New Methods and Models in Functional Data Analysis

by

Peijun Sang

B.Sc., Zhejiang University, 2010
M.Sc., University of British Columbia, 2014

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

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2018

Approval

Name:	**Peijun Sang**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	**New Methods and Models in Functional Data Analysis**
Examining Committee:	**Chair:** Jinko Graham Professor
	Jiguo Cao Senior Supervisor Associate Professor
	Liangliang Wang Supervisor Assistant Professor
	Richard Lockhart Internal Examiner Professor Department of Statistics and Actuarial Science
	Nancy Heckman External Examiner Professor Department of Statistics University of British Columbia
Date Defended:	**July 23, 2018**

Abstract

Functional data analysis (FDA) plays an important role in analyzing function-valued data such as growth curves, medical images and electromagnetic spectrum profiles, etc. Since dimension reduction can be achieved for infinite-dimensional functional data via functional principal component analysis (FPCA), this technique has attracted substantial attention. We develop an easy-to-implement algorithm to perform FPCA and find that this algorithm compares favorably with traditional methods in numerous applications. Knowing how random functions interact is critical to studying mechanisms like gene regulations and event-related brain activation. A new approach is proposed to calibrate dynamical correlations of random functions and we apply this approach to quantify functional connectivity from medical images. Scalar-on-function regression, which is widely used to characterize the relationship between a functional covariate and a scalar response, is an important ingredient of FDA. We propose several new scalar-on-function regression models and investigate their properties from both theoretical and practical perspectives.

Keywords: functional data analysis; functional principal component analysis; dynamical correlation; sparse; semiprametric additive models; quantile regression

Dedication

To my wife, Siqi Hong, whose love and encouragement give me the best support. To my dear parents, thanks for your unconditional love.

Acknowledgements

I would like to thank my advisor Dr. Jiguo Cao for bringing me to the field of functional data analysis, and for your patience and encouragement throughout my Ph.D studies. I would have never got chances to discover so many interesting statistical problems without insightful discussions with you.

Thanks to the Department of Statistics and Actuarial Science at Simon Fraser University. I appreciate the wonderful environment you have created for us. I could never imagine a better place where I could get more help and caring.

I would like to give a special thanks to Dr. Richard Lockhart and Dr. Tom Loughin. For Richard, I would like to appreciate your kind help and guidance for my research. I would also like to thank you for making me aware of the importance of being a rigorous statistician. Last but least, without your support, I would never get a chance to go to the distinguished institution for my academic career. For Tom, you are always there sharing me with your experiences when I turn to you for help.

I would also like to thank all members in my thesis examining committee for taking your time to read my thesis. Your invaluable comments and questions have helped me a lot in trying to improve my thesis.

There are still numerous names that I need to mention. I would like to give my thanks to everyone who has helped me, no matter in academia or other aspects in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis mainly focuses on developing new methodologies in functional data analysis (FDA) for applications in environmental science, genetics, and neuroscience. Functional data analysis refers to studying curves or functions with multiple measurements over a continuum like space or time. With the development of modern computing technology, functional data are frequently encountered in the above fields. Functional principal component analysis (FPCA) and regression models with functional covariates and a scalar response are two main problems of interest in FDA. Both of them are covered in the thesis. In addition, we investigate methods to calibrate dynamical correlations between random functions; this provides another perspective to look at gene regulation networks and functional connectivity of brain signals at different locations.

A critical problem in FDA is to analyze functional data by dimension reduction since functional data are intrinsically infinite dimensional. In this regard, there has been extensive research on FPCA, which is an extension of the principal component analysis in multivariate data analysis. More specifically, FPCA explores major sources of variability in a sample of random curves by finding functional principal components (FPCs) that maximize curve variation. Consequently, the top few FPCs play a dominant role in capturing the variability in the random curves. Classical FPCA employs flexible functions such as B-spline or Fourier basis functions to represent top FPCs. These flexible representations usually yield FPCs without analytic formulae; hence users often find it onerous to understand and interpret FPCs. To address this issue, we developed the parametric FPCA method to represent FPCs with simple functions such as polynomials in Chapter 2. Numerical studies demonstrate that the proposed method is competitive in terms of explaining similar variations of functional data in comparison with the classical FPCA method.

[25] proposed the concept of dynamical correlation, which extends the basic idea of the correlation coefficient to functional data and reflects the dynamical feature of associations between random functions. A new approach is proposed in Chapter 3 to implement statistical inference for the dynamical correlation. Compared with the bootstrap approach suggested by [25], our approach is considerably more efficient in computation. Furthermore,

simulation studies demonstrate that the confidence interval of the dynamical correlation obtained from our new approach has a more accurate coverage probability than its counterpart from the bootstrap approach for random functions with irregular observations. This approach is applied to real examples such as gene regulations, functional connectivity in brain imaging data and air pollutants.

Functional regression models are widely used to predict a scalar response given functional covariates. In Chapter 4, we propose a functional semiparametric additive model for the effects of a functional covariate and several scalar covariates on a scalar response. The effect of the functional covariate is modeled nonparametrically, while a linear form is adopted to model the effects of the scalar covariates. This strategy can enhance flexibility in modeling the effect of the functional covariate and maintain interpretability for the effects of scalar covariates simultaneously. We develop the method for estimating the functional semiparametric additive model by smoothing and selecting non-vanishing components for the functional covariate. Asymptotic properties of our method are also established. Two simulation studies are implemented to compare our method with various conventional methods. We demonstrate our method with two real applications.

Due to simplicity and easy interpretation, the functional linear model (FLM) is extensively studied in literature and frequently applied in practice. However, this specific linear form may lead to an inadequate fit of the relationship between a functional covariate and a scalar response. To address this issue, we propose a sparse functional additive model in Chapter 5. The effect of the functional predictor is represented in a nonparametric additive form, where the arguments are the scaled functional principal component scores. Component selection and smoothing are considered when fitting the model to reduce the variability and enhance the prediction accuracy, while providing an adequate fit. The convergence rate of the proposed estimator is established. Simulation studies show that the proposed estimation method compares favourably with various conventional methods in terms of prediction accuracy and component selection. The advantage of our proposed model and the estimation method is further demonstrated in two real data examples.

In contrast to the models that focus on the conditional mean structure of a scalar response in Chapter 4 and Chapter 5, a regression model characterizing the conditional quantiles is discussed in Chapter 6. This new model connects the conditional quantile of a response with a linear functional of a functional covariate using a nonparametric adaptive link function. We propose a generalized profiling approach to estimate both the unknown linear functional and the link function in the model. Simulation studies show that, compared with approaches modelling the conditional mean, the proposed model is more robust to extremely large outcomes in the response. In addition, when applying them to predict the maximal intraday PM_{10} concentrations, we find that the quantile regression model outperforms its counterpart in prediction accuracy.

Chapter 2

Parametric Functional Principal Component Analysis

2.1 Introduction

Functional data analysis has received considerable attention in diverse areas of applications where the data are random curves ([85]; [30]). One major tool in functional data analysis is functional principal component analysis (FPCA). FPCA explores major sources of variability in a sample of random curves by finding functional principal components (FPCs) that maximize curve variation. Consequently, the top few FPCs explain most of the variability in the random curves. Each curve can be approximated by a linear combination of the top few FPCs. In other words, the infinite dimensional curves are projected to a low-dimensional space defined by the top FPCs. This powerful dimension reduction feature also promotes the popularity of FPCA.

Asymptotic properties of FPCA have been studied and discussed at length. For example, [20] built asymptotic theories for the principal component analysis of a vector of random functions by treating the covariance as a linear operator from a separable Hilbert space. Functional analysis is also an excellent tool to boost theoretical developments of FPCA, including the work by [9], [67] and [11], among many others.

FPCA has not only gained considerable breakthroughs in theoretical developments, but also achieved great success in applications. [116] applied FPCA in functional linear regression models for longitudinal data; [87] employed FPCA to analyze paired functional data with complex variations within and across individuals. A general assumption in FPCA is that the observed data are dense and regularly spaced. For sparse and irregularly spaced data, [115] proposed estimating the principal component scores via conditional expectation (in short PACE), which recovers the individual trajectories by exploiting information from all curves.

As suggested by [86], FPCA is more appealing if we control the roughness of FPCs to achieve some degree of smoothness. Three methods have been proposed as far as we know.

The first method smooths functional data in the first step, and then conducts the regular FPCA procedure on the smoothed functional data ([84]; [10]). The second method adds a roughness penalty term on the FPCs in the optimization criterion of FPCA ([81]; [96]). The third method first smooths the variance-covariance function of functional data, and then conducts eigenanalysis of the smoothed variance-covariance function ([115]).

FPCs explain the major variation of the curves and project the infinite-dimensional curves to low-dimensional spaces; therefore it is important to interpret FPCs accurately. Conventional methods provide flexible estimates of FPCs without analytic formulae; hence users often find it onerous to understand and interpret FPCs. This bottleneck of FPCA has gained recent attention. [62] proposed a penalty-based method to derive smooth FPCs that are nonzero only in intervals where curves display major variation; while [15] considered a localized version of FPCA to achieve both interpretability and functionality.

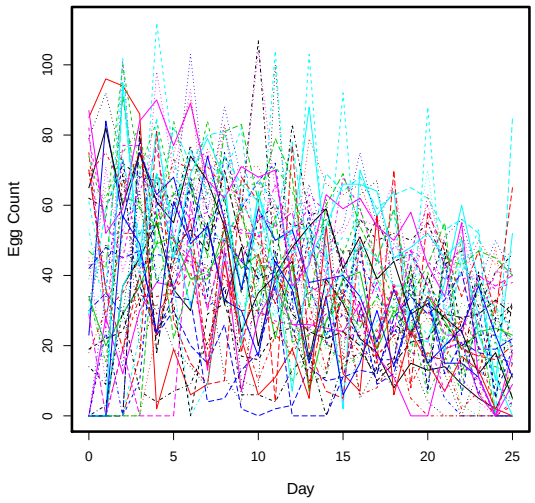


Figure 2.1: The number of eggs laid by 50 medflies in 25 days.

In the line of interpretable FPCA, our work is motivated by observing the following fact: although functional data may be complicated, the top few FPCs often display simple trends in most applications we know of. For instance, medfly data have been discussed and analyzed in substantial literature (e.g., [90]; [70]). This dataset consists of records of number of eggs laid by Mediterranean fruit flies in 25 days. Figure 2.1 displays the number of eggs laid by 50 flies across 25 days. FPCA can be employed to explore the major variation of 50 curves. Figure 2.2 shows the top three FPCs obtained from the smooth FPCA method ([86]). The top three FPCs explain about 97.8% of total variations among 50 curves. Even though the top three FPCs display simple trends, it is still challenging for users to understand and interpret them because they are given numerically without parametric forms.

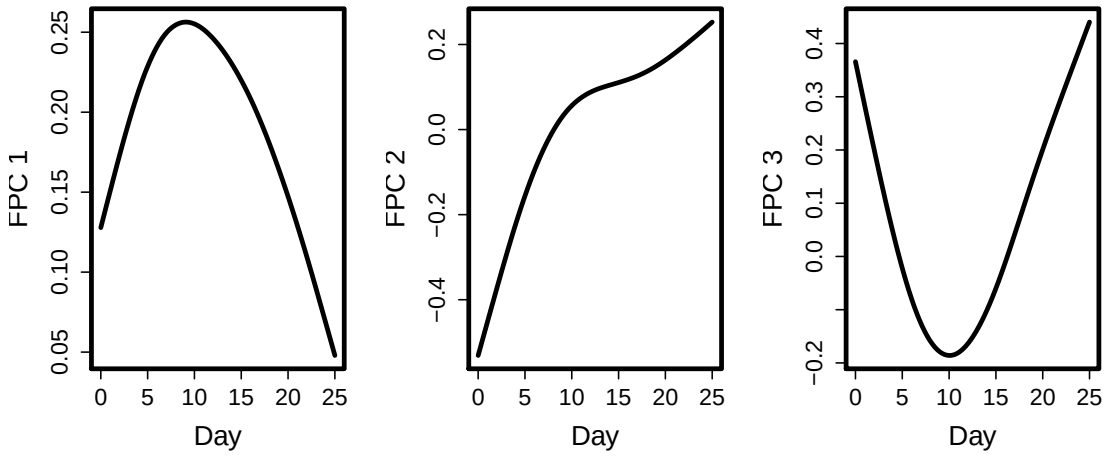


Figure 2.2: The top three FPCs obtained from the nonparametric FPCA method for the medfly data.

The above phenomenon is also found in many other applications of FPCA such as additional applications introduced in Section 2.3. Therefore, we propose to use simple parametric functions to approximate the top FPCs, which possess both simple shapes and easy interpretations for users. We call this method parametric FPCA. Although many simple parametric functions can be used, we find that for most practical purposes polynomial functions suffice for approximating the top FPCs.

In comparison with the conventional nonparametric FPCA method, our parametric FPCA method has three major advantages. The first advantage is that the estimated FPCs have a closed-form expression, which helps to understand and interpret the FPCs. The second advantage is that the parametric FPCA method is more robust to outlier curves than the conventional FPCA method, which will be justified in our simulation studies in Section 2.4. The third advantage is that the FPCs estimated by the parametric FPCA method are always smooth, so it is unnecessary to add a roughness penalty on the FPCs. Thus, parametric FPCA allows us to circumvent the smoothing parameter selection procedure when estimating the FPCs.

The rest of this chapter is organized as follows. In Section 2.2, after presenting a brief review of the conventional nonparametric FPCA method, we propose the parametric FPCA method for analyzing regularly spaced and dense functional data and then for analyzing irregularly spaced and sparsely observed functional data. In Section 2.3, three real applications are presented to demonstrate the utility of parametric FPCA. In Section 2.4, a simulation study is conducted to justify the robustness of parametric FPCA. Section 2.5 concludes the chapter.

2.2 Parametric FPCA

In this section, we first review the conventional nonparametric functional principal component analysis (FPCA), with emphasis on regularized FPCA. Then we propose the parametric FPCA method and illustrate how to carry out this method for densely observed and regularly spaced functional data. Next we provide procedures to implement parametric FPCA to analyze sparsely observed and irregularly spaced functional data. In the end we propose an approach to choose the degree of polynomials when implementing parametric FPCA.

2.2.1 Nonparametric FPCA

In classical functional data analysis, FPCA is widely used to explain the major variations in curves. Suppose we have a square integrable stochastic process $X(t)$, $t \in \mathcal{I}$, with mean $\mathrm{E}(X(t)) = \mu(t)$ and covariance function $\mathrm{Cov}(X(s), X(t)) = G(s,t)$. Mercer's Theorem states that $G(s,t)$ has an orthogonal expansion in $L^2(\mathcal{I})$:

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t), \tag{2.1}$$

where $\psi_k(t)$ and λ_k are eigenfunctions and eigenvalues of the covariance function with the order $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The eigenfunctions $\{\psi_k(t)\}_{k=1}^{\infty}$ satisfy

$$\int \psi_k^2(t)\,\mathrm{d}t = 1, \text{ and } \int \psi_j(t)\psi_k(t)\,\mathrm{d}t = 0 \text{ for any } j \neq k. \tag{2.2}$$

Let $x_i(t), i = 1, \ldots, n$ be the sample curves of the stochastic process $X(t)$. The Karhunen-Loève expansion of $x_i(t)$ is :

$$x_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \psi_k(t),$$

where $\{\psi_k(t)\}_{k=1}^{\infty}$ are called the functional principal components (FPCs), and $\{\xi_{ik}\}_{k=1}^{\infty}$ are the corresponding FPC scores. All sample curves $x_i(t)$ can be well approximated by the first K FPCs, provided that these K FPCs explain most of the variability in the sample curves. In practice, we choose the value of K such that the first K FPCs can explain at least 85% of total variability in the sample curves, as suggested by [59]. Since all sample curves can be projected to the finite K-dimensional space expanded by the first K FPCs, it is of great interest to estimate these FPCs.

Usually, FPCs are represented nonparametrically as a linear combination of some flexible basis functions such as B-spline or Fourier basis. The estimation procedure requires numerical integration and calculation of a high dimensional inverse matrix to achieve high precision for densely observed functional data. In addition, the nonparametric estimate may be excessively wiggly or locally variable due to a large number of basis functions employed

in the representation of the FPCs. [86] argued that it would be more appealing to accommodate smoothness when estimating FPCs. They proposed two different approaches to obtain smooth estimates for FPCs. One method directly smooths FPCs by penalizing their roughness; the second method first smooths the functional data and then computes the corresponding FPCs from the smoothed data. The latter approach is used in all applications and simulations in this chapter, where we call it as the nonparametric FPCA method. In the proposed parametric FPCA method, we also adopt the latter approach: smooth raw functional data and then apply parametric FPCA to the smoothed functional data.

2.2.2 Parametric FPCA for Dense Functional Data

Now we propose how to carry out parametric FPCA for densely observed and regularly spaced functional data. We find that the top few FPCs often display some simple trends in most applications we know of. In particular, the first FPC is often close to a constant over time, which reflects the mean level of the sample curves; the second FPC is close to a straight line over time, which usually crosses the x-axis and reflects a change of the sample curves between two time intervals; and the third FPC is close to a quadratic curve, crossing the x-axis twice and representing the change of the sample curves among three time intervals.

Motivated by the simple shapes of the top few FPCs in many situations, we propose to approximate the top K FPCs using a simple parametric form. Meanwhile, we hope that the parametric FPCs can still explain most variations of sample curves, which will be assessed by our three applications in Section 2.3. Any appropriate parametric form can be used to represent the top K FPCs. In this chapter, we assume that the top K FPCs are given in the following polynomial forms of degree p ($p \geq K-1$):

$$\psi_k(t) = b_{k0} + b_{k1}t + \cdots + b_{kp}t^p, \quad k = 1, \ldots, K, \tag{2.3}$$

where the coefficients b_{kj} are chosen to satisfy the constraints (2.2). The choice of the degree of polynomials p will be discussed in Section 2.2.4.

Based on (2.1) and (2.2), all FPCs $\psi_k(t)$, $k = 1, \ldots, K$, satisfy the following eigenequation

$$\int_{\mathcal{I}} G(s,t)\psi_k(s)\,\mathrm{d}s = \lambda_k \psi_k(t), \tag{2.4}$$

where $G(s,t)$ is the covariance function of $X(t)$, i.e., $G(s,t) = \mathrm{Cov}\left(X(s), X(t)\right)$. In the rest of this subsection, we show how to estimate the coefficients b_{kj} for the FPCs in (2.3) based on the above eigenequation.

The parametric FPCs in (2.3) can be expressed in a matrix notation $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{\phi}(t) = (1, t, \ldots, t^p)'$ and $\boldsymbol{b}_k = (b_{k0}, \ldots, b_{kp})'$. Plugging this expression into (2.4), it

follows

$$\int_{\mathcal{I}} G(s,t)\psi_k(s)\,\mathrm{d}s = \int_{\mathcal{I}} G(s,t)\boldsymbol{\phi}'(s)\boldsymbol{b}_k\,\mathrm{d}s = \left\{\int_{\mathcal{I}} G(s,t)\boldsymbol{\phi}(s)\,\mathrm{d}s\right\}'\boldsymbol{b}_k = \lambda_k\boldsymbol{\phi}'(t)\boldsymbol{b}_k.$$

The above eigenequation holds for all $t \in \mathcal{I}$. We choose $M+1$ equally spaced time points, $t_0 < t_1 < \cdots < t_M$ on $\mathcal{I}$, where t_0 and t_M are the two endpoints of $\mathcal{I}$, respectively. Then for any t_m, $m = 0, \ldots, M$, the eigenequation is

$$\left\{\int_{\mathcal{I}} G(s,t_m)\boldsymbol{\phi}(s)\,\mathrm{d}s\right\}'\boldsymbol{b}_k = \lambda_k\boldsymbol{\phi}'(t_m)\boldsymbol{b}_k.$$

Let $\boldsymbol{\Phi}$ be an $(M+1)\times(p+1)$ matrix with the (m,j)-th entry $\boldsymbol{\Phi}_{mj} = \phi_j(t_m)$, where $\phi_j(t) = t^{j-1}$, and $\boldsymbol{A}$ be an $(M+1)\times(p+1)$ matrix with the (m,j)-th entry

$$\boldsymbol{A}_{mj} = \int_{\mathcal{I}} G(s,t_m)\phi_j(s)\,\mathrm{d}s\,.$$

Then the eigenequation can be written in the matrix form

$$\boldsymbol{A}\boldsymbol{b}_k = \lambda_k\boldsymbol{\Phi}\boldsymbol{b}_k.$$

Therefore,

$$(\boldsymbol{\Phi}'\boldsymbol{A})\boldsymbol{b}_k = \lambda_k\boldsymbol{\Phi}'\boldsymbol{\Phi}\boldsymbol{b}_k. \tag{2.5}$$

Defining $\boldsymbol{c}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{b}_k$, Equation (2.5) can be expressed in terms of $\boldsymbol{c}_k$:

$$(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\boldsymbol{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k = \lambda_k\boldsymbol{c}_k.$$

Therefore, $\boldsymbol{c}_k$ is the eigenvector of the symmetric matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\boldsymbol{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}$ and λ_k is the eigenvalue of this matrix. Here the matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\boldsymbol{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}$ is symmetric since we approximate the integral for $\boldsymbol{A}_{mj}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$ be the K eigenvalues of this matrix, and $\boldsymbol{c}_k$ be the corresponding eigenvector of this matrix. Then $\boldsymbol{b}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$, and the k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$. When $k_1 \neq k_2$, the orthogonality of $\boldsymbol{c}_{k_1}$ and $\boldsymbol{c}_{k_2}$ implies the approximate orthogonality of FPCs $\psi_{k_1}(t)$ and $\psi_{k_2}(t)$, which can be verified as follows:

$$\begin{aligned}
\int_{\mathcal{I}} \psi_{k_1}(t)\psi_{k_2}(t)\,\mathrm{d}t &= \int_{\mathcal{I}} \boldsymbol{b}_{k_1}'\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)\boldsymbol{b}_{k_2}\,\mathrm{d}t \\
&= \int_{\mathcal{I}} \boldsymbol{c}_{k_1}'(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_{k_2}\,\mathrm{d}t \\
&\approx \boldsymbol{c}_{k_1}'(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\left\{\frac{L}{M}(\boldsymbol{\Phi}'\boldsymbol{\Phi})\right\}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_{k_2} \\
&= 0,
\end{aligned}$$

where L is the length of $\mathcal{I}$.

Algorithm 1 details our parametric FPCA method for dense functional data.

Algorithm 1 : **Parametric FPCA for dense functional data**

Step 1: Smooth the original functional data.

Step 2: Obtain the sample variance-covariance function for $G(t_\ell, t_m)$:

$$\hat{G}(t_\ell, t_m) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{x}_i(t_\ell) - \bar{x}(t_\ell)\}\{\hat{x}_i(t_m) - \bar{x}(t_m)\}, \tag{2.6}$$

where $\hat{x}_i(t)$ is the smooth estimate for each functional data by using the smooth spline method ([109]), and $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i(t)$.

Step 3: Employ the rectangle rule to approximate the entries in the (m, j)-th entry of the matrix $\boldsymbol{A}$

$$\boldsymbol{A}_{mj} = \int_{\mathcal{I}} G(s, t_m)\phi_j(s)\,\mathrm{d}s \approx \frac{L}{M}\sum_{\ell=0}^{M}\hat{G}(t_\ell, t_m)\phi_j(t_\ell)\,,$$

where L is the length of $\mathcal{I}$, and $\hat{G}(s,t)$ is the sample covariance function estimated by (2.6).

Step 4: Find eigenvalues and eigenvectors of the matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'A)(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}$. Let $\boldsymbol{c}_k, k = 1, \ldots, K$, be the eigenvectors of this matrix. The k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{b}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$.

2.2.3 Parametric FPCA for sparse functional data

Sometimes the functional data only have sparse observations, and the time points when the observations are made are irregularly spaced ([115]). Our parametric FPCA method can also be extended to conduct FPCA for irregularly spaced and sparsely observed functional data. The FPCs obtained under these conditions also have simple parametric forms and straightforward interpretations, unlike the nonparametric estimation method proposed in [115].

Let Y_{ij} denote the jth observation of $X_i(t)$ at time point t_{ij}, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. It's natural to assume that the observation Y_{ij} made at time t_{ij} contains some measurement error. Thus we consider the model:

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}, \tag{2.7}$$

where ϵ_{ij} denotes the measurement error with mean 0 and variance σ^2. In addition, ϵ_{ij} are assumed to be i.i.d and independent of $X_i(t_{ij})$. The mean curve $\mu(t)$ of the functional data $X_i(t)$ can be estimated using local linear regression ([28]) from the pooled data of all

subjects. Denote the corresponding estimator as $\hat{\mu}(t)$, $t \in \mathcal{I}$. Note that in Model (2.7),

$$\mathrm{Cov}\,(Y_{ij}, Y_{il}) = \mathrm{Cov}\,(X_{ij}(t_{ij}), X_{il}(t_{il})) + \sigma^2 \delta(t_{ij} = t_{il}),$$

where $\delta(t = s) = 1$ if $t = s$; and $\delta(t = s) = 0$ if $t \neq s$. Define

$$G_i(t_{ij}, t_{il}) = \{Y_{ij} - \hat{\mu}(t_{ij})\}\{Y_{il} - \hat{\mu}(t_{il})\}.$$

We can pool $\{G_i(t_{ij}, t_{il}) : t_{ij} \neq t_{il}, i = 1, \ldots, n\}$ together to estimate the covariance function $G(s, t)$. The diagonal elements are eliminated from $G_i(\cdot, \cdot)$ since they account for additional variance of noises. As suggested by [115], a local linear estimator can be employed to estimate the covariance function, where the two-dimensional tuning parameters can be chosen based on leave-one-curve-out cross-validation to smooth the covariance function. Then we can obtain FPCs for the sparse functional data with the following Algorithm 2:

Algorithm 2 : **Parametric FPCA for sparse functional data**

Step 1: Estimate the mean curve $\mu(t)$ using the local linear regression.
Step 2: Estimate the covariance function using the local linear regression method. Denote the estimator as $\hat{G}(s, t)$.
Step 3: Employ the rectangle rule to approximate the entries in the (m, j)-th entry of the matrix $\boldsymbol{A}$

$$\boldsymbol{A}_{mj} = \int_{\mathcal{I}} G(s, t_m)\phi_j(s)\,\mathrm{d}s \approx \frac{L}{M}\sum_{\ell=0}^{M} \hat{G}(t_\ell, t_m)\phi_j(t_\ell)\,,$$

where L is the length of $\mathcal{I}$, and $\hat{G}(s, t)$ is the estimated covariance function.

Step 4: Find eigenvalues and eigenvectors of the matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\boldsymbol{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}$. Let $\boldsymbol{c}_k, k = 1, \ldots, K$, be the eigenvectors of this matrix. The k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{b}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$.

2.2.4 Choosing the Degree of Polynomials p

A practical consideration when employing parametric FPCA is to choose p, the degree of polynomials. Polynomials with a smaller p yield a less flexible, less accurate but more interpretable and robust estimate of FPCs. On the other hand, a more accurate and flexible but less interpretable and robust estimate can be obtained when choosing a larger p. The degree of polynomials p therefore controls the trade off between flexibility and interpretability. The optimal choice of p may depend on the context of the study. In this chapter, we suggest to choose p by comparing the distance between the first K FPCs estimated using parametric FPCA and nonparametric FPCA for each given p. To account for different importance of each FPC, a weighted sum is adopted. To be more specific, we define the weighted distance

of the FPCs estimated using parametric FPCA and nonparametric FPCA:

$$\mathrm{J}(p) = \sum_{k=1}^{K} w_k \int_{\mathcal{I}} |\hat{\psi}_k^{\mathrm{P}}(t) - \hat{\psi}_k^{\mathrm{NP}}(t)| dt,$$

where the weight $w_k = \lambda_k / \sum_{k=1}^{K} \lambda_k$, and $\hat{\psi}_k^{\mathrm{P}}(t)$ and $\hat{\psi}_k^{\mathrm{NP}}(t)$ are the estimated k-th FPC using the parametric FPCA method and the nonparametric FPCA method, respectively. A plot of $\mathrm{J}(p)$ for a variety of p values shows the influence of degree of polynomials. In our experience, we recommend to choose the point at which $\mathrm{J}(p)$ levels off from a practical perspective. We adopt this strategy to choose p in the following applications and simulation studies. Another way to compare the spaces spanned by eigenfunctions obtained from the parametric and nonparametric FPCA methods is to redefine the eigenfunctions with the parametric method to minimize the distance between them and counterparts with the nonparametric method.

2.3 Applications

In this section, we compare the proposed parametric FPCA method with the nonparametric FPCA method using three application examples. Subsection 2.3.2 is an application on sparse functional data, and the rest are applications on dense functional data. The advantage of the parametric FPCA method is that the estimated parametric FPCs have closed-form expressions, which helps to understand and interpret the FPCs. The main risk in using parametric FPCA is that the estimated FPCs may be significantly different from those obtained using nonparametric FPCA, and the estimated parametric FPCs may not explain enough variability of the functional data. We will show that this risk is insignificant in six application examples. When functional data are very bumpy, the risk is still relatively small, which is demonstrated in the last application.

2.3.1 Analysis of Medfly Data

The medfly data is fairly popular among researchers interested in functional data analysis ([39]). This dataset catalogs the number of eggs laid by 50 Mediterranean fruit flies over 25 days, which is assumed to be related to the smooth process that controls fertility. We are interested in exploring modes of variability in eggs laid at each stage, which can reflect the variability of the underlying process governing fertility. Figure 2.1 displays the profiles of the number of eggs laid by 50 medflies in 25 days, in which substantial wiggles and spikes are observable. First we use the smoothing spline method to smooth the original data. Using cubic B-splines, we put one knot at each day and choose a value of 100 for the smoothing parameter since it minimizes generalized cross-validation (GCV). We then estimate FPCs with both the parametric and nonparametric FPCA methods from the smoothed functional data. Nonparametric FPCA suggests that the first two FPCs can explain over 92% of the total variability. So we choose to estimate two FPCs for the medfly data. Figure 2.3 shows

that the weighted distance of the FPCs estimated using parametric FPCA and nonparametric FPCA levels off at $p = 3$. So the parametric FPCA method chooses cubic polynomials to estimate FPCs. The top two FPCs estimated using parametric FPCA are:

$$\hat{\psi}_1(t) = 0.122 + 0.030t - 2.0 \times 10^{-3}t^2 + 2.5 \times 10^{-5}t^3,$$
$$\hat{\psi}_2(t) = -0.561 + 0.110t - 6.1 \times 10^{-3}t^2 + 1.2 \times 10^{-4}t^3.$$

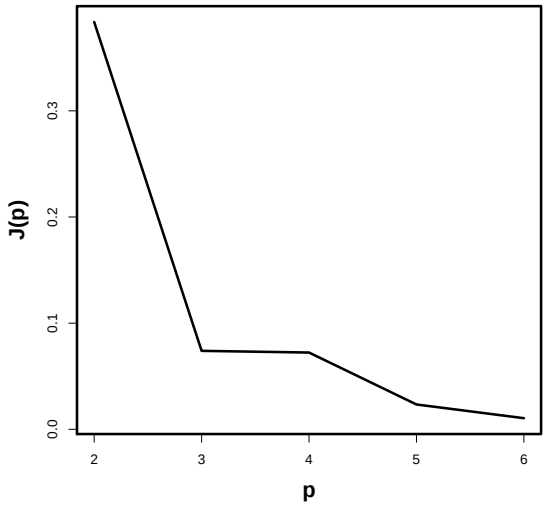


Figure 2.3: The weighted distance J(p) of the FPCs estimated using parametric FPCA and nonparametric FPCA when the degree of polynomials p varies for the medfly data .

Medfly	Method	FPC 1	FPC 2		Total
	Parametric FPCA	62.20%	29.44%		91.64%
	Nonparametric FPCA	62.21%	29.49%		91.70%
CD4	Method	FPC 1			Total
	Parametric FPCA	84.71%			84.71%
	Nonparametric FPCA	85.11%			85.11%
DTI	Method	FPC 1	FPC 2	FPC 3	Total
	Parametric FPCA	58.80%	16.00%	6.00%	80.70%
	Nonparametric FPCA	60.42%	17.64%	7.48%	85.27%

Table 2.1: Comparison of variations explained by the leading FPCs estimated using parametric FPCA and nonparametric FPCA for all three application cases.

Table 2.1 summarizes the comparison of variations explained by the top two FPCs estimated using the parametric and nonparametric FPCA methods. Their performances are very close in terms of the proportions of total variations explained by the top two FPCs. Figure 2.4 displays the shapes of the top two FPCs. Both of them are almost identical when estimated using these two methods.

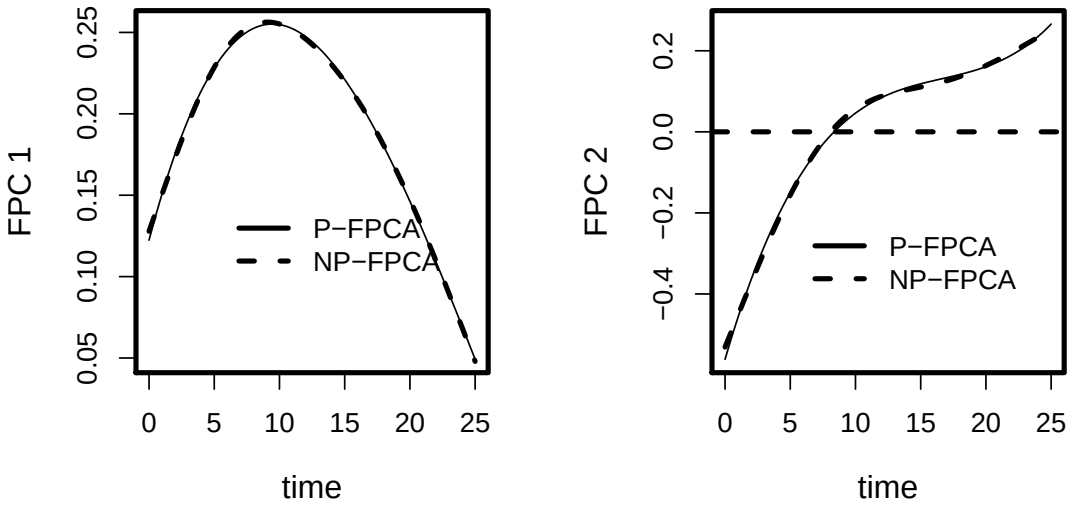


Figure 2.4: The top two FPCs estimated using nonparametric FPCA and parametric FPCA for the medfly data. P-FPCA stands for parametric FPCA; while NP-FPCA stands for nonparametric FPCA.

Using the parametric FPCA method, the first FPC explains around 62.2% of total variability in the medfly egg data. It is positive over the whole time interval and can be interpreted as a weighted average of all values of each curve in the whole time interval. The second FPC explains around 29.4% of total variability in the medfly egg data. It is negative in [1, 8] and positive in [8, 25], which may be interpreted as the change of egg numbers laid after the eighth day.

2.3.2 Analysis of longitudinal CD4 Counts

The dataset has CD4 cell counts of 283 homosexual men who became HIV-positive between 1984 and 1991 from the Multicenter AIDS Cohort Study. A more detailed description of this study can be found in [52]. Figure 2.5 displays the trajectories of CD4 percentage of the 283 homosexual men. For each subject, the measurements are sparsely taken and irregularly spaced, but the measurement time points pooled across all subjects are dense. We have chosen this example to illustrate how parametric FPCA is applied to sparse functional

data, and how its performance compares with the nonparametric FPCA. For convenience, we first normalize the range of the sampling time points to $[0, 1]$ and then fit a local linear regression model to smooth the mean and covariance functions of CD4 cell counts. Next we estimate FPCs for the smoothed covariance function using the nonparametric and parametric methods. We estimate the nonparametric FPCs with the PACE method proposed by [115], and the parametric FPCs with the parametric FPCA method introduced in Section 2.3. The nonparametric method suggests that the first FPC can explain around 85% of total variability in the data. So we choose to estimate one FPC for the CD4 data. The weighted distance of the FPC estimated using parametric FPCA and nonparametric FPCA levels off at $p = 2$. So the parametric FPCA method chooses quadratic polynomials to estimate FPCs. The top FPC estimated using the parametric FPCA method is:

$$\hat{\psi}_1(t) = 0.521 + 1.288t - 0.563t^2.$$

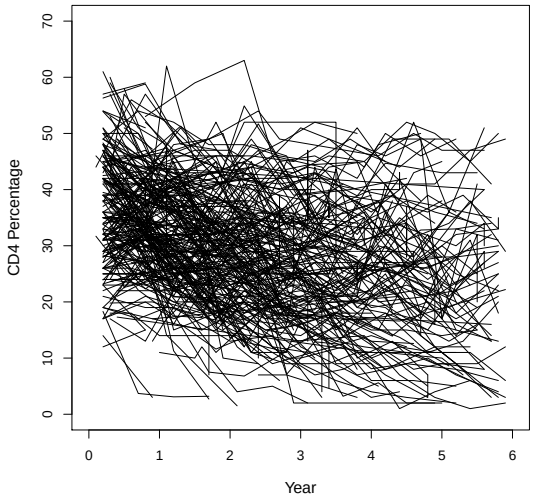


Figure 2.5: The CD4 percentage trajectories of 283 homosexual men.

Table 2.1 summarizes the comparison between the nonparametric and parametric FPCA methods in terms of the proportions of the variability of the longitudinal CD4 cell counts explained by the first FPC. It turns out that the FPC obtained using the parametric FPCA method has captured the variations of the functional data almost as well as the nonparametric FPCA method. Figure 2.6 presents the shapes of the first FPC obtained from these different FPCA methods. Looking at the result we see that there is little disagreement between the nonparametric and parametric methods, with respect to the first FPC.

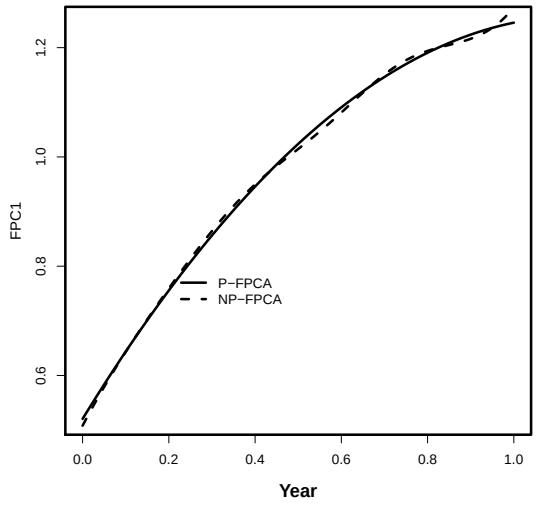


Figure 2.6: The first FPC estimated using nonparametric FPCA and parametric FPCA for CD4 cell counts. P-FPCA stands for parametric FPCA; while NP-FPCA stands for nonparametric FPCA.

The estimated first FPC using parametric FPCA can be interpreted as follows. It plays a dominating role in explaining variability in the CD4 counts data. Since the first FPC is positively valued over the whole time interval, it can reflect the “average" of all (smoothed) profiles in the CD4 sample.

2.3.3 Analysis of Diffusion Tensor Imagining (DTI) data

DTI reveals microscopic details about the architecture of the white matter tracts by measuring the three-dimensional directions of water diffusion in the brain [5]. An R package “**Refund**" provides fractional anisotropy (FA) tract profiles for the corpus callosum (CCA) and the right corticospinal tract of 42 healthy controls and 340 patients with multiple sclerosis.

Figure 2.7 displays the profiles of the CCA sampled at 93 locations in 42 controls. A few fast oscillations can be observed within each of the 42 controls. Cubic B-spline basis functions with one knot at each measurement location were adopted to fit a nonparametric regression to smooth individual profiles. The value of smoothing parameter λ is set to 31, and both parametric and nonparametric FPCA methods are carried out on the smoothed DTI data for comparison. Nonparametric FPCA suggests that the first three FPCs can explain over 85% of total variability in the smoothed CCA profiles. So we choose to estimate three FPC for the DTI data. Since we find that the weighted distance of the FPC estimated using parametric FPCA and nonparametric FPCA levels off at $p = 4$, the parametric FPCA method chooses quartic polynomials to estimate FPCs. The parametric FPCA estimates of

the top three FPCs are given by:

$$\hat{\psi}_1(t) = 0.060 + 9.7 \times 10^{-3}t - 4.8 \times 10^{-4}t^2 + 7.9 \times 10^{-6}t^3 - 4.2 \times 10^{-8}t^4,$$
$$\hat{\psi}_2(t) = 0.225 - 0.032t + 1.7 \times 10^{-3}t^2 - 3.3 \times 10^{56}t^3 + 2.0 \times 10^{-7}t^4,$$
$$\hat{\psi}_3(t) = 0.353 - 0.028t + 4.6 \times 10^{-4}t^2 - 6.7 \times 10^{-7}t^3 - 1.6 \times 10^{-8}t^4.$$

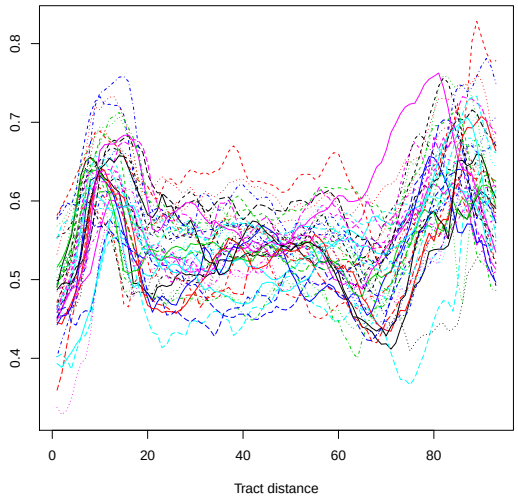


Figure 2.7: Profiles of the corpus callosum across 42 healthy controls.

Since a great amount of variability exists both within and between subjects, we expect that the nonparametric FPCA may slightly outperform the parametric FPCA since the nonparametric basis functions offer greater flexibility. Table 2.1 confirms this. This is the price paid to utilize the simpler representations provided by the parametric FPCA. Figure 2.8 displays the shapes of the FPCs estimated from both nonparametric and parametric FPCA. Not surprisingly, more wiggles are observed in the FPCs obtained from nonparametric FPCA, even though regularization has been imposed to control the roughness of FPCs. The FPCs obtained from parametric FPCA can be treated as smoothed versions of the counterparts from nonparametric FPCA.

Due to the existence of substantial fluctuations, the first FPC estimated using parametric FPCA cannot explain the variance of the sample as well as those in previous examples. The shape of it, however, is still quite stable: positive over the whole time interval. Therefore it can still be regarded as a weighted average of all values of each curve in the sample. A considerable decrease in explaining the variance of the DTI data does not occur for the second FPC, which still accounts for about 16.0% of total variability in the DTI data. As expected, there is one change in sign: positive in [0, 53.5] and negative in [53.5, 86.1].

Accordingly, the second FPC reveals the change of CCA after the time 53.5 if we neglect the fact that it is positive in [86.1, 92]. This will not make an evident difference since the neglected time interval is very short and thus makes minor contribution to the whole process. Not surprisingly, the third FPC is inferior in explaining total variability in the DTI data and less unvarying in terms of shape. It is negative in [17.3, 65.9], and positive elsewhere; it can therefore be interpreted as the difference in CCA during the interval [17.3, 65.9] and other time periods.

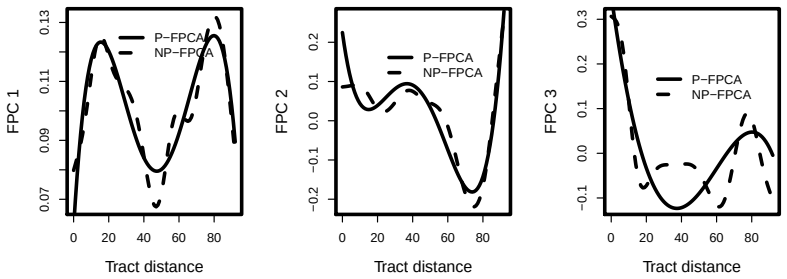


Figure 2.8: The first three FPCs estimated from NP-FPCA and P-FPCA for DTI data. NP-FPCA stands for nonparametric FPCA; while P-FPCA stands for parametric FPCA.

2.4 Simulation Study

To compare the parametric FPCA method and the nonparametric FPCA method, we conduct a simulation study. Since the true FPCs are known prior to simulation, we compare the bias, standard error and the squared root of the mean squared error (RMSE) of the estimated FPCs using each method.

We choose the top three FPCs estimated from the medfly data using the nonparametric FPCA method, which are shown in Figure 2.2, and the corresponding eigenvalues together with the mean curve of the smoothed functional data to generate random curves in the simulation. More specifically, the random curves are generated as

$$X_i(t) = \mu(t) + \xi_{i1}\psi_1(t) + \xi_{i2}\psi_2(t) + \xi_{i3}\psi_3(t), \quad i = 1, \ldots, n$$

where $\mu(t)$ denotes the mean curve, $\xi_{ij} \sim N(0, \lambda_j)$, $j = 1, 2, 3$, with λ_1, λ_2 and λ_3 denoting the largest three eigenvalues, respectively, and $\psi_1(t)$, $\psi_2(t)$ and $\psi_3(t)$ denote the top three FPCs estimated from the medfly data using nonparametric FPCA. Since the data are generated from the FPCs estimated using nonparametric FPCA, the nonparametric FPCA method should outperform the parametric FPCA method. On the other hand, parametric FPCA turns out to be more robust in comparison with nonparametric FPCA when the functional data are contaminated by outlier curves.

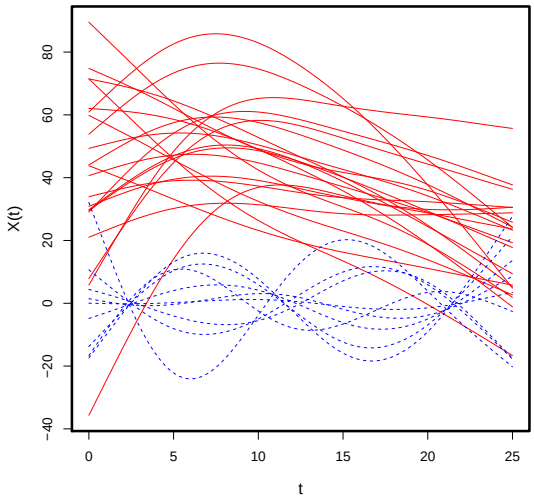


Figure 2.9: The trajectories of 21 true curves and 9 contaminated curves randomly selected from 1 of the 200 simulations. The solid (——) and dashed (– – –) lines represent the true curves and contaminated curves, respectively.

We generate $n = 200$ curves in total; each curve is sampled from $m = 100$ regular grid points on $[0, 25]$. Then we compare the performances of nonparametric and parametric FPCA in two scenarios: outlier curves are absent and present. In the first scenario, all 200 curves have no outlier curves. In the second scenario, 30% of these 200 curves are selected randomly to be outlier curves. More specifically, the outlier curves are assumed to be generated from the linear combination of the fourth and fifth FPCs estimated from the medfly data using the nonparametric FPCA method with corresponding eigenvalues scaled to make the variabilities of the outliers comparable with the variabilities of $X(t)$. Figure 2.9 presents the trajectories of 21 true curves and 9 contaminated cures randomly selected from one simulated dataset in the second scenario. Both parametric FPCA and nonparametric FPCA are conducted on the sample of 200 curves. To assess the performance of these two FPCA approaches in both scenarios, 100 simulation replicates are conducted to estimate the bias, standard error and RMSE of the FPCs estimated using both methods.

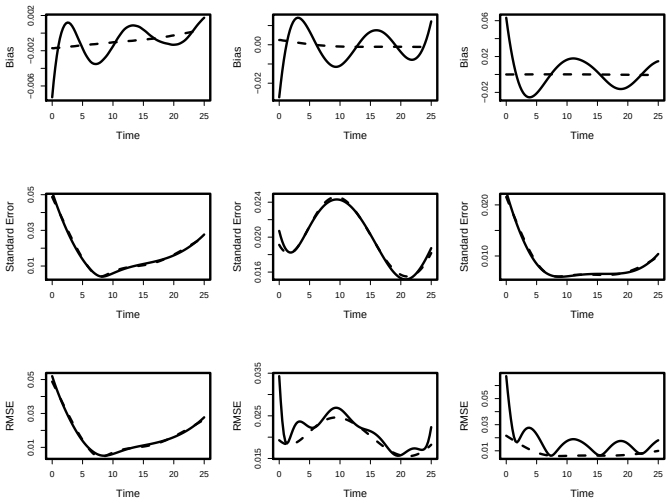


Figure 2.10: The estimated bias, standard error and RMSE of the top three FPCs estimated from parametric FPCA and nonparametric FPCA when there is no outlier curve. In each panel, the solid (——) and dashed (− − −) lines represent the FPC estimated from parametric FPCA and nonparametric FPCA, respectively. The first, second and third columns correspond to the first, second and third FPCs, respectively.

Figures 2.10 and 2.11 summarize the estimated bias, standard error and root mean squared error (RMSE) of the top three FPCs estimated from both nonparametric FPCA and parametric FPCA in these two scenarios, respectively. When the 200 curves have no outlier curves, nonparametric FPCA has smaller bias and RMSE than parametric FPCA for all three FPCs. This is not surprising since nonparametric FPCA, compared with parametric FPCA, is more flexible, thus more effective in capturing some local features such as rapid fluctuations of true FPCs. On the other hand, when curves are contaminated with outlier curves in the second scenario, Figure 2.11 shows that parametric FPCA compares favourably with nonparametric FPCA. In the presence of outlier curves, the two approaches have a similar performance in terms of bias. But parametric FPCA leads to a much more stable estimate of the three FPCs in comparison with nonparametric FPCA. Although nonparametric FPCA is able to capture features from both contaminated and non-contaminated curves with a large number of basis functions, the flexibility of nonparametric FPCA results in more unrobust estimates of FPCs. In summary, the parametric FPCA estimates are more robust than their nonparametric counterparts in the presence of outlier curves. Furthermore, when RMSE is used as the criterion to assess the performance of the estimated FPCs, parametric FPCA yields more accurate FPC estimates when the functional data are contaminated with outlier curves.

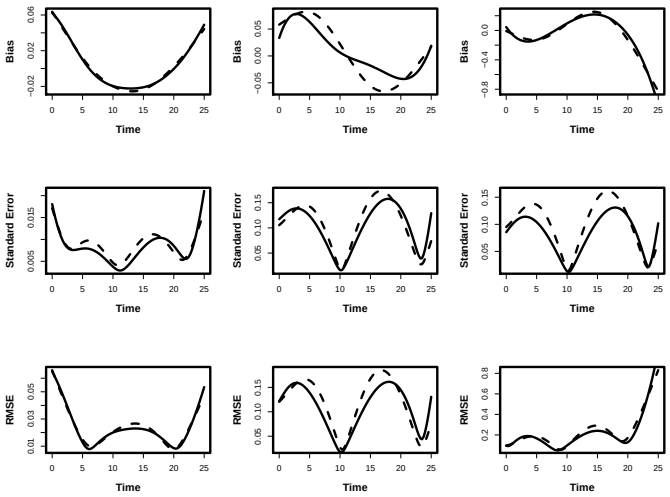


Figure 2.11: The estimated bias, standard error and RMSE of the top three FPCs estimated from parametric FPCA and nonparametric FPCA when 30% of curves are contaminated with outliers. In each panel, the solid (——) and dashed (− − −) lines represent the FPC estimated from parametric FPCA and nonparametric FPCA, respectively. The first, second and third columns correspond to the first, second and third FPCs, respectively.

2.5 Conclusions

FPCA is a powerful tool to detect major sources of variation in functional data. Even when the functional data displays great variability and is highly oscillatory, the top FPCs often still have simple trends and may be approximated by simple parametric functions. We propose a parametric FPCA method which is able to estimate the top FPCs with some parametric functions for either dense or sparse functional data.

Our parametric FPCA method is demonstrated with three applications in a variety of fields. The performance of the parametric FPCA method is satisfactory in terms of explaining similar variations of functional data in comparison with the more complicated nonparametric FPCA method. In addition, the estimated FPCs using these two FPCA approaches are very similar as well. An advantage of parametric FPCA is that compared with FPCs estimated from nonparametric FPCA, the ones from parametric FPCA are simple parametric functions; thus they are considerably easier to understand and interpret. Last but not least, as shown in the simulation study in Section 2.4, the FPCs estimated from the parametric FPCA are more robust compared with the nonparametric FPCA counterparts, when a small proportion of curves are contaminated with outlier curves.

Although in many applications the performance of parametric FPCA is comparable with that of the nonparametric FPCA, it should also be noted that there exist cases when nonparametric FPCA may outperform parametric FPCA, particularly when great variability

is presented within and between curves. Take for instance the DTI data in Section 2.3.3, where the nonparametric FPCA might be more appealing since the basis functions have greater flexibility, which can better capture the local variability of the FPCs.

Chapter 3

Weighted Empirical Likelihood Inference for Dynamical Correlations

3.1 Introduction

With the development of techniques in collecting data, analysis of data with complex structures has become increasingly popular in statistical research. Functional data analysis has attracted extensive attention in fields like image analysis, disease diagnostic and gene regulation ([85], [57], [3]). Particularly, analysis of univariate functional data analysis has been focused on for a long time. Among them, functional regression and functional principal component analysis are two particularly important problems for researchers. For a more comprehensive view of them, refer to monographs like [86], [34], [31] and [47]. In contrast, multivariate functional data have not been studied popularly.

Dependence modeling plays a significant role in multivariate data analysis. Dependence structure among stochastic processes likewise is worth studying for its own merit in functional data analysis. [45] proposed the rank correlation coefficient between two random functions in cluster analysis. [25] considered the dynamical correlation of two random functions, which can be regarded as an extension of the correlation coefficient in multivariate data. [113] developed singular valued decomposition for pairs of functional data.

In this chapter, we focus on the dynamical correlation, which summarizes the correlation of two functional variables over their domain. There have been substantial applications related to this concept in literature. For instance, a graphical Gaussian model for functional data was developed by [75] based on the dynamical correlation. [63] applied the dynamical correlation in a psychological study. The dynamical correlation was employed in [6] to study the mechanism when $CD8^+$ T cells are activated in HIV-infected individuals.

[25] proposed a sample average estimator for dynamical correlation. However, the asymptotical variance of this estimator is not tractable. Thus they recommended a bootstrap procedure to construct a confidence interval for dynamical correlation based on the av-

erage estimator. Since a local linear regression is used to smooth functional data in the pre-processing step, both fitted values and residuals must be resampled in their proposed bootstrap procedure. This leads to a computationally intensive procedure for inferring dynamical correlation, which may restrict its applications in practice. Another concern of this estimator is that each subject makes an equal contribution to the final estimator. This may not be an optimal treatment in practice since subjects with dense observations may provide a more accurate estimate of dynamical correlation in comparison with subjects with less dense observations. An estimator which can adjust weights on each subject is, therefore, more appealing than a simple average of the estimate from each subject. Actually [25] also considered such an estimator in applications under the bootstrap-based inference framework, even though this estimator was not thoroughly investigated in the paper. Furthermore, we find in simulation studies that this bootstrap procedure is unstable and cannot provide a reliable confidence interval for dynamical correlation when the functional data are irregularly spaced.

Motivated by these observations, we propose a weighted average estimator for the dynamical correlation which can adaptively adjust weights on each subject. The weights are chosen via the weighted empirical likelihood. This new estimator can be regarded as an alternative to that mentioned in applications of [25]. The statistical inference framework, however, is different from the bootstrap-based analysis proposed by [25]. The main advantage of this method is that we do not need to estimate the standard error of the estimator for statistical inference since the test statistic itself is self-normalized under some regularity conditions. This improves computational efficiency substantially for constructing confidence intervals, in comparison with the bootstrap-based method proposed by [25]. This improvement will be demonstrated in our simulation studies presented in Section 3.4. The second advantage of our method is that our proposed weighted empirical likelihood is possible to yield a confidence interval with a more accurate coverage probability than its counterpart by the bootstrap method when functional data are irregularly spaced. Our simulation studies show that the coverage probability of the proposed confidence interval is still close to the nominal level even when the functional data are irregularly designed. In summary, the main advantage of our method in statistical inference is well reflected when there exist great variations in the number of observations across subjects. Note that our method cannot accommodate the case when functional data have sparse observations for every subject.

The chapter is organized as follows. The definition of dynamical correlation is reviewed in Section 3.2. Then a new method of constructing confidence intervals for dynamical correlation based on the weighted empirical likelihood is proposed. For comparison, we introduce the bootstrap method for constructing confidence intervals for dynamical correlation proposed by [25] later. Some theoretical properties of the proposed inference tool are given in Section 3.3. The proofs are deferred to Appendix B. Section 3.4 compares the performances of the weighted empirical likelihood-based method and the bootstrap method in associated

confidence intervals for dynamical correlation via simulation studies. The proposed method is illustrated with three applications in Section 3.5. Section 3.6 concludes the chapter.

3.2 Confidence Intervals for Dynamical Correlation

We propose the weighted empirical likelihood method to estimate confidence intervals for the dynamical correlation of two random functions. First, we introduce the definition of the dynamical correlation of two random functions. We then introduce our method. Last, the point estimator and the bootstrap confidence interval for dynamical correlation proposed by [25] are reviewed.

3.2.1 Dynamical Correlation

The dynamical correlation was firstly proposed by [25] to use a single measure to describe the correlation between two longitudinal curves. As noted by them, it is simpler and more efficient, compared with functional canonical correlation ([43], [86]), which is well defined provided restrictive assumptions.

Let $f_j, j = 1, \ldots, p$, be p random functions defined over a compact interval $\mathcal{I}$. Suppose they belong to $L^2(dw)$, the collection of all square integrable functions with respect to a measure $dw = wdt$. That is, $\mathrm{E}\{\int_{\mathcal{I}} f_j^2(t)w(t)dt\} < \infty$, $j = 1, \ldots, p$, where the weight function $w(t)$ defined over $\mathcal{I}$ satisfies (i) $w(t) \geq 0,\ t \in \mathcal{I}$, (ii) $\int_{\mathcal{I}} w(t)dt = 1$, (iii) $\int_{\mathcal{I}} w^2(t)dt < \infty$. For any two functions $f, g \in L^2(dw)$, the inner product of them is defined by $\langle f, g\rangle = \int_{\mathcal{I}} f(t)g(t)dw$.

These random functions can be expressed as

$$f_j(t) = \mu_{0,j} + \mu_j(t) + \xi_{0,j} + \sum_{l=1}^{\infty} \xi_{l,j}\phi_l(t),\ j = 1, \ldots, p, \tag{3.1}$$

where the fixed intercept $\mu_{0,j} = \mathrm{E}(\langle f_j, 1\rangle)$, $\mu_j(t)$ is the fixed mean function with $\mu_j(t) \in L^2(dw)$ and $\langle \mu_j(t), 1\rangle = 0$, and $\xi_{0,j}$ serves as a random intercept. Let $\phi_0(t) \equiv 1$ and we assume that the fixed functions, $\{\phi_l(t)\}_{l=0}^{\infty}$, constitute a complete orthonormal basis of the space $L^2(dw)$; that is, $\langle \phi_i, \phi_{i'}\rangle = 0$ if $i \neq i'$ and 1 otherwise. The other random components, $\xi_{l,j}$'s, are the coefficients when the random function f_j is represented in terms of this orthonormal basis. We assume that $\xi_{l,j}$'s are uncorrelated with $\xi_{0,j}$ for $l \geq 1$, and they satisfy $\mathrm{E}(\xi_{l,j}) = 0,\ \mathrm{Var}(\xi_{l,j}) = \sigma_{l,j}^2 < \infty,\ l \geq 0$ and $0 < \sum_{l=0}^{\infty} \sigma_{l,j}^2 < \infty$. Note that the expansion given in (3.1) is not the Karhunen-Loève expansion of $f_j(t)$. The random coefficients, $\xi_{l,j}$'s ($l \geq 1$), are not required to be uncorrelated with each other. Furthermore, the fixed functions $\mu_j(t)$ and $\phi_l(t)$'s are assumed to be twice continuously differentiable over $\mathcal{I}$.

Let $M_j = \langle f_j, 1\rangle$. Then $M_j = \mu_{0,j} + \xi_{0,j}$ and $f_j(t) - M_j = \mu_j(t) + \sum_{l=1}^{\infty} \xi_{l,j}\phi_l(t)$. The standardized f_j's are defined as

$$f_j^\star(t) = \frac{f_j(t) - M_j - \mu_j(t)}{[\int_{\mathcal{I}}\{f_j(t) - M_j - \mu_j(t)\}^2 dw]^{\frac{1}{2}}}. \tag{3.2}$$

The dynamical correlation between f_{j_1} and f_{j_2}, $1 \le j_1, j_2 \le p$, is defined as the expected inner product of $f_{j_1}^\star$ and $f_{j_2}^\star$,

$$\rho_{j_1 j_2} = \mathrm{E}\,\langle f_{j_1}^\star, f_{j_2}^\star\rangle.$$

As argued by [25], the dynamical correlation satisfies $|\rho_{j_1 j_2}| \le 1$ for any two random functions which are square integrable with respect to dw.

3.2.2 Confidence Interval via the Weighted Empirical Likelihood

As mentioned in Section 3.1, [25] proposed an estimator which averages the point estimate for each subject to estimate dynamical correlation. They demonstrated that the performance of this estimator is satisfactory when random functions are densely and regularly observed. However, they did not explore its performance when random functions are irregularly observed.

Random functions are more often observed at different time points for different subjects, which is called the irregular design in this chapter. For simplicity, we assume that for the same subject, p random functions are observed at the same time points, though the method presented in the following can be extended to deal with more general cases as well. More specifically, we have observations $\{(f_1(t_{i1}), \ldots, f_p(t_{i1})), \ldots, (f_1(t_{in_i}), \ldots,$
$f_p(t_{in_i}))\}$ for the i-th subject, $i = 1, \ldots, n$, where n denotes the number of subjects. Furthermore, we assume that these observations may also have measurement errors.

Now we introduce a novel approach to construct confidence intervals for dynamical correlation of two random functions under the irregular design. Inspired by the idea of [25], we first employ local linear regression to smooth each random function for all subjects. For $i = 1, \ldots, n$, $1 \le j_1 \ne j_2 \le p$, let $f_{i,j_1}^S(t)$ and $f_{i,j_2}^S(t)$ denote smoothed $f_{i,j_1}(t)$ and $f_{i,j_2}(t)$ with local linear smoothing, respectively. To define a sample version of (3.2), we introduce $\tilde{f}_{i,j}^S(t) = f_{i,j}^S(t) - 1/n\sum_{i=1}^n f_{i,j}^S(t)$ and $\tilde{M}_{i,j}^S = \langle \tilde{f}_{i,j}^S, 1\rangle$, $i = 1, \ldots, n$, $j = j_1, j_2$. Hence $\tilde{f}_{i,j}^S(t) - \tilde{M}_{i,j}^S = f_{i,j}^S(t) - M_{i,j}^S - 1/n\sum_{i=1}^n (f_{i,j}^S(t) - M_{i,j}^S)$. The standardized random function in the ith subject, defined by

$$\hat{f}_{i,j}^{S\star}(t) = \frac{\tilde{f}_{i,j}^S(t) - \tilde{M}_{i,j}^S}{\left\{\int_{\mathcal{I}} (\tilde{f}_{i,j}^S(t) - \tilde{M}_{i,j}^S)^2 dw\right\}^{\frac{1}{2}}}, \tag{3.3}$$

therefore provides a reasonable estimate of $f_j^\star(t)$ defined in (3.2). As a result, the point estimate of dynamical correlation of random functions f_{j_1} and f_{j_2} provided by the ith

subject can be defined as

$$\hat{\rho}^S_{i,j_1j_2} = \langle \hat{f}^{S\star}_{i,j_1}, \hat{f}^{S\star}_{i,j_2} \rangle. \tag{3.4}$$

Using (3.4), we are able to obtain a point estimate of dynamical correlation of each pair of random functions for each subject. In the Estimation section, [25] estimated dynamical correlation by averaging these point estimates. However, this may not be an optimal strategy to make use of all $\hat{\rho}^S_{i,j_1j_2}$'s. Since different subjects have different numbers of measurements, their contributions to the final estimate of dynamical correlation may not be equally important. As a result, a simple average of $\hat{\rho}^S_{i,j_1j_2}$'s may not be an optimal choice. [25] pointed out and addressed this issue in applications by using weighted averages to estimate dynamical correlation.

A natural problem about choosing appropriate weights arises. Another interesting problem is to construct a valid confidence interval for dynamical correlation based on $\hat{\rho}^S_{i,j_1j_2}$'s. To solve these problems, we borrow the idea of the weighted empirical likelihood proposed by [110]. The log weighted empirical likelihood is defined as $l(F) = T_n \sum_{i=1}^n c_i\{\log(p_i) - np_i\}$, where c_i is the weight put on the ith subject, $T_n = n/\sum_{i=1}^n c_i$ and p_i is probability put on $\hat{\rho}^S_{i,j_1j_2}$ satisfying $\sum_{i=1}^n p_i = 1$. Obviously $l(F)$ achieves its maximum when $p_1 = p_2 = \cdots = p_n = 1/n$. The log likelihood ratio is then as

$$\begin{aligned} l(F) - \max_F l(F) &= T_n \sum_{i=1}^n c_i\{\log(p_i) - np_i\} - T_n \sum_{i=1}^n c_i\{-\log(n) - 1\} \\ &= T_n \sum_{i=1}^n c_i\{\log(np_i) - np_i + 1\}. \end{aligned}$$

Given the additional constraint that $\mathrm{E}\,(\hat{\rho}^S_{i,j_1j_2}) \approx \rho_{j_1j_2}$ under some regularity conditions provided in Theorem 2 in [25], the profile log weighted empirical likelihood ratio is defined as

$$\tilde{R}(\rho) = T_n \sum_{i=1}^n c_i\{\log(np_i) - np_i + 1\}, \tag{3.5}$$

where p_i's maximize $l(F)$ subject to $p_i \geq 0, i = 1, \ldots, n$, $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i \hat{\rho}^S_{i,j_1j_2} = \rho$. To account for the effect of different numbers of observations for different subjects, we choose

$$c_i = \frac{1}{n_i} \tag{3.6}$$

in (3.5) when applying the weighted empirical likelihood. This specific choice actually follows [110], who suggested that c_i should be proportional to the variance of the estimate from ith subjection. [110] suggested χ^2_1 to calibrate the asymptotic distribution of the log weighted empirical likelihood ratio for this choice. Likewise, we propose a $1-\alpha$ confidence interval for the true dynamical correlation $\rho_{j_1j_2}$: $\{\rho : -2\tilde{R}(\rho) < \chi^2_1(1-\alpha)\}$, where $\chi^2_1(1-\alpha)$ denotes the $1-\alpha$ quantile of χ^2 distribution of one degree of freedom. Numerical studies presented

in Section 3.4 show that χ_1^2 approximates $-2\tilde{R}(\rho_{j_1j_2})$ reasonably well. As shown in [76] and [110], the logarithms of both the unweighted and weighted empirical likelihood ratio converge to a chi-squared distribution under some regularity conditions. But as argued in [110], these two methods have different performances in statistical inference for samples with a small size. The superiority of the weighted method is demonstrated via a simulation study in [110], which shows that even though the lengths are similar, the confidence interval generated from the weighted empirical likelihood has a coverage probability closer to the nominal level than its counterpart, especially when the sample size is small.

Next we consider another scenario in which random functions are observed over a common grid of time points across all subjects. This scenario is called the regular design in this chapter. We assume that observations consist of $\{(f_{i,1}(t_k), \ldots, f_{i,p}(t_k)),\ i = 1, \ldots, n,\ k = 1, \ldots, m\}$, where $t_1, t_2, \ldots, t_m$ are m distinct points of $\mathcal{I}$. Without loss of generality, we assume that $t_1 < t_2 \ldots < t_m$. Using the same idea as in irregular design, the corresponding log weighted empirical likelihood ratio becomes

$$\tilde{R}(\rho) = \sum_{i=1}^{n} \log(np_i), \tag{3.7}$$

where p_i's maximize $\prod_{i=1}^n p_i$ subjects to $p_i \geq 0, i = 1, \ldots n$, $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i \hat{\rho}^S_{i,j_1j_2} = \rho$. If $\hat{\rho}^S_i$'s were identically and independently distributed, $\tilde{R}(\rho)$ is actually the same as the log empirical likelihood ratio defined by [76]. We employ χ_1^2 to calibrate the asymptotic distribution of $-2\tilde{R}(\rho_{j_1j_2})$ as well. This chi-square calibration will be investigated both theoretically and numerically later. Numerical studies show that $\{\rho : -2\tilde{R}(\rho) < \chi_1^2(1-\alpha)\}$ is a reasonable $1-\alpha$ confidence interval for $\rho_{j_1j_2}$.

3.2.3 Confidence Intervals for Dynamical Correlation via Bootstrap

The point estimator for dynamical correlation between $f_{j_1}(t)$ and $f_{j_2}(t)$ proposed by [25] is given by:

$$\hat{\rho}_{j_1j_2} = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}^S_{i,j_1j_2}. \tag{3.8}$$

They considered confidence intervals for dynamical correlation based on $\hat{\rho}^S_{i,j_1j_2}$'s as well. Provided some regularity conditions and assuming $\mu_j(t)$ is known, $j = j_1, j_2$, they proved that the estimator in (3.8) is asymptotically normal if $0 < |\rho_{j_1j_2}| < 1$. Based on this theoretical result, a confidence interval for the dynamical correlation would be immediately available if we are able to find a plausible estimate of the variance of the estimator. However, they argued that the variance of $\hat{\rho}_{j_1j_2}$ is rather involved and there's no straightforward estimate of it. They hence suggested employing the bootstrap to obtain a confidence interval. To account for the uncertainty introduced by the pre-smoothing step via a local linear smoother, they proposed to sample the fitted values and residuals from the pre-smoothing

step separately. Additionally, a relatively large number, $B = 500$ of bootstrap samples were recommended to obtain a plausible confidence interval. To implement the above procedure to compute the point estimator and the bootstrap confidence interval for dynamical correalation, a package called **dynCorr** was developed by [24].

There are several issues associated with the estimator $\hat{\rho}_{j_1 j_2}$ and the bootstrap strategy, which may flaw their applications. First of all, each $\hat{\rho}^S_{i,j_1 j_2}$ makes equal contribution to $\hat{\rho}_{j_1 j_2}$ in (3.8), the final estimator of $\rho_{j_1 j_2}$. As pointed out in Section 3.2.2, this may not be an optimal treatment under irregular design. Noting this fact, [25] suggested using a weighted average of $\hat{\rho}^S_{i,j_1 j_2}$'s to estimate $\rho_{j_1 j_2}$ to accommodate the irregular design in applications. Second, the bootstrap procedure was proposed to account for the uncertainty in the pre-smoothing step and to estimate the variance of the average point estimator. But the proposed bootstrap procedure is computationally intensive when the pre-smoothing step is included and data are resampled $B = 500$ times. Last but not least, the performance of the bootstrap method is not satisfactory under irregular design. The last issue will be demonstrated in numerical studies presented in Section 3.4. The weighted empirical likelihood, however, can circumvent estimating the variance since the log weighted likelihood ratio statistics itself is self-normalized with the chi-square as the limiting distribution. Unlike the bootstrap method, this chi-square calibration based on the weighted empirical likelihood results in a confidence interval with an appealing coverage probability even under the irregular design.

3.3 Theoretical Properties

Let $\rho_{j_1 j_2}$ denote the true dynamical correlation between $f_{j_1}(t)$ and f_{j_2}, $1 \le j_1, j_2 \le p$, and $j_1 \ne j_2$. Then we have

Theorem 1 *Under the assumptions listed in Appendix A, if* $0 < |\rho_{j_1 j_2}| < 1$, *and both* $\mathrm{E}\left(f_{j_1}(t)\right)$ *and* $\mathrm{E}\left(f_{j_2}(t)\right)$ *are known, then* $-2\tilde{R}(\rho_{j_1 j_2}) \to \chi^2(1)$ *in distribution, where* $\tilde{R}(\cdot)$ *is defined in* (3.7).

The proof is deferred to Appendix B.

3.4 Simulation Studies

In this section, several simulation scenarios are considered to compare the confidence intervals for dynamical correlation of two random functions based on the bootstrap method and the weighted empirical likelihood. Our main concern consists of computational time and coverage probabilities of these types of confidence intervals.

We generate n identically and independently copies of two random functions, f_1 and f_2, on $[0, 1]$. Here $n = 50$ or 100. Let $Y_{ij}(t)$ denote the measurement of f_j of ith copy at time t. More specifically, $Y_{ij}(t) = f_{ij}(t) + e_{ij}(t),\ i = 1, \ldots, n,\ j = 1, 2$, where

1. The random function is defined as $f_{ij}(t) = 1+\sum_{l=0}^{2} \xi_{l,j}\phi_l(t)$. In other words, $\mu_j(t) \equiv 0$ and $\mu_{0,j} = 1$ for $j = 1, 2$. In addition, there are only three random components in both random functions.

2. The fixed orthogonal functions ϕ_l's take the following form: $\phi_0(t) = 1$, $\phi_1(t) = 2\sqrt{3}(t-1/2)$ and $\phi_2(t) = 6\sqrt{5}(t-1/2)^2 - \sqrt{5}/2$.

3. The random components are generated from a centered multivariate Gaussian distribution with a covariance matrix as follows: $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, where each block submatrix is defined as $\Sigma_{11} = \text{diag}(1, 1/2, 1/3)$, $\Sigma_{12} = \text{diag}(1/3, 1/4, 1/6)$ and $\Sigma_{22} = \text{diag}(1/2, 1/3, 1/4)$.

4. $e_{ij}(t) \sim N(0, 1/4)$ denotes the measurement error of the observation of f_j from the ith copy evaluated at time t. These measurement errors are assumed to be independent of each other.

Under this setup, the true dynamical correlation between f_1 and f_2 is 0.5. The above simulation setups are the same as in [25].

To compare the performances of the bootstrap method and the weighted empirical likelihood method in associated confidence intervals for dynamical correlation, both irregular and regular designs are considered. In both designs, a grid of equally-spaced bandwidths $\{.01, .035, \ldots, .36\}$ is chosen in the local linear smoother in the pre-smoothing step. The weight function $w(t)$ is taken to be 1 over $[0, 1]$. To evaluate the real coverage probabilities of the 1 - α confidence intervals for dynamical correlation generated from both the bootstrap method and the weighted empirical likelihood, $K = 100$ simulation trials are run. Here we focus on $\alpha = 0.05$, namely 95% confidence intervals for dynamical correlation.

Irregular Design: In the first scenario, the ith copy of random functions are observed on n_i of these 100 time points, where n_i is uniformly sampled from $\{25, 26, \ldots, 100\}$. Table 3.1 summarizes the coverage probabilities of these two types of confidence intervals for $\alpha = 0.05$ across the 15 bandwidth choices under the irregular design. It can be observed that there's a remarkable gap between these two methods in terms of coverage probability. The weighted empirical likelihood method is robust to the choice of bandwidths and the coverage probabilities of the confidence intervals obtained from most bandwidth choices are close to the nominal level. In contrast, the performance of the bootstrap method is not satisfactory at all, no matter from the perspective of robustness or from coverage probability.

Data	n	Method	h														
			.01	.035	.06	.085	.11	.135	.16	.185	.21	.235	.26	.285	.31	.335	.36
irregular	50	B	.90	.79	.70	.67	.63	.63	.60	.62	.59	.66	.63	.57	.67	.63	.59
		E	.33	.88	.91	.96	.93	.91	.94	.96	.91	.94	.97	.91	.95	.95	.96
	100	B	.47	.58	.49	.43	.41	.38	.40	.39	.42	.35	.44	.33	.46	.35	.43
		E	.12	.87	.97	.96	.93	.94	.92	.93	.96	.94	.91	.98	.91	.95	.96

Table 3.1: Summary of coverage probabilities of two confidence intervals under the irregular design: “B" and “E" stand for the bootstrap and the weighted empirical likelihood based confidences, respectively. The bandwidth used in the local linear smoother is denoted by h and the coverage probabilities are calculated from the 100 Monte Carlo simulations.

Regular Design: In the second scenario, the observational time points for both random functions are chosen to be $m = 100$ equidistant points on $[0, 1]$. Under the regular design, Table 3.2 indicates that both the bootstrap confidence interval and the empirical likelihood-based confidence interval are robust to the choice of the bandwidth in the local linear smoother. Furthermore, the coverage probabilities of both of these two confidence intervals are quite close to the nominal level, even though the bootstrap method is slightly better in some exceptional choices of bandwidth.

Data	n	Method	h														
			.01	.035	.06	.085	.11	.135	.16	.185	.21	.235	.26	.285	.31	.335	.36
regular	50	B	.92	.98	.94	.98	.99	.95	.96	.98	.94	.94	.98	.94	.95	.96	.94
		E	.79	.95	.96	.88	.95	.91	.98	.96	.97	.99	.95	.94	.94	.92	.91
	100	B	.83	.94	.97	.96	.97	.96	.94	.96	.95	.95	.94	.94	.94	.94	.94
		E	.63	.94	.97	.95	.94	.90	.97	.93	.97	.96	.94	.99	.93	.89	.93

Table 3.2: Summary of coverage probabilities of two confidence intervals under the regular design. “B" and “E" stand for the bootstrap and the weighted empirical likelihood based confidences, respectively. The bandwidth used in the local linear smoother is denoted by h and the coverage probabilities are calculated from the 100 Monte Carlo simulations.

A further comparison between these two methods is made in terms of computational time. According to Table 3.3, the weighted empirical likelihood method is considerably more efficient than the bootstrap method in computation. The bootstrap method spends substantial time on estimating the variability of the average estimator of dynamical correlation by resampling both the smoothed random functions and residuals when a local linear smoother is involved. Nevertheless, this step can be circumvented in the weighted empirical likelihood method. As pointed out by one referee, it should be noted that the bootstrap method is able to account for the uncertainty in the pre-smoothing step, while the weighted empirical likelihood method ignores this source of uncertainty in constructing confidence intervals for dynamical correlation.

Data	n	Method	Time (s)
irregular	50	B	205
		E	0.44
	100	B	387
		E	0.85
regular	50	B	214
		E	0.55
	100	B	450
		E	1

Table 3.3: Summary of average computation time of two confidence intervals: “B" and “E" stand for the bootstrap and the weighted empirical likelihood based confidences, respectively. The average computation times are calculated from the 100 Monte Carlo simulations and all bandwidths considered in Table 3.1 and Table 3.2.

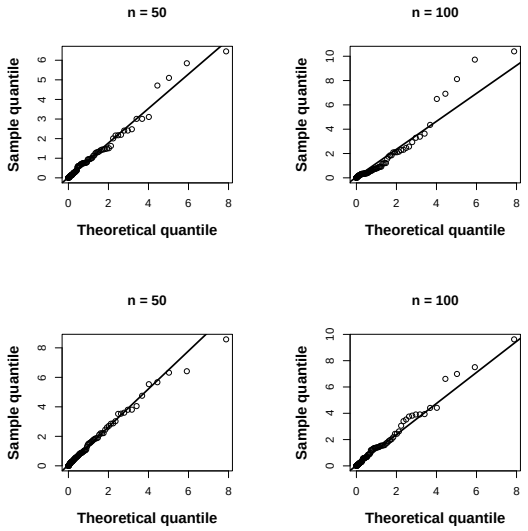


Figure 3.1: Q-Q plots of the weighted empirical likelihood ratio statistics. The upper and bottom panels are for the irregular and regular designs, respectively. The left and right panels correspond to $n = 50$ and $n = 100$, respectively.

To demonstrate that χ^2_1 provides a reasonable calibration when constructing confidence intervals for dynamical correlation based on the weighted empirical likelihood method, Figure 3.1 compares the empirical distribution of the weighted empirical likelihood ratio test statistics with χ^2_1 in the four different cases. For all four panels in Figure 3.1, $h = 0.21$ is used in the local linear smoothing. Due to the robustness of the weighted empirical likelihood method, this specific choice should not play a critical role in determining shapes of these four panels. It turns out that χ^2_1 can provide a satisfactory approximation to the

distribution of the empirical likelihood ratio test statistic even for a moderate number of subjects for both irregular and regular designs. In the top right panel, five points with a theoretical quantile greater than 4 deviate slightly from the straight line. This might seem to be a serious issue when compared with the top left panel, in which only 50 independent copies of random functions are generated. However, quantiles greater than 4 for the χ^2_1 distribution correspond to quantiles greater than 2 for the $N(0,1)$ distribution; thus they are greater than 95% quantile of $N(0,1)$. Usually an extreme quantile cannot be estimated accurately due to data sparsity near the extreme quantile.

3.5 Applications

In this section, we apply the proposed empirical likelihood-based method to analyze the dynamical correlation between random functions in three applications.

3.5.1 Dynamical Correlation of Air Pollutants

Dynamical correlation between air pollutants is treated in the first example. The dataset is obtained from the **NMMAPSdata** package ([79]), which contains daily mortality, air pollution, and weather data for the study of national morbidity, mortality, and air pollution. Six air pollutants, including PM_{10}, $PM_{2.5}$, SO_2, O_3, NO_2, and CO, are measured in 108 cities of America from 1987 to 2000. There has been extensive research on the adverse effect of $PM_{2.5}$ and NO_2 on health of human beings (see [2], [49] and [19]). [91] argued that $PM_{2.5}$ is one of the most concerning air pollutants which have an adverse impact on human health. Actually limited recordings of $PM_{2.5}$ were available before the 2000s in many regions ([79], [56], [93]). Thus it is crucial to understand the relationships between $PM_{2.5}$ and other regularly monitored air pollutants like PM_{10} and NO_2 ([56]). According to [17], there is a statistically significant association between the concentrations of $PM_{2.5}$ and NO_2 after removing autocorrelation. Furthermore, they also claim that this association varies from season to season. [56] showed that, for both ambient and kerbsite locations in Mumbai, there's a strong positive correlation between $PM_{2.5}$ and NO_2. It indicates that NO_2 and $PM_{2.5}$ share a common origin. A two-step method was proposed in [17]: a time series filter was employed to account for the temporal correlation within both NO_2 and $PM_{2.5}$ measurements, and then the Pearson correlation analysis was carried out on the residuals to estimate their associations. [56], however, conducted a regression analysis directly to estimate the correlations between these air pollutants. Neither methods treated repeated measurements of NO_2 and/or $PM_{2.5}$ as a sample of a smooth random function and considered the correlation between two random functions, even though the former one can allow for the temporal correlations. In this study, we are interested in investigating the association between these two air pollutants from a novel perspective: dynamical correlation, which is directly based

on the repeated measurements of them while accounting the temporal correlation within each of them.

Since from 1999 daily measurements for $PM_{2.5}$ are available in most cities ([79]), we choose measurements in 2000 for analysis for the consideration of fewer missing data. Cities without any measurement of $PM_{2.5}$ or NO_2 are excluded; then in total, we have measurements from 66 cities in the year of 2000. The left panel in Figure 3.2 displays the standardized profiles of $PM_{2.5}$ or NO_2 from one randomly selected city. These two air pollutants show a strong positive correlation in fall and winter and somewhat weak correlation in spring and summer according to this single city. Statistical inference based on the weighted empirical likelihood suggests that there's a moderate positive dynamical correlation between these two air pollutants if we aggregate the association in each season. In particular, the 95% confidence interval for the dynamical correlation is (0.38, 0.44). This conclusion does not contradict the mechanism how $PM_{2.5}$ is generated. On the one hand, gases like NO_2 and $PM_{2.5}$ may be emitted from the same source ([56]) and the former may serve as precursors to secondary $PM_{2.5}$ formation ([17]). On the other hand, correlations among these air pollutants exhibit strong seasonality. These may explain why there exists a moderate positive dynamical correlation between $PM_{2.5}$ and NO_2.

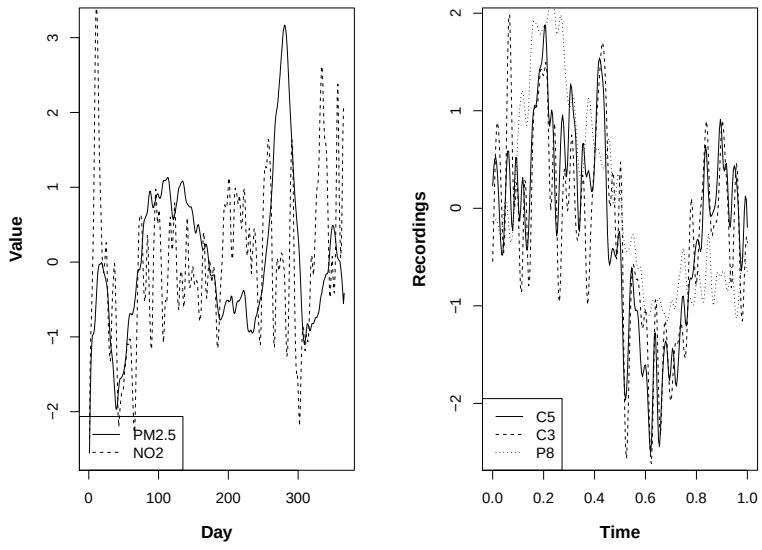


Figure 3.2: Left: Standardized profiles of $PM_{2.5}$ or NO_2 in 2000 from one randomly selected city. Right: Standardized profiles of three electrodes from one randomly selected subject.

3.5.2 Dynamical Correlation of EEG Signals

[120] studied the object recognition process by measuring electroencephalogram (EEG) signals using pictures of objects. In their experiment, 64 electrodes are placed on subject's

scalps to investigate EEG correlates of genetic predisposition to alcoholism. There are 122 subjects in this study, 77 of them in the alcoholic group and the remaining in the control group. Every subject is measured in 120 trials. In each trial, 256 observations were taken at each scalp in one second when a stimulus was shown for a subject. Each subject was exposed to a single stimulus or two stimuli. The two stimuli can be in a matched or non-matched condition. This dataset is available at the following link: `https://archive.ics.uci.edu/ml/datasets/eeg+database`.

In neuroimaging studies, there has been extensive research on functional connectivity, which refers to the statistical association or interactions between different brain regions ([74]). For instance, when studying a EEG dataset, [66] found that significant functional connectivity is identified between short-range sites (neighboring electrodes) considerably more often than between electrodes in long distances, when an individual is not performing an explicit task. Though using different measures to quantify functional connectivity, [83] drew a similar conclusion when applying the partial-correlation based framework to analyze the EEG data mentioned in the last paragraph. We're interested in studying functional connectivity from the perspective of dynamical correlation. To examine whether the dynamical correlation between short-range electrodes is stronger than that between long-range electrodes, we choose three electrodes: C5, C3, and P8. The first two electrodes are neighboring sites located in the central region, while the last electrode is in the temporal region and situated a long distance away from the first two. According to [68], the central region is in charge of functional integration of the tactile task while the temporal region is the integration centre for auditory attention.

To alleviate the effect of an individual's status (alcoholic or non-alcoholic) on functional connectivity between two electrodes, our study is focused on the alcoholic group. For each subject in the alcoholic group, we average recordings under the single stimulus condition across 120 trials at each observational time point. Our primary interest here is to explore dynamical correlation among the recordings of these three electrodes, which can serve as another measure of functional connectivity. Since P8 is situated a long distance away from both C5 and C3, which are neighboring electrodes, we suspect that the dynamical correlation between C5 and C3 are the most significant, compared with the other two counterparts. The right panel in Figure 3.2 depicts the standardized profile, defined in (3.3), of these three electrodes from one randomly selected subject. It can be observed that C5 and C3 maintain synchronization almost across the whole time course while disagreement appears between P8 and C5 or C3. With the empirical likelihood method, we find that the 95% confidence intervals for ρ_{12}, ρ_{13} and ρ_{23} are $(0.77, 0.86)$, $(0.36, 0.51)$ and $(0.29, 0.44)$, where ρ_{12}, ρ_{13} and ρ_{23} denote the dynamical correlation of C5 and C3, C5 and P8, C3 and P8, respectively. This conclusion is not surprising to us since electrodes must maintain synchronization to govern our brain to make a response to a stimulus. The electrodes closer to each other may be more consistent compared with electrodes far apart. Additionally, this result demonstrates

the hypothesis that functional connectivity between short-range electrodes is stronger than that between long-range electrodes from the perspective of dynamical correlation of random functions. A similar result is obtained from using the bootstrap method to compute confidence intervals: (0.77, 0.86), (0.35, 0.51) and (0.28, 0.44) for ρ_{12}, ρ_{13} and ρ_{23}, respectively. The consistency between these methods suggests that in the regular design, ignoring the uncertainty in the pre-smoothing step does not impose a notable impact on confidence intervals obtained from the weighted empirical likelihood method.

3.5.3 Dynamical Correlation of Gene Expressions

According to [111], the activation of T lymphocytes (T-cells) plays a key role in generation of an immune response. Two experiments regarding the response of a human T-cell line to phorbol myristyl acetate and ionomicin treatment were conducted by [88] to model the gene regulation network during T-cell activation. They collected the expression of 88 genes across 10 times points (0, 2, 4, 6, 8, 18, 24, 32, 48, 72 hr) and 34 replicates were obtained for each gene. This is a considerably less dense design compared with the EEG example. Only 58 genes were retained for final analysis after removing those genes whose expressions were lack of reproducibility between these two experiments.

[111] applied sparse additive ordinary differential equation models to identify a dynamic network among these genes. Both [88] and [111] found that FYB (gene 45) is one of the genes which regulate expressions of many other genes. Furthermore, [111] found that the gene FYB always has a positive regulation on gene 9 regardless of the expression level of gene 9. Dynamical correlation provides another perspective to look at the association of gene expressions during T-cell activation. We apply both the empirical likelihood method and the bootstrap method to compute the 95% confidence interval for the dynamical correlation of gene 9 and gene 45. The confidence interval given by the empirical likelihood method is (0.16, 0.52), while its counterpart is (0.10, 0.51). This finding is consistent with that in [111]. In this example, the random functions are considerably less densely observed, the confidence interval obtained from the bootstrap method is slightly wider than the that from the empirical likelihood method. The primary reason might be that the bootstrap method accounts for the uncertainty in smoothing each curve from only a few observations; this uncertainty has a non-negligible impact on the interval estimate of the dynamical correlation.

3.6 Conclusions and Discussion

This chapter proposes a new method based on the weighted empirical likelihood to infer the dynamical correlation between random functions. Compared with the bootstrap method by [25], the proposed method is considerably more efficient in computation due to the self-normalized property of the weighted empirical likelihood ratio statistic. Another reason

is that the bootstrap method spends substantial time on sampling both fitted values and residuals to account for uncertainty in the pre-smoothing step. Even though this source of uncertainty is ignored in the weighted empirical likelihood method, simulation studies demonstrate that the weighted empirical likelihood method is possible to yield a confidence interval with a more accurate coverage probability than that from the bootstrap method when the random functions of interest are irregularly spaced. But whether it is crucial to incorporate this source in the weighted empirical likelihood method in real datasets needs to carefully investigated in future, especially when functional data are irregularly observed. Furthermore, the proposed method accounts for the differing effect of each pair of random functions on the final estimator of dynamical correlation, rather than simply averaging these effects. Hence it compares favorably with the bootstrap method in statistical inference of dynamical correlation when there's a remarkable variation in the number of observations of functional data across subjects. In light of these advantages, it is worthwhile to explore applications of this methodology in other similar scenarios. But it should be noted that the proposed method cannot be directly applied to sparse functional data, in which only a few observations are available for every subject. For sparse functional data, we suggest to first recover the trajectory of each functional data from sparse observations. We will investigate this problem in our future work.

At first glance, dynamical correlation is not a powerful tool to reveal the dynamical feature of associations between random functions, since it summarizes correlations of two random functions over the domain of these functions with a single quantity. However, as pointed out by one referee, the dynamical feature can be discovered by adaptively choosing the weight function $w(t)$. In addition, to better reveal dynamical features of random functions and accommodate lagged relationships, the definition of dynamical correlation was extended to incorporate lag terms in [25]. Dynamical functional connectivity has attracted attention from numerous researchers in neuroimaging studies. It would be interesting work to compare dynamical correlation with varying weight functions with the framework of dynamical functional connectivity. Additionally, a closer look at time-varying correlations between random functions is also worthy of study, especially from the perspective of dynamical functional connectivity.

3.7 Appendix: Theoretical Proofs

3.7.1 Appendix A: Technical Assumptions

In this section, we lay out regularity conditions which can guarantee that the difference between the smoothed random function $f_{i,j}^S$ and the true underlying random function $f_{i,j}$ is negligible. Suppose the observations satisfy $Y_{i,j}(t_k) = f_{i,j}(t_k) + e_{i,j}(t_k)$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $k = 1, \ldots, m$. We assume that the random errors $e_{i,j}(t_k)$ are identically and independently distributed with $\mathrm{E}\{e_{i,j}(t_k)\} = 0$ and $\mathrm{E}\{e_{i,j}^2(t_k)\} = \sigma^2$. Let h denote the

bandwidth used in the local linear smoother in the pre-smoothing step. All limits below are taken given the number of subjects, n diverges.

Condition 1. Observational time points $\{t_k\}_{k=1}^m$ follow a density g, which is equi-continuous and bounded away from 0. Furthermore, the support of g is $\mathcal{I}$, which is compact.

Condition 2. The random errors satisfy $\mathrm{E}\{|e_{i,j}(t_k)|^s\} < \infty$ for some $s > 2$.

Condition 3. The bandwidth h and the number of observations satisfy $\underline{\lim}\, mh^2 > 0$ and $\underline{\lim} \left(\frac{mh}{\log m}\right)^{\frac{1}{2}} m^{-\frac{2}{s-\eta}} > 0$ for any η with $0 < \eta < 2$.

Condition 4. The random functions $f_{i,j}$ are twice differentiable. Furthermore, the second derivatives are equi-continuous satisfying $\sup |f''_{i,j}(t)| = O_p(1)$, where the supremum is taken over $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $t \in \mathcal{I}$.

Condition 5. The number of observations for each random function satisfies

$$m \to \infty, \quad n\left(\frac{\log m}{m}\right)^{\frac{4}{5}} \to 0.$$

Condition 6. The bandwidth h satisfies

$$\sqrt{n}h^2 \to 0, \quad n\left(\frac{\log m}{mh}\right)^{\frac{1}{2}} \to 0.$$

3.7.2 Appendix B: Proofs

When $\mu_{j_1}(t)$ and $\mu_{j_2}(t)$ are known or unknown but a constant, let $\tilde{f}^{\star}_{i,j_1}$ and $\tilde{f}^{\star}_{i,j_2}$ be the standardized curve for $f_{j_1}(t)$ and $f_{j_2}(t)$, respectively, for subject i. Then define

$$\rho_{i,j_1j_2} = \int \tilde{f}^{*}_{i,j_1}(t)\tilde{f}^{*}_{i,j_2}(t)w(t)\,\mathrm{d}t.$$

Thus $\rho_{1,j_1j_2}, \ldots, \rho_{n,j_1j_2}$, are identically and independently distributed with $E(\rho_{i,j_1j_2}) = \rho_{j_1j_2}$. $\mathrm{Var}(\rho_{i,j_1j_2}) = \delta^2 < \infty$. Furthermore, [25] showed $0 < \mathrm{Var}(\rho_{i,j_1j_2}) < 1$ given $0 < |\rho_{j_1j_2}| < 1$. In [77], the profile empirical likelihood function of θ based on ρ_{i,j_1j_2}'s is defined as

$$\tilde{L}(\theta) = \sup\{\prod_{i=1}^{n}(np_i) : p_i \geq 0, i = 1, \ldots, n;\ \sum_{i=1}^{n} p_i = 1;\ \sum_{i=1}^{n} p_i\rho_{i,j_1j_2} = \theta\} \tag{3.9}$$

As shown in Theorem 2.2 in [77], $-2\log\tilde{L}(\rho_{j_1j_2})$ converges in distribution to $\chi^2(1)$ as n diverges, if $0 < |\rho_{j_1j_2}| < 1$. We list some important steps to verify this result; for details, see [77].

The weights which maximize $\tilde{L}(\rho_{j_1j_2})$ can be written as

$$\tilde{w}_i = \frac{1}{n}\frac{1}{1 + \tilde{\lambda}(\rho_{i,j_1j_2} - \rho_{j_1j_2})},$$

where $\tilde{\lambda}$ satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho_{i,j_1j_2}-\rho_{j_1j_2}}{1+\lambda(\rho_{i,j_1j_2}-\rho_{j_1j_2})}=0$$

Hence,

$$|\tilde{\lambda}|=O_p(n^{-1/2});\qquad \max_{1\le i\le n}|\rho_{i,j_1j_2}-\rho_{j_1j_2}|=o_p(n^{1/2});\qquad |\bar{\rho}_{j_1j_2}-\rho_{j_1j_2}|=O_p(n^{-1/2}), \tag{3.10}$$

where $\bar{\rho}_{j_1j_2}=\frac{1}{n}\sum_{i=1}^{n}\rho_{i,j_1j_2}$.

With (3.2), for the smoothed random functions via a local linear smoother we can obtain the corresponding standardized curves, denoted by $\tilde{f}^S_{i,j_1}(t)$ and $\tilde{f}^S_{i,j_2}(t)$. Our main result concerns the asymptotic distribution of log-empirical likelihood function constructed based on the smoothed version of ρ_{i,j_1j_2}, denoted by ρ^S_{i,j_1j_2}, which is defined as

$$\rho^S_{i,j_1j_2}=\int \tilde{f}^S_{i,j_1}(t)\tilde{f}^S_{i,j_2}(t)w(t)\,\mathrm{d}t,\ i=1,\ldots,n.$$

Specifically, the corresponding profile empirical likelihood function of θ is

$$L(\theta)=\sup\{\prod_{i=1}^{n}(np_i):p_i\ge 0,i=1,\ldots,n;\ \sum_{i=1}^{n}p_i=1;\ \sum_{i=1}^{n}p_i\rho^S_{i,j_1j_2}=\theta\} \tag{3.11}$$

and

Let $\bar{\rho}^S_{j_1j_2}=\frac{1}{n}\sum_{i=1}^{n}\rho^S_{i,j_1j_2}$, $V^S_{j_1j_2}=\frac{1}{n}\sum_{i=1}^{n}(\rho^S_{i,j_1j_2}-\rho_{j_1j_2})^2$ and $V_{j_1j_2}=\frac{1}{n}\sum_{i=1}^{n}(\rho_{i,j_1j_2}-\rho_{j_1j_2})^2$. When $\mu_j(t)$ is constant or known for $j=j_1,j_2$, under the assumptions in Appendix A, from the proof of Theorem 2 in [25], we have

$$|\bar{\rho}^S_{j_1j_2}-\bar{\rho}_{j_1j_2}|=o_p(n^{-\frac{1}{2}}),\qquad \max_{1\le i\le n}|\rho^S_{i,j_1j_2}-\rho_{i,j_1j_2}|=o_p(n^{-1/2}) \tag{3.12}$$

The weights maximizing $L(\rho_{j_1j_2})$ can be written as

$$w_i=\frac{1}{n}\frac{1}{1+\lambda(\rho^S_{i,j_1j_2}-\rho_{j_1j_2})},$$

where $\lambda=\lambda(\rho_{j_1j_2})$ satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho^S_{i,j_1j_2}-\rho_{j_1j_2}}{1+\lambda(\rho^S_{i,j_1j_2}-\rho_{j_1j_2})}=0.$$

Let $M_n=\max_{1\le i\le n}|\rho^S_{i,j_1j_2}-\rho_{j_1j_2}|$. Based on (3.10) and (3.12), we obtain

$$|\bar{\rho}^S_{j_1j_2}-\rho_{j_1j_2}|=O_p(n^{-1/2}),\qquad M_n=o_p(n^{1/2}). \tag{3.13}$$

The next step is to verify that the order of λ is $O_p(n^{-1/2})$ as well. After some simple algebra, it can shown that

$$|\lambda| \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})^2}{1+\lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})} = sgn(\lambda)(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2}).$$

Note that $w_i > 0$, i.e., $1+\lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2}) > 0$. Thus

$$\begin{aligned} |\lambda| V^S_{j_1j_2} &\le |\lambda| \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})^2}{1+\lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})}(1+|\lambda| M_n) \\ &= sgn(\lambda)(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2})(1+|\lambda| M_n). \end{aligned}$$

Since $V_{j_1j_2} = \delta^2 + o_p(1)$, $V^S_{j_1j_2} - V_{j_1j_2} = o_p(1)$, and (3.13), it follows that

$$|\lambda|(\delta^2 + o_p(1)) = O_p(n^{-1/2}),$$

and hence,

$$|\lambda| = O_p(n^{-1/2}).$$

It follows that

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\rho^S_{i,j_1j_2} - \rho_{j_1j_2}}{1+\lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})} \\ &= \bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2} - \lambda V^S_{j_1j_2} + o_p(n^{-1/2}). \end{aligned}$$

Thus, $\lambda = (V^S_{j_1j_2})^{-1}(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2}) + o_p(n^{-1/2})$.

Plugging in the expression of w_i maximizing $L(\rho_{j_1j_2})$, we have

$$\begin{aligned} -2\log L(\rho_{j_1j_2}) &= 2\sum_{i=1}^{n} \log\{1+\lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})\} \\ &= 2\left\{\sum_{i=1}^{n} \lambda(\rho^S_{i,j_1j_2} - \rho_{j_1j_2}) - \frac{\lambda^2(\rho^S_{i,j_1j_2} - \rho_{j_1j_2})^2}{2}\right\} + o_p(1) \\ &= 2n(V^S_{j_1j_2})^{-1}(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2})^2 - n(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2})^2 (V^S)^{-1}_{j_1j_2} + o_p(1) \\ &= n(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2})^2 (V^S_{j_1j_2})^{-1} + o_p(1) \end{aligned} \tag{3.14}$$

As shown in [77],

$$-2\log \tilde{L}(\rho_{j_1j_2}) = \frac{n(\bar{\rho}_{j_1j_2} - \rho_{j_1j_2})^2}{V_{j_1j_2}} + o_p(1). \tag{3.15}$$

Based on (3.12), (3.14) and (3.15), it follows that

$$\begin{aligned} -2\log L(\rho_{j_1j_2}) &= n(\bar{\rho}^S_{j_1j_2} - \rho_{j_1j_2})^2 (V^S_{j_1j_2})^{-1} + o_p(1) \\ &= -2\log \tilde{L}(\rho_{j_1j_2}) + o_p(1) \\ &\to \chi^2(1) \qquad \text{in distribution.} \end{aligned}$$

□

Chapter 4

Sparse Estimation for Functional Semiparametric Additive Models

4.1 Introduction

High dimensional data sets of large volume and complex structure are rapidly emerging in various fields. Functional data analysis, due to its great flexibility and wide applications in dealing with high dimensional data, has received considerable attention. One important problem in functional data analysis is functional linear regression (FLR). One type of FLR models the relationship between a functional covariate and a univariate scalar response of interest. Due to potential lack of fit with FLR models, [72] proposed functional additive models (FAM), in which a scalar response depends on an additive form of the functional principal component (FPC) scores of a functional covariate. A local linear smoother was employed to estimate each component in the additive form and consistency was established for this estimator.

However, in many cases, not only functional covariates but also some scalar covariates may play a role in explaining the variation of response. For instance, the Tecator dataset (see Section 4.4.1 for a more detailed description), which consists of three contents (fat, water, and protein) and 100-channel spectral trajectories of absorbance, has been analyzed with various models, where the response of interest is one of the three contents. Previous studies have focused on regressing the response on the spectral trajectories, which can be viewed as a functional covariate. [121], for example, employed a regularized functional additive model, where scaled FPC scores are treated as covariates to predict the protein content. However, pairwise scatter plots of the three contents suggest that the other two contents are highly correlated with the protein content as well; thus it may be beneficial to add them into the regression model. In light of this fact, we aim to build a model which can incorporate the effects of both the spectral trajectories and the fat and water contents on the prediction of the protein content.

Motivated by the above example, we propose a functional semiparametric additive model (FSAM) to describe the relationship between a functional covariate, a finite number of scalar covariates and a response variable of interest. In this model, the effect of a functional covariate is represented by its scaled leading FPC scores while scalar covariates are modeled linearly. As a result, this model enables us to acquire flexibility in calibrating the effect of the functional covariate while retaining easy interpretation of the effects of the scalar covariates. There are two main difficulties associated with this new model: the first one is the model estimation and the second concerns theoretical properties. Obviously, the estimation of the effect of a functional covariate may affect that of scalar covariates and vice versa. To address this issue, we propose an iterative updating algorithm, which is similar in spirit to the EM algorithm, to account for the interdependence between these two estimated effects. In addition, only the nonparametric effect of the functional covariate needs to be regularized; this adds additional difficulties in estimation. In the theoretical aspect, we aim to establish consistency for the parametric part and the nonparametric part, respectively. Separating these two effects is more difficult than developing theoretical properties with only a nonparametric part as in a FAM.

A semiparametric additive model (sometimes described under alternative names like partially linear model) can be viewed as a special version of a generalized additive model in which the mean response is assumed to have a linear relationship with one or more of the covariates, but the relation with other covariates cannot be easily modelled in a parametric form ([97]; [92]). Numerous methods have been proposed to fit such models. The method of penalized least squares ([105]; [27]; [44]) has played a major role in this regard. [14] employed a piecewise polynomial to approximate the nonparametric part and developed asymptotic properties of the least squares estimator of the coefficients in the parametric part. [28] estimated the nonparametric part using a local polynomial and derived asymptotic properties of their estimators as well. A comprehensive review of different approaches to fitting a semiparametric additive model can be seen in [41].

For the case when both a functional covariate and scalar covariates are involved to predict the mean response, [95] considered a functional partially linear model in which the effect of a functional covariate is modeled via a finite-dimensional linear combination of principal component scores. A similar model was proposed by [64] to model the quantile function of the response variable. Even though both papers derived asymptotic properties of their estimators, they did not consider selection of functional principal components for the functional covariate. [55] extended the above work to the situation when multiple functional covariates and high-dimensional scalar covariates are encountered. The effect of each functional covariate is represented via a truncated linear combination of FPC scores, the truncation level of which is allowed to increase as sample size increases. To identify important features, reduce variability and enhance interpretability, they proposed to combine regularization of each functional covariate with a penalty on high-dimensional scalar covari-

ates. [50] proposed a general regression framework which considered a functional response and two functional covariates.

This chapter has three main contributions. First, in comparison with previous work on functional partial linear regression, our model allows for a more general representation in terms of the effect of a functional covariate. In addition, using a special regularization scheme, our method can select the non-vanishing functional principal components for a functional covariate. Last but not least, we derive asymptotic properties of the estimator.

The remainder of this chapter is organized as follows. Section 4.2 introduces FSAM and our method to estimate FSAM using a special regularization scheme and implementing an iterative updating algorithm. Section 4.3 evaluates the finite sample performance of our proposed estimation method in comparison with three alternative methods using some simulation studies. In Section 4.4, our method is demonstrated in two real examples. Some asymptotic results for the proposed estimation method are provided in Section 4.5. Section 4.6 concludes this chapter. Theoretical proofs are given in the Appendix.

4.2 Model and Estimation Method

4.2.1 Functional semiparametric additive model

Let $X(t)$ denote a square integrable stochastic process on a domain $\mathcal{I} = [0, T]$ and Y denote a scalar random variable. A functional regression model characterizes the relationship between the scalar response Y and the random function $X(t)$. A typical example is the functional linear model: $Y = \int_{\mathcal{I}} X(t)\beta(t)\,\mathrm{d}t$, where $\beta(t)$ is a square integrable function on $[0, T]$ as well.

To account for the effect of some scalar predictors in a functional regression model, several functional partial linear models have been proposed (see [95]; [64]; and [55]). In their work, the effect of the functional predictor is modelled nonparametrically while a linear form is adopted to model the effect of scalar predictors. For instance, [95] considered the following model:

$$E(Y|X, \boldsymbol{z}) = \int_{\mathcal{I}} X(t)\beta(t)\,\mathrm{d}t + \boldsymbol{z}^\top \boldsymbol{\alpha}, \tag{4.1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$ is a p-dimensional scalar covariate and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top \in \mathbb{R}^p$ is the corresponding coefficient vector.

Let $m(t)$ and $G(s, t)$ denote the mean function and covariance function of $X(t)$, respectively. The covariance function $G(s, t)$ can be expressed as $G(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$, where $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenvalues of G, satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\{\psi_k(t)\}_{k=1}^{\infty}$ are the corresponding orthonormal eigenfunctions, which satisfy $\int \psi_j \psi_k \,\mathrm{d}t = 1$ if $j = k$ and 0 otherwise. Then the process $X(t)$ admits the Karhunen-Loève expansion:

$$X(t) = m(t) + \sum_{k=1}^{\infty} \xi_k \psi_k(t), \tag{4.2}$$

where $\xi_k = \int_{\mathcal{I}}(X(t) - m(t))\psi_k(t)dt$ is the uncorrelated functional principal component (FPC) score. In addition, $E(\xi_k\xi_{k'}) = \lambda_k$ if $k = k'$ and 0 otherwise. Replacing $X(t)$ in (4.1) with the expression given in (4.2), we have

$$E(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty}\xi_k b_k + \boldsymbol{z}^\top\boldsymbol{\alpha},$$

where $b = \int_{\mathcal{I}}\beta(t)m(t)\,\mathrm{d}t$ and $b_k = \int_{\mathcal{I}}\beta(t)\psi_k(t)\,\mathrm{d}t$.

To allow for greater flexibility, the additive components with respect to the FPC scores $\{\xi_k\}_{k=1}^{\infty}$ in the above equation can take a more general form. Motivated by the idea of a generalized additive model ([42]), we consider

$$E(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty}f_k(\xi_k) + \boldsymbol{z}^\top\boldsymbol{\alpha}.$$

This model without scalar predictors is previously studied in [72] and [121] to describe the relationship between a scalar response and a functional predictor .

For convenience of regularization on each component f_k, we first scale the FPC scores to [0, 1]. One possible approach is to treat ξ_k as having a $N(0, \lambda_k)$ distribution and apply the cumulative distribution function (cdf) of $N(0, \lambda_k)$ to ξ_k, i.e., $\zeta_k = \Phi(\xi_k/\sqrt{\lambda_k})$, where Φ is the cdf of $N(0, 1)$. Other cumulative distributions could be employed for scaling, but we focus solely the Gaussian case here. The corresponding additive model becomes:

$$E(Y|X, \boldsymbol{z}) = b + \sum_{k=1}^{\infty}f_k(\zeta_k) + \boldsymbol{z}^\top\boldsymbol{\alpha}. \tag{4.3}$$

In addition to making the following regularization scheme easier to implement, there are two main reasons why we consider transferring FPC scores to a compact domain. The first reason concerns theoretical derivations. When each function in a functional space can be represented in terms of spline basis functions like B-spline bases or reproducing kernel functions, assuming the domain is compact can simplify theoretical derivations. Such examples can found in [65] and [98], etc. The second reason explains why this treatment is reasonable. Let h_j denote the transformation: $h(\xi_j) = \zeta_j$ and g_j denote the function with the argument being ζ_j. Note that if h_j is any strictly monotone, continuous map from $\mathbb{R}$ to $(0, 1)$, we may write $f_j = g_j \circ h_j$ with $g_j = f_j \circ h_j^{-1}$.

We assume that there exists an integer, d, which is large enough that $f_k \equiv 0$ when $k > d$. This amounts to assuming that only some of the FPC scores of the functional predictor are relevant to the response. Our truncated model is then given as:

$$E(Y|X, \boldsymbol{z}) = b + \sum_{j=1}^{d}f_j(\zeta_j) + \boldsymbol{z}^\top\boldsymbol{\alpha}. \tag{4.4}$$

In practice we choose d initials in such a way that at least 99.9% of variability in $X(t)$ can be explained by the first d FPCs. Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_d)^\top$.

We also assume that the effect of each transformed score, f_j, $j = 1, \ldots, d$, is a smooth function. In this chapter, we call the effect of $X(t)$, $f(\boldsymbol{\zeta}) = b + \sum_{j=1}^{d} f_j(\zeta_j)$, the nonparametric part of the model and the linear combination $\boldsymbol{z}^\top \boldsymbol{\alpha}$ the parametric part of the model. In addition, f_j's are called nonparametric components. Model (4.4) is called a functional semiparametric additive model (FSAM) in this chapter.

4.2.2 Estimation method

The objective is to propose an estimation method which can select and estimate nonparametric components that are relevant to the response while estimating the effects of scalar covariates in Model (4.4).

[121] considered a special case when $\boldsymbol{\alpha}$ is known to be $\mathbf{0}$ in Model (4.4); to select and smooth non-vanishing nonparametric components in the estimation of the nonparametric part, they apply the COmponent Selection and Smoothing Operator (COSSO) proposed by [60]. We provide a brief review of COSSO next. Let H be the lth-order Sobolev space on $[0, 1]$, defined by

$$H([0,1]) = \{h | h^{(\nu)} \text{ is absolutely continuous for } \nu = 0, 1, \ldots, l-1; h^{(l)} \in L^2\}.$$

Then H is a reproducing kernel Hilbert space (RKHS) equipped with the squared norm

$$||h||^2 = \sum_{\nu=0}^{l-1} \left\{ \int_0^1 h^{(\nu)}(t)\, \mathrm{d}t \right\}^2 + \int_0^1 \{h^{(l)}(t)\}^2\, \mathrm{d}t. \tag{4.5}$$

For a more detailed introduction to this RKHS, one can refer to Chapter 2.3 in [40]. H can be decomposed as $H = \{1\} \oplus \bar{H}$, where elements of $\bar{H}$ have been centered. For example, take $h(t) = t$. Then $h(t) = 1/2 + t - 1/2$ and $t - 1/2 \in \bar{H}$. Assuming that $f_j \in \bar{H},\ j = 1, \ldots, d$, then $f(\boldsymbol{\zeta})$ lies in the subspace $\mathscr{F}^d = \{1\} \oplus \sum_{j=1}^{d} \bar{H}$. This assumption addresses the issue of identifiability for nonparametric components in Model (4.4). The COSSO regularization, applied to functions in the RKHS, is used to select and smooth non-vanishing components when estimating f. Suppose the data consists of n independent and identically distributed triples $\{X_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$. When $\boldsymbol{\alpha}$ in Model (4.4) are known to be $\mathbf{0}$, then the COSSO estimate of f is defined by minimizing

$$Q(f) = \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{\zeta}_i)\}^2 + \tau^2 J(f), \tag{4.6}$$

where $J(f) = \sum_{j=1}^{d} ||P^j f||$ with each $P^j f$ denoting the projection of f onto $\bar{H}$ with the argument being the jth component of $\boldsymbol{\zeta}$, and τ denotes a tuning parameter which controls

the trade-off between fidelity to the data and complexity of the model. Another way to look at $P^j f$ is $P^j f = f_j - \int f_j$.

If $P^j f$ is linear for $1 \leq j \leq d$, then the minimizer of $Q(f)$ is the LASSO estimate. However, the sum of the seminorms of $P^j f$, that is $J(f)$, rather than the L_1 norm of the coefficient vector, is penalized in $Q(f)$ in general scenarios. More specifically, if we represent each $P^j f$ with a linear combination of the reproducing kernel functions, $||P^j f||$'s are not differentiable with respect to the coefficients. This fact makes minimization of $Q(f)$ an intricate problem. [60] argued that introducing an ancillary parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$, can ease the minimization task greatly. As shown in Lemma 2 of [60], minimization of (4.6) is equivalent to minimizing

$$H(f, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{\zeta}_i)\}^2 + \lambda_0 \sum_{j=1}^{d} \theta_j^{-1} ||P^j f||^2 + \lambda \sum_{j=1}^{d} \theta_j \tag{4.7}$$

with respect to f and $\boldsymbol{\theta}$, when $f \in \mathscr{F}^d, \theta_j \geq 0,\ j = 1, \ldots, d$ and $\lambda = \tau^4/(4\lambda_0)$. In (4.7), both λ_0 and λ are nonnegative tuning parameters, which control the smoothness and selection of the estimated nonparametric part, respectively. If $\theta_j = 0$, then the minimizer satisfies $||P^j f|| = 0$, indicating that f_j, the jth component in the nonparametric part, vanishes. The outline of the algorithm is given as follows.

Generally speaking, the FPC scores cannot be observed directly; thus estimating the first d FPC scores for each trajectory X_i is indispensable to estimate the nonparametric part later. Trajectories are usually recorded at a grid of time points, which can be different across subjects, and they are often observed with measurement errors. To address these issues when estimating FPC scores, we can employ regularized FPCA, proposed by [86], or PACE, proposed by [116]. Then the estimated scaled FPC scores, denoted as $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})$, can be obtained by applying the CDF of a normal distribution with specific variance to the estimated FPC scores. Now we can implement COSSO. Let R_j denote the $n \times n$ matrix with the $(s,\ t)$ entry $R(\hat{\zeta}_{sj},\ \hat{\zeta}_{tj})$, where $R(\cdot, \cdot)$ is the reproducing kernel of $\bar{H}$, and $R_{\boldsymbol{\theta}}$ for the matrix $\sum_{j=1}^{d} \theta_j R_j$. For fixed λ_0 and λ, the minimizer of (4.7) has the form $f(\boldsymbol{\zeta}) = b + \sum_{i=1}^{n} c_i \sum_{j=1}^{d} \theta_j R(\zeta_j, \hat{\zeta}_{ij})$. Thus $\boldsymbol{f} = (f(\hat{\boldsymbol{\zeta}}_1), \ldots, f(\hat{\boldsymbol{\zeta}}_n))^\top = \mathbf{1}_n b + R_{\boldsymbol{\theta}} \boldsymbol{c}$, where $\boldsymbol{c} = (c_1, \ldots, c_n)^\top$ and $\mathbf{1}_n$ is the vector of ones of length n. Then the penalty term has $\sum_{j=1}^{d} \theta_j^{-1} ||P^j f||^2 = \sum_{j=1}^{d} \theta_j \boldsymbol{c}^\top R_j \boldsymbol{c} = \boldsymbol{c}^\top R_{\boldsymbol{\theta}} \boldsymbol{c}$. Now (4.7) becomes

$$\min_{b, \boldsymbol{c}, \boldsymbol{\theta} \geq \mathbf{0}_d} \frac{1}{n}(\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}} \boldsymbol{c})^\top (\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}} \boldsymbol{c}) + \lambda_0 \boldsymbol{c}^T R_{\boldsymbol{\theta}} \boldsymbol{c} + \lambda \mathbf{1}_d^\top \boldsymbol{\theta}, \tag{4.8}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{0}_d$ denotes the vector consisting of d zeros.

To solve (4.8), we alternatively solve for the pair $(b, \boldsymbol{c})$ with $\boldsymbol{\theta}$ fixed and then solve for $\boldsymbol{\theta}$ with $(b, \boldsymbol{c})$ fixed. More specifically,

1. When $\boldsymbol{\theta}$ is fixed, solving (4.8) is equivalent to solving the standard smoothing spline

$$\min_{b,\boldsymbol{c}} ||\boldsymbol{y} - \mathbf{1}_n b - R_{\boldsymbol{\theta}}\boldsymbol{c}||^2 + n\lambda_0 \boldsymbol{c}^T R_{\boldsymbol{\theta}} \boldsymbol{c}. \tag{4.9}$$

   The solution of (4.9) is similar to a smoothing spline estimate and can be found in [106].

2. With $(b, \boldsymbol{c})$ fixed, (4.8) becomes

$$\min_{\boldsymbol{\theta}\geq \mathbf{0}} (\boldsymbol{v} - G\boldsymbol{\theta})^T (\boldsymbol{v} - G\boldsymbol{\theta}) + n\lambda \mathbf{1}_S \boldsymbol{\theta}, \tag{4.10}$$

   where $\boldsymbol{v} = \boldsymbol{y} - (1/2)n\lambda_0 \boldsymbol{c} - \mathbf{1}_n b$ and G is the $n \times d$ matrix with the jth column being $R_j \boldsymbol{c}$. [60] suggested considering an equivalent optimization problem: for some $M \geq 0$, find

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{v} - G\boldsymbol{\theta}||^2, \text{ subject to } \mathbf{1}^T\boldsymbol{\theta} \leq M, \boldsymbol{\theta} \geq \mathbf{0}_d. \tag{4.11}$$

   The tuning parameter M in (4.11) is equivalent to λ in (4.10). Alternatively, the optimization problem (4.10) can be addressed directly using **glmnet** in **R** with the lower bound of the parameters to be estimated set as 0.

Only when the effect of the scalar covariates $\boldsymbol{z}$ can be removed or is known, can the above algorithm be implemented. Now we take the unknown effect of $\boldsymbol{z}$ into consideration as well. The estimate of $g = f(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top \boldsymbol{z}$ is defined as

$$\hat{g}_n \in \underset{g(\boldsymbol{\zeta},\boldsymbol{z}) = b + \sum_{j=1}^d f_j(\zeta_j) + \boldsymbol{\alpha}^\top \boldsymbol{z}}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^n \{y_i - g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\}^2 + \tau^2 J(g) \right],$$

where $f_j \in \bar{H}$ and $J(g)$ is set to be $J(f) = \sum_{j=1}^d ||P^j f||$. Note that the regularization suggested above penalizes only the nonparametric part, while neglecting the effect of the parametric part $\boldsymbol{\alpha}^\top \boldsymbol{z}$. Difficulties arise when we apply COSSO directly to estimate the nonparametric part f in (4.4) since the effect of scalar predictor $\boldsymbol{z}$ needs to be accounted for as well. If the coefficient, $\boldsymbol{\alpha}$, of $\boldsymbol{z}$ were known, then a slight modification of COSSO would suffice to deal with the estimation problem: replace $\boldsymbol{y}$ in (4.9) with $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha}$ and $\boldsymbol{v}$ in (4.10) with $\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - (1/2)n\lambda_0 \boldsymbol{c} - \mathbf{1}_n b$, where $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$. However, $\boldsymbol{\alpha}$ is unknown as well and the estimate of the nonparametric part f depends on the value of $\boldsymbol{\alpha}$, which poses a bottleneck when implementing COSSO to estimate f. To deal with this problem, we propose an iterative updating algorithm to estimate both the nonparametric and parametric parts and to select and smooth the non-vanishing components in f.

After estimating $\boldsymbol{\zeta}_i$'s via regularized FPCA or PACE, the target function to be minimized can be written as:

$$Q^\star(f, \boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\hat{\boldsymbol{\zeta}}_i) - \boldsymbol{\alpha}^\top \boldsymbol{z}_i\}^2 + \tau^2 \sum_{j=1}^{d} ||P^j f||,$$

where $f(\hat{\boldsymbol{\zeta}}_i) = b + \sum_{j=1}^{d} f_j(\hat{\zeta}_{ij})$ denotes the the nonparametric part evaluated at the estimated transformation $\hat{\boldsymbol{\zeta}}_i$. We aim to look for $\hat{f} \in \mathscr{F}^d$ and $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$, which can minimize the target function $Q^\star$. As illustrated above, minimizing $Q^\star$ is equivalent to another minimization problem:

$$\min_{\boldsymbol{\alpha}, b, \boldsymbol{c}, \boldsymbol{\theta} \geq \mathbf{0}_d} \frac{1}{n}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - \mathbf{1}_n b - R_{\boldsymbol{\theta}}\boldsymbol{c})^\top (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - \mathbf{1}_n b - R_{\boldsymbol{\theta}}\boldsymbol{c}) + \lambda_0 \boldsymbol{c}^T R_{\boldsymbol{\theta}} \boldsymbol{c} + \lambda \mathbf{1}_d^\top \boldsymbol{\theta}. \quad (4.12)$$

Algorithm 3 outlines the steps to solve (4.12).

Algorithm 3 : **Iterative updating for regularized functional semiparametric additive model**

Step 1: Start with an initial value of $\boldsymbol{\alpha}$, say $\hat{\boldsymbol{\alpha}}^{(0)}$, and an initial value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}^{(0)}$.

Step 2: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$ and $\hat{\boldsymbol{\theta}}^{(m)}$ to obtain estimates $\hat{b}^{(m+1)}$ and $\hat{\boldsymbol{c}}^{(m+1)}$ by solving (4.9), in which $\boldsymbol{y}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)}$.

Step 3: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$, $\hat{b}^{(m+1)}$ and $\hat{\boldsymbol{c}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\theta}}^{(m+1)}$ by solving (4.11), in which $\boldsymbol{v}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)} - (1/2)n\lambda_0 \hat{\boldsymbol{c}}^{(m+1)} - \mathbf{1}_n \hat{b}^{(m+1)}$.

Step 4: Use the estimate $\hat{b}^{(m+1)}$, $\hat{\boldsymbol{c}}^{(m+1)}$ and $\hat{\boldsymbol{\theta}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\alpha}}^{(m+1)}$ by solving a least squares problem.

Step 5: Repeat steps 2, 3 and 4 until $||\hat{\boldsymbol{\alpha}}^{(m+1)} - \hat{\boldsymbol{\alpha}}^{(m)}|| < \epsilon$, where ϵ is a pre-determined tolerance value.

The fitting method presented above is called Functional Semiparametric Additive Model via COmponent Selection and Smoothing Operator (FSAM-COSSO) in this chapter. Minimization of (4.12) turns out to be a convex problem. Our numerical studies show that this algorithm can converge in a few steps with reasonable initial estimates for both $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$; these are taken to be the ordinary least squares estimate and $\mathbf{0}_d$, respectively.

4.2.3 Tuning Parameter Selection

Cross-validation (CV) or generalized cross-validation (GCV) can be employed to choose the tuning parameters. The following adaptive tuning scheme is a slight modification of the proposal by [60]:

1. In step 1 of Algorithm 3, the initial value of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}^{(0)}$ is chosen as $\mathbf{1}_d$. We employ GCV or CV to choose the tuning parameter λ_0 when addressing the smoothing spline problem. In the following updating steps, λ_0 is fixed to be the chosen value.

2. A grid of points in a reasonable range are chosen as candidates for M. CV is employed to choose the "optimal" value of M. More specifically, the whole data set is randomly split into G folds. The optimal M is chosen as the value which can minimize

$$CV(M) = \frac{1}{n}\sum_{g=1}^{G}(\hat{\boldsymbol{y}}_g^{(-g)} - \boldsymbol{y}_g)^\top(\hat{\boldsymbol{y}}_g^{(-g)} - \boldsymbol{y}_g),$$

where $\hat{\boldsymbol{y}}_g^{(-g)}$ denotes the predicted values for the gth fold of the data when it is removed and the model is fitted using the other $G-1$ folds of the data.

Remark: In this case, GCV may not be an appropriate approach for selecting M. The primary reason is that we need a well-defined hat (smoothing) matrix to apply GCV to select tuning parameters. However, when shrinkage methods such as LASSO are employed, this matrix may not be well-defined, particularly in high-dimension problems.

4.3 Simulation Studies

In this section, two simulation studies are conducted to evaluate the finite sample performance of our proposed approach and compare it with other alternative methods.

4.3.1 FSAM With Scalar Covariates

A functional covariate is generated from the first 20 Fourier basis functions with eigenvalues $\lambda_k = ab^k,\ k = 1, \ldots, 20$; we take $a = 31.6$ and $b = 0.5$. More specifically,

$$X(t) = \mu(t) + \sum_{k=1}^{20}\xi_k\psi_k(t) + e(t),$$

where $\mu(t) = t + \sin(t)$ denotes the mean function of $X(t)$, $\xi_k \sim N(0, \lambda_k)$, ψ_k's denote the Fourier basis functions and the measurement error $e(t)$ follows $N(0, 0.01)$, independently of all ξ_k's. We generate $n = 1000$ independent curves in total; each curve is sampled at 200 equally spaced points between 0 and 10. The corresponding scaled FPC scores, ζ_{ik}'s, are defined as $\zeta_{ik} = \Phi(\xi_{ik}/\sqrt{\lambda_k}),\ i = 1, \ldots, n,\ k = 1, \ldots, 20$. Then the response variable y is generated from the following model

$$y_i = 1.2 + f_1(\zeta_{i1}) + f_2(\zeta_{i2}) + f_4(\zeta_{i4}) + \boldsymbol{z}_i^\top\boldsymbol{\alpha}_0 + \epsilon_i,\ i = 1, \ldots, n.$$

In this model,

- $f_1(x) = xe^x - 1,\ f_2(x) = \cos(2\pi x)$ and $f_4(x) = 3(x - \frac{1}{4})^2 - \frac{7}{16}$. They have a common domain [0, 1]. The nonparametric part is $f(\boldsymbol{\zeta}) = 1.2 + f_1(\zeta_1) + f_2(\zeta_2) + f_4(\zeta_4)$, where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{21})^\top$; in other words, the non-vanishing nonparametric components are f_1, f_2 and f_4.

- $\boldsymbol{z}_i = (z_{i1}, z_{i2})^\top$ is independent of $X_i(t)$; the two components of $\boldsymbol{z}_i$ are independently generated from Uniform[0, 1] distribution. $\boldsymbol{\alpha}_0 = (-1, 2)^\top$;

- ϵ_i is independent of both $X_i(t)$ and $\boldsymbol{z}_i$ and is generated from $N(0, 1)$.

The signal-to-noise ratio, defined as $\text{var}(f(\boldsymbol{\zeta}))/\text{var}(\epsilon)$, is around 1.75 under this setup.

Among the 1000 data points $\{X_i(t), \boldsymbol{z}_i, y_i\}_{i=1}^{1000}$, 200 are randomly selected as the training set and the remaining 800 data points are treated as the test set. We used PACE to estimate FPC scores and then choose so that the first d fitted scores explain 99.9% of the variability in sample curves of the variability of the training set. We find d is around 20 in all simulation replicates. Let $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})^\top$ denote the estimate of $\boldsymbol{\zeta}_i$. Then different methods are fitted to the triple $(\hat{\boldsymbol{\zeta}}_i, \boldsymbol{z}_i, y_i)$, where $i \in$ training set. The proposed method in this chapter, FSAM-COSSO, is implemented to fit Model (4.4) to estimate and select non-vanishing nonparametric components as well as estimating the coefficient vector of the scalar covariate $\boldsymbol{z}$. In simulation studies and real data applications presented in Section 4.4, we take the order of the Sobolev space to be $l = 2$. But the algorithm proposed below can be extended to more general cases.

MARS ([36]) fits an additive model for $(\hat{\boldsymbol{\zeta}}_i, y_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_0)$, assuming that the coefficients in the parametric part are known to be $(-1, 2)^\top$ and $y_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_0$ is the new response. As a comparison, two types of extended FAMs are considered as well. The FSAM-GAMS model denotes a saturated model in which $(\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{id})^\top$ are fitted by a generalized additive model (GAM) while z_{i1}, z_{i2} are fitted in a linear form. FSAM-COSSO differs from FSAM-GAMS in that the latter does not take component selection into consideration. In the second extended FAM, assuming that ζ_1, ζ_2 and ζ_4 are known to be the only non-vanishing features and the expressions of f_1, f_2, f_4 are known as well, a multiple linear regression is fitted on $\{f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4}), \boldsymbol{z}_i, y_i\}$, in which y_i denotes the response and the explanatory variables consist of $f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4}), z_{i1}, z_{i2}$; this model is called FSAM-GAM1 in this chapter. The FSAM-PFLR model employs a linear combination of $\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{im}$, where m denotes the number of retained FPCs, to represent the effect of $X_i(t)$ on y_i. It is a modified version of the partial functional linear regression proposed by [95], where the effect of a functional predictor is represented by a linear combination of the original FPC scores. The tuning parameter m is chosen based on AIC, as suggested in [95]. To investigate the effect of using $\hat{\boldsymbol{\zeta}}_i$'s on estimation of each f_j, we also implement the proposed method with true scores, $\boldsymbol{\zeta}_i$'s, to fit the model. This method is denoted by FSAM-COSSO1 in the chapter.

To assess the performance of the above methods, 1000 simulation replicates are conducted to estimate the mean squared predicted errors (MSPE) on the test set, which is

defined as $\sum_i (y_i - \hat{y}_i)^2/n_0$ with n_0 equal to the size of the test set. Besides prediction accuracy, we can also compare the performance of the methods from the perspective of model fitting; particularly, the number of selected nonparametric components, the frequency with which each $\hat{\zeta}_k$ is selected, and the bias and standard error (SE) of the estimates of $\boldsymbol{\alpha}$ are reported for each method as well.

Model	Counts with the model size												
	1	2	3	4	5	6	7	8	9	10	11	12	13
MARS	0	0	20	71	140	170	197	181	127	69	16	6	3
FSAM-GAMS	0	0	169	261	238	161	98	42	19	9	2	1	0
FSAM-PFLR	5	533	312	98	40	8	4	0	0	0	0	0	0
FSAM-COSSO	0	5	593	202	113	54	26	7	0	0	0	0	0
FSAM-COSSO1	0	1	709	149	73	39	22	6	1	0	0	0	0

Table 4.1: Summary of the number of selected nonparametric components over the 1000 simulations for each model. Model size indicates the number of nonparametric components selected in the model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). Here we implement the function **gam** in the R package **mgcv** to fit FSAM-GAMS. The corresponding p-values of nonparametric components are available from the function **summary.gam**. This selection rule applies to FSAM-PFLR as well, where the p-value is available from the function **lm**.

Table 4.1 and Table 4.2 summarize the number and frequency of nonparametric components selected in each method over the 1000 simulations, respectively. FSAM-COSSO in most cases selects the correct number of nonparametric components. In contrast, FSAM-GAMS and MARS are prone to retain some irrelevant nonparametric components, which results in more complex models and hence greater variance. Since AIC is employed to select the number of retained FPCs, FSAM-PFLR tends to yield a model with a relatively small size. As a result, even though it is less likely for irrelevant features to be selected, FSAM-PFLR suffers from frequently ignoring relevant features. Furthermore, FSAM-COSSO not only selects relevant factors $(\zeta_1, \zeta_2, \zeta_4)$ in almost every simulation but also retains irrelevant features considerably less often compared with MARS and FSAM-GAMS. The similarity between FSAM-COSSO and FSAM-COSSO1 suggests that replacing the true scores with estimates would make little difference in component selection. Table 4.1 and Table 4.2 therefore demonstrate that FSAM-COSSO enables us to better discover the nonparametric relationship between the functional covariate $X(t)$ and the response when a model is given in the form of (4.4).

Model	Frequency of each nonparametric factor									
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$	$\hat{f}_6$	$\hat{f}_7$	$\hat{f}_8$	$\hat{f}_9$	$\hat{f}_{10}$
MARS	1000	1000	356	1000	274	226	233	235	211	244
FSAM-GAMS	1000	1000	232	997	155	100	120	106	103	114
FSAM-PFLR	1000	981	231	1000	154	100	121	98	92	83
FSAM-COSSO	1000	1000	100	993	59	30	36	22	30	31
FSAM-COSSO1	1000	1000	34	999	26	29	40	32	32	25
	$\hat{f}_{11}$	$\hat{f}_{12}$	$\hat{f}_{13}$	$\hat{f}_{14}$	$\hat{f}_{15}$	$\hat{f}_{16}$	$\hat{f}_{17}$	$\hat{f}_{18}$	$\hat{f}_{19}$	$\hat{f}_{20}$
MARS	210	208	196	226	222	240	236	228	217	249
FSAM-GAMS	105	108	105	106	96	101	126	111	114	125
FSAM-PFLR	63	75	65	84	70	60	63	59	18	42
FSAM-COSSO	31	39	38	36	29	40	50	48	57	55
FSAM-COSSO1	31	25	29	50	26	33	32	29	31	32

Table 4.2: Summary of frequency of each nonparametric component selected over the 1000 simulations for each model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). This selection rule applies to FSAM-PFLR as well.

Table 4.3 compares the above methods in terms of the estimated bias and SE of the estimates of $\boldsymbol{\alpha}$ and of the prediction accuracy which is represented by MSPE. Since there are two elements in α, each entry in Table 4.3 denotes estimated values for these two elements. To be specific we estimate the bias of an estimate $\hat{\theta}$ of a parameter θ with true value θ_0 by $\text{bias}(\hat{\theta}) = \sum_{i=1}^{1000}(\hat{\theta}_i - \theta_0)/1000$ and $\text{SE}(\hat{\theta}) = \sqrt{\sum_{i=1}^{1000}(\hat{\theta}_i - \bar{\hat{\theta}})^2/999}$, in which $\hat{\theta}_i$ denotes the estimated θ in the ith simulation and $\bar{\hat{\theta}}$ is average of $\hat{\theta}$ over the 1000 simulations. FSAM-COSSO compares favorably with the other three competitors except FSAM-GAM1 and FSAM-COSSO1 in terms of prediction accuracy, even under the assumption that $\boldsymbol{\alpha}$ are known to be $\boldsymbol{\alpha}_0$ in MARS. In addition, the point estimator of $\boldsymbol{\alpha}$ obtained from FSAM-COSSO is more stable than its counterparts from the other three competitors. Even though FSAM-GAM1 outperforms FSAM-COSSO with respect to prediction accuracy and/or the bias and SE of estimated $\boldsymbol{\alpha}$, in practice we usually have no sufficient evidence to point out non-vanishing nonparametric components in advance, let alone the closed forms of these components. The boxplot in Figure 4.1 provides a more detailed comparison of prediction errors among the six methods over the 1000 simulations; it shows that FSAM-COSSO has a substantial advantage in prediction when the underlying model is given in the form of (4.4) but unknown. The fact that FSAM-COSSO1 outperforms FSAM-GAM1 in prediction further indicates that the proposed algorithm is effective in discovering predictive features of the response.

Model	Bias	SE	MSPE
MARS	–	–	1.33
FSAM-GAM1	(-0.025, -0.021)	(0.255, 0.272)	1.15
FSAM-GAMS	(0.032, -0.047)	(0.282, 0.308)	1.41
FSAM-PFLR	(0.034, -0.031)	(0.303, 0.319)	1.67
FSAM-COSSO	(-0.026, -0.028)	(0.262, 0.283)	1.20
FSAM-COSSO1	(-0.022, -0.018)	(0.249, 0.250)	1.11

Table 4.3: Summary of estimated bias and standard error (SE) of estimated $\boldsymbol{\alpha}$ using each method, and mean squared prediction errors (MSPE). The above statistics are calculated over the 1000 simulations. Note that the column of MSPE corresponds to average of MSPE over the 1000 simulations.

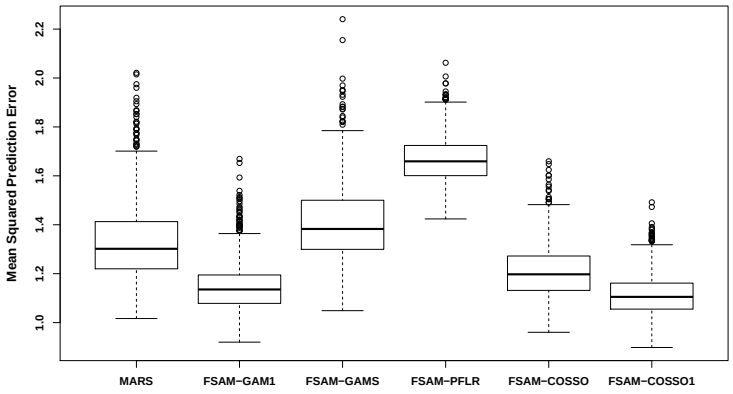


Figure 4.1: Mean squared prediction errors of each method over 1000 simulations.

In a randomly selected trial, 20 FPCs are retained initially such that over 99.9% of the variability in the curves can be captured. Figure 4.2 illustrates how cross-validation is employed to choose the tuning parameter M. Choosing the value of M which can minimize the cross-validation error, FSAM-COSSO correctly selects the three non-vanishing nonparametric components. In addition, the other three panels in Figure 4.2 display the estimated nonparametric components obtained from using the estimated scores and the true scores, as well as the true nonparametric components. It shows that estimates from these two methods are close to the true nonparametric functions and there is little disagreement between the two. This observation demonstrates that replacing true scores with the estimates has little impact on estimation of the nonparametric components.

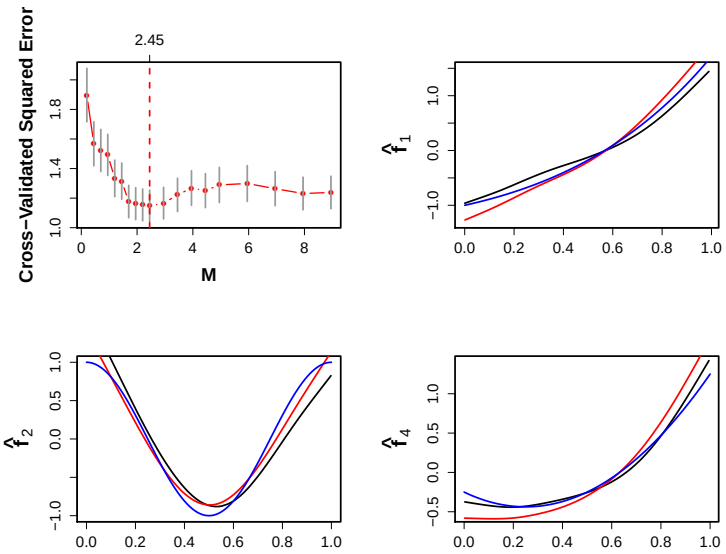


Figure 4.2: The top left panel shows how the cross validation errors change across a range of plausible values for the tuning parameter M. The other 3 panels compare the estimated nonparametric components and the true underlying nonparametric components ($f_k,\ k = 1, 2, 4$). The blue lines denote the true nonparametric components, while the black and red lines represent the estimated nonparametric components from FSAM-COSSO and FSAM-COSSO1, respectively.

4.3.2 FSAM Without Scalar Covariates

We also generate data in the same set up as in Section 4.3.1, except that the coefficient vector for the scalar covariate $\boldsymbol{z}$, $\boldsymbol{\alpha}_0$, is now set to $(0, 0)^\top$. This is essentially the model discussed in [121]. Besides the methods employed in Section 4.3.1, we also apply a method which regresses the scalar response y against $\hat{\boldsymbol{\zeta}}$ with COSSO regularization. This method is called FSAM-GAM2 in this chapter. We also fit the data using the FSAM-GAM1 method, which estimates FSAM (4.4) by assuming that ζ_1, ζ_2 and ζ_4 are known to be the only non-vanishing features, the parametric expressions of f_1, f_2, f_4 are known, and the coefficients of the scalar covariate $\boldsymbol{z}$ are known to be 0. In other words, the FSAM-GAM1 method is essentially a a multiple linear regression model with y_i as the response and $f_1(\hat{\zeta}_{i1}), f_2(\hat{\zeta}_{i2}), f_4(\hat{\zeta}_{i4})$ as the explanatory variables.

Table 4.4 summarizes the number of nonparametric components selected by each method over the 1000 Monte Carlo runs. There is only a slight difference between FSAM-COSSO and FSAM-GAM2 in terms of selecting relevant components, which suggests that our proposed method can still perform well in component selection even if there is actually no effect from scalar covariates. Table 4.5 further compares these methods in a more delicate way by providing the frequency of each component selected across these 1000 Monte Carlo runs. Likewise in the scenario when there are scalar covariates involved in the model, FSAM-COSSO shows great advantages in retaining irrelevant components far less often compared

with MARS and FSAM-PFLR. In addition, the reason why FSAM-PFLR compares favorably with FSAM-COSSO in retaining irrelevant components is that AIC tends to select a relatively small number of scaled FPC scores.

Model	Counts with the model size												
	1	2	3	4	5	6	7	8	9	10	11	12	13
MARS	0	0	20	71	140	170	197	181	127	69	16	6	3
FSAM-GAMS	0	0	169	261	238	161	98	42	19	9	2	1	0
FSAM-PFLR	5	533	312	98	40	8	4	0	0	0	0	0	0
FSAM-COSSO	0	3	615	202	93	49	27	8	3	0	0	0	0
FSAM-GAM2	0	6	738	168	52	26	3	3	4	0	0	0	0

Table 4.4: Summary of the number of selected nonparametric components over the 1000 simulations for each model. Model size indicates the number of nonparametric components selected in the model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). Here we implement the function **gam** in the R package **mgcv** to fit FSAM-GAMS. The corresponding p-values of nonparametric components are available from the function **summary.gam**. This selection rule applies to FSAM-PFLR as well, where the p-value is available from the function **lm**.

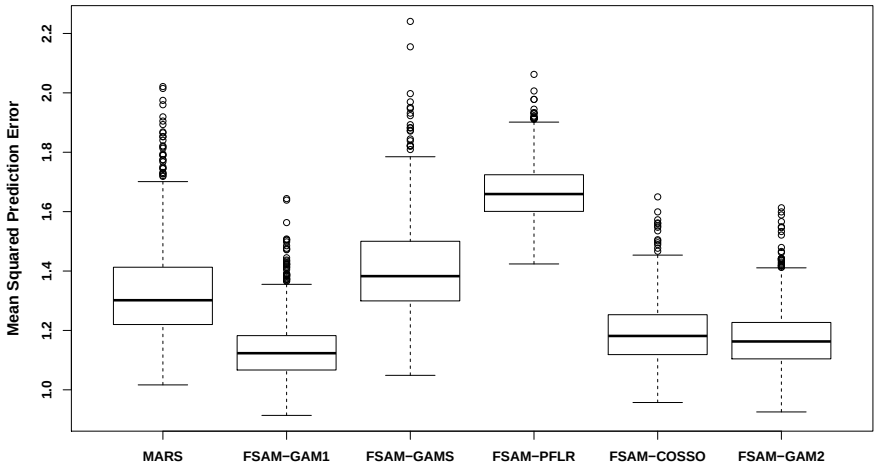


Figure 4.3: Mean squared prediction errors of each method over 1000 simulations.

More remarkable distinctions among these methods can be found in Figure 4.3, which depicts the mean squared predictions on the test data over these 1000 Monte Carlo runs. The performance of the proposed method, FSAM-COSSO is slightly inferior to those of FSAM-GAM1 and FSAM-GAM2, but much better than the other methods in prediction accuracy. FSAM-GAM1 and FSAM-GAM2 each know the model structure to some extent

in advance. That's why they can achieve greater prediction accuracy. Our method, however, does not assume that the linear part or the nonparametric part is known. Thus our proposed method is highly competitive in prediction compared with other fitting methods.

Model	Frequency of each nonparametric factor									
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$	$\hat{f}_6$	$\hat{f}_7$	$\hat{f}_8$	$\hat{f}_9$	$\hat{f}_{10}$
MARS	1000	1000	356	1000	274	226	233	235	211	244
FSAM-GAMS	1000	1000	232	997	155	100	120	106	103	114
FSAM-PFLR	1000	990	232	998	156	101	121	106	98	110
FSAM-COSSO	1000	1000	102	995	52	22	43	31	27	34
FSAM-GAM2	1000	999	79	992	34	13	28	17	16	20
	$\hat{f}_{11}$	$\hat{f}_{12}$	$\hat{f}_{13}$	$\hat{f}_{14}$	$\hat{f}_{15}$	$\hat{f}_{16}$	$\hat{f}_{17}$	$\hat{f}_{18}$	$\hat{f}_{19}$	$\hat{f}_{20}$
MARS	210	208	196	226	222	240	236	228	217	249
FSAM-GAMS	105	108	105	106	96	101	126	111	114	125
FSAM-PFLR	96	105	104	95	91	78	105	104	91	58
FSAM-COSSO	31	38	39	33	34	41	43	39	47	47
FSAM-GAM2	16	18	14	23	18	14	27	16	21	30

Table 4.5: Summary of frequency of each nonparametric component selected over the 1000 simulations for each model. In FSAM-GAMS we only retain the significant nonparametric components (p-value less than 0.05). This selection rule applies to FSAM-PFLR as well.

4.4 Real Data Applications

In this section, the proposed method (FSAM-COSSO) and several alternative methods are applied to analyze two real datasets: the Tecator data and attention deficit hyperactivity disorder (ADHD) data.

4.4.1 Tecator data

The Tecator data are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. The dataset consists of 240 meat samples; a 100-channel spectrum of absorbance (negative base 10 logarithms of the transmittance measured by the spectrometer) is recorded for each sample along with the percentages of three components of the meat: moisture (water), fat and protein. The three contents are determined by analytic chemistry. There has been extensive research on how to predict the contents using the spectrum of absorbance (see [104]; [23]; [38]; [121]; etc.). The objective of this study is to examine the effect of the spectral trajectories and the fat and water content of the meat sample on the protein contents by

fitting Model (4.1). Compared with traditional methods, this serves as a nondestructive method to carry out a quantitative determination of protein content.

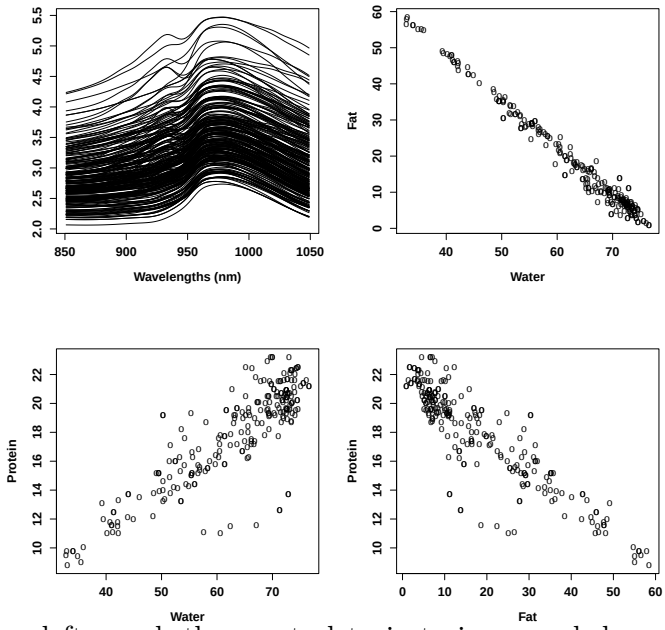


Figure 4.4: The top left panel: the spectral trajectories recorded on the Tecator Infratec Food and Feed Analyzer. The other three panels depict the scatter plots among the three contents.

In Model (4.1), the response of primary interest, Y, is the protein content; both the spectral trajectories denoted as $X(t)$, and the fat and water contents denoted as $\boldsymbol{z}$, are considered as explanatory variables for predicting the protein content, differing from the method called the component selection and estimation for functional additive model (CSE-FAM) in [121] where only spectral trajectories were taken into consideration for predicting the protein content. After applying PACE to the spectral trajectories to obtain estimated $\boldsymbol{\zeta}_i$'s, FSAM-COSSO is implemented to fit Model (4.4) to estimate and select non-vanishing nonparametric components for the response as well as estimating the effect of scalar covariates on the response. The top left panel of Figure 4.4 presents the spectral trajectories of the 240 meat samples. To assess the performance of each method, 185 out of the 240 meat samples are randomly selected from the training set and the remaining 55 samples constitute the test set.

As suggested by [121], the first 20 FPCs, accounting for over 99.9% of the total variability in the spectral trajectories, are initially retained to avoid neglecting some relevant FPCs. In addition, pairwise scatter plots among the three contents, illustrated in Figure 4.4, suggest a substantial multicollinearity between the fat and water contents and a linear relationship between the protein content and the fat and/or water content. Therefore, only the fat content is used in the parametric part when predicting the protein content. We then apply

FSAM-COSSO to estimate and select nonparametric components while estimating the effect of the fat content on the prediction of the protein content. The tuning parameter λ_0 in the iterative updating algorithm is selected by sixfold cross-validation, which gives $\lambda_0 = 2.57 \times 10^{-5}$. Fivefold cross-validation suggests that 13 is an optimal choice for M.

The estimated nonparametric components are displayed in Figure 4.5, which shows the 15 nonparametric components $\{\hat{f}_1, \ldots, \hat{f}_8, \ \hat{f}_{11}, \hat{f}_{13}, \ldots, \hat{f}_{18}\}$ are selected from the 20 components. The estimated coefficient of the fat content is -0.19, corresponding to the fact that the protein content is negatively correlated with the fat content indicated in the bottom right panel of Figure 4.4.

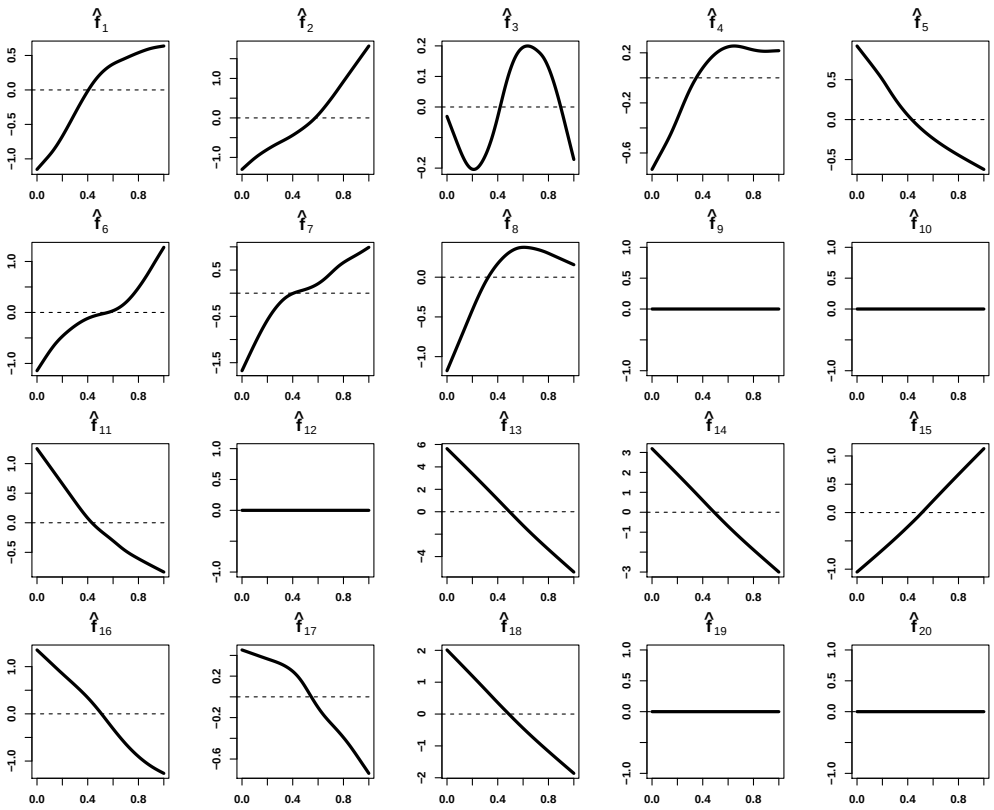


Figure 4.5: Non-vanishing nonparametric components estimated for the functional semiparametric additive model (4) from the Tecator data. Out of total 20 nonparametric components, 15 nonparametric components are selected.

MSPE and quasi-R^2 ([121]) are calculated on the test data set to compare various methods; the latter is defined by

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

The last 10 FPCs actually explain less than 0.01% of the total variability in the spectral curves. However, they play a critical role in predicting the protein content, justifying why a sufficiently large number of principal components should be retained initially. To demonstrate the importance of the last 10 FPCs, the same model (FSAM-COSSO) is fit where only the first 10 FPCs are initially retained. The model neglecting the last 10 FPCs appears to be considerably inferior to that model retaining 20 FPCs initially in terms of both MPSE and R^2.

We also fit several alternative methods such as MARS and FSAM-FAMS to discover the relationship between the protein content and the explanatory variables and compare these models with FSAM-COSSO in prediction accuracy and the ability to explain the variability in the response. The effect of FPCs retained initially on the prediction of the protein content is examined as well in the alternative methods. Furthermore, to investigate how the prediction accuracy can be enhanced by incorporating the fat content in prediction models, the models are fitted by regressing the protein content only on the initially retained $\hat{\boldsymbol{\zeta}}_i$'s and then compared to the corresponding models where both $\hat{\boldsymbol{\zeta}}_i$'s and the fat content are considered as explanatory variables.

	$d = 20$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.52	0.71	0.84	0.73	0.83	1.18
R^2	0.97	0.96	0.95	0.96	0.95	0.93
	$d = 10$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.92	1.99	1.35	1.42	0.97	1.01
R^2	0.95	0.88	0.92	0.92	0.94	0.94

Table 4.6: Summary of prediction error and proportion of variance explained on the test set of each model. FAM represents the functional additive model ([72]) where only $\hat{\boldsymbol{\zeta}}_i$'s are considered as explanatory variables. MARS_0 denotes the MARS model considering only $\hat{\boldsymbol{\zeta}}_i$'s as explanatory variables while neglecting the effect of the fat content. $d = 10$ and $d = 20$ indicate that 10 and 20 leading FPCs are initially retained, respectively.

Table 4.6 summarizes prediction errors and proportions of variance explained on the test set of all methods. It can be observed from the table that retaining a sufficiently large number of FPCs initially can ameliorate prediction accuracy to a great extent, even though the last 10 FPCs only make a negligible contribution to capturing the variance of the spectral curves. In addition, accounting for the effect of the fat content in each of the above models outperforms its counterpart, which does not incorporate the effect of the fat content, in terms of prediction accuracy and explaining variability in the response variable.

Last but not least, the proposed method demonstrates its far superior ability to predict the response and explain its variance compared with other methods.

4.4.2 ADHD data

The attention deficit hyperactivity disorder (ADHD) is the most prevalent neurodevelopmental disorder in school-age children and adolescents ([29]). The key symptoms of ADHD comprise inattention, hyperactivity and impulsivity. Due to lack of objective measurements in diagnosis, there have been critical concerns for appropriate diagnosis of ADHD, which is associated with substantial social and economic costs ([35]). The data were obtained from the ADHD-200 Sample Initiative Project, which aimed to seek objective biological tools like neuroimaging signals to aid diagnosis of ADHD. Our analysis is based on the data collected from the New York University (NYU) Child Study Center, one of the eight sites in the project. The dataset consists of two main parts. The first part is filtered preprocessed resting state data using anatomical automatic labelling atlas, which parcellates brain into 116 Regions of Interests (ROIs). In each region, the mean blood-oxygen-level dependent (BOLD) signal was recorded at 172 equally spaced time points. The second part is composed of phenotypic features like gender, handedness, diagnosis of ADHD, medication status, ADHD index and IQ measures. A more detailed description of the data can be found at [7]. Our objective is to use the BOLD signals and phenotypic features to predict the ADHD index, a measure which can reflect the overall level of ADHD symptoms ([18]).

We focus on 120 subjects in our analysis after removing measurements which failed in quality control. The functional predictor is taken as the BOLD signals of 91st-108th regions, because they are parcellations of the cerebellum region, which was found to play a role in the development of ADHD ([8]). To compare the prediction performance of each method, these 120 subjects are randomly divided into a training set with 100 subjects and a test set with the other 20 subjects. Following this rule, we randomly split the data to the training and test set for 100 times.

Table 4.7 summarizes the mean squared prediction error across the 100 splits for each method. FSAM-COSSO turns out to be substantially superior to other methods in terms of prediction accuracy. In addition, accounting for effects of phenotypic features is able to improve prediction accuracy greatly for each method. Moreover, for methods other than FSAM-COSSO, retaining a large number of FPC scores initially may impair prediction of the ADHD index as may be seen by comparing the upper part with the lower part of Table 4.7. The primary reason for this might be that the BOLD signal is not a strong predictor of the ADHD index; thus incorporating more FPC scores would add considerable prediction variabilities while making little contribution to reducing bias. However, the negligible difference in prediction accuracy of FSAM-COSSO between these two scenarios suggests that the proposed method manages to reduce variances via component selection and thus achieve a better trade-off between bias and variance. As a result, the proposed method can still

achieve a satisfactory performance in prediction even though a large number of irrelevant FPC scores are retained initially.

	99.9%					
	FSAM-COSSO	CSEFAM	FSAM-PFLR	FLR	MARS	MARS_0
MSPE	49.24	196.66	84.52	216.78	88.67	332.51
	85%					
	FSAM-COSSO	CSEFAM	FSAM-PFLR	FLR	MARS	MARS_0
MSPE	49.27	194.07	84.52	216.78	67.44	247.48

Table 4.7: Summary of prediction error on the test set of each model. FLR represents the functional linear model where only $\hat{\boldsymbol{\zeta}}_i$'s are considered as explanatory variables. For both FSAM-PFLR and FLR, the number of retained FPCs is chosen via AIC. MARS_0 denotes the MARS model considering only $\hat{\boldsymbol{\zeta}}_i$'s as explanatory variables while neglecting other phenotypic features. 99.9% and 85% indicate that the first d FPCs initially retained can explain 99.9% and 85% of variability of curves in the training set, respectively.

4.5 Asymptotic Properties

The following theorem is given only in the case when the true scaled FPC scores ($\boldsymbol{\zeta}$) are known. It would be desirable to establish the corresponding theorem with the estimated scaled FPC scores, when the functional data is densely observed and may even be contaminated with measurement errors. This problem was considered by [121], where there were no scalar convariates. Thus it would be natural for us to follow their ideas to develop the theorem. Unfortunately we are not able to follow the proof of their Lemma 2, where a dual problem is used to show that the penalty term $J(f)$ is bounded by a constant independent of sample size. This statement is essential to show that the derivative of the estimated function is uniformly bounded for each argument. We are thus not able to derive [121]'s subsequent results without a bounded derivative. For this reason we present the following theorem assuming the true scaled scores are known. We observe, however, that simulation studies show that there would not be remarkable differences in performances when the true scores are replaced by estimates. Thus, although we expect that our theorem should extend to estimated scores, we have not succeeded in doing so.

We now set out conditions to establish Theorem 1 given that the true scaled FPC scores are known. Suppose that n i.i.d samples are generated from the following model

$$y_i = f_0(\boldsymbol{\zeta}_i) + \boldsymbol{\alpha}_0^\top \boldsymbol{z}_i + \epsilon_i,$$

where $\boldsymbol{\zeta}_i \in [0,1]^d$, $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$, and f_0 is assumed to be an element of $\mathscr{F}^d = \{1\} \oplus \sum_{j=1}^d \bar{H}$ with $H = \{1\} \oplus \bar{H}$ being the lth-order Sobolev space on $[0,1]$ with the norm defined in

(4.5). Note that the estimated nonparametric part $\hat{f}_n$ and the estimated coefficient $\hat{\boldsymbol{\alpha}}$ in the parametric part are defined as the solution of the following minimization problem:

$$(\hat{f}_n, \hat{\boldsymbol{\alpha}}) = \underset{f \in \mathscr{F}^d, \boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{\zeta}_i) - \boldsymbol{\alpha}^\top \boldsymbol{z}_i\}^2 + \tau_n^2 J(f) \right].$$

Consequently, the estimate of the conditional expectation function of y, $g_0(\boldsymbol{\zeta}, \boldsymbol{z}) = f_0(\boldsymbol{\zeta}) + \boldsymbol{\alpha}_0^\top \boldsymbol{z}$, is defined as $\hat{g}_n = \hat{f}_n + \hat{\boldsymbol{\alpha}}^\top \boldsymbol{z}$. The empirical norm of g is defined as $||g||_n = \sqrt{\frac{1}{n}\sum_{i=1}^n g^2(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)}$ for $g \in \mathscr{G} = \{g : g(\boldsymbol{\zeta}, \boldsymbol{z}) = f(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top \boldsymbol{z}, f \in \mathscr{F}^d, \boldsymbol{\alpha} \in \mathbb{R}^p\}$. Define $h(\boldsymbol{\zeta}) = \mathsf{E}(\boldsymbol{z}|\boldsymbol{\zeta})$ and $\boldsymbol{z}^\star = \boldsymbol{z} - h(\boldsymbol{\zeta})$. Let $\Lambda_{\min}(\boldsymbol{A})$ and $\Lambda_{\max}(\boldsymbol{A})$ denote the minimal and maximal eigenvalues of a matrix $\boldsymbol{A}$, respectively.

The following assumptions are needed.

(A.1) Both $\boldsymbol{\zeta}$ and $\boldsymbol{z}$ are statistically independent of ϵ. Furthermore, $\mathsf{E}(\epsilon) = 0$ and $\max_{1 \le j \le p} \mathsf{E}(|\boldsymbol{z}_{(j)}|) < \infty$, where $\boldsymbol{z}_{(j)}$ denotes the jth component of $\boldsymbol{z}$.

(A.2) $\Lambda_{\max}[\mathrm{var}\{h(\boldsymbol{\zeta})\}] < \infty$ and $0 < \Lambda_{\min}\{\mathrm{var}(\boldsymbol{z}^\star)\} \le \Lambda_{\max}\{\mathrm{var}(\boldsymbol{z}^\star)\} < \infty$.

Obviously $0 < \Lambda_{\min}\{\mathrm{var}(\boldsymbol{z})\} \le \Lambda_{\max}\{\mathrm{var}(\boldsymbol{z})\} < \infty$ under (A.2).

(A.3) ϵ_i's are (uniformly) sub-Gaussian, i.e., there exist some constant K and σ_0^2, such that

$$K^2(\mathsf{E}e^{\epsilon_i^2/K^2} - 1) \le \sigma_0^2.$$

(A.4) The tuning parameter τ_n satisfies

$$\tau_n = o(1)$$

as $n \to \infty$.

The main result is the following

Theorem 1 Provided that Assumptions (A.1)-(A.3) hold, then (i) if $0 < J(f_0) < \infty$, and $\tau_n^{-1} = n^{\frac{l}{2l+1}}\{J(f_0)\}^{\frac{2l-1}{4l+2}}$, we have $||\hat{g}_n - g_0||_n = O_P(n^{-\frac{l}{2l+1}})\{J(f_0)\}^{\frac{1}{2l+1}}$ and $J(\hat{f}_n) = J(f_0)O_P(1)$; (ii) if f_0 is a constant, i.e., $J(f_0) = 0$, and $\tau_n^{-1} = n^{1/4}$, then we have $||\hat{g}_n - g_0||_n = O_P(n^{-\frac{1}{2}})$ and $J(\hat{f}_n) = O_P(n^{-\frac{1}{2}})$.

(A.5) The support of $\boldsymbol{z}$ is compact in $\mathbb{R}^p$.

Corollary 1 Besides Assumptions (A.1)-(A.3), if Assumption (A.5) is assumed to hold as well, then in either (i) or (ii),

$$||\hat{f}_n - f_0||_n = O_P(n^{-\frac{l}{2l+1}}), \quad ||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||_E = O_P(n^{-\frac{l}{2l+1}}),$$

where $||\cdot||_E$ denotes the Euclidean norm of a vector.

Proofs of Theorem 1 and Corollary 1 are given in the Appendix.

4.6 Conclusions and Discussions

Semiparametric additive models are known to possess the flexibility of a nonparametric model and the interpretability of a parametric model. In this chapter, we propose a functional semiparametric additive model in which a scalar response is regressed on a functional covariate and finite dimensional scalar covariates. To achieve flexibility and interpretability simultaneously, the effect of the functional covariate on the mean response is modeled in the framework of FAM, where the additive components are functions of scaled FPC scores, and a linear relationship is assumed between the mean response and the scalar covariates. We also develop an estimation method to estimate both the nonparametric and parametric parts in the proposed model.

The estimation procedure consists of three important steps. First, FPCA or PACE is employed to estimate FPC scores of the functional covariate which may be subject to measurement errors. Second, we adopt a special regularization scheme (COSSO) to penalize the additive components to smooth and select non-vanishing components. Third, to address the issue of interdependence between the estimated nonparametric part and parametric part, we propose an iterative updating algorithm, which is similar in spirit to the EM algorithm.

We show that choosing a sufficiently large number of FPCs is essential. On the one hand, this can account for a great proportion of variability in the functional covariate. On the other hand, retaining a sufficiently large number of FPCs can to a great extent circumvent neglecting predictive FPC scores with small variances, since there is no guarantee that leading FPC scores are necessarily more relevant to the response. The importance of retaining a sufficiently large number of FPCs is demonstrated via the application to the Tecator data, where retaining a smaller number of FPCs results in substantially greater prediction errors. The applications also show that incorporating the effect of scalar covariates can enhance prediction accuracy compared with models that only account for the effect of the functional covariate when the scalar covariates are predictive of the response variable.

The asymptotic theory is based on the assumption that the true scaled FPC scores are known, but in practice these are unavailable. We provide an algorithm with respect to estimating FPC scores from observed curves which may be subject to measurement errors and then estimating both nonparametric and parametric parts in the model using estimated FPC scores. It would be very nice to extend the theory to this case where true scaled FPC scores are not observable. The simulation study suggests that the estimates are still quite close to the true nonparametric and parametric parts when FPC scores are estimated. Even though this work focuses on regressing a scalar response on a functional covariate and another finite dimensional covariate, the methodology can be extended to accommodate other scenarios. For example, the framework may be extended to fit a generalized functional semiparametric additive model in which the distribution of the response variable of interest belongs to an exponential family. In addition, more than one functional covariate can be

investigated in future work, in which we are more concerned about choosing relevant ones from multiple functional covariates. The model may be further extended to allow for high dimensional scalar covariates. In the model proposed in this chapter, the dimension of scalar covariates p is fixed. If $p = p_n$ is allowed to increase when sample size n increases, we may also need to penalize the parametric part to choose relevant scalar covariates. In this case, regularization of both the nonparametric part and the parametric part is necessary to improve interpretability and reduce variability.

The proposed functional semiparametric additive model can be regarded as a natural extension of the work by [95], which represented the effect of a functional covariate by a linear combination of leading FPC scores. We, however, employ an additive structure of FPC scores to model the effect of a functional covariate. As pointed out by one referee, there might be concerns in both the additive structure and the method of dimension reduction associated with this model. [107] considered a special case of functional generalized additive models, which are easy to interpret since they do not impose the condition that a scalar response depends on a functional covariate via an additive form of FPC scores. [94] extended the idea to function-on-function regression models. Another additive model was proposed to model the relationship between a functional response and high dimensional scalar covariates in [4]. Concerning dimension reduction, we project a functional covariate onto the directions of FPCs and then retain the leading FPC scores in the model. A possible concern is that this procedure does not take into account the response information. To address this issue, [58] proposed nonlinear sufficient dimension reduction based on the relationship between the functional covariate and the scalar response. It would be worthwhile to explore both the generalized functional additive model and sufficient dimension reduction in functional regression models with scalar covariates.

4.7 Appendix: Proofs

Consider the regression model:

$$y_i = f_0(\boldsymbol{\zeta}_i) + \boldsymbol{\alpha}_0^\top \boldsymbol{z}_i + \epsilon_i,$$

where $f_0(\boldsymbol{\zeta}) = b_0 + \sum_{j=1}^d f_{0j}(\zeta_j)$ with $f_{0j} \in \bar{H}$ and $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$. Write $g(\boldsymbol{\zeta}, \boldsymbol{z}) = a + \tilde{f}(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} = a + \sum_{j=1}^d \tilde{f}_j(\zeta_j) + \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}}$ such that $\sum_{i=1}^n \tilde{f}_j(\zeta_{ij}) = 0, j = 1, \ldots, d$ and $\tilde{\boldsymbol{z}} = \boldsymbol{z} - \bar{\boldsymbol{z}}$ which satisfies $\sum_{i=1}^n \tilde{z}_{is} = 0, s = 1, \ldots, p$, where $\bar{\boldsymbol{z}}$ denotes the sample mean of $\boldsymbol{z}_i$'s and $\tilde{\boldsymbol{z}}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{ip})^\top$ is the evaluation of $\tilde{\boldsymbol{z}}$ at the data point $\boldsymbol{z}_i, i = 1, \ldots, n$. Similarly, write $g_0(\boldsymbol{\zeta}) = a_0 + \tilde{f}_0(\boldsymbol{\zeta}) + \boldsymbol{\alpha}_0^\top \tilde{\boldsymbol{z}} = a_0 + \sum_{j=1}^d \tilde{f}_{0j}(\zeta_j) + \boldsymbol{\alpha}_0^\top \tilde{\boldsymbol{z}}$ such that $\sum_{i=1}^n \tilde{f}_{0j}(\zeta_{ij}) = 0, j = 1, \ldots, d$, and $\hat{g}(\boldsymbol{\zeta}, \boldsymbol{z}) = \hat{a} + \hat{f}(\boldsymbol{\zeta}) + \hat{\boldsymbol{\alpha}}^\top \tilde{\boldsymbol{z}} = \sum_{j=1}^d \hat{f}_j(\zeta_j) + \hat{\boldsymbol{\alpha}}^\top \tilde{\boldsymbol{z}}$.

Remark 1 The above decomposition for f_0 as sum of a_0 and $\tilde{f}_0$ is different from that as an element of $\{1\} \oplus \sum_{j=1}^d \bar{H}$. This difference applies to the decomposition of $\hat{f}$ as well. The

latter representation is given for the sake of identifiability and useful for entropy calculation, which will be illustrated in Lemma 2.

Let $J(g) = J(f)$; that is, the penalty ignores the linear components. Then since $\hat{g}$ minimizes the target function

$$\begin{aligned}
L(g) &= \frac{1}{n}\sum_{i=1}^{n}\{g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i) - y_i\}^2 + \tau_n^2 J(g) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{a + \sum_{j=1}^{d}\tilde{f}_j(\zeta_{ij}) + \boldsymbol{\alpha}^\top\tilde{\boldsymbol{z}}_i - a_0 - \sum_{j=1}^{d}\tilde{f}_{0j}(\zeta_{ij}) - \boldsymbol{\alpha}_0^\top\tilde{\boldsymbol{z}}_i - \epsilon_i\right\}^2 + \tau_n^2 J(g) \\
&= (a-a_0)^2 - \left(\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\right)(a-a_0) + \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{d}\tilde{f}_j(\zeta_{ij}) + \tilde{\boldsymbol{z}}_i^\top\boldsymbol{\alpha} - \sum_{j=1}^{d}\tilde{f}_{0j}(\zeta_{ij}) - \tilde{\boldsymbol{z}}_i^\top\boldsymbol{\alpha}_0 - \epsilon_i\right\}^2 \\
&\quad + \tau_n^2 J(g),
\end{aligned}$$

the estimated intercept $\hat{a}$ in $\hat{g}$ must satisfy $\hat{a} = a_0 + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$, which implies $\hat{a} - a_0 = O_P(n^{-\frac{1}{2}})$. From now on, we consider the target function

$$\tilde{L}(\tilde{f}, \boldsymbol{\alpha}|\boldsymbol{\zeta}_i, \tilde{\boldsymbol{z}}_i) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{f}(\boldsymbol{\zeta}_i) + \boldsymbol{\alpha}^\top\tilde{\boldsymbol{z}}_i - \tilde{f}_0(\boldsymbol{\zeta}_i) - \boldsymbol{\alpha}_0^\top\tilde{\boldsymbol{z}}_i - \epsilon_i\right\}^2 + \tau_n^2 J(\tilde{f}). \tag{4.13}$$

The solution is denoted as $\hat{g}_n = \hat{f}_n + \hat{\boldsymbol{\alpha}}^\top\tilde{\boldsymbol{z}}$, which is an estimate of $g_0 = \tilde{f}_0 + \boldsymbol{\alpha}_0^\top\tilde{\boldsymbol{z}}$.

Let $\mathscr{F}^d = \{f : f \in 1 \oplus \left(\bigoplus_{j=1}^{d}\bar{H}\right),\ J(f) < \infty\}$, where $J(f) = \sum_{j=1}^{d}||P^j f||$ with P^j denoting the orthogonal projection from $\mathscr{F}$ onto $\bar{H}$. Therefore, the conditional expectation, g_0 is an element of

$$\mathscr{G} = \left\{g : g(\boldsymbol{\zeta}, \boldsymbol{z}) = \sum_{j=1}^{d} f_j(\zeta_j) + \boldsymbol{\alpha}^\top\tilde{\boldsymbol{z}}, \boldsymbol{\alpha} \in \mathbb{R}^p, \sum_{j=1}^{d} f_j \in \mathscr{F}^d, \sum_{i=1}^{n} f_j(\zeta_{ij}) = 0\right\},$$

under the assumption that $J(f_0) < \infty$. Following [65], for $g(\boldsymbol{\zeta}, \boldsymbol{z}) = f(\boldsymbol{\zeta}) + \boldsymbol{\alpha}^\top\tilde{\boldsymbol{z}} \in \mathscr{G}$, $J(g)$ is set to be $J(f)$; thus $J(g_0) < \infty$. Now consider two subsets of $\mathscr{G}$, $\mathscr{G}_1 = \{g_1 : g_1(\boldsymbol{\zeta}, \boldsymbol{z}) = \sum_{j=1}^{d} f_j(\zeta_j), f_j'\text{s satisfy } \sum_{i=1}^{n} f_j(\zeta_{ij}) = 0, g_1 \in \mathscr{F}^d\}$ and $\mathscr{G}_2 = \{g_2 : g_2(\boldsymbol{\zeta}, \boldsymbol{z}) = \boldsymbol{\alpha}^\top\tilde{\boldsymbol{z}}, \boldsymbol{\alpha} \in \mathbb{R}^p\}$. Every element of $\mathscr{G}$ can be written as sum of two elements, one from each of $\mathscr{G}_1$ and $\mathscr{G}_2$.

Before stating a proposition that will be employed later, we first introduce some notation and the concept of entropy. Let Q be the joint distribution of $\boldsymbol{\zeta}$ and $\boldsymbol{z}$ and Q_n the corresponding empirical distribution. Obviously the support of Q is $\mathscr{X} = [0,1]^d \times \mathbb{R}^p$. For any function g supported on $\mathscr{X}$, if $\int |g|^2 \, \mathrm{d}Q < \infty$, then define

$$||g||_{2,Q} = \left(\int |g|^2 \, \mathrm{d}Q\right)^{\frac{1}{2}}.$$

We refer to $||\cdot||_{2,Q}$ as the $L_2(Q)$ metric; similarly we can define the $L_2(Q_n)$ metric or the $||\cdot||_n$ metric by replacing Q with Q_n. We can now define the entropy of $\mathscr{G}$ with respect to the $||\cdot||_n$ metric. For any $\delta > 0$, we can find a collection of functions $g_1, \ldots, g_N$, such that for each $g \in \mathscr{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that $||g - g_j||_n \leq \delta$. Let $N(\delta, \mathscr{G}, ||\cdot||_n)$ be the smallest value of N for which such a covering by balls with radius δ exists. Then $H(\delta, \mathscr{G}, ||\cdot||_n) = \log\{N(\delta, \mathscr{G}, ||\cdot||_n)\}$ is called the δ-entropy of $\mathscr{G}$ (for the $||\cdot||_n$ metric). Similarly, we can define the entropy of $\mathscr{G}$ for other metrics like the $||\cdot||_\infty$ metric. It is trivial that $H(\delta, \mathscr{G}, ||\cdot||_n) \leq H(\delta, \mathscr{G}, ||\cdot||_\infty)$. For distinction, we write $||\cdot||_\infty$ to denote the supremum norm of a function, $||\cdot||_E$ to denote the Euclidean norm of a vector, $||\cdot||$ to denote the Sobolev norm defined in the RKHS, $||\cdot||_{2,Q}$ to denote the $L_2(Q)$ metric and $||\cdot||_n$ to denote the $L_2(Q_n)$ metric. Following the notation on page 167 from [102], for any $g \in \mathscr{G}$,

$$||g||_n^2 = \frac{1}{n}\sum_{i=1}^n g^2(\boldsymbol{\zeta}_i, \boldsymbol{z}_i), \quad (\epsilon, g)_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i), \quad ||y - g||_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i))^2.$$

For readability, we refer to the assumptions (A.1) - (A.4) as Assumption 1, Assumption 2, etc.

Assumption 1 *Both $\boldsymbol{\zeta}$ and $\boldsymbol{z}$ are statistically independent of ϵ. Furthermore,* $\mathrm{E}(\epsilon) = 0$ *and* $\max_{1 \leq j \leq p} \mathrm{E}(|\boldsymbol{z}_{(j)}|) < \infty$*, where $\boldsymbol{z}_{(j)}$ denotes the jth component of $\boldsymbol{z}$.*

Assumption 2 $\Lambda_{\max}[\mathrm{Var}\{h(\boldsymbol{\zeta})\}] < \infty$ *and* $0 < \Lambda_{\min}\{\mathrm{Var}(\boldsymbol{z}^\star)\} \leq \Lambda_{\max}\{\mathrm{Var}(\boldsymbol{z}^\star)\} < \infty$.

Assumption 3 *ϵ_i's are (uniformly) sub-Gaussian, i.e., there exist some constants K and σ_0^2, such that*

$$K^2(\mathrm{E}\, e^{\epsilon_i^2/K^2} - 1) \leq \sigma_0^2.$$

Assumption 4 *The tuning parameter τ_n satisfies*

$$\tau_n = o(1)$$

as $n \to \infty$.

Assumption 2 implies that $0 < \Lambda_{\min}\{\mathrm{Var}(\boldsymbol{z})\} \leq \Lambda_{\max}\{\mathrm{Var}(\boldsymbol{z})\} < \infty$. The sample variance-covariance matrix of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ is denoted as $\boldsymbol{S}_{\boldsymbol{z}}^2$, i.e., $\boldsymbol{S}_{\boldsymbol{z}}^2 = \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i^\top$.

Lemma 1 If $\hat{g}_n$ is the minimizer of $L(g)$ and Assumptions 3 and 4 are met, then there exists a constant σ not depending on n, such that

$$||\hat{g}_n - g_0||_n \leq \sigma$$

almost surely for sufficiently large n. Furthermore, if Assumptions 1 and 2 are satisfied as well, then almost surely

$$||\hat{g}_n||_n \leq R$$

for some positive constant R (independent of n) as long as n is sufficiently large.

Proof: Since $\hat{g}_n$ minimizes $L(g)$, then it must satisfy

$$||y - \hat{g}_n||_n^2 + \tau_n^2 J(\hat{g}_n) \leq ||y - g_0||_n^2 + \tau_n^2 J(g_0);$$

thus

$$||\hat{g}_n - g_0||_n^2 + \tau_n^2 J(\hat{g}_n) \leq 2(\epsilon, \hat{g}_n - g_0)_n + \tau_n^2 J(g_0). \tag{4.14}$$

From Assumption 3, we have $\mathrm{E}\,(\epsilon_1^2) < \infty$. Then $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 = O(1)$ almost surely. By the Cauchy-Schwarz inequality, it follows

$$||\hat{g}_n - g_0||_n^2 \leq ||\hat{g}_n - g_0||_n O(1) + o(1).$$

Therefore, there exist positive constants σ and R such that, almost surely, for all large n,

$$||\hat{g}_n - g_0||_n \leq \sigma.$$

Additionally, since f_0 is a continuous function defined on $[0,1]^d$ and $\Lambda_{\max}\{\mathrm{Var}(\boldsymbol{z})\}$ is finite, we see that $\mathrm{E}\,\{g_0(\boldsymbol{\zeta}, \boldsymbol{z})\}^2 < \infty$. By the strong law of large numbers, we have almost surely, for all large n,

$$||\hat{g}_n||_n \leq ||\hat{g}_n - g_0||_n + ||g_0||_n \leq R.$$

□

Note that we incorporate the estimated intercept $\hat{a}$ in $\hat{g}$. Actually this will not make a difference if we remove $\hat{a}$ from $\hat{g}$ given the fact $|\hat{a} - a_0| = O_P(n^{-\frac{1}{2}})$ as shown above. Denote $B_n(g_0, \sigma) = \{g \in \mathscr{G} : ||g - g_0||_n \leq \sigma\}$. Due to Lemma 1, we restrict our attention to $B_n(g_0, \sigma)$ from now on. It follows that $\sup_{g \in B_n(g_0,\sigma)} ||g||_n \leq R$, with a similar argument to that used in showing Lemma 1. Let $\mathscr{G}'$ denote $B_n(g_0, \sigma) \cap \{g \in \mathscr{G} : J(g) \leq C\}$, where C is a positive constant. Correspondingly, let $\mathscr{G}_1' = \mathscr{G}' \cap \mathscr{G}_1$ and $\mathscr{G}_2' = \mathscr{G}' \cap \mathscr{G}_2$.

Proposition 1 Under Assumptions 1, 2 and 3, there exist constants T_0 and C_0, both of which are independent of n, such that

$$\mathbf{P}\left\{\sup_{g \in \mathscr{G}^\top} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)|}{||g||_n^{1-\frac{1}{2l}}} \geq T\right\} \leq 2\exp\left(-\frac{T^2}{C_0^2}\right), \tag{4.15}$$

for all $T \geq T_0$.

To prove Proposition 1, we need the following lemmas.

Lemma 2 Assuming that Assumption 1 and 2 are met, then there exists a positive constant A, which does not depend on n, such that the entropy of $\mathscr{G}'$ satisfies

$$H(\delta, \mathscr{G}', ||\cdot||_n) \leq A\delta^{-\frac{1}{l}}, \quad \forall \delta > 0$$

for sufficiently large n.

Proof: First we study the entropy of $\mathscr{G}_1'$. As shown in Lemma A.1 in [60],

$$H(\delta, \{g_1 : g_1(\boldsymbol{\zeta}) = \sum_{j=1}^{d} f_j(\zeta_j), f_j\text{s satisfy } \sum_{i=1}^{n} f_j(\zeta_{ij}) = 0, J(g_1) \leq 1\}, ||\cdot||_\infty) \leq A_0 d^{(l+1)/l}\delta^{-\frac{1}{l}},$$

for all $\delta > 0$, $n \geq 1$ and some $A_0 > 0$ not depending on δ, n or d. Therefore it can be claimed that

$$H(\delta, \mathscr{G}_1', ||\cdot||_\infty) \leq A_1\delta^{-\frac{1}{l}} \quad \forall \delta > 0, \tag{4.16}$$

where A_1 is a positive constant not depending on n or δ.

For any g writing as $g(\boldsymbol{\zeta}, \boldsymbol{z}) = \sum_{j=1}^{d} f_j(\zeta_j) + \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} \in \mathscr{G}'$, where $g_1(\boldsymbol{\zeta}, \boldsymbol{z}) = \sum_{j=1}^{d} f_j(\zeta_j)$ satisfies $J(g_1) \leq C$, then we have $||g - g_0||_n \leq \sigma$ and $\sum_{j=1}^{d} ||f_j - \tilde{f}_{0j}||_\infty \leq 2dC$, based on Lemma A.1 in [60]. $\boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}}$ therefore satisfies, for all large n,

$$\begin{aligned} ||\boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} - \boldsymbol{\alpha}_0^\top \tilde{\boldsymbol{z}}||_n &= \left\| \left\{ g(\boldsymbol{\zeta}, \boldsymbol{z}) - \sum_{j=1}^{d} f_j \right\} - \left\{ g_0(\boldsymbol{\zeta}, \boldsymbol{z}) - \sum_{j=1}^{d} \tilde{f}_{0j} \right\} \right\|_n \\ &\leq ||g - g_0||_n + \sum_{j=1}^{d} ||f_j - \tilde{f}_{0j}||_n \\ &\leq 2\sigma + 2dC. \end{aligned} \tag{4.17}$$

As a result, for any $q(\boldsymbol{\zeta}, \boldsymbol{z}) = \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} \in \mathscr{G}_2'$, $||q||_n \leq M$ holds for some constant M and sufficiently large n, based on the triangular inequality and the fact that $||\boldsymbol{\alpha}_0 \tilde{\boldsymbol{z}}||_n$ is finite for sufficiently large n. It is from the fact that for sufficiently large n, $||\boldsymbol{\alpha}_0^\top \tilde{\boldsymbol{z}}||_n^2 \leq (\Lambda_{\max}(\mathrm{Var}(\boldsymbol{z})) + \epsilon)||\boldsymbol{\alpha}_0||_E^2$ holds almost surely for any given $\epsilon > 0$ and $\Lambda_{\max}(\mathrm{Var}(\boldsymbol{z}))$ is finite if Assumption 2 is met. Additionally, $||\boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} - \boldsymbol{\alpha}_0^\top \tilde{\boldsymbol{z}}||_n^2 > (\Lambda_{\min}(\mathrm{Var}(\boldsymbol{z})) - \epsilon)||\boldsymbol{\alpha} - \boldsymbol{\alpha}_0||_E^2$ holds almost surely for any given $\epsilon > 0$. Therefore, $||\boldsymbol{\alpha} - \boldsymbol{\alpha}_0||_E \leq C_a$ for some constant C_a and any $g_2(\boldsymbol{\zeta}, \boldsymbol{z}) = \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}} \in \mathscr{G}_2^\top$. It follows that $H(\delta, \mathscr{G}_2', ||\cdot||_n) \leq A_2 \log\left(\frac{1}{\delta}\right)$ almost surely for some constant A_2 not dependent on n, when n is sufficiently large.

As pointed out earlier, every element in $\mathscr{G}'$ can be written as sum of two elements from $\mathscr{G}_1'$ and $\mathscr{G}_2^\top$, respectively. Consequently, $H(\delta, \mathscr{G}', ||\cdot||_n) \leq H(\delta/2, \mathscr{G}_1', ||\cdot||_n) + H(\delta/2, \mathscr{G}_2', ||\cdot||_n) \leq A\delta^{-\frac{1}{l}}$, for sufficiently large n and some positive A, which is independent of n. □

Lemma 3 Assume that Assumption 3 is met. Then for all $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top \in \mathbb{R}^n$ and $a > 0$,

$$\mathbf{P}\left(\left|\sum_{i=1}^n \epsilon_i \gamma_i\right| \geq a\right) \leq 2\exp\left\{-\frac{a^2}{8(K^2+\sigma_0^2)\sum_{i=1}^n \gamma_i^2}\right\}.$$

Proof: See the proof of Lemma 8.2 of [102]. □

Lemma 4 Assuming that Assumptions 1, 2 and 3 are met, then for some constant B depending only on K and σ_0, and for any $\delta > 0$, we have

$$\mathbf{P}\left\{\sup_{g\in\mathscr{G}^\top}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right| \geq \delta\right\} \leq 2\exp\left(-\frac{n\delta^2}{B^2R^2}\right),$$

where $\sup_{g\in B_n(g_0,\sigma)} ||g||_n \leq R$, as long as n is sufficiently large.

Proof: Let for each $i = 0, 1, \ldots$, T_i be a $2^{-i}R$-covering set of $\mathscr{G}'$, i.e., for each $g \in \mathscr{G}'$ there is a $g^i \in T_i$ such that $||g - g^i||_n \leq 2^{-i}R,\ i = 0, 1, \ldots$. Without loss of generality, we assume that $T_i \subset \mathscr{G}',\ i = 0, 1, \ldots$. Note that

$$\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \left\{g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i) - g^S(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right\}\right| \leq \sqrt{\mathrm{E}\,(\epsilon_i^2)}||g - g^S||_n \text{ almost surely}$$

for sufficiently large n, applying the strong law of large numbers. The inequality above implies that, as long as a sufficiently large S is chosen, then almost surely, we have $\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \left\{g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i) - g^S(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right\}\right| \leq \delta/2$ for sufficiently large n. Therefore, it suffices to prove an exponential inequality for

$$\mathbf{P}\left\{\sup_{g\in T}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right| \geq \delta/2\right\},$$

where $T = \cup_{i=1}^\infty T_i$.

Since $\sup_{g\in B_n(g_0,\sigma)} ||g||_n \leq R$, T_0 can be chosen as $\{0\}$. For any $j \in N^+$, $g^j = \sum_{i=1}^j (g^i - g^{(i-1)})$. Let $C_2^2 = 8(K^2+\sigma_0^2)$. Since for any $g \in T$, $\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right| \leq \sum_{j=1}^\infty \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (g^j - g^{(j-1)})\right|$, we have that for any nonnegative sequence $\{\eta_j\}$ satisfying $\sum_{j=1}^\infty \eta_j \leq 1$,

$$\begin{aligned}
\mathbf{P} &= \mathbf{P}\left\{\sup_{g\in T}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{\zeta}_i, \boldsymbol{z}_i)\right| \geq \delta/2\right\} \\
&\leq \sum_{j=1}^\infty \mathbf{P}\left\{\sup_{g\in T}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (g^j - g^{(j-1)})\right| \geq \eta_j\delta/2\right\} \\
&\leq 2\sum_{j=1}^\infty \exp\left\{2H(2^{-j}R, \mathscr{G}', ||\cdot||_n) - \frac{n\delta^2\eta_j^2}{36C_2^2 2^{-2j}R^2}\right\}.
\end{aligned}$$

The last expression comes from the fact that

$$||g^j - g^{(j-1)}||_n \leq ||g^j - g||_n + ||g - g^{(j-1)}||_n \leq 2^{-j}R + 2^{-j+1}R = 3(2^{-j}R)$$

and Lemma 3.

As shown in Lemma 2, $H(\delta, \mathscr{G}', ||\cdot||_n) \leq A\delta^{-\frac{1}{l}}\ \forall \delta > 0$, for some $A > 0$ independent of n. We have

$$\sqrt{n}\delta \geq 24C_2 \left(\int_0^R H^{\frac{1}{2}}(x, \mathscr{G}', ||\cdot||_n)dx \vee R \right),$$

for sufficiently large n. We choose

$$\eta_j = \frac{12C_2 2^{-j} R H^{\frac{1}{2}}(2^{-j}R, \mathscr{G}', ||\cdot||_n)}{\sqrt{n}\delta} \vee \frac{2^{-j}\sqrt{j}}{2E},$$

where $E = \sum_{j=1}^{\infty} 2^{-j}\sqrt{j}$. Then

$$\sum_{j=1}^{\infty} \eta_j \leq \sum_{j=1}^{\infty} \frac{12C_2 2^{-j} R H^{\frac{1}{2}}(2^{-j}R, \mathscr{G}', ||\cdot||_n)}{\sqrt{n}\delta} + \sum_{j=1}^{\infty} \frac{2^{-j}\sqrt{j}}{2E} \leq \frac{1}{2} + \frac{1}{2} = 1.$$

Note that $\eta_j \geq \frac{12C_2 2^{-j} R H^{1/2}(2^{-j}R, \mathscr{G}', ||\cdot||_n)}{\sqrt{n}\delta}$. Plugging this into the expression of $\mathbf{P}$, it follows

$$\begin{aligned}
\mathbf{P} &\leq 2\sum_{j=1}^{\infty} \exp\left\{ 2H(2^{-j}R, \mathscr{G}', ||\cdot||_n) - \frac{n\delta^2\eta_j^2}{36C_2^2 2^{-2j}R^2} \right\} \\
&\leq 2\sum_{j=1}^{\infty} \exp\left(\frac{n\delta^2\eta_j^2}{72C_2^2 2^{-2j}R^2} - \frac{n\delta^2\eta_j^2}{36C_2^2 2^{-2j}R^2} \right) \\
&= 2\sum_{j=1}^{\infty} \exp\left(-\frac{n\delta^2\eta_j^2}{72C_2^2 2^{-2j}R^2} \right) \\
&\leq 2\sum_{j=1}^{\infty} \exp\left(-\frac{n\delta^2}{72C_2^2 2^{-2j}R^2} \frac{2^{-2j}j}{4E^2} \right) \ \text{(since } \eta_j \geq \frac{2^{-j}\sqrt{j}}{2E}\text{)} \\
&= 2\sum_{j=1}^{\infty} \exp\left(-\frac{n\delta^2 j}{288C_2^2 E^2 R^2} \right) \\
&\leq 2\exp\left(-\frac{n\delta^2}{B^2R^2} \right) \ \text{for some } B > 0.
\end{aligned}$$

□

Proof of Proposition 1: T_0 is defined as $\sup\left\{(2^{-j}R)^{\frac{1}{2l}-1}\times 24C_2\left(\frac{2l}{2l-1}A(2^{-j+1}R)^{\frac{2l-1}{2l}}\vee 2^{-j+1}R\right),\ j=1,2\ldots,\right\}$. Then for $T\geq T_0$,

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{g\in\mathscr{G}^\top}\frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i g(\boldsymbol{\zeta}_i,\boldsymbol{z}_i)|}{||g||_n^{1-\frac{1}{2l}}}\geq T\right\}\\
&\leq\sum_{j=1}^{\infty}\mathbf{P}\left\{\sup_{g\in\mathscr{G}^\top,2^{-j}R<||g||_n\leq 2^{-j+1}R}|\frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i g(\boldsymbol{\zeta}_i,\boldsymbol{z}_i)|\geq T(2^{-j}R)^{1-\frac{1}{2l}}\right\}\\
&\leq 2\sum_{j=1}^{\infty}\exp\left\{-\frac{T^2(2^{-j}R)^{2-\frac{1}{l}}}{B^2R^2}\right\}\ \text{(using Lemma 4)}\\
&\leq 2\exp\left(-\frac{T^2}{C_0^2}\right),
\end{aligned}$$

for some constant $C_0>0$. □

Proof of Theorem 1: By Lemma 2, for sufficiently large n, we have

$$H\left(\delta,\left\{\frac{g-g_0}{J(g_0)+J(g)}:g\in B_n(g_0,\sigma)\right\},||\cdot||_n\right)<A\delta^{-\frac{1}{l}},\quad\forall\delta>0$$

for some constant A not depending on n. Now we can apply Proposition 1 to the class $\left\{\frac{g-g_0}{J(g_0)+J(g)}:g\in B_n(g_0,\sigma)\right\}$. Consequently,

$$\frac{(\epsilon,\hat{g}_n-g_0)_n}{||\hat{g}_n-g_0||_n^{1-\frac{1}{2l}}(J(g_0)+J(\hat{g}_n))^{\frac{1}{2l}}}=O_P(n^{-\frac{1}{2}}).\tag{4.18}$$

Incorporating (4.18) in (4.14), we have

$$||\hat{g}_n-g_0||_n^2+\tau_n^2J(\hat{g}_n)\leq O_P(n^{-\frac{1}{2}})||\hat{g}_n-g_0||_n^{1-\frac{1}{2l}}\left\{J(g_0)+J(\hat{g}_n)\right\}^{\frac{1}{2l}}+\tau_n^2J(g_0).$$

If $O_P(n^{-\frac{1}{2}})||\hat{g}_n-g_0||_n^{1-\frac{1}{2l}}\left\{J(g_0)+J(\hat{g}_n)\right\}^{\frac{1}{2l}}<\tau_n^2J(g_0)$, it follows that

$$||\hat{g}_n-g_0||_n^2+\tau_n^2J(\hat{g}_n)\leq 2\tau_n^2J(g_0);\tag{4.19}$$

otherwise,

$$||\hat{g}_n-g_0||_n^2+\tau_n^2J(\hat{g}_n)\leq O_P(n^{-\frac{1}{2}})||\hat{g}_n-g_0||_n^{1-\frac{1}{2l}}\left\{J(g_0)+J(\hat{g}_n)\right\}^{\frac{1}{2l}}.\tag{4.20}$$

Next we will verify the result for separated cases. For the case of inequality (4.19), it is trivial that

$$||\hat{g}_n-g_0||_n=O_P(\tau_n),\qquad J(\hat{g}_n)=O_P(1)J(g_0).\tag{4.21}$$

For the case of inequality (4.20), there are two possibilities.

(i) If $J(\hat{g}_n) \geq J(g_0)$, it follows that $||\hat{g}_n - g_0||_n^2 + \tau_n^2 J(\hat{g}_n) \leq O_P(n^{-\frac{1}{2}})||\hat{g}_n - g_0||_n^{1-\frac{1}{2l}} \{J(\hat{g}_n)\}^{\frac{1}{2l}}$. Then $\{J(\hat{g}_n)\}^{\frac{1}{2l}} \leq O_P(n^{-\frac{1}{4l-2}})||\hat{g}_n - g_0||_n^{\frac{1}{2l}} \tau_n^{-\frac{2}{2l-1}}$. Thus

$$||\hat{g}_n - g_0||_n^2 \leq O_P(n^{-\frac{1}{2}})||\hat{g}_n - g_0||_n^{1-\frac{1}{2l}} \{J(\hat{g}_n)\}^{\frac{1}{2l}} \leq O_P(n^{-\frac{l}{2l-1}})||\hat{g}_n - g_0||_n \tau_n^{-\frac{2}{2l-1}}.$$

In other words,

$$||\hat{g}_n - g_0||_n = O_P(n^{-\frac{l}{2l-1}})\tau_n^{-\frac{2}{2l-1}}, \quad J(\hat{g}_n) = O_P(n^{-\frac{2l}{2l-1}})\tau_n^{-\frac{4l+2}{2l-1}}. \tag{4.22}$$

(ii) If $J(\hat{g}_n) < J(g_0)$, it follows that $||\hat{g}_n - g_0||_n^2 + \tau_n^2 J(\hat{g}_n) \leq O_P(n^{-\frac{1}{2}})||\hat{g}_n - g_0||_n^{1-\frac{1}{2l}} \{J(g_0)\}^{\frac{1}{2l}}$. After some simple algebra, we have

$$||\hat{g}_n - g_0||_n = O_P(n^{-\frac{l}{2l+1}}) \{J(g_0)\}^{\frac{1}{2l+1}}, \quad J(\hat{g}_n) = J(g_0)O_P(1). \tag{4.23}$$

When $J(g_0) = J(f_0) > 0$ and $\tau_n^{-1} = n^{\frac{l}{2l+1}} \{J(f_0)\}^{\frac{2l-1}{4l+2}}$, then we obtain the same result from (4.21), (4.22) and (4.23). To be more specific, $||\hat{g}_n - g_0||_n = O_P(n^{-\frac{l}{2l+1}}) \{J(f_0)\}^{\frac{1}{2l+1}}$ and $J(\hat{f}_n) = J(f_0)O_P(1)$. When $J(g_0) = J(f_0) = 0$, then both $O_P(n^{-\frac{1}{2}})||\hat{g}_n - g_0||_n^{1-\frac{1}{2l}} \{J(g_0) + J(\hat{g}_n)\}^{\frac{1}{2l}} < \tau_n^2 J(g_0)$ and $J(\hat{g}_n) < J(g_0)$ are impossible, which indicate that we only need to consider (4.22) under this circumstance. When $\tau_n^{-1} = n^{1/4}$, $||\hat{g}_n - g_0||_n = O_P(n^{-\frac{1}{2}})$ and $J(\hat{f}_n) = O_P(n^{-\frac{1}{2}})$. □

In Corollary 1, we only need to show that $\hat{f}_n$ and $\hat{\boldsymbol{\alpha}}$ defined above satisfy Corollary 1 as well since the estimated intercept $\hat{a}$ converges to a_0 with a rate of $O_P(n^{-\frac{1}{2}})$, as indicated at the very beginning. To prove Corollary 1, we need to quantify the ratio of $||\cdot||_n$ and $||\cdot||_{2,Q}$ norm for both $\hat{f}_n$ and $\hat{g}_n$. Entropy with bracketing is an important tool in studying magnitude of the ratio. Let $N_B(\delta, \mathscr{G}, ||\cdot||_{2,Q})$ be the smallest value of N for which there exist pairs of functions $\{[g_j^L, g_j^U]\}_{j=1}^N$ such that $||g_j^U - g_j^L||_{2,Q} \leq \delta$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathscr{G}$, there exists $j = j(g) \in \{1, \ldots, N\}$ such that $g_j^L \leq g \leq g_j^U$. Then $H_B(\delta, \mathscr{G}, ||\cdot||_{2,Q}) = \log N_B(\delta, \mathscr{G}, ||\cdot||_{2,Q})$ is called the δ-entropy with bracketing of $\mathscr{G}$ (for the $L_2(Q)$ metric). Following lemmas are needed to compute the ratio of $||g||_n$ and $||g||_{2,Q}$ for any $g \in \mathscr{G}$.

Lemma 5 For all $\delta > 0$, $H_B(\delta, \mathscr{G}, ||\cdot||_{2,Q}) \leq H(\delta/2, \mathscr{G}, ||\cdot||_\infty)$

Proof: See Lemma 2.1 of [102]. □

Lemma 6 Let $\mathscr{A}$ denote a collection of functions defined on $\mathscr{X}$. Suppose that $\mathscr{A}$ is uniformly bounded, i.e., $\sup_{a \in \mathscr{A}} ||a||_\infty \leq M$ for some constant M, and that for some $0 < \nu < 2$,

$\sup_{\delta>0} \delta^{\nu} H_B(\delta, \mathscr{A}, ||\cdot||_{2,Q}) < \infty$. Then for all $\eta > 0$ there exists a constant C such that

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{a\in\mathscr{A}, ||a||_{2,Q} > Cn^{-1/(2+\nu)}} \left| \frac{||a||_n}{||a||_{2,Q}} - 1 \right| > \eta \right) = 0.$$

See Theorem 2.3 of [65] or [101], Lemma 6.3.4.

Before proving the corollary, we restate the extra assumption.

Assumption 5 *The support of $\boldsymbol{z}$ is compact in $\mathbb{R}^p$.*

Proof of Corollary 1 We first need to show that $\hat{g}_n$ is bounded. As shown above, $||\frac{\hat{f}_n}{1+J(\hat{f}_n)}||_\infty \leq C$ for some constant C and $J(\hat{f}_n) = O_P(1)$ with an suitable τ_n. Therefore, $||\hat{f}_n||_\infty = O_p(1)$. Additionally, provided that Assumption (5) holds, then $||\hat{\boldsymbol{\alpha}}^\top \tilde{\boldsymbol{z}}||_\infty$ is bounded in probability as well, since it has been shown that $\hat{\boldsymbol{\alpha}} = O_P(1)$. Thus $||\hat{g}_n||_\infty \leq ||\hat{f}_n||_\infty + ||\hat{\boldsymbol{\alpha}}^\top \tilde{\boldsymbol{z}}||_\infty = O_P(1)$ in Lemma 2. We henceforth consider a subset of $\mathscr{G}$, $\{g : g \in \mathscr{G}, ||g||_\infty \leq C, J(g) \leq C\}$, which is still denoted as $\mathscr{G}'$. Similarly, let $\mathscr{G}_1'$ denote $\{g_1 : g_1(\boldsymbol{\zeta}, \boldsymbol{z}) = \sum_{j=1}^d f_j(\zeta_j), \sum_{i=1}^n f_j(\zeta_{ij}) = 0, j = 1, \ldots, d, ||g_1||_\infty \leq C_f, J(g_1) \leq C\}$, where C_f is a positive constant, and $\mathscr{G}_2'$ for $\{g_2 : g_2(\boldsymbol{\zeta}, \boldsymbol{z}) = \boldsymbol{\alpha}^\top \tilde{\boldsymbol{z}}, ||\boldsymbol{\alpha}||_E \leq C_{\boldsymbol{\alpha}}\}$ with $C_{\boldsymbol{\alpha}}$ being a positive constant that does not depend on $\boldsymbol{\alpha}$.

Next, we shall provide a uniform bound for both $||g_1||_n/||g_1||_{2,Q}$, $g_1 \in \mathscr{G}_1'$ and $||g||_n/||g||_{2,Q}$, $g \in \mathscr{G}'$. For the former one, Lemma 5.6 of [102] is employed. As shown in Lemma 2, $H(\delta, \mathscr{G}_1', ||\cdot||_n) \leq A\delta^{-\frac{1}{l}}$ for some constant A. Take $\delta_n = (2A)^{l/(2l+1)} n^{-l/(2l+1)}$ and $H(\delta) = \delta^{-\frac{1}{l}}$. Then $n\delta_n^2 \to \infty$, and $n\delta_n^2 = 2A\delta_n^{-\frac{1}{l}} = 2AH(\delta_n)$ for all n. Thus we have

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{g_1\in\mathscr{G}_1'} \frac{||g_1||_n}{||g_1||_{2,Q} \vee \delta_n} > 14 \right) \leq \limsup_{n\to\infty} 4\mathbf{P}\left\{ \sup_{u>0} \frac{H(u, \mathscr{G}_1', ||\cdot||_n)}{H(u)} > A \right\} = 0 \quad (4.24)$$

Inequality (4.24) implies that with probability arbitrarily close to 1,

$$||\hat{f}_n - \tilde{f}_0||_n^2 \leq \max\left\{ 196||\hat{f}_n - \tilde{f}_0||_{2,Q}^2, O(n^{-2l/(2l+1)}) \right\}, \quad (4.25)$$

for sufficiently large n. We use Lemma 6 to derive a uniform bound on $||g||_n/||g||_{2,Q}$ for $g \in \mathscr{G}'$. Based on Lemma 5 and combining (4.16), $H_B(\delta, \mathscr{G}_1', ||\cdot||_{2,Q}) \leq H(\delta/2, \mathscr{G}_1', ||\cdot||_\infty)$ $\leq A_1\delta^{-\frac{1}{l}}$ for some constant A_1. Since the support of $\boldsymbol{z}$ is compact, it is straightforward that $H_B(\delta, \mathscr{G}_2', ||\cdot||_{2,Q}) \leq H(\delta/2, \mathscr{G}_2', ||\cdot||_\infty) \leq A_2 \log(1/\delta)$ for some constant A_2. Therefore, $H_B(\delta, \mathscr{G}', ||\cdot||_{2,Q}) \leq A\delta^{-\frac{1}{l}}$ for some constant A. Taking $\mathscr{A} = \mathscr{G}'$ and $\nu = \frac{1}{l}$, then the condition $\sup_{\delta>0} \delta^{\nu} H_B(\delta, \mathscr{A}, ||\cdot||_{2,Q}) < \infty$ is satisfied in Lemma 6. We can derive from Lemma 6, that with probability arbitrarily close to 1,

$$||\hat{g}_n - g_0||_{2,Q}^2 \leq \max(\eta_1||\hat{g}_n - g_0||_n^2, O(n^{-2l/(2l+1)})) = O_P(n^{-2l/(2l+1)}), \quad (4.26)$$

for some constant η_1 and sufficiently large n.

Note that

$$\begin{aligned}
||\hat{g}_n - g_0||_{2,Q}^2 &= ||\hat{f}_n(\boldsymbol{\zeta}) - \tilde{f}_0(\boldsymbol{\zeta}) + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top (\boldsymbol{z} - \bar{\boldsymbol{z}})||_{2,Q}^2 \\
&= ||\hat{f}_n(\boldsymbol{\zeta}) - \tilde{f}_0(\boldsymbol{\zeta}) + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top \{\boldsymbol{z}^\star + h(\boldsymbol{\zeta}) - \bar{\boldsymbol{z}}\}||_{2,Q}^2 \\
&= ||\hat{f}_n(\boldsymbol{\zeta}) - \tilde{f}_0(\boldsymbol{\zeta}) + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top \{h(\boldsymbol{\zeta}) - \bar{\boldsymbol{z}}\}||_{2,Q}^2 + ||(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top \boldsymbol{z}^\star||_{2,Q}^2 \\
&= O_P(n^{-2l/(2l+1)}). \qquad (4.27)
\end{aligned}$$

The last equation holds according to (4.26). Since $||(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top \boldsymbol{z}^\star||_{2,Q}^2 \geq \Lambda_{\min}\{\mathrm{Var}(\boldsymbol{z}^\star)\}\, ||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||_E^2$, $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||_E = O_P(n^{-l/(2l+1)})$ based on (4.27) when Assumption (2) is met.

Now we can verify the consistency of $\hat{f}_n$. Take $C^\star = \max_{1 \leq j \leq p} |z_j|$. Then $C^\star < \infty$ when Assumption (5) is satisfied. Given that $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||_E = O_P(n^{-l/(2l+1)})$ and

$$\begin{aligned}
||\hat{g}_n - g_0||_n^2 &= ||\hat{f}_n - \tilde{f}_0 + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top \tilde{\boldsymbol{z}}||_n^2 \\
&\geq ||\hat{f}_n - \tilde{f}_0||_n^2 + \frac{2}{n}\sum_{i=1}^{n} \left\{\hat{f}_n(\boldsymbol{\zeta}_i) - \tilde{f}_0(\boldsymbol{\zeta}_i)\right\} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^\top (\boldsymbol{z}_i - \bar{\boldsymbol{z}}) \\
&\geq ||\hat{f}_n - \tilde{f}_0||_n^2 - 4C^\star ||\hat{f}_n - \tilde{f}_0||_n ||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||_E
\end{aligned}$$

we have

$$||\hat{f}_n - \tilde{f}_0||_n^2 \leq 4C^\star ||\hat{f}_n - \tilde{f}_0||_n O_P(n^{-l/(2l+1)}) + ||\hat{g}_n - g_0||_n^2$$

Therefore, in either Case (i), $0 < J(f_0) < \infty$ and $\tau_n^{-1} = n^{\frac{l}{2l+1}} \{J(f_0)\}^{\frac{2l-1}{4l+2}}$, or Case (ii), $J(f_0) = 0$, and $\tau_n^{-1} = n^{1/4}$,

$$||\hat{f}_n - \tilde{f}_0||_n = O_P(n^{-l/(2l+1)})$$

□

Chapter 5

Sparse Functional Additive Models

5.1 Introduction

Functional data analysis has become an important tool for dealing with data collected over multiple time points, spatial locations, or other continua. A fundamental problem in functional data analysis is how to model the relationship between a scalar response of interest and a functional predictor. For instance, the Tecator data (see Section 5.5.2) consists of 240 meat samples; each of them comprises the spectrum of absorbance, and three contents: water, fat and protein. Researchers have been concerned about how to use the spectrum of absorbance, which can be treated as a functional predictor, to predict one of the three contents. Functional linear regression (FLR) is a conventional and interpretable model for predicting a scalar response from a functional predictor.

In FLR, the relationship between a scalar response and a functional predictor is modelled in a linear form. Hence the key problem in fitting FLR is to estimate the coefficient function of the functional predictor. There has been extensive research to address this problem. For example, [70] considered representing the coefficient function in terms of Fourier basis functions or the eigenfunctions of the estimated covariance function of the functional predictor. The coefficients of the Fourier basis functions were then obtained from solving a functional estimating equation. [86] suggested using spline basis functions to represent the coefficient function and then solving a regularized regression problem, in which the roughness of the spline representation is penalized to obtain a smooth estimate of the coefficient function. [61] proposed a local sparse estimator for the coefficient function to enhance the interpretability of FLR. A comprehensive introduction to FLR can be found in [47] and [69].

Although the studies aforementioned have proposed various estimating methods to fit a FLR model and established some appealing properties of the corresponding estimators, in practice, applications of FLR are sometimes restricted due to its simple linear form. Similar to the multiple linear model, which in some cases may not adequately describe the relationship between a scalar response and scalar covariates, FLR can also suffer from an

inadequate flexibility for modelling the relationship between a scalar response and a functional predictor. This phenomenon has been noted by many researchers. For instance, [114] extended the FLR model to the case when the scalar response depends on a polynomial of the functional predictor and they mainly focused on the quadratic case. [13] considered using a nonparametric link to connect the scalar response and the functional linear form. A class of flexible functional nonlinear regression models were proposed by [71] by using continuously additive models to characterize the relationship between a functional predictor and a scalar response. Nonlinear and/or nonparametric functional regression models can somewhat address the issue of inadequate fit caused by FLR (see [13], [71], [72]). However, they can lead to other issues such as over-flexibility and a lack of stability ([121]). [89] summarized some of main approaches of regressing a scalar response on a functional predictor. In this chapter, we aim to propose a functional regression model which can achieve a satisfactory trade-off between flexibility and simplicity.

[121] proposed an extended functional additive model, in which the scalar response of interest depends on a transformation of the leading functional principal component (FPC) scores. They assumed that some additive components were vanishing and the nonvanishing components were smooth functions for the sake of simplicity and interpretability while retaining flexibility. To achieve this goal, they adopted the regularization scheme of component selection and smoothing operator (COSSO) proposed by [60], which can select and smooth components simultaneously. While this model turns out to achieve a better trade-off between flexibility and simplicity compared with many other functional regression models, the estimation procedure seems to suffer from several drawbacks. First, only estimation consistency is guaranteed for the proposed estimator. Whether selection consistency holds for this estimator remains an open problem. Another drawback is associated with computational complexity. As noted by [119], when a full basis is employed, the complexity of the algorithm is $O(n^3)$, where n is the sample size. To reduce the computational burden, [119] suggested using a subset basis algorithm instead, which was computationally much more efficient than the full basis algorithm. [121] seemed to ignore this computational issue when implementing COSSO to fit the proposed model. The computational complexity is demonstrated in simulation studies.

To overcome the drawbacks of the method proposed by [121], we propose a method to estimate the extended functional additive model. In contrast to representing nonparametric additive components in the framework of RKHS ([121]), we use B-spline basis functions to represent these components, which are easier to understand and implement. Then selecting nonzero components is equivalent to selecting nonzero coefficients of the B-spline basis functions. The group LASSO method ([118]) has been shown to perform well when selecting grouped variables for accurate prediction in both theory and application. Because an additive component corresponds to a vector of coefficients, which can be treated as a group of variables, we employ the group LASSO method to select nonzero vectors of coefficients.

The adaptive group LASSO method is then applied to allow for different shrinkages for different vectors of coefficients. This modification can yield a more accurate estimate of the coefficient vectors, which then leads to a better estimate for the additive components. This method enables us to attain our goal of obtaining a parsimonious model via component selection.

Nevertheless, the estimated nonzero components can be wiggly due to the representation of the additive components using a large number of B-spline basis functions. It may impair predictive performance, which will be demonstrated in simulation studies in Section 5.4. Thus we suggest refining the selected components via smoothing splines. This extra smoothing step can improve the prediction accuracy of the estimator obtained from the adaptive group LASSO, which will be demonstrated in our simulation studies.

This chapter makes three main contributions. First, compared with traditional FLR models, our proposed model provides a better trade-off between flexibility and simplicity in modelling the effect of a functional predictor. By selecting and smoothing nonzero components, our proposed method obtains an estimator which has better prediction accuracy. Second, unlike the COSSO regularization scheme adopted in [121], we employ group LASSO to select components and use the smoothing spline method to smooth nonzero components. As a result, our proposed estimation method is easy to understand and implement. Last but not least, we give both theoretical and empirical demonstrations of the selection consistency and estimation consistency of our proposed estimator, while [121] only provided a theoretical proof of the estimation consistency of their estimator.

The remainder of this chapter is organized as follows. Section 5.2 introduces a sparse functional additive model and our method to estimate the additive components in the model. Section 5.3 establishes the selection consistency and the estimation consistency of our proposed estimator. The finite-sample performance of the estimator is investigated empirically in Section 5.4, where we conduct simulation studies to compare our proposed estimator and other conventional methods. In Section 5.5, our method is demonstrated by analyzing two real data examples. In Section 5.6, we give some conclusions about our method. The procedures to estimate the FPC scores and proofs of the main results in Section 5.3 are provided in the Appendix.

5.2 Model and Estimation Method

5.2.1 Sparse Functional Additive Model

Suppose that $\{X_i(t), y_i\}_{i=1}^n$ are independent and identically distributed (iid) observations from $\{X(t), Y\}$, where $X(t)$ is a random function and Y is a scalar random variable. We assume $X(t)$ to be a square integrable stochastic process over a compact interval $\mathcal{I} = [0, T]$, i.e., $\mathrm{E}\{\int_{\mathcal{I}} X^2(t)dt\} < \infty$. Let $m(t)$ and $G(s,t)$ denote the mean function and covariance function of $X(t)$, respectively. According to Mercer's theorem, $G(s,t)$ can be represented

as $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where λ_k is a nonnegative eigenvalue and $\phi_k(t)$ is the corresponding eigenfunction. For the sake of identifiability, we postulate that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Additionally, $\{\phi_k\}_{k=1}^{\infty}$ is assumed to be a complete orthonormal basis of the space $L^2(\mathcal{I})$, the collection of all square integrable functions on $\mathcal{I}$. Then the stochastic process $X(t)$ admits the Karhunen-Loève expansion:

$$X(t) = m(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \tag{5.1}$$

where $\xi_k = \int_{\mathcal{I}} (X(t) - m(t))\phi_k(t)dt,\ k = 1, \ldots,$ is called the k-th FPC score. The FPC score satisfies $\mathrm{E}(\xi_k \xi_{k'}) = \lambda_k$ if $k = k'$ and 0 otherwise.

In FLR, Y is treated as the response and $X(t)$ is the functional predictor. Furthermore, the relationship between Y and $X(t)$ is modelled in a linear form:

$$y_i = \int_{\mathcal{I}} X_i(t) b(t) dt + \epsilon_i,$$

where ϵ_i's denote random errors with mean 0 and variance σ_ϵ^2. Given the representation of $X(t)$ in (5.1), we have $y_i = a + \sum_{k=1}^{\infty} b_k \xi_{ik} + \epsilon_i$, where $a = \int_{\mathcal{I}} m(t)b(t)dt$, ξ_{ik} denotes the k-th FPC score of $X_i(t)$, and $b_k = \int_{\mathcal{I}} \phi_k(t)b(t)dt,\ k \geq 1$. To address the curse of dimensionality, a truncated model is usually adopted such that Y only depends on the first d FPC scores. In other words, we get a truncated linear model: $y_i = a + \sum_{j=1}^{d} b_j \xi_{ij} + \epsilon_i$. In practice, d is chosen as the smallest number of FPCs which can explain over 99.9% of the total variability of the functional predictor $X(t)$. As noted by [121], this choice can, to some extent, circumvent neglecting those FPC scores which play a negligible role in capturing the variability of the functional predictor but are relevant in predicting the response. This truncated model is slightly restrictive since an explicit parametric form is assumed between the response and the leading FPC scores. The linearity assumption is likely to be violated in substantial practical scenarios.

In light of the idea proposed by [42], and the fact that ξ_j's are mutually uncorrelated, a nonparametric functional additive model was proposed by [72] to describe the relationship between the response and the first d FPC scores

$$y_i = a + \sum_{j=1}^{d} f_j(\xi_{ij}) + \epsilon_i, \tag{5.2}$$

where we call f_j the j-th component in the nonparametric functional additive model.

FPC scores usually cannot be observed directly. Therefore we need firstly to estimate FPC scores from the observed functional data which may be subject to measurement errors. We assume that $W_{ij} = X_i(t_{ij}) + e_{ij}$, where W_{ij} denotes the observation of the process $X_i(t)$ made at time point $t_{ij},\ j = 1, \ldots, N_i,\ i = 1, \ldots, n$ and e_{ij} denotes the measurement error

and is assumed to be independent of $X_i(t)$. Functional principal component analysis (FPCA) is implemented to estimate FPC scores, denoted by $\hat{\xi}_{ij}$'s. The details of this procedure can be found in the Appendix.

We first scale the FPC scores to [0, 1] via a transformation function F. One possible strategy is to apply the cumulative distribution function (cdf) $F(z|\lambda_j)$ of the Normal$(0, \lambda_j)$ on ξ_j, where λ_j is the eigenvalue of the covariance function $G(s,t)$, and $\lambda_j = Var(\xi_j)$. We define ζ_j to be the j-th scaled FPC score: $\zeta_j = F(\xi_j|\lambda_j),\ j = 1, \ldots, d$. Then the estimated scaled FPC scores are given as $\hat{\zeta}_{ij} = F(\hat{\xi}_{ij}|\hat{\lambda}_j), j = 1, \ldots, d, i = 1, \ldots, n$. Assumption B in Section 5.3 gives more general conditions on the transformation function, F. We still use f_j for the j-th component in the nonparametric functional additive model when ξ_j's are replaced by ζ_j's.

The nonparametric functional additive model (5.2) can now be expressed as

$$y_i = a + \sum_{j=1}^{d} f_j(\zeta_{ij}) + \epsilon_i. \tag{5.3}$$

To make the model identifiable, we assume that $\mathrm{E}\{f_j(\zeta_j)\} = 0,\ j = 1, \ldots, d$. Models with a parsimonious structure are preferable in practice. Thus we assume that some components, f_j's are vanishing and the rest of the components are nonzero and smooth. Model (5.3) is called a sparse functional additive model in this chapter.

B-spline functions, due to their nice properties ([21]), have been widely used in estimating unknown functions (see [98], [99], [48], etc). In this chapter we also employ B-spline functions to estimate the additive components in Model (5.3). We present here a brief overview of B-splines. For more information, see [21]. Let $0 = \tau_0 < \tau_1 < \cdots < \tau_{L_n} < \tau_{L_n+1} = 1$ be the breakpoints which separate the interval [0, 1] into $L_n + 1$ subintervals. We assume that $L_n = O(n^{\alpha})$, where $0 < \alpha < 0.5$, and define $\delta_n = \max_{0 \le m \le L_n} |\tau_{m+1} - \tau_m| = O(n^{-\alpha})$. Let c_1 be a constant, independent of n, such that $\delta_n < c_1 \min_{0 \le m \le L_n} |\tau_{m+1} - \tau_m|$. Let $\mathscr{S}_n$ be the space of polynomial splines of order l, which is one more than the degree of polynomials, on [0, 1] consisting of functions s satisfying: (i) s is a polynomial of order l at each subinterval $[\tau_m, \tau_{m+1}], m = 0, \ldots, L_n$; (ii) for $0 \le l^{\star} \le l - 2$, the $l^{\star}$-th order derivative of s is continuous on [0, 1]. Then there exist $m_n = L_n + l$ normalized B-spline basis functions $\{B_k, 1 \le k \le m_n\}$ bounded by 0 and 1 on [0, 1], such that any $f \in \mathscr{S}_n$ can be written as

$$f_j(x) = \sum_{k=1}^{m_n} \beta_{jk} B_k(x), \quad j = 1, 2, \ldots, d, \tag{5.4}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})^{\top}$ is the spline coefficient vector. Now selecting nonzero components $f_j(\cdot)$ for Model (5.3) amounts to selecting nonzero $\boldsymbol{\beta}_j$.

5.2.2 Group LASSO

Accounting for the fact that $\mathrm{E}\{f_j(\zeta_j)\} = 0,\ j = 1, \ldots, d$, we define $\psi_{jk}(x) = B_k(x) - \frac{1}{n}\sum_{i=1}^{n} B_k(\hat{\zeta}_{ij}),\ k = 1, \ldots, m_n,\ j = 1, \ldots, d$. For brevity, $\psi_{jk}(x)$ is denoted by $\psi_k(x)$ without causing any confusion. Thus $\sum_{i=1}^{n}\psi_k(\hat{\zeta}_{ij}) = 0,\ j = 1, \ldots, d$. The estimated intercept in Model (5.3) is given as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Let $\boldsymbol{Z}_{ij} = (\psi_1(\hat{\zeta}_{ij}), \ldots, \psi_{m_n}(\hat{\zeta}_{ij}))^T$, $\boldsymbol{Z}_j = (\boldsymbol{Z}_{1j}, \ldots, \boldsymbol{Z}_{nj})^T$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_d)$. Similarly, define $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T)^T$, where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})^T$, and $\boldsymbol{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$. Nonzero $\boldsymbol{\beta}_j$'s in Model (5.3) can be selected and estimated using the group LASSO ([118]), in which the corresponding estimate $\tilde{\boldsymbol{\beta}}$ minimizes

$$D_1(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_1\sum_{j=1}^{d}||\boldsymbol{\beta}_j||_2. \tag{5.5}$$

In (5.5), the positive tuning parameter λ_1 determines the magnitude of shrinkage and $||\cdot||_2$ denotes the Euclidean norm of a vector in $\mathbb{R}^{m_n}$. If $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jm_n})^T$, then the corresponding estimate of f_j is denoted by $\tilde{f}_j$, which equals to $\sum_{k=1}^{m_n}\tilde{\beta}_{jk}\psi_k(x)$. Cross validation is employed to choose an "optimal" λ_1, which is chosen to minimize the cross-validation error.

5.2.3 Adaptive Group LASSO

The group LASSO method penalizes each $\boldsymbol{\beta}_j$ equally in (5.5), which may not be an optimal treatment. To account for different impact on the response of different ζ_j's, we propose an adaptive group LASSO method, which is similar in spirit to the adaptive LASSO method proposed by [122]. More explicitly, a weight vector $(w_1, \ldots w_d)$, which can produce different shrinkages for different $\boldsymbol{\beta}_j$'s, is introduced. Given $\tilde{\boldsymbol{\beta}}$ estimated from group LASSO, for $j = 1, \ldots, d$, w_j is set to be $||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$ if $||\tilde{\boldsymbol{\beta}}_j||_2 > 0$ and ∞ otherwise. Then the adaptive group LASSO estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}$, is obtained by minimizing

$$D_2(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_2\sum_{j=1}^{d}w_j||\boldsymbol{\beta}_j||_2, \tag{5.6}$$

where λ_2 denotes a penalty parameter that can be determined by cross validation. Then the corresponding estimate of $f_j(x)$, denoted by $\hat{f}_j(x)$, can be represented in terms of $\boldsymbol{\psi}_j(x) = (\psi_{j1}(x), \ldots, \psi_{jm_n}(x))^T$, i.e., $\hat{f}_j(x) = \widehat{\boldsymbol{\beta}}_j^T\boldsymbol{\psi}_j(x),\ j = 1, \ldots d$. If $\widehat{\boldsymbol{\beta}}_j = \mathbf{0}$ for some j, then the estimate, $\hat{f}_j$, is also zero.

5.2.4 Smoothing Spline Method

When a large number of B-spline basis functions are employed to estimate f_j, then the adaptive group LASSO estimate may be wiggly. Further smoothing for nonzero estimates obtained from adaptive group LASSO is indispensable if it is the case. This concern was

also discussed in [111]. To allow for different roughness penalties for different nonzero components, we propose a smoothing spline method. The weight is defined as $w_j = ||\widehat{\boldsymbol{\beta}}_j||_2^{-1}$, where $j \in S$, and $S = \{j : \widehat{\boldsymbol{\beta}}_j \neq \mathbf{0}\}$ is the set of nonzero components. In particular, the updated estimate of $\boldsymbol{\beta}_j$ is obtained from the smoothing spline method by minimizing

$$D_3(\boldsymbol{\beta}) = (\boldsymbol{y} - \sum_{j\in S} \boldsymbol{Z}_j\boldsymbol{\beta}_j)^T(\boldsymbol{y} - \sum_{j\in S} \boldsymbol{Z}_j\boldsymbol{\beta}_j) + \lambda_3 \sum_{j\in S} w_j \int_0^1 \{f_j''(\zeta_j)\}^2 d\zeta_j, \tag{5.7}$$

where λ_3 denotes the smoothing parameter. The roughness penalty term $\int_0^1 \{f_j''(\zeta_j)\}^2 d\zeta_j = \boldsymbol{\beta}_j^T \boldsymbol{Q}_j \boldsymbol{\beta}_j$, where $\boldsymbol{Q}_j$ is an $m_n \times m_n$ penalty matrix with the pq-th element $\boldsymbol{Q}_j^{pq} = \int_0^1 B_p''(\zeta_j) B_q''(\zeta_j) d\zeta_j$. When the second derivative of $f_j(\zeta_j)$ does not exist, the penalty matrix $\boldsymbol{Q}_j$ can be replaced by the difference matrix introduced by [26]. Minimization of (5.7) is equivalent to a classical smoothing spline problem except that there is a weight vector in this problem. Let $\sum_{j\in S} \boldsymbol{Z}_j\boldsymbol{\beta}_j = \boldsymbol{Z}_S\boldsymbol{\beta}$, where $\boldsymbol{Z}_S = (\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_{|S|}}) \in \mathbb{R}^{n\times m_n|S|}$, $i_1, \ldots, i_{|S|}$ are all elements of S and $|S|$ denotes the cardinality of the set S. Let $\boldsymbol{Q} = \text{diag}(w_{i_1}\boldsymbol{Q}_{i_1}, \ldots, w_{i_{|S|}}\boldsymbol{Q}_{i_{|S|}})$. Then the estimate of $\boldsymbol{\beta}$, still denoted by $\widehat{\boldsymbol{\beta}}$, is given as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}_S^T\boldsymbol{Z}_S + \lambda_3\boldsymbol{Q})^{-1}\boldsymbol{Z}_S^T\boldsymbol{y}$. The corresponding estimate of f_j is $\hat{f}_j = \widehat{\boldsymbol{\beta}}_j^T \boldsymbol{\psi}_j(x),\ j \in S$.

The smoothing parameter λ_3 can be determined by the generalized cross-validation (GCV) measure. For a given λ_3, the corresponding measure can be expressed as

$$\text{GCV}(\lambda_3) = \frac{n \cdot \text{SSE}}{(n - df(\lambda_3))^2},$$

where $\text{SSE} = (\boldsymbol{y} - \boldsymbol{Z}_S\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}_S\widehat{\boldsymbol{\beta}})$ and $df(\lambda_3) = \text{trace}(\boldsymbol{Z}_S(\boldsymbol{Z}_S^T\boldsymbol{Z}_S + \lambda_3\boldsymbol{Q})^{-1}\boldsymbol{Z}_S^T)$. The optimal smoothing parameter is chosen to minimize the GCV measure. Our whole estimating procedure is called Components Selection and Smoothing in a sparse Functional Additive Model (CSS-FAM in short) in this chapter.

5.3 Theoretical Properties

To ensure that the estimated scaled FPC scores, $\hat{\boldsymbol{\zeta}}$, are consistent estimators of the true scaled FPC scores, we need to impose some regularity conditions on the design of the functional predictor $X(t)$. The following conditions follow [121]. As stated in Section 5.2.1, $\{t_{ij}, j = 1, \ldots, N_i; i = 1, \ldots, n\} \subset \mathcal{I}$ denote the time points when the functional predictor $X_i(t)$ is observed. We assume that $t_{i0} = 0$ and $t_{iN_i} = T$ for each $X_i(t)$. Let $\mathcal{I}_\tau = [-\tau, T + \tau]$ for some $\tau > 0$, and let h_i and $K(\cdot)$ denote the bandwidth and the kernel function used in smoothing the i-th trajectory, respectively. Note that the same kernel function is employed in the local linear smoother for each trajectory, when estimating FPC scores. Below is a list of regularity conditions which can guarantee that the estimated FPC scores and eigenvalues of the covariance function of $X(t)$ converge in probability to the corresponding population values.

Assumption A

(A1) $X(t)$ has a continuous second derivative on $\mathcal{I}_d$ with probability 1 and for $k = 0, 2$, $\int \mathrm{E}\{X^{(k)}(t)\}^4 dt < \infty$. The measurement errors e_{ij}'s of $X_i(t)$ satisfy $\mathrm{E}(e_{ij}^4) < \infty$ and they are identically and independently distributed.

(A2) We define T_n to be the lower bound of the number of observations for each trajectory $X_i(t)$. As $n \to \infty$, $T_n \to \infty$. Let $\triangle_i$ denote the largest time difference of two consecutive observations for each trajectory $X_i(t)$, i.e., $\triangle_i = \max\{t_{ij} - t_{i,j-1} : j = 1, \ldots, N_i\}$. The maximal value of these satisfies $\max_i \triangle_i = O(T_n^{-1})$.

(A3) There is a sequence $b_n \to 0$, such that $c_1 b_n \leq \min_i h_i \leq \max_i h_i \leq c_2 b_n$ for some constants $c_2 \geq c_1 > 0$ as $n \to \infty$. In addition, b_n and T_n satisfy that $(T_n b_n)^{-1} + b_n^4 + T_n^{-2} = O(n^{-1})$.

(A4) The kernel function $K(\cdot)$ has a compact support and satisfies $|K(s) - K(t)| \leq C|s - t|$ for $s,\ t$ in its domain and some positive constant C.

For Model (5.3), let A_1 and A_0 denote the set of non-vanishing and vanishing components, respectively; i.e., $A_1 = \{j : f_j \neq 0, j = 1, \ldots, d\}$ and $A_0 = \{j : f_j \equiv 0, 1 \leq j \leq d\}$. Regarding the transformation function $F(x|\lambda)$, a cdf with variance λ, we make the following assumptions.

Assumption B

(B1) The transformation function $F(x|\lambda)$ is differentiable at x and λ. Furthermore, there exist a positive constant C and a negative constant γ, such that $\frac{\partial F(x|\lambda)}{\partial x} \leq C\lambda^\gamma$ and $\frac{\partial F(x|\lambda)}{\partial \lambda} \leq C\lambda^\gamma |x|$.

(B2) The cdf of each scaled score ζ_j is absolutely continuous and there exist positive constants C_1 and C_2 such that the probability density function of ζ_j, g_j, satisfies $C_1 \leq g_j(x) \leq C_2$ for $x \in [0, 1]$ and $j \in A_1$.

Assumption (B1) is from [121] as well. Together with Assumptions (A1)-(A4), it can guarantee that the $\hat{\zeta_j}$ is a consistent estimator of ζ_j, $1 \leq j \leq d$. Assumption (B2) is a standard assumption in nonparametric additive models according to [98].

Define $||f||_2 = \{\int_0^1 f^2(x)dx\}^{1/2}$ whenever the integral is finite. Let $L > 0$, r be a nonnegative integer, and $\nu \in (0, 1]$ such that $\rho = r + \nu > 0.5$. Let $\mathscr{F}$ be the class of functions h on $[0, 1]$ whose r-th derivative exists and satisfies the Hölder condition with exponent ν: $|h^{(r)}(s) - h^{(r)}(t)| \leq L|s - t|^\nu$ for any $0 \leq s, t \leq 1$. Other standard assumptions for additive nonparametric models (see [48]) include:

Assumption C

(C1) $\min_{j \in A_1} ||f_j|| \geq c_f$ for some $c_f > 0$.

(C2) The random variables $\epsilon_1, \ldots \epsilon_n$ are iid with mean 0 and variance σ_ϵ^2. Furthermore, the tail probability satisfies $P(|\epsilon_1| > x) \leq K \exp(-Cx^2)$, for $\forall x \geq 0$ and some constants C and K.

(C3) $\mathrm{E}\{f_j(\zeta_j)\} = 0$ and $f_j \in \mathscr{F}$, $j \in A_1$.

The following proposition explains why it is reasonable to employ B-spline functions to approximate each nonparametric component f_j in Model (5.3). To guarantee that B-spline functions in $\mathscr{S}_n$ can provide a satisfactory approximation of functions in $\mathscr{F}$, throughout the chapter we assume that l, the order of polynomial functions in $\mathscr{S}_n$, satisfies $l > \max\{r, 1\}$. Write the centered version of $\mathscr{S}_n$ as

$$\mathscr{S}_{nj}^0 = \left\{ f_{nj} : f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\psi_k(x), (\beta_{j1}, \ldots, \beta_{jm_n}) \in \mathbb{R}^{m_n} \right\}, \ j = 1, \ldots, d,$$

where ψ_k's are the centered spline basis functions defined in Section 5.2.2.

Proposition 2 Suppose that $f \in \mathscr{F}$ and $\mathrm{E}\, f(\zeta_j) = 0$. Then under Assumptions A and B, there exists an $f_{nj} \in \mathscr{S}_{nj}^0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 = O_p(m_n^{-2\rho} + n^{-1}).$$

if $m_n = O(n^{\alpha})$ with $0 < \alpha < 0.5$.

Let $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_{m_n}(x))^T$ for $x \in [0, 1]$. Proposition 2 implies that, uniformly over $j \in \{1, \ldots, d\}$, there exists $\boldsymbol{\beta}_j \in \mathbb{R}^{m_n}$, such that $\frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{\beta}_j^T\boldsymbol{\psi}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 = O_p(m_n^{-2\rho} + n^{-1})$ under Assumptions A and B, provided $m_n = O(n^{\alpha})$. Furthermore, we can take $\boldsymbol{\beta}_j = \mathbf{0}$ for $j \in A_0$. Denote $\{j : \tilde{\boldsymbol{\beta}}_j \neq \mathbf{0}\}$ and $\{j : \tilde{\boldsymbol{\beta}}_j = \mathbf{0}\}$ as $\tilde{A}_1$ and $\tilde{A}_0$, respectively. Theorem 2 establishes the selection consistency and estimation consistency of $\tilde{\boldsymbol{\beta}}_j$'s obtained from the group LASSO step.

Theorem 2 Suppose that Assumptions A, B and C hold and $\lambda_1 \geq$
$C\sqrt{n\log(m_n)}$ for some sufficiently large constant C. Then it follows that

(i) If $m_n \to \infty$ as $n \to \infty$ with rate satisfying $m_n = o(n^{1/6})$ and $(\lambda_1^2 m_n^2)/n^2 \to 0$ as $n \to \infty$, then all the nonzero $\boldsymbol{\beta}_j$, $j \in A_1$, are selected with probability converging to 1.

(ii) If $m_n = o(n^{1/6})$, then $\sum_{j=1}^{d}||\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j||_2^2 = O_p\left(\frac{m_n^2 \log m_n}{n}\right) + O_p\left(\frac{m_n^2\lambda_1^2}{n^2}\right)$
$+ O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho-1}}\right)$.

Theorem 3 further illustrates that the estimated functions obtained from the group LASSO step, $\tilde{f}_j$'s, also enjoy selection consistency and estimation consistency.

Theorem 3 Suppose that Assumptions A, B and C hold and $\lambda_1 \geq$
$C\sqrt{n\log(m_n)}$ for some sufficiently large constant C. Then we have

(i) If $m_n \to \infty$ as $n \to \infty$ with rate satisfying $m_n = o(n^{1/6})$ and $(\lambda_1^2 m_n)/n^2 \to 0$ as $n \to \infty$, then in the group LASSO step, all the nonzero additive components f_j's, $j \in A_1$, are selected with probability converging to 1.

(ii) If $m_n = o(n^{1/6})$, then $||\tilde{f}_j - f_j||_2^2 = O_p\left(\frac{m_n \log m_n}{n}\right) + O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho}}\right) + O_p\left(\frac{m_n\lambda_1^2}{n^2} + \frac{m_n}{n}\right)$, $j \in A_1 \cup \tilde{A}_1$.

For two (positive) sequences $\{a_n\}$ and $\{b_n\}$, if $\frac{a_n}{b_n}$ is bounded away from 0 and ∞, then denote $a_n \sim b_n$. The following corollary can be directly derived from Theorem 3.

Corollary 2 Suppose that Assumptions A, B and C hold. If $m_n \sim n^{1/(2\rho+1)}$ and $\lambda_1 \sim \sqrt{n \log(m_n)}$, then

(i) If $\rho > \frac{5}{2}$, then in the group LASSO step, all the nonzero additive components f_j, $j \in A_1$, are selected with probability converging to 1.

(ii) If $\rho > \frac{5}{2}$, then $||\tilde{f}_j - f_j||_2^2 = O_p(n^{-2\rho/(2\rho+1)} \log m_n), \qquad j \in A_1 \cup \tilde{A}_1$.

Theorem 4 states that the adaptive group LASSO yields an estimate which is also consistent in both selection and estimation. Furthermore, it illustrates that this estimate compares favorably with that given by the group LASSO with respect to estimation accuracy.

Theorem 4 Suppose that Assumptions A, B and C hold and $m_n \sim n^{1/(2\rho+1)}$, where $\rho > 5/2$. If the tuning parameters satisfy $\lambda_1 \sim \sqrt{n \log(m_n)}$, $\lambda_2 \leq O(n^{\frac{1}{2}})$, $\frac{\lambda_2}{n^{(8\rho+3)/(8\rho+4)}} = o(1)$ and $\frac{n^{1/(4\rho+2)}\sqrt{\log(m_n)}}{\lambda_2} = o(1)$, then we have

(i) With probability approaching 1, the nonzero components, i.e., $\{f_j, j \in A_1\}$ are selected and $||\hat{f}_j||_2 = 0,\ j \in A_0$.

(ii) $\sum_{j \in A_1} ||\hat{f}_j - f_j||_2^2 = O_p(n^{-2\rho/(2\rho+1)})$.

5.4 Simulation Studies

In this section we use simulated examples to illustrate several properties of our proposed estimator, and compare our method with several conventional methods commonly used in practice.

We simulate data as follows. In each simulation replicate, we generate n curves and the observations are made at $m = 200$ equally spaced points in [0, 10]. In our simulation studies, we set $n = 100$ or 500. To accommodate measurement errors, the observation at t_j $(j = 1, \ldots, m)$ is generated as $W_{ij} = X_i(t_j) + e_{ij}$, where $\{X_i(t)\}_{i=1}^n$ are i.i.d samples of a stochastic process $X(t)$ and e_{ij} are i.i.d normals with mean 0 and variance 0.1. For $k = 1, \ldots, 20$, let $\lambda_k = 31.5 \times 0.6^k$ denote the k-th eigenvalue of the covariance function of $X(t)$. The corresponding k-th eigenfunction is the k-th Fourier basis function, denoted by $\phi_k(t)$. Then $X_i(t) = m(t) + \sum_{k=1}^{20} \xi_{ik}\phi_k(t)$, where $m(t) = t + \sin t$ denotes the mean function of $X(t)$ and $\{\xi_{ik}\}_{k=1}^{20}$ are independently sampled from $N(0, \lambda_k)$. The scaled score ζ_{ik} is defined as the uniform score of ξ_{ik}, i.e., $\zeta_{ik} = \Phi(\xi_{ik}/\sqrt{\lambda_k}), k = 1, \ldots, 20,\ i = 1, \ldots, n$, where Φ denotes the cdf of a standard normal distribution. The response variable is generated from Model (5.3): $y_i = a + f_1(\zeta_{i1}) + f_2(\zeta_{i2}) + f_4(\zeta_{i4}) + \epsilon_i$. We set the true intercept to $a = 1.2$, and the true components to $f_1(x) = x\exp(x) - 1$, $f_2(x) = \cos(2\pi x)$ and $f_4(x) = 3\left(x - \frac{1}{4}\right)^2 - \frac{7}{16},\ x \in [0, 1]$. The random errors ϵ_i's are independently sampled from a normal distribution with mean 0 and variance 0.67. The signal-to-noise ratio is defined

as $\mathrm{Var}\{f_1(\zeta_1) + f_2(\zeta_2) + f_4(\zeta_4)\}/\mathrm{Var}(\epsilon)$, and we set the signal-to-noise ratio to be approximately 2. We estimate the model by fitting n randomly generated training observations, and evaluate its performance on 200 randomly generated test observations. The simulation is implemented for 100 simulation replicates.

Besides our proposed method CSS-FAM, we also fit the data with three conventional models including MARS ([36]), two extended functional additive models (FAM) proposed by [72] and the component selection and estimation for the functional additive model (CSE-FAM) in [121]. More specifically, MARS is fitted using the function **earth** in the R package **earth** and the variables which enter the final model are examined by the function **evimp**. In the first extended FAM, denoted by FAM, the response variable y is fitted with a multiple linear regression where the covariates are $f_1(\hat{\zeta}_1)$, $f_2(\hat{\zeta}_2)$ and $f_4(\hat{\zeta}_4)$. In other words, FAM assumes to know the true model structure with three true covariates $f_1(\hat{\zeta}_1)$, $f_2(\hat{\zeta}_2)$ and $f_4(\hat{\zeta}_4)$. The second extended FAM, denoted by S-FAM, considers a saturated model to incorporate the first d FPC scores such that they can explain over 99.9% of the total variability in the smoothed sample curves. The value of d is 15, 16 or 17 in all simulation replicates. We employ the function **gam** in the R package **mgcv** to fit such a model in which the arguments of additive components are $\hat{\zeta}_j$, $j = 1, \ldots, d$. Then p-values of all terms in the model are available from the function **summary.gam**. Only the significant nonparametric components (p-value < 0.05) are retained in computing the true positive (TP) rate and the false positive (FP) rate. We also consider an alternative method for estimating Model (5.3) by only using the two steps of group LASSO and adaptive group LASSO, which is denoted by AGL-FAM.

Table 5.1 summarizes the comparison between these six methods in 1000 simulation replicates. It suggests that compared with CSE-FAM, CSS-FAM has similar performance in prediction when the sample size $n = 100$ and 500. Both of them outperform the other three methods except FAM in prediction accuracy, and are slightly inferior compared with FAM, which assumes the true components are known. These arguments are further demonstrated in Figure 5.1. This suggests that the extra adaptive smoothing spline step can increase the prediction accuracy when adaptive group LASSO yields a wiggly estimate. Concerning the quality of estimating nonparametric components, CSS-FAM can rival CSE-FAM as well, since both of them yield estimates which are reasonably close to the true nonparametric components. In addition, the residual sum of squares (RSS) for each component estimated using CSS-FAM is much smaller than that using AGL-FAM, indicating that smoothing spline enables us to obtain a smoother and more accurate estimate of nonparametric components. Therefore, we suggest employ CSS-FAM considering its good performance in both estimating components and prediction of the response.

Statistics	n	Methods					
		MARS	FAM	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
MSPE	100	1.15 (.25)	.92 (.17)	1.15 (.21)	1.00 (.19)	1.25 (0.23)	1.01 (0.20)
	500	.78 (.09)	.73 (.08)	.77 (.09)	.74 (.08)	.80 (.09)	.74 (.08)
RSS($\hat{f}_1$)	100	-	2.6 (4.9)	3.6 (4.7)	2.6 (4.1)	13.4 (6.9)	3.8 (5.7)
($\times 10^{-2}$)	500	-	0.4 (0.5)	0.6 (0.6)	0.5 (0.4)	2.5 (0.9)	0.6 (0.4)
RSS($\hat{f}_2$)	100	-	6.8 (10.5)	11.2 (10.0)	18.1 (13.6)	12.8 (6.8)	8.1 (13.8)
($\times 10^{-2}$)	500	-	0.5 (0.7)	1.9 (1.3)	2.9 (1.5)	3.1 (1.4)	1.9 (1.3)
RSS($\hat{f}_4$)	100	-	6.7 (10.3)	4.0 (3.3)	4.6 (5.7)	14.3 (7.2)	5.9 (11.2)
($\times 10^{-2}$)	500	-	0.7 (1.1)	0.7 (0.5)	0.5 (0.4)	2.3 (1.0)	0.5 (0.7)
TP%	100	99.1 (.05)	-	98.2 (.08)	95.7 (.12)	94.7 (.17)	94.7 (.10)
	500	100 (.00)	-	100 (.0)	100 (.0)	100 (.0)	100 (.0)
FP%	100	20.4 (.12)	-	13.7 (.11)	3.8 (.07)	0.9 (.03)	0.9 (.03)
	500	29.0 (.14)	-	8.9 (.08)	3.0 (.07)	< 0.01 (.003)	<0.01 (.003)
Time	100	.01 (.03)	< 0.01	.39 (.21)	2.87 (.23)	0.48 (.04)	2.40 (.10)
(seconds)	500	.02 (.06)	< 0.01	2.88 (2.28)	117.2 (5.91)	3.57 (0.27)	11.4 (2.77)

Table 5.1: Summary statistics for evaluating six methods. MSPE refers to the mean squared prediction error on the test data; the residual sum of squares (RSS) for each estimated component $\hat{f}_j$ is defined as: $\text{RSS}(\hat{f}_j) = \int_0^1 (\hat{f}_j(x) - f_j(x))^2 dx$; TP% and FP% stand for the true positive and false positive rates in percentage, respectively. The point estimate for each measure is averaged over 100 simulation replicates, and the corresponding estimated standard error is given in parenthesis.

Table 5.1 also compares these methods from the perspective of variable selection, where the true positive (TP) rate and the false positive (FP) rate are employed for assessment. We find that although CSS-FAM and AGL-FAM perform slightly worse than the other models in correctly selecting nonzero variables for a relatively small sample, this tiny gap vanishes when the sample size increases. Furthermore, CSS-FAM and AGL-FAM dominate the other methods in not selecting irrelevant components, regardless of how large or small the sample size is. The other methods mistakenly select irrelevant variables substantially more often than CSS-FAM or AGL-FAM, especially when the sample size is relatively small. The computational time for each method is recorded in Table 5.1 as well. Obviously CSE-FAM is the most computationally intensive method if a full basis is employed. This is a serious issue in implementations particularly when the sample size is large, as mentioned in Section 5.1. In comparison, the proposed method, CSS-FAM, can still be implemented within around 11 seconds even when the training data set consists of 500 curves.

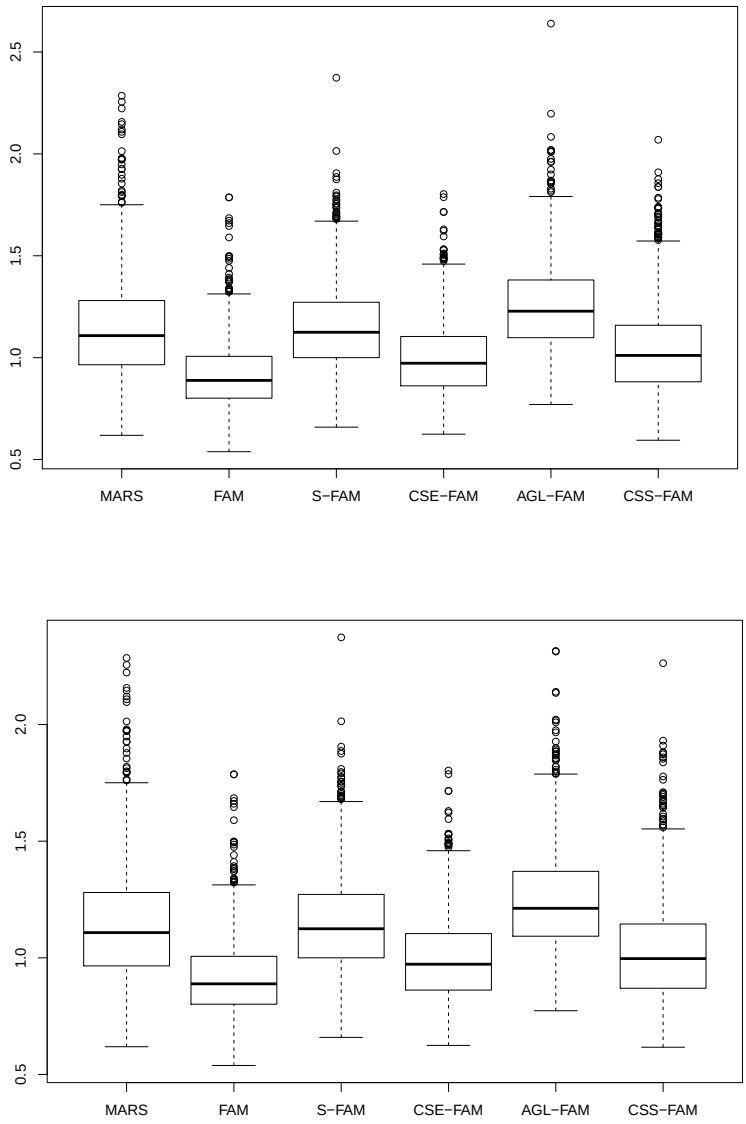


Figure 5.1: Boxplots of prediction errors across the 1000 replications for different approaches. The top figure is for $n = 100$ while the bottom is for $n = 500$.

Figure 5.2 illustrates the estimation details for one randomly selected simulation replicate of $n = 500$. After estimating the scaled FPC score, we fit group LASSO on the training data, as shown in (5.5). The top left panel in Figure 5.2 describes how the 5-fold cross-validation error changes with λ_1. The optimal λ_1 is chosen to minimize the 5-fold cross-validation error. Like the top left one, the top middle panel explains how to choose the optimal λ_2 for the adaptive group LASSO step in (5.6) based on 5-fold cross validation. The top right panel shows how to choose the optimal smoothing parameter (λ_3) by minimizing GCV in the smoothing spline step. The bottom three panels in Figure 5.2 illustrates

the effects of the extra smoothing spline step on the estimation of the nonparametric components after using adaptive group LASSO. The adaptive group LASSO method may lead to an excessively wiggly estimate for each nonzero nonparametric component. Smoothing spline can control the roughness appropriately and hence yield a smoother and more accurate estimate.

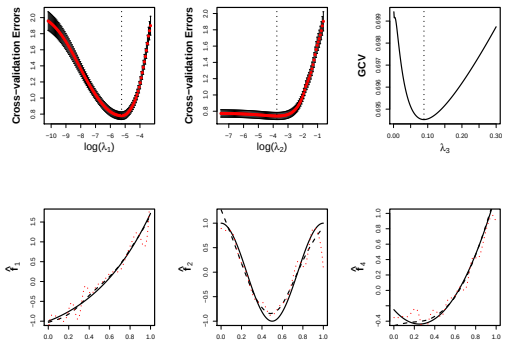


Figure 5.2: The top three panels illustrate how the cross-validation errors and GCV change with the tuning parameter λ_1, λ_2 and λ_3 in group LASSO, adaptive group LASSO and smoothing spline, respectively. The bottom three panels compare the estimated nonparametric components and the true underlying nonparametric components ($f_k,\ k = 1, 2, 4$). Dashed lines (- - -) and dotted lines ($\cdots$) represent the estimated nonparametric components with and without smoothing spline after the adaptive group LASSO fitting, respectively; while solid lines (—) represent the true underlying nonparametric components.

5.5 Applications

In this section, we fit the sparse functional additive model (5.3) using our proposed method (CSS-FAM), together with several conventional models considered in the simulation studies, to analyze two real data sets. Besides the models considered in the simulation, we fit a multiple linear model to investigate whether a functional linear model can adequately characterize the relationship between the scalar response and the functional predictor in these two examples. The covariates in the multiple linear model are the first d FPC scores. We choose the truncation level d in the same way as for Model (5.2). This multiple linear model is actually a special case of Model (5.2): each additive component taking a linear form. LASSO ([100]) is implemented when fitting the mulitple linear model in these two examples to obtain a more parsimonious model and reduce variability. This estimating method is called LAF in this chapter. In the air pollution data, the trajectories of the functional predictor for some subjects are sparsely observed. In contrast, in the Tecator data, the functional predictor is regularly spaced and densely observed across all subjects. In each example, we randomly divide the whole data set into a training set and a test set, and the training set is used to fit each model while the test set is used for evaluation. All these

models are compared with respect to the mean squared prediction errors calculated on the test set.

5.5.1 Air Pollution Data

It is of great interest to study the association between air pollution and respiratory system diseases ([112]). In this example we investigate this association using the sparse functional additive model (5.3). More specifically, in this chapter we study the effect of fine particulate matter (PM 2.5) on the rate of death caused by respiratory diseases (RESP).

The data consists of the median of the daily observations of PM 2.5 measured in micrograms per cubic meter ($\mu g/m^3$) from 1987 to 2000 across 108 cities of the United States. Some negative values appear due to detrending the time series in order to make them centered around 0 (see Page 42 in [78]). This data is obtained from the **NMMAPSdata** package ([79]), which was originally assembled for the national morbidity, mortality, and air pollution study. In this chapter we focus on exploring the relationship between the daily observations of PM 2.5 and the RESP death rate in the year 2000, since considerably fewer missing values are present during that year. There are 7 cities that have no data for PM 2.5 in that year, so we remove these 7 cities from the data set. Among the 101 remaining cites, we randomly select 80 cities as the training set. The test set consists of the trajectories of PM 2.5 in the remaining 21 cities.

Six methods are used to predict the RESP death rate using the trajectories of PM 2.5. The first row of Table 5.2 displays the mean squared prediction errors (MSPEs) of these six methods. Our proposed CSS-FAM method has the lowest prediction error among the six methods. Compared with MARS and S-FAM, LAF achieves better performance in prediction. A possible reason is that even though MARS and S-FAM models the relationship between the functional predictor PM 2.5 and the response in a more flexible way than the linear model, variabilities in these models are not well-controlled. LAF, however, reduces the variability to a great extent by variable selection. Furthermore, since CSS-FAM and CSE-FAM combine nonparametric modelling and selection of components, they can not only provide a more adequate characterization of the relationship between the functional predictor and the response than FLR, but also result in a model with less variability and better prediction. In addition, the obvious distinction between CSS-FAM and AGL-FAM suggests that the further smoothing of the estimate from the adaptive group LASSO method via smoothing splines can improve the prediction accuracy.

We use 15 cubic B-spline basis functions to represent the additive components in Model (5.3) when applying our proposed CSS-FAM method. In the group LASSO step, the optimal value of the tuning parameter λ_1 is 3×10^{-6}, which is determined by a 5-fold cross validation. Another 5-fold cross-validation suggests that $\lambda_2 = 1.5 \times 10^{-10}$ is an optimal choice of the tuning parameter in the adaptive LASSO step. As a result, 3 non-vanishing components, $\{\hat{f}_5, \hat{f}_8, \hat{f}_{11}\}$, are selected. These raw estimates turn out to be excessively wiggly. We

therefore refine these raw estimates via smoothing splines, in which the optimal choice of the smoothing parameter λ_3 determined by GCV is 1.2×10^{-5}.

5.5.2 Tecator Data

The Tecator data are recorded for 240 meat samples on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. Each of them consists of a 100-channel spectrum of absorbance, and the percentages of three components of the meat: moisture (water), fat and protein. The spectrum records the negative base 10 logarithm of the transmittance measured by the spectrometer. The percentage of three meat components are determined by analytic chemistry. As demonstrated by a large body of research (see [104], [38], [121]), the spectrum of absorbance is highly predictive of the percentage of these three meat components. We aim to study the effect of the spectral trajectories of the meat sample on the protein content by using the sparse functional additive model (5.3).

The protein content, denoted by Y, is the response variable of primary interest; the functional predictor $X(t)$ denotes the spectrum of absorbance. FPCA is implemented to estimate FPC scores and then to obtain the scaled FPC scores, denoted by $\hat{\boldsymbol{\zeta}} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_d)$. [121] suggested that the first $d = 20$ should be retained in order to achieve satisfactory prediction accuracy, even though the first 10 FPCs explain more than 99.9% of total variability in the smoothed sample curves. To compare the performance of various methods with respect to the prediction accuracy, the 240 meat samples are divided into a training sample and a test sample. According to the dataset's original description (`http://lib.stat.cmu.edu/datasets/tecator`), the 240 meat samples have been divided into three parts: the training sample consists of the 172 meat samples, the following 43 meat samples form the test samples and the last 25 meat samples are for extrapolation use and should be ignored. We, however, randomly choose 187 meat samples to train the model; the test set comprises the remaining 53 meat samples.

The comparison among the six models with respect to prediction accuracy is shown in Table 5.2. Obviously CSS-FAM outperforms the other methods in terms of prediction. In particular, the difference between CSS-FAM and LAF implies that a linear model cannot adequately characterize the relationship between the protein content and the spectrum of absorbance of the meat samples. CSS-FAM can nevertheless achieve a better trade-off between flexibility and simplicity compared with other methods. Additionally, the poor performance of AGL-CSS, especially when compared with CSS-FAM, suggests that the extra smoothing spline step in the proposed algorithm considerably enhances prediction accuracy.

	Methods					
	MARS	LAF	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
Air Pollution ($\times 10^{-8}$)	1.68	1.68	1.77	1.61	1.68	1.49
Tecator	0.99	0.66	0.56	0.55	0.92	0.51

Table 5.2: Mean squared prediction errors on the test data for six methods for the two data sets.

In AGL-FAM, 10 cubic B-spline basis functions are employed to represent the nonparametric components in the sparse functional additive model (5.3). A 5-fold cross validation suggests that $\lambda_1 = 0.002$ is an optimal choice of the penalty parameter in the group LASSO step and $\lambda_2 = 0.011$ minimizes the 5-fold cross-validation error in the adaptive LASSO step. As a result, 14 non-vanishing components, $\{\hat{f}_1, \ldots, \hat{f}_9, \hat{f}_{11}, \hat{f}_{16}, \hat{f}_{17}, \hat{f}_{19}, \hat{f}_{20}\}$, are selected from the 20 components. This finding is slightly inconsistent with the conclusion drawn in [121], where they argued that $\{\hat{f}_1, \ldots, \hat{f}_8, \hat{f}_{10}, \hat{f}_{13}, \hat{f}_{16}, \hat{f}_{17}\}$ are non-vanishing components. To refine these estimated components, smoothing spline is employed and the optimal choice of smoothing parameter, $\lambda_3 = 0.001$, is chosen to minimize the GCV measure.

5.6 Conclusions and Discussion

Compared with traditional FLR, the sparse functional additive model (5.3) proposed in this chapter provides a more flexible description of the relationship between a scalar response and a functional predictor. To achieve sparseness, we employ the group LASSO penalty to select and estimate nonzero components in the nonparametric additive model, thereby reducing variability and enhancing interpretability.

The estimation procedure consists of several important techniques. FPCA is employed to estimate FPC scores and eigenvalues of the covariance function of the functional predictor. Then we use B-spline basis functions to represent the nonparametric additive components in the sparse functional additive model (5.3). The use of the group LASSO penalty enables us to achieve the goal of selecting and estimating nonzero components. To obtain a better estimate of the coefficient vectors, we adopt the idea of the adaptive group LASSO to provide different shrinkages for different components. Considering the fact that the estimated components given by the adaptive LASSO may not be smooth, since a large number of B-spline basis functions are used to represent them, we propose using smoothing splines to further refine the estimated nonzero components obtained from the group LASSO step. Simulation studies demonstrate that this smoothing step can improve both estimation of the additive components and prediction of the response.

In contrast to other methods, we theoretically justify that our proposed estimator enjoys both selection consistency and estimation consistency. These consistency results are also demonstrated by simulation studies. Two real data applications show that the proposed

model together with the estimating method provides an appealing tool in predicting a scalar response from a functional predictor, compared with other conventional methods.

Even though this chapter discusses only regressing a scalar response on a functional covariate, the methodology can be extended to accommodate other scenarios. For example, this framework can be extended to explore the relationship between a scalar response, whose distribution belongs to the exponential family, and a functional predictor. In addition, in this work the truncation level d, such that the first d FPCs can explain over 99.9% of total variability in the functional predictor, is assumed to be fixed. From the theoretical perspective, it is worthwhile to investigate the properties of the corresponding estimator when d is allowed to increase with the sample size in future work. Lastly, compared with the result of the application to the meat sample in last chapter, we find that the prediction accuracy becomes worse when the fat content is ignored in the prediction of the protein content. This implies that we should extend the current framework to allow for scalar covariates which might be predictive of a scalar response. As to the computational aspect, we can borrow the idea of iteratively updating to address the issue of interdependence between the estimation of the nonparametric part and that of the parametric part. Issues such as whether the algorithm can converge and multiple minimum demand further investigation.

5.7 Appendix

5.7.1 Appendix A: Estimation of FPC scores

When the functional predictor $X_i(t)$ has sparse observations, we employ the principal component analysis by conditional expectation (PACE) algorithm proposed by [116] to obtain estimated FPC scores denoted by $\hat{\xi}_{ij},\ j = 1, \ldots, d,\ i = 1, \ldots, n$. PACE first estimates the mean function of $X(t)$ via local linear regression and the corresponding estimator is denoted by $\hat{m}(t)$. Then the raw estimate of covariance evaluated at (t_{ij}, t_{il}), $G_i(t_{ij}, t_{ij}) = (W_{ij} - \hat{m}(t_{ij}))(W_{il} - \hat{m}(t_{il}))$ is pooled. To account for the fact that $\mathrm{Cov}\,(W_{ij}, W_{il}) = \mathrm{Cov}\,(X_i(t_{ij}), X_i(t_{il})) + \delta(t_{ij} = t_{il})$ where $\delta(t = s) = 1$ if $t = s$ and 0 otherwise, a local linear regression is employed to smooth the non-diagonal estimates, $\{G_i(t_{ij}, t_{ij}), t_{ij} \neq t_{il}, i = 1, \ldots, n\}$. Let $\hat{G}(s,t)$ denote the smoothed covariance function and let $\hat{\lambda}_j$'s and $\hat{\phi}_j(t)$'s denote the corresponding eigenvalues and eigenfunctions, respectively. They satisfy $\int_{\mathcal{I}} \hat{G}(s,t)\hat{\phi}_j(t)dt = \hat{\lambda}_j\hat{\phi}_j(s)$. The estimated FPC scores, $\hat{\xi}_{ij}$'s, are estimated via conditional expectation.

When the functional predictor $X_i(t)$ has dense observations, we smooth each trajectory $X_i(t)$ via the local linear regression rather than smoothing the mean function, and denote the smoothed trajectory by $\hat{X}_i(t)$, $i = 1, \ldots, n$. The estimated mean function and covariance function are then given by $\hat{m}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{X}_i(t)$ and $\hat{G}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i(s) - \hat{m}(s))(\hat{X}_i(t) - \hat{m}(t))$, respectively. The estimated eigenvalues $\hat{\lambda}_j$ and eigenfunctions $\hat{\phi}_j(t)$ of

$\hat{G}$ satisfy $\int_{\mathcal{I}} \hat{G}(s,t)\hat{\phi}_j(t)dt = \hat{\lambda}_j\hat{\phi}_j(s)$ as well. Unlike PACE, the FPC scores in the scenario are calculated as $\hat{\xi}_{ij} = \int_{\mathcal{I}}(\hat{X}_i(s) - \hat{m}(s))\hat{\phi}_j(s)ds$, $j = 1, \ldots, d$, $i = 1, \ldots, n$.

5.7.2 Appendix B: Proofs

We follow the main ideas of [48] to prove Theorems 1-3, but great effort needs to be taken to tackle the difficulties caused by using the estimated scaled FPC scores (rather than the true scaled FPC scores), to estimate the unknown parameters and functions. Before proving Proposition 1, we first present the following result which establishes consistency of the estimated scaled FPC scores.

Lemma 7 Suppose that Assumptions A and B1 hold. We have

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{d}|\hat{\zeta}_{ij} - \zeta_{ij}|\right\}^2 = O_p(n^{-1}).$$

Proof: See Lemma 2 in [121]. □

B.1. Proof of Proposition 1:
For any function f on [0, 1], let $||f||_\infty = \sup_{x\in[0,1]}|f(x)|$. For $f \in \mathscr{F}$ and $\mathrm{E}\{f(\zeta_j)\} = 0$, there exists $f_j' \in \mathscr{S}_n$, such that $||f - f_j'||_\infty = O(m_n^{-\rho})$; see [22] or Lemma 5 of [98]. Let $f_{nj} = f_j' - n^{-1}\sum_{i=1}^n f_j'(\hat{\zeta}_{ij})$. Then $f_{nj} \in \mathscr{S}_{nj}^0$ and we next prove that f_{nj} satisfies $\frac{1}{n}\sum_{i=1}^n\{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 = O_p(m_n^{-2\rho} + n^{-1})$. Since $(x+y)^2 \le 2(x^2+y^2)$ $\forall x, y \in \mathbb{R}$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 &= \frac{1}{n}\sum_{i=1}^{n}\{f_{nj}(\hat{\zeta}_{ij}) - f_j'(\hat{\zeta}_{ij}) + f_j'(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 \\
&\le 2n^{-1}\sum_{i=1}^{n}\{f_{nj}(\hat{\zeta}_{ij}) - f_j'(\hat{\zeta}_{ij})\}^2 + \{f_j'(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 \\
&= 2\left\{\frac{1}{n}\sum_{i=1}^{n}f_j'(\hat{\zeta}_{ij})\right\}^2 + O(m_n^{-2\rho}) \\
&= 2\left[\frac{1}{n}\sum_{i=1}^{n}\{f_j'(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}) + f(\hat{\zeta}_{ij}) - f(\zeta_{ij}) + f(\zeta_{ij})\}\right]^2 + O(m_n^{-2\rho}) \\
&\le 6E_1 + 6E_2 + 6E_3 + O(m_n^{-2\rho}), \qquad (5.8)
\end{aligned}$$

where $E_1 = \left[\sum_{i=1}^n\{f_j'(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}/n\right]^2$, $E_2 = \left[\sum_{i=1}^n\{f(\hat{\zeta}_{ij}) - f(\zeta_{ij})\}/n\right]^2$ and $E_3 = \{\sum_{i=1}^n f(\zeta_{ij})/n\}^2$. Obviously, $E_1 = O(m_n^{-2\rho})$ and $E_3 = O_p(n^{-1})$ by Chebyshev's

inequality. As for E_2, there are two different cases. (i) If $r = 0$, then

$$\begin{aligned} E_2 &\leq \left\{ n^{-1} \sum_{i=1}^{n} |f(\hat{\zeta}_{ij}) - f(\zeta_{ij})| \right\}^2 \\ &\leq \left\{ L n^{-1} \sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}|^{\nu} \right\}^2 \\ &\leq \left[L n^{-1} \left\{ \sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}|^2 \right\}^{\nu/2} n^{1-\nu/2} \right]^2 \\ &= O_p(n^{-\nu}). \end{aligned}$$

The second inequality holds due to $f \in \mathscr{F}$ and the third inequality is obtained by Hölder's inequality. In this case $0.5 < \rho = \nu \leq 1$, we have $E_2 = O_p(n^{-\nu}) = O_p(n^{-\rho}) = O_p(m_n^{-2\rho})$. Therefore, $\sum_{i=1}^{n} \{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2/n = O_p(m_n^{-2\rho})$ according to (5.8). (ii) In the second case, $r \geq 1$, so f is continuously differentiable over $[0, 1]$, which implies that the first derivative of f is bounded some positive constant M. By Lemma 7,

$$\begin{aligned} E_2 &\leq \left\{ n^{-1} \sum_{i=1}^{n} |f(\hat{\zeta}_{ij}) - f(\zeta_{ij})| \right\}^2 \\ &\leq M^2 n^{-2} \left\{ \sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}| \right\}^2 \\ &\leq M^2 n^{-1} \left\{ \sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}|^2 \right\} \\ &= O_p(n^{-1}). \end{aligned}$$

It follows that $\sum_{i=1}^{n} \{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2/n = O_p(m_n^{-2\rho} + n^{-1})$ by (5.8). Hence in both cases, $\sum_{i=1}^{n} \{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2/n = O_p(m_n^{-2\rho} + n^{-1})$.

B.2. Some useful lemmas:
To establish estimation and selection consistency of the proposed estimator, we need the following lemmas.

Lemma 8 Let $T_{jk} = n^{-\frac{1}{2}} m_n^{\frac{1}{2}} \sum_{i=1}^{n} \psi_k(\hat{\zeta}_{ij}) \epsilon_i$, $j = 1, \ldots, d,\ k = 1, \ldots, m_n$ and $T_n = \max_{1 \leq j \leq d, 1 \leq k \leq m_n} |T_{jk}|$. Under Assumptions A, B and C2 in Section 5.3, then $T_n = O_p(\sqrt{\log m_n})$, if $m_n = o(n^{1/4})$.

Proof: Let $\hat{S}_{njk}^2 = \sum_{i=1}^{n} \psi_k^2(\hat{\zeta}_{ij})$ and $S_{njk}^2 = \sum_{i=1}^{n} \psi_k^2(\zeta_{ij})$. Conditional on $\hat{\zeta}_{ij}$'s, then T_{jk}'s are sub-gaussian. Let $\hat{S}_n^2 = \max_{1 \leq j \leq d, 1 \leq k \leq m_n} \hat{S}_{njk}^2$ and $S_n^2 = \max_{1 \leq j \leq d, 1 \leq k \leq m_n} S_{njk}^2$. Based on maximal inequalities on sub-gaussian random variables ([103]), we have

$$\mathrm{E}\left\{ T_n | (\hat{\zeta}_{ij}, i = 1, \ldots, n, j = 1, \ldots, d) \right\} \leq C_1 n^{-\frac{1}{2}} m_n^{\frac{1}{2}} \sqrt{\log m_n} \hat{S}_n$$

for some constant $C_1 > 0$. It follows

$$T_n = O_p(n^{-\frac{1}{2}} m_n^{\frac{1}{2}} \sqrt{\log m_n} \hat{S}_n). \tag{5.9}$$

Now we study the properties of $\hat{S}_n$, from which the order of T_n is immediately available. Since $\hat{S}_{njk}^2 = (\hat{S}_{njk}^2 - S_{njk}^2) + S_{njk}^2$, $\hat{S}_n^2 \leq S_n^2 + \max_{1\leq j\leq d, 1\leq k\leq m_n}(\hat{S}_{njk}^2 - S_{njk}^2)$. On the one hand,

$$\mathrm{E}\,(S_n^2) \leq \sqrt{2C_2 m_n^{-1} n \log m_n} + C_2 \log m_n + C_2 n m_n^{-1}$$

holds for a sufficiently large constant C_2 according to the proof of Lemma 2 in [48]. Since $m_n = O(n^\alpha)$ with $\alpha < 0.5$,

$$S_n^2 = O_p(n m_n^{-1}). \tag{5.10}$$

On the other hand, from (5) of [82], we have $||\frac{dB_k(x)}{dx}||_\infty = O(m_n)$ when the order of the B-spline basis functions, l, satisfies $l \geq 2$ and $||B_k||_\infty \leq 1$. Both of them hold uniformly for $1 \leq k \leq m_n$. It follows that

$$\begin{aligned}
\hat{S}_{njk}^2 - S_{njk}^2 &= \sum_{i=1}^n \left\{ B_k(\hat{\zeta}_{ij}) - n^{-1} \sum_{i=1}^n B_k(\hat{\zeta}_{ij}) \right\}^2 - \left\{ B_k(\zeta_{ij}) - n^{-1} \sum_{i=1}^n B_k(\zeta_{ij}) \right\}^2 \\
&= \sum_{i=1}^n B_k^2(\hat{\zeta}_{ij}) - B_k^2(\zeta_{ij}) + n^{-1} \left\{ \sum_{i=1}^n B_k(\zeta_{ij}) \right\}^2 - n^{-1} \left\{ \sum_{i=1}^n B_k(\hat{\zeta}_{ij}) \right\}^2 \\
&= \sum_{i=1}^n \{B_k(\hat{\zeta}_{ij}) + B_k(\zeta_{ij})\}\{B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})\} \\
&\qquad + n^{-1} \left[\sum_{i=1}^n \{B_k(\hat{\zeta}_{ij}) + B_k(\zeta_{ij})\} \right] \left[\sum_{i=1}^n \{B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})\} \right] \\
&\leq 4 \sum_{i=1}^n |B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})| \\
&= O\left(m_n \sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}| \right) \\
&= O\left(m_n \sqrt{n} \sqrt{\sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}|^2} \right) \\
&= O_p(m_n \sqrt{n})
\end{aligned} \tag{5.11}$$

uniformly over $1 \leq j \leq d$ and $1 \leq k \leq m_n$. The first inequality is obtained using the Cauchy-Schwarz inequality and the last equation is a direct application of Lemma 7. Combining (5.10) and (5.11), we have $\hat{S}_n^2 = O_p(m_n\sqrt{n} + n m_n^{-1})$. $\hat{S}_n = O_p(n^{\frac{1}{2}} m_n^{-\frac{1}{2}})$ if $m_n = o(n^{1/4})$. As a result, $T_n = O_p(\sqrt{\log m_n})$. □

Given a matrix Σ, let $\Lambda_{\min}(\Sigma)$ and $\Lambda_{\max}(\Sigma)$ denote the smallest and largest eigenvalues of Σ, respectively. Let $\tilde{\boldsymbol{Z}}_{ij} = \{\psi_1(\zeta_{ij}), \ldots, \psi_{m_n}(\zeta_{ij})\}^T$ and $\tilde{\boldsymbol{Z}}_j = (\tilde{\boldsymbol{Z}}_{1j}, \ldots, \tilde{\boldsymbol{Z}}_{nj})^T$. For a subset of $\{1, \ldots, d\}$, A, $\boldsymbol{Q}_A = \frac{\boldsymbol{Z}_A^T \boldsymbol{Z}_A}{n} \in \mathbb{R}^{|A|m_n \times |A|m_n}$ and $\tilde{\boldsymbol{Q}}_A = \frac{\tilde{\boldsymbol{Z}}_A^T \tilde{\boldsymbol{Z}}_A}{n} \in \mathbb{R}^{|A|m_n \times |A|m_n}$, where $\boldsymbol{Z}_A$ (or $\tilde{\boldsymbol{Z}}_A$) represents the matrix by stacking $\boldsymbol{Z}_j$ (or $\tilde{\boldsymbol{Z}}_j$), $j \in A$ by column. For any $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, $||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^p x_i^2}$.

Lemma 9 *Suppose $m_n = o(n^{1/6})$. Under Assumptions A, B and C3 in Section 5.3, then given a nonempty subset of $\{1, \ldots, d\}$, A, with probability converging to 1,*

$$C_3 m_n^{-1} \leq \Lambda_{\min}(\boldsymbol{Q}_A) \leq \Lambda_{\max}(\boldsymbol{Q}_A) \leq C_4 m_n^{-1}$$

for some positive constants C_3 and C_4.

Proof: We first lay out some facts about matrix theory, which are useful in determining the magnitude of eigenvalues. For an $m \times n$ matrix $\boldsymbol{G}$,

$$||\boldsymbol{G}||_2 = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \frac{||\boldsymbol{G}x||_2}{||x||_2} \text{ and } \quad ||\boldsymbol{G}||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |g_{ij}|,$$

where g_{ij} is the (i,j)th entry of $\boldsymbol{G}$. For a symmetric matrix $\boldsymbol{G}$, $\Lambda_{\max}(\boldsymbol{G}) \leq ||\boldsymbol{G}||_2$. The following inequalities will be used later to prove the lemma. Let $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ be two $n \times n$ symmetric matrices.

1. Weyl's inequality ([108]):

$$\Lambda_{\min}(\boldsymbol{G}_1) - \Lambda_{\max}(\boldsymbol{G}_2 - \boldsymbol{G}_1) \leq \Lambda_{\min}(\boldsymbol{G}_2),$$
$$\Lambda_{\max}(\boldsymbol{G}_2) \leq \Lambda_{\max}(\boldsymbol{G}_1) + \Lambda_{\max}(\boldsymbol{G}_2 - \boldsymbol{G}_1).$$

   Thus $|\Lambda_{\min}(\boldsymbol{G}_1) - \Lambda_{\min}(\boldsymbol{G}_2) \leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2$ and $|\Lambda_{\max}(\boldsymbol{G}_1) - \Lambda_{\max}(\boldsymbol{G}_2)|$ $\leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2$.

2. The Gershgorin circle theorem ([37]):

$$||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2 \leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_1.$$

We are ready to prove Lemma 9 now. Denote $\gamma_{k_1 k_2}^{(j_1, j_2)} = \sum_{i=1}^n \{\psi_{k_1}(\hat{\zeta}_{ij_1})$ $\psi_{k_2}(\hat{\zeta}_{ij_2}) - \psi_{k_1}(\zeta_{ij_1})\psi_{k_2}(\zeta_{ij_2})\}/n$, $1 \leq k_1,\ k_2 \leq m_n$, $1 \leq j_1,\ j_2 \leq d$. Based on the definition of ψ_k's, it's obvious that uniformly over $1 \leq k \leq m_n$, $||\frac{d\psi_k(x)}{dx}||_\infty = O(m_n)$ and $||\psi_k||_\infty \leq$

$||B_k||_\infty + |n^{-1}\sum_{i=1}^n B_k(\hat{\zeta}_{ij})| \leq 2$. Then according to Lemma 7, we have

$$\begin{aligned}
|\gamma_{k_1k_2}^{(j_1,j_2)}| &= \frac{1}{n}|\sum_{i=1}^n \{\psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\hat{\zeta}_{ij_2}) - \psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\zeta_{ij_2}) \\
&\qquad + \psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\zeta_{ij_2}) - \psi_{k_1}(\zeta_{ij_1})\psi_{k_2}(\zeta_{ij_2})\}| \\
&= O(m_n n^{-1})\sum_{i=1}^n (|\hat{\zeta}_{ij_1} - \zeta_{ij_1}| + |\hat{\zeta}_{ij_2} - \zeta_{ij_2}|) \\
&= O(m_n n^{-\frac{1}{2}})\sqrt{\sum_{i=1}^n (|\hat{\zeta}_{ij_1} - \zeta_{ij_1}| + |\hat{\zeta}_{ij_2} - \zeta_{ij_2}|)^2} \\
&= O_p(m_n n^{-\frac{1}{2}}).
\end{aligned}$$

The third equation follows directly from the Cauchy-Schwarz inequality. Note that $\gamma_{k_1k_2}^{(j_1,j_2)} = O_p(m_n n^{-\frac{1}{2}})$ holds uniformly over $1 \leq j_1,\ j_2 \leq d$ and $1 \leq k_1,\ k_2 \leq m_n$. For a nonempty subset of $\{1,\ldots,d\}$, A, without loss of generality, we can assume that $A = \{1,\ldots,q\}$, where $q \leq d$. Then we have

$$\begin{aligned}
|\Lambda_{\min}(\boldsymbol{Q}_A) - \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A)| &\leq ||\boldsymbol{Q}_A - \tilde{\boldsymbol{Q}}_A||_2 \\
&\leq ||\boldsymbol{Q}_A - \tilde{\boldsymbol{Q}}_A||_1 \\
&= O_p\left(\sum_{j_1=1}^q \sum_{k_1=1}^{m_n} |\gamma_{k_1k_2}^{(j_1,j_2)}|\right) \\
&= O_p(m_n^2 n^{-\frac{1}{2}}).
\end{aligned}$$

This leads to $|\Lambda_{\min}(\boldsymbol{Q}_A) - \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A)| = o_p(m_n^{-1})$ if $m_n = o(n^{1/6})$.

On the other hand, by Lemma 3 of [48], there exists positive constants C_5 and C_6 such that $C_5 m_n^{-1} \leq \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A) \leq \Lambda_{\max}(\tilde{\boldsymbol{Q}}_A) \leq C_6 m_n^{-1}$ with probability converging to 1. Therefore, we can find a positive constant C_3, satisfying $\Lambda_{\min}(\boldsymbol{Q}_A) \geq C_3 m_n^{-1}$ with probability converging to 1. Similarly, we can prove that, with probability tending to 1, there exists a constant C_4 such that $\Lambda_{\max}(\boldsymbol{Q}_A) \leq C_4 m_n^{-1}$. □

B.3. Proof of Theorem 1:
Since $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \ldots, \tilde{\boldsymbol{\beta}}_d^T)^T$ minimizes (2.5), it follows that

$$(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}) + \lambda_1 \sum_{j=1}^d ||\tilde{\boldsymbol{\beta}}_j||_2 \leq (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^d ||\boldsymbol{\beta}_j||_2. \tag{5.12}$$

For any set $A \subset \{1,\ldots,d\}$, $\boldsymbol{\beta}_A$ denotes the long vector obtained by stacking vectors $\boldsymbol{\beta}_j,\ j \in A$. Let $A_2 = A_1 \cup \tilde{A}_1$. Then (5.12) can be written as

$$||(\boldsymbol{y} - \boldsymbol{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{A_2})||_2^2 + \lambda_1 \sum_{j \in A_2} ||\tilde{\boldsymbol{\beta}}_j||_2 \leq ||(\boldsymbol{y} - \boldsymbol{Z}_{A_2}\boldsymbol{\beta}_{A_2})||_2^2 + \lambda_1 \sum_{j \in A_2} ||\boldsymbol{\beta}_j||_2. \tag{5.13}$$

Let $\boldsymbol{u}_n = \boldsymbol{y} - \boldsymbol{Z}_{A_2}\boldsymbol{\beta}_{A_2}$ and $\boldsymbol{v}_n = \boldsymbol{Z}_{A_2}(\tilde{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}_{A_2})$. After some simple algebra, we can write (5.13) as $\boldsymbol{v}_n^T\boldsymbol{v}_n - 2\boldsymbol{u}_n\boldsymbol{v}_n \leq \lambda_1 \sum_{j\in A_2}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||)$. Note that

$$\begin{aligned}\sum_{j\in A_2}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||_2) &\leq \sum_{j\in A_1}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||_2)\\ &\leq \sum_{j\in A_1}||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2\\ &\leq \sqrt{|A_1|}||\boldsymbol{\beta}_{A_1} - \tilde{\boldsymbol{\beta}}_{A_1}||_2\\ &\leq \sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2.\end{aligned}$$

Hence (5.13) can be simplified as

$$\boldsymbol{v}_n^T\boldsymbol{v}_n - 2\boldsymbol{u}_n^T\boldsymbol{v}_n \leq \lambda_1\sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2. \tag{5.14}$$

Let $\boldsymbol{u}_n^\star = \boldsymbol{Z}_{A_2}(\boldsymbol{Z}_{A_2}^T\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}_{A_2}^T\boldsymbol{u}_n$. In other words, $\boldsymbol{u}_n^\star$ represents the projection of $\boldsymbol{u}_n$ onto the space spanned by the columns of $\boldsymbol{Z}_{A_2}$. The matrix inverse of $\boldsymbol{Z}_{A_2}^T\boldsymbol{Z}_{A_2}$ exists due to Lemma 9, which indicates that the $\Lambda_{\min}(\boldsymbol{Z}_{A_2}^T\boldsymbol{Z}_{A_2})$ is positive with probability approaching to 1. Applying the Cauchy-Schwarz inequality, we have

$$2|\boldsymbol{u}_n^T\boldsymbol{v}_n| = 2|(\boldsymbol{u}_n^\star)^T\boldsymbol{v}_n| \leq 2||\boldsymbol{u}_n^\star||_2||\boldsymbol{v}_n||_2 \leq 2||\boldsymbol{u}_n^\star||_2^2 + \frac{1}{2}||\boldsymbol{v}_n||_2^2, \tag{5.15}$$

where the last inequality is from the fact that $ab \leq a^2 + b^2/4$ for $\forall a,\ b \in \mathbb{R}$. Plugging (5.15) into (5.14), we have

$$\boldsymbol{v}_n^T\boldsymbol{v}_n \leq 4||\boldsymbol{u}_n^\star||_2^2 + 2\lambda_1\sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2. \tag{5.16}$$

By Lemma 9, $\boldsymbol{v}_n^T\boldsymbol{v}_n \geq nC_3m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2$ for some positive constant C_3, with probability approaching 1. Since $2ab \leq a^2 + b^2\ \forall a, b \in \mathbb{R}$, it can be derived from (5.16) that with probability approaching 1,

$$\begin{aligned}nC_3m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 &\leq 4||\boldsymbol{u}_n^\star||_2^2 + 2\lambda_1\sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2\\ &\leq 4||\boldsymbol{u}_n^\star||_2^2 + \frac{(2\lambda_1\sqrt{|A_1|})^2}{2nC_3m_n^{-1}} + \frac{n}{2}C_3m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2.\end{aligned}$$

Therefore,

$$||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 \leq \frac{8||\boldsymbol{u}_n^\star||_2^2}{nC_3m_n^{-1}} + \frac{4\lambda_1^2|A_1|}{n^2C_3^2m_n^{-2}}, \tag{5.17}$$

with probability approaching 1.

Next we study the magnitude of $||\boldsymbol{u}_n^\star||_2^2$. Denote the ith component in $\boldsymbol{u}_n$ as u_i, $i = 1, \ldots, n$. Then u_i can be expressed as

$$\begin{aligned} u_i &= y_i - \bar{y} - \sum_{j \in A_2} \boldsymbol{Z}_{ij}\boldsymbol{\beta}_j \\ &= y_i - a - \sum_{j=1}^{d} f_j(\zeta_{ij}) + a - \bar{y} + \sum_{j=1}^{d} f_j(\zeta_{ij}) - \sum_{j=1}^{d} f_j(\hat{\zeta}_{ij}) \\ &+ \sum_{j=1}^{d} f_j(\hat{\zeta}_{ij}) - \sum_{j \in A_2} \boldsymbol{Z}_{ij}\boldsymbol{\beta}_j \\ &= E_{1i} + E_{2i} + E_{3i} + E_{4i}, \end{aligned}$$

where $E_{1i} = \epsilon_i$, $E_{2i} = a - \bar{y}$, $E_{3i} = \sum_{j=1}^{d}\{f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij})\}$ and $E_{4i} = \sum_{j \in A_2}\{f_j(\hat{\zeta}_{ij}) - \boldsymbol{Z}_{ij}\boldsymbol{\beta}_j\}$. Let $\boldsymbol{E}_j = (E_{j1}, \ldots, E_{jn})^T$, $1 \le j \le 4$ and $\boldsymbol{P}_{A_2} = \boldsymbol{Z}_{A_2}(\boldsymbol{Z}_{A_2}^T\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}_{A_2}^T$. Then we have

$$||\boldsymbol{u}_n^\star||_2 = ||\boldsymbol{P}_{A_2}\boldsymbol{u}_n||_2^2 \le 4||\boldsymbol{\epsilon}^\star||_2^2 + 4||\boldsymbol{E}_2||_2^2 + 4||\boldsymbol{E}_3||_2^2 + 4||\boldsymbol{E}_4||_2^2, \tag{5.18}$$

where $\boldsymbol{\epsilon}^\star$ is the projection of $\boldsymbol{E}_1$ onto the space spanned by columns in $\boldsymbol{Z}_{A_2}$. Obviously $||\boldsymbol{E}_2||_2^2 = O_p(1)$ and $||\boldsymbol{E}_4||_2^2 = O_p(1 + nm_n^{-2\rho})$ by Proposition 1. For $||\boldsymbol{E}_3||_2^2$, there are two different cases. (i) If $r = 0$, by Hölder's inequality, it follows that

$$\begin{aligned} ||\boldsymbol{E}_3||_2^2 &= \sum_{i=1}^{n}\left[\sum_{j=1}^{d}\left\{f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij})\right\}\right]^2 \\ &\le L^2\sum_{i=1}^{n}\left(\sum_{j=1}^{d}|\zeta_{ij} - \hat{\zeta}_{ij}|^\rho\right)^2 \\ &\le L^2 d^{2-2\rho}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}|\zeta_{ij} - \hat{\zeta}_{ij}|\right)^{2\rho} \\ &\le L^2 d^{2-2\rho} n^{1-\rho}\left\{\sum_{i=1}^{n}\left(\sum_{j=1}^{d}|\zeta_{ij} - \hat{\zeta}_{ij}|\right)^2\right\}^\rho \\ &= O_p(n^{1-\rho}). \end{aligned} \tag{5.19}$$

The last equation is based on Lemma 7. (ii) If $r \ge 1$, then f_j's are continuously differentiable on [0,1]. Then there exists a positive constant M such that $\max_{1 \le j \le d}|f_j(s) - f_j(t)| \le M|s - t|$

$\forall s,\ t \in [0,1]$. Therefore,

$$
\begin{aligned}
||\boldsymbol{E}_3||_2^2 &= \sum_{i=1}^{n}\left[\sum_{j=1}^{d}\{f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij})\}\right]^2 \\
&\leq M^2 \sum_{i=1}^{n}\left(\sum_{j=1}^{d}|\zeta_{ij} - \hat{\zeta}_{ij}|\right)^2 \\
&= O_p(1), \qquad (5.20)
\end{aligned}
$$

where the last equation is obtained from Lemma 7. For $||\boldsymbol{\epsilon}^\star||_2^2$, we have $||\boldsymbol{\epsilon}^\star||_2^2 \leq ||(\boldsymbol{Z}_{A_2}^T\boldsymbol{Z}_{A_2})^{-\frac{1}{2}}\boldsymbol{Z}_{A_2}^T\boldsymbol{\epsilon}||_2^2 \leq \frac{m_n}{C_3 n}||\boldsymbol{Z}_{A_2}^T\boldsymbol{\epsilon}||_2^2$ with probability approaching 1. By Lemma 8, we have

$$
\begin{aligned}
\max_{A:|A|\leq d}||\boldsymbol{Z}_A^T\boldsymbol{\epsilon}||_2^2 &= \max_{A:|A|\leq d}\sum_{j\in A}||\boldsymbol{Z}_j^T\boldsymbol{\epsilon}||_2^2 \\
&\leq d m_n \max_{1\leq j\leq d, 1\leq k\leq m_n}\left|\sum_{i=1}^{n}\psi_k(\hat{\zeta}_{ij})\epsilon_i\right|^2 \\
&= O_p(n\log m_n).
\end{aligned}
$$

Hence,

$$||\boldsymbol{\epsilon}^\star||_2^2 = O_p(m_n \log m_n). \qquad (5.21)$$

In both cases ($r = 0$ or $r \geq 1$), combining (5.17), (5.18), (5.19) (or (5.20)) and (5.21), we have

$$||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 = O_p\left(\frac{m_n^2\log m_n}{n}\right) + O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho-1}}\right) + O_p\left(\frac{\lambda_1^2 m_n^2}{n^2}\right). \qquad (5.22)$$

This completes the proof of (ii) of Theorem 1.

Now we go back to prove part (i). For $j \in A_1$, $||f_j||_2 \geq c_f$ under Assumption C1. If $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})$, then from the proof of Proposition 1, we have $||f_j - f_j'||_2 = O(m_n^{-\rho})$, where $f_j'(x) = \sum_{k=1}^{m_n} B_k(x)\beta_{jk}$. Therefore, $||f_j'||_2 \geq ||f_j||_2 - ||f_j - f_j'||_2$ by the triangle inequality, which leads to $||f_j'||_2 \geq 0.5c_f$ when n is sufficiently large. On the other hand, based on (12) of [99], $C_7 m_n^{-1}||\boldsymbol{\beta}_j||_2^2 \leq ||f_j'||_2^2 \leq C_8 m_n^{-1}||\boldsymbol{\beta}_j||_2^2$ for some positive constants C_7 and C_8. Combining the above facts, we have $||\boldsymbol{\beta}_j||_2^2 \geq C_8^{-1}m_n||f_j'||_2^2 \geq 0.25C_8^{-1}m_n c_f^2$, $j \in A_1$. If $\tilde{\boldsymbol{\beta}}_j = 0$ for some $j \in A_1$, then for such j, we have $||\boldsymbol{\beta}_j||_2^2 = ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \geq 0.25C_8^{-1}m_n c_f^2$. However, by part (ii), we have $||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 = o_p(1)$ since $m_n^2\log(m_n)/n \to 0$ and $(\lambda_1^2 m_n^2)/n^2 \to 0$. Therefore, $||\tilde{\boldsymbol{\beta}}_j||_2 > 0\ \forall j \in A_1$ with probability approaching 1. This completes the proof of part (i). □

B.4. Proof of Theorem 2:

Part (i) is immediately available from the definition of $\tilde{f}_j$ and $\tilde{\boldsymbol{\beta}}_j$ and part (i) of Theorem 1.

Based on properties of B-spline functions ([22]), there exists $f_j' \in \mathscr{S}_n$ such that $||f_j - f_j'||_\infty = O(m_n^{-\rho})$. Let $f_{nj}(x) = f_j'(x) - n^{-1}\sum_{i=1}^n f_j'(\hat{\zeta}_{ij})$, As shown in Proposition 1, $f_{nj}(x)$ satisfies $\forall x \in [0,1]$

$$|f_{nj}(x) - f_j'(x)|^2 = \left| n^{-1}\sum_{i=1}^n f_j'(\hat{\zeta}_{ij}) \right|^2 = O_p(m_n^{-2\rho} + n^{-1}).$$

Then it follows that

$$||f_j - f_{nj}||_2^2 = O_p(m_n^{-2\rho} + n^{-1}). \tag{5.23}$$

Let $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})$ and $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jm_n})$. Then $f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} B_k(x) - n^{-1}\sum_{i=1}^n \sum_{k=1}^{m_n} \beta_{jk} B_k(\hat{\zeta}_{ij})$ and $\tilde{f}_j(x) = \sum_{k=1}^{m_n} \tilde{\beta}_{jk} B_k(x) - n^{-1}\sum_{i=1}^n \sum_{k=1}^{m_n} \tilde{\beta}_{jk} B_k(\hat{\zeta}_{ij})$. Therefore,

$$\begin{aligned} ||f_{nj} - \tilde{f}_j||_2^2 &= \int_0^1 \{f_{nj}(x) - \tilde{f}_j(x)\}^2 dx \\ &\leq 2E_1 + 2E_2, \end{aligned} \tag{5.24}$$

where $E_1 = \int_0^1 \left\{\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) B_k(x)\right\}^2 dx$ and $E_2 = \left\{\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^n B_k(\hat{\zeta}_{ij})\right\}^2$. By (12) of [99], we have

$$E_1 \leq C_9 m_n^{-1} ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \tag{5.25}$$

for some positive constant C_9. The second term E_2 can be written as

$$\begin{aligned} E_2 &= \left[\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^n \{B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})\} + \sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^n B_k(\zeta_{ij})\right]^2 \\ &\leq 2E_3 + 2E_4, \end{aligned} \tag{5.26}$$

where $E_3 = \left[\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^n \{B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})\}\right]^2$ and $E_4 = \left\{\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^n B_k(\zeta_{ij})\right\}^2$. Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} E_3 &\leq ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \sum_{k=1}^{m_n} \left\{ n^{-1} \sum_{i=1}^n |B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})| \right\}^2 \\ &\leq O(m_n^3) n^{-2} \left(\sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}| \right)^2 ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \\ &\leq O(m_n^3) n^{-1} \left(\sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}|^2 \right) ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \\ &= O_p(m_n^{-1}) ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2, \end{aligned} \tag{5.27}$$

where the second inequality holds since the derivative of B-spline basis functions satisfies $||B_k'||_\infty = O(m_n)$ uniformly over $1 \leq k \leq m_n$, and the third inequality follows from Lemma

7. Next we deal with E_4. Let $h(x) = \sum_{k=1}^{m_n}(\beta_{jk} - \tilde{\beta}_{jk})B_k(x)$ for $x \in [0,1]$. Thus $h(x) \in \mathscr{S}_n$. As shown in Lemma 1 of [48], h satisfies $n^{-1}\sum_{i=1}^{n} h(\zeta_{ij}) - \mathrm{E}\, h(\zeta_{ij}) = O_p(n^{-\frac{1}{2}} m_n^{\frac{1}{2}})$. Hence

$$\begin{aligned}
E_4 &= \left\{ n^{-1}\sum_{i=1}^{n} h(\zeta_{ij}) \right\}^2 \\
&= \left[n^{-1}\sum_{i=1}^{n} \{h(\zeta_{ij}) - \mathrm{E}\, h(\zeta_{ij})\} + \mathrm{E}\, h(\zeta_{ij}) \right]^2 \\
&\leq O_p(n^{-1}m_n) + 2\{\mathrm{E}\, h(\zeta_{ij})\}^2 \\
&= O_p(n^{-1}m_n) + 2\left\{ \int_0^1 h(x) g_j(x) dx \right\}^2 \\
&\leq O_p(n^{-1}m_n) + 2C_2^2 \int_0^1 h^2(x) dx \\
&\leq O_p(n^{-1}m_n) + 2C_2^2 C_9 m_n^{-1} ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2, \qquad (5.28)
\end{aligned}$$

where the second inequality is by the Cauchy-Schwarz inequality and the last one is based on (5.25). Combining (5.24)- (5.28), we have

$$||f_{nj} - \tilde{f}_j||_2^2 = O_p(m_n^{-1}||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 + n^{-1}m_n). \qquad (5.29)$$

Since $||\tilde{f}_j - f_j||_2^2 \leq 2||f_j - f_{nj}||_2^2 + 2||\tilde{f}_j - f_{nj}||_2^2$, (5.23) and (5.29) lead to

$$||\tilde{f}_j - f_j||_2^2 = O_p(m_n^{-2\rho} + n^{-1}) + O_p(m_n^{-1}||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 + n^{-1}m_n).$$

Additionally, by part (ii) of Theorem 1, it follows that for $j \in A_1 \cup \tilde{A}_1$,

$$||\tilde{f}_j - f_j||_2^2 = O_p\left(\frac{m_n \log m_n}{n}\right) + O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho}}\right) + O_p\left(\frac{m_n \lambda_1^2}{n^2} + \frac{m_n}{n}\right).$$

This completes the proof of part (ii). □

The proof of Theorem 3 follows the proof of Corollary 2 in [48] with a similar change made to prove the part (ii) of Theorem 1, and is omitted here.

Chapter 6

Functional Single-index Quantile Regression Models

6.1 Introduction

With the development of technology, data of complex structures like growth curves, medical images and spectral data are becoming more and more popular in practice. Functional data is one important type of them. Methods for dealing with functional data have been rapidly developed and applied in fields such as agriculture, medical research and astronomy. Among them, using functional covariates to predict a scalar response plays an important role. In this chapter, we are interested in prediction of the maximal concentration of an atmospheric particulate matter called PM_{10} given the intraday PM_{10} profile of last day. More details about the data can be found in Section 6.4.

In recent years regression models with functional covariates and a scalar response has attracted substantial attention. They are referred to as scalar-on-function regression models. One can refer to [89] for more details. The most tractable one within the family is the functional linear model, which assumes that a scalar response depends linearly on a functional covariate in the sense of linear operators. But this linear assumption is susceptible to model misspecification, and thus applications of this model are restricted in practice. This concern has been raised by numerous researchers (see [13]; [71]; [72], etc.). Fully nonparametric models associated with a functional Nadayara-Watson estimator ([34]) and a functional local linear estimator ([32]) were proposed to address this issue. One concern of the fully nonparametric models is that it is onerous to interpret the effect of a functional covariate after fitting. Moreover, [13] argued that their convergence rates are quite slow and thus seeking methods that make better use of data structures is indispensable from both theoretical and practical viewpoints.

A critical idea for progress in this direction is to link a response with a linear functional of a functional covariate with some functions; it is in a similar spirit to the generalized linear model. [51] assumed that the link function was known while [70] considered a monotone but

unkonwn link function. However, as pointed out by [13], these link-function models are lack of adequate flexibility due to restrictive assumptions imposed on the link function. They, therefore, proposed extensions of these models to allow for nonparametric adaptive link functions. A faster convergence rate than that of fully nonparametric models was established in [13]. Moreover, they found that the proposed adaptive link-function model enjoyed better prediction accuracy than the models with restricted link functions in a real example.

The models introduced above mainly focus on the conditional mean structure. Thus they may not be effective tools if the conditional distribution of a response given a functional covariate is of primary interest. Conditional quantile regression models, however, have been justified to be useful in this regard. For instance, [16] considered using a monotone link function to connect the conditional distribution with a linear functional of a functional covariate; this is an indirect way to model the conditional quantile function ([53]). The proposed model was applied to a growth data set to obtain detailed knowledge of the conditional distribution of girls' heights at age 18 given their growth history between ages 1 and 12. Several methods, though not many, were proposed in literature to directly model the conditional quantile function. [53] introduced a functional principal component based approach to fit a functional linear quantile regression model, which in his opinion is a benchmark model in estimating conditional quantiles given functional covariates. In contrast, this model was fitted with smoothing splines in [12]. [33] modelled the conditional quantile in a fully nonparametric manner.

In light of the issues of linear models and full nonparametric models in estimation of the conditional mean function, we aim to seek appropriate models when the target becomes the conditional quantile function. Since the link-function regression for modeling the conditional mean function achieves good performances in both theories and applications, as shown in [13], we propose to connect the conditional quantile function with a linear functional of a functional covariate using a nonparametric adaptive link function. To our best knowledge, this is the first nonparametric link-function based framework to directly model the conditional quantile function when a functional covariate is given. To fit this model, we adopt a generalized profiling method, where two nested levels of optimization with different target functions are implemented to estimate the linear functional and the link function, respectively. Simulation studies show that the proposed estimator, compared with its counterpart of the conditional mean function, is more robust to extremely large outcomes. Furthermore, when applying these two models to predict the maximal value of PM_{10} concentrations in the real example, we find that the estimator of the conditional median outperforms its counterpart in prediction accuracy.

The remainder of the chapter is organized as follows. In Section 6.2 we introduce the functional single index quantile regression model and the generalized profiling method to fit the model. Simulation results for evaluating finite sample performances of the proposed estimator are reported in Section 6.3. We apply the new model to the PM_{10} data and

compare its prediction accuracy with that of the functional single index mean regression model proposed by [13] in Section 6.4. Section 6.5 concludes the chapter.

6.2 Model and Estimation

6.2.1 Model

Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ are identical and independent copies of (X, Y), where the covariate X is a random function defined on a compact interval $\mathcal{I} = [0, T]$ and the response Y is a scalar variable. We further postulate that X belongs to $L^2[0, T]$, a collection of all square integrable functions defined over $[0, T]$. A general model for characterizing the relationship between X and Y can be expressed as follows:

$$Y = m(X) + \text{error}, \tag{6.1}$$

where m is a functional defined on $L^2[0, T]$. In the literature of Scalar-on-Function regression, the most commonly used model assumes that m is linear. In other words, $m(X) = a + \int_{\mathcal{I}} X(t)\beta(t)dt$ for some $a \in \mathbb{R}$ and $\beta \in L^2[0, T]$. As with multiple linear models, functional linear models may suffer from inadequate fit and misspecification. These drawbacks of the linear assumption restrict its applications in practice.

Numerous approaches have been proposed to address the issues of functional linear models. For instance, [34] considered a functional Nadayara-Watson estimator of m and [32] proposed a functional local linear estimator. These fully nonparametric models and estimators can avoid model misspecifications and thus can be widely applied to predict the responses using the functional covariate. However, on the one hand, an interpretation of such models is onerous in practice. On the other hand, their performances in prediction are not satisfactory since convergence rates of them are slow ([13]). To enhance interpretability and prediction accuracy of nonparametric models, [1] and [13] proposed functional single index models:

$$Y = f\left\{\int_{\mathcal{I}} X(t)\beta(t)dt\right\} + \text{error},$$

where f is an unknown link function and β is an unknown index function satisfying $\int_{\mathcal{I}} \beta_\tau^2(t)dt < \infty$.

The conditional mean structure of Y is the target of the models mentioned above. Therefore, they are not effective tools to characterize the entire conditional distribution of Y given the functional covariate X. To achieve this goal, we propose a functional single index model for the conditional quantile of Y. More specifically, the conditional quantile of Y given X is can be expressed as follows:

$$Y = f_\tau\left\{\int_{\mathcal{I}} X(t)\beta_\tau(t)dt\right\} + \epsilon, \tag{6.2}$$

where the random error satisfies $P(\epsilon \leq 0|X(t)) = \tau$ for every $\tau \in (0,1)$. Namely, we assume that $f_\tau\{\int_{\mathcal{I}} X(t)\beta_\tau(t)dt\}$ is the τth conditional quantile of Y given X. This quantile regression model is able to accommodate both nonlinearity in the functional covariate and heteroscedasticity in the error term. For instance, assuming that $Y = \{\int_{\mathcal{I}} X(t)\beta(t)dt\}^2\epsilon$, where $\epsilon \sim N(0,1)$, the conditional τth quantile of Y given X is $\{\int_{\mathcal{I}} X(t)\beta(t)dt\}^2\Phi^{-1}(\tau)$, where Φ^{-1} denotes the quantile function of the standard normal distribution.

To address the issue of model identifiability, we assume that $\beta_\tau(0) = 1$. As is known to us, quantile regression models enjoy robustness when the response variable is skewed or heavy-tailed distributed, in contrast to mean regression models. This robust property will be investigated in both simulation studies and a real data illustration.

6.2.2 Estimation

Fitting the model (6.2) requires estimating both the index function β_τ and the link function f_τ. We first treat estimation of β_τ; the B-spline basis functions are employed for this purpose. A brief review of B-spline functions is presented as follows. Let $0 = t_0 < t_1 < \cdots < t_N = T$ be the breakpoints which separate the interval [0, T] into N subintervals. Let $\mathscr{S}_n$ denote the space of polynomial splines of degree l on $[0,T]$; it consists of functions s satisfying: (i) s is a polynomial of degree l in each subinterval $[\tau_m, \tau_{m+1}], m = 0,\ldots,N-1$; (ii) for $0 \leq l^\star \leq l-1$, the $l^\star$-th order derivative of s is continuous in $[0,T]$. Then there exist $M = N + l$ normalized B-spline basis functions $\{B_k, 1 \leq k \leq M\}$ bounded by 0 and 1 in [0, T], such that any $g \in \mathscr{S}_n$ can be written as

$$g(t) = \sum_{j=1}^{M} b_j B_k(t)$$

for $t \in [0,T]$. According to [21], the index function $\beta_\tau(t)$ can be approximated by functions in $\mathscr{S}_n$ reasonably well under some mild conditions. In particular, there exists a vector $\boldsymbol{\theta}_\tau \in \mathbb{R}^M$ such that $\beta_\tau(t) \approx \boldsymbol{\theta}_\tau^\top \boldsymbol{B}(t)$, where $\boldsymbol{B}(t) = (B_1(t),\ldots,B_M(t))^\top$. To ensure that $\beta_\tau(0) = 1$, we just need to take the first component of $\boldsymbol{\theta}_\tau$ to be 1.

Now we move to estimate the link function f_τ; the quantile spline ([73]) is considered. Let $a = \sup_X \int_0^T X(t)\beta_\tau(t)dt$ and $b = \inf_X \int_0^T X(t)\beta_\tau(t)dt$. In empirical applications, we take the maximal and minimal values of $\{\boldsymbol{\theta}_\tau^\top \int_0^T X_i(t)\boldsymbol{B}(t)dt,\ 1 \leq i \leq n\}$ as the definition of a and b for each given $\boldsymbol{\theta}_\tau$. Suppose that f_τ belongs to the Sobolev space $\mathscr{F} = \{h \mid h^{(\nu)} \text{ is absolutely continuous for } \nu = 0,1;\ J(h) < \infty\}$, where

$$J^2(h) = \int_a^b \{h^{(2)}(t)\}^2 dt.$$

Let $\rho_\tau(s) = \tau s - sI(s < 0)$ be the quantile loss function. The estimators of the spline coefficient vector $\boldsymbol{\theta}_\tau$ and the unknown function f_τ are obtained by minimizing the following

regularized empirical risk function:

$$L_\tau(f_\tau, \boldsymbol{\theta}_\tau) = n^{-1} \sum_{i=1}^{n} \rho_\tau\{Y_i - f_\tau(\boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i)\} + \lambda J^2(f_\tau), \tag{6.3}$$

where $\boldsymbol{u}_i = \left\{\int_0^T X_i(t)B_1(t)dt, \ldots, \int_0^T X_i(t)B_M(t)dt\right\}^\top$ and $\lambda > 0$ denotes a smoothing parameter.

Remark: If X_i's are not fully observed, we suggest using smoothing techniques such as local smoothers to obtain a full trajectory for each X_i.

A generalized profiling approach is adopted to combine these two estimation tasks. Estimation of $\boldsymbol{\theta}_\tau$ and estimation of f_τ are two nested levels of optimization with different target functions. In the inner level, we estimate the spline coefficient vector $\boldsymbol{\theta}_\tau$ given an estimate of f_τ by minimizing

$$\tilde{L}_\tau(\boldsymbol{\theta}_\tau | f_\tau) = n^{-1} \sum_{i=1}^{n} \rho_\tau\{Y_i - f_\tau(\boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i)\}. \tag{6.4}$$

This implies that the estimate of $\boldsymbol{\theta}_\tau$ is an implicit function of f_τ; it is denoted by $\widehat{\boldsymbol{\theta}_\tau}(f_\tau)$. We implement the *nlm* function to minimize $\tilde{L}_\tau$ with respect to $\boldsymbol{\theta}_\tau$ to obtain this function. In the outer level, we minimize the target function L_τ in (6.3) with respect to f_τ, since the estimate of the spline coefficient vector $\boldsymbol{\theta}_\tau$ is an implicit function of f_τ. To address this minimization problem, we resort to the representer theorem ([54]). The link function f_τ is an element of $\mathscr{F}$, which is a reproducing kernel Hilbert space when equipped with appropriate inner products (see Chapter 2.3 in [40]). Without loss of generality, we assume that $a = 1$ and $b = 0$. Consider transforming the argument of f_τ, x, to [0, 1] first by $(x - b)/(a - b)$ in general cases. By the representer theorem, given $\boldsymbol{\theta}_\tau$, the minimizer of (6.3) admits the following representation:

$$\widehat{f_\tau}(x) = c + dx + \sum_{i=1}^{n} c_i R_1(x, \boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i),$$

where $c, d, c_i \in \mathbb{R}$ and $R_1(x, y) = \int_0^1 (x-u)_+(y-u)_+ du$ with $(x)_+ = \max(0, x)$. Then solving the minimization problem becomes looking for c, d and c_i's which minimize L_τ in (6.3). This procedure can be implemented with the *qsreg* function in the R package **fields**.

6.2.3 Tuning Parameter Selection

There are two tuning parameters in estimation of β_τ and f_τ: the number of B-spline basis functions, M, and the smoothing parameter λ in (6.3). A small M may not be able to provide an adequate approximation of β_τ, while a large M may lead to an excessively

wiggly estimate of β_τ. The smoothing parameter λ controls the trade off between fidelity to the data and complexity of the link function f_τ.

We employ the generalized approximate cross-validation (GACV) proposed by [117] to choose these two tuning parameters. For each pair (M, λ), the GACV is defined by

$$\text{GACV}(M, \lambda) = \frac{\sum_{i=1}^{n} \rho_\tau\{Y_i - \widehat{f_\tau}(\widehat{\boldsymbol{\theta}_\tau}^\top \boldsymbol{u}_i)\}}{n - tr(H)}, \tag{6.5}$$

where $\widehat{f_\tau}$ and $\widehat{\boldsymbol{\theta}_\tau}$ denote the point estimator of f_τ and $\boldsymbol{\theta}_\tau$ by minimizing (6.3) with the generalized profiling approach, and H denotes a hat matrix with the (i, j)th entry $\partial\widehat{f_\tau}(\widehat{\boldsymbol{\theta}_\tau}^\top \boldsymbol{u}_i)/\partial Y_j$. We select the pair of M and λ which can minimize the GACV as the "optimal" tuning parameters.

6.3 Simulation Studies

In this section we aim to study the finite sample performance of the functional single index quantile regression model (6.2).

The functional covariates are identically and independently generated as:

$$X_i(t) = t + \sum_{k=1}^{4} \phi_k(t)\xi_{ik}(t), \quad i = 1, \ldots, n,$$

where $\phi_1(t) = \sin(2\pi t)/\sqrt{2}$, $\phi_2(t) = \cos(2\pi t)/\sqrt{2}$, $\phi_3(t) = \sin(4\pi t)/\sqrt{2}$, $\phi_4(t) = \sin(4\pi t)/\sqrt{2}$ and $\xi_{ik} \sim N(0, \lambda_k)$ are independent with $\lambda_k = 0.5^{k-1}$, $k = 1, 2, 3, 4$. These covariates are sampled at 100 equally spaced points between 0 and 1. The responses are generated as follows:

$$Y_i = f\left\{\int_0^1 X_i(t)\beta(t)dt\right\} + \epsilon_i, \quad i = 1, \ldots, n, \tag{6.6}$$

where $\beta(t) = 2t^2 + 0.25t + 1$ and $f(x) = 0.66\exp(x^2)$. The random errors, ϵ_i's, are generated in three ways: (i) $\epsilon_i \sim$ Laplace$(\mu = 0, b = 1)$, (ii) $\epsilon_i \sim N(0, 1)$, (iii) $\epsilon_i \sim 0.85 \cdot t_{\nu=3}$, for better assessment of the finite sample performance of the proposed estimator in different scenarios. The signal-to-noise ratios are all around 2 in these designs,.

To assess the performance of the proposed estimator in prediction, we compare it with the estimator for the functional single index model proposed in [13]. Our estimator models the conditional quantiles of Y_i's, while the estimator by [13] concerns the conditional mean. Therefore, we denote our estimator by FSiQ and their estimator by FSiM throughout the chapter. In each simulation replicate we randomly generate $n = 600$ independent copies of (X_i, Y_i), of which 500 are used for training and the rest are for testing. We repeat the procedure 100 times for evaluating variations in predictions across simulation replicates.

In these three designs, the conditional median of Y_i is $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\}$, which is also the conditional mean of Y_i. We compare these two estimators in terms of prediction accuracy. More specifically, after estimating both the index function and the link function using the training data, we will then compute the root mean squared errors on the test data:

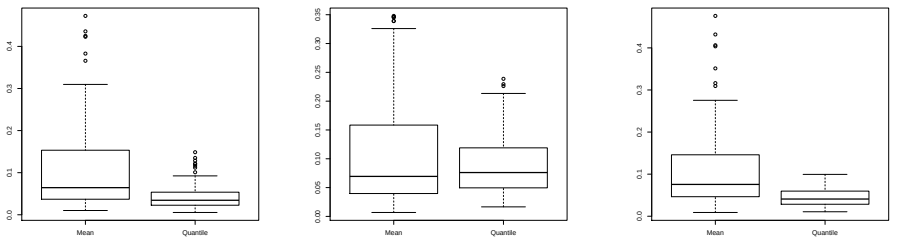
$$\text{RMSE} = \sqrt{\frac{\sum_{i\in\text{test}}\left[f\left\{\int_0^1 X_i(t)\beta(t)dt\right\} - \hat{f}\left\{\int_0^1 X_i(t)\hat{\beta}(t)dt\right\}\right]^2}{100}}.$$

In addition, performances in estimation of the index function β are compared as well. In each simulation replicate, the root integrated squared error for β is computed:

$$\text{RISE} = \sqrt{\int_0^1 \{\beta(t) - \hat{\beta}(t)\}^2 dt}.$$

Distribution	RMSE		RISE	
	FSiQ	FSiM	FSiQ	FSiM
Laplace	.197 (.071)	.304 (.147)	.123 (.049)	.125 (.029)
Normal	.288 (.087)	.300 (.136)	.118 (.049)	.138 (.045)
Student t	.203 (.051)	.311 (.136)	0.122 (.056)	.132 (.040)

Table 6.1: Summary of the averages and the standard errors (in brackets) of RMSEs and RISEs across the 100 replicates in the three designs.



Laplace distribution Normal distribution Student t distribution

Figure 6.1: Boxplots of RMSEs across the 100 replicats. In each panel, the left side denotes results from FSiM while the right side is for FSiQ.

Table 6.1 summarizes the comparison of the averages and the standard errors of RMSEs and RIMSEs of these two estimators in the three designs. We find that the conditional distribution of the response has a negligible effect on estimation of the index function. However, there exist remarkable differences in the comparison of predictions between these two methods for different designs. When the conditional distribution is light tailed, there's little difference in predictions between these two estimators, as can be seen from the row corresponding to "Normal" in Table 6.1 and Figure 6.1. But obvious distinctions can be seen

when the conditional distribution becomes heavy tailed. These distinctions imply that when extremely large values occur frequently in the response, the estimator from the functional single index mean regression model is less accurate than its counterpart from the quantile regression model. In other words, the functional single index quantile regression model is more robust to extremely large outcomes.

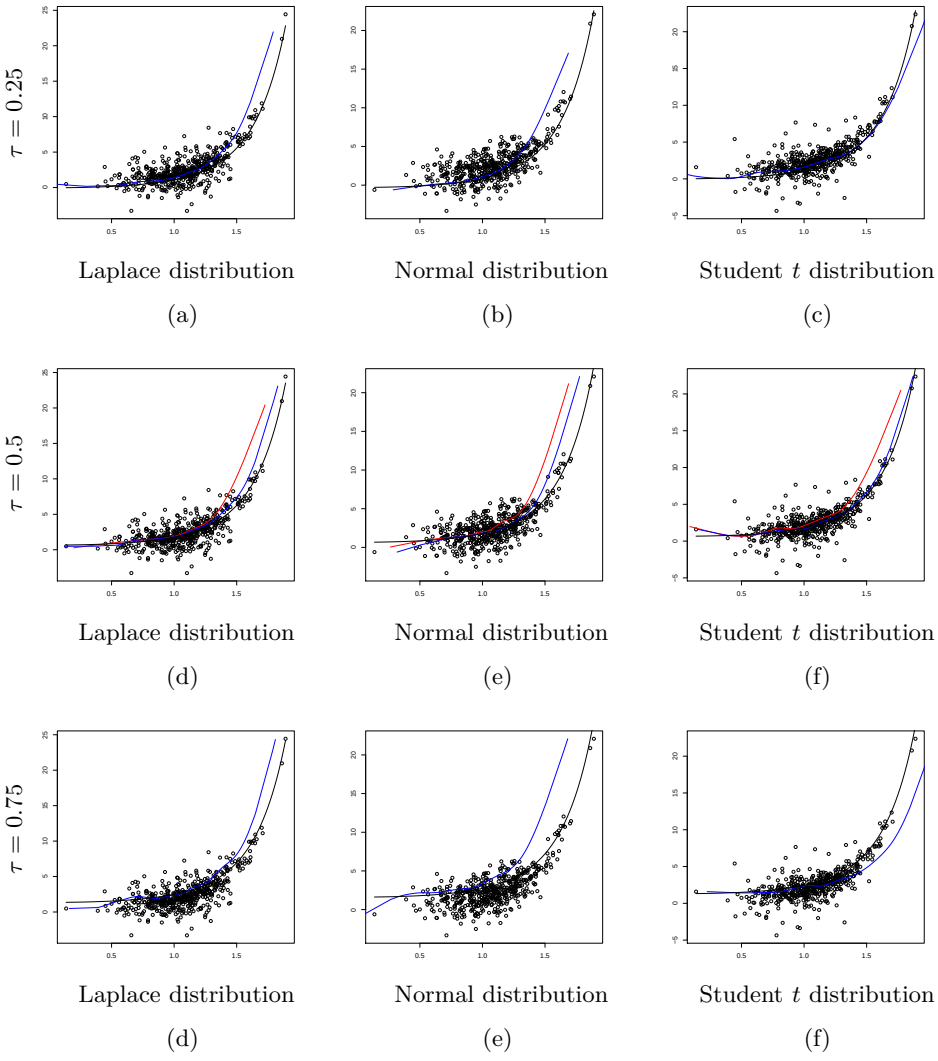


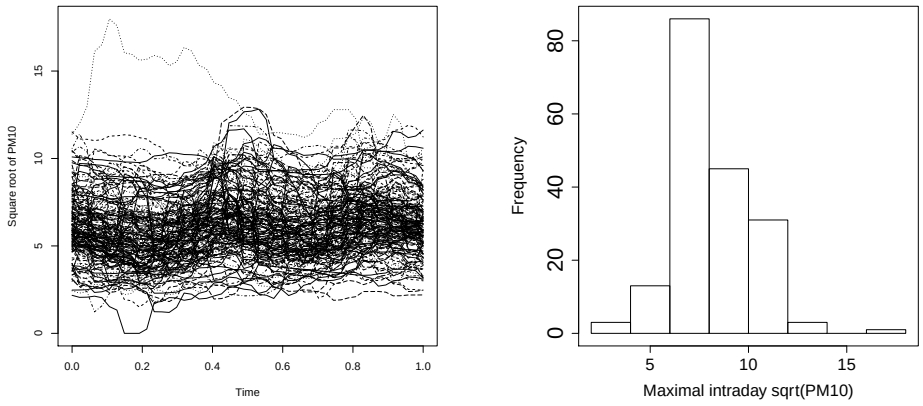
Figure 6.2: Estimators of the link function for $\tau = 0.25, 0.5, 0.75$ in the three designs, respectively. In each panel, the black line denotes the true link function while the blue line represents the estimator from FsiQ. In the middle row, the red lines denote the estimated link function from FSiM.

In addition to the robust property of the quantile regression model, we are also interested in whether the proposed estimation scheme is able to provide a reasonable estimate of conditional quantiles. Figure 6.2 displays the estimated conditional quantiles from one randomly selected simulation replicate. Three quantile levels ($\tau = 0.25, 0.5, 0.75$) are con-

sidered in Figure 6.2; the estimated conditional quantile functions are reasonably close to the true ones in all three designs. To provide a direct comparison of the two models, we present the estimated conditional mean functions in the middle row of Figure 6.2. Fitting these two models yields consistent estimators when the conditional distribution is light tailed. The estimator of FSiQ, however, is considerably closer to the true conditional median (mean) function in the Student t case. This suggests that the performance of FSiM might be impaired when extremely large values are frequently observed in the response.

6.4 Real Data Illustration

PM_{10} is defined to be subtypes of atmospheric particles, such as suspended particulate matter, thoracic and respirable particles and inhalable coarse particles, with a diameter greater than 2.5 and less than 10 micrometers. [80] argued that these subtypes are so small that they are able to penetrate the respiratory tract of humans. As a result, the normal functioning of the organism might be impaired by them. According to [80], if the maximal value of 24 hour moving average of PM_{10} concentrations is above 240 $\mu g/m^3$, humans' health may be in serious risk. There has been extensive research on modelling and forecasting PM_{10} since its concentration is closely related to human healty. However, considerably less attention has been focused on modelling the maximal value of daily PM_{10} concentrations.

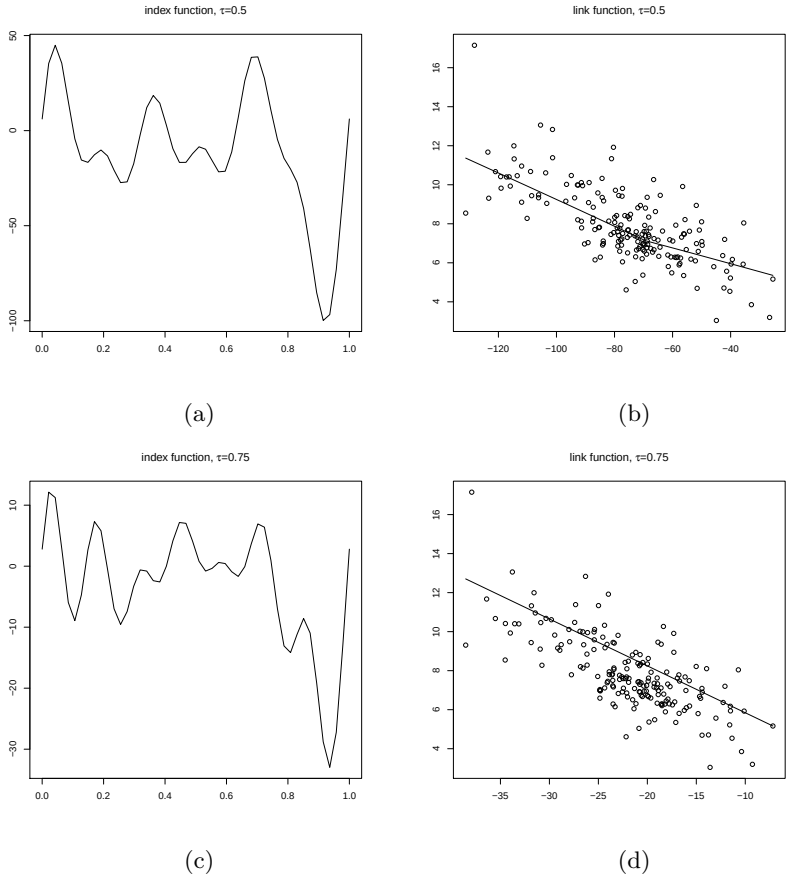


(a) (b)

Figure 6.3: (a) Profiles of half-hourly square root of PM_{10} concentrations from October 1st, 2010 to March 31st, 2011. (b)Histogram of maximal values of square root of intraday PM_{10} concentrations.

[46] carried out a dynamic functional principal component analysis of PM_{10} data collected in Graz, Austria. This data set consists of half-hourly measurements of PM_{10} concentrations (in $\mu g/m^3$) from October 1st, 2010 to March 31st, 2011. They suggested performing a square-root transformation to stabilize the variance and mitigate heavy-tailed observa-

tions. The profiles of half-hourly measurements of PM_{10} concentrations after the square-root transformation are displayed in the left panel of Figure 6.3, where the time scale has been transformed to [0, 1]. Heavy-tailed observations still exist after the transformation. This is further justified by the right panel of Figure 6.3; it depicts the histogram of maximal values of square root of intraday PM_{10} concentrations, sqrt(PM_{10}) in short. We are interested in using the current intraday profile to predict the maximal sqrt(PM_{10}) of the next day. 15 Fourier basis functions are employed to transform the raw observations into functional data, as suggested by [46]. This transformation can be implemented with the *Data2fd* function in the **fda** package.



(a) (b)

(c) (d)

Figure 6.4: The estimated index functions and link functions. Left two panels: the estimated index functions for $\tau = 0.5$ and $\tau = 0.75$, respectively. Right two panels: the estimated link functions for $\tau = 0.5$ and $\tau = 0.75$, respectively.

We use the GACV proposed in Section 6.2.3 to select both the number of B-spline basis functions and the smoothing parameter λ to fit the functional single index quantile regression. The estimators of the index function and the link function are obtained by minimizing L_τ is (6.3) with the proposed generalized profiling method. The estimated index functions and link functions are shown in Figure 6.4. We find that the shape of the index

function for $\tau = 0.5$ is different from that for $\tau = 0.75$. This suggests that the functional single index mean regression is inadequate to characterize the conditional distribution of the maximal value of intraday sqrt(PM_{10}) concentrations.

As shown in simulation studies, FSiQ compares favorably with FSiM in estimation of the target function when the scalar response is heavy-tailed distributed. We further compare these two methods in terms of prediction accuracy in this real example. Five-fold cross-validation is adopted to compute the mean squared errors (MSPEs) defined by

$$\text{MSPE} = \frac{\sum_{i \in \text{test set}} \left[Y_i - \hat{f} \left\{ \int_0^1 X_i(t) \hat{\beta}(t) dt \right\} \right]^2}{|\text{test set}|},$$

where |test set| denotes the cardinal number of a test set and $\hat{f}$ and $\hat{\beta}$ are obtained by taking $\tau = 0.5$ in FSiQ. The whole data set is randomly splitted 100 times and the distributions of the corresponding MSPEs obtained from FSiM and FSiQ are displayed in Figure 6.5. Obviously FSiQ outperforms FSiM in predicting the maximal intraday sqrt(PM_{10}) of the next day with the profile of the current day. A possible reason for this remarkable gap in prediction accuracy is that the distribution of the maximal intraday sqrt(PM_{10}) is skewed and some extremely large values exist in the response; these can be found in the right panel of Figure 6.3.

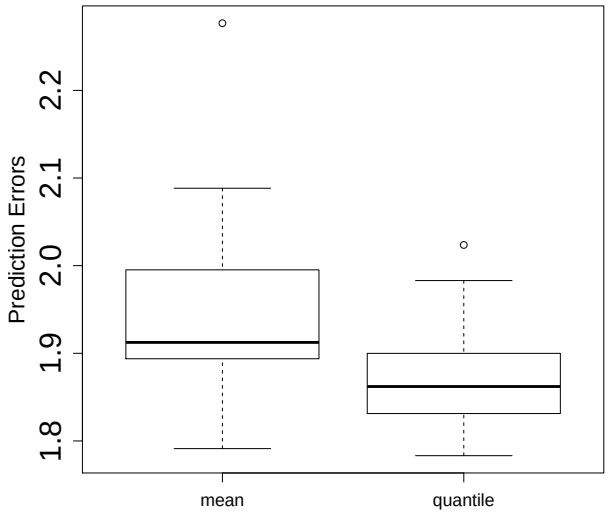


Figure 6.5: Boxplots of mean squared prediction errors (MSPE) of FSiM and FSiQ across the 100 random splits. The left side and the right side represent MSPEs for FSiM and FSiQ, respectively.

6.5 Conclusion

Scalar-on-function regression models play a significant role in characterizing the relationship between a scalar response and a functional covariate. There has been an extensive research on modelling the conditional mean structure of a scalar response given a functional covariate. These models such as functional single index regression models and fully nonparametric functional models are widely applied in practice for prediction. However, they are not able to provide an adequate characterization of the conditional distribution of the scalar response. Moreover, their prediction accuracy might be impaired when the distribution of the scalar response is skewed or heavy tailed.

In the chapter, we have proposed a functional single index quantile regression model which concerns the conditional quantile of a scalar response. It, therefore, enables us to investigate the conditional distribution of the response in details. A generalized profiling method is proposed to fit this model. This method consists of two nested levels of optimization with different target functions. In the inner level, we use B-spline basis functions to approximate the index function and represent the estimated coefficient vector in terms of the link function. In the outer level, we resort to the representer theorem to simplify the minimization task that searchs a minimizer of a regularized empirical risk function from an infinite dimensional functional space. Simulation studies are conducted to compare the finite sample performances of the proposed estimator and that for the functional single index mean regression model. We find that the performance of the proposed estimator is less susceptible to extremely large outcomes compared with its counterpart. Their performances in prediction accuracy is further compared in an real example, which concerns using the current intraday PM_{10} concentrations to predict the maximal value of the next day. The comparison result demonstrates that the proposed estimator is superior to its counterpart in terms of prediction accuracy when the scalar response is skewed distributed and contains numerous extremely large values.

Chapter 7

Conclusions and Future Work

A new approach to implement functional principal component analysis (FPCA) was proposed in Chapter 2. With the aid of polynomial functions, the proposed approach, called parametric FPCA, yields functional principal components that have closed forms and thus are easier to interpret. Moreover, we found that the paramerric FPCA is more robust than the traditional nonparametric FPCA when outlier curves exist. It is worthwhile to study this new FPCA approach in terms of recovering trajectories from noisy functional data and dimension reduction in regression models of functional data.

We proposed a new approach to calibrate dynamical correlations between random functions in Chapter 3. The new framework for statistical inference for the dynamical correlation relies on the weighted empirical likelihood. As a result, the test statistic, under the null hypothesis, can be calibrated using χ^2 distribution, which is self-normalized. Therefore the procedure of estimating the variance of the test statistics with the bootstrap approach can be avoided in the new approach; it is the main reason why our approach is considerably more efficient in computation. Though the dynamical correlation can reflect the dynamical feature of associations between random functions by adaptively choosing the weight function, developing a new framework to reveal this dynamical feature in a more explicit manner is worthwhile to consider in quantifying dynamical functional connectivity of brain signals. It also provides another tool to determine time-varying networks in gene regulations.

In Chapter 4, a semiparametric framework was proposed to model the effects of a functional covariate and several scalar covariates on a scalar response. The effect of the functional covariate is modelled in a nonparametric manner while the effect of scalar covariates is modelled linearly. A regularization scheme was considered to achieve sparsity and smoothness in estimation of the nonparametric part. Our asymptotic theory is established upon the assumption the true FPC scores are known; it is impossible in practice. We, therefore, are interested in establishing the theorems based on the estimated FPC scores, rather than the true scores, in future work. Additionally, We are interested in extending the framework to the scenarios where responses of interest are from the exponential family.

To tackle the issue of inadequate fit of functional linear regression, we assume that the conditional mean of a scalar response depends on an additive form of the leading FPC scores of a functional covariate in Chapter 5. Some of the nonparametric components are penalized to be vanishing in estimation to enhance prediction accuracy and interpretability of the model. We propose using the adaptive group LASSO method to select relevant components and smoothing splines to refine estimator of those relevant components. One of the limitation is that we assume that the number of nonparametric components in the model is independent of sample size. From theoretical viewpoints, it would be worthwhile to allow for this number diverges as sample size increases. Additionally, even though the model is considerably less restrictive than the functional linear regression, there's no justification why the scalar response depends on the leading FPC scores of the functional covariance in an additive manner. A model which can accommodate more general forms in calibrating the effects of leading FPC scores on the scalar response would be desirable.

We proposed a functional single index model to estimate the conditional quantile of a scalar response. Compared with functional mean regression models, the proposed model have two main advantages. The first advantage is that a full picture of the conditional distribution would be available from the quantile regression model. The second advantage is that the performance of the quantile regression model is more robust to skewness or extremely large values in the response. In our framework, only one index function and one link function are allowed. We are interested in developing algorithms to allow for multiple indices to enhance flexibility and prediction accuracy of the current model in future work.

Bibliography

[1] Ahmed Ait-Saïdi, Frederic Ferraty, Rabah Kassa, and Philippe Vieu. Cross-validated estimations in the single-functional index model. *Statistics*, 42(6):475–494, 2008.

[2] RW Atkinson, S Kang, HR Anderson, IC Mills, and HA Walton. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*, 69:660–665, 2014.

[3] Veerabhadran Baladandayuthapani, Bani K Mallick, Mee Young Hong, Joanne R Lupton, Nancy D Turner, and Raymond J Carroll. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64(1):64–73, 2008.

[4] Rina Foygel Barber, Matthew Reimherr, and Thomas Schill. The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11(1):1351–1389, 2017.

[5] Peter J Basser, James Mattiello, and Denis LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

[6] Sonia Bastidas, Frederik Graw, Miranda Z Smith, Herbert Kuster, Huldrych F Günthard, and Annette Oxenius. CD8+ T cells are activated in an antigen-independent manner in HIV-infected individuals. *The Journal of Immunology*, 192(4):1732–1744, 2014.

[7] Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The Neuro Bureau ADHD-200 Preprocessed repository. *Neuroimage*, 144:275–286, 2017.

[8] PC Berquin, JN Giedd, LK Jacobsen, SD Hamburger, AL Krain, JL Rapoport, and FX Castellanos. Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. *Neurology*, 50(4):1087–1093, 1998.

[9] Philippe Besse. PCA stability and choice of dimensionality. *Statistics & Probability Letters*, 13(5):405–410, 1992.

[10] Philippe C Besse, Hervé Cardot, and Frédéric Ferraty. Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis*, 24(3):255–270, 1997.

[11] Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149. New York: Springer, 2000.

[12] Hervé Cardot, Christophe Crambes, and Pascal Sarda. Quantile regression when the covariates are functions. *Nonparametric Statistics*, 17(7):841–856, 2005.

[13] Dong Chen, Peter Hall, and Hans-Georg Müller. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747, 2011.

[14] Hung Chen. Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, 16(1):136–146, 1988.

[15] Kehui Chen and Jing Lei. Localized functional principal component analysis. *Journal of the American Statistical Association*, 110:1266–1275, 2015.

[16] Kehui Chen and Hans-Georg Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):67–89, 2012.

[17] Daniel P Connell, Jeffrey A Withum, Stephen E Winter, Robert M Statnick, and Richard A Bilonick. The Steubenville Comprehensive Air Monitoring Program (SCAMP): associations among fine particulate matter, co-pollutants, and meteorological conditions. *Journal of the Air & Waste Management Association*, 55(4):481–496, 2005.

[18] C Keith Conners, Drew Erhardt, and Elizabeth P Sparrow. Conners' adult ADHD rating scales (CAARS): technical manual. Multi-helath Systems, Toronto, 1999.

[19] Dan L Crouse, Paul A Peters, Perry Hystad, Jeffrey R Brook, Aaron van Donkelaar, Randall V Martin, Paul J Villeneuve, Michael Jerrett, Mark S Goldberg, C Arden Pope III, et al. Ambient $PM_{2.5}$, O_3, and NO_2 exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environmental health perspectives*, 123(11):1180–1186, 2015.

[20] Jacques Dauxois, Alain Pousse, and Yves Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, 12(1):136–154, 1982.

[21] C De Boor. *A practical guide to splines*. New York: Springer-Verlag, 2001.

[22] Carl De Boor. On uniform approximation by splines. *Journal of Approximation Theory*, 1(2):219–235, 1968.

[23] Olivier Devos and Ludovic Duponchel. Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 107(1):50–58, 2011.

[24] Joel Dubin, Mike Li, Dandi Qiao, and Hans-Georg MÃijller. *dynCorr: Dynamic Correlation Package*, 2017. R package version 1.1.0.

[25] Joel A Dubin and Hans-Georg Müller. Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881, 2005.

[26] Paul HC Eilers and Brian D Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102, 1996.

[27] Robert F Engle, Clive WJ Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986.

[28] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Chapman and Hall/CRC, New York, 1996.

[29] Heidi M Feldman and Michael I Reiff. Attention deficit–hyperactivity disorder in children and adolescents. *New England Journal of Medicine*, 370(9):838–846, 2014.

[30] F Ferraty and Ph Vieu. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametric Statistics*, 16:111–125, 2004.

[31] Frédéric Ferraty. *Recent advances in functional data analysis and related topics*. Springer, New York, 2011.

[32] Frédéric Ferraty, Ingrid Van Keilegom, and Philippe Vieu. On the Validity of the Bootstrap in Non-Parametric Functional Regression. *Scandinavian Journal of Statistics*, 37(2):286–306, 2010.

[33] Frédéric Ferraty, Abbes Rabhi, and Philippe Vieu. Conditional quantiles for dependent functional data with application to the climatic El Niño phenomenon. *Sankhyā: The Indian Journal of Statistics*, 67:378–398, 2005.

[34] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, New York, 2006.

[35] Polly Christine Ford-Jones. Misdiagnosis of attention deficit hyperactivity disorder: "normal behaviour" and relative maturity. *Paediatrics & Child Health*, 20(4):200–202, 2015.

[36] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer, New York, 2001.

[37] Semyon Aranovich Gershgorin. Uber die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk (in German)*, (6):749–754, 1931.

[38] Jeff Goldsmith and Fabian Scheipl. Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis*, 70:362–372, 2014.

[39] Spencer Graves, Giles Hooker, and J Ramsay. *Functional data analysis with R and MATLAB*. Springer, New York, 2009.

[40] Chong Gu. *Smoothing spline ANOVA models*. New York: Springer, 2013.

[41] Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially linear models*. Springer, New York, 2000.

[42] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[43] Guozhong He, Hans-Georg Müller, and Jane-Ling Wang. Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis*, 85(1):54–77, 2003.

[44] Nancy E Heckman. Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2):244–248, 1986.

[45] Nancy E Heckman and Ruben H Zamar. Comparing the shapes of regression functions. *Biometrika*, 87(1):135–144, 2000.

[46] Siegfried Hörmann, Łukasz Kidziński, and Marc Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):319–348, 2015.

[47] Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*. Springer, New York, 2012.

[48] Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282, 2010.

[49] Ayaz Hyder, Hyung Joo Lee, Keita Ebisu, Petros Koutrakis, Kathleen Belanger, and Michelle Lee Bell. PM2. 5 exposure and birth outcomes: use of satellite-and monitor-based data. *Epidemiology*, 25(1):58–67, 2014.

[50] Andrada E Ivanescu, Ana-Maria Staicu, Fabian Scheipl, and Sonja Greven. Penalized function-on-function regression. *Computational Statistics*, 30(2):539–568, 2015.

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

[52] Richard A Kaslow, David G Ostrow, Roger Detels, John P Phair, B Frank Polk, and Charles R Rinaldo. The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2):310–318, 1987.

[53] Kengo Kato. Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136, 2012.

[54] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[55] Dehan Kong, Kaijie Xue, Fang Yao, and Hao H Zhang. Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159, 2016.

[56] Rakesh Kumar and Abba Elizabeth Joseph. Air pollution concentrations of $PM_{2.5}$, PM_{10} and NO_2 at ambient and kerbsite and their correlation in Metro City–Mumbai. *Environmental Monitoring and Assessment*, 119(1-3):191–199, 2006.

[57] Xiaoyan Leng and Hans-Georg Müller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2005.

[58] Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095, 2017.

[59] Haocheng Li, John Staudenmayer, and Raymond J Carroll. Hierarchical functional data with mixed continuous and binary measurements. *Biometrics*, 70(4):802–811, 2014.

[60] Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

[61] Zhenhua Lin, Jiguo Cao, Liangliang Wang, and Haonan Wang. Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26:306–318, 2017.

[62] Zhenhua Lin, Liangliang Wang, and Jiguo Cao. Interpretable functional principal component analysis. *Biometrics*, 72(3):846–854, 2016.

[63] Siwei Liu, Yang Zhou, Richard Palumbo, and Jane-Ling Wang. Dynamical correlation: A new method for quantifying synchrony with multivariate intensive longitudinal data. *Psychological methods*, 21(3):291–308, 2016.

[64] Ying Lu, Jiang Du, and Zhimeng Sun. Functional partially linear quantile regression model. *Metrika*, 77(2):317–332, 2014.

[65] Enno Mammen and Sara van de Geer. Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, 25(3):1014–1035, 1997.

[66] William J Marshall, Christine L Lackner, Paul Marriott, Diane L Santesso, and Sidney J Segalowitz. Using phase shift granger causality to measure directed connectivity in eeg recordings. *Brain connectivity*, 4(10):826–841, 2014.

[67] André Mas. Local functional principal component analysis. *Complex Analysis and Operator Theory*, 2(1):135–167, 2008.

[68] Robert Miller and Alexey M Ivanitsky. *Complex brain functions: conceptual advances in Russian neuroscience*. CRC Press, London, 2003.

[69] Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

[70] Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33:774–805, 2005.

[71] Hans-Georg Müller, Yichao Wu, and Fang Yao. Continuously additive models for nonlinear functional regression. *Biometrika*, 100(3):607–622, 2013.

[72] Hans-Georg Müller and Fang Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008.

[73] Doug Nychka, Gerry Gray, Perry Haaland, David Martin, and Michael O'connell. A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*, 90(432):1171–1178, 1995.

[74] Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston. *Handbook of Neuroimaging Data Analysis*. CRC Press, London, 2016.

[75] Rainer Opgen-Rhein and Korbinian Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *RevStat*, 4(1):53–65, 2006.

[76] Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[77] Art B Owen. *Empirical likelihood*. Chapman and Hall, London, 2001.

[78] Roger D Peng and Francesca Dominici. *Statistical methods for environmental epidemiology with R*. New York: Springer, 2008.

[79] Roger D Peng and Leah J Welty. The NMMAPSdata package. *R news*, 4(2):10–14, 2004.

[80] Patricio Perez and Jorge Reyes. Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. *Atmospheric Environment*, 36(28):4555–4561, 2002.

[81] S Pezzulli and BW Silverman. Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8:1–16, 1993.

[82] Jana Prochazkova. Derivative of B-spline function. In *Proceedings of the 25th Conference on Geometry and Computer Graphics*. Prague, Czech Republic, 2005.

[83] Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2017.1390466, 2017.

[84] James O Ramsay and CJ Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:539–572, 1991.

[85] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, New York, 2002.

[86] James O Ramsay and Bernard W Silverman. *Functional Data Analysis*. New York: Springer, second edition, 2005.

[87] Jim O Ramsay. Functional components of variation in handwriting. *Journal of the American Statistical Association*, 95(449):9–15, 2000.

[88] Claudia Rangel, John Angus, Zoubin Ghahramani, Maria Lioumi, Elizabeth Sotheran, Alessia Gaiba, David L Wild, and Francesco Falciani. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

[89] Philip T Reiss, Jeff Goldsmith, Han Lin Shang, and R Todd Ogden. Methods for Scalar-on-Function Regression. *International Statistical Review*, 85(2):228–249, 2017.

[90] John A Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14(3):631–648, 2004.

[91] Hannah Ritchine and Max Roser. Air Pollution. *OurWorldInData*, 2018.

[92] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Cambridge University Press, Cambridge, 2003.

[93] Federico M San Martini, Christa A Hasenkopf, and David C Roberts. Statistical analysis of PM2.5 observations from diplomatic facilities in China. *Atmospheric Environment*, 110:174–185, 2015.

[94] Fabian Scheipl, Ana-Maria Staicu, and Sonja Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015.

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

[96] Bernard W Silverman et al. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

[97] Paul Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):413–436, 1988.

[98] Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

[99] Charles J Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986.

[100] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[101] Sara A van de Geer. Regression analysis and empirical processes. *CWI Tracts*, 45:1–161, 1988.

[102] Sara A van de Geer. *Empirical Processes in M-estimation*. Cambridge: Cambridge Univ. Press, 2000.

[103] Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. New York: Springer, 1996.

[104] J-P Vila, Véréne Wagner, and Pascal Neveu. Bayesian nonlinear model selection and neural networks: a conjugate prior approach. *IEEE Transactions on neural networks*, 11(2):265–278, 2000.

[105] Grace Wahba. *Cross validated spline methods for the estimation of multivariate functions from data on functionals*. University of Wisconsin, Department of Statistics, 1984.

[106] Grace Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[107] Xiao Wang and David Ruppert. Optimal prediction in an additive functional model. *Statistica Sinica*, 25(2):567–589, 2015.

[108] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[109] Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

[110] Changbao Wu. Weighted empirical likelihood inference. *Statistics & probability letters*, 66(1):67–79, 2004.

[111] Hulin Wu, Tao Lu, Hongqi Xue, and Hua Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716, 2014.

[112] Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian. The impact of PM2. 5 on the human respiratory system. *Journal of thoracic disease*, 8(1):69–74, 2016.

[113] Wenjing Yang, Hans-Georg Müller, and Ulrich Stadtmüller. Functional singular component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):303–324, 2011.

[114] Fang Yao and Hans-Georg Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010.

[115] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[116] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

[117] Ming Yuan. GACV for quantile smoothing splines. *Computational statistics & data analysis*, 50(3):813–829, 2006.

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

[119] Hao Helen Zhang and Yi Lin. Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, 16(3):1021–1041, 2006.

[120] Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang, and Ann Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995.

[121] Hongxiao Zhu, Fang Yao, and Hao Helen Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):581–603, 2014.

[122] Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.