymbol{b}_0^{\mathrm{T}}(u_1) \\ \cdots \\ \boldsymbol{b}_0^{\mathrm{T}}(u_2) \\ \cdots \\ \boldsymbol{b}_0^{\mathrm{T}}(u_r) \\ \cdots\end{bmatrix}$ and $\boldsymbol{M}_2 = \begin{bmatrix}\hat{\boldsymbol{\xi}}_1^{\mathrm{T}}\hat{\boldsymbol{P}}(u_1)\boldsymbol{Q} \\ \hat{\boldsymbol{\xi}}_2^{\mathrm{T}}\hat{\boldsymbol{P}}(u_1)\boldsymbol{Q} \\ \cdots \\ \hat{\boldsymbol{\xi}}_1^{\mathrm{T}}\hat{\boldsymbol{P}}(u_2)\boldsymbol{Q} \\ \cdots \\ \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q} \\ \cdots\end{bmatrix}$. The matrix $\boldsymbol{M}_1$ is a $n_A n$ by $n_{knots}$ matrix. It is easy to see that rank$(\boldsymbol{M}_1) \leq (n_A \wedge C|\Delta|^{-1})$ and rank$(\boldsymbol{M}_2) \leq m$. Therefore, rank$(\boldsymbol{Z}) \leq m + (n_A \wedge C|\Delta|^{-1})$. Then by Sard's theorem, almost surely, Card$(I_3) \leq m + (n_A \wedge C|\Delta|^{-1})$.

Combined with (2.18) and (2.19), we have almost surely,

$$
\begin{aligned}
&E_n\left[\frac{1}{n_A}\sum_{r=1}^{n_A}\left\{\left(u_r-\mathbb{1}\left\{Y_i-\boldsymbol{Z}_i(u_r)\hat{\boldsymbol{\Gamma}}\le 0\right\}\right)\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\right\}\right]\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\\
=&\frac{c}{n_A n}\sum_{(i,r)\in I_3}\left(u_r-\mathbb{1}\left\{Y_i-\boldsymbol{Z}_{\boldsymbol{i}}(u_r)\hat{\boldsymbol{\Gamma}}\le 0\right\}\right)\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right),\\
\le&\frac{c\{m+(n_A\wedge C|\Delta|^{-1})\}}{n_A n}\max_{i,r}\left|\left(u_r-\mathbb{1}\left\{Y_i-\boldsymbol{Z}_i(u_r)\hat{\boldsymbol{\Gamma}}\le 0\right\}\right)\hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\right|\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_{l^2}\\
&+\left\|\left(2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}+2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_{l^2},\\
\le&\frac{c\{m+(n_A\wedge C|\Delta|^{-1})\}}{n_A n}\max_{i\in 1,\ldots,n}\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}\max_{u_r\in A}\sigma_{\max}\{\hat{\boldsymbol{P}}(u_r)\}\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_{l^2}\\
&+\left\|\left(2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}+2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_{l^2},
\end{aligned}
$$

where $\sigma_{\max}(x)$ is the largest singular value of a matrix $x$.

### 2.7.1 Proof of Theorem 1

Denote the minimizer of (2.1) by $(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\theta}})$. Define $\hat{\beta}(t,u)=\boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\hat{\boldsymbol{\theta}}$,

$$
A_0=\{r\in(1,\ldots,n_A):\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|_{l^2}\ge M\sqrt{m/n},\ \text{for some constant } M>0\},
$$

and

$$
A_1=\{r\in(1,\ldots,n_A):\left\|\hat{\beta}_{u_r}(t)-\beta_{u_r}(t)\right\|_{L^2}\ge M\kappa_m^{-1/2}m^{1/2}n^{-1/2},\ \text{for some constant } M>0\}.
$$

Then by Lemma 4, we have for some $M>0$

$$
\begin{aligned}
&-n_A^{-1}\sum_{r=1}^{n_A}E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right]\\
=&-n_A^{-1}\sum_{r\in(A\setminus A_0)}E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right]\\
&-n_A^{-1}\sum_{r\in A_0}E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right],\\
\ge&-n_A^{-1}\sum_{r\in(A\setminus A_0)}E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right]\\
&+n_A^{-1}\left\{M(1-o_p(1))m^{1/2}n^{-1/2}-O_p\left(m^{1/2}n^{-1/2}\right)\right\}\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|_{l^2}.
\end{aligned}
\tag{2.20}
$$

For the first term in (2.20), almost surely

$$\begin{aligned}
&\left|-n_A^{-1}\sum_{r\in(A\setminus A_0)} E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right]\right|\\
&\le n_A^{-1}\sum_{r\in(A\setminus A_0)} E_n\left[\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\left\|\hat{\boldsymbol{\eta}}_i^{m+1}\right\|\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|\right]\\
&= n_A^{-1}n^{-1}\sum_{(i,r)\in(A\setminus A_0)\cap I_3}\left\{u-\mathbb{1}\left(Y_i\le\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\left\|\hat{\boldsymbol{\eta}}_i^{m+1}\right\|\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|\\
&= o_p\left(n_A^{-1}n^{-1}m|\Delta|^{-1}\right)\times O\left(m^{1/2}n^{-1/2}\right)=o_p\left(n_A^{-1}n^{-3/2}|\Delta|^{-1}m^{3/2}\right). \qquad (2.21)
\end{aligned}$$

On the other hand, by standard matrix theory,

$$\begin{aligned}
\max_{u_r\in A}\sigma_{\max}\{\hat{\boldsymbol{P}}(u_r)\} &\le \max_{u_r\in A}\left\|\hat{\boldsymbol{P}}(u_r)\right\|_F,\\
&\le\left\{cm\max_{u_r\in A}\sum_{b_j(t,u_r)\neq 0}\left(\int_{\Delta_j}b_j(t,u_r)\hat{\phi}_k(t)dt\right)^2\right\}^{1/2},\\
&= cm^{1/2}|\Delta|^{1/2}.
\end{aligned}$$

Then, for the first term of (2.10), we have

$$\begin{aligned}
&\max_{i\in1,\ldots,n}\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}\max_{u_r\in A}\sigma_{\max}\{\hat{\boldsymbol{P}}(u_r)\}\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_{l^2}\\
\le& o_p\left((\log n)^{-1}n^{1/2}m^{-1/2}\right)cm^{1/2}|\Delta|^{-1/2}\left\|\hat{\beta}-\beta^*\right\|_{L^2(\boldsymbol{\Omega})},\\
=& o\left((\log n)^{-1}n^{1/2}|\Delta|^{-1/2}\right)\left\|\hat{\beta}(t,u)-\sum_{k=1}^m\beta_k(u)\hat{\phi}_k(t)+\sum_{k=1}^m\beta_k(u)\hat{\phi}_k(t)-\beta^*(t,u))\right\|_{L^2(\boldsymbol{\Omega})},\\
\lesssim& o\left((\log n)^{-1}n^{1/2}|\Delta|^{-1/2}\right)\left\{n_A^{-1/2}\left(\sum_{r=1}^{n_A}\kappa_m^{-1}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}^2\right)^{1/2}+O(|\Delta|^{d+1})\right\},
\end{aligned} \qquad (2.22)$$

provided that $\max_{1\le i\le n}\left\|\hat{\boldsymbol{\xi}}_i\right\|_{l^2}=o((\log n)^{-1}n^{1/2}m^{-1/2})$ [41]. By Lemma 2 and [82],

$$\left\|\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right\|_{l^2}\le C|\Delta|^{-2}\left\|\hat{\boldsymbol{\theta}}\right\|_{l^2}=C|\Delta|^{-3}\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}.$$

By Lemma 5,

$$\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}\approx n_A^{-1/2}n_T^{-1/2}\left\|\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\right\|_{l^2}\le O_p(m^{3/4}n^{-3/4}|\Delta|^{(d+1)/2}\vee m^{1/2}n^{-1/2})+C.$$

In addition, $\left\|\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\right\| \leq \Gamma_{max}(\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}})\left\|\hat{\boldsymbol{\theta}}\right\|_{l^2} = C|\Delta|n_A n_T\left\|\hat{\beta}(t,u)\right\|_{L^2(\boldsymbol{\Omega})}$. Then we conclude that

$$\begin{aligned}
&\left\|\left(2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}+2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*}\right)\right\|_{l^2}\\
\leq&\left\|2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\right\|_{l^2}+\left\|2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\hat{\boldsymbol{\theta}}\right\|_{l^2}\\
&+\left\|\left(2\lambda_{1,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}+2\lambda_{2,n}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\right)\boldsymbol{\theta}^{*}\right\|_{l^2}\\
=&O_p(m^{-1/2}n|\Delta|^{d+1})(1+o(1))
\end{aligned}\tag{2.23}$$

Given $|\Delta| = o\left(m^{(1+\alpha)/(2d+2)}n^{-3/(2d+2)}\right)$, from (2.22), if

$$n_A^{-1/2}\left(\sum_{r=1}^{n_A}\kappa_m^{-1}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}^2\right)^{1/2} \leq O_p(|\Delta|^{d+1}) = o_p\left(m^{(1+\alpha)/2}n^{-1/2}\right),$$

then by Lemma 5, we can conclude $\left\|\hat{\beta}(t,u)-\beta(t,u)\right\|_{L^2(\boldsymbol{\Omega})} \approx O_p\left(\kappa_m^{-1/2}m^{1/2}n^{-1/2}\right)$. If not, then

$$\begin{aligned}
(2.22) &= o_p\left((\log n)^{-1}n^{1/2}|\Delta|^{-1/2}\right)\left(n_A^{-1/2}\sum_{r=1}^{n_A}\kappa_m^{-1/2}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}^2+O(|\Delta|^{d+1})\right)^{1/2},\\
&= o_p\left((\log n)^{-1}n_A^{-1/2}n^{1/2}|\Delta|^{-1/2}\kappa_m^{-1/2}\right)\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}(1+o(1)).
\end{aligned}\tag{2.24}$$

By Lemma 6, (2.23) and (2.24) , almost surely

$$\begin{aligned}
&-n_A^{-1}\sum_{r=1}^{n_A}E_n\left[\left\{u-\mathbb{1}\left(Y_i\leq\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\hat{\boldsymbol{d}}^{m+1}(u_r)\right)\right\}\hat{\boldsymbol{\eta}}_i^{m+1}\cdot\left(\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right)\right]\\
&\leq o_p\left(n_A^{-3/2}|\Delta|^{-3/2}n^{-1/2}\kappa_m^{-1/2}\right)\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}+O_p(m^{-1/2}n|\Delta|^{d+1})(1+o(1))
\end{aligned}\tag{2.25}$$

Then by Lemma 4, (2.20), (2.21), (2.25) and $n_A^{-1}|\Delta|^{-1}m^{(\alpha-1)/3} = o(1)$, we have almost surely

$$\begin{aligned}
&\left\{M(1-o_p(1))m^{1/2}n^{-1/2}-O_p\left(m^{1/2}n^{-1/2}\right)\right\}\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|_{l^2}\\
&\leq o_p\left(n_A^{-3/2}|\Delta|^{-3/2}n^{-1/2}\kappa_m^{-1/2}\right)\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^m(u_r)-\boldsymbol{d}^m(u_r)\right\|_{l^2}+O_p(m^{-1/2}n|\Delta|^{d+1})(1+o(1)),
\end{aligned}\tag{2.26}$$

which implies that

$$\sum_{r\in A_0}\left\|\hat{\boldsymbol{d}}^{m+1}(u_r)-\boldsymbol{d}^{m+1}(u_r)\right\|_{l^2} = O_p(n_A n^{1/2}|\Delta|^{d+1}).$$

Based on the definition of $A_0$ and $A_1$, we have $A_1 \subset A_0$ and $|A_0| = o_p(m^{-1-\alpha}n^{-1/2}n_A)$. Then by Lemma 4,

$$\left\|\hat{\beta}(t,u) - \sum_{k=1}^{m}\beta_k(u)\phi_k(t)\right\|_{L^2(\mathbf{\Omega})} \lesssim n_A^{-1/2}\left(\sum_{r=1}^{n_A}\kappa_m^{-1}\left\|\hat{\boldsymbol{d}}^m(u_r) - \boldsymbol{d}^m(u_r)\right\|_{l^2}^2\right)^{1/2} = O_p(\kappa_m^{-1/2}m^{1/2}n^{-1/2}).$$

By condition A5,

$$\left\|\beta(t,u) - \sum_{k=1}^{m}\beta_k(u)\phi_k(t)\right\|_{L^2(\mathbf{\Omega})} = O(m^{-(2\zeta+1)/2}).$$

Therefore, we conclude that

$$\left\|\hat{\beta}(t,u) - \beta(t,u)\right\|_{L^2(\mathbf{\Omega})} \approx O_p\left(\kappa_m^{-1/2}m^{1/2}n^{-1/2} \vee m^{-2\zeta+1}\right).$$

### 2.7.2 Proof of Theorem 2

Assume that $p_0$ is known and let $m = p_0$. Under the condition A5, and by Lemma 1 and Lemma 3, there exist $\boldsymbol{\gamma}_0^*$ and $\boldsymbol{\theta}^*$ such that

$$\sup_{(t,u)\in\mathscr{T}\times\mathscr{A}}|\beta(t,u) - \boldsymbol{B}^{\mathrm{T}}(t,u)\boldsymbol{Q}\boldsymbol{\theta}^*| \le C_1|\Delta|^{d+1},$$
$$\sup_{u\in\mathscr{A}}|c(u) - \boldsymbol{b}_0^{\mathrm{T}}(u)\boldsymbol{\gamma}_0^*| \le C_2|\Delta|^{d+1},$$

for some constant $C_1$ and $C_2$. Let $\mathbf{\Gamma}^* = (\boldsymbol{\gamma}_0^{*\mathrm{T}}, \boldsymbol{\theta}^{*\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\delta} = \sqrt{n}(\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_0^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}}$. Let $\boldsymbol{P}(u_r)$ denote the $p_0$ by $J$ matrix with the $(k,j)$-entry being $\int_{(t,u_r)\in\Delta_j}\phi_k(t)b_j(t,u)dt$, where $j = 1,\ldots,J$ is the index for the bivariate splines basis. Then define

$$\psi_u(x) = u - \mathbb{1}(x \le 0), \quad \hat{\omega}_{i,r} = y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^*, \quad \text{and} \quad \omega_{i,r} = y_i - \boldsymbol{Z}_i^*(u_r)\mathbf{\Gamma}^*,$$

where

$$\boldsymbol{Z}_i^*(u_r) = \left[\boldsymbol{b}_0^{\mathrm{T}}(u_r), \boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{Q}\right].$$

We further define $L_n^{0,1}(\mathbf{\Gamma},\boldsymbol{\delta})$ and $L_n^{0,2}(\mathbf{\Gamma},\boldsymbol{\delta})$ as follow,

$$\begin{aligned}L_n^{0,1}(\mathbf{\Gamma},\boldsymbol{\delta}) &= \{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\boldsymbol{Z}_i(u_r)\frac{\boldsymbol{\delta}}{\sqrt{n}}\left\{\mathbb{1}(y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \le 0) - u_r\right\}\\ &= -\{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\boldsymbol{Z}_i(u_r)\frac{\boldsymbol{\delta}}{\sqrt{n}}\psi_{u_r}(\hat{\omega}_{i,r}),\end{aligned}$$

and

$$L_n^{0,2}(\boldsymbol{\Gamma}) = (nn_A)^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \{\mathbb{1}\left(Y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le t\right) - \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le 0)\}\, dt$$
$$= n_A^{-1} \sum_{r=1}^{n_A} G_r(\boldsymbol{\Gamma}, \boldsymbol{\delta}),$$

where

$$G_r(\boldsymbol{\Gamma}, \boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \{\mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le t) - \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le 0)\}\, dt.$$

**Lemma 7.** *Under the conditions of Theorem 1,*

$$L_n^{0,2}(\boldsymbol{\Gamma}, \boldsymbol{\delta}) = \frac{1}{2n} \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{\delta} + o_p(1),$$

*where*

$$\boldsymbol{\Sigma}_1 = n_A^{-1} \sum_{r=1}^{n_A} E\left[f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \boldsymbol{Z}_i^{\mathrm{T}}(u_r) \boldsymbol{Z}_i(u_r)\right].$$

*Proof*: By Taylor expansion,

$$\begin{aligned}
E\left[G_r(\boldsymbol{\Gamma}, \boldsymbol{\delta})\right] &= E\left[E\left[G_r(\boldsymbol{\Gamma}, \boldsymbol{\delta}) \mid \hat{\boldsymbol{\xi}}_i, i = 1, \ldots, n\right]\right] \\
&= E\left[n^{-1} \sum_{i=1}^{n} \int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \{F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* + t) - F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*)\}\, dt\right] \\
&= n^{-1} \sum_{i=1}^{n} E\left[\int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \{f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*)t + R_{i,r}(t)\}\, dt\right] \\
&= n^{-1} \sum_{i=1}^{n} E\left[\frac{f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*)}{2} t^2 \Big|_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} + R_{i,r}(t)\right] \\
&= n^{-1} \sum_{i=1}^{n} E\left[\frac{1}{2n} f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{Z}_i^{\mathrm{T}}(u_r) \boldsymbol{Z}_i(u_r) \boldsymbol{\delta}\right] + \sum_{i=1}^{n} E\left[R_{i,r}(t)\right] \\
&= \frac{1}{2n} \boldsymbol{\delta}^{\mathrm{T}} \left\{n^{-1} \sum_{i=1}^{n} E\left[f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \boldsymbol{Z}_i^{\mathrm{T}}(u_r) \boldsymbol{Z}_i(u_r)\right]\right\} \boldsymbol{\delta} + \sum_{i=1}^{n} E\left[R_{i,r}(t)\right],
\end{aligned}$$

where $\{R_{i,r}(t)\}$ are the remainder terms of Taylor expansion. Regarding these remainder terms $\{R_{i,r}(t)\}$, we have

$$\sum_{i=1}^{n} E\left[R_{i,r}(t)\right] \le \text{constant} \cdot n^{-3/2} \sum_{i=1}^{n} E\left[\left(\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}\right)^3\right] \le O(n^{-1/2}) E\left[\|\boldsymbol{Z}_i(u_r)\|_{l^2}^3\right] \|\boldsymbol{\delta}\|_{l^2}^3 = o(1).$$

Then, we can conclude that

$$E\left[L_n^{0,2}(\boldsymbol{\Gamma})\right] = \frac{1}{2n} \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{\delta} + o(1).$$

Next, we need to prove $E\left[L_n^{0,2}(\mathbf{\Gamma})^2\right] = o(1)$.

$$
\begin{aligned}
E\left[\left(L_n^{0,2}(\mathbf{\Gamma})\right)^2\right] =& E\left[E\left[\left(L_n^{0,2}(\mathbf{\Gamma})\right)^2 \mid \hat{\boldsymbol{\xi}}_i, i=1,\ldots,n\right]\right] \\
\leq& E\Bigg[E\Bigg[(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\Bigg\{\int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq t) \\
& - \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq 0)dt\Bigg\}^2 \mid \hat{\boldsymbol{\xi}}_i, i=1,\ldots,n\Bigg]\Bigg] \\
\leq& E\Bigg[(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\int_0^{|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}|} |\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}| E\Bigg[\{\mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq t) \\
& - \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq 0)\}^2 dt \mid \hat{\boldsymbol{\xi}}_i, i=1,\ldots,n\Bigg]\Bigg] \\
=& E\Bigg[(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}| \int_0^{|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}|} E\Bigg[|\mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq t) \\
& - \mathbb{1}(Y_i - \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* \leq 0)|dt \mid \hat{\boldsymbol{\xi}}_i, i=1,\ldots,n\Bigg]\Bigg].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E\left[\left(L_n^{0,2}(\mathbf{\Gamma})\right)^2\right] \leq& E\Bigg[(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}| \\
& \int_0^{|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}|} E\big[\mathbb{1}(\boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* < Y_i < \boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* + t)\big]dt \mid \hat{\boldsymbol{\xi}}_i, i=1,\ldots,n\Bigg]\Bigg] \\
=& E\Bigg[(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}| \\
& \int_0^{|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}|} \{F_i(\boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^* + t) - F_i(\boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^*)\}\,dt\Bigg] \\
=& (n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n} E\left[\frac{|\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}|}{2n} f_i(\boldsymbol{Z}_i(u_r)\mathbf{\Gamma}^*)\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{Z}_i^{\mathrm{T}}(u_r)\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}\right] \\
& + (n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n} E\left[R_{i,r}(t)\right] \\
=& o(1).
\end{aligned}
$$

Define

$$
\boldsymbol{V}_0 = -(n_A n)^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\boldsymbol{Z}_i^*(u_r)\psi_{u_r}(\omega_{i,r}),
$$

and let $\tilde{\boldsymbol{Z}}_i$ and $\boldsymbol{\psi}$ denote $(\boldsymbol{Z}_i^{\mathrm{T}}(u_1),\ldots,\boldsymbol{Z}_i^{\mathrm{T}}(u_{n_A}))$ and $\left(\psi_{u_1},\ldots,\psi_{u_{n_A}}\right)$ respectively.

**Lemma 8.** *Under the conditions of Theorem 1 and A6,*

$$
\sqrt{n}\boldsymbol{V}_0 \to N(\mathbf{0}, \boldsymbol{U}_2),
$$

*in distribution, where*

$$\boldsymbol{U}_2 = n_A^{-2} E\left[\tilde{\boldsymbol{Z}}_i^{\mathrm{T}} \boldsymbol{U}_1 \tilde{\boldsymbol{Z}}_i\right].$$

*Proof*: First notice that

$$\begin{aligned}
&\operatorname{cov}\left(\psi_{u_r}(\omega_{i,r}), \psi_{u_r}(\omega_{j,r})\right) \\
&= \begin{cases} 0, & i \neq j \\ u_{r'} - u_r u_{r'} + O\left(\left|\int \sum_{k=1}^{p_0} \xi_{i,k}\phi_k(t)\left(\beta(t,u_r) - \beta^*(t,u_r)\right) dt\right|\right), & \text{otherwise} \end{cases} \\
&= \begin{cases} 0, & i \neq j \\ u_r \wedge u_{r'} - u_r u_{r'} + O_p(|\Delta|^{d+1}), & \text{otherwise} \end{cases}
\end{aligned}$$

Since $E\left[\psi_{u_r}(\omega_{i,r}) \mid \boldsymbol{\xi}_i\right] = u_r - F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*) = O_p(|\Delta|^{d+1})$, then $E\left[\boldsymbol{V}_0 \mid \boldsymbol{\xi}_i\right] = O_p(|\Delta|^{d+1})$. On the other hand, $Var\left(\boldsymbol{V}_0 \mid \boldsymbol{\xi}_i\right) = \operatorname{var}\left(n^{-1}\sum_{i=1}^n n_A^{-1}\tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\boldsymbol{\psi}\right) = (n_A n)^{-2}\sum_{i=1}^n \tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\boldsymbol{U}_1\tilde{\boldsymbol{Z}}_i$, where $\boldsymbol{U}_1$ is a matrix with its $(r,r')$-entry being $u_r \wedge u_{r'} - u_r u_{r'}$. Then the co-variance matrix of $\boldsymbol{V}_0$ is given by

$$\begin{aligned}
\operatorname{var}(\boldsymbol{V}_0) =& \operatorname{var}\left(E\left(\boldsymbol{V}_0 \mid \boldsymbol{\xi}_i\right)\right) + E\left(\operatorname{var}\left(\boldsymbol{V}_0 \mid \boldsymbol{\xi}_i\right)\right) \\
=& O_p(|\Delta|^{d+1}) + n^{-1}\left[n_A^{-2} E\left[\tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\boldsymbol{U}_1\tilde{\boldsymbol{Z}}_i\right]\right] \\
=& O_p(|\Delta|^{d+1}) + n^{-1}\boldsymbol{U}_2.
\end{aligned}$$

Regarding $\boldsymbol{U}_2$, we have

$$\begin{aligned}
\boldsymbol{U}_2 \leq& \frac{2}{n_A^2} E\left[\tilde{\boldsymbol{Z}}_i^{\mathrm{T}}\tilde{\boldsymbol{Z}}_i\right] \leq \frac{2}{n_A^2} n_A \max_r E\left\|\boldsymbol{Z}_i(u_r)\right\|_2^2 \\
=& \frac{2}{n_A}\max_r\left\{\left\|\boldsymbol{b}_0(u_r)\right\|_2^2 + E\left\|\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{Q}\right\|_2^2\right\} \\
=& O\left(n_A^{-1}|\Delta|^{-1}\right).
\end{aligned}$$

Based on $n_A n|\Delta|^{d+2} = o(1)$, we just prove that

$$\sqrt{n}\boldsymbol{V}_0 \to N(\boldsymbol{0}, \boldsymbol{U}_2)$$

in distribution.

Define

$$\begin{aligned}
L_n^0(\boldsymbol{\Gamma}) &= \{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{b}_0^{\mathrm{T}}(u_r)\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r)\boldsymbol{Q}\boldsymbol{\theta}\right) \\
&= \{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}\right), \qquad (2.27)
\end{aligned}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\theta})^{\mathrm{T}}$, and $P_\lambda(\boldsymbol{\Gamma}) = \lambda_{1,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}_{A,T}\boldsymbol{B}_{A,T}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta} + \lambda_{2,n}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{\theta}$. Recall that $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* + \boldsymbol{\delta}/\sqrt{n}$. Then minimizing

$$L_n(\boldsymbol{\Gamma}) = L_n^0(\boldsymbol{\Gamma}) + P_\lambda(\boldsymbol{\Gamma})$$

is equivalent to minimizing

$$\begin{aligned} LL_n(\boldsymbol{\delta}) = &\{nn_A\}^{-1} \sum_{r=1}^{n_A}\sum_{i=1}^{n} \left\{\rho_{u_r}\left(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n} - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*\right) - \rho_{u_r}\left(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*\right)\right\} \\ &+ P_\lambda(\boldsymbol{\delta}/\sqrt{n} + \boldsymbol{\Gamma}^*). \end{aligned} \tag{2.28}$$

Applying the Knight's identity [42],

$$\rho_u(x-y) - \rho_u(x) = y\left\{\mathbb{1}(x \le 0) - u\right\} + \int_0^y \left\{\mathbb{1}(x \le t) - \mathbb{1}(x \le 0)\right\} dt,$$

on the first term of (2.28), then we have

$$\begin{aligned} L_n^0(\boldsymbol{\Gamma}) = &\{nn_A\}^{-1} \sum_{r=1}^{n_A}\sum_{i=1}^{n} \Bigg\{ \boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}\left\{\mathbb{1}(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le 0) - u_r\right\} \\ &+ \int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \left\{\mathbb{1}(y_i - \boldsymbol{Z}_i(u_r) \le t) - \mathbb{1}(y_i - \boldsymbol{Z}_i(u_r) \le 0)\right\} dt \Bigg\} \\ = &\{nn_A\}^{-1} \sum_{r=1}^{n_A}\sum_{i=1}^{n} \boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}\left\{\mathbb{1}(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le 0) - u_r\right\} \\ &+ \{nn_A\}^{-1} \sum_{r=1}^{n_A}\sum_{i=1}^{n} \int_0^{\boldsymbol{Z}_i(u_r)\boldsymbol{\delta}/\sqrt{n}} \left\{\mathbb{1}(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le t) - \mathbb{1}(y_i - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^* \le 0)\right\} dt \\ = &L_n^{0,1}(\boldsymbol{\Gamma}) + L_n^{0,2}(\boldsymbol{\Gamma}). \end{aligned}$$

From Lemma 7 and Lemma 8, under the conditions $\lambda_{1,n} \asymp n_A^{-1} n_T^{-1} m^{-1/2} n |\Delta|^{d+1}$, and $\lambda_{2,n} = o(\lambda_{1,n} n_A n_T |\Delta|^4)$, we have

$$\begin{aligned}
LL_n(\boldsymbol{\delta}) =& - \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i(u_r) \frac{1}{\sqrt{n}} \boldsymbol{\delta} \psi_{u_r}(\hat{\omega}_{i,r}) \\
&+ \frac{1}{2n} \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{\delta} + E\left[R(t)\right] + P_\lambda(\frac{1}{\sqrt{n}} \boldsymbol{\delta} + \boldsymbol{\Gamma}^*) \\
=& - \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i(u_r) \frac{1}{\sqrt{n}} \boldsymbol{\delta} \psi_{u_r}(\hat{\omega}_{i,r}) \\
&+ \frac{1}{2n} \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{\delta} + \boldsymbol{\delta}^{\mathrm{T}} \left[ \lambda_{1,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{B}_{A,T} \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} \end{bmatrix} + \lambda_{2,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{Q} \end{bmatrix} \right] \boldsymbol{\delta} \\
&+ \frac{2}{\sqrt{n}} \boldsymbol{\delta}^{\mathrm{T}} \left[ \lambda_{1,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{B}_{A,T} \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} \end{bmatrix} + \lambda_{2,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{Q} \end{bmatrix} \right] \boldsymbol{\Gamma}^* + o(1) \\
=& \boldsymbol{V} \boldsymbol{\delta} (1 + o_p(1)) + \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Sigma}_2 \boldsymbol{\delta} + o(1),
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{V} =& - \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \boldsymbol{Z}_i(u_r) \psi_{u_r}(\hat{\omega}_{i,r}), \\
\boldsymbol{\Sigma}_2 =& \frac{1}{2n} \boldsymbol{\Sigma}_1 + \lambda_{1,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{B}_{A,T} \boldsymbol{B}_{A,T}^{\mathrm{T}} \boldsymbol{Q} \end{bmatrix} + \lambda_{2,n} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{Q} \end{bmatrix}.
\end{aligned}$$

Then by the convexity lemma [63] and quadratic approximation lemma [18], we have $\hat{\boldsymbol{\delta}} = -\frac{1}{2} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{V}^{\mathrm{T}} + o_p(1)$. Let $\tilde{\boldsymbol{B}}(t,u) = (\mathbf{0}_{1 \times n_B}, \boldsymbol{B}^{\mathrm{T}}(t,u) \boldsymbol{Q})^{\mathrm{T}}$, and $\beta^*(t,u) = \boldsymbol{B}^{\mathrm{T}}(t,u) \boldsymbol{Q} \boldsymbol{\theta}^*$. Then, we have $\hat{\beta}(t,u) - \beta^*(t,u) = \tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u) \hat{\boldsymbol{\delta}} / \sqrt{n}$. Therefore,

$$\begin{aligned}
\hat{\beta}(t,u) - \beta(t,u) &= \tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u) \hat{\boldsymbol{\delta}} / \sqrt{n} + \beta^*(t,u) - \beta(t,u) \\
&= -\frac{1}{2n} \tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u) \boldsymbol{\Sigma}_2^{-1} \sqrt{n} \boldsymbol{V} + \beta^*(t,u) - \beta(t,u) + o_p(1).
\end{aligned}$$

Now, we need to prove the asymptotic normality of $\boldsymbol{V}$. We first decompose $\boldsymbol{V}_0 - \sqrt{n}\boldsymbol{V}$ into three terms and then calculate the order for each of them.

$$
\begin{aligned}
&\{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i(u_r) \psi_{u_r}(\hat{\omega}_{i,r}) - \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i^*(u_r) \psi_{u_r}(\omega_{i,r}) \\
=&\{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \{\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)\} \{\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})\} \\
&+ \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i^*(u_r) \{\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})\} \\
&+ \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \{\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)\} \psi_{u_r}(\omega_{i,r}) \\
=&\boldsymbol{V}_1 + \boldsymbol{V}_2 + \boldsymbol{V}_3.
\end{aligned}
$$

First notice that

$$
\begin{aligned}
&E\left[\psi_{u_r}(\hat{\omega}_{i,r}) \middle| \boldsymbol{Z}_i(u_r), \boldsymbol{Z}_i^*(u_r)\right] \\
=&u_r - F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \\
=&u_r - F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*) + F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*) - F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \\
=&o_p(1) + f_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) \left(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^* - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*\right) + O\left(\left(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^* - \boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*\right)^2\right).
\end{aligned}
$$

Then for the first term $\boldsymbol{V}_1 = \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \{\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)\} \{\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})\}$,

$$
\begin{aligned}
E|\boldsymbol{V}_1| \leq& E\left[\{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} |\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)| \, |\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})|\right] \\
=&E\left[\{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} |\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)| \, E\left[|\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})| \mid \boldsymbol{Z}_i, \boldsymbol{Z}_i^*\right]\right] \\
=&E\left[\{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} |F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) - F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)| \, |\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)|\right]
\end{aligned}
$$

Therefore,

$$\begin{aligned}
E|\boldsymbol{V}_1| \leq & E\left[ \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} |F_i(\boldsymbol{Z}_i(u_r)\boldsymbol{\Gamma}^*) - F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)|^2 \right]^{1/2} \\
& \times E\left[ \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} |\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)|^2 \right]^{1/2} \\
\leq & O\left( \max_r E\left[ (\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r))^2 \right]^{1/2} \right) \\
= & O\left( \max_r E\left[ \left\{ \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}} \boldsymbol{P}(u_r) \right\}^2 \right]^{1/2} \right) \\
= & O\left( \max_r E\left[ \left\{ \hat{\boldsymbol{\xi}}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) + \boldsymbol{\xi}_i^{\mathrm{T}} \hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}} \boldsymbol{P}(u_r) \right\}^2 \right]^{1/2} \right)
\end{aligned}$$

According to [29] and [30], we have

$$\left\| \hat{\phi}_k - \phi_k \right\|_{L^2(\boldsymbol{\Omega})} \leq \text{constant} \cdot s_k^{-1} n^{-1/2},$$

and for any $c > 0$,

$$E\left\| \hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i \right\|^c \leq \text{constant} \cdot s_{p0}^{-c} n^{-c/2},$$

where $s_k = \min_{r \leq k} (\kappa_r - \kappa_{r+1})$. Therefore, we can conclude that

$$E|\boldsymbol{V}_1| = O(n^{-1/2}|\Delta|).$$

Similarly,

$$\begin{aligned}
E\left[\boldsymbol{V}_1^2\right] \leq & E\left[ \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \{\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)\}^2 \{\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})\}^2 \right] \\
\leq & \max_r E\left[ (\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r))^4 \right]^{1/2} \max_r E\left[ (\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r}))^4 \right]^{1/2} \\
= & \text{constant} \cdot n^{-1} |\Delta|^2.
\end{aligned}$$

Then we have

$$\boldsymbol{V}_1 = O_p(n^{-1/2}|\Delta|).$$

For the second term $\boldsymbol{V}_2 = \{nn_A\}^{-1} \sum_{r=1}^{n_A} \sum_{i=1}^{n} \boldsymbol{Z}_i^*(u_r) \{\psi_{u_r}(\hat{\omega}_{i,r}) - \psi_{u_r}(\omega_{i,r})\}$, we first define

$$R_r(\boldsymbol{t}_r) = \sum_{i=1}^{n} \boldsymbol{Z}_i^*(u_r) \{\psi_{u_r}(\omega_{i,r} - \boldsymbol{\xi}_i^{\mathrm{T}} \boldsymbol{t}_r) - \psi_{u_r}(\omega_{i,r})\}$$

for any vector such that $\|\boldsymbol{t}_r\| \leq C_t$ with some constant $C_t$. According to [48], we have $\sup \|R_r(\boldsymbol{t}_r) - E\left[R_r(\boldsymbol{t}_r)\right]\| = O_p\left(n^{1/2}\log\left(n\right)\|\boldsymbol{t}_r\|^{1/2}\right)$, and therefore,

$$\begin{aligned} E\left[R_r(\boldsymbol{t}_r)\right] =& \sum_{i=1}^{n} E\left[\boldsymbol{Z}_i^*(u_r)\left\{F_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*) - F_i\left(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^* - [\boldsymbol{b}_0(u_r)^{\mathrm{T}}, \boldsymbol{\xi}_i^{\mathrm{T}}]\,\boldsymbol{t}_r\right)\right\}\right] \\ =& -nE[\boldsymbol{Z}_i^*(u_r)f_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)\left([\boldsymbol{b}_0(u_r)^{\mathrm{T}}, \boldsymbol{\xi}_i^{\mathrm{T}}]\,\boldsymbol{t}_r\right)] + O\left(nE\left[\left([\boldsymbol{b}_0(u_r)^{\mathrm{T}}, \boldsymbol{\xi}_i^{\mathrm{T}}]\,\boldsymbol{t}_r\right)^2\boldsymbol{Z}_i^*(u_r)\right]\right). \end{aligned} \tag{2.29}$$

Define $\boldsymbol{\Sigma}_{c,r} = E[f_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)\boldsymbol{b}_0\boldsymbol{b}_0^{\mathrm{T}}]$ and $\boldsymbol{\Sigma}_{3,r} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}(u_r)E[f_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}]$. Then from (2.29), we have

$$R_r^{\mathrm{T}}(\boldsymbol{t}_r) = -n\left[\boldsymbol{\Sigma}_{c,r}, \boldsymbol{\Sigma}_{3,r}\right]\boldsymbol{t}_r + O\left(n\|\boldsymbol{t}_r\|_2^2\right) + O_p\left(n^{1/2}\log\left(n\right)\|\boldsymbol{t}_r\|_2^{1/2}\right).$$

By [48], we know that there exists a random matrix $\boldsymbol{C}_\xi$ such that

$$\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i = n^{-1/2}\boldsymbol{C}_\xi\boldsymbol{\xi}_i + O_p(n^{-1}).$$

The dimension of $\boldsymbol{C}_\xi = (c_{k,k'})$ is $p_0$ by $p_0$ where $c_{k,k'} = 0$ if $k = k'$ and $c_{k,k'} = n^{-1/2}(\kappa_k - \kappa_{k'})^{-1}\sum_{i=1}^{n}\xi_{ik}\xi_{ik'}$ if $k \neq k'$. Then,

$$\begin{aligned} \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{P}(u_r) &= \hat{\boldsymbol{\xi}}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r) + \boldsymbol{\xi}_i^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r) - \boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{P}(u_r) \\ &= n^{-1/2}\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{C}_\xi^{\mathrm{T}}\hat{\boldsymbol{P}}(u_r) + O_p\left(n^{-1}|\Delta|\right) + O_p\left(n^{-3/4}|\Delta|\right) \\ &= n^{-1/2}\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{C}_\xi^{\mathrm{T}}\left(\boldsymbol{P}(u_r) + O_p\left(n^{-3/4}|\Delta|\right)\right) + O_p\left(n^{-1}|\Delta|\right) + O_p\left(n^{-3/4}|\Delta|\right) \\ &= n^{-1/2}\boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{C}_\xi^{\mathrm{T}}\boldsymbol{P}(u_r) + o_p(1). \end{aligned}$$

Choose $\boldsymbol{t}_r$ as $\boldsymbol{t}_r^{\mathrm{T}} = \left[\boldsymbol{0}, n^{-1/2}\boldsymbol{\Gamma}^{*\tau}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}(u_r)\boldsymbol{C}_\xi\right]$, and define $\boldsymbol{\Sigma}_3 = n_A^{-1}\sum_{r=1}^{n_A}\boldsymbol{\Sigma}_{3,r}\boldsymbol{C}_\xi^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{Q}$. Then,

$$\begin{aligned} \boldsymbol{V}_2 =& \{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\boldsymbol{Z}_i^*(u_r)\left\{\psi_{u_r}(\omega_{i,r} - [\boldsymbol{b}_0(u_r)^{\mathrm{T}}, \boldsymbol{\xi}_i^{\mathrm{T}}]\,\boldsymbol{t}_r) - \psi_{u_r}(\omega_{i,r})\right\} \\ =& -n^{-1/2}n_A^{-1}\sum_{r=1}^{n_A}\boldsymbol{\Sigma}_{3,r}\boldsymbol{C}_\xi^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{\Gamma}^*\boldsymbol{Q}\left(1 + o_p(1)\right) \\ =& -n^{-1/2}\boldsymbol{\Sigma}_3\boldsymbol{\Gamma}^*\left(1 + o_p(1)\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_3\boldsymbol{\Gamma}^* =& n_A^{-1}\sum_{r=1}^{n_A}\boldsymbol{\Sigma}_{3,r}\boldsymbol{C}_\xi^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{Q}\boldsymbol{\Gamma}^* \\ =& n_A^{-1}\sum_{r=1}^{n_A}f_i(\boldsymbol{Z}_i^*(u_r)\boldsymbol{\Gamma}^*)\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}(u_r)E\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right]\boldsymbol{C}_\xi^{\mathrm{T}}\boldsymbol{P}(u_r)\boldsymbol{Q}\boldsymbol{\Gamma}^*. \end{aligned}$$

We know that $E\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}\right]$ is a diagonal matrix with the main diagonal entries being the variances of functional principal component scores. For $\boldsymbol{C}_\xi$, its main diagonal are zeros and its off diagonal entries are of order $O_p(1)$ because under the conditions A2 and A3, we have

$$(\kappa_k - \kappa_{k'})^{-2} E\left[(\xi_{ik}\xi_{ik'})^2\right] \leq (\kappa_k - \kappa_{k'})^{-2}\left(E\left[\xi_{ik}^4\right] E\left[\xi_{ik'}^4\right]\right)^{1/2} = O(1).$$

On the other hand, recall that for each $r$, $\boldsymbol{P}(u_r)$ is a $p_0$ by $J$ matrix with the $(k,j)$-entry being $\int_{(t,u_r)\in\Delta_j}\phi_k(t)b_j(t,u)dt$, where $j = 1,\ldots,J$ is the index for the bivariate splines basis and $J = O(|\Delta|^{-1})$. By the choice of $\boldsymbol{\Gamma}^*$, we know that

$$\boldsymbol{P}(u_r)\boldsymbol{Q}\boldsymbol{\Gamma}^* = \begin{bmatrix}\beta_1(u_r)\\ \ldots \\ \beta_{p_0}(u_r)\end{bmatrix},$$

where $\beta_k(u) = \int \beta(t,u)\phi_k(t)dt$, $k = 1,\ldots,p_0$. Given that $p_0$ is finite, we can conclude that

$$\boldsymbol{V}_2 = O_p(n^{-1/2}|\Delta|).$$

For the third term $\boldsymbol{V}_3 = \{nn_A\}^{-1}\sum_{r=1}^{n_A}\sum_{i=1}^{n}\{\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r)\}\psi_{u_r}(\omega_{i,r})$, since

$$E\left[\psi_{u_r}(\omega_{i,r}) \mid \boldsymbol{\xi}_i\right] = O_p(|\Delta|^{d+1}),$$
$$E\left[\psi_{u_r}^2(\omega_{i,r}) \mid \boldsymbol{\xi}_i\right] = u_r - u_r^2 + O_p(|\Delta|^{d+1}),$$

then we have

$$E\left[\boldsymbol{V}_3\right] = O_p(|\Delta|^{d+1}),$$

and

$$E\left[\boldsymbol{V}_3^2\right] \leq n_A^{-1}\sum_{r=1}^{n_A} u_r(1-u_r)E\left[(\boldsymbol{Z}_i(u_r) - \boldsymbol{Z}_i^*(u_r))^2\right] = O_p(n^{-1}|\Delta|^2),$$

which implies that

$$\boldsymbol{V}_3 = O_p(n^{-1/2}|\Delta|).$$

Based on the above calculation and Lemma 8, under the conditions of Theorem 1, by Slutsky's theorem we have

$$\sqrt{n}\boldsymbol{V} \to N(\boldsymbol{0}, \boldsymbol{U}_2/n) \tag{2.30}$$

in distribution. Finally we prove the asymptotic normality of $\hat{\beta}(t,u) - \beta(t,u)$. Recall that $\hat{\beta}(t,u) - \beta(t,u)$ admits the following decomposition,

$$\hat{\beta}(t,u) - \beta(t,u) = -\frac{1}{2n}\tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}_2^{-1}\sqrt{n}\boldsymbol{V} + \beta^*(t,u) - \beta(t,u) + o_p(1).$$

Under the conditions of Theorem 1, the bias of $\hat{\beta}(t,u) - \beta(t,u)$ is asymptotically negligible. Then by (2.30) and Slutsky's theorem, we can conclude that

$$\frac{1}{\sqrt{\sigma_\beta(t,u)}}\left(\hat{\beta}(t,u) - \beta(t,u)\right) \to N(0,1)$$

in distribution, where $\sigma_\beta(t,u) = \tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(t,u)$.

### 2.7.3 Proof of Theorem 3

Define $\vartheta(t,u) = \sigma_\beta(t,u)^{-1/2}\tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{U}_2^{1/2}n^{-1/2}\sum_{i=1}^n \tilde{\boldsymbol{Z}}_i$, where $\tilde{\boldsymbol{Z}}_1,\ldots,\tilde{\boldsymbol{Z}}_n \overset{iid}{\sim} N(0,\boldsymbol{I})$ and the dimension of $\tilde{\boldsymbol{Z}}_i$ is same as the dimension of $\boldsymbol{Z}_i$, $i = 1,\ldots,n$. Note that $\vartheta(t,u)$ is a Gaussian random field with $E\left[\vartheta(t,u)\right] = 0$ and $Var\left[\vartheta(t,u)\right] = 1$ for any $(t,u)$, and its covariance function is given by

$$Cov\left[\vartheta(t,u),\vartheta(t',u')\right] = \sigma_\beta^{-1/2}(t,u)\sigma_\beta^{-1/2}(t',u')\tilde{\boldsymbol{B}}^{\mathrm{T}}(t,u)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(t',u').$$

For part (1), similar to the proofs for Theorem 3 in [56], Theorem 3 in [28], and Theorem 5 in [3], by the strong approximation theorem [14], we can prove that

$$\sup_{t,u}\left|\sigma_\beta^{-1/2}(t,u)\left\{\hat{\beta}(t,u) - \beta(t,u)\right\} - \vartheta(t,u)\right| = o_p(1). \tag{2.31}$$

For part (2), based on the triangulation, we first partition $\boldsymbol{\Omega}$,the domain of $\beta(t,u)$, into $M$ triangles with vertices $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_{J_M}$. Then we can construct the simultaneous confidence regions (SCRs) for the estimator $\hat{\beta}(t,u)$ over a subset of $\Omega$, $\Omega_s = (\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{J_M})$.

For any $\boldsymbol{v}_j$ and $\boldsymbol{v}_{j'} \in \Omega_s$, we notice that

$$\left|Cov\left(\vartheta(\boldsymbol{v}_j),\vartheta(\boldsymbol{v}_{j'})\right)\right| = \begin{cases} 0, & \Delta_j \neq \Delta_{j'} \\ 1, & j = j' \\ \sigma_\beta^{-1/2}(\boldsymbol{v}_j)\sigma_\beta^{-1/2}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'}), & j \neq j', \Delta_j = \Delta_{j'}. \end{cases}$$

By definition of $\sigma_\beta(t,u)$, for any $v_j \in \Omega_s$, we have $\sigma_\beta(\boldsymbol{v}_j) = \mathrm{tr}\left\{\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\boldsymbol{v}_j)\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right\}$, and therefore, $\lambda_{min}\left(\boldsymbol{\Sigma}\right)\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_j)\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right) \leq \sigma_\beta(\boldsymbol{v}_j) \leq \lambda_{max}\left(\boldsymbol{\Sigma}\right)\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_j)\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right)$, where $\lambda_{min}(\cdot)$ and

$\lambda_{max}(\cdot)$ represent the minimum and maximum eigenvalues of the matrix. Then,

$$\begin{aligned}
&\sigma_\beta^{-1/2}(\boldsymbol{v}_j)\sigma_\beta^{-1/2}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'}) \\
=&\sigma_\beta^{-1/2}(\boldsymbol{v}_j)\sigma_\beta^{-1/2}(\boldsymbol{v}_{j'})\mathrm{tr}\left(\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\right) \\
=&\sigma_\beta^{-1/2}(\boldsymbol{v}_j)\sigma_\beta^{-1/2}(\boldsymbol{v}_{j'})\mathrm{tr}\left(\boldsymbol{\Sigma}\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right) \\
\leq&\sigma_\beta^{-1/2}(\boldsymbol{v}_j)\sigma_\beta^{-1/2}(\boldsymbol{v}_{j'})\lambda_{max}\left\{\boldsymbol{\Sigma}\right\}\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right) \\
\leq&\lambda_{min}\left\{\boldsymbol{\Sigma}\right\}^{-1}\lambda_{max}\left\{\boldsymbol{\Sigma}\right\} \\
&\times\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_j)\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right)^{-1/2}\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_{j'})\right)^{-1/2}\mathrm{tr}\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\right) \\
=&\lambda_{min}\left\{\boldsymbol{\Sigma}\right\}^{-1}\lambda_{max}\left\{\boldsymbol{\Sigma}\right\}\frac{\tilde{\boldsymbol{B}}^{\mathrm{T}}(\boldsymbol{v}_j)\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})}{\|\tilde{\boldsymbol{B}}(\boldsymbol{v}_j)\|\|\tilde{\boldsymbol{B}}(\boldsymbol{v}_{j'})\|} \\
\leq&\lambda_{min}\left\{\boldsymbol{\Sigma}\right\}^{-1}\lambda_{max}\left\{\boldsymbol{\Sigma}\right\}.
\end{aligned}$$

Since this upper bound does not depend on the location of $\boldsymbol{v}_j$ and $\boldsymbol{v}_{j'}$, then there must exist constants $c_1$ and $c_2$ such that

$$\lambda_{min}\left\{\boldsymbol{\Sigma}\right\}^{-1}\lambda_{max}\left\{\boldsymbol{\Sigma}\right\}\leq c_1c_2^{J_M}\leq c_1c_2^{j-j'},$$

for any $1\leq j,j'\leq J_M$. Now combined with Lemma 1 in [57], we can conclude that for any $a\in(0,1)$,

$$\lim_{n\to\infty}P\left\{\sup_j|\vartheta(\boldsymbol{v}_j)|\leq Q_\beta(a)\right\}=1-a,$$

where $Q_\beta(a)=(2\log J_M)^{1/2}-(2\log J_M)^{-1/2}\left\{\log(-0.5\log(1-a))+0.5\left[\log\right.\right.$ $\left.\left.(\log J_M)+\log 4\pi\right]\right\}$.

# Chapter 3

# Convolution Smoothing Based Locally Sparse Estimation for Functional Quantile Regression

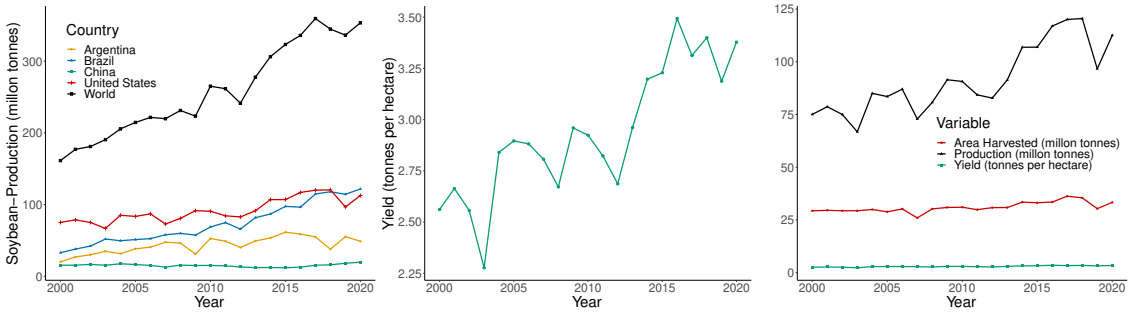
## 3.1. Introduction

In agriculture, crop yield is a key focus worldwide because of its direct connection to the global needs for food, feed, and fuel. In addition, crops are highly liquid in the futures market, so crop price fluctuations can directly affect the stability of financial markets. As one of the most important crops around the world, more than three-quarters of soybeans are used to feed livestock, and only a small percentage (about 7%) of global soybeans are used for typical soybean products such as tofu and soy milk. Meanwhile, the growing appetite for meat, dairy and soybean oil results in a rapidly increasing demand for soybean as shown in Figure 3.1-(a). There are two main ways to increase production, one is to expand the amount of land to grow soybean and the other is to improve soybean yields (increasing per area harvest). Taking data from the United States as an example, it is clear that the impressive improvement in soybean yields (Figure 3.1-(b)) is not able to keep up with the increasing demand for soybean production (Figure 3.1-(c)), which makes the government have to devote additional land to production. However, many scientists think increasing harvested area is a major underlying cause of deforestation. Therefore, it is urgent to improve soybean yields due to increasing product demand and environmental protection.

Soybean is a crop with high demands on the natural environment and resources, especially on temperature and water. Our study focuses on studying how soybean yield is affected by temperature, precipitation and irrigation. Our objective is to identify the time regions when there is a significant effect of daily temperature on the annual soybean yields. Motivated by this problem, we propose the following novel locally sparse semi-parametric functional quantile model,

$$Q_\tau(Y|\boldsymbol{Z}, \boldsymbol{X}(t)) = \boldsymbol{Z}^T\boldsymbol{\alpha}_\tau + \int_0^{\mathcal{T}} \boldsymbol{X}^T(t)\boldsymbol{\beta}_\tau(t)dt, \tag{3.1}$$

where $Q_\tau(Y|\boldsymbol{Z}, \boldsymbol{X}(t))$ is the $\tau$-th conditional quantile of the scalar response $Y$ given $\boldsymbol{X}(t) = (X_1(t), \ldots, X_m(t))^T$ and $\boldsymbol{Z} = (Z_0, \ldots, Z_d)^T$ for a fixed quantile level $\tau \in (0, 1)$. $\boldsymbol{\beta}_\tau(\cdot) = (\beta_{\tau,1}(\cdot), \ldots, \beta_{\tau,m}(\cdot))^T$ is a vector of functional coefficients and $\boldsymbol{\alpha}_\tau$ is a $d \times 1$ vector of coefficients. We set $Z_0 \equiv 1$ and use $\alpha_{\tau,1}$ to denote the intercept throughout this project. In our soybean application, $Y$ is the annual soybean yield per unit, $X_1(t)$ and $X_2(t)$ are the daily maximum and minimum temperature, respectively, and $Z_1$ and $Z_2$ are the annual precipitation and the ratio of irrigated area. All of these observations are measured at county level in Kansas.




(a) (b) (c)

Figure 3.1: (a) Soybean production data from 2000 to 2020 in the world and the four highest soybean-producing countries, namely the United States, Brazil, Argentina, and China. (b) Soybean yield in the United States from 2000 to 2020. (c) Comparison of soybean production and yield and area harvested in the United States from 2000 to 2020. These data are published by *the Food and Agriculture Organization of the United Nations*.

The functional coefficient $\beta_{\tau,l}(\cdot)$ is assumed to be locally sparse, which means that $\beta_{\tau,l}(t) = 0$ in some regions $\mathcal{N}$, where $\mathcal{N}$ is a subset of the whole time domain $[0, \mathcal{T}]$. The local sparsity of $\boldsymbol{\beta}_\tau(\cdot)$ can depict the dynamic dependence of the functional covariates on the $\tau$-th conditional quantile of the scalar response $Y$. Especially, when the identified locally sparse region $\mathcal{N}$ is identical to the domain $[0, \mathcal{T}]$, then the corresponding functional variable is not significantly related to the $\tau$-th conditional quantile of $Y$.

The proposed sparse semi-parametric functional quantile model (3.1) includes a variety of functional models as special cases. For example, when the identified sparse regions for all functional coefficient $\beta_{\tau,l}(t), l = 1, \ldots, m$, equal to $[0, \mathcal{T}]$, then model (3.1) becomes the classic quantile regression [44]. If $\alpha_{\tau,l} = 0$ for all $l = 1, \ldots, d$, and no functional coefficient $\beta_{\tau,l}$ has a locally sparse region, model (3.1) is reduced to the functional quantile regression (FQR) with functional covariates only [6, 11, 41]. Various partially functional quantile regression models are also special cases of the proposed model. For instance, [86] studied a model with multiple functional covariates and a finite number of scalar covariates. [85] consider a partially functional quantile regression with a functional covariate and high-dimensional scalar covariates. [55] proposed a functional partially linear model with multiple functional covariates and ultrahigh-dimensional scalar covariates, and imposed two nonconvex penalties to select the significant functional and scalar covariates. To the best of

our knowledge, no work has studied the local sparsity structure for functional quantile regression models, although it is important in various applications.

Most existing works that consider local sparsity feature are based on the functional linear models [80, 39, 94]. For example, [50] proposed a functional generalization of ordinary SCAD [19], called fSCAD, to obtain a locally sparse estimator for the slope function under univariate scalar-on-function regression. [22] and [49] applied fSCAD [50] to the multiple outputs functional linear regression. In [27], they used the B-spline expansion and group bridge penalty [35] to identify the non-zero region close to the endpoint. In addition, [62] developed a new regularization method for functional linear discriminant analysis to induce zero regions.

This project has four major contributions. First, we introduce a semi-parametric functional quantile model with a locally sparse structure. Second, we propose a Convolution smoothing based Locally Sparse Estimation (CLoSE) method to do three tasks simultaneously, including selecting significant functional covariates, identifying locally sparse regions to improve the interpretability of the model, and estimating functional coefficients in nonzero regions. In order to overcome the computational difficulty in minimizing the non-differentiable quantile loss combined with the non-convex fSCAD penalty, we choose to replace the standard quantile loss function with a smoothed version using convolution-type smoothing method. [24, 32, 76]. Third, we establish the oracle property of the proposed estimator, and derive a simultaneous confidence band (SCB) for the functional coefficients. The last but not the least, the analysis of the soybean data from Kansas shows that the time period in which the daily temperature has a significant impact on soybean yields varies with the quantile levels. In particular, we find that the significant functional predictors are different for different quantile levels.

## 3.2. Convolution Smoothing based Locally Sparse Estimation

### 3.2.1 Convolution-type Smoothing Approach

Suppose that $\{(\boldsymbol{Z}_i, \boldsymbol{X}_i(t), Y_i, t \in [0, \mathcal{T}]\}_{i=1}^n$ is an independent and identically distributed random sample from $(\boldsymbol{Z}, \boldsymbol{X}(t), Y, t \in [0, \mathcal{T}])$. Then the locally sparse FQR estimator can be obtained by minimizing the following loss function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau, \gamma_l, \lambda_l) =& \frac{1}{n}\sum_{i=1}^{n} \rho_\tau \left( Y_i - \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau - \int_0^{\mathcal{T}} \boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt \right) \\ &+ \sum_{l=1}^{m} \gamma_l \left\| \boldsymbol{\mathcal{D}}^q \beta_{\tau,l} \right\|_2^2 \\ &+ \text{Locally sparse penalty},\end{aligned} \tag{3.2}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile check function, and $I(\cdot)$ is an indicator function. But, the first-order derivative of $\rho(u)$, $\psi_\tau(u) = \dot{\rho}(u) = \tau - I(u < 0)$, is not a smooth function. In the loss

function (3.2), the second term is a convex roughness penalty with the $q$th-order differential operator $\boldsymbol{\mathcal{D}}^q$ and the non-negative tuning parameters $\gamma_l, l = 1, \ldots, m$, which control the smoothness of the estimated coefficient functions. For the third term, we choose the fSCAD penalty [50], which is a concave penalty, to simultaneously identify the zero regions of the estimated functional coefficients and select the significant functional covariates.

The loss function (3.2) has no closed-form solution. Iterative procedures are often adopted to find the optimal solution. However, in the iterative algorithm for minimizing the loss function (3.2), the combination of a second-order non-differentiable check function and a concave locally sparse penalty brings computation difficulty. More specifically, local quadratic approximation (LQA) is a commonly used strategy for the optimization that involves fSCAD or SCAD penalty [19, 50]. Such an algorithm usually requires the calculation of gradient and Hessian matrix, which is not available for (3.2) because of the non-differentiable check function. To make the computation fast and stable, we consider an alternative convolution smoothed quantile loss function [24, 32, 76].

Let $e(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) = Y - \boldsymbol{Z}^T\boldsymbol{\alpha}_\tau - \int_0^{\mathcal{T}} \boldsymbol{X}^T(t)\boldsymbol{\beta}_\tau(t)dt$. Denote the conditional cumulative distribution function (CDF) and density function of $e(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$ given $\boldsymbol{Z}$ and $\boldsymbol{X}(t)$ as $F_{e|\boldsymbol{Z},\boldsymbol{X}}(\cdot)$ and $f_{e|\boldsymbol{Z},\boldsymbol{X}}(\cdot)$, respectively. Then the population quantile loss can be expressed as

$$\mathbb{E}[\rho_\tau(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)] = \int \rho_\tau(u) dF_{e|\boldsymbol{Z},\boldsymbol{X}}(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau). \tag{3.3}$$

When the unknown CDF $F_{e|\boldsymbol{Z},\boldsymbol{X}}(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$ in (3.3) is replaced by the empirical distribution function $\widehat{F}(t; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) = 1/n \sum_{i=1}^n I(e_i(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) \leq u\}$ of the residuals $e_i(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) = Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \int_0^{\mathcal{T}} \boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt$, we can obtain the standard quantile loss, that is, the first term in (3.2). However, the empirical distribution function $\widehat{F}(t; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$ is a discontinuous function. Thus, a kernel smoothing estimator of $F_{e|\boldsymbol{Z},\boldsymbol{X}}(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$ is a better choice [24, 76]. The kernel density estimator is $\widehat{f}_h(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) = 1/n \sum \mathcal{K}_h(u - e_i(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)) = (nh)^{-1} \sum \mathcal{K}((u - e_i(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau))/h)$, where $\mathcal{K}(\cdot)$ is a kernel function, $h$ is a bandwidth and $\mathcal{K}_h(u) = 1/h\mathcal{K}(u/h)$. The corresponding kernel smoothing CDF estimator is $\widehat{F}_h(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau) = n^{-1} \sum G_h(t - e_i(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau))$, where $G_h(u) = G(u/h)$ and $G(u) = \int_{-\infty}^u \mathcal{K}(v)dv$. When we replace the unknown CDF by the kernel smoothing estimator of the CDF $\widehat{F}_h(u; \boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$, we derive a new loss function with the convolution smoothed quantile loss:

$$\begin{aligned}
\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau, \gamma_l, \lambda_l) =& \frac{1}{n} \sum_{i=1}^n (\rho_\tau * \mathcal{K}_h) \left( Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \int_0^{\mathcal{T}} \boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt \right) \\
&+ \sum_{l=1}^m \gamma_l \left\| \boldsymbol{\mathcal{D}}^q \beta_{\tau,l} \right\|_2^2 \\
&+ \sum_{l=1}^m \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} p_{\lambda_l}(|\beta_{\tau,l}(t)|)dt,
\end{aligned} \tag{3.4}$$

where $(\rho_\tau * \mathcal{K}_h)(u) = \int_{-\infty}^{\infty} \rho_\tau(v)\mathcal{K}_h(v - u)dv$ and $*$ is the convolution operator. The smoothed quantile loss is a globally convex function of $u$, which has first and second-order derivatives to make

the computation faster. Here $p_{\lambda_l}(u)$ is the SCAD penalty [19] defined as

$$p_{\lambda_l}(u) = \begin{cases} \lambda_l u & 0 \le u \le \lambda_l \\ -\dfrac{u^2 - 2a\lambda_l u + \lambda_l^2}{2(a-1)} & \lambda_l \le u \le a\lambda_l \\ \dfrac{(a+1)\lambda_l^2}{2} & u \ge a\lambda_l \end{cases},$$

with the non-negative tunning parameters $a$.

### 3.2.2 Estimation Procedure

We use B-spline basis functions to represent the functional coefficients $\beta_{\tau,l}(t), l = 1, \dots m$. We set the order to be $p+1$ and place $K+1$ equally spaced knots $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = \mathcal{T}$ in the domain $[0, \mathcal{T}]$ to define a set of B-spline basis functions. Then the functional coefficient $\beta_{\tau,l}(t)$ can be approximated by the B-spline basis functions:

$$\beta_{\tau,l}(t) \approx \boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}, \tag{3.5}$$

where $\boldsymbol{\mathcal{B}}(t) = (B_1(t), \dots, B_{K+p}(t))^T$ is the vector of B-spline basis functions and $\boldsymbol{\theta}_{\tau,l} = (\theta_{\tau,l,1}, \dots, \theta_{\tau,l,K+p})^T$ is the corresponding coefficients. Denote $\boldsymbol{U}_i = \int_o^{\mathcal{T}} \boldsymbol{X}_i(t) \otimes \boldsymbol{\mathcal{B}}(t)dt$ and $\boldsymbol{\theta}_\tau^T = (\boldsymbol{\theta}_{\tau,1}^T, \cdots, \boldsymbol{\theta}_{\tau,m}^T)$, then the first term of $\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \beta_\tau, \gamma, \lambda)$ can be written as $n^{-1}\sum_{i=1}^n (\rho_\tau * \mathcal{K}_h)(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau)$. The roughness penalty in (3.4) can be expressed as

$$\sum_{l=1}^m \gamma_l \left\| \left( \boldsymbol{\mathcal{B}}^{(q)}(t) \right)^T \boldsymbol{\theta}_{\tau,l} \right\|_2^2 = \sum_{l=1}^m \gamma_l \boldsymbol{\theta}_{\tau,l}^T \boldsymbol{V} \boldsymbol{\theta}_{\tau,l} = \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau, \tag{3.6}$$

where $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \cdots, \gamma_m)$, $\boldsymbol{V} = \int_0^{\mathcal{T}} [\boldsymbol{\mathcal{B}}^{(q)}(t)][\boldsymbol{\mathcal{B}}^{(q)}(t)]^T dt$.

According to Theorem 1 in [50], as $K \to \infty$, the fSCAD penalty term can be approximated by

$$\sum_{l=1}^m \frac{K}{\mathcal{T}} \int_0^{\mathcal{T}} p_{\lambda_l}(|\beta_{\tau,l}(t)|)dt \approx \sum_{l=1}^m \sum_{j=1}^K p_{\lambda_l} \left( \sqrt{\frac{K}{\mathcal{T}} \int_{t_{j-1}}^{t_j} \beta_{\tau,l}^2(t)dt} \right). \tag{3.7}$$

Plugging the spline approximation into $\int_{t_{j-1}}^{t_j} \beta_\tau^2(t)dt$, we can get the matrix representation, $\int_{t_{j-1}}^{t_j} \beta_{\tau,l}^2(t)dt = \boldsymbol{\theta}_{\tau,l}^T \boldsymbol{W}_j \boldsymbol{\theta}_{\tau,l}$ , where $\boldsymbol{W}_j$ is an $(K+p)$ by $(K+p)$ matrix with entries $w_{uv} = \int_{t_{j-1}}^{t_j} B_u(t)B_v(t)dt$ if $j \le u, v \le j+p$ and zero otherwise. Thus the fSCAD penalty (3.7) can be rewritten as

$$\sum_{l=1}^m \sum_{j=1}^K p_{\lambda_l} \left( \sqrt{\frac{K}{\mathcal{T}} \int_{t_{j-1}}^{t_j} \beta_{\tau,l}^2(t)dt} \right) = \sum_{l=1}^m \sum_{j=1}^K p_{\lambda_l} \left( \sqrt{\frac{K}{\mathcal{T}} \boldsymbol{\theta}_{\tau,l}^T \boldsymbol{W}_j \boldsymbol{\theta}_{\tau,l}} \right). \tag{3.8}$$

Then, combined with (3.5), (3.6) and (3.8), we can rewrite the loss function (3.4) as follows,

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau, \gamma_l, \lambda_l) \approx & \frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau * \mathcal{K}_h\right)\left(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau\right) \\ & + \boldsymbol{\theta}_\tau^T(\boldsymbol{\Gamma}\otimes\boldsymbol{V})\boldsymbol{\theta}_\tau \\ & + \sum_{l=1}^{m}\sum_{j=1}^{K} p_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\boldsymbol{\theta}_{\tau,l}^T\boldsymbol{W}_j\boldsymbol{\theta}_{\tau,l}}\right).\end{aligned} \tag{3.9}$$

The SCAD penalty $p_{\lambda_l}(\cdot)$ is not differentiable, which brings difficulty in the optimization. There are two ways to approximate $p_{\lambda_l}(\cdot)$, local quadratic approximation (LQA; [19]) and local linear approximation (LLA; [95]). [50] found that LLA does not work well with $L^q$ norm of functions and they suggested using LQA in their paper. Therefore, we also choose the LQA to approximate the fSCAD penalty. When $u \approx u^{(0)}$, the LQA of the SCAD function $p_{\lambda_l}(u)$ is:

$$\begin{aligned}p_{\lambda_l}(|u|) &\approx p_{\lambda_l}(|u^{(0)}|) + \frac{1}{2}\frac{\dot{p}_{\lambda_l}(|u^{(0)}|)}{|u^{(0)}|}(u^2 - u^{(0)2}) \\ &= \frac{1}{2}\frac{\dot{p}_{\lambda_l}(|u^{(0)}|)}{|u^{(0)}|}u^2 + p_{\lambda_l}(|u^{(0)}|) - \frac{1}{2}\frac{\dot{p}_{\lambda_l}(|u^{(0)}|)}{|u^{(0)}|}u^{(0)2} \\ &\triangleq \frac{1}{2}\frac{\dot{p}_{\lambda_l}(u^{(0)})}{|u^{(0)}|}u^2 + R_1(u^{(0)}),\end{aligned} \tag{3.10}$$

where $\dot{p}_{\lambda_l}(u)$ is the first order derivative of $p_{\lambda_l}(u)$ and $R_1(u^{(0)})$ is a constant which only depends on $u^{(0)}$.

Then given some initial estimator $\widehat{\boldsymbol{\theta}}_\tau^{(0)} = (\widehat{\boldsymbol{\theta}}_{\tau,1}^{(0)T}, \cdots, \widehat{\boldsymbol{\theta}}_{\tau,m}^{(0)T})^T$, when $\boldsymbol{\theta}_\tau \approx \widehat{\boldsymbol{\theta}}_\tau^{(0)}$, we have

$$\begin{aligned}\sum_{l=1}^{m}\sum_{j=1}^{K} p_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\boldsymbol{\theta}_{\tau,l}^T\boldsymbol{W}_j\boldsymbol{\theta}_{\tau,l}}\right) &\approx \frac{1}{2}\sum_{l=1}^{m}\sum_{j=1}^{K}\frac{\dot{p}_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}\right)}{\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}}\frac{\boldsymbol{\theta}_{\tau,l}^T\boldsymbol{W}_j\boldsymbol{\theta}_{\tau,l}}{\mathcal{T}/K} + R_2(\widehat{\boldsymbol{\theta}}_\tau^{(0)}) \\ &\approx \frac{1}{2}\sum_{l=1}^{m}\sum_{j=1}^{K}\frac{\dot{p}_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}\right)}{\sqrt{\frac{\mathcal{T}}{K}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}}\boldsymbol{\theta}_{\tau,l}^T\boldsymbol{W}_j\boldsymbol{\theta}_{\tau,l} + R_2(\widehat{\boldsymbol{\theta}}_\tau^{(0)}),\end{aligned} \tag{3.11}$$

where

$$\begin{aligned}R_2(\widehat{\boldsymbol{\theta}}_\tau^{(0)}) = & \sum_{l=1}^{m}\sum_{j=1}^{K} p_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}\right) \\ & - \frac{1}{2}\sum_{l=1}^{m}\sum_{j=1}^{K}\dot{p}_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}\right)\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}},\end{aligned}$$

only depends on the initial estimator $\widehat{\boldsymbol{\theta}}_\tau^{(0)}$. Denote

$$\boldsymbol{W}_{\tau,l}^{(0)} = \frac{1}{2}\sum_{j=1}^{K} \frac{\dot{p}_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}\right)}{\sqrt{\frac{\mathcal{T}}{K}\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)T}\boldsymbol{W}_j\widehat{\boldsymbol{\theta}}_{\tau,l}^{(0)}}}\boldsymbol{W}_j, \quad \text{and} \quad \boldsymbol{W}_\tau^{(0)} = \begin{pmatrix} \boldsymbol{W}_{\tau,1}^{(0)} & & \\ & \ddots & \\ & & \boldsymbol{W}_{\tau,m}^{(0)} \end{pmatrix},$$

we can get

$$\sum_{l=1}^{m}\sum_{j=1}^{K} p_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}}\boldsymbol{\theta}_{\tau,l}^T\boldsymbol{W}_j\boldsymbol{\theta}_{\tau,l}}\right) \approx \boldsymbol{\theta}_\tau^T\boldsymbol{W}_\tau^{(0)}\boldsymbol{\theta}_\tau + R_2(\widehat{\boldsymbol{\theta}}_\tau^{(0)}).$$

Thus we can express the loss function (3.9) as follows,

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau, \gamma_l, \lambda_l \mid \widehat{\boldsymbol{\theta}}_\tau^{(0)}) \approx & \frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau * \mathcal{K}_h\right)\left(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau\right) \\ & + \boldsymbol{\theta}_\tau^T(\boldsymbol{\Gamma}\otimes\boldsymbol{V})\boldsymbol{\theta}_\tau \\ & + \boldsymbol{\theta}_\tau^T\boldsymbol{W}_\tau^{(0)}\boldsymbol{\theta}_\tau + R_2(\widehat{\boldsymbol{\theta}}_\tau^{(0)}).\end{aligned} \tag{3.12}$$

Denote $\mathcal{L}_0^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) = 1/n\sum_{i=1}^{n}\left(\rho_\tau * \mathcal{K}_h\right)\left(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau\right)$, we can get the first-order and second-order derivatives of $\mathcal{L}_0^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)$ respectively, that is,

$$\begin{aligned}\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) &= 1/n\sum_{i=1}^{n}\left\{G_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau - Y_i) - \tau\right\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T, \\ \ddot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) &= 1/n\sum_{i=1}^{n}\mathcal{K}_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau - Y_i)(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T).\end{aligned}$$

With the smoothed quantile loss and LQA on the fSCAD penalty, we are able to calculate its gradient and Hessian matrix. Then an iterative Newton–Raphson-type algorithm can be used to solve the optimization problem. More specifically, for a fixed $\boldsymbol{W}_\tau^{(k)}$, we use a Newton–Raphson type algorithm to solve the minimization of $\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau, \gamma_l, \lambda_l \mid \widehat{\boldsymbol{\theta}}_\tau^{(k)})$ with respect to $(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)$. After it converges, we update $\boldsymbol{W}_\tau^{(k)}$ to $\boldsymbol{W}_\tau^{(k+1)}$ and then minimize $\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau, \gamma_l, \lambda_l \mid \widehat{\boldsymbol{\theta}}_\tau^{(k+1)})$. We repeat this procedure until the sequences of minimizers $(\boldsymbol{\alpha}_\tau^{(j)}, \boldsymbol{\theta}_\tau^{(j)})$ converge. Note that in practice, the term $\dot{p}_{\lambda_l}(u^{(0)})/|u^{(0)}|$ in (3.10) can go to infinity if $|u^{(0)}|$ is very small, which can cause the algorithm to be unstable. Following [36], in our algorithm we use a perturbed version of LQA for $p_{\lambda_l}(\cdot)$. We summarize the computational details in Algorithm 1. Note that for each iteration, $\eta_{j,k}$ is chosen such that the objective function decreases after the update.

Let $S_1$ and $S_2$ denote the candidate sets for the tuning parameters $\lambda_l$ and $\gamma_l$, respectively. We tune these parameters based on the following strategy. For a given pair $(\lambda_l, \gamma_l)$, we first fit the model (3.12) to identify the estimated null and nonnull subregions. Next, we remove the information on the functional covariates within the null subregions and refit the model without the fSCAD penalty. Then, we apply the Bayesian information criterion (BIC) [47] to find the best pair $(\lambda_l, \gamma_l)$.

**Algorithm 1:** Algorithm for CLoSE Method

**Input :** Data, quantile level $\tau$, bandwidth $h$

**Initialization**:

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_\tau^{(0)} \\ \widehat{\boldsymbol{\theta}}_\tau^{(0)} \end{pmatrix} = \text{argmin}_{\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau} \left\{ \mathcal{L}_0^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) + \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau \right\},$$

where the tuning parameters $\gamma_l$ in the matrix $\boldsymbol{\Gamma}$ are selected by cross validation;

**while** *not converged* **do**

  **while** *not converged* **do**

    Let $\widehat{\boldsymbol{\alpha}}_\tau^{(0,k_2)} = \widehat{\boldsymbol{\alpha}}_\tau^{(k_2)}$ and $\widehat{\boldsymbol{\theta}}_\tau^{(0,k_2)} = \widehat{\boldsymbol{\theta}}_\tau^{(k_2)}$;

    Update $\widehat{\boldsymbol{\alpha}}_\tau^{(k_1,k_2)}$ and $\widehat{\boldsymbol{\theta}}_\tau^{(k_1,k_2)}$ as follows:

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_\tau^{(k_1,k_2)} \\ \widehat{\boldsymbol{\theta}}_\tau^{(k_1,k_2)} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_\tau^{(k_1-1,k_2)} \\ \widehat{\boldsymbol{\theta}}_\tau^{(k_1-1,k_2)} \end{pmatrix} - \eta_{k_1,k_2} \left\{ \boldsymbol{D}_2^{(k_1-1,k_2)} \right\}^{-1} \boldsymbol{D}_1^{(k_1-1,k_2)},$$

    where

$$\boldsymbol{D}_1^{(k_1-1,k_2)} = \dot{\mathcal{L}}_0^* \left( \widehat{\boldsymbol{\alpha}}_\tau^{(k_1-1,k_2)}, \widehat{\boldsymbol{\theta}}_\tau^{(k_1-1,k_2)} \right) + 2 \begin{pmatrix} \mathbf{0}_{d \times d} & \\ & \boldsymbol{W}_\tau^{(k_2-1)} + \boldsymbol{\Gamma} \otimes \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_\tau^{(k_1-1,k_2)} \\ \widehat{\boldsymbol{\theta}}_\tau^{(k_1-1,k_2)} \end{pmatrix},$$

$$\boldsymbol{D}_2^{(k_1-1,k_2)} = \ddot{\mathcal{L}}_0^* \left( \widehat{\boldsymbol{\alpha}}_\tau^{(k_1-1,k_2)}, \widehat{\boldsymbol{\theta}}_\tau^{(k_1-1,k_2)} \right) + 2 \begin{pmatrix} \mathbf{0}_{d \times d} & \\ & \boldsymbol{W}_\tau^{(k_2-1)} + \boldsymbol{\Gamma} \otimes \boldsymbol{V} \end{pmatrix}.$$

    Let $k_1 := k_1 + 1$.

  **end**

  The limit is denoted as $\widehat{\boldsymbol{\alpha}}_\tau^{(k_2)}$ and $\widehat{\boldsymbol{\theta}}_\tau^{(k_2)}$

  Update $\boldsymbol{W}_\tau^{(k_2)}$;

  Let $k_2 := k_2 + 1$;

**end**

**Output:** The final estimators $\widehat{\boldsymbol{\alpha}}_\tau$ and $\widehat{\boldsymbol{\theta}}_\tau$

## 3.3. Large Sample Properties

For any vector, we shall use $\|\cdot\|_2$ to denote the Euclidean norm. The null region and nonnull region of $\beta_{\tau,l}(t), l = 1,\ldots,m$ are denoted by $\mathcal{N}(\beta_{\tau,l})$ and $\mathcal{S}(\beta_{\tau,l})$, respectively, where $\mathcal{N}(\beta_{\tau,l}) = \{t \in [0,\mathcal{T}] : \beta_{\tau,l}(t) = 0\}$ and $\mathcal{N}(\beta_{\tau,l}) = \{t \in [0,\mathcal{T}] : \beta_{\tau,l}(t) \neq 0\}$. $\boldsymbol{\mathcal{B}}_{\tau,l,1}(t)$ denotes the $K^*_{\tau,l}$ dimensional sub-vector of $\boldsymbol{\mathcal{B}}(t)$ such that each $B_j(t)$ in $\boldsymbol{\mathcal{B}}_{\tau,l,1}(t)$ has a support inside $\mathcal{S}(\beta_{\tau,l})$. Let $\boldsymbol{U}^*_\tau$ and $\boldsymbol{U}^*_{\tau,i}$ be a $K^*_\tau = \sum_{l=1}^m K^*_{\tau,l}$ dimensional vector of the corresponding elements of $\boldsymbol{U}$ and $\boldsymbol{U}_i$, respectively, associated with $\boldsymbol{\mathcal{B}}_{\tau,1,1},\ldots,\boldsymbol{\mathcal{B}}_{\tau,m,1}$. We also define $\boldsymbol{Z}^*_\tau = (\boldsymbol{Z}^T, \boldsymbol{U}^{*T}_\tau)^T$, $\boldsymbol{\Sigma}_{\tau,1} = E\left\{\boldsymbol{Z}^*_\tau \boldsymbol{Z}^{*T}_\tau\right\}$ and $\boldsymbol{\Sigma}_{\tau,2} = E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\boldsymbol{Z}^*_\tau \boldsymbol{Z}^{*T}_\tau\right\}$.

### 3.3.1 Conditions

To establish the asymptotic results of the estimated parameters, and the oracle properties of the estimated functional coefficients, we first need to give some Conditions.

**Definition 1.** (i) For the functional predictor $X_l(t), l = 1,\ldots,m$, it holds that $\|X_l\|_2$ is almost surely bounded, where $\|X_l\|_2^2 = \int_0^{\mathcal{T}} X_l^2(t)dt$. Moreover, $\rho_{min}(\boldsymbol{U}\boldsymbol{U}^T)/K^{-1}$ and $\rho_{max}(\boldsymbol{U}\boldsymbol{U}^T)/K^{-1}$ are bounded away from 0 and $\infty$ as $n \to \infty$, where $\rho_{\min}(\boldsymbol{A})$ and $\rho_{\max}(\boldsymbol{A})$ denote the minimal and maximal eigenvalues of the matrix $\boldsymbol{A}$, respectively. (ii) For the scalar predictors, the components of $\boldsymbol{Z}$ have bounded support. $E(\boldsymbol{Z}\boldsymbol{Z}^T)$ is positive definite and has eigenvalues bounded away from zero.

**Definition 2.** Let $\upsilon$ be a nonnegative integer, and $\kappa \in (0,1]$ such that $r = \upsilon + \kappa \geq p+1$. We assume the unknown functional coefficient $\beta_{\tau,l}(\cdot) \in \mathcal{H}^{(r)}(\mathcal{S}(\beta_{\tau,l}))$, which is the class of function $f$ on $\mathcal{S}(\beta_{\tau,l})$ whose $\upsilon$th deriative exists and satisfies a Lipschitz condition of order $\kappa$: $|f^{(\upsilon)}(t) - f^{(\upsilon)}(s)| \leq C_\upsilon |s-t|^\kappa$, for $s,t \in \mathcal{S}(\beta_{\tau,l})$ and some constant $C_\upsilon > 0$.

**Definition 3.** The conditional density function $f_{e|\boldsymbol{Z},\boldsymbol{X}}(\cdot)$ is bounded away from zero, and second times continuously differentiable.

**Definition 4.** Suppose $\mathcal{K}(\cdot)$ is a symmetric, bounded, continuous and non-negative function integrating to one, which means that $\mathcal{K}(u) = \mathcal{K}(-u)$, $\mathcal{K} \geq 0$ and $\int_{-\infty}^{\infty} \mathcal{K}(u) = 1$ for all $u \in \mathbb{R}$. Furthermore, $\mathcal{K}(\cdot)$ is second-order continuously differentiable and bounded.

**Definition 5.** Let $\delta_j = k_{j+1} - k_j$ and $\delta = \max_{0\leq j\leq K}(k_{j+1} - k_j)$. There exists a constant $M > 0$, such that

$$\delta / \min_{0\leq j\leq K}(k_{j+1} - k_j) \leq M, \quad \max_{0\leq j\leq K-1} |\delta_{j+1} - \delta_j| = o(K^{-1}). \tag{3.13}$$

**Definition 6.** The number of knots $K = o(\sqrt{n})$ and $K = \omega(n^{1/(2r+1)})$, where $K = \omega(n^{1/(2r+1)})$ means $K/n^{1/(2r+1)} \to \infty$ as $n \to \infty$.

**Definition 7.** For the roughness penalty, we assume tuning parameter $\gamma_l, l = 1\ldots,m$ satisfies that $\gamma_l = o(n^{1/2}K^{1/2-2q})$, where $q \leq p$.

**Definition 8.** The positive bandwidth $h$ satisfies $nh^4 \to 0$ and $hK^r \to 0$.

**Definition 9.** For the fSCAD penalty term, $\lambda_l \to 0$ as $n \to \infty$, (i) $\max_l \sqrt{\int_{\mathcal{S}(\beta^0_{\tau,l})} \dot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}|)^2 dt} = o(n^{-1/2}K^{-1})$, and $\max_l \sqrt{\int_{\mathcal{S}(\beta^0_{\tau,l})} \ddot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}|)^2 dt} = o(1)$. (ii) $K^{1/2}n^{-1/2}\lambda_l^{-1} \to 0$ and $\liminf_{n\to\infty} \liminf_{u\to 0^+} \dot{p}_{\lambda_l}(u)\lambda_l^{-1} > 0$.

**Remark 4.** Condition 1-(i) gives some moment conditions for functional predictors and is essential to obtain the oracle property of the estimators. For the scalar predictors, Condition 1-(ii) imposes some moment conditions. Condition 2 is about the smoothness of the functional coefficients $\beta_{\tau,l}(t)$, which has been widely used in the literature of nonparametric estimation. The common conditions on the conditional density function of $e(\boldsymbol{\alpha}_\tau, \boldsymbol{\beta}_\tau)$ in the quantile regression context is given in Condition 3. Condition 4 is a necessary condition on the kernel function, which is also required in [24], [32] and [76]. Condition 5 gives the assumption about the knots used in B-spline approximation, which implies that $\delta \sim K^{-1}$, i.e., $\delta$ and $K^{-1}$ are rate-wise equivalent from (3.13). Condition 6 is imposed to make the spline approximation bias asymptotically negligible. Condition 7 can make the shrinkage bias negligible brought by roughness penalty. Condition 8 ensures that the convolution smoothing has an asymptotically negligible bias on the estimator of $\boldsymbol{\alpha}^0_\tau$ and $\beta_{\tau,l}$. Condition 9-(i) is the regularity condition to ensure the bias brought by the sparsity penalty is asymptotically negligible [20], and Condition 9-(ii) is used to obtain the functional oracle property of $\hat{\beta}_{\tau,l}(t)$.

### 3.3.2 Functional Oracle Property and Asymptotic Normality

**Theorem 1.** *Under Conditions 1-9, there exists a local minimizer $(\widehat{\boldsymbol{\alpha}}_\tau, \widehat{\boldsymbol{\beta}}_\tau)$ of (3.4) such that $\|\widehat{\boldsymbol{\alpha}}_\tau - \boldsymbol{\alpha}_\tau\|_2 = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\|_2 = O_p(n^{-1/2}K^{1/2})$.*

From this theorem, it is clear that there exists a root-$n$ consistent estimator $\widehat{\boldsymbol{\alpha}}_\tau$ and a root-$n/M$ consistent estimator $\widehat{\boldsymbol{\beta}}_\tau(t)$. We then give the functional oracle property of $\widehat{\boldsymbol{\beta}}_\tau(t)$ and the asymptotic result of $\widehat{\boldsymbol{\alpha}}_\tau$.

**Theorem 2 (Functional Oracle Property).** *If Conditions 1-9 hold, as $n \to \infty$:*

*(i)* ***Locally Sparsity****: For every $t$ not in the support of $\beta_{\tau,l}(t)$, we have $\widehat{\beta}_{\tau,l}(t) = 0$ with probability tending to one.*

*(ii)* ***Asymptotic Normality****: For $t$ such that $\beta_{\tau,l}(t) \neq 0$ we have*

$$\sigma^{-1/2}_{\tau,l}(t)(\widehat{\beta}_{\tau,l}(t) - \beta_{\tau,l}(t)) \xrightarrow{d} \mathbb{G}(t), \tag{3.14}$$

*where $\sigma_{\tau,l}(t) = \tau(1-\tau)\boldsymbol{\Lambda}_{\tau,l}(t)\boldsymbol{\Sigma}^{-1}_{\tau,2}(\boldsymbol{\Sigma}_{\tau,1}/n)\boldsymbol{\Sigma}^{-1}_{\tau,2}\boldsymbol{\Lambda}^T_{\tau,l}(t)$, $\boldsymbol{\Lambda}_{\tau,l}(t) = \boldsymbol{\xi}_l^T\widetilde{\boldsymbol{\mathcal{B}}}_\tau(t)(\mathbf{0}_{K^*_\tau \times d}, \boldsymbol{I}_{K^*_\tau})$, $\boldsymbol{\xi}_l$ is a $m \times 1$ unit vector in which the $l$-th element is 1, $\boldsymbol{I}_d$ be a $d \times d$ identity matrix, $\mathbf{0}_{d\times K^*_\tau}$ is a $d \times K^*_\tau$ zero matrix,*

$$\widetilde{\boldsymbol{\mathcal{B}}}_{\tau,1}(t) = Bdiag(\boldsymbol{\mathcal{B}}^T_{\tau,1,1}(t), \cdots \boldsymbol{\mathcal{B}}^T_{\tau,m,1}(t)),$$

*and $\mathbb{G}(t)$ is a Gaussian random process with mean 0 defined on $S(\beta_{\tau,l})$ with the covariance function*

$$\mathcal{C}(t,s) = \tau(1-\tau)\sigma_{\tau,l}^{-1/2}(t)\sigma_{\tau,l}^{-1/2}(s)\mathbf{\Lambda}_{\tau,l}(t)\mathbf{\Sigma}_{\tau,2}^{-1}(\mathbf{\Sigma}_{\tau,1}/n)\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Lambda}_{\tau,l}^{T}(s).$$

*Consequently, for any $a \in (0,1)$,*

$$\lim_{n\to\infty} P\left\{\sup_{t\in\mathcal{S}_\varepsilon(\beta_{\tau,l})}\left|\sigma_{\tau,l}^{-1/2}(t)\left\{\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t)\right\}\right| \le \mathcal{Q}_{\tau,l}(a)\right\} = 1-a,$$

*where $\mathcal{S}_\varepsilon(\beta_{\tau,l})$ as a subset of $\mathcal{S}(\beta_{\tau,l})$ becomes denser as $n\to\infty$, $\mathcal{Q}_{\tau,l}(a) = (2\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|)^{1/2}$ $-(2\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|)^{-1/2}\{\log(-0.5\log(1-a)) + 0.5\,[\log(\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|) + \log(4\pi)]\}$, and $|\mathcal{S}_\varepsilon(\beta_{\tau,l})|$ denote the cardinality of the set $|\mathcal{S}_\varepsilon(\beta_{\tau,l})|$. Then an asymptotic $100(1-a)\%$ simultaneous confidence band (SCB) for $\beta_{\tau,l}(t)$ over $\mathcal{S}_\varepsilon(\beta_{\tau,l})$ is given by $\widehat{\beta}_{\tau,l}(t) \pm \sigma_{\tau,l}^{1/2}(t)\mathcal{Q}_{\tau,l}(a)$.*

**Theorem 3 (Asymptotic Normality of the Estimators of Parameters).** *If Conditions 1-9 hold, as $n\to\infty$, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_\tau - \boldsymbol{\alpha}_\tau^0) \xrightarrow{d} \mathbb{N}(\mathbf{0}, \tau(1-\tau)\boldsymbol{\mathcal{I}}_\tau\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Sigma}_{\tau,1}\mathbf{\Sigma}_{\tau,2}^{-1}\mathcal{I}_\tau^T), \tag{3.15}$$

*where $\boldsymbol{\alpha}_\tau^0$ is the true parameters and $\boldsymbol{\mathcal{I}}_\tau = (\boldsymbol{I}_d, \mathbf{0}_{d\times K_\tau^*})$.*

**Remark 5.** Theorem 2 shows that the estimators $\widehat{\boldsymbol{\beta}}_\tau(t)$ possess the functional version oracle property, which makes the fitted model simpler and more interpretable. From Theorem 2, for any given fixed point $t$ such that $\beta_{\tau,l}(t) \neq 0$, we have $\sigma_{\tau,l}^{-1/2}(t)(\widehat{\beta}_{\tau,l}(t) - \beta_{\tau,l}(t)) \xrightarrow{d} \mathbb{N}(\mathbf{0}, 1)$, then the asymptotic $100(1-a)\%$ point-wise confidence band (PCB) of $\beta_{\tau,l}$ is $\widehat{\beta}_{\tau,l}(t) \pm \sigma_{\tau,l}^{1/2}(t)z_a$ and the width of SCB is inflated by $\mathcal{Q}_{\tau,l}(a)/z_a$, where $z_a$ is the $a$-quantile of the standard normal distribution. Theorem 3 establishes the asymptotic normality of the parameters $\alpha_{\tau,l}$ and its corresponding $100(1-a)\%$ point-wise confidence interval is $\widehat{\alpha}_{\tau,l} \pm n^{-1/2}\sqrt{\tau(1-\tau)\widetilde{\boldsymbol{\xi}}_l^T\boldsymbol{\mathcal{I}}_\tau\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Sigma}_{\tau,1}\mathbf{\Sigma}_{\tau,2}^{-1}\mathcal{I}_\tau^T\widetilde{\boldsymbol{\xi}}_l}z_a$, where $\widetilde{\boldsymbol{\xi}}_l$ is a $d\times 1$ unit vector in which the $l$-th element is 1 and $l = 1,\ldots,d$.

### 3.3.3 Wild Bootstrap

However, $\mathbf{\Sigma}_{\tau,1}$ and $\mathbf{\Sigma}_{\tau,2}$ in the covariance function of $\widehat{\beta}_{\tau,l}(t)$ and the covariance matrix of $\widehat{\boldsymbol{\alpha}}_\tau$ are both unknown. Especially, in the covariance function and matrix, the unknown conditional density function of the error given the scalar covariates and infinite-dimensional functional covariates is very difficult to estimate. So, before we construct the corresponding SCBs and point-wise confidence intervals, we need to obtain the accurate estimators of the covariance function and covariance matrix. Adopting the ideas of the wild bootstrap procedure for classical quantile regression in [23] and refitted wild bootstrap for the high-dimensional quantile regression in [12], we propose a modified wild bootstrap method for the sparse semi-parametric functional quantile regression to estimate the covariance function of $\widehat{\beta}_{\tau,l}(t)$ and the covariance matrix of $\widehat{\boldsymbol{\alpha}}_\tau$ simultaneously. We randomly split

the original data set into two even parts and carry out this wild bootstrapping using the following steps:

***Step* 1.** We first minimize the objective function (3.9) based on the first part dataset to obtain the estimators $\widetilde{\boldsymbol{\alpha}}_{\tau,I}$ and $\widetilde{\boldsymbol{\theta}}_{\tau,I}$.

***Step* 2.** Based on the non-zero coefficient identified by the estimator $\widetilde{\boldsymbol{\theta}}_{\tau,I}$ and the second part dataset, we can obtain the estimators $(\widehat{\boldsymbol{\alpha}}_{\tau,II}^T, \widehat{\boldsymbol{\theta}}_{\tau,II}^T)^T$ by the following equation

$$(\widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II}) = \arg\min_{\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau} \frac{1}{n_1} \sum_{i=1}^{n_1} (\rho_\tau * \mathcal{K}_h)\left(Y_i - \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau\right) + \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau, \tag{3.16}$$

where $n_1$ is the sample size of the second dataset and $\widehat{\boldsymbol{\theta}}_{\tau,II}$ includes those zero coefficients identified in ***Step* 1**.

***Step* 3.** Independently generate weights $w_i$ such that

(1) there are two positive constants $c_1$ and $c_2$ satisfying $\sup w \in \mathbb{W} : w \leq 0 = -c_1$ and $\inf w \in \mathbb{W} : w \geq 0 = c_1$, where $\mathbb{W}$ is the support of $w$;

(2) $E\left[w_i^{-1} I(w_i > 0)\right] = -E\left[w_i^{-1} I(w_i < 0)\right] = 1/2$ and $E\left[|w_i|\right] < \infty$;

(3) the $\tau$th quantile of $w$ is zero.

***Step* 4.** Calculate the residuals based on the second dataset: $\hat{e}_i = Y_i - \boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}_{\tau,II} - \boldsymbol{U}_i^T \hat{\boldsymbol{\theta}}_{\tau,II}$. Then used the second part data set to obtain the bootstrapped samples denoted by $Y_i^{(b)} = \boldsymbol{Z}_i^T \hat{\boldsymbol{\alpha}}_\tau + \boldsymbol{U}_i^T \hat{\boldsymbol{\theta}}_\tau + e_i^{(b)}$, where $e_i^{(b)} = w_i |\hat{e}_i|$.

***Step* 5.** Resolve the objective function (3.16) based on the bootstrapped samples and denote the estimate by $\hat{\boldsymbol{\alpha}}_\tau^{(b)}$ and $\hat{\boldsymbol{\theta}}_\tau^{(b)}$.

***Step* 6.** Repeat ***Step* 2 - *Step* 5** $B$ times and then estimate the variance function $\sigma_{\tau,l}(t)$ of $\hat{\beta}_{\tau,l}(t)$ from the 100 estimations at each $t$. A confidence interval for $\boldsymbol{\alpha}_\tau$ can also be established from the bootstrap estimations of $\boldsymbol{\alpha}_\tau$.

**Theorem 4.** *Under the Conditions given in Section 3.3.1, using the above wild bootstrap procedure, we have*

$$\sup_{u \in \mathbb{R}} \left| P\left\{ \sup_{t \in \mathcal{S}_\varepsilon(\beta_{\tau,l})} \sqrt{\frac{n}{K}} (\widehat{\beta}_{\tau,l}^{(b)}(t) - \widehat{\beta}_{\tau,l}(t)) \leq u \right\} - P\left\{ \sup_{t \in \mathcal{S}_\varepsilon(\beta_{\tau,l})} \sqrt{\frac{n}{K}} (\widehat{\beta}_{\tau,l}(t) - \beta_{\tau,l}(t)) \leq u \right\} \right| \to 0,$$

$$\sup_{v \in \mathbb{R}} \left| P(n^{1/2}(\widehat{\alpha}_{\tau,l}^{(b)} - \widehat{\alpha}_{\tau,l}) \leq v) - P(n^{1/2}(\widehat{\alpha}_{\tau,l} - \alpha_{\tau,l}^0) \leq v) \right| \to 0.$$

Theorem 4 shows that the estimated variance function based on the above-mentioned wild bootstrap method is consistent, which makes it possible to conduct statistical inferences without estimating the unknown terms in the covariance function and covariance matrix, such as the error density function.

## 3.4. Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of our proposed CLoSE method in comparison with the smoothed quantile loss (SQL) method. The difference between the SQL method and the CLoSE method is that the SQL method only uses the combination of the smoothed quantile loss and the roughness penalty to estimate $\beta_\tau(t)$, while the CLoSE method contains the additional local sparse penalty in the objective function to estimate $\beta_\tau(t)$.

We consider synthetic data generated from the following semi-parametric functional quantile model:

$$Y_i = \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau + \int_{-1}^{1} X_i^T(t)\beta_\tau(t)dt + e_i(\tau), \qquad i = 1, \ldots, n,$$

where $X_i(t)$ is a Wiener process, $Z_{i1}, Z_{i2} \overset{iid}{\sim} N(0, 0.1)$, $\boldsymbol{Z}_i = (1, Z_{i1}, Z_{i2})^T$, $\boldsymbol{\alpha}_\tau = (0, 1, 1)^T$, $\beta_\tau(t) = \sin(2\pi t)\mathbb{1}\{-0.5 \le t \le 0.5\}$, and $e_i(\tau) = e_i - F_e^{-1}(\tau)$, where $F_e(\cdot)$ is the CDF of $e_i$. We consider the following two scenarios to generate $e_i$.

- **Scenario I:** $e_i$ are i.i.d from the normal distribution, namely, $e_i \overset{iid}{\sim} N(0, 0.02)$. Under this scenario, the distribution of the random error is symmetric.
- **Scenario II:** $e_i$ are i.i.d from a heavy-tailed distribution, Cauchy distribution, namely, $e_i \overset{iid}{\sim} C(0, 0.01)$.

In the convolution smoothing loss function, we use the Gaussian kernel $\mathcal{K}(u) = (2\pi)^{-1/2}e^{-u^2/2}$, and bandwidth $h = ((K + p) \times m + d)/n)^{2/5}$ [32]. Then the corresponding smoothed loss function is $(\rho_\tau * \mathcal{K}_h)(u) = (h/2)l_G(u/h) + (\tau - 1/2)u$ where $l_G(u) = (2/\pi)^{1/2}e^{-u^2/2} + u(1 - 2\Phi(u))$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The simulation is repeated for 100 times.

Regarding the estimation for $\beta_\tau(t)$, we consider two aspects of the performance of the proposed estimator $\hat{\beta}_\tau(t)$: (i) identification of null and nonnull regions and (ii) the difference between $\hat{\beta}_\tau(t)$ and $\beta_\tau(t)$. Let $T$ denote the set of time points where $X(t)$ is evaluated. Define the null region $\mathcal{N}(\beta_\tau) = \{t \in T : \beta_\tau(t) = 0\}$ and the nonnull region $\mathcal{S}(\beta_\tau) = \{t \in T : \beta_\tau(t) \neq 0\}$. Define the true discovery rate (TDR) for $\mathcal{N}(\beta_\tau)$ and the false discovery rate (FDR) for $\mathcal{S}(\beta_\tau)$ as follows

$$\text{TDR} = \frac{|\mathcal{N}(\hat{\beta}_\tau) \cap \mathcal{N}(\beta_\tau)|}{|\mathcal{N}(\beta_\tau)|}, \quad \text{FDR} = \frac{|\mathcal{N}(\hat{\beta}_\tau) \cap \mathcal{S}(\beta_\tau)|}{|\mathcal{S}(\beta_\tau)|}.$$

We use TDR and FDR to evaluate the performance of the proposed method on identifying the null and nonnull regions of $\beta_{\tau,l}(t)$ for $\tau = 0.1, \ldots, 0.8$ and $l = 1, \ldots, m$. Ideally, we want a larger TDR

and a smaller FDR. A larger TDR means more null regions of functional predictors are correctly detected. A smaller FDR means less nonnull region of functional predictors are mistakely identified as the null region.

The simulation results on TDR and FDR are summarized in Table 3.1. Table 3.1 shows that the TDR is overall satisfying for all chosen quantiles under both scenarios. The FDR of the Normal scenario is much better than that of the Cauchy scenario. This may be due to the fact that the Cauchy distribution is heavy-tailed and the variance of the error cannot be controlled. But under both scenarios, FDR decreases as the sample sizes increase for all chosen quantiles, which is desirable.

Table 3.1: True discovery rate (TDR) for the null region $\mathcal{N}(\beta_\tau)$ and false discovery rate (FDR) for the nonnull region $\mathcal{S}(\beta_\tau)$ using the convolution smoothing based locally sparse estimation (CLoSE) method when the errors are simulated from the Normal or Cauchy distribution. Here $n$ denotes the sample size and $\tau$ denotes the quantile.

| | | Normal | | Cauchy | |
|---|---|---|---|---|---|
| n | $\tau$ | TDR | FDR | TDR | FDR |
| 500 | 0.2 | 98.5% | 4.2% | 89.3% | 17.2% |
| | 0.3 | 99.2% | 4.3% | 95.1% | 13.9% |
| | 0.4 | 98.9% | 3.3% | 88.6% | 17.5% |
| | 0.5 | 98.7% | 2.8% | 98.4% | 10.7% |
| | 0.6 | 98.7% | 3.1% | 92.5% | 17.6% |
| | 0.7 | 99.0% | 3.5% | 96.1% | 17.2% |
| | 0.8 | 98.6% | 3.8% | 86.9% | 16.3% |
| 1000 | 0.2 | 99.3% | 2.5% | 95.8% | 12.2% |
| | 0.3 | 98.9% | 2.7% | 98.6% | 11.6% |
| | 0.4 | 95.9% | 1.9% | 99.9% | 11.5% |
| | 0.5 | 99.5% | 2.1% | 99.6% | 7.8% |
| | 0.6 | 96.1% | 1.5% | 99.7% | 10.6% |
| | 0.7 | 98.6% | 2.0% | 97.5% | 12.6% |
| | 0.8 | 98.5% | 2.4% | 93.4% | 13.6% |

Table 3.2 shows the $L_2$-norm of the difference between the estimator $\hat{\beta}_\tau(t)$ and the true function $\beta_\tau(t)$ from two different methods. It shows that the performance of both methods is improved as the sample size increases while the proposed method always outperforms the SQL method without the locally sparse regularization. The advantage of the proposed method is especially significant when the sample size is small. Figure 3.2 displays the estimator $\hat{\beta}_{0.8}(t)$ under the Normal scenario using the CLoSE method and the SQL method. It shows that the CLoSE method can obtain a strictly zero estimate for the functional coefficient $\beta_{0.8}(t)$ in the null regions, while $\hat{\beta}_{0.8}(t)$ estimated from the SQL method is always nonzero in the null regions.

Table 3.3 shows the performance of the estimator for the parametric components $\boldsymbol{\alpha}_\tau$ using the CLoSE method. The bias and standard error of the estimators for $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$ both decrease as the sample size increases under two scenarios for all chosen quantiles.

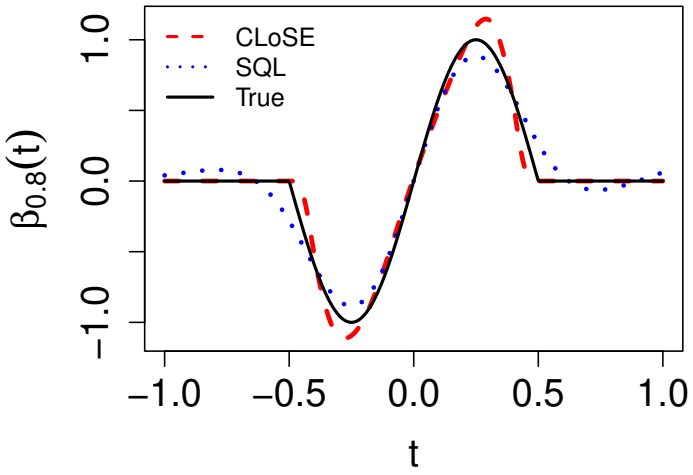



Figure 3.2: The estimator $\hat{\beta}_{0.8}(t)$ in one simulation replicate using the convolution smoothing based locally sparse estimation (CLoSE) method (red solid line) and the smoothed quantile loss (SQL) method (black dashed line) when the errors are simulated from the normal distribution when the sample size $n = 500$. The true $\beta_{0.8}(t)$ is displayed as the blue dotted line.

Table 3.2: The $L_2$-norm of the difference between the estimator $\hat{\beta}_\tau(t)$ and the true function $\beta_\tau(t)$ $\|\hat{\beta}_\tau(t,u) - \beta_\tau(t)\|_2$ using the convolution smoothing based locally sparse estimation (CLoSE) method and the smoothed quantile loss (SQL) method when the errors are simulated from the normal or Cauchy distribution. Here $n$ denotes the sample size and $\tau$ denotes the quantile.

| | | Normal | | Cauchy | |
|---|---|---|---|---|---|
| n | $\tau$ | SQL | CLoSE | SQL | CLoSE |
| 500 | 0.2 | 0.058 | 0.028 | 0.110 | 0.097 |
| | 0.3 | 0.037 | 0.027 | 0.090 | 0.086 |
| | 0.4 | 0.039 | 0.026 | 0.102 | 0.091 |
| | 0.5 | 0.036 | 0.023 | 0.076 | 0.046 |
| | 0.6 | 0.043 | 0.028 | 0.103 | 0.092 |
| | 0.7 | 0.039 | 0.024 | 0.093 | 0.088 |
| | 0.8 | 0.064 | 0.026 | 0.105 | 0.088 |
| 1000 | 0.2 | 0.043 | 0.026 | 0.075 | 0.070 |
| | 0.3 | 0.041 | 0.026 | 0.069 | 0.057 |
| | 0.4 | 0.036 | 0.028 | 0.065 | 0.062 |
| | 0.5 | 0.029 | 0.022 | 0.057 | 0.035 |
| | 0.6 | 0.037 | 0.027 | 0.065 | 0.058 |
| | 0.7 | 0.044 | 0.025 | 0.071 | 0.059 |
| | 0.8 | 0.044 | 0.026 | 0.082 | 0.075 |

Table 3.3: Biases and standard errors (SEs) of the estimator for $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$ using the convolution smoothing based locally sparse estimation (CLoSE) method when the errors are simulated from the normal or Cauchy distribution. Here $n$ denotes the sample size, and $\tau$ denotes the quantile.

| $n$ | $\tau$ | Normal | | | | Cauchy | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Bias($\alpha_{1,\tau}$) $\times 10^2$ | SE($\alpha_{1,\tau}$) $\times 10^2$ | Bias($\alpha_{2,\tau}$) $\times 10^2$ | SE($\alpha_{2,\tau}$) $\times 10^2$ | Bias($\alpha_{1,\tau}$) $\times 10^2$ | SE($\alpha_{1,\tau}$) $\times 10^2$ | Bias($\alpha_{2,\tau}$) $\times 10^2$ | SE($\alpha_{2,\tau}$) $\times 10^2$ |
| 500 | 0.2 | 0.05 | 0.90 | 0.11 | 0.99 | 0.09 | 4.40 | -0.92 | 4.36 |
| | 0.3 | 0.04 | 0.87 | 0.12 | 0.99 | 0.24 | 3.65 | -0.88 | 3.62 |
| | 0.4 | 0.04 | 0.90 | 0.12 | 0.97 | 0.31 | 3.37 | -0.80 | 3.19 |
| | 0.5 | 0.05 | 0.89 | 0.11 | 0.97 | 0.34 | 3.34 | -0.85 | 3.09 |
| | 0.6 | 0.04 | 0.87 | 0.11 | 0.99 | 0.50 | 3.52 | -0.92 | 3.15 |
| | 0.7 | 0.05 | 0.88 | 0.11 | 0.98 | 0.49 | 3.85 | -0.99 | 3.43 |
| | 0.8 | 0.04 | 0.88 | 0.12 | 0.98 | 0.75 | 4.66 | -1.20 | 4.07 |
| 1000 | 0.2 | -0.07 | 0.63 | -0.14 | 0.68 | -0.09 | 2.34 | 0.30 | 2.80 |
| | 0.3 | -0.06 | 0.62 | -0.15 | 0.66 | -0.03 | 1.97 | 0.16 | 2.37 |
| | 0.4 | -0.05 | 0.65 | -0.14 | 0.68 | 0.04 | 2.03 | 0.10 | 2.34 |
| | 0.5 | -0.06 | 0.61 | -0.15 | 0.66 | 0.02 | 1.82 | -0.02 | 2.19 |
| | 0.6 | -0.08 | 0.63 | -0.15 | 0.65 | 0.03 | 2.01 | -0.10 | 2.36 |
| | 0.7 | -0.06 | 0.62 | -0.15 | 0.63 | 0.01 | 2.14 | -0.29 | 2.50 |
| | 0.8 | -0.05 | 0.64 | -0.15 | 0.66 | 0.03 | 2.66 | -0.51 | 3.05 |

## 3.5. Real Data Analysis

Climate factors such as temperature and rainfall have significant effects on soybean germination and growth. In North America, these climate factors can account for 15% variation of the soybean yield [79]. For the stability of soybean production, it is important to keep tracking these factors over the growing season. If we can figure out when and how the temperatures combined with other environmental and non-environmental factors influence the soybean yield, then we may obtain a better soybean planting and harvesting strategy. We can also make some interventions when the temperature is too low or too high. In addition, to have a more thorough understanding of this relationship, analyzing the conditional quantiles of soybean yield should be more meaningful than analyzing the conditional mean only. For these reasons, in this analysis, we want to identify the impact of daily minimum and daily maximum temperatures on the soybean yield for different quantiles.

Kansas in the United States has 4.7 million acres of soybean planted, producing 200 million bushels, and rank the 10th in the United States in terms of soybean yield. The data on soybean yield and other related non-environmental variables in Kansas state between 1991 and 2006 are collected and organized from the United States Department of Agriculture (USDA) website (`https://quickstats.nass.usda.gov/`). The corresponding measurements on climate variables are collected from National Oceanic and Atmospheric Administration (NOAA) website (`https://www.ncei.noaa.gov/products`) for Kansas within the same time range.

To be more specific, the soybean yield-related data collected from the USDA website contains the county level annual soybean yield (measured in bushels per acre), the size of harvested land and the size of irrigated area among each harvested land. The climate data collected from the NOAA website contains daily minimum temperatures, daily maximum temperatures and annual precipitations at the climate station level. To link the observations from different websites together, we first identify the latitudes and longitudes of each climate station and the center of each county of Kansas. Next we label the location of each climate station by comparing its distance to all the county centers. More specifically, for each climate station, the closest county center is its label of location. To obtain county level daily minimum and maximum temperature and annual precipitation observations, we average the corresponding climate station level observations over all the climate stations within each county. In this way, we integrate all the observations at the county level.

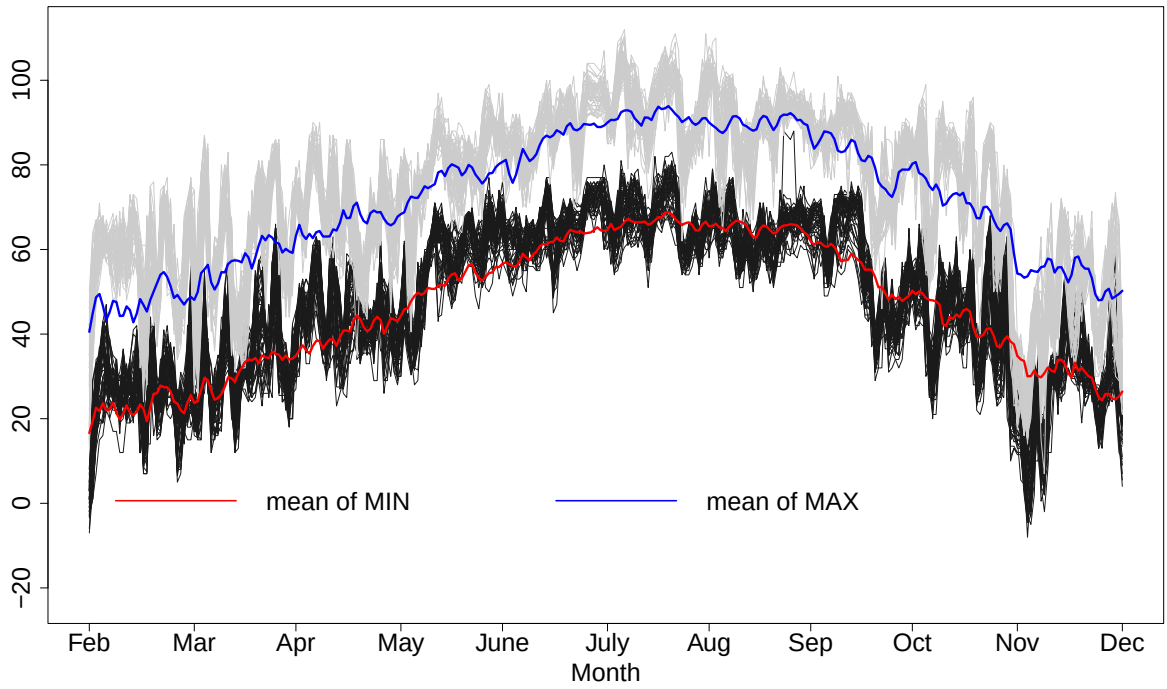



Figure 3.3: A sample of daily minimum and maximum temperature curves of counties in Kansas. The unit of the y-axis is the Fahrenheit temperature scale.

In the following analysis, we treat the annual precipitation, and the ratio of the size of the irrigated area over the size of harvested land of each county as two scalar predictors and we treat daily minimum temperature and daily maximum temperature curves as two functional predictors. Figure 3.3 displays a sample of the daily minimum temperature and daily maximum temperature curves of Kansas.

Let $X_1(t)$ and $X_2(t)$ denote the daily maximum temperature and daily minimum temperature between February and November respectively. Let $Z_1$ and $Z_2$ denote the annual precipitation and

the ratio of irrigated area of each county in Kansas. Let $Y$ denote the annual soybean yield of each county in Kansas. The model we want to investigate is the following,

$$Q_\tau(Y|X_1(t), X_2(t), Z_1, Z_2) = c(\tau) + \alpha_{1,\tau} Z_1 + \alpha_{2,\tau} Z_2 + \int_{\mathcal{T}} X_1(t)\boldsymbol{\beta}_{1,\tau}(t)dt + \int_{\mathcal{T}} X_2(t)\boldsymbol{\beta}_{2,\tau}(t)dt.$$

We fit the model at three different quantiles $\tau = 0.25, 0.5, 0.75$, which represent three different scenarios of the soybean yield: the "worst" case, the normal case and the "best" case. We apply the bootstrap procedure mentioned above to compute the simultaneous confident bands for $\beta_{1,\tau}$ and $\beta_{2,\tau}$, and the 90%-confidence intervals for $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$.

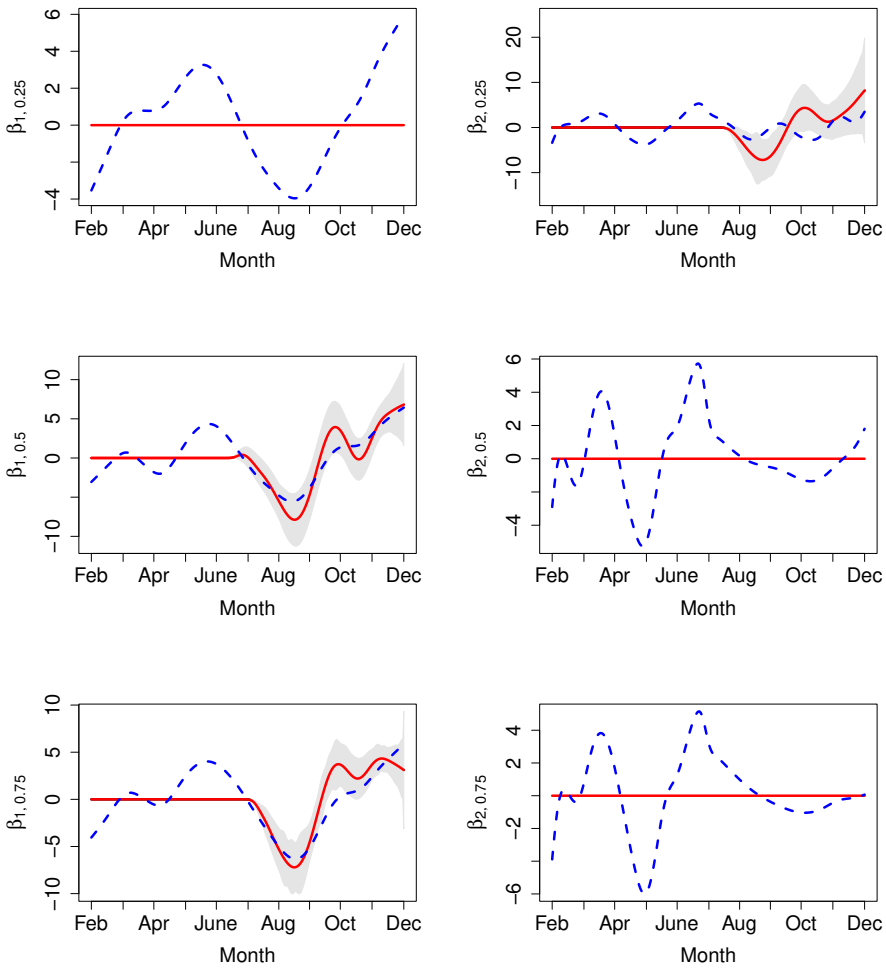



Figure 3.4: The estimated slope functions using the convolution smoothing based locally sparse estimation (CLoSE) method (red solid line) and the smoothed quantile loss (SQL) method (blue dashed line) at different quantile levels, $\tau = 0.25, 0.50$ and $0.75$, from the soybean dataset of Kansas. The gray areas are the corresponding 95% simultaneous confidence bands.

Figure 3.4 displays the simultaneous confident bands for $\beta_{1,\tau}$ and $\beta_{2,\tau}$ with $\tau = 0.25, 0.5$ and $0.75$. Specifically, Figure 3.4-(a) and Figure 3.4-(b) show that the daily maximum temperature has no influence on the $25\%$ quantile of the soybean yield. Only the daily minimum temperature after late July matters for this "worst" scenario of the soybean yield. From Figure 3.4-(d) and Figure 3.4-(f), we observe that the daily minimum temperature has no effect on the soybean yield for the 50% and 75% quantiles of the soybean yield. On the other hand, Figure 3.4-(c) and Figure 3.4-(e) show that the daily maximum temperatures play an important role in these two quantiles of the soybean yield. More specifically, from Figure 3.4-(c) and Figure 3.4-(e), we can observe that regarding the 50% and 75% quantiles of the soybean yield, the maximum temperature during the hot summer has a negative effect and the maximum temperature during the fall and winter have a positive influence. This may be due to the fact that before reaching $30°C$ $(86°F)$, the increasing air temperature can lead to an increase in soybean yield. But after reaching $30°C$ $(86°F)$, a higher temperature is negatively related to the soybean yield [72]. The most active soybean planting dates of different counties of Kansas range from mid-May to late June. This can explain why we observe the sparsity of $\hat{\beta}_{1,\tau}(t)$ and $\hat{\beta}_{2,\tau}(t)$ between February and late June. In addition, comparing the estimated functions from the proposed method and the SQL method, we notice that the SQL method tends to use both functional predictors: daily minimum and daily maximum temperature curves. This may be due to the high correlation of these two functional predictors. The proposed CLoSE method tends to use only one of them, which implies that the two functional predictors may contain repetitive information and therefore using only one of them in the model should be enough.

Table 3.4 summarizes 95%-confidence intervals for $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$. We observe that for all three quantiles, the proportion of irrigated areas of each county is significant to the annual soybean yield, while the annual precipitation is not significant for any quantiles.

Table 3.4: The estimate and 95% Confidence Intervals (CI) for $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$ at different quantile levels, $\tau = 0.25, 0.50$ and $0.75$, using the convolution smoothing based locally sparse estimation (CLoSE) method from the soybean dataset of Kansas.

| $\tau$ | $\alpha_{1,\tau}$ | | $\alpha_{2,\tau}$ | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| 0.25 | 19.26 | [-94.23, 84.60] | 15.19 | [13.03, 23.69] |
| 0.50 | 5.36 | [-56.55, 14.34] | 16.69 | [14.33, 19.42] |
| 0.75 | 7.23 | [-32.79, 28.42] | 15.17 | [13.22, 18.34] |

## 3.6. Conclusion and Discussion

In this project, we propose a locally sparse semi-parametric functional quantile model to study the dynamic dependence of functional covariates on scalar response. A convolution smoothing based locally sparse estimation (CLoSE) method is developed to identify the locally sparse regions of the

functional coefficients and estimate the parameters and the functional coefficients on the non-null regions. We also establish the functional version of the oracle property of the functional coefficients and the asymptotic properties of the estimated parameters. The standard quantile loss function is convex, which means that it is relatively easy to optimize using standard methods. However, we need to add a concave penalty to keep the locally sparse structure of the functional coefficients. Then the loss function becomes more complicated, and optimization becomes more difficult. The CLoSE method addresses this problem by first smoothing the quantile loss function. This step smooths out the function and makes it easier to optimize. The CLoSE method has been shown to be effective in estimating the locally sparse semi-parametric functional quantile model in our simulation studies and real application.

When the data size is large-scale, the computation burden brought by the fSCAD penalty and the selection of several tuning parameters is heavy even if we use the convolution-smoothed quantile loss. Improving the computation efficiency when analyzing large-scale functional data is indeed an important topic worth investigating in future research. One possible approach is to develop new algorithms that can handle large-scale data more efficiently. For example, one could consider using parallel computing techniques or developing distributed algorithms that can run on clusters of computers.

## 3.7. Some Lemmas

To prove our theorems in the manuscripts, we start by proving the following lemmas.

**Lemma 1.** *Under Condition 2, $\beta^0_{\tau,l}(t) - \beta_{\tau,l}(t) = b_a(t) + o(K^{-r})$, where $\beta^0_{\tau,l}(t) = \mathcal{B}^T(t)\boldsymbol{\theta}^0_{\tau,l}$ is the best B-spline function in approximating $\beta_{\tau,l}$, and $b_a(t) = O(k^{-r})$ is the spline approximation bias.*

**Proof** The proof of this lemma can be found in [2].

**Lemma 2.** *Under Condition 1 and 5,*

*(i) there exists constants $C_G > c_G > 0$ such that*

$$c_G K^{-1} \leq \rho_{min}\left(\boldsymbol{U}\boldsymbol{U}^T\right) \leq \rho_{max}\left(\boldsymbol{U}\boldsymbol{U}^T\right) \leq C_G K^{-1},$$

*where $\rho_{min}$ and $\rho_{max}$ denote the smallest and largest eigenvalues of a matrix, respectively.*

*(ii) we can have $\|\boldsymbol{U}\|_\infty = O(K^{-1})$.*

**Lemma 3.** *Under Condition 5, we can get*

*(i) $\|\boldsymbol{V}\|_\infty = O(K^{2q-1})$;*

*(ii) for any non-zero vector $\boldsymbol{u}$, there are two positive constants $c_D < C_D$ such that $c_D K^{2q-1}\|\boldsymbol{\mu}\|_2^2 \leq \boldsymbol{\mu}^T\boldsymbol{V}\boldsymbol{\mu} \leq C_D K^{2q-1}\|\boldsymbol{\mu}\|_2^2$.*

**Lemma 4.** *If Conditions 1, 2, 3, 4 and 8 hold, for any $\delta > 0$, there exists a constant $C$ such that*

$$Pr(\|\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)\|_2 < C\sqrt{1/n}) > 1 - \delta, \tag{3.17}$$

*holds for all sufficiently large $K$ and $n$.*

**Proof** Recall that

$$\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) = \frac{1}{n}\sum_{i=1}^{n}\left\{G_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - Y_i) - \tau\right\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T.$$

Note that each B-spline basis function $B_k(t)$ is non-negative, nonzero over no more than $p+1$ consecutive subintervals and $\sum_{k=1}^{K+p} B_k(t) = 1$ for all $t \in [0, \mathcal{T}]$, therefore, we can get $\int_0^{\mathcal{T}} B_k^2(t) = O(K^{-1})$. Then, with Condition 1, we have

$$\|\boldsymbol{U}_i\|_\infty = O(K^{-1/2}), \qquad and \qquad \|\boldsymbol{U}_i\|_2 = O(1).$$

For $G_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - Y_i)$, we have

$$\begin{aligned}
&G_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - Y_i)\\
=&G_h\left(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - Y_i + \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&G_h\left(-e_i + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right),
\end{aligned}$$

by Lagrange mean value theorem, we can get

$$\begin{aligned}
&G_h\left(-e_i + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&G_h(-e_i) + \dot{G}_h\left(-e_i + \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&G_h(-e_i) + \mathcal{K}_h\left(-e_i + \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right),
\end{aligned}$$

where $\boldsymbol{\beta}_\tau^*(t)$ lies between $\boldsymbol{\beta}_\tau(t)$ and $\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0$.

Then $\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)$ can be expressed as:

$$\begin{aligned}
\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) =&\frac{1}{n}\sum_{i=1}^{n}\left\{\mathcal{K}_h\left(-e_i + \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\right\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\\
&+\frac{1}{n}\sum_{i=1}^{n}\{G_h(-e_i) - \tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\\
&\triangleq \boldsymbol{I}_1 + \boldsymbol{I}_2.
\end{aligned}$$

First, we will prove that $\|\boldsymbol{I}_1\|_2 = O_p(K^{-r}(h^2 + (nh)^{-1/2}))$. By Lemma A.1 and Condition 1, we have $\left|\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - \int_0^{\mathcal{T}} \boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right| = O_p(K^{-r})$. Then, by the weak law of large numbers and Condition 1, we can get $\|\boldsymbol{I}_1\|_2 = O_p(K^{-r}(h^2 + (nh)^{-1/2}))$.

Moreover, we want to prove the bound of the term $\boldsymbol{I}_2$. It follows from Condition 3 that

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n} G_h(-e_i) \mid \boldsymbol{Z}_i, \boldsymbol{X}_i(t)\right) &= E(G_h(-e_i) \mid \boldsymbol{Z}_i, \boldsymbol{X}_i(t)) \\
&= \int G(-\varepsilon/h) f_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) d\varepsilon \\
&= \frac{1}{h}\int \mathcal{K}(-\varepsilon/h) F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) du \\
&= \int \mathcal{K}(-u) F_{e|\boldsymbol{Z},\boldsymbol{X}}(hu) d\varepsilon \\
&= \int \mathcal{K}(u)\left\{F_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + huf_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + O(h^2)\right\} du \\
&= \tau + O(h^2),
\end{aligned}$$

then, using Condition 1, it is easy to have

$$\begin{aligned}
E\left\{\frac{1}{n}\sum_{i=1}^{n}\{G_h(-e_i) - \tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right\} &= E\left\{E\left\{\frac{1}{n}\sum_{i=1}^{n}\{G_h(-e_i) - \tau\} \mid \boldsymbol{Z}_i, \boldsymbol{X}_i(t)\right\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right\} \\
&= O(h^2).
\end{aligned} \tag{3.18}$$

For any matrix $\boldsymbol{A}$, let $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^T$. Given that

$$\begin{aligned}
Cov\left\{\frac{1}{n}\sum_{i=1}^{n}\{G_h(-e_i) - \tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right\} &= \frac{1}{n}Cov\left\{\{G_h(-e_i) - \tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right\} \\
&= \frac{1}{n}E\left\{\{G_h(-e_i) - \tau\}^2\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\} \\
&= \frac{1}{n}E\left\{G_h^2(-e_i)\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\} \\
&\quad - 2\tau \times \frac{1}{n}E\left\{G_h(-e_i)\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\} \\
&\quad + \tau^2\frac{1}{n}E\left\{\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\}.
\end{aligned}$$

Denote $\widetilde{\mathcal{K}}(u) = 2G(u)\mathcal{K}(u)$, $c_K = \int u\widetilde{\mathcal{K}}(u)du$ and note that $\int \widetilde{\mathcal{K}}(u)du = 1$. Similarly, using integration by parts and a change of variable, we have

$$\begin{aligned} E(G_h^2(-\varepsilon_i) \mid \boldsymbol{Z}_i, \boldsymbol{X}_i(t)) &= \int G_h^2(-\varepsilon) f_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)d\varepsilon \\ &= \frac{1}{h}\int \widetilde{\mathcal{K}}(-\varepsilon/h) F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)d\varepsilon \\ &= \tau - h f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\int u\widetilde{\mathcal{K}}(u)du + O(h^2) \\ &= \tau + O(h). \end{aligned}$$

According to the bounded condition of $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i(t)$ (Condition 1), we can get

$$\begin{aligned} E\left\{G_h^2(-e_i)\left\|\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\|_\infty\right\} &= E\left\{\left\|\left((\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right)^{\otimes 2}\right\|_\infty E(G_h^2(-e_i) \mid \boldsymbol{Z}_i, \boldsymbol{X}_i(t))\right\} \\ &= \tau + O(h). \end{aligned} \tag{3.19}$$

Combining (3.18) and (3.19), we cant get $\left\|Cov\left\{n^{-1}\sum_{i=1}^n \{G_h(-e_i)-\tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T\right\}\right\|_2 = 1/n(\tau(1-\tau)+O(h)) = O(1/n)$. Then $\|\boldsymbol{I}_2\|_2 = O_p(h^2+\sqrt{1/n})$ and $\|\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)\|_2 = O_p(K^{-r}(h^2+(nh)^{-1/2}) + h^2 + \sqrt{1/n}) = O_p(\sqrt{1/n})$.

**Lemma 5.** *If Conditions 1, 2, 3, 4 and 8 hold hold, we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{E\left((\boldsymbol{Z}^T, \boldsymbol{U}^T)^T\right)^{\otimes 2}\right\}^{-1/2} \{G_h(-e_i)-\tau\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T \xrightarrow{d} \mathbb{N}(\boldsymbol{0}, \tau(1-\tau)\boldsymbol{I}_{d+m\times(K+p)}). \tag{3.20}$$

**Proof** Denote $\boldsymbol{\mathcal{M}}_i = (\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T$, $\boldsymbol{\mathcal{M}} = E\left\{\left((\boldsymbol{Z}^T, \boldsymbol{U}^T)^T\right)^{\otimes 2}\right\}$, and

$$\boldsymbol{G}_{\tau,ni}^* = \frac{1}{\sqrt{n}}\boldsymbol{\mathcal{M}}^{-1/2}\{G_h(-e_i)-\tau\}\boldsymbol{\mathcal{M}}_i,$$

then

$$\sum_{i=1}^n \boldsymbol{G}_{\tau,ni}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\mathcal{M}}^{-1/2}\{G_h(-e_i)-\tau\}\boldsymbol{\mathcal{M}}_i.$$

From the proof of Lemma A.4, we can get $E(G_h(-e_i)-\tau) = O(h^2)$ and $Var(G_h(-e_i)-\tau) = \tau(1-\tau)+O(h)$, then

$$E(\boldsymbol{G}_{\tau,ni}^*) = O(h^2 n^{-1/2}), \qquad Cov(\boldsymbol{G}_{\tau,ni}^*) = \frac{1}{n}\tau(1-\tau)\boldsymbol{\mathcal{M}}^{-1/2}E\left(\boldsymbol{\mathcal{M}}_i\boldsymbol{\mathcal{M}}_i^T\right)\boldsymbol{\mathcal{M}}^{-1/2} + O(h/n).$$

Correspondingly,

$$E\left(\sum_{i=1}^n \boldsymbol{G}_{\tau,ni}^*\right) = O(h^2 n^{-1/2}), \qquad Cov\left(\sum_{i=1}^n \boldsymbol{G}_{\tau,ni}^*\right) = \tau(1-\tau)\boldsymbol{I} + O(h).$$

By Cramér–Wold Theorem, to establish the asymptotic normality of the vector $\sum_{i=1}^{n} \boldsymbol{G}^{*}_{\tau,ni}$, we first need to prove the asymptotic normality of $\sum_{i=1}^{n} \boldsymbol{\mu}^T \boldsymbol{G}^{*}_{\tau,ni}$ for any unit vector $\boldsymbol{\mu}$, where $\mu_j, j = 1, \ldots, d + m \times (K + p)$ is the $j$-th column of the identity matrix $I_{d+m\times(K+p)}$.

Denote $\Phi_i = n^{-1/2}\boldsymbol{\mu}^T \boldsymbol{\mathcal{M}}^{-1/2}\boldsymbol{\mathcal{M}}_i$, then $\boldsymbol{\mu}^T \boldsymbol{G}^{*}_{\tau,ni} = \Phi_i G_h(-e_i) - \tau)$. Next, we will check the following condition, which ensures Lindeberg's condition is both sufficient and necessary:

$$\max_{1\le i\le n} \frac{Var(\Phi_i G_h(-e_i) - \tau))}{\sum_{i=1}^{n} Var(\Phi_i G_h(-e_i) - \tau))} = \max_{1\le i\le n} \frac{E(\Phi_i^2)}{\sum_{i=1}^{n} E(\Phi_i^2)} \to 0, \tag{3.21}$$

as $n \to \infty$. By the definition of $\Phi_i$, we have

$$\begin{aligned}
E(\Phi_i^2) &= \frac{1}{n}\boldsymbol{\mu}^T \boldsymbol{\mathcal{M}}^{-1/2} E(\boldsymbol{\mathcal{M}}_i \boldsymbol{\mathcal{M}}_i^T)\boldsymbol{\mathcal{M}}^{-1/2}\boldsymbol{\mu} \\
&= \frac{1}{n} tr\left\{\boldsymbol{\mu}^T \boldsymbol{\mathcal{M}}^{-1/2} E(\boldsymbol{\mathcal{M}}_i \boldsymbol{\mathcal{M}}_i^T)\boldsymbol{\mathcal{M}}^{-1/2}\boldsymbol{\mu}\right\} \\
&= \frac{1}{n} tr\left\{\boldsymbol{\mu}\boldsymbol{\mu}^T \boldsymbol{\mathcal{M}}^{-1/2} E(\boldsymbol{\mathcal{M}}_i \boldsymbol{\mathcal{M}}_i^T)\boldsymbol{\mathcal{M}}^{-1/2}\right\} \\
&= \frac{1}{n} tr\left\{\boldsymbol{\mathcal{M}}^{-1} E(\boldsymbol{\mathcal{M}}_i \boldsymbol{\mathcal{M}}_i^T)\right\} \\
&\le \frac{1}{n} tr\left\{\boldsymbol{\mathcal{M}}^{-1}\right\} tr\left\{E(\boldsymbol{\mathcal{M}}_i \boldsymbol{\mathcal{M}}_i^T)\right\} \\
&\le C_\Phi / n,
\end{aligned}$$

then we can obtain

$$\max_{1\le i\le n} \frac{E(\Phi_i^2)}{\sum_{i=1}^{n} E(\Phi_i^2)} = O\left(\frac{1}{n}\right) \to 0.$$

Thus condition (3.21) holds, which means that $\sum_{i=1}^{n} \boldsymbol{\mu}^T \boldsymbol{G}^{*}_{\tau,ni}$ has a asymptotic normal distribution. By Cramér–Wold Theorem and the above results, it is straightforward to prove that $\sum_{i=1}^{n} \boldsymbol{G}^{*}_{\tau,ni} = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^{-1/2} \{G_h(-e_i) - \tau\} \boldsymbol{\mathcal{M}}_i$ has an asymptotic multivariate normal distribution

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^{-1/2} \{G_h(-e_i) - \tau\} \boldsymbol{\mathcal{M}}_i \xrightarrow{d} \mathbb{N}(\boldsymbol{0}, \tau(1-\tau)\boldsymbol{I}_{d+m\times(K+p)}),$$

thus (3.20) holds and the proof is completed.

**Lemma 6.** *Under Conditions 1, 2, 3, 4 and 8, we have*

$$\ddot{\mathcal{L}}^{*}_0(\boldsymbol{\alpha}^0_\tau, \boldsymbol{\theta}^0_\tau) = E\left\{ f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \end{pmatrix}^{\otimes 2} \right\} + O_p\left(\frac{1}{\sqrt{nh}}\right), \tag{3.22}$$

**Proof** Recall that

$$\ddot{\mathcal{L}}^{*}_0(\boldsymbol{\alpha}^0_\tau, \boldsymbol{\theta}^0_\tau) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}^0_\tau + \boldsymbol{U}_i^T \boldsymbol{\theta}^0_\tau - Y_i) \begin{pmatrix} \boldsymbol{Z}_i \\ \boldsymbol{U}_i \end{pmatrix}^{\otimes 2}.$$

For $\mathcal{K}_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-Y_i)$, we have

$$\begin{aligned}
&\mathcal{K}_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-Y_i)\\
=&\mathcal{K}_h\left(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-Y_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&\mathcal{K}_h\left(-e_i+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right),
\end{aligned}$$

by Lagrange mean value theorem, we can get

$$\begin{aligned}
&\mathcal{K}_h\left(-e_i+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&\mathcal{K}_h(-e_i)+\dot{\mathcal{K}}_h\left(-e_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\\
=&\mathcal{K}_h(-e_i)+\frac{1}{h^2}\dot{\mathcal{K}}_h\left(-e_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right),
\end{aligned}$$

where $\dot{\mathcal{K}}_h(\cdot)=\dot{\mathcal{K}}(\cdot/h)$, $\beta_\tau^*(t)$ lies between $\beta_\tau(t)$ and $\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0$.

Then $\ddot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)$ can be expressed as:

$$\begin{aligned}
\ddot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)=&\frac{1}{nh^2}\sum_{i=1}^n\left\{\dot{\mathcal{K}}_h\left(-e_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\right\}\begin{pmatrix}\boldsymbol{Z}_i\\\boldsymbol{U}_i\end{pmatrix}^{\otimes 2}\\
&+\frac{1}{n}\sum_{i=1}^n\mathcal{K}_h(-e_i)\begin{pmatrix}\boldsymbol{Z}_i\\\boldsymbol{U}_i\end{pmatrix}^{\otimes 2}\\
\triangleq&\boldsymbol{I}_1+\boldsymbol{I}_2.
\end{aligned}$$

**(1) First, we consider the first term $\boldsymbol{I}_1$.** $\left|\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right|=O(K^{-r})$, and

$$\frac{1}{nh^2}\sum_{i=1}^n\dot{\mathcal{K}}_h\left(-e_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)=O_p\left(\frac{1}{h}+\frac{1}{\sqrt{nh^3}}\right),$$

then

$$\boldsymbol{I}_1=o_p\left(\frac{1}{\sqrt{nh}}\right),\tag{3.23}$$

with the Condition $hK^r\to\infty$.

**(2) We consider the second term $\boldsymbol{I}_2$.** For $n^{-1}\sum_{i=1}^{n}\mathcal{K}_h(-e_i)$, using change of variable, we have

$$\begin{aligned}
E\left\{\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_h(-e_i)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\} &= E\left\{\mathcal{K}_h(-e_i)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\}\\
&= \int \mathcal{K}_h(-\varepsilon)f_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)d\varepsilon\\
&= \int \mathcal{K}_h(-u)f_{e|\boldsymbol{Z},\boldsymbol{X}}(hu)du\\
&= \int \mathcal{K}_h(u)\left[f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+O(h)\right]du\\
&= f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+O(h^2),
\end{aligned}$$

$$\begin{aligned}
E\left\{\mathcal{K}^2(-e_i/h)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\} &= \int \mathcal{K}^2(-\varepsilon/h)f_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)d\varepsilon\\
&= h\int \mathcal{K}^2(-u)f_{e|\boldsymbol{Z},\boldsymbol{X}}(hu)du\\
&= O(h),
\end{aligned}$$

moreover,

$$\begin{aligned}
Var\left\{\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_h(-e_i)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\} &= \frac{1}{nh^2}Var\left\{\mathcal{K}(-e_i/h)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\}\\
&= \frac{1}{nh^2}E\left\{\mathcal{K}^2(-e_i/h)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\}\\
&\quad -\frac{1}{nh^2}\left\{E\left\{\mathcal{K}(-e_i/h)\mid \boldsymbol{Z}_i,\boldsymbol{X}_i(t)\right\}\right\}^2\\
&= O(n^{-1}h^{-1}).
\end{aligned}$$

Then, we have

$$\boldsymbol{I}_2=\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_h(-e_i)\begin{pmatrix}\boldsymbol{Z}_i\\ \boldsymbol{U}_i\end{pmatrix}^{\otimes 2}=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}+O_p\left(h^2+\frac{1}{\sqrt{nh}}\right). \tag{3.24}$$

Combining (3.23) and (3.24), we can get that

$$\ddot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}+O_p\left(\frac{1}{\sqrt{nh}}\right),$$

with Condition $nh^4\to 0$. Then the proof is completed.

## 3.8. Proofs in Section 3.3

### 3.8.1 Proof of Theorem 1

We want to prove that for any $\delta > 0$, there exists a sufficiently large constant $\Delta$ such that for sufficiently large $n$,

$$Pr\left\{\inf_{\substack{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^0\|_2=\Delta\sqrt{n}\\ \|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|_2\le\Delta K/\sqrt{n}}} \mathcal{L}^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)-\mathcal{L}^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)>0\right\}\ge 1-\delta. \tag{3.25}$$

(3.25) implies with probability at least $1-\delta$ that there is a local minimizer $(\widehat{\boldsymbol{\alpha}}_\tau,\widehat{\boldsymbol{\theta}}_\tau)$ such that $\|\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\|_2=O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\theta}}_\tau-\boldsymbol{\theta}_\tau^0\|_2=O_p(n^{-1/2}K)$.

By Taylor expansion, we have

$$\begin{aligned}
&\mathcal{L}^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)-\mathcal{L}^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)\\
=&\mathcal{L}_1^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)-\mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)+\sum_{l=1}^m\frac{K}{\mathcal{T}}\left[\int_0^{\mathcal{T}}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt-\int_0^{\mathcal{T}}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right]\\
\ge&\mathcal{L}_1^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)-\mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)+\sum_{l=1}^m\frac{K}{\mathcal{T}}\left[\int_{\mathcal{S}(\beta_{\tau,l}^0)}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt-\int_{\mathcal{S}(\beta_{\tau,l}^0)}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right]\\
\approx&\left\{\dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)\right\}^T\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}+\frac{1}{2}\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}^T\ddot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}\\
&\quad+\sum_{l=1}^m\frac{K}{\mathcal{T}}\left[\int_{\mathcal{S}(\beta_{\tau,l}^0)}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt-\int_{\mathcal{S}(\beta_{\tau,l}^0)}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right]\\
=&I+II+III.
\end{aligned} \tag{3.26}$$

**(1) We consider the first term $I$ in (3.26).** Recall

$$\dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)=\dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^0)+2\begin{pmatrix}\mathbf{0}_{d\times d} & \\ & \boldsymbol{\Gamma}\otimes\boldsymbol{V}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau^0\end{pmatrix},$$

For this term, with Lemma A.4 and the Condition 7, then

$$
\begin{aligned}
|I| &= \left| \left\{ \dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) \right\}^T \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right| \\
&\leq \left| \left\{ \dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) \right\}^T \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right| + 2 \left| \left\{ \begin{pmatrix} \mathbf{0}_{d\times d} & \\ & \boldsymbol{\Gamma} \otimes \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right\}^T \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right| \\
&= \left| \left\{ \dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) \right\}^T \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right| + 2 \left| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T (\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0) \right| \\
&= \left| \left\{ \dot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) \right\}^T \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \right| + 2 \left\| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \left\{ \boldsymbol{Z}_i^T (\boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0) + \boldsymbol{U}_i^T (\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0) \right\} \right| \\
&\quad + 2 \left\| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \left\{ \boldsymbol{Z}_i^T (\boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0) \right\} \right| \\
&\quad + \left| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \left\{ \boldsymbol{U}_i^T (\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0) \right\} \right| \\
&\quad + 2 \left\| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \boldsymbol{Z}_i^T \right\|_2 \left\| \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \right\|_2 \\
&\quad + \left\| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \boldsymbol{U}_i^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 \\
&\quad + 2 \left\| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2
\end{aligned}
$$

With the bounded Condition 1 and Lemma A.4, we have

$$
\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \boldsymbol{Z}_i^T \right\|_2 &= C \left| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \right| \\
&= O(n^{-1/2}).
\end{aligned}
$$

Thus, we have

$$
\left\| \frac{1}{n} \sum_{i=1}^n \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \boldsymbol{Z}_i^T \right\|_2 \left\| \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \right\|_2 = O_p(n^{-1}) = o_p(K/n), \quad (3.27)
$$

as the Condition $\left\| \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \right\|_2 = O_p(n^{-1/2})$.

Note that $\|\boldsymbol{U}_i\|_2 = O(1)$ and Lemma A.4, it has

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) - \tau \right\} \boldsymbol{U}_i^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 = O_p(K/n), \tag{3.28}$$

with the Condition $\left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2 = O_p(K/n^{1/2})$.

As $\sup_l \gamma_l = o(n^{-1/2} K^{1/2-2d})$ with Condition 7 , $\|\boldsymbol{\theta}_\tau^0\|_2 = O(K^{1/2})$ and Lemma A.3, we can obtain

$$\left\| \boldsymbol{\theta}_\tau^{0T} (\boldsymbol{\Gamma} \otimes \boldsymbol{V})^T \right\|_2 \leq \sup_l (\gamma_l) \lambda_{\max}(\boldsymbol{V}) \|\boldsymbol{\theta}_\tau^0\|_2 = o(K/n). \tag{3.29}$$

Combining (3.27), (3.28) and (3.29), we have

$$I = O_p(K/n). \tag{3.30}$$

**(2) For the second term $II$.**

$$\begin{aligned}
&\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix}^T \ddot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0) \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \\
=&\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix}^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) \begin{pmatrix} \boldsymbol{Z}_i \\ \boldsymbol{U}_i \end{pmatrix}^{\otimes 2} \right\} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \\
&+ (\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0)^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V})(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0)
\end{aligned}$$

From Lemma A.6, we have

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) \begin{pmatrix} \boldsymbol{Z}_i \\ \boldsymbol{U}_i \end{pmatrix}^{\otimes 2} = E \left\{ f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \end{pmatrix}^{\otimes 2} \right\} + O_p(\frac{1}{\sqrt{nh}}).$$

Based on Conditions 3 and 4, we can also have

$$\begin{aligned}
&\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix}^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau^0 - Y_i) \begin{pmatrix} \boldsymbol{Z}_i \\ \boldsymbol{U}_i \end{pmatrix}^{\otimes 2} \right\} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} \\
=&\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix}^T E \left\{ f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \end{pmatrix}^{\otimes 2} \right\} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} + O_p(\frac{1}{\sqrt{nh}} \frac{K^2}{n}) \\
=&\frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix}^T E \left\{ f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \end{pmatrix}^{\otimes 2} \right\} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \end{pmatrix} + o_p\left(\frac{K}{n}\right).
\end{aligned} \tag{3.31}$$

And, we have

$$(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0)^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V})(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0) \leq \sup_l (\gamma_l) \rho_{\max}(\boldsymbol{V}) \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0 \right\|_2^2 = o_p\left(\frac{K}{n}\right). \tag{3.32}$$

Then, (3.31) and (3.32) lead to

$$II = \frac{1}{2}\begin{pmatrix}\boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0\end{pmatrix}^T E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z} \\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}\begin{pmatrix}\boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}_\tau^0 \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0\end{pmatrix} + o_p\left(\frac{K}{n}\right). \tag{3.33}$$

**(3) Next, we consider the third term $III$.** By the Taylor Expansion, we can get

$$\begin{aligned}
&\sum_{l=1}^{m}\frac{K}{\mathcal{T}}\left[\int_{\mathcal{S}(\beta_{\tau,l}^0)} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt - \int_{\mathcal{S}(\beta_{\tau,l}^0)} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right] \\
\approx &\begin{pmatrix}\frac{K}{\mathcal{T}}\nabla\left[\int_{\mathcal{S}(\beta_{\tau,1}^0)} p_{\lambda_1}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,1}^0\right|)dt\right] \\ \vdots \\ \frac{K}{\mathcal{T}}\nabla\left[\int_{\mathcal{S}(\beta_{\tau,m}^0)} p_{\lambda_m}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,m}^0\right|)dt\right]\end{pmatrix}^T\left(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0\right) + \frac{1}{2}\left(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0\right)^T \\
&\begin{bmatrix}\frac{K}{\mathcal{T}}\nabla^2\left[\int_{\mathcal{S}(\beta_{\tau,1}^0)} p_{\lambda_1}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,1}^0\right|)dt\right] & & \\ & \ddots & \\ & & \frac{K}{\mathcal{T}}\nabla^2\left[\int_{\mathcal{S}(\beta_{\tau,m}^0)} p_{\lambda_m}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,m}^0\right|)dt\right]\end{bmatrix}\left(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\tau^0\right) \\
=&III_1 + III_2
\end{aligned} \tag{3.34}$$

For the bound of $\nabla\left[\int_{\mathcal{S}(\beta_{\tau,l}^0)} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right]$ and all $l = 1, 2, ..., m$ and $j = 1, 2, ..., K+p$, using Hölder inequality, we have

$$\begin{aligned}
\left|\frac{\partial}{\partial\theta_{\tau,l,j}}\int_{\mathcal{S}(\beta_{\tau,l}^0)} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right| &= \left|\int_{\mathcal{S}(\beta_{\tau,l}^0)}\frac{\partial}{\partial\theta_{\tau,l,j}^0}p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)dt\right| \\
&= \left|\int_{\mathcal{S}(\beta_{\tau,l}^0)}\dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0\right|)B_j(t)dt \quad \text{sgn}(\theta_{\tau,l,j}^0)\right| \\
&\leq \sqrt{\int_{\mathcal{S}(\beta_{\tau,l}^0)}\dot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0|)^2dt}\sqrt{\int_{\mathcal{S}(\beta_{\tau,l}^0)}B_j^2(t)dt} \\
&\leq \|B_j\|_2\sqrt{\int_{\mathcal{S}(\beta_{\tau,l}^0)}\dot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}^0|)^2dt} \\
&= O(K^{-1/2}n^{-1/2}K^{-1}) = O(n^{-1/2}K^{-3/2}),
\end{aligned}$$

with the condition $\max_l \sqrt{\int_{\mathcal{S}(\beta^0_{\tau,l})} \dot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}|)^2 dt} = o(n^{-1/2}K^{-1})$ (Condition 9-(i)). Then, we can also have

$$
\begin{aligned}
|III_1| \leq & \left\| \begin{pmatrix} \frac{K}{\mathcal{T}} \nabla \left[ \int_{\mathcal{S}(\beta^0_{\tau,1})} p_{\lambda_1}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,1}\right|) dt \right] \\ \vdots \\ \frac{K}{\mathcal{T}} \nabla \left[ \int_{\mathcal{S}(\beta^0_{\tau,m})} p_{\lambda_m}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,m}\right|) dt \right] \end{pmatrix}^T \right\|_2 \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}^0_\tau \right\|_2 \\
= & O_p(K/n).
\end{aligned} \tag{3.35}
$$

For the bound of $\nabla^2 \left[ \int_{\mathcal{S}(\beta^0_{\tau,l})} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|) dt \right]$ and all $l = 1, 2, ..., m$ and $j, k = 1, 2, ..., K + p$, using Hölder inequality, we have

$$
\begin{aligned}
\left| \frac{\partial}{\partial \theta_{\tau,l,j} \partial \theta_{\tau,l,k}} \int_{\mathcal{S}(\beta^0_{\tau,l})} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|) dt \right| &= \int_{\mathcal{S}(\beta^0_{\tau,l})} \nabla^2 p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|) dt \\
&= \int_{\mathcal{S}(\beta^0_{\tau,l})} \ddot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|) \left[ B_j(t) \circ \text{sgn}(\theta^0_{\tau,l,j}) \right] \left[ B_k(t) \circ \text{sgn}(\theta^0_{\tau,l,k}) \right] dt \\
&\leq \sqrt{\int_{\mathcal{S}(\beta^0_{\tau,l})} \ddot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}|)^2 dt} \sqrt{\int_0^{\mathcal{T}} B_j^2(t) B_k^2(t) dt} \\
&\leq \sqrt{\int_{\mathcal{S}(\beta^0_{\tau,l})} \ddot{p}_{\lambda_l}(|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}|)^2 dt} \sqrt{\int_0^{\mathcal{T}} B_j(t) B_k(t) dt} \\
&= o(K^{-1}),
\end{aligned}
$$

with Condition 9-(i), where $\circ$ denotes the entry-wise product of two vectors and the last equation is based on the property $\sup_{j,k} | < B_j, B_k > | = O(K^{-1})$. Therefore,

$$
\begin{aligned}
|III_2| \leq & \frac{1}{2} \left\| \boldsymbol{\theta}_\tau - \boldsymbol{\theta}^0_\tau \right\|_2^2 \max_{l,j,k} \left| \frac{\partial}{\partial \theta_{\tau,l,j} \partial \theta_{\tau,l,k}} \int_{\mathcal{S}(\beta^0_{\tau,l})} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|) dt \right| \\
= & o_p(K/n).
\end{aligned} \tag{3.36}
$$

With the above results, we have

$$
III = O_p(K/n). \tag{3.37}
$$

Combining (3.26), (3.30), (3.33) and (3.37), we finally have

$$
\begin{aligned}
& \mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) - \mathcal{L}^*(\boldsymbol{\alpha}^0_\tau, \boldsymbol{\theta}^0_\tau) \\
\geq & I + II + III \\
= & O_p(K/n) + \frac{1}{2} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}^0_\tau \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}^0_\tau \end{pmatrix}^T E \left\{ f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \end{pmatrix}^{\otimes 2} \right\} \begin{pmatrix} \boldsymbol{\alpha}_\tau - \boldsymbol{\alpha}^0_\tau \\ \boldsymbol{\theta}_\tau - \boldsymbol{\theta}^0_\tau \end{pmatrix} \\
> & 0,
\end{aligned} \tag{3.38}
$$

where the positive term

$$\begin{aligned}
&\frac{1}{2}\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}^T E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}\\
\geq& C_1\rho_{\min}(\boldsymbol{U}\boldsymbol{U}^T)\left\|\begin{pmatrix}\boldsymbol{\alpha}_\tau-\boldsymbol{\alpha}_\tau^0\\ \boldsymbol{\theta}_\tau-\boldsymbol{\theta}_\tau^0\end{pmatrix}\right\|_2^2\\
=& O_p(K/n),
\end{aligned}$$

with Condition 1 and Lemma A.2.

Then, (3.26) holds, which means there is a local minimizer $(\widehat{\boldsymbol{\alpha}}_\tau, \widehat{\boldsymbol{\theta}}_\tau)$ of $\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)$ such that $\|\widehat{\boldsymbol{\alpha}}_\tau - \boldsymbol{\alpha}_\tau^0\|_2 = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_\tau^0\|_2 = O_p(n^{-1/2}K)$. Moreover, with $\widehat{\beta}_{\tau,l}(t) = \boldsymbol{\mathcal{B}}^T(t)\widehat{\boldsymbol{\theta}}_{\tau,l}$, we have

$$\begin{aligned}
\left\|\widehat{\beta}_{\tau,l}-\beta_{\tau,l}\right\|_2 \leq& \left\|\widehat{\beta}_{\tau,l}-\beta_{\tau,l}^0\right\|_2 + \left\|\beta_{\tau,l}^0-\beta_{\tau,l}\right\|_2\\
=& \left\|(\widehat{\boldsymbol{\theta}}_{\tau,l}-\boldsymbol{\theta}_{\tau,l}^0)^T\boldsymbol{\mathcal{B}}(t)\right\|_2 + \left\|\beta_{\tau,l}^0-\beta_{\tau,l}\right\|_2\\
=& O_p(n^{-1/2}K^{1/2}) + O(K^{-r})\\
=& O_p(n^{-1/2}K^{1/2}),
\end{aligned} \tag{3.39}$$

with the condition $K/n^{1/(2r+1)} \to \infty$ (Condition 6).

Then the proof of Theorem 1 is completed.

### 3.8.2 Proof of Theorem 2

**Proof of (i) of Theorem 2**

To prove the first part of Theorem 1, we need to show that $\widehat{\beta}_{\tau,l}(t) = 0$ for all $t \in \mathcal{N}(\beta_{\tau,l})$ with probability tending to one.

As stated in Section 3, $\boldsymbol{\mathcal{B}}_{\tau,l,1}(t)$ denotes the $K_{\tau,l}^*$ dimensional sub-vector of $\boldsymbol{\mathcal{B}}(t)$ such that each $B_j(t)$ in $\boldsymbol{\mathcal{B}}_{\tau,l,1}(t)$ has a support inside $\mathcal{S}(\beta_{\tau,l})$. We also further denote the $\boldsymbol{\mathcal{B}}_{\tau,l,2}(t)$ as the sub-vector of $\boldsymbol{\mathcal{B}}(t)$ such that the support of each $B_j(t)$ in $\boldsymbol{\mathcal{B}}_{\tau,l,2}(t)$ belongs to $\mathcal{N}(\beta_{\tau,l})$. Let $\mathcal{A}_{\tau,l,j}$ consist of indices $k$ such that $B_j(t)$ belongs to $\boldsymbol{\mathcal{B}}_{\tau,l,j}(t)$, and $F_{\tau,l,j}$ denote the union of supports of those basis functions in $\boldsymbol{\mathcal{B}}_{\tau,l,j}(t), j = 1, 2$. By the local support property of B-spline basis functions, $F_{\tau,l,1}$ converges to $\mathcal{S}(\beta_{\tau,l})$ and $F_{\tau,l,2}$ converges to $\mathcal{N}(\beta_{\tau,l})$ as $K \to \infty$.

Fix a $\theta_{\tau,l,k}$ such that $k \in \mathcal{A}_{\tau,l,2}$, recall that

$$\frac{\partial\mathcal{L}^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)}{\partial\theta_{\tau,l,k}} = \frac{\partial\mathcal{L}_1^*(\boldsymbol{\alpha}_\tau,\boldsymbol{\theta}_\tau)}{\partial\theta_{\tau,l,k}} + \frac{K}{\mathcal{T}}\text{sgn}(\theta_{\tau,l,k})\int_0^{\mathcal{T}}\dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)B_k(t)dt,$$

for the first term of RHS, we take Taylor expansion, then

$$\begin{aligned}\frac{\partial \mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)}{\partial \theta_{\tau,l,k}} =& \frac{\partial \mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)}{\partial \theta_{\tau,l,k}} + \sum_{l'=1}^{m}\sum_{j=1}^{K+p} \frac{\partial^2 \mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)}{\partial \theta_{\tau,l,k}\partial \theta_{\tau,l',j}}(\theta_{\tau,l',j} - \theta^0_{\tau,l',j}) \\ &+ \frac{K}{\mathcal{T}}\text{sgn}(\theta_{\tau,l,k}) \int_0^{\mathcal{T}} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)B_k(t)dt,\end{aligned} \tag{3.40}$$

where

$$\frac{\partial \mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)}{\partial \theta_{\tau,l,k}} = \frac{1}{n}\sum_{i=1}^{n}\left\{G_h(\boldsymbol{Z}_i^T\boldsymbol{\alpha}_\tau^0 + \boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0 - Y_i) - \tau\right\}U_{i,lk} + 2\gamma_l \boldsymbol{V}_{k\cdot}\boldsymbol{\theta}^0_{\tau,l},$$

$U_{i,lk} = \int_0^{\mathcal{T}} X_{il}(t)B_k(t)dt$ and $\boldsymbol{V}_{k\cdot}$ denotes the $k$th row of $\boldsymbol{V}$.

Below we will consider each term in (3.40).

(1) By Lemma A. 4, $|U_{i,lk}| = O(K^{-1/2})$, $\|\boldsymbol{V}_{k\cdot}\|_\infty = O(K^{2q-1})$ and Conditions 1-8, we have

$$\frac{\partial \mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)}{\partial \theta_{\tau,l,k}} = O_p(n^{-1/2}K^{-1/2}). \tag{3.41}$$

(2) Based on the Conditions 1-8, Lemma A.6 and assumption $\left|\theta_{\tau,l',j} - \theta^0_{\tau,l',j}\right| = O_p(\sqrt{K/n})$, we can get

$$\sum_{l'=1}^{m}\sum_{j=1}^{K+p} \frac{\partial^2 \mathcal{L}_1^*(\boldsymbol{\alpha}_\tau^0, \boldsymbol{\theta}_\tau^0)}{\partial \theta_{\tau,l,k}\partial \theta_{\tau,l',j}}(\theta_{\tau,l',j} - \theta^0_{\tau,l',j}) = O_p(K^{1/2}n^{-1/2}). \tag{3.42}$$

(3) Let $\mathcal{S}_k$ denote the support of $B_k(t)$, then, in the last term,

$$\begin{aligned}&\left|\int_0^{\mathcal{T}} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt - \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)dt\right| \\ =&\left|\int_{\mathcal{S}(\beta_{\tau,l})} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt + \int_{\mathcal{N}(\beta_{\tau,l})} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt - \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)dt\right| \\ \leq&\left|\int_{\mathcal{S}(\beta_{\tau,l})} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt\right| + \left|\int_{\mathcal{N}(\beta_{\tau,l})} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt - \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)dt\right| \\ =&o(n^{-1/2}K^{-1}) + O(K^{-1}) \\ =&o(1),\end{aligned}$$

with the Condition $\int_{\mathcal{S}(\beta_{\tau,l})} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt = o(n^{-1/2}K^{-1})$. Then we have

$$\int_0^{\mathcal{T}} \dot{p}_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}_{\tau,l}\right|)dt = \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)dt + o(1). \tag{3.43}$$

Combining (3.40), (3.41), (3.42) and (3.43), we finally have

$$\begin{aligned}\frac{\partial \mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)}{\partial \theta_{\tau,l,k}} &= \frac{K}{\mathcal{T}} \lambda_l \text{sgn}(\theta_{\tau,l,k}) \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)\lambda_l^{-1} B_k(t)dt + O_p(K^{1/2}n^{-1/2}) \\ &= \lambda_l \left\{ \frac{K}{\mathcal{T}} \lambda_l \text{sgn}(\theta_{\tau,l,k}) \int_{\mathcal{S}_k} \dot{p}_{\lambda_l}(|B_k(t)\theta_{\tau,l,k}|)\lambda_l^{-1} B_k(t)dt + O_p(K^{1/2}n^{-1/2}\lambda_l^{-1}) \right\}. \end{aligned} \tag{3.44}$$

If Condition 9-(ii) holds, $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \dot{p}_{\lambda_l}(\theta)\lambda_l^{-1} > 0$, and $K^{1/2}n^{-1/2}\lambda_l^{-1} \to 0$. Besides, $B_k(t)$ is non-negative, the sign of the derivative is completely determined by that of $\theta_{\tau,l,k}$. Now, since $(\widehat{\boldsymbol{\alpha}}_\tau, \widehat{\boldsymbol{\theta}}_\tau)$ minimizes $\mathcal{L}^*(\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau)$, we must have $\widehat{\theta}_{\tau,l,k} = 0$ for $k \in \mathcal{A}_{\tau,l,2}$. with probability tending to one. Similar with [50], it is easy to prove that the union $\widehat{F}_{\tau,l,2}$ of supports of basis functions associated to $\widehat{\boldsymbol{\theta}}_{\tau,l,2}$ equals to $F_{\tau,l,2}$ in probability. Therefore $\widehat{F}_{\tau,l,2}$ converges to $\mathcal{N}(\beta_{\tau,l})$. This completes the proof of the first part of Theorem 2.

**Proof of (ii) of Theorem 2**

We divide $\boldsymbol{\theta}^0_{\tau,l}$ into two parts: $\boldsymbol{\theta}^{0(1)}_{\tau,l}$ that consists of those $\theta_{\tau,l,k}$ such that $k \in \mathcal{A}_{\tau,l,1}$, and $\boldsymbol{\theta}^{0(2)}_{\tau,l}$ that consists of those $\theta_{\tau,l,k}$ such that $k \in \mathcal{A}_{\tau,l,2}$. Then we define $\boldsymbol{\theta}^{0(1)}_\tau = \left( \left(\boldsymbol{\theta}^{0(1)}_{\tau,1}\right)^T, \cdots, \left(\boldsymbol{\theta}^{0(1)}_{\tau,m}\right)^T \right)^T$ and $\boldsymbol{\theta}^{0(2)}_\tau = \left( \left(\boldsymbol{\theta}^{0(2)}_{\tau,1}\right)^T, \cdots, \left(\boldsymbol{\theta}^{0(2)}_{\tau,m}\right)^T \right)^T$. Correspondingly, we also divide $\widehat{\boldsymbol{\theta}}_{\tau,l}$ into two parts, namely, $\widehat{\boldsymbol{\theta}}^{(1)}_{\tau,l}$ and $\widehat{\boldsymbol{\theta}}^{(2)}_{\tau,l}$, then define $\widehat{\boldsymbol{\theta}}^{(1)}_\tau = \left( \left(\widehat{\boldsymbol{\theta}}^{(1)}_{\tau,1}\right)^T, \cdots, \left(\widehat{\boldsymbol{\theta}}^{(1)}_{\tau,m}\right)^T \right)^T$ and $\widehat{\boldsymbol{\theta}}^{(2)}_\tau = \left( \left(\widehat{\boldsymbol{\theta}}^{(2)}_{\tau,1}\right)^T, \cdots, \left(\widehat{\boldsymbol{\theta}}^{(2)}_{\tau,m}\right)^T \right)^T$. From the proof of (i), we can get that each element of $\widehat{\boldsymbol{\alpha}}_\tau$ and $\widehat{\boldsymbol{\theta}}^{(1)}_\tau$ stay away from zero when $n$ is sufficiently large. At the same time, $\widehat{\boldsymbol{\theta}}^{(2)}_\tau = \mathbf{0}$ with probability tending to one. Thus, $\left( \widehat{\boldsymbol{\alpha}}^T_\tau, \left(\widehat{\boldsymbol{\theta}}^{(1)}_\tau\right)^T \right)^T$ satisfies

$$\dot{\mathcal{L}}^*_1(\widehat{\boldsymbol{\alpha}}_\tau, \widehat{\boldsymbol{\theta}}^{(1)}_\tau) + \begin{pmatrix} \mathbf{0}_{d\times 1} \\ \left[ \frac{K}{\mathcal{T}} \frac{\partial}{\partial \theta_{\tau,l,k}} \int_0^{\mathcal{T}} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\widehat{\boldsymbol{\theta}}_{\tau,l}\right|)dt \right]^{l=1,\ldots,m}_{k\in\mathcal{A}_{\tau,l,1}} \end{pmatrix} = 0.$$

Then applying the Taylor expansion to $\frac{K}{\mathcal{T}} \frac{\partial}{\partial \theta_{\tau,l,k}} \int_0^{\mathcal{T}} p_{\lambda_l}(\left|\boldsymbol{\mathcal{B}}^T(t)\widehat{\boldsymbol{\theta}}_{\tau,l}\right|)dt$, we can get that

$$\begin{aligned}\frac{K}{\mathcal{T}} \frac{\partial}{\partial \theta_{\tau,l,k}} \int_0^{\mathcal{T}} p_{\lambda_l}\left(\left|\boldsymbol{\mathcal{B}}^T(t)\widehat{\boldsymbol{\theta}}_{\tau,l}\right|\right) dt =& \frac{K}{\mathcal{T}} \text{sgn}(\theta^0_{\tau,l,k}) \int_0^{\mathcal{T}} \dot{p}_{\lambda_l}\left(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|\right) B_k(t)dt \\ &+ \frac{K}{\mathcal{T}} \int_0^{\mathcal{T}} \ddot{p}_{\lambda_l}\left(\left|\boldsymbol{\mathcal{B}}^T(t)\boldsymbol{\theta}^0_{\tau,l}\right|\right) B_k^2(t)dt(\widehat{\theta}_{\tau,l,k} - \theta^0_{\tau,l,k}) \\ &= o_p(\widehat{\theta}_{\tau,l,k} - \theta^0_{\tau,l,k}),\end{aligned}$$

by the Condition 9-(i). Then, we can have

$$\dot{\mathcal{L}}^*_1(\widehat{\boldsymbol{\alpha}}_\tau, \widehat{\boldsymbol{\theta}}^{(1)}_\tau) + \begin{pmatrix} \mathbf{0}_{d\times 1} \\ o_p(\widehat{\boldsymbol{\theta}}^{(1)}_\tau - \boldsymbol{\theta}^{0(1)}_\tau) \end{pmatrix} = 0.$$

By Taylor expansion and the above results, we have

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\\ \widehat{\boldsymbol{\theta}}_\tau^{(1)}-\boldsymbol{\theta}_\tau^{0(1)}\end{pmatrix}=-\left\{\ddot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})\right\}^{-1}\sqrt{n}\dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})\left\{1+o_p(1)\right\}. \tag{3.45}$$

According to the proof of Lemma A.4, we have

$$\begin{aligned}\sqrt{n}\dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)}) =&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\mathcal{K}_h\left(-e_i+\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau^*(t)dt\right)\left(\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-\int_0^{\mathcal{T}}\boldsymbol{X}_i^T(t)\boldsymbol{\beta}_\tau(t)dt\right)\right\}(\boldsymbol{Z}_i^T,\boldsymbol{U}_{\tau,i}^{*T})^T\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{G_h(-e_i)-\tau\right\}(\boldsymbol{Z}_i^T,\boldsymbol{U}_{\tau,i}^{*T})^T+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{G_h(-e_i)-\tau\right\}(\boldsymbol{Z}_i^T,\boldsymbol{U}_{\tau,i}^{*T})^T+o_p(1).\end{aligned}$$

Similar with the proof of Lemma A.5, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{E\left((\boldsymbol{Z}^T,\boldsymbol{U}_\tau^{*T})^T\right)^{\otimes 2}\right\}^{-1/2}\left\{G_h(-e_i)-\tau\right\}(\boldsymbol{Z}_i^T,\boldsymbol{U}_{\tau,i}^{*T})^T\xrightarrow{d}\mathbb{N}(\boldsymbol{0},\tau(1-\tau)\boldsymbol{I}_{d+K_\tau^*}),$$

and

$$\boldsymbol{\Sigma}_{\tau,1}^{-1/2}\sqrt{n}\dot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})\xrightarrow{d}\mathbb{N}(\boldsymbol{0},\tau(1-\tau)\boldsymbol{I}_{d+K_\tau^*}), \tag{3.46}$$

where $\boldsymbol{\Sigma}_{\tau,1}=E\left\{\left((\boldsymbol{Z}^T,\boldsymbol{U}_\tau^{*T})^T\right)^{\otimes 2}\right\}$.

Similar with the proof of Lemma A.6, we have

$$\begin{aligned}\ddot{\mathcal{L}}_0^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})&=\frac{1}{n}\sum_{i=1}^n\mathcal{K}_h(\boldsymbol{Z}_i\boldsymbol{\alpha}_\tau^0+\boldsymbol{U}_i^T\boldsymbol{\theta}_\tau^0-Y_i)\begin{pmatrix}\boldsymbol{Z}_i\\ \boldsymbol{U}_{\tau,i}^*\end{pmatrix}^{\otimes 2}\\ &=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}_\tau^*\end{pmatrix}^{\otimes 2}\right\}+O_p(\frac{1}{\sqrt{nh}}),\end{aligned}$$

and

$$\begin{aligned}\ddot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})&=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}_\tau^*\end{pmatrix}^{\otimes 2}\right\}+o_p\left(\frac{1}{\sqrt{nK}}\right)+O_p(\frac{1}{\sqrt{nh}})\\ &=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}_\tau^*\end{pmatrix}^{\otimes 2}\right\},\end{aligned}$$

then

$$\ddot{\mathcal{L}}_1^*(\boldsymbol{\alpha}_\tau^0,\boldsymbol{\theta}_\tau^{0(1)})\xrightarrow{P}\boldsymbol{\Sigma}_{\tau,2},\qquad\boldsymbol{\Sigma}_{\tau,2}=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}_\tau^*\end{pmatrix}^{\otimes 2}\right\}. \tag{3.47}$$

Thus, by (3.46), (3.47) and Lutsky's theorem, we can get that

$$\sqrt{n}\left(\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Sigma}_{\tau,1}\boldsymbol{\Sigma}_{\tau,2}^{-1}\right)^{-1/2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\\ \widehat{\boldsymbol{\theta}}_\tau^{(1)}-\boldsymbol{\theta}_\tau^{0(1)}\end{pmatrix}\xrightarrow{d}\mathbb{N}(\mathbf{0},\tau(1-\tau)\boldsymbol{I}) \tag{3.48}$$

Finally we prove the asymptotic normality of $\sqrt{n/K}(\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t))$. Recall that $\sqrt{n/K}(\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t))$ admits the following decomposition,

$$\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t)=\frac{1}{\sqrt{n}}\boldsymbol{\Lambda}_{\tau,l}(t)\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\\ \widehat{\boldsymbol{\theta}}_\tau^{(1)}-\boldsymbol{\theta}_\tau^{0(1)}\end{pmatrix}+\beta_{\tau,l}^0(t)-\beta_{\tau,l}(t).$$

From Lemma A.1, for the B-spline approximation error, we have $\beta_{\tau,l}^0(t)-\beta_{\tau,l}(t)=O(K^{-r})$. Under Condition 6, $nK^{-(2r+1)}\to 0$, the bias of $\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t)$ is asymptotically negligible. Then by (3.48) and Slutsky's theorem, we can conclude that for any given point $t$ such $\beta_{\tau,l}\neq 0$, we have

$$\sigma_{\tau,l}^{-1/2}(t)\left(\widehat{\beta}_{\tau,l}(t)-\beta(t)\right)\to N(0,1), \tag{3.49}$$

in distribution, where $\sigma_{\tau,l}(t)=n^{-1}\tau(1-\tau)\boldsymbol{\Lambda}_{\tau,l}(t)\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Sigma}_{\tau,1}\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Lambda}_{\tau,l}^T(t)=O(K/n)$.

(3.49) shows the point-wise asymptotic normality and we can construct the corresponding confidence interval. However, in many studies, it is desirable to obtain simultaneous confidence bands (SCB) for the varying coefficient functions.

Define

$$\Theta_{\tau,l}(t)=\sigma_{\tau,l}^{-1/2}(t)n^{-1}\sqrt{\tau(1-\tau)}\boldsymbol{\Lambda}_{\tau,l}(t)\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Sigma}_{\tau,1}^{1/2}\sum_{ß=1}^{n}\boldsymbol{\zeta}_i,\qquad \boldsymbol{\zeta}_{\tau,l,i}\overset{iid}{\sim}\mathbb{N}(\mathbf{0},\boldsymbol{I}_{d+K_\tau^*}).$$

It is easy to show that $\Theta_{\tau,l}(t)$ is a Gaussian process with $E(\Theta_{\tau,l}(t))=0$ and $Var(\Theta_{\tau,l}(t))=1$. And, the covariance function of the Gaussian process is

$$\mathcal{C}(t,s)=n^{-1}\tau(1-\tau)\sigma_{\tau,l}^{-1/2}(t)\sigma_{\tau,l}^{-1/2}(s)\boldsymbol{\Lambda}_{\tau,l}(t)\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Sigma}_{\tau,1}\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Lambda}_{\tau,l}^T(s).$$

Similar to the proofs for Theorem 3 in [56], and Theorem 5 in [3], by the strong approximation theorem [14], we can prove that

$$\sup_{t\in\mathcal{S}(\beta_{\tau,l})}\left|\sigma_{\tau,l}^{-1/2}(t)\left\{\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t)\right\}-\Theta_{\tau,l}(t)\right|=o_p(\sqrt{K/n}). \tag{3.50}$$

Then for $t\in\mathcal{S}(\beta_{\tau,l})$, it can be proved that

$$\sigma_{\tau,l}^{-1/2}(t)(\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t))\xrightarrow{d}\mathbb{G}(t), \tag{3.51}$$

where $\mathbb{G}(t)$ is a Gaussian random process with mean 0 defined on $\mathcal{S}(\beta_{\tau,l})$ with the covariance function $\mathcal{C}(t,s)$.

Denote the left endpoint and right endpoint of $\mathcal{S}(\beta_{\tau,l})$ as $\mathcal{S}^L(\beta_{\tau,l})$ and $\mathcal{S}^R(\beta_{\tau,l})$, respectively. Then we partition $\mathcal{S}(\beta_{\tau,l})$ into $\widetilde{K}_{\tau,l}+1$ equally spaced intervals with $\mathcal{S}^L(\beta_{\tau,l})=\nu_0<\nu_1<\cdots<\nu_{\widetilde{K}_{\tau,l}}<\nu_{\widetilde{K}_{\tau,l}+1}=\mathcal{S}^R(\beta_{\tau,l})$, where $\widetilde{K}_{\tau,l}\to\infty$. We construct the SCB for $\beta_{\tau,l}(t)$ over a subset of $\mathcal{S}(\beta_{\tau,l})$, namely, $\mathcal{S}_\varepsilon(\beta_{\tau,l})=\left\{\nu_0,\ldots,\nu_{\mathcal{N}_{\tau,l}+1}\right\}$, and $\mathcal{S}_\varepsilon(\beta_{\tau,l})$ becomes denser as $n\to\infty$.

And, for any $0\le j<j^{'}\le\mathcal{N}_{\tau,l}+1$, the covariance function of the Gaussian process $\Theta_{\tau,l}(t)$ is

$$\begin{aligned}\left|Cov\left(\Theta_{\tau,l}(\nu_j),\Theta_{\tau,l}(\nu_{j'})\right)\right| &= \left|n^{-1}\sigma_{\tau,l}^{-1/2}(\nu_j)\sigma_{\tau,l}^{-1/2}(\nu_{j'})\mathbf{\Lambda}_{\tau,l}(\nu_j)\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Sigma}_{\tau,1}\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Lambda}_{\tau,l}^T(\nu_{j'})\right|\\ &\propto K^{-1}\left|\mathbf{\Lambda}_{\tau,l}(\nu_j)\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Sigma}_{\tau,1}\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Lambda}_{\tau,l}^T(\nu_{j'})\right|\\ &\propto \left|\mathbf{\Lambda}_{\tau,l}(\nu_j)\mathbf{\Lambda}_{\tau,l}^T(\nu_{j'})\right|\\ &= \left|\boldsymbol{\mathcal{B}}_{\tau,l}^T(\nu_j)\boldsymbol{\mathcal{B}}_{\tau,l}(\nu_{j'})\right|\\ &= \left|\sum_{k=1}^{K+p}B_k(\nu_j)B_k(\nu_{j'})\right|.\end{aligned}$$

Since the B-spline basis function has a local support property, that is, each B-spline basis function is nonzero over no more than $p+1$ consecutive sub-intervals [15], we can have $\left|Cov\left(\Theta_{\tau,l}(\nu_j),\Theta_{\tau,l}(\nu_{j'})\right)\right|\le C_B^{-|j-j^{'}|}$, with some positive constant $C_B$. Now combined with Lemma 1 in [57], we can conclude that for any $a\in(0,1)$,

$$\lim_{n\to\infty}P\left\{\sup_{t\in\mathcal{S}_\varepsilon(\beta_{\tau,l})}\left|\sigma_{\tau,l}^{-1/2}(t)\left\{\widehat{\beta}_{\tau,l}(t)-\beta_{\tau,l}(t)\right\}\right|\le\mathcal{Q}_{\tau,l}(a)\right\}=1-a,$$

where

$$\begin{aligned}\mathcal{Q}_{\tau,l}(a)=&\left(2\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|\right)^{1/2}\\ &-\left(2\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|\right)^{-1/2}\left\{\log(-0.5\log(1-a))+0.5\left[\log(\log|\mathcal{S}_\varepsilon(\beta_{\tau,l})|)+\log(4\pi)\right]\right\}.\end{aligned}$$

Then an asymptotic $100(1-a)\%$ simultaneous confidence band (SCB) for $\beta_{\tau,l}(t)$ over $\mathcal{S}_\varepsilon(\beta_{\tau,l})$ is given by $\widehat{\beta}_{\tau,l}(t)\pm\sigma_{\tau,l}^{1/2}(t)\mathcal{Q}_{\tau,l}(a)$. The proof of (ii) of Theorem 2 is completed.

### 3.8.3 Proof of Theorem 3

We want to prove the asymptotic normality of $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0)$, and $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0)$ can be expressed as

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0)=\boldsymbol{\mathcal{I}}_\tau\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\\ \widehat{\boldsymbol{\theta}}_\tau^{(1)}-\boldsymbol{\theta}_\tau^{0(1)}\end{pmatrix},$$

then by (3.48), we can have

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0\right)\xrightarrow{d}\mathbb{N}(\mathbf{0},\tau(1-\tau)\boldsymbol{\mathcal{I}}_\tau\mathbf{\Sigma}_{\tau,2}^{-1}\mathbf{\Sigma}_{\tau,1}\mathbf{\Sigma}_{\tau,2}^{-1}\boldsymbol{\mathcal{I}}_\tau^T),\tag{3.52}$$

where $\left\|Cov(\widehat{\boldsymbol{\alpha}}_\tau-\boldsymbol{\alpha}_\tau^0)\right\|_\infty=O(1/n)$. Then we complete the proof of Theorem 3.

### 3.8.4 Proof of Theorem 4

We choose to randomly split the original dataset into two parts $Data_I$ and $Data_{II}$, where $Data_I = \{(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i(t), t \in [0, \mathcal{T}), n_1 + 1 \leq i \leq n\}$, $Data_{II} = \{(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i(t), t \in [0, \mathcal{T}), 1 \leq i \leq n_1\}$, $n_1 = [n/2]$ and $[v]$ denotes the integer not greater than the positive number $v$. Denote the estimators based on $Data_I$ as $\widetilde{\boldsymbol{\alpha}}_{\tau,I}$ and $\widetilde{\boldsymbol{\theta}}_{\tau,I}$, which is the solution of the following objective function

$$\begin{aligned}(\widetilde{\boldsymbol{\alpha}}_{\tau,I}, \widetilde{\boldsymbol{\theta}}_{\tau,I}) = &\arg\min_{\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (\rho_\tau * \mathcal{K}_h)\left(Y_i - \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau\right) \\ &+ \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau + \sum_{l=1}^{m} \sum_{j=1}^{K} p_{\lambda_l}\left(\sqrt{\frac{K}{\mathcal{T}} \boldsymbol{\theta}_{\tau,l}^T \boldsymbol{W}_j \boldsymbol{\theta}_{\tau,l}}\right),\end{aligned} \tag{3.53}$$

where $n_2 = n - n_1$.

Define the set $\widehat{S}_{\boldsymbol{\theta},I} = \{j : \widetilde{\theta}_{\tau,I,j} \neq 0, 1 \leq j \leq m(K+p)\}$ and $\widehat{\boldsymbol{T}}_{\boldsymbol{\theta},I} = \{\boldsymbol{v} \in \mathbb{R}^{m(K+p)} : v_j = 0, \forall j \in \widehat{S}^c_{\boldsymbol{\theta},I}\}$, where $\widehat{S}^c_{\boldsymbol{\theta},I}$ is the complement of the set $\widehat{S}_{\boldsymbol{\theta},I}$. To make the notation consistent, we denote the estimators based on the $Data_{II}$ as $\widetilde{\boldsymbol{\alpha}}_{\tau,II}$ and $\widetilde{\boldsymbol{\theta}}_{\tau,II}$, which is from

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II}) = &\arg\min_{\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau \in \widehat{\boldsymbol{T}}_{\boldsymbol{\theta},I}} \frac{1}{n_1} \sum_{i=1}^{n_1} (\rho_\tau * \mathcal{K}_h)\left(Y_i - \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau\right) \\ &+ \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau,\end{aligned} \tag{3.54}$$

then $\widehat{\boldsymbol{\theta}}_{\tau,II}$ includes those zero terms obtained from (3.53). Similar to the proof in [23] and [12], it is easy to have $\widetilde{\boldsymbol{\alpha}}_{\tau,I} \to \boldsymbol{\alpha}_\tau^0$, $\widehat{\boldsymbol{\alpha}}_{\tau,II} \to \boldsymbol{\alpha}_\tau^0$, $\widetilde{\boldsymbol{\theta}}_{\tau,I} \to \boldsymbol{\theta}_\tau^0$, and $\widehat{\boldsymbol{\theta}}_{\tau,II} \to \boldsymbol{\theta}_\tau^0$ in probability.

Recall that $\widehat{\boldsymbol{\alpha}}_\tau^{(b)}$ $\widehat{\boldsymbol{\theta}}_\tau^{(b)}$ are the estimates which satisfy:

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_\tau^{(b)}, \widehat{\boldsymbol{\theta}}_\tau^{(b)}) = &\arg\min_{\boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau} \frac{1}{n_1} \sum_{i=1}^{n_1} (\rho_\tau * \mathcal{K}_h)\left(Y_i^{(b)} - \boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau - \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau\right) \\ &+ \boldsymbol{\theta}_\tau^T (\boldsymbol{\Gamma} \otimes \boldsymbol{V}) \boldsymbol{\theta}_\tau.\end{aligned} \tag{3.55}$$

Based on the bootstrapped sample, denote the following derivative function

$$\begin{aligned}\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) = &\left\{G_h(\boldsymbol{Z}_i^T \boldsymbol{\alpha}_\tau + \boldsymbol{U}_i^T \boldsymbol{\theta}_\tau - Y_i^{(b)}) - \tau\right\} (\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T \\ &+ 2 \begin{pmatrix} \mathbf{0}_{d \times d} & \\ & \boldsymbol{\Gamma} \otimes \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_\tau \\ \boldsymbol{\theta}_\tau \end{pmatrix}\end{aligned} \tag{3.56}$$

and the estimator $(\widehat{\boldsymbol{\alpha}}_\tau^{(b)}, \widehat{\boldsymbol{\theta}}_\tau^{(b)})$ based on the bootstrapped dataset is the solution to the equation:

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \boldsymbol{\mathcal{G}}(Y_i^{(b)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\theta}_\tau) = 0. \tag{3.57}$$

Consider the following identity

$$
\begin{aligned}
&E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})\right\}\\
=&-\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}\\
&+\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}-E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}\\
=&-\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})\\
&+\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}-E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}.
\end{aligned}
\tag{3.58}
$$

**(1) First, we want to show that**

$$
\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}-E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}=o_p(1),
\tag{3.59}
$$

it suffices by Chebyshev's inequality to show that

$$
\frac{1}{n_1}Cov\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}\to 0.
\tag{3.60}
$$

We have

$$
\begin{aligned}
&Cov\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_\tau^{(b)},\widehat{\boldsymbol{\theta}}_\tau^{(b)})\right\}\\
=&Var\left\{G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_\tau^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_\tau^{(b)}-Y_i^{(b)})\right\}(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)\\
=&E\left[G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_\tau^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_\tau^{(b)}-Y_i^{(b)})\right]^2(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)\\
&-\left\{E\left[G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_\tau^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_\tau^{(b)}-Y_i^{(b)})\right]\right\}^2(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T).
\end{aligned}
\tag{3.61}
$$

Let $\Delta=\boldsymbol{Z}_i^T(\boldsymbol{\alpha}_\tau-\widehat{\boldsymbol{\alpha}}_{\tau,II})+\boldsymbol{U}_i^T(\boldsymbol{\theta}_\tau-\widehat{\boldsymbol{\theta}}_{\tau,II})$ and $\widehat{\Delta}^{(b)}=\boldsymbol{Z}_i^T(\widehat{\boldsymbol{\alpha}}_{\tau,II}-\widehat{\boldsymbol{\alpha}}_\tau^{(b)})+\boldsymbol{U}_i^T(\widehat{\boldsymbol{\theta}}_{\tau,II}-\widehat{\boldsymbol{\theta}}_\tau^{(b)})$.
Note that

$$
\begin{aligned}
&\frac{\partial}{\partial e_i}\left[G_h(-w_i|\widehat{e}_i|)-G_h(-w_i|\widehat{e}_i|-\widehat{\Delta}^{(b)})\right]\\
=&\frac{\partial}{\partial e_i}\left[G_h(-w_i|e_i+\Delta|)-G_h(-w_i|e_i+\Delta|-\widehat{\Delta}^{(b)})\right]\\
=&-w_i\text{sgn}(e_i+\Delta)\left[\mathcal{K}_h(-w_i|e_i+\Delta|)-\mathcal{K}_h(-w_i|e_i+\Delta|-\widehat{\Delta}^{(b)})\right]
\end{aligned}
\tag{3.62}
$$

then when covariates are given, by integration by parts, we have

$$
\begin{aligned}
& E\left[G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau}^{(b)}-Y_i^{(b)})\right] \\
=& E\left[G_h(-w_i|\widehat{e}_i|)-G_h(-w_i|\widehat{e}_i|-\widehat{\Delta}^{(b)})\right] \\
=& E\left\{w_i\int \operatorname{sgn}(\varepsilon+\Delta)F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\left[\mathcal{K}_h(-w_i|e_i+\Delta|)-\mathcal{K}_h(-w_i|e_i+\Delta|-\widehat{\Delta}^{(b)})\right]d\varepsilon\right\} \\
=& E\left\{w_i\int \operatorname{sgn}(\varepsilon+\Delta)F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\mathcal{K}_h(-w_i|e_i+\Delta|)d\varepsilon\right\} \\
& - E\left\{w_i\int \operatorname{sgn}(\varepsilon+\Delta)F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\mathcal{K}_h(-w_i|e_i+\Delta|-\widehat{\Delta}^{(b)})d\varepsilon\right\}.
\end{aligned}
\tag{3.63}
$$

When $w_i>0$, using a change of variable, we have

$$
\begin{aligned}
& w_i\int \operatorname{sgn}(\varepsilon+\Delta)F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\mathcal{K}_h(-w_i|e_i+\Delta|)d\varepsilon \\
=& \int_{\Delta}^{+\infty} w_i\int F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\mathcal{K}_h(-w_i(e_i+\Delta))d\varepsilon \\
& -\int_{-\infty}^{\Delta} w_i\int F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon)\mathcal{K}_h(w_i(e_i+\Delta))d\varepsilon \\
=& \int_{0}^{-\infty} -F_{e|\boldsymbol{Z},\boldsymbol{X}}(-hv/w_i-\Delta)\mathcal{K}(v)dv-\int_{-\infty}^{0}F_{e|\boldsymbol{Z},\boldsymbol{X}}(hv/w_i-\Delta)\mathcal{K}(v)dv \\
=& \int_{-\infty}^{0} F_{e|\boldsymbol{Z},\boldsymbol{X}}(-hv/w_i-\Delta)\mathcal{K}(v)dv-\int_{-\infty}^{0}F_{e|\boldsymbol{Z},\boldsymbol{X}}(hv/w_i-\Delta)\mathcal{K}(v)dv \\
\approx& \int_{-\infty}^{0}\left[F_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)(-hv/w_i-\Delta)\right]\mathcal{K}(v)dv \\
& -\int_{-\infty}^{0}\left[F_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)(hv/w_i-\Delta)\right]\mathcal{K}(v)dv \\
=& -2h/w_i f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\int_{-\infty}^{0} v\mathcal{K}(v)dv+o_p(1),
\end{aligned}
$$

and

$$
\begin{aligned}
& w_i \int \operatorname{sgn}(\varepsilon+\Delta) F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) \mathcal{K}_h(-w_i|e_i+\Delta| - \widehat{\Delta}^{(b)}) d\varepsilon \\
=& \int_{\Delta}^{+\infty} w_i \int F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) \mathcal{K}_h(-w_i(e_i+\Delta) - \widehat{\Delta}^{(b)}) d\varepsilon \\
& - \int_{-\infty}^{\Delta} w_i \int F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) \mathcal{K}_h(w_i(e_i+\Delta) - \widehat{\Delta}^{(b)}) d\varepsilon \\
=& \int_{0}^{-\infty} -F_{e|\boldsymbol{Z},\boldsymbol{X}}(-(hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \mathcal{K}(v) dv - \int_{-\infty}^{0} F_{e|\boldsymbol{Z},\boldsymbol{X}}((hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \mathcal{K}(v) dv \\
=& \int_{-\infty}^{0} F_{e|\boldsymbol{Z},\boldsymbol{X}}(-(hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \mathcal{K}(v) dv - \int_{-\infty}^{0} F_{e|\boldsymbol{Z},\boldsymbol{X}}((hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \mathcal{K}(v) dv \\
\approx& \int_{-\infty}^{0} \left[ F_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)(-(hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \right] \mathcal{K}(v) dv \\
& - \int_{-\infty}^{0} \left[ F_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)((hv+\widehat{\Delta}^{(b)})/w_i - \Delta) \right] \mathcal{K}(v) dv \\
=& -2h/w_i f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \int_{-\infty}^{0} v\mathcal{K}(v) dv - 2\widehat{\Delta}^{(b)} w_i^{-1} f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \int_{-\infty}^{0} \mathcal{K}(v) dv + o_p(1) \\
=& -2h/w_i f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \int_{-\infty}^{0} v\mathcal{K}(v) dv - \widehat{\Delta}^{(b)} w_i^{-1} f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + o_p(1)
\end{aligned}
$$

Similarly, when $w_i < 0$, we have

$$
\begin{aligned}
& w_i \int \operatorname{sgn}(\varepsilon+\Delta) F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) \mathcal{K}_h(-w_i|e_i+\Delta|) d\varepsilon \\
=& -2h/w_i f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \int_{-\infty}^{0} v\mathcal{K}(v) dv + o_p(1),
\end{aligned}
$$

and

$$
\begin{aligned}
& w_i \int \operatorname{sgn}(\varepsilon+\Delta) F_{e|\boldsymbol{Z},\boldsymbol{X}}(\varepsilon) \mathcal{K}_h(-w_i|e_i+\Delta| - \widehat{\Delta}^{(b)}) d\varepsilon \\
=& -2h/w_i f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) \int_{-\infty}^{0} v\mathcal{K}(v) dv + \widehat{\Delta}^{(b)} w_i^{-1} f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + o_p(1).
\end{aligned}
$$

Combing these results and (3.63), we have

$$
E\left[ G_h(\boldsymbol{Z}_i^T \widehat{\boldsymbol{\alpha}}_{\tau,II} + \boldsymbol{U}_i^T \widehat{\boldsymbol{\theta}}_{\tau,II} - Y_i^{(b)}) - G_h(\boldsymbol{Z}_i^T \widehat{\boldsymbol{\alpha}}_{\tau}^{(b)} + \boldsymbol{U}_i^T \widehat{\boldsymbol{\theta}}_{\tau}^{(b)} - Y_i^{(b)}) \right] = \widehat{\Delta}^{(b)} f_{e|\boldsymbol{Z},\boldsymbol{X}}(0) + o_p(\widehat{\Delta}^{(b)}). \tag{3.64}
$$

Let $w_i|e_i+\Delta^*|$ lies between $w_i|\widehat{e}_i|$ and $w_i|\widehat{e}_i|-\widehat{\Delta}^{(b)}$, then by Lagrange mean value theorem and Condition 4, we can get

$$\begin{aligned}
&E\left[G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau}^{(b)}-Y_i^{(b)})\right]^2\\
=&E\left[G_h(-w_i|\widehat{e}_i|)-G_h(-w_i|\widehat{e}_i|-\widehat{\Delta}^{(b)})\right]^2\\
=&E\left[-w_i\text{sgn}(e_i+\Delta^*)\mathcal{K}_h(-w_i|e_i+\Delta^*|)\widehat{\Delta}^{(b)}\right]^2\\
=&o_p(1/h).
\end{aligned}\tag{3.65}$$

By Conditions 1, 3, 4, and 8, combined (3.61), (3.63) and (3.65), (3.60) holds. So we can show that

$$n_1^{-1}\sum_{i=1}^{n_1}\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)},\widehat{\boldsymbol{\theta}}_{\tau}^{(b)})\right\}-E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)},\widehat{\boldsymbol{\theta}}_{\tau}^{(b)})\right\}=o_p(1).$$

**(2) Second, we want to prove**

$$\sqrt{n_1}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_{\tau,II}-\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)}\\ \widehat{\boldsymbol{\theta}}_{\tau,II}-\widehat{\boldsymbol{\theta}}_{\tau}^{(b)}\end{pmatrix}=E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}^{-1}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})+o_p(n^{-1/2}).$$

When covariates are given and Conditions 6-9 hold, we have

$$\begin{aligned}
&E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)},\widehat{\boldsymbol{\theta}}_{\tau}^{(b)})-\boldsymbol{\mathcal{G}}(Y_i^{(b)},\widehat{\boldsymbol{\alpha}}_{\tau,II},\widehat{\boldsymbol{\theta}}_{\tau,II})\right\}\\
=&E\left[G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau}^{(b)}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau}^{(b)}-Y_i^{(b)})-G_h(\boldsymbol{Z}_i^T\widehat{\boldsymbol{\alpha}}_{\tau,II}+\boldsymbol{U}_i^T\widehat{\boldsymbol{\theta}}_{\tau,II}-Y_i^{(b)})\right](\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)^T+o_p(1)\\
=&E\left[G_h(-w_i|\widehat{e}_i|-\widehat{\Delta}^{(b)})-G_h(-w_i|\widehat{e}_i|)\right](\boldsymbol{Z}_i^T,\boldsymbol{U}_i^T)^T+o_p(1).
\end{aligned}\tag{3.66}$$

For any vector $\boldsymbol{u}_1$, $\boldsymbol{u}_2\in\mathbb{R}^d$ and $\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^{m\times(K+p)}$, similar with the proof of (3.63), we have

$$\begin{aligned}
&E\left[G_h(-w_i|e_i-\boldsymbol{Z}_i^T\boldsymbol{u}_1-\boldsymbol{U}_i^T\boldsymbol{v}_1|-\boldsymbol{Z}_i^T\boldsymbol{u}_2-\boldsymbol{U}_i^T\boldsymbol{v}_2)-G_h(-w_i|e_i-\boldsymbol{Z}_i^T\boldsymbol{u}_1-\boldsymbol{U}_i^T\boldsymbol{v}_1|)\right]\\
=&-(\boldsymbol{Z}_i^T\boldsymbol{u}_2+\boldsymbol{U}_i^T\boldsymbol{v}_2)E\left[\mathcal{K}_h(-w_i|e_i-\boldsymbol{Z}_i^T\boldsymbol{u}_1-\boldsymbol{U}_i^T\boldsymbol{v}_1|)\right]\\
=&-(\boldsymbol{Z}_i^T\boldsymbol{u}_2+\boldsymbol{U}_i^T\boldsymbol{v}_2)\left[f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+O(\boldsymbol{Z}_i^T\boldsymbol{u}_1+\boldsymbol{U}_i^T\boldsymbol{v}_1)\right].
\end{aligned}\tag{3.67}$$

When $\|\boldsymbol{u}\|_2=O_p(n^{-1/2})$ and $\|\boldsymbol{u}\|_2=O_p(n^{-1/2}K)$, we have

$$\begin{aligned}
&E\left[G_h(-w_i|e_i-\boldsymbol{Z}_i^T\boldsymbol{u}_1-\boldsymbol{U}_i^T\boldsymbol{v}_1|-\boldsymbol{Z}_i^T\boldsymbol{u}_2-\boldsymbol{U}_i^T\boldsymbol{v}_2)-G_h(-w_i|e_i-\boldsymbol{Z}_i^T\boldsymbol{u}_1-\boldsymbol{U}_i^T\boldsymbol{v}_1|)\right]\\
=&-(\boldsymbol{Z}_i^T\boldsymbol{u}_2+\boldsymbol{U}_i^T\boldsymbol{v}_2)f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)+o_p(n^{-1/2}).
\end{aligned}\tag{3.68}$$

By (3.66), (3.67) and (3.68), we have

$$\begin{aligned} & E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_\tau^{(b)}, \widehat{\boldsymbol{\theta}}_\tau^{(b)}) - \boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II})\right\} \\ = & - E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)\right\}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_{\tau,II} - \widehat{\boldsymbol{\alpha}}_\tau^{(b)} \\ \widehat{\boldsymbol{\theta}}_{\tau,II} - \widehat{\boldsymbol{\theta}}_\tau^{(b)}\end{pmatrix} + o_p(n_1^{-1/2}). \end{aligned} \tag{3.69}$$

Inserting (3.69) to (3.58) and by (3.59), we further have

$$\begin{aligned} & E\left\{\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_\tau^{(b)}, \widehat{\boldsymbol{\theta}}_\tau^{(b)}) - \boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II})\right\} + o_p(1) \\ = & - E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)\right\}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_{\tau,II} - \widehat{\boldsymbol{\alpha}}_\tau^{(b)} \\ \widehat{\boldsymbol{\theta}}_{\tau,II} - \widehat{\boldsymbol{\theta}}_\tau^{(b)}\end{pmatrix} \\ = & - \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II}) + o_p(1). \end{aligned} \tag{3.70}$$

Then, we have

$$\sqrt{n_1}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_{\tau,II} - \widehat{\boldsymbol{\alpha}}_\tau^{(b)} \\ \widehat{\boldsymbol{\theta}}_{\tau,II} - \widehat{\boldsymbol{\theta}}_\tau^{(b)}\end{pmatrix} = E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z} \\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n_1}\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II}) + o_p(n_1^{-1/2}), \tag{3.71}$$

where

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\boldsymbol{\mathcal{G}}(Y_i^{(b)}, \widehat{\boldsymbol{\alpha}}_{\tau,II}, \widehat{\boldsymbol{\theta}}_{\tau,II}) = \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left\{G_h(-w_i|\widehat{e}_i|) - \tau\right\}(\boldsymbol{Z}_i^T, \boldsymbol{U}_i^T)^T + o_p(1). \tag{3.72}$$

For $G_h(-w_i|\widehat{e}_i|)$, when covariates are given, we have

$$E(G_h(-w_i|\widehat{e}_i|)) = \tau - E(w_i^{-1})2hf_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\int_{-\infty}^{0} v\mathcal{K}(v)dv + o_p(1) = \tau + o_p(1),$$

and

$$E(G_h^2(-w_i|\widehat{e}_i|)) = \tau + o_p(1),$$

then asymptotic mean and variance of $\{G_h(-w_i|\widehat{e}_i|) - \tau\}$ are 0 and $\tau(1-\tau)$, respectively.

Finally, we can have

$$\sqrt{n}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right)^{-1/2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau^{(b)} - \widehat{\boldsymbol{\alpha}}_{\tau,II} \\ \widehat{\boldsymbol{\theta}}_\tau^{(b)} - \widehat{\boldsymbol{\theta}}_{\tau,II}\end{pmatrix} \xrightarrow{d} \mathbb{N}(\boldsymbol{0}, \tau(1-\tau)\boldsymbol{I}), \tag{3.73}$$

where

$$\boldsymbol{\Sigma}_1 = E\left\{\left((\boldsymbol{Z}^T, \boldsymbol{U}^T)^T\right)^{\otimes 2}\right\}, \quad \boldsymbol{\Sigma}_2 = E\left\{f_{e|\boldsymbol{Z},\boldsymbol{X}}(0)\begin{pmatrix}\boldsymbol{Z}\\ \boldsymbol{U}\end{pmatrix}^{\otimes 2}\right\}.$$

Correspondingly, we have

$$\sqrt{n}\left(\boldsymbol{\Sigma}_{\tau,2}^{-1}\boldsymbol{\Sigma}_{\tau,1}\boldsymbol{\Sigma}_{\tau,2}^{-1}\right)^{-1/2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_\tau^{(b)} - \widehat{\boldsymbol{\alpha}}_{\tau,II}\\ \widehat{\boldsymbol{\theta}}_\tau^{(b),(1)} - \widehat{\boldsymbol{\theta}}_{\tau,II}^{(1)}\end{pmatrix} \xrightarrow{d} \mathbb{N}(\boldsymbol{0}, \tau(1-\tau)\boldsymbol{I}). \tag{3.74}$$

This completes the proof of Theorem 4 by Theorem 2-3 and the asymptotic result above.

# Chapter 4

# Locally Sparse Estimation in Simultaneous Functional Quantile Regression

## 4.1. Introduction

In this project, our primary goal remains understanding how soybean yield is influenced by key factors such as temperature, precipitation, and machinery level on farms. Additionally, we aim to identify specific time regions during the growing season when temperature has a significant impact on soybean yields.

Building upon the previous chapter's work, we now propose an extension of the locally sparse estimation method introduced in Chapter 3. In the previous chapter, we focused on locally sparse estimation for a single quantile at a time during model fitting. However, in this project, we aim to expand this method to perform locally sparse estimation for multiple quantiles simultaneously. By doing so, we can gain a more comprehensive understanding of how soybean yield varies across different quantiles.

The locally sparse estimation method is particularly valuable in this project, as it helps us identify critical time regions during the growing season when temperature significantly influences soybean yields. By focusing on quantile-specific effects, we can pinpoint periods when temperature has a substantial impact on yield outcomes, beyond traditional mean-based analyses.

Given the importance of analyzing crop yield dynamics, we continue using the same soybean yield data set as in Chapter 3. In this project, we consider the following functional quantile regression model to study the relationship between the predictors and soybean yields,

$$Q_Y(u|\boldsymbol{Z}, X) = \boldsymbol{Z}^\tau \boldsymbol{\alpha}(u) + \int_0^T \beta(t,u)X(t)dt, \tag{4.1}$$

where $Y$ is a scalar response, $\boldsymbol{Z}$ is a vector of scalar predictors and $X(t)$ is a functional predictor defined over $[0, T]$.

In our soybean yield study, $Y$ represents the annual soybean yield, which is the response variable we are interested in modeling and predicting. $X(t)$ denotes the daily average temperature, which is a functional predictor that varies with time $t$ throughout the growing season. $Z_1$ represents the annual precipitation, which is a scalar predictor representing the amount of rainfall received during the year. $Z_2$ is the ratio of the irrigated area of each county in Kansas. It serves as another scalar predictor, describing the proportion of land that is irrigated for soybean cultivation in each county.

In the functional quantile regression model (4.1), each entry of the vector $\boldsymbol{\alpha}(u)$ corresponds to a varying coefficient that characterizes the influence of a scalar predictor on the $u$-quantile of the response variable $Y$. This varying coefficient allows us to capture how the impact of the scalar predictors changes across different quantiles of the soybean yield distribution. The function $\beta(t, u)$ is a bivariate slope function that is indexed by both time $t$ and quantile $u$. It describes the dynamic influence of the functional predictor $X(t)$ on the quantiles of the soybean yield $Y$.

Reviewing the literature [72, 79], we know that there exists a comfortable range of temperature for soybean growth, where small fluctuations of temperature have no influence on soybean yields. This observation serves as a strong motivation for introducing the concept of local sparsity for the functional coefficient $\beta(t, u)$ in the functional quantile regression model (4.1).

The local sparsity refers to the property of having sub-regions within the domain of $\beta(t, u)$ where $\beta(t, u) = 0$. In other words, there are specific time intervals (sub-regions) during the growing season where the effect of the daily average temperature on some quantiles of soybean yield is non-existent. These sub-regions correspond to the comfortable temperature range for soybean growth, where temperature fluctuations do not substantially impact the crop's performance for some specific quantiles. More formally, we assume that there exists a collection of sub-regions $\mathcal{N} \subset [0, T] \times (0, 1)$, and $\beta(t, u) = 0$ for all $(t, u) \in \mathcal{N}$.

By incorporating local sparsity into the model, we can accurately describe the dynamic dependence between quantiles of soybean yield and the functional predictor, daily average temperature. More specifically, this property allows the model to focus on the periods when temperature has a meaningful influence on soybean growth and yield outcomes, while disregarding the sub-regions where temperature fluctuations have no effect, and to provide more precise and meaningful insights into how temperature impacts soybean yield across different quantiles.

Functional quantile regression is indeed an important and popular tool for data analysis, particularly in applications related to environmental science. It serves as an extension of classic quantile regression [44] and allows for more flexible and comprehensive modeling of data involving functional predictors. Researchers have developed various variants of functional quantile regression to address different scenarios and accommodate diverse data characteristics. For instance, [86] studied a model with multiple functional covariates and a finite number of scalar covariates. [85] consider a partially functional quantile regression with a functional covariate and high-dimensional scalar covariates. [55] proposed a functional partially linear model with multiple functional covariates and ultrahigh-dimensional scalar covariates.

Indeed, many existing works that consider local sparsity in the context of functional data analysis have focused on functional linear models. These models involve functional predictors and have proven to be valuable in various applications, [80, 39, 94]. For example, [50] proposed a functional generalization of ordinary SCAD [19], called fSCAD, to obtain the locally sparse estimator for univariate scalar-on-function regression. [22] and [49] applied fSCAD [50] to the multiple outputs functional linear regression.

The conventional strategy of fitting the functional quantile regression model (4.1) by fixing the quantile $u$ and estimating each entry of $\boldsymbol{\alpha}(u)$ as a scalar parameter, along with estimating $\beta(t, u)$ as a univariate function of $t$, does offer advantages in reducing the dimensionality of the parameter space. However, it also introduces two major limitations: lack of smoothness in the estimation of the slope function $\beta(t, u)$ and non-monotonicity of the estimated conditional quantile function of the response $Y$.

When the model is fitted separately for different quantiles, the resulting estimator for the slope function $\beta(t, u)$ is not guaranteed to be smooth in the direction of the quantile $u$. This lack of smoothness can lead to erratic behavior of the estimated $\beta(t, u)$ across quantiles, which may not accurately capture the true underlying relationship between the functional predictor and response variable. In addition, fitting the model for different quantiles separately may lead to non-monotonic conditional quantile estimation, while the true conditional quantile function of $Y$ should be monotonically non-decreasing in $u$. The crossing quantiles can result in invalid interpretations.

## 4.2. Estimation

Let $Y$ denote a scalar response. Let $\boldsymbol{Z} = (1, Z_1, \ldots, Z_p)$ be a vector of scalar predictors and $X(t)$ be a functional predictor over $[0, T]$. For any quantile $u \in (0, 1)$, we propose the following functional linear quantile regression model,

$$Q_Y(u|\boldsymbol{Z}, X) = \boldsymbol{Z}^\tau \boldsymbol{\alpha}(u) + \int_0^T \beta(t, u)X(t)dt, \tag{4.2}$$

where $\boldsymbol{\alpha}(u) = (\alpha_0(u), \ldots, \alpha_p(u))^\tau$ and $\beta(t, u)$ are unknown functions, and $\beta(t, u)$ exhibits the local sparsity property.

Estimating the infinite-dimensional functions $\boldsymbol{\alpha}(u)$ and $\beta(t, u)$ directly can be challenging in practice. To address this issue, a commonly used strategy is to first find finite-dimensional approximations for these unknown functions and then estimate the coefficients of these approximations using the data. When the univariate functions in $\boldsymbol{\alpha}(u)$ are assumed to be smooth enough, a popular approximation method is B-splines, where the entries of $\boldsymbol{\alpha}(u)$ are represented as linear combinations of B-spline basis functions with unknown coefficients,

$$\alpha_k(u) = \sum_{j=1}^{n_b} \eta_{k,j} b_j(u) := \boldsymbol{b}(u)^\tau \boldsymbol{\eta}_k,$$

where $\boldsymbol{b}(u) = (b_1(u), \ldots, b_{n_b}(u))^\tau$ is the vector of B-splines over $(0,1)$ and $\{\eta_{k,j}\}_{j=1}^{n_b}$ are the unknown coefficients.

In the context of the proposed functional quantile regression model with locally sparse slope function $\beta(t,u)$, the estimation involves identifying the subset $S_0 = \{(t,u) \in [0,T] \times (0,1) : \beta(t,u) = 0\}$, which represents the sub-regions of the domain of $\beta(t,u)$ where $\beta(t,u)$ is assumed to be zero. Let $\hat{S}_0$ denote the identified $S_0$ using the data. The next step is to estimate the function $\beta(t,u)$ on the complement set $[0,T] \times (0,1) \backslash \hat{S}_0$. This complement set consists of the regions where the function $\beta(t,u)$ is non-zero, and these are the regions of interest.

Similarly, when $\beta(t,u)$ is smooth enough, we approximate $\beta(t,u)$ using bivariate splines. Regarding the bivariate spline basis, there are various choices available for approximation purposes. Two commonly used options are tensor products of B-splines [75, 64, 91] and bivariate Bernstein polynomials over the triangulation [46]. In this project, the choice is made to use bivariate Bernstein polynomials over a triangulation for approximating the unknown slope function $\beta(t,u)$ in (4.6). Each bivariate Bernstein polynomial has its support on a single triangle of the triangulation, and multiple polynomials are defined over each triangle. This local support property ensures that we can use the triangles to estimate the subset $S_0$.

Suppose that $\mathscr{U}$ is the interval containing the quantiles of interest. Let $S_d^r(\Delta)$ denote the linear space spanned by bivariate splines defined over a triangulation $\Delta$, where $r$ is the smoothness condition of this space and $d$ is the degree of the splines. Our goal is to find a function $s(t,u) \in S_d^r(\Delta)$ that can effectively approximate the slope function $\beta(t,u)$ on the domain $[0,T] \times \mathscr{U}$. To make our writing and proofs in the subsequent sections more clear, we use $\{b_j(t,u)\}_{j=1}^{n_B}$ to denote the Bernstein polynomials over the triangulation $\Delta = \{\Lambda_1, \ldots, \Lambda_M\}$, where $j = 1, \ldots, n_B$ is the index for the polynomials. The Bernstein polynomials are locally defined on each triangle of the triangulation $\Delta$. More specifically, $(d+2)(d+1)/2$ Bernstein polynomials are defined on each triangle of $\Delta$. Therefore, the relationship between $n_B$ and $M$ is $n_B = (d+2)(d+1)M/2$ In addition, for each bivariate basis function $b_j(t,u)$, we denote its support by $\Delta_j$, which is a specific triangle of $\Delta$ with $\Delta_j =$ the triangle of $\Delta$ that is the support of $b_j(t,u)$. In other words, $b_j(t,u) \neq 0$ for $(t,u) \in \Delta_j$, and $b_j(t,u) = 0$ for $(t,u) \notin \Delta_j$. If two Bernstein polynomials $b_j(t,u)$ and $b_k(t,u)$ are associated with the same triangle, then $\Delta_j$ and $\Delta_k$ are identical.

The function $s(t,u) \in S_d^r(\mathscr{T} \times \mathscr{U})$ that approximates $\beta(t,u)$ can be written as a linear combination of Bernstein polynomials $\{B_j(t,u)\}_{j=1}^{n_B}$. Then on the domain $[0,T] \times \mathscr{U}$, we have the approximation

$$\beta(t,u) \approx s(t,u) = \sum_{j=1}^{n_B} \gamma_j B_j(t,u) := \boldsymbol{B}(t,u)^\tau \boldsymbol{\gamma} \in S_d^r(\Delta), \tag{4.3}$$

where $\boldsymbol{B}(t,u) = (B_1(t,u), \ldots, B_{n_B}(t,u))^\tau$ is the vector of bivariate splines over $\mathscr{T} \times \mathscr{U}$ and $\{\gamma_j\}_{j=1}^{n_B}$ are the corresponding coefficients.

To ensure the desired smoothness of the approximation for $\beta(t,u)$, such as continuity or continuity of derivatives, it is necessary to impose linear constraints on the coefficients $\boldsymbol{\gamma}$. These constraints can be represented as $\boldsymbol{H}\boldsymbol{\gamma} = 0$, where $\boldsymbol{H}$ is a matrix of linear constraints. By incorporating these

constraints, we can enforce the desired smoothness property in the estimation. However, imposing linear constraints on the coefficients may lead to complications in solving the following optimization problem. To overcome this, a common approach is to use QR decomposition to remove the linear constraints. For a given $\boldsymbol{H}$, by QR decomposition, we have

$$\boldsymbol{H}^{\mathrm{T}} = (\boldsymbol{Q}^*, \boldsymbol{Q}) \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix}, \tag{4.4}$$

where $(\boldsymbol{Q}^*, \boldsymbol{Q})$ is a matrix with orthogonal columns and $\boldsymbol{R}$ is a upper triangle matrix with nonzero diagonal elements. Based on the decomposition (4.4), the constraints $\boldsymbol{H}\boldsymbol{\gamma} = \boldsymbol{0}$ can be removed by rewriting $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma} = \boldsymbol{Q}\boldsymbol{\theta}.$$

Then, the integral $Q_Y(u|X)$ can be approximated by

$$\begin{aligned} Q_Y(u|X, \boldsymbol{Z}) &\approx \boldsymbol{Z}^\tau \boldsymbol{D}(u)\boldsymbol{\eta} + \int \boldsymbol{B}(t,u)^\tau \boldsymbol{Q}\boldsymbol{\theta} X(t) dt \\ &= \boldsymbol{Z}^\tau \boldsymbol{D}(u)\boldsymbol{\eta} + \boldsymbol{A}(u)^\tau \boldsymbol{Q}\boldsymbol{\theta}, \end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_0^\tau, \ldots, \boldsymbol{\eta}_p^\tau)^\tau$, $\boldsymbol{A}(u)^\tau = \int \boldsymbol{B}(t,u)^\tau X(t) dt$, and

$$\boldsymbol{D}(u) = \begin{pmatrix} \boldsymbol{b}(u)^\tau & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{b}(u)^\tau & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \boldsymbol{0} & \boldsymbol{b}(u)^\tau \end{pmatrix}$$

In the proposed estimation procedure, the goal is to estimate the univariate functions $\boldsymbol{\alpha}(u)$ and the bivariate function $\beta(t,u)$ by considering all the quantiles of interest simultaneously. Combining multiple quantile regression models in the estimation process has been widely recognized for its advantages. Several papers have discussed and proposed various methods for joint estimation of quantile regression models such as [96, 40, 93, 31].

For a real-valued random variable $Y$, the minimizer of $E\{\rho_u(Y-u)\}$ is the $u$-quantile of $Y$, where $\rho_u(x) = x\,(u - \mathbb{1}\{x < 0\})$ is called quantile loss function or check function [44]. Assume that we observe independent and identically distributed data pairs $\{y_i, z_i, x_i(t)\}_{i=1}^n$ as the realizations of $\{Y, \boldsymbol{Z}, X(t)\}$. To perform the quantile regression, we consider a set of quantiles of interest denoted by $U \in \mathscr{U}$. These quantiles are assumed to be uniformly distributed within the interval $\mathscr{U}$. The cardinality of $U$, denoted by $n_U$, represents the number of quantiles of interest. Then, based on the approximation (4.3), a reasonable estimation method of the unknown approximation coefficients is to minimize the following loss function with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$:

$$\frac{1}{n n_U} \sum_{r=1}^{n_U} \sum_{i=1}^{n} \rho_{u_r} \left( y_i - \boldsymbol{z}_i^\tau \boldsymbol{D}(u_r)\boldsymbol{\eta} - \boldsymbol{A}_i(u_r)^\tau \boldsymbol{Q}\boldsymbol{\theta} \right) \tag{4.5}$$

where $\boldsymbol{A}_i(u_r) = \int \boldsymbol{B}(t,u_r)^\tau x_i(t)dt$.

However, after approximating $\beta(t,u)$ using bivariate splines, the resulting design matrix involving $\boldsymbol{z}_i$, $\boldsymbol{b}(u_r)$, and $A_i(u_r)$ may become ill-conditioned, leading to a highly oscillatory estimator for $\beta(t,u)$ when directly minimizing (4.5). Consequently, the estimation of $\beta(t,u)$ may not be as smooth as assumed by the model. To address this issue, a common approach is to add a penalty term to (4.5) during the minimization procedure. This addition aims to enhance numerical stability and control the smoothness of the estimation for $\beta(t,u)$. One particularly effective tool for this purpose is the roughness penalty. Detailed discussions on roughness penalty in the context of functional data analysis can be found in [8], [68], [67] and [5]. For a smooth bivariate function $s(t,u)$, the roughness penalty $R(s)$ is defined as

$$R(s) = \sum_{\Lambda\in\Delta}\int_\Lambda \sum_{d_1+d_2=2}\binom{2}{d_1}\left[\nabla_t^{d_1}\nabla_u^{d_2}s(t,u)\right]^2 dtdu. \tag{4.6}$$

For the bivariate spline approximation $\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{\theta}$, the roughness penalty of $\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{\theta}$(4.6) can be further written as a quadratic form with a positive semi-definite matrix $\boldsymbol{G}$,

$$R(\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{\theta}) = \boldsymbol{\theta}^\tau\boldsymbol{G}\boldsymbol{\theta}.$$

With the help of the roughness penalty $R(\cdot)$, we propose to estimate the unknown coefficients $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ in (4.6) by minimizing the following objective function

$$\frac{1}{nn_U}\sum_{r=1}^{n_U}\sum_{i=1}^{n}\rho_{u_r}\left(y_i - \boldsymbol{z}_i^\tau\boldsymbol{D}(u_r)\boldsymbol{\eta} - \boldsymbol{A}_i(u_r)^\tau\boldsymbol{Q}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta}^\tau\boldsymbol{G}\boldsymbol{\theta}. \tag{4.7}$$

### 4.2.1 Locally Sparse Estimation for $\beta(t,u)$

Our primary objective is to obtain a locally sparse estimator for $\beta(t,u)$, which helps identify the inactive regions of the functional predictor $X(t)$ across different quantiles of the scalar response $Y$. To achieve this goal, we employ a group lasso-type method.

By utilizing the triangulation-based spline approximation, we can introduce penalization on the $L_2$-norm of the approximation of $\beta(t,u)$ (4.3) at the triangle level. This enables us to promote the local sparsity of the target function $\beta(t,u)$ at the level of triangles. In essence, the group lasso-type penalties encourages the bivariate splines over certain triangles to have zero coefficients, which can help identify the inactive regions of the functional predictor $X(t)$ for specific quantiles of $Y$.

Denote the minimizer of (4.7) as $(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}})$ and define

$$w_j = \left\|\tilde{\boldsymbol{\gamma}}_{[j]}\right\|_2^{-\eta},$$

where $\tilde{\gamma} = \boldsymbol{Q}\tilde{\boldsymbol{\theta}}$ and $\tilde{\gamma}_{[j]}$ denote the entries of $\tilde{\gamma}$ corresponding to the triangle $\Delta_j$. Then we minimize the following objective function with respect to $(\boldsymbol{\eta}, \boldsymbol{\theta})$

$$\frac{1}{nn_U}\sum_{r=1}^{n_U}\sum_{i=1}^{n}\rho_{u_r}(y_i - \boldsymbol{z}_i^\tau \boldsymbol{D}(u_r)\boldsymbol{\eta} - \boldsymbol{A}_i(u_r)^\tau \boldsymbol{Q}\boldsymbol{\theta}) + \lambda_1 \boldsymbol{\theta}^\tau \boldsymbol{G}\boldsymbol{\theta} + \lambda_2 \sum_{j=1}^{J} w_j \left\|\gamma_{[j]}\right\|_2, \tag{4.8}$$

where $\gamma = \boldsymbol{Q}\boldsymbol{\theta}$ and $\gamma_{[j]}$ denote the entries of $\gamma$ corresponding to the triangle $\Delta_j$.

### 4.2.2 Parameter Tuning

In the estimation procedure, the tuning parameter $\lambda$ in (4.7) is selected using 10-fold cross-validation. After fixing the tuning parameter $\lambda$, the weights $w_j$ can be computed. These weights are essential for the group lasso-type penalties used to achieve local sparsity in the estimation of $\beta(t, u)$. After we obtain $w_j$, we need to further tune $\lambda_1$ and $\lambda_2$ in (4.8). For this step, we still use 10-fold cross validations. During each validation run, we first identify the triangles with zero coefficients based on the $\hat{\gamma}$ obtained from minimizing (4.8), where $\hat{\gamma} = \boldsymbol{Q}\hat{\boldsymbol{\theta}}$. Then we refit the model only using the active regions of the functional predictor identified by $\hat{\gamma}$ without group lasso-type penalties. To ensure that the identified inactive regions of the functional predictor (triangles with zero coefficients) are properly excluded from the model fitting, we need to impose some linear constraints on $\gamma$ during the estimation procedure,

$$\boldsymbol{K}\gamma = \boldsymbol{0},$$

where the matrix $\boldsymbol{D}$ consists of 0 and 1. Specifically, through the above linear constraints, we can ensure that the coefficients corresponding to the triangles with zero coefficients in $\hat{\gamma}$ remain zero throughout the model fitting. In summary, we use the following procedure to select tuning parameters $(\lambda_1, \lambda_2)$.

- Let $\Lambda_1$ and $\Lambda_2$ denote the candidate sets for $\lambda_1$ and $\lambda_2$ respectively. For any pair of $(\lambda_1, \lambda_2)$, $\lambda_1 \in \Lambda_1$, $\lambda_2 \in \Lambda_2$, we compute the minimizer of (4.8) denoted as $(\hat{\boldsymbol{\eta}}, \hat{\gamma})$.
- Given $(\hat{\boldsymbol{\eta}}, \hat{\gamma})$, we can obtain the constraints matrix $\boldsymbol{D}$ corresponding to $\hat{\gamma}$ and then define $\tilde{\boldsymbol{H}}^\tau = (\boldsymbol{K}^\tau, \boldsymbol{H}^\tau)^\tau$.
- Consider the following minimization problem

$$\min_{\boldsymbol{\eta},\boldsymbol{\theta}} (nn_U)^{-1} \sum_{u_r \in U}\sum_{i=1}^{n} \rho_{u_r}(y_i - \boldsymbol{z}_i^\tau \boldsymbol{D}(u_r)\boldsymbol{\eta} - \boldsymbol{A}_i(u_r)^\tau \boldsymbol{Q}\boldsymbol{\theta}) + \lambda_3 \boldsymbol{\theta}^\tau \boldsymbol{G}\boldsymbol{\theta} \tag{4.9}$$
$$\text{subject to} \quad \tilde{\boldsymbol{H}}\gamma = \boldsymbol{0}.$$

  Let $(\tilde{\boldsymbol{\eta}}, \tilde{\gamma})$ denote the minimizer of (4.9). From the definition of $\tilde{\boldsymbol{H}}$, we know that $\tilde{\gamma}$ and $\hat{\gamma}$ share the same sparsity structure. Run a 10-fold cross validation to select $\lambda_3$ in (4.9), and use the validation loss associated with the selected $\lambda_3$ as the validation loss for $(\lambda_1, \lambda_2)$.
- Use the proposed validation loss to select the best pairs $(\lambda_1, \lambda_2)$.

## 4.3. Theoretical Results

Define

$$\Gamma_X f = \int_s cov(s,t) f(s,u) ds,$$

and

$$\|f(t,u)\|_{\Gamma_X,2} = \left( \int_u \int_t \Gamma_X f \times f dt du \right)^{1/2}.$$

To derive the asymptotic property of the proposed estimator $\hat{\beta}(t,u)$ by minimizing (4.7), we assume that the following hypotheses are satisfied.

(A1) $\{Y_i, \boldsymbol{Z}_i, X_i(t)\}_{i=1}^n$ are i.i.d.

(A2) $\|X\| \leq C_0 < \infty$ a.s.

(A3) The functions in $\boldsymbol{\alpha}(u)$ are supposed to have $p'$th derivatives $\boldsymbol{\alpha}^{(p')}(u)$ such that

$$\left| \alpha_k^{(p')}(t) - \alpha_k^{(p')}(s) \right| \leq C \left| t - s \right|^{\vartheta}, \quad s,t \in (0,1),$$

where $C_1 > 0$ and $\vartheta \in [0,1]$. In what follows, we set $\nu = p' + \vartheta$.

(A4) The eigenvalues of $\Gamma_X$ are positive.

(A5) $\beta(t,u) \in W_q^{d+1}(\mathscr{T} \times \mathscr{U})$.

(A6) The random variable $\epsilon_i$ defined by $\epsilon_i = Y_i - \boldsymbol{Z}_i^\tau \boldsymbol{\alpha}(u) - \int \beta(t,u) X_i(t) dt$ has a density function $f$ continuous and bounded below uniformly by a strictly positive constant at 0.

**Theorem 1.** *Under the assumptions A1-A6, suppose $n_b \sim |\Delta|^{-(d+1)/\nu}$, and $n_U$ is large enough, then*

$$\|\hat{\beta}(t,u) - \beta(t,u)\|_{\Gamma_X,2} = O_p\left( |\Delta|^{d+1} + \frac{1}{n\lambda|\Delta|^2} + \lambda \right),$$

*except on a set whose probability goes to zero as $n$ goes to infinity.*

## 4.4. Simulation Studies

In this section, we perform simulation studies to assess the finite sample performance of the proposed method and the conventional method [6] for estimating the slope function $\beta(t,u)$. We consider three different scenarios to evaluate the performance under varying levels of complexity.

Scenario I and Scenario II both assume that the true slope function $\beta(t,u)$ is independent of $u$ and can be represented as a univariate function of $t$ only. This setting is commonly used in the literature of functional quantile regression. In Scenario III, we introduce a more intricate true slope function $\beta(t,u)$, which is a bivariate function of both $t$ and $u$, adding complexity to the simulation setup.

### 4.4.1 Data Generating Models

**Scenario I.** In this scenario, the realizations of the $\{X(t), \boldsymbol{Z}, Y\}$ are generated from

$$\begin{aligned} Y_i &= \boldsymbol{Z}_i^T \boldsymbol{\alpha} + \int_0^1 X_i(t)\rho_1(t)dt + e_i, \qquad i = 1, \ldots, n, \\ \rho_1(t) &= 20(t-0.5)^2 \mathbb{1}\{t \leq 0.5\}, \end{aligned}$$

where $X_i(t)$ is a Wiener process, $Z_{i,1}, Z_{i,2} \sim N(0, 0.0.5)$, $\boldsymbol{Z}_i = (1, Z_{i,1}, Z_{i,2})^\tau$, $\boldsymbol{\alpha} = (0, 1, 1)^\tau$, and $e_i \sim N(0, 0.05)$. $X_i(t)$, $Z_{i,1}$, $Z_{i,2}$ and $e_i$ are all independent. Under this setting, the underlying slope function $\beta(t, u)$ in model (4.6) is given by

$$\beta(t, u) = \rho_1(t) = 20(t-0.5)^2 \mathbb{1}\{t \leq 0.5\}.$$

**Scenario II.** In this scenario, the realizations of the $\{X(t), \boldsymbol{Z}, Y\}$ are generated from

$$\begin{aligned} Y_i &= \boldsymbol{Z}_i^T \boldsymbol{\alpha} + \int_0^1 X_i(t)\rho_1(t)dt + e_i, \qquad i = 1, \ldots, n, \\ \rho_1(t) &= \sin(2\pi(t-0.25)) \mathbb{1}\{0.25 \leq t \leq 0.75\}, \end{aligned}$$

where $X_i(t)$ is a Wiener process, $Z_{i,1}, Z_{i,2} \sim N(0, 0.0.5)$, $\boldsymbol{Z}_i = (1, Z_{i,1}, Z_{i,2})^\tau$, $\boldsymbol{\alpha} = (0, 1, 1)^\tau$, and $e_i \sim N(0, 0.05)$. $X_i(t)$, $Z_{i,1}$, $Z_{i,2}$ and $e_i$ are all independent. Under this setting, the underlying slope function $\beta(t, u)$ in model (4.6) is given by

$$\beta(t, u) = \rho_1(t) = \sin(2\pi(t-0.25)) \mathbb{1}\{0.25 \leq t \leq 0.75\}.$$

**Scenario III.** In this scenario, the realizations of the pair $\{X(t), \boldsymbol{Z}, Y\}$ are generated from

$$\begin{aligned} Y &= \boldsymbol{Z}_i^T \boldsymbol{\alpha} + \int_0^1 \rho_1(t)X(t)dt + \sigma(X)\epsilon, \quad \sigma(X) = \int_0^1 \rho_2(t)(X(t)+4)dt, \\ \rho_1(t) &= 20(t-0.5)^2 \mathbb{1}\{t \leq 0.5\}, \quad \rho_2(t) = (t-0.5)^2 \mathbb{1}\{t \leq 0.5\} \end{aligned}$$

where $Z_{i,1}, Z_{i,2} \sim N(0, 0.0.5)$, $\boldsymbol{Z}_i = (1, Z_{i,1}, Z_{i,2})^\tau$, $\boldsymbol{\alpha} = (0, 1, 1)^\tau$, and $e_i \sim N(0, 1)$. Regarding the $X_i(t)$ in the above data generating model, for each $i$, we first generate $X_i(t)$ from Wiener process. If $X_i(t)$ satisfies $X_i(t) + 4 \geq 0$, then we keep it as the functional predictor of the $i$-th sample otherwise we will generate a new $X_i(t)$ for the $i$-th sample. We repeat this process until $X_i(t)$ satisfies $X_i(t) + 4 \geq 0$. $X_i(t)$, $Z_{i,1}$, $Z_{i,2}$ and $e_i$ are all independent. Under this setting, the underlying slope function $\beta(t, u)$ in model (4.6) is given by

$$\beta(t, u) = \rho_1(t) + \rho_2(t)Q_e(u) = 20(t-0.5)^2 \mathbb{1}\{t \leq 0.5\} + (t-0.5)^2 Q_e(u),$$

where $Q_e(u)$ is the quantile function of a standard Normal distribution.

### 4.4.2 Simulation Results

Table 4.1: The averaged squared $L_2$-norm of the difference between the estimator $\hat{\beta}(t,u)$ and the true function $\beta(t,u)$, $\|\hat{\beta}(t,u)-\beta(t,u)\|_2^2$ using the proposed method with locally sparse estimation (LS-SFQR), the proposed method without locally sparse estimation (SFQR) and the conventional estimation method (FQR).

| | $n=300$ | | | $n=500$ | | |
|---|---|---|---|---|---|---|
| | $\|\hat{\beta}(t,u)-\beta(t,u)\|_2^2\times 10^2$ | | | $\|\hat{\beta}(t,u)-\beta(t,u)\|_2^2\times 10^2$ | | |
| Scenario | LS-SFQR | SFQR | FQR | LS-SFQR | SFQR | FQR |
| I | 0.59 | 0.65 | 1.62 | 0.44 | 0.36 | 0.98 |
| II | 3.31 | 4.95 | 6.95 | 2.51 | 3.38 | 4.41 |
| III | 2.40 | 3.75 | 4.16 | 1.81 | 2.17 | 2.76 |

We use the averaged squared $L_2$-norm of the difference between the estimator $\hat{\beta}(t,u)$ and the true function $\beta(t,u)$, to evaluate the estimation error,

$$\|\hat{\beta}(t,u)-\beta(t)\|_2^2 \approx \frac{1}{n_U n_T}\sum_{r=1}^{n_U}\sum_{j=1}^{n_T}\left\{\hat{\beta}(t_j,u_r)-\beta((t_j,u_r))\right\}^2,$$

where $t_j$ are the time points when the functional predictor is discretely observed. The results are displayed in Table 4.1 from three different methods. LS-SFQR is the proposed method that simultaneously fits the model for multiple quantiles while employing locally sparse estimation for $\beta(t,u)$. The objective function for LS-SFQR is given by (4.8), where locally sparse penalties are applied to achieve a locally sparse estimator for $\beta(t,u)$. On the other hand, SFQR is another proposed method that simultaneously fits the model for multiple quantiles but without using locally sparse estimation for $\beta(t,u)$. Instead, it minimizes the objective function (4.7) without the locally sparse penalties. FQR is the method proposed in [6], which considers only one quantile for each fitting. It estimates $\beta(t,u)$ for different quantiles separately. This method does not consider the smoothness of $\beta(t,u)$.

Table 4.1 reveals two important insights. Firstly, it demonstrates the advantage of using multiple quantiles simultaneously for model fitting when comparing the estimation errors of the two proposed methods with the conventional method (FQR). Both LS-SFQR and SFQR outperform the conventional method in terms of estimation accuracy. This highlights the benefit of considering the quantiles jointly rather than fitting the model with each quantile separately.

Secondly, when comparing the estimation errors between the methods "LS-SFQR" and "SFQR", it becomes evident that the estimation efficiency can be further improved by employing the locally sparse approach. By identifying and screening out irrelevant sub-regions from the domain of the functional predictor before fitting the model, LS-SFQR achieves a more efficient estimation of the slope function $\beta(t,u)$. This locally sparse estimation leads to a more accurate representation of the dynamic dependence between the quantiles of the response variable $Y$ and the functional predictor $X(t)$.

In summary, fitting the model (4.6) for multiple quantiles the simultaneously provides substantial benefits over the conventional single-quantile approach. Moreover, incorporating the locally sparse estimation technique in the proposed LS-SFQR method further enhances the estimation efficiency and accuracy. These findings reinforce the significance of considering multiple quantiles jointly and leveraging the sparsity structure of the slope function $\beta(t, u)$ in functional quantile regression analysis.

## 4.5. Real Data Analysis

Indeed, temperature plays a crucial role in soybean growth during all stages, including the vegetative and reproductive phases. Both high and low temperatures can have adverse effects on soybean germination and growth. Therefore, understanding the relationship between daily temperature and soybean yield at different quantiles is vital for gaining a comprehensive understanding of soybean growth dynamics.

In this study, we aim to investigate the influence of daily temperatures on soybean yield across various quantiles while controlling for other environmental factors. By accounting for factors such as annual precipitation and the ratio of irrigated area, we can better understand the independent contribution of temperature to soybean yield.

Quantile regression allows us to analyze the relationship between variables at different points of the distribution, going beyond traditional mean-based regression analysis. By examining multiple quantiles, we can identify how temperature affects soybean yield under various conditions, including both favorable and unfavorable scenarios.

The data collected from the United States Department of Agriculture (USDA) website and the National Oceanic and Atmospheric Administration (NOAA) website provides valuable information on soybean yield and related environmental variables in Kansas between 1991 and 2006. The data covers an important agricultural period and geographical region, as Kansas is one of the leading states in soybean production.

In the subsequent analysis, we consider three predictors for investigating the annual soybean yield of each county in Kansas. The functional predictor, denoted as $X(t)$, represents the average of daily minimum and daily maximum temperatures. We specifically consider the temperature data for February and November, allowing us to investigate the influence of temperature during both the colder and warmer months on soybean yield. Figure 4.1 displays a sample of the daily average temperature measurements of Kansas. The two scalar predictors, annual precipitation and the ratio of irrigated area to harvested land (Z2), denoted as $Z_1$ and $Z_2$ provide additional environmental information that can impact soybean growth. Finally, the response variable, denoted as $Y$, represents the annual soybean yield per bushel for each county in Kansas. Then the model we want to investigate is as follows:

$$Q_Y(u|X(t), Z_1, Z_2) = \alpha_0(u) + \alpha_1(u)Z_1 + \alpha_2(u)Z_2 + \int_{\mathcal{T}} \beta(t, u)X(t)dt. \tag{4.10}$$

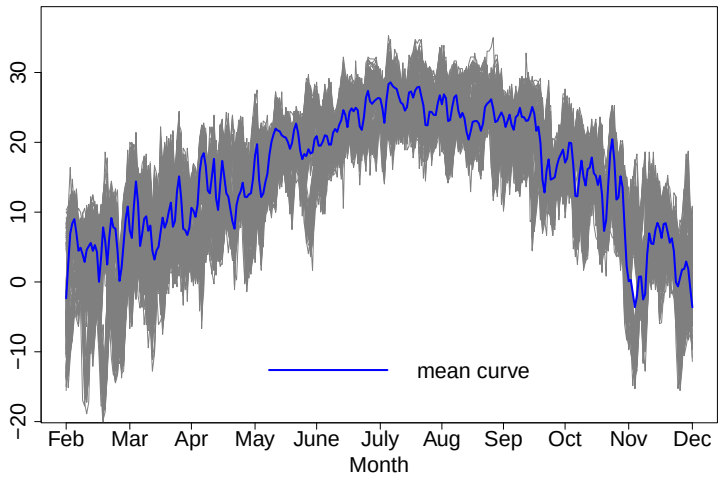



Figure 4.1: A sample of daily average temperature curves of counties in Kansas. The unit of the y-axis is the Fahrenheit temperature scale.

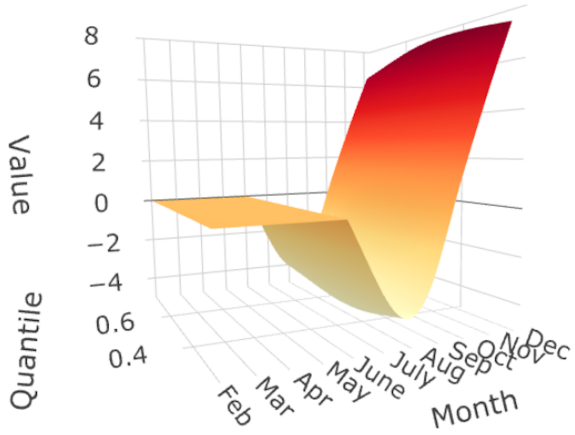



Figure 4.2: Estimated slope function $\beta(t, u)$ of (4.10) using the Kansas soybean yield data set.

We fit the model with 17 quantiles uniformly distributed between 25% and 0.75%. Based on the estimated slope function $\hat{\beta}(t, u)$ displayed in Figure 4.2, we can observe several trends. First, the daily average temperature in Kansas during before June is found to be comfortable for soybean germination and growth since the temperature before June has no influence on the soybean yield for all the quantiles of interest. Second, between June and September, $\hat{\beta}(t, u)$ is consistently negative for all quantiles between 25% and 75%. This indicates that the temperature of the late summer in Kansas is relatively high, negatively affecting soybean growth and resulting in a decrease in soybean yield. The last but not the least, after September, $\hat{\beta}(t, u)$ becomes positive for all the quantiles. This suggests that the winter temperature in Kansas is too low for soybean growth, also having negative impact on soybean yield for the entire year. Figure 4.3 provides another illustration of $\hat{\beta}(t, u)$ for

three specific quantiles: 25%, 50%, and 75% and further emphasizes the changing patterns of the temperature's influence on soybean yield across different quantiles.

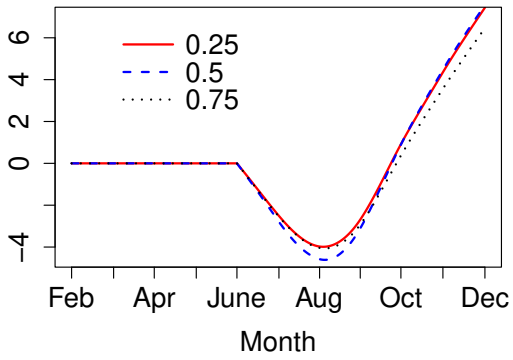



Figure 4.3: Estimated slope function $\beta(t, u)$ of (4.10) for three quantiles: 25%, 50% and 75%.

Overall, the analysis of $\hat{\beta}(t, u)$ provides valuable insights into the dynamic relationship between daily average temperature and soybean yield in Kansas. The findings highlight the critical periods during which temperature can significantly affect soybean growth and yield, and can aid in developing strategies to optimize soybean production in the region.

## 4.6. Conclusion and Discussion

In this project, we introduce a novel approach for estimating the bivariate slope function $\beta(t, u)$ in the functional quantile regression model . The goal is to investigate and understand the dynamic relationship between functional predictors and scalar response variables. Our proposed method employs a locally sparse estimation technique, which allows us to identify and focus on specific regions of interest in the domain of the functional predictor where the influence on the response variable is significant.

For future research, we aim to extend this work to a more general model that can handle multiple functional predictors, which would enhance the applicability of the method to a wider range of real-world applications. However, this extension poses a challenge due to the large number of tuning parameters involved. In practice, tuning these parameters effectively can be difficult. Addressing this challenge would require developing efficient parameter tuning strategies for functional quantile regression.

## 4.7. Proofs in Section 3

**Proof.** Define

$$\boldsymbol{W}_{i,u} = [\boldsymbol{z}_i^\intercal \boldsymbol{D}(u), \boldsymbol{A}_i(u)^\intercal \boldsymbol{Q}] \,.$$

Then

$$\begin{aligned}&\frac{1}{nn_U}\sum_{r=1}^{n_U}\sum_{i=1}^{n}\rho_{u_r}\left(y_i-\boldsymbol{z}_i^\tau\boldsymbol{D}(u_r)\boldsymbol{\eta}-\boldsymbol{A}_i(u_r)^\tau\boldsymbol{Q}\boldsymbol{\theta}\right)+\lambda\boldsymbol{\theta}^\tau\boldsymbol{G}\boldsymbol{\theta}\\=&\frac{1}{nn_U}\sum_{r=1}^{n_U}\sum_{i=1}^{n}\rho_{u_r}\left(y_i-\boldsymbol{W}_{i,u_r}^\tau\boldsymbol{\xi}\right)+\lambda\boldsymbol{\theta}^\tau\boldsymbol{G}\boldsymbol{\theta},\end{aligned}$$

where $\boldsymbol{\xi}=(\boldsymbol{\eta}^\tau,\boldsymbol{\theta}^\tau)^\tau$. We further define

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{A}_1(u_1)\\\cdots,\\\boldsymbol{A}_n(u_1)\\\cdots,\\\boldsymbol{A}_n(u_{n_U})\end{pmatrix}$$

and

$$\boldsymbol{C}_\lambda=\frac{1}{n_Un}\boldsymbol{A}\boldsymbol{A}^\tau+\lambda\boldsymbol{\theta}^\tau\boldsymbol{G}\boldsymbol{\theta}.$$

**Lemma 1.** *When $\lambda=O(|\Delta|^4)$, there exists a positive constants $c_1$ such that*

$$\sigma_{min}\left\{\boldsymbol{C}_\lambda\right\}\geq c_1\lambda|\Delta|^2,$$

*except on an event whose probability tends to $0$ as $n\to\infty$.*

**Proof.** Suppose the dimension of $\boldsymbol{G}$ is $n_\theta\times n_\theta$. Let $K(\boldsymbol{G})=\{\boldsymbol{\theta}\in\mathbb{R}^{n_\theta}:\boldsymbol{G}\boldsymbol{\theta}=\boldsymbol{0}\}$, and let $\boldsymbol{v}\in\mathbb{R}^{n_\theta}$ with $\|\boldsymbol{v}\|=1$. We can decompose $\boldsymbol{v}$ into

$$\boldsymbol{v}=\boldsymbol{v}_1+\boldsymbol{v}_2,$$

where $\boldsymbol{v}_1\in K(\boldsymbol{G})$ and $\boldsymbol{v}_2\in K(\boldsymbol{G})^\perp$. Then

$$\begin{aligned}\frac{1}{nn_U}\boldsymbol{v}^\tau\boldsymbol{Q}^\tau\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\tau\boldsymbol{Q}\boldsymbol{v}+\lambda\boldsymbol{v}\boldsymbol{G}\boldsymbol{v}&=\frac{1}{n_U}\sum_{u\in U}E\left(\sum_j\left[\boldsymbol{v}^\tau\boldsymbol{Q}\right]_j\int B_j^\tau(t,u)X(t)dt\right)^2+\lambda\boldsymbol{v}^\tau\boldsymbol{G}\boldsymbol{v}\\&=\frac{1}{n_U}\sum_{u\in U}\langle\Gamma_X\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{v},\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{v}\rangle+\lambda\boldsymbol{v}^\tau\boldsymbol{G}\boldsymbol{v}\\&\geq\frac{1}{n_U}\sum_{u\in U}\langle\Gamma_X\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{v}_1,\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{v}_1\rangle+\lambda\boldsymbol{v}_2^\tau\boldsymbol{G}\boldsymbol{v}_2,\\&=\frac{1}{n_U}\sum_{u\in U}h_{\boldsymbol{v}_1}(u)+\lambda\boldsymbol{v}_2^\tau\boldsymbol{G}\boldsymbol{v}_2,\\&=\int_u h_{\boldsymbol{v}_1}(u)du+O(n_U^{-2}\|\boldsymbol{v}_1\|^2)+\lambda\boldsymbol{v}_2^\tau\boldsymbol{G}\boldsymbol{v}_2,\end{aligned}$$

where the inner product $\langle \cdot, \cdot \rangle$ is defined as $\langle f_1(t,u), f_2(t,u) \rangle = \int_t f_1(t,u) f_2(t,u) dt$, and $h_{\boldsymbol{v}_1}(u) = \langle \Gamma_X \boldsymbol{B}(t,u)^\tau \boldsymbol{Q} \boldsymbol{v}_1, \boldsymbol{B}(t,u)^\tau \boldsymbol{Q} \boldsymbol{v}_1 \rangle$.

Let $f(t,u) = \boldsymbol{B}(t,u)^\tau \boldsymbol{Q} \boldsymbol{v}_1$. Since $\boldsymbol{G}\boldsymbol{v}_2 = 0$, then $\nabla_t^i \nabla_u^j f(t,u) = 0$ for any non-negative integers $i$ and $j$ with $i + j \leq m$, which implies when $u$ is fixed, as a function of $t$, $f(t,u)$ is a polynomial with degree less than $m$. We assume that all eigenvalues of $\Gamma_X$ are positive, therefore for any $g \in \mathcal{P}_m$, there exists a constant $C_3$ such that

$$\langle \Gamma_X g, g \rangle \geq C_3 \|g\|^2.$$

Then

$$\int_u h_{\boldsymbol{v}_1}(u) du \geq C_3 \int_u \int_t f(t,u)^2 dt du \geq C_4 |\Delta|^2 \|\boldsymbol{v}_1\|^2. \tag{4.11}$$

On the other hand, for any $\boldsymbol{v}_2 \in K(\boldsymbol{G})^\perp$ there exists a constant $C_5$ such that

$$\lambda \boldsymbol{v}_2^\tau \boldsymbol{G} \boldsymbol{v}_2 \geq C_5 \lambda |\Delta|^2 \|\boldsymbol{v}_2\|^2. \tag{4.12}$$

Combine (4.11) and (4.12), we conclude that

$$\frac{1}{n n_U} \boldsymbol{v}^\tau \boldsymbol{Q}^\tau \tilde{\boldsymbol{A}} \tilde{\boldsymbol{A}}^\tau \boldsymbol{Q} \boldsymbol{v} + \lambda \boldsymbol{v} \boldsymbol{G} \boldsymbol{v} \geq c_1 \lambda |\Delta|^2 |\boldsymbol{v}\|^2.$$

Since

$$\begin{aligned}
&\frac{1}{n n_U} \left\{ \boldsymbol{v}^\tau \left( \boldsymbol{Q}^\tau \tilde{\boldsymbol{A}} \tilde{\boldsymbol{A}}^\tau \boldsymbol{Q} \right) \boldsymbol{v} - \boldsymbol{v}^\tau \left( \boldsymbol{Q}^\tau \boldsymbol{A} \boldsymbol{A}^\tau \boldsymbol{Q} \right) \boldsymbol{v} \right\} \\
=& \boldsymbol{v}^\tau \left\{ \frac{1}{n_U} \sum_{u \in U} \langle \Gamma_X \boldsymbol{B}(t,u)^\tau \boldsymbol{Q}, \boldsymbol{B}(t,u)^\tau \boldsymbol{Q} \rangle - \langle \Gamma_n \boldsymbol{B}(t,u)^\tau \boldsymbol{Q}, \boldsymbol{B}(t,u)^\tau \boldsymbol{Q} \rangle \right\} \boldsymbol{v}, \\
\leq& \frac{1}{n_U} \sum_{u \in U} \|\Gamma_X - \Gamma_n\| \sup_j \sum_k |\langle B_k(t,u), B_j(t,u) \rangle| \, \|\boldsymbol{v}\|^2
\end{aligned}$$

where $\Gamma_n f(t,u) = \int_t \widehat{cov}(s,t) f(s,u) ds$ and $\widehat{cov}(s,t) = n^{-1} \sum_{i=1}^n X_i(s) X_i(t)$, and $\binom{d+2}{2}$ Bernstein polynomial basis functions are defined over each triangle, then

$$\sup_j \sum_k |\langle B_k(t,u), B_j(t,u) \rangle| = O(|\Delta|).$$

According to [6], with some $\delta \in (0,1)$, we have

$$\|\Gamma_X - \Gamma_n\| = o_p(n^{(\delta-1)/2}),$$

which implies

$$\frac{1}{n n_U} \left\{ \boldsymbol{v}^\tau \left( \boldsymbol{Q}^\tau \tilde{\boldsymbol{A}} \tilde{\boldsymbol{A}}^\tau \boldsymbol{Q} \right) \boldsymbol{v} - \boldsymbol{v}^\tau \left( \boldsymbol{Q}^\tau \boldsymbol{A} \boldsymbol{A}^\tau \boldsymbol{Q} \right) \boldsymbol{v} \right\} = o_p(n^{(\delta-1)/2} |\Delta| \|\boldsymbol{v}\|^2). \tag{4.13}$$

Then we can conclude that $\left(\frac{1}{nn_U}\boldsymbol{v}^\tau\boldsymbol{Q}^\tau\boldsymbol{A}\boldsymbol{A}^\tau\boldsymbol{Q}\boldsymbol{v}+\lambda\boldsymbol{v}\boldsymbol{G}\boldsymbol{v}\right)$ is non-singular except on an event whose probability goes to 0 as $n\to\infty$.

**Lemma 2.** *For a given set $[a,b]\subset(0,1)$, there exist constants $c_1$ and $c_2$ such that*

$$\sup_{u\in[a,b]}\max_{i=1\ldots n}|\langle\boldsymbol{B}(t,u)^\tau\left(\boldsymbol{Q}^\tau\boldsymbol{A}_X(u)\boldsymbol{A}_X(u)^\tau\boldsymbol{Q}+\lambda\boldsymbol{G}\right)^{-1/2}\boldsymbol{\theta},X_i\rangle|\leq\lambda^{-1/2}\|\boldsymbol{\theta}\|.$$

**Proof.** For any $u\in[a,b]$, we have

$$\begin{aligned}&|\langle\boldsymbol{B}(t,u)^\tau\left\{(nn_U)^{-1}\boldsymbol{Q}^\tau\boldsymbol{A}\boldsymbol{A}^\tau\boldsymbol{Q}+\lambda\boldsymbol{G}\right\}^{-1/2}\boldsymbol{\theta},X_i\rangle|^2\\ \leq&\langle\boldsymbol{B}(t,u)^\tau,X_i\rangle\left\{(nn_U)^{-1}\boldsymbol{Q}^\tau\boldsymbol{A}\boldsymbol{A}^\tau\boldsymbol{Q}+\lambda\boldsymbol{G}\right\}^{-1}\langle\boldsymbol{B}(t,u),X_i\rangle\|\boldsymbol{\theta}\|^2\\ \leq&\lambda^{-1}\|\boldsymbol{\theta}\|^2,\end{aligned}$$

which achieves the proof of Lemma 2.

Under the assumptions of Theorem 1 and according to [15, 46], there exists $\boldsymbol{\eta}^*=(\boldsymbol{\eta}_0^*,\ldots,\boldsymbol{\eta}_p^*)$ and $\boldsymbol{\theta}^*$ such that

$$\sup_u|\alpha_k(u)-\alpha_k^*(u)|=O(n_b^{-\nu}),$$

and

$$\sup_{t,u}|\beta(t,u)-\beta^*(t,u)|=O(|\Delta|^{d+1}),$$

where $d$ is the degree of Bernstein polynomials, $\alpha_k^*(u)=\boldsymbol{b}(u)^\tau\boldsymbol{\eta}_k^*$, and $\beta^*(t,u)=\boldsymbol{B}(t,u)^\tau\boldsymbol{Q}\boldsymbol{\theta}^*$. Let $\Sigma_Z$ be the covariance matrix of $\boldsymbol{Z}_i$ and

$$\boldsymbol{C}_\eta=\frac{1}{n_U}\sum_{u\in U}\boldsymbol{D}(u)^\tau\Sigma_Z\boldsymbol{D}(u).$$

We further define

$$\begin{aligned}f_{i,u}(\boldsymbol{d})=&\rho_u\left(Y_i-\boldsymbol{Z}_i^\tau\boldsymbol{D}(u)\left(\boldsymbol{C}_\eta^{-1/2}\boldsymbol{d}_\alpha+\boldsymbol{\eta}^*\right)-\boldsymbol{A}_i(u)^\tau\left(\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta+\boldsymbol{\theta}^*\right)\right),\\ &+\lambda\left(\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta+\boldsymbol{\theta}^*\right)^\tau\boldsymbol{G}\left(\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta+\boldsymbol{\theta}^*\right)\\ =&\rho_u\left(\epsilon_{u,i}-\boldsymbol{Z}_i^\tau\boldsymbol{D}(u)\boldsymbol{C}_\eta^{-1/2}\boldsymbol{d}_\alpha-\boldsymbol{A}_i(u)^\tau\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta-R_{u,i}\right)\\ &+\lambda\left(\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta+\boldsymbol{\theta}^*\right)^\tau\boldsymbol{G}\left(\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta+\boldsymbol{\theta}^*\right),\end{aligned}$$

where $\boldsymbol{d}=(\boldsymbol{d}_\alpha^\tau,\boldsymbol{d}_\theta^\tau)^\tau$, $\epsilon_{u,i}=Y_i-\boldsymbol{Z}_i^\tau\boldsymbol{\alpha}(u)-\int\beta(t,u)X_i(t)dt$, and $R_{u,i}=\boldsymbol{Z}_i^\tau\left(\boldsymbol{\alpha}^*(u)-\boldsymbol{\alpha}(u)\right)+\langle\beta^*(t,u)-\beta(t,u),X_i\rangle$.

**Lemma 3.** *Let $T_n$ denote the set of random variables $\{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n,X_1(t),\ldots,X_n(t)\}$. For any $\epsilon>0$, there exists $L$ such that*

$$\lim_{n\to\infty}P\left\{\sup_{\|\boldsymbol{d}\|_2=1}\sum_u\sum_i f_{i,u}(L\delta_n\boldsymbol{d})-f_{i,u}(\boldsymbol{0})-\mathbb{E}\left[f_{i,u}(L\delta_n\boldsymbol{d})-f_{i,u}(\boldsymbol{0})\right]>\epsilon\delta_n^2nn_U|T_n\right\}=0,$$

*where* $\delta_n = \sqrt{1/(n\lambda|\Delta|^2) + \lambda}$

**Proof.**

$$
\begin{aligned}
&\sup_{\|\boldsymbol{d}\|_2 \le 1} \sum_u \sum_i f_{i,u}(L\delta_n \boldsymbol{d}) - f_{i,u}(\mathbf{0}) - \mathbb{E}\left[f_{i,u}(L\delta_n \boldsymbol{d}) - f_{i,u}(\mathbf{0}) | T_n\right] \\
=&\frac{1}{2} \sup_{\|\boldsymbol{d}\|_2 = 1} \sum_u \sum_i \left| \epsilon_{u,i} - L\delta_n \left( \boldsymbol{Z}_i^\tau \boldsymbol{D}(u) \boldsymbol{C}_\eta^{-1/2} \boldsymbol{d}_\alpha + \boldsymbol{A}_i(u)^\tau \boldsymbol{C}_\lambda^{-1/2} \boldsymbol{d}_\theta \right) - R_{u,i} \right| \\
&- |\epsilon_{u,i} - R_{u,i}| \\
&- \mathbb{E}\Bigg( \left| \epsilon_{u,i} - L\delta_n \left( \boldsymbol{Z}_i^\tau \boldsymbol{D}(u) \boldsymbol{C}_\eta^{-1/2} \boldsymbol{d}_\alpha - \boldsymbol{A}_i(u)^\tau \boldsymbol{C}_\lambda^{-1/2} \boldsymbol{d}_\theta \right) - R_{u,i} \right| \\
&- |\epsilon_{u,i} - R_{u,i}| \, | T_n \Bigg) \\
:=&\frac{1}{2} \sup_{\|\boldsymbol{d}\|_2 \le 1} \sum_u \sum_i J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d}) | T_n).
\end{aligned}
$$

The set $D = \{\boldsymbol{d} \in \mathbb{R}^{(p+1)n_b + n_\theta} : \|\boldsymbol{d}\| \le 1\}$ is compact and therefore, it can be covered by finite number of open balls. More specifically, for some $r > 0$, $D = \cup_{k=1}^{K_n} D_k$ with $\text{diam}(D_k) = r$, and $K_n = r^{-(p+1)n_b - n_\theta}$. For $1 \le k \le K_n$, we use $\boldsymbol{\omega}_k = (\boldsymbol{\omega}_{k,\alpha}, \boldsymbol{\omega}_{k,\theta})$ to denote the element belongs to $D_k$, where $\boldsymbol{\omega}_{k,\eta}$ denotes the entries of $\boldsymbol{\omega}_k$ corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\omega}_{k,\theta}$ denote the entries of $\boldsymbol{\omega}_k$ corresponding to $\boldsymbol{\theta}$. The minimum eigenvalue of $\boldsymbol{C}_\eta$ is bounded by constant when the covariace matrix $\Sigma_Z$ is not singular, and a lower bound for the minimum eigenvalue of $\boldsymbol{C}_\lambda$ is given by Lemma 2. Then we have

$$
\begin{aligned}
&\min_{k=1,\ldots,K_n} \sum_u \sum_i \left| [J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d}) | T_n)] - [J_{u,i}(\boldsymbol{\omega}_k) - \mathbb{E}(J_{u,i}(\boldsymbol{\omega}_k) | T_n)] \right| \\
&\le 2L\delta_n \min_{k=1,\ldots,K_n} \sum_u \sum_i \left| \boldsymbol{Z}_i^\tau \boldsymbol{D}(u) \boldsymbol{C}_\eta^{-1/2} \left\{ \boldsymbol{d}_\alpha - \boldsymbol{\omega}_{k,\alpha} \right\} \right. \\
&\quad \left. + \boldsymbol{A}_i(u)^\tau \boldsymbol{C}_\lambda^{-1/2} \left\{ \boldsymbol{d}_\theta - \boldsymbol{\omega}_{k,\theta} \right\} \right| \\
&\le C_6 n n_U L\delta_n \lambda^{-1/2} \min_{k=1,\ldots,K_n} \|\boldsymbol{d} - \boldsymbol{w}_k\| \\
&\le C_6 n n_U L\delta_n \lambda^{-1/2} r.
\end{aligned}
$$

On the other hand, by Lemma 2, we get

$$
\sup_{\boldsymbol{d} \in D} |J_{u,i}(\boldsymbol{d})| \le C_6 L\delta_n \lambda^{-1/2}.
$$

In addition,

$$\begin{aligned}\sum_u \sum_i \text{Var}\left(J_{u,i}(\boldsymbol{d})|\boldsymbol{T}_n\right) &\leq \sum_u \sum_i L\delta_n^2 \left(\boldsymbol{Z}_i^\tau \boldsymbol{D}(u)\boldsymbol{C}_\eta^{-1/2}\boldsymbol{d}_\alpha - \boldsymbol{A}_i(u)^\tau \boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta\right)^2 \\ &= nn_U L^2\delta_n^2\|\boldsymbol{d}\|^2 - nn_U L^2\delta_n^2\lambda \boldsymbol{d}_\theta^\tau \boldsymbol{C}_\lambda^{-1/2}\boldsymbol{G}\boldsymbol{C}_\lambda^{-1/2}\boldsymbol{d}_\theta \\ &\leq nn_U L^2\delta_n^2.\end{aligned}$$

Then,

$$\begin{aligned}&P\left[\sup_{\|\boldsymbol{d}\|_2\leq 1}\sum_u\sum_i J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d})|T_n) \geq t_\epsilon\right] \\ &= P\left[\exists\ k_0\ \ s.t.\ \boldsymbol{d}\in D_{k_0}\ \text{ and }\ \sum_u\sum_i J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d})|T_n) \geq t_\epsilon\right] \\ &\leq P\left[\max_{k=1,\ldots,K_n}\sum_u\sum_i J_{u,i}(\boldsymbol{\omega}_k) - \mathbb{E}(J_{u,i}(\boldsymbol{\omega}_k)|T_n) \geq t_\epsilon - 2C_6 nn_U L\delta_n\lambda^{-1/2}r\right] \\ &\leq K_n P\left[\sum_u\sum_i J_{u,i}(\boldsymbol{\omega}_k) - \mathbb{E}(J_{u,i}(\boldsymbol{\omega}_k)|T_n) \geq t_\epsilon - 2C_6 nn_U L\delta_n\lambda^{-1/2}r\right].\end{aligned}$$

By Bernstein inequality,

$$P\left[\sum_u\sum_i J_{u,i}(\boldsymbol{w}_k) - \mathbb{E}(J_{u,i}(\boldsymbol{w}_k)|T_n) \geq t'_\epsilon\right] \leq \exp\left\{-\frac{{t'_\epsilon}^2/2}{nn_U L^2\delta_n^2 + C_6 L\delta_n\lambda^{-1/2}t'_\epsilon/3}\right\},$$

where $t'_\epsilon = t_\epsilon - 2C_6 nn_U L\delta_n\lambda^{-1/2}r$.

Recall that $K_n = r^{-(p+1)n_b - n_\theta}$, and $n_\theta \sim |\Delta|^{-2}$. Then under the condition $n_b = O(n_\theta)$, we have

$$\begin{aligned}&P\left[\sup_{\|\boldsymbol{d}\|_2\leq 1}\sum_u\sum_i J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d})|T_n) \geq t_\epsilon\right] \\ &\leq r^{-(p+1)n_b - n_\theta}\exp\left\{-\frac{{t'_\epsilon}^2/2}{nn_U L^2\delta_n^2 + C_6 L\delta_n\lambda^{-1/2}t'_\epsilon/3}\right\} \\ &= \exp\left\{C_7|\Delta|^{-2}\log\left(r^{-1}\right) - \frac{{t'_\epsilon}^2/2}{nn_U L^2\delta_n^2 + C_6 L\delta_n\lambda^{-1/2}t'_\epsilon/3}\right\}.\end{aligned}$$

Let $r = C_6^{-1}n^{-1/2}n_U^{-1/2}|\Delta|^{-3}$, $L = \frac{\epsilon\delta_n^2}{4C_6\delta_n\lambda^{-1/2}r}$, and $t_\epsilon = \epsilon nn_U\delta_n^2$. Then as $n\to\infty$,

$$P\left[\sup_{\|\boldsymbol{d}\|_2\leq 1}\sum_u\sum_i J_{u,i}(\boldsymbol{d}) - \mathbb{E}(J_{u,i}(\boldsymbol{d})|T_n) \geq \epsilon nn_U\delta_n^2\right] \to 0$$

**Lemma 4.** *$\forall \epsilon > 0$, there exists $L$ such that when $n$ is large enough,*

$$P\left\{\inf_{\|\boldsymbol{d}\|_2=1}\sum_u\sum_i \mathbb{E}\left[f_{i,u}(L\delta_n\boldsymbol{d}) - f_{i,u}(\mathbf{0})\right] > nn_U\delta_n^2 | T_n\right\} > 1-\epsilon.$$

**Proof.** The proof follows from a small modification of Lemma 5 in [6].

**Lemma 5.** *Suppose*

$$\frac{1}{nn_U}\sum_{u\in U}\sum_{i=1}^n \langle\hat{\beta}(t,u) - \beta(t,u)^*, X_i\rangle^2 + \lambda\|\hat{\beta}(t,u)^{(m)} - \beta(t,u)^{*(m)}\|_2^2 = O_p(a_n),$$

*with $a_n \to 0$ when $n \to \infty$. Under assumptions*

$$\|X\| \le C_0 < +\infty \quad a.s.$$

*and the eigenvalues of $\Gamma_X$ are strictly positive, then*

$$\|\hat{\beta}(t,u) - \beta(t,u)^*\|_{\Gamma_X,2}^2 = O_p(a_n).$$

**Proof.** The proof follows from a small modification of Lemma 2 in [6].

Let $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}})$ be the minimizer of (4.7), and define

$$\hat{\beta}(t,u) = \boldsymbol{B}(t,u)^\tau \boldsymbol{Q}\hat{\boldsymbol{\theta}}$$

Then, under the assumption that $\boldsymbol{Z}_i$ and $X_i(t)$ are independent, we have

$$\begin{aligned}
&P\Big\{\frac{1}{nn_U}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^\tau \boldsymbol{D}(u)^\tau \boldsymbol{Z}\boldsymbol{Z}^\tau \boldsymbol{D}(u)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \\
&\quad + \frac{1}{nn_U}\sum_{u\in U}\sum_{i=1}^n \langle\hat{\beta}(t,u) - \beta^*(t,u), X_i(t)\rangle + \lambda\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^\tau \boldsymbol{G}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \ge L^2\delta_n^2\Big\} \\
&= P\Big\{\frac{1}{n_U}\sum_{u\in U}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^\tau \boldsymbol{D}(u)^\tau \boldsymbol{\Sigma}_Z \boldsymbol{D}(u)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \\
&\quad + \frac{1}{nn_U}\sum_{u\in U}\sum_{i=1}^n \langle\hat{\beta}(t,u) - \beta^*(t,u), X_i(t)\rangle^2 + \lambda\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^\tau \boldsymbol{G}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \ge L^2\delta_n^2\Big\} \\
&= P\left\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^\tau \boldsymbol{C}_\eta (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^\tau \boldsymbol{C}_\lambda \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \ge L^2\delta_n^2\right\} \\
&\ge P\left\{\inf_{\|\boldsymbol{d}\|\ge L\delta_n}\sum_{u\in U}\sum_{i=1}^n f_{i,u}(\boldsymbol{d}) > \sum_{u\in U}\sum_{i=1}^n f_{i,u}(\boldsymbol{C}^{1/2}\hat{\boldsymbol{\xi}} - \boldsymbol{C}^{1/2}\boldsymbol{\xi}^*)\right\} \qquad (4.14)
\end{aligned}$$

where $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_\eta & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_\lambda \end{pmatrix}$, $\hat{\boldsymbol{\xi}} = \left(\hat{\boldsymbol{\eta}}^\tau, \hat{\boldsymbol{\theta}}^\tau\right)^\tau$, and $\boldsymbol{\xi}^* = (\boldsymbol{\eta}^{*\tau}, \boldsymbol{\theta}^{*\tau})^\tau$.

By Lemma 3 and Lemma 4, for any positive $\epsilon$, there exists $L$ such that

$$P\left\{\inf_{\|\boldsymbol{d}\|=1}\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(L\delta_n\boldsymbol{d})-\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(\mathbf{0})>nn_U\delta_n^2\right\}>1-\epsilon. \tag{4.15}$$

Since $f_{i,u}(\cdot)$ is a strictly convex function, then (4.15) implies

$$P\left\{\inf_{\|\boldsymbol{d}\|\geq 1}\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(L\delta_n\boldsymbol{d})-\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(\mathbf{0})>nn_U\delta_n^2\right\}>1-\epsilon. \tag{4.16}$$

On the other hand, by definition of $f_{i,u}(\cdot)$ and $\hat{\boldsymbol{\xi}}$,

$$\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(\boldsymbol{d})$$

is minimized at the point $\boldsymbol{d}=\boldsymbol{C}^{1/2}\hat{\boldsymbol{\xi}}-\boldsymbol{C}^{1/2}\boldsymbol{\xi}^*$, therefore,

$$\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(\mathbf{0})\geq\sum_{u\in U}\sum_{i=1}^{n}f_{i,u}(\boldsymbol{C}^{1/2}\hat{\boldsymbol{\xi}}-\boldsymbol{C}^{1/2}\boldsymbol{\xi}^*). \tag{4.17}$$

By (4.15), (4.16) and (4.17), we can conclude that for any $\epsilon>0$, there exists $L$ such that

$$(4.14)>1-\epsilon.$$

In other words,

$$\frac{1}{nn_U}\sum_{u\in U}\sum_{i=1}^{n}\langle\hat{\beta}(t,u)-\beta^*(t,u),X_i(t)\rangle^2+\lambda\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)^{\tau}\boldsymbol{G}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)=O_p(\delta_n).$$

Then by Lemma 5, we conclude that

$$\|\hat{\beta}(t,u)-\beta(t,u)^*\|_{\Gamma_X,2}^2=O_p(\delta_n).$$

In addition,

$$\sup_{t,u}|\beta(t,u)-\beta^*(t,u)|=O(|\Delta|^{d+1}).$$

Then we achieve the proof of Theorem 1.

# Chapter 5

# A Semi-Parametric Functional Generalized Linear Model with Density Ratio Structure

## 5.1. Introduction

Soybean holds significant global importance due to its vital role in meeting food, feed, and fuel demands. A significant portion, approximately 77%, of global soy production is used for livestock feed to support meat and dairy production. The remaining soy is utilized for various purposes, including fuel, industrial applications, and vegetable oils. A smaller proportion, around 7%, is directly consumed by humans in the form of tofu and soy milk.

The germination and growth of soybeans are heavily reliant on environmental resources, particularly temperature and water availability. As a result, these environmental factors have a substantial impact on soybean yield at the end of the growing season. In order to study the relationship between soybean yield and its environmental determinants, we have collected data on soybean yield along with two related environmental variables, daily temperature and machinery levels on farms in Kansas state over a period spanning from 1991 to 2006.

By analyzing this data set, we aim to gain insights into the factors influencing soybean yield and how the temperature and water resources, play a crucial role in determining the crop's final yield. This study could provide valuable information for farmers, policymakers, and researchers seeking to optimize soybean production and address food security and sustainability challenges.

To understand how these environmental factors influence soybean yield, we propose a novel semi-parametric functional generalized linear model (FGLM) to model the relationship between the county-level soybean yield, considered as a continuous response variable, and the scalar predictor, machinery level, along with the functional predictor, daily temperature, across multiple years. This model allows for a flexible and interpretable representation of the relationship between the response and predictors. By considering the response as a continuous variable, we can effectively capture the variability in soybean yield across different counties and years.

The Functional Generalized Linear Model (FGLM) [69, 38] is an extension of the Generalized Linear Model (GLM) [58] that directly models the relationship between a single scalar response variable, which follows any member of the exponential family of distributions, and a functional predictor. For a random variable $Y$ with density

$$p(y;\eta,\phi) = \exp\left\{\frac{y\eta - b(\eta)}{a(\phi)} + c(y,\phi)\right\},$$

the FGLM models the relationship between a functional predictor $X(t)$ and response $Y$ as

$$h(\mu) = \beta_0 + \int_T X(t)\beta(t)dt, \tag{5.1}$$

where $\mu = \mathbb{E}(Y|\eta,\phi) = b'(\eta)$ and $h(\cdot)$ is a link function. Depending on the distribution of $Y$, different link functions are used in the FGLM. For instance, when $Y$ follows a Gaussian distribution, the identity function is commonly used as the link function. On the other hand, if $Y$ is binary, the logit function is typically used as the link function.

Model (5.1) is indeed a valuable tool in environmental science applications, particularly when dealing with functional observations such as daily temperature and humidity commonly seen in agriculture and forestry studies. However, it is important to note certain challenges and limitations that arise when using this model in real-world scenarios.

One significant concern of the model (5.1) is the need to specify the functions $a(\phi)$, $b(\eta)$, and $c(y,\phi)$ in a way that $\exp\{c(y,\phi)\}$ represents a valid density function of some underlying distribution. When the underlying distribution is misspecified, it can lead to difficulties in accurately estimating the functional coefficients $\beta(t)$ in model (5.1). This could potentially impact the reliability and interpretability of the results obtained from the model.

Furthermore, agricultural data sets often comprise samples gathered over various years or from diverse locations. These samples might be considered to stem from distinct populations due to differences in environmental conditions and management practices. In such scenarios, directly applying model (5.1) to each sample separately may not be the most efficient approach, especially when the number of observations in each sample is limited. This limitation is particularly relevant in the specific context of this project, where the soybean yield data might have a relatively small number of observations in each population. For example, the data collected for the target application includes only a restricted number of environmental predictors, such as temperature and machinery level, while other crucial factors like humidity and sunshine are not accounted for, which suggests that the observations from different years in Kansas should be treated as originating from different populations, given the distinct and potentially significant variations in climate conditions over time. In addition, the data set provides only a relatively small sample size for most years between 1991 and 2006, with only 20 to 40 annual measurements available for most of the years. Such small sample sizes can pose challenges when attempting to fit a model accurately, particularly for semi-parametric models.

To address these issues and effectively utilize the available data, we propose a novel semi-parametric functional generalized linear with unspecified baseline distributions connected through a density ratio structure [1]. Let $\boldsymbol{Z}$ be a vector of scalar predictors and $X(t)$ be a functional predictor. Specifically, to overcome the potential model misspecification limitation of (5.1), we propose employing a semi-parametric functional generalized linear model for each individual year's data,

$$f(y|\boldsymbol{z}, x(t)) = \frac{\exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma} + \langle x, \beta\rangle\right) y\right\} g(y)}{\int \exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma} + \langle x, \beta\rangle\right) y\right\} g(y) dy}, \tag{5.2}$$

where the $g(y)$ is a density function of some unknown baseline distribution $G(y)$, and $f(y|\boldsymbol{z}, x(t))$ denotes the condition density function of $Y$ given $\boldsymbol{Z} = \boldsymbol{z}$ and $X(t) = x(t)$. In model (5.2), the vector $\boldsymbol{\gamma}$, the function $\beta(t)$ and the density $g(y)$ are all unknown parameters, and we use $\langle x, \beta\rangle$ to denote the inner product of $x(t)$ and $\beta(t)$, i.e. $\langle x, \beta\rangle = \int_0^T \beta(t)x(t)dt$. The proposed model (5.2) is robust again model misspecification compared with the conventional FGLM (5.1).

In order to address the remaining limited sample size issue, we further propose incorporating the concept of a density ratio model (DRM) [1] into the framework (5.2). Suppose we have $m + 1$ populations with distribution functions $G_r(y)$ and density functions $g_r(y)$, $y = 0, \ldots, m$. The DRM models the ratios of density functions $g_r(y)$ through the following framework,

$$\frac{g_r(y)}{g_0(y)} = \exp\{\alpha_r + \boldsymbol{\theta}_r^\tau \boldsymbol{q}(y)\}, \tag{5.3}$$

where $\boldsymbol{q}(y)$ is a vector of basis functions corresponding to the distribution family of $g_r(y)$, and $\alpha_r$ are normalizing constants. The flexibility of the DRM is a key advantage, making it a valuable tool for modeling a wide variety of data distributions. For example, selecting $\boldsymbol{q}(y) = (y, y^2)$ allows the DRM to cover Normal distribution families, while $\boldsymbol{q}(y) = (y, \log(y))$ covers Gamma distribution families. In practice, researchers can use a long vector $\boldsymbol{q}(y)$ to cover a wide range of distribution families and employ model selection criterion such as AIC or BIC to decide the best model. A detailed discussions on using the density ratio structure to link multiple populations can be found in [1, 65, 10, 90]. By incorporating the density ratio structure in the model (5.4), we can account for the variations in the underlying baseline distributions of soybean yield across different years. In addition, this approach allows us for simultaneous estimation of parameters associated with different populations using the pooled data, effectively leveraging information from all samples.

By combining the framework (5.2) and (5.3), in this project we focus on the following semi-parametric functional generalized linear model,

$$\begin{aligned} f_k(y|\boldsymbol{z}, x(t)) &= \frac{\exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma}_k + \langle x, \beta_k\rangle\right) y\right\} g_k(y)}{\int \exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma}_k + \langle x, \beta_k\rangle\right) y\right\} g_k(y) dy}, \quad k = 0, \ldots, m \\ \frac{g_r(y)}{g_0(y)} &= \exp\{\alpha_r + \boldsymbol{\theta}_r^\tau \boldsymbol{q}(y)\}, \quad r = 1, \ldots, m \end{aligned} \tag{5.4}$$

where $f_k(y|\boldsymbol{z}, x(t))$ is the conditional density of $Y$ and $g_k(y)$ is the density of some unknown baseline distribution $G_k(y)$ for a specific year $k$. The unknown densities $(g_0(y), \ldots, g_m(y))$ are linked through the DRM structure. The proposed framework (5.4) is advantageous compared with the conventional FGLM (5.1) in two aspects. It uses unspecified baseline distributions, which is robust against model misspecification, and it allows for simultaneous estimation of parameters associated with different populations using the pooled data, effectively leveraging information from all samples.

When dealing with data from only one population, model (5.2) can be regarded as an extension of the proportional likelihood ratio model (PLRM) proposed in prior literature [52],

$$f(y|\boldsymbol{x}) = \frac{\exp\{\boldsymbol{z}^\tau \boldsymbol{\beta} y\}\, g(y)}{\int \exp\{\boldsymbol{z}^\tau \boldsymbol{\beta} y\}\, g(y) dy}. \tag{5.5}$$

Model (5.5) is a commonly used approach to assess nonlinear monotone relationships between a scalar response and finite-dimensional predictors. This model has connections with several other well-known statistical models and methods, making it a versatile and powerful tool for modeling complex relationships in various fields. The connections between the model (5.5) with GLM [58], density ratio models [65], biased sampling models [26], single-index models [37] and exponential tilt regression models [70] are extensively discussed in [52].

In addition to these connections, the model (5.5) offers several advantages over other modeling techniques such as generalized estimating equations, generalized additive models, and fully nonparametric models [52]. It provides a parsimonious and interpretable way to capture nonlinear relationships between the outcome variable and covariates, making it suitable for practical applications, and it has been also extended and adapted to various domains, such as engineering, biomedical research, and survival analysis [53, 33, 34, 61]. These extensions and new models based on the model (5.5) have demonstrated their utility and efficacy in analyzing diverse types of data.

Overall, the model (5.5) stands as a powerful and flexible tool for modeling nonlinear relationships in a wide range of applications, making it an essential method in statistical modeling and data analysis.

## 5.2. Estimation

### 5.2.1 Empirical likelihood based on the proposed model

Suppose the baseline population distribution functions, $G_k(y)$, of $m+1$ samples with sample sizes $n_k$ satisfy

$$\frac{g_k(y)}{g_0(y)} = \exp\{\alpha_k + \boldsymbol{\theta}_k \boldsymbol{q}(y)\} \quad k = 1, \ldots, m, \tag{5.6}$$

where $g_k(y)$ denotes the density function of $G_k(y)$. For notational simplicity, we set $\alpha_0 = 0$ and $\boldsymbol{\theta}_0 = \mathbf{0}$. Assume i.i.d. data $\{\boldsymbol{z}_{k,i}, x_{k,i}(t), y_{k,i}\}$, $k = 0, \ldots, m; i = 1, \ldots, n_k$ from distributions that

satisfy the model,

$$f_k(y|\boldsymbol{z}, x(t)) = \frac{\exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma}_k + \langle x, \beta_k \rangle\right) y\right\} g_k(y)}{\int \exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma}_k + \langle x, \beta_k \rangle\right) y\right\} g_k(y) dy}, \quad k = 0, \ldots, m. \tag{5.7}$$

In this project, we use maximum empirical likelihood (EL) method to fit the model (5.7). Empirical likelihood under DRM and PLRM can be found be in [25, 65, 52]. For convenience, we first introduce some notations. Define

$$A(\boldsymbol{z}, x(t); \boldsymbol{\gamma}, \beta(t), G(y)) = \int \exp\left\{\left(\boldsymbol{z}^\tau \boldsymbol{\gamma} + \langle x, \beta \rangle\right) y\right\} dG(y).$$

The contribution to the likelihood function of one observation from population $k$, $(\boldsymbol{z}_{k,i}, x_{k,i}(t), y_{k,i})$ is given by

$$lik(\boldsymbol{\gamma}_k, \beta_k, G_k) = \exp\left\{\left(\boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}, \beta_k \rangle\right) y_{k,i}\right\} \Delta G_k(y_{k,i}) / A(\boldsymbol{z}_{k,i}, x_{k,i}(t); \boldsymbol{\gamma}_k, \beta_k(t), G_k),$$

where $\Delta G_k(y) = G_k(y+) - G_k(y-)$. Then the log-likelihood for $(\boldsymbol{\gamma}_k, \beta_k, G_k)$ is

$$\sum_{i=1}^{n_k} \left\{\boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}, \beta_k \rangle\right\} y_{k,i} + \sum_{i=1}^{n_k} \log \Delta G_k(y_{k,i}) - \sum_{i=1}^{n_k} \log A(\boldsymbol{z}_{k,i}, x_{k,i}(t); \boldsymbol{\gamma}_k, \beta_k(t), G_k).$$

Under the DRM assumption on the baseline distributions $G_k(y)$, the log-likelihood function of $(\boldsymbol{\gamma}_k, \beta_k, G_k)$ can also be viewed as a function of $(\boldsymbol{\gamma}_k, \beta_k, \alpha_k, \boldsymbol{\theta}_k, G_0)$. Then the log-likelihood function of $(\boldsymbol{\gamma}_k, \beta_k, \alpha_k, \boldsymbol{\theta}_k, G_0)$ can be written as

$$\begin{aligned} l_n(\boldsymbol{\gamma}_k, \beta_k, \alpha_k, \boldsymbol{\theta}_k, G_0) = & \sum_{i=1}^{n_k} \left\{\boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}, \beta_k \rangle\right\} y_{k,i} + \sum_{i=1}^{n_k} \log \Delta G_0(y_{k,i}) \\ & + \sum_{i=1}^{n_k} \alpha_k + \boldsymbol{q}(y)^\tau \boldsymbol{\theta}_k - \sum_{i=1}^{n_k} \log A(\boldsymbol{z}_{k,i}, x_{k,i}(t); \boldsymbol{\gamma}_k, \beta_k, G_k). \end{aligned} \tag{5.8}$$

From the empirical likelihood point of view, the distributions $G_k(y)$ can be treated as discrete distributions. Let $p_{k,i}$ denote the mass that the distribution function $G_0(y)$ assigns at $y_{k,i}$. Then the log-likelihood function (5.8) can be further written as

$$\begin{aligned} l_n(\boldsymbol{\gamma}_k, \beta_k, \alpha_k, \boldsymbol{\theta}_k, G_0) = & \sum_{i=1}^{n_k} \left\{\boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}(t), \beta_k(t) \rangle\right\} y_{k,i} + \sum_{i=1}^{n_k} \log p_{k,i} \\ & + \sum_{i=1}^{n_k} \alpha_k + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_k \\ & - \sum_{i=1}^{n_k} \log \left( \sum_{r=0}^{m} \sum_{j=1}^{n_r} p_{r,j} \exp\left\{\left(\boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}(t), \beta_k(t) \rangle\right) y_{r,j} + \alpha_k + \boldsymbol{q}(y_{r,j})^\tau \boldsymbol{\theta}_k \right\} \right), \end{aligned}$$

under the constraints

$$\begin{aligned}
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} p_{k,i} = 1,\\
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} \{\alpha_r + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_r\} p_{k,i} = 1, \quad r = 1, \ldots, m,\\
&p_{k,i} > 0, \quad k = 0, \ldots, m, \quad i = 1, \ldots, n_k.
\end{aligned}$$

Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^\tau, \ldots, \boldsymbol{\gamma}_m^\tau)^\tau$, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_m(t))^\tau$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\tau$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\tau, \ldots, \boldsymbol{\theta}_m^\tau)^\tau$, and $\boldsymbol{p} = \{p_{k,i}\}_{k=0,\ldots,m;i=1,\ldots,n_k}^\tau$. Then the log-likelihood function of $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p})$ is given by

$$\begin{aligned}
l_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p}) = &\sum_{k=0}^{m}\sum_{i=1}^{n_k} \left\{ \left( \boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}(t), \beta_k(t) \rangle \right) y_{k,i} \right\} + \sum_{k=0}^{m}\sum_{i=1}^{n_k} \log p_{k,i} + \sum_{k=0}^{m}\sum_{i=1}^{n_k} \alpha_k + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_k \\
&- \sum_{k=0}^{m}\sum_{i=1}^{n_k} \log \left( \sum_{r=0}^{m}\sum_{j=1}^{n_r} p_{r,j} \exp \left\{ \left( \boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \langle x_{k,i}(t), \beta_k(t) \rangle \right) y_{r,j} + \alpha_k + \boldsymbol{q}(y_{r,j})^\tau \boldsymbol{\theta}_k \right\} \right),
\end{aligned} \tag{5.9}$$

$$\begin{aligned}
&s.t.\\
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} p_{k,i} = 1,\\
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} \{\alpha_r + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_r\} p_{k,i} = 1, \quad r = 1, \ldots, m,\\
&p_{k,i} > 0, \quad k = 0, \ldots, m, \quad i = 1, \ldots, n_k.
\end{aligned}$$

To estimate the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p})$, a direct maximization of the log-likelihood function (5.9) is challenging due to the infinite-dimensional nature of the unknown functions $\boldsymbol{\beta}$, which belong to an infinite-dimensional function space. To address this issue, a common approach is to find finite-dimensional approximations for the unknown functions $\boldsymbol{\beta}$. This involves approximating the target functions by a set of basis functions with finite-dimensional coefficients. This way, we estimate the coefficients of the approximation rather than the target functions directly, making the optimization more tractable. In this project, we propose to use B-spline basis functions as the approximation for the unknown functions in $\boldsymbol{\beta}(t)$. B-splines provide a flexible and finite dimensional representation for approximating smooth functions. By employing B-spline basis functions, we can effectively reduce the dimensionality of the problem and estimate the finite-dimensional coefficients instead of the original infinite-dimensional functions.

In the next sub-section, we will introduce penalized B-spline estimators for $\beta_k(t)$, which will enable us to estimate the coefficients of the B-spline approximations, and, consequently, the unknown functions $\boldsymbol{\beta}(t)$.

### 5.2.2 Penalized B-spline estimator for $\beta_k(t)$

For non-parametric estimation problems, the function to be estimated is usually assumed to be sufficiently smooth, enabling its expansion in a basis representation. B-spline is a popular and effective choice for such approximation tasks. For model (5.7), we employ B-splines to expand all $m+1$ unknown functions $\beta_0(t), \ldots, \beta_m(t)$:

$$\beta_k(t) \approx \sum_{j=1}^{n_b} c_{k,j} b_j(t) = \boldsymbol{b}(t)^\tau \boldsymbol{c}_k, \quad k = 0, \ldots, m \tag{5.10}$$

where $\boldsymbol{b}(t) = (b_1, \ldots b_{n_b}(t))^\tau$ are B-spline basis functions and $\boldsymbol{c}_k = (c_{k,1}, \ldots, c_{k,n_b})$ are unknown coefficients. With the approximation (5.10), the inner product between $\beta_k(t)$ and $x_{k,i}(t)$ can be approximately expressed as

$$\int \beta_k(t) x_{k,i}(t) dt \approx \sum_{j=1}^{n_b} c_{k,j} \int b_j(t) x_{k,i}(t) dt := \boldsymbol{u}_{k,i}^\tau \boldsymbol{c}_k,$$

where $\boldsymbol{u}_{k,i} = (\int b_1(t) x_{k,i}(t) dt, \ldots, \int b_{n_b}(t) x_{k,i}(t) dt)^\tau$. Then the log-likelihood function defined in (5.9) can be approximately rewritten as:

$$\begin{aligned} l_n(\boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p}) = & \sum_{k=0}^{m} \sum_{i=1}^{n_k} \left\{ \boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \boldsymbol{u}_{k,i}^\tau \boldsymbol{c}_k \right\} y_{k,i} + \sum_{k=0}^{m} \sum_{i=1}^{n_k} \log p_{k,i} + \sum_{k=1}^{m} \sum_{i=1}^{n_k} \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_k \\ & - \sum_{k=0}^{m} \sum_{i=1}^{n_k} \log \left( \sum_{r=0}^{m} \sum_{j=1}^{n_r} p_{r,j} \exp \left\{ \left( \boldsymbol{z}_{k,i}^\tau \boldsymbol{\gamma}_k + \boldsymbol{u}_{k,i}^\tau \boldsymbol{c}_k \right) y_{r,j} + \boldsymbol{q}(y_{r,j})^\tau \boldsymbol{\theta}_k \right\} \right), \end{aligned} \tag{5.11}$$

$$\begin{aligned} & s.t. \\ & \sum_{k=0}^{m} \sum_{i=1}^{n_k} p_{k,i} = 1, \\ & \sum_{k=0}^{m} \sum_{i=1}^{n_k} \{\alpha_r + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_r\} p_{k,i} = 1, \quad r = 1, \ldots, m, \\ & p_{k,i} > 0, \quad k = 0, \ldots, m, \quad i = 1, \ldots, n_k. \end{aligned}$$

In order to obtain a good approximation for an unknown function using B-splines, the number of basis functions, denoted as $n_b$, is often chosen to be large, typically depending on the sample size $n$. However, a large value for $n_b$ can lead to an ill-conditioned design matrix derived from the B-spline approximation, causing numerical stability issues. To address this problem, one effective solution is to introduce penalty terms on the coefficients $\boldsymbol{c}_k$. This penalty term acts as a regularization mechanism, controlling the smoothness of the estimated functions and mitigating the ill-conditioning problem caused by a large number of basis functions.

In the context of Functional Data Analysis (FDA), a widely used choice for the penalty is the roughness penalty [69],

$$\text{Pen}\left\{\boldsymbol{b}(t)^\tau \boldsymbol{c}_k\right\} = \| \left\{\boldsymbol{b}(t)^\tau \boldsymbol{c}_k\right\}^{(2)} \|_2^2 := \boldsymbol{c}_k^\tau \boldsymbol{Q}_k \boldsymbol{c}_k,$$

where $\boldsymbol{Q}_k$ is a symmetric and positive semi-definite matrix. This roughness penalty seeks to penalize the deviation of the estimated function from being too wiggly or oscillatory, encouraging the estimation to favor smoother functions. By incorporating roughness penalties on the coefficient vectors $\boldsymbol{c}_k$, the estimation process achieves more stable and reliable results, improving the numerical stability of the model while ensuring a well-behaved approximation for the unknown functions using B-splines.

Then we proposed to estimate the parameters by minimizing the following penalized minimization criterion with respect to $(\boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p})$,

$$\begin{aligned}
&\min - l_n(\boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p}) + \sum_{k=0}^{m} \lambda_k \boldsymbol{c}_k^\tau \boldsymbol{Q}_k \boldsymbol{c}_k, \qquad (5.12)\\
&s.t.\\
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} p_{k,i} = 1,\\
&\sum_{k=0}^{m}\sum_{i=1}^{n_k} \{\alpha_r + \boldsymbol{q}(y_{k,i})^\tau \boldsymbol{\theta}_r\} p_{k,i} = 1, \quad r = 1, \ldots, m,\\
&p_{k,i} > 0, \quad k = 0, \ldots, m, \quad i = 1, \ldots, n_k,
\end{aligned}$$

where $\lambda_k$ are tuning parameters. By minimizing this penalized criterion, we get a balance between fitting the model well and achieving smooth and stable estimates of the unknown functions.

### 5.2.3 Parameter tuning strategy for $\lambda_k$

The proposed approach involves using cross-validation for parameter tuning in model (5.12). We define the index sets for training and test data from population $k$ as $I_k$ and $I_k^*$, respectively. Define

$$S_k = \{(\boldsymbol{z}_{k,i}, x_{k,i}(t), y_{k,i}) : i \in I_k\}, \quad k = 0, \ldots, m,$$

and

$$S_k^* = \{(\boldsymbol{z}_{k,i}, x_{k,i}(t), y_{k,i}) : i \in I_k^*\}, \quad k = 0, \ldots, m.$$

To maintain the ratio of sample sizes from different populations in the cross validation procedure, we ensure that

$$\frac{|I_k|}{|I_0|} = \frac{n_k}{n_0}, \quad k = 1, \ldots, m,$$

when constructing the training sets $S_k$.

For each validation run, we fit model (5.12) using the training set $\cup_{k=0}^{m} S_k$ to obtain the parameter estimates $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}})$. Then, we compute the negative log-likelihood of the test sets $\cup_{k=0}^{m} S_k^*$ as the validation loss. We use this validation loss to tune the parameters $\lambda_k$, $k = 1, \ldots, m$, using a 10-fold cross-validation procedure. The procedure involves iterating through different combinations of $\lambda_k$, $k = 1, \ldots, m$, and selecting the one that results in the smallest validation loss.

To calculate the log-likelihood of $S_k^*$, we need to compute $p_{k,i} = P_{\hat{G}_k(y)}(Y = y_{k,i})$ for $(k, i) \in I_k^*$ based on $\hat{G}_k(y)$ estimated from $\cup_{k=0}^{m} S_k$. Since for any $y_{k,i} \in S_k^*$, $y_{k,i} \notin \cup_{k=0}^{m} S_k$ then $\hat{p}_{k,i} = P_{\hat{G}_k(y)}(Y = y_{k,i}) = 0$, which is not reasonable. To address the issue of calculating the log-likelihood for $S_k^*$, where $y_{k,i} \in S_k^*$ and $y_{k,i} \notin \cup_{k=0}^{m} S_k$, we propose the following procedure to approximate $p_{k,i}$ for $(k, i) \in I_k^*$.

The main concept is to first derive a continuous density estimation for $g_k(y)$ based on $\hat{G}_k(y)$, the estimated distribution function. This continuous density approximation can be obtained using kernel density estimation or any suitable non-parametric density estimation technique. In this project, we choose to use the kernel density estimation approach. Define a kernel smoothed EL estimator for the density function $g_k(y)$ of $G_k(y)$ as

$$\hat{g}_k^{ksel}(y) = \frac{1}{h} \sum_{r=0}^{m} \sum_{i \in I_r} \hat{p}_{r,i} \exp\left\{\hat{\alpha}_k + \boldsymbol{q}(y_{r,i})^\tau \hat{\boldsymbol{\theta}}_k\right\} K\left(\frac{y - y_{k,i}}{h}\right), \tag{5.13}$$

where $K(\cdot)$ is a kernel function and $h$ is the bandwidth.

Next, we use a weighted average of the probability mass of observations around the target data point $y_{k,i}$ to approximate $p_{k,i}$. This means that for each $y_{k,i} \in S_k^*$, we calculate a weighted average of probabilities assigned to nearby data points in $\cup_{k=0}^{m} S_k$ based on the derived density estimation. Specifically, For any $y_{k,i} \in S_k^*$, we define a neighbourhood for $y_{k,i}$ as

$$N_\eta(y_{k,i}) := \{y_{r,j} : |y_{r,j} - y_{k,i}| \leq \eta, \quad (r, j) \in S_r, \quad r = 0, \ldots, m\}.$$

Then we estimate $p_{k,i}$ for $y_{k,i} \in S_k^*$ by

$$\tilde{p}_{k,i} = \frac{1}{|N_\eta(y_{k,i})|} \sum_{y_{r,j} \in N_\eta(y_{k,i})} \frac{\hat{g}_k^{ksel}(y_{k,i})}{\hat{g}_k^{ksel}(y_{r,j})} \hat{p}_{r,j} \exp\left\{\hat{\alpha}_k + \boldsymbol{q}(y_{r,j}) \hat{\boldsymbol{\theta}}_k\right\}, \quad k = 0, \ldots, m. \tag{5.14}$$

By employing this approach, we can reasonably approximate $p_{k,i}$ for $(k, i) \in I_k^*$ and calculate the log-likelihood of the test set $S_k^*$. This procedure ensures that we consider the information available in the training set while evaluating the likelihood of the testing set, even for data points that are not directly observed in the training set. Overall, this method provides a sound and reliable way to compute the log-likelihood and assess the model's performance during the cross-validation process.

We suggest using a 10-fold cross-validation process to tune the parameters $(\lambda_0, \ldots, \lambda_m)$ for the model. Here are the steps involved:

1. Split the data into 10 subsets for the cross-validation process. Each subset serves as the testing set once while the remaining nine subsets are used as the training set.

2. Given the set of parameters $(\lambda_0, \ldots, \lambda_m)$, estimate the model parameters $(\boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{p})$ using the maximum empirical likelihood method on the training set $\cup_{k=0}^{m} S_k$. The estimators are denoted as $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}})$.

3. Use formulas (5.13) and (5.14) to calculate $\tilde{p}_{k,i}$ for each observation $y_{k,i}$ in the testing set $S_k^*$ based on the estimated distribution function $\hat{G}_k(y)$. Compute the negative log-likelihood of the test set $\cup_{k=0}^{m} S_k^*$ as the validation loss.

4. Repeat the above steps 10 times for each combination of tuning parameters $(\lambda_0, \ldots, \lambda_m)$.

5. Find the combination of $(\lambda_0, \ldots, \lambda_m)$ that results in the smallest validation loss. This combination of parameters is considered optimal and will be used for the final model estimation.

By using the 10-fold cross-validation, we can tune the regularization parameters $(\lambda_0, \ldots, \lambda_m)$ and identify the combination that leads to the best model performance, helping us achieve a well-calibrated model for the data.

## 5.3. Simulation Studies

### 5.3.1 Data Generating Models

**Scenario I.** In the first scenario, the data sets are generated from the following model:

$$Y_{k,i} = \int_{[0,1]} \rho_k(t) X_{k,i}(t) dt + \epsilon_{k,i}, \quad k = 0, 1, \tag{5.15}$$

where $X_{k,i}(t)$ is generated from a Wiener process, $\epsilon_{k,i} \sim N(0, \sigma_k^2)$ with $\sigma_0 = 0.1$ and $\sigma_1 = 0.09$, and

$$\rho_0(t) = 10(t - 0.5)^2, \quad \rho_1(t) = 8(t - 0.4)^2.$$

**Scenario II.** In the second scenario, the data sets are generated from the following model:

$$Y_{k,i} = \nu_{k,1} Z_{1,k,i} + \nu_{k,2} Z_{2,k,i} + \int_{[0,1]} \rho_k(t) X_{k,i}(t) dt + \epsilon_{k,i}, \quad k = 0, 1, 2 \tag{5.16}$$

where $X_{k,i}(t)$ is generated from a Wiener process; for $k = 0, 1, 2$, $Z_{1,k,i} \sim Unif(-0.5, 0.5)$, $Z_{2,k,i} \sim Unif(-0.5, 0.5)$, $\epsilon_{k,i} \sim N(\mu_k, \sigma_k^2)$ with $(\mu_0, \mu_1, \mu_2) = (0, 0, 0.05)$ and $(\sigma_0, \sigma_1, \sigma_2) = (0.1, 0.1, 0.2)$; and

$$\boldsymbol{\nu}_0 = (0.2, 0.4)^\tau, \quad \boldsymbol{\nu}_1 = (0.4, 0.2)^\tau, \quad \boldsymbol{\nu}_2 = (0.2, 0.4)^\tau;$$

$$\rho_k(t) = 10(t - 0.5)^2, \quad k = 0, 1, 2.$$

**Scenario III.** In the third scenario, the data sets are generated from the following model:

$$Y_{k,i}|\boldsymbol{Z}_{k,i}, X_{k,i}(t) \sim TN(\mu_{k,i}, \sigma_k), \quad \mu_{k,i} = \boldsymbol{\nu}_k^\tau \boldsymbol{Z}_{k,i} + \int_0^1 \rho_k(t) X_{k,i}(t) dt \quad k = 0, 1, 2 \quad (5.17)$$

where $TN(\mu_{k,i}, \sigma_k)$ denotes a truncated Normal r.v. over $[0, \infty)$ with $(\sigma_0, \sigma_1, \sigma_2) = (0.1, 0.1, 0.15)$; $X_{k,i}(t)$ is generated from a Wiener process; $Z_{1,k,i} \sim Unif(-0.5, 0.5)$, $Z_{2,k,i} \sim Unif(-0.5, 0.5)$;

$$\boldsymbol{\nu}_0 = (0.2, 0.4)^\tau, \quad \boldsymbol{\nu}_1 = (0.4, 0.2)^\tau, \quad \boldsymbol{\nu}_2 = (0.1, 0.1)^\tau,$$

$$\rho_k(t) = 10(t - 0.5)^2, \quad k = 0, 1, 2.$$

### 5.3.2 Simulation Results

Data generating model (5.15) indicates that the conditional density of $Y_{k,i}$ is also Normal. Then model (5.15) can also be formulated in the format as we proposed in (5.4) with

$$\beta_0(t) = \rho_0(t)/\sigma_0^2, \quad \beta_1(t) = \rho_1(t)/\sigma_1^2.$$

We perform simulations using model (5.15) under two different settings: $(n_0, n_1) = (100, 200)$ and $(n_0, n_1) = (200, 200)$. Each setting is repeated 100 times to ensure reliable results. As the $L_2$-norms of the true functions $\beta_0(t)$ and $\beta_1(t)$ are relatively large, we adopt the relative integrated squared error (RISE) to assess the performance of the proposed penalized B-spline estimators. RISE is defined as follows:

$$\text{RISE} := \frac{\|\hat{\beta}_k(t) - \beta_k(t)\|_{L2}^2}{\|\beta_k(t)\|_{L2}^2}.$$

Under the proposed method, we utilize the true basis function $\boldsymbol{q}(y) = (y, y^2)$ for the DRM framework. In the cross-validation procedure for tuning parameters, we employ the Epanechnikov kernel,

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \leq 1\},$$

for the kernel smoothed EL density estimator, represented as $\hat{f}_k^{ksel}(y)$.

The simulation results for Scenario I are presented in Table (5.1). We compare two methods: EL, which employs the maximum empirical likelihood method to fit model (5.2) using only one sample ($k = 0$ or $k = 1$), and DRM-EL, the proposed method that incorporates data from multiple populations through model (5.4). EL method can be considered as a special case of the proposed method with a single population ($m = 1$), where there is no density ratio structure. On the other hand, DRM-EL leverages the pooled samples from both $k = 0$ and $k = 1$ for model fitting.

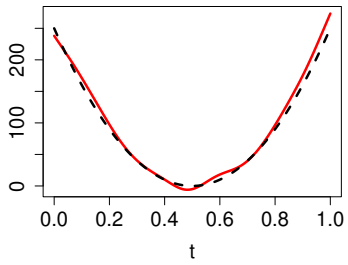
From the estimation errors in Table (5.1), we observe the clear advantage of using combined samples for model fitting in DRM-EL. The efficiency of the estimations improves with a larger sample size, thanks to the density ratio structure employed in model (5.4). By considering multiple populations, the proposed method harnesses more information from the pooled samples, leading to

more accurate results compared to EL method, especially when the sample size is increased. This demonstrates the effectiveness of the DRM approach in enhancing estimation efficiency and making better use of available data.

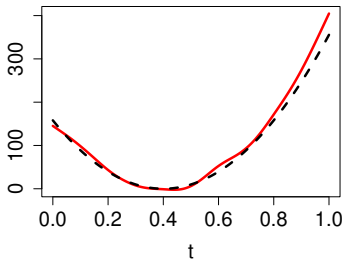
Figure (5.1) presents the averaged estimations for $\beta_0(t)$ and $\beta_1(t)$ over 100 repetitions using EL and DRM-EL respectively under the setting $(n_0, n_1) = (200, 200)$. We can observe that the bias of the estimations obtained from DRM-EL is smaller than that obtained from EL, which provides another illustration for the strength of using the density ratio structure to pool information across samples.

| Methods | $(n_0, n_1)$ | RISE($\hat{\beta}_0$) | RISE($\hat{\beta}_1$) | SE($\hat{\theta}_{1,1}$) | Bias($\hat{\theta}_{1,1}$) | SE($\hat{\theta}_{1,2}$) | Bias($\hat{\theta}_{1,2}$) |
|---|---|---|---|---|---|---|---|
| EL | $(100, 200)$ | 0.263 | 0.129 | | | | |
| DRM-EL | $(100, 200)$ | 0.224 | 0.104 | 1.447 | 0.035 | 11.915 | 2.068 |
| EL | $(200, 200)$ | 0.101 | 0.129 | | | | |
| DRM-EL | $(200, 200)$ | 0.059 | 0.054 | 1.171 | 0.034 | 9.058 | -1.513 |

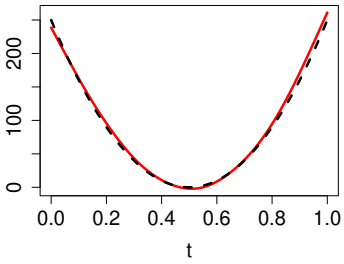
Table 5.1: Simulation results based on Scenario I for the two proposed methods: EL and DRM-EL.



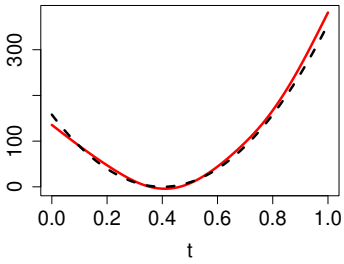

(a) $\hat{\beta}_0(t)$ obtained from EL method




(b) $\hat{\beta}_1(t)$ obtained from EL method




(c) $\hat{\beta}_0(t)$ obtained from DRM-EL method




(d) $\hat{\beta}_1(t)$ obtained from DRM-EL method

Figure 5.1: Averaged $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ (red solid line) obtained from two proposed methods: EL and DRM-EL based on 100 repetitions under the setting $(n_0, n_1) = (200, 200)$; the true $\beta_0(t)$ and $\beta_1(t)$ (black dashed line).

In both Scenario II and Scenario III, we perform simulations with the same setup, having $(n_0, n_1, n_2) = (100, 50, 50)$, and conduct 100 repetitions for each scenario. For the basis function $\boldsymbol{q}(y)$ in the proposed method, we stick to the true basis function $\boldsymbol{q}(y) = (y, y^2)$ in both scenarios. During the cross-validation procedure for tuning parameters, we utilize the Epanechnikov kernel for the kernel smoothed EL density estimator, denoted as $\hat{f}_k^{ksel}(y)$. The data generating model (5.16) and (5.17) both can be formulated in the format of (5.4) with

$$\beta_k(t) = \rho_k(t)/\sigma_k^2, \quad \boldsymbol{\gamma}_k = \boldsymbol{\nu}_k/\sigma_k^2, \quad k = 0, 1, 2,$$

where $\boldsymbol{\nu}_k = (\nu_{k,1}, \nu_{k,2})^\tau$.

The primary objective in both Scenario II and Scenario III is to compare the estimation efficiency of the parameters $\boldsymbol{\theta}_k$ and $\beta_k(t)$ obtained from the proposed method DRM-EL and the Maximum Likelihood Estimation (MLE) based on the Gaussian distribution family. While the Gaussian distribution family is correctly specified in Scenario II, it is misspecified in Scenario III.

To implement the MLE method for FGLM, we employ the R function "fregre.glm" from the package "fda.usc". This function expands the unknown functions $\beta_k(t)$ using B-splines and then fits the model using MLE. The number of B-splines for approximation is chosen based on the Akaike Information Criterion (AIC) criterion, also integrated into the "fregre.glm" function.

The simulation results for both Scenario II and 3 are summarized in Table (5.2) and Table (5.3). In Table (5.2), we observe that regardless of whether the model is misspecified or not, the estimation errors of $\beta_k(t)$ obtained from MLE are significantly larger than those from the proposed method in both scenarios. The reason behind this discrepancy is that the MLE method, as implemented by the R package "fregre.glm", lacks a penalty term to control the smoothness of the estimation. Although the AIC criterion helps select a smaller number of B-spline basis functions, the resulting $\beta_k(t)$ estimations are still undersmoothed.

Under Scenario II, the mean squared errors (MSE) of $\hat{\boldsymbol{\gamma}}_k$ derived from MLE are usually slightly smaller than those from the proposed method, which is expected. But the difference is not large. This could be attributed to the fact that the proposed method, leveraging the DRM structure in (5.4), can use a larger sample to fit the model compared to MLE. However, under Scenario III, where the model is misspecified, MLE performs significantly worse than the proposed method. Scenario III serves as an example, emphasizing the importance of considering semi-parametric methods, as in the proposed approach, to estimate the baseline distribution of FGLM when the model is misspecified.

## 5.4. Real Data Analysis

In the United States, Kansas boasts a substantial soybean cultivation, with approximately 4.7 million acres dedicated to soybean planting. This productive effort yields an impressive output of 200 million bushels, securing Kansas's position as the 10th highest soybean-yielding state in the country.

The scientific research cited from [79, 72] highlights the paramount significance of temperature and water availability throughout all stages of soybean growth, encompassing both the vegetative

| Setting | Methods | RISE($\hat{\beta}_0$) | RISE($\hat{\beta}_1$) | RISE($\hat{\beta}_2$) |
|---|---|---|---|---|
| Scenario II | DRM-EL | 0.12 | 0.13 | 0.16 |
| | MLE | 3.05 | 5.78 | 10.29 |
| Scenario III | DRM-EL | 0.20 | 0.22 | 0.47 |
| | MLE | 4.59 | 14.01 | 16.05 |

Table 5.2: Simulation results on the estimation of $\beta_k(t)$ under Scenario II and 3. DRM-EL denotes the proposed method based on model (5.4) and MLE denotes the maximum likelihood estimation method.

| Setting | Methods | MSE($\hat{\gamma}_{0,1}, \hat{\gamma}_{0,2}$) | MSE($\hat{\gamma}_{1,1}, \hat{\gamma}_{1,2}$) | MSE($\hat{\gamma}_{2,1}, \hat{\gamma}_{2,2}$) |
|---|---|---|---|---|
| Scenario II | DRM-EL | $(0.025, 0.015)$ | $(0.011, 0.027)$ | $(0.037, 0.076)$ |
| | MLE | $(0.018, 0.022)$ | $(0.022, 0.022)$ | $(0.042, 0.025)$ |
| Scenario III | DRM-EL | $(0.020, 0.027)$ | $(0.027, 0.030)$ | $(0.011, 0.032)$ |
| | MLE | $(0.075, 0.031)$ | $(0.056, 0.070)$ | $(0.030, 0.044)$ |

Table 5.3: Simulation results on the estimation of $\boldsymbol{\gamma}_k$ under Scenario II and 3. DRM-EL denotes the proposed method based on model (5.4) and MLE denotes the maximum likelihood estimation method.

and reproductive phases. During the vegetative stages, soybeans demonstrate a reasonable tolerance to drought stress, proving to be less susceptible than corn crops until the late reproductive stages.

Nonetheless, the simultaneous occurrence of water scarcity and heat stress poses a significant threat in many regions of Kansas during the soybean growing season. Such conditions can lead to increased flower and pod abortion, ultimately resulting in reduced soybean yields. Furthermore, excessively low temperatures also have adverse consequences, negatively impacting soybean germination and overall growth. As a result, it becomes evident that temperature and water availability exert substantial influence over the soybean yield as the growing season draws to a close.

In following analysis, we use the data on soybean yield and two environmental predictors in four specific years: 1993, 1998, 2001, and 2004. These years were chosen because they provide observations with reasonable sample sizes and cover the main time span of the data. Let us define the variables used in the model:

- $X(t)$ represents the daily average temperature between June and November, which are displayed in Fig (5.2).
- $Z$ represents the ratio of irrigated area for each county in Kansas.
- $Y$ represents the annual soybean yield per unit for each county in Kansas.

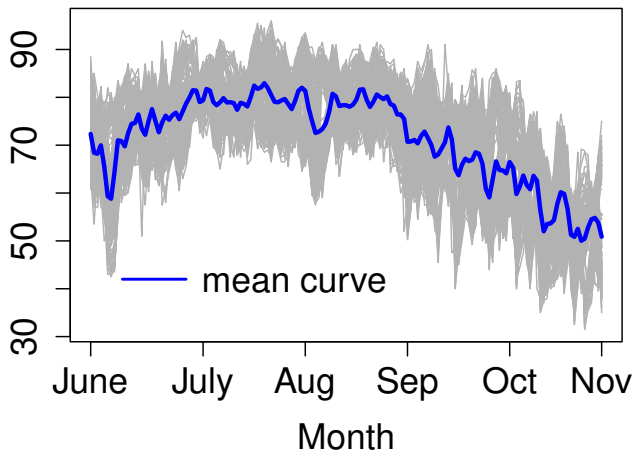



Figure 5.2: The daily average temperature curves of counties of Kansas in 1993, 1998, 2001 and 2004.

To keep things concise, we will use the year index $k$ to refer to the specific year under consideration. The following model is fitted to the target data set,

$$f_k(y|z,x(t)) = \frac{\exp\left\{\left(\gamma_k z + \langle x, \beta_k\rangle\right) y\right\} g_k(y)}{\int \exp\left\{\left(\gamma_k z + \langle x, \beta_k\rangle\right) y\right\} g_k(y) dy}, \quad k = 0,1,2,3 \tag{5.18}$$
$$\frac{g_r(y)}{g_0(y)} = \exp\{\alpha_r + \boldsymbol{\theta}_r^\tau \boldsymbol{q}(y)\}, \quad r = 1,2,3.$$

The vector $\boldsymbol{q}(y)$ of the density ratio structure in (5.18) is chosen as

$$\boldsymbol{q}(y) = (y, y^2, y^3, \log(y))^\tau,$$

which covers the Normal and Gamma distribution families. We also tried the vector $\boldsymbol{q}(y)$ with more polynomials terms and we observe very similar results. Therefore, in this section we report the estimation results based on this $\boldsymbol{q}(y)$

The estimated $\beta_k(t)$, $k = 0,1,2,3$ are displayed in Fig (5.3). According to Fig (5.3), we can observe the season effect of temperature and the impact of irrigation system on soybean yield. The estimated functions $\hat{\beta}_k(t)$ for all four years exhibit a consistent pattern. During the summer, all $\hat{\beta}_k(t)$ are negative, indicating that the temperature at this time is too hot for soybean growth. In contrast, during the winter, all $\hat{\beta}_k(t)$ become positive, suggesting that the temperature during this period is too cold for soybean growth. The turning point from negative to positive $\hat{\beta}_k(t)$ occurs around late July and early August. In addition, the estimated values for $\gamma_k$ are given by

$$(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.45, 0.56, 1.07, 2.09),$$

which indicates the positive effect of irrigation system on soybean growth. This observation emphasizes the importance of irrigation in promoting soybean growth and yield in Kansas.

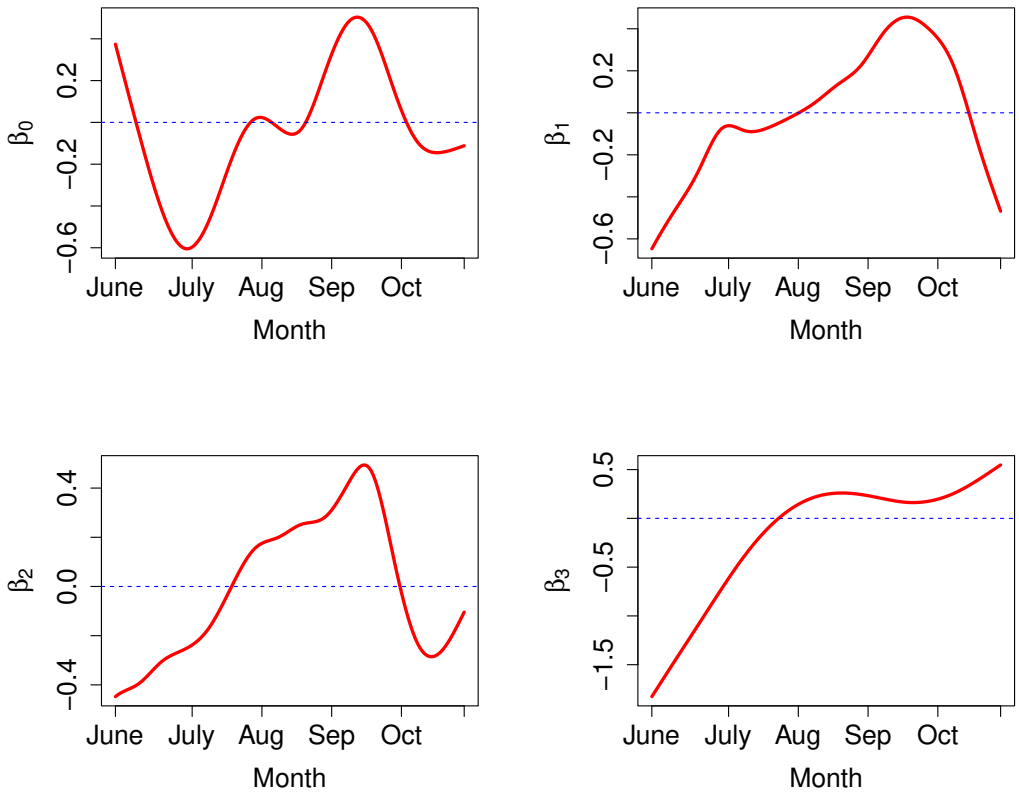



Figure 5.3: The estimated $\beta_k(t)$ using the proposed DRM-EL method based on the soybean data set of Kansas.

## 5.5. Conclusion and Discussion

The proposed semi-parametric functional generalized linear model (FGLM) with density ratio structure offers valuable flexibility for modeling data from different populations simultaneously, particularly for continuous responses. However, one limitation is that the current DRM-EL method based on (5.4) can only handle continuous response variables, as the density ratio model (DRM) is specifically defined for continuous random variables. For future research, we would like to investigate how to further modify the framework (5.4) to incorporate categorical or other discrete responses.

Note that in the real data analysis of the this project and the previous project, the response variables that related to the soybean yield are both modeled linearly in the daily average temperature $X(t)$ so that decreasing the summer temperature is estimated to improve the yield no matter how cold the summer is. In reality, this cannot be the case. This is another limitation of the current model

when it is applied on the target application. We also aim to solve this issue in the future research by allowing non-linear relationship between the response and the functional predictors.

# Bibliography


[1] JA Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

[2] DL Barrow and PW Smith. Asymptotic properties of best $l_2[0,1]$ approximation by splines with variable knots. *Quarterly of Applied Mathematics*, 36(3):293–304, 1978.

[3] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019.

[4] Howard D Bondell, Brian J Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

[5] J. Cao and J. O. Ramsay. Linear mixed effects modeling by parameter cascading. *Journal of the American Statistical Association*, 105:365–374, 2010.

[6] Hervé Cardot, Christophe Crambes, and Pascal Sarda. Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics*, 17(7):841–856, 2005.

[7] Hervé Cardot, Christophe Crambes, and Pascal Sarda. Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In *Statistical methods for biostatistics and related fields*, pages 221–243. Springer, New York, 2007.

[8] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591, 2003.

[9] Menglu Che, Linglong Kong, Rhonda C Bell, and Yan Yuan. Trajectory modeling of gestational weight: a functional principal component analysis approach. *PloS one*, 12(10):e0186761, 2017.

[10] Jiahua Chen and Yukun Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669 – 1692, 2013.

[11] Kehui Chen and Hans-Georg Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):67–89, 2012.

[12] Chao Cheng, Xingdong Feng, Jian Huang, and Xu Liu. Regularized projection score estimation of treatment effects in high-dimensional quantile regression. *Statistica Sinica*, 32(1):23–41, 2022.

[13] Victor Chernozhukov, Ivan Fernández-Val, and Alfred Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

[14] Miklos Csörgo and Pál Révész. *Strong Approximations in Probability and Statistics*. Academic press, 2014.

[15] C. de Boor. *A Practical Guide to Splines*. Springer, New York, 2001.

[16] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A Practical Guide to Splines*, volume 27. Springer, New York, 1978.

[17] RA El-Attar, M Vidyasagar, and SRK Dutta. An algorithm for l_1-norm minimization with application to nonlinear l_1-approximation. *SIAM Journal on Numerical Analysis*, 16(1):70–86, 1979.

[18] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and its Applications*. CRC Press, New York, 1996.

[19] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[20] Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[21] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

[22] Kuangnan Fang, Xiaochen Zhang, Shuangge Ma, and Qingzhao Zhang. Smooth and locally sparse estimation for multiple-output functional linear regression. *Journal of Statistical Computation and Simulation*, 90(2):341–354, 2020.

[23] Xingdong Feng, Xuming He, and Jianhua Hu. Wild bootstrap for quantile regression. *Biometrika*, 98(4):995–999, 2011.

[24] Marcelo Fernandes, Emmanuel Guerre, and Eduardo Horta. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39(1):338–357, 2021.

[25] Konstantinos Fokianos, Benjamin Kedem, Jing Qin, and David A Short. A semiparametric approach to the one-way layout. *Technometrics*, 43(1):56–65, 2001.

[26] Peter B Gilbert, Subhash R Lele, and Yehuda Vardi. Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika*, 86(1):27–43, 1999.

[27] Tianyu Guan, Zhenhua Lin, and Jiguo Cao. Estimating truncated functional linear models with a nested group bridge approach. *Journal of Computational and Graphical Statistics*, 29(3):620–628, 2020.

[28] Xin Guan, Hua Liu, Jinhong You, and Yong Zhou. Estimation and inference for dynamic single-index varying-coefficient models. *Statistica Sinica*, 2021. DOI:10.5705/ss.202019.0467.

[29] Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006.

[30] Peter Hall and Mohammad Hosseini-Nasab. Theory for high-order bounds in functional principal components analysis. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 146, page 225. Cambridge University Press, 2009.

[31] Qianchuan He, Linglong Kong, Yanhua Wang, Sijian Wang, Timothy A Chan, and Eric Holland. Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits. *Computational Statistics & Data Analysis*, 95:222–239, 2016.

[32] Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.

[33] Alan Huang and Paul J Rathouz. Proportional likelihood ratio models for mean regression. *Biometrika*, 99(1):223–229, 2012.

[34] Chiung-Yu Huang and Jing Qin. Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100(4):877–888, 2013.

[35] Jian Huang, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

[36] David R Hunter and Runze Li. Variable selection using MM algorithms. *The Annals of Statistics*, 33(4):1617–1642, 2005.

[37] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993.

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

[39] Gareth M James, Jing Wang, and Ji Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108, 2009.

[40] Bo Kai, Runze Li, and Hui Zou. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of statistics*, 39(1):305–332, 2011.

[41] Kengo Kato. Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136, 2012.

[42] Keith Knight. Limiting distributions for l1 regression estimators under general conditions. *The Annals of statistics*, 26(2):755–770, 1998.

[43] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[44] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[45] Roger Koenker and Olga Geling. Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468, 2001.

[46] Ming-Jun Lai and Larry L Schumaker. *Spline Functions on Triangulations*. Cambridge University Press, Cambridge, 2007.

[47] Eun Ryung Lee, Hohsuk Noh, and Byeong U Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

[48] Meng Li, Kehui Wang, Arnab Maity, and Ana-Maria Staicu. Inference in functional linear quantile regression. *Journal of Multivariate Analysis*, 190:104985, 2022.

[49] Yang Li, Fan Wang, Mengyun Wu, and Shuangge Ma. Integrative functional linear model for genome-wide association studies with multiple traits. *Biostatistics*, 23(2):574–590, 2022.

[50] Zhenhua Lin, Jiguo Cao, Liangliang Wang, and Haonan Wang. Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26(2):306–318, 2017.

[51] Yufeng Liu and Yichao Wu. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*, 23(2):415–437, 2011.

[52] Xiaodong Luo and Wei Yann Tsai. A proportional likelihood ratio model. *Biometrika*, 99(1):211–222, 2012.

[53] Xiaodong Luo and Wei Yann Tsai. Moment-type estimators for the proportional likelihood ratio model with longitudinal data. *Biometrika*, 102(1):121–134, 2015.

[54] Haiqiang Ma, Ting Li, Hongtu Zhu, and Zhongyi Zhu. Quantile regression for functional partially linear model in ultra-high dimensions. *Computational Statistics & Data Analysis*, 129:135–147, 2019.

[55] Haiqiang Ma, Ting Li, Hongtu Zhu, and Zhongyi Zhu. Quantile regression for functional partially linear model in ultra-high dimensions. *Computational Statistics & Data Analysis*, 129:135–147, 2019.

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

[57] Shujie Ma and Lijian Yang. A jump-detecting procedure based on spline estimation. *Journal of Nonparametric Statistics*, 23(1):67–81, 2011.

[58] Peter McCullagh. *Generalized linear models*. Routledge, 2019.

[59] Y. Nie and J. Cao. Sparse functional principal component analysis in a new regression framework. *Computational Statistics & Data Analysis*, 152:107016, 2020.

[60] Y. Nie, L. Wang, B. Liu, and J. Cao. Supervised functional principal component analysis. *Statistics and Computing*, 28:713–723, 2018.

[61] Yang Ning, Tianqi Zhao, and Han Liu. A likelihood ratio framework for high-dimensional semiparametric regression. *The Annals of Statistics*, 45(6):2299 – 2327, 2017.

[62] Juhyun Park, Jeongyoun Ahn, and Yongho Jeon. Sparse functional linear discriminant analysis. *Biometrika*, 109(1):209–226, 2022.

[63] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[64] Hartmut Prautzsch, Wolfgang Boehm, and Marco Paluszny. *Bézier and B-spline Techniques*. Springer Science & Business Media, 2002.

[65] Jing Qin and Biao Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

[66] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, New York, 2002.

[67] James O.. Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer, New York, 2009.

[68] James O Ramsay and Bernhard W Silverman. *Functional data analysis*. Springer, New York, second edition, 2005.

[69] JO Ramsay and BW Silverman. *Functional Data Analysis (2nd ed.)*. Springer Series in Statistics. New York: Springer., 2005.

[70] Paul J Rathouz and Liping Gao. Generalized linear models with unspecified reference distribution. *Biostatistics*, 10(2):205–218, 2009.

[71] P. Sang, L. Wang, and J. Cao. Parametric functional principal component analysis. *Biometrics*, 73:802–810, 2017.

[72] Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37):15594–15598, 2009.

[73] H. Shi, J. Dong, L. Wang, and J. Cao. Functional principal component analysis for longitudinal data with informative dropout. *Statistics in Medicine*, 40:712–724, 2021.

[74] A. Soliman, V. De Sanctis, R. Elalaily, and S. Bedair. Advances in pubertal growth and factors influencing it: Can we increase pubertal growth? *Indian Journal of Endocrinology and Metabolism*, 18(Suppl 1):S53–S62, 2014.

[75] C. J. Stone, M. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, 25(4):1371–1470, 1997.

[76] Kean Ming Tan, Lan Wang, and Wen-Xin Zhou. High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):205–233, 2022.

[77] Qingguo Tang and Linglong Kong. Quantile regression in functional linear semiparametric model. *Statistics*, 51(6):1342–1358, 2017.

[78] R.D. Tuddenham and M.M. Snyder. Physical growth of california boys and girls from birth to eighteen years. *University of California Publications in Child Development*, 1:183–364, 1954.

[79] Elisabeth Vogel, Markus G Donat, Lisa V Alexander, Malte Meinshausen, Deepak K Ray, David Karoly, Nicolai Meinshausen, and Katja Frieler. The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, 14(5):054010, 2019.

[80] Lifeng Wang, Guang Chen, and Hongzhe Li. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

[81] Yafei Wang, Linglong Kong, Bei Jiang, Xingcai Zhou, Shimei Yu, Li Zhang, and Giseon Heo. Wavelet-based lasso in functional linear quantile regression. *Journal of Statistical Computation and Simulation*, 89(6):1111–1130, 2019.

[82] Yueying Wang, Guannan Wang, Li Wang, and R. Todd Ogden. Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. *Biometrics*, 76:427–437, 2020.

[83] Yuanshan Wu, Yanyuan Ma, and Guosheng Yin. Smoothed and corrected score approach to censored quantile regression with measurement errors. *Journal of the American Statistical Association*, 110(512):1670–1683, 2015.

[84] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[85] Fang Yao, Shivon Sue-Chee, and Fan Wang. Regularized partially functional quantile regression. *Journal of Multivariate Analysis*, 156:39–56, 2017.

[86] Dengdeng Yu, Linglong Kong, and Ivan Mizera. Partial functional linear quantile regression for neuroimaging data analysis. *Neurocomputing*, 195:74–87, 2016.

[87] Shan Yu, Guannan Wang, Li Wang, Chenhui Liu, and Lijian Yang. Estimation and inference for generalized geoadditive models. *Journal of the American Statistical Association*, 115(530):761–774, 2020.

[88] Shan Yu, Guannan Wang, Li Wang, and Lijian Yang. Multivariate spline estimation and inference for image-on-scalar regression. *Statistica Sinica*, 31(3):1–39, 2021.

[89] Ming Yuan and Yi 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.

[90] Archer Gong Zhang and Jiahua Chen. Density ratio model with data-adaptive basis function. *Journal of Multivariate Analysis*, 191:105043, 2022.

[91] X. Zhang, J. Cao, and R. J. Carroll. Estimating varying coefficients for partial differential equation models. *Biometrics*, 73:949–959, 2017.

[92] Zhengwu Zhang, Xiao Wang, Linglong Kong, and Hongtu Zhu. High-dimensional spatial quantile function-on-scalar regression. *Journal of the American Statistical Association*, 2021. DOI:10.1080/01621459.2020.1870984.

[93] Zhibiao Zhao and Zhijie Xiao. Efficient regressions via optimally combining quantile information. *Econometric Theory*, 30(6):1272–1314, 2014.

[94] Jianhui Zhou, Nae-Yuh Wang, and Naisyin Wang. Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23(1):25–50, 2013.

[95] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.

[96] Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.