ROBUST STATISTICAL MODELING IN FUNCTIONAL LINEAR REGRESSION

YAN ZHANG

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

September, 2024

Abstract

Functional linear regression is a prominent field within the domain of functional data analysis, with extensive applications in various domains such as biomedical studies, brain imaging, and chemometrics. However, despite the abundance of literature on functional linear regression, limited attention has been devoted to addressing outliers or heavy-tailed distributions in the data. Consequently, robust statistical analysis remains an underdeveloped practice in this area. The primary objective of this dissertation is to enhance the utilization of robust methods for modeling functional linear regression by primarily focusing on robust estimation techniques, hypothesis testing procedures that are resilient to outliers or heavy-tailed distributions, and robust variable selection methods.

First, we consider the problem of robust estimation in partial functional linear models under RKHS framework. The theoretical properties of robust estimation simulation studies are discussed in this chapter. Furthermore, two real data examples are presented to illustrate the performance of the robust procedure.

Then, we extend three robust tests: Wald-type, the likelihood ratio-type and F-type in functional linear models. Meanwhile, we investigate the theoretical properties of these robust testing procedures and assess the finite sample properties through the numerical simulation.

Finally, we propose a robust variable selection method in multiple functional linear regression and present a novel algorithm for identifying significant functional predictors using a robust group variable inflation factor (VIF) selection procedure. Our methodology is validated through rigorous simulation studies as well as its application to real-world data.

To ensure the cohesiveness of this dissertation, Chapter 1 provides an introduction to the research background, mathematical foundations, and primary motivations underlying this study. Chapter 2 presents a comprehensive overview of basis expansion methods for functional data analysis. Lastly, Chapter 6 concludes this dissertation by offering potential avenues for future research.

Acknowledgements

Though it is impossible to express my gratitude to all individuals who supported me, I wish to extend my deepest appreciation to several special people who have helped and encouraged me in compiling this dissertation.

I would like to begin by expressing my deepest gratitude towards my supervisor, Professor Yuehua Wu. Her constant support and strong encouragement have been invaluable during the development of these ideas. Without her guidance, care, patience, and persistent help, this dissertation would not have been possible.

I would like to express my sincere appreciation to Professor Xin Gao and Professor Michael Chen for their invaluable guidance and support as esteemed members of my supervisory committee. I am deeply grateful to all the faculty members and staff in the Department of Mathematics and Statistics at York University. Furthermore, I extend my heartfelt thanks to Professor Yuejiao Fu, Professor Ada Chan, Primrose Miranda, Steven Chen, Ann-Marie Carless, Susan Raniey, and Wenrui Cui for their generous assistance throughout this research endeavor. Finally, I am indebted to my

dear friends Yixin Zhang, Dongwei Wei, and Jiacheng Wang for their unwavering encouragement and unwavering support.

Last but not least, I deeply thank my parents, whose love and guidance are with me in whatever I pursue. They are always supporting me with their best wishes. I wish to thank my loving and supportive families, who provide unending inspiration. To my parents, Bosong and Jingmin, my husband Xiaohua, and my beloved children, Grayson and Lauren.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Functional Data Analysis

In recent decades, there has been a significant increase in the continuous recording of data over a time interval or intermittently at discrete time points. This growth can be attributed to major advancements in data collection technology, which have led to the emergence of the revolutionary concept of "Big Data". The distinguishing feature of these datasets is their infinite-dimensional structure, where the primary unit of observation can be conceptualized as a curve or more generally as a function. Such data can be referred to as *functional data*, giving rise to the rapid development of a novel field in statistics known as *functional data analysis* (FDA). From a mathematical perspective, they can inherently be considered as realizations of functional random variables or trajectories of suitable stochastic processes. Failure to consider the functional characteristics of this type of data when employing conventional multivariate methods for analysis can lead to significant challenges, including the curse

of dimensionality, collinearity issues, and loss of valuable information. In such scenarios, specific statistical techniques are indispensable for effectively handling and extracting relevant underlying information.

The presence of functional data is pervasive across diverse scientific disciplines, underscoring their natural occurrence. For instance, econometrics encompasses curves representing financial assets, energy research involves curves depicting electricity demand, and environmental science incorporates curves illustrating pollutant levels. Moreover, advancements in technology have expanded the scope of this field to encompass other domains such as spatial and imaging domains, as well as genomic locations. A plethora of methodologies and applications including both parametric and nonparametric approaches have been comprehensively summarized in Ramsay and Silverman (2005), Ramsay and Silverman (2007), Ferraty and Vieu (2006), among numerous other scholarly works.

1.1.1 What Is Functional Data?

The fundamental unit of functional data analysis is a function, where one or multiple functions are recorded for each subject in a random sample (Wang et al., 2016). In the context of functional data, the optimal units of the observation are functions defined on a continuous domain, with each function being sampled on a

discrete grid. This grid can be either dense or sparse, regular or irregular, and may vary across the sampled functions (Morris, 2015). The basic idea in FDA is to regard functional data as realizations of an underlying stochastic process. In practical applications involving real data, if a random variable $X(t)$ is observed on a discrete grid $t \in \mathcal{T}$, where $\mathcal{T}$ typically represents a compact interval on the real line and t usually denotes time, it can be assumed that $\{X(t) : t \in \mathcal{T}\}$ constitutes a functional random variable when the time instants t are sufficiently close. It should also be noted that the independent parameter denoted by t, often conceptualized as time, has the flexibility to represent any other parameter. For example, in a spectrometer data set, the response is determined by wavelengths, while in a chemometric study on material weight decay under applied heat, the mass loss curve with temperature would be recorded.

Definition 1.1 Given a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, a **functional random variable** $\{X(t) : t \in \mathcal{T}\}$ is a $\mathcal{B}$-measurable mapping from sample space Ω to an infinite dimensional space or functional space $\mathcal{F}$. $\mathcal{B}$ is typically a σ-algebra.

Definition 1.2 Given a functional random variable X, a n-length **functional random sample** of X is a set of independent and identically distributed functional random variables $\{X_i, i = 1, 2, \cdots, n\}$ with the same distribution as X.

Definition 1.3 An instance of the functional random variable will be called as

functional data. And a data set $X_1, X_2, \cdots, X_n$ containing n functional random variables is a functional data set. Here we note that the random variable X being a function must have a defined support $\mathcal{S} = \{X(t) : t \in \mathcal{T}\}$. The support set $\mathcal{S}$ can be from uni-dimensional or multidimensional set of positive reals.

Typically, functional data is observed and recorded on a grid, which may or may not be predefined. It is essential to consider the spacing between the grid points. If the data is collected at regular intervals, it will be aligned on an equally spaced grid, referred to as balanced functional data. However, in many experimental scenarios, the grid points are not predetermined, resulting in irregularly recorded data. Despite being functionally natured, such datasets are termed imbalanced functional data. Although we acknowledge that imbalanced datasets are prevalent in natural phenomena, this work does not delve into their specifics; instead, all applications considered in this document assume an equally spaced grid.

1.1.2 Some Real Data Examples

In terms of intricacy, functional data can be broadly classified into two categories based on their origin's simplicity-the first generation and the next generation. The first generation functional data set consists of curves, which is most common when a single dataset is considered. On the other hand, the next generation functional data

is derived from complex data objects and involves more profound analysis problems (Wang et al., 2016). These datasets may exhibit various structural elements such as correlation, repeated measurements, or other inherent characteristics. Our research primarily focuses on analyzing the first generation functional data, and all real examples provided in this section fall within this category. Several of these examples have been extensively examined throughout our studies.

Example 1.1 Handwriting Data

The first example comes from **fda** package (Ramsay and Silverman, 2005). The dataset consists of 20 samples of the word "fda" written in cursive. Each sample includes 1401 pairs of (x, y) coordinates, representing the pen's position over time.

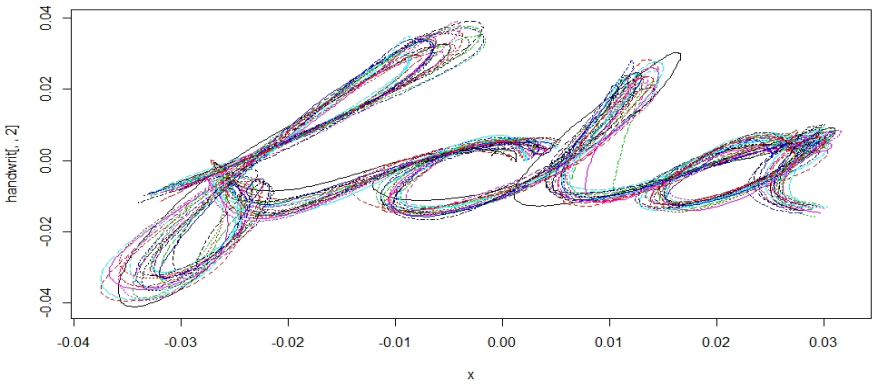

Figure 1.1: Records of 20 cursive samples for writing "fda"

Example 1.2 Canadian Weather Data

The second example is derived from the Canadian Weather dataset, which encompasses averaged daily temperature and precipitation records at 35 distinct locations across Canada spanning from 1960 to 1994. Functional data analysis methods applied to the Canadian weather data have been extensively explored in numerous seminal works. By loading the R package **fda**, we can extract the mean temperature curves as demonstrated below.

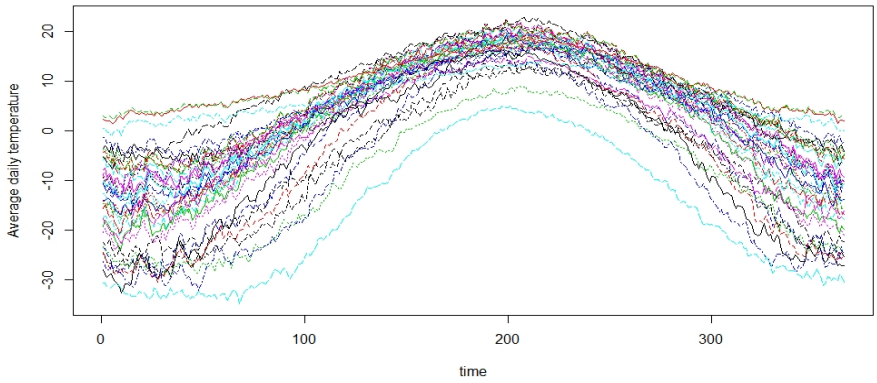

Figure 1.2: Canadian Weather Data: 35 mean temperature estimated curves

Example 1.3 Diffusion Tensor Data

The third example pertains to the Diffusion Tensor Imaging (DTI) study data, which were collected at Johns Hopkins University and the Kennedy-Krieger Institute. This data-set is available in R package **refund**. Diffusion tensor imaging (DTI) tractog-

raphy is a magnetic resonance imaging technique that enables the quantification of water's restricted diffusion within tissue, facilitating the generation of neural tract images. It allows the study of white-matter tracts by measuring the diffusivity of water in the brain:within white-matter tracts, water diffuses anisotropically along their direction, while elsewhere it diffuses isotropically. Fractional anisotropy (FA) measures water molecule diffusion at each voxel and profiles along both corpus callosum and right corticospinal tracts.

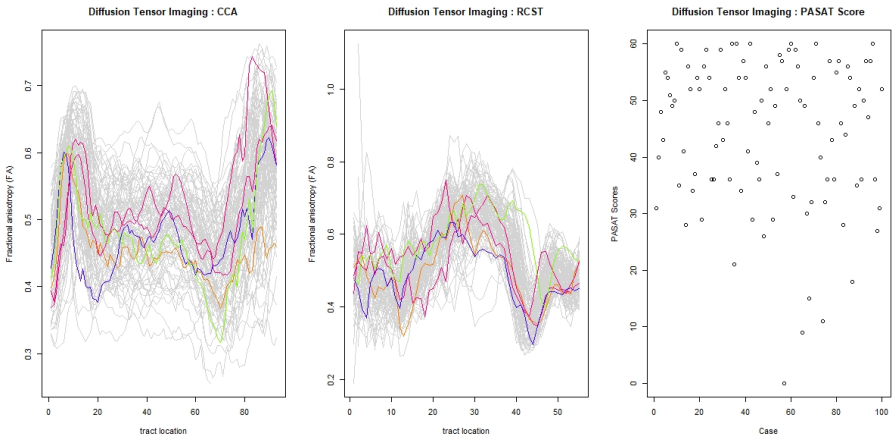

Figure 1.3: Fractional anisotropy profiles along corpus callosum (CC) and the right corticospinal tract(RCST). The associated Paced Auditory Serial Addition Test scores (PASAT) of the 100 multiple sclerosis patients.

Example 1.4 Near-infrared Spectroscopy Data

Quantitative NIR (near-infrared spectroscopy) data was obtained from Osborne et al.

(1984). This dataset comprises 72 sample sets, with variations in the standard recipe to encompass a wide range for each of the four constituents: fat, sucrose, dry flour, and water. The measurements for these constituents are expressed as percentages. Spectra were collected between 1100 and 2498 nanometers (nm), with increments of 2 nm, resulting in densely observed functional predictors on a grid of 700 points.

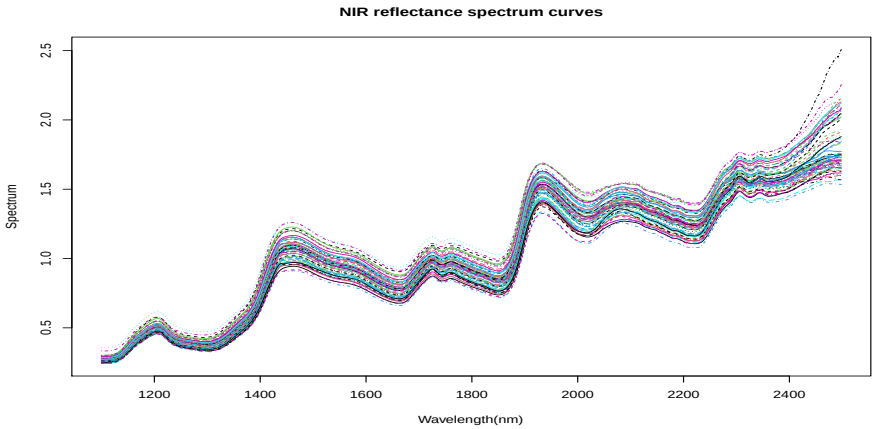

Figure 1.4: NIR(near-infrared reflectance) spectrum measured from 1100 to 2498 nanometers (nm) in 2 nm increments.

Example 1.5 Fat Content Spectrometric Data

The Fat Content Spectrometric (FCS) data pertains to a sample comprising 215 finely chopped meat pieces. This dataset is a component of the Tecator dataset, which can be accessed through the R package **fds** and is also available on the website (*http://lib.stat.cmu.edu/datasets*). Each sample encompasses a 100-channel ab-

sorbance spectrum within the wavelength range of 850-1050 nm. Every spectrum in the database corresponds to an analytical chemistry-derived content description of the meat sample, encompassing percentages denoting fat, water, and protein (Rossi et al., 2005).

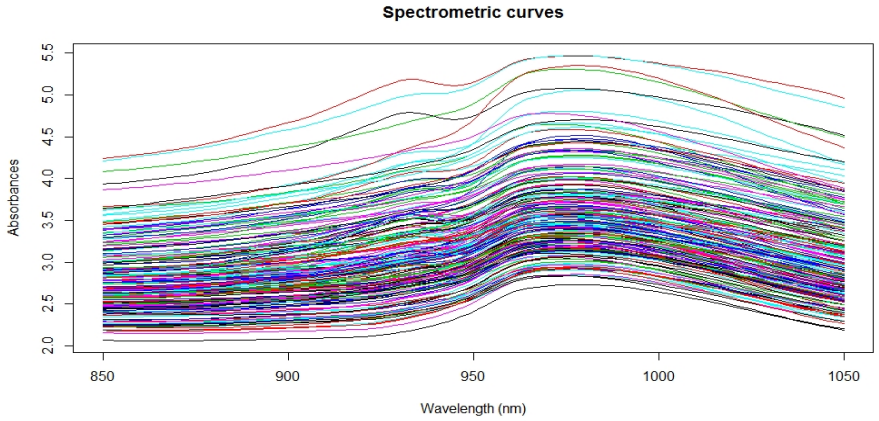

Figure 1.5: Fat content spectrometric data: absorbance trajectories of 215 meat samples measured over 100 equally spaced wavelengths between 850 nm and 1050 nm.

1.1.3 Functional Linear Regression

The concept of *Functional linear regression* (FLR) was initially introduced by Ramsay and Dalzell (1991) and further refined in its common form by Hastie and Mallows (1993), which is a powerful tool in functional data analysis, utilized for modeling the relationships between functional predictors and responses. Functional

linear regression aims to extend the applicability of linear models into the functional domain, which can be seen as a continuous counterpart of multivariate linear regression.The FLR methodology finds applications in diverse domains, including finance (for the prediction of stock prices based on historical data), medicine (in the analysis of growth curves), and environmental science (for modeling pollution levels) (Ramsay and Silverman, 2005; Müller, 2005; Morris, 2015).

The issue of functional linear regression modeling is an active area of research, and it can be categorized into three types based on the nature of the variables: (1) scalar-on-function models where scalar responses are modeled with functional covariates, (2) function-on-function models where functional responses are modeled with functional covariates, and (3) function-on-scalar models where functional responses are modeled with a set of scalar covariates. Our research primarily focuses on scalar-on-function linear models, which can be defined as follows:

$$Y = \alpha + \int_{\mathcal{T}} \beta(t)X(t)dt + \varepsilon, \tag{1.1}$$

where Y is a scalar response and $X(t)$ is a potential functional covariate with a compact set $\mathcal{T} \subset \mathbb{R}$, $\beta(t)$ is a coefficient function to be estimated, α is an unknown intercept, and ϵ is a random error being independent with $X(t)$. In the field of functional predictor regression, most methodological advancements have adhered to the overarching framework proposed by Ramsay and Silverman (2005), employing

diverse choices of basis functions and regularization techniques. Similar to traditional linear models, a key challenge lies in estimating the coefficient function $\beta(t)$. Initially, one could adopt a naive approach such as ordinary least squares (OLS) or impose certain constraints. However, dimension reduction and the selection of an appropriate basis for expansion pose significant challenges in the analysis of functional linear regression.

1.2 Robust Statistics

1.2.1 Why Robust Statistics Is Needed?

The field of robust statistics, in a broad and informal sense, addresses the inherent limitations associated with numerous assumptions commonly employed in statistical analysis, such as normality, linearity, and independence. One rationale behind this approach is the presence of outliers; these outliers represent data points that exhibit significant deviation from the majority of observations and have the potential to substantially distort traditional statistical methodologies even when only a single outlier is present. Another reason is that idealized model assumptions often fail to capture the empirical nature of many real-world phenomena. For instance, a widely adopted assumption in statistical modeling is that observed data conforms to a normal (Gaussian) distribution. This assumption has served as the fundamental

framework for classical methodologies in regression, analysis of variance, and multivariate regression. However, this assumption can be easily violated in practical applications where the actual distributions of data exhibit heavy tails. In such cases, estimates based on classical methods may suffer from significantly reduced statistical efficiency or substantial bias.

Consequently, robust statistical procedures have been developed as adaptations of classical methods to accommodate minor deviations from the assumed conditions. This objective remains valid not only when the data strictly adhere to a given distribution but also when they approximately conform to it, as previously described. In briefly, Huber and Ronchetti (2011) gave a relatively narrow definition for "robustness": *robustness signifies insensitive to small deviations from the assumptions or being less influenced by outliers.* Naturally, there exist fundamental tools employed for assessing the robustness of an estimator, namely sensitivity curve, influence function, and breakdown point (value). However, we will not delve into the complexity of their definitions and theoretical backgrounds, as the scope of our work does not cover this aspect.

1.2.2 M-estimation

Robust regression is a crucial tool for analyzing data contaminated with outliers, enabling the detection of outliers and providing resistant results in their presence. Numerous methods have been developed to tackle these challenges, including M-estimator, L-estimator, and R-estimator. These robust estimators extend the concept of maximum likelihood estimation (MLE) proposed by Huber (1964). Let $\mathcal{X}$ denote the sample space and Θ represent the parameter space. Assume an estimator $T_n = T_n(x_1, x_2, \cdots, x_n) \in \Theta$ satisfies the optimal equation

$$\min \sum_{i=1}^{n} \rho(x_i; T_n), \tag{1.2}$$

where $\rho(\cdot)$ is a properly chosen function on $\mathcal{X} \times \Theta$, then T_n is called an M-estimator. If the loss function ρ has a derivative $\psi(x, \theta) = \dfrac{\partial \rho(x, \theta)}{\partial \theta}$, the estimate T_n satisfies the implicit equation

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0. \tag{1.3}$$

Typically, equations (1.2) and (1.3) are not always equivalent; however, equation (1.3) is often valuable in the pursuit of a solution for equation (1.2). The maximum likelihood estimator is also an M-estimator, corresponding to $\rho(x, \theta) = -\ln f_\theta(x)$, and $f(\cdot)$ is the density function.

Remark 1.1 The well known families of robust loss function are Huber's, Tukey's

and Hampel's families. Unlike the standard least square loss function $\rho(x) = x^2$, these loss functions have the property that $\psi(x) = \rho'(x)$ is bounded.

- Huber's family of loss function is given by (Huber, 1964)

$$\rho(x) = \begin{cases} \dfrac{x^2}{2}, & if\ |x| \leqslant c, \\ c\left(|x| - \dfrac{c}{2}\right), & if\ |x| > c, \end{cases} \tag{1.4}$$

  with threshold parameter $c > 0$. This is a convex, but not strictly convex loss function.

- Tukey's bisquare family of loss function is given by (Beaton and Tukey, 1974)

$$\rho(x) = \begin{cases} \dfrac{c^2}{6}\left[1 - \left(1 - \left(\dfrac{x}{c}\right)^2\right)^3\right], & if\ |x| \leqslant c, \\ \dfrac{c^2}{6}, & if\ |x| > c, \end{cases} \tag{1.5}$$

  with parameter $c > 0$. This loss function is not convex but can better deal with extreme outliers.

- Hampel's family of loss function is given by (Hampel, 1974)

$$\rho(x) = \begin{cases} \dfrac{x^2}{2}, & if\ |x| \leqslant a, \\ a|x| - \dfrac{a^2}{2}, & if\ a < |x| \leqslant b, \\ \dfrac{a}{2}\left[2b - a + (|x| - b)\left(1 + \dfrac{c - |x|}{c - b}\right)\right], & if\ b < |x| \leqslant c, \\ \dfrac{a}{2}(b + c - a), & if\ |x| > c, \end{cases} \tag{1.6}$$

with non-negative parameters $0 < a < b < c$. The loss function also exhibits non-convexity, wherein the contribution to the loss remains unchanged when the observations are far from the center (Wilcox, 2011).

In summary, robust procedures are designed to effectively handle contaminated data by exhibiting less sensitivity towards outliers and deviations from distributional assumptions on random errors. Firstly, robust methods employ diverse loss functions to mitigate the impact of outliers, encompassing concerns related to outliers in covariate variables. For instance, according to the Huber's loss function, it can be observed that for values of $|x| \leqslant c$, the Huber loss exhibits characteristics similar to the squared error loss, thereby demonstrating sensitivity towards small errors. Conversely, when $|x| > c$, the loss function becomes linear, which reduces the influence of large residuals (outliers). In contrast to the abrupt corner at zero in the absolute error loss, the Huber loss exhibits smoothness throughout, thereby offering potential advantages for optimization algorithms. Secondly, robust methods typically have a high breakdown point value, indicating their ability to withstand substantial contamination without compromising accuracy. Lastly, robust methodologies often employ flexible models that are not heavily reliant on stringent assumptions regarding the distribution of data, thereby enhancing their adaptability to real-world datasets which frequently exhibit imperfections and complexities.

1.3 Some Mathematical Foundations of Functional Data Analysis

In this section, we present an exposition on the mathematical underpinnings of functional data analysis. The contents of this segment have been succinctly summarized in Hsing and Eubank (2015).

1.3.1 Vector and Functional Space

Firstly, this section aims to elucidate some fundamental concepts that are indispensable for introducing the definition of functional space.

Definition 1.4 A **metric** on a set $\mathbb{M}$ is a function d: $\mathbb{M} \times \mathbb{M} \to \mathbb{R}$ that satisfies:

1. $d(x, y) \geqslant 0$, **Non-negativity**.
2. $d(x, y) = 0$, if $x = y$, **Identity of indiscernibl**e.
3. $d(x, y) = d(y, x)$, **Symmetry**.
4. $d(x, z) \leqslant d(x, y) + d(y, z)$, **Triangle inequality**.

Definition 1.5 Let $(\mathbb{M}, d)$ be a metric space with $E \subset \mathbb{M}$. Then, E is said to be **open** if for every $e \in E$ there exists an $\epsilon > 0$ such that $\{x \in \mathbb{M} : d(x, e) < \epsilon\} \subset E$.

That is, if every point in E has a neighborhood contained in E, E is open. Otherwise, E is **closed**.

Definition 1.6 (Countable and Dense)

1. The **closure** $\bar{E}$ of $E \subset \mathbb{M}$ is the smallest closed set in $\mathbb{E}$ that contains E.

2. A set $E \subset \mathbb{E}$ is **dense** in $\mathbb{E}$, if $\bar{E} = \mathbb{E}$.

3. A set E is **countable** if there exists an injective (one-to-one) function f from E to the natural numbers $N = \{0, 1, 2, 3, \cdots\}$.

4. A metric space $\mathbb{E}$ is **separable space** if it has a **countable**, **dense** subset.

Definition 1.7 A vector space $\mathbb{V}$ is a set of elements, referred to as vectors, for which two operations have been defined: addition and scalar multiplication. Given vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{R}$, the addition and multiplication operations are assumed to satisfy

1. $\boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v}_2 + \boldsymbol{v}_1$.

2. $\boldsymbol{v}_1 + (\boldsymbol{v}_2 + \boldsymbol{v}_3) = (\boldsymbol{v}_1 + \boldsymbol{v}_2) + \boldsymbol{v}_3$.

3. $a_1(a_2)\boldsymbol{v} = (a_1 a_2)\boldsymbol{v}$.

4. $a(\boldsymbol{v}_1 + \boldsymbol{v}_2) = a\boldsymbol{v}_1 + a\boldsymbol{v}_2$, and $(a_1 + a_2)\boldsymbol{v} = a_1\boldsymbol{v} + a_2\boldsymbol{v}$.

5. $1\boldsymbol{v} = \boldsymbol{v}$.

In addition, there is a unique element $\mathbf{0}$ with the property that $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$ for every $\boldsymbol{v} \in \mathbb{V}$ and corresponding to each element $\boldsymbol{v}$ there is another element $-\boldsymbol{v}$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}$.

1.3.2 Operators and Random Elements in a Hilbert Space

Definition 1.8 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be normed linear spaces (vector spaces) over the same scalar field $\mathbb{F}$. A mapping $\mathcal{L}$ defined over a linear subspace $\mathcal{D}_\mathcal{L}$ of $\mathcal{H}_1$ ($\mathcal{D}_\mathcal{L} \subseteq \mathcal{H}_1$), and taking values in $\mathcal{H}_2$ is said to be a **linear operator** if

$$\mathcal{L}(\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2) = \alpha_1 \mathcal{L}\boldsymbol{x}_1 + \alpha_2 \mathcal{L}\boldsymbol{x}_2, \tag{1.7}$$

for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}_\mathcal{L}$ and scalar $\alpha_1, \alpha_2 \in \mathbb{F}$, $\mathcal{D}_\mathcal{L}$ is the domain of the operator $\mathcal{L}$.

Definition 1.9 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be normed linear spaces (vector spaces) with norm $\|\cdot\|_{\mathcal{H}_i}, i = 1, 2$. A linear operator $\mathcal{L}$ is called **bounded** if there exist a finite constant C, such that

$$\|\mathcal{L}\boldsymbol{x}\|_{\mathcal{H}_2} \leqslant C \|\boldsymbol{x}\|_{\mathcal{H}_1}, \tag{1.8}$$

for all $\boldsymbol{x} \in \mathcal{D}_\mathcal{L}$. Then C is called a bound of $\mathcal{L}$.

Remark 1.2 From Definition 1.9, if a linear operator is bounded, it implies that $\sup_{\substack{\boldsymbol{x} \in \mathcal{D} \\ \|\boldsymbol{x}\|_{\mathcal{H}_1} \leqslant 1}} \|\mathcal{L}\boldsymbol{x}\|_{\mathcal{H}_2} < \infty$.

Example 1.6 (Identity operator) Let $\mathcal{H}$ be a Hilbert space. The identity operator $\mathcal{I} : \mathcal{H} \to \mathcal{H}$ defined by $\mathcal{I}\boldsymbol{x} = \boldsymbol{x}$, $\boldsymbol{x} \in \mathcal{H}$, is linear and bounded with $\|\mathcal{I}\| = 1$ when $\mathcal{H} \neq \{\mathbf{0}\}$.

Example 1.7 (Integral operator) Let $(\Omega, \mathcal{A}, \mu)$ is a $\sigma-$finite measurable space, and K be some square integral function on $\mathcal{A} \times \mathcal{A}$. The linear mapping defined by

$$(\mathcal{L}\boldsymbol{x})(s) = \int_{\Omega} K(s,t)\boldsymbol{x}(t)d\mu(t), \tag{1.9}$$

is a bounded linear operator. The function $K(s,t)$ is called a kernel function.

Theorem 1.1 Suppose $\mathcal{L} : \mathbb{F}^n \to \mathbb{F}^m$ is a linear operator. Then there exists a unique $m \times n$ matrix $A_{m\times n}$ such that $\mathcal{L}\boldsymbol{x} = A\boldsymbol{x}$, $\forall \boldsymbol{x} \in \mathbb{F}^n$.

Proof. Suppose $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ is a standard basis of $\mathbb{F}^n$, $\epsilon_1, \epsilon_2, \cdots, \epsilon_m$ is standard basis of $\mathbb{F}^m$, then $\boldsymbol{x} = \epsilon_1 x_1 + \cdots + \epsilon_n x_n$. $\mathcal{L}\boldsymbol{x} = \mathcal{L}(\epsilon_1 x_1 + \cdots + \epsilon_n x_n) = x_1\mathcal{L}\epsilon_1 + \cdots + x_j\mathcal{L}\epsilon_j + x_n\mathcal{L}\epsilon_n$. Note that $\mathcal{L}\epsilon_j \in \mathbb{R}^m$, $j = 1, \cdots, n$, then we have $\mathcal{L}\epsilon_j = a_{1j}\epsilon_1 + \cdots + a_{mj}\epsilon_m$, $j = 1, 2, \cdots, n$. Therefore, we can arrange these scalars in an $m \times n$ matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \text{ and } \mathcal{L}\boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = A\boldsymbol{x}. \tag{1.10}$$

□

Example 1.8 Let $\mathcal{L} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $\mathcal{L}\boldsymbol{x} = (ax_1 + bx_2 + cx_3, dx_1 + ex_2 + fx_3)$ for some $a, b, c, d, e, f \in \mathbb{R}$. Then, with respect to the canonical basis of $\mathbb{R}^3$ given by $(1, 0, 0), (0, 1, 0), (0, 0, 1)$, the corresponding matrix is

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}.$$

Example 1.9 Let $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $\mathcal{L}\boldsymbol{x} = (x_2, x_1 + 2x_2, x_1 + x_2)^T$. Then, with respect to the standard basis, we have $\mathcal{L}(1, 0)^T = (0, 1, 1)^T, \mathcal{L}(0, 1)^T = (1, 2, 1)^T$, and the corresponding matrix is

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}.$$

Definition 1.10 Tensor product Let vector $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)' \in \mathbb{R}^n$ and $\boldsymbol{u} = (u_1, u_2, \cdots, u_m)' \in \mathbb{R}^m$, then Tensor product (outer product) of $\boldsymbol{v}$ and $\boldsymbol{u}$ is given by

$$\boldsymbol{v} \otimes \boldsymbol{u} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \otimes \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} = \begin{bmatrix} v_1u_1 & v_1u_2 & \cdots & v_1u_m \\ v_2u_1 & v_2u_2 & \cdots & v_2u_m \\ \vdots & \vdots & \ddots & \vdots \\ v_nu_1 & v_2u_2 & \cdots & v_nu_m \end{bmatrix} = \boldsymbol{v}\boldsymbol{u}^T. \tag{1.11}$$

When considering functional data, the smoothness of a function mapping from one metric space to another becomes a crucial aspect, as it determines how the value

of the function changes with variations in its argument. Continuity is regarded as the fundamental form of smoothness.The mathematical theory of functional data analysis is expounded in greater detail in (Hsing and Eubank, 2015).

1.3.3 Karhunen–Loève Decomposition

In the theory of stochastic processes, the Karhunen–Loève theorem (named after Kari Karhunen and Michel Loève), also known as the Kosambi–Karhunen–Loève theorem, provides a representation of a stochastic process as an infinite linear combination of orthogonal functions, similar to a Fourier series representation of a function on a bounded interval. This theorem is closely associated with the widely employed principal component analysis (PCA) technique in multivariate scenarios. Furthermore, this principle serves as the fundamental basis for the theory of functional principal component analysis (FPCA, hereinafter).

Theorem 1.2 (Mercer's Theorem, Shorack and Wellner (2009))

Let $K(s,t)$ be a symmetric, non-negative definite, continuous function on $(a,b) \times (a,b)$. There exists a countable sequence of functions $\phi_j(t)$ and a sequence of positive real numbers λ_j such that for any $s,t \in (a,b)$, the kernel function can be expressed as

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t)$$

where the value λ_j and the function ϕ_j satisfies the integral eigenvalue equation:

$$\int_a^b K(s,t)\phi_j(s)dt = \lambda_j\phi_j(t).$$

The quantity λ_j is referred to as an eigenvalue of the kernel function K, while the corresponding function f_j is known as its associated eigenfunction. Furthermore, it holds that the series $\sum_{j=1}^{\infty} \lambda_j$ converges. One of its main applications is to find convenient ways for expressing stochastic processes through the Karhunen–Loève expansion.

Theorem 1.3 (Karhunen–Loève Decomposition, Wang (2008))

Suppose $X(t)$ is a stochastic process indexed by t in a finite interval $\mathcal{T}$ and take some values in $L^2(\Omega, P)$ for some probability space (Ω, P). Without loss of generality, we assume that $E(X(t)) = \mu(t) = 0$, and covariance function $K(s,t) = Cov[X(t), X(s)] = E(X(s)X(t))$. Then, the process $X(t)$ can be written down as:

$$X(t) = \sum_{i=1}^{\infty} \xi_i\phi_i(t), \quad \text{where} \quad \xi_i = \int_{\mathcal{T}} X(t)\phi_i(t)dt,$$

and $E(\xi_i) = 0$, $Cov(\xi_i\xi_j) = \lambda_i\delta_{ij}$, and $Var(\xi_i) = \lambda_i$.

Proof. (a) If we let $\xi_i = \int_{\mathcal{T}} X(t)\phi_i(t)dt$, then we have

$$\begin{aligned} E(\xi_i) =& E(\int_{\mathcal{T}} X(t)\phi_i(t)dt) \\ =& \int_{\mathcal{T}} E(X(t))\phi_i(t)dt \\ =& 0. \end{aligned}$$

$$\begin{aligned} E(\xi_i\xi_j) =& E(\int_{\mathcal{T}}\int_{\mathcal{T}} X(t)\phi_i(t)X(s)\phi_j(s)dtds) \\ =& \int_{\mathcal{T}}\int_{\mathcal{T}} E(X(t)X(s))\phi_i(t)\phi_j(s)dtds \\ =& \int_{\mathcal{T}}(\int_{\mathcal{T}} K(t,s)\phi_i(t)dt)\phi_j(s)ds \\ =& \lambda_i \int_{\mathcal{T}} \phi_i(s)\phi_j(s)ds \\ =& \lambda_i\delta_{ij}. \end{aligned}$$

Note that $\phi(.)$ is orthonormal and δ_{ij} is Kronecker Delta, therefore we have $Var(\xi_i) = \lambda_i \times 1 = \lambda_i$. Also, we note that

$$\begin{aligned} E(X(t)\xi_i) =& E(X(t)\int_{\mathcal{T}} X(s)\phi_i(s)ds) \\ =& \int_{\mathcal{T}} E(X(t)X(s))\phi_i(s)ds \\ =& \int_{\mathcal{T}} K(t,s)\phi_i(s)ds \\ =& \lambda_i\phi_i(t). \end{aligned}$$

Hence,

$$
\begin{aligned}
&E(X(t)-\sum_{i=1}^{n}\xi_i\phi_i)^2\\
=&E(X(t)^2-2\sum_{i=1}^{n}\xi_i\phi_i(t)X(t)+\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j\phi_i(t)\phi_j(t))\\
=&E(X(t)^2)-2\sum_{i=1}^{n}\phi_i(t)E(X(t)\xi_i)+\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_i(t)\phi_j(t)E(\xi_i\xi_j)\\
=&E(X(t)^2)-2\sum_{i=1}^{n}\phi_i(t)\lambda_i\phi_i(t)+\sum_{i=1}^{n}\lambda_i\phi_i^2(t)\\
=&E(X(t)^2)-\sum_{i=1}^{n}\lambda_i\phi_i^2(t)\to 0,\quad n\to\infty,
\end{aligned}
$$

uniformly by Mercer's theorem (there exist $\phi_j(s)$ and λ_j such that $E(X(t)^2) = \sum_{i=1}^{\infty}\lambda_i\phi_i(t)\phi_i(t)$). Then, $X(t)=\sum_{i=1}^{\infty}\xi_i\phi_i(t)$.

(b) Conversely, if $X(t)=\sum_{i=1}^{\infty}\xi_i\phi_i(t)$, then

$$
\begin{aligned}
K(s,t)=&E(X(s),X(t))\\
=&\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}\phi_j(s)\phi_i(t)E(\xi_i\xi_j)\\
=&\sum_{j=1}^{\infty}\lambda_j\phi_j(s)\phi_j(t).
\end{aligned}
$$

Hence,

$$\begin{aligned}\int_{\mathcal{T}} K(s,t)\phi_i(t)dt &= \int_{\mathcal{T}} \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)\phi_i(t)dt \\ &= \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \int_{\mathcal{T}} \phi_j(t)\phi_i(t)dt \\ &= \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\delta_{ij} \\ &= \lambda_i \phi_i(s).\end{aligned}$$

□

Remark 1.3 In the Karhunen–Loève expansion, a stochastic process $X(t), t \in \mathcal{T}$ is represented by a sequence of independent simple random variables ξ_i, $i \in \mathcal{N}$. Here, $\phi_i, i = 1, 2, \cdots$ represents a set of orthonormal basis functions, while ξ_i can be regarded as principal component scores.

1.4 Our Objective and the Structure of the Dissertation

1.4.1 The Objective of Our Work

Motivated by practical applications, the primary objective of our study is to present innovative contributions to robust methodology in functional data analysis, encompassing both novel techniques and tools. These advancements are built upon

the principles of traditional robust statistics that have recently been extended to accommodate functional data.

1.4.2 The Structure of Our Work

The focus of our research primarily revolves around the issue of functional linear models with scalar responses. A concise overview of their content is as follows:

- **Chapter 1** provides a concise introduction to FDA. It introduces the concept of functional data and presents several real-world examples, some of which will be suitable for illustrating the methods proposed in subsequent chapters. Additionally, it covers relevant background information such as pre-processing techniques, robust statistical estimation, dimension reduction methods for functional regression, and mathematical concepts in FDA.

- **Chapter 2** provides a comprehensive overview of the most widely used basis expansion methods in functional data analysis (FDA), and conducts a comparative analysis of these methods using some simulated functional linear models.

- **Chapter 3** primarily focuses on the problem of robust estimation in partial functional linear models within the framework of reproducing kernel Hilbert spaces (RKHS, hereinafter). We investigate the theoretical properties of robust estimation for a partially functional linear regression model that incorporates

both functional predictors and multivariate predictors. Moreover, our simulation studies compare the performance of classic and robust procedures under three different contamination schemes. Furthermore, we present two real data examples to illustrate the effectiveness of the robust procedure.

- **Chapter 4** addresses the issue of robust hypothesis testing in functional linear regression, focusing on extending three robust tests: Wald-type, likelihood ratio-type, and F-type to functional linear models with a scalar dependent variable and a functional covariate. We thoroughly investigate the theoretical properties of these robust testing procedures and evaluate their finite sample properties through numerical simulations.

- **Chapter 5** investigates the robust variable selection method in a multiple functional linear regression. We propose a robust group variance inflation factor (VIF) procedure. A novel selection algorithm based on α investing rule are presented. Our methodology has been rigorously validated through some simulation studies and its application to a real-world data.

- **Chapter 6** provides a comprehensive overview of the principal findings of this dissertation, along with posing some unresolved questions that necessitate further investigation.

Finally, we list some notations used throughout the rest of this dissertation. $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are the $L^2(\mathcal{T})$ inner product and norm, respectively. $\otimes$ denotes the tensor product operator of two norm space. For two positive sequences a_n and b_n, $a_n \asymp b_n$ means that a_n/b_n is bounded away from 0 and ∞ as $n \to \infty$. $\mathbb{1}(\cdot)$ is an indicator function, and "$\xrightarrow{d}$" represents convergence in distribution.

2 Basis Expansion Methods in Functional Data Analysis

2.1 Representing Functional Data

The representation of data plays a crucial role in the initial stage of functional data analysis, where a suitable basis is chosen to represent discrete data in a functional manner. Therefore, the selection of appropriate basis functions assumes paramount importance in conducting accurate and reliable functional data analysis. These basis functions should closely resemble the characteristics of the real data, enabling an precise representation of the function with minimal reliance on numerous basis terms. (D. B. Clarkson and Ramsay, 2005; Ramsay and Silverman, 2005, 2007)

Typically, we are provided with paired data (y_{ij}, t_{ij}) $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, N)$, where n represents the number of observations and N denotes the number of discrete gird points. If it is evident that these data exhibit a functional structure, our objective involves the estimation of the latent function $x_i(t)$. For example, considering

only one record ($j = 1, 2, \cdots, n$) as $y_j = x(t_j) + \epsilon_j$, if the data (y_j, t_j) have the functional structure:

$$y_j = c_1 + c_2 t_j + c_3 {t_j}^2 + c_4 {t_j}^3 + \ldots + \epsilon_j,$$

which implies the latent function $x(t) = c_1 + c_2 t + c_3 t^2 + c_4 t^3 + \ldots$.

For functional data analysis, our primary goal is to represent data recorded as a continuous function via basis expansion

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{\Phi}(t)\mathbf{c},$$

where $\mathbf{c}$ is K-vector of coefficients and $\mathbf{\Phi}$ is K-vector of basis functions.

2.2 Representation via Basis Expansion

The utilization of dimension reduction techniques is imperative in the context of functional data, given its infinite dimensionality. These techniques serve as indispensable tools for this purpose. A commonly employed approach involves expanding random functions into suitable basis functions to effectively reduce the dimensions of functional coefficients and predictors. Consequently, selecting an appropriate basis system for observed data becomes a critical initial decision that must be made. However, the current state of research lacks an automated method for determining the optimal basis system based on the given data; thus, it becomes necessary to select a

specific basis system depending on distinct characteristics of the data at hand. For instance, Fourier basis can be employed for periodic data, B-splines basis for non-periodic data, and wavelets basis for capturing discontinuous or rapidly changing behavior in the data. In this section, we will provide a comprehensive summary of these popular methods used for basis expansion.

2.2.1 Spline Method

The most popular functional basis expansion method is using polynomial segments jointed end to end, which is spline functions. Spline functions are the most common choice of approximation system for non-periodic functional data or parameters. Assume the coefficient function can be expanded by a B-spline basis. Let $\mathbf{B}(t) = (B_1(t), B_2(t), \cdots, B_{k_n+1+l}(t))^T$ be a set of B-spline basis function of order $l+1$ with k_n knots. Then, the coefficient function $\beta(t)$ can be approximated by

$$\beta(t) \approx \mathbf{b}^T \mathbf{B}(t).$$

If we assume only k_n basis functions affect the shape of the estimate, then $\beta(t) = \sum_{k=1}^{k_n} b_k B_k(t) + \sum_{k=k_n+1}^{\infty} b_k B_k(t) =: \sum_{k=1}^{k_n} b_k B_k(t) + \delta(t)$, where $\delta(t)$ is the truncation error

and the parameter k_n can be viewed as a tuning parameter.

$$\begin{aligned}\int_{\mathcal{T}} \beta(t)X_i(t) &= \sum_{k=1}^{k_n} b_k \int_{\mathcal{T}} B_k(t)X_i(t)dt + \int_{\mathcal{T}} \delta(t)X_i + \epsilon_i \\ &= \sum_{k=1}^{k_n} b_k\xi_{ik} + +\delta_i + \epsilon_i,\end{aligned}$$

where $\xi_{ik} = \langle B_k, X_i(t)\rangle = \int_{\mathcal{T}} B_k(t)X_i(t)dt$ can be viewed as components of X_i on the basis. Note that it is necessary to assume that $k_n \to \infty$ as $n \to \infty$, so that we can obtain an asymptotically consistent estimation of $\beta(t)$.

The parameter k_n in the previous section could be viewed as a tuning parameter, adjusting k_n adjusts the smoothness of the resulting estimator of $\beta(t)$. From this perspective, it is often more desirable to smooth $\beta(t)$ by using a roughness penalty term. The roughness penalty approach transfers the control of smoothness from k_n to the smoothing parameter λ and a differential operator L,

$$P_\lambda(\alpha, \beta) = \sum_{i=1}^{n}\{Y_i - \alpha - \int_{\mathcal{T}} \beta(t)X_i(t)dt\}^2 + \lambda \int_{\mathcal{T}} \left[(L\beta)(t)\right]^2 dt.$$

The objective is to enforce smoothness by penalizing excessively rough functions with a penalty term $\lambda \int_{\mathcal{T}} \left[(L\beta)(t)\right]^2 dt$. A commonly employed option for L is the second derivative, denoted as $(L\beta)(t) = \beta''(t)$, which is commonly referred to as a smooth penalty. Various approaches for selecting the smoothing parameter λ include cross validation (CV), information criteria, and restricted maximum likelihood (REML). The B-spline bases of order 4 are illustrated in Figure 2.1-2.3, showcasing

cases with 5, 10, and 20 bases respectively.

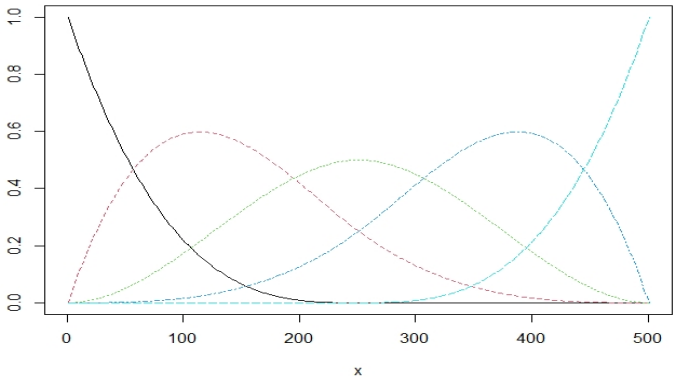

Figure 2.1: 5 B-spline bases of order 4

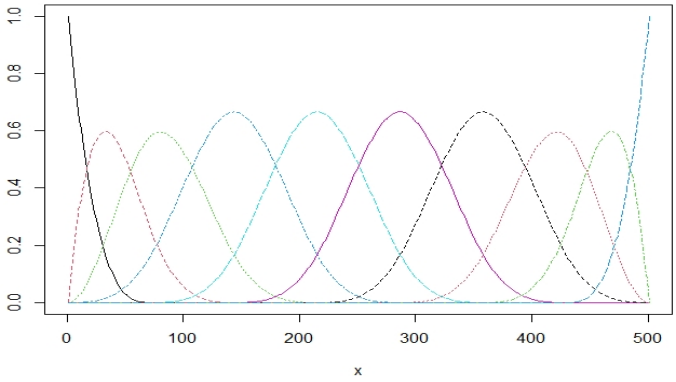

Figure 2.2: 10 B-spline bases of order 4

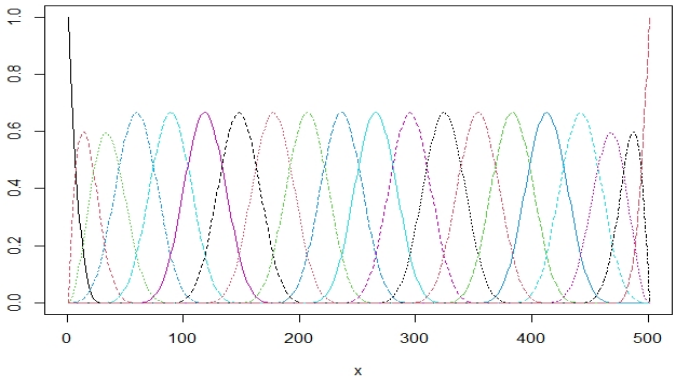

Figure 2.3: 20 B-spline bases of order 4

2.2.2 FPCA Method

The utilization of a data-driven basis provides enhanced flexibility, allowing us to represent functions as principal components analysis basis functions across various closed intervals $\mathcal{T}_j$. Without loss of generality, we assume that the mean function of $X(t)$ is $\mathrm{E}X(t) = 0$ and the covariance function is $\Sigma(s,t) = \mathrm{Cov}(X(s), X(t))$. Then by Mercer's Theorem, we can obtain the spectral decomposition

$$\Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t),$$

where λ_k is the eigenvalues with non-increasing order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$, $\sum_{k=1}^{\infty} \lambda_k < \infty$, and ϕ_k's are the corresponding orthonormal eigenfunctions. Use the Karhunen-Loève

representation, we can obtain

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

where $\{\xi_k = \int_{\mathcal{T}} X_(t)\phi_k(t)dt, k \geqslant 1\}$ are principal component scores with $\mathrm{E}(\xi_k) = 0, \mathrm{Var}(\xi_k) = \mathrm{E}(\xi_k^2) = \lambda_k$, and $\mathrm{E}(\xi_k \xi_{k'}) = 0$, for $k \neq k'$. Based on the same orthonormal basis functions $\{\phi_k(t), k = 1, 2, \ldots\}$, the coefficient function can be expanded by $\beta(t) = \sum_{k=1}^{\infty} b_k \phi_k(t)$. Then, a scalar-on-function linear model can be expressed as the following form:

$$Y = \alpha + \sum_{k=1}^{\infty} b_k \int_{\mathcal{T}} X(t)\phi_k(t)dt + \varepsilon = \alpha + \sum_{k=1}^{\infty} \xi_k b_k + \varepsilon. \tag{2.1}$$

To deal with the problem of infinite sum in the model 2.1, we approximate the model via a truncating parameters k_n. Then the model 2.1 can be approximated by

$$Y = \alpha + \sum_{k=1}^{k_n} \xi_k \beta_k + \varepsilon. \tag{2.2}$$

Remark 2.1 The choice of truncated value k_n in very important in FPCA method. In practice, there are some empirical choice of this value, such as PVE (Percentage of Variance Explained) method, leading PCs (Principal Components) method (Cardot et al., 2003; Kong et al., 2013; Swihart et al., 2014), CV (Cross-Validation) criterion (Qingguo, 2017), and information (AIC or BIC) criterion (Kato et al., 2012).

2.2.3 RKHS Method

Cai and Yuan (2012) highlighted potential issues among with the effectiveness of eigenfunction basis and the selection of truncation number of eigenvalue in the FPCA approach. Therefore, they discussed the functional linear regression problem within the framework of RKHS, where functional data are considered as realizations of random variables that take values in an RKHS. In this context, if we consider the slope function $\beta(t)$ to be in an RKHS $\mathcal{H}$, which is a subspace of the Hilbert space consisting of square integrable functions with reproducing kernel K defined on a compact set $\mathcal{T} \subset \mathbb{R}$. Without loss of generality, $\mathcal{T} = [0, 1]$, $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a real, symmetric, square integrable, and non-negative defined function and the canonical example of $\mathcal{H}$ is the Sobolev spaces. Yuan and Cai (2010) also defined the penalized least-squares approach to functional linear regression with taking the second order Sobolev space $\mathcal{H} = W_2^2([0, 1]) = \{\beta : [0, 1] \to \mathbb{R} | \beta, \beta' \text{are absolutely continuous and } \beta'' \in \mathcal{L}_2\}$ and the penalty function $J(\beta) = \int_{\mathcal{T}} [\beta''(t)]^2 dt$. Thus, the penalized least squares (PLS) criterion is

$$\frac{1}{n} \sum_{i=1}^{n} [Y_i - \alpha - \int_{\mathcal{T}} X_i(t)\beta(t)dt]^2 + \lambda J(\beta). \tag{2.3}$$

Yuan and Cai (2010) demonstrated that the minimization of equation (2.3) is well-defined and can be easily computed using the representer theorem (Wahba, 1990).

More details can be found in Yuan and Cai (2010) and Cai and Yuan (2012).

2.2.4 Wavelet Method

It is well known that the majority of applications of wavelets in statistical data analysis are in the area of nonlinear regression and function estimation. Wavelets, referred to as "small waves", are mathematical functions that satisfy specific criteria. The term wavelet originates from the condition of their integral being zero, oscillating both above and below the x-axis. Similar to sines and cosines in Fourier analysis, wavelets serve as fundamental units for representing other functions. By fixing the mother wavelet $\psi(t)$, a family of wavelets can be generated through translations and dilations $\psi(\frac{t-b}{a})$. For a comprehensive understanding of wavelet theory and its statistical applications, we refer readers to the books by Benedetto (1993) and Nason (2008).

Definition 2.1 A wavelet system in $\mathcal{L}^2(\mathbb{R})$ is a collection of functions of the form

$$\phi_k(x) = \phi(x-k), \quad \psi_{k,j}(x) = 2^{j/2}\psi(2^j x - k),$$

where $j = 0, 1, 2, \cdots$, $k = 0, \pm 1, \pm 2, \cdots$, and the two functions $\psi, \phi \in \mathcal{L}^2(\mathbb{R})$ have compact supports. $\psi(x)$ denotes the wavelet function (also called the mother function or primary wavelet), $\phi(x)$ denotes a scaling function (also called the father wavelet).

Various types of wavelets exist, including smooth wavelets, compactly supported wavelets, mathematically simple wavelets, and wavelets with short associated filters. We introduce three popular wavelets, which are *Haar wavelet* , *Shannon wavelet* and *Mexican-hat wavelet*. Details can be found in Morettin et al. (2017).

✦ **Haar Wavelet (Daubechies wavelet, order=1)**

The Haar mother wavelet is a mathematical function defined by

$$\psi(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}), \\ -1, & x \in [\frac{1}{2}, 1), \\ 0, & \text{otherwise.} \end{cases}$$

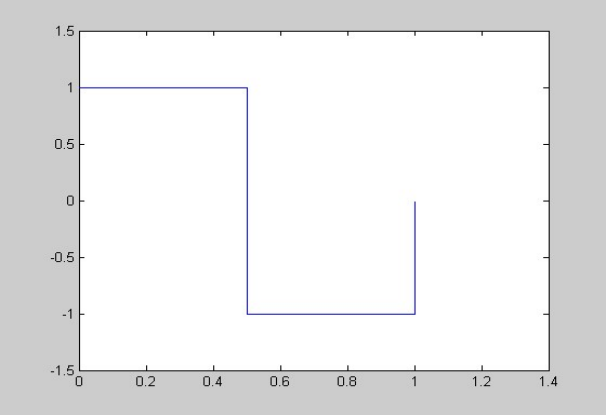

Figure 2.4: Haar Wavelet

✦ **Shannon Wavelet**

The Shannon mother wavelet is a mathematical function defined by

$$\psi(x) = \frac{\sin 2\pi x - \cos \pi x}{\pi(x - \frac{1}{2})}.$$

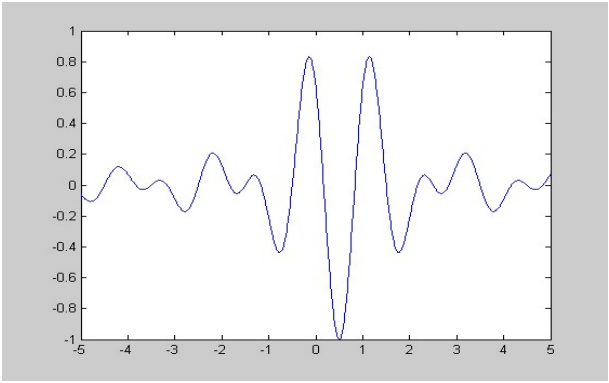

Figure 2.5: Shannon Wavelet

✦ **Mexican-hat wavelet (Ricker wavelet)**

The Ricker wavelet, also known as the Mexican-hat wavelet, is derived from the negative normalized second derivative of a Gaussian function and can be defined as follows:

$$\psi(x) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{x^2}{\sigma^2}\right)e^{-\frac{x^2}{2\sigma^2}}.$$

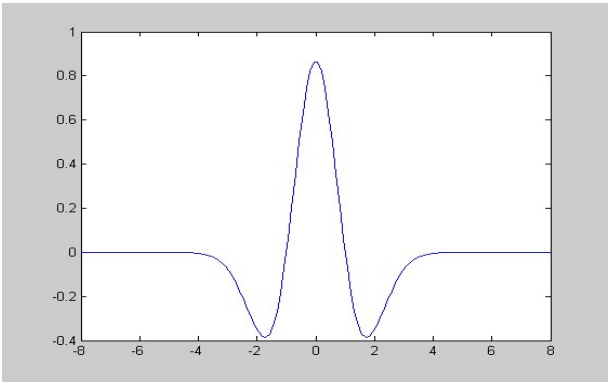

Figure 2.6: Mexican-hat Wavelet

2.2.5 Comparative Analysis

The functional basis expansion methods will be compared in this section through a series of simulated examples. Our experiments focus on the scalar-on-function model, as described below:

$$Y_i = \alpha + \int_0^1 \beta(t) X_i(t) dt + \varepsilon_i,$$

where $\varepsilon_i \sim N(0,1),\ i = 1, 2, \cdots, n$.

The majority of the examples presented below are derived from various research literature on functional data analysis, and we provide a concise summary of these illustrative instances. Examples 2.1-2.3 generate samples based on Fourier basis functions, while Examples 2.4-2.5 utilize power functions and Examples 2.6 employ Gaussian basis functions to generate samples. Coefficient functions were estimated using various methods including B-spline, FPCA, RKHS, and wavelet-based methods. It is worth noting that in the RKHS approach, our kernel function is defined as $K(s,t) = \sum_{k=1}^{50} \frac{2}{(k\pi)^4} \cos(k\pi s)\cos(k\pi t)$ with $\lambda = 10^{-6}$. In wavelet method, we utilize Daubechies 10 (db10) wavelet as shown in Figure 2.7.

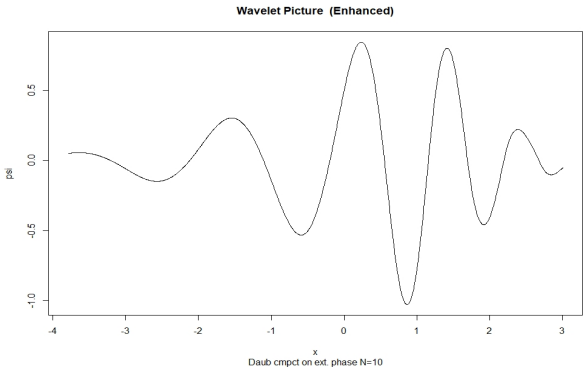

Figure 2.7: Daubechies 10 (db10) wavelet function.

Example 2.1

In the first example, the setting details of $X(t)$ and $\beta(t)$ are presented as follows (Qingguo, 2017; Su et al., 2017). We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.8-2.10.

- $\phi_k(t), k = 1, 2, \cdots, 5$ are the five Fourier basis functions, which are $\phi_1(t) = 1, \phi_2(t) = \sqrt{2}\sin(2\pi t), \phi_3(t) = \sqrt{2}\cos(2\pi t), \phi_4(t) = \sqrt{2}\sin(4\pi t)$, and $\phi_5(t) = \sqrt{2}\cos(4\pi t)$;

- $X_i(t) = \sum_{k=1}^{5} \xi_{ik}\phi_k(t)$, where $\xi_{ik} \sim N(0, \lambda_k)$ with $\lambda_k = k^{-a}$, $a = 1.1$, $k = 1, 2, \cdots, 5$;

- $\beta(t) = \sum_{k=1}^{5} b_k\phi_k(t)$.

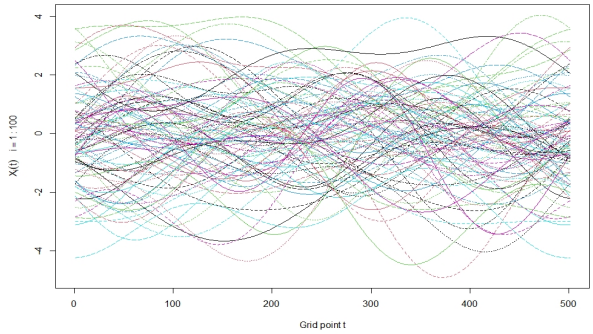

Figure 2.8: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.1.

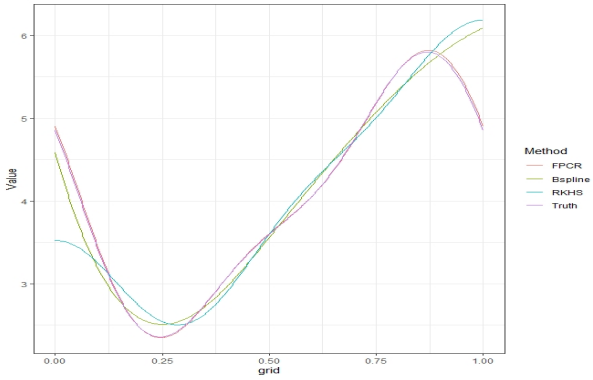

Figure 2.9: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.1.

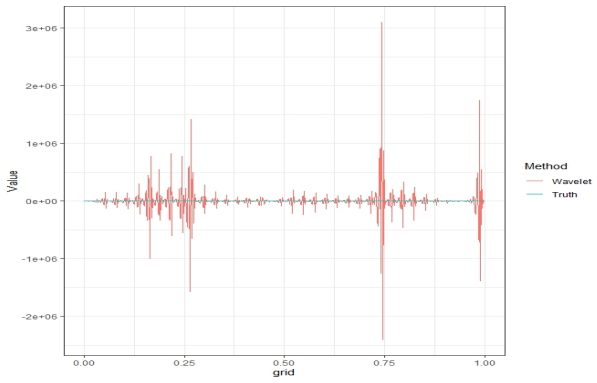

Figure 2.10: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.1.

Example 2.2 In the second example, the setting details for $X(t)$ and $\beta(t)$ are presented as follows (Hall et al., 2007; Yuan and Cai, 2010). We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.11-2.13.

- $\phi_1(t) = 1, \phi_k = \sqrt{2}\cos(k\pi t),\ k = 2, \cdots, 50$ are the basis functions;
- $X(t) = \sum_{k=1}^{50} k^{-v} U_k \phi_k(t)$, where $U_k \sim U(-\sqrt{3}, \sqrt{3})$. For these coefficients, the eigenvalues of the covariance function are k^{-2v}. We take $v = 0.6$ to regulate the decaying rate of eigenvalues;
- $\beta(t) = \sum_{k=1}^{50} b_k \phi_k(t)$, where $b_1 = 0.5$, and $b_k = 4(-1)^{k-1}k^{-2},\ k > 1$.

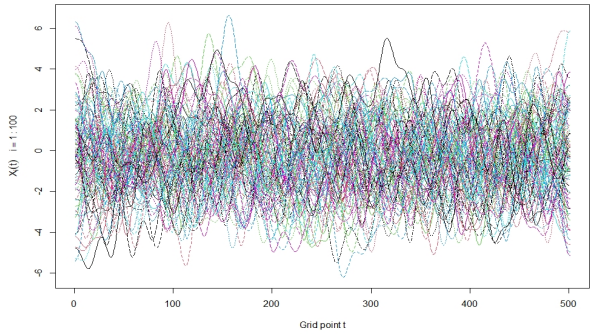

Figure 2.11: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.2.

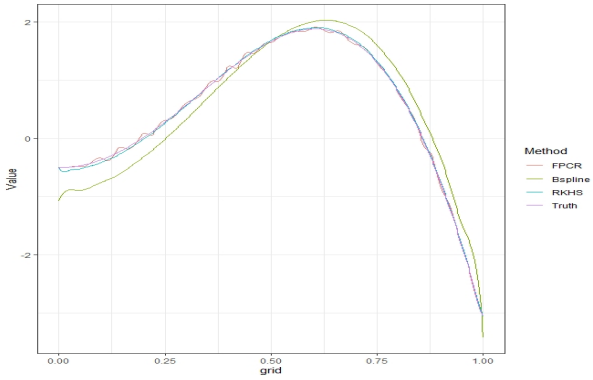

Figure 2.12: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.2.

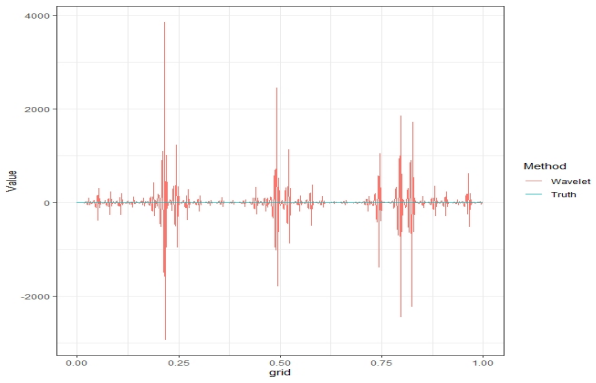

Figure 2.13: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.2.

Example 2.3 In the third example, the setting details of $X(t)$ and $\beta(t)$ are presented as follows:

- $\phi_1(t) = 1, \phi_k = \sqrt{2}\cos(k\pi t),\ k = 2, \cdots, 50$, are the basis functions for the predictors;
- $\psi_1(t) = 1, \psi_k = \sqrt{2}\sin(k\pi t), k = 2, \cdots, 50$, are the basis functions for the slope function;
- $X_i(t) = \sum_{k=1}^{50} k^{-v} U_k \phi_k(t),\ v = 0.6$, where $U_k \sim U(-\sqrt{3}, \sqrt{3})$;
- $\beta(t) = \sum_{k=1}^{50} b_k \psi_k(t)$, where $b_1 = 0.5$, and $b_k = 4(-1)^{k-1}k^{-2},\ k > 1$.

We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.14-2.16.

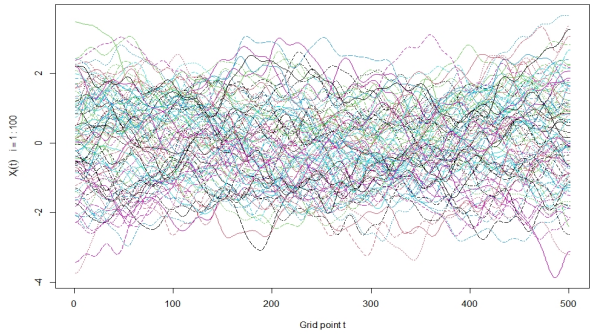

Figure 2.14: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.3.

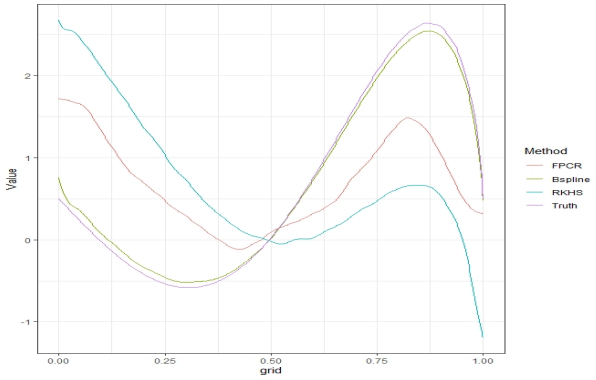

Figure 2.15: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.3.

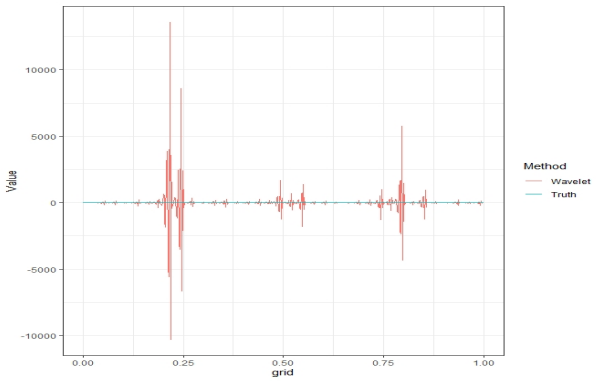

Figure 2.16: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.3.

Example 2.4 In the fourth example, the setting details of $X(t)$ and $\beta(t)$ are presented as follows:

- $\phi_1(t) = 1, \phi_k = t^{k-1},\ k = 2, \cdots, 5$, are the basis functions for the predictors and the slope function;
- $X_i(t) = \sum_{k=1}^{5} k^{-v} U_k \phi_k,\ v = 1.1$, where $U_k \sim U(-\sqrt{3}, \sqrt{3})$;
- $\beta(t) = \sum_{k=1}^{5} b_k \phi_k(t)$, where $b_1 = 0.5$, and $b_k = 4(-1)^{k+1} k^{-2},\ k > 1$.

We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.17-2.19.

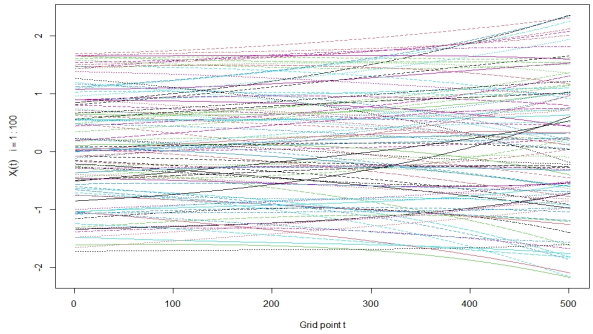

Figure 2.17: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.4.

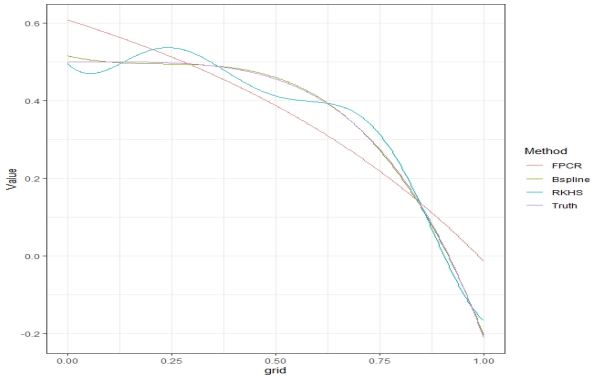

Figure 2.18: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.4.

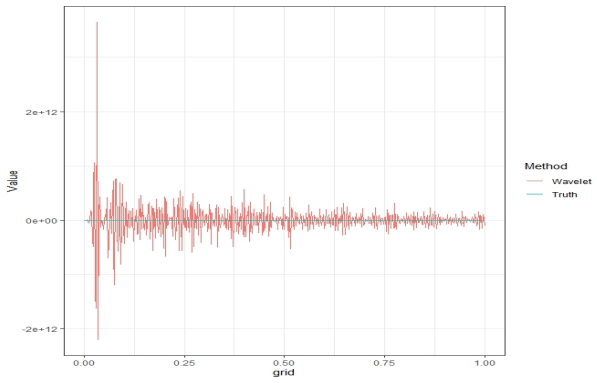

Figure 2.19: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.4.

Example 2.5 In the fifth example, the setting details of $X(t)$ and $\beta(t)$ are presented as follows:

- $\phi_1(t) = 1, \phi_k = t^{k-1},\ k = 2, \cdots, 5$ are the basis functions for the predictors;
- $\psi_1(t) = 1, \psi_k = \sqrt{2}\cos(k\pi t),\ k = 2, \cdots, 5$ are the basis functions for the slope function;
- $X_i(t) = \sum_{k=1}^{5} k^{-v} U_k \phi_k,\ v = 1.1$, where $U_k \sim U(-\sqrt{3}, \sqrt{3})$;
- $\beta(t) = \sum_{k=1}^{5} b_k \psi_k(t)$, where $b_1 = 0.5$, and $b_k = 4(-1)^{k+1} k^{-2},\ k > 1$.

We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.20-2.22.

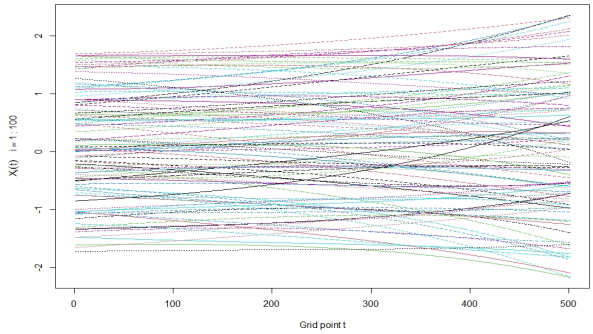

Figure 2.20: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.5.

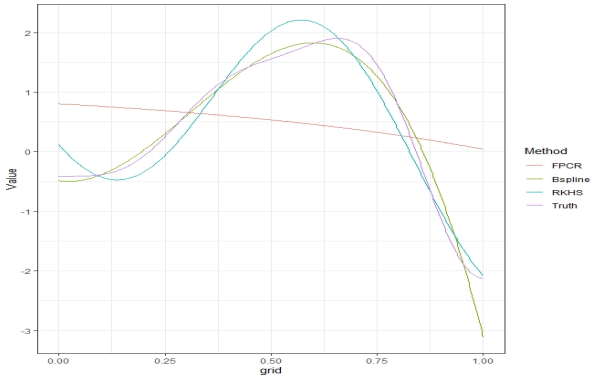

Figure 2.21: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.5.

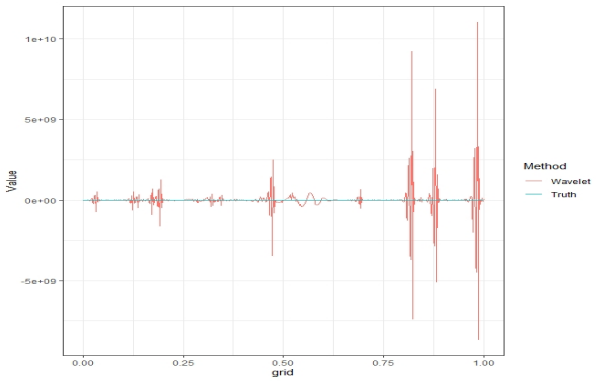

Figure 2.22: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.5.

Example 2.6 In the final example, Gaussian functions are utilized as the basis functions for the predictor, and their corresponding curves are illustrated in Figure 2.23.

- $\phi_k = \exp\{-(t-\mu_k)^2/2\sigma^2\}$ are the basis functions for the predictors, where $\sigma = 0.1$ and $\mu_k = (k-1)\sigma$ for $k = 1, 2, \cdots, 11$;
- $\psi_1(t) = 1, \psi_k = \sqrt{2}\cos(k\pi t),\ k = 2, \cdots, 11$ are Fourier basis functions for the slope function;
- $\alpha = 2,\ X_i(t) = \sum_{k=1}^{11} k^{-v} U_k \phi_k,\ v = 1.1$, where $U_k \sim U(-\sqrt{3}, \sqrt{3})$;
- $\beta(t) = \sum_{k=1}^{11} b_k \psi_k(t)$, where $b_1 = 0.5$, and $b_k = 4(-1)^{k+1}k^{-2},\ k > 1$.

We illustrate several trajectories of covariate functions $X_i(t)$ along with the estimation of the coefficient function $\beta(t)$ in Figure 2.24-2.26.

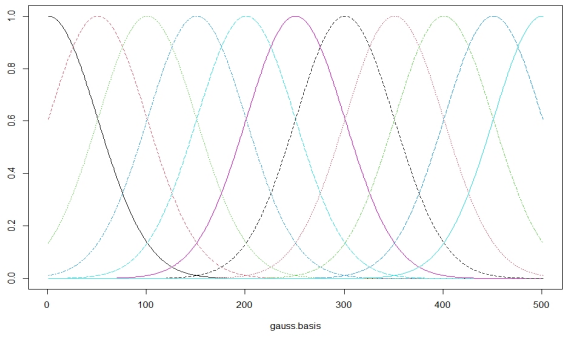

Figure 2.23: Gaussian basis functions, $\sigma = 0.1$.

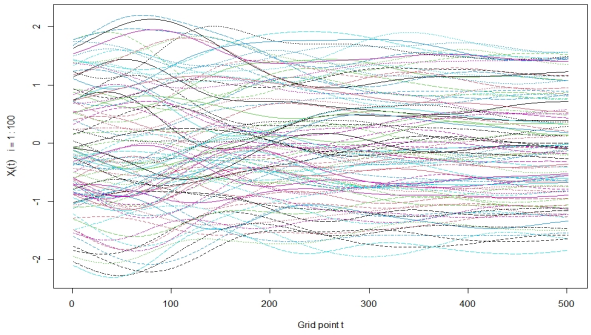

Figure 2.24: Some trajectories of covariate functions $X_i(t)$, $i = 1, \cdots, 100$, in Example 2.6.

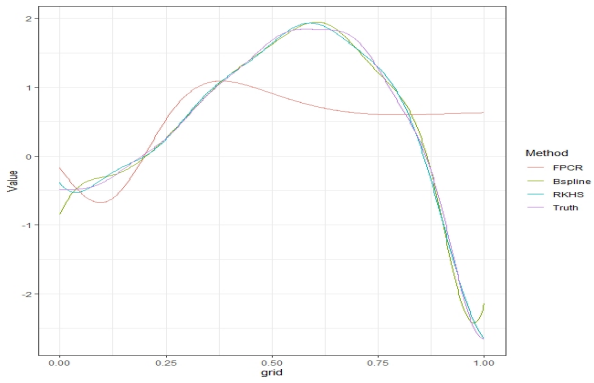

Figure 2.25: The estimated coefficient function $\hat{\beta}(t)$ based on the different approaches in Example 2.6.

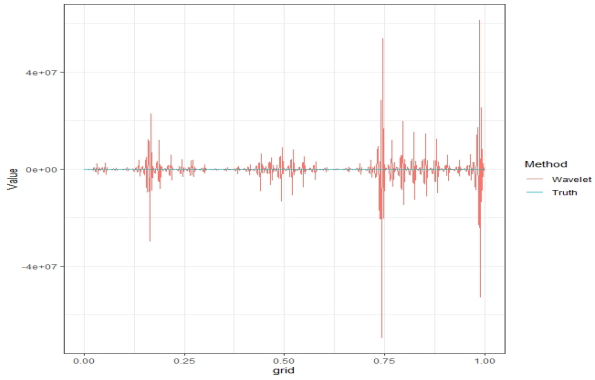

Figure 2.26: Wavelet transform of the estimated coefficient $\hat{\beta}(t)$ and true coefficient $\beta(t)$ in Example 2.6.

3 Robust Estimation for Partially Functional Linear Regression under an RKHS Framework

3.1 Introduction

In recent decades, advancements in technology have facilitated the recording of increasingly intricate high-dimensional data. Some of these data exhibit a functional structure and are commonly referred to as functional data, such as Diffusion Tensor data (Goldsmith et al., 2012; Kong et al., 2013; Su et al., 2017), EEG and EMG recording data (Rügamer et al., 2018), and and the renowned Canadian Weather dataset (Ramsay and Silverman, 2005). A comprehensive overview of various methodologies and applications, encompassing both parametric and nonparametric approaches, can be found in notable works by Ramsay and Silverman (2005), Ramsay and Silverman (2007), Ferraty and Vieu (2006). Functional linear regression is one popular analysis tool that explores the linear impact of functional predictors on scalar or function responses, with coefficients typically represented as functions.

For an extensive review of regression methodologies, theoretical properties, and computational techniques, refer to Morris (2015) and Wang et al. (2016).

However, in practice, it is expected to collect information that simultaneously contains functional and nonfunctional data. Therefore, to enhance the interpretation of the functional regression model in datasets with mixed data, one can incorporate a finite-dimensional vector-valued random variable into the function-valued random variable within the model. This approach is commonly known as the partially functional linear model (PFLM, hereinafter) (Shin, 2009). Recently, several research studies have been conducted to address the theoretical and practical concerns associated with this model. For instance, Kong et al. (2016b) proposed a method that simultaneously achieves variable selection and estimation for PFLM, establishing the consistency and oracle properties of their approach. Cai et al. (2020) considered a robust estimation procedure for PFLM using a modified Huber's function based on B-spline basis expansion. Yu et al. (2020) presented a robust estimation procedure for PFLM by employing modal regression to fit the slope function with B-splines, obtaining convergence rates and asymptotic normality of the estimators.

Similar to ordinary linear regression, the primary challenge in PFLM lies in estimating coefficients. The predominant approach for addressing this issue is functional principal component analysis, in which eigenfunctions of the covariance function of

the random predictor are utilized as basis functions to represent the functional predictor and the unknown coefficients. However, as highlighted by Cai and Yuan (2012), the FPCA approach has potential concerns regarding the effectiveness of eigenfunction basis and selection of truncation number for eigenvalues. The simulation results presented in Cai and Yuan (2012) demonstrate that the reproducing kernel Hilbert spaces approach outperforms the FPCA method in cases where there is imperfect alignment between the reproducing kernel function K and the covariance function C of the predictor $X(t)$. Consequently, recent research endeavors have increasingly focused on functional linear regression within RKHS framework, considering functional data as realizations of random variables that assume values in an RKHS (Yuan and Cai, 2010; Shin and Lee, 2016; Sun et al., 2018; Cui et al., 2020).

Motivated by the works of Shin and Lee (2016) and Cui et al. (2020), our objective in this study is to investigate penalized M-estimation for the partially functional linear model within the framework of RKHS. Specifically, we extend the application of RKHS approach to partially functional linear regression associated with robust M-estimation. In this process, we utilize well-established families of robust loss functions and incorporate a preliminary estimator for residual scale. These estimations are both scale equivariant and robust against high-leverage outliers. Our primary contribution lies in achieving simultaneous robust estimation for both the

functional coefficient $\beta(t)$ and the multivariate coefficient $\boldsymbol{\theta}$ within the framework of RKHS. Theoretically, we achieve a convergence rate for prediction errors as well as the asymptotic normality in estimating $\boldsymbol{\theta}$ under certain regularized conditions. In terms of computation, we employ an iteratively re-weighted least-squares method for parameter estimation and devise an efficient algorithm to implement our robust procedure. Simulation studies and two real data applications demonstrate the superior performance of our robust approach.

The subsequent sections of this chapter are organized as follows. Section 3.2 presents a comprehensive description of the partially functional linear model and introduces the penalized robust M-estimation procedure within the framework RKHS. In Section 3.3, we delve into the asymptotic normality of the multivariate linear component and discuss the convergence rate of prediction errors for the functional part. Simulation studies and real data examples are showcased in Section 3.4-3.5. Finally, our conclusions, along with all proofs and additional simulation results are presented in Section 3.6 and 3.7, respectively.

3.2 Model and Estimation

3.2.1 Robust Partially Functional Linear Regression

We consider the following partially functional linear regression

$$Y = \alpha + \boldsymbol{z}^T\boldsymbol{\theta} + \int_{\mathcal{T}} \beta(t)X(t)dt + \varepsilon, \tag{3.1}$$

where α is the intercept term, $X(t)$ is functional predictor with a compact set $\mathcal{T} \subset \mathbb{R}$, $\boldsymbol{z} = (z_1, z_2, \cdots, z_p)^T$ is the p-dimensional vector predictor in addition to the functional predictor $X(t)$, and Y is the scalar response random variable defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is an unknown p-dimensional parameter vector, $\beta(t)$ is a square integrable coefficient function on $\mathcal{T}$, and ε with $\mathrm{E}\varepsilon = 0$ and $\mathrm{E}\varepsilon^2 = \sigma_\varepsilon^2$ is a random error independent of $X(t)$ and $\boldsymbol{z}$. Without loss of generality, we assume that $\mathcal{T} = [0, 1]$ throughout the chapter, and $\mathrm{E}[X(t)] = 0$, $\mathrm{E}X(t)^2 < \infty$ for all $t \in \mathcal{T}$.

Given independently and identically distributed data $(x_i, y_i, \boldsymbol{z}_i)$, $i = 1, 2, \ldots, n$, we denote the underlying true parameters by α_0, β_0 and $\boldsymbol{\theta}_0$. Our goal is to investigate the asymptotic properties of M-estimators for the coefficient function β as well as the coefficient $\boldsymbol{\theta}$ in multivariate part under the RKHS framework. Suppose that the coefficient function $\beta_0(t)$ is in a Hilbert space $\mathcal{H}$ with the reproducing kernel K, where $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a real, symmetric, square integrable, and nonnegative definite

function. And K is also a nonnegative definite operator on $\mathcal{L}_2$, which implies that $Kf(\cdot) = \int_{\mathcal{T}} K(s,\cdot)f(s)ds$. Using the regularization method, we define the robust M-estimation of $(\alpha_0, \beta_0, \boldsymbol{\theta}_0)$ by solving the following optimization problem

$$\min_{\alpha,\beta,\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \rho\Big(\frac{y_i - \alpha - \boldsymbol{z}_i^T\boldsymbol{\theta} - \int_{\mathcal{T}} \beta(t)x_i(t)dt}{\hat{\sigma}_\varepsilon}\Big) + n\lambda J(\beta) \right], \tag{3.2}$$

where ρ is a loss function, $\hat{\sigma}_\varepsilon$ is a preliminary residual scale estimate, $J(\beta)$ is a penalty function on β, and $\lambda > 0$ is a regularization parameter. A typical choice of the penalty function is $J(\beta) = \int_{\mathcal{T}}[\beta''(t)]^2 dt$ to penalize the roughness of $\beta(t)$. Furthermore, our M-estimation of coefficients is scale equivariant since we consider a preliminary residual scale estimate. The popular robust scale estimate is the normalized median absolute deviation (MAD) (Yohai and Maronna, 1979) and can be computed through the residuals from the initial fit, i.e., $\hat{\sigma}_\varepsilon = 1.483 \cdot (\text{med}|r_i - \text{med}(r_i)|)$.

3.2.2 Computation Details

First, we define $\mathcal{H} = W_2^2[0,1] = \{\beta(t) : [0,1] \to \mathbb{R} | \beta, \beta' \text{ are absolutely continuous and } \beta'' \in \mathcal{L}_2\}$, which is the Sobolev space of order 2 (Yuan and Cai, 2010). Then, based on the representer theorem for smooth splines (Wahba, 1990), it suffices to consider the form $\beta(t) = d_1 + d_2 t + \sum_{i=1}^{n} c_i \int_{\mathcal{T}} x_i(s)K(s,t)ds = d_1 + d_2 t + \sum_{i=1}^{n} c_i\xi_i$, with $\xi_i = \int_{\mathcal{T}} x_i(s)K(s,t)ds$ and some $d_1, d_2 \in \mathbb{R}$, $c_i \in \mathbb{R}$, $i = 1, 2, \cdots, n$. Hence, we obtain $\int_{\mathcal{T}}[\beta''(t)]^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \xi_i, \xi_j \rangle_{\mathcal{H}}$ with $\langle \xi_i, \xi_j \rangle_{\mathcal{H}} = \int_{\mathcal{T}}\int_{\mathcal{T}} x_i(s)K(s,t)x_i(t)dsdt$,

and $\int_{\mathcal{T}} \beta(t)x_i(t)dt = \sum_{l=1}^{2} d_l\langle x_i, \vartheta_l\rangle + \sum_{i=1}^{n} c_i\langle \xi_i, \xi_j\rangle_{\mathcal{H}}$ with $\vartheta_1(t) = 1$ and $\vartheta_2(t) = t$. The minimization problem in (2.2) can be equivalently reformulated as follows:

$$\begin{aligned}\arg\min_{\alpha,\beta,\boldsymbol{\theta}} \sum_{i=1}^{n} \rho\Big(\frac{y_i - \alpha - \boldsymbol{z}_i^T\boldsymbol{\theta} - \sum_{l=1}^{2} d_l\langle x_i, \vartheta_l\rangle - \sum_{i=1}^{n} c_i\langle \xi_i, \xi_j\rangle_{\mathcal{H}}}{\hat{\sigma}_\varepsilon}\Big) \\ + n\lambda \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \xi_i, \xi_j\rangle_{\mathcal{H}}.\end{aligned} \tag{3.3}$$

The minimizer of the optimization problem (3.3) can be obtained through an iteratively re-weighted least squares (IRWLS) algorithm, similar to the approach proposed by Shin et al. (2016). Let $Y = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{Z} = (\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_n^T)^T$. The penalized M-estimation criterion of (3.3) can be reformulated in matrix form as follows:

$$\rho\left(\frac{Y - \alpha\mathbf{1} - \boldsymbol{Z}\boldsymbol{\theta} - T\boldsymbol{d} - \Sigma\boldsymbol{c}}{\hat{\sigma}_\varepsilon}\right) + n\lambda\boldsymbol{c}^T\Sigma\boldsymbol{c}, \tag{3.4}$$

where $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^T$, $\boldsymbol{d} = (d_1, d_2)^T$, $\mathbf{1} = (1, 1, \ldots, 1)^T$ is n-dimensional constant vector, $\Sigma = (\Sigma_{ij})$ is a $n \times n$ matrix with $\Sigma_{ij} = \langle \xi_i, \xi_j\rangle_{\mathcal{H}}$, and $T = (T_{il})$ is an $n \times 2$ matrix with $T_{il} = \langle x_i, \vartheta_l\rangle = \int_{\mathcal{T}} x_i(t)t^{l-1}dt$ for $l = 1, 2$.

The notation $Q = [\mathbf{1}, \boldsymbol{Z}, T]$ and $\boldsymbol{b} = (\alpha, \boldsymbol{\theta}^T, \boldsymbol{d}^T)^T$ are introduced for the sake of notational convenience. Consequently, the aforementioned criterion can be simplified as follows:

$$\rho\left(\frac{Y - Q\boldsymbol{b} - \Sigma\boldsymbol{c}}{\hat{\sigma}_\varepsilon}\right) + n\lambda\boldsymbol{c}^T\Sigma\boldsymbol{c}. \tag{3.5}$$

The solution to minimizing criterion (3.5) can be obtained by taking derivatives with

respect to $\boldsymbol{b}$ and $\boldsymbol{c}$, setting them equal to 0. This is equivalent to determining the minimizer of the penalized weighted least-squares criterion, expressed as:

$$(Y - Q\boldsymbol{b} - \Sigma\boldsymbol{c})^T W (Y - Q\boldsymbol{b} - \Sigma\boldsymbol{c}) + n\lambda \boldsymbol{c}^T \Sigma \boldsymbol{c}, \tag{3.6}$$

where $W = \text{diag}(w_1, w_2, \ldots, w_n)$, and $w_i = w_i(r_i) = \frac{1}{\hat{\sigma}_\varepsilon^2}\psi(r_i)/r_i, \psi = \rho'$ with $r_i = \frac{1}{\hat{\sigma}_\varepsilon}(Y_i - \alpha - \boldsymbol{z}_i^T\boldsymbol{\theta} - \sum_{l=1}^{2} d_l\langle x_i, \vartheta_l\rangle_{\mathcal{L}_2} - \sum_{i=1}^{n} c_i\langle \xi_i, \xi_j\rangle_{\mathcal{H}})$. Then, The minimizer of (3.6) can be determined as follows:

$$\begin{aligned} \hat{\boldsymbol{b}} =& (Q^T M^{-1} Q)^{-1} Q^T M^{-1} Y, \\ \hat{\boldsymbol{c}} =& M^{-1}(I_n - Q(Q^T M^{-1} Q)^{-1} Q^T M^{-1}) Y, \end{aligned} \tag{3.7}$$

with $M = \Sigma + n\lambda W^{-1}$. Note that W is not well defined when some w_is are zeros. For instance, in the case of Tukey's bisquare loss function, there may exist certain values of $\psi(r_i)$ that can be equal to zero for some i. In such cases, we remove the entries or rows corresponding to $\{i : w_i = 0\}$ from Y, Q, Σ and W. Then, the minimizer can be modified by

$$\begin{aligned} \tilde{\boldsymbol{b}} =& (\widetilde{Q}^T \widetilde{M}^{-1} \widetilde{Q})^{-1} \widetilde{Q}^T \widetilde{M}^{-1} \widetilde{Y}, \\ \tilde{\boldsymbol{c}}_2 =& \widetilde{M}^{-1}(I_{n_2} - \widetilde{Q}(\widetilde{Q}^T \widetilde{M}^{-1} \widetilde{Q})^{-1} \widetilde{Q}^T \widetilde{M}^{-1}) \widetilde{Y}, \\ \tilde{\boldsymbol{c}}_1 =& \mathbf{0}_{n_1}. \end{aligned} \tag{3.8}$$

Here, we denote $n = n_1 + n_2$ with $n_1 = \#\{i : w_i = 0\}$ and $n_2 = \#\{i : w_i \neq 0\}$. Subsequently, $\widetilde{Y}, \widetilde{Q}, \widetilde{M}$ (as well as $\widetilde{\Sigma}$ and $\widetilde{W}$) are redefined vector or matrices by excluding the entries or rows corresponding to $\{i : w_i = 0\}$. In the given formula,

$\mathbf{0}_{n_1}$ represents a zero vector of dimension n_1, and $\tilde{\boldsymbol{c}}_1$ denotes the sub-vector of $\tilde{\boldsymbol{c}}$ with entries corresponding to $\{i : w_i = 0\}$. We now briefly describe the algorithm as follows:

(a) Obtain an initial estimate $\widehat{\Theta}^0 = (\hat{\boldsymbol{b}}^0, \hat{\boldsymbol{c}}^0)$ through LS estimator (details in Yuan and Cai (2010)).

(b) For $\widehat{\Theta}^t = (\hat{\boldsymbol{b}}^t, \hat{\boldsymbol{c}}^t)$, compute the residuals r_i^t for $i = 1, 2, \cdots, n$ and the weight W^t, then update the estimate $\widehat{\Theta}^{t+1} = (\hat{\boldsymbol{b}}^{t+1}, \hat{\boldsymbol{c}}^{t+1})$ as follows:

$$\begin{aligned} \hat{\boldsymbol{b}}^{t+1} =& (Q^T(M^t)^{-1}Q)^{-1}Q^T(M^t)^{-1}Y, \\ \hat{\boldsymbol{c}}^{t+1} =& (M^t)^{-1}(I_n - Q(Q^T(M^t)^{-1}Q)^{-1}Q^T(M^t)^{-1})Y. \end{aligned} \tag{3.9}$$

If there exists $w_i = 0$ for some loss function in the t step, the update of $\widehat{\Theta}$ is adjusted as $\widetilde{\Theta}^{t+1} = (\tilde{\boldsymbol{b}}^{t+1}, \tilde{\boldsymbol{c}}^{t+1})$, where $\tilde{\boldsymbol{b}}^{t+1}$ and $\tilde{\boldsymbol{c}}^{t+1}$ are modified versions of estimators defined in equation (3.8).

(c) Repeat step (b) until the estimate converges, yielding $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{c}})$. Subsequently, the final estimator of β is obtained as $\hat{\beta} = T\hat{\boldsymbol{d}} + \Sigma\hat{\boldsymbol{c}}$.

3.2.3 Tuning Parameter Selection

The selection of the tuning parameter plays a pivotal role in enhancing the performance of regularized estimators across various smoothing methods. In our study, we employ the generalized cross-validation criterion (GCV) to ascertain the optimal

value of the tuning parameter λ, thereby effectively mitigating computational costs (Yuan and Cai, 2010). Given that the fitted value $\widehat{Y}$ serves as a linear predictor of the response Y, i.e., $\widehat{Y} = H_\lambda Y$, we choose λ as a minimizer of the weighted version of GCV score

$$GCV(\lambda) = \frac{1}{n}\frac{(\widehat{Y}-Y)^T W(\widehat{Y}-Y)}{(1-tr(H_\lambda)/n)^2}, \tag{3.10}$$

where the hat matrix H_λ has the form $H_\lambda = \left[\Sigma + n\lambda W^{-1}M^{-1}Q(Q^TM^{-1}Q)^{-1}Q^T\right]M^{-1}$. Additionally, if $w_i = 0$ exists with some loss function, the hat matrix is adjusted as $\widetilde{H}_\lambda = \left[\widetilde{\Sigma} + n\lambda\widetilde{W}^{-1}\widetilde{M}^{-1}\widetilde{Q}(\widetilde{Q}^T\widetilde{M}^{-1}\widetilde{Q})^{-1}\widetilde{Q}^T\right]\widetilde{M}^{-1}$ with notations defined as in equation (3.8).

3.3 Assumptions and Theoretical Results

In this section, we aim to establish the theoretical properties of M-estimation for $\beta(t)$ and $\boldsymbol{\theta}$ within the RKHS framework. Before presenting the main theoretical results, we introduce some technical assumptions required for our asymptotic properties. All the proofs are given in the Appendix.

(C1) The predictors $X(t)$ and $\boldsymbol{z}$ are independent and have finite fourth moments, i.e., $E\|X\|^4 < \infty$, $E\|\boldsymbol{z}\|^4 < \infty$. The noise is uncorrelated with predictors, as indicated by the conditions $E(X\varepsilon) = 0$ and $E(\boldsymbol{z}\varepsilon) = 0$.

(C2) For $j = 1, 2, \cdots, p$, $\mathrm{E}(\boldsymbol{z}_j|X(\cdot))$ is a continuous linear function, and there exists a function $g_j(\cdot) \in \mathcal{H}$, such that $\mathrm{E}(\boldsymbol{z}_j|X(\cdot)) = \langle X(\cdot), g_j(\cdot)\rangle$. We denote $u_{ij} = z_{ij} - \mathrm{E}(\boldsymbol{z}_j|x_i(\cdot)) = z_{ij} - \langle x_i(\cdot), g_j(\cdot)\rangle$, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Furthermore, we assume both $E(\boldsymbol{z}\boldsymbol{z}^T)$ and $E(\boldsymbol{u}_i\boldsymbol{u}_i^T)$ with $\boldsymbol{u}_i = (u_{i1}, u_{i2}, \ldots, u_{ip})^T$ are positive definite matrix .

(C3) The eigenvalues k_j of reproducing kernel K of $\mathcal{H}$ satisfy $k_j \asymp j^{-2\alpha}$ for some $\alpha > 1/2$, and the eigenvalues μ_j of the covariance function C for X satisfy $\mu_j \asymp j^{-2r}$ for some $r > 1/2$.

(C4) For any square integrable function f, there exists some constant $c > 0$ such that

$$E\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^4 \leqslant c\left[E\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^2\right]^2.$$

(C5) The function $\rho(\cdot)$ is convex and nonmonotone. Let $\psi = \rho' \in C^2(-\infty, \infty)$, where $C^2(-\infty, \infty)$ represents the space of functions that are twice continuously differentiable. Additionally, we have $\sup|\psi''(\cdot)| < \infty$, $E\psi(\varepsilon_i) = 0, E\psi'(\varepsilon_i) \neq 0$ and $Var(\psi^{(j)}) < \infty$, $j = 0, 1$.

The assumptions (C1)-(C2) are commonly employed in partially functional linear models, which align with the assumptions made in previous studies such as Yuan and Cai (2010) and Shin (2009). Assumption (C2) bears resemblance to equations

(18) and (19) in Shin (2009), which is essential for addressing the linear component associated with the vector predictor part in a partially functional linear model. Assumption (C3) takes into account the decay rate of eigenvalues for both the kernel function K and the covariance function C. Assumption (C4) pertains to the fourth moment of a linear functional of $X(t)$. Assumption (C5), as observed in works like Cox (1983) and Shin and Lee (2016), is commonly considered when examining M-type smoothing splines. Consequently, we obtain following asymptotic proprieties about the M-estimators of coefficients $\beta(t)$ and $\boldsymbol{\theta}$.

Theorem 3.1 Suppose tuning parameter $\lambda \asymp n^{-(2\alpha+2r)/(2\alpha+2r+1)}$. Under the assumptions (C1)-(C5), we have

$$\|\hat{\beta} - \beta_0\|_C^2 = O_p(n^{-\frac{(2\alpha+2r)}{(2\alpha+2r)+1}}). \tag{3.11}$$

where for some function f, $\|f\|_C^2 = \int_{\mathcal{T}} \int_{\mathcal{T}} f(s)C(s,t)f(t)dsdt = \|C^{1/2}f\|^2$.

Note that $\|\hat{\beta} - \beta_0\|_C^2$ measures the prediction errors for any random function X^* possessing the same distribution as X and independent of the sample $X_1, X_2, \cdots, X_n$. Thus, we obtain the same convergence rate as reported in previous studies by Yuan and Cai (2010) and Cui et al. (2020).

Theorem 3.2 For parameter $\boldsymbol{\theta}$, under the conditions in Theorem 3.1, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_2\boldsymbol{\Delta}_1^{-1}), \tag{3.12}$$

where $\boldsymbol{\Delta}_1 = E[\psi'(\varepsilon_i)\boldsymbol{u}_i\boldsymbol{u}_i^T], \boldsymbol{\Delta}_2 = E[\psi^2(\varepsilon_i)(\boldsymbol{u}_i\boldsymbol{u}_i^T)]$.

Remark 3.1 Let ϕ_j be the eigenfunction of the covariance function C corresponding to its eigenvalues μ_j, and ψ_j be the eigenfunction of the reproducing kernel function K corresponding to its eigenvalues k_j. The subsequent proof assumes perfect alignment between K and C, implying that they share a common ordered set of eigenfunctions (Cai and Yuan, 2012). Consequently, we have $\tau_j \asymp j^{-(2\alpha+2r)}$ and τ_j represents the jth largest eigenvalue of $K^{1/2}CK^{1/2}$. In other words, $K^{1/2}CK^{1/2}$ can be expressed as a sum over all j, where each term is given by $\tau_j e_j(t) \otimes e_j(s)$. Here, it should be noted that the sequence of eigenvalues satisfies $\tau_1 \geqslant \tau_2 \geqslant \cdots > 0$, with $\lim_{j\to\infty} \tau_j = 0$, and the set of functions $\{e_j\}$ are orthonormalized.

3.4 Simulation Studies

In this section, we investigate the finite-sample performance of penalized M-estimators for partially functional linear under an RKHS framework. The simulation setting in our study closely resembles the experimental conditions described in Yuan and Cai (2010) and Zhou et al. (2016), albeit with certain modifications aimed at introducing outliers into the datasets. Without loss of generality, we generated simulation data from the following model

$$y_i = \alpha + \boldsymbol{z}_i^T\boldsymbol{\theta} + \int_{\mathcal{T}} \beta(t)x_i(t)dt + \varepsilon_i, \tag{3.13}$$

where $\alpha = 2$. We draw samples $\boldsymbol{z}_i$ from the multivariate normal distribution $\mathcal{MN}(\mathbf{0}, \Sigma)$ with $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, and $\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (1, 2)^T$. The slope function $\beta(t)$ is given by $\sum_{j=1}^{50} b_j \phi_j(t)$, where $b_1 = 0.5, \phi_1(t) = 1$, and $b_j = 4(-1)^{j+1} j^{-2}, \phi_j(t) = \sqrt{2}\cos(j\pi t)$, $j > 1$. The functional predictors $x_i(t)$ is generated as follows:

$$x_i(t) = \sum_{j=1}^{50} j^{-v} U_{ij} \phi_j(t),$$

with the independent random variables $U_{ij} \sim U(-\sqrt{3}, \sqrt{3})$. It is not hard to see that each U_{ij} had zero mean and unit variance. In these settings, the eigenvalues of the covariance function C are given by j^{-2v}. We choose values of $v = 0.6$ and 1.2 to control the rate at which the eigenvalues decay. Additionally, following Cai and Yuan (2012), we assume that the reproducing kernel function is defined as presented below and illustrated in Figure 3.1.

$$K(s, t) = \sum_{j=1}^{50} \frac{2}{(j\pi)^4} \cos(j\pi s) \cos(j\pi t). \tag{3.14}$$

In our simulation design, we consider a dense model where the observed points on each curve are uniformly distributed with an equal spacing of $m = 301$ points within the interval $[0, 1]$. To examine the efficacy of the robust methodology, we have devised three scenarios inspired by Boente et al. (2020). For each scenario, we conduct 200 replicated trials with sample sizes of $n = 100$, 300, and 500, respectively. The first scenario considers random errors that are independently and

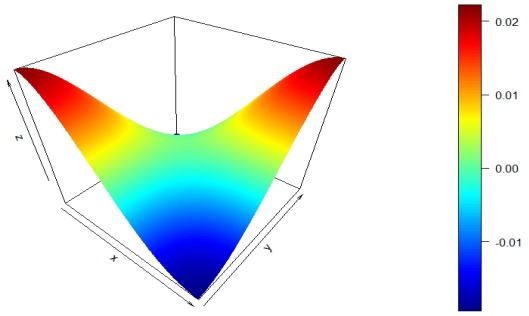

Figure 3.1: Kernel function based on Yuan and Cai (2010)

identically distributed according to a standard normal distribution with zero mean and unit variance. The second scenario includes outliers in the response variables, which primarily affect the estimation of $\boldsymbol{\theta}_0$. The third scenario involves high-leverage outliers in the functional variables, which typically have a significant impact on the estimation of coefficient function β_0. Furthermore, within this particular scenario, we have devised three distinct settings to deliberately introduce contamination into the functional predictors.

- ✦ Scenario I: The errors ε_i, $i = 1, 2, \ldots, n$ follow the standard normal distribution $N(0, 1)$.

- ✦ Scenario II: We assume that the random errors ε_i, $i = 1, 2, \ldots, n$ are draw from a mixed normal distribution $0.9N(0, 1) + 0.1N(10, 0.25)$ or a heavy-tailed

distribution $t(3)$.

- Scenario III: We consider high-leverage outliers by contaminating the functional covariates $x_i(t)$ and the random errors ε_i simultaneously. The distribution of random errors ε_i of the following settings are identical to that observed in Scenario II. Specifically, we first draw random samples w_i form Bernoulli distribution Bernoulli(0.1) and design three different settings with the diverse contaminated $x_i^c(t)$.

  - Setting 1:

$$x_i^c(t) = \begin{cases} x_i(t), & w_i = 0, \\ \sum_{j=1}^{50} j^{-v} U_{ij}^c \phi_j(t), \text{ where } U_{i1}^c \sim U(0, 12), & \\ \text{and } U_{ij}^c \sim U(-\sqrt{3}, \sqrt{3}), j \neq 1, & w_i = 1. \end{cases}$$

  - Setting 2:

$$x_i^c(t) = \begin{cases} x_i(t), & w_i = 0, \\ \sum_{j=1}^{50} j^{-v} U_{ij}^c \phi_j(t), \text{ where } U_{i2}^c \sim U(0, 12), & \\ \text{and } U_{ij}^c \sim U(-\sqrt{3}, \sqrt{3}), j \neq 2, & w_i = 1. \end{cases}$$

  - Setting 3: We randomly select an index m_i from 1 to 3. If $m_i = 1$, the value of $X_i^c(t)$ remains the same as in setting 1. If $m_i = 2$, the value of

$X_i^c(t)$ remains the same as in setting 2. Otherwise,

$$x_i^c(t) = \begin{cases} x_i(t), & w_i = 0, \\ \sum_{j=1}^{50} j^{-v} U_{ij}^c \phi_j(t), \text{where } U_{ij}^c \sim U(0, 12), j = 1, 2, & \\ \text{and } U_{ij}^c \sim U(-\sqrt{3}, \sqrt{3}), j \neq 1, 2, & w_i = 1. \end{cases}$$

As an illustration of the aforementioned simulation design for contamination data, we present a selection of 50 randomly sampled non-contaminated functional covariates $x_i(t)$ alongside their contaminated counterparts $x_i^c(t)$ in Figures 3.2-3.4. The trajectories of non-contaminated $x_i(t)$ and contaminated $x_i^c(t)$ under setting 1 of Scenario III are illustrated in Figure 3.2, while Figures 3.3-3.4 present similar trajectory patterns for random samples under settings 2-3 of Scenario III, respectively.

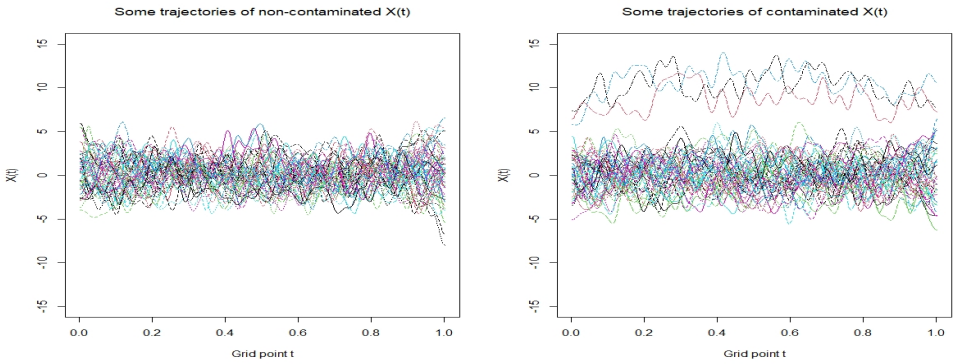

Figure 3.2: Some trajectories of non-contaminated and contaminated functional predictors (Scenario III, Setting 1).

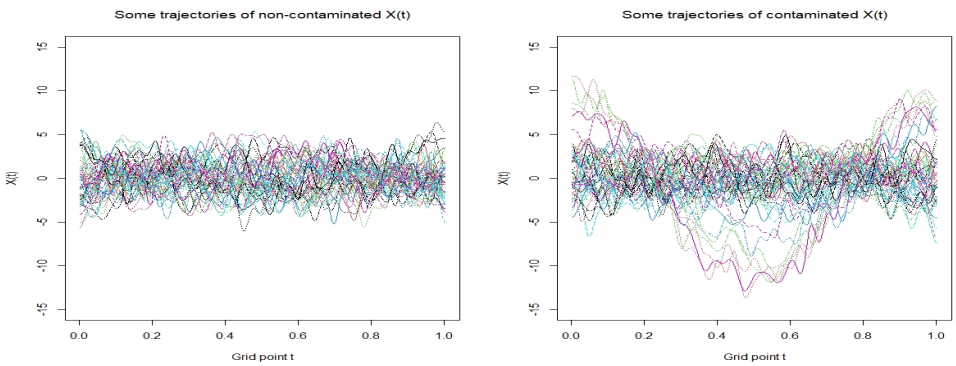

Figure 3.3: Some trajectories of non-contaminated and contaminated functional predictors (Scenario III, Setting 2).

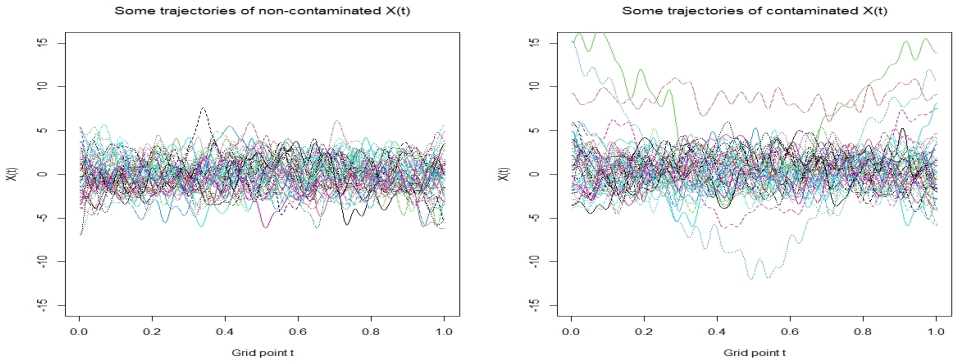

Figure 3.4: Some trajectories of non-contaminated and contaminated functional predictors (Scenario III, Setting 3).

To evaluate the performance of our robust estimation, the mean integrated squared error (MISE) for the functional coefficient $\hat{\beta}$ is used as an evaluation criterion in Table

3.1, given by

$$\text{MISE} = \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}}(\hat{\beta}(t) - \beta(t))dt,$$

where N is the number of Monet Carlo replications. Meanwhile, the mean square error (MSE) for $\widehat{\boldsymbol{\theta}}$ was further examined and documented in Table 3.2. To conserve space, we exclusively present simulation results for $v = 0.6$, while additional numerical findings ($v = 1.2$) are provided in the Appendix. According to Table 3.1, it is evident that the classical estimators exhibit vulnerability to contamination. In the absence of outliers, all estimates of $\hat{\beta}$ demonstrate comparable performance. However, when confronted with contaminated data sets, the mean integrated squared error (MISE) of $\hat{\beta}$ estimated using least squares (LS) is observed to be higher in comparison to robust methodologies such as least absolute deviations (LAD), Huber, Bisquare, and Hampel. Similarly, based on the simulation results presented in Table 3.2, it can be observed that the magnitude of the mean estimation error squared (MES) for $\widehat{\boldsymbol{\theta}}$ obtained through classical least squares (LS) method is comparatively higher than that obtained by the robust methods.

Table 3.1: MISE of $\hat{\boldsymbol{\beta}}(t)$ based on RKHS approach with different sample size and error distributions, where λ is selected by generalized cross-validation and $v = 0.6$.

Sample size	Scenario	Setting	Distribution of Errors	Method				
				LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	**0.0375**	0.0444	0.0404	0.0405	0.0379
	Scenario II		mixed normal	**0.2098**	0.1095	0.1571	0.1420	0.1720
			$t(3)$	**0.0942**	0.0835	0.0721	0.0834	0.0724
	Scenario III	Setting 1	mixed normal	**0.2253**	0.1138	0.1639	0.1517	0.1994
			$t(3)$	**0.0975**	0.0896	0.0839	0.0695	0.0740
		Setting 2	mixed normal	**0.2250**	0.2069	0.1879	0.2003	0.1787
			$t(3)$	**0.0956**	0.0821	0.0752	0.0811	0.0695
		Setting 3	mixed normal	**0.2430**	0.1357	0.1681	0.2077	0.2132
			$t(3)$	**0.0835**	0.0813	0.0766	0.0818	0.0655
$n = 300$	Scenario I		normal	**0.0145**	0.0218	0.0127	0.0106	0.0153
	Scenario II		mixed normal	**0.0909**	0.0304	0.0396	0.0414	0.0480
			$t(3)$	**0.0269**	0.0181	0.0182	0.0180	0.0227
	Scenario III	Setting 1	mixed normal	**0.0945**	0.0380	0.0416	0.0450	0.0410
			$t(3)$	**0.0323**	0.0321	0.0196	0.0199	0.0228
		Setting 2	mixed normal	**0.0675**	0.0337	0.0426	0.0489	0.0429
			$t(3)$	**0.0276**	0.0256	0.0213	0.0243	0.0280
		Setting 3	mixed normal	**0.0689**	0.0470	0.0411	0.0677	0.0457
			$t(3)$	**0.0328**	0.0356	0.0278	0.0324	0.0277
$n = 500$	Scenario I		normal	**0.0093**	0.0125	0.0087	0.0102	0.0084
	Scenario II		mixed normal	**0.0628**	0.0172	0.0204	0.0241	0.0234
			$t(3)$	**0.0198**	0.0184	0.0130	0.0148	0.0155
	Scenario III	Setting 1	mixed normal	**0.0480**	0.0304	0.0247	0.0216	0.0249
			$t(3)$	**0.0211**	0.0156	0.0139	0.0128	0.0154
		Setting 2	mixed normal	**0.0480**	0.0287	0.0234	0.0260	0.0274
			$t(3)$	**0.0192**	0.0201	0.0116	0.0183	0.0192
		Setting 3	mixed normal	**0.0573**	0.0287	0.0424	0.0537	0.0277
			$t(3)$	**0.0214**	0.0271	0.0249	0.0279	0.0210

Table 3.2: MSE of $\hat{\boldsymbol{\theta}}$ based on RKHS approach with different sample size and error distributions, where λ is selected by generalized cross-validation and $v = 0.6$.

Sample size	Scenario	Setting	Distribution of Errors	Parameters									
				θ_1					θ_2				
				LS	LAD	Huber	Bisquare	Hampel	LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	**0.0168**	0.0240	0.0184	0.0149	0.0163	**0.0067**	0.0109	0.0088	0.0096	0.0074
	Scenario II		mixed normal	**0.1553**	0.0485	0.0609	0.0401	0.0456	**0.0744**	0.0264	0.0220	0.0197	0.0226
			$t(3)$	**0.0381**	0.0298	0.0260	0.0322	0.0309	**0.0226**	0.0180	0.0137	0.0130	0.0158
	Scenario III	Setting 1	mixed normal	**0.1341**	0.0729	0.0472	0.0338	0.0347	**0.0500**	0.0399	0.0229	0.0197	0.0240
			$t(3)$	**0.0370**	0.0353	0.0250	0.0314	0.0285	**0.0190**	0.0161	0.0138	0.0131	0.0153
		Setting 2	mixed normal	**0.1427**	0.0948	0.0498	0.0448	0.0387	**0.0684**	0.0290	0.0222	0.0249	0.0160
			$t(3)$	**0.0516**	0.0347	0.0246	0.0336	0.0302	**0.0216**	0.0167	0.0162	0.0151	0.0156
		Setting 3	mixed normal	**0.2480**	0.0542	0.0624	0.0640	0.1052	**0.1423**	0.0234	0.0368	0.0407	0.0534
			$t(3)$	**0.0443**	0.0335	0.0283	0.0379	0.0339	**0.0206**	0.0160	0.0144	0.0168	0.0159
$n = 300$	Scenario I		normal	**0.0043**	0.0055	0.0038	0.0043	0.0042	**0.0017**	0.0031	0.0020	0.0026	0.0018
	Scenario II		mixed normal	**0.0404**	0.0111	0.0080	0.0060	0.0061	**0.0182**	0.0054	0.0041	0.0036	0.0033
			$t(3)$	**0.0133**	0.0087	0.0055	0.0071	0.0063	**0.0066**	0.0038	0.0033	0.0033	0.0031
	Scenario III	Setting 1	mixed normal	**0.0402**	0.0148	0.0079	0.0057	0.0054	**0.0153**	0.0090	0.0040	0.0029	0.0036
			$t(3)$	**0.0120**	0.0109	0.0074	0.0075	0.0052	**0.0059**	0.0057	0.0032	0.0036	0.0030
		Setting 2	mixed normal	**0.0273**	0.0123	0.0087	0.0060	0.0060	**0.0195**	0.0047	0.0037	0.0033	0.0035
			$t(3)$	**0.0120**	0.0092	0.0066	0.0068	0.0079	**0.0057**	0.0035	0.0030	0.0038	0.0033
		Setting 3	mixed normal	**0.0360**	0.0085	0.0059	0.0111	0.0068	**0.0205**	0.0059	0.0027	0.0063	0.0030
			$t(3)$	**0.0116**	0.0108	0.0079	0.0079	0.0065	**0.0060**	0.0048	0.0035	0.0032	0.0037
$n = 500$	Scenario I		normal	**0.0022**	0.0033	0.0023	0.0023	0.0022	**0.0011**	0.0018	0.0013	0.0013	0.0010
	Scenario II		mixed normal	**0.0197**	0.0052	0.0043	0.0030	0.0033	**0.0102**	0.0028	0.0023	0.0019	0.0015
			$t(3)$	**0.0056**	0.0053	0.0040	0.0028	0.0039	**0.0022**	0.0021	0.0017	0.0019	0.0018
	Scenario III	Setting 1	mixed normal	**0.0194**	0.0097	0.0047	0.0029	0.0032	**0.0137**	0.0048	0.0022	0.0017	0.0020
			$t(3)$	**0.0052**	0.0034	0.0034	0.0036	0.0037	**0.0034**	0.0019	0.0018	0.0013	0.0020
		Setting 2	mixed normal	**0.0214**	0.0054	0.0040	0.0028	0.0031	**0.0112**	0.0024	0.0020	0.0017	0.0017
			$t(3)$	**0.0069**	0.0050	0.0034	0.0038	0.0036	**0.0035**	0.0026	0.0020	0.0022	0.0021
		Setting 3	mixed normal	**0.0191**	0.0054	0.0057	0.0057	0.0031	**0.0112**	0.0024	0.0039	0.0025	0.0015
			$t(3)$	**0.0080**	0.0052	0.0038	0.0042	0.0045	**0.0033**	0.0027	0.0024	0.0020	0.0021

Table 3.3: MAPE (Mean Absolute Percentage Error) and SE (Standard Error, in the brackets) of prediction errors over 500 trails with different sample size and error distributions, where λ is selected by generalized cross-validation and $v = 0.6$.

Sample size	Scenario	Setting	Distribution of Errors	Method				
				LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	**0.0445**(0.2183)	0.0525(0.2566)	0.0446(0.2186)	0.0451(0.2188)	0.0445(0.2183)
	Scenario II		mixed normal	**0.0958**(0.5908)	0.0895(0.5751)	**0.0862**(0.5597)	0.0859(0.5627)	0.0888(0.5798)
			$t(3)$	**0.1271**(0.3369)	0.0954(0.4045)	0.1090(0.3284)	0.0784(0.3347)	0.1144(0.3311)
	Scenario III	Setting 1	mixed normal	**0.2068**(0.5224)	0.1124(0.4926)	0.1429(0.4898)	0.1972(0.4993)	0.1680(0.5016)
			$t(3)$	**0.0537**(0.2804)	0.0337(0.2795)	0.0517(0.2728)	0.0524(0.2763)	0.0500(0.2730)
		Setting 2	mixed normal	**0.0674**(0.5437)	0.0658(0.5684)	0.0604(0.5108)	0.0590(0.5206)	0.0625(0.5235)
			$t(3)$	**0.0575**(0.3252)	0.0538(0.2968)	0.0503(0.2886)	0.0507(0.2956)	0.0505(0.2901)
		Setting 3	mixed normal	**0.0654**(0.5481)	0.0588(0.5175)	0.0590(0.5162)	0.0591(0.5231)	0.0545(0.5090)
			$t(3)$	**0.0356**(0.3053)	0.0334(0.3101)	0.0332(0.2964)	0.0334(0.3012)	0.0333(0.2985)
$n = 300$	Scenario I		normal	**0.0260**(0.0893)	0.0273(0.0948)	0.0259(0.0897)	0.0259(0.0899)	0.0260(0.0894)
	Scenario II		mixed normal	**0.1006**(0.2712)	0.0754(0.2609)	0.0795(0.2623)	0.0790(0.2619)	0.0782(0.2619)
			$t(3)$	**0.0548**(0.1477)	0.0469(0.1477)	0.0414(0.1441)	0.0449(0.1446)	0.0452(0.1445)
	Scenario III	Setting 1	mixed normal	**0.0720**(0.2740)	0.0533(0.2659)	0.0551(0.2651)	0.0528(0.2654)	0.0555(0.2660)
			$t(3)$	**0.0425**(0.1491)	0.0404(0.1518)	0.0364(0.1469)	0.0369(0.1475)	0.0358(0.1470)
		Setting 2	mixed normal	**0.0618**(0.2771)	0.0529(0.2673)	0.0517(0.2679)	0.0513(0.2684)	0.0512(0.2687)
			$t(3)$	**0.0349**(0.1495)	0.0344(0.1505)	0.0327(0.1467)	0.0326(0.1472)	0.0322(0.1468)
		Setting 3	mixed normal	**0.0653**(0.2751)	0.0521(0.2693)	0.0560(0.2695)	0.0566(0.2711)	0.0524(0.2657)
			$t(3)$	**0.0447**(0.1491)	0.0450(0.1538)	0.0443(0.1483)	0.0451(0.1490)	0.0444(0.1478)
$n = 500$	Scenario I		normal	**0.0413**(0.0672)	0.0422(0.0717)	0.0427(0.0688)	0.0431(0.0691)	0.0426(0.0682)
	Scenario II		mixed normal	**0.0934**(0.2064)	0.0794(0.2023)	0.0798(0.2024)	0.0779(0.2025)	0.0780(0.2025)
			$t(3)$	**0.0497**(0.1108)	0.0486(0.1095)	0.0474(0.1091)	0.0471(0.1094)	0.0469(0.1089)
	Scenario III	Setting 1	mixed normal	**0.0648**(0.2070)	0.0579(0.2042)	0.0578(0.2022)	0.0573(0.2022)	0.0578(0.2025)
			$t(3)$	**0.0389**(0.1162)	0.0373(0.1129)	0.0371(0.1124)	0.0369(0.1127)	0.0370(0.1123)
		Setting 2	mixed normal	**0.0671**(0.2079)	0.0570(0.2037)	0.0582(0.2030)	0.0577(0.2031)	0.0577(0.2033)
			$t(3)$	**0.0671**(0.2079)	0.0570(0.2037)	0.0582(0.2030)	0.0577(0.2031)	0.0577(0.2033)
		Setting 3	mixed normal	**0.0714**(0.2077)	0.0571(0.2058)	0.0583(0.2037)	0.0549(0.2036)	0.0576(0.2044)
			$t(3)$	**0.0331**(0.1148)	0.0316(0.1184)	0.0327(0.1147)	0.0328(0.1151)	0.0326(0.1147)

Additionally, we provide the mean absolute percentage error (MAPE) and standard error (SE) of prediction for both LS and robust methods through 500 replicated experiments in Table 3.3. The MAPE of prediction errors based on robust methods for contaminated data is observed to be lower than that obtained using the classical least squares (LS) method, as demonstrated in Table 3.3. For instance, the utilization of the Huber function in Scenario II, where ε_i follows a mixed normal distribution, results in a reduction of the MAPE of prediction errors from 0.0958 (using LS method) to 0.0862 (using Huber method), with a sample size of $n = 100$.

3.5 Real Data Examples

3.5.1 Near-infrared Spectroscopy Data

Quantitative NIR (Near-infrared reflectance) spectroscopy data is commonly employed for the analysis of diverse materials, including food, beverages, and pharmaceutical products. In this study, we utilize the biscuit dough dataset from Osborne et al. (1984), which is readily available in the R package **ppls**. Previous analyses have been conducted by Brown et al. (2001), Mas and Pumo (2009), and Luo and Qi (2015), among others. In this data set, 72 sample sets are made up, with the standard recipe varied to provide a large range for each of the four constituents: fat, sucrose, dry flour and water. The four constituents are measured as percent-

age. There are spectra measured from 1100 and 2498 nanometers (nm) in 2 nm increments, providing densely observed functional predictors on a grid of 700 points. Our goal is to predict the fat content of a cookie sample from the functional spectra measurements as well as dry flour and water. One important feature of this data set is that it contains outliers (Osborne et al., 1984; Brown et al., 2001); therefore, robust approaches are expected to perform better than the least squares approach. Our proposed framework is given by the equation

$$y_i = \boldsymbol{z}_i^T \theta + \int_{\mathcal{T}} \beta(t) x_i(t) dt + \varepsilon_i, \tag{3.15}$$

where the response variable y_i represents fat content, $\boldsymbol{z}_i$ denotes the two dimensional vector encompassing dry flour and water, and x_i is the functional spectra measurements in the interval $\mathcal{T} = [1100, 2498]$. NIR(near-infrared) reflectance spectrum measured from 1100 to 2498 nanometers for all samples are represented in Figure 3.5.

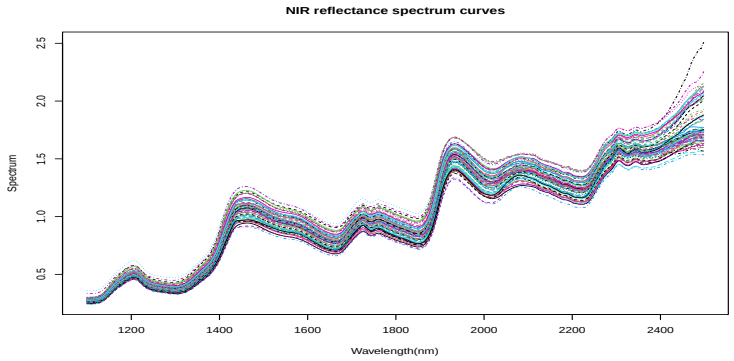

Figure 3.5: NIR(near-infrared reflectance) spectrum measured from 1100 to 2498 nanometers (nm) in 2 nm increments.

The mean absolute percentage error (MAPE) and standard error (SE) of the LS, LAD, Huber, Bisquare, and Hampel estimations based on the RKHS approach are presented in Table 3.4. The kernel function used in our numerical experiment in Section 3.4 remains consistent. In Table 3.4, the value of N in the splits column indicates the number of times we iteratively split the data into training and testing sets, followed by calculating the mean absolute percentage error (MAPE) for predictions. As expected, estimators employing robust methods such as LAD, Huber, Bisquare, and Hampel exhibit lower MAPE values compared to those utilizing the least squares (LS) method.

Table 3.4: MAPE (Mean Absolute Percentage Error) and SE (Standard Error, in the brackets) of prediction errors with different randomly splits for analysis of NIR spectroscopy data.

Splits N	Method				
	LS	LAD	Huber	Bisquare	Hampel
$N = 50$	**0.0255**(0.1607)	0.0147(0.1196)	0.0182(0.1336)	0.0140(0.1173)	0.0198(0.1399)
$N = 100$	**0.0252**(0.1574)	0.0145(0.1162)	0.0177(0.1302)	0.0138(0.1144)	0.0194(0.1366)
$N = 200$	**0.0253**(0.1599)	0.0147(0.1202)	0.0179(0.1328)	0.0142(0.1184)	0.0193(0.1385)
$N = 500$	**0.0253**(0.1637)	0.0151(0.1254)	0.0182(0.1382)	0.0144(0.1236)	0.0199(0.1447)

3.5.2 Appliances Energy Prediction Data

The second real-life example we considered is the dataset for predicting energy consumption of appliances, which can be accessed from the UCI Machine Learning Repository website (*https://archive.ics.uci.edu*). This dataset was collected as part of a comprehensive study aimed at investigating the intricate relationships between appliances energy consumption and various predictors, meticulously recorded every 10 minutes over a span of approximately 4.5 months. In this data set, the house temperature and humidity conditions were monitored using a ZigBee wireless sensor network, while the weather data from the nearest airport weather station (Chievres Airport, Belgium) was obtained from a public data set provided by Reliable Prognosis (rp5.ru) (Candanedo et al., 2017). We are interested to fit a model to predict the

energy consumption of appliances along with the outside weather data from the nearest weather station. The modeling framework we are considered is

$$y_i = \alpha + \boldsymbol{z}_i^T \boldsymbol{\theta} + \int_{\mathcal{T}} \boldsymbol{\beta}(t) x_i(t) dt + \varepsilon_i, \tag{3.16}$$

where y_i is the daily average consumption, $x_i(t)$ is the outside temperature observed every 10 minutes from the weather station, $\boldsymbol{z}_i = (z_{1i}, z_{2i})^T$ is the two dimensional vector encompassing pressure and wind speed.

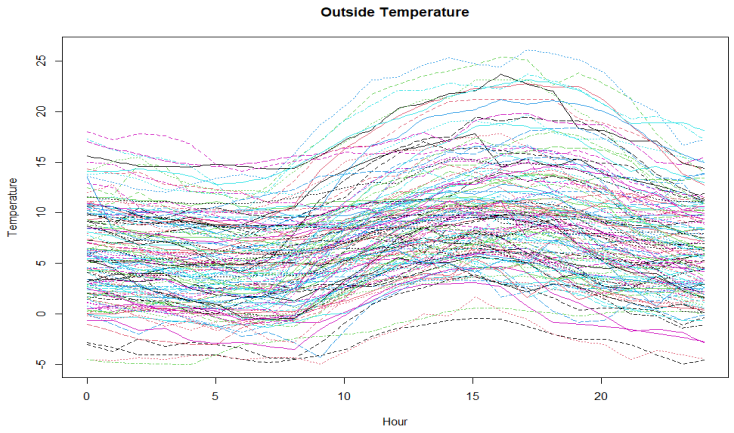

Figure 3.6: Outside temperature observed every 10 minutes during 136 days.

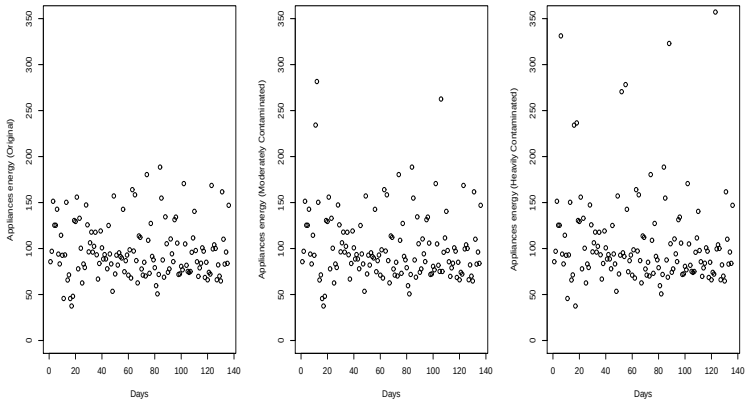

Figure 3.7: Appliances energy consumption in the original data (left panel), which in the mildly contaminated data (middle panel) and the moderately contaminated data (right panel).

Figure 3.6 shows the original outside temperature observed every 10 minutes (i.e., $x_i(t)$) from the nearest weather station during a period of 136 days. The contaminate setting by introducing mild and moderate levels of contamination to the original data, as illustrated in Figure 3.7. Specifically, we randomly select a sample of 2% (mild) and 5% (moderate) from original y_i, and adjust the remaining values by adding the maximum value of y_i. Table 3.5 reports the mean absolute percentage error (MAPE) and standard deviation (SD) for both the original and contaminated data sets, respectively. The results indicate that robust estimation provides more accurate predictions compared to classical methods. For instance, in

datasets with moderate contamination, the mean absolute percentage error (MAPE) of classical least squares method (0.3574) is comparatively higher than that of robust estimation methods, particularly LAD and Huber.

Table 3.5: MAPE (Mean Absolute Percentage Error) and SE (Standard Error, in the brackets) of prediction errors with 500 randomly splits for analysis results of appliances energy data.

	Method				
	LS	LAD	Huber	Bisquare	Hampel
Original Data	**0.2708**(3.7214)	0.2466(3.7390)	0.2496(3.7540)	0.2709(3.7536)	0.2698(3.7743)
Mildly contaminated Data	**0.3053**(3.8461)	0.2497(3.7156)	0.2509(3.7239)	0.3052(3.8218)	0.3047(3.8315)
Moderately contaminated Data	**0.3574**(4.0307)	0.2579(3.7480)	0.2580(3.7467)	0.3476(3.9366)	0.3464(3.9765)

3.6 Conclusions

In this chapter, we investigate the theoretical properties of robust estimation for the partially functional linear regression model within the framework of RKHS. We derive convergence rates for functional prediction errors and establish asymptotic normality of the multivariate covariate variable. By making appropriate assumptions, we demonstrate that our robust estimation achieves the same convergence rate of prediction errors as classic least squares estimation in functional linear models. Through simulation studies, we illustrate that our proposed estimators exhibit

robustness and possess desirable statistical properties in finite-sample scenarios. Furthermore, application of our method on two real datasets demonstrates that the robust M-estimators remain reliable even when atypical observations are present in the functional explanatory variables.

3.7 Appendix

3.7.1 Proofs

This section encompasses the formal proofs of lemmas and theorems pertaining to our asymptotic findings. Firstly, we assume that $\alpha = 0$ and $\sigma_\varepsilon = 1$ for the sake of simplicity in presenting the proof. We consider a decomposition of $\mathcal{H}$ as $\mathcal{H} = \mathcal{H}_0 \otimes \mathcal{H}_1$, where $\mathcal{H}_0 := \{\beta \in \mathcal{H} : J(\beta) = 0\}$ and $\mathcal{H}_1$ is its orthogonal complement in $\mathcal{H}$. Since $J(\beta) = \int_{\mathcal{T}} [\beta''(t)]^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \xi_i, \xi_j \rangle_{\mathcal{H}}$ with $\langle \xi_i, \xi_j \rangle_{\mathcal{H}} = \int_{\mathcal{T}} \int_{\mathcal{T}} x_i(s) K(s,t) x_i(t) ds dt$, we can denote $J(\beta) = \|\beta\|^2_{\mathcal{H}_1} = \|\beta\|^2_{\mathcal{H}}$ for $\beta \in \mathcal{H}_1$. Then, we employ the profiling technique to derive the asymptotic properties of the estimator for the functional parameter β, which is commonly utilized in semi-parametric statistics. Specifically, we consider β as a function of $\boldsymbol{\theta}$, denoted by $\beta(\boldsymbol{\theta})$. Subsequently, given any specific value of $\boldsymbol{\theta}$, the optimization problem can be reduced to

$$\min_{\beta(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \rho\big(y_i - \boldsymbol{z}_i^T \boldsymbol{\theta} - \langle \beta(\boldsymbol{\theta}), x_i \rangle\big) + n\lambda \|\beta(\boldsymbol{\theta})\|^2_{\mathcal{H}} \right], \tag{3.17}$$

and the minimizer is denoted by $\hat{\beta}(\boldsymbol{\theta})$. Then, the estimation of $\hat{\boldsymbol{\theta}}$ can be derived subsequently from:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho\big(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\hat{\beta}(\boldsymbol{\theta}), x_i\rangle\big). \tag{3.18}$$

Finally, the estimators of β and $\boldsymbol{\theta}$ are updated using the expressions $\hat{\beta}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$, respectively. Our proof closely follows the methodology employed in Cui et al. (2020).

We now list some notations and properties that will be used in the following proof. For any operator $\mathcal{F}$, we denote $\mathcal{F}^T$ as its adjoint operator. For any operator $\mathcal{F}$, $\|\mathcal{F}\|_{op}$ stands for its usual operator norm, that is, $\|\mathcal{F}\|_{op} = \sup_{\{h:\|h\|_{\mathcal{L}_2}=1\}} \|\mathcal{F}h\|_{\mathcal{L}_2}$. The trace norm of an operator $\mathcal{F}$ is $\mathrm{tr}(\mathcal{F}) = \sum_j \langle(\mathcal{F}^T\mathcal{F})^{1/2}e_j, e_j\rangle$ for any orthonormal basis $\{e_j\}$. $\|\cdot\|_{HS} = (\sum_{j,k}\langle\mathcal{F}e_j, e_k\rangle^2)^{1/2}$ stands for the Hilbert-Schmidt norm (i.e., Frobenius norm). In particular, if $\mathcal{F}$ is a Hilbert-Schmidt norm and $\mathcal{G}$ is a bounded operator, we have the property $\|\mathcal{F}\mathcal{G}\|_{HS} \leqslant \|\mathcal{F}\|_{HS} \cdot \|\mathcal{G}\|_{op}$.

Proof of Theorem 3.1

For a given $\boldsymbol{\theta}$, we rewrite the criterion of taking derivatives (3.17) with respect to $\beta(\boldsymbol{\theta})$ and setting it equal to zero, it is easy to show that the solution to (3.17) is equivalently to the minimizer of the following penalized weighted least-squares criterion:

$$\sum_{i=1}^{n} w_i\big(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\beta(\boldsymbol{\theta}), x_i\rangle\big)^2 + n\lambda\|\beta(\boldsymbol{\theta})\|_{\mathcal{H}}^2,$$

where $w_i = \psi(\varepsilon_i)/\varepsilon_i$. We note that the reproducing Hilbert space $\mathcal{H}$ is the range of $K^{1/2}$ in $L^2(\mathcal{T})$. We define $\delta(\boldsymbol{\theta}) = K^{-1/2}\beta(\boldsymbol{\theta})$ and $\delta_0(\boldsymbol{\theta}) = K^{-1/2}\beta_0(\boldsymbol{\theta})$. Then we have $\|\beta(\boldsymbol{\theta})\|_{\mathcal{H}} = \|\delta(\boldsymbol{\theta})\|$ and $\|C^{1/2}(\hat{\beta}(\boldsymbol{\theta}) - \beta_0(\boldsymbol{\theta}))\|^2 = \|\Gamma^{1/2}(\hat{\delta}(\boldsymbol{\theta}) - \delta_0(\boldsymbol{\theta}))\|^2$ with the linear operator $\Gamma := K^{1/2}CK^{1/2}$. Thus, the penalized criterion can be rewritten as

$$\sum_{i=1}^{n} w_i\big(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\delta(\boldsymbol{\theta}), K^{1/2}x_i\rangle\big)^2 + n\lambda\|\delta(\boldsymbol{\theta})\|^2,$$

Then, we have

$$\begin{aligned}\hat{\delta}(\boldsymbol{\theta}) =& \frac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n} K^{1/2}x_i(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta}) \\ =& \frac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n} K^{1/2}x_i(\langle\delta_0(\boldsymbol{\theta}), K^{1/2}x_i\rangle + \varepsilon_i),\end{aligned}$$

where $\Gamma_n = K^{1/2}C_nK^{1/2}$ is a simple moment estimator of Γ, and $C_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i(s)\otimes x_i(t)$ is a simple moment estimator of covariance function $C = \mathrm{E}(\boldsymbol{x}\otimes\boldsymbol{x})$. Here, $W_n = \mathrm{diag}(w_1, w_2, \ldots, w_n)$ is the weighted matrix. In order to expression our proof more concisely, we denote $\hat{\delta}(\boldsymbol{\theta})$ as $\hat{\delta}$ and $\delta_0(\boldsymbol{\theta})$ as δ_0. Thus, we have

$$\begin{aligned}\Gamma^{1/2}(\hat{\delta} - \delta_0) =& \frac{1}{n}\Gamma^{1/2}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n} K^{1/2}x_i\langle\delta_0, K^{1/2}x_i\rangle_{\mathcal{L}_2} \\ &+ \frac{1}{n}\Gamma^{1/2}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n}\varepsilon_i K^{1/2}x_i - \Gamma^{1/2}\delta_0 \\ =& \Gamma^{1/2}(\Gamma_n(\Gamma_n + \lambda W_n^{-1})^{-1} - I)\delta_0 + \Gamma^{1/2}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i K^{1/2}x_i}{n}\end{aligned}$$

$$
\begin{aligned}
&= -\lambda\Gamma^{1/2}W_n^{-1}(\Gamma_n+\lambda W_n^{-1})^{-1}\delta_0+\Gamma^{1/2}(\Gamma_n+\lambda W_n^{-1})^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i K^{1/2}x_i}{n}\\
&= -\lambda\Gamma^{1/2}(\Gamma_n W_n+\lambda I)^{-1}\delta_0+\Gamma^{1/2}(\Gamma_n+\lambda W_n^{-1})^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i K^{1/2}X_i}{n}\\
&:= M_1+M_2.
\end{aligned}
$$

First, we can decompose the term M_1 based on the fact that $B^{-1}-A^{-1}=B^{-1}(A-B)A^{-1}$ for $A=\Gamma W_n+\lambda I$ and $B=\Gamma_n W_n+\lambda I$, thus we obtain that

$$
\begin{aligned}
M_1 =& -\lambda\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\delta_0-\lambda\Gamma^{1/2}\left[(\Gamma_n W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\right]\delta_0\\
=& -\lambda\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\delta_0-\lambda\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\delta_0\\
&-\lambda\Gamma^{1/2}(\Gamma_n W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\delta_0\\
:=& M_{11}+M_{12}+M_{13}.
\end{aligned}
$$

Define $\mathcal{N}_1(\lambda):=\sum_{j\geqslant 1}\frac{v_j^2}{w_j\tau_j+\lambda}$ with $\delta_0=\sum_{j\geqslant 1}v_j e_j$. Note that $\{\tau_j\}$ and $\{e_j\}$ for $j\geqslant 1$ are eigenvalues and the orthonormalized eigenfunctions of linear operator Γ. Then, according to the definition of $\mathcal{N}_1(\lambda)$ and the fact that $\tau_j\to 0$ as $j\to\infty$, we can deduce that $\|M_{11}\|^2=\|-\lambda\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\delta_0\|^2=\lambda^2\mathcal{N}_1(\lambda)$.

For M_{12}, we define $\mathcal{N}_2(\lambda):=tr((\Gamma W_n+\lambda I)^{-1}\Gamma)=\sum_{j\geqslant 1}\frac{\tau_j}{w_j\tau_j+\lambda}$. Then by Lemma 3.1, we have

$$
\begin{aligned}
\|M_{12}\|^2 \leqslant& \|M_{11}\|^2\|(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1/2}\|_{HS}^2\|(\Gamma W_n+\lambda I)^{-1/2}\|^2\\
&= O_p(\lambda^2\mathcal{N}_1(\lambda)\cdot\frac{\mathcal{N}_2(\lambda)}{n}\cdot\frac{1}{\lambda})=O_p(\frac{\lambda\mathcal{N}_1(\lambda)\mathcal{N}_2(\lambda)}{n}).
\end{aligned}
$$

For M_{13}, by the property that for two Hilbert–Schmidt operators A and B, $\|AB\|_{HS} \leqslant \|A\|_{HS}\|B\|_{op}$, it follows that

$$\begin{aligned}\|M_{13}\|^2 \leqslant & \|\lambda\Gamma^{1/2}W_n^{-1}(\Gamma_n+\lambda W_n^{-1})^{-1}(\Gamma-\Gamma_n)\|_{HS}^2 \cdot \|(\Gamma+\lambda W_n^{-1})^{-1}\|_{op}^2 \cdot \|(\Gamma-\Gamma_n)(\Gamma+\lambda W_n^{-1})^{-\frac{1}{2}}\|_{HS}^2 \\ & \cdot \|(\Gamma+\lambda W_n^{-1})^{-\frac{1}{2}}\delta_0\|_{op}^2 \\ = & O_p\Big(\frac{\mathcal{N}_2(\lambda)}{n\lambda}\cdot\frac{\mathcal{N}_2(\lambda)}{n}\cdot\mathcal{N}_1(\lambda)\Big) = O_p\Big(\frac{\mathcal{N}_1(\lambda)\mathcal{N}_2^2(\lambda)}{n^2\lambda}\Big).\end{aligned}$$

where $\|\cdot\|_{op}$ stands for the usual operator norm, that is, $\|\mathcal{F}\|_{op} = \sup_{\{h:\|h\|=1\}}\|\mathcal{F}h\|$ for an operator $\mathcal{F}$.

Secondly, employing a similar decomposition as described above, we express M_2 as the sum of two terms: $M_{21} = \Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i K^{1/2}x_i}{n}$ and $M_{22} = \Gamma^{1/2}(\Gamma_n W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i K^{1/2}x_i}{n}$ respectively. We proceed by considering the conditional expectation of $\|M_{21}\|^2$ with $\tau_j \to 0$ as $j\to\infty$, and we obtain

$$\begin{aligned}\mathrm{E}(\|M_{21}\|^2|x_i) &= \mathrm{E}\left[\mathrm{tr}(M_{21}\otimes M_{21})|x_i\right] \\ &= \frac{1}{n}\mathrm{tr}\Big(\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\mathrm{E}\left[\varepsilon_i^2(K^{1/2}x_i\otimes K^{1/2}x_i)\right]\Gamma^{1/2}(\Gamma W_n+\lambda I)^{-1}\Big) \\ &= O_p\Big(\frac{1}{n}\mathrm{tr}(\Gamma^2(\Gamma W_n+\lambda I)^{-2})\Big) \\ &= O_p\Big(\frac{1}{n}\sum_{j\geqslant 1}\frac{\tau_j^2}{(\tau_j w_j+\lambda)^2}\Big) \\ &\leqslant O_p\Big(\frac{\mathcal{N}_2(\lambda)}{n}\Big).\end{aligned}$$

For simplicity, we denote $\mathcal{S} = (\Gamma_n W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}$. By the

property $\mathrm{tr}(\mathcal{F}^T\mathcal{F}) = \mathrm{tr}(\mathcal{F}\mathcal{F}^T) = \|\mathcal{F}\|_{HS}^2$ for some finite operator $\mathcal{F}$, we obtain that

$$\begin{aligned}
\mathrm{E}(\|M_{22}\|^2|x_i) &= \mathrm{E}\left[\mathrm{tr}(M_{22}\otimes M_{22})|x_i\right] \\
&= \frac{1}{n}\mathrm{tr}\big(\mathcal{S}^T\mathrm{E}\left[\varepsilon_i^2(K^{1/2}x_i\otimes K^{1/2}x_i)\right]\mathcal{S}\big) \\
&= O_p\big(\frac{1}{n}\|\Gamma_n^{1/2}\mathcal{S}\|_{HS}^2\big) \\
&= O_p\big(\frac{1}{n}\|\Gamma_n^{1/2}(\Gamma_n W_n+\lambda I)^{-1}(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\|_{HS}^2\big) \\
&\leqslant O_p\big(\frac{1}{n}\|\Gamma_n^{1/2}(\Gamma_n W_n+\lambda I)^{-1}\|_{op}^2\cdot\|(\Gamma W_n-\Gamma_n W_n)(\Gamma W_n+\lambda I)^{-1}\|_{HS}^2\big) \\
&\leqslant O_p(\frac{1}{n\lambda})\cdot O_p(\frac{\mathcal{N}_2(\lambda)}{n}).
\end{aligned}$$

By integrating all the aforementioned five terms, we obtain

$$\|\Gamma^{1/2}(\hat{\delta}-\delta_0)\|^2 = O_p(\lambda^2\mathcal{N}_1(1+\frac{\mathcal{N}_2}{n\lambda}+\frac{\mathcal{N}_2^2}{n^2\lambda^2})+\frac{\mathcal{N}_2}{n}(1+\frac{1}{n\lambda})).$$

Since $\lambda \asymp n^{-(2\alpha+2r)/(2\alpha+2r+1)}$, we have $\dfrac{\mathcal{N}_2}{n\lambda}\to 0$ and $\dfrac{1}{n\lambda}\to 0$. Then, by Lemma 3.2,

$$\begin{aligned}
\|\hat{\beta}-\beta_0\|_C^2 &= \|C^{1/2}(\hat{\beta}-\beta_0)\|^2 = \|\Gamma^{1/2}(\hat{\delta}-\delta_0)\|^2 \\
&= O_p(\lambda^2\mathcal{N}_1+\frac{\mathcal{N}_2}{n}) = O_p(n^{-(2\alpha+2r)/(2\alpha+2r+1)}).
\end{aligned}$$

Since we assume that $\boldsymbol{\theta}$ is known, the convergence rate of $\hat{\beta}(\boldsymbol{\theta})$ is actually same as that in purely linear regression. Thus, we complete the proof of Theorem 3.1. □

Lemma 3.1 For a positive constant c, we have $\mathrm{E}\|(\Gamma W_n+\lambda I)^{-1/2}(\Gamma W_n-\Gamma_n W_n)\|_{\mathcal{HS}}^2 = O(\dfrac{\mathcal{N}_2(\lambda)}{n})$.

Proof. Assume that, for a random predictor, we have $K^{1/2}\boldsymbol{x} = \sum_{j=1}^{\infty} \xi_j e_j$ and $E(\xi_j^2) = \tau_j$,. Then, by the definition of Hilbert-Schmidt norm and the tensor product,

$$
\begin{aligned}
&\|(\Gamma W_n + \lambda I)^{-1/2}(K^{1/2}\boldsymbol{x} \otimes K^{1/2}\boldsymbol{x})W_n\|_{HS}^2 \\
&= \sum_{j,k} \langle (\Gamma W_n + \lambda I)^{-1/2}(K^{1/2}\boldsymbol{x} \otimes K^{1/2}\boldsymbol{x})W_n e_j, e_k\rangle^2 \\
&= \sum_{j,k} \xi_j^2 \langle (\Gamma W_n + \lambda I)^{-1/2} W_n K^{1/2}\boldsymbol{x}, e_k\rangle^2 \\
&= \sum_{j,k} \xi_j^2 \langle K^{1/2}\boldsymbol{x}, (\Gamma W_n + \lambda I)^{-1/2} W_n e_k\rangle^2 \\
&= \sum_{j,k} \xi_j^2 \langle K^{1/2}\boldsymbol{x}, w_j e_k/\sqrt{w_j\tau_j + \lambda}\rangle^2 \\
&= \sum_{j,k} \frac{(w_j \xi_j \xi_k)^2}{w_j\tau_j + \lambda}.
\end{aligned}
$$

Then, $\mathrm{E}\|(\Gamma W_n + \lambda I)^{-1}((K^{1/2}\boldsymbol{x} \otimes K^{1/2}\boldsymbol{x})W_n)\|_{\mathcal{HS}}^2 = \frac{1}{n}\sum_{j,k} \frac{w_j^2 \tau_j \tau_k}{w_j\tau_j + \lambda} \leqslant \frac{c}{n}\sum_j \frac{\tau_j}{w_j\tau_j + \lambda} = c\frac{\mathcal{N}_2(\lambda)}{n}$, which implies $\mathrm{E}\|(\Gamma W_n + \lambda I)^{-1/2}(\Gamma W_n - \Gamma_n W_n)\|_{\mathcal{HS}}^2 = O(\frac{\mathcal{N}_2(\lambda)}{n})$. □

Lemma 3.2 For a positive constant c, we have $\mathcal{N}_1(\lambda) = O(\lambda^{-1})$, and $\mathcal{N}_2(\lambda) = \lambda^{-1/(2\alpha+2r)}$.

Proof. We note that $\sum_{j=1} v_j^2 < \infty$,$\sum_{j=1} w_j^2 < \infty$ and $\tau_j \to 0$,

$$\begin{aligned}\mathcal{N}_1(\lambda) =& \sum_{j\geqslant 1} \frac{v_j^2}{w_j\tau_j + \lambda} \\ \leqslant& c \max_j \frac{1}{w_j\tau_j + \lambda} \\ \leqslant& c\lambda^{-1}.\end{aligned}$$

Let $J = \lambda^{-(2\alpha+2r)}$. Then, $\tau_j/(w_j\tau_j + \lambda) \leqslant 1$ for $j \leqslant J$ and $\tau_j/(w_j\tau_j + \lambda) \geqslant \tau_j/\lambda$ for $j > J$,

$$\begin{aligned}\mathcal{N}_2(\lambda) =& \sum_{j=1} \frac{\tau_j}{w_j\tau_j + \lambda} \\ \leqslant& c\sum_{j=1}^{J} 1 + c\lambda^{-1} \sum_{j\geqslant J+1} j^{-(2\alpha+2r)} = O(\lambda^{-(2\alpha+2r)}).\end{aligned}$$

□

Proof of Theorem 3.2

To obtain the asymptotic normality of $\boldsymbol{\theta}$, we denote $\hat{\delta}(\boldsymbol{\theta}) = K^{-1/2}\hat{\beta}(\boldsymbol{\theta})$, then our objective function is

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\hat{\delta}(\boldsymbol{\theta}), K^{1/2}x_i\rangle). \tag{3.19}$$

Note that $\hat{\delta}(\boldsymbol{\theta}) = \dfrac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1} \sum\limits_{i=1}^{n} K^{1/2}x_i(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta})$, we plug it into (3.19) and obtain that $\widehat{\boldsymbol{\theta}}$ is the minimizer of

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\frac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1} \sum_{i=1}^{n} K^{1/2}x_i(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta}), K^{1/2}x_i\rangle).$$

Taking derivative of $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and setting them equal to 0, it can be shown that the solution to the resulting estimating equations is the minimizer of weighted least-squares criterion:

$$\begin{aligned}\widetilde{Q}(\boldsymbol{\theta}) =& \sum_{i=1}^{n} w_i(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta} - \langle\frac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n} K^{1/2}x_i(y_i - \boldsymbol{z}_i^T\boldsymbol{\theta}), K^{1/2}x_i\rangle)^2 \\ =& \sum_{i=1}^{n} w_i(\varepsilon_i + \langle\delta_0, K^{1/2}x_i\rangle - \boldsymbol{z}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \langle\frac{1}{n}(\Gamma_n + \lambda W_n^{-1})^{-1}\sum_{i=1}^{n} K^{1/2}x_i(\varepsilon_i + \\ & \langle\delta_0, K^{1/2}x_i\rangle - \boldsymbol{z}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)), K^{1/2}x_i\rangle)^2,\end{aligned}$$

where $w_i = \psi(\varepsilon_i)/\varepsilon_i$. By solving the equation $\dfrac{\partial\widetilde{Q}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}$, we obtained the following results:

$$\begin{aligned}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 =& \frac{1}{n}\Big(\frac{\sum_{i=1}^{n} w_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T}{n}\Big)^{-1}\Big(\sum_{i=1}^{n} w_i\boldsymbol{\eta}_i(\varepsilon_i + \langle\delta_0, K^{1/2}x_i\rangle - \frac{1}{n}\langle(\Gamma_n + \lambda W_n^{-1})^{-1} \\ & \sum_{i=1}^{n} K^{1/2}x_i(\varepsilon_i + \langle\delta_0, K^{1/2}x_i\rangle), K^{1/2}x_i\rangle\Big),\end{aligned}$$

where $\boldsymbol{\eta}_i = \boldsymbol{z}_i - \langle(\Gamma_n + \lambda W_n^{-1})^{-1}\dfrac{\sum_{i=1}^{n} K^{1/2}x_i\boldsymbol{z}_i}{n}, K^{1/2}x_i\rangle$.

Based on the assumption (C4), there exists a function $g_j \in \mathcal{H}$, such that $\mathrm{E}(\boldsymbol{z}_j|X(\cdot)) = \langle X(\cdot), g_j(\cdot)\rangle$ for $j = 1, 2, \cdots, p$. Let $\boldsymbol{g}_0 = (g_{01}, g_{02}, \cdots, g_{0p})^T$ be the minimizer obtained by the robust optimization problem $\min_{\boldsymbol{g}} \mathrm{E}[\rho(\|\boldsymbol{z}_j - \langle X, \boldsymbol{g}\rangle\|_{\mathcal{H}}]$, and define $\boldsymbol{\gamma}_0 = K^{-1/2}\boldsymbol{g}_0$. Thus, $\widehat{\boldsymbol{\gamma}} = K^{-1/2}\widehat{\boldsymbol{g}} = (\Gamma_n + \lambda W_n^{-1})^{-1}\dfrac{\sum_{i=1}^{n} K^{1/2}x_i\boldsymbol{z}_i}{n}$ is a robust estimator of $\boldsymbol{\gamma}_0$ based on RKHS approach. Subsequently, we can represent

$\boldsymbol{\eta}_i = \boldsymbol{z}_i - \langle K^{1/2}x_i, \widehat{\boldsymbol{\gamma}}\rangle = \boldsymbol{z}_i - \langle x_i, \widehat{\boldsymbol{g}}\rangle$, and obtained

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T - \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)^T \\
&= \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \widehat{\boldsymbol{g}}\rangle)(\boldsymbol{z}_i - \langle x_i, \widehat{\boldsymbol{g}}\rangle)^T - \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)^T \\
&= \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)(\boldsymbol{z}_i - \langle x_i, \widehat{\boldsymbol{g}}\rangle)^T + \sum_{i=1}^{n} w_i \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle (\boldsymbol{z}_i - \langle x_i, \widehat{\boldsymbol{g}})^T \\
&- \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)^T \\
&= \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)\langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle^T + \sum_{i=1}^{n} w_i \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle + \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle)^T \\
&= \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)\langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle^T + \sum_{i=1}^{n} w_i \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)^T \\
&+ \sum_{i=1}^{n} w_i \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle^T \\
&= O_p(\sqrt{n}\|\boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\|) + O_p(n\|C_n^{1/2}(\boldsymbol{g}_0 - \widehat{\boldsymbol{g}})\|^2) = o_p(n). \qquad (3.20)
\end{aligned}
$$

Since $w_i(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)x_i$ has a zero mean within the framework of RKHS, the first and second terms above are of order $O_p(\sqrt{n}\|\boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\|)$, while the third term is of order

$O_p(n\|C_n^{1/2}(\boldsymbol{g}_0 - \widehat{\boldsymbol{g}})\|^2)$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \boldsymbol{\eta}_i \varepsilon_i = \sum_{i=1}^{n} \psi(\varepsilon_i)\boldsymbol{\eta}_i \\
&= \sum_{i=1}^{n} \psi(\varepsilon_i)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle) + \sum_{i=1}^{n} \psi(\varepsilon_i)(\langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle) \\
&= \sum_{i=1}^{n} \psi(\varepsilon_i)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle) + \langle \sum_{i=1}^{n} \psi(\varepsilon_i) x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle \\
&= \sum_{i=1}^{n} \psi(\varepsilon_i)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle) + O_p(\sqrt{n}\|\boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\|) \\
&= \sum_{i=1}^{n} \psi(\varepsilon_i)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle) + o_p(n). \qquad (3.21)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \boldsymbol{\eta}_i (\langle \delta_0, K^{1/2} x_i\rangle - \frac{1}{n}\langle (\Gamma_n + \lambda W_n^{-1})^{-1} \sum_{i=1}^{n} K^{1/2} x_i (\varepsilon_i + \langle \delta_0, K^{1/2} x_i\rangle - \boldsymbol{z}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)), K^{1/2} x_i\rangle) \\
&= \sum_{i=1}^{n} w_i \boldsymbol{\eta}_i (\langle \delta_0, K^{1/2} x_i\rangle - \langle \hat{\delta}^*, K^{1/2} x_i\rangle) \\
&= \sum_{i=1}^{n} w_i \boldsymbol{\eta}_i \langle \delta_0 - \hat{\delta}^*, K^{1/2} x_i\rangle \\
&= \sum_{i=1}^{n} w_i (\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)\langle \delta_0 - \hat{\delta}^*, K^{1/2} x_i\rangle + \sum_{i=1} w_i \langle x_i, \boldsymbol{g}_0 - \widehat{\boldsymbol{g}}\rangle \langle \delta_0 - \hat{\delta}^*, K^{1/2} x_i\rangle \\
&= O_p(\sqrt{n}\|\delta_0 - \hat{\delta}^*\|) + O_p(n\|\Gamma_n^{1/2}(\boldsymbol{g}_0 - \widehat{\boldsymbol{g}})\| \cdot \|\delta_0 - \hat{\delta}^*\|) = o_p(n). \qquad (3.22)
\end{aligned}$$

where $\hat{\delta}^*$ is simply the RKHS estimator of δ in a purely functional linear model.

According to (3.20)-(3.22), the dominant term in $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is

$$\begin{aligned}
&\frac{1}{n}(\sum_{i=1}^{n} \frac{\psi(\varepsilon_i)}{\varepsilon_i}(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle)^T)^{-1} \sum_{i=1}^{n} \psi(\varepsilon_i)(\boldsymbol{z}_i - \langle x_i, \boldsymbol{g}_0\rangle) \\
&= \frac{1}{n}(\sum_{i=1}^{n} \frac{\psi(\varepsilon_i)}{\varepsilon} \boldsymbol{u}_i \boldsymbol{u}_i^T)^{-1} \sum_{i=1}^{n} \psi(\varepsilon_i)\boldsymbol{u}_i.
\end{aligned}$$

Denote $\boldsymbol{\Delta}_1 = E[\psi'(\varepsilon_i)\boldsymbol{u}_i\boldsymbol{u}_i^T], \boldsymbol{\Delta}_2 = E[\psi^2(\varepsilon_i)(\boldsymbol{u}_i\boldsymbol{u}_i^T)]$. Then, by the central limit theorem and condition (C5), we have the asymptotic normality of $\hat{\boldsymbol{\theta}}$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_2\boldsymbol{\Delta}_1^{-1}).$$

□

3.7.2 Additional Simulation Results

Table 3.6: MISE of $\hat{\boldsymbol{\beta}}(t)$ based on RKHS approach with different sample size and error distributions, where λ is selected by generalized cross-validation, $v = 1.2$.

Sample size	Scenario	Setting	Distribution of Error	Method				
				LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	0.0375	0.0544	0.0404	0.0405	0.0379
	Scenario II		mixed normal	0.2436	0.1095	0.1571	0.1420	0.1720
			$t(3)$	0.0906	0.0835	0.0721	0.0834	0.0724
	Scenario III	Setting 1	mixed normal	0.2253	0.1138	0.1639	0.1517	0.1994
			$t(3)$	0.0975	0.0896	0.0839	0.0695	0.0740
		Setting 2	mixed normal	0.2250	0.2069	0.1879	0.2003	0.1787
			$t(3)$	0.0956	0.0821	0.0752	0.0811	0.0695
		Setting 3	mixed normal	0.2430	0.1357	0.1681	0.2077	0.2132
			$t(3)$	0.0835	0.0813	0.0766	0.0818	0.0655
$n = 300$	Scenario I		normal	0.0145	0.0218	0.0127	0.0106	0.0153
	Scenario II		mixed normal	0.0909	0.0304	0.0396	0.0414	0.0480
			$t(3)$	0.0269	0.0181	0.0182	0.0180	0.0227
	Scenario III	Setting 1	mixed normal	0.0945	0.0380	0.0416	0.0450	0.0410
			$t(3)$	0.0323	0.0321	0.0196	0.0199	0.0228
		Setting 2	mixed normal	0.0675	0.0337	0.0426	0.0489	0.0429
			$t(3)$	0.0276	0.0256	0.0213	0.0243	0.0280
		Setting 3	mixed normal	0.0689	0.0470	0.0411	0.0677	0.0457
			$t(3)$	0.0328	0.0356	0.0278	0.0324	0.0277
$n = 500$	Scenario I		normal	0.0093	0.0125	0.0087	0.0102	0.0084
	Scenario II		mixed normal	0.0628	0.0172	0.0204	0.0241	0.0234
			$t(3)$	0.0198	0.0184	0.0130	0.0148	0.0155
	Scenario III	Setting 1	mixed normal	0.0480	0.0304	0.0247	0.0216	0.0249
			$t(3)$	0.0211	0.0156	0.0139	0.0128	0.0154
		Setting 2	mixed normal	0.0480	0.0287	0.0234	0.0260	0.0274
			$t(3)$	0.0192	0.0201	0.0116	0.0183	0.0192
		Setting 3	mixed normal	0.0573	0.0287	0.0424	0.0537	0.0277
			$t(3)$	0.0214	0.0271	0.0249	0.0279	0.0210

Table 3.7: MSE of $\hat{\boldsymbol{\theta}}$ based on RKHS approach with different sample size and error distributions, where λ is selected by generalized cross-validation, $v = 1.2$.

Sample size	Scenario	Setting	Distribution of Error	Parameters									
				θ_1					θ_2				
				LS	LAD	Huber	Bisquare	Hampel	LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	0.0168	0.0240	0.0184	0.0149	0.0163	0.0067	0.0109	0.0088	0.0096	0.0074
	Scenario II		mixed normal	0.1476	0.0485	0.0609	0.0401	0.0456	0.0672	0.0264	0.0220	0.0197	0.0226
			$t(3)$	0.0381	0.0298	0.0260	0.0322	0.0309	0.0226	0.0180	0.0137	0.0130	0.0158
	Scenario III	Setting 1	mixed normal	0.1341	0.0729	0.0472	0.0338	0.0347	0.0500	0.0399	0.0229	0.0197	0.0240
			$t(3)$	0.0370	0.0353	0.0250	0.0314	0.0285	0.0190	0.0161	0.0138	0.0131	0.0153
		Setting 2	mixed normal	0.1427	0.0948	0.0498	0.0448	0.0387	0.0684	0.0290	0.0222	0.0249	0.0160
			$t(3)$	0.0516	0.0347	0.0246	0.0336	0.0302	0.0216	0.0167	0.0162	0.0151	0.0156
		Setting 3	mixed normal	0.2480	0.0542	0.0624	0.0640	0.1052	0.1423	0.0234	0.0368	0.0407	0.0534
			$t(3)$	0.0443	0.0335	0.0283	0.0379	0.0339	0.0206	0.0160	0.0144	0.0168	0.0159
$n = 300$	Scenario I		normal	0.0043	0.0055	0.0038	0.0043	0.0042	0.0017	0.0031	0.0020	0.0026	0.0018
	Scenario II		mixed normal	0.0404	0.0111	0.0080	0.0060	0.0061	0.0182	0.0054	0.0041	0.0036	0.0033
			$t(3)$	0.0133	0.0087	0.0055	0.0071	0.0063	0.0066	0.0038	0.0033	0.0033	0.0031
	Scenario III	Setting 1	mixed normal	0.0402	0.0148	0.0079	0.0057	0.0054	0.0153	0.0090	0.0040	0.0029	0.0036
			$t(3)$	0.0120	0.0109	0.0074	0.0075	0.0052	0.0059	0.0057	0.0032	0.0036	0.0030
		Setting 2	mixed normal	0.0273	0.0123	0.0087	0.0060	0.0060	0.0195	0.0047	0.0037	0.0033	0.0035
			$t(3)$	0.0120	0.0092	0.0066	0.0068	0.0079	0.0057	0.0035	0.0030	0.0038	0.0033
		Setting 3	mixed normal	0.0360	0.0085	0.0059	0.0111	0.0068	0.0205	0.0059	0.0027	0.0063	0.0030
			$t(3)$	0.0116	0.0108	0.0079	0.0079	0.0065	0.0060	0.0048	0.0035	0.0032	0.0037
$n = 500$	Scenario I		normal	0.0022	0.0033	0.0023	0.0023	0.0022	0.0011	0.0018	0.0013	0.0013	0.0010
	Scenario II		mixed normal	0.0197	0.0052	0.0043	0.0030	0.0033	0.0102	0.0028	0.0023	0.0019	0.0015
			$t(3)$	0.0056	0.0053	0.0040	0.0028	0.0039	0.0022	0.0021	0.0017	0.0019	0.0018
	Scenario III	Setting 1	mixed normal	0.0194	0.0097	0.0047	0.0029	0.0032	0.0137	0.0048	0.0022	0.0017	0.0020
			$t(3)$	0.0052	0.0034	0.0034	0.0036	0.0037	0.0034	0.0019	0.0018	0.0013	0.0020
		Setting 2	mixed normal	0.0214	0.0054	0.0040	0.0028	0.0031	0.0112	0.0024	0.0020	0.0017	0.0017
			$t(3)$	0.0069	0.0050	0.0034	0.0038	0.0036	0.0035	0.0026	0.0020	0.0022	0.0021
		Setting 3	mixed normal	0.0191	0.0054	0.0057	0.0057	0.0031	0.0112	0.0024	0.0039	0.0025	0.0015
			$t(3)$	0.0080	0.0052	0.0038	0.0042	0.0045	0.0033	0.0027	0.0024	0.0020	0.0021

Table 3.8: MAPE (Mean Absolute Percentage Error) and SE (Standard Error, in the brackets) of prediction errors over 500 simulation with different sample size and error distributions, where λ is selected by generalized cross-validation, $v = 1.2$.

Sample size	Scenario	Setting	Distribution of Error	Method				
				LS	LAD	Huber	Bisquare	Hampel
$n = 100$	Scenario I		normal	0.0445(0.2183)	0.0525(0.2566)	0.0446(0.2186)	0.0451(0.2188)	0.0445(0.2183)
	Scenario II		mixed normal	0.0958(0.5908)	0.0895(0.5751)	0.0862(0.5597)	0.0859(0.5627)	0.0888(0.5798)
			$t(3)$	0.1271(0.3369)	0.0954(0.4045)	0.1090(0.3284)	0.0784(0.3347)	0.1144(0.3311)
	Scenario III	Setting 1	mixed normal	0.2068(0.5224)	0.1124(0.4926)	0.1429(0.4898)	0.1972(0.4993)	0.1680(0.5016)
			$t(3)$	0.0537(0.2804)	0.0337(0.2795)	0.0517(0.2728)	0.0524(0.2763)	0.0500(0.2730)
		Setting 2	mixed normal	0.0674(0.5437)	0.0658(0.5684)	0.0604(0.5108)	0.0590(0.5206)	0.0625(0.5235)
			$t(3)$	0.0575(0.3252)	0.0538(0.2968)	0.0503(0.2886)	0.0507(0.2956)	0.0505(0.2901)
		Setting 3	mixed normal	0.0654(0.5481)	0.0588(0.5175)	0.0590(0.5162)	0.0591(0.5231)	0.0545(0.5090)
			$t(3)$	0.0356(0.3053)	0.0334(0.3101)	0.0332(0.2964)	0.0334(0.3012)	0.0333(0.2985)
$n = 300$	Scenario I		normal	0.0260(0.0893)	0.0273(0.0948)	0.0259(0.0897)	0.0259(0.0899)	0.0260(0.0894)
	Scenario II		mixed normal	0.1006(0.2712)	0.0754(0.2609)	0.0795(0.2623)	0.0790(0.2619)	0.0782(0.2619)
			$t(3)$	0.0548(0.1477)	0.0469(0.1477)	0.0414(0.1441)	0.0449(0.1446)	0.0452(0.1445)
	Scenario III	Setting 1	mixed normal	0.0720(0.2740)	0.0533(0.2659)	0.0551(0.2651)	0.0528(0.2654)	0.0555(0.2660)
			$t(3)$	0.0425(0.1491)	0.0404(0.1518)	0.0364(0.1469)	0.0369(0.1475)	0.0358(0.1470)
		Setting 2	mixed normal	0.0618(0.2771)	0.0529(0.2673)	0.0517(0.2679)	0.0513(0.2684)	0.0512(0.2687)
			$t(3)$	0.0349(0.1495)	0.0344(0.1505)	0.0327(0.1467)	0.0326(0.1472)	0.0322(0.1468)
		Setting 3	mixed normal	0.0653(0.2751)	0.0521(0.2693)	0.0560(0.2695)	0.0566(0.2711)	0.0524(0.2657)
			$t(3)$	0.0447(0.1491)	0.0450(0.1538)	0.0443(0.1483)	0.0451(0.1490)	0.0444(0.1478)
$n = 500$	Scenario I		normal	0.0413(0.0672)	0.0422(0.0717)	0.0427(0.0688)	0.0431(0.0691)	0.0426(0.0682)
	Scenario II		mixed normal	0.0934(0.2064)	0.0794(0.2023)	0.0798(0.2024)	0.0779(0.2025)	0.0780(0.2025)
			$t(3)$	0.0497(0.1108)	0.0486(0.1095)	0.0474(0.1091)	0.0471(0.1094)	0.0469(0.1089)
	Scenario III	Setting 1	mixed normal	0.0648(0.2070)	0.0579(0.2042)	0.0578(0.2022)	0.0573(0.2022)	0.0578(0.2025)
			$t(3)$	0.0389(0.1162)	0.0373(0.1129)	0.0371(0.1124)	0.0369(0.1127)	0.0370(0.1123)
		Setting 2	mixed normal	0.0671(0.2079)	0.0570(0.2037)	0.0582(0.2030)	0.0577(0.2031)	0.0577(0.2033)
			$t(3)$	0.0671(0.2079)	0.0570(0.2037)	0.0582(0.2030)	0.0577(0.2031)	0.0577(0.2033)
		Setting 3	mixed normal	0.0714(0.2077)	0.0571(0.2058)	0.0583(0.2037)	0.0549(0.2036)	0.0576(0.2044)
			$t(3)$	0.0331(0.1148)	0.0316(0.1184)	0.0327(0.1147)	0.0328(0.1151)	0.0326(0.1147)

4 Robust Hypothesis Testing in Functional Linear Regression

4.1 Introduction

In the studies of functional data, a common problem is to explore the relationship between functional covariates and functional or scalar response, which is an active area of research in FDA. A popular method to address this problem is the functional linear model (FLM, hereinafter) analysis, which was first introduced by (Ramsay and Dalzell, 1991). The basic idea of FLM is to extend the classical linear model to functional data, and one may reference the book of Ramsay and Silverman (2005) for the details.

Due to the infinite dimensionality of the functional covariates, the testing problem about the association of the covariates and the response is of great interest in the research area of FLM. The popular strategy to reduce the dimension is to represent the functional covariates and coefficient function by linear combinations of a set of

pre-determined basis functions, such as B-splines, Fourier and wavelet bases or eigen-basis from the functional principal analysis. After such pre-treatment, the testing problem of functional data has transformed into a testing problem under the classical linear model. Numerous works in literature have used this strategy. For example, Cardot et al. (2003) studied the hypothesis testing in the functional linear model with the scalar response. They introduced two test statistics based on the norm of the empirical cross-covariance operators and used penalized B-spline estimator of the coefficient function in the simulation studies. Kong et al. (2016a) re-expressed the functional linear model as a standard linear model through functional principal component analysis method and considered the classical tests such as Wald, score, likelihood ratio and F test. Su et al. (2017) proposed a new method of selecting principal components (PCs) for constructing the Wald-type test statistics to avoid the sensitive performance of the power in the classical test statistics with varying thresholds.

In this chapter, we studies robust hypothesis testing procedures in the functional linear model. Our motivation of this study is to use M-estimators in the testing procedure, which has been studied in the classical linear regression (Maronna et al., 2019) but has not been studied in functional data analysis. Our objective is to show that the robust versions of Wald-type, the likelihood ratio-type, and F-type

tests can be used in functional data analysis. The proposed testing procedures are scale equivariant since we consider a residual scale estimator such that they have the protection against high-leverage outliers. Moreover, our simulation results and real data studies show that the robust test procedures are more stable in contaminated datasets and have the higher estimated power values than those based on the classical testing procedure.

This chapter is organized as follows. Section 4.2 describes the FLM and introduces the robust testing procedures for testing whether the coefficient function is zero. The asymptotic properties of the proposed procedures are also established in Section 4.2. Simulation studies are performed in Section 4.3. Real data examples and conclusions are given in Sections 4.4-4.5, respectively. All proofs and some additional numerical simulation results are shown in Appendix.

4.2 Methodology

4.2.1 Model Specification

We consider the FLM in which the response of interest is scalar while the covariate is a function. Let y be a real-valued random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, $X(t) \in L^2(\mathcal{T})$ is functional covariate defined on a compact support $\mathcal{T}$.

Suppose that the mean and covariance function of $X(t)$ are $\mu(t) = \mathrm{E}(X(t))$ and $\Sigma(s,t) = \mathrm{Cov}(X(s) - \mu(s), X(t) - \mu(t))$, respectively. Then our FLM with scalar response is expressed as follows:

$$y = \alpha_0 + \langle X, \beta \rangle + \varepsilon_0 = \alpha_0 + \int_{\mathcal{T}} \beta(t) X(t) dt + \varepsilon_0, \tag{4.1}$$

where $\beta(t)$ is a smooth square integrable coefficient function on $\mathcal{T}$, α_0 is an unknown intercept and ε_0 is a random error with zero mean and variance $\sigma^2_{\varepsilon_0}$. Denote centered functional covariate $\widetilde{X}(t) = X(t) - \mu(t)$ and $\alpha = \alpha_0 + \int_{\mathcal{T}} \beta(t)\mu(t)dt$, then the equivalent model is

$$y = \alpha + \int_{\mathcal{T}} \beta(t) \widetilde{X}(t) dt + \varepsilon_0. \tag{4.2}$$

One of the popular techniques for dimension reduction in FDA is the functional principal component analysis (FPCA, hereinafter). Suppose that the spectral decom-

position of the covariance function is $\Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where $\{\lambda_k,\ k \geqslant 1\}$ is a set of eigenvalues with non-increasing order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$, $\sum_{k=1}^{\infty} \lambda_k < \infty$, and $\{\phi_k(\cdot), k \geqslant 1\}$ are the corresponding orthonormal eigenfunctions. Use the Karhunen-Loève representation, we have

$$\widetilde{X}(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t),$$

where $\{\xi_k = \int_{\mathcal{T}} \widetilde{X}(t)\phi_k(t)dt, k \geqslant 1\}$ are principal component scores with $\mathrm{E}(\xi_k) = 0, \mathrm{Var}(\xi_k) = \mathrm{E}(\xi_k^2) = \lambda_k$, and $\mathrm{E}(\xi_k \xi_{k'}) = 0$, for $k \neq k'$.

Use the same basis functions, the coefficient function can be expanded by $\beta(t) = \sum_{k=1}^{\infty} \beta_k \phi_k(t)$, where $\beta_k = \int_{\mathcal{T}} \beta(t)\phi_k(t)dt$ is the unknown basis coefficient. Then model (4.2) can be equivalently written as

$$y = \alpha + \sum_{k=1}^{\infty} \beta_k \int_{\mathcal{T}} \widetilde{X}(t)\phi_k(t)dt + \varepsilon_0 = \alpha + \sum_{k=1}^{\infty} \xi_k \beta_k + \varepsilon_0. \tag{4.3}$$

This model is still impractical because of its infinite sum. In many existing methods, the main technique is to approximate the model by truncating $\sum_{k=1}^{\infty} \xi_k \beta_k$ to $\sum_{k=1}^{k_n} \xi_k \beta_k$, where k_n is a finite number and increases with the sample size n. The truncation value k_n can be seen as a "cut-off" level. In practice, there are some empirical choices of the "cut-off" level, such as PVE (Percentage of Variance Explained, hereinafter) method, equivalently leading PCs (Principal Components, hereinafter) method (Cardot et al., 2003; Kong et al., 2013; Swihart et al., 2014), CV (Cross-Validation, hereinafter)

criterion (Qingguo, 2017), and information (AIC or BIC) criterion (Kato et al., 2012). The truncated model, which employs a finite number of terms k_n, can be expressed as follows:

$$y = \alpha + \sum_{k=1}^{k_n} \xi_k \beta_k + \varepsilon, \tag{4.4}$$

where $\varepsilon = \sum_{k=k_n+1}^{\infty} \beta_k \xi_k + \varepsilon_0$ have zero mean and large variance $\sigma_\varepsilon^2 = \sum_{k=k_n+1}^{\infty} \lambda_k \beta_k^2 + \sigma_{\varepsilon_0}^2$(Su et al., 2017). Our robust testing procedures in the next section are based on the above truncated model. In the later simulation studies, we report the results based on the different PVE threshold levels.

4.2.2 Testing Procedure Based on M-estimation

The statistical problem of our interest is to test whether there has relationship between the covariate function $X(t)$ and the scalar response y. Therefore, the null hypothesis is to test if the coefficient function $\beta(t)$ is equal to zero:

$$H_0 : \beta(t) = 0 \quad \text{for any } t \in \mathcal{T} \quad vs. \quad H_a : \beta(t) \neq 0 \quad \text{for } t \text{ in some open subset in } \mathcal{T}. \tag{4.5}$$

For alternative hypothesis $H_a : \beta(t) \neq 0$, we can consider a certain known function $\beta(t) = \beta_a(t) \neq 0$ for t in some open subset in $\mathcal{T}$. Then, under our truncated FLM model (4.4), the hypothesis testing problem (4.5) is converted to the following

testing problem:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_{k_n} = 0 \quad vs. \\ H_a &: \beta_k^a \neq 0,\ \beta_k^a = \int_{\mathcal{T}} \beta_a(t)\phi_k(t)dt, \quad \text{for at least one } k, 1 \leqslant k \leqslant k_n. \end{aligned} \tag{4.6}$$

Let $\mathcal{G} = \{(x_i(t), y_i) : i = 1, 2, \cdots, n\}$ be independent realizations of $(X(t), y)$ generated by the model (4.1) and the errors ε_i are independent and identically distributed with finite variance and zero mean. The empirical version of $\Sigma(s,t)$ is

$$\widehat{\Sigma}(s,t) = \frac{1}{n}\sum_{i=1}^{n}[x_i(s) - \bar{x}(s)][x_i(t) - \bar{x}(t)] = \sum_{k}^{\infty} \hat{\lambda}_k \hat{\phi}_k(s)\hat{\phi}_k(t), \qquad s, t \in \mathcal{T}$$

where $\bar{x}(\cdot) = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i(\cdot)$, and $\{(\hat{\lambda}_k, \hat{\phi}_k), \hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant 0,\ k \geqslant 1\}$ are eigenvalue and eigenfunction pairs of $\widehat{\Sigma}(s,t)$.

Denote $\widetilde{x}_i(t) = x_i(t) - \bar{x}(t)$, $\hat{\xi}_{ik} = \displaystyle\int_{\mathcal{T}} \widetilde{x}_i(t)\hat{\phi}_k(t)dt$, and $\hat{\lambda}_k = E\hat{\xi}_{ik}^2$. Based on FPCA method, the approximate solution of estimator $\hat{\beta}(t)$ is $\sum\limits_{k=1}^{k_n} \hat{\beta}_k \hat{\phi}_k(t)$, where $\hat{\alpha}$ and $\hat{\beta}_k$ are obtained by solving the following minimization problem

$$\min \sum_{i=1}^{n} \rho((y_i - \alpha - \sum_{k=1}^{k_n} \hat{\xi}_{ik}\beta_k)/\hat{\sigma}_\varepsilon) \tag{4.7}$$

with $\rho(\cdot)$ being the loss function.

Remark 4.1 In equation (4.7), $\hat{\sigma}_\varepsilon$ is a preliminary scale estimation of errors. A popular approach for estimating σ_ε is to use the normalized MAD (median absolute deviation about the median) of residuals, say, $\hat{\sigma}_\varepsilon = \dfrac{1}{0.675}\text{MAD}(r_i)$ with $\text{MAD}(r_i) = \text{median}\left\{|r_i - \text{median}(r_i)|\right\}$(Holland and Welsch, 1977; Yohai, 1974).

Remark 4.2 The loss function $\rho(\cdot)$ is a suitable function which typically is convex, symmetric about 0, and bounded. In our simulation studies, we choose Turkey's bisquare function ($\rho(x) = \frac{c^2}{6}\{1 - \left[1 - (\frac{x}{c})^2\right]^3\}\mathbb{1}_{\{|x|\leqslant c\}} + \frac{c^2}{6}\mathbb{1}_{\{|x|>c\}},\ c > 0$), where the tuning parameter $c = 4.685$. Even though this loss function is not convex, it works well in dealing with extreme outliers. The additional simulation results based on Huber's loss function are given in the supplementary material.

Let ψ be the derivative of ρ. Then, (4.7) is equivalent to the following local M-estimation equation

$$\sum_{i=1}^{n} \psi((y_i - \alpha - \sum_{k=1}^{k_n} \hat{\xi}_{ik}\beta_k)/\hat{\sigma}_\varepsilon)\hat{\xi}_{ik} = 0. \tag{4.8}$$

Denote $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top$ with $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{k_n})^\top$ being unknown parameter vector, and $\hat{\boldsymbol{\theta}}_R = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_R^\top)^\top$ with $\hat{\boldsymbol{\beta}}_R = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_{k_n})^\top$ being the solution of the estimating equation (4.8). Note that only the slope function but not intercept α is of our interest, then $\hat{\boldsymbol{\beta}}_R$ is M-estimator in model (4.4). Thus, our robust Wald-type test statistic is defined as

$$T_{RW} = \hat{\boldsymbol{\beta}}_R^\top (Var(\hat{\boldsymbol{\beta}}_R))^{-1} \hat{\boldsymbol{\beta}}_R, \tag{4.9}$$

where $Var(\hat{\boldsymbol{\beta}}_R)$ is $k_n \times k_n$ matrix $\hat{\tau}^2(\widehat{\Xi}^\top\widehat{\Xi})^{-1}$ in which $\widehat{\Xi} = [\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_i, \ldots, \hat{\boldsymbol{\xi}}_n]^\top$, $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, \hat{\xi}_{i2}, \cdots, \hat{\xi}_{ik_n})^\top, i = 1, 2, \ldots, n$, and $\hat{\tau}$ is a consistent estimator of the asymptotic covariance matrix standardizing constant $\tau = \sigma_\varepsilon^2 \frac{E\psi^2(\varepsilon/\sigma_\varepsilon)}{(E\psi'(\varepsilon/\sigma_\varepsilon))^2}$. Denote $r_i = y_i - \widehat{y}_i$.

Then, $\hat{\tau}$ can be estimated by

$$\hat{\tau} = \hat{\sigma}_{\varepsilon}^2 \frac{n^{-1}\sum_{i=1}^{n}\psi^2(r_i/\hat{\sigma}_{\varepsilon})}{(n^{-1}\sum_{i=1}^{n}\psi'(r_i/\hat{\sigma}_{\varepsilon}))^2}\frac{n}{n-(k_n+1)}, \tag{4.10}$$

where $\hat{\sigma}_{\varepsilon}$ is previously given scale parameter estimator (see Remark 1).

Next, we consider the likelihood ratio type test statistic. Let $\tilde{\boldsymbol{\theta}}_R = (\tilde{\alpha}, \mathbf{0}_{k_n}^{\top})^{\top}$ be the M-estimator satisfying equation (4.7) but computed under the null model, where $\mathbf{0}_{k_n}^{\top} = (0, 0, \ldots, 0)^{\top}$ is k_n-dimensional zero vector. Then based on a pseudolikelihood ratio, a robust likelihood ratio type test statistic is given as

$$T_{RL} = \sum_{i=1}^{n}\rho\left(\frac{r_i(\tilde{\boldsymbol{\theta}}_R)}{\hat{\sigma}_{\varepsilon}}\right) - \sum_{i=1}^{n}\rho\left(\frac{r_i(\hat{\boldsymbol{\theta}}_R)}{\hat{\sigma}_{\varepsilon}}\right). \tag{4.11}$$

Finally, we define the F test in terms of the residual sum of squares under the full and null models. According to Schrader and Hettmansperger (1980), a robust F type test statistic can be constructed as

$$T_{RF} = \{k_n\lambda\}^{-1}\Big\{\sum_{i=1}^{n}\rho\left(\frac{r_i(\tilde{\boldsymbol{\theta}}_R)}{\hat{\sigma}_{\varepsilon}}\right) - \sum_{i=1}^{n}\rho\left(\frac{r_i(\hat{\boldsymbol{\theta}}_R)}{\hat{\sigma}_{\varepsilon}}\right)\Big\}, \tag{4.12}$$

where $\lambda = \dfrac{E\psi^2(\varepsilon/\sigma_{\varepsilon})}{2E\psi'(\varepsilon/\sigma_{\varepsilon})}$, and $\hat{\lambda} = \dfrac{1}{(n-(k_n+1))}\dfrac{\sum_{i=1}^{n}\psi^2(r_i/\hat{\sigma}_{\varepsilon})}{2n^{-1}\sum_{i=1}^{n}\psi'(r_i/\hat{\sigma}_{\varepsilon})}$. Similar to the robust likelihood ratio type test, the computation of F test statistic also requires the fitting of both the full and null models, while Wald test only requires the fitting of the full model.

Kong et al. (2016a) pointed out that the testing problems considered in the truncated functional linear model were analogue to the multivariate covariates situation but with a few differences: (1) The number of principal components selected (i.e. the truncation number) k_n that diverges with the sample size n is not known and needs to be pre-estimated based on some criterions; (2) the covariates $\{\xi_{ik}, i = 1, 2, \ldots, n, k = 1, 2, \ldots, k_n\}$ (i.e. the principal components) are estimated through the empirical covariance function and can not be observed directly; (3) the previously scale parameter estimation $\hat{\sigma}_\varepsilon$ is dependent on the truncation number k_n.

4.2.3 Asymptotic Distributions under the Null and Alternatives

As discussed in the last section, our robust testing procedures are based on the truncated model (4.4), and the selected truncation level k_n may diverge as $n \to \infty$. In the following, we first make some assumptions required for our theoretical developments and then present the results of asymptotic distributions of the test statistics under both null and alternative hypotheses. Note that we use a generic constant C whose value may be changed from line to line throughout these assumptions.

(A1) $X(t)$ have finite fourth moment, i.e., $\int_{\mathcal{T}} E[X(t)]^4 dt < \infty$, and for each k, $E(\xi_k^4) < C\lambda_k^2$ for some constant C.

(A2) The eigenvalues $\lambda_k,\ k = 1, 2, \cdots$, satisfy that

$$C^{-1}k^{-a} \leqslant \lambda_k \leqslant Ck^{-a} \qquad \text{and} \qquad \lambda_k - \lambda_{k+1} \geqslant C^{-1}k^{-a-1} \text{ for some } a > 1.$$

For the Fourier coefficients β_k , there exist constant C and $b > a/2 + 1$ such that $|\beta_k| \leqslant Ck^{-b}$ for all $k \geqslant 1$.

(A3) The number of principal components selected, k_n, satisfies that $k_n \to \infty$ as $n \to \infty$, and $k_n \asymp n^{1/(a+2b)}$, which means that the ratio $\dfrac{k_n}{n^{1/(2a+b)}}$ is bounded away from both zero and infinity.

(A4) The function $\rho(\cdot)$ is convex and nonmonotone, and $E\psi(\varepsilon_i|X_i) = 0, E(\psi^2(\varepsilon_i)|X_i) < \infty$ for $i = 1, 2, \cdots, n$.

(A5) There exists constant $0 < C < \infty$, such that $\sup\limits_{i\in n} E([\psi(\varepsilon_i + u) - \psi(\varepsilon_i)]^2|X_i) < C|u|$, as $u \to 0$.

(A6) There exist a positive function $g(X_i)$ with $0 < g(X_i) < \infty$ and some constant $0 < C, S < \infty$ such that$|E(\psi(\varepsilon_i + v|X_i) - g(X_i)v| \leqslant Cv^2$ for $|v| \leqslant S$.

(A7) $h_n = o(k_n^{-\frac{2}{3}})$, where $h_n = \max_{i\leqslant n} h_{ii}$ with h_{ii} being the ith diagonal element of the projection matrix $H = \Xi(\Xi^\top\Xi)^{-1}\Xi$.

Assumption (A1) is a common moment condition imposed for the estimation of functional predictor (Kato et al., 2012), which ensures that $X(t)$ has light tails

so that the empirical covariance has $\sqrt{n}$ consistency. Assumptions (A2)-(A3) are adapted from Hall et al. (2007), which implies that all λ_k's are positive and guarantees the identification of the coefficient function. The second part of Assumption (A2) requires that the spacing between adjacent eigenvalues not be small, which ensures identifiability of the eigenfunctions. Assumption (A3) allows k_n to diverge, and gives the order of the truncation number k_n. Combined with Assumption (A2), (A3) implies that the divergence of the number of functional principal components with n depends on the spacing between adjacent eigenvalues (Kong et al., 2016a). Assumptions (A4)-(A7) are some regularity conditions about the score function $\psi(\cdot)$ and robust estimates, which are adapted from Yohai and Maronna (1979) and Huber et al. (1973).

Theorem 4.1 Assume that $X_i(t) \in L^2(\mathcal{T})$ for $i = 1, 2, \cdots, n$, $\lambda = \dfrac{E\psi^2(\varepsilon/\sigma_\varepsilon)}{2E\psi'(\varepsilon/\sigma_\varepsilon)}$, and assumptions (A1) -(A7) hold. Then, under the null hypothesis H_0: $\beta(t) = 0$ for any $t \in \mathcal{T}$, we have that:

(1) $(T_{RW} - k_n)/\sqrt{2k_n} \xrightarrow{d} N(0, 1)$,

(2) $(\lambda^{-1}T_{RL} - k_n)/\sqrt{2k_n} \xrightarrow{d} N(0, 1)$,

(3) $(k_n T_{RF} - k_n)/\sqrt{2k_n} \xrightarrow{d} N(0, 1)$,

as $k_n \to \infty$.

Theorem 4.2 Assume that $X_i(t) \in L^2(\mathcal{T})$ for $i = 1, 2, \cdots, n$, $\lambda = \dfrac{E\psi^2(\varepsilon/\sigma_\varepsilon)}{2E\psi'(\varepsilon/\sigma_\varepsilon)}$,

and the assumptions (A1)-(A7)hold. Then, under alternative hypothesis H_a: $\beta(t) = \beta_a(t) \neq 0$ for some unknown real valued function $\beta_a(t)$ defined in $\mathcal{T}$, we have that:

(1) $\dfrac{T_{RW} - (k_n + \eta_n)}{\sqrt{2(k_n + 2\eta_n)}} \xrightarrow{d} N(0,1)$,

(2) $\dfrac{\lambda^{-1}T_{RL} - (k_n + \frac{1}{2}\eta_n)}{\sqrt{2(k_n + \eta_n)}} \xrightarrow{d} N(0,1)$,

(3) $\dfrac{k_n T_{RF} - (k_n + \frac{1}{2}\eta_n)}{\sqrt{2(k_n + \eta_n)}} \xrightarrow{d} N(0,1)$,

as $k_n \to \infty$, where $\eta_n = \dfrac{n}{\hat{\tau}^2} \sum_{k=1}^{k_n} \lambda_k (\beta_k^a)^2$ and $\beta_k^a = \int_{\mathcal{T}} \beta_a(t)\phi_k(t)dt$.

Under the null hypothesis and truncated functional linear model, the distributions of these test statistics are similar to their counterparts in multiple regression. In particular, if the truncated number k_n is fixed, the null distribution of Wald type and likelihood type test will behave like $\chi^2_{k_n}$ and the null distribution of F test will behave like F distribution $F_{k_n, n-(k_n+1)}$. The proofs of both theories are given in the Appendix.

4.3 Simulation Studies

In this section, we study the performance of the robust version of Wald-type, the likelihood type and F-type tests in terms of Type I errors and powers in the functional linear model. According to (4.1), we simulate data with three different settings based on three different models. Meanwhile, we add a baseline covariate

$Z_i \sim U[0,5]$ in the model as a potential con-founder. For simplicity of presentation, we consider a scenario in which the functional covariates $x_i(\cdot)$, $i = 1, 2, \cdots, n$, are generated by five Fourier basis functions, i.e.,

$$x_i(t) = \sum_{k=1}^{5} \xi_{ik}\phi_k(t),$$

where $\xi_{ik} \sim N(0, \lambda_k)$ with $\lambda_k = k^{-1.1}$, $k = 1, 2, \cdots, 5$. The five orthonormal Fourier basis functions are $\phi_1(t) = 1, \phi_2(t) = \sqrt{2}\sin(2\pi t), \phi_3(t) = \sqrt{2}\cos(2\pi t), \phi_4(t) = \sqrt{2}\sin(4\pi t)$, and $\phi_5(t) = \sqrt{2}\cos(4\pi t)$. The scalar response variable y_i is generated by the following model:

$$y_i = -2 + 0.5Z_i + \int_0^1 \beta(t)x_i(t)dt + \varepsilon_i \tag{4.13}$$

where $Z_i \sim U[0,5]$, and the coefficient function $\beta(t)$ is formed by the same basis functions for $x_i(t)$. Here, $\beta(t) = a\sum_{k=1}^{5} \beta_k\phi_k(t)$ with $a = 0, 0.03, 0.06, 0.12$ for controlling the magnitude of $\beta(\cdot)$ across the support. In our simulation studies, we consider the true coefficient parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top$ in three different settings. The details of these settings are as follows:

- Setting 1: $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^\top$, which means that the functional effects are the same across the five directions.

- Setting 2: $\boldsymbol{\beta} = (1, 0.7, 0.5, 0.3, 0)^\top$, which means that the direction with the smallest variation has a null effect on the response.

- Setting 3: $\boldsymbol{\beta} = (0, 0.7, 0.5, 0.3, 1)^\top$, which means that the direction with the smallest variation has a strong effect on the response.

A total of 2000 datasets, each with $n = 1000$ observations, are generated under these three settings. We study the performance of the powers under both classical and robust testing procedures for each setting. The robust testing procedures are mainly implemented in "RobStatTM" R packages (Maronna et al., 2019). For the above three settings, we consider two types of sampling designs for the functional covariates: dense model and sparse model. Based on our numerous simulations, we find out that the weight of random errors have effect on the dense model. Therefore, we increase the weight of the errors term in the dense model and design the following three models:

- Model I (Dense model 1): The weight on the random errors ε_is is 1, and the 501 grid points on each curve are equally spaced points in the interval $[0, 1]$.

- Model II (Dense model 2): The weight on the random errors ε_is is $\sqrt[3]{m}$, where $m = 501$ is the number of grid points on each curve that are the same as in the model I.

- Model III (Sparse model): The numbers of points per curve t_i are small and varies across subjects, which is chosen randomly from a discrete uniform dis-

tribution on $\{3, 4, 5\}$. Each curve is assumed to be observed at t_i point which are randomly sampling form the set of 501 equally spaced points in $[0, 1]$.

To study the proposed robust hypothesis testing procedure, the error ε_i can be distributed from a heavy-tailed distribution or a contaminated normal distribution. For a simple presentation, we only report the simulation results of the three models with errors being Cauchy $(0, 1)$ distributed. The simulation results are displayed in the following Tables 4.1-4.3 and Figures 4.1-4.3. For concise presentation, the results corresponding to five selected threshold levels, i.e., 0.65, 0.75, 0.85, 0.95, and 0.99, are given in the tables, while the estimated Type I error rates and power curves based on robust Wald type testing are shown in Figures 4.1-4.3.

Table 4.1: Simulation results based on the classical (T^c) and robust (T^r) methods for Model I (Dense model 1) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.049	0.053	0.049	0.053	0.049	0.053	0.050	0.048	0.050	0.044	0.050	0.045	0.041	0.057	0.040	0.055	0.041	0.055
0	0.75	0.042	0.048	0.042	0.049	0.042	0.049	0.043	0.046	0.042	0.050	0.043	0.051	0.049	0.046	0.049	0.047	0.049	0.048
0	0.85	0.045	0.052	0.045	0.050	0.046	0.050	0.053	0.047	0.051	0.046	0.053	0.047	0.054	0.043	0.053	0.042	0.054	0.044
0	0.95	0.050	0.056	0.050	0.056	0.050	0.057	0.057	0.046	0.056	0.044	0.057	0.046	0.053	0.040	0.053	0.045	0.053	0.047
0	0.99	0.054	0.054	0.052	0.051	0.054	0.054	0.058	0.050	0.057	0.049	0.058	0.050	0.051	0.054	0.050	0.056	0.051	0.057
0.03	0.65	0.932	1.000	0.932	1.000	0.932	1.000	0.912	1.000	0.912	1.000	0.912	1.000	0.794	1.000	0.794	1.000	0.794	1.000
0.03	0.75	0.922	1.000	0.922	1.000	0.922	1.000	0.908	1.000	0.908	1.000	0.908	1.000	0.808	1.000	0.808	1.000	0.808	1.000
0.03	0.85	0.926	1.000	0.926	1.000	0.926	1.000	0.912	1.000	0.912	1.000	0.912	1.000	0.826	1.000	0.823	1.000	0.826	1.000
0.03	0.95	0.928	1.000	0.927	1.000	0.928	1.000	0.911	1.000	0.911	1.000	0.911	1.000	0.856	1.000	0.855	1.000	0.856	1.000
0.03	0.99	0.928	1.000	0.928	1.000	0.928	1.000	0.893	1.000	0.893	1.000	0.893	1.000	0.826	1.000	0.826	1.000	0.826	1.000
0.06	0.65	0.960	1.000	0.960	1.000	0.960	1.000	0.954	1.000	0.954	1.000	0.954	1.000	0.907	1.000	0.906	1.000	0.907	1.000
0.06	0.75	0.965	1.000	0.965	1.000	0.965	1.000	0.957	1.000	0.957	1.000	0.957	1.000	0.911	1.000	0.911	1.000	0.911	1.000
0.06	0.85	0.967	1.000	0.967	1.000	0.967	1.000	0.956	1.000	0.956	1.000	0.956	1.000	0.903	1.000	0.903	1.000	0.903	1.000
0.06	0.95	0.966	1.000	0.966	1.000	0.966	1.000	0.951	1.000	0.951	1.000	0.951	1.000	0.916	1.000	0.911	1.000	0.916	1.000
0.06	0.99	0.966	1.000	0.966	1.000	0.966	1.000	0.956	1.000	0.956	1.000	0.956	1.000	0.920	1.000	0.920	1.000	0.920	1.000
0.12	0.65	0.985	1.000	0.985	1.000	0.985	1.000	0.979	1.000	0.979	1.000	0.979	1.000	0.949	1.000	0.949	1.000	0.949	1.000
0.12	0.75	0.983	1.000	0.983	1.000	0.983	1.000	0.978	1.000	0.978	1.000	0.978	1.000	0.959	1.000	0.959	1.000	0.959	1.000
0.12	0.85	0.975	1.000	0.975	1.000	0.975	1.000	0.979	1.000	0.978	1.000	0.979	1.000	0.943	1.000	0.943	1.000	0.943	1.000
0.12	0.95	0.980	1.000	0.980	1.000	0.980	1.000	0.975	1.000	0.975	1.000	0.975	1.000	0.963	1.000	0.963	1.000	0.963	1.000
0.12	0.99	0.982	1.000	0.982	1.000	0.980	1.000	0.979	1.000	0.979	1.000	0.979	1.000	0.961	1.000	0.961	1.000	0.961	1.000

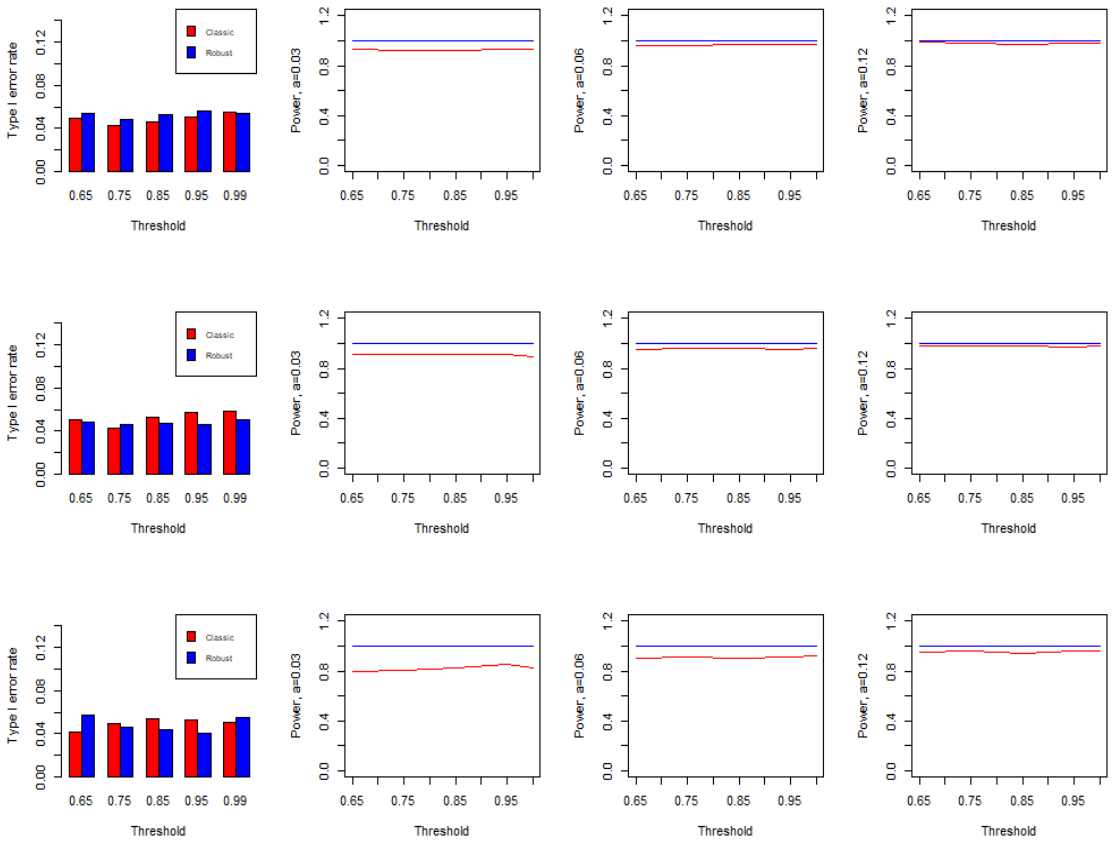

Figure 4.1: The estimated Type I error rates (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model I (Dense model 1). The results in the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is corresponding to the robust Wald type testing procedure.

Table 4.2: Simulation results based on the classical (T^c) and robust (T^r) methods for Model II (Dense model 2) with random errors $\sqrt[3]{m}\,\varepsilon_i$, where $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.050	0.053	0.049	0.053	0.050	0.054	0.054	0.055	0.054	0.055	0.054	0.056	0.044	0.042	0.044	0.040	0.044	0.045
0	0.75	0.045	0.057	0.044	0.057	0.045	0.058	0.057	0.051	0.055	0.050	0.057	0.050	0.059	0.046	0.059	0.047	0.059	0.047
0	0.85	0.058	0.055	0.057	0.054	0.058	0.054	0.0565	0.046	0.056	0.045	0.056	0.045	0.046	0.046	0.045	0.047	0.046	0.047
0	0.95	0.048	0.054	0.047	0.050	0.048	0.051	0.046	0.052	0.046	0.052	0.047	0.053	0.047	0.051	0.046	0.050	0.047	0.051
0	0.99	0.049	0.049	0.046	0.053	0.049	0.054	0.049	0.058	0.048	0.058	0.049	0.059	0.051	0.054	0.051	0.051	0.051	0.052
0.03	0.65	0.463	1.000	0.462	1.000	0.462	1.000	0.437	1.000	0.436	1.000	0.437	1.000	0.163	1.000	0.162	1.000	0.163	1.000
0.03	0.75	0.477	1.000	0.475	1.000	0.477	1.000	0.423	1.000	0.423	1.000	0.424	1.000	0.189	1.000	0.189	1.000	0.189	1.000
0.03	0.85	0.472	1.000	0.470	1.000	0.472	1.000	0.393	1.000	0.392	1.000	0.393	1.000	0.171	1.000	0.169	1.000	0.171	1.000
0.03	0.95	0.486	1.000	0.486	1.000	0.486	1.000	0.393	1.000	0.390	1.000	0.393	1.000	0.213	1.000	0.212	1.000	0.213	1.000
0.03	0.99	0.499	1.000	0.497	1.000	0.500	1.000	0.410	1.000	0.409	1.000	0.410	1.000	0.184	1.000	0.183	1.000	0.184	1.000
0.06	0.65	0.700	1.000	0.699	1.000	0.700	1.000	0.677	1.000	0.677	1.000	0.677	1.000	0.385	1.000	0.357	1.000	0.358	1.000
0.06	0.75	0.712	1.000	0.711	1.000	0.712	1.000	0.671	1.000	0.670	1.000	0.671	1.000	0.401	1.000	0.400	1.000	0.401	1.000
0.06	0.85	0.704	1.000	0.703	1.000	0.704	1.000	0.642	1.000	0.641	1.000	0.642	1.000	0.403	1.000	0.399	1.000	0.403	1.000
0.06	0.95	0.709	1.000	0.709	1.000	0.709	1.000	0.644	1.000	0.644	1.000	0.644	1.000	0.470	1.000	0.468	1.000	0.470	1.000
0.06	0.99	0.722	1.000	0.721	1.000	0.722	1.000	0.652	1.000	0.650	1.000	0.652	1.000	0.482	1.000	0.481	1.000	0.483	1.000
0.12	0.65	0.855	1.000	0.855	1.000	0.855	1.000	0.835	1.000	0.835	1.000	0.835	1.000	0.627	1.000	0.627	1.000	0.627	1.000
0.12	0.75	0.849	1.000	0.849	1.000	0.849	1.000	0.829	1.000	0.829	1.000	0.829	1.000	0.675	1.000	0.673	1.000	0.675	1.000
0.12	0.85	0.860	1.000	0.859	1.000	0.860	1.000	0.813	1.000	0.811	1.000	0.813	1.000	0.654	1.000	0.653	1.000	0.654	1.000
0.12	0.95	0.853	1.000	0.852	1.000	0.852	1.000	0.816	1.000	0.815	1.000	0.816	1.000	0.712	1.000	0.711	1.000	0.712	1.000
0.12	0.99	0.844	1.000	0.843	1.000	0.844	1.000	0.809	1.000	0.807	1.000	0.809	1.000	0.705	1.000	0.703	1.000	0.705	1.000

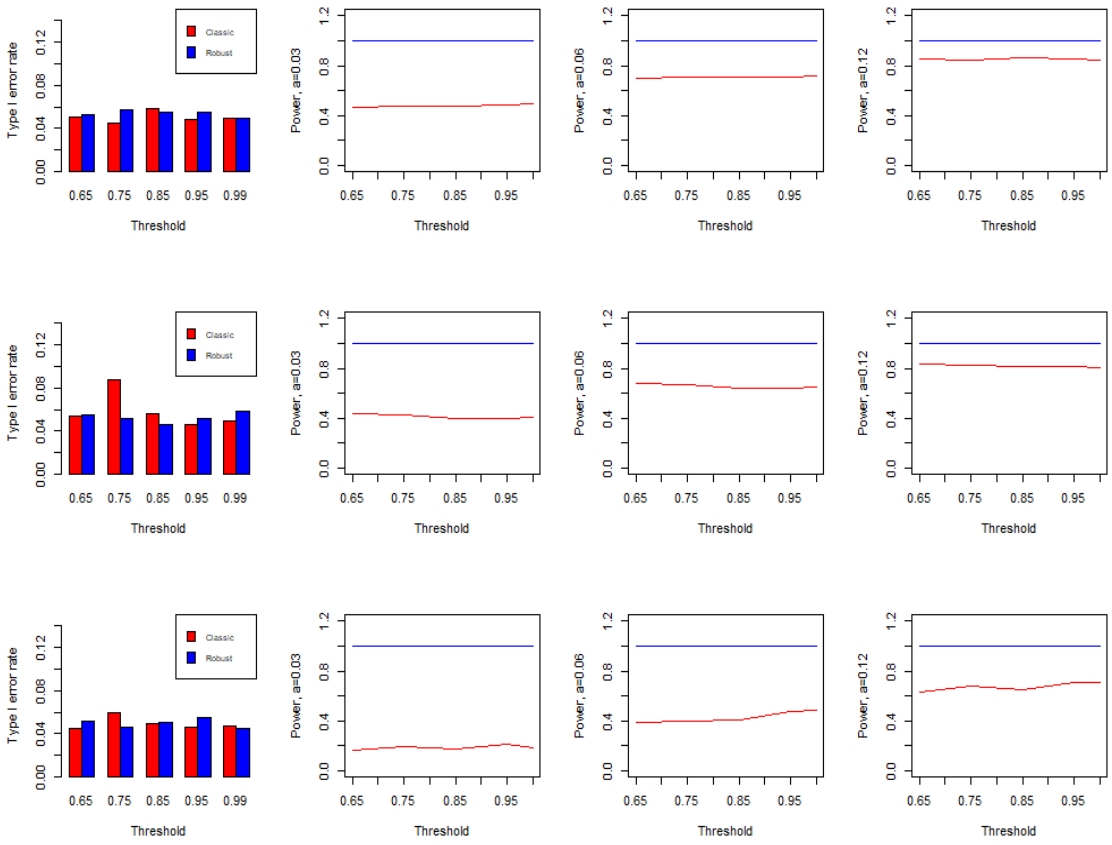

Figure 4.2: The estimated Type I error rates (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model II (Dense model 2). The results in the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is corresponding to the robust Wald type testing procedure.

Table 4.3: Simulation results based on the classical (T^c) and robust (T^r) methods for Model III (Sparse model) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.045	0.051	0.045	0.052	0.045	0.053	0.044	0.050	0.043	0.047	0.044	0.048	0.055	0.054	0.050	0.050	0.050	0.050
0	0.75	0.052	0.042	0.052	0.041	0.052	0.041	0.058	0.049	0.058	0.048	0.058	0.048	0.058	0.048	0.058	0.047	0.058	0.048
0	0.85	0.053	0.055	0.053	0.053	0.053	0.054	0.060	0.049	0.059	0.047	0.060	0.047	0.065	0.051	0.065	0.052	0.065	0.052
0	0.95	0.058	0.058	0.055	0.058	0.058	0.060	0.075	0.049	0.074	0.049	0.075	0.049	0.067	0.053	0.066	0.055	0.067	0.057
0	0.99	0.090	0.047	0.082	0.046	0.090	0.046	0.083	0.054	0.082	0.053	0.083	0.054	0.091	0.056	0.090	0.055	0.092	0.056
0.03	0.65	0.063	0.612	0.062	0.602	0.063	0.603	0.057	0.522	0.056	0.516	0.057	0.517	0.050	0.158	0.049	0.158	0.050	0.158
0.03	0.75	0.054	0.624	0.053	0.619	0.054	0.620	0.055	0.519	0.054	0.512	0.055	0.515	0.051	0.154	0.051	0.149	0.051	0.150
0.03	0.85	0.063	0.651	0.062	0.638	0.063	0.640	0.058	0.491	0.057	0.482	0.058	0.485	0.060	0.168	0.059	0.166	0.060	0.168
0.03	0.95	0.072	0.667	0.071	0.656	0.072	0.659	0.073	0.433	0.071	0.425	0.073	0.428	0.075	0.171	0.074	0.168	0.076	0.169
0.03	0.99	0.080	0.612	0.080	0.608	0.080	0.615	0.086	0.373	0.084	0.371	0.086	0.376	0.074	0.164	0.072	0.162	0.074	0.165
0.06	0.65	0.062	0.992	0.061	0.992	0.062	0.992	0.063	0.988	0.063	0.988	0.063	0.988	0.059	0.437	0.058	0.429	0.059	0.430
0.06	0.75	0.073	0.996	0.073	0.995	0.073	0.995	0.064	0.989	0.063	0.986	0.064	0.987	0.063	0.515	0.063	0.506	0.063	0.506
0.06	0.85	0.073	0.998	0.072	0.997	0.073	0.997	0.062	0.983	0.062	0.980	0.062	0.980	0.060	0.557	0.059	0.558	0.060	0.560
0.06	0.95	0.073	0.999	0.072	0.999	0.073	0.999	0.072	0.978	0.072	0.974	0.072	0.974	0.077	0.644	0.075	0.636	0.077	0.638
0.06	0.99	0.102	0.997	0.099	0.995	0.102	0.995	0.081	0.955	0.080	0.952	0.082	0.953	0.089	0.587	0.089	0.587	0.090	0.590
0.12	0.65	0.101	1.000	0.100	1.000	0.101	1.000	0.089	1.000	0.089	1.000	0.089	1.000	0.051	0.906	0.051	0.902	0.051	0.902
0.12	0.75	0.110	1.000	0.108	1.000	0.110	1.000	0.096	1.000	0.095	1.000	0.096	1.000	0.070	0.949	0.069	0.944	0.070	0.944
0.12	0.85	0.114	1.000	0.111	1.000	0.114	1.000	0.090	1.000	0.090	1.000	0.090	1.000	0.073	0.986	0.072	0.982	0.073	0.982
0.12	0.95	0.111	1.000	0.111	1.000	0.111	1.000	0.109	1.000	0.108	1.000	0.109	1.000	0.077	0.998	0.077	0.998	0.077	0.998
0.12	0.99	0.126	1.000	0.124	1.000	0.126	1.000	0.118	1.000	0.117	1.000	0.118	1.000	0.101	0.997	0.099	0.996	0.101	0.997

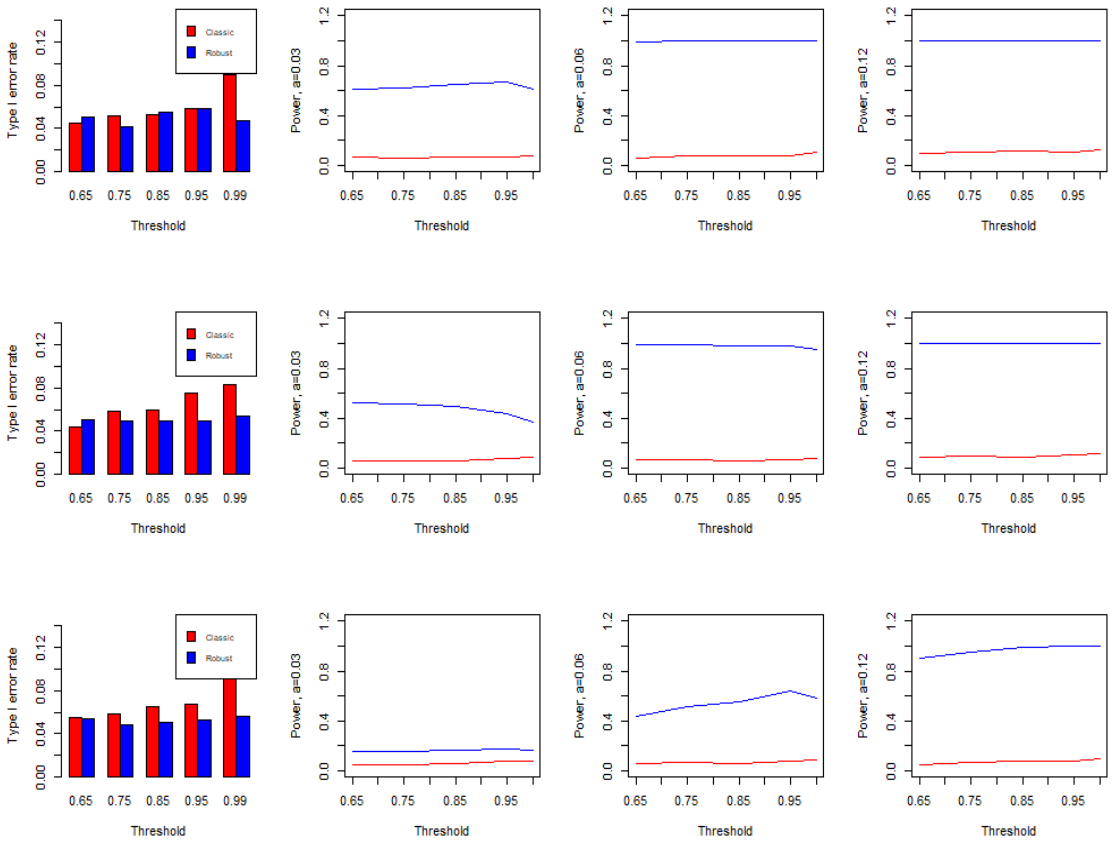

Figure 4.3: The estimated Type I error rates (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model III (Sparse model). The results for the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is corresponding to the robust Wald type testing procedure.

The simulation results shown in Table 4.1 associated with Figure 4.1 are for the model I. The results show that the estimated Type I error rates in most settings are close to the nominal significance level of 0.05. The power performance based on the classical method is worse than that of the robust method; this is particularly obvious in the third setting. The simulation results shown in Table 4.2 (or Figure 4.2) and Table 4.3 (or Figure 4.3) are for Models II and III, respectively. As with Model I, the type I error rates are close to 0.05, except we observe inflated Type I error rates for the sparse model (Model III). Obviously, the power values based on the classical method are less than those based on the robust testing procedures. All these results indicate that the robust procedures are more stable and consistent than the classical ones in the hypothesis testing problem. The results also show that the high threshold of PVE does not guarantee a better power both in the robust and classical testing procedures. For example, in Table 4.2, under the setting 1 with a=0.12, the power decreases from 0.855 to 0.844 (based on the classical Wald test) with $\gamma = 0.65$ increasing to $\gamma = 0.99$. In Table 4.3, under the first setting with a=0.03, the power drops from 0.667 to 0.612 (based on the robust Wald type test) with threshold γ increasing from 0.95 to 0.99. This is a trade-off between the degrees of freedom and the information included in the selected PCs. Based on our extensive experiments, $\gamma = 0.95$ is the most suitable threshold when implementing both the classical and robust testing procedures. To

further demonstrate the performance of the robust hypothesis testing procedure, we also report the extensive simulation results in Parts 1-3 of Appendix. Part 1 centers on the dense model with errors $\frac{1}{3}m \times \varepsilon_i$, where ε_i is Cauchy $(0, 1)$ distributed or the contaminated normal $0.9N(0, 1) + 0.1N(1, 9^2)$ distributed. Part 2 contributes to a small sample size ($n = 200$). Part 3 focuses on Huber's loss function. The results there yield the conclusions similar to those for large sample size, different type of contaminated error distributions, and loss function.

4.4 Real Data Examples

4.4.1 Diffusion Tensor Imaging Data

The first real example we consider is the Diffusion Tensor Imaging (DTI) study data. This dataset has been considered in literature including (Goldsmith et al., 2012; Kong et al., 2013; Su et al., 2017) and is available in the R package **refund**. Diffusion tensor imaging (DTI) tractography is a magnetic resonance imaging technique that measures the restricted diffusion of water in tissue to produce neural tract images. It allows the study of white-matter tracts by measuring the diffusivity of water in the brain. In white-matter tracts, water diffuses anisotropically in the direction of the tract, while elsewhere water diffuses isotropically. The diffusion of water molecules at each voxel is measured by fractional anisotropy (FA) and profiles along

corpus callosum and right corticospinal tracts. Using measurements of diffusivity along several gradients, DTI can provide relatively detailed images of white-matter anatomy in the brain (Goldsmith et al., 2011).

In this study, 100 multiple scleroses (MS) patients and 42 healthy controls were observed at multiple visits. For each subject and each visit, FA along with the corpus callosum (CC) and the right corticospinal tract (RCST) were recorded. There were 93 locations along the corpus callosum (CC) and 55 tracts profiles from RCST (Goldsmith et al., 2011, 2012). Since the Auditory Serial Addition Test was only recorded to multiple sclerosis patients and the missingness along RCST is very large, the interest of our study here is to test the association between the Paced Auditory Serial Addition Test scores (PASAT) and corpus callosum (CC) in multiple sclerosis group. Although some multiple sclerosis patients visited several times, we use the data obtained at the baseline visit in our study. The modeling framework we assumed is $y_i = \alpha + \int_{\mathcal{T}_{CCA}} \beta_{CCA}(t) X_i(t) dt + \varepsilon_i$, and we want to test the hypothesis $H_0 : \beta_{CCA}(t) = 0$ for all $t \in \mathcal{T}_{CCA} = [0, 93]$.

We first explore the relationship between FA profiles and PASAT scores. The randomly selected five samples (left panel) and estimated mean FA profiles (middle panel) along CC tracts and PASAT scores (right panel) for all subjects are presented in Figure 4.4. It is difficult to conclude that the FA profiles along CC tracts have a

significant impact on PASAT scores intuitively; a scalar-on-function linear model is applied to study the relationship between PASAT scores and the FA profiles along CC tracts. We use both classical and robust testing procedures with different threshold levels of PVE: 65%, 75%, 85%, and 95%. All the results are presented in Table 4.4. We also report the estimated coefficient function $\hat{\beta}(t)$ based on both classical and robust methods in Figure 4.5, respectively.

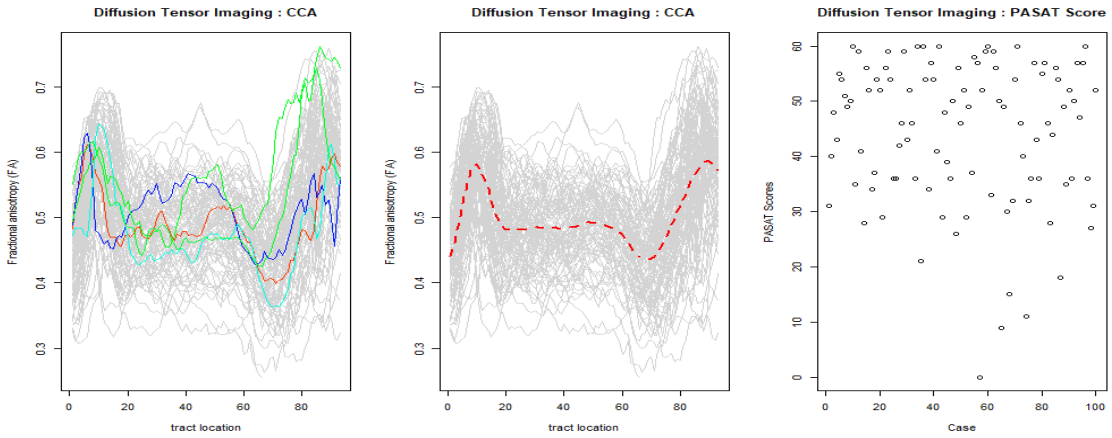

Figure 4.4: Fractional anisotropy profiles along corpus callosum (CC) and the associated Paced Auditory Serial Addition Test scores (PASAT) of the 100 multiple sclerosis patients.

From Table 4.4, we can see that there are inconsistent conclusions for classical hypothesis testing procedures with different choices of the threshold. If we choose significant level $\alpha = 0.05$, we can conclude that there is a significant association

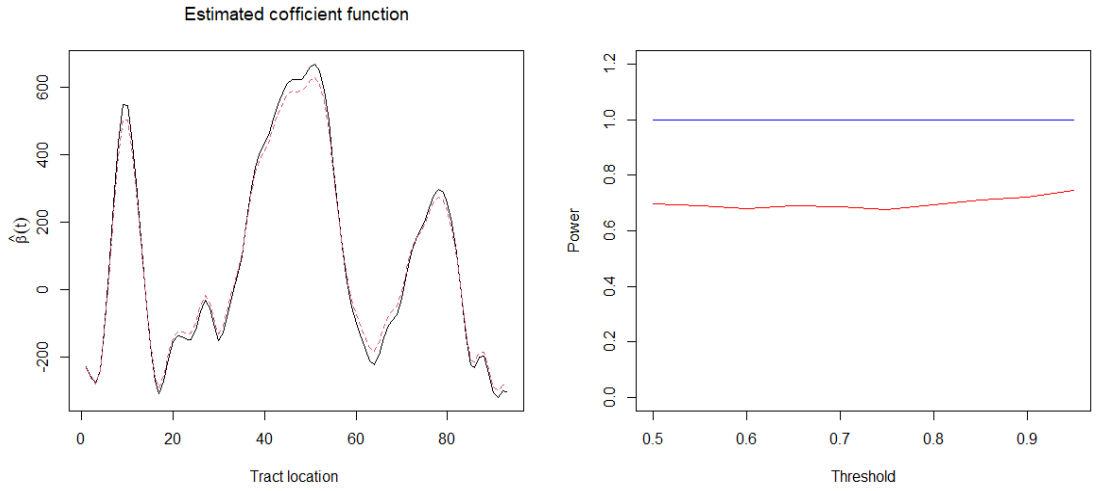

Figure 4.5: Left Panel: Estimated coefficient function $\hat{\beta}_{CCA}(t)$ in CCA area based on both robust estimation and OLS estimation methods. The threshold for selecting PCs is 0.95. The black solid line is estimated $\hat{\beta}_{CCA}(t)$ based on the robust estimation, the red dashed line is estimated $\hat{\beta}_{CCA}(t)$ based on the OLS estimation. Right panel: Power curve, the red line is based on the classical Wald testing, while the blue line is corresponding to the robust Wald type testing.

between FA profiles along CC tracts and the PASAT scores if the threshold $\gamma = 0.65, 0.75, 0.85$. But if $\gamma = 0.95$, we cannot draw such a conclusion because the p-value is larger than 0.05, which means that different thresholds lead to opposite conclusions based on the classical testing methods. On the other hand, the robust testing procedures provide a consistent conclusion for different threshold levels, i.e.,

Table 4.4: Testing results based on both classical (T^c) and robust (T^r) methods for the DTI data with FA profiles in corpus callosum (CC). The number of PCs are selected based on various threshold choices.

		p-value (T^c)			p-value (T^r)		
γ	Number of PCs	Wald	F	Likelihood	Wald	F	Likelihood
0.65	2	0.0014	0.0021	0.0017	0.0003	0.0002	0.0001
0.75	3	0.0037	0.0053	0.0043	0.0009	0.0006	0.0003
0.85	5	0.0198	0.0260	0.0204	0.0044	0.0024	0.0012
0.95	9	0.0708	0.0876	0.0630	0.0316	0.0162	0.0094

the FA profiles along CC tracts is significantly associated with PASAT scores.

Furthermore, we mimic the DTI data by a simulation study to explain the inconsistency of the classical testing procedures. First, we estimated coefficient function based on the classical and robust methods as well as the functional regression along CC tracts. Then, we conducted a simulation study by setting the true parameters to be those estimated from the DTI data. To better explain the advantage of the robust testing procedure, we contaminate the mimic data by adding an error term $\epsilon_i \sim$ Cauchy(0,1) to producing "vertical outliers". For a simple presentation, we only display the estimated power curve based on the classical and robust Wald tests with different thresholds in Figure 4.5 (right panel). For the robust testing method, the power performance is better than that with the classical testing procedure. The

good performance of power based on the robust method is demonstrated by relatively stableness for the different threshold levels.

4.4.2 Fat Content Spectrometric Data

The Fat Content Spectrometric (FCS) data consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food and Feed Analyzer. This data set is a part of the Tecator data set, which is available in the R package "fds" and can also be found at the website (*http://lib.stat.cmu.edu/datasets*). Each sample consists of a 100-channel absorbance spectrum of the 850–1050 nm wavelength range. Each spectrum in the database is associated with a content description of the meat sample, obtained by analytic chemistry, which is the percentage of fat, water and protein (Rossi et al., 2005). Our interest is to study the association between the fat percentage value and the recorded 100-channel absorbance spectrum.

To investigate the robustness of the proposed testing procedure, we contaminate the fat percentage by the collective outliers. We notice that some values of the fat percentage in a segment are small contextually. In fact, there are 20 samples (sample 50 to sample 70) in this segment. We reset the values there to be two, five, ten and twenty multiples of the original ones so that the data are contaminated. We compare the classical and robust hypothesis testing procedures in the following table

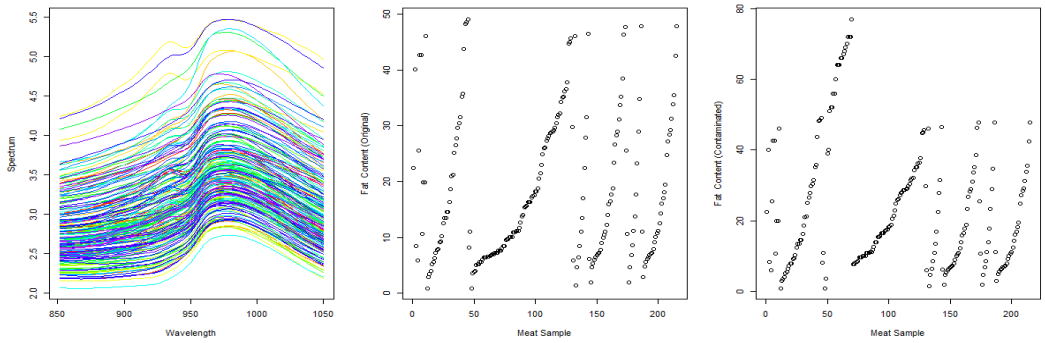

Figure 4.6: FCS data: absorbance trajectories of 215 meat samples measured over 100 equally spaced wavelengths between 850 and 1050 nm (left panel). Percentage values of fat in the original data (middle panel) as well as in the contaminated data (right panel).

and use 0.95 as the threshold for selecting PC numbers. The 100-channel absorbance spectrum in the 850-1050 nm wavelength range and fat content percentage are shown in Figure 4.6.

From Table 4.5, it is obviously that the robust testing procedure is more effective than the classical method, especially for contaminated data. If we choose 0.05 as our significant level, the p-values based on the two testing procedures $\ll 0.05$, which means that there is a significant association between the fat percentage value and 100-channel absorbance spectrum. On the other hand, when we reset some values to be ten or twenty multiples of the original ones, the p-value is larger than 0.05

Table 4.5: Testing results based on the classical (T^c) and robust (T^r) methods for the Fat Content Spectrometric (FCS) data. The threshold for selecting PC numbers is 0.95.

	p-value (T^c)			p-value (T^r)		
	Wald	F	Likelihood	Wald	F	Likelihood
Original data	5.0223e − 13	8.8070e − 12	6.6661e − 12	5.6623e − 12	3.5630e − 12	1.6198e − 13
Contaminated data (2 multilples)	7.1982e − 12	7.6327e − 11	5.9218e − 118	5.8154e − 11	4.5510e − 11	3.8232e − 11
Contaminated data (5 multilples)	1.4890e − 6	2.8187e − 6	2.4648e − 6	5.1957e − 6	5.0819e − 6	28578e − 6
Contaminated data (10 multilples)	0.1225	0.1240	0.1218	1.5136e − 4	3.4521e − 5	2.3234e − 5
Contaminated data (20 multilples)	0.3130	0.3140	0.3113	2.3493e − 9	1.4487e − 9	2.4934e − 10

for using the classical procedure, which shows that there is no significant association between the fat percentage value and the 100-channel absorbance spectrum. This conclusion is unappealing in practice as contradictory conclusions could be drawn when a dataset is contaminated.

4.5 Conclusions and Discussion

Functional linear model (FLM) has been a prevalent tool to describe the dynamic data of infinite-dimensional covariates on scalar/ functional responses. There is a rich literature on the performance of a classical test based on the FPCA method under FLM. However, these testing procedures are not robust to outliers. In this paper, we propose three robust testing procedures (Wald, F and Likelihood-type)

and investigate their asymptotic properties. The number of PCs is determined by the PVE threshold. Three simulation models, along with three different settings, are designed for studying the performance of testing power. Under FLM, our numerical studies show that the power of the classical testing procedure is lower than that of the robust procedures if a dataset contains outliers. In addition, the sparse functional linear model is more sensitive to outliers than the dense model. Furthermore, the robust testing approach guarantees a higher power and is more stable for a contaminated dataset. The two real data examples also demonstrate the good performance of the robust testing procedures.

There are several directions for future research. In this work, we only discuss the scalar-on-function functional linear model. Extending the method to the function-on-function linear or the generalized functional linear models will be interesting. On the other hand, the establishment of theoretical properties for robust global hypothesis testing in the functional linear model is still challenging. That will be most of the interest in our future work.

4.6 Appendix

4.6.1 Proofs

Proof of Theorem 4.1

Assumptions (A1)-(A3) guarantee the consistency of the estimated ege Under the truncated functional linear model 4.4, if assumptions (A4)-(A7) hold, the M-estimator $\hat{\boldsymbol{\beta}}_R$ of $\boldsymbol{\beta}$ has asymptotic normality with $\mathcal{N}_{k_n}(\boldsymbol{\beta}, \tau^2(\Xi^\top\Xi)^{-1})$, where $\Xi = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n]^\top$ and even if $k_n \to \infty$ (Yohai and Maronna, 1979). Under the null hypothesis and $k_n = o(n)$, the robust Wald type statistic T_{RW} follows a centered chi-square distribution with degree of freedom k_n . Recalling that a quadratic form of normal distribution has a chi-square distribution, then we can write $T_{RW} = \sum_{k=1}^{k_n} A_{1k}^2$, where A_{1k}'s are i.i.d standard normal random variables. By the central limit theorem, we have $(\sum_{k=1}^{k_n} A_{1k}^2 - k_n)/\sqrt{2k_n} \xrightarrow{D} N(0,1)$ when $k_n \to \infty$. Thus, $(T_{RW} - k_n)/\sqrt{2k_n} \xrightarrow{D} N(0,1)$ as $n \to \infty$.

Denote $D(R) = \sum_{i=1}^{n} \rho\left(\frac{r_i(\tilde{\boldsymbol{\theta}}_R)}{\tilde{\sigma}_\varepsilon}\right)$ and $D(F) = \sum_{i=1}^{n} \rho\left(\frac{r_i(\hat{\boldsymbol{\theta}}_R)}{\hat{\sigma}_\varepsilon}\right)$. Then $D(R)$ is the minimum values of equation (4.7) under the restriction H_0, and $D(F)$ is the minimum values of equation (4.7) without constraints. Based on the theorem 1 in Schrader and Hettmansperger (1980), if the null hypothesis is true, $\lambda^{-1}(D(R) - D(F)) \xrightarrow{D} \chi^2_{k_n}$. Then, similar with proof of (1), we can denote $\lambda^{-1}(D(R) - D(F)) = \sum_{k=1}^{k_n} A_{2k}^2$, where

A_{2k}'s are i.i.d standard normal random variables. By the central limit theorem, we have $(\sum_{k=1}^{k_n} A_{2k}^2 - k_n)/\sqrt{2k_n} \xrightarrow{D} N(0,1)$ as $n \to \infty$. Thus, $(\lambda^{-1}T_{RL} - k_n)/\sqrt{2k_n} \xrightarrow{D} N(0,1)$ as $n \to \infty$. For T_{RF}, note that $k_n T_{RF} = \lambda^{-1}T_{RL}$, then under null hypothesis, $(k_n T_{RF} - k_n)/\sqrt{2k_n} \xrightarrow{D} N(0,1)$ can be obtained easily. □

Proof of Theorem 4.2

We know that under the truncated model (4.4), the alternative hypothesis H_a: $\beta(t) = \beta_a(t) \neq 0$ is equivalent to $H_a : \beta_k^a \neq 0$, $\beta_k^a = \int_{\mathcal{T}} \beta_a(t)\phi_k(t)dt$, $1 \leqslant k \leqslant k_n$. According to Yohai and Maronna (1979) and under the assumptions(A1)-(A7), it can be shown that,

$$\frac{\hat{\beta}_k - \beta_k^a}{\hat{\tau}/\sqrt{\boldsymbol{\xi}_k^T \boldsymbol{\xi}_k}} \xrightarrow{D} N(0,1), \text{ for } 1 \leqslant k \leqslant k_n$$

independently. Note that $E(\xi_{ik}) = 0$ and $E(\xi_{ik}^2) = \lambda_k, i = 1, 2, \ldots, n$, then $\boldsymbol{\xi}_k^T \boldsymbol{\xi}_k = \|\boldsymbol{\xi}_k\|^2 = n\lambda_k$, where $\boldsymbol{\xi}_k = (\xi_{1k}, \xi_{2k}, \ldots, \xi_{nk})^T$. Therefore, according to (Muirhead, 2005, p.20-22), the robust Wald test statistic T_{RW} follows a noncentral χ^2 distribution with degree of freedom k_n and the noncentral parameter $\eta_n = \frac{n}{\hat{\tau}^2} \sum_{k=1}^{k_n} \lambda_k (\beta_k^a)^2$ (Su et al., 2017). Based on a simple calculation, we have the asymptotic distribution under H_a,

$$\frac{T_{RW} - (k_n + \eta_n)}{\sqrt{2(k_n + 2\eta_n)}} \xrightarrow{D} N(0,1)$$

as $k_n \to \infty$.

For robust likelihood test statistic T_{RL}, according to the theorem 2 in Schrader and

Hettmansperger (1980), under the alternative hypothesis H_a, $\lambda^{-1}(D(R)-D(F))$ has a asymptotic noncentral χ^2 distribution with degree of freedom k_n and the noncentral parameter $\frac{1}{2}\eta_n$. Then continue with a simple calculation, we have the asymptotic distribution under H_a,

$$\frac{T_{RL} - (k_n + \frac{1}{2}\eta_n)}{\sqrt{2(k_n + \eta_n)}} \xrightarrow{D} N(0,1)$$

as $k_n \to \infty$. Similarly, for robust F test statistic $T_{RF} = \{k_n\lambda\}^{-1}(D(R) - D(F))$, then $k_n T_{RF}$ has a asymptotic noncentral χ^2 distribution with degree of freedom k_n and the noncentral parameter $\frac{1}{2}\eta_n$. That implies,

$$\frac{k_n T_{RF} - (k_n + \frac{1}{2}\eta_n)}{\sqrt{2(k_n + \eta_n)}} \xrightarrow{D} N(0,1)$$

as $k_n \to \infty$. □

4.6.2 Additional Simulation Results

Part I: Simulation results for dense model with errors $\frac{1}{3}m \times \epsilon_i$.

Table 4.6: Simulation results based on the classical (T^c) and robust (T^r) methods for Model II*, the dense model, with random errors $\frac{1}{3}m \times \varepsilon_i$, where $\varepsilon_i \sim$ Cauchy$(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.044	0.056	0.043	0.053	0.044	0.053	0.051	0.057	0.050	0.056	0.051	0.057	0.052	0.045	0.052	0.046	0.052	0.047
0	0.75	0.044	0.053	0.043	0.051	0.044	0.051	0.049	0.047	0.049	0.047	0.049	0.048	0.052	0.046	0.051	0.046	0.052	0.047
0	0.85	0.047	0.049	0.046	0.047	0.047	0.048	0.045	0.055	0.044	0.055	0.045	0.056	0.053	0.053	0.051	0.049	0.053	0.051
0	0.95	0.053	0.044	0.052	0.048	0.053	0.048	0.051	0.055	0.050	0.056	0.051	0.056	0.044	0.050	0.044	0.051	0.044	0.051
0	0.99	0.059	0.052	0.058	0.048	0.059	0.048	0.049	0.054	0.049	0.054	0.049	0.054	0.047	0.049	0.047	0.050	0.047	0.050
0.03	0.65	0.044	0.436	0.044	0.423	0.044	0.423	0.051	0.402	0.051	0.388	0.051	0.290	0.046	0.108	0.043	0.101	0.046	0.102
0.03	0.75	0.053	0.433	0.053	0.423	0.053	0.425	0.041	0.341	0.041	0.326	0.041	0.329	0.048	0.112	0.048	0.112	0.048	0.112
0.03	0.85	0.052	0.444	0.050	0.434	0.052	0.436	0.044	0.292	0.044	0.291	0.044	0.273	0.053	0.104	0.051	0.103	0.053	0.105
0.03	0.95	0.051	0.449	0.050	0.435	0.051	0.441	0.050	0.276	0.050	0.270	0.050	0.273	0.054	0.116	0.054	0.114	0.054	0.115
0.03	0.99	0.056	0.451	0.055	0.444	0.056	0.447	0.048	0.269	0.047	0.067	0.048	0.270	0.052	0.120	0.051	0.121	0.052	0.122
0.06	0.65	0.058	0.960	0.058	0.955	0.058	0.956	0.052	0.920	0.051	0.910	0.052	0.911	0.053	0.296	0.053	0.282	0.053	0.285
0.06	0.75	0.059	0.975	0.059	0.972	0.059	0.972	0.047	0.909	0.045	0.900	0.047	0.900	0.051	0.311	0.051	0.309	0.051	0.312
0.06	0.85	0.060	0.977	0.059	0.973	0.060	0.973	0.050	0.879	0.050	0.865	0.050	0.868	0.046	0.307	0.046	0.298	0.046	0.302
0.06	0.95	0.051	0.977	0.054	0.972	0.055	0.973	0.053	0.868	0.052	0.852	0.053	0.855	0.051	0.417	0.050	0.407	0.051	0.411
0.06	0.99	0.050	0.984	0.048	0.982	0.050	0.982	0.066	0.874	0.065	0.864	0.066	0.865	0.044	0.0418	0.044	0.408	0.044	0.411
0.12	0.65	0.074	1.000	0.073	1.000	0.074	1.000	0.071	1.000	0.070	1.000	0.071	1.000	0.050	0.811	0.050	0.802	0.050	0.803
0.12	0.75	0.072	1.000	0.072	1.000	0.072	1.000	0.062	1.000	0.062	1.000	0.062	1.000	0.056	0.884	0.055	0.876	0.056	0.878
0.12	0.85	0.077	1.000	0.077	1.000	0.077	1.000	0.071	1.000	0.071	1.000	0.071	1.000	0.047	0.876	0.046	0.857	0.047	0.858
0.12	0.95	0.076	1.000	0.075	1.000	0.076	1.000	0.063	1.000	0.063	1.000	0.063	1.000	0.056	0.971	0.056	0.967	0.056	0.967
0.12	0.99	0.074	1.000	0.073	1.000	0.074	1.000	0.072	1.000	0.070	1.000	0.072	1.000	0.047	0.974	0.045	0.967	0.047	0.968

Table 4.7: Simulation results based on the classical (T^c) and robust (T^r) methods for Model IV, the dense model, with random errors $\frac{1}{3}m \times \varepsilon_i$, where $\epsilon_i \sim 0.9N(0,1) + 0.1N(1,9^2)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.046	0.058	0.045	0.059	0.046	0.059	0.049	0.056	0.049	0.058	0.049	0.058	0.047	0.048	0.046	0.048	0.047	0.048
0	0.75	0.050	0.055	0.049	0.056	0.050	0.057	0.049	0.056	0.049	0.055	0.049	0.056	0.046	0.052	0.046	0.055	0.046	0.055
0	0.85	0.051	0.051	0.050	0.055	0.051	0.054	0.047	0.051	0.044	0.050	0.047	0.051	0.051	0.044	0.050	0.044	0.052	0.044
0	0.95	0.054	0.058	0.053	0.059	0.054	0.059	0.049	0.047	0.048	0.049	0.049	0.050	0.044	0.048	0.044	0.049	0.044	0.049
0	0.99	0.044	0.052	0.042	0.051	0.044	0.052	0.040	0.050	0.040	0.051	0.040	0.051	0.051	0.056	0.050	0.056	0.051	0.057
0.03	0.65	0.172	0.789	0.171	0.791	0.172	0.793	0.139	0.709	0.139	0.728	0.129	0.728	0.063	0.180	0.063	0.180	0.063	0.181
0.03	0.75	0.152	0.837	0.151	0.840	0.152	0.840	0.130	0.674	0.128	0.676	0.130	0.678	0.067	0.191	0.066	0.193	0.067	0.194
0.03	0.85	0.167	0.832	0.165	0.832	0.168	0.834	0.118	0.655	0.117	0.656	0.118	0.658	0.070	0.182	0.070	0.183	0.070	0.185
0.03	0.95	0.145	0.849	0.145	0.853	0.145	0.854	0.120	0.621	0.118	0.623	0.120	0.626	0.077	0.241	0.077	0.246	0.070	0.246
0.03	0.99	0.154	0.851	0.150	0.853	0.155	0.855	0.112	0.596	0.111	0.601	0.112	0.605	0.074	0.238	0.075	0.243	0.074	0.246
0.06	0.65	0.529	1.000	0.528	1.000	0.529	1.000	0.466	0.999	0.464	0.999	0.466	0.999	0.128	0.563	0.127	0.566	0.128	0.566
0.06	0.75	0.539	1.000	0.538	1.000	0.539	1.000	0.414	1.000	0.413	1.000	0.414	1.000	0.115	0.634	0.114	0.636	0.115	0.638
0.06	0.85	0.571	1.000	0.569	1.000	0.571	1.000	0.382	1.000	0.380	1.000	0.382	1.000	0.122	0.646	0.115	0.652	0.122	0.655
0.06	0.95	0.575	1.000	0.572	1.000	0.575	1.000	0.343	0.999	0.342	0.999	0.342	0.999	0.142	0.813	0.140	0.819	0.142	0.820
0.06	0.99	0.565	1.000	0.563	1.000	0.566	1.000	0.347	1.000	0.343	1.000	0.347	1.000	0.152	0.817	0.150	0.819	0.152	0.821
0.12	0.65	0.984	1.000	0.984	1.000	0.984	1.000	0.966	1.000	0.966	1.000	0.966	1.000	0.334	0.990	0.333	0.990	0.334	0.990
0.12	0.75	0.988	1.000	0.988	1.000	0.988	1.000	0.955	1.000	0.955	1.000	0.955	1.000	0.395	0.998	0.392	0.998	0.395	0.998
0.12	0.85	0.994	1.000	0.994	1.000	0.994	1.000	0.949	1.000	0.949	1.000	0.949	1.000	0.392	0.999	0.389	0.999	0.392	0.999
0.12	0.95	0.995	1.000	0.995	1.000	0.995	1.000	0.929	1.000	0.927	1.000	0.930	1.000	0.504	1.000	0.500	1.000	0.505	1.000
0.12	0.99	0.994	1.000	0.993	1.000	0.994	1.000	0.938	1.000	0.937	1.000	0.939	1.000	0.510	1.000	0.507	1.000	0.510	1.000

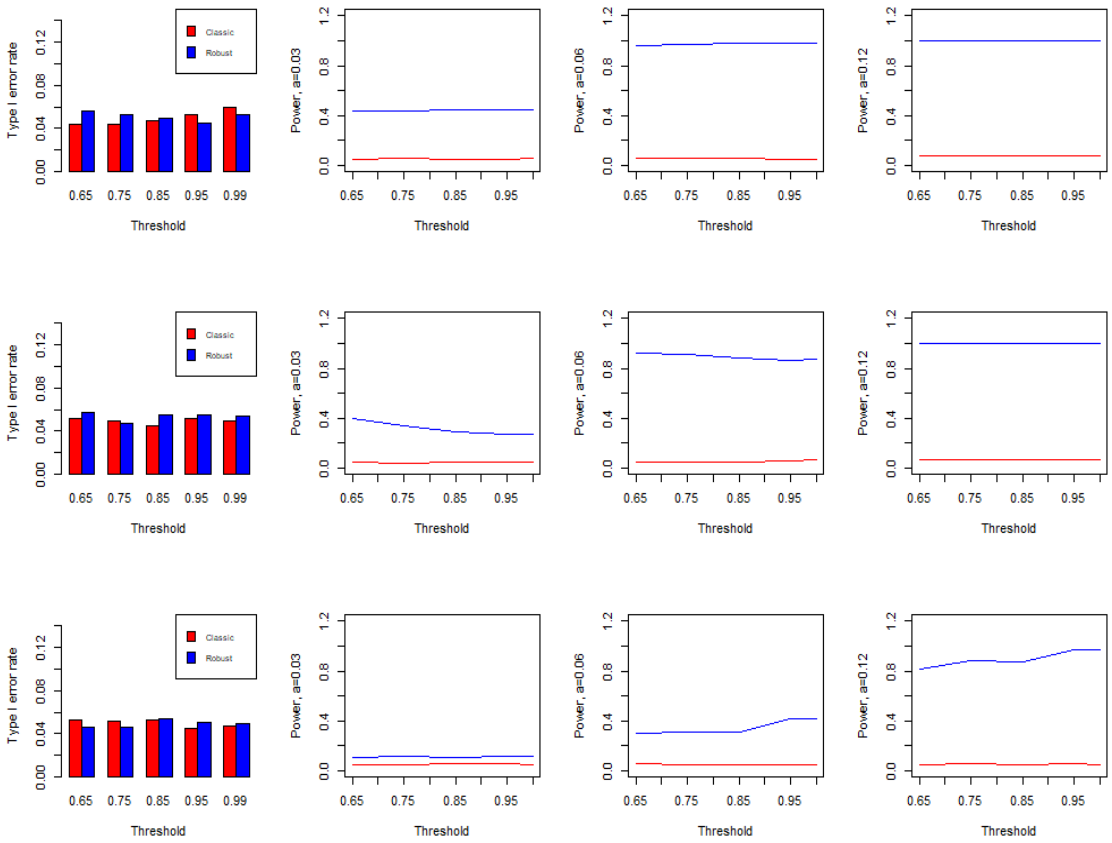

Figure 4.7: The estimated Type I errors (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model II*. The results in the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is corresponding to the robust Wald type testing procedure.

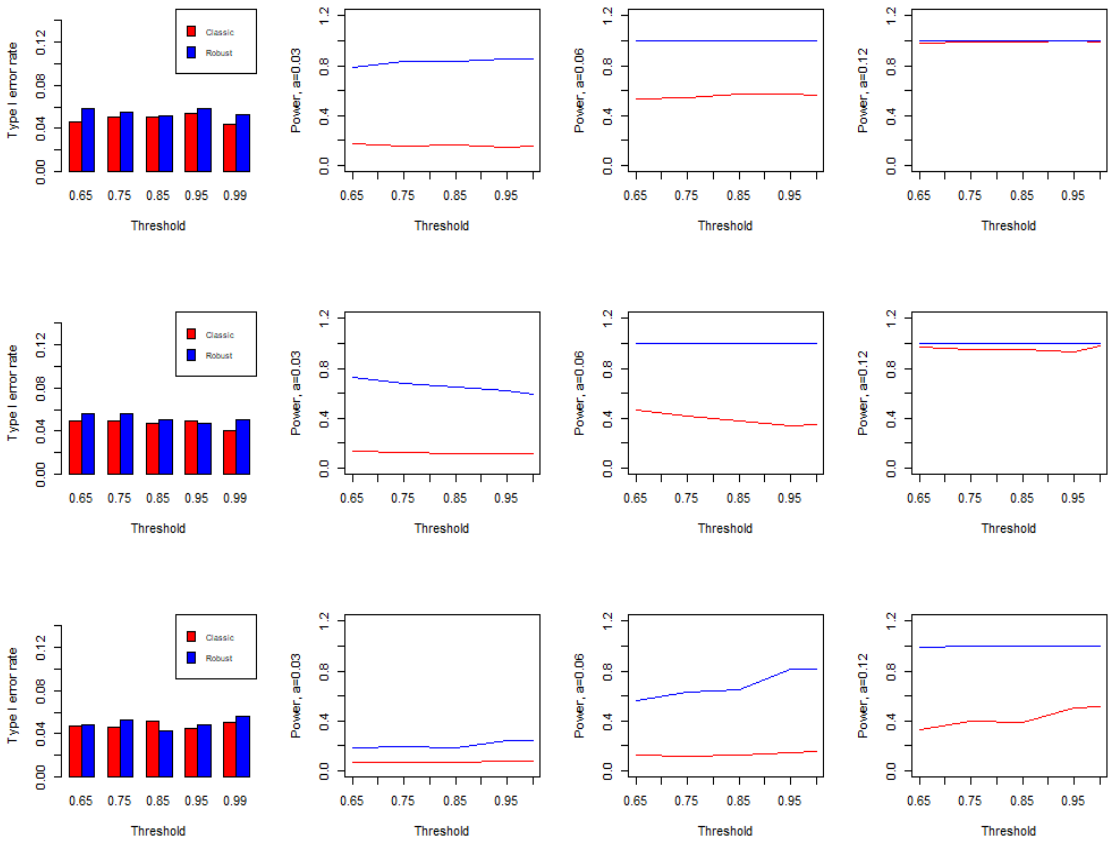

Figure 4.8: The estimated Type I errors (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model IV. The results in the settings 1-3 are aligned from the top to the bottom. In addition, red line is based on classic Wald testing procedure, , while the blue line is corresponding to the robust Wald type testing procedure.

Part II: Simulation results for small sample size ($n = 200$).

Table 4.8: Simulation results based on the classical (T^c) and robust (T^r) methods for Model I (Dense model 1) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.054	0.055	0.053	0.058	0.054	0.061	0.048	0.053	0.045	0.059	0.048	0.061	0.049	0.055	0.048	0.054	0.049	0.056
0	0.75	0.054	0.054	0.052	0.055	0.054	0.059	0.055	0.054	0.050	0.053	0.055	0.056	0.059	0.047	0.056	0.048	0.059	0.053
0	0.85	0.052	0.046	0.048	0.051	0.052	0.055	0.050	0.054	0.047	0.052	0.050	0.058	0.048	0.058	0.046	0.060	0.048	0.062
0	0.95	0.050	0.050	0.044	0.056	0.051	0.061	0.059	0.054	0.053	0.059	0.060	0.062	0.056	0.059	0.050	0.066	0.056	0.065
0	0.99	0.050	0.050	0.044	0.056	0.051	0.061	0.050	0.044	0.045	0.051	0.049	0.053	0.056	0.049	0.052	0.055	0.056	0.059
0.03	0.65	0.929	1.000	0.928	1.000	0.929	1.000	0.914	1.000	0.914	1.000	0.914	1.000	0.789	1.000	0.789	1.000	0.789	1.000
0.03	0.75	0.926	1.000	0.924	1.000	0.926	1.000	0.920	1.000	0.919	1.000	0.920	1.000	0.824	1.000	0.823	1.000	0.824	1.000
0.03	0.85	0.920	1.000	0.919	1.000	0.920	1.000	0.912	1.000	0.911	1.000	0.912	1.000	0.812	1.000	0.810	1.000	0.812	1.000
0.03	0.95	0.921	1.000	0.919	1.000	0.921	1.000	0.899	1.000	0.898	1.000	0.899	1.000	0.846	1.000	0.841	1.000	0.846	1.000
0.03	0.99	0.922	1.000	0.922	1.000	0.922	1.000	0.901	1.000	0.897	1.000	0.901	1.000	0.842	1.000	0.838	1.000	0.842	1.000
0.06	0.65	0.956	1.000	0.956	1.000	0.956	1.000	0.961	1.000	0.961	1.000	0.961	1.000	0.897	1.000	0.895	1.000	0.897	1.000
0.06	0.75	0.965	1.000	0.965	1.000	0.965	1.000	0.950	1.000	0.950	1.000	0.950	1.000	0.901	1.000	0.899	1.000	0.901	1.000
0.06	0.85	0.960	1.000	0.960	1.000	0.960	1.000	0.958	1.000	0.957	1.000	0.958	1.000	0.915	1.000	0.912	1.000	0.915	1.000
0.06	0.95	0.957	1.000	0.957	1.000	0.957	1.000	0.950	1.000	0.949	1.000	0.950	1.0000	0.911	1.000	0.909	1.000	0.912	1.000
0.06	0.99	0.965	1.000	0.964	1.000	0.965	1.000	0.948	1.000	0.947	1.000	0.948	1.000	0.917	1.000	0.914	1.000	0.917	1.000
0.12	0.65	0.982	1.000	0.982	1.000	0.982	1.000	0.975	1.000	0.975	1.000	0.975	1.000	0.946	1.000	0.946	1.000	0.946	1.000
0.12	0.75	0.988	1.000	0.988	1.000	0.988	1.000	0.973	1.000	0.973	1.000	0.973	1.000	0.950	1.000	0.949	1.000	0.950	1.000
0.12	0.85	0.982	1.000	0.982	1.000	0.982	1.000	0.981	1.000	0.981	1.000	0.981	1.000	0.960	1.000	0.959	1.000	0.960	1.000
0.12	0.95	0.982	1.000	0.986	1.000	0.987	1.000	0.976	1.000	0.976	1.000	0.976	1.000	0.955	1.000	0.953	1.000	0.955	1.000
0.12	0.99	0.982	1.000	0.981	1.000	0.982	1.000	0.977	1.000	0.977	1.000	0.977	1.000	0.959	1.000	0.959	1.000	0.959	1.000

Table 4.9: Simulation results based on the classical (T^c) and robust (T^r) methods for Model II (Dense model 2) with random errors $\sqrt[3]{m}\varepsilon_i$, where $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.056	0.050	0.054	0.052	0.057	0.054	0.048	0.049	0.046	0.050	0.048	0.052	0.055	0.052	0.054	0.054	0.055	0.057
0	0.75	0.052	0.052	0.049	0.053	0.052	0.055	0.054	0.047	0.050	0.047	0.054	0.050	0.056	0.050	0.050	0.054	0.056	0.059
0	0.85	0.052	0.049	0.048	0.057	0.052	0.061	0.060	0.055	0.057	0.057	0.061	0.059	0.056	0.050	0.049	0.054	0.056	0.059
0	0.95	0.054	0.054	0.051	0.058	0.054	0.061	0.058	0.047	0.056	0.054	0.058	0.058	0.060	0.046	0.055	0.051	0.060	0.055
0	0.99	0.057	0.052	0.052	0.060	0.058	0.060	0.052	0.061	0.047	0.062	0.052	0.067	0.054	0.052	0.053	0.056	0.055	0.053
0.03	0.65	0.472	1.000	0.468	1.000	0.475	1.000	0.436	1.000	0.432	1.000	0.436	1.000	0.157	0.990	0.155	0.991	0.157	0.991
0.03	0.75	0.480	1.000	0.477	1.000	0.472	1.000	0.428	1.000	0.424	1.000	0.429	1.000	0.178	1.000	0.171	1.000	0.178	1.000
0.03	0.85	0.488	1.000	0.486	1.000	0.488	1.000	0.412	1.000	0.405	1.000	0.413	1.000	0.177	1.000	0.167	1.000	0.178	1.000
0.03	0.95	0.477	1.000	0.489	1.000	0.477	1.000	0.396	1.000	0.395	1.000	0.398	1.000	0.207	1.000	0.198	1.000	0.208	1.000
0.03	0.99	0.496	1.000	0.488	1.000	0.497	1.000	0.392	1.000	0.383	1.000	0.393	1.000	0.201	1.000	0.193	1.000	0.203	1.000
0.06	0.65	0.696	1.000	0.695	1.000	0.696	1.000	0.686	1.000	0.684	1.000	0.687	1.000	0.361	1.000	0.354	1.000	0.361	1.000
0.06	0.75	0.715	1.000	0.711	1.000	0.715	1.000	0.662	1.000	0.655	1.000	0.662	1.000	0.403	1.000	0.398	1.000	0.404	1.000
0.06	0.85	0.715	1.000	0.710	1.000	0.715	1.000	0.642	1.000	0.637	1.000	0.643	1.000	0.404	1.000	0.397	1.000	0.405	1.000
0.06	0.95	0.697	1.000	0.695	1.000	0.697	1.000	0.645	1.000	0.639	1.000	0.645	1.000	0.480	1.000	0.472	1.000	0.480	1.000
0.06	0.99	0.720	1.000	0.715	1.000	0.720	1.000	0.638	1.000	0.633	1.000	0.640	1.000	0.462	1.000	0.454	1.000	0.462	1.000
0.12	0.65	0.841	1.000	0.840	1.000	0.840	1.000	0.846	1.000	0.845	1.000	0.846	1.000	0.619	1.000	0.615	1.000	0.619	1.000
0.12	0.75	0.845	1.000	0.842	1.000	0.845	1.000	0.842	1.000	0.840	1.000	0.843	1.000	0.672	1.000	0.671	1.000	0.675	1.000
0.12	0.85	0.857	1.000	0.856	1.000	0.858	1.000	0.827	1.000	0.825	1.000	0.827	1.000	0.648	1.000	0.643	1.000	0.648	1.000
0.12	0.95	0.863	1.000	0.860	1.000	0.864	1.000	0.799	1.000	0.798	1.000	0.799	1.000	0.706	1.000	0.703	1.000	0.707	1.000
0.12	0.99	0.836	1.000	0.834	1.000	0.836	1.000	0.812	1.000	0.808	1.000	0.812	1.000	0.699	1.000	0.693	1.000	0.699	1.000

Table 4.10: Simulation results based on the classical (T^c) and robust (T^r) methods for Model III (Sparse model) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ stand for the power values under alternative hypotheses.

		Setting 1						Setting 2						Setting 3					
		Wald		F		Likelihood		Wald		F		Likelihood		Wald		F		Likelihood	
a	γ	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r	T^c	T^r
0	0.65	0.046	0.053	0.044	0.053	0.046	0.055	0.053	0.050	0.048	0.052	0.053	0.055	0.057	0.053	0.054	0.051	0.057	0.054
0	0.75	0.051	0.055	0.050	0.055	0.051	0.057	0.059	0.051	0.057	0.054	0.059	0.056	0.045	0.065	0.044	0.068	0.046	0.065
0	0.85	0.049	0.054	0.046	0.054	0.049	0.057	0.055	0.047	0.052	0.048	0.055	0.050	0.058	0.054	0.052	0.057	0.058	0.060
0	0.95	0.065	0.055	0.059	0.057	0.065	0.060	0.073	0.053	0.069	0.057	0.073	0.060	0.068	0.062	0.066	0.064	0.069	0.060
0	0.99	0.084	0.052	0.079	0.056	0.086	0.062	0.075	0.062	0.071	0.070	0.077	0.070	0.088	0.052	0.083	0.054	0.091	0.059
0.03	0.65	0.051	0.170	0.050	0.166	0.051	0.617	0.057	0.122	0.053	0.122	0.057	0.125	0.068	0.066	0.064	0.067	0.068	0.070
0.03	0.75	0.056	0.158	0.053	0.166	0.056	0.174	0.057	0.123	0.054	0.129	0.057	0.132	0.051	0.058	0.048	0.062	0.051	0.064
0.03	0.85	0.060	0.143	0.056	0.148	0.060	0.157	0.057	0.127	0.051	0.130	0.057	0.136	0.059	0.077	0.056	0.083	0.059	0.086
0.03	0.95	0.071	0.143	0.066	0.150	0.072	0.158	0.065	0.115	0.061	0.118	0.067	0.128	0.069	0.069	0.065	0.069	0.069	0.075
0.03	0.99	0.087	0.145	0.084	0.156	0.090	0.168	0.092	0.094	0.086	0.099	0.092	0.109	0.090	0.077	0.083	0.090	0.091	0.090
0.06	0.65	0.063	0.477	0.062	0.470	0.063	0.478	0.058	0.315	0.056	0.410	0.058	0.415	0.052	0.105	0.051	0.105	0.052	0.110
0.06	0.75	0.064	0.490	0.060	0.485	0.064	0.492	0.064	0.400	0.058	0.399	0.064	0.406	0.054	0.117	0.051	0.121	0.054	0.124
0.06	0.85	0.071	0.476	0.069	0.481	0.071	0.495	0.061	0.384	0.058	0.390	0.061	0.403	0.056	0.133	0.053	0.136	0.057	0.145
0.06	0.95	0.080	0.500	0.074	0.509	0.081	0.524	0.079	0.375	0.077	0.330	0.079	0.341	0.077	0.131	0.073	0.139	0.077	0.149
0.06	0.99	0.103	0.459	0.096	0.480	0.105	0.501	0.097	0.271	0.088	0.287	0.098	0.301	0.091	0.131	0.084	0.142	0.091	0.157
0.12	0.65	0.109	0.901	0.103	0.899	0.109	0.901	0.092	0.925	0.089	0.919	0.092	0.924	0.063	0.308	0.058	0.309	0.063	0.317
0.12	0.75	0.112	0.936	0.107	0.935	0.112	0.936	0.098	0.921	0.094	0.921	0.099	0.924	0.070	0.360	0.067	0.358	0.070	0.365
0.12	0.85	0.115	0.955	0.108	0.948	0.115	0.951	0.092	0.921	0.087	0.916	0.092	0.918	0.061	0.397	0.056	0.405	0.061	0.415
0.12	0.95	0.118	0.959	0.112	0.963	0.118	0.965	0.099	0.875	0.096	0.876	0.099	0.881	0.077	0.445	0.069	0.458	0.078	0.472
0.12	0.99	0.135	0.947	0.129	0.952	0.136	0.956	0.089	0.853	0.084	0.864	0.091	0.873	0.094	0.426	0.090	0.444	0.095	0.464

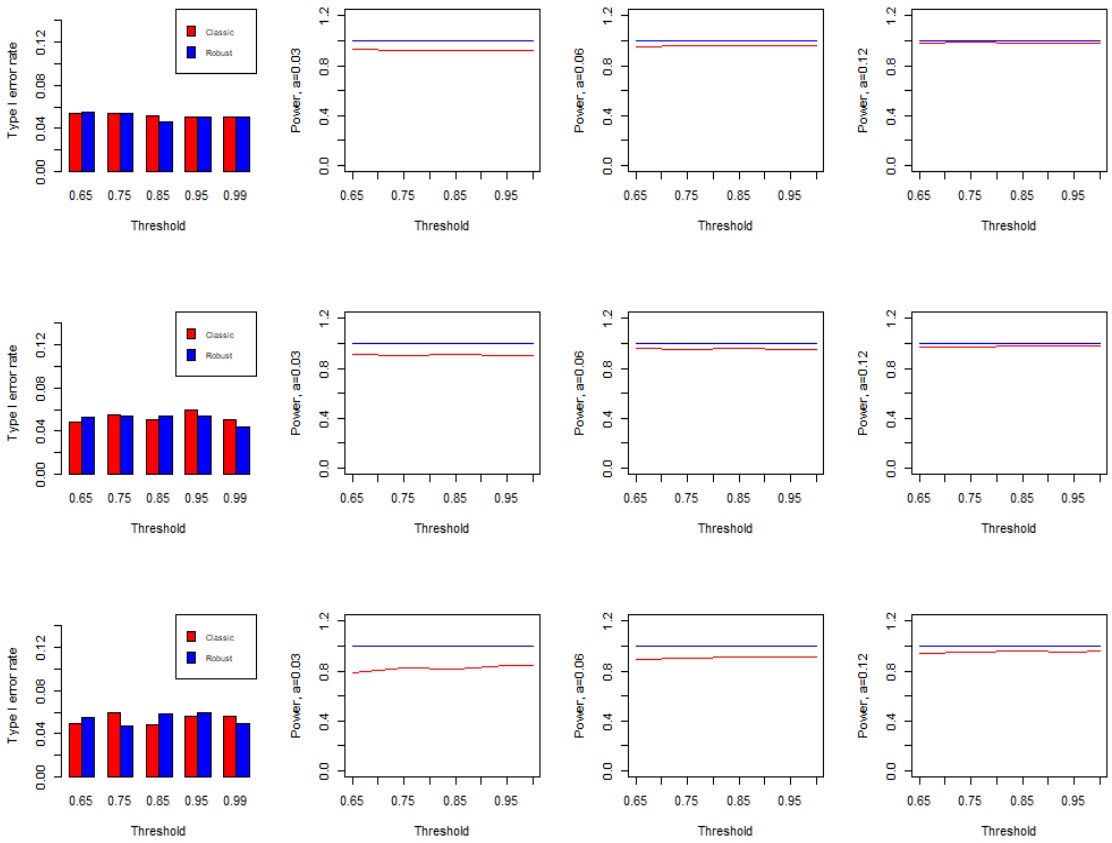

Figure 4.9: The estimated Type I errors (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model I with a small sample size ($n = 200$). The results in the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is based on the robust Wald type testing procedure.

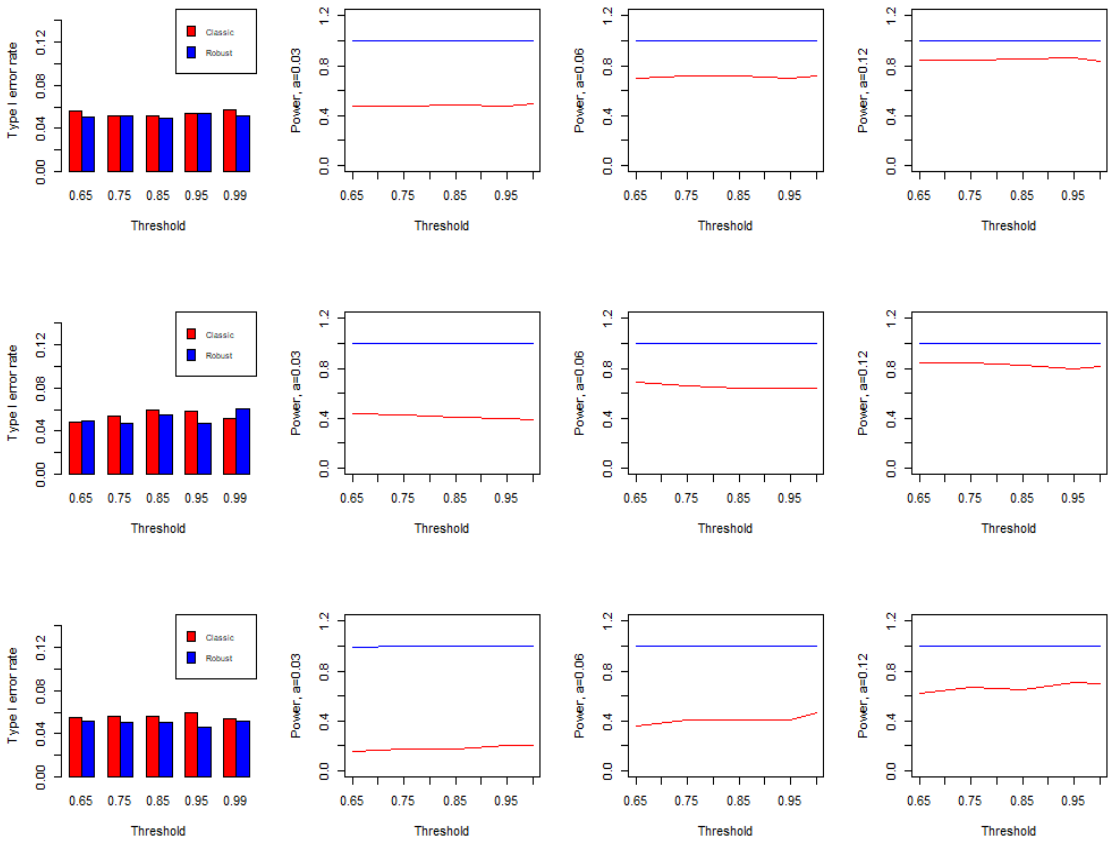

Figure 4.10: The estimated Type I errors (depicted as the height of the bars) and power curves are shown from the left to the right panels for Model II with a small sample size ($n = 200$). The results in the settings 1-3 are aligned from the top to the bottom. In addition, the red line is based on the classical Wald testing procedure, while the blue line is corresponding to the robust Wald type testing procedure.

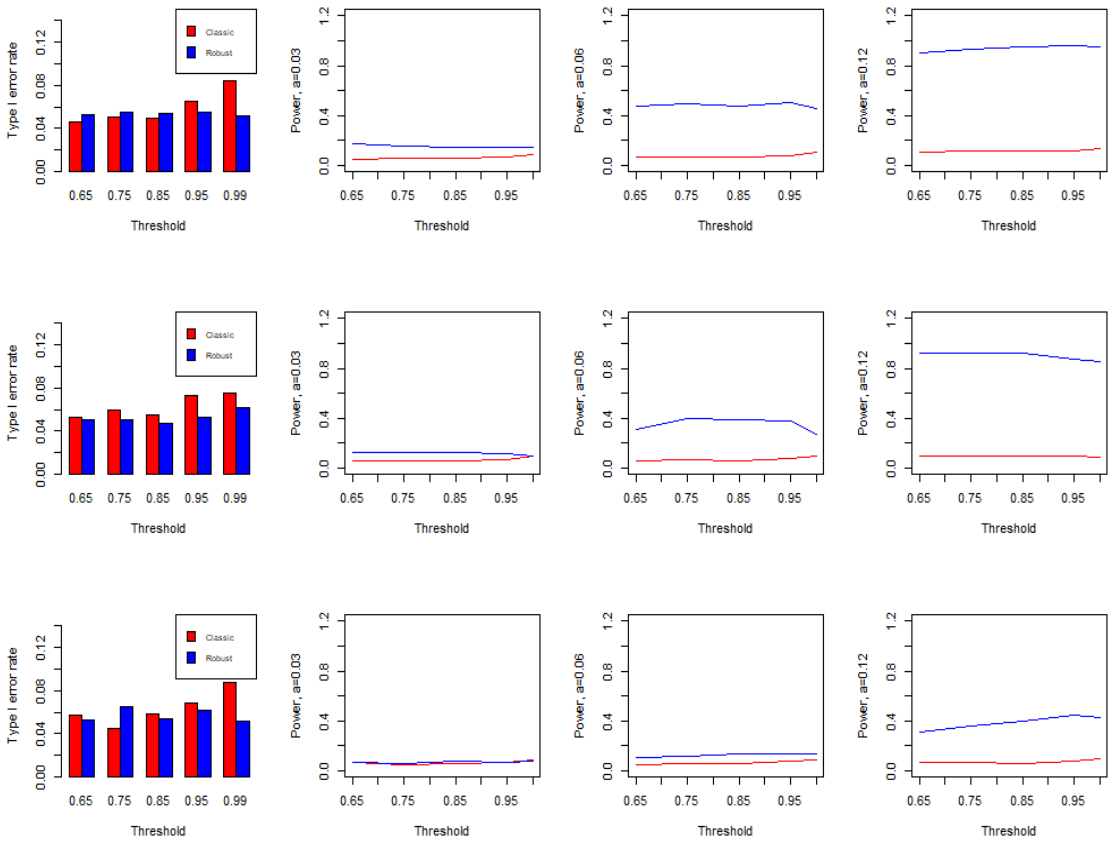

Figure 4.11: The estimated Type I errors (depicted as the height of the bars) and power curves are shown from the left to right panel according to Model III for small sample size ($n = 200$). The results in setting 1-3 are aligned from the top to bottom. In addition, red line is based on classic Wald testing procedure, blue line is based on robust Wald type testing procedure.

Part III: Simulation results based on Huber's loss function.

Table 4.11: Simulation results based on the Huber's loss function for Model I (Dense model 1) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1			Setting 2			Setting 3		
a	γ	Wald	F	Likelihood	Wald	F	Likelihood	Wald	F	Likelihood
0	0.65	0.052	0.056	0.056	0.054	0.052	0.052	0.047	0.045	0.045
0	0.75	0.052	0.053	0.055	0.050	0.045	0.045	0.059	0.063	0.065
0	0.85	0.046	0.054	0.055	0.052	0.053	0.053	0.055	0.053	0.054
0	0.95	0.052	0.047	0.049	0.049	0.054	0.056	0.042	0.051	0.052
0	0.99	0.049	0.053	0.055	0.053	0.059	0.061	0.055	0.049	0.051
0.03	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.12: Simulation results based on the Huber's loss function for Model II (Dense model 2) with random errors $\sqrt[3]{m}\varepsilon_i$, where $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ are the power values under alternative hypotheses.

		Setting 1			Setting 2			Setting 3		
a	γ	Wald	F	Likelihood	Wald	F	Likelihood	Wald	F	Likelihood
0	0.65	0.051	0.049	0.050	0.057	0.047	0.048	0.047	0.055	0.055
0	0.75	0.046	0.045	0.048	0.051	0.056	0.056	0.050	0.059	0.060
0	0.85	0.048	0.053	0.054	0.055	0.056	0.056	0.054	0.058	0.059
0	0.95	0.054	0.051	0.052	0.051	0.050	0.051	0.044	0.053	0.053
0	0.99	0.052	0.058	0.058	0.051	0.050	0.052	0.057	0.051	0.051
0.03	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.06	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.65	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.12	0.99	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.13: Simulation results based on the Huber's loss fucntion for Model III (Sparse model) with random errors $\varepsilon_i \sim$ Cauchy $(0, 1)$. The rows with $a = 0$ stand for the Type I error rates of both methods under different settings, while the rows with $a > 0$ stand for the power values under alternative hypotheses.

		Setting 1			Setting 2			Setting 3		
a	γ	Wald	F	Likelihood	Wald	F	Likelihood	Wald	F	Likelihood
0	0.65	0.047	0.048	0.048	0.063	0.049	0.049	0.051	0.051	0.053
0	0.75	0.052	0.051	0.052	0.053	0.047	0.048	0.053	0.058	0.058
0	0.85	0.046	0.049	0.049	0.051	0.051	0.052	0.057	0.048	0.048
0	0.95	0.053	0.053	0.054	0.051	0.053	0.054	0.053	0.049	0.049
0	0.99	0.049	0.046	0.046	0.056	0.050	0.051	0.056	0.057	0.058
0.03	0.65	0.529	0.570	0.573	0.441	0.464	0.465	0.126	0.126	0.127
0.03	0.75	0.544	0.561	0.562	0.430	0.465	0.466	0.138	0.154	0.155
0.03	0.85	0.570	0.604	0.606	0.433	0.464	0.466	0.142	0.151	0.152
0.03	0.95	0.570	0.570	0.601	0.375	0.394	0.397	0.157	0.162	0.165
0.03	0.99	0.575	0.582	0.592	0.283	0.328	0.330	0.132	0.142	0.145
0.06	0.65	0.978	0.987	0.987	0.959	0.973	0.973	0.376	0.401	0.403
0.06	0.75	0.991	0.995	0.9995	0.966	0.976	0.976	0.416	0.465	0.466
0.06	0.85	0.986	0.994	0.994	0.965	0.973	0.973	0.432	0.469	0.471
0.06	0.95	0.998	0.998	0.998	0.945	0.949	0.950	0.569	0.600	0.601
0.06	0.99	0.994	0.998	0998	0.914	0.935	0.935	0.522	0.538	0.539
0.12	0.65	1.000	1.000	1.000	1.000	1.000	1.000	0.860	0.882	0.882
0.12	0.75	1.000	1.000	1.000	1.000	1.000	1.000	0.926	0.935	0.935
0.12	0.85	1.000	1.000	1.000	1.000	1.000	1.000	0.969	0.975	0.975
0.12	0.95	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.996	0.996
0.12	0.99	1.000	1.000	1.000	1.000	1.000	1.000	0.988	0.994	0.994

5 Robust Variable Selection via Group VIF Regression in Functional Multiple Linear Models

5.1 Introduction

Functional data are commonly sampled discretely over a continuous domain, particularly in the context of time. It is typically assumed that there exists an underlying curve describing the data. The high dimensionality and multicollinearity among functional predictors pose challenges in model selection for functional data, potentially leading to erroneous scientific conclusions. Similar to linear and generalized linear regression analysis, variable selection plays a crucial role in functional data analysis, thus emphasizing the necessity of conducting variable selection on functional covariates. The prevailing approach to tackle this issue involves a two-stage procedure. The initial step entails formulating the functional regression model within a conventional multiple model framework, thereby addressing the challenge of infinite dimensionality inherent in functional data. Subsequently, a group se-

lection method is employed in high-dimensional models to identify the influential functional predictors. Several studies have been published on these issues. Lian (2013) proposed functional group-SCAD method. Zhao et al. (2012) investigated the wavelet-based LASSO approach for regressing scalars on functions, examining its asymptotic convergence and finite-sample performance through simulations and real-data applications

However, the efficacy of these methods is compromised in the presence of outliers; therefore, it is imperative to employ a robust variable selection method that exhibits resistance to outliers in functional regression. Huang et al. (2016) proposed a robust variable selection procedure with data-driven basis functional principal components and LAD loss function for functional linear regression when there are multiple functional predictors. Pannu and Billor (2017) proposed a robust functional predictor selection method, the LAD-group LASSO, for a functional linear regression model, since group LASSO selects grouped variables rather than individual variables. Subsequently, an adaptive version of the LAD-group LASSO, known as LAD-agLASSO, was proposed to enhance model accuracy in the presence of outliers in response variables (Pannu and Billor, 2022).

Although the aforementioned research works have primarily focused on penalized methods, their efficacy may be limited when dealing with large data sets due to the

exponential increase in potential models and computationally expensive implementations optimized for cross-validation criteria. Motivated by the proposed algorithm for handling group structure data using the group variance inflation factor (VIF, hereinafter) regression (Ding et al., 2023), we propose a robust group VIF procedure to address the issue of selecting functional covariates in the functional multiple linear regression model.

The primary objective of this chapter is to develop a robust variable selection procedure that exhibits resistance to outliers in the context of functional multiple linear regression. In Section 5.2, we present the methodology and algorithm for functional multiple linear regression along with its reformulation. Section 5.3 includes numerical studies and corresponding results. An real data example, further studies, and conclusions are provided in Sections 5.4-5.5, respectively.

5.2 Methodology

5.2.1 Reformulation of Functional Multiple Linear Regression

Let y be a real-valued random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, $X_j(t) \in L^2(\mathcal{T}_j)$ be a functional covariate defined on a compact support $\mathcal{T}_j$, $j = 1, 2, \ldots, J$. Without loss of generality, we assume that both the response variable and predictors are centered. Subsequently, we consider the following functional multiple

linear regression model

$$y = \sum_{j=1}^{J} \int_{\mathcal{T}_j} \beta_j(t) X_j(t) dt + \varepsilon_0, \tag{5.1}$$

where $\beta_j(t)$ is the functional coefficient, and ε_0 are independently and identically distributed random error with zero mean and variance $\sigma^2_{\varepsilon_0}$.

It is well known that the inherent problem of functional data analysis is how to overcome the infinite dimensionality of functional predictors and the regression coefficient functions. Since the functional regressors and coefficient functions belong to $L^2(\cdot)$ spaces, they can be approximated by a sufficient large number of basis functions $\{\phi_{jk}\}_{k=1}^{\infty}$ of these spaces (Ramsay and Silverman, 2005). Given its enhanced flexibility, we opt to represent the functional predictors $X_j(t)$ and coefficients $\beta_j(t)$ in equation (5.1) using FPCA basis functions.

Denote the covariance function $\text{Cov}(X_j(s), X_j(t)) = \Sigma_j(s,t)$, $j = 1, 2, \ldots, J$, we can employ Mercer's Theorem to derive the spectral decomposition of the covariance function by

$$\Sigma_j(s,t) = \sum_{k=1}^{\infty} \lambda_{jk} \phi_{jk}(s) \phi_{jk}(t),$$

where λ_{jk} is the eigenvalues with non-increasing order $\lambda_{j1} \geqslant \lambda_{j2} \geqslant \cdots \geqslant 0$, $\sum_{k=1}^{\infty} \lambda_{jk} < \infty$ and $\{\phi_{jk},\, k \geqslant 1\}$ are the corresponding orthonormal eigenfunctions. According to

the Karhunen-Loève representation, we have

$$X_j(t) = \sum_{k=1}^{\infty} \xi_{jk}\phi_{jk}(t),$$

where $\{\xi_{jk} = \int_{\mathcal{T}_j} X_j(t)\phi_{jk}(t)dt, k \geqslant 1 \text{ and } j = 1, 2, \cdots, J\}$ are functional principal component scores with $\mathrm{E}(\xi_{jk}) = 0, \mathrm{Var}(\xi_{jk}) = \lambda_{jk}$, and $\mathrm{E}(\xi_{jk}\xi_{jk'}) = 0$, for $k \neq k'$. Then, based on the same FPCA basis functions $\{\phi_{j1}(t), \phi_{j2}(t), \ldots, \ j = 1, 2, \ldots, J\}$, we have

$$\beta_j(t) = \sum_{k=1}^{\infty} \beta_{jk}\phi_{jk}(t),$$

where $\beta_{jk} = \int_{\mathcal{T}_j} \beta_j(t)\phi_{jk}(t)dt, k \geqslant 1 \text{ and } j = 1, 2, \cdots, J$ are the unknown basis coefficients. Therefore, the model (5.1) can be expressed in the following form:

$$y = \sum_{j=1}^{J}\sum_{k=1}^{\infty} \beta_{jk} \int_{\mathcal{T}_j} X_j(t)\phi_{jk}(t)dt + \varepsilon_0 = \sum_{j=1}^{J}\sum_{k=1}^{\infty} \xi_{jk}\beta_{jk} + \varepsilon_0. \tag{5.2}$$

In practical applications, the regression model is usually approximated by a set of finite basis functions $\{\phi_{jk}\}_{k=1}^{k_j}$, for $j = 1, 2, \ldots, J$. The choice of the "cut-off value" k_j, i.e., the truncation parameter, is related to the characteristics of the different bases. Thus, the functional regression model (5.2) can be approximated by

$$y \approx \sum_{j=1}^{J}\sum_{k=1}^{k_j} \xi_{jk}\beta_{jk} + \varepsilon = \sum_{j=1}^{J} \mathbf{z}_j^T\boldsymbol{\beta}_j + \varepsilon, \tag{5.3}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jk_j})^T$, $\mathbf{z}_j = (\xi_{j1}, \xi_{j2}, \ldots, \xi_{jk_j})^T$ for $j = 1, 2, \cdots, J$, and $\varepsilon = \sum_{j=1}^{J}\sum_{k=k_j+1}^{\infty} \xi_{jk}\beta_{jk} + \varepsilon_0$ have zero mean and large variance $\sigma_\varepsilon^2 = \sum_{j=1}^{J}\sum_{k=k_j+1}^{\infty} \lambda_{jk}\beta_{jk}^2 + \sigma_{\varepsilon_0}^2$(Su et al., 2017).

Remark 5.1 The choice of truncation parameters k_j, $j = 1, 2, \ldots, J$ is very important in functional data analysis. In practice, there are some empirical choice of this value, such as PVE (Percentage of Variance Explained) method, leading PCs (Principal Components) method (Cardot et al., 2003; Kong et al., 2013; Swihart et al., 2014), CV (Cross-Validation) criterion (Qingguo, 2017), and information (AIC or BIC) criterion (Kato et al., 2012). In our simulation study, we adopt the PVE method to select k_j because of its computational efficiency and convenience. The percentage level used in our numerical studies is set at 95%.

Suppose that we have n observations $\{(x_{ij}(t), y_i), i = 1, 2, \ldots, n, j = 1, 2, \ldots, J\}$, the spectral decomposition of the empirical version of $\Sigma(s,t)$ is

$$\widehat{\Sigma}_j(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}(s)x_{ij}(t) = \sum_{k}^{\infty} \hat{\lambda}_{jk}\hat{\phi}_{jk}(s)\hat{\phi}_{jk}(t), \qquad s,t \in \mathcal{T}_j,$$

where $\{(\hat{\lambda}_{jk}, \hat{\phi}_{jk}), \hat{\lambda}_{j1} \geqslant \hat{\lambda}_{j2} \geqslant \cdots \geqslant 0,\ k \geqslant 1\}$ are eigenvalue and eigenfunction pairs of $\widehat{\Sigma}_j(s,t)$. Then, the regression model (5.3) can be reformulated as follows:

$$y_i \approx \sum_{j=1}^{J}\sum_{k=1}^{k_{jn}} \hat{\xi}_{ijk}\beta_{jk} + \varepsilon_i = \sum_{j=1}^{J} \widehat{\mathbf{z}}_{ij}^{\top}\boldsymbol{\beta}_j + \varepsilon_i, \tag{5.4}$$

where $\widehat{\mathbf{z}}_{ij} = (\hat{\xi}_{ij1}, \hat{\xi}_{ij2}, \ldots, \hat{\xi}_{ijk_j})^{\top}$ and $\hat{\xi}_{ijk} = \int_{\mathcal{T}_j} x_{ij}(t)\hat{\phi}_{jk}(t)dt$. Furthermore, we define $\mathbf{Z}_j = (\widehat{\mathbf{z}}_{1j}^{\top}, \widehat{\mathbf{z}}_{2j}^{\top}, \ldots, \widehat{\mathbf{z}}_{nj}^{\top})^{\top}$, and obtain the following linear regression model with J groups:

$$\boldsymbol{y} = \sum_{j=1}^{J} \mathbf{Z}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{5.5}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\mathbf{Z}_j$ is $n \times k_j$ design matrix, $\boldsymbol{\beta}_j$ is an unknown k_j-dimensional coefficient vector associated with the jth group $\mathbf{Z}_j$, and $\boldsymbol{\varepsilon}$ is $n \times 1$ error vector with mean $\mathbf{0}$ and variance matrix $\Sigma_{\boldsymbol{\varepsilon}}$.

5.2.2 Robust Group VIF

In this section, we propose a robust group variable inflation factor (VIF) selection procedure to identify significant functional predictors based on the innovative approach proposed by Ding et al. (2023) and Dupuis and Victoria-Feser (2013). First, we denote the index set of nonzero regression coefficient vectors as $\mathcal{G} = \{j : \boldsymbol{\beta} j \neq \mathbf{0},\ j = 1, \ldots, J\} \subseteq \{1, \ldots, J\}$ and $\sum_{j \in \mathcal{G}} k_j \leqslant n$. Let $\mathbf{Z}_{\mathcal{G}}$ be an $n \times (\sum_{j \in \mathcal{G}} k_j)$ matrix spanned by the predictors $\mathbf{Z}_j, j \in \mathcal{G}$ and we assume that $\mathbf{Z}_{\mathcal{G}}^{\top} \mathbf{Z}_{\mathcal{G}}$ is invertible. Our goal is to select all the group predictors $\mathbf{Z}_j$ corresponding to the set $\mathcal{G}$ in the model. To introduce our robust group VIF procedure, we initially consider a weighted least squares estimator

$$\widehat{\boldsymbol{\beta}}^{w} = (\mathbf{X}^{w\top} \mathbf{X}^{w})^{-1} \mathbf{X}^{w\top} \boldsymbol{y}^{w},$$

where $\mathbf{X}^{w} = \operatorname{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})\mathbf{X}$ and $\boldsymbol{y}^{w} = \operatorname{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})\boldsymbol{y}$. Note that the weights w_i, which are dependent on the data, incorporate information from observations to mitigate the impact of outliers.

Remark 5.2 The weighted slope estimator $\widehat{\boldsymbol{\beta}}^w$ is a robust estimator computed using the estimating equation

$$\sum_{i=1}^{n} w_i(r_i, k) r_i \boldsymbol{x}_i = \mathbf{0},$$

where $r_i = (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})/\sigma$. When $w_i(r_i, k) = \min\{1, k/|r_i|\}$ with $k = 1.345$, it becomes Huber estimators, see Huber and Ronchetti (2011).

Denote $\mathbf{Z}^w_{\mathcal{M}} = (\mathbf{Z}^w_{j_1}, \ldots, \mathbf{Z}^w_{j_M})$ with $\mathbf{Z}^w_{j_m} = \text{diag}(\sqrt{w_{ij_m}})\mathbf{Z}_{j_m}$, then the assumption on $\mathbf{Z}_{\mathcal{G}}^\top \mathbf{Z}_{\mathcal{G}}$ implies that $(\mathbf{Z}^w_{\mathcal{M}})^\top \mathbf{Z}^w_{\mathcal{M}}$ is invertible. We consider the following stepwise regression model:

$$\boldsymbol{y}^w = \sum_{m=1}^{M} \mathbf{Z}^w_{j_m} \boldsymbol{\beta}^w_{j_m} + \mathbf{Z}^w_{new} \boldsymbol{\beta}^w_{new} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, \sigma^2_{\mathcal{M}} \boldsymbol{I}), \tag{5.6}$$

where $\mathbf{Z}^w_{j_m}$, $j_m \in \mathcal{G}$, $m = 1, \ldots, M$, are $n \times j_m$ group predictors that have been added to the model and $\mathcal{M} = \{j_1, j_2, \cdots, j_m\} \subseteq \mathcal{G}$. $\mathbf{Z}^w_{new} = \text{diag}(\sqrt{w_{i(j_{new})}})\mathbf{Z}_{j_{new}}$ is a new $n \times j_{\text{new}}$ group predictor, $\boldsymbol{\beta}^w_{j_m}$ is j_m-dimensional unknown parameter vectors associated with the mth group $\mathbf{Z}^w_{j_m}$. We need to decide whether the new group predictor $\mathbf{Z}^w_{\text{new}}$ should be added to the model or not, that is, we need to test if $\boldsymbol{\beta}^w_{\text{new}} = \mathbf{0}$. The incorporation of $\mathbf{Z}^w_{\text{new}}$ into $\mathcal{G}$ implies that the incorporation of $\mathbf{Z}_{\text{new}}$ into $\mathcal{G}$ is also feasible.

Let $\boldsymbol{r}^w = (\boldsymbol{I} - \mathbf{H}^w_{\mathcal{M}})\boldsymbol{y}^w$ be the residual vector of projecting $\boldsymbol{y}^w$ on $\mathbf{Z}^w_{\mathcal{M}}$ with $\mathbf{H}^w_{\mathcal{M}} = \mathbf{Z}^w_{\mathcal{M}}(\mathbf{Z}^{w\top}_{\mathcal{M}} \mathbf{Z}^w_{\mathcal{M}})^{-1} \mathbf{Z}^{w\top}_{\mathcal{M}}$. In light of the stagewise algorithm (Tibshirani, 2015),

we only need to regress the new group predictor $\mathbf{Z}^w_{\text{new}}$ on the residual $\boldsymbol{r}^w$. Motivated by Lin et al. (2011) and Ding et al. (2023) , we consider the following stagewise regression:

$$\boldsymbol{r}^w = \mathbf{Z}^w_{\text{new}}\boldsymbol{\gamma}^w_{new} + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \tilde{\sigma}^2 I). \tag{5.7}$$

Let $\widehat{\boldsymbol{\beta}}^w_{new}$ be the least squares estimate of $\boldsymbol{\beta}^w_{new}$ in the model (5.6), and $\widehat{\boldsymbol{\gamma}}^w_{new}$ be the least squares estimate of $\boldsymbol{\gamma}^w_{new}$ in the model (5.7) with $(\mathbf{Z}^w_{new})^\top \mathbf{Z}^w_{new}$ invertible. In light of Ding et al. (2023), we have the following theorem.

Theorem 5.1 Under models (5.6) and (5.7),

$$\widehat{\boldsymbol{\gamma}}^w_{new} = (\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})^{-1}\boldsymbol{\Lambda}^2_w \widehat{\boldsymbol{\beta}}^w_{new}, \tag{5.8}$$

where $\boldsymbol{\Lambda}^2_w = \mathbf{Z}^{w\top}_{new}(\boldsymbol{I} - \mathbf{H}^w_{\mathcal{M}})\mathbf{Z}^w_{new}$.

Proof. Denote $\widetilde{\mathbf{Z}^w} = [\mathbf{Z}^w_{\mathcal{M}}, \mathbf{Z}^w_{\text{new}}]$, where $\mathbf{Z}^w_{\mathcal{M}} = (\mathbf{Z}^w_{j_1}, \ldots, \mathbf{Z}^w_{j_M})$, then

$$\widetilde{\mathbf{Z}^w}^\top \widetilde{\mathbf{Z}^w} = \begin{pmatrix} \mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}} & \mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{new} \\ \mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{\mathcal{M}} & \mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new} \end{pmatrix},$$

and

$$(\widetilde{\mathbf{Z}^w}^\top \widetilde{\mathbf{Z}^w})^{-1} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ -\boldsymbol{\Lambda}^{-2}_w \mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{\mathcal{M}}(\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}})^{-1} & \boldsymbol{\Lambda}^{-2}_w \end{pmatrix},$$

where $\boldsymbol{A}_{11} = (\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}})^{-1} + (\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}})^{-1}\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{new}\boldsymbol{\Lambda}^{-2}_w\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{\mathcal{M}}(\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}})^{-1}$, $\boldsymbol{A}_{12} = -(\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{\mathcal{M}})^{-1}\mathbf{Z}^{w\top}_{\mathcal{M}}\mathbf{Z}^w_{new}\boldsymbol{\Lambda}^{-2}_w$, $\boldsymbol{\Lambda}^2_w = (\mathbf{Z}^w_{new})^\top(\boldsymbol{I} - \mathbf{H}^w_{\mathcal{M}})\mathbf{Z}^w_{new}$.

Note that $\widehat{\boldsymbol{\beta}}^w_{new}$ is a vector consists of the last p_{new} elements of $(\widetilde{\mathbf{Z}^w}^\top \widetilde{\mathbf{Z}^w})^{-1}\widetilde{\mathbf{Z}^w}^\top \boldsymbol{y}^w$. Then we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^w_{new} &= -\boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}\mathbf{H}^w_{\mathcal{M}}\boldsymbol{y}^w + \boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}\boldsymbol{y}^w \\
&= \boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}(\boldsymbol{I} - \mathbf{H}^w_{\mathcal{M}})\boldsymbol{y}^w \\
&= \boldsymbol{\Lambda}_w^{-2}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})^{-1}\mathbf{Z}^{w\top}_{new}\boldsymbol{r}^w \\
&= \boldsymbol{\Lambda}_w^{-2}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\widehat{\boldsymbol{\gamma}}^w_{new},
\end{aligned}$$

which completes the proof. □

By Theorem 5.1, estimate $\widehat{\boldsymbol{\gamma}}^w_{new}$ is simply a linear transformation of $\widehat{\boldsymbol{\beta}}^w_{new}$. Thus, both of hypothesis tests $\boldsymbol{\beta}^w_{new} = \mathbf{0}$ and $\boldsymbol{\gamma}^w_{new} = \mathbf{0}$ can be used to find out whether or not the new predictor $\mathbf{Z}^w_{new}$ contributes to the linear model. Based on Theorem 5.1, we have following conclusion about $\widehat{\boldsymbol{\beta}}^w_{new}$.

Theorem 5.2 Under models (5.6) and (5.7),

$$\widehat{\boldsymbol{\beta}}^w_{new} = \boldsymbol{\Lambda}_w^{-2}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\widehat{\boldsymbol{\gamma}}^w_{new} \overset{d}{\sim} N(\boldsymbol{\beta}^w_{new}, \sigma^2_{\mathcal{M}}\boldsymbol{\Lambda}_w^{-2}). \tag{5.9}$$

Proof. By the proof of Theorem 5.1, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^w_{new} =& \boldsymbol{\Lambda}_w^{-2}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\widehat{\boldsymbol{\gamma}}^w_{new} \\
=& \boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}(\boldsymbol{I}-\mathbf{H}^w_{\mathcal{M}})\boldsymbol{y}^w \\
=& \boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}(\boldsymbol{I}-\boldsymbol{H}^w_{\mathcal{M}})(\sum_{m=1}^{M}\mathbf{Z}^w_{j_m}\boldsymbol{\beta}^w_{j_m}+\mathbf{Z}^w_{\text{new}}\boldsymbol{\beta}^w_{\text{new}}+\boldsymbol{\varepsilon}) \\
=& \boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}(\boldsymbol{I}-\boldsymbol{H}^w_{\mathcal{M}})(\mathbf{Z}^w_{\text{new}}\boldsymbol{\beta}^w_{\text{new}}+\boldsymbol{\varepsilon}) \\
=& \boldsymbol{\beta}^w_{\text{new}}+\boldsymbol{\Lambda}_w^{-2}\mathbf{Z}^{w\top}_{new}(\boldsymbol{I}-\mathbf{H}^w_{\mathcal{M}})\boldsymbol{\varepsilon}.
\end{aligned}$$

Note that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2_{\mathcal{M}}\boldsymbol{I})$ and $\boldsymbol{I}-\mathbf{H}^w_{\mathcal{M}}$ is an idempotent symmetric matrix. We have $\widehat{\boldsymbol{\beta}}^w_{new} \sim N_{p_{\text{new}}}(\boldsymbol{\beta}^w_{\text{new}}, \sigma^2_{\mathcal{M}}\boldsymbol{\Lambda}_w^{-2})$, which completes the proof. □

To test the null hypothesis

$$H_0 : \boldsymbol{\beta}^w_{\text{new}} = \mathbf{0},$$

under H_0 and (5.9), a test statistic is given by

$$T_w = \widehat{\boldsymbol{\gamma}}^{w\top}_{new}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\boldsymbol{\Lambda}_w^{-2}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\widehat{\boldsymbol{\gamma}}^w_{new}/\sigma^2_{\mathcal{M}} \overset{d}{\sim} \chi^2_{p_{new}}.$$

Similar to Lin et al. (2011), $\sigma^2_{\mathcal{M}}$ can be estimated by the mean square error (MSE) under the null hypothesis H_0, which can prevent over-fitting or generating selection bias (Foster and Stine, 2004). Note that $\chi^2_{p_{new}}$ represents the χ-squared distribution with p_{new} degrees of freedom. By Theorem 5.1, we have

$$(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})^{-1}\boldsymbol{\Lambda}_w^2 = \boldsymbol{I}_{p_{new}} - (\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})^{-1}\mathbf{Z}^{w\top}_{new}\mathbf{H}^w_{\mathcal{M}}\mathbf{Z}^w_{new}.$$

Our robust group VIF algorithm is given below

(1) Obtain the residuals $\boldsymbol{r}^w_{\mathcal{M}} = (\boldsymbol{I} - \mathbf{H}^w_{\mathcal{M}})\boldsymbol{y}^w$.

(2) Compute $\widehat{\boldsymbol{\gamma}}^w_{new} = (\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})^{-1}\mathbf{Z}^{w\top}_{new}\boldsymbol{r}^w_{\mathcal{M}}$, and $\widehat{\sigma}_{\mathcal{M}} = \text{MAD}(\boldsymbol{r}^w_{\mathcal{M}})$.

(3) Compute the test statistic

$$\begin{aligned}\widehat{T}_w =& \widehat{\boldsymbol{\gamma}}^{w\top}_{new}(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\boldsymbol{\Lambda}^{-2}_w(\mathbf{Z}^{w\top}_{new}\mathbf{Z}^w_{new})\widehat{\boldsymbol{\gamma}}^w_{new}/\widehat{\sigma}^2_{\mathcal{M}} \\ =& (\mathbf{Z}^{w\top}_{new}\boldsymbol{r}^w_{\mathcal{M}})^\top\boldsymbol{\Lambda}^{-2}_w\mathbf{Z}^{w\top}_{new}\boldsymbol{r}^w_{\mathcal{M}}/\widehat{\sigma}^2_{\mathcal{M}}. \end{aligned} \tag{5.10}$$

(4) If $\widehat{T}_w > \chi^2_{p_{\text{new}}}(\alpha)$, then the new predictor $\mathbf{Z}^w_{new}$ is added to the model, where $\chi^2_{p_{\text{new}}}(\alpha)$ represents the upper α quantile of a χ-squared distribution with p_{new} degrees of freedom. It implies that we can choose $\mathbf{Z}_{new}$ from the set $\mathcal{G}$.

In light of Lin et al. (2011), we incorporate the above procedure with a stagewise regression algorithm using an α-investing rule (Foster and Stine, 2008). A more detailed algorithm can be found in the Appendix.

Remark 5.3 The α-investing rule is an adaptive procedure which controls the false discovery rate (FDR) bound by comparing a threshold α_i with p-value of a test statistic and adjusting dynamically such that it can control over-fitting when searching for new features (Foster and Stine, 2008). Denote so-called α wealth by u. We typically set the initial value of u to be 0.05, which is an allowance for the type I error. Then in the ith testing, at level α_i, if a rejection occurs, the current wealth u_i earns a pay-out $\triangle u$; otherwise, it will be reduced by $\alpha_i/(1-\alpha_i)$.

5.3 Numerical Experiments

5.3.1 Simulation Studies

In this section, simulation studies are performed to evaluate the finite sample performance of our proposed method. Without loss of generality, we assume $\mathcal{T}_j = [0,1]$, $j = 1, 2, \ldots, J$ and have n observations $\{(X_{ij}(t), y_i, \varepsilon_i),\ i = 1, 2, \ldots, n\}$ for our simulation experiments. In the following simulation studies, we consider the dense models where the functional covariates are observed at a set of 301 equidistant points in the interval $[0,1]$.

Example 5.1 The model setting in our first example is similar to that in Lian (2013) except for the error term. For $1 \leqslant j \leqslant J = 5$, the predictor function $X_{ij}(t) = \sum_{k=1}^{50} \xi_{ijk}\phi_k(t)$ with $\xi_{ijk} \sim N(0, k^{-2})$ are independent predictors. The basis functions are Fourier basis functions consisting of $\phi_1 \equiv 1$, and $\phi_k(t) = \sqrt{2}\cos(k\pi t)$ for $k > 1$. Thus, the true model is given by

$$y_i = -2 + \int_0^1 \beta_1(t)X_{i1}(t)dt + \int_0^1 \beta_3(t)X_{i3}(t)dt + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{5.11}$$

For the coefficient functions $\beta_1(t)$ and $\beta_3(t)$, we set $\boldsymbol{\beta}_1 = (-2, 1, 2, -1, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_3 = (1, -1, 0.5, -0.5, 1, 0, \ldots, 0)^T$, and $\beta_j(t) = \sum_{k=1}^{50} \beta_{jk}\phi_k(t)$, $j = 1, 3$. We take $\beta_2(t) = \beta_4(t) = \beta_5(t) = 0$ to set up the model where the response variable is only

depend on the $X_1(t)$ and $X_3(t)$. The following distributions of ε_i are considered: (1) the standard normal distribution $N(0,1)$; (2) t-distribution with degree freedom 3; (3) the standard Cauchy distribution.

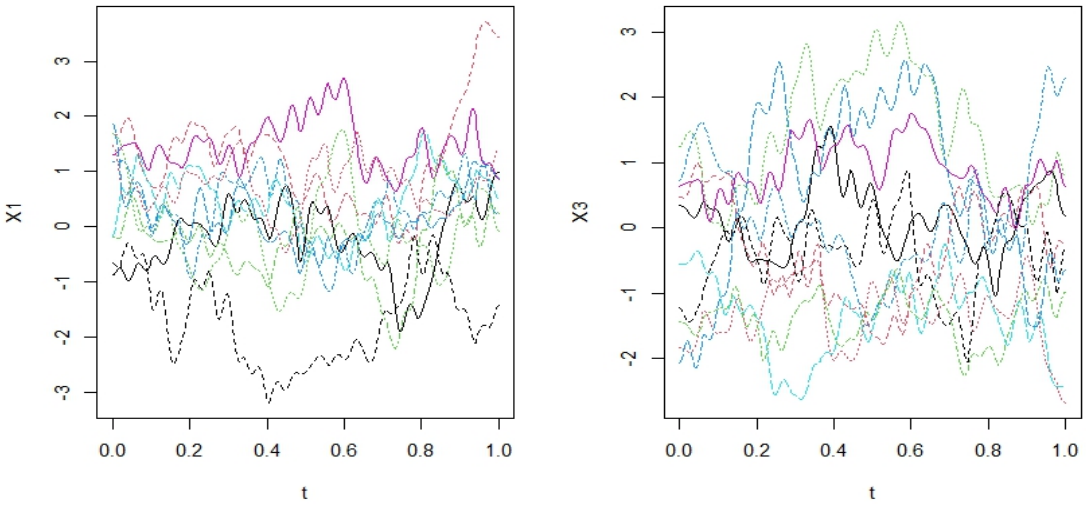

Figure 5.1: Some trajectories of covariates function $X(t)$ in Example 1.

Example 5.2 The model setting in our second example is similar to that in Matsui and Konishi (2011), but we contaminate data in order to examine the robustness of our proposed method. We set $J = 5$ and generated functional predictors X_{ij} as

follows:

$$X_{i1} = \cos(2\pi(t - a_1)) + a_2 t, \quad a_1 \sim N(0,1), a_2 \sim N(1, 2^2), \quad \mathcal{T} = [0,1].$$

$$X_{i2} = \sin(2\pi(t + a_1)) + a_2 t, \quad a_1 \sim N(0,1), a_2 \sim N(1, 2^2), \quad \mathcal{T} = [0,1].$$

$$X_{i3} = b_1 \cos(2t) + b_2 t, \quad b_1 \sim N(1, 2^2), b_2 \sim N(0,1), \quad \mathcal{T} = [0,1].$$

$$X_{i4} = b_1 \sin(2t) + b_2 t, \quad b_1 \sim N(1, 2^2), b_2 \sim N(0,1), \quad \mathcal{T} = [0,1].$$

$$X_{i5} = c_1 t^3 + c_2 t^2 + c_3 t + c_4, \quad c_1 \sim N(1, 2^2),\ c_2 \sim N(-1, 2^2),\ c_3 \sim N(2, 2^2),$$
$$c_4 \sim N(2,1),\ \mathcal{T} = [0,1].$$

Our true model is

$$y_i = \int_0^1 \beta_1(t) X_{i1}(t) dt + \int_0^1 \beta_3(t) X_{i3}(t) dt + \int_0^1 \beta_5(t) X_{i5}(t) dt + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where the functional coefficient $\beta_j(t)$ are given by

$$\beta_1(t) = \sin 2\pi t, \beta_3(t) = \sin \pi t, \beta_5(t) = 2t, \beta_2(t) = \beta_4(t) = 0.$$

To demonstrate the efficacy of our proposed methodology, we intentionally introduce contamination into the dataset in the following manner.

- ✦ Scenario I: Clean functional data. The error ε_i are distributed from the standard normal distribution $N(0,1)$.

- ✦ Scenario II: Contaminated scalar response y. In order to create outliers in response y, the following distributions of ε_i are considered: (1) mixed normal

distribution $0.7N(0,1)+0.3N(1,3^2)$; (2) $t(3)$ distribution; (3) standard Cauchy distribution Cauchy$(0,1)$.

- ✦ Scenario III: Contaminated functional predictors $X(t)$. This process is carried out as described by Fraiman and Muniz (2001). In this scenario, ε_i are distributed from $N(0,1)$.

  - Setting 1: Asymmetric contamination in which $X_j^*(t) = X_j(t) + \nu M$, where ν is a random variables sampled from Bernoulli (q) with different contamination fractions $q = 0.05, 0.1, 0.2$. M is the size of the contamination (a constant, say, $M = 15$).

  - Setting 2: Symmetric contamination in which $X_j^*(t) = X_j(t) + \nu\tau M$, where ν and M are same defined as in Setting 1. τ is a random variables independent of ν and taken values 1 and -1 with probability 0.5.

To demonstrate the robustness of our proposal, we compare the model selected by using our proposed robust method with those obtained by applying alternative group variable selection methods. Note that the term "Correct %" means that the percentage of times the correct model is selected, "mFDR %" denotes the percentage of the empirical marginal false discovery rate (mFDR), and the estimate of mFDR

is defined as the empirical mFDR is defined as

$$\text{mFDR} = \frac{\text{FP}}{\text{FP} + \text{TP} + \eta}$$

where FP is average number of false discoveries, TP is the average number of true discoveries, and $\eta = 10$ is chosen following Lin et al. (2011) and Dupuis and Victoria-Feser (2013). Given the focus of our project on robust methodology, we exclusively compare the advantages of robust group procedures over non-robust group procedures. Consequently, the robust group VIF (rgVIF) method is compared with the group VIF (gVIF), group SCAD (gSCAD), and group MCP (gMCP) methods For gSCAD and gMCP, ten-fold cross-validation is employed to select their penalty parameters. For each setting, the simulation results are summarized in Table 5.1-5.2 based on the analysis of 1000 replicates with sample size $n = 500, 700, 1000$, respectively.

Table 5.1: Simulation results based on the robust group VIF (rgVIF), group VIF(gVIF), group SCAD (gSCAD), and group MCP(gMCP) with different distributions in Example 5.1.

	rgVIF			gVIF			gSCAD			gMCP		
Normal(0,1)												
n	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
500	**93.5**	0.4862	0.4228	92.0	0.4232	0.5547	71.0	6.6947	2.2197	68.5	4.4205	2.2955
700	**94.5**	0.9651	0.6558	84.5	0.8920	0.9048	69.0	5.4597	2.3315	78.0	3.1946	1.9834
1000	**96.0**	0.4881	1.1992	86.0	0.7444	1.5995	67.0	6.6947	2.2574	75.5	2.9126	2.2469
$t(3)$												
n	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
500	**93.3**	0.8281	0.3974	89.5	0.8674	0.4655	76.5	5.3030	3.2058	80.5	2.0568	3.3455
700	**95.5**	0.7250	0.7582	90.0	0.5223	0.8759	70.0	6.2793	2.3017	82.0	2.3438	2.3519
1000	**94.6**	0.5685	1.2365	88.0	0.6458	1.6327	68.5	6.4109	2.5579	71.0	3.7304	2.4828
Cauchy(0,1)												
n	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
500	**80.1**	0.7543	0.1043	31.0	0.3850	0.4540	12.0	18.1294	3.3582	25.5	4.9703	0.3547
700	**87.4**	0.5149	0.2205	30.0	0.4876	0.7192	12.0	17.6872	3.9788	27.0	4.9424	4.2259
1000	**86.5**	0.5838	0.4430	31.0	0.4888	1.2205	10.5	16.2401	3.4241	21.0	4.6466	3.2857

Table 5.2: Simulation results based on the robust group VIF (rgVIF), group VIF(gVIF), group SCAD (gSCAD), and group MCP(gMCP) with different distributions in Example 5.2.

	rgVIF			gVIF			gSCAD			gMCP		
Scenario I												
n	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
500	**95.0**	0.3356	0.4652	96.0	0.3964	0.3934	76.0	2.7102	0.4506	81.2	1.9471	0.2625
700	**98.3**	0.3021	0.4485	95.0	0.4943	0.5196	79.0	2.6087	0.4523	81.0	1.9127	0.3326
1000	**98.5**	0.4023	0.9721	93.5	0.6417	0.7352	83.0	2.1321	0.3408	83.3	1.8095	0.4474
Scenario II												
n=1000	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
mixed normal	**98.6**	0.1968	0.3521	96.1	0.2778	0.5625	72.3	3.0491	0.6360	72.8	3.2568	0.8965
$t(3)$	**98.7**	0.1071	0.3516	93.3	0.3438	0.4703	74.3	2.8041	0.6453	81.0	1.5166	0.6182
Cauchy (0,1)	**92.0**	0.5436	0.3582	55.9	0.1528	0.9067	19.7	9.6296	0.7149	37.8	3.3210	0.7094
Scenario III												
n=1000	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)	Correct %	mFDR %	time(s)
Setting 1: $q = 0.05$	**85.6**	0.7563	0.3258	45.6	0.4561	0..9282	23.5	0.5632	2.1365	35.6	2.9563	1.2231
Setting 1: $q = 0.10$	**82.3**	0.8202	0.5681	51.3	0.5162	1.0362	25.6	0.6545	1.9856	38.7	3.3221	0.9856
Setting 1: $q = 0.20$	**80.5**	0.8823	0.3452	55.6	0.5563	0.8821	30.3	0.7526	2.1546	36.8	3.3321	1.0982
Setting 2: $q = 0.05$	**82.3**	0.7856	0.5623	46.5	0.4567	0.7823	35.6	2.1203	3.2154	30.5	3.5687	1.3256
Setting 2: $q = 0.10$	**79.5**	0.8325	0.3589	42.3	0.5456	0.8235	32.6	2.3516	2.985	32.5	3.2216	1.5623
Setting 2: $q = 0.20$	**80.5**	0.9546	0.3689	39.6	0.5236	0.4526	30.2	2.0545	2.7786	33.2	3.1456	1.2156

The simulation results presented in Table 5.1-5.2 clearly demonstrate the superior performance of the "rgVIF" method in terms of both "Correct%" and "mFDR%". These findings strongly suggest that robust group VIF exhibits enhanced efficiency

and stability when handling contaminated data or datasets containing outliers. For example, in Table 5.1, the values of the term "Correct%" obtained using our robust group selection method ("rgVIF") are higher compared to other non-robust methods when applied to models with heavy-tailed distributed errors (such as $t(3)$ distribution or the standard Cauchy distribution). Similarly, results presented in Table 5.2 demonstrate the superiority of the robust procedure.

5.3.2 A Real Data Example

In this section, we present the proposed robust VIF variable selection method for multiple functional linear regression by analyzing weather data collected from 79 stations in Japan, which is available in Chronological Scientific Table 2005. The dataset comprises monthly and annual total observations averaged from 1971 to 2000, including: (1) monthly average temperatures (TEMP); (2) monthly maximum temperature (MAX.TEMP); (3) monthly minimum temperature (MIN.TEMP); (4) atmospheric pressure (PRESSURE); (5) daylight duration (DAYLIGHT); (6) humidity levels (HUMIDITY); (7) annual total precipitation (PRECIPITATION). Given that the data has been collected over a span of 12 months, it can be regarded as functional data. The box plot in Figure 5.2 illustrates the distribution of the response variable "PRECIPITATION" and indicates the presence of outliers. Figures

5.3-5.5 depict the trajectories of the predictor functions, with each graph displaying 79 curves representing measurements from different predictor variables at 79 stations. The group of curves exhibits the presence of a few functional outliers, which are trajectories that deviate from the rest in all six predictor variables.

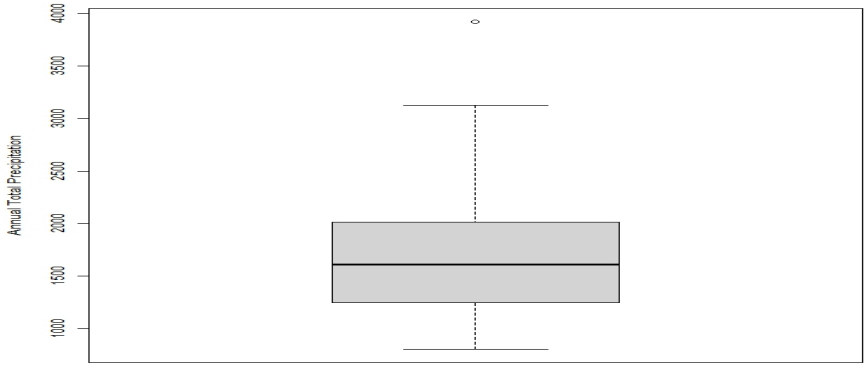

Figure 5.2: Annual total precipitation, outliers in response variable.

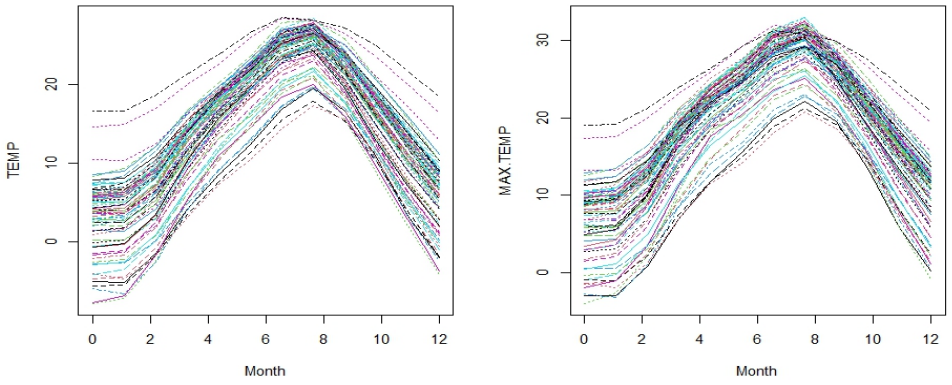

Figure 5.3: Some trajectories of covariates "Temp" and "MAX.TEMP".

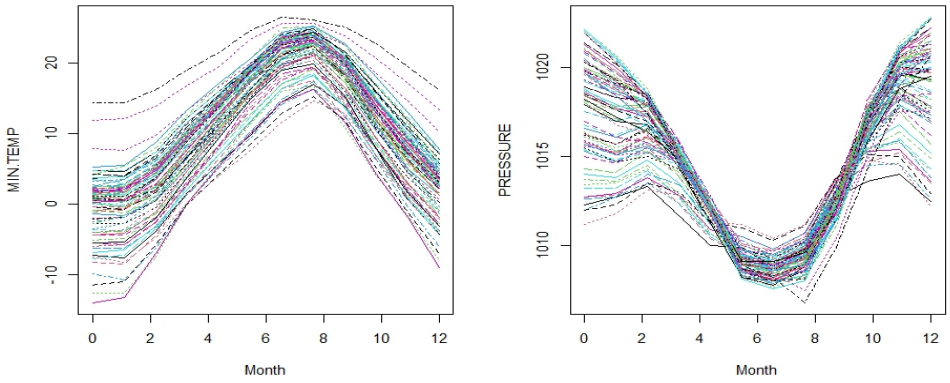

Figure 5.4: Some trajectories of covariates “MIN.TEMP” and “ PRESSURE”.

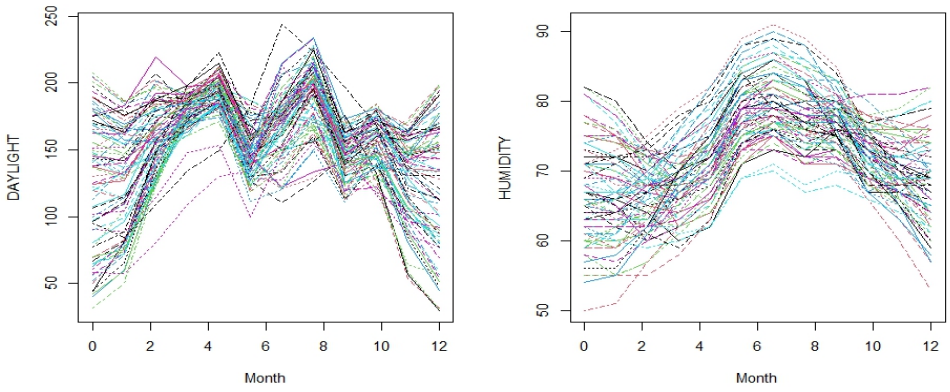

Figure 5.5: Some trajectories of covariates “DAYLIGHT” and “HUMIDITY”.

Our objective is to identify the variables that demonstrate a significant correlation with annual total precipitation (PRECIPITATION). To accomplish this, we employ a group variable selection procedure for this functional dataset. Based on the

outcomes presented in Table 5.3, the robust VIF selection method identifies more appropriate covariates. Consequently, the predictors such as MAX.TEMP PRESSURE, DAYLIGHT, and HUMIDITY are selected while those variables like TEMP and MIN.TEMP are excluded from the model. The result indicates no statistically significant association between TEMP/MIN.TEMP and PRECIPITATION.

Table 5.3: Functional variable selection results for the weather data. For each entry, the vector displays the selected groups via different methods.

Methods	rgVIF	gVIF	gLASSO	gSCAD	gMCP
group selected	(2,4,5,6)	(1,2,4)	(1,2,3,4,5,6)	(2,4,5,6)	(2,4,5)

5.4 Conclusions

We propose a robust method for variable selection in multiple functional linear regression models, particularly when outliers are present. Our approach is primarily based on the group VIF (Variance Inflation Factor) technique, which was previously introduced in Ding et al. (2023). This method utilizes a forward stagewise procedure and offers computational efficiency due to its simple iterations. We extend this procedure to handle robust estimation within the context of functional data analysis. An effective robust algorithm is proposed for functional variable selection, which ensures

predictive accuracy and stability. Through simulation studies and real data applications, we demonstrate the superior performance of our robust procedure compared to other group variable selection methods in handling contaminated data. However, we acknowledge that our current research does not consider situations where the number of functional predictors J diverges, which should be an attractive area for future investigation. We believe that this issue can be addressed through the utilization of shrinkage estimation and variable selection methods, which will be pursued as part of our future research agenda.

5.5 Appendix

Input data $\boldsymbol{y}$ (centered), $Z_1, Z_2, \ldots$;

Set $u_0 = 0.05$, $\triangle u = 0.05$, $f = 0$, $\mathcal{C} = \varnothing$, $j = 1$

Initialization $Z_{\mathcal{C}} = \mathbf{1}$, $Z_{\mathcal{C}}^w = \text{diag}(\sqrt{w_{i\mathcal{C}}^0})Z_{\mathcal{C}}$, $\boldsymbol{y}^w = \text{diag}(\sqrt{w_{i\mathcal{C}}^0})\boldsymbol{y}$, Where $w_{i\mathcal{C}}^0$ is computed using Huber's weight function in Remark 5.2, in which $\widehat{\sigma}_0 = 1.483\text{med}|\tilde{\boldsymbol{r}}_0 - \text{med}\,(\tilde{\boldsymbol{r}}_0)\,|$, $\tilde{\boldsymbol{r}}_0 = \boldsymbol{y} - \bar{\boldsymbol{y}}$.

Repeat{ **Set** $\alpha_j = u_j/(1 + j - f)$

Obtain test statistic $\widehat{T}_w$ by (5.10)

If $\widehat{T}_w > \chi^2_{p_j}(\alpha_j)$

Then $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{j\}$,$u_{j+1} \leftarrow u_j + \triangle u$, $f = j$,

Update $Z_{\mathcal{C}} = \left[Z_{\mathcal{C}}, Z_j\right]$,$Z_{\mathcal{C}}^w = \text{diag}(\sqrt{w_{i\mathcal{C}}})Z_{\mathcal{C}}$, $\boldsymbol{y}^w = \text{diag}(\sqrt{w_{i\mathcal{C}}})\boldsymbol{y}$

and $w_{i\mathcal{C}}$ is computed with $\boldsymbol{r} = (\boldsymbol{y}^w - Z_{\mathcal{C}}^w\widehat{\boldsymbol{\beta}}^w)/\widehat{\sigma}$

using $\widehat{\boldsymbol{\beta}}^w = [\left(Z_{\mathcal{C}}^w\right)^\top Z_{\mathcal{C}}^w]^{-1}\left(Z_{\mathcal{C}}^w\right)^\top \boldsymbol{y}^w$, $\widehat{\sigma} = 1.483\text{med}|\tilde{\boldsymbol{r}} - \text{med}\,(\tilde{\boldsymbol{r}})\,|$, $\tilde{\boldsymbol{r}} = \boldsymbol{y}^w - Z_{\mathcal{C}}^w\widehat{\boldsymbol{\beta}}^w$

Else $u_{j+1} \leftarrow u_j - \alpha_j/(1 - \alpha_j)$.

End if

$j = j + 1$. }

Until all candidate group predictors have been considered, **Output** $\mathcal{C}$.

Algorithm 1: Robust Group VIF algorithm

6 Conclusions and Future Work

6.1 Conclusions and Remarks

The primary objective of this study is to address robust methodologies for functional linear models, which holds significant importance for several reasons. First, functional data is gaining popularity because it accurately captures the intrinsic functional properties of the various phenomena being studied. Second, traditional methods applicable to multivariate data do not always extend directly to the context of functional data analysis; therefore, it is necessary to consider specific solutions to address this issue. Finally, given that functional observations may be affected by unnatural variation, and given that inference is often performed on relatively small datasets, it becomes critical to couple inference procedures with robust methods. Within this context, we have developed some robust methods for functional linear regression, with a particular focus on scalar-on-function models.

In Chapter 2, we explore the popular basis expansion methods in functional

data analysis and demonstrate their effectiveness through simulation examples for estimating coefficient functions.

In Chapter 3, we investigate the robust estimation problem in partial functional linear models under the framework of Reproducing Kernel Hilbert Spaces (RKHS). We derive the theoretical properties of robust estimation and conduct simulation studies using real data examples.

In Chapter 4, we address the issue of robust hypothesis testing in functional linear regression by extending three robust tests: Wald-type, likelihood ratio-type, and F-type tests within the framework of functional linear models.

In Chapter 5, we develop the robust functional variable selection procedure for multiple functional linear regression by proposing a robust version of the group VIF method and applying it to functional data analysis.

The above research demonstrate the significance and potential applications of robust methods in functional data analysis through simulations and examples. These findings will lay a solid foundation for our future work.

6.2 Future Work

Functional data analysis is a fascinating and constantly evolving area in statistics. In this dissertation, we introduce a series of robust functional linear regression meth-

ods. It is important to note that these methods are not limited to the applications discussed in the previous chapters; however, our main focus is on scalar-on-function linear models. The techniques outlined in Chapters 3 to 5 may also be extended to function-on-function or function-on-scalar linear models. Furthermore, these resilient methods may also be applied to generalized functional regression, whose theoretical properties pose potential challenges for future research. Last but not least, we acknowledge that in practical scenarios, complete functional data are usually unattainable. This may involve partially observed covariates or missing responses. While our method and algorithm may not be applicable in this situation, we believe it necessitates further investigation and the development of relevant asymptotic theory.

Bibliography

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.

Benedetto, J. J. (1993). *Wavelets: mathematics and applications*, volume 13. CRC press.

Boente, G., Salibian-Barrera, M., and Vena, P. (2020). Robust estimation for semi-functional linear regression models. *Computational Statistics & Data Analysis*, 152:107041.

Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216.

Cai, X., Xue, L., and Lu, F. (2020). Robust estimation with a modified huber's loss for partial functional linear models based on splines. *Journal of the Korean Statistical Society*, 49(4):1214–1237.

Candanedo, L. M., Feldheim, V., and Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, 140:81–97.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics*, 30(1):241–255.

Cox, D. D. (1983). Asymptotics for m-type smoothing splines. *The Annals of Statistics*, 11(2):530–551.

Cui, X., Lin, H., and Lian, H. (2020). Partially functional linear regression in reproducing kernel hilbert spaces. *Computational Statistics & Data Analysis*, 150:106978.

D. B. Clarkson, C. Fraley, C. G. and Ramsay, J. (2005). *S+Functional data analysis,User's Manual for Windows*. Springer.

Ding, H., Zhang, Y., and Wu, Y. (2023). A novel group vif regression for group variable selection with application to multiple change-point detection. *Journal of Applied Statistics*, 50(2):247–263.

Dupuis, D. J. and Victoria-Feser, M.-P. (2013). Robust vif regression with application to variable selection in large data sets. *The Annals of Applied Statistics*, 7(1):319–341.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99(466):303–313.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2):429–444.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

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

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469.

Hall, P., Horowitz, J. L., et al. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Hastie, T. and Mallows, C. (1993). [a statistical view of some chemometrics regression tools]: Discussion. *Technometrics*, 35(2):140–143.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons.

Huang, L., Zhao, J., Wang, H., and Wang, S. (2016). Robust shrinkage estimation and selection for functional multiple linear model through lad loss. *Computational Statistics & Data Analysis*, 103:384–400.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Statistics*, 35(1):73–101.

Huber, P. J. et al. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Huber, P. J. and Ronchetti, E. M. (2011). *Robust statistics*. John Wiley & Sons.

Kato, K. et al. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136.

Kong, D., Staicu, A.-M., and Maity, A. (2016a). Classical testing in functional linear models. *Journal of nonparametric statistics*, 28(4):813–838.

Kong, D., Xue, K., Yao, F., and Zhang, H. H. (2016b). Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159.

Kong, X., Deng, H., Yan, F., Kim, J., Swisher, J. A., Smit, B., Yaghi, O. M., and Reimer, J. A. (2013). Mapping of functional groups in metal-organic frameworks. *Science*, 341(6148):882–885.

Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statistica Sinica*, 23(1):51–74.

Lin, D., Foster, D. P., and Ungar, L. H. (2011). Vif regression: a fast regression algorithm for large data. *Journal of the American Statistical Association*, 106(493):232–247.

Luo, R. and Qi, X. (2015). Sparse wavelet regression with multiple predictive curves. *Journal of Multivariate Analysis*, 134:33–49.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Mas, A. and Pumo, B. (2009). Functional linear regression with derivatives. *Journal of Nonparametric Statistics*, 21(1):19–40.

Matsui, H. and Konishi, S. (2011). Variable selection for functional regression models via the l1 regularization. *Computational Statistics & Data Analysis*, 55(12):3304–3310.

Morettin, P. A., Pinheiro, A., and Vidakovic, B. (2017). *Wavelets in functional data analysis*. Springer.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.

Muirhead, R. J. (2005). *Aspects of multivariate statistical theory*. John Wiley & Sons.

Müller, H.-g. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240.

Nason, G. (2008). *Wavelet methods in statistics with R*. Springer Science & Business Media.

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35(1):99–105.

Pannu, J. and Billor, N. (2017). Robust group-lasso for functional regression model. *Communications in statistics-simulation and computation*, 46(5):3356–3374.

Pannu, J. and Billor, N. (2022). Robust sparse functional regression model. *Communications in Statistics-Simulation and Computation*, 51(9):4883–4903.

Qingguo, T. (2017). M-estimation for functional linear regression. *Communications in Statistics-Theory and Methods*, 46(8):3782–3800.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, 64:183–210.

Rügamer, D., Brockhaus, S., Gentsch, K., Scherer, K., and Greven, S. (2018). Boosting factor-specific functional historical models for the detection of synchronization in bioelectrical signals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 67(3):621–642.

Schrader, R. M. and Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, 67(1):93–101.

Shin, H. (2009). Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139(10):3405–3418.

Shin, H. and Lee, S. (2016). An rkhs approach to robust functional linear regression. *Statistica Sinica*, 26(1):255–272.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. Society for Industrial and Applied Mathematics.

Su, Y.-R., Di, C.-Z., and Hsu, L. (2017). Hypothesis testing in functional linear models. *Biometrics*, 73(2):551–561.

Sun, X., Du, P., Wang, X., and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611.

Swihart, B. J., Goldsmith, J., and Crainiceanu, C. M. (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, 56(4):483–493.

Tibshirani, R. J. (2015). A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.*, 16(1):2543–2588.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Wang, L. (2008). *Karhunen-Loeve expansions and their applications.* PhD thesis, London School of Economics and Political Science (United Kingdom).

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.

Yohai, V. J. (1974). Robust estimation in the linear model. *The Annals of Statistics*, pages 562–567.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, 7(2):258–268.

Yu, P., Zhu, Z., Shi, J., and Ai, X. (2020). Robust estimation for partial functional linear regression model based on modal regression. *Journal of Systems Science and Complexity*, 33(2):527–544.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based lasso in functional linear regression. *Journal of computational and graphical statistics*, 21(3):600–617.

Zhou, J., Du, J., and Sun, Z. (2016). M-estimation for partially functional linear regression model based on splines. *Communications in Statistics-Theory and Methods*, 45(21):6436–6446.