Two-sample Inference, Order Determination, and Data Integration for Functional Data

by

Chi Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2025

© Chi Zhang 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:
Luo Xiao
Associate Professor, Department of Statistics
North Carolina State University

Supervisor(s):
Peijun Sang
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo

Yingli Qin
Associate Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal Member:
Pengfei Li
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Mu Zhu
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal-External Member: Stanko B. Dimitrov
Professor, Department of Management Sciences
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The chapters of this thesis describe and contain Chi Zhang's work, under the supervision of Professors Peijun Sang and Yingli Qin, that has appeared in the following preprints and published works.

Chapter 2: Zhang, C., Sang, P., and Qin, Y. (2025). Two-sample Inference for Sparse Functional Data. *Electronic Journal of Statistics*, 19(1):792–864.

Chapter 3: Zhang, C., Sang, P., and Qin, Y. (2025). Order Determination for Functional Data. *arXiv:2503.03000*.

Chapter 4: Zhang, C., Sang, P., and Qin, Y. (2025). Learning Shared and Source-specific Subspaces across Multiple Data Sources for Functional Data. *arXiv:2510.13010*.

Abstract

Functional data analysis has gained increasing prominence in modern statistics, largely due to advancements in data collection technologies. It provides a nonparametric framework for analyzing discrete observations obtained from realizations of a continuous random function, often defined over time or space. In this thesis, we focus on three distinct problems, each reflecting a different aspect of functional data analysis.

In Chapter 2, we address the problem of comparing mean functions between two groups of sparse functional data within the framework of a reproducing kernel Hilbert space. The proposed method is well-suited for sparsely and irregularly sampled functional data. Traditional approaches often assume homogeneous covariance structures across groups, an assumption that is difficult to justify in practice. To circumvent this limitation, we first develop a novel linear approximation for the mean estimator, which naturally leads to its desirable pointwise limiting distributions. Furthermore, we establish the weak convergence of the mean estimator, enabling the construction of a test statistic for the mean differences. The finite-sample performance of our method is demonstrated through extensive simulations and two real-world applications.

In Chapter 3, we study the problem of determining the number of eigenpairs to retain in functional principal component analysis—a problem commonly referred to as order determination. When a covariance function admits a finite representation, the challenge becomes estimating the rank of the corresponding covariance operator. While this problem is straightforward when the full trajectories of functional data are available, in practice, functional data are typically collected discretely and are subject to measurement error contamination. Such contamination introduces a ridge in the empirical covariance function, obscuring the true rank. We develop a novel procedure to identify the true rank of the covariance operator by leveraging the information of eigenvalues and eigenfunctions. By incorporating smoothing techniques to accommodate the nonparametric nature of functional data, the method is applicable to functional data collected at random, subject-specific points. Extensive simulation studies demonstrate the excellent performance of our approach across a wide range of settings, outperforming commonly used information-criterion-based methods and maintaining effectiveness even in high-noise scenarios. We

further illustrate our method with two real-world data examples.

In Chapter 4, we investigate the integration of multi-source functional data to extract a subspace that captures the variation shared across sources. In practice, data collection procedures often follow source-specific protocols. Directly averaging sample covariance operators across sources implicitly assumes homogeneity, which may lead to biased recovery of both shared and source-specific variation structures. To address this issue, we propose a projection-based data integration method that explicitly separates the shared and source-specific subspaces. The method first estimates source-specific projection operators via smoothing to accommodate the nonparametric nature of functional data. The shared subspace is then isolated by examining the eigenvalues of the averaged projection operator across all sources. If a source-specific subspace is of interest, we re-project the associated source-specific covariance estimator onto the subspace orthogonal to the estimated shared subspace, and estimate the source-specific subspace from the resulting projection. We further establish the asymptotic properties of both the shared and source-specific subspace estimators. Extensive simulation studies demonstrate the effectiveness of the proposed method across a wide range of settings. Finally, we illustrate its practical utility with an example of air pollutants data.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Peijun Sang and Yingli Qin, without whose guidance this thesis would not have been possible. We met almost every week over the past four years. Peijun has offered many invaluable suggestions from his unique perspective, encouraging me to grow as an independent researcher, whereas Yingli has always listened to my questions and concerns with patience, kindness, and thoughtfulness. They have set excellent examples of how to become successful researchers, and their unwavering support has guided me through every stage of my doctoral studies.

I would also like to extend my sincere thanks to the professors who served as my committee members: Pengfei Li, Mu Zhu, Stanko B. Dimitrov, and Luo Xiao. I am deeply grateful for the time and effort you devoted to reviewing my thesis. Your valuable insights and constructive comments have greatly improved this work.

There are several other professors who have generously shared their advice on research, life, and beyond. My master's supervisor, Gregory Rice, definitely deserves a special mention. I have always enjoyed the walks we took together, as well as our conversations about research and many other topics beyond it. I would also like to express my gratitude to Professors Qinglong Tian, Aukosh Jagannath, Paul Marriott, and Kun Liang for their valuable suggestions on research and for sharing their experiences with me.

I gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Government of Ontario for their financial support.

My deep appreciation goes to Mary Lou Dufton for her support and encouragement throughout my doctoral journey. I would never have become a recipient of the NSERC scholarships without her encouragement.

I started my journey at the University of Waterloo as a master's student, and I have had the privilege of meeting numerous wonderful people along the way. Hao-an Hung and Peiwei Wang—I will always remember the days and nights we spent together, whether on projects or playing games. I would also like to thank my (semi-)roommates—Weijia Cai, Keqi Li, Shaowen Liu, and Zhaoxin Ye—for all the joyful moments we shared. Chi-Kuang Yeh, Jingyue Huang, Jianchu Chen, and Yimiao Zhao—the adventures of capturing

Pokémon and raiding bosses with you have added many colorful memories to my doctoral life. Xisi Zhang, Tianyi Pan, Henan Xu, Yicheng Qin, and Yixing Zhao—playing soccer with you has always been truly enjoyable. I also extend my heartfelt appreciation to Meixi Chen, Jason Hou-Liu, Hwanwoo Kim, Kecheng Li, Wenyuan Li, Xianwei Li, Yifan Sun, Zhaohan Sun, Mingwei Xu, Liyuan Zheng, and Ziqian Zhuang for every warm moment we shared together.

I am deeply indebted to Xiyue Han, Dingding Hu, Wei Liang, Jiyu Wang, Yiming Xu, Qi Yao, and Xianliang Zhao for their willingness to listen to my abstract nonsense and for enlightening me with research ideas.

Lastly, I would like to thank my family for their unconditional support throughout my life. In particular, I am deeply grateful to my wife, Ran. Despite the distance, she has always been there for me with her patience, encouragement, and unwavering love. Words cannot adequately express how much you all mean to me.

Dedication

For the path I have walked, and the world I have yet to reach.

行远自迩，念昔如初

Table of Contents

Examining Committee ii

Author's Declaration iv

Statement of Contributions v

Abstract vi

Acknowledgements viii

Dedication x

List of Figures xv

List of Tables xviii

1 Introduction 1

1.1 Functional Data Analysis 1

1.2 Functional Principal Component Analysis 4

1.3 Outline of the Thesis 7

2 Two-sample Inference for Sparse Functional Data **10**

2.1 Introduction . . . 10

2.2 Model and Estimation . . . 12

2.2.1 Problem Setup . . . 12

2.2.2 Estimation of Mean Functions . . . 14

2.3 Theoretical Properties . . . 18

2.3.1 General Assumptions . . . 18

2.3.2 Reproducing Kernel Hilbert Space . . . 19

2.3.3 Functional Bahadur Representation . . . 22

2.3.4 Pointwise Asymptotic Distribution . . . 24

2.3.5 Weak Convergence . . . 28

2.3.6 Bootstrap Calibration . . . 32

2.4 Simulation Studies . . . 32

2.4.1 Simulation Settings . . . 34

2.4.2 Estimation Performance . . . 35

2.4.3 Pointwise Inference Performance . . . 36

2.4.4 Global Inference Performance . . . 38

2.5 Real Data Examples . . . 41

2.5.1 Diffusion Tensor Imaging for Fractional Anisotropy Tract Profiles . 41

2.5.2 Beijing Air Pollutants Data . . . 44

2.A Theoretical Justification . . . 47

2.A.I Proofs for the Lemmas in Section 2.3 . . . 48

2.A.II Proof of Theorem 2.1 . . . 60

2.A.III Proof of Theorem 2.2 . . . 62

2.A.IV Proof of Theorem 2.4 . . . 66
2.A.V Proofs of Theorems in Section 2.3.6 . . . 76
2.A.VI Technical Lemmas . . . 79
2.B Additional Simulation Results . . . 90
2.B.I Out-of-sample Performance . . . 90
2.B.II Additional Global Testing Results . . . 92

3 Order Determination for Functional Data 99
3.1 Introduction . . . 99
3.2 Functional Principal Component Analysis . . . 102
3.2.1 Data Structure and Model Assumptions . . . 102
3.2.2 Estimation of the Model Components . . . 103
3.3 Functional Ladle Estimator . . . 104
3.4 Simulation Studies . . . 109
3.5 Real Data Examples . . . 112
3.5.1 Capital Bike-share Data . . . 113
3.5.2 Beijing Air Pollutants Data . . . 115
3.A Additional Details and Numerical Results . . . 118
3.A.I Estimated Eigenfunctions . . . 118
3.A.II Additional Simulation Results . . . 118
3.A.III Additional Real Data Analyses . . . 123

4 Learning Shared and Source-specific Subspaces across Multiple Data Sources for Functional Data 126
4.1 Introduction . . . 126

4.2 Estimation via Data Integration 129
4.2.1 Standard FPCA 129
4.2.2 Multi-source Model 131
4.2.3 Multi-source FPCA 134
4.3 Statistical Theory 136
4.4 Simulation 142
4.5 Real Data Illustration 146
4.A Proof of the Theorems in Section 4.3 149
4.A.I Proof of Theorem 4.1 149
4.A.II Proof of Theorem 4.2 153
4.B Proposition 4.1 and its Proof 156
4.C Technical Lemmas 157

5 Concluding Remarks and Future Work **160**

References **163**

List of Figures

2.1 In every panel, the red dotted-dash line represents the nominal coverage probability. The actual pointwise coverage probabilities for $N_{\max} = 10$ and $N_{\max} = 18$ are represented by black solid and blue long-dashed lines, respectively. The number of subjects in group 2, denoted by n_2, with values of 100, 200, and 400, arranged from the top to the bottom. 37

2.2 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.1. 39

2.3 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.2. 39

2.4 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.3. 40

2.5 The FA profiles along the CC tract for all participants are shown. The graph features a highlighted control subject represented by a red dot-dashed line and a case subject denoted by a blue dashed line. 42

2.6 The mean difference between cases and controls (black solid line), along with 95% pointwise confidence intervals (blue dot-dashed lines) over the tract of CC. 43

2.7 PM2.5 readings (weekly aggregate data) from 10 randomly selected weeks from summer (solid lines) and winter (dashed lines) in Huairou, Beijing. . . 44

2.8 Estimated mean differences (black solid line) and 95% confidence intervals (blue dashed lines) of hourly readings of $PM_{2.5}$ between winter and summer in Huairou. The vertical brown dash line indicates 14:00 at the local time. 46

2.9 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.1. 93

2.10 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.2. 93

2.11 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.3. 94

2.12 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.1. 95

2.13 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.2. 95

2.14 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.3. 96

2.15 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.1. 97

2.16 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.2. 97

2.17 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.3. 98

3.1 The boxplot for $\cos\theta := \langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle$ for $\nu = 1, 2, \ldots, 5$, over 1000 simulation runs, where the true rank is 2. 106

3.2 The leftmost plot presents $g(\ell)$, which illustrates the proportion of variation explained by each $\hat{\lambda}_\nu$. The middle panel shows $f(\ell)$, capturing the (normalized) variability between $\widehat{B}_{1,\ell}$ and $\widehat{B}_{2,\ell}$. The rightmost panel displays $h(\ell)$, which integrates information from both estimated eigenvalues and estimated eigenfunctions. In each panel, a vertical dotted line marks the true rank. . 108

3.3 The number of rentals over 24 hours for all days in the training set. 114

3.4 The left panel displays the first three estimated eigenfunctions derived from $\mathcal{D}_1$, while the right panel presents those from $\mathcal{D}_2$. Here the solid line represents $\hat{\phi}_1$, the dash line is $\hat{\phi}_2$, and the long dash line is $\hat{\phi}_3$, for both panels. . 119

3.5 The histogram of total rental time (in hours) 123

3.6 The estimated slope function, $\hat{\beta}(t)$, for the bike-sharing dataset, using the first four estimated eigenfunctions. 124

3.7 The plot shows the top three estimated eigenfunctions for the bike-sharing data, where the first eigenfunction is presented in solid line, the second eigenfunction is in dash line, and the third eigenfunction is the dotted line. 124

3.8 The 5th estimated eigenfunction (solid line) and the estimated 6th eigenfunction (dash line) for the bike-sharing data. 125

4.1 Data are generated from model (4.4) and Example 4.1. The solid line represents the true shared eigenfunction across two sites. The dashed line shows the estimated shared eigenfunction obtained using the proposed method in Section 4.2.3. The dotted line corresponds to the eigenfunction estimated from the averaged covariance operator. 127

4.2 Scree plot of $\widehat{\boldsymbol{P}}_w$ based on Example 4.1 with $m_1 = m_2 = 3$. Data are generated from two sources, each with 200 subjects and 25 observations per subject. 135

4.3 The left panel shows the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{winter}}$ for five cities in the winter months, and the right panel demonstrates the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{summer}}$ in the summer months. Both projected operators are calculated based on equation (4.7). 147

4.4 The left panel displays the leading two estimated eigenfunctions of $\widehat{\boldsymbol{P}}_{\text{winter}}$, and the right panel presents those of $\widehat{\boldsymbol{P}}_{\text{summer}}$. In both panels, the solid and dashed lines represent $\hat{\psi}_1$ and $\hat{\psi}_2$, respectively, obtained from the corresponding $\widehat{\boldsymbol{P}}_s$. 148

List of Tables

2.1 Mean and standard error (in parenthesis) of IMSEs across 1000 Monte Carlo runs with $\delta = 0.5$. 36

2.2 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ 40

2.3 Prediction errors and standard error (in parenthesis) of OOSPs across 1000 Monte Carlo runs for group 2 with $\mu(t) = (2t - 0.3)^3 + 0.5t$. 91

2.4 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ 92

2.5 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ 94

2.6 Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ 96

3.1 Comparison of order determination for Gaussian random processes under irregular design with $d = 3$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 111

3.2 Comparison of order determination for Gaussian random processes with $d = 3$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 112

3.3 The estimated rank of the bike sharing data with averaged prediction errors under 500 runs. The estimated rank is the mode of estimated ranks. 115

3.4 The estimated rank of the Beijing air pollutants data and the classification accuracy for different methods in winter workdays under 500 runs. The estimated rank is the mode of estimated ranks. 117

3.5 Comparison of order determination for Gaussian random processes under irregular design with $d = 6$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 119

3.6 Comparison of order determination for Gaussian random processes with $d = 6$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 120

3.7 Comparison of order determination for non-Gaussian random processes under irregular design with $d = 3$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. . 120

3.8 Comparison of order determination for non-Gaussian random processes with $d = 3$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 121

3.9 Comparison of order determination for non-Gaussian random processes under irregular design with $d = 6$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. . 121

3.10 Comparison of order determination for Gaussian random processes with $d = 6$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations. 122

4.1 Mean and standard error (in parenthesis) comparison of the norm differences for the *shared* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only. 144

4.2 Mean and standard error (in parenthesis) comparison of the norm differences for the *source-specific* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only. 145

4.3 Mean and standard error (in parenthesis) comparison of the norm differences for the *entire* space across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only. 145

Chapter 1

Introduction

Functional data analysis (FDA) has gained prominence in modern data analysis. The object under analysis is a continuous random function, often defined over time or space. The inherently infinite-dimensional nature of functional data distinguishes it from classical multivariate data defined in a finite-dimensional Euclidean space. A typical functional dataset consists of a set of realizations of a Hilbertian random element or real-valued random process. Comprehensive treatments of the subject are available in monographs by Ramsay and Silverman (2005) and Kokoszka and Reimherr (2017). This dissertation addresses several topics in FDA, including inference for mean differences, dimension reduction and data integration.

1.1 Functional Data Analysis

Let X be a random element of a separable Hilbert space $\mathcal{H}$, defined on a probability space $(\Omega, \mathcal{F}, \mathfrak{P})$, where $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the induced norm $\|\cdot\|_{\mathcal{H}}$. Suppose X is integrable, that is, $\mathbb{E}\|X\|_{\mathcal{H}} < \infty$. Then the mean element $\mathbb{E}X \in \mathcal{H}$ of X is defined by

$$\mathbb{E}\langle X, f \rangle_{\mathcal{H}} = \langle \mathbb{E}X, f \rangle_{\mathcal{H}}$$

for any $f \in \mathcal{H}$. In addition, if $\mathbb{E}\|X\|_{\mathcal{H}}^2 < \infty$, a condition commonly referred to as square-integrability, then we can define the covariance operator of X as

$$\boldsymbol{G} = \mathbb{E}(X - \mathbb{E}X) \otimes (X - \mathbb{E}X).$$

Here, $f \otimes g(h) = \langle f, h \rangle_{\mathcal{H}} g$ for any $f, g, h \in \mathcal{H}$ (Horváth and Kokoszka, 2012; Hsing and Eubank, 2015).

Given an i.i.d. sample $X_1, X_2, \ldots X_n$ of X, a canonical estimate of $\mathbb{E}X$ is the sample mean $\widehat{\mathbb{E}X} = n^{-1} \sum_{i=1}^n X_i$, and a canonical estimate of the covariance operator is the sample covariance $\widehat{\boldsymbol{G}} = n^{-1} \sum_{i=1}^n (X_i - \widehat{\mathbb{E}X}) \otimes (X_i - \widehat{\mathbb{E}X})$. The Hilbert-space central limit theorem ensures $\widehat{\mathbb{E}X}$ is a root-n consistent estimator of $\mathbb{E}X$ (Billingsley, 1968), and Dauxois et al. (1982) further established that $\widehat{\boldsymbol{G}}$ is root-n consistent for $\boldsymbol{G}$ under the Hilbert-Schmidt norm when $\mathbb{E}\|X\|_{\mathcal{H}}^4 < \infty$.

Beyond viewing X as an abstract random element in $\mathcal{H}$, in FDA, one typically assumes $X = \{X(t) \mid t \in \mathcal{T}\}$ is a square-integrable stochastic process defined over a compact set $\mathcal{T}$. Thus, it is natural to consider $\mathcal{L}^2(\mathcal{T})$, defined as

$$\mathcal{L}^2(\mathcal{T}) \equiv \left\{ f : \mathcal{T} \to \mathbb{R} \;\middle|\; \int_{\mathcal{I}} f^2(t)\, \mathrm{d}t < \infty \right\}.$$

The space $\mathcal{L}^2(\mathcal{T})$ is a separable Hilbert space when endowed with the inner product $\langle f, g \rangle_2 = \int_{t \in \mathcal{T}} f(t) g(t)\, \mathrm{d}t$, where the induced norm $\|f\|_2 = \sqrt{\langle f, f \rangle_2}$. Define $\mu(t) = \mathbb{E}X(t)$ and $G(t, s) = \mathbb{E}\{X(t) - \mu(t)\}\{X(s) - \mu(s)\}$ for any $t, s \in \mathcal{T}$. If X is mean-square continuous, then both $\mu(t)$ and $G(t, s)$ are continuous. Furthermore, assuming X is jointly measurable, it follows that X is also a random element in $\mathcal{L}^2(\mathcal{T})$ (Hsing and Eubank, 2015). Under this setting, $\mu \in \mathcal{H}$ is referred to as the mean function, coinciding with the mean element $\mathbb{E}X$ of X. Likewise, the integral operator induced by $G(t, s)$ coincides with the covariance operator of X; the kernel $G(t, s)$ is therefore termed the covariance function. More precisely,

$$\boldsymbol{G}(f)(\cdot) = \int_{s \in \mathcal{T}} G(\cdot, s) f(s)\, \mathrm{d}s \tag{1.1}$$

for any $f \in \mathcal{H}$. Hence, the previous results for the Hilbert-space-valued random sample

apply directly to stochastic processes X and their sample paths $X_1, \ldots, X_n$. For example, the sample mean function $\hat{\mu}(\cdot) = n^{-1}\sum_{i=1}^{n} X_i(\cdot)$ is root-n consistent for $\mu(\cdot)$. In addition to the covariance operator, the covariance function can be estimated by $\widehat{G}(t,s) = n^{-1}\sum_{i=1}^{n}\{X_i(t) - \hat{\mu}(t)\}\{X_i(s) - \hat{\mu}(s)\}$. Throughout this dissertation, we assume X is a square-integrable stochastic process defined on $\mathcal{T} = [0,1]$ with continuous mean and covariance functions, and that X is jointly measurable.

The above asymptotic results require the entire trajectory of each X_i is *fully observed*. In practice, however, functional data are often *partially observed*, meaning that only discrete measurements—possibly contaminated by random errors—are available for each sample path. When each X_i has N_i observed points, the following model is commonly adopted:

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \text{ for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, N_i. \tag{1.2}$$

Here, Y_{ij} denotes the jth observation of X_i taken at T_{ij}, and ε_{ij} represents the associated random error. This model is widely used in modelling longitudinal observations under the framework of functional data; see Yao et al. (2005b), Li and Hsing (2010), Zhang and Wang (2016), Zhou et al. (2024) and references therein.

To estimate the mean and covariance functions of X, the choice of estimation method and the corresponding optimal convergence rate depend on the distributions of the T_{ij}'s, and on the magnitude of $\bar{N}$ relative to n, where $\bar{N} = n^{-1}\sum_{i=1}^{n} N_i$. There are generally two settings for how the observation times are distributed. One is referred to as *common design*, in which the observations are sampled at the same locations across all curves. The other is referred to as *random design*, where T_{ij} are independently sampled from $\mathcal{T}$. Cai and Yuan (2011) illustrated the phase transition phenomenon in estimating mean function as $\bar{N}$ changes relative to n for both designs. Zhang and Wang (2016) further categorized functional data under random design into three types: "sparse", "dense", and "ultra-dense". Specifically, for sparse functional data—where each subject is observed at only a few irregular time points—the estimators of the mean and covariance functions generally cannot achieve the root-n rate of convergence. A typical example of sparse functional data arises in longitudinal studies, where only a limited number of irregularly spaced observations per subject are available. In contrast, for dense and ultra-dense functional

data, where the number of observations per subject increases much faster relative to the sample size, the estimators of the mean and covariance functions can attain the root-n rate. Moreover, in the ultra-dense regime, the bias induced by smoothing becomes asymptotically negligible.

For dense or ultra-dense functional data, a two-step procedure is commonly employed in estimation. Individual trajectories are first reconstructed through a pre-smoothing step, which applies nonparametric smoothing methods such as smoothing splines (Rice and Silverman, 1991), local polynomial regression (Zhang and Chen, 2007), and B-splines (Cardot, 2000). Subsequent analyses, such as mean estimation, are then performed using the estimated sample paths.

For sparse functional data, however, the pre-smoothing approach becomes inappropriate because of the limited number of observations per subject, which leads to insufficient information for accurate trajectory reconstruction. In such sparse settings, Yao et al. (2005b) suggested aggregating data across all subjects to estimate the mean and covariance functions using local linear regression. This approach has been proven to ensure consistency and uniform convergence rates for estimating both the mean and covariance functions. Cai and Yuan (2011) established the minimax rate for the mean function estimate via smoothing splines under both common and random designs. Furthermore, Zhang and Wang (2016) established convergence rates and pointwise limiting distributions for the mean and covariance estimators using local linear regression, considering various weighting schemes.

1.2 Functional Principal Component Analysis

Besides from estimating mean and covariance functions, dimension reduction plays a pivotal role in FDA. It transforms infinite-dimensional random functions into finite-dimensional random vectors and thereby facilitates subsequent analyses. Moreover, dimension reduction is commonly adopted as a way to regularize the covariance operator, which is required in functional linear regression (Hall and Horowitz, 2007; Zhou et al., 2023) and functional generalized linear models (Dou et al., 2012). Among these techniques, functional principal

component analysis (FPCA) is one of the most widely used. Analogous to principal component analysis for multivariate data, FPCA aims to extract a parsimonious subspace that captures the dominant modes of variation around the mean function in a data-adaptive manner.

Let $\mathcal{H} = \mathcal{L}^2(\mathcal{T})$ denote the space of square-integrable functions on $\mathcal{T}$, equipped with inner product $\langle f, g \rangle_2 = \int_{t \in \mathcal{T}} f(t)g(t)\,\mathrm{d}t$. Suppose X is a continuous and square-integrable stochastic process defined on $\mathcal{T}$. For a fixed m, FPCA seeks an m-dimensional subspace $\mathcal{H}_0 \subset \mathcal{H}$ such that the projection of $\widetilde{X} = X - \mu$ onto $\mathcal{H}_0$ provides the optimal mean-squared approximation to $\widetilde{X}$. More precisely, FPCA identifies m orthonormal functions $e_1, e_2, \ldots, e_m \in \mathcal{H}$ that minimize the expected mean-squared distance between $\widetilde{X}$ and its m-dimensional projection, which is formalized as follows:

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P} \in \mathcal{P}_m}{\arg\min}\, \mathbb{E}\|\widetilde{X} - \boldsymbol{P}\widetilde{X}\|_2^2. \tag{1.3}$$

Here, $\mathcal{P}_m = \{\boldsymbol{P} \in \mathcal{B}(\mathcal{H}) \mid \boldsymbol{P}^2 = \boldsymbol{P}, \mathrm{rank}(\boldsymbol{P}) = m, \boldsymbol{P} \text{ is self-adjoint}\}$, and $\mathcal{B}(\mathcal{H})$ denotes the set of bounded linear operators on $\mathcal{H}$. An operator $\boldsymbol{A}$ is self-adjoint if $\langle \boldsymbol{A}f, g \rangle_2 = \langle f, \boldsymbol{A}g \rangle_2$ for any $f, g \in \mathcal{H}$. The rank of $\boldsymbol{A}$ is the dimension of its image space. The definition of self-adjoint and rank can be treated as the generalization of symmetric and rank for a matrix. In addition, any $\boldsymbol{P} \in \mathcal{P}_m$ can be written as $\boldsymbol{P} = \sum_{\nu=1}^{m} e_\nu \otimes e_\nu$.

For two operators $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{B}(\mathcal{H})$, define the Hilbert–Schmidt inner product as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{HS}} = \sum_\nu \langle \boldsymbol{A}(e_\nu), \boldsymbol{B}(e_\nu) \rangle_2$, where $\{e_\nu : \nu = 1, 2, \ldots\}$ is any orthonormal basis of $\mathcal{H}$. This definition does not depend on the choice of basis and induces the Hilbert–Schmidt norm $\|\boldsymbol{A}\|_{\mathrm{HS}}^2 = \langle \boldsymbol{A}, \boldsymbol{A} \rangle_{\mathrm{HS}}$. With this notation, the optimization problem in (1.3) can be equivalently written as

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P} \in \mathcal{P}_m}{\arg\min}\, \mathbb{E}\|\widetilde{X} - \boldsymbol{P}\widetilde{X}\|_2^2 - \mathbb{E}\|\widetilde{X}\|_2^2 = \underset{\boldsymbol{P} \in \mathcal{P}_m}{\arg\max}\, \langle \boldsymbol{G}, \boldsymbol{P} \rangle_{\mathrm{HS}}\,, \tag{1.4}$$

where $\boldsymbol{G}$ is the covariance operator given in equation (1.1). As the covariance function is

continuous and positive semi-definite, Mercer's theorem implies that

$$G(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s)\phi_\nu(t), \quad t, s \in \mathcal{T}, \tag{1.5}$$

where (λ_ν, ϕ_ν) is the νth eigenvalue-eigenfunction pair of $\boldsymbol{G}$ satisfying $\boldsymbol{G}\phi_\nu = \lambda_\nu\phi_\nu$, and $\{\phi_\nu\}_{\nu\geq 1}$ forms an orthonormal basis of $\mathcal{H}$. Moreover, $\boldsymbol{G}$ is bounded, compact, and self-adjoint. The compactness of $\boldsymbol{G}$ implies $\lim_{\nu\to\infty} \lambda_\nu = 0$. Throughout this dissertation, we further assume $\lambda_1 > \lambda_2 > \ldots > 0$ for brevity. Combining equations (1.4) and (1.5), it follows immediately that $e_i = \phi_i$ for $i = 1, \ldots, m$.

In addition, the Karhunen–Loève expansion yields that

$$X(t) = \mu(t) + \sum_{\nu=1}^{\infty} \xi_\nu \phi_\nu(t), \tag{1.6}$$

where $\xi_\nu = \langle X - \mu, \phi_\nu \rangle_2$ are uncorrelated, mean-zero random variables with variance λ_ν. These ξ_ν's are the *functional principal component scores*, often simply termed *scores*. The eigenfunctions of $\boldsymbol{G}$ are also referred to as *functional principal components* (FPCs). In addition, X can be approximated using the leading m eigenfunctions of $\boldsymbol{G}$:

$$X^m(t) = \mu(t) + \widetilde{\boldsymbol{P}}\widetilde{X} = \mu(t) + \sum_{\nu=1}^{m} \xi_\nu \phi_\nu(t) \tag{1.7}$$

with $\widetilde{\boldsymbol{P}} = \sum_{\nu=1}^{m} \phi_\nu \otimes \phi_\nu$. Once an estimate $\widehat{G}$ of the covariance function is obtained, the eigenpairs can be computed by performing eigen-decomposition of the discretization of $\widehat{G}$; see, for example, Chapter 8 of Ramsay and Silverman (2005).

Although the basis functions in equation (1.6) can be replaced by alternative function bases such as B-splines, this particular expansion is optimal in capturing the maximal variance of X in the $\mathcal{L}^2$ sense, as implied by equation (1.3). This optimality has prompted extensive research into the properties of estimated eigenpairs. For instance, Hall and Hosseini-Nasab (2006) presented the asymptotic expansions of estimated eigenpairs for fully observed functional data. For any fixed m that is independent of the sample size, Hall and

Horowitz (2007) developed the optimal $\mathcal{L}^2$ convergence rate of estimated eigenfunctions $\hat{\phi}_\nu$ for sparse functional data. Further, when m is allowed to diverge with respect to n, Zhou et al. (2023) established error bounds for $\hat{\phi}_\nu$ that yield consistency and the convergence rate of the estimated slope function in a functional linear regression model.

1.3 Outline of the Thesis

This dissertation concentrates on examining functional data from three perspectives. First, it addresses inference for mean differences between two groups of sparse functional data. Second, it investigates the selection of the number of FPCs, that is, determining m in equation (1.7). Lastly, it develops a procedure for extracting shared and source-specific subspaces across multiple functional data sources. To streamline the presentation of this dissertation, the literature review relevant to each research topic will be included in its respective chapter.

In Chapter 2, we develop a testing procedure for comparing mean functions between two groups of sparse functional data within the reproducing-kernel Hilbert-space (RKHS). Conventional approaches, which are built upon FPCA, usually assume a homogeneous covariance structure across groups. This assumption implies that both groups share the same eigenbasis, allowing the sample paths to be projected onto common directions; see equation (1.6). Nonetheless, justifying this assumption in real-world scenarios can be challenging. To eliminate the need for a homogeneous covariance structure, we consider a pth order Sobolev space with higher regularity condition than $\mathcal{L}^2$. With a suitably defined inner product, the space becomes an RKHS, and we are able to develop the *functional Bahadur representation* (FBR) for a mean estimator obtained by smoothing splines. This representation is essentially a first-order approximation for the mean estimator, directly resulting in a pointwise limiting distribution. Moreover, we establish weak convergence of the mean estimator, which enables the construction of a test statistic for the mean difference. We further develop a multiplier bootstrap procedure with theoretical justification to calibrate the inference for the mean difference. Finally, the finite-sample performance of the proposed test is demonstrated through extensive simulations and two real-world

applications.

In Chapter 3, our objective is to determine the number of eigenpairs to retain, a problem known as order determination. The number of retained components m controls the bias-variance trade-off in approximating X. Higher-order FPCs often exhibit significant variability, making their interpretation difficult. This leads to the pragmatic assumption of treating higher-order terms as noise, effectively assuming that the covariance operator has finite rank. Under this scenario, correctly identifying the true rank can improve the performance of the subsequent analysis, such as functional linear regression or functional classification. In essence, the finite rank assumption transforms the order determination problem into estimating the rank of the covariance operator. When functional data are discretely observed and contaminated by measurement errors, the contamination creates a confounding issue: the true rank of the covariance operator is obscured by the presence of noise, which effectively adds a ridge to the covariance function. To address this challenge, we propose a novel rank-estimation procedure that jointly utilizes information from both the estimated eigenvalues and eigenfunctions. By incorporating the nonparametric nature of functional data through smoothing techniques, the method is applicable to functional data collected at random, subject-specific points. The numerical studies showcase excellent performance of our method under various simulation settings. The method is further illustrated by two real-world examples.

In Chapter 4, we investigate the extraction of shared and source-specific subspaces when multiple functional data sources are available. The heterogeneity across sources—often arising from site-specific data collection protocols—poses substantial challenges for analyses. A naive approach that averages the site-wise sample covariance functions and performs eigen-decomposition on the resulting average implicitly assumes homogeneity across sources. However, dominant variation patterns unique to specific sources can severely bias the identification of the shared variation structure. To mitigate this issue, we propose to focus exclusively on the eigenfunctions for subspace recovery. Specifically, during data integration, we discard the eigenvalues and average the projection operators generated by estimated eigenfunctions from each source. The eigenvalues of the resulting averaged projection operator quantify the frequency with which each eigenfunction recurs across sources, effectively down-weighting those associated with dominant but unshared subspaces. If the

source-specific subspace is of interest, we then re-project the estimated covariance operator of that source onto the orthogonal complement of the shared subspace. In addition, we provide theoretical justification for both the shared and source-specific estimators. The convergence rate of the estimator of the shared projection operator highlights the gain in estimation efficiency achieved by effectively leveraging information from multiple data sources. Extensive numerical studies confirm accurate recovery of the shared space and substantial efficiency gains achieved through data integration. The method is further illustrated using a real-world multi-site dataset.

Chapter 5 summarizes the insights and methodologies presented in Chapters 2 - 4, outlines potential avenues for future research, and provides concluding remarks on the theoretical and practical implications of this dissertation.

Chapter 2

Two-sample Inference for Sparse Functional Data

2.1 Introduction

This chapter focuses on statistical inference for two-sample mean functions. Well-established methods are available in the context of dense functional data. Under the assumption that the number of observations grows faster than the number of subjects, the approximation error incurred by the pre-processing becomes negligible in subsequent analysis (Zhang and Chen, 2007). This is the primary reason why various statistical inference approaches exist for two-sample mean functions in the context of dense functional data. For example, the methods proposed by Zhang et al. (2010) and Zhang and Liang (2014), which were initially intended for fully observed data, suggest that utilizing pre-smoothed individual trajectories in test statistics is asymptotically equivalent to using fully observed data.

However, conducting inference for two-sample mean functions in the context of sparse functional data remains relatively unexplored. In cases with only a few irregularly spaced observations per subject, recovering trajectories from such sparse data can result in significant estimation errors, which cannot be overlooked in subsequent analysis. Consequently, developing a statistically sound procedure for mean function inference becomes more chal-

lenging. Existing methods, to our knowledge, either consider a polynomial-type null hypothesis (Staicu et al., 2014) or assume a homogeneous covariance structure for the two groups (Pomann et al., 2016; Wang, 2021).

Testing if a mean function is a polynomial can be overly restrictive when prior knowledge about the mean function is lacking. Additionally, determining the polynomial order can be challenging. On the other hand, the homogeneous covariance assumption allows us to project individual trajectories onto basis functions derived from pooled data. These basis functions are eigenfunctions of the common covariance function, known as functional principal components in the FDA literature. Without assuming a common covariance structure, trajectories from different groups are projected onto different directions, making comparisons of mean functions based on projections misleading. However, verifying this assumption can be quite challenging in itself. For example, Pigoli et al. (2014) proposed an infinite-dimensional version of the Procrustes size-and-shape distance for covariance operators. However, the theoretical properties of this test statistic are not yet fully established. As a result, the calculation of p-values using this statistic depends on re-sampling techniques, such as permutations.

In this chapter, we propose a novel procedure for testing the mean difference based on sparse functional data within the RKHS framework. The proposed method does not entail a homogeneous covariance for the two groups of functional data. We begin by aggregating data within each group, enabling us to borrow information from data across all subjects to obtain a group-wise mean function estimator via smoothing splines. We then develop a *functional Bahadur representation* for this estimator, which is crucial for establishing the asymptotic properties of the mean estimator. This representation was first developed by Shang and Cheng (2013), and we adopt the same term throughout the chapter. Specifically, it allows us to represent the estimation error as a summation of i.i.d. random elements plus higher-order terms that are negligible in the asymptotic analysis. With the aid of this representation, we establish pointwise asymptotic distributions and weak convergence for the mean estimator within each group. Under the independence assumption of these two groups, we subsequently develop pointwise and global tests for the mean difference between the two groups. To evaluate finite-sample performance of our proposed method, we conduct extensive simulation studies to investigate its performance in estimating the

mean function and testing the mean difference between the two groups.

The remainder of this chapter is organized as follows. In Section 2.2, we outline the construction of a mean function estimator via smoothing splines for sparse functional data. In Section 2.3, we develop asymptotic results for the mean estimator, and design a bootstrap-based algorithm to implement the testing procedure with statistical guarantees. Detailed proofs are relegated to the Appendix. We showcase finite-sample performance of the proposed method under various settings in Section 2.4. Finally, we consider two real applications of the proposed method in Section 2.5: diffusion tensor imaging data and Beijing air quality data.

2.2 Model and Estimation

In Section 2.2.1, we introduce our model, formalize the hypothesis of detecting mean differences in two groups of functional data, and review existing methods. Subsequently, in Section 2.2.2, we propose a smoothing-spline-based estimator for the mean function within each group under the RKHS framework, leading to the immediate availability of a mean difference estimator.

2.2.1 Problem Setup

In this chapter, we consider a two-sample problem involving data generated from the variant of model (1.2) as follows:

$$Y_{gij} = X_{gi}(T_{gij}) + \varepsilon_{gij}, \quad T_{gij} \in \mathcal{T} = [0,1], \tag{2.1}$$

where $g = 1, 2$, $i = 1, 2, \ldots, n_g$ and $j = 1, 2, \ldots, N_{gi}$. Here g is the group index.

(i) Let Y_{gij} denote the jth noisy observation of the ith random function X_{gi};

(ii) Let T_{gij}'s denote i.i.d. observation times with a common group-wise probability density defined over $\mathcal{T}$;

(iii) Let ε_{gij}'s denote i.i.d. random errors with mean zero and variance $\sigma^2_{g,\varepsilon}$;

(iv) Let N_{gi}'s denote numbers of observations, assumed to be i.i.d. within group g.

We further assume that every ε_{gij} is independent of the random process X_g and the observation time T_{gij}; N_{gi}'s are independent of all other random objects in the model. Notice that N_{gi}'s characterize sparseness of the observations in group g. This model structure has been widely adopted in the FDA literature, including Yao et al. (2005b), Li and Hsing (2010), Zhang and Wang (2016), Wang (2021) and references therein.

For $g = 1, 2$, we assume $\mathbb{E}\{X_g(t)\} = \mu_g(t)$ for all $t \in \mathcal{T}$. In this chapter, we focus on two types of hypothesis testing problems. The first is commonly referred to as pointwise testing, where the null hypothesis is defined for every specific time point $t^* \in \mathcal{T}$:

$$H_0 : \mu_1(t^*) = \mu_2(t^*), \quad \text{VS} \quad H_1 : \mu_1(t^*) \neq \mu_2(t^*). \tag{2.2}$$

The second type is global testing, which formulates the testing problem for the entire domain $\mathcal{T}$:

$$H_0 : \mu_1(t) = \mu_2(t), \ \forall t \in \mathcal{T} \quad \text{VS} \quad H_1 : \mu_1(t) \neq \mu_2(t), \text{ for some } t \in \mathcal{T}. \tag{2.3}$$

The hypothesis testing problem described in (2.2) is focused on testing mean differences at specific time points. However, the global test may not align with the pointwise test even for all $t \in \mathcal{T}$, at the same significance level. Furthermore, in contrast to (2.2), the testing problem in (2.3) is more complex due to the infinite-dimensional nature of the parameter of interest.

Addressing this problem with a finite number of observations entails dimension reduction techniques. Functional principal components analysis (FPCA) plays a pivotal role in this regard, which is introduced in Chapter 1.2. Recall that, in classical FPCA, the covariance function $\mathbb{C}\text{ov}(X_{gi}(t), X_{gi}(s)) \triangleq c_g(t, s)$ can be orthogonally expanded using a sequence of non-increasing and non-negative eigenvalues θ_{gk} and their corresponding eigenfunctions ϕ_{gk}: $c_g(t, s) = \sum_k \theta_{gk}\phi_{gk}(t)\phi_{gk}(s)$. Additionally, the Karhunen–Loéve expansion given in

equation (1.6) for each group is as follows:

$$X_{gi}(t) = \mu_g(t) + \sum_{k=1}^{\infty} \xi_{gik}\phi_{gk}(t), \tag{2.4}$$

where $\xi_{gik} = \int_{\mathcal{T}}\{X_{gi}(t) - \mu_g(t)\}\phi_{gk}(t)\,\mathrm{d}t$ is the kth fPC score satisfying:

(i) $\mathbb{E}\xi_{gik} = 0$;

(ii) $\mathbb{C}\mathrm{ov}(\xi_{gik_1}, \xi_{gik_2}) = \theta_{gk}\delta_{k_1k_2}$, which equals θ_{gk} if $k_1 = k_2$, and 0 otherwise.

Here $k, k_1, k_2 \geq 1$, $g = 1, 2$, and $i = 1, \ldots, n_g$.

Assuming $c_1(t, s) = c_2(t, s)$, the eigenfunctions for both groups are identical. Under the null hypothesis in (2.3), the difference between the groups of random functions arises from the randomness of FPC scores as described in (2.4). In practice, we usually consider a truncated version of (2.4), by retaining only the first p eigenfunctions estimated from Y_{gij}'s within each group. Interested readers can refer to Ramsay and Silverman (2005) and Yao et al. (2005b) for further details. Here p can be determined via cross-validation or criteria such as the Bayesian information (Yao et al., 2005b). Such a truncation converts the infinite-dimensional inference problem to a finite-dimensional one. Following this strategy, Wang (2021) constructed a χ^2_p test statistic by assuming joint normality of scores and random errors. To compare the distributions of these groups of functional data, Pomann et al. (2016) proposed the Anderson–Darling test for comparing the distribution of the two groups of FPC scores. It should be noted that the homogeneous covariance assumption is vital for FPCA-based methods to perform dimension reduction, although verifying it in real-world scenarios can be challenging.

2.2.2 Estimation of Mean Functions

In the FDA literature, it is common to assume that discrete observations are taken from sample paths that are realizations of a random process, typically with certain smoothness

properties (Wang et al., 2016). In this chapter, we assume that almost surely the sample path of X_g for $g = 1, 2$ resides in an mth order Sobolev space, which is defined as

$$\mathcal{H} := H^m(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R} \mid f, f^{(1)}, f^{(2)}, \ldots, f^{(m-1)} \text{ are absolutely continuous and } f^{(m)} \in \mathcal{L}^2(\mathcal{T})\},$$

where $f^{(v)}$ denotes the vth order derivative of f and

$$\mathcal{L}^2(\mathcal{T}) = \left\{ f : \mathcal{T} \to \mathbb{R} \;\middle|\; \int_{\mathcal{T}} f^2(t)\, \mathrm{d}t < \infty \right\},$$

denotes the collection of square integrable functions on $\mathcal{T}$. This assumption also appears in Cai and Yuan (2011). Furthermore, as demonstrated in the same reference, the mean function μ_g belongs to $\mathcal{H}$, as indicated by the following inequality:

$$\int_{\mathcal{T}} \{\mu_g^{(m)}(t)\}^2 \, \mathrm{d}t \leq \mathbb{E}\left[\int_{\mathcal{T}} \{X_g^{(m)}(t)\}^2 \, \mathrm{d}t \right].$$

Choosing an appropriate mean function estimator hinges on the subject sampling frequency within the same group. For fully observed data $X_{g1}, \ldots, X_{gn_g}$, the sample mean is a natural choice. However, in practice, observations are often collected discretely over fixed or random grids. Estimating the mean function for dense functional data commonly involves a two-step method. Firstly, a pre-smoothing procedure is employed to reconstruct individual trajectories denoted as $\tilde{X}_{gi}$ via non-parametric techniques like smoothing splines (Rice and Silverman, 1991), local polynomial regression (Zhang and Chen, 2007), and B-splines (Cardot, 2000). The subsequent analysis then utilizes $\tilde{X}_{gi}$ in place of X_{gi} in the sample mean estimator.

The pre-smoothing approach becomes inappropriate with limited observations per subject due to insufficient information for accurate trajectory recovery. In sparse settings, Yao et al. (2005b) suggested aggregating data across all subjects to estimate the mean function through local linear regression. They established consistency and uniform convergence rates for both mean and covariance estimators. Furthermore, Zhang and Wang (2016) provided convergence rates and pointwise limiting distributions for the mean estimator us-

ing local linear regression under various weighting schemes. Additionally, they presented a systematic framework for differentiating dense functional data from sparse functional data, which is based on the ratio of the average number of observations per subject to the total number of subjects.

Inspired by the aggregation idea proposed in Yao et al. (2005b), we introduce a novel mean function estimator within the RKHS framework. Specifically, we compute the smoothing splines estimate for μ_g in each group g by minimizing a loss function $\ell(f, \lambda_g)$ defined as follows:

$$\begin{aligned}\hat{\mu}_{\lambda_g} &= \arg\min_{f \in \mathcal{H}} \ell(f, \lambda_g) \\ &:= \arg\min_{f \in \mathcal{H}} \left[\frac{1}{2M_g} \sum_{i=1}^{n_g} \sum_{j=1}^{N_{gi}} \left\{ Y_{gij} - f(T_{gij}) \right\}^2 + \frac{\lambda_g}{2} J(f, f) \right], \end{aligned} \tag{2.5}$$

where $M_g = \sum_{i=1}^{n_g} N_{gi}$ denotes the total number of observations across n_g subjects in group g, $J(f, f) = \int_{\mathcal{T}} \{f^{(m)}(t)\}^2 \, \mathrm{d}t$ is the roughness penalty, and $\lambda_g > 0$ denotes the smoothing parameter controlling the trade-off between fidelity to the data and roughness of the estimated mean function. It is important to point out that we assign the same weight to each observation in (2.5), while Cai and Yuan (2011) assigned the same weight to each subject. They showed that their proposed mean estimator attained the minimax convergence rate, but did not develop any inferential tools for the mean function.

When $\mathcal{H}$ is equipped with the following (squared) norm

$$\|f\|_R^2 = \sum_{\nu=0}^{m-1} \left\{ \int_{\mathcal{T}} f^{(\nu)}(t) \, \mathrm{d}t \right\}^2 + \int_{\mathcal{T}} \left\{ f^{(m)}(t) \right\}^2 \, \mathrm{d}t, \tag{2.6}$$

$\mathcal{H}$ becomes an RKHS (Gu, 2013, Chapter 2.3). By the well-known representer theorem for an RKHS (Wahba, 1990), the mean function estimator, i.e., the solution to the optimization problem (2.5), has a finite-dimensional representation. This result is summarized in the following proposition. Let $\mathcal{H}_0$ denote the null space of J, that is, $\mathcal{H}_0 = \{f \in \mathcal{H} \mid J(f, f) = 0\}$, and $\mathcal{H}_1$ the orthogonal complement of $\mathcal{H}_0$ in $\mathcal{H}$. Suppose the basis functions of $\mathcal{H}_0$ are $\{\phi_1, \phi_2, \ldots, \phi_m\}$ and $R(s, t)$ denotes the reproducing kernel of $\mathcal{H}_1$ associated with the

norm defined in (2.6).

Proposition 2.1. *The solution to the minimization problem* (2.5) *can be expressed as*

$$\hat{\mu}_{\lambda_g}(t) = \sum_{k=1}^{m} d_{gk}\phi_k(t) + \sum_{i=1}^{n}\sum_{j=1}^{N_{gi}} c_{gij}R(t_{gij}, t), \tag{2.7}$$

for some coefficients $d_{g1}, d_{g2}, \ldots, d_{gk}$ *and* $c_{g11}, c_{g12}, \ldots, c_{gn_gN_{gn}}$.

A suitable choice for ϕ_k and $R(\cdot,\cdot)$ is $\phi_k = t^{k-1}, k = 1, 2, \ldots, m$ and

$$R(s,t) = \frac{1}{(m!)^2}B_m(t)B_m(s) - \frac{1}{(2m)!}B_{2m}(|t-s|),$$

respectively. Here B_l stands for the lth Bernoulli polynomials. For further details, please refer to Chapter 2.3 of Gu (2013).

The choice of λ_g has a great impact on the performance of mean function estimation and inference for the mean function difference. A commonly used data-driven method for choosing λ is through *generalized cross-validation* (GCV). Since it can be shown that c_{gij} and d_{gk} linearly depend on $\boldsymbol{y}_g$ in (2.7), we express $\hat{\boldsymbol{\mu}}_{\lambda_g} = S(\lambda_g)\boldsymbol{y}_g$, where

$$\hat{\boldsymbol{\mu}}_{\lambda_g} = (\hat{\mu}_{\lambda_g}(t_{g11}), \hat{\mu}_{\lambda_g}(t_{g12}), \ldots, \hat{\mu}_{\lambda_g}(t_{gn_gN_{n_g}}))',$$

and

$$\boldsymbol{y}_g = (y_{g11}, y_{g12}, \ldots, y_{gn_gN_g})'.$$

Here $S(\lambda_g)$ is the smoothing matrix. To determine λ_g, we minimize the GCV score defined as

$$\mathrm{GCV}(\lambda_g) = \frac{n^{-1}\|\hat{\boldsymbol{\mu}}_{\lambda_g} - \boldsymbol{y}_g\|_2^2}{[n^{-1}\mathrm{tr}\{I - S(\lambda_g)\}]^2}, \tag{2.8}$$

where $\|\cdot\|_2$ denotes the vector Euclidean norm. The choice of λ_g is crucial for this procedure, and more details can be found in Craven and Wahba (1978) and Chapter 3.2 of Gu (2013).

2.3 Theoretical Properties

Before presenting the main results, we establish necessary notations and assumptions in Section 2.3.1 that define the data-generating process for sparse functional data and ensure estimation consistency. Section 2.3.2 introduces a new inner product that makes $\mathcal{H}$ an RKHS with desirable properties. With the aid of these properties, we establish the functional Bahadur representation (FBR) in Section 2.3.3. This representation leads to the pointwise limiting distributions (Section 2.3.4) and weak convergence (Section 2.3.5) for the mean estimator of each group; they are then extended to the estimated mean difference. Subscripts are omitted when assumptions and properties apply equally to both groups.

2.3.1 General Assumptions

Assumption 2.1. *For $i = 1, 2, \ldots, n, j = 1, 2, \ldots, N_i$, (Y_{ij}, T_{ij}) generated by model (2.1), are copies of (Y, T) with a bounded joint density function $f(y, t)$. The marginal density function of T, denoted as $\pi(t)$, is bounded away from both infinity and zero.*

Assumption 2.2. $\mathbb{E}(\varepsilon) = 0$ *and* $\mathbb{V}\mathrm{ar}(\varepsilon) = \sigma_\varepsilon^2 < \infty$. *Furthermore, $\varepsilon_{ij}, j = 1, \ldots, N_i, i = 1, \ldots, n$ are i.i.d. copies of ε and are independent of T_{ij}.*

Assumption 2.3. *The number of observations for the ith subject, $N_i, i = 1, \ldots, n$, are i.i.d. copies of a positive integer-valued random variable N satisfying $\mathbb{E}(N) = \mu_N < \infty$ and $\mathbb{V}\mathrm{ar}(N) = \sigma_N^2 < \infty$. Furthermore, $\mathbb{P}(N > 1) > 0$. Additionally, for $i = 1, \ldots, n$, N_i is independent of T_{ij} and ε_{ij}.*

Assumption 2.4. *The sample path of the stochastic process X in model (2.1) has an mth-order derivative, which is denoted by $X^{(m)}$, almost surely. Moreover, $X^{(m)} \in \mathcal{L}_2(\mathcal{T})$, i.e., $\mathbb{E}[\int_{\mathcal{T}} \{X_g^{(m)}(t)\}^2 dt] < \infty$.*

These assumptions are common in the FDA literature. For example, Assumptions 2.1-2.3 were adopted by Yao et al. (2005b) for FPCA in the context of sparse functional data. Assumption 2.1 ensures that observation Y_{ij}'s are randomly and evenly spread on the domain $\mathcal{T}$ for each subject, facilitating reasonable estimators of the mean function

when observations are aggregated. Assumption 2 ensures that the second moment of the error term is finite. The assumption of independent errors within the same subject can be relaxed. For instance, one could assume $\varepsilon(t)$ is a square-integrable and weakly stationary process. In such a case, an additional term accounting for the within-subject variability would appear in the variance expressions provided in Theorems 2.2 and 2.4. Assumption 2.3 characterizes the sparseness of the observations. To see this, let $M = \sum_{i=1}^{n} N_i$. Then we have

$$\mathbb{E}(M) = \mathbb{E}\left(\sum_{i=1}^{n} N_i\right) = n\mu_N = O(n) \text{ as } \mu_N < \infty.$$

Further, we can show that $M = O_p(n)$ by the weak law of large numbers. This indicates that the total number of observations per group scales with the sample size, meaning the number of observations per subject does not increase with respect to the sample size. Assumption 2.4 is introduced for theoretical convenience, which ensures continuity and boundedness of sample paths and the covariance function in the compact domain. This assumption with $m = 2$ was considered in Zhu et al. (2014) and Kong et al. (2016). Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iN_i})'$ and $\boldsymbol{T}_i = (T_{i1}, T_{i2}, \ldots, T_{iN_i})'$. Assumptions 2.2 and 2.3 indicate $\{(\boldsymbol{Y}_i, \boldsymbol{T}_i, N_i)\}_{i=1}^{n}$ are i.i.d in model (2.1). The independence of subjects allows for the application of the central limit theorem, to establish asymptotic results for the mean estimator.

2.3.2 Reproducing Kernel Hilbert Space

To test hypotheses in (2.2) and (2.3), we develop the FBR for a first-order approximation to the mean estimator. The FBR entails a properly defined RKHS, where the mean estimator resides. Following Shang and Cheng (2013), for $m > 1/2$, we consider $\mathcal{H} = H^m(\mathcal{T})$ as an RKHS, endowed with the inner product:

$$\langle f, g \rangle = \mathbb{E}\left\{f(T)g(T)\right\} + \lambda J(f, g). \tag{2.9}$$

Henceforth, $\|f\|$ represents the norm induced by (2.9) for any $f \in \mathcal{H}$, and K stands for the corresponding reproducing kernel of $\mathcal{H}$. Consequently, K, serving as a function from

$\mathcal{T} \times \mathcal{T}$ to $\mathbb{R}$, satisfies the following:

$$\langle K_t, f \rangle = f(t) \text{ for any } f \in \mathcal{H} \text{ and } t \in \mathcal{T},$$

where $K_t(\cdot) = K(t, \cdot) \in \mathcal{H}$. Moreover, we introduce a positive self-adjoint operator $W_\lambda : \mathcal{H} \to \mathcal{H}$ satisfying

$$\langle W_\lambda f, g \rangle = \lambda J(f, g) \text{ for any } f, g \in \mathcal{H}.$$

Define $V(f, g)$ as $\mathbb{E}\{f(T)g(T)\}$; thus, the inner product $\langle f, g \rangle$ defined in (2.9) becomes the sum of $V(f, g)$ and $\langle W_\lambda f, g \rangle$. It is evident that if T is uniformly distributed over $[0, 1]$, then $V(f, f)$ simplifies to the squared L_2 norm.

Suppose $\{a_\nu\}_{\nu=1}^\infty$ and $\{b_\nu\}_{\nu=1}^\infty$ are two positive sequences. Denote $a_\nu \asymp b_\nu$ if

$$0 < \underline{\lim}_{\nu\to\infty} a_\nu/b_\nu \leq \overline{\lim}_{\nu\to\infty} a_\nu/b_\nu < \infty.$$

Denote the sup-norm of $g \in \mathcal{H}$ and the L_2 norm of $f \in \mathcal{L}_2(\mathcal{T})$ as

$$\|g\|_{\sup} = \sup_{t\in\mathcal{T}} |g(t)| \text{ and } \|f\|_2 = \left\{ \int_{t\in\mathcal{T}} f^2(t)\, \mathrm{d}t \right\}^{1/2},$$

respectively. Finally, let $\sum_v = \sum_{v=1}^\infty$ for convenience.

Assumption 2.5. *There exists a sequence of eigenfunctions $h_v \in \mathcal{H}$ such that $\sup_{v\in\mathbb{N}_+} \|h_v\|_{\sup} < \infty$ and a non-decreasing sequence of eigenvalues $\gamma_v \asymp v^{2m}$ that satisfy*

$$V(h_u, h_v) = \delta_{uv} \quad \textit{and} \quad J(h_u, h_v) = \gamma_u \delta_{uv}, \qquad u, v \in \mathbb{N}.$$

Particularly, every $g \in \mathcal{H}$ admits a Fourier expansion $g = \sum_v V(g, h_v)h_v$.

Assumption 2.5 is pivotal in establishing the asymptotic results, as it establishes the connection between the two bilinear forms V and J. Such assumptions are widely adopted in the FDA literature to develop consistency and the minimax convergence rate of estimators; see Shang and Cheng (2013) and Sun et al. (2018) for example. This assumption leads

to an analytic expression for any function in $\mathcal{H}$, and plays an important role in determining upper bounds for several key functions. The following two lemmas follow immediately from Assumption 2.5 and are crucial for subsequent theoretical developments; for example, see Remark 2.2.

Lemma 2.1. *Under Assumption 2.5, for any $g \in \mathcal{H}$, we have*

$$\|g\|^2 = \sum_u |V(g, h_u)|^2 (1 + \lambda\gamma_u), \;\; K_t(\cdot) = \sum_v \frac{h_v(t)}{1 + \lambda\gamma_v} h_v(\cdot),$$

and $W_\lambda h_v(\cdot) = \lambda\gamma_v/(1 + \lambda\gamma_v)h_v(\cdot)$.

Lemma 2.2. *Under Assumption 2.5, $\|W_\lambda f\|^2 \leq \lambda J(f, f)$ and $\|K_t\|^2 \leq C_K^2 h^{-1}$, for any $f \in \mathcal{H}$ and $t \in \mathcal{T}$, where C_K is a universal constant that is not associated with t.*

We present the following lemma that confirms the existence of the eigen-system within $\mathcal{H}$ as outlined in Assumption 2.5. A similar result can be found in Shang and Cheng (2013).

Lemma 2.3. *Suppose Assumption 2.1 holds, and the density function of T, $\pi(t)$, has up to $(2m-1)$th order continuous derivatives on $\mathcal{T}$. Then the eigenvalues γ_v and the corresponding eigenfunctions h_v, solved from the following ordinary differential equation systems,*

$$(-1)^m h_v^{2m}(\cdot) = \gamma_v \pi(\cdot) h_v(\cdot),$$
$$h_v^j(0) = h_v^j(1) = 0, \;\; j = m, m+1, \ldots, 2m-1,$$

satisfy Assumption 2.5 if we normalize h_v such that $V(h_v, h_v) = 1$.

Finally, we derive the first and second-order Fréchet derivatives of $\ell(f, \lambda)$ in (2.5) with respect to f as detailed below:

$$\begin{aligned} D\ell(f,\lambda)\Delta g &= -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}\left\langle K_{T_{ij}}, \Delta g\right\rangle + \langle W_\lambda f, \Delta g\rangle \\ &= \langle S_M(f) + W_\lambda f, \Delta g\rangle \\ &= \langle S_{M,\lambda}(f), \Delta g\rangle, \end{aligned}$$

and

$$D^2\ell(f,\lambda)\Delta g\Delta h = \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\langle K_{T_{ij}},\Delta g\right\rangle\left\langle K_{T_{ij}},\Delta h\right\rangle + \langle W_\lambda \Delta g,\Delta h\rangle$$
$$\triangleq \langle DS_{M,\lambda}(f)\Delta g,\Delta h\rangle,$$

where

$$\begin{aligned} S_{M,\lambda}(f) &= -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij}-f(T_{ij})\}K_{T_{ij}} + W_\lambda f \triangleq S_M(f) + W_\lambda f, \\ DS_{M,\lambda}(f)\Delta g &= \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\Delta g(T_{ij})K_{T_{ij}} + W_\lambda \Delta g. \end{aligned} \tag{2.10}$$

2.3.3 Functional Bahadur Representation

In this section, we present the FBR for $\hat{\mu}_\lambda$, the solution to (2.5). For ease of notation, let $h = \lambda^{1/2m}$, where λ is the smoothing parameter in (2.5). We define the set

$$\mathcal{G} = \left\{g \in \mathcal{H} : \|g\|_{\sup} \le 1, J(g,g) \le C_K^{-2}h\lambda^{-1}\right\},$$

where C_K is specified in Lemma 2.2. For the formulation of the FBR, we introduce an empirical process $Z_M(g)$ for any $g \in \mathcal{G}$ and $t \in \mathcal{T}$,

$$Z_M(g)(t) = \frac{1}{\sqrt{M}}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\psi_n(Y_{ij},T_{ij};g)K_{T_{ij}}(t) - E\left\{\psi_n(Y,T;g)K_T(t)\right\}\right].$$

Here $\psi_n(T;g)$ is a real-valued function defined on $\mathcal{T}\times\mathcal{G}$, which may depend on n, the total number of subjects in the sample. Provided certain smoothness conditions are met for ψ_n, we can establish the following concentration inequality.

Lemma 2.4. *Suppose that ψ_n satisfies the following Lipschitz continuous condition:*

$$|\psi_n(T;f) - \psi_n(T;g)| \le C_K^{-1}h^{1/2}\|f-g\|_{\sup}$$

for any $f, g \in \mathcal{G}$, where C_K is specified in Lemma 2.2. Then we have

$$\lim_{n\to\infty} \Pr\left(\sup_{g\in\mathcal{G}} \frac{\|Z_M(g)\|}{h^{-(2m-1)/(4m)}\|g\|_{\sup}^{1-1/(2m)} + n^{-1/2}} \le (5\log\log M)^{1/2} \right) = 1.$$

Based on Lemma 2.4, we can proceed to establish the consistency of $\hat{\mu}_\lambda$ as described below.

Lemma 2.5. *Suppose Assumption 2.1 - 2.5 hold. Further, we assume*

$$h = o(1) \text{ and } \{nh^{3-1/(2m)}\}^{-1/2}(\log\log n)^{1/2} = o(1).$$

Then $\|\hat{\mu}_\lambda - \mu\| = O_p(d_n)$, where $d_n = h^m + (nh)^{-1/2}$. In particular, $\hat{\mu}_\lambda$ achieves the optimal convergence rate if $h \asymp n^{-1/(2m+1)}$.

We are now ready to introduce the crucial technical instrument FBR for $\hat{\mu}_\lambda$.

Theorem 2.1 (functional Bahadur representation)**.** *Suppose all conditions in Lemma 2.5 are satisfied. Then we have*

$$\left\|\hat{\mu}_\lambda - \mu + S_{M,\lambda}(\mu)\right\| = O_p(a_n),$$

where $a_n = n^{-1/2}h^{-(6m-1)/4m}\{h^m + (nh)^{-1/2}\}(\log\log n)^{1/2}$.

Recall that $S_{M,\lambda}(f)$ is defined in (2.10) for any $f \in \mathcal{H}$ as

$$S_{M,\lambda}(f) = -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}K_{T_{ij}} + W_\lambda f.$$

The first term can be treated as the average of i.i.d. random variables indexed by subjects. The second term is the bias induced by the penalty term. Theorem 2.1 provides a first-order approximation to the difference between the mean estimator and the population mean.

2.3.4 Pointwise Asymptotic Distribution

In this subsection we develop pointwise asymptotic distributions for $\hat{\mu}_\lambda$ based on the FBR established in Theorem 2.1. Subsequently, the pointwise asymptotic distribution for $\hat{\mu}_{\lambda_1} - \hat{\mu}_{\lambda_2}$ can be readily obtained due to the independence between the two groups. We introduce the following notation for convenience:

$$\begin{aligned} I_1^{uv} &= \iint_{t,s\in\mathcal{T}} c(t,s)h_u(t)h_v(s)\pi(t)\pi(s)dtds, \\ I_2^{uv} &= \int_{t\in\mathcal{T}} c(t,t)h_u(t)h_v(t)\pi(t)dt, \\ \mathcal{S}_i(t,s) &= N_i \sum_{u,v} \frac{h_u(t)h_v(s)}{(1+\lambda\gamma_u)(1+\lambda\gamma_v)} \left\{(N_i-1)I_1^{uv} + I_2^{uv} + \sigma_\varepsilon^2\delta_{uv}\right\}, i \le n, \end{aligned} \tag{2.11}$$

where $\{h_u \mid u \in \mathbb{N}\}$ are the basis functions of $\mathcal{H}$ defined in Assumption 2.5. The following assumption ensures that the tail probability of the random objects in model (2.1) decays at an exponential rate.

Assumption 2.6. *There exist some positive constants C_ε, C_X and C_N such that*

$$(i)\ \mathbb{E}\left\{\exp(C_X\|X\|)\right\} < \infty,\ (ii)\ \mathbb{E}\left\{\exp(C_\varepsilon|\varepsilon|)\right\} < \infty,\ (iii)\ \mathbb{E}\left\{\exp(C_N N)\right\} < \infty$$

for X, ε, and N in model (2.1). Furthermore, there exists a constant $C > 0$ such that for any $t \in \mathcal{T}$, (iv) $\mathbb{E}[\{X(t)-\mu(t)\}^4] \le C(\mathbb{E}[\{X(t)-\mu(t)\}^2])^2$.

Proposition 3.2 in Shang and Cheng (2015) demonstrates that part (i) of Assumption 2.6 is satisfied if X is a Gaussian process with a square-integrable mean function, for any $C_X \in (0, 1/4)$. Part (ii) is met if ε is normally distributed, which was adopted in Yao et al. (2005b). If N follows a Poisson distribution, part (iii) is fullfilled. The fourth moment condition (iv) is commonly seen in literature; see Theorem 8.3.5 in Hsing and Eubank (2015) for example. In particular, C can be taken as 3 in part (iv) when $X(t)$ is a Gaussian process. Additionally, ε has a finite fourth moment due to (ii).

Under Assumption 2.6, we can readily derive the pointwise asymptotic distribution of $\hat{\mu}_\lambda$ using Theorem 2.1. We define the biased mean function as $\mu_b = (\mathrm{id} - W_\lambda)\mu$, where id

is the identity operator in $\mathcal{H}$ that maps any function within $\mathcal{H}$ to itself. The bias term comes from the penalization of the function's smoothness in (2.5). Indeed, the mean of the asymptotic distribution in the following theorem is μ_b, and the process of bias removal is elaborated in Remark 2.2.

Theorem 2.2. *Assume all conditions in Lemma 2.5 and Assumption 2.6 are satisfied. We also assume,*

$$\lim_{n\to\infty} h \sum_{i=1}^{n} \mathbb{E}\left\{\frac{n}{M^2}\mathcal{S}_i(t,t)\right\} \triangleq \sigma_t^2, \ \forall t \in \mathcal{T}. \tag{2.12}$$

Additionally, if $a_n = o(n^{-1/2})$, where a_n is specified in Theorem 2.1, and

$$\log^2(n)\exp\left\{-\frac{(nh)^{1/2}}{\log(n)}\right\} = o(h^2),$$

then we have

$$\frac{(nh)^{1/2}\left\{\hat{\mu}_\lambda(t) - \mu_b(t)\right\}}{\sigma_t} \xrightarrow{d} N(0,1).$$

Remark 2.1. *To better appreciate the variance term as outlined in (2.12), we consider a specific case where $N_i = m_0$ almost surely, where m_0 is a fixed integer no less than one. A direct computation leads to the following result:*

$$\sigma_t^2 = \lim_{n\to\infty} h \sum_{u,v} \frac{h_u(t)h_v(t)}{(1+\lambda\gamma_u)(1+\lambda\gamma_v)}\left(\frac{m_0-1}{m_0}I_1^{uv} + \frac{1}{m_0}I_2^{uv} + \frac{\sigma_\varepsilon^2}{m_0}\delta_{uv}\right).$$

Notice that, the term $\{h\sigma_\varepsilon^2\sum_u h_u^2(t)/(1+\lambda\gamma_u)^2\}/m_0$ reflects the variability resulting from the application of smoothing splines within the RKHS framework, a phenomenon frequently noted in the literature. This can be seen in Theorem 3.5 of Shang and Cheng (2013) and Theorem 3.3 of Hao et al. (2021), for instance. The remaining terms account for the variability introduced by the stochastic nature of X and T.

If the bias term, $W_\lambda(\mu)$, is negligible in comparison to the variance term, we have the following result for $\hat{\mu}(t)$.

Corollary 2.1. *Suppose the conditions in Theorem 2.2 hold and*

$$\lim_{n\to\infty}(nh)^{1/2}(W_\lambda\mu)(t) = -b_t = 0.$$

Then we have

$$\frac{(nh)^{1/2}\{\hat{\mu}_\lambda(t) - \mu(t)\}}{\sigma_t} \xrightarrow{d} N(0,1),$$

where σ_t^2 is defined in (2.12).

Remark 2.2. *It should be noted that even when $\lim_{n\to\infty}(nh)^{1/2}(W_\lambda\mu)(t)$ exists, the bias term $-b_t$ may not necessarily be zero. In the context of nonparametric regression using smoothing splines, Shang and Cheng (2013) investigated appropriate conditions under which the bias term becomes negligible. The key idea is to ensure that $b_t/\sigma_t = o(1)$, i.e., the squared bias is dominated by the variance term. A sufficient condition to achieve this is*

$$\sum_v V^2(\mu, h_v)\gamma_v^2 < \infty \text{ and } nh^{4m} = o(1).$$

Given this condition, it can be shown that:

$$\begin{aligned}
|W_\lambda\mu(t)| &= \left|W_\lambda \sum_v V(\mu, h_v)h_v(t)\right| \\
&\leq \sum_v \left|V(\mu, h_v)\right|\left|\frac{\lambda\gamma_v}{1+\lambda\gamma_v}h_v(t)\right| \\
&\leq \lambda \sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\left\{\sum_v V^2(\mu, h_v)\gamma_v^2\right\}^{1/2}\left\{\sum_v \frac{1}{(1+\lambda\gamma_v)^2}\right\}^{1/2} \\
&\leq O(\kappa_n), \text{ where } \kappa_n = h^{2m-1/2}\left\{\sum_v V^2(\mu, h_v)\gamma_v^2\right\}^{1/2}.
\end{aligned}$$

Denote N_0 as the smallest integer satisfying the condition

$$\frac{3\sigma_t^2}{4} \geq h\sum_{i=1}^{n}\mathbb{E}\left\{\frac{n}{M^2}\mathcal{S}_i(t,t)\right\}.$$

Then for $n \geq N_0$, it follows that:

$$\begin{aligned}\frac{b_t^2}{\sigma_t^2} &\leq \frac{3nhW_\lambda^2\mu(t)}{4h\sum_{i=1}^n \mathbb{E}\left\{n\mathcal{S}_i(t,t)/M^2\right\}} \\ &\leq C_0 nh^{4m}\sum_v V^2(\mu, h_v)\gamma_v^2 = o(1).\end{aligned}$$

We now establish the asymptotic distribution for the estimator of the mean difference, under the assumption that the two groups are independent. The sample sizes (and corresponding penalty parameters) for the two groups are denoted as n_1 (λ_1) and n_2 (λ_2), respectively. Let $h_g = \lambda_g^{1/(2m)}$ for $g = 1, 2$. The following theorem is a direct result of applying Theorem 2.2 and Corollary 2.1.

Theorem 2.3. *Suppose the conditions in Corollary 2.1 are satisfied. Additionally, we assume*

$$\lim_{n_g\to\infty}(n_g h_g)^{1/2}(W_{\lambda_g}\mu_g)(t) = 0,$$

and

$$\lim_{n_g\to\infty} h_g \sum_{i=1}^n \mathbb{E}\left\{\frac{n_g}{M_g^2}\mathcal{S}_{g,i}(t,t)\right\} = \sigma_{g,t}^2, \forall t \in \mathcal{T},\ g = 1, 2.$$

In this context, $\mathcal{S}_{g,i}(t,s)$ is the group-specific variant of $\mathcal{S}_i(t,s)$ as outlined in (2.11). Here, the general terms $c(t,s), \pi(t)$, and N_i are replaced by their respective group-specific equivalents $c_g(t,s), \pi_g(t)$ and N_{gi} for $g = 1, 2$. Furthermore, if

$$a_{g,n} = o(n_g^{-1/2}) \text{ and } \log^2(n_g)\exp\left\{-(n_g h_g)^{1/2}/\log(n_g)\right\} = o(h_g^2),$$

where $a_{g,n}$ is the group-specific counterpart of a_n as defined in Theorem 2.1, we have for any $t \in \mathcal{T}$,

$$\left(\frac{\sigma_{1,t}^2}{n_1 h_1} + \frac{\sigma_{2,t}^2}{n_2 h_2}\right)^{-1/2} \left[\{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\} - \{\mu_1(t) - \mu_2(t)\}\right] \xrightarrow{d} N(0,1), \tag{2.13}$$

where $\sigma_{g,t}^2$ denotes the pointwise variance for group g as outlined previously.

Remark 2.3. *Suppose all notation is the same as in Theorem 2.3, and all associated*

conditions are satisfied. The pointwise hypothesis testing outlined in (2.2) is conducted as follows. Let

$$\widehat{\mathfrak{T}}_{pt}(t) = \left(\frac{\sigma_{1,t}^2}{n_1 h_1} + \frac{\sigma_{2,t}^2}{n_2 h_2}\right)^{-1/2} \{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\}$$

For every fixed time point $t^* \in \mathcal{T}$, under H_0, we reject the null hypothesis at level α if

$$\mathbb{P}\left\{|\widehat{\mathfrak{T}}_{pt}(t^*)| > q_{\alpha/2}\right\} < \alpha,$$

where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2) \times 100\%$ quantile of a standard normal distribution. Under H_α,

$$\begin{aligned}\widehat{\mathfrak{T}}_{pt}(t) = &\left(\frac{\sigma_{1,t}^2}{n_1 h_1} + \frac{\sigma_{2,t}^2}{n_2 h_2}\right)^{-1/2} [\hat{\mu}_1(t) - \hat{\mu}_2(t) - \{\mu_1(t^*) - \mu_2(t^*)\}] \\ &+ \left(\frac{\sigma_{1,t}^2}{n_1 h_1} + \frac{\sigma_{2,t}^2}{n_2 h_2}\right)^{-1/2} \{\mu_1(t^*) - \mu_2(t^*)\}.\end{aligned}$$

Note that, $n_g h_g \to \infty$ as $n \to \infty$ by the condition $\{nh^{3-1/(2m)}\}^{-1}(\log\log n) = o(1)$. Therefore, $\widehat{\mathfrak{T}}_{pt}(t) \to \infty$ if $n \to \infty$ under H_α.

2.3.5 Weak Convergence

This subsection is dedicated to establishing weak convergence of $\hat{\mu}_\lambda - \mu$ as a random function in $\mathcal{H}$, extending it to two groups for a global test for the mean difference. Without loss of generality, we assume that T_{ij} follows a uniform distribution on $[0, 1]$.

We define a covariance operator $\mathcal{C} : \mathcal{H} \to \mathcal{H}$ as follows:

$$\mathcal{C}(f) = \mathbb{E}\left\{\langle X - \mu, f\rangle_2 (X - \mu)\right\},$$

where $\langle f, g\rangle_2 = \int_{\mathcal{T}} f(t)g(t)dt$ for any $f, g \in \mathcal{L}^2(\mathcal{T})$. Clearly, we have

$$\langle \mathcal{C}(h_u), h_v\rangle_2 = \iint_{t,s\in\mathcal{T}} c(t,s)h_u(t)h_v(s)dtds = I_1^{uv}, \tag{2.14}$$

where I_1^{uv} is defined in (2.11) as T is uniformly distributed on [0, 1]. The proof of weak convergence relies on the decay rate of the integral (2.14) as the indices $u, v \to \infty$.

In the FDA literature, it is common to assume that the covariance operator is a trace-class operator; see Hall and Horowitz (2007) and Zhou et al. (2023) for examples. If $h_v(t)$ coincides with eigenfunctions of the covariance operator $\mathcal{C}$ for $v \geq 1$, the associated eigenvalue, which is defined as

$$\theta_v = \langle \mathcal{C}(h_v), h_v \rangle_2 = \iint_{t,s \in \mathcal{T}} c(t,s) h_v(t) h_v(s) dt ds,$$

is usually assumed to have a polynomial decay rate, i.e., $\theta_v \leq c v^{-\tau}$ with some universal constant c and $\tau > 1$. This decay rate of θ_v ensures that $\mathcal{C}$ is a trace-class operator. Therefore, under this assumption, $\mathcal{C}$ is also guaranteed to be a Hilbert-Schmidt operator. Moreover, the Hilbert-Schmidt norm of $\mathcal{C}$ does not depend on the choice of orthonormal basis functions. Besides that, Assumption 2.5 implies $\{h_v\}_{v=1}^{\infty}$ forms a basis in $\mathcal{L}^2(\mathcal{T})$ when T follows a uniform distribution on $[0, 1]$. Given these considerations, we introduce the following assumption.

Assumption 2.7. *The Hilbert-Schmidt norm of the covariance operator is finite, i.e.,* $\|\mathcal{C}\|_{\text{HS}}^2 = \sum_v \|\mathcal{C}(h_v)\|_2^2 < \infty$. *Further,*

$$\iint_{t,s \in \mathcal{T}} c(t,s) h_u(t) h_v(s) dt ds \leq u^{-\tau/2} v^{-\tau/2} \text{ and}$$
$$\int_{t \in \mathcal{T}} c(t,t) h_u(t) h_v(t) dt \leq u^{-\tau/2} v^{-\tau/2} \text{ for some } \tau > 1.$$

To establish the weak convergence for $\hat{\mu}_\lambda - \mu$, we denote $C(\mathcal{T})$ as the collection of all continuous functions on $\mathcal{T}$, which is equipped with the distance metric

$$\rho(f, g) = \sup_{t \in \mathcal{T}} |f(t) - g(t)|$$

for any $f, g \in C(\mathcal{T})$.

Theorem 2.4. *Under Assumption 2.7, and considering all the conditions outlined in The-*

orem 2.2, along with the sufficient condition in Remark 2.2, and also

$$\lim_{n\to\infty} h\sum_{i=1}^{n}\mathbb{E}\left\{\frac{n}{M^2}\mathcal{S}_i(t,s)\right\} = C_Z(t,s), \tag{2.15}$$

we have

$$\mathbb{S}_n(t) = (nh)^{1/2}\{\hat{\mu}_\lambda(t) - \mu(t)\}_{t\in\mathcal{T}} \rightsquigarrow \{Z(t)\}_{t\in\mathcal{T}}$$

in $C(\mathcal{T})$, where Z denotes a Gaussian process with zero mean and covariance function C_Z as described in (2.15).

The subsequent theorem is an extension of Theorem 2.4 and employs the same notation as those used in Theorem 2.3.

Theorem 2.5. *Suppose all conditions outlined in Theorem 2.4 are satisfied and*

$$\lim_{n_1,n_2\to\infty}\frac{n_1h_1}{n_2h_2} = r$$

for some positive real number r. Additionally, we assume

$$\lim_{n_g\to\infty} h_g\sum_{i=1}^{n_g}\mathbb{E}\left\{\frac{n_g}{M_g^2}\mathcal{S}_{g,i}(t,s)\right\} = C_{Z_g}(t,s), \quad g = 1,2. \tag{2.16}$$

Then,

$$\mathbb{S}_d(t) = \left\{\frac{(n_1h_1)(n_2h_2)}{n_1h_1+n_2h_2}\right\}^{1/2}\{(\hat{\mu}_{\lambda_1} - \hat{\mu}_{\lambda_2})(t) - (\mu_1-\mu_2)(t)\}_{t\in\mathcal{T}} \rightsquigarrow \{Z_d(t)\}_{t\in\mathcal{T}}$$

in $C(\mathcal{T})$, where Z_d is a Gaussian process with zero mean and covariance function

$$C_d = \frac{1}{r+1}C_{Z_1} + \frac{r}{r+1}C_{Z_2},$$

where C_{Z_1} and C_{Z_2} are defined in (2.16).

Remark 2.4. *For global testing of $\mu_1(t)-\mu_2(t)$ as outlined in (2.3), we employ the Cramér-von Mises criterion (Anderson, 1962). Indeed, Theorem 2.5 and the continuous mapping*

theorem allow us to establish the following:

$$\begin{aligned}
&\frac{(n_1h_1)(n_2h_2)}{n_1h_1+n_2h_2}\int_{t\in\mathcal{T}}\{(\hat{\mu}_{\lambda_1}-\hat{\mu}_{\lambda_2})(t)-(\mu_1-\mu_2)(t)\}^2\,dt\\
\xrightarrow{d}&\int_{t\in\mathcal{T}}\{Z_d(t)\}^2\,dt\\
=&\int_{t\in\mathcal{T}}\left\{\sum_{\nu}\eta_\nu^{1/2}\zeta_\nu\varphi_\nu(t)\right\}^2dt\\
=&\sum_{\nu}\eta_\nu\chi_\nu^2(1)\triangleq\chi_w^2,
\end{aligned}$$

where (η_ν,φ_ν)*'s stand for eigenvalue-eigenfunction pairs of the covariance function of* Z_d*. Additionally,* $\chi_\nu^2(1)$*'s are i.i.d.* χ^2 *distributed random variables, each with one degree of freedom. The series comprising these weighted chi-squared random variables can be approximated either through truncation or using a Welch–Satterthwaite type* χ^2 *approximation (Kokoszka et al., 2017).*

Remark 2.5. *We can construct the global testing outlined in (2.3). Suppose all conditions in Theorem 2.5 are satisfied. Let*

$$\widehat{\mathfrak{T}}_{glo}^2=\frac{(n_1h_1)(n_2h_2)}{n_1h_1+n_2h_2}\int_{t\in\mathcal{T}}\{(\hat{\mu}_{\lambda_1}-\hat{\mu}_{\lambda_2})(t)\}^2\,dt.$$

We reject the null hypothesis at level α *when* $\widehat{\mathfrak{T}}_{glo}^2>q_{\chi_w^2,\alpha}$*, where* $q_{\chi_w^2,\alpha}$ *is the* $(1-\alpha)\times100\%$ *quantile of a* χ_w^2 *defined in Remark 2.4. Under* H_α*, we have*

$$\begin{aligned}
\widehat{\mathfrak{T}}_{glo}^2=&\frac{(n_1h_1)(n_2h_2)}{n_1h_1+n_2h_2}\int_{t\in\mathcal{T}}[\hat{\mu}_{\lambda_1}(t)-\hat{\mu}_{\lambda_2}(t)-\{\mu_1(t)-\mu_2(t)\}]^2dt\\
&+\frac{(n_1h_1)(n_2h_2)}{n_1h_1+n_2h_2}\int_{t\in\mathcal{T}}\{\mu_1(t)-\mu_2(t)\}^2\,dt.
\end{aligned}$$

The above calculation implies that, under H_α*,* $\widehat{\mathfrak{T}}_{glo}^2\to\infty$ *when* $n\to\infty$*.*

2.3.6 Bootstrap Calibration

Ideally, the asymptotic results established in Sections 2.3.4 and 2.3.5 can be directly employed to construct pointwise confidence intervals and test the global null. Nonetheless, implementing these results in practice entails solving ODEs as indicated in Lemma 2.3 and estimating nuisance parameters, such as the covariance function of X. On the one hand, our numerical studies show that the convergence rates of the limits in equation (2.12) and equation (2.15) are slow. On the other hand, the computational time increases exponentially as two-dimensional smoothing is required to estimate the covariance function. To circumvent these technical challenges and enhance computational efficiency, we develop a multiplier bootstrap procedure for performing the inference task. Algorithm 1 presents the details to implement the bootstrap method to construct pointwise confidence intervals for $\mu_1(t) - \mu_2(t)$ and conduct hypothesis testing for (2.3).

Because two groups of data are independent of each other, we only need to establish the validity of the multiplier bootstrap procedure for each group. The following two theorems establish statistical guarantees for the bootstrap procedure, where the group index g is omitted, and $\hat{\mu}_{\lambda,b}$ is the bootstrap mean estimator defined in Algorithm 1.

Theorem 2.6. *Under the conditions of Theorem 2.2, we have for any $t \in \mathcal{T}$,*

$$\frac{(nh)^{1/2}\left\{\hat{\mu}_{\lambda,b}(t) - \hat{\mu}_{\lambda}(t)\right\}}{\sigma_t} \to N(0,1).$$

Theorem 2.7. *Under the conditions of Theorem 2.4, we have*

$$\mathbb{S}_{n,b}(t) = (nh)^{1/2}\{\hat{\mu}_{\lambda,b}(t) - \hat{\mu}_{\lambda}(t)\}_{t\in\mathcal{T}} \rightsquigarrow \{Z(t)\}_{t\in\mathcal{T}}$$

2.4 Simulation Studies

Throughout this section, we consider $\mathcal{H} = H^2(\mathcal{T})$ and $J(f,f) = \int_{\mathcal{T}}\{f''(t)\}^2\,\mathrm{d}t$. We showcase the finite sample performance of the proposed method, henceforth referred to as SS. Details regarding the simulation settings are provided in Section 2.4.1. We then elucidate

Algorithm 1 Proposed multiplier bootstrap procedure

1. Generate i.i.d. bootstrap weights $U_{i,b}^g$ independent of the data $\{T_{gij}, Y_{gij}\}$, where $1 \le i \le n_g, 1 \le b \le B$, and $g = 1, 2$. Here, $U_{i,b}^g = 1 + 2^{-1/2}$ with probability $2/3$ and $U_{i,b}^g = 1 + 2^{1/2}$ with probability $1/3$.

2. The bth bootstrap estimator for group g is computed as follows:

$$\hat{\mu}_{\lambda_g,b} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{2M_g} \sum_{i=1}^{n_g} \sum_{j=1}^{N_{gi}} U_{i,b}^g \{Y_{gij} - f(T_{gij})\}^2 + \frac{\lambda_g}{2} J(f,f) \right],$$

   for $b = 1, 2, \ldots, B$. This optimization problem can be solved by applying a weighted variant of Proposition 2.1; see Chapter 3.2.4 in Gu (2013).

3.a For each $1 \le b \le B$, let

$$\Delta_b(t) = \hat{\mu}_{\lambda_1,b}(t) - \hat{\mu}_{\lambda_2,b}(t) - \{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\}, \quad t \in \mathcal{T}.$$

   Denote $\hat{\sigma}_{\Delta_t}$ as the standard error of $\{\Delta_b(t) \mid b = 1, 2, \ldots, B\}$.

3.b For each b, we construct a variance-adjusted test statistic (Wang et al., 2025)

$$\hat{\kappa}_b^2 = \int_{\mathcal{T}} \left[\frac{\hat{\mu}_{\lambda_1,b}(t) - \hat{\mu}_{\lambda_2,b}(t) - \{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\}}{\hat{\sigma}_{\Delta_t}} \right]^2 dt.$$

4.a The $(1-\alpha) \times 100\%$ confidence interval of $\mu_1(t) - \mu_2(t)$ at every specific $t \in \mathcal{T}$ is as follows:

$$\{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\} \pm q_{1-\alpha/2} \hat{\sigma}_{\Delta_t}, \tag{2.17}$$

   where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2) \times 100\%$ quantile of a standard normal distribution.

4.b At a significance level of α, we reject the null hypothesis in (2.3) if $\hat{\kappa} > q_{1-\alpha}$, where

$$\hat{\kappa}^2 = \int_{\mathcal{T}} \left\{ \frac{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)}{\hat{\sigma}_{\Delta_t}} \right\}^2 dt,$$

   and $q_{1-\alpha}$ is the $(1-\alpha) \times 100\%$ quantile of $\{\hat{\kappa}_b\}_{b=1}^B$.

the estimation performance of our method in Section 2.4.2. Finally, the performance of our proposed method in both pointwise and global testing is thoroughly evaluated in Sections 2.4.3 and 2.4.4, respectively.

2.4.1 Simulation Settings

Data are simulated from model (2.1) for each group, where $\varepsilon_g \sim N(0, \sigma^2_{g,\varepsilon})$. Specifically, $\sigma^2_{1,\varepsilon} = 0.09$ and $\sigma^2_{2,\varepsilon} = 0.04$. For $i = 1, \ldots, n_g$, we generate the random function $X_{gi}(t)$ utilizing the Karhunen–Loéve expansion:

$$X_{gi}(t) = \mu_g(t) + \sum_{k=1}^{\infty} \xi_{gik}\varphi_{gk}(t), \ \ i = 1, \ldots, n_g.$$

In this context, ξ_{gik}'s, which are FPC scores, follow $N(0, \theta_{gk})$. Additionally, φ_{gk} represents the eigenfunction of the covariance function of X_{gi}, corresponding specifically to the kth eigenvalue denoted as θ_{gk}.

To examine the impact of the homogeneous covariance assumption on mean difference detection, we keep the eigensystem for group 1 constant across all simulation settings, while various configurations are explored for the eigensystem of group 2. In particular, for group 1, the eigenvalues are set as $\theta_{11} = 1$, $\theta_{12} = 0.25$, $\theta_{13} = 0.09$, $\theta_{14} = 0.05$, and $\theta_{1k} = 0$ for $k \geq 5$. Additionally, the eigenfunctions for group 1, $\varphi_{1k}(t)$, are set to be $\sin(k\pi t)$ for every $t \in \mathcal{T}$ and for each $k \geq 1$. Regarding the eigensystem of group 2, we consider the following three settings.

1. In setting *c.1*, group 2's eigensystem is identical to that of group 1, meaning $\gamma_{2k} = \gamma_{1k}$ and $\varphi_{2k}(t) = \varphi_{1k}(t)$ for every k and all $t \in \mathcal{T}$.

2. In setting *c.2*, group 2 has different eigenvalues from group 1, but retains the same eigenfunctions. The eigenvalues are set as $\theta_{21} = 0.81, \theta_{22} = 0.36, \theta_{23} = 0.09, \theta_{24} = 0.01$, and $\theta_{2k} = 0$ for $k \geq 5$.

3. In setting *c.3*, the eigensystem of group 2 is completely distinct from that of group 1. Specifically, the eigenvalues are $\theta_{21} = 0.64, \theta_{22} = 0.36, \theta_{23} = 0.16, \theta_{24} = 0.04, \theta_{25} =$

0.01, and $\theta_{2k} = 0$ for $k \geq 6$. The eigenfunctions are set as $\varphi_{2k} = \cos(k\pi t)$ for every $k \geq 1$ and all $t \in \mathcal{T}$.

In terms of the mean functions for these two groups, we consider the following design:

$$\begin{aligned} \mu_1(t) &= (2t - 0.3)^3 + 0.5t, \quad \text{and} \\ \mu_2(t) - \mu_1(t) &= \delta n_2^{-1/4} \left\{ e^t - (2t-1)^3 - 1 \right\}, \end{aligned} \tag{2.18}$$

where δ, ranging in the set $\{0, 0.5, 1\}$, represents the size of the mean difference between the two groups.

To assess the performance of the proposed estimator and testing procedure, we consider different combinations of subject counts in the two groups: specifically, $n_1 = 200$, and n_2 varies among $\{100, 200, 400\}$. The number of observations per subject in both groups is determined by random sampling from a discrete uniform distribution over the set $\{2, 3, \ldots, N_{\max}\}$, where $N_{\max} \in \{6, 10, 14, 18\}$. The larger $N_{\max}$ results in, on average, a greater number of observations per subject. Comparing these settings allows us to evaluate the performance of our method under varying levels of data sparsity.

The observation times for each group are sampled randomly from a uniform distribution over the interval $[0, 1]$. For each of the aforementioned settings, we perform 1000 Monte Carlo simulations. For both the pointwise testing and the global testing outlined in Algorithm 1, we choose to use 300 bootstrap samples, denoted as $B = 300$.

2.4.2 Estimation Performance

In evaluating the estimation efficacy of SS, we compare it against PACE (Yao et al., 2005b), a widely adopted method for estimating the mean function in the context of sparsely observed functional data. The performance of both methods is assessed using the Integrated Mean Squared Error (IMSE), defined as follows:

$$\int_{\mathcal{T}} [\{\mu_1(t) - \mu_2(t)\} - \{\hat{\mu}_{\lambda_1}(t) - \hat{\mu}_{\lambda_2}(t)\}]^2 dt.$$

Table 2.1: Mean and standard error (in parenthesis) of IMSEs across 1000 Monte Carlo runs with $\delta = 0.5$.

Setting	$N_{\max}$	$n_1 = 200, n_2 = 100$		$n_1 = 200, n_2 = 200$		$n_1 = 200, n_2 = 400$	
		SS	PACE	SS	PACE	SS	PACE
c.1	10	0.020(0.019)	0.021(0.018)	0.013(0.010)	0.014(0.010)	0.011(0.008)	0.011(0.008)
	18	0.018(0.016)	0.018(0.015)	0.012(0.010)	0.012(0.010)	0.009(0.008)	0.009(0.008)
c.2	10	0.019(0.015)	0.020(0.015)	0.014(0.010)	0.014(0.010)	0.010(0.008)	0.010(0.007)
	18	0.017(0.015)	0.017(0.015)	0.011(0.009)	0.011(0.008)	0.009(0.008)	0.009(0.008)
c.3	10	0.020(0.012)	0.022(0.012)	0.014(0.009)	0.015(0.009)	0.011(0.007)	0.011(0.007)
	18	0.019(0.012)	0.019(0.011)	0.012(0.008)	0.012(0.007)	0.009(0.007)	0.009(0.006)

We also compare our method to PACE in terms of out-of-sample performance, with detailed results provided in Appendix 2.B.I. It turns out that these two methods have similar out-of-sample performance.

Table 2.1 presents the average values and standard errors of the IMSEs for both methods, calculated over 1000 simulation runs under various settings. It is evident that the proposed method attains estimation accuracy comparable to that of PACE. As anticipated, the accuracy of both methods improves with an increase in n_2. These numerical findings underscore the effectiveness of aggregating observations from all subjects in estimating the mean function.

2.4.3 Pointwise Inference Performance

To evaluate the performance of the pointwise confidence intervals outlined in Algorithm 1, we compute their actual coverage probabilities at values of $t_p \in \{0, 0.02, \ldots, 1\}$. For these pointwise confidence intervals, the nominal coverage probability is set as $1 - \alpha = 0.95$. Figure 2.1 illustrates the actual coverage probabilities of these intervals under different simulation settings.

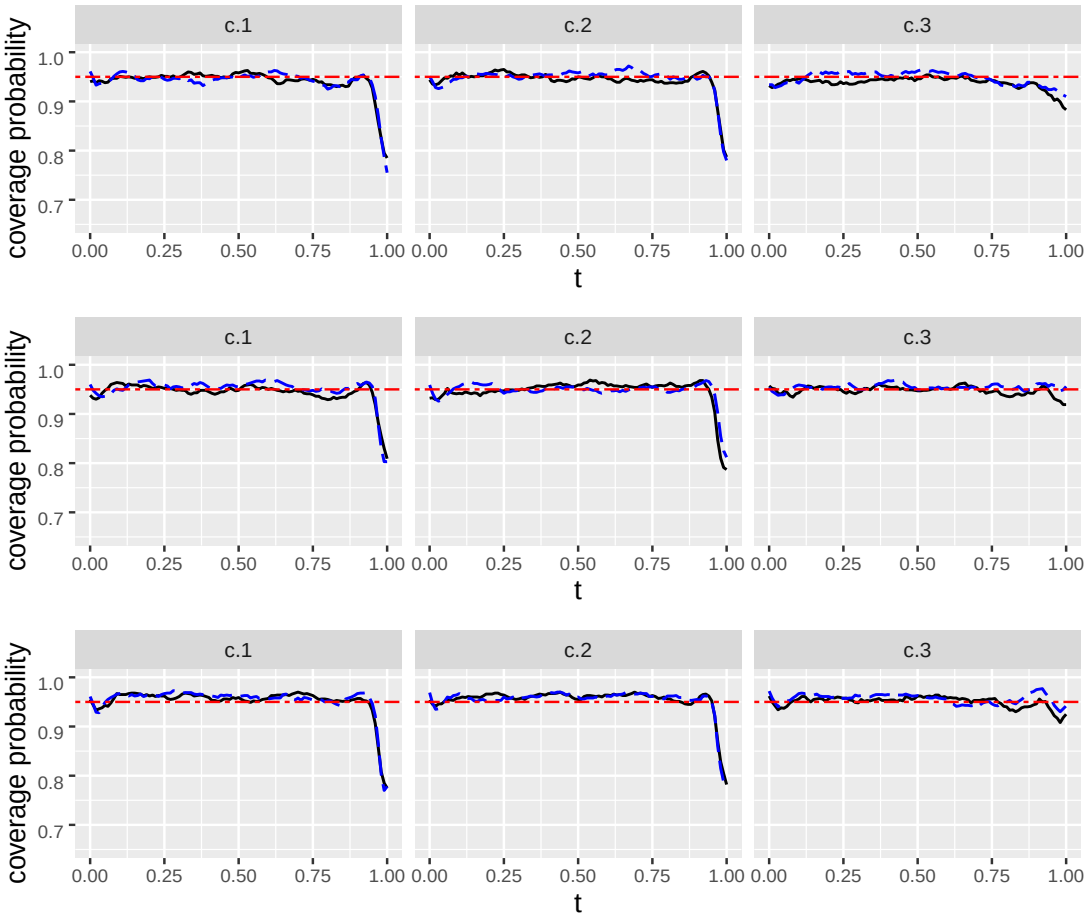


Figure 2.1: In every panel, the red dotted-dash line represents the nominal coverage probability. The actual pointwise coverage probabilities for $N_{\max} = 10$ and $N_{\max} = 18$ are represented by black solid and blue long-dashed lines, respectively. The number of subjects in group 2, denoted by n_2, with values of 100, 200, and 400, arranged from the top to the bottom.

Our observations reveal that the actual coverage probabilities of our proposed method closely align with the nominal level, regardless of the groups having a common covariance structure or not. The only deviation from this pattern occurs near the boundary, where the performance drops. This is likely due to a boundary effect, which implies that the scarcity of observations near the boundary leads to less reliable estimates.

2.4.4 Global Inference Performance

In this subsection, we evaluate the performance of the global testing procedure introduced in Remark 2.4, where the implementation follows Algorithm 1. Our approach is compared with the FPCA-type test statistic (hereinafter referred to as SH), as proposed in Wang (2021), which has shown superior performance against several alternatives. Specifically, SH assumes the covariance functions of these two groups share common eigenfunctions. It estimates these eigenfunctions by aggregating data from both groups. Each individual curve is then projected onto the space spanned by these estimated eigenfunctions. Assuming that the projected scores within each group follow a multivariate normal distribution, Wang (2021) developed a test statistic that asymptotically follows a χ^2 distribution, with the degrees of freedom equal to p, the dimensionality of the projection space. On the suggestion of a referee, we also compare to the functional additive mixed models proposed by Scheipl et al. (2015) (hereinafter referred to as pffr) to estimate the mean function only explained by the group indicator.

Figures 2.2 - 2.4 display the rejection rates at a 5% significance level for these two methods across different simulation designs with sampling frequency $N_{\max} = 10$, and Table 2.2 presents more details that may not be able to obtained directly from these figures. Generally, the power of all methods improves with a larger number of observations and/or a greater magnitude of mean difference between the two groups. This increase in observations can be attributed to either a higher sampling frequency or a greater number of subjects in the second group. The Type-I error of the proposed testing procedure aligns closely with the pre-determined significance level under various settings. In contrast, both SH and pffr tend to produce inflated type I errors. Moreover, our method tends to be more conservative in rejecting the null hypothesis when there is no mean difference ($\delta = 0$) or

Figure 2.2: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.1.

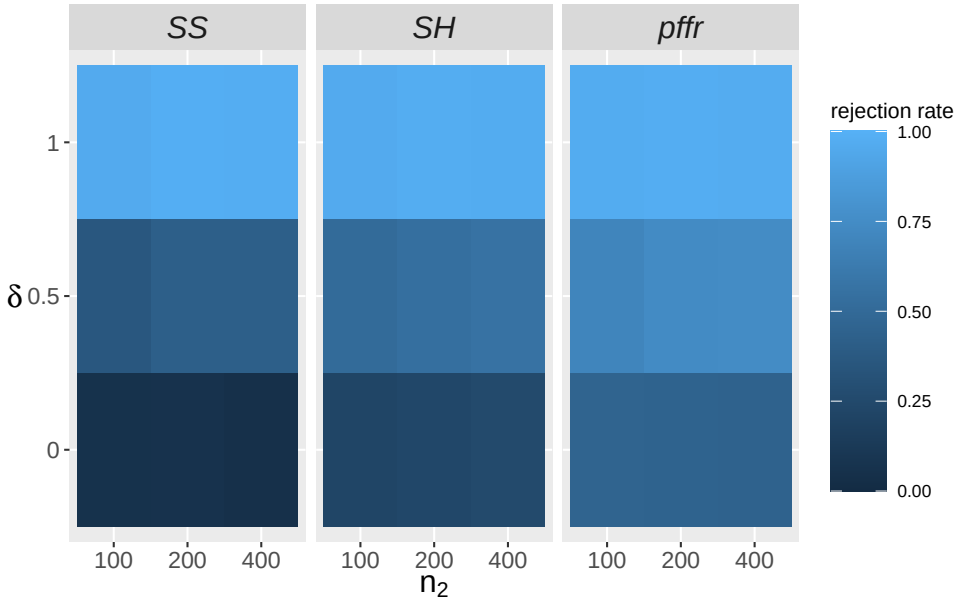


Figure 2.3: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.2.

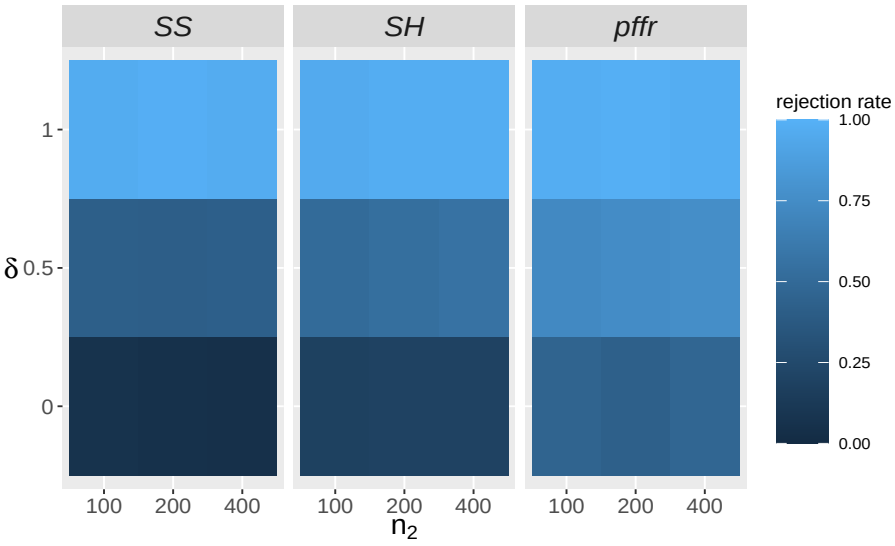


Figure 2.4: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$ with $n_1 = 100$ under setting c.3.

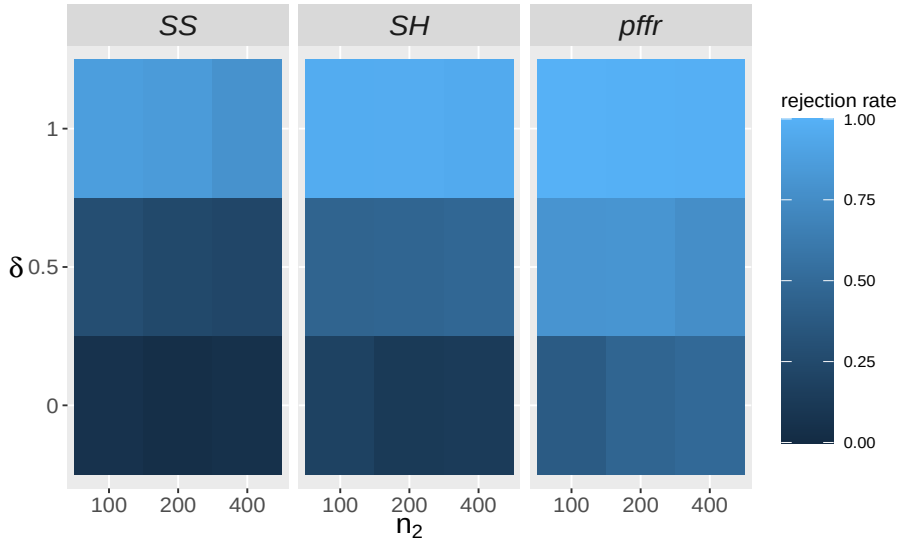


Table 2.2: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 10$

Setting	Test	$n_1 = 200, n_2 = 100$			$n_1 = 200, n_2 = 200$			$n_1 = 200, n_2 = 400$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
c.1	**SS**	0.056	0.358	0.955	0.062	0.414	0.979	0.046	0.417	0.976
	SH	0.213	0.506	0.954	0.232	0.538	0.970	0.252	0.560	0.963
	pffr	0.453	0.696	0.963	0.446	0.737	0.976	0.437	0.743	0.966
c.2	**SS**	0.066	0.416	0.964	0.053	0.411	0.983	0.042	0.420	0.966
	SH	0.180	0.506	0.954	0.193	0.536	0.974	0.192	0.559	0.973
	pffr	0.453	0.727	0.974	0.421	0.748	0.986	0.466	0.758	0.975
c.3	**SS**	0.071	0.297	0.876	0.070	0.274	0.836	0.055	0.227	0.781
	SH	0.202	0.458	0.947	0.157	0.457	0.950	0.138	0.451	0.951
	pffr	0.387	0.801	0.997	0.459	0.805	0.990	0.494	0.759	0.985

the mean difference is relatively small ($\delta = 0.5$), especially when the covariance structures of the two groups are distinct. Additional simulation results for other values of $N_{\max}$ are available in Appendix 2.B.II.

2.5 Real Data Examples

2.5.1 Diffusion Tensor Imaging for Fractional Anisotropy Tract Profiles

The first dataset we focus on originates from a neuroimaging study on individuals with multiple sclerosis (MS). MS is recognized as an immune-mediated inflammatory ailment, manifesting alterations in white-matter tracts. Diffusion tensor imaging (DTI), a type of magnetic resonance imaging, is employed to examine these white-matter tracts.

DTI facilitates the generation of detailed images of white-matter anatomy within the brain by assessing the overall directionality of water diffusion. A commonly employed measure is fractional anisotropy (FA), which characterizes the degree of directional diffusion of water. This study is centered on the analysis of FA profiles parameterized by distance along the tract, the corpus callosum (CC) in particular, to investigate brain structural differences between MS patients and healthy controls.

This dataset consists of 42 control subjects and 100 cases and is publicly available via the `refund` package (Goldsmith et al., 2023). Our analysis focuses on data from the baseline visit of the patients. For each individual, FA readings are recorded at 93 distinct locations along the CC. These readings are then interpreted as regularly spaced observations from a latent smooth function specific to each subject, with the presence of random noise in the measurements.

Figure 2.5 displays trajectories of FA profiles for both the healthy and control groups. While FA observations are typically collected using a regular grid sampling approach, there are instances of missing data in some subjects, as illustrated by the curve marked with a blue dashed line in Figure 2.5. More detailed information regarding the design,

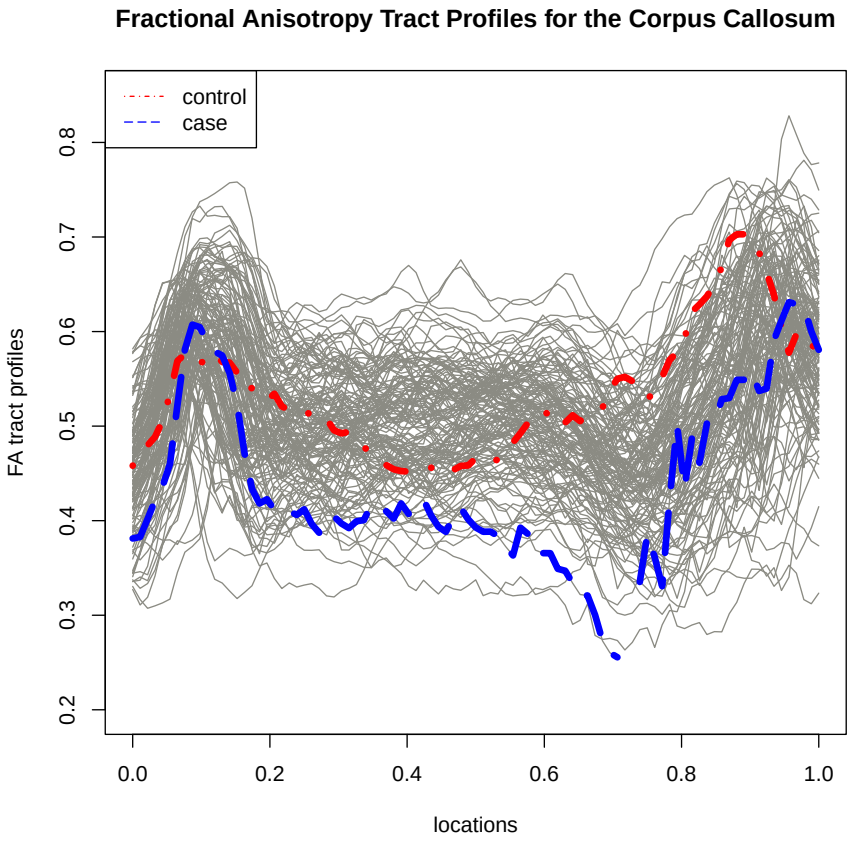


Figure 2.5: The FA profiles along the CC tract for all participants are shown. The graph features a highlighted control subject represented by a red dot-dashed line and a case subject denoted by a blue dashed line.

methodologies, and rationale behind the medication used in this study is available in Ozturk et al. (2010).

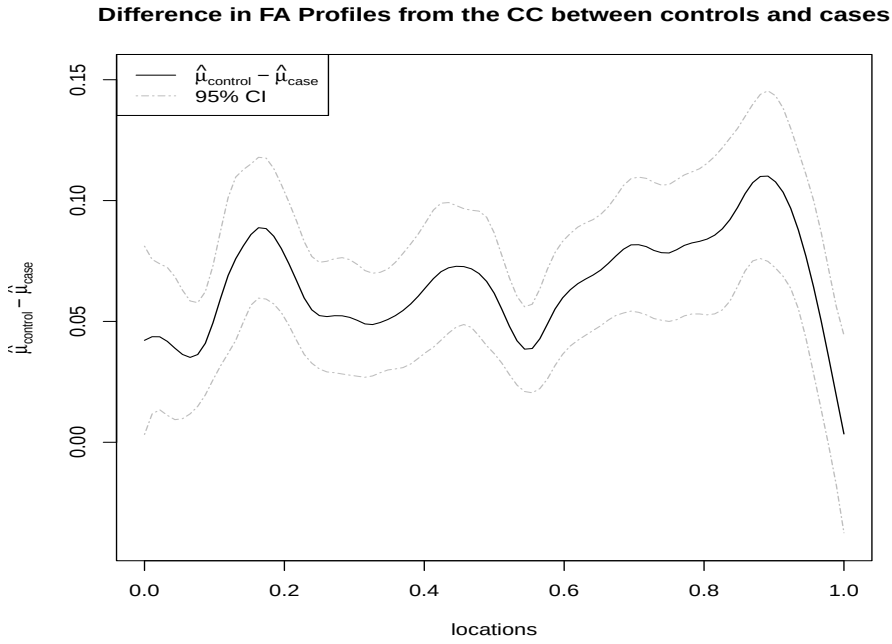


Figure 2.6: The mean difference between cases and controls (black solid line), along with 95% pointwise confidence intervals (blue dot-dashed lines) over the tract of CC.

To showcase the performance of the proposed method, we deliberately 'sparsify' the data, which originally consists of regularly spaced and densely collected observations. For each subject, the number of observations is randomly chosen to be between 2 and 18, with each number being equally likely. The estimated difference in mean functions between the control and case groups, along with 95% pointwise confidence intervals, is shown in Figure 2.6. This illustration uncovers a complex pattern that is not readily apparent from just observing Figure 2.5. It identifies specific locations along the CC where there are notable differences in white-matter between healthy individuals and MS patients. Furthermore, the null hypothesis of identical mean FA profiles across the two groups is rejected at a 5% significance level. This result corroborates existing research, which indicates considerable neuronal loss in the CC of MS patients at the advanced stages, as referenced in (Ozturk

et al., 2010; Pomann et al., 2016). Given that this analysis exclusively utilizes baseline data, it suggests that our method could be effective for early diagnosis of MS.

2.5.2 Beijing Air Pollutants Data

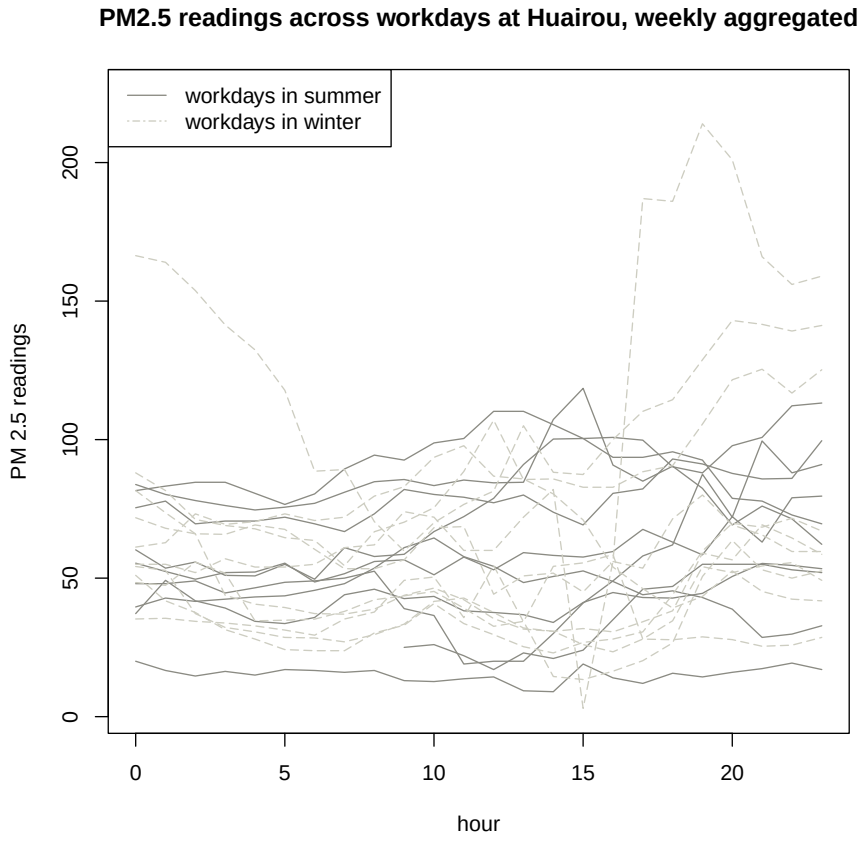


Figure 2.7: PM2.5 readings (weekly aggregate data) from 10 randomly selected weeks from summer (solid lines) and winter (dashed lines) in Huairou, Beijing.

Due to rapid industrialization and urbanization, numerous cities in China are experiencing severe air pollution, with $PM_{2.5}$ being a major pollutant. $PM_{2.5}$ refers to fine particulate matter (PM) with an aerodynamic diameter of less than 2.5 μm. These fine

particles can penetrate the circulatory system directly, leading to the development of cardiovascular and respiratory diseases, and even lung cancers, as reported in studies like (Pope III et al., 2002; Hoek et al., 2013; Lelieveld et al., 2015).

Our investigation focuses on hourly air pollution data from Huairou, located near the northern part of Beijing, collected over a five-year period from March 2013 to February 2017. This data is accessible through the UCI repository (https://doi.org/10.24432/C5RK5G). The high levels of $PM_{2.5}$ observed in this area can be attributed to factors like industrial emissions, human activities, and winter heating, as discussed in Liang et al. (2015) and Zhang et al. (2023). To analyze air pollution patterns in Huairou, we start by consolidating the daily data into weekly averages, reducing data dependency. Our analysis is focused on workday data, from Monday to Friday, specifically during the summer months (June, July, and August) and winter months (December, January, February). From this weekly dataset, we calculate average hourly readings. As a result, for each week within these summer or winter periods, the $PM_{2.5}$ data comprise 24 hourly averages.

Lastly, we 'sparsify' the dataset by randomly selecting between 2 and 10 observations per week, with each number being equally likely. Figure 2.7 illustrates these sparsified $PM_{2.5}$ readings for workdays, showcasing data from ten randomly selected weeks in both the summer and winter seasons. Notably, the $PM_{2.5}$ levels during certain hours in the winter months appear to be exceptionally high.

Figure 2.8 displays the estimated mean difference in $PM_{2.5}$ levels between summer and winter, along with 95% pointwise confidence intervals. Notably, this difference is more pronounced during nighttime, which could be due to the prevalent use of coal for heating in the winter. As the day progresses, the $PM_{2.5}$ disparity gradually lessens, suggesting that daytime human activities become a more significant source of $PM_{2.5}$ emissions. As expected, a global test confirms a significant difference in the average $PM_{2.5}$ levels between summer and winter, aligning with the findings in Liang et al. (2015).

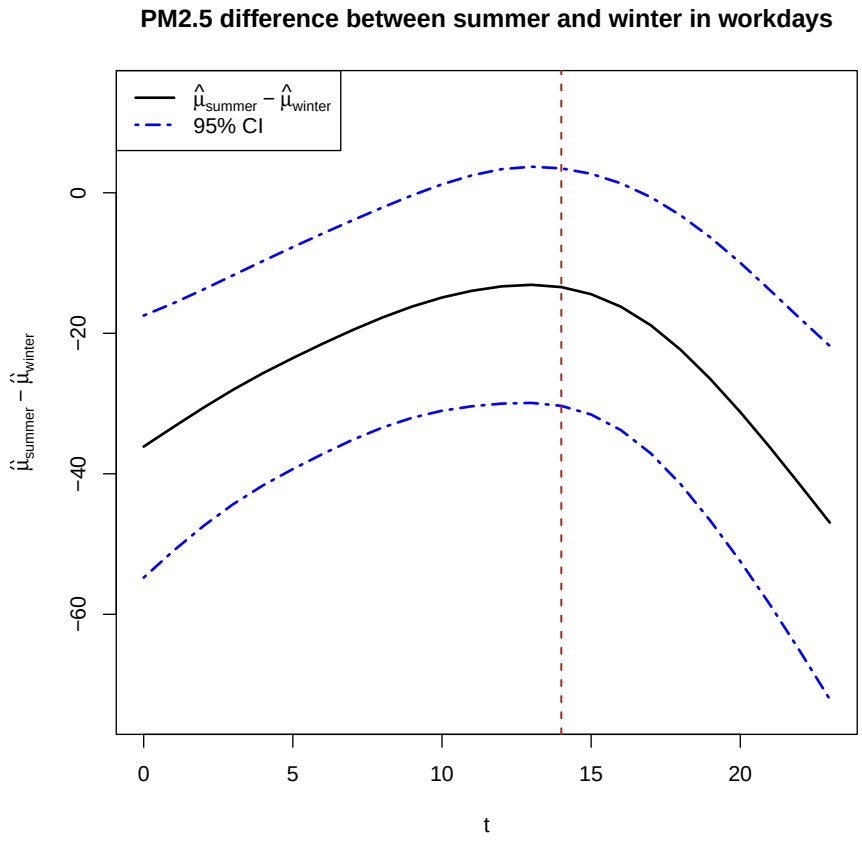


Figure 2.8: Estimated mean differences (black solid line) and 95% confidence intervals (blue dashed lines) of hourly readings of $PM_{2.5}$ between winter and summer in Huairou. The vertical brown dash line indicates 14:00 at the local time.

2.A Theoretical Justification

Here is how this section is organized.

(i) Section 2.A.I contains the proofs for the lemmas in Section 2.3.

(ii) The proof of Theorem 2.1, which establishes the functional Bahadur representation (FBR), is found in Section 2.A.II.

(iii) Section 2.A.III presents the proof of Theorem 2.2, discussing the asymptotic pointwise normality based on FBR.

(vi) The proof of Theorem 2.4, addressing the weak convergence results for the mean estimator, is detailed in Section 2.A.IV.

(v) Section 2.A.V contains the justification of the validity of the bootstrap procedure.

(vi) Section 2.A.VI includes various lemmas that provide technical support for the proofs.

Throughout this appendix, we represent insignificant constants using $c_1, c_2, c_3, \ldots$. On the other hand, we denote important constants related to a specific variable by subscripting them with a letter, such as C_x for constants associated with x. For example, C_K represents a constant related to the maximum norm of the kernel function K in a RKHS, where the inner product is defined in (2.9).

2.A.I Proofs for the Lemmas in Section 2.3

Proof of Lemma 2.1. According to Assumption 2.5, for any $g \in \mathcal{H}$,

$$\begin{aligned}
\|g\|^2 &= \langle g, g \rangle \\
&= \left\langle \sum_u V(g, h_u) h_u, \sum_v V(g, h_v) h_v \right\rangle \\
&= \sum_u \sum_v V(g, h_u) V(g, h_v) \langle h_u, h_v \rangle \\
&= \sum_u \sum_v V(g, h_u) V(g, h_v) (\delta_{uv} + \lambda \gamma_u \delta_{uv}) \\
&= \sum_u |V(g, h_u)|^2 (1 + \lambda \gamma_u).
\end{aligned}$$

Additionally, Assumption 2.5 suggests $K_t = \sum_v V(K_t, h_v) h_v$ as $K_t \in \mathcal{H}$. Thus,

$$\begin{aligned}
h_v(t) &= \langle K_t, h_v \rangle \\
&= \left\langle \sum_u V(K_t, h_u) h_u, h_v \right\rangle \\
&= \sum_u V(K_t, h_u) \langle h_u, h_v \rangle \\
&= V(K_t, h_v)(1 + \lambda \gamma_v).
\end{aligned}$$

Therefore,

$$V(K_t, h_v) = h_v(t)/(1 + \lambda \gamma_v) \text{ and } K_t = \sum_v \frac{h_v(t)}{1 + \lambda \gamma_v} h_v.$$

Regarding $W_\lambda h_v$, Assumption 2.5 ensures that

$$\langle W_\lambda h_v, h_u \rangle = \lambda J(h_v, h_u) = \lambda \gamma_v \delta_{vu}.$$

On the other hand, $W_\lambda h_v = \sum_k V(W_\lambda h_v, h_k) h_k$. Therefore,

$$\begin{aligned}\langle W_\lambda h_v, h_u \rangle &= \left\langle \sum_k V(W_\lambda h_v, h_k) h_k, h_u \right\rangle \\ &= \sum_k V(W_\lambda h_v, h_k) \langle h_k, h_u \rangle \\ &= V(W_\lambda h_v, h_u)(1 + \lambda\gamma_u).\end{aligned}$$

In summary,

$$W_\lambda h_v = \sum_u V(W_\lambda h_v, h_u) h_u = \sum_u \frac{\lambda\gamma_v \delta_{vu}}{1 + \lambda\gamma_u} h_u = \frac{\lambda\gamma_v}{1 + \lambda\gamma_v} h_v.$$

□

Leveraging the results provided in the previous lemma, we can determine the upper bounds for various critical functions outlined in Lemma 2.2. This step is crucial in simplifying our forthcoming development of the FBR.

Proof of Lemma 2.2. According to Assumption 2.5, for any $f \in \mathcal{H}$, we have $f = \sum_v V(f, h_v) h_v$. Therefore, in conjunction with Lemma 2.1,

$$W_\lambda f = \sum_v V(f, h_v) W_\lambda h_v = \sum_v V(f, h_v) \frac{\lambda\gamma_v}{1 + \lambda\gamma_v} h_v.$$

Subsequently,

$$\begin{aligned}\|W_\lambda f\|^2 &= \langle W_\lambda f, W_\lambda f \rangle \\ &= \left\langle \sum_u V(f, h_u) \frac{\lambda\gamma_u}{1 + \lambda\gamma_u} h_u, \sum_v V(f, h_v) \frac{\lambda\gamma_v}{1 + \lambda\gamma_v} h_v \right\rangle \\ &= \sum_u |V(f, h_u)|^2 \frac{\lambda^2\gamma_u^2}{1 + \lambda\gamma_u}.\end{aligned}$$

Furthermore,

$$\begin{aligned}\langle W_\lambda f, f\rangle &= \left\langle \sum_u V(f,h_u)\frac{\lambda\gamma_u}{1+\lambda\gamma_u}h_u, \sum_v V(f,h_v)h_v \right\rangle \\ &= \sum_u |V(f,h_u)|^2\lambda\gamma_u.\end{aligned}$$

Since λ and γ_u's are all positive, we then arrive at this conclusion,

$$\|W_\lambda f\|^2 \le \langle W_\lambda f, f\rangle = \lambda J(f,f), \text{ for any } f \in \mathcal{H}.$$

Finally, as per Assumption 2.5, $\gamma_v \asymp v^{2m}$, implying the existence of constants c_0 and V_1 such that $\gamma_v \ge c_0 v^{2m}$ for all $v \ge V_1$. Furthermore, we can identify constants c_1, c_2, and V_2 such that $c_2 \ge \lambda V_2^{2m} \ge c_1$. Let's define $V = \max(V_1, V_2)$. Therefore, we can establish the following upper bound for the the norm of K_t:

$$\begin{aligned}\|K_t\|^2 &= \left\langle \sum_v \frac{h_v(t)}{1+\lambda\gamma_v}h_v(\cdot), \sum_v \frac{h_u(t)}{1+\lambda\gamma_u}h_u(\cdot) \right\rangle \\ &= \sum_v \frac{h_v^2(t)}{1+\lambda\gamma_v} \\ &\le \left(\sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\right)^2 \sum_v \frac{1}{1+\lambda\gamma_v^{2m}} \\ &\le \left(\sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\right)^2 \left(\sum_{v=1}^{V}\frac{1}{1+\lambda\gamma_v^{2m}} + \sum_{v=V}^{\infty}\frac{1}{1+\lambda\gamma_v^{2m}}\right) \\ &\le \left(\sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\right)^2 \left(V + \sum_{v=V}^{\infty}\frac{1}{\lambda c_0 v^{2m}}\right) \\ &\le \left(\sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\right)^2 \left(V + \frac{1}{\lambda}\frac{c_4}{V^{2m-1}}\right) \\ &\le \left(\sup_{v\in\mathbb{N}}\|h_v\|_{\sup}\right)^2 c_5 V^{-1} \le C_K^2 h^{-1}.\end{aligned}$$

□

Proof of Lemma 2.3. This is a direct consequence of Proposition 2.2 as stated in Shang and Cheng (2013). □

Proof of Lemma 2.4. This proof is directly derived from the proof of Lemma 3.2 presented in Shang and Cheng (2013). □

The upcoming proof is a vital step in deriving the FBR, which is a first-order approximation for the smoothing-spline mean estimator from (2.5). Let's introduce some key notations. Recall that $D\ell(f,\lambda)\Delta g = \langle S_{M,\lambda}(f), \Delta g\rangle$ where

$$\begin{aligned} S_{M,\lambda}(f) &= -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}K_{T_{ij}} + W_\lambda f \in \mathcal{H} \\ &\triangleq S_M(f) + W_\lambda f. \end{aligned}$$

Furthermore, we define

$$\mathbb{E}\{S_{M,\lambda}(f)(t)\} \triangleq S_\lambda(f)(t) \text{ and } \mathbb{E}\{S_M(f)(t)\} \triangleq S(f)(t) \text{ for all } t \in \mathcal{T}.$$

For ease of notation, let's define the following integrals:

$$\begin{aligned} I_{t,s} &= \int_{t,s\in\mathcal{T}} c(t,s)\pi(t)\pi(s)\,\mathrm{d}t\,\mathrm{d}s, \\ I_{t,t} &= \int_{t\in\mathcal{T}} c(t,t)\pi(t)\pi(t)\,\mathrm{d}t\,\mathrm{d}s, \\ \mathcal{G}_1(s,t) &= \int_{(u,v)\in\mathcal{T}^2} c(u,v)K(u,t)K(v,s)\pi(u)\pi(v)\,\mathrm{d}u\,\mathrm{d}v, \\ \mathcal{G}_2(s,t) &= \int_{\mathcal{T}} c(u,u)K(u,t)K(u,s)\pi(u)\,\mathrm{d}u, \\ \mathcal{G}_3(s,t) &= \int_{\mathcal{T}} K(u,t)K(u,s)\pi(u)\,\mathrm{d}u. \end{aligned} \tag{2.19}$$

It's evident that $|I_{t,s}|$ and $|I_{t,t}|$ have upper bounds, as $\pi(t)$ and $c(t,s)$ have upper bounds by Assumptions 2.1 and 2.4, respectively.

Proof of Lemma 2.5. The proof consists of the following two parts.

1. The first part aims to find f_λ s.t. $S_\lambda(f_\lambda) = 0$ where

$$S_\lambda(f)(t) = \mathbb{E}\{S_{M,\lambda}(f)(t)\} \text{ for all } t \in \mathcal{T}.$$

   It should be noted that $S_{M,\lambda}(f)$, as per the Riesz representation theorem, is the unique representer determined by the first Fréchet derivative of the loss function $\ell(f, \lambda)$ in (2.5). Here $S_\lambda(f_\lambda)$ represents the expectation of this representer. This implies that f_λ can serve as an estimator when "$n \to \infty$". The discrepancy between f_λ and the true mean function μ reflects the bias introduced by the penalty term in the model outlined in (2.1).

2. The second part aims to bound the distance between f_λ and the finite-sample mean estimator $\hat{\mu}_\lambda$. This distance illustrates the impact of having only a finite number of data points.

Both parts utilize the following Banach contraction mapping theorem (CMT):

Suppose a mapping $T : \mathcal{X} \to \mathcal{X}$ *satisfies*

$$\|T(x) - T(y)\| \leq k\|x - y\| \text{ for some } k \in [0, 1),$$

where $\mathcal{X} \subseteq \mathcal{H}$ *is a closed set. Then, there is a unique* x_0 *s.t.* $T(x_0) = x_0$.

For more details on the CMT, readers are directed to (Rudin, 1986, Chapter 9, The Contraction Principle).

Part 1: The existence of f_λ.

Denote $B_1 \triangleq B_1(r_1) := \{g \in \mathcal{H} \mid \|g\| \leq r_1\}$, where

$$r_1 = \sqrt{\lambda\{J(\mu, \mu) + 1\}} = h^m\sqrt{\{J(\mu, \mu) + 1\}}$$

and $T_1(\delta) = \delta - S_\lambda(\mu + \delta)$ for any $\delta \in \mathcal{H}$. By the triangle inequality, we have

$$\|T_1(\delta)\| = \|\delta - S_\lambda(\mu + \delta)\| \leq \|\delta - S_\lambda(\mu + \delta) + S_\lambda(\mu)\| + \|S_\lambda(\mu)\|.$$

Note that $\|S_\lambda(\mu)\| = \|S(\mu) + W_\lambda \mu\|$, where

$$\begin{aligned} S(\mu) &= -\mathbb{E}[\{Y - \mu(T)\}K_T] \\ &= -\mathbb{E}\left(\mathbb{E}\Big[\{Y - \mu(T)\}K_T \mid T\Big]\right) \\ &= 0. \end{aligned}$$

Due to Lemma 2.2, we have

$$\|S_\lambda(\mu)\| = \|W_\lambda \mu\| \leq \sqrt{\lambda J(\mu, \mu)} < r_1. \tag{2.20}$$

Additionally, $DS_\lambda(\mu)\delta = \delta$ and $D^2 S_\lambda(\mu + ss'\delta) = 0$ as $DS_\lambda(f) = \text{id}$ for any $f \in \mathcal{H}$ by Lemma 2.7. Thus,

$$\|\delta - S_\lambda(\mu + \delta) + S_\lambda(\mu)\| = \left\|\delta - (DS_\lambda(\mu)\delta + \int_{\mathcal{T}}\int_{\mathcal{T}} sD^2 S_\lambda(\mu + ss'\delta)\delta\delta \,\mathrm{d}s \,\mathrm{d}s')\right\| = 0. \tag{2.21}$$

Combining (2.20) and (2.21), we can establish that $T_1(B_1) \subseteq B_1$ for every $\delta \in B_1$.

To conclude that T_1 is a contraction, we consider the following:

$$\begin{aligned} T_1(\delta_1) - T_1(\delta_2) &= \delta_1 - \delta_2 - S_\lambda(\mu + \delta_1) + S_\lambda(\mu + \delta_2) \\ &= \delta_1 - \delta_2 + \mathbb{E}\left[\{Y - (\mu + \delta_1)(T)\}\, K_T\right] - W_\lambda(\mu + \delta_1) \\ &\quad - \mathbb{E}\left[\{Y - (\mu + \delta_2)(T)\}\, K_T\right] + W_\lambda(\mu + \delta_2) \\ &= \delta_1 - \delta_2 + \mathbb{E}\left[\{(\delta_2 - \delta_1)(T)\}\, K_T\right] + W_\lambda(\delta_2 - \delta_1). \end{aligned}$$

Therefore, for any $g \in \mathcal{H}$,

$$\begin{aligned} &\langle T_1(\delta_1) - T_1(\delta_2), g\rangle \\ &= \langle \delta_1 - \delta_2, g\rangle + \mathbb{E}\left[\{(\delta_2 - \delta_1)(T)\}\, \langle K_T, g\rangle\right] + \langle W_\lambda(\delta_2 - \delta_1), g\rangle \\ &= \langle \delta_1 - \delta_2, g\rangle + \langle \delta_2 - \delta_1, g\rangle \\ &= 0. \end{aligned}$$

The calculations above indicate that

$$T_1(\delta_1) - T_1(\delta_2) = 0 \text{ for any } \delta_1, \delta_2 \in B_1.$$

Thus, T_1 is a contraction defined on B_1.

Concluding Part 1, we apply the CMT, leading to the assertion that there exits a δ_λ such that $T_1(\delta_\lambda) = \delta_\lambda$. This implies $S_\lambda(\mu + \delta_\lambda) = 0$. Defining $f_\lambda = \mu + \delta_\lambda$, we can then conclude that

$$\|\mu - f_\lambda\| \leq r_1 = O(h^m).$$

Part 2:

We will define a bounded linear operator T_2 such that

$$T_2(\delta) = \delta - S_{M,\lambda}(f_\lambda + \delta) \text{ for any } \delta \in \mathcal{H}.$$

It is apparent that

$$\|T_2(\delta)\| \leq \|S_{M,\lambda}(f_\lambda)\| + \|\delta - S_{M,\lambda}(f_\lambda + \delta) + S_{M,\lambda}(f_\lambda)\|,$$

with f_λ derived from Part 1.

To establish an appropriate radius of $B_2 \triangleq B(r_2)$ so that T_2 is a contraction on B_2, we will consider $\mathbb{E}\{\|S_{M,\lambda}(f_\lambda)\|^2\}$. Recalling $\mathbb{E}\{S_{M,\lambda}(f)\} = S_\lambda(f)$, we define

$$O_{ij} = \{Y_{ij} - f_\lambda(T_{ij})\} K_{T_{ij}} \text{ and } D_{ij} = O_{ij} - \mathbb{E}(O_{ij}).$$

It has been shown in Part 1 that $S_\lambda(f_\lambda) = 0$. Thus, $\mathbb{E}\{\|S_{M,\lambda}(f_\lambda)\|^2\}$ reflects the (expected) square of the radius of the ball centred at zero. Since $\mathbb{E}\{S_{M,\lambda}(f)(t)\} = S_\lambda(f)(t)$, it is natural to consider how $S_{M,\lambda}(f_\lambda)$ behaviour around $S_\lambda(f)(t)$. The above observations

motivates us to consider the following:

$$
\begin{aligned}
&\mathbb{E}\{\|S_{M,\lambda}(f_\lambda)\|^2\} \\
&= \mathbb{E}\{\|S_{M,\lambda}(f_\lambda) - S_\lambda(f_\lambda)\|^2\} \\
&= \mathbb{E}\{\|S_M(f_\lambda) - S(f_\lambda)\|^2\} \text{ where } S(f) \text{ is calculated in (2.44)} \\
&= \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\Big\{O_{ij} - \mathbb{E}(O_{ij})\Big\}\right\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{M^2}\Big\|\sum_{i=1}^{n}\{\mathcal{W}_i - \mathbb{E}(\mathcal{W}_i)\}\Big\|^2\right] \\
&\leq \mathbb{E}\left\{\frac{n}{M^2}\|\mathcal{W}_1 - \mathbb{E}(\mathcal{W}_1)\|^2\right\} \\
&\leq \mathbb{E}\left\{\frac{n}{M^2}\|\mathcal{W}_1\|^2\right\}, \qquad (2.22)
\end{aligned}
$$

where $\mathcal{W}_i = \sum_{j=1}^{N_i} O_{ij}$ for $i = 1, 2, \ldots, n$. The expectation of this norm is calculated as follows:

$$
\begin{aligned}
\mathbb{E}\left(\Big\|\sum_{j=1}^{N_1} O_{1j}\Big\|^2\right) &= \sum_{k,l}\mathbb{E}\left\{\langle O_{1k}, O_{1l}\rangle\right\} \\
&= \sum_{k,l}\mathbb{E}\Big[\{Y_{1k} - f_\lambda(T_{1k})\}\{Y_{1l} - f_\lambda(T_{1l})\}\left\langle K_{T_{1k}}, K_{T_{1l}}\right\rangle\Big] \\
&\leq \sup_{t\in\mathcal{T}}\|K_t\|^2\sum_{k,l}\mathbb{E}\Big[\{Y_{1k} - f_\lambda(T_{1k})\}\{Y_{1l} - f_\lambda(T_{1l})\}\Big], \qquad (2.23)
\end{aligned}
$$

where the expectation in the final step can be reformulated as follows

$$
\begin{aligned}
&\mathbb{E}\big[\{Y_{1k} - f_\lambda(T_{1k})\}\{Y_{1l} - f_\lambda(T_{1l})\}\big] \\
&= \mathbb{E}\big[\{Y_{1k} - \mu(T_{1k}) + \mu(T_{1k}) - f_\lambda(T_{1k})\}\{Y_{1l} - \mu(T_{1l}) + \mu(T_{1l}) - f_\lambda(T_{1l})\}\big] \\
&= \mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{Y_{1l} - \mu(T_{1l})\}] + \mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{\mu(T_{1l}) - f_\lambda(T_{1l})\}] \\
&+ \mathbb{E}[\{Y_{1l} - \mu(T_{1l})\}\{\mu(T_{1k}) - f_\lambda(T_{1k})\}] \\
&+ \mathbb{E}[\{\mu(T_{1k}) - f_\lambda(T_{1k})\}\{\mu(T_{1l}) - f_\lambda(T_{1l})\}]. \qquad (2.24)
\end{aligned}
$$

Observe that both terms involving cross products are zero. For instance,

$$\begin{aligned}
&\mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{\mu(T_{1l}) - f_\lambda(T_{1l})\}] \\
&= \mathbb{E}\Big(\mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{\mu(T_{1l}) - f_\lambda(T_{1l})\} \mid T]\Big) \\
&= \mathbb{E}\Big(\{\mu(T_{1l}) - f_\lambda(T_{1l})\}\mathbb{E}\{Y_{1k} - \mu(T_{1k}) \mid T\}\Big) \\
&= 0. \qquad (2.25)
\end{aligned}$$

Additionally,

$$\mathbb{E}[\{\mu(T_{1k}) - f_\lambda(T_{1k})\}\{\mu(T_{1l}) - f_\lambda(T_{1l})\}] = \mathbb{E}^2\{\mu(T) - f_\lambda(T)\}$$

because T_{1k} and T_{1l} are i.i.d. for all observations, regardless of whether they are from the same or different subjects. Furthermore, based on the conclusions from Part 1, we can establish the following results:

$$\begin{aligned}
\mathbb{E}^2\{\mu(T) - f_\lambda(T)\} &\leq \mathbb{E}[\{\mu(T) - f_\lambda(T)\}^2] \\
&= V(\mu - f_\lambda, \mu - f_\lambda) \\
&\leq \|\mu - f_\lambda\|^2 \\
&= r_{1N}^2. \qquad (2.26)
\end{aligned}$$

The remaining term to consider is $\mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{Y_{1l} - \mu(T_{1l})\}]$. Note that,

$$\begin{aligned}
&\mathbb{E}[\{Y_{1k} - \mu(T_{1k})\}\{Y_{1l} - \mu(T_{1l})\}] \\
&= \mathbb{E}\left(\mathbb{E}\Big[\{Y_{1k} - \mu(T_{1k})\}\{Y_{1l} - \mu(T_{1l})\} \mid T_{1k}, T_{1l}\Big]\right) \\
&= \mathbb{E}\{c(T_{1k}, T_{1l}) + \delta_{kl}\sigma_\varepsilon^2\} \\
&= \iint_{t,s\in\mathcal{T}} c(t,s)\pi(t)\pi(s)dtds + \sigma_\varepsilon^2\delta_{kl} \\
&= I_{t,s} + \sigma_\varepsilon^2\delta_{kl} \\
&< \infty, \qquad (2.27)
\end{aligned}$$

where $c(t,s)$ is the covariance function of $X(t)$, and $\delta_{kl} = 1$ if $k = l$ and 0 otherwise.

Combining the results from equations (2.22) - (2.27), we have

$$\mathbb{E}\left(\Big\|\sum_{j=1}^{N_1} O_{ij}\Big\|^2\right) \leq \sup_{t\in\mathcal{T}}\|K_t\|^2\Big[N_1^2\mathbb{E}^2\{\mu(T) - f_\lambda(T)\} + N_1\sigma_\varepsilon^2 + N_1\mathbb{E}\left\{c(T,T)\right\} + (N_1^2 - N_1)\mathbb{E}\left\{c(T_1,T_2)\right\}\Big].$$

Additionally, we have

$$\begin{aligned}
&\mathbb{E}\left\{\frac{n}{M^2}\mathbb{E}\Big(\|\mathcal{W}_1\|^2\Big)\right\} \\
&\leq \mathbb{E}\left\{\sup_{t\in\mathcal{T}}\|K_t\|^2\frac{n}{M^2}\left\{N_1^2 r_{1N}^2 + N_1\sigma_\varepsilon^2 + N_1 I_{t,t} + (N_1^2 - N_1)I_{t,s}\right\}\right\} \\
&\leq \frac{nC_K^2 h^{-1}}{n^2}\left\{(\sigma_N^2 + \mu_N^2)h^{2m} + \mu_N\sigma_\varepsilon^2 + \mu_N I_{t,t} + (\sigma_N^2 + \mu_N^2 - \mu_N)I_{t,s}\right\} \\
&= O((nh)^{-1}),
\end{aligned}$$

where the final step is based on the established bounds for $I_{t,t}$ and $I_{t,s}$. Consequently, this leads to the following:

$$\mathbb{E}\{\|S_{M,\lambda}(f_\lambda)\|^2\} = O((nh)^{-1}),$$

which implies $\|S_{M,\lambda}(f_\lambda)\|^2 = O_p((nh)^{-1})$.

This result enables us to choose a radius for B_2,

$$B_2(r_2) \equiv \{f \in \mathcal{H} \mid \|f\| \leq r_2\}, \text{ where } r_2 = 2C_{B_2}(nh)^{-1/2}.$$

Therefore, we ascertain that, with a probability approaching one,

$$\|S_{M,\lambda}(f_\lambda)\| < C_{B_2}(nh)^{-1/2}.$$

The next step is to handle $\|\delta - S_{M,\lambda}(f_\lambda + \delta) + S_{M,\lambda}(f_\lambda)\|$. As in Part 1, using Taylor

expansion results in

$$\|\delta - S_{M,\lambda}(f_\lambda + \delta) + S_{M,\lambda}(f_\lambda)\| \leq \|\delta - DS_{M,\lambda}(f_\lambda)\delta\| + \left\| \iint_{s,s' \in \mathcal{T}} sD^2 S_{M,\lambda}(f_\lambda + ss'\delta)\delta\delta \, \mathrm{d}s \, \mathrm{d}s' \right\|.$$

The latter term becomes zero as $D^2 S_{M,\lambda}(f) = 0$ for any $f \in \mathcal{H}$. Observing that $DS_\lambda(f_\lambda) = \mathrm{id}$, we can then rewrite this as

$$\begin{aligned} \|\delta - DS_{M,\lambda}(f_\lambda)\delta\| &= \left\| DS_\lambda(f_\lambda)\delta - DS_{M,\lambda}(f_\lambda)\delta \right\| \\ &= \left\| \frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left[\delta(T_{ij}) K_{T_{ij}} - \mathbb{E}\{\delta(T)K(T)\} \right] \right\|. \end{aligned} \tag{2.28}$$

Let $\psi_n(T, \tilde{\delta}) = C_K^{-1} h^{1/2} \tilde{\delta}(T)$ where $\tilde{\delta}(T) = C_K^{-1} h^{1/2} \|\delta\|^{-1} \delta(T)$ for any $\delta \in B_2$. Note that,

$$\begin{aligned} \|\tilde{\delta}\|_{\sup} &= C_K^{-1} h^{1/2} \|\delta\|^{-1} \|\delta\|_{\sup} \\ &\leq C_K^{-1} h^{1/2} \|\delta\|^{-1} C_K h^{-1/2} \|\delta\| \leq 1, \end{aligned}$$

where the final line employs Lemma 2.6. Additionally,

$$J(\tilde{\delta}, \tilde{\delta}) \leq \lambda^{-1} \|\tilde{\delta}\|^2 = \lambda^{-1} C_K^{-2} h \|\delta\|^{-2} \|\delta\|^2 \leq \lambda^{-1} C_K^{-2} h.$$

Thus, $\tilde{\delta}(T) \in \mathcal{G}$. Furthermore,

$$|\psi_n(T, \tilde{\delta}_1) - \psi_n(T, \tilde{\delta}_2)| \leq C_K^{-1} h^{1/2} \|\tilde{\delta}_1 - \tilde{\delta}_2\|_{\sup},$$

which satisfies the Lipschitz condition in Lemma 2.4. Therefore, for any $\delta \in B_2$ and sufficiently large n, we have the following result:

$$\left\| \sum_{i=1}^{n} \sum_{j=1}^{N_i} \psi_n(T_{ij}, \tilde{\delta}) K_{T_{ij}} - \mathbb{E}\{\psi_n(T, \tilde{\delta}) K(T)\} \right\| \leq c_1 \left\{ M^{1/2} h^{-\frac{2m-1}{4m}} + \left(\frac{M}{n} \right)^{1/2} \right\} (5 \log \log M)^{1/2}.$$

Note that $\psi_n(T, \tilde{\delta}) = C_K^{-2}\|\delta\|^{-1}h\delta(T)$. Therefore,

$$\begin{aligned}(2.28) &\leq \sqrt{5}c_1C_K^2(Mh)^{-1}\left\{h^{-\frac{2m-1}{4m}}M^{1/2} + \left(\frac{M}{n}\right)^{1/2}\right\}(\log\log M)^{1/2}\|\delta\| \\ &\leq o_p(1)\|\delta\|.\end{aligned}$$

Therefore, $T_2(B_2) \subseteq B_2$ with a probability that approaches one as sample size increases infinitely. To conclude Part 2, it's necessary to verify that T_2 is a contraction. Note that,

$$\begin{aligned}&S_{M,\lambda}(f_\lambda + \delta_1) - S_{M,\lambda}(f_\lambda + \delta_2) \\ &= S_{M,\lambda}(f_\lambda) + DS_{M,\lambda}(f_\lambda)\delta_1 + \int_0^1\int_0^1 sD^2S_{M,\lambda}(f_\lambda + ss'\delta_1)\delta_1\delta_1 \,\mathrm{d}s\,\mathrm{d}s' \\ &- \left\{S_{M,\lambda}(f_\lambda) + DS_{M,\lambda}(f_\lambda)\delta_2 + \int_0^1\int_0^1 sD^2S_{M,\lambda_2}(f_\lambda + ss'\delta_2)\delta_2\delta_2 \,\mathrm{d}s\,\mathrm{d}s'\right\} \\ &= DS_{M,\lambda}(f_\lambda)(\delta_1 - \delta_2).\end{aligned}$$

For any $\delta_1, \delta_2 \in B_2$. If δ_1 and δ_2 are identical, then $\|T_2(\delta_1) - T_2(\delta_2)\|$ equals zero. Otherwise, when $\|\delta_1 - \delta_2\| > 0$,

$$\begin{aligned}\|T_2(\delta_1) - T_2(\delta_2)\| &= \|\delta_1 - \delta_2 - \{S_{N,\lambda}(f_\lambda + \delta_1) - S_{M,\lambda}(f_\lambda + \delta_2)\}\| \\ &\leq \|(\delta_1 - \delta_2) - DS_{M,\lambda}(f_\lambda)(\delta_1 - \delta_2)\| \\ &= \|DS_\lambda(f_\lambda)(\delta_1 - \delta_2) - DS_{M,\lambda}(f_\lambda)(\delta_1 - \delta_2)\|.\end{aligned}$$

The same arguments used for equation (2.28) apply here. Let $u = \delta_1 - \delta_2$ and $\tilde{u} = C_K^{-1}h^{1/2}\|u\|^{-1}u(T)$, while keeping $\psi(T, f)$ unchanged. It is straightforward to verify that $\tilde{u} \in \mathcal{G}$. Therefore,

$$\begin{aligned}&\|u - DS_{M,\lambda}(f_\lambda)u\| \\ &\leq \sqrt{5}c_2C_K^2(Mh)^{-1}\left\{h^{-\frac{2m-1}{4m}}M^{1/2} + \left(\frac{M}{n}\right)^{1/2}\right\}(\log\log M)^{1/2}\|u\| \\ &= o_p(1)\|u\|.\end{aligned}$$

This suggests that, in probability, T_2 is indeed a contraction mapping on B_2.

Consequently, by applying the fixed point theorem, it follows that there exists a $\delta_{M,\lambda} \in B_2$ such that $T_2(\delta_{M,\lambda}) = \delta_{M,\lambda}$. This means that $S_{M,\lambda}(f_\lambda + \delta_{M,\lambda}) = 0$. Additionally, $S_{M,\lambda}(\hat{\mu}_\lambda) = 0$. Therefore, we conclude that $f_\lambda + \delta_{M,\lambda} = \hat{\mu}_\lambda$, which implies $\|f_\lambda - \hat{\mu}_\lambda\| \leq r_2 = O((nh)^{-1})$.

To summarize, Part 1 establishes that $\|f_\lambda - \mu\| = O(h^m)$, and Part 2 demonstrates that $\|f_\lambda - \hat{\mu}\| = O_p((nh)^{-1})$, where μ is the true mean function. Combining these two results with the triangle inequality leads to

$$\|\hat{\mu} - \mu\| = O_p(h^m + (nh)^{-1/2}).$$

In particular, the optimal convergence rate is achieved when $h = O(n^{1/(2m+1)})$, or equivalently, when $\lambda = O(n^{2m/(2m+1)})$. □

2.A.II Proof of Theorem 2.1

We can now establish a key theorem for this chapter, the *functional Bahadur representation*. This theorem provides a first-order approximation of the mean estimator.

Proof of Theorem 2.1. Let's represent the true mean function as μ, and denote $\hat{\mu}_\lambda$ as the estimator derived from equation (2.5). Additionally, define $\mu_\delta = \hat{\mu}_\lambda - \mu$. First, notice that

$$\begin{aligned} &S_{M,\lambda}(\hat{\mu}_\lambda) - S_\lambda(\hat{\mu}_\lambda) - \{S_{M,\lambda}(\mu) - S_\lambda(\mu)\} \\ &= S_M(\hat{\mu}_\lambda) - S(\hat{\mu}_\lambda) - \{S_M(\mu) - S(\mu)\}. \end{aligned} \tag{2.29}$$

Given that $S_{M,\lambda}(\hat{\mu}_\lambda) = 0$, the left-hand-side of (2.29) becomes

$$\begin{aligned} \text{LHS of (2.29)} &= -S_\lambda(\mu) - DS_\lambda(\mu)\mu_\delta - \iint_{s,s' \in \mathcal{T}} sD^2 S_\lambda(\mu + ss'\mu_\delta)\mu_\delta\mu_\delta \, \mathrm{d}s \, \mathrm{d}s' \\ &\quad - \{S_{M,\lambda}(\mu) - S_\lambda(\mu)\} \\ &= -\{DS_\lambda(\mu)\mu_\delta + S_{M,\lambda}(\mu)\} \\ &= -\{\mu_\delta + S_{M,\lambda}(\mu)\}, \end{aligned}$$

where $D^2 S_\lambda(f) = 0$ as $DS_\lambda(f) = \mathrm{id}$ for any $f \in \mathcal{H}$.

The calculation above suggests that to complete the proof, it is sufficient to establish an upper bound for the norm of the right-hand-side of (2.29). Note that,

$$\begin{aligned}\text{RHS of (2.29)} &= -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\{Y_{ij} - \hat{\mu}_\lambda(T_{ij})\}K_{T_{ij}} - \{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}\right] \\ &\quad - \left(\mathbb{E}[\{Y - \hat{\mu}_\lambda(T)\}K_T] - \mathbb{E}[\{Y - \mu(T)\}K_T]\right) \\ &= \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\mu_\delta(T_{ij})K_{T_{ij}} - \mathbb{E}\{\mu_\delta(T)K_T\}\right]. \end{aligned} \tag{2.30}$$

Therefore, we can utilize the same reasoning as in Lemma 2.5. Specifically, define

$$\tilde{\mu}_\delta(T) = C_K^{-1}h^{1/2}\|\mu_\delta\|^{-1}\mu_\delta(T) \text{ and } \psi(T, f) = C_K^{-1}h^{1/2}f(T).$$

It's straightforward to verify that $\tilde{\mu}_\delta \in \mathcal{G}$, and $\psi(T, f)$ satisfies the Lipschitz condition outlined in Lemma 2.4. Thus, with a probability approaching one,

$$\begin{aligned}&\left\|\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\psi(T_{ij}, \tilde{\mu}_\delta)K_{T_{ij}} - \mathbb{E}\{\psi(T, \tilde{\mu}_\delta)K_T\}\right]\right\| \\ &\leq c_1\left\{M^{1/2}h^{-\frac{2m-1}{4m}} + \left(\frac{M}{n}\right)^{1/2}\right\}(5\log\log M)^{1/2}.\end{aligned}$$

It's worth noting that

$$\psi(T, \tilde{\mu}_\delta) = C_K^{-2}h\|\mu_\delta\|^{-1}\mu_\delta(T),$$

which implies

$$(2.30) \leq \sqrt{5}c_1C_K^2(Mh)^{-1}\left\{M^{1/2}h^{-\frac{2m-1}{4m}} + \left(\frac{M}{n}\right)^{1/2}\right\}(\log\log M)^{1/2}\|\mu_\delta\|.$$

We know that $\|\mu_\delta\| \leq c_2\{h^m + (nh)^{-1/2}\}$ with a probability approaching one, as indicated

by Lemma 2.5. Therefore,

$$(2.29) = O_p(a'_n),$$

where

$$\begin{aligned} a'_n &= \frac{h^m + (nh)^{-1/2}}{Mh} \left\{ M^{1/2} h^{-\frac{2m-1}{4m}} + \left(\frac{M}{n}\right)^{1/2} \right\} (5 \log\log M)^{1/2} \\ &\leq O_p(n^{-1/2} h^{-(6m-1)/4m} \{h^m + (nh)^{-1/2}\} (\log\log n)^{1/2}) \\ &= O_p(a_n). \end{aligned}$$

This concludes the proof for the theorem. □

2.A.III Proof of Theorem 2.2

As discussed in the main text, FBR offers a first-order approximation for the smoothing-spline estimators. Given that $\hat{\mu}_\lambda = \mu + S_{M,\lambda}(\mu) + o_p(1)$ where

$$S_{M,\lambda}(f) = -\frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{Y_{ij} - f(T_{ij})\} K_{T_{ij}} + W_\lambda f,$$

we can apply Lindeberg's Central Limit Theorem (CLT) to deduce the limiting distribution at any $t \in \mathcal{T}$. Let's now present the detailed proof.

Proof of Theorem 2.2. Let's consider $r_\delta(t) = \hat{\mu}_\lambda(t) - \mu_b(t) + S_M(\mu)(t)$, where $\mu_b(t) = (\mathrm{id} - W_\lambda)\mu(t)$ and

$$S_M(\mu)(t) = -\frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{Y_{ij} - \mu(T_{ij})\} K_{T_{ij}}(t) \text{ for any } t \in \mathcal{T}.$$

We can re-write r_δ as follows:

$$r_\delta = \hat{\mu}_\lambda - \mu_b + S_M(\mu) = (\hat{\mu}_\lambda - \mu) + S_{M,\lambda}(\mu).$$

According to Theorem 2.1 and the assumptions in Theorem 2.2, we know that $\|r_\delta\| = O_p(a_n) = o_p(n^{-1/2})$. Regarding $S_{M,\lambda}(\mu)$, define $D_{ij}(\mu) \triangleq \{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}$. Then,

$$\begin{aligned}
\mathbb{E}\big[\|S_M(\mu)\|^2\big] &= \mathbb{E}\left[\Big\|-\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij}-\mu(T_{ij})\}K_{T_{ij}}\Big\|^2\right] \\
&= \mathbb{E}\left\{\frac{1}{M^2}\sum_i\sum_j\sum_k\sum_l \langle D_{ik}(\mu), D_{jl}(\mu)\rangle\right\} \\
&= \mathbb{E}\left[\frac{1}{M^2}\sum_{i,j,k,l}\mathbb{E}\{\langle D_{ik}(\mu), D_{jl}(\mu)\rangle \mid N_i, N_j\}\right] \\
&\le \sup_{t\in\mathcal{T}}\|K_t\|^2\mathbb{E}\left(\frac{1}{M^2}\sum_{i,j,k,l}\mathbb{E}[\{Y_{ik}-\mu(T_{ik})\}\{Y_{jl}-\mu(T_{jl})\}]\right) \\
&= \sup_{t\in\mathcal{T}}\|K_t\|^2\mathbb{E}\left(\frac{1}{M^2}\sum_{i,k,l}\mathbb{E}[\{Y_{ik}-\mu(T_{ik})\}\{Y_{il}-\mu(T_{il})\}]\right) \\
&= \sup_{t\in\mathcal{T}}\|K_t\|^2\mathbb{E}\left[\frac{1}{M^2}\sum_{i=1}^{n}\Big\{N_i(N_i-1)I_{t,s}+N_i\big(I_{t,t}+\sigma_\varepsilon^2\big)\Big\}\right] \\
&\le \sup_{t\in\mathcal{T}}\|K_t\|^2\mathbb{E}\left\{\frac{nN(N-1)}{M^2}I_{t,s}+\frac{nN}{M}\big(I_{t,t}+\sigma_\varepsilon^2\big)\right\} \\
&\le C_K^2(nh)^{-1}\left\{\Big(\mu_N^2+\sigma_N^2\Big)I_{t,s}+\mu_N(I_{t,t}+\sigma_\varepsilon^2)\right\} \\
&= O((nh)^{-1}).
\end{aligned}$$

Observe that the final step takes advantage of the bounded nature of $I_{t,s}$ and $I_{t,t}$. Thus, $S_M(\mu) = O_p((nh)^{-1/2})$, and $a_n = o_p(S_M(\mu))$. Additionally, it can be readily verified that

$$\begin{aligned}
|(nh)^{1/2}r_\delta(t)| &= (nh)^{1/2}|\langle K_t, r_\delta\rangle| \\
&\le (nh)^{1/2}\|K_t\|\|r_\delta\| \\
&= O_p((nh)^{1/2}h^{-1/2}a_n) \\
&= o_p(1).
\end{aligned}$$

Therefore,

$$(nh)^{1/2}\{\hat{\mu}_\lambda(t) - \mu_b(t)\} = \frac{(nh)^{1/2}}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t) + (nh)^{1/2}r_\delta(t)$$
$$= \frac{(nh)^{1/2}}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t) + o_p(1).$$

The limiting distribution of the first term on the RHS of the above equation can be established by the Lindeberg's CLT. Specifically, let

$$S_n(t) = (nh)^{1/2}\left[\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)\right]$$
$$= \sum_{i=1}^{n}\frac{(nh)^{1/2}}{M}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)$$
$$\triangleq \sum_{i=1}^{n}\widetilde{\Theta}_i(t).$$

It is important to note that $\mathbb{E}\left\{\widetilde{\Theta}_i(t)\right\} = 0$ and the $\widetilde{\Theta}_i(t)$'s are independent. Moreover, the point-wise variance is shown as follows:

$$\mathbb{E}\left\{\widetilde{\Theta}_i^2(t)\right\} = \mathbb{E}\left[\frac{nh}{M^2}\sum_{j=1}^{N_i}\sum_{k=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}\{Y_{ik} - \mu(T_{ik})\}K_{T_{ij}}(t)K_{T_{ik}}(t)\right]$$
$$= \mathbb{E}\left(\frac{nh}{M^2}\sum_{j,k=1}^{N_i}K_{T_{ij}}(t)K_{T_{ik}}(t)\mathbb{E}\left[\{Y_{ij} - \mu(T_{ij})\}\{Y_{ik} - \mu(T_{ik})\} \mid T_{ij}, T_{ik}\right]\right)$$
$$= \mathbb{E}\left[\frac{nh}{M^2}\sum_{j,k=1}^{N_i}K_{T_{ij}}(t)K_{T_{ik}}(t)\left\{c(T_{ij}, T_{ik}) + \delta_{jk}\sigma_\varepsilon^2\right\}\right]$$
$$= \mathbb{E}\left[\frac{nh}{M^2}\left\{(N_i^2 - N_i)\mathcal{G}_1(t,t) + N_i\mathcal{G}_2(t,t) + \sigma_\varepsilon^2 N_i\mathcal{G}_3(t,t)\right\}\right] = h\mathbb{E}\left\{\frac{n}{M^2}S_i(t,t)\right\},$$

where the final line is based on the assumption stated in Theorem 2.2 and Lemma 2.8.

Thus,

$$\mathbb{V}\mathrm{ar}\left\{\sum_{i=1}^{n}\widetilde{\Theta}_i(t)\right\} = h\sum_{i=1}^{n}\mathbb{E}\left\{n\mathcal{S}_i(t,t)/M^2\right\} = \sigma_t^2 + o(1).$$

The application of Lindeberg's conditions to the triangular array $\{\widetilde{\Theta}_i(t)\}_{i=1}^n$ results

$$\begin{aligned}
&\frac{1}{\sigma_t^2}\sum_{i=1}^{n}\mathbb{E}\left[\widetilde{\Theta}_i^2(t)\mathbb{1}_{\left\{|\widetilde{\Theta}_i(t)|>e\right\}}\right] \\
&\leq \frac{1}{\sigma_t^2}\sum_{i=1}^{n}\left[\mathbb{E}\left\{\widetilde{\Theta}_i^4(t)\right\}\right]^{1/2}\left\{\mathbb{P}(|\widetilde{\Theta}_i(t)|>e)\right\}^{1/2} \\
&\leq \frac{1}{\sigma_t^2}\sum_{i=1}^{n}\left[\mathbb{E}\left\{\left\langle\widetilde{\Theta}_i, K_t\right\rangle^4\right\}\right]^{1/2}\left\{\mathbb{P}(|\widetilde{\Theta}_i(t)|>e)\right\}^{1/2}. \qquad (2.31)
\end{aligned}$$

Note that, according to Lemma 2.9 and Lemma 2.2, we can establish that

$$\mathbb{E}\left\{\left\langle\widetilde{\Theta}_i, K_t\right\rangle^4\right\} \leq \mathbb{E}\left(\|\widetilde{\Theta}\|^4\|K_t\|^4\right) = c_0^2C_K^4(nh)^{-2}.$$

Regarding the probability, it is indicated in Lemma 2.10 that

$$\begin{aligned}
\mathbb{P}\left\{\sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i| > e\right\} &\leq c_1\log^2(n)\exp\left\{-\frac{c_2(nh)^{1/2}}{\log(n)}\right\} + 2\mathbb{P}\left\{N_i > c_3\log(n)\right\} \\
&\leq c_1\log^2(n)\exp\left\{-\frac{c_1(nh)^{1/2}}{\log(n)}\right\} + 2\mathbb{E}\left\{\exp(C_NN_i)\right\}n^{-c_3C_N}.
\end{aligned}$$

Putting all the aforementioned results together and choosing $c_3 > C_N^{-1} + 1$, we can establish that

$$\begin{aligned}
(2.31) &\leq n\left(c_0C_K^2\frac{1}{nh}\right)\left[c_1\log^2(n)\exp\left\{-\frac{c_2(nh)^{1/2}}{\log(n)}\right\} + \frac{2\mathbb{E}\left\{\exp(C_NN_i)\right\}}{n^{c_3C_N}}\right]^{1/2} \\
&\leq c_0C_K^2\left[c_1\frac{\log^2(n)}{h^2}\exp\left\{-\frac{c_2(nh)^{1/2}}{\log(n)}\right\} + \frac{2\mathbb{E}\left\{\exp(C_NN_i)\right\}}{n^{c_3C_N}h^2}\right]^{1/2} \\
&= o(1),
\end{aligned}$$

given that $c_3 C_N > 1$ and $\log^2(n)\exp\left\{-(nh)^{1/2}/\log(n)\right\} = o(h^2)$. Thus, for any $t \in \mathcal{T}$,

$$(nh)^{1/2}\left\{\hat{\mu}(t) - \mu(t)\right\} \xrightarrow{d} N(0, \sigma_t^2),$$

by the Lindeberg's CLT and the symmetry of the standard normal distribution. □

2.A.IV Proof of Theorem 2.4

Proof. The proof consists of three parts.

1. The first part demonstrates the kernel of the limiting process.
2. The second part establishes the asymptotic tightness of the $\mathbb{S}_n(t)$.
3. The third part verifies the weak convergence of the finite-dimensional marginals of $\mathbb{S}_n$.

Part 1:

Note that,

$$\begin{aligned}
\mathbb{S}_n(t) &= (nh)^{1/2}\left\{\hat{\mu}(t) - \mu(t) \ + S_{M,\lambda}(\mu)(t) - W_\lambda(\mu)(t) - S_M(\mu)(t)\right\} \\
&= (nh)^{1/2}\left\{\hat{\mu}(t) - \mu(t) + S_{M,\lambda}(\mu)(t)\right\} - (nh)^{1/2} W_\lambda(\mu)(t) \\
&- (nh)^{1/2} S_M(\mu)(t) \\
&\triangleq I_{1,n}(t) + I_{2,n}(t) + I_{3,n}(t). \qquad (2.32)
\end{aligned}$$

According to Theorem 2.1,

$$\begin{aligned}
\sup_{t\in\mathcal{T}}|I_{1,n}(t)| &= (nh)^{1/2}\left\langle K_t, \hat{\mu} - \mu + S_{M,\lambda}(\mu)\right\rangle \\
&\leq (nh)^{1/2}\|K_t\|\|\hat{\mu} - \mu - S_{M,\lambda}(\mu)\| \\
&= o_p(1).
\end{aligned}$$

As outlined in Remark 2.2, if the sufficient condition specified therein is met,

$$\sup|I_{2,n}(t)| = O((nh)^{1/2}\kappa_n) = o(1).$$

As for the $I_{3,n}(t)$, recall that

$$S_M(\mu)(t) = -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}\, K_{T_{ij}}(t).$$

First, let's rewrite the leading term $I_{3,n}$ in (2.32),

$$I_{3,n}(t) = \sum_{i=1}^{n}\frac{(nh)^{1/2}}{M}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t) \triangleq \sum_{i=1}^{n}\widetilde{\Theta}_i(t),$$

where $\widetilde{\Theta}_i(t) = (nh)^{1/2}\Theta_i(t)/M$ and $\Theta_i(t) = \sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)$.

Furthermore, we have

$$\begin{aligned}
\mathbb{E}\left\{\Theta_i \mid N_i\right\} &= \sum_{j=1}^{N_i}\mathbb{E}\left[\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)\right] \\
&= \sum_{j=1}^{N_i}\mathbb{E}\Big[K_{T_{ij}}(t)\mathbb{E}\left\{Y_{ij} - \mu(T_{ij}) \mid T_{ij}\right\}\Big] \\
&= 0.
\end{aligned}$$

A similar argument can be used to compute the expectation of $I_{3,n}(t)$, yielding

$$\begin{aligned}
\mathbb{E}\{I_{3,n}(t)\} &= (nh)^{1/2}\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}\, K_{T_{ij}}(t)\right] \\
&= (nh)^{1/2}\mathbb{E}\left(\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\Big[K_{T_{ij}}(t)\mathbb{E}\left\{Y_{ij} - \mu(T_{ij}) \mid T_{ij}\right\}\Big]\right) \\
&= 0,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{M^2}\sum_{i=1}^{n}\sum_{\substack{k=1\\k\neq i}}^{n}\sum_{j=1}^{N_i}\sum_{l=1}^{N_k}\{Y_{ij}-\mu(T_{ij})\}\{Y_{kl}-\mu(T_{kl})\}K_{T_{ij}}(t)K_{T_{kl}}(s)\Big]\\
&=\mathbb{E}\Big(\frac{1}{M^2}\sum_{\substack{i,k=1\\i\neq k}}^{n}\sum_{j=1}^{N_i}\sum_{l=1}^{N_k}K_{T_{ij}}(t)K_{T_{kl}}(s)\mathbb{E}\big[\{Y_{ij}-\mu(T_{ij})\}\{Y_{kl}-\mu(T_{kl})\}\mid T_{ij},T_{kl}\big]\Big)\\
&=0.
\end{aligned}$$

Therefore, the covariance between $I_{3,n}(t)$ and $I_{3,n}(s)$ is as follows:

$$\begin{aligned}
&\mathbb{E}\{I_{3,n}(t)I_{3,n}(s)\}\\
&=\mathbb{E}\Big(\frac{nh}{M^2}\Big[\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij}-\mu(T_{ij})\}K_{T_{ij}}(t)\Big]\Big[\sum_{k=1}^{n}\sum_{l=1}^{N_k}\{Y_{kl}-\mu(T_{kl})\}K_{T_{kl}}(s)\Big]\Big)\\
&=\mathbb{E}\Big[\frac{nh}{M^2}\sum_{i=1}^{n}\sum_{j,l=1}^{N_i}\{Y_{ij}-\mu(T_{ij})\}\{Y_{kl}-\mu(T_{kl})\}K_{T_{ij}}(t)K_{T_{il}}(s)\Big]\\
&=\mathbb{E}\Big(\frac{nh}{M^2}\sum_{i=1}^{n}\sum_{j,l=1}^{N_i}K_{T_{ij}}(t)K_{T_{il}}(s)\mathbb{E}\Big[\{Y_{ij}-\mu(T_{ij})\}\{Y_{kl}-\mu(T_{kl})\}\mid T_{ij},T_{il}\Big]\Big)\\
&=\mathbb{E}\Big[\frac{nh}{M^2}\sum_{i=1}^{n}\sum_{j,l=1}^{N_i}K_{T_{ij}}(t)K_{T_{il}}(s)\big\{c(T_{ij},T_{il})+\delta_{jl}\sigma_\varepsilon^2\big\}\Big]\\
&=\mathbb{E}\Big(\frac{nh}{M^2}\sum_{i=1}^{n}\sum_{j,l=1}^{N_i}\mathbb{E}\Big[K_{T_{ij}}(t)K_{T_{il}}(s)\big\{c(T_{ij},T_{il})+\delta_{jl}\sigma_\varepsilon^2\big\}\mid N_i\Big]\Big)\\
&=\mathbb{E}\Big(\frac{nh}{M^2}\sum_{i=1}^{n}\Big[(N_i^2-N_i)\mathcal{G}_1(s,t)+N_i\{\mathcal{G}_2(s,t)+\sigma_\varepsilon^2\mathcal{G}_3(s,t)\}\Big]\Big)\\
&\triangleq C_Z(t,s)+o(1),
\end{aligned}$$

where $\mathcal{G}_i(t,s), i=1,2,3$, are defined in (2.19). The final two lines are derived using the assumption stated in Theorem 2.4 and Lemma 2.8.

Part 2:

The objective of this part is to establish the equicontinuity of $(nh)^{1/2}\{\hat{\mu}(t) - \mu(t)\}$. Given that the supremum of both $I_{1,n}(t)$ and $I_{2,n}(t)$ are $o(1)$, and the equicontinuity can be readily demonstrated for them, we focus our attention on the leading term, $I_{3,n}$. In other words, we aim to show, for any $e > 0$,

$$\lim_{\delta \to 0} \lim_{n \to \infty} \mathbb{P}\left\{ \sup_{d(s,t) \leq \delta} |I_{3,n}(t) - I_{3,n}(s)| > e \right\} = 0. \tag{2.33}$$

To achieve this, we will apply Lemma A.1 from Kley et al. (2016). To do so, we introduce a metric function $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ defined as $d(s,t) = |t - s|$ for any $s, t \in \mathcal{T}$, and a Orlicz function $\Psi(x) = x^2$. The Orlicz norm for any real-valued random variable U is defined as follows,

$$\|U\|_\Psi = \inf\{c > 0 \mid \mathbb{E}\{\Psi(|U|/c)\} \leq 1\}\}.$$

Notice that the square of the Orlicz norm of U corresponds to the second moment of U. Therefor, we can express $\|I_{3,n}(t) - I_{3,n}(s)\|_\Psi^2 = \mathbb{E}[\{I_{3,n}(t) - I_{3,n}(s)\}^2]$. Given that $\mathbb{E}\{\Theta_i(t)\} = 0$ for all i and $t \in \mathcal{T}$, and from Part 1, $\mathbb{E}\left\{\Theta_i(t)\Theta_j(s)\right\} = \mathbb{E}\left\{\Theta_i(t)\right\}\mathbb{E}\left\{\Theta_j(s)\right\} = 0$ for all $i \neq j$ and any $(t,s) \in \mathcal{T} \times \mathcal{T}$. Furthermore,

$$\begin{aligned}
&\mathbb{E}\left\{|I_{3,n}(t) - I_{3,n}(s)|^2\right\} \\
&= \mathbb{E}\left[\frac{nh}{M^2}\Big|\sum_{i=1}^{n} \Theta_i(t) - \Theta_i(s)\Big|^2\right] \\
&= \mathbb{E}\left[\frac{nh}{M^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\Theta_i(t) - \Theta_i(s)\}\{\Theta_j(t) - \Theta_j(s)\}\right] \\
&= \mathbb{E}\left[\frac{nh}{M^2}\sum_{i=1}^{n}\{\Theta_i^2(t) - 2\Theta_i(t)\Theta_i(s) + \Theta_i^2(s)\}\right] \\
&= \mathbb{E}\left[\frac{nh}{M^2}\sum_{i=1}^{n}\left\{\Theta_i^2(t) - 2\Theta_i(t)\Theta_i(s) + \Theta_i^2(s)\right\}\right].
\end{aligned} \tag{2.34}$$

Let us examine each term separately. To begin, $\mathbb{E}\{\Theta_i^2(t) \mid N_i\}$ can be further simplified as

follows:

$$
\begin{aligned}
&\mathbb{E}\{\Theta_i^2(t) \mid N_i\} \\
&= \mathbb{E}\Bigg(\Big[\sum_{j=1}^{N_i} \{Y_{ij} - \mu(T_{ij})\} K_{T_{ij}}(t)\Big] \Big[\sum_{k=1}^{N_i} \{Y_{ik} - \mu(T_{ik})\} K_{T_{ik}}(t)\Big] \;\Big|\; N_i \Bigg) \\
&= \mathbb{E}\Big[\sum_{j,k=1}^{N_i} K_{T_{ij}}(t) K_{T_{ik}}(t) \{c(T_{ij}, T_{ik}) + \sigma_\varepsilon^2 \delta_{jk}\} \mid N_i \Big] \\
&= (N_i^2 - N_i)\mathcal{G}_1(t,t) + N_i \mathcal{G}_2(t,t) + N_i \sigma_\varepsilon^2 \mathcal{G}_3(t,t). \qquad (2.35)
\end{aligned}
$$

Secondly,

$$
\begin{aligned}
&\mathbb{E}\{\Theta_i(t)\Theta_i(s) \mid N_i\} \\
&= \mathbb{E}\Bigg(\Big[\sum_{j=1}^{N_i} \{Y_{ij} - \mu(T_{ij})\} K_{T_{ij}}(t)\Big] \Big[\sum_{k=1}^{N_i} \{Y_{ik} - \mu(T_{ik})\} K_{T_{ik}}(s)\Big] \;\Big|\; N_i \Bigg) \\
&= \mathbb{E}\Bigg[\sum_{j,k=1}^{N_i} K_{T_{ij}}(t) K_{T_{ik}}(s) \{\mathbb{C}\mathrm{ov}(T_{ij}, T_{ik}) + \sigma_\varepsilon^2 \delta_{jk}\} \;\Big|\; N_i \Bigg] \\
&= (N_i^2 - N_i)\mathcal{G}_1(t,s) + N_i \mathcal{G}_2(t,s) + N_i \sigma_\varepsilon^2 \mathcal{G}_3(t,s). \qquad (2.36)
\end{aligned}
$$

Therefore, by combining (2.34) through (2.36), Lemma 2.8, and the condition $M \geq n$, we derive the following result:

$$
\begin{aligned}
&\mathbb{E}\left\{|I_{3,n}(t) - I_{3,n}(s)|^2\right\} \\
&\leq h\Big[(\mu_{N^2} - \mu_N)\Delta_{\mathcal{G}_1}(t,s) + \mu_N \Delta_{\mathcal{G}_2}(t,s) + \mu_N \sigma_\varepsilon^2 \Delta_{\mathcal{G}_3}(t,s)\Big] \\
&\leq \left\{C_{\mathcal{G}_1}(\mu_{N^2} - \mu_N) + C_{\mathcal{G}_2}\mu_N\right\} h^\tau |t-s| + \mu_N \sigma_\varepsilon^2 C_{\mathcal{G}_3} |t-s|^2 \\
&\triangleq c_1 h^\tau |t-s| + c_2 |t-s|^2.
\end{aligned}
$$

In the equation above, μ_{N^2} represents the second moment of N.

It's worth noting that $d(s,t) = |t-s|$ is a well-defined metric. Additionally, we define

$\bar{\eta} = h/2$. Thus, for sufficient large n, when $d(s,t) \geq \bar{\eta}$, we can establish the following inequality:

$$\begin{aligned}
&\left(c_1 h^\tau |t-s| + c_2 |t-s|^2\right) |t-s|^{-2} \\
&= c_1 h^\tau |t-s|^{-1} + c_2 \\
&\leq 2c_1 h^{\tau-1} + c_2 \\
&\leq 2c_1 + c_2.
\end{aligned}$$

Therefore, there exists a universal constant $C_\Psi^2 = 2c_1 + c_2$ such that

$$\|I_{3,n}(t) - I_{3,n}(s)\|_\Psi \leq C_\Psi |t-s| = C_\Psi d(s,t),$$

where C_Ψ depends on $d(s,t)$ and $\Psi(x)$ when $d(s,t) \geq \bar{\eta}$ for sufficient large n.

To handle the case where $d(s,t) \leq \bar{\eta}/2$, we rely on the Bernstein inequality. Therefore, we need an almost sure upper bound for $|\widetilde{\Theta}_i(t) - \widetilde{\Theta}_i(s)|$.

According to Lemma 2.9, we can establish the following:

$$\sup_{t,s\in\mathcal{T}} |\widetilde{\Theta}_i(t) - \widetilde{\Theta}_i(s)| \leq 2 \sup_{t\in\mathcal{T}} |\widetilde{\Theta}_i(t)| \leq 2C_{\widetilde{\Theta}} \frac{\log^2 n}{n^{1/2} h}.$$

Furthermore, when $d(s,t) \leq \bar{\eta}$,

$$\begin{aligned}
&\mathbb{E}\{|\widetilde{\Theta}_i(t) - \widetilde{\Theta}_i(s)|^2\} \\
&= \mathbb{E}\left\{ \frac{nh}{M} |\Theta_i(t) - \Theta_i(s)|^2 \right\} \\
&= \mathbb{E}\left[\frac{nh}{M^2} \left\{ \Theta_i(t)^2 + \Theta_i(s)^2 - 2\Theta_i(t)\Theta_i(s) \right\} \right] \\
&\leq \frac{h}{n} \mathbb{E}\Big[(\mu_{N^2} - \mu_N) \Delta_{G_1}(t,s) + \mu_N \Delta_{\mathcal{G}_2}(t,s) + \sigma_\varepsilon^2 \mu_N \Delta_{\mathcal{G}_3}(t,s) \Big] \\
&\leq \frac{c_1 h^\tau |t-s| + c_2 |t-s|^2}{n} \\
&\leq \frac{c_3 h^2}{n}.
\end{aligned}$$

Let $\mathcal{D}(\epsilon, d)$ represent the associated ϵ-packing number in the metric space $(\mathcal{T}, d)$ where d is the previously defined metric. Notice that $\mathcal{D}(\xi, d) \lesssim \xi^{-1}$, and according to Lemma A.1 in Kley et al. (2016), there exists a set $\tilde{\mathcal{T}}$ containing at most $\mathcal{D}(\bar{\eta}, d)$ points such that, for any $\eta > \bar{\eta}$ and $\delta, e > 0$, as $n \to \infty$,

$$
\begin{aligned}
(2.33) &\leq c_4 \left\{ \int_{\bar{\eta}/2}^{\eta} \mathcal{D}^{1/2}(\zeta, d) d\zeta + (\delta + 2\bar{\eta}) \mathcal{D}(\eta, d) \right\}^2 \\
&\quad + \mathbb{P} \left\{ \sup_{\substack{d(s,t) \leq \bar{\eta} \\ s,t \in \tilde{\mathcal{T}}}} |I_{3,n}(t) - I_{3,n}(s)| > e/2 \right\} \\
&\leq c_5(\eta + \delta^2 \eta^{-2}) + \mathcal{D}(\bar{\eta}, d) \sup_{d(s,t) \leq \bar{\eta}, (s,t) \in \mathcal{T}^2} \mathbb{P}\left\{ |I_{3,n}(t) - I_{3,n}(s)| > e/4 \right\} \\
&\leq c_5(\eta + \delta^2 \eta^{-2}) \\
&\quad + c_6 \bar{\eta}^{-1} \exp\left[- \frac{e^2/32}{\sum_{i=1}^n \mathbb{E}\left\{ |\widetilde{\Theta}_i(t) - \widetilde{\Theta}_i(s)|^2 \right\} + c_7 e (\log^2 n)/(n^{1/2} h)} \right] \\
&\leq c_5(\eta + \delta^2 \eta^{-2}) + c_6 \bar{\eta}^{-1} \exp\left[- \frac{e^2/32}{c_3 \bar{\eta}^2 + c_7 e (\log^2 n)/(n^{1/2} h)} \right] \\
&\leq o(1).
\end{aligned}
$$

The final line is obtained by setting $\eta = \delta^{2/3}$ and $\delta \to 0$.

Regarding the remaining $I_{1,n}$ and $I_{2,n}$ in (2.32), we obtain the following:

$$
\begin{aligned}
&\lim_{\delta \to 0} \lim_{n \to \infty} \mathbb{P} \left\{ \sup_{d(s,t) \leq \delta} |I_{1,n}(t) + I_{2,n}(t) - I_{1,n}(s) - I_{2,n}(s)| > e \right\} \\
&\leq \lim_{\delta \to 0} \lim_{n \to \infty} \mathbb{P} \left\{ \sup_{d(s,t) \leq \delta} |I_{1,n}(t) - I_{1,n}(s)| > e/2 + \sup_{d(s,t) \leq \delta} |I_{2,n}(t) - I_{2,n}(s)| > e/2 \right\} \\
&= 0,
\end{aligned}
$$

where the final step is due to $\sup_t |I_{1,n}(t)| = o_p(1)$ and $\sup_t |I_{2,n}(t)| = o(1)$ in Part 1.

This establishes the equicontinuity of the random process

$$(nh)^{1/2}\{\hat{\mu}(t) - \mu(t)\}.$$

The asymptotic tightness is supported by Theorem 1.5.4 and Theorem 1.5.7 in van der Vaart and Wellner (1996).

Part 3:

To establish the weak convergence of the finite-dimensional marginal distribution of $\mathbb{S}_n(t)$, we will employ the Cramér-Wold device. This involves demonstrating that for any $q \in \mathbb{N}$, $(b_1, b_2, \ldots, b_q)$, and any $t_1, t_2, \ldots, t_q \in \mathcal{T}$, the following holds:

$$\sum_{k=1}^{q} b_k \mathbb{S}_n(t_k) \xrightarrow{d} \sum_{k=1}^{q} b_k Z(t_k) \tag{2.37}$$

Adopting the same notation as in Part 1, for any $t \in \mathcal{T}$, we find that:

$$\begin{aligned}\sum_{k=1}^{q} b_k \mathbb{S}_n(t) &= \sum_{k=1}^{q} b_k \left\{I_{1,n}(t) + I_{2,n}(t) + I_{3,n}(t)\right\} \\ &= \sum_{k=1}^{q} b_k \left\{I_{1,n}(t) + I_{2,n}(t)\right\} + \sum_{k=1}^{q} b_k I_{3,n}(t).\end{aligned}$$

Once more, we focus on the leading term $I_{3,n}$, as the first two terms have been shown to be negligible in Part 1. Note that:

$$\begin{aligned}\sum_{k=1}^{q} b_k I_{3,n}(t_k) &= \sum_{k=1}^{q} b_k \sum_{i=1}^{n} \widetilde{\Theta}_i(t_k) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{q} b_k \widetilde{\Theta}_i(t_k) \\ &\triangleq \sum_{i=1}^{n} \mathcal{U}_{i,q},\end{aligned}$$

where $\mathcal{U}_{i,q} = \sum_{k=1}^{q} b_k \widetilde{\Theta}_i(t_k)$.

Clearly, $\mathbb{E}(U_{i,q}) = 0$ as $\mathbb{E}\{\widetilde{\Theta}_i(t)\} = 0$ for all i and $t \in \mathcal{T}$. Thus,

$$\begin{aligned}\mathbb{E}\{\mathcal{U}_{i,q}^2\} &= \mathbb{E}\left[\left\{\sum_{k=1}^{q} b_k \widetilde{\Theta}_i(t_k)\right\}\left\{\sum_{l=1}^{q} b_l \widetilde{\Theta}_i(t_l)\right\}\right] \\ &= \sum_{k=1}^{q}\sum_{l=1}^{q} b_k b_l \mathbb{E}\left\{\widetilde{\Theta}_i(t_k)\widetilde{\Theta}_i(t_l)\right\} \\ &= \sum_{k=1}^{q}\sum_{l=1}^{q} b_k b_l C_Z(t_k, t_l) + o(1),\end{aligned}$$

where the last line follows the same calculation as in Part 1.

When

$$\sum_{k=1}^{q}\sum_{l=1}^{q} b_k b_l C_Z(t_k, t_l) = 0,$$

it implies that $\sum_{k=1}^{q} b_k Z(t_k)$ has a degenerate distribution with a point mass at 0 and (2.37) immediately follows. However, if

$$\sum_{k=1}^{q}\sum_{l=1}^{q} b_k b_l C_Z(t_k, t_l) \neq 0,$$

we shall examine the limiting distribution of $\sum_{i=1}^{n} \mathcal{U}_{i,q}$.

Firstly, $\mathcal{U}_{i,q}$ are independent with respect to the index i for all $q \in \mathbb{N}$. The variance of $\mathcal{U}_{i,q}$ is finite. Thus, the proof is concluded if we can verify the Lindeberg's condition.

Notice that $C_{b,q} = \sum_{k=1}^{q} |b_k| > 0$. For any $e > 0$,

$$\begin{aligned}&\sum_{i=1}^{n} \mathbb{E}\left[\mathcal{U}_{i,q}^2 \mathbb{1}_{\{|\mathcal{U}_{i,q}| > e\}}\right] \\ &\leq \sum_{i=1}^{n} \left\{\mathbb{E}(\mathcal{U}_{i,q}^4)\right\}^{1/2} \left\{\mathbb{P}(|\mathcal{U}_{i,q}| > e)\right\}^{1/2}, \end{aligned} \tag{2.38}$$

where the final line is derived using the Cauchy-Schwarz inequality. It's important to note

that:

$$\begin{aligned}
\mathbb{P}(|\mathcal{U}_{i,q}| > e) &= \mathbb{P}\left\{\left|\sum_{k=1}^{q} b_k \widetilde{\Theta}_i(t_k)\right| > e\right\} \\
&\leq \mathbb{P}\left\{\sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i(t)||\sum_{k=1}^{q} b_k| > e\right\} \\
&\leq \mathbb{P}\left\{\sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i(t)| > \frac{e}{C_{b,q}}\right\} \\
&\overset{a.s.}{\leq} c_0 \log^2(n) \exp\left\{-c_1 \frac{e(nh)^{1/2}}{\log(n)}\right\} + 2\mathbb{P}\left\{N_i > c_2 \log(n)\right\}, \qquad (2.39)
\end{aligned}$$

where the upper bound is obtained from Lemma 2.10. As for the other term, let's first rewrite $\mathcal{U}_{i,q}$:

$$\begin{aligned}
\mathcal{U}_{i,q} &= \sum_{k=1}^{q} b_k \widetilde{\Theta}_i(t_k) \\
&= \sum_{k=1}^{q} b_k \left\langle \widetilde{\Theta}_i, K_{t_k} \right\rangle \\
&\leq \|\widetilde{\Theta}_i\| \|K_{t_k}\| \sum_{k=1}^{q} |b_k| \\
&\leq C_K h^{-1/2} C_{b,q} \|\widetilde{\Theta}_i\|.
\end{aligned}$$

Therefore, according to Lemma 2.9,

$$E(\mathcal{U}_{i,q}^4) = E(C_K^4 h^{-2} C_{b,q}^4 \|\widetilde{\Theta}_i\|^4) \leq c_3^2 C_K^4 C_{b,q}^4 (nh)^{-2}. \qquad (2.40)$$

Combining (2.38) through (2.40), and employing the same calculations used for (2.31) in the proof of Theorem 2.2, we obtain:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left[\mathcal{U}_{i,q}^2 \mathbb{1}_{\{|\mathcal{U}_{i,q}|>e\}}\right] = o(1).$$

Finally, by applying the Lindeberg's CLT, we will have (2.37) established. □

2.A.V Proofs of Theorems in Section 2.3.6

This subsection illustrates the validity of the bootstrap procedure discussed. Even though two separated theorems are stated in Section 2.3.6, we will only provide the proofs for the Theorem 2.7. As been pointed out in Figure 1 from Bücher and Kojadinovic (2019), Theorem 2.6 can be understood as a special case of Theorem 2.7, and the proofs for these two theorems are very similar if not identical.

Proof. To begin, we introduce the following notation. Let $\mathbb{S}_{n,b}$ denote the bth bootstrap process obtained from Algorithm 1 for $b = 1, 2, \ldots, B$, where B denotes the bootstrap size. Let $\mathrm{BL}_1\{\mathrm{C}[0,1]\}$ denote the collection of all uniformly Lipschitz functionals. i.e.,

$$\mathrm{BL}_1\{\mathrm{C}[0,1]\} := \{\mathfrak{h} : \mathrm{C}[0,1] \to [-1,1] \mid \\ |\mathfrak{h}(g_1) - \mathfrak{h}(g_2)| \leq \|g_1 - g_2\|_{\sup} \text{ for any } g_1, g_2 \in \mathrm{C}[0,1]\}.$$

It is sufficient to show that the

$$\sup_{\mathfrak{h} \in \mathrm{BL}_1\{\mathrm{C}[0,1]\}} |\mathbb{E}_{\mathcal{D}}\{\mathfrak{h}(\mathbb{S}_{n,1})\} - \mathbb{E}\{\mathfrak{h}(\mathbb{S}_n)\}|, \tag{2.41}$$

where $\mathbb{E}_{\mathcal{D}}$ denotes the conditional expectation given data $\mathcal{D} = \{T_{ij}, Y_{ij} \mid i = 1, 2, \ldots, n, j = 1, 2, \ldots, N_i\}$. For more details, please see Chapter 23 (or more specifically, Theorem 23.7) in van der Vaart (1998). Lemma 3.1 in Bücher and Kojadinovic (2019) implies equation (2.41) is equivalent to prove, for any positive integer $B \geq 2$, as $n \to \infty$,

$$(\mathbb{S}_n, \mathbb{S}_{n,1}, \mathbb{S}_{n,2}, \ldots, \mathbb{S}_{n,B}) \rightsquigarrow (Z, Z_1, Z_2, \ldots, Z_B), \tag{2.42}$$

where $Z_1, Z_2, \ldots, Z_B$ are i.i.d. copies of Z in Theorem 2.4. We have shown $\mathbb{S}_n \rightsquigarrow Z$ in Theorem 2.4. Because Z and Z_i's are Gaussian processes, we only need to show $\mathbb{S}_{n,l} \rightsquigarrow Z_l$ for any $b = 1, 2, \ldots, B$. The main idea of this proof is similar to the proof of Theorem 2.4. That is, we will show the finite dimensional convergence and asymptotic tightness of $\mathbb{S}_{n,b}$.

For bth bootstrap sample, we have

$$\hat{\mu}_b = \arg\min_{f\in\mathcal{H}} \left[\frac{1}{2M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} U_{i,b} \{Y_{ij} - f(T_{ij})\}^2 + \frac{\lambda}{2} J(f,f) \right],$$

and

$$S^b_{M,\lambda}(f) = -\frac{1}{M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} U_{i,b} \{Y_{ij} - f(T_{ij})\} K_{T_{ij}} + W_\lambda f.$$

Because $\mathbb{E}U_{i,b} = 1$ and these bootstrap weights are independent of the random components in model (2.1). These facts imply we can obtain an variant of Lemma 2.7, i.e. $DS^b_\lambda(f) = \mathrm{id}$ for any $f \in \mathcal{H}$ and obtain the same results in Theorem 2.5 and 2.1.

Besides that, direct calculations yield that

$$\begin{aligned} S_{n,b}(t) &= (nh)^{1/2} \{\hat{\mu}_b(t) - \hat{\mu}(t)\} \\ &= (nh)^{1/2} \left\{\hat{\mu}_b - \mu + S^b_{M,\lambda}(\mu)(t)\right\} \\ &\quad - (nh)^{1/2} \{\hat{\mu} - \mu + S_{M,\lambda}(\mu)(t)\} \\ &\quad + (nh)^{1/2} \left\{S_{M,\lambda}(\mu)(t) - S^b_{M,\lambda}(\mu)(t)\right\} \\ &\triangleq H_{1,b}(t) + H_2(t) + H_{3,b}(t), \end{aligned}$$

where

$$H_{3,b} = \frac{(nh)^{1/2}}{M} \sum_{i=1}^{n} \sum_{j=1}^{N_i} (U_{i,b} - 1)\{Y_{ij} - \mu(T_{ij})\} K_{T_{ij}}.$$

Lemma 2.1 implies $\|\hat{\mu} - \mu + S_{M,\lambda}(\mu)\| = O_p(a_n)$ and we have further assumed $a_n = o_p(n^{-1/2})$. This assumption is one of the crucial conditions for Theorem 2.2 and 2.4. Because

$$\begin{aligned} \sup_{t\in\mathcal{T}} |H_2(t)| &= \sup_{t\in\mathcal{T}} (nh)^{1/2} |\langle K_t, \hat{\mu} - \mu + S_{M,\lambda}(\mu)\rangle| \\ &\le (nh)^{1/2} \|K_t\| \|\hat{\mu} - \mu - S_{M,\lambda}(\mu)\| \\ &= o_p(1). \end{aligned}$$

The same result holds for $H_{1,b}(t)$ as $\|\hat{\mu}_b - \mu + S^b_{M,\lambda}(\mu)\| = O_p(a_n)$. Therefore, we need

to consider the leading term H_3^b, and the validity of equation (2.42) relies on the finite dimensional convergence of $\mathbb{H} = (H_{3,1}, H_{3,2}, \ldots, H_{3,B})^\top$ and asymptotic tightness of $H_{3,b}$.

To begin, for arbitrary $R \in \mathbb{N}$, and $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_R \in \mathbb{R}^B$, we want to show that, for sufficient large n,

$$\sum_{r=1}^{R} \boldsymbol{c}_r^\top \mathbb{H}(t_r) \to \sum_{r=1}^{R} \boldsymbol{c}_r^\top \boldsymbol{Z}(t_r), \tag{2.43}$$

where $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_B)^\top$. Note that,

$$\begin{aligned}
\sum_{r=1}^{R} \boldsymbol{c}_r \mathbb{H}(t_r) &= \sum_{r=1}^{R} \sum_{b=1}^{B} c_{rb} H_{3,b}(t_r) \\
&= \sum_{i=1}^{n} \sum_{b=1}^{B} \sum_{r=1}^{R} \sum_{j=1}^{N_i} c_{rb} \frac{(nh)^{1/2}}{M} (U_{i,b} - 1) \left\{ Y_{ij} - \mu(T_{ij}) \right\} K_{T_{ij}}(t_r) = \sum_{i=1}^{n} \mathfrak{U}_i(t_r),
\end{aligned}$$

where

$$\begin{aligned}
\mathfrak{U}_i &= \sum_{b=1}^{B} \sum_{r=1}^{R} \sum_{j=1}^{N_i} c_{rb} \frac{(nh)^{1/2}}{M} (U_{i,b} - 1) \left\{ Y_{ij} - \mu(T_{ij}) \right\} K_{T_{ij}}(t_r) \\
&= \sum_{b=1}^{B} \sum_{r=1}^{R} c_{rb} (U_{i,b} - 1) \widetilde{\Theta}_i(t_r),
\end{aligned}$$

Here, $\widetilde{\Theta}_i(t_r)$ is defined in equation (2.45), which includes all random elements in model (2.1). Notice that $\mathbb{E}U_{i,b} = 1$ and $\mathbb{E}\left\{(U_{i,b} - 1)^2\right\} = 1$. Moreover, $U_{i,b}$'s are independent of all random elements in model (2.1), we can easily conclude $\mathbb{E}\mathfrak{U}_i = 0$. As for the second moment of $\mathfrak{U}_i$, we have

$$\begin{aligned}
\mathbb{E}\mathfrak{U}_i^2 &= \sum_{b,b'} \sum_{r,r'} c_{rb} c_{r'b'} \mathbb{E} \left\{ (U_{i,b} - 1)(U_{i,b'} - 1) \widetilde{\Theta}_i(t_r) \widetilde{\Theta}_i(t_{r'}) \right\} \\
&= \sum_{b=1}^{B} \sum_{r,r'} c_{rb} c_{r'b} \mathbb{E} \left\{ (U_{i,b} - 1)^2 \right\} \mathbb{E} \left\{ \widetilde{\Theta}_i(t_r) \widetilde{\Theta}_i(t_{r'}) \right\} \\
&= \sum_{b=1}^{B} \sum_{r,r'} c_{rb} c_{r'b} \mathbb{E} \left\{ \widetilde{\Theta}_i(t_r) \widetilde{\Theta}_i(t_{r'}) \right\}
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^{n}\mathfrak{U}_i\Big) &= \sum_{i=1}^{n}\mathbb{E}\mathfrak{U}_i^2 \\ &= \sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{r,r'} c_{rb}c_{r'b}\mathbb{E}\left\{\widetilde{\Theta}_i(t_r)\widetilde{\Theta}_i(t'_r)\right\} \\ &= C_Z(t_r,t'_r)+o_p(1),\end{aligned}$$

where the last line comes from calculating the covariance kernel in Theorem 2.4 as $\sum_{i=1}^{n}\mathbb{E}\mathfrak{U}_i^2 = \mathbb{E}\{I_{3,n}(t)I_{3,n}(s)\}$. The finite dimensional distributions can be proved by using Lindeberg's condition. Note that, $|U_{i,b}-1|\leq\sqrt{2}$ almost surely. Therefore, for arbitrary fixed $B\in\mathbb{N}$, the Lindeberg condition naturally holds as $|\mathfrak{U}_i|\leq 2B|\sum_{r=1}^{R}c_{rb}\widetilde{\Theta}_i(t_r)|$.

As for the asymptotic tightness of $H_{3,b}(t)$, note that, the asymptotic tightness proved for $I_{3,n}(t)$ in Theorem 2.4 relies on equation (2.33). Because $|U_{i,b}-1|\leq\sqrt{2}$, we have $H_{3,b}(t)\leq\sqrt{2}I_{3,n}(t)$, i.e. the ratio between $H_{3,b}(t)$ and $\sqrt{2}I_{3,n}(t)$ is bounded above by a constant almost surely. This fact implies the asymptotic tightness of $H_{3,b}$ for all $b=1,2,\ldots,B$.

Note that,

$$\frac{(nh)^{1/2}}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij}-\mu(T_{ij})\}K_{T_{ij}} = I_{3,n}.$$

The asymptotic tightness is proved in Theorem 2.4. By Theorem 1.5.4 in van der Vaart and Wellner (1996), Lemma 3.1 in Bücher and Kojadinovic (2019), and Theorem 2.4, equation (2.43) is valid for any positive integer $B\geq 2$, as $n\to\infty$. This conclude the proof.

□

2.A.VI Technical Lemmas

This subsection contains several technical lemmas that are used throughout all the proofs of the main theorems. An immediate result, following from Lemma 2.1 and 2.2, establishes a connection between the supremum norm and the norm induced by the inner product

defined in (2.9). It's important to note that

$$|f(t)|^2 = \langle K_t, f\rangle^2 \leq \|K_t\|^2\|f\|^2$$

and $\|K_t\|^2 \leq C_K^2 h^{-1}$ as shown in Lemma 2.2. Thus, we have the following lemma:

Lemma 2.6. *For any $f \in \mathcal{H}$, there exists a constant C_K such that*

$$|f(t)| \leq C_K h^{-1/2}\|f\|$$

for any $t \in \mathcal{T}$, where C_K, as defined in Lemma 2.2, does not depend on the choice of f or t. In other words, $\|f\|_{\sup} \leq C_K h^{-1/2}\|f\|$.

The next lemma is a direct result of computing the Frechét derivatives of $\ell(f, \lambda)$.

Lemma 2.7. *$DS_\lambda(f) = \mathrm{id}$ where* id *is the identity operator on $\mathcal{H}$ for any $f \in \mathcal{H}$.*

Proof. Recall that $D\ell(f,\lambda)\Delta g = \langle S_{M,\lambda}(f), \Delta g\rangle$ where

$$S_{M,\lambda}(f) = -\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}K_{T_{ij}} + W_\lambda f \in \mathcal{H}.$$

Let's define $\mathbb{E}\{S_{M,\lambda}(f)\} \triangleq S_\lambda(f)$ and $\mathbb{E}\{S_M(f)\} \triangleq S(f)$. It can be shown that

$$\begin{aligned}
\mathbb{E}\{S_M(f)\} &= \mathbb{E}\left\{-\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}K_{T_{ij}}\right\} \\
&= \mathbb{E}\left[-\mathbb{E}\left\{\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - f(T_{ij})\}K_{T_{ij}}\,\middle|\, N_i, \text{ for } i = 1, 2, \ldots n\right\}\right] \\
&= \mathbb{E}\left[-\{Y_{11} - f(T_{11})\}K_{T_{11}}\right] \\
&= S(f). \qquad (2.44)
\end{aligned}$$

Notice that $S_\lambda(f) = S(f) + W_\lambda f$. Therefore, we have $DS_\lambda(f) = DS(f) + W_\lambda$. Additionally,

the second order Fréchet derivative w.r.t f is

$$\begin{aligned} D^2\ell(f,\lambda)\Delta g\Delta h &= \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\langle K_{T_{ij}},\Delta g\right\rangle\left\langle K_{T_{ij}},\Delta h\right\rangle + \langle W_\lambda \Delta g,\Delta h\rangle \\ &\triangleq DS_{M,\lambda}(f)\Delta g\Delta h \\ &= \text{const w.r.t. } f. \end{aligned}$$

Therefore,

$$DS_{M,\lambda}(f)g = \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{N_i} g(T_{ij})K_{T_{ij}} + W_\lambda g.$$

Similar to the previous calculation, it can be shown that

$$DS_\lambda(f)g = \mathbb{E}[g(T)K_T] + W_\lambda g.$$

Therefore, for any $f,g,h\in\mathcal{H}$, we can obtain:

$$\begin{aligned} \langle DS_\lambda(f)g,h\rangle &= \mathbb{E}[\langle K_T,g\rangle\langle K_T,h\rangle] + \langle W_\lambda g,h\rangle \\ &= \mathbb{E}[g(T)h(T)] + \lambda J(g,h) \\ &= \langle g,h\rangle. \end{aligned}$$

This concludes that $DS_\lambda(f) = \text{id}$ for any $f\in\mathcal{H}$. □

Lemma 2.8. *Suppose that Assumption 2.1, Assumption 2.4 and Assumption 2.5 hold, then the following functions defined in (2.19) can be expressed as*

$$\begin{aligned} \mathcal{G}_1(s,t) &= \sum_{i,j}\frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)}I_1^{ij}, \\ \mathcal{G}_2(s,t) &= \sum_{i,j}\frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)}I_2^{ij}, \\ \mathcal{G}_3(s,t) &= \sum_{i}\frac{h_i(t)h_i(s)}{(1+\lambda\gamma_i)^2}. \end{aligned}$$

Here, $\{h_v\}_{v\in\mathbb{N}_+}$ of $\mathcal{H}$ are basis functions defined in Assumption 2.5. In particular, if $m \geq 1$ for $\mathcal{H} = H^m$ and Assumption 2.7 holds, then

$$\Delta_{\mathcal{G}_1}(t,s) = |\mathcal{G}_1(t,t) + \mathcal{G}_1(s,s) - 2\mathcal{G}_1(t,s)| \leq C_{\mathcal{G}_1} h^{\tau-1}|t-s|,$$

$$\Delta_{\mathcal{G}_2}(t,s) = |\mathcal{G}_2(t,t) + \mathcal{G}_2(s,s) - 2\mathcal{G}_2(t,s)| \leq C_{\mathcal{G}_2} h^{\tau-1}|t-s|,$$

$$\Delta_{\mathcal{G}_3}(t,s) = |\mathcal{G}_3(t,t) + \mathcal{G}_3(s,s) - 2\mathcal{G}_3(t,s)| \leq C_{\mathcal{G}_3} h^{-1}|t-s|^2.$$

Proof. Note that Lemma 2.1 implies that $K(t,\cdot) = \sum_i h_i(t)h_i(\cdot)/(1+\lambda\gamma_i)$. Let's start by considering $\mathcal{G}_3(s,t)$ first.

$$\begin{aligned}
\mathcal{G}_3(s,t) &= \int_{u\in\mathcal{T}} \left\{\sum_i \frac{h_i(t)}{1+\lambda\gamma_i} h_i(u)\right\}\left\{\sum_j \frac{h_j(s)}{1+\lambda\gamma_j} h_j(u)\right\}\pi(u)du \\
&= \sum_{i,j} \frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} \int_{u\in\mathcal{T}} h_i(u)h_j(u)\pi(u)du \\
&= \sum_{i,j} \frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} V(h_i,h_j) \\
&= \sum_i \frac{h_i(t)h_i(s)}{(1+\lambda\gamma_i)^2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\Delta_{\mathcal{G}_3}(t,s) &= |\mathcal{G}_3(t,t) + \mathcal{G}_3(s,s) - 2\mathcal{G}_3(t,s)| \\
&= \sum_i \frac{\{h_i(t) - h_i(s)\}^2}{(1+\lambda\gamma_i)^2} \\
&= \sum_i \frac{\{h_i^{(1)}(x)\}^2}{(1+\lambda\gamma_k)^2}(t-s)^2 \quad \text{for some } x \text{ between } t \text{ and } s \\
&\leq C_{\mathcal{G}_3} h^{-1}|t-s|^2,
\end{aligned}$$

where the second-to-last line uses the Mean Value Theorem, and the last line relies on the

boundedness of the first derivative of h_i, for all i, as $h_i \in \mathcal{H} = H^m$. Furthermore,

$$\begin{aligned}
\mathcal{G}_1(s,t) &= \int_{(u,v)\in\mathcal{T}^2} c(u,v) \left\{ \sum_i \frac{h_i(t)}{1+\lambda\gamma_i} h_i(u) \right\} \left\{ \sum_j \frac{h_j(s)}{1+\lambda\gamma_j} h_j(u) \right\} \pi(u) du \\
&= \sum_{i,j} \frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} \int_{(u,v)\in\mathcal{T}^2} c(u,v) h_i(u) h_j(v) \pi(u)\pi(v) du dv \\
&= \sum_{i,j} \frac{h_i(t)h_j(s)}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} I_1^{ij}.
\end{aligned}$$

Note that

$$|\mathcal{G}_1(t,t) + \mathcal{G}_1(s,s) - 2\mathcal{G}_1(t,s)| \le |\mathcal{G}_1(t,t) - \mathcal{G}_1(t,s)| + |\mathcal{G}_1(s,s) - \mathcal{G}_1(s,t)|.$$

Moreover,

$$\begin{aligned}
|\mathcal{G}_1(t,t) - \mathcal{G}_1(t,s)| &\le \sum_{i,j} \frac{|h_i(t)\{h_j(s) - h_j(t)\} I_1^{ij}|}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} \\
&= \sum_{i,j} \frac{|h_i(t) h_j^{(1)}(x) I_1^{ij}|}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} |t-s| \\
&\le c_1 \sum_{i,j} \frac{\int_{(u,v)\in\mathcal{T}^2} c(u,v) h_i(u) h_j(v) du dv}{(1+\lambda\gamma_i)(1+\lambda\gamma_j)} |t-s| \\
&\le c_2 |t-s| \left\{ \sum_i \frac{i^{-\tau}}{(1+\lambda\gamma_i)^2} \right\}^{1/2} \left\{ \sum_j \frac{j^{-\tau}}{(1+\lambda\gamma_j)^2} \right\}^{1/2} \\
&\le C_{\mathcal{G}_1} |t-s| h^{\tau-1}.
\end{aligned}$$

Thus, we have $\Delta_{\mathcal{G}_1}(t,s) \le C_{\mathcal{G}_1}|t-s|h^{\tau-1}$. The same argument applies to $\Delta_{\mathcal{G}_2}(t,s)$. □

The following technical lemma aims to establish the properties of

$$\widetilde{\Theta}_i(t) = \frac{(nh)^{1/2}}{M} \Theta_i(t), \tag{2.45}$$

where $\Theta_i(t) = \sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)$ for $i = 1, 2, \ldots, n$, and μ is the true mean function. These results facilitate the derivation of point-wise normality and weak convergence.

Lemma 2.9. *Suppose Assumptions 2.1 - 2.6 hold, then, for* $i = 1, 2, \ldots, n$,

$$(i)\ \sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i(t)| \overset{a.s}{\leq} c_1\frac{\log^2(n)}{n^{1/2}h} \text{ and } (ii)\ \mathbb{E}\left(\|\widetilde{\Theta}_i\|^4\right) \leq c_2 n^{-2}.$$

Proof. Let $c_1 > C_\varepsilon^{-1}$, then

$$\begin{aligned}\sum_n \Pr(|\varepsilon_n| \geq c_1\log(n)) &\leq \sum_n \frac{\mathbb{E}\{\exp(C_\varepsilon|\varepsilon|)\}}{n^{c_1 C_\varepsilon}} \\ &= E\{\exp(C_\varepsilon|\varepsilon|)\}\sum_n n^{-c_1 C_\varepsilon} \\ &< \infty,\end{aligned}$$

where the final line arises from Assumptions 2.6. Consequently, we have $|\varepsilon| \leq (c_1 \log n)$ almost surely. For the random function X and the number of observations per subject N, similar arguments apply, leading to $\|X\| \leq (c_2 \log n)$ and $|N| \leq (c_3 \log n)$ almost surely, where $c_2 > C_X^{-1}$ and $c_3 > C_N^{-1}$.

Note that

$$\sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i(t)| = \frac{(nh)^{1/2}}{M}\sup_{t\in\mathcal{T}}|\Theta(t)|.$$

As a result, we can obtain the following:

$$\begin{aligned}\sup_{t\in\mathcal{T}}|\Theta_i(t)| &= \sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\{Y_{ij} - \mu(T_{ij})\}K_{T_{ij}}(t)\}\right| \\ &= \sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\{Y_{ij} - X_i(T_{ij}) + X_i(T_{ij}) - \mu(T_{ij})\}K_{T_{ij}}(t)\}\right| \\ &\leq \sup_{t\in\mathcal{T}}\sum_{j=1}^{N_i}|\varepsilon_{ij}K_{T_{ij}}(t)| + \sup_{t\in\mathcal{T}}\sum_{j=1}^{N_i}\left|\{X_i(T_{ij}) - \mu(T_{ij})\}K_{T_{ij}}(t)\right|. \qquad (2.46)\end{aligned}$$

The first term on the RHS of (2.46) can be computed as follows:

$$\begin{aligned}
\sup_{t\in\mathcal{T}}\sum_{j=1}^{N_i}|\varepsilon_{ij}K_{T_{ij}}(t)| &\leq (\sum_{j=1}^{N_i}|\varepsilon_{ij}|)\sup_{t\in\mathcal{T}}|K_T(t)| \\
&\leq C_K h^{-1/2}\|K_t\|\sum_{j=1}^{N_i}|\varepsilon_{ij}| \text{ by Lemma 2.6} \\
&\overset{a.s.}{\leq} C_h\left\{c_1\log(n)\right\}\left\{c_3\log(n)\right\}h^{-1} \\
&\leq c_4 h^{-1}\log^2(n). \qquad (2.47)
\end{aligned}$$

The second term on the RHS of (2.46) can be analyzed using a similar argument. Notice that

$$|X_i(T_{ij})-\mu(T_{ij})| = \left\langle X_i-\mu, K_{T_{ij}}\right\rangle \leq \|X_i-\mu\|\|K_{T_{ij}}\| \overset{a.s.}{<} c_2 C_K h^{-1/2}\log(n).$$

Additionally,

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\sum_{j=1}^{N_i}|\{X_i(T_{ij})-\mu(T_{ij})\}K_{T_{ij}}(t)| \\
&\leq \sum_{j=1}^{N_i}\left|\{X_i(T_{ij})-\mu(T_{ij})\}\right|\sup_{t\in\mathcal{T}}|K_{T_{ij}}(t)| \\
&\leq C_K h^{-1/2}\|K_t\|\sum_{j=1}^{N_i}\left|\{X_i(T_{ij})-\mu(T_{ij})\}\right| \text{ by Lemma 2.6} \\
&\leq C_K^2 h^{-1}\Big(\sum_{j=1}^{N_i}\left\langle X_i-\mu, K_{T_{ij}}\right\rangle\Big) \\
&\leq C_K^2 h^{-1} N_i\|X_i-\mu\|\|K_t\| \\
&\overset{a.s.}{\leq} c_2 C_K^3 h^{-3/2}\log(n)N_i \\
&\overset{a.s.}{\leq} c_5 h^{-3/2}\log^2(n). \qquad (2.48)
\end{aligned}$$

Because (2.48) dominates (2.47) when $h = o(1)$, for sufficient large n, we can obtain

$$(2.46) \overset{a.s.}{\leq} c_6 \frac{\log^2 n}{h^{3/2}}.$$

Immediately, it can be shown that

$$\sup_{t\in\mathcal{T}} |\widetilde{\Theta}(t)| \overset{a.s.}{\leq} c_6 \frac{(nh)^{1/2}}{n} \frac{\log^2(n)}{h^{3/2}} = c_6 \frac{\log^2(n)}{n^{1/2}h}.$$

As for the second part of the lemma, consider that:

$$\begin{aligned}
&\left\langle \widetilde{\Theta}_i, \widetilde{\Theta}_i \right\rangle \\
&= \frac{nh}{M^2} \sum_{j=1}^{N_i} \sum_{k=1}^{N_i} \left\{ Y_{ij} - \mu(T_{ij}) \right\} \left\{ Y_{ik} - \mu(T_{ik}) \right\} \left\langle K_{T_{ij}}, K_{T_{ik}} \right\rangle \\
&\leq \frac{1}{n} \sum_{j,k=1}^{N_i} \left\{ Y_{ij} - \mu(T_{ij}) \right\} \left\{ Y_{ik} - \mu(T_{ik}) \right\} \\
&= \frac{1}{n} \sum_{j,k=1}^{N_i} \left\{ \varepsilon_{ij} + \delta_{ij} \right\} \left\{ \varepsilon_{ik} + \delta_{ik} \right\},
\end{aligned}$$

where $\delta_{ij} = X_i(T_{ij}) - \mu(T_{ij})$. We can see that:

$$(i)\ \mathbb{E}(\delta_{ij}) = 0, (ii)\ \mathbb{E}(\delta_{ij}^2) = I_{t,t}, (iii)\ \mathbb{E}(\delta_{ij}\delta_{ik}) = I_{t,s}, j \neq k,$$

where $I_{t,s}, I_{t,t}$ are defined as in (2.19) for all $j, k = 1, 2, \ldots, N_i$. Thus,

$$\begin{aligned}
\mathbb{E}\left(\|\widetilde{\Theta}_i\|^4 \right) &= \mathbb{E}\left(\left\langle \widetilde{\Theta}_i, \widetilde{\Theta}_i \right\rangle^2 \right) \\
&\leq \frac{1}{n^2} \mathbb{E}\left[\sum_{j,k,u,v=1}^{N_i} (\varepsilon_{ij} - \delta_{ij})(\varepsilon_{ik} - \delta_{ik}) \left\{ \varepsilon_{iu} - \delta_{iu} \right\} \left\{ \varepsilon_{iv} - \delta_{iv} \right\} \right] \\
&= \frac{1}{n^2} \mathbb{E}\left\{ N_i \varepsilon_{i1}^4 + (N_i^2 - N_i) \varepsilon_{i1}^2 \delta_{i1} \delta_{i2} + N_i \varepsilon_{i1}^2 \delta_{i1}^2 + N_i \delta_{i1}^4 \right\}. \qquad (2.49)
\end{aligned}$$

Note that there always exists an ϵ such that $\epsilon^4 \leq \exp(C_\varepsilon \epsilon)$. Thus,

$$\begin{aligned}\mathbb{E}(\varepsilon_{i1}^4) &= \int_{\varepsilon \leq \epsilon} \varepsilon^4 h_\varepsilon(\varepsilon) d\varepsilon + \int_{\varepsilon > \epsilon} \varepsilon^4 h_\varepsilon(\varepsilon) d\varepsilon \\ &\leq \epsilon^4 + \mathbb{E}\left\{\exp(C_\varepsilon |\varepsilon|)\right\} \\ &= O(1) \text{ by Assumption 2.6,}\end{aligned}$$

where h_ε is the density function of ε.

Furthermore,

$$\begin{aligned}\mathbb{E}[\{X_i(T_{ij}) - \mu(T_{ij})\}^4] &\leq C(\mathbb{E}[\{X_i(T_{ij}) - \mu(T_{ij})\}^2])^2 \\ &\leq C I_{t,t}^2 \\ &= O(1),\end{aligned}$$

where the last line uses the the marginal density function for T, and $I_{t,t}$ is bounded from above by Assumption 2.1 and Assumption 2.4. Therefore, combining the above two results, and Assumption 2.1 through 2.4, we can conclude that

$$(2.49) = O(1/n^2).$$

□

The final technical lemma calculates the decay rate for the supremum of $\widetilde{\Theta}_i$. A general statement without restrictions on h is given below. This lemma is applied when one needs to verify the Lindeberg's condition.

Lemma 2.10. *Suppose Assumptions 2.1 - 2.6 hold, then for any $e > 0$ and $i = 1, 2, \ldots, n$, we have*

$$\mathbb{P}\left\{\sup_{t \in \mathcal{T}} |\widetilde{\Theta}_i| > e\right\} \overset{a.s.}{\leq} c_0 \log^2(n) \exp\left\{-c_1 \frac{e(nh)^{1/2}}{\log(n)}\right\} + 2\mathbb{P}\left\{N_i > c_2 \log(n)\right\},$$

where $c_2 > C_N^{-1} + 1$ for C_N in Assumption 2.6.

Proof. This proof is a repetitive one that involves the repeated use of the Borel-Cantelli Lemma and the Markov inequality.

Notice that we first decompose $\widetilde{\Theta}$ into two parts, which is the same as in (2.46). Before delving into the detailed calculations, we first establish that for any random variable T whose marginal distribution is bounded between 0 and infinity, as stated in Assumption 2.1, the kernel function $K_T(t)$ is bounded on the order of $O(h^{-1})$. Hence,

$$\sup_{t\in\mathcal{T}}|K_T(t)| = \|K_t\|_{\sup} \leq C_K h^{-1/2}\|K_t\| \leq C_K^2 h^{-1},$$

where the final step uses Lemma 2.6. Let $X_{ij} \triangleq X_i(T_{ij})$, we have

$$\begin{aligned}
&\mathbb{P}(\sup_{t\in\mathcal{T}}|\widetilde{\Theta}_i(t)| > e) \\
&= \mathbb{P}\left[\sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\{Y_{ij}-\mu(T_{ij})\}K_{T_{ij}}(t)\right| > \frac{eM}{(nh)^{1/2}}\right] \\
&\leq \mathbb{P}\left[\sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\{Y_{ij}-X_{ij}+X_{ij}-\mu(T_{ij})\}K_{T_{ij}}(t)\right| \geq \frac{en}{(nh)^{1/2}}\right] \\
&\leq \mathbb{P}\left[\sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\varepsilon_{ij}K_{T_{ij}}(t)\right| \geq \frac{en}{2(nh)^{1/2}}\right] \qquad (2.50) \\
&+ \mathbb{P}\left[\sup_{t\in\mathcal{T}}\left|\sum_{j=1}^{N_i}\{X_{ij}-\mu(T_{ij})\}K_{T_{ij}}(t)\right| \geq \frac{en}{2(nh)^{1/2}}\right]. \qquad (2.51)
\end{aligned}$$

Note that both the error term ε and X decay at an exponential rate, almost surely, according to Assumption 2.6.

Specifically, for some $c_1^{'} > C_\varepsilon^{-1}$ and $c_2^{'} > C_N^{-1}$,

$$|\varepsilon| \overset{a.s.}{\leq} c_1^{'}\log(n) \text{ and } \|X\| \overset{a.s.}{\leq} c_2^{'}\log(n).$$

The detailed computations are outlined in Lemma 2.9. For conciseness, this part specif-

ically elaborates on the detailed analysis of (2.50).

$$
\begin{aligned}
(2.50) &\leq \mathbb{P}\left\{\sup_{t\in\mathcal{T}}|K_T(t)|\Big(\sum_{j=1}^{N_i}|\varepsilon_{ij}|\Big) \geq \frac{en^{1/2}}{2h^{1/2}}\right\} \\
&\leq \mathbb{P}\left\{C_K^2 h^{-1}\sum_{j=1}^{N_i}|\varepsilon_{ij}| \geq \frac{en^{1/2}}{2h^{1/2}}\right\} \\
&\leq \mathbb{P}\left\{\sum_{j=1}^{N_i}|\varepsilon_{ij}| \geq \frac{e(nh)^{1/2}}{2C_K^2} \mid N_i \leq c_1\log(n)\right\}\mathbb{P}\left\{N_i \leq c_1\log(n)\right\} \\
&+ \mathbb{P}\left\{\sum_{j=1}^{N_i}|\varepsilon_{ij}| \geq \frac{e(nh)^{1/2}}{2C_K^2} \mid N_i > c_1\log(n)\right\}\mathbb{P}\left\{N_i > c_1\log(n)\right\} \\
&\leq \mathbb{P}\left\{\sum_{j=1}^{N_i}|\varepsilon_{ij}| \geq \frac{e(nh)^{1/2}}{2C_K^2} \mid N_i \leq c_1\log(n)\right\} + \mathbb{P}\left\{N_i > c_1\log(n)\right\} \\
&\triangleq p_1 + p_2, \qquad (2.52)
\end{aligned}
$$

where $c_1 > C_N^{-1} + 1$ with C_N being the constant defined in Assumption 2.6. We will first delve into the analysis of p_1.

$$
\begin{aligned}
&\mathbb{P}\left\{\sum_{j=1}^{N_i}|\varepsilon_{ij}| \geq \frac{e(nh)^{1/2}}{2C_K^2} \mid N_i \leq c_1\log(n)\right\} \\
&\leq \sum_{j=1}^{N_i}\mathbb{P}\left\{|\varepsilon_{ij}| \geq \frac{e(nh)^{1/2}}{2C_K^2 N_i} \mid N_i \leq c_1\log(n)\right\} \\
&\leq N_i \sum_{k=1}^{\lceil c_1\log(n)\rceil}\mathbb{P}\left\{N_i = k \mid N_i \leq c_1\log(n)\right\}\mathbb{P}\left\{|\varepsilon_{i1}| \geq \frac{e(nh)^{1/2}}{2C_K^2 c_1\log(n)} \;\Big|\; N_i \leq c_1\log(n)\right\} \\
&\overset{a.s.}{\leq} 2c_1\log^2(n)\mathbb{E}\left\{\exp(C_\varepsilon|\varepsilon_{ij}|)\right\}\exp\left\{-\frac{eC_\varepsilon(nh)^{1/2}}{2C_K^2 c_1\log(n)}\right\}.
\end{aligned}
$$

The final line uses Assumption 2.6 and the Markov inequality. In this context, $\lceil\cdot\rceil$ represents the smallest integer that is great than or equal to $c_1\log(n)$. Thus, we establish the following

result:

$$(2.52) \overset{a.s.}{\leq} c_2 \log^2(n) \exp\left\{-c_3 \frac{e(nh)^{1/2}}{\log(n)}\right\} + \mathbb{P}\left\{N_i > c_1 \log(n)\right\}.$$

The same argument is applicable to (2.51). The only difference lies in the calculation process.

$$\Pr\left\{|X_i(T_{ij} - \mu(T_{ij})| \geq e(nh)^{1/2}/2C_K^2 c_1 \log(n) \mid N_i \leq c_1 \log(n)\right\}.$$

Note that $|X_i(T_{ij}) - \mu(T_{ij})| \leq \|X - \mu\|_{\sup} \leq h^{-1/2}\|X - \mu\|$. Thus,

$$\begin{aligned}(2.51) &\overset{a.s.}{\leq} 2c_1 \log^2(n) \mathbb{E}\left\{\exp(C_\varepsilon |\varepsilon_{ij}|)\right\} \exp\left\{-\frac{eC_\varepsilon n^{1/2} h}{2C_K^2 c_1 \log(n)}\right\} \\ &\leq c_4 \log^2(n) \exp\left\{-c_5 \frac{en^{1/2}h}{\log(n)}\right\} + \mathbb{P}\left\{N_i > c_1 \log(n)\right\}.\end{aligned}$$

It's important to note that $nh \geq nh^2$ given that $h = o(1)$. Therefore, in the asymptotic sense, we obtain the following results:

$$\mathbb{P}(\sup_{t \in \mathcal{T}} |\widetilde{\Theta}_i(t)| > e) \overset{a.s.}{\leq} c_7 \log^2(n) \exp\left\{-c_8 \frac{(nh)^{1/2}}{\log(n)}\right\} + 2\mathbb{P}\left\{N_i > c_1 \log(n)\right\}.$$

□

2.B Additional Simulation Results

2.B.I Out-of-sample Performance

In this section, the out-of-sample performance (OOSP) of the proposed method, compared with PACE from Yao et al. (2005b), is presented in Table 2.3. Note that the OOSP is well-defined only if the data from the two groups are collected over a common grid. Otherwise, $Y_{1ij} - Y_{2i'j'}$ may not be comparable, as responses could be collected at different observation times, or one subject may have more observations than another. Consequently, we present the OOSP results only for the second group. Specifically, let $\mathcal{D}$ denote the entire dataset,

from which 10% is randomly selected as the test set. Define

$$\text{OOSP} = \sum_{i=1}^{n'} \sum_{j=1}^{N_i} \left\{ Y'_{ij} - \hat{\mu}(T'_{ij}) \right\}^2,$$

where $\hat{\mu}$ is derived from the training set, and $\{(Y'_{ij}, T'_{ij})\}$ is taken from the test set. The results are summarized in Table 2.3.

Table 2.3: Prediction errors and standard error (in parenthesis) of OOSPs across 1000 Monte Carlo runs for group 2 with $\mu(t) = (2t - 0.3)^3 + 0.5t$.

Setting	$N_{\max}$	$n_1 = 200, n_2 = 100$		$n_1 = 200, n_2 = 200$		$n_1 = 200, n_2 = 400$	
		SS	PACE	SS	PACE	SS	PACE
c.1	6	0.770(0.084)	0.771(0.085)	0.741(0.039)	0.741(0.039)	0.743(0.020)	0.743(0.020)
	10	0.753(0.082)	0.753(0.082)	0.748(0.041)	0.748(0.041)	0.736(0.019)	0.736(0.019)
	14	0.748(0.068)	0.748(0.068)	0.736(0.034)	0.736(0.034)	0.737(0.017)	0.737(0.017)
	18	0.743(0.073)	0.743(0.073)	0.739(0.034)	0.739(0.034)	0.736(0.016)	0.736(0.016)
c.2	6	0.694(0.064)	0.695(0.065)	0.676(0.033)	0.676(0.033)	0.681(0.015)	0.681(0.015)
	10	0.707(0.063)	0.707(0.063)	0.667(0.027)	0.666(0.027)	0.680(0.014)	0.680(0.014)
	14	0.678(0.055)	0.679(0.055)	0.679(0.026)	0.679(0.026)	0.677(0.014)	0.677(0.014)
	18	0.685(0.052)	0.685(0.053)	0.681(0.024)	0.681(0.024)	0.686(0.013)	0.686(0.013)
c.3	6	0.657(0.052)	0.660(0.053)	0.644(0.024)	0.645(0.024)	0.648(0.013)	0.649(0.013)
	10	0.651(0.046)	0.652(0.046)	0.657(0.023)	0.657(0.023)	0.645(0.011)	0.646(0.011)
	14	0.659(0.043)	0.660(0.044)	0.652(0.023)	0.653(0.023)	0.650(0.011)	0.650(0.011)
	18	0.660(0.040)	0.661(0.041)	0.650(0.021)	0.650(0.021)	0.647(0.010)	0.647(0.010)

2.B.II Additional Global Testing Results

In this section, we provide additional global testing results of the proposed method and the competing methods for $N_{\max} \in \{6, 14, 18\}$. These results are shown in Figures 2.9-2.17. Corresponding numerical tables are presented in Tables 2.4 - 2.6.

Table 2.4: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$

Setting	Test	$n_1 = 200, n_2 = 100$			$n_1 = 200, n_2 = 200$			$n_1 = 200, n_2 = 400$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
c.1	**SS**	0.079	0.314	0.924	0.064	0.371	0.944	0.047	0.350	0.941
	SH	0.230	0.438	0.889	0.244	0.521	0.921	0.278	0.510	0.915
	pffr	0.340	0.590	0.927	0.306	0.623	0.949	0.325	0.596	0.943
c.2	**SS**	0.072	0.360	0.930	0.043	0.367	0.950	0.053	0.371	0.945
	SH	0.218	0.452	0.906	0.236	0.513	0.943	0.257	0.547	0.944
	pffr	0.313	0.614	0.940	0.309	0.606	0.964	0.343	0.625	0.953
c.3	**SS**	0.078	0.302	0.852	0.056	0.237	0.797	0.047	0.229	0.729
	SH	0.229	0.441	0.907	0.149	0.419	0.905	0.140	0.430	0.890
	pffr	0.300	0.690	0.967	0.318	0.685	0.975	0.354	0.700	0.955

Figure 2.9: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.1.

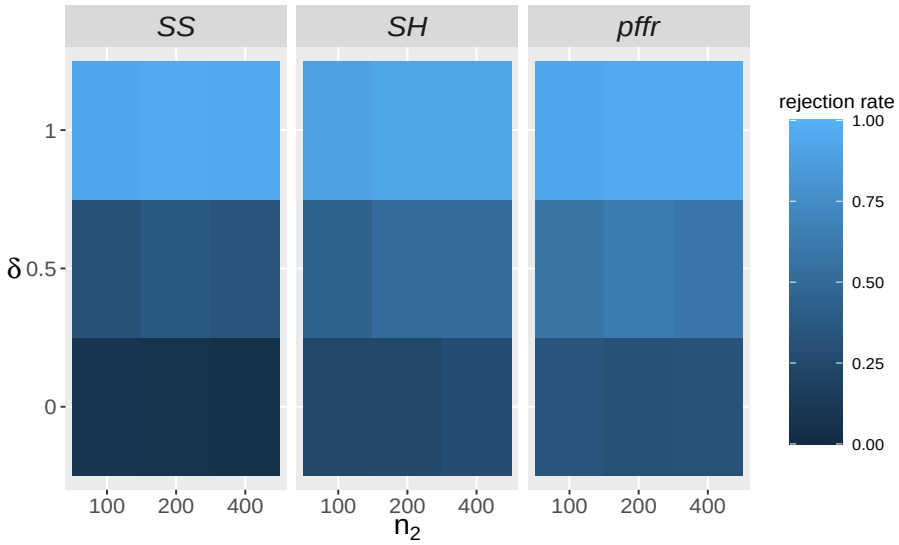


Figure 2.10: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.2.

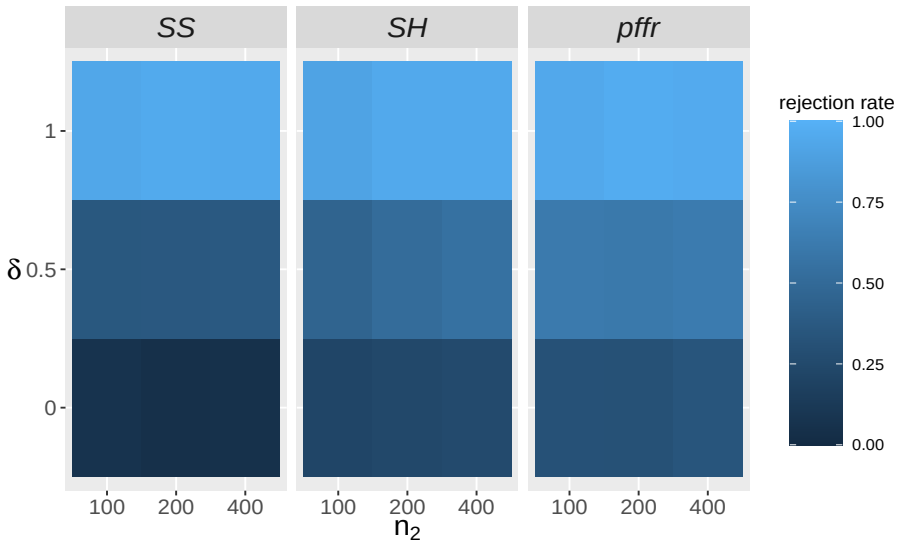


Figure 2.11: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 6$ with $n_1 = 100$ under setting c.3.

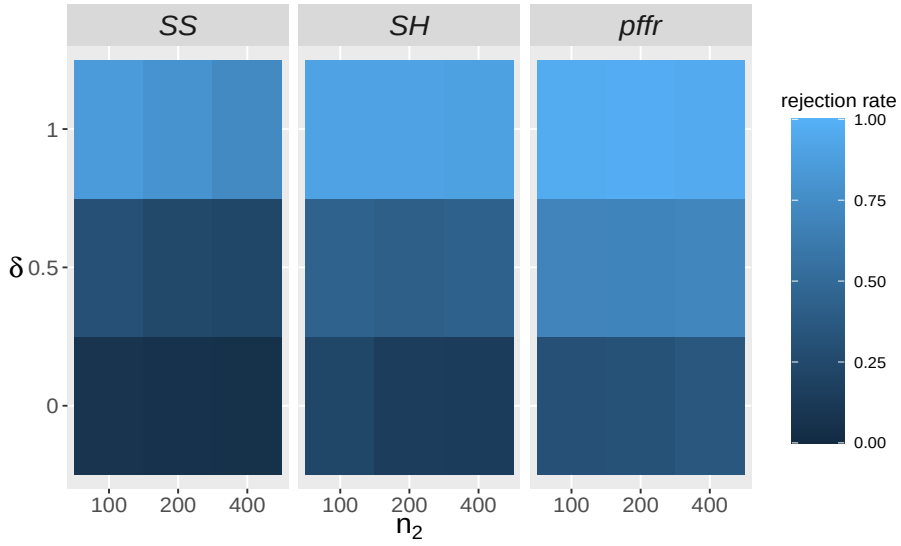


Table 2.5: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$

Setting	Test	$n_1 = 200, n_2 = 100$			$n_1 = 200, n_2 = 200$			$n_1 = 200, n_2 = 400$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
c.1	**SS**	0.057	0.405	0.974	0.053	0.453	0.989	0.047	0.446	0.991
	SH	0.206	0.527	0.955	0.232	0.577	0.974	0.251	0.563	0.972
	pffr	0.533	0.778	0.984	0.534	0.814	0.988	0.524	0.791	0.988
c.2	**SS**	0.062	0.428	0.985	0.053	0.472	0.992	0.034	0.433	0.991
	SH	0.187	0.520	0.975	0.187	0.595	0.980	0.191	0.592	0.982
	pffr	0.533	0.789	0.989	0.580	0.812	0.989	0.558	0.819	0.992
c.3	**SS**	0.049	0.278	0.878	0.052	0.253	0.854	0.036	0.224	0.807
	SH	0.209	0.531	0.968	0.125	0.530	0.970	0.133	0.510	0.964
	pffr	0.515	0.858	0.999	0.538	0.875	0.999	0.586	0.841	0.990

Figure 2.12: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.1.

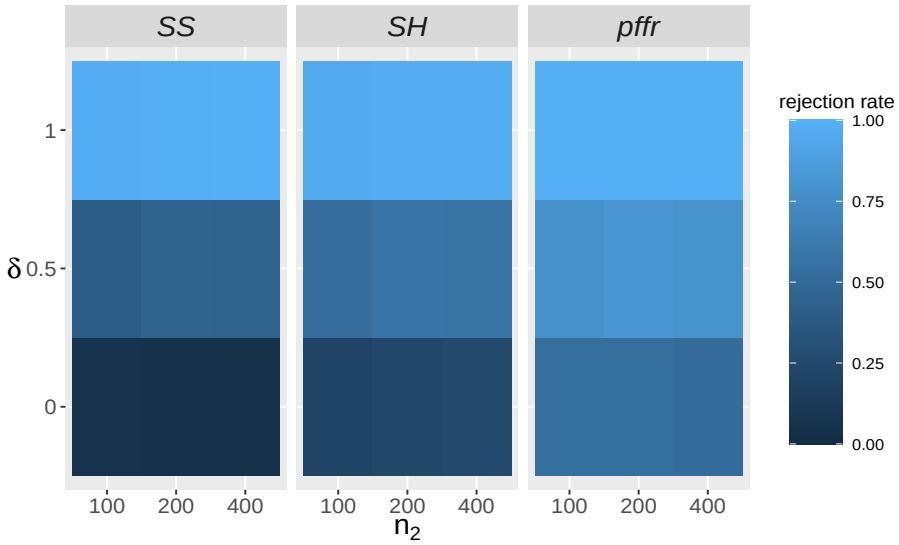


Figure 2.13: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.2.

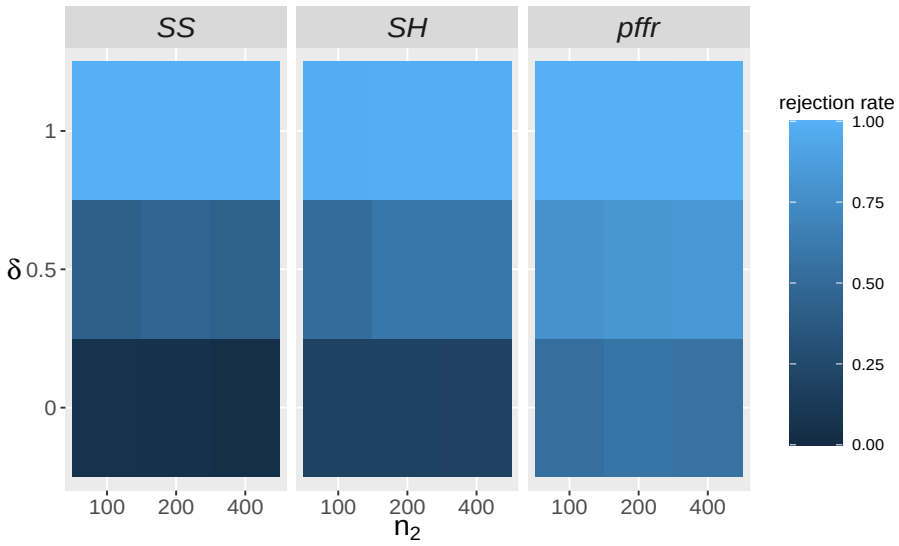


Figure 2.14: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 14$ with $n_1 = 100$ under setting c.3.

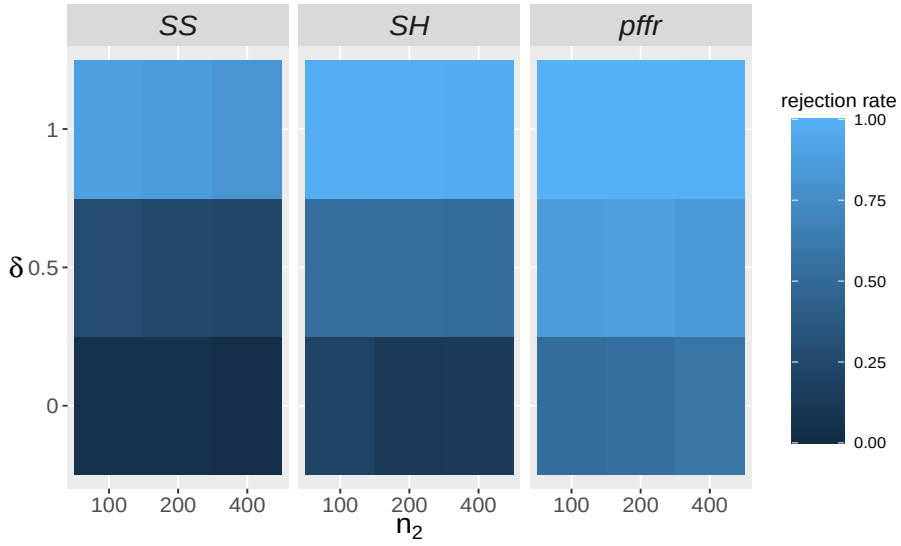


Table 2.6: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$

Setting	Test	$n_1 = 200, n_2 = 100$			$n_1 = 200, n_2 = 200$			$n_1 = 200, n_2 = 400$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
c.1	**SS**	0.048	0.425	0.989	0.051	0.472	0.998	0.047	0.438	0.991
	SH	0.184	0.566	0.975	0.216	0.586	0.974	0.214	0.608	0.973
	pffr	0.572	0.837	0.992	0.603	0.853	0.996	0.624	0.851	0.993
c.2	**SS**	0.057	0.454	0.993	0.044	0.504	0.994	0.037	0.494	0.994
	SH	0.170	0.533	0.974	0.192	0.617	0.979	0.175	0.599	0.985
	pffr	0.605	0.831	0.992	0.607	0.869	0.990	0.638	0.872	0.991
c.3	**SS**	0.044	0.300	0.899	0.048	0.221	0.863	0.041	0.249	0.802
	SH	0.210	0.558	0.982	0.128	0.518	0.987	0.130	0.559	0.977
	pffr	0.600	0.920	1.000	0.642	0.901	0.999	0.637	0.884	0.993

Figure 2.15: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.1.

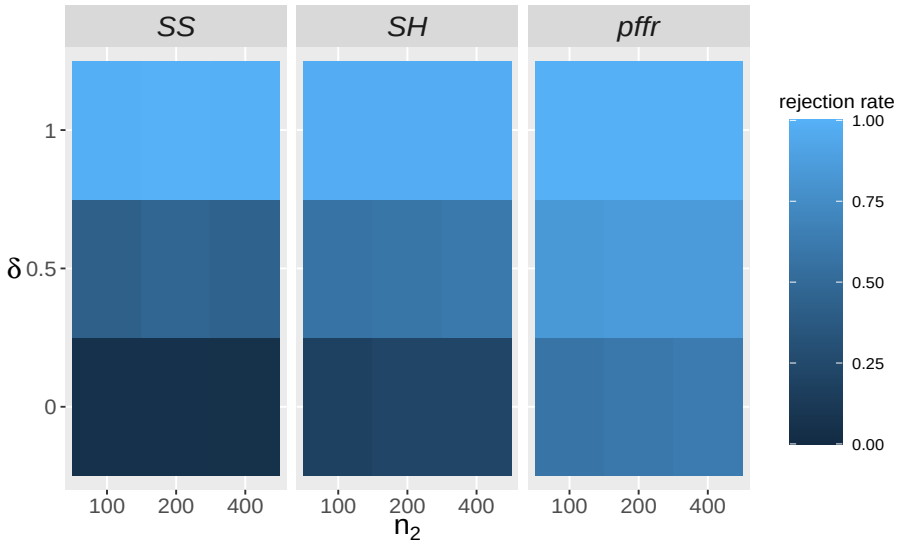


Figure 2.16: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.2.

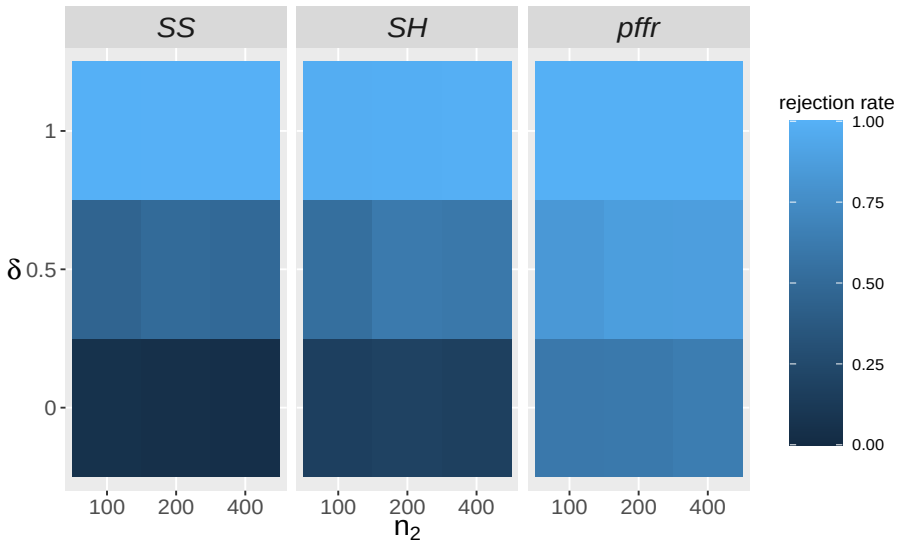
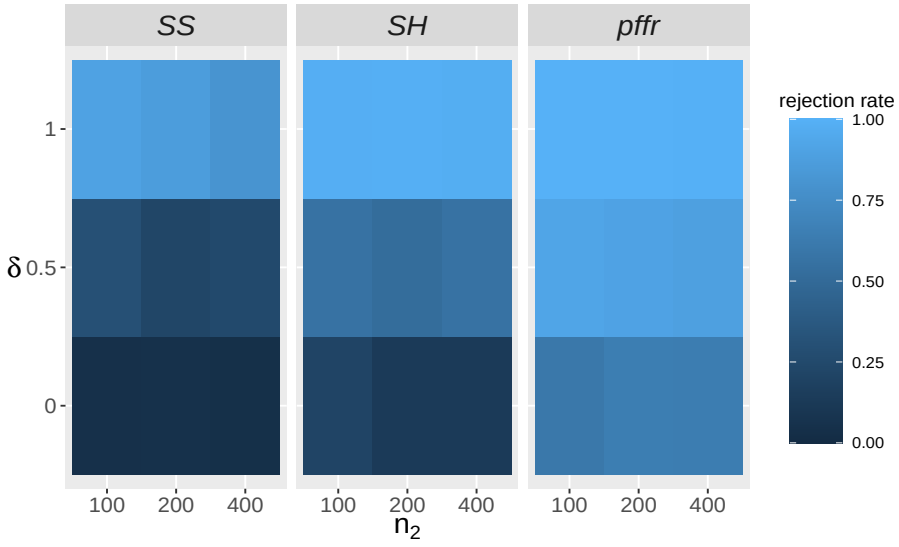


Figure 2.17: Rejection rates across 1000 Monte Carlo runs at a 5% significance level with $N_{\max} = 18$ with $n_1 = 100$ under setting c.3.



Chapter 3

Order Determination for Functional Data

3.1 Introduction

Analogous to principal component analysis (PCA), a subtlety in functional principal component analysis (FPCA) is the selection of the number of functional principal components (FPCs) to retain, known as order determination, which is crucial for effective dimension reduction and approximation of functional data. The choice of the truncation level d in equation (1.6) directly influences the quality of the functional data representation, impacting subsequent analyses such as regression, classification, and clustering. An inadequate selection of d can lead to either overfitting or underfitting, thereby affecting the interpretability and predictive performance of the models built upon the reduced data.

When determining the order in FPCA, the goal is to identify the optimal number of FPCs that capture the essential features of the functional data. The existence of an "optimal" truncation level implicitly assumes that X can be represented by the leading d eigenfunctions of $\boldsymbol{G}$ defined in (1.1), treating higher-order terms as noise. This is equivalent to assuming that the underlying covariance operator has finite rank d, and this assumption has been adopted in Li et al. (2013) and Charkaborty and Panaretos (2022). This effec-

tively transforms the order determination problem to estimating the rank of the covariance operator. When functional data are fully observed without any measurement error, the rank can be estimated in a straightforward manner. This follows from the fact that the estimated covariance operator is a linear combination of observed trajectories, which themselves can be expressed as linear combinations of eigenfunctions associated with non-zero eigenvalues via the Karhunen-Loève expansion, see equation (1.6). However, in practice, functional data are often observed discretely and contaminated by measurement errors. The errors add a ridge to the emprical covariance function, complicating the estimation of the rank of the true covariance operator.

Estimating the rank of the covariance operator from contaminated functional data can be approached through heuristic methods such as scree plots and the fraction of variance explained (FVE), which normally require a subjective pre-specified threshold. A more deliberate approach is to separate the effect of measurement errors from the true covariance function. Hall and Vial (2006) proposed a "low-noise" model, assuming that the noise variance diminishes as the sample size increases. However, this method still requires a subjective threshold as a stopping criterion. Charkaborty and Panaretos (2022) proposed an alternative approach by approximating the infinite-dimensional functional space with a finite-dimensional matrix space. They argued that the corruption of the diagonal entries of the covariance matrix by measurement errors does not affect the rank estimation. Thus, they disregarded the diagonal entries of the sample covariance matrix and subsequently filled them via matrix completion based on a modified Frobenius distance, yielding a matrix with the same rank as the discretized covariance function.

Another approach to removing the impact of measurement errors entails smoothing techniques that leverage the continuity of the covariance function. Once a smoothed covariance function is obtained, the problem can be framed as a model selection problem, allowing the use of information criterion (IC)-based methods, such as the Akaike information criterion (AIC) (Yao et al., 2005b) or the Bayesian information criterion (BIC) (Zhou et al., 2022). However, a direct application of classical IC-based techniques to functional data tends to favour selecting an excessive number of FPCs, or equivalently, overestimating the rank. This issue may arise from the nonparametric nature of the data, where each FPC comprises both a variance parameter and a nonparametric function (Li et al., 2013). To

address this, Li et al. (2013) introduced modified penalty terms in place of those used in AIC and BIC. The penalty term for the adjusted BIC method depends on eigenvalues approaching zero, which pose a challenge, since small eigenvalue estimates can be unreliable in practice. The modified penalty term for AIC is derived under a Gaussian assumption for densely observed functional data. Furthermore, AIC-based methods require that the true model is within the set of candidate models, disregarding potential estimation bias (Hurvich et al., 1998; Li et al., 2013). This assumption, however, is fundamentally flawed when nonparametric smoothing is applied to estimate the mean and covariance functions, as bias is inherently introduced by those smoothing methods when the harmonic mean of the number of observations is not sufficient large compare to the sample size (Zhang and Wang, 2016).

When estimating the rank of the covariance operator of functional data, many existing methods often rely solely on estimated eigenvalues (Hall and Vial 2006, the modified BIC in Li et al. 2013, and Charkaborty and Panaretos 2022), without incorporating information from estimated eigenfunctions. In this chapter, inspired by the work of Luo and Li (2016) on determining the number of principal components for multivariate data, we propose a novel procedure for estimating the rank of the covariance operator by integrating information from both estimated eigenvalues and eigenfunctions. To the best of our knowledge, no existing FDA method integrates information from both estimated eigenvalues and eigenfunctions for rank estimation. Furthermore, unlike AIC-type methods (e.g., Yao et al. 2005b and the modified AIC in Li et al. 2013), our method does not require a distribution assumption to estimate the FPC scores. The key observation is that the variability of the estimated eigenfunctions increases sharply when their index exceeds the true rank d, while the corresponding estimated eigenvalues exhibit a steep drop.

Unlike the multivariate setting studied by Luo and Li (2016), we need to account for the intrinsic infinite-dimensional nature of functional data and sparse observations that are contaminated by measurement errors. In particular, we estimate the mean and covariance functions by applying a local linear smoother to aggregated observations. Instead of using the entire dataset to estimate eigenfunctions, we partition the subjects into two disjoint subsets and estimate their eigenfunctions separately to assess the variability of these eigenfunction estimates. Combining this variability assessment with eigenvalues estimated from

the complete dataset, we develop the Functional Ladle Estimator (FLE) to determine the rank of the covariance operator. The numerical studies showcase excellent performance of our method under various simulation settings, whereas IC-based methods are sensitive to the choice of simulation settings. When applying our method as well as IC-based methods to two real-world applications, FLE also displays great advantages in estimating the order.

The remainder of this chapter is organized as follows. Section 3.2 introduces the data generation process and provides an overview of the FPCA estimation procedures. Section 3.3 provides a detailed description of FLE. In Section 3.4 we conduct extensive simulation studies to compare the performance of our method with that of some alternative methods in various settings, and in Section 3.5 we apply FLE and some alternatives to two real-world datasets: bike-sharing and air pollution data.

3.2 Functional Principal Component Analysis

3.2.1 Data Structure and Model Assumptions

Let $X(t)$ be a continuous and square-integrable stochastic process defined on a compact interval $\mathcal{T} = [0, 1]$, with continuous mean function $\mu(t)$ and covariance function $G(s,t) = \mathbb{E}\{X(s) - \mu(s)\}\{X(t) - \mu(t)\}$. The covariance function defines an operator from $\mathcal{L}^2(\mathcal{T})$ to $\mathcal{L}^2(\mathcal{T})$: $(\boldsymbol{G}f)(s) = \int_0^1 G(s,t)f(t)dt$ for any $f \in \mathcal{L}^2(\mathcal{T})$. Furthermore, the covariance function can be represented as

$$G(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s)\phi_\nu(t), \quad t, s \in \mathcal{T}, \tag{3.1}$$

where (λ_ν, ϕ_ν) is the νth eigenvalue-eigenfunction pair of $\boldsymbol{G}$ satisfying $\boldsymbol{G}\phi_\nu = \lambda_\nu\phi_\nu$ and these ϕ_ν's form an orthonormal basis of $\mathcal{L}^2(\mathcal{T})$. Without loss of generality, we assume that $\lambda_1 > \lambda_2 > \ldots > 0$.

As noted in Section 3.1, we assume that the covariance operator has finite rank d, implying $\lambda_\nu = 0$ for all $\nu > d$ in (3.1). We say the dimensionality of X is d under this

assumption. Consequently, the Karhunen–Loève expansion of $X(t)$ reduces to

$$X(t) = \mu(t) + \sum_{\nu=1}^{d} \xi_\nu \phi_\nu(t), \quad t \in \mathcal{T}, \tag{3.2}$$

where $\xi_\nu = \int_{\mathcal{T}} \{X(t) - \mu(t)\}\phi(t)dt$, $\nu = 1, 2, \ldots, d$ are uncorrelated zero-mean random variables with variance λ_ν. Given n i.i.d. sample paths of X, we assume that the responses Y_{ij} are observed at discrete time points T_{ij} from X_i, subject to additive measurement errors as in model (1.2):

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, N_i. \tag{3.3}$$

Here, N_i's can be random or fixed, and T_{ij}'s are subject-specific (potentially random) observation times in $\mathcal{T}$. Measurement errors ε_{ij} are independent random variables with mean zero and variance σ_ε^2. Moreover, N, T, ε and X are independent. For convenience, we define $R(T_{ij}, T_{il}) = \mathbb{C}\text{ov}(Y_{ij}, Y_{il}) = G(T_{ij}, T_{il}) + \sigma_\varepsilon^2 \delta_{jl}$, where $\delta_{jl} = 1$ if $j = l$, and 0 otherwise.

3.2.2 Estimation of the Model Components

The proposed method is based on the estimation of the eigenpair (λ_ν, ϕ_ν), which are normally obtained by performing the spectral decomposition on the discretization of the smoothed covariance function (Rice and Silverman, 1991). The selection of an appropriate method to estimate the mean function $\mu(\cdot)$ and the covariance function $G(\cdot, \cdot)$ depends on the sampling rate and sampling scheme of X; see Cai and Yuan (2011) and Zhang and Wang (2016) for a more detailed discussion. Here we estimate $\mu(\cdot)$ and $G(\cdot, \cdot)$ by pooling observations across subjects, following the approach of Yao et al. (2005b). In particular, we employ a local linear smoother to estimate $\mu(\cdot)$, where $\hat{\mu}(t) = \hat{a}_0$ is given by

$$(\hat{a}_0, \hat{a}_1) = \arg\min_{a_0, a_1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} K_1\left(\frac{T_{ij} - t}{h_\mu}\right) \{Y_{ij} - a_0 - a_1(T_{ij} - t)\}^2,$$

where $K(\cdot)$ is a symmetric kernel function, and h_μ denotes the bandwidth for the estimation of $\mu(\cdot)$. We adopt a similar method to estimate $G(\cdot,\cdot)$. Specifically, $\hat{b}_0 = \widehat{G}(s,t)$ is determined by using a local linear surface smoothing technique as follows:

$$(\hat{b}_0, \hat{b}_1, \hat{b}_2) = \underset{b_0,b_1,b_2}{\arg\min} \sum_{i=1}^{n} \sum_{1\le j\neq l\le N_i} \left\{ \widehat{R}_i(T_{ij}, T_{il}) - b_0 - b_1(T_{ij}-t) - b_2(T_{il}-s) \right\}^2 \times K\left(\frac{T_{ij}-t}{h_G}\right) K\left(\frac{T_{il}-s}{h_G}\right),$$

where $\widehat{R}_i(T_{ij}, T_{il}) = \{Y_{ij} - \hat{\mu}(T_{ij})\}\{Y_{il} - \hat{\mu}(T_{il})\}$, with h_G being the bandwidth for estimating $G(\cdot,\cdot)$. Data-driven methods are often used to determine bandwidths for estimating mean and covariance functions. For example, Yao et al. (2005b) considered leave-one-curve-out cross-validation. Li et al. (2013) utilized generalized cross-validation (GCV), where the optimal bandwidth is selected to minimize the GCV error.

The eigenpairs $\{\lambda_\nu, \phi_\nu(\cdot)\}_{\nu=1}^{L}$ can be estimated by performing an eigen-decomposition on the discretized estimated covariance function; see Chapter 8 of Ramsay and Silverman (2005) for more details. Let $L < n$ be a prespecified number that is larger than d, denoting the maximum index in the search range for d. In practice, L can be chosen as the total number of non-negative eigenvalues of the discretized $\widehat{G}$ or determined by setting a sufficiently large threshold for the FVE, such as 99.99% (Zhu et al., 2014), i.e.,

$$L = \min\left\{ \nu : \frac{\sum_{j=1}^{\nu} \hat{\lambda}_j}{\sum_{j\ge 1} \hat{\lambda}_j} \ge .9999 \right\}.$$

3.3 Functional Ladle Estimator

To effectively leverage the variability pattern of the estimated eigenfunctions, we develop the following procedure to estimate the rank of the covariance operator $\boldsymbol{G}$. As a preliminary step, we employ a sample-splitting strategy, randomly partitioning the entire dataset into

two disjoint subsets, $\mathcal{D}_1$ and $\mathcal{D}_2$. Without loss of generality, we denote

$$\mathcal{D}_1 = \{(Y_{ij}, T_{ij}) \mid i \leq n/2;\ j \leq N_i;\ i, j \in \mathbb{N}_+\},$$
$$\mathcal{D}_2 = \{(Y_{ij}, T_{ij}) \mid n/2 < i \leq n;\ j \leq N_i;\ i, j \in \mathbb{N}_+\}.$$

For $g = 1, 2$, let $\widehat{G}_g$ and $\{\hat{\lambda}_{g,\nu}, \hat{\phi}_{g,\nu}(\cdot)\}_{\nu=1}^{L}$ denote the estimated covariance function and the corresponding eigenpairs derived from subset $\mathcal{D}_g$, using the techniques introduced in Section 3.2.2. The sampling splitting strategy enables us to quantify the variability in eigenfunction estimates, as detailed below. This strategy is also adopted in Zhou et al. (2023).

Under mild conditions, for $\nu \leq d$, the estimated eigenfunctions $\hat{\phi}_{g,\nu}$ converge in probability to ϕ_ν in both the $\mathcal{L}^2(\mathcal{T})$ and the $\mathcal{L}^\infty(\mathcal{T})$ norms as the sample size increases for $g = 1, 2$ (Zhou et al., 2024). The consistent estimation of ϕ_ν implies that $\langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle \approx 1$, where $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$ for any $f, g \in L^2(\mathcal{T})$. However, for $\nu > d$, $\hat{\phi}_{g,\nu}$ is no longer a consistent estimator for ϕ_ν as $\lambda_\nu = 0$. It should be noted that ϕ_ν is not well defined at the population level since $\lambda_\nu = 0$ for $\nu > d$. However, at the sample level, we may still obtain $\hat{\phi}_\nu$ from the eigendecomposition of $\widehat{G}$ even for $\nu > d$. These estimates arise from the measurement errors in model (3.3), as well as the estimation errors introduced by local linear smoothing. Although $\widehat{G}$ consistently estimates G under mild conditions (Zhou et al., 2024), the re-normalization step in the spectral decomposition on $\widehat{G}$ amplifies the variability of $\hat{\phi}_\nu$'s with index $\nu > d$. Therefore, $\langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle$ could be much smaller than 1 for $\nu > d$ with high probability. Figure 3.1 displays the box plots of $\langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle$ across 1000 simulation runs for different values of ν, where the true rank of $\boldsymbol{G}$ is set to 2. More details on the data generation process for Figure 3.1 are available in 3.A.I. Figure 3.1 justifies our assertion about the phase transition of $\langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle$ from close to 1 to much smaller than 1.

This observation motivates us to introduce a Gram matrix $\boldsymbol{U}_\ell$ for $\ell \leq L$, which is defined as

$$\boldsymbol{U}_\ell = (u_{ij})_{\ell \times \ell}, \quad u_{ij} = \left\langle \hat{\phi}_{1,i}, \hat{\phi}_{2,j} \right\rangle, \quad \text{for } i, j \leq \ell.$$

Then we define $f_{\boldsymbol{U}}(\ell) = 1 - |\det(\boldsymbol{U}_\ell)|$ for $1 \leq \ell \leq L$ to quantify the variability of the space spanned by the first ℓ estimated eigenfunctions.

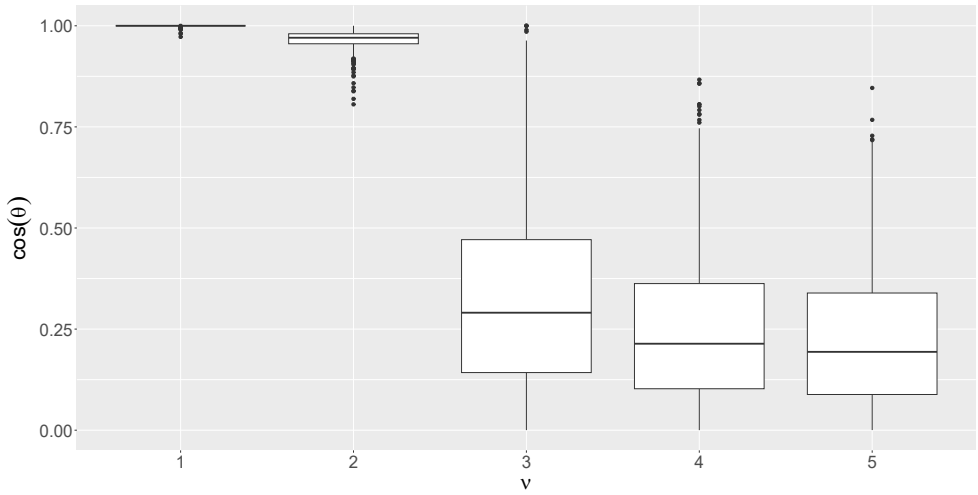


Figure 3.1: The boxplot for $\cos\theta := \langle \hat{\phi}_{1,\nu}, \hat{\phi}_{2,\nu} \rangle$ for $\nu = 1, 2, \ldots, 5$, over 1000 simulation runs, where the true rank is 2.

Intuitively, $\boldsymbol{U}_\ell$ converges to the identity matrix in probability when $\ell \leq d$. In contrast, when $\ell > d$, let us partition the matrix $\boldsymbol{U}_\ell$ into 4 submatrices as follows:

$$\boldsymbol{U}_\ell = \left[\begin{array}{c|c} \mathbf{M}_1 & \mathbf{E}_1 \\ \hline \mathbf{E}_2 & \mathbf{M}_2 \end{array}\right], \tag{3.4}$$

where $\mathbf{M}_1$ is a $d \times d$ matrix and $\mathbf{M}_2$ is an $(\ell - d) \times (\ell - d)$ matrix. By Corollary 5.1.5 in Hsing and Eubank (2015) and Theorem 4.2 in Zhang and Wang (2016), $u_{ij} \approx 0$ with high probability when $i > d, j \leq d$ or $i \leq d, j > d$, leading to $\det(\boldsymbol{U}_\ell) \approx \det(\mathbf{M}_2)$. Meanwhile, the diagonal entries of $\mathbf{M}_2$ are much smaller than 1, which implies a much smaller $|\det(\boldsymbol{U}_\ell)|$.

To find $f_{\boldsymbol{U}}(\ell)$, we evaluate $\hat{\phi}_{g,\nu}(t)$ at a fixed grid $\boldsymbol{t} = (t_1, t_2, \ldots, t_m)^\top$, where the grid points are equally spaced with size δ, i.e., $\delta = t_{i+1} - t_i$ for $i = 1, 2, \ldots, m-1$. Let

$$\widehat{\boldsymbol{B}}_{g,\ell} = \left\{ \hat{\phi}_{g,1}(\boldsymbol{t}), \hat{\phi}_{g,2}(\boldsymbol{t}), \ldots, \hat{\phi}_{g,\ell}(\boldsymbol{t}) \right\}$$

for $g = 1, 2$. Provided that δ is sufficiently small, each entry of $\delta \widehat{\boldsymbol{B}}_{1,\ell}^\top \widehat{\boldsymbol{B}}_{2,\ell}$ is a Riemann sum of the corresponding entry in $\boldsymbol{U}_\ell$, and thus, $1 - |\det(\delta \widehat{\boldsymbol{B}}_{1,\ell}^\top \widehat{\boldsymbol{B}}_{g,\ell})| := \hat{f}_{\boldsymbol{U}}(\ell)$ serves as an approximation of $f_{\boldsymbol{U}}(\ell)$, quantifying the discrepancy between the columns space of $\widehat{\boldsymbol{B}}_{1,\ell}$

and $\widehat{\boldsymbol{B}}_{2,\ell}$. When $\ell \leq d$, the eigenfunctions ϕ_ν are consistently estimated by $\hat{\phi}_{g,\nu}$'s for all $\nu \leq \ell$ and $g = 1, 2$, leading to a minor discrepancy. However, for $\ell > d$, the estimates of ϕ_ν become increasingly dominated by noise for $\nu > d$, leading to a significant increase in discrepancy between $\widehat{\boldsymbol{B}}_{1,\ell}$ and $\widehat{\boldsymbol{B}}_{2,\ell}$. To stabilize numerical performance, we re-normalize $\hat{f}_{\boldsymbol{U}}$ as

$$f(\ell) = \frac{\hat{f}_{\boldsymbol{U}}(\ell)}{1 + \sum_{\ell=1}^{L} \hat{f}_{\boldsymbol{U}}(\ell)}, \quad \text{for } \ell = 1, 2, \ldots, L. \tag{3.5}$$

Regarding eigenvalues, it is anticipated that $\hat{\lambda}_\nu$, estimated from the full dataset, will exhibit a steep decline at $\nu = d + 1$, transitioning from relatively large values observed at $\nu \leq d$. Similar to defining $f(\ell)$, we normalize the eigenvalues and define a function

$$g(\ell) = \frac{\hat{\lambda}_\ell}{\sum_{\ell=1}^{L} \hat{\lambda}_\ell}, \quad \text{for } \ell = 1, 2, \ldots, L. \tag{3.6}$$

Combining the trends of $f(\ell)$ and $g(\ell)$, we construct a function characterized by a 'V' shape, incorporating information from both the estimated eigenvalues and eigenfunctions. Specifically, the functional ladle estimator (FLE) is defined as

$$h(\ell) = f(\ell) + g(\ell), \quad \text{for } \ell = 1, 2, \ldots, L,$$

with the components $f(\ell)$ and $g(\ell)$ specified in (3.5) and (3.6), respectively. In particular, $h(\ell)$ is expected to attain its minimum value around $\ell = d$. This intuition behind using $h(\ell)$ to estimate the rank is demonstrated in Figure 3.2. From the leftmost panel of Figure 3.2, which only accounts for the estimated eigenvalues, we may estimate the rank of $\widehat{G}$ as two or three. However, incorporating the variability of the estimated eigenfunctions, as shown in the middle panel of Figure 3.2, suggests that the most plausible estimate is $\hat{d} = 2$. Algorithm 2 provides the details for implementing our method to estimate the rank of functional data.

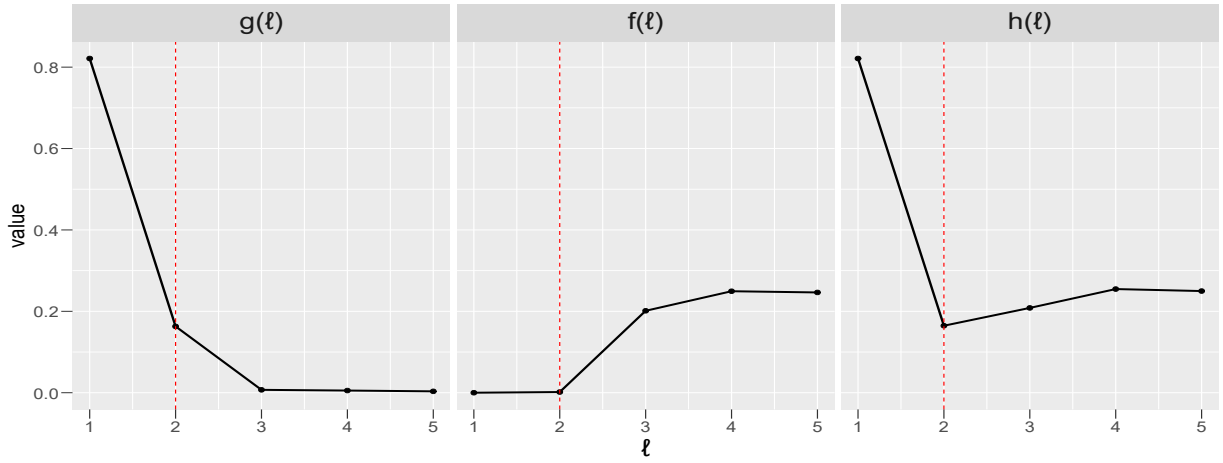


Figure 3.2: The leftmost plot presents $g(\ell)$, which illustrates the proportion of variation explained by each $\hat{\lambda}_\nu$. The middle panel shows $f(\ell)$, capturing the (normalized) variability between $\widehat{B}_{1,\ell}$ and $\widehat{B}_{2,\ell}$. The rightmost panel displays $h(\ell)$, which integrates information from both estimated eigenvalues and estimated eigenfunctions. In each panel, a vertical dotted line marks the true rank.

Algorithm 2 Proposed order determination procedure

1. Randomly split the data set $\mathcal{D}$ into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$.
2. Obtain $\hat{\mu}_g, \widehat{G}_g, \{\hat{\lambda}_{g,\nu}, \hat{\phi}_{g,\nu}\}_{\nu=1}^{L}$ based on the methods discussed in Section 3.2 for each subset $\mathcal{D}_g, g = 1, 2$, where L is a pre-specified number greater than d.
3. Approximate $f_{\boldsymbol{U}}(\ell)$ by $1 - |\det(\delta \widehat{\boldsymbol{B}}_{1,\ell}^\top \widehat{\boldsymbol{B}}_{2,\ell})|$ and obtain $f(\ell)$ by equation (3.5) for all $\ell \leq L$.
4. Obtain $\hat{\lambda}_\nu$ from $\mathcal{D}$ for all $\nu \leq L$, and compute $g(\ell)$ based on equation (3.6).
5. Obtain the estimate of the rank by $\hat{d} = \underset{\ell \leq L}{\arg\min}\ h(\ell) = \underset{\ell \leq L}{\arg\min}\ \{f(\ell) + g(\ell)\}$.

3.4 Simulation Studies

In this section, we perform simulation studies to investigate the finite sample performance of the proposed method. We generate data $\{(Y_{ij}, T_{ij}) \mid i = 1, 2, \ldots, n, j = 1, 2, \ldots, N_i\}$ based on the Karhunen–Loève expansion in (3.2) and model (3.3), where the true mean function is given by $\mu(t) = t + 10\exp\{-(t - 1/2)^2\}$ for $t \in \mathcal{T}$. To showcase the performance under various sampling frequencies, we consider the number of observations per subject $N_i = m \in \{11, 26, 51\}$, where $m = 11$ and 51 are referred to as sparse and dense functional data, respectively. The intermediate case, $m = 26$, represents a transitional state between sparse and dense, hereby referred to as neither. Observation times T_{ij} are uniformly generated over $\mathcal{T}$, referred to as irregular. In addition, for the dense case alone, we also examine a scenario where the data are collected at regularly spaced time points, referred to as regular.

Throughout this section, eigenfunctions are given by

$$\phi_1(t) = 1,\ \phi_{2k}(t) = \sqrt{2}\sin(2k\pi t),\ \phi_{2k+1}(t) = \sqrt{2}\cos(2k\pi t) \quad k \in \mathbb{N}.$$

Regarding the true dimension d, eigenvalues $\lambda_\nu, \nu \in \mathbb{N}$, and the number of curves n, we consider the following scenarios.

- In the *simple* setting, we consider $n \in \{100, 200\}$, and $\lambda_\nu = (4 - \nu)^2$ for $\nu = 1, 2, 3$, and $\lambda_\nu = 0$ for $\nu \geq 4$. Hence, $d = 3$.
- In the *complex* setting, we consider $n \in \{200, 300\}$, and $\lambda_\nu = (7 - \nu)^2$ for $\nu = 1, 2, \ldots, 6$, and $\lambda_\nu = 0$ for all $\nu \geq 7$. Thus, $d = 6$.

Additionally, we consider two designs for FPC scores: Gaussian, where scores $\xi_{i\nu} \sim N(0, \lambda_\nu)$, and non-Gaussian, where scores $\xi_{i\nu}$ are generated from the centered exponential distribution with variance λ_ν for all $\nu \leq d$, introducing skewness into the functional data. The variance of measurement error, σ_ε^2, is set to $\{0.1, 0.5, 1, 4\}$. Notably, the largest value of σ_ε^2 exceeds the smallest nonzero eigenvalue in each scenario mentioned above, while the smallest σ_ε^2 mimics the low-noise regime described in Hall and Vial (2006).

For irregularly sampled data, we estimate model components using the methods described in Section 3.2.2. If observations are observed at a regularly dense grid, i.e.,

$T_{ij} = t_j$ for all $j = 1, 2, \ldots, m$, we estimate the mean function using the sample mean $\hat{\mu}(t_j) = n^{-1} \sum_{i=1}^{n} Y_{ij}$. For the covariance function, we first estimate the raw covariance function by $\widehat{R}(t_j, t_k) = n^{-1} \sum_{i=1}^{n} \{Y_{ij} - \hat{\mu}(t_j)\}\{Y_{ik} - \hat{\mu}(t_k)\}$. Thus, $\widehat{G}(t_j, t_k) = \widehat{R}(t_j, t_k) - \hat{\sigma}_{\varepsilon}^2 \mathbf{I}_m$, where $\hat{\sigma}_{\varepsilon}^2$ can be estimated by a difference-based estimator (Rice, 1984), and $\mathbf{I}_m$ denotes the $m \times m$ identity matrix.

The bandwidths used for mean and covariance estimation follows from the procedure described in Section 3.2.2, except in the scenario where $m = 51$ observations are irregularly spaced. In this particular scenario, we set $h_G = n^{-1/5}/6$ (approximately 0.06 when $n = 200$ and 0.05 when $n = 300$) to reduce the computational cost, as the h_G's selected by GCV range between 0.04 and 0.07.

Each simulation setting is repeated 500 times, and we report the percentage of times the true rank is identified. We compare our proposed method (FLE) with several commonly used approaches for selecting the rank of functional data, particularly focusing on IC-based methods. We include pseudo-AIC from Yao et al. (2005b) (denoted as AIC_{Yao}) and pseudo-BIC implemented in R package *fdapace* (denoted as BIC_{PACE}). Besides that, we consider the modified AIC and BIC proposed by Li et al. (2013), denoted by AIC_{Li} and BIC_{Li}, respectively. Simulation results for Gaussian processes with $d = 3$ are summarized in Tables 3.1 and 3.2. Additional simulation results, including those for non-Gaussian processes and $d = 6$, are provided in Appendix 3.A.II.

The simulation results demonstrate that FLE is highly competitive across all simulation settings. In particular, when data are regularly and densely observed from a Gaussian process, FLE consistently identifies the true rank. Even in sparse settings with $m = 11$, this method remains robust, achieving an accuracy of at least 86% across all scenarios. Moreover, the performance improves as m or n increases, which implies that our method performs better when more aggregated observations are available. However, the performance of the proposed method can deteriorate when both the sample size and the number of observation per subject is small, for example, see Table 3.5. In this scenario, the information available for each subsample to estimate the eigenfunction is limited, thereby the resulting eigenfunction estimators are not reliable anymore.

In contrast, IC-based methods exhibit high instability. For example, the accuracy of

Table 3.1: Comparison of order determination for Gaussian random processes under irregular design with $d = 3$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	100				200			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 11$	**FLE**		0.89	0.89	0.90	0.86	0.92	0.92	0.91	0.90
	AIC_{Li}		0.80	0.90	0.98	0.83	0.75	0.94	0.99	0.97
	BIC_{Li}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AIC_{Yao}		0.11	0.07	0.07	0.10	0.01	0.01	0.00	0.01
	BIC_{PACE}		0.31	0.23	0.24	0.32	0.10	0.04	0.03	0.06
$m = 26$	**FLE**		0.96	0.97	0.95	0.96	0.96	0.98	0.97	0.96
	AIC_{Li}		0.27	0.72	0.91	1.00	0.36	0.93	0.97	1.00
	BIC_{Li}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AIC_{Yao}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BIC_{PACE}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

AIC_{Li} ranges from 20% to 100% depending on the value of σ_ε^2 for dense functional data. This instability is evident in every setting that we have examined. Moreover, the performance of IC-based methods that are developed based on likelihood functions is highly sensitive to the values of n, m and σ_ε^2. Our numerical analysis reveals that these methods perform well only for specific parameter combinations. Furthermore, although BIC_{Li} does not depend on the likelihood function, it requires accurate estimation of eigenvalues, particularly for those with indices close to the true rank d. Consequently, its performance deteriorates under the setting of irregular and sparse observations or a high noise variance σ_ε^2, since accurate estimation of eigenvalues with indices close to d becomes quite challenging.

Table 3.2: Comparison of order determination for Gaussian random processes with $d = 3$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	100				200			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 51$, irregular	**FLE**		0.97	0.96	0.98	0.98	0.97	0.98	0.99	0.98
	AIC_{Li}		0.19	0.74	0.93	1.00	0.25	0.92	0.99	1.00
	BIC_{Li}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	AIC_{Yao}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BIC_{PACE}		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$m = 51$, regular	**FLE**		1.00	1.00	1.00	1.000	1.00	1.00	1.00	1.00
	AIC_{Li}		0.36	0.01	0.00	0.00	0.75	0.21	0.05	0.00
	BIC_{Li}		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	AIC_{Yao}		0.04	0.00	0.00	0.00	0.06	0.00	0.00	0.00
	BIC_{PACE}		0.04	0.00	0.00	0.00	0.06	0.00	0.00	0.00

3.5 Real Data Examples

In this section, we apply our proposed method to two real-world examples. In reality, we can never know the true order of functional data, thus we cannot evaluate the accuracy of our method in estimating the dimension of functional data directly. Instead, we illustrate its effectiveness indirectly. In the literature on functional linear regression and functional data classification, FPC-based methods have received extensive attention; see Yao et al. (2005a), Hall and Horowitz (2007), Delaigle and Hall (2012), Dai et al. (2017) and references therein. More specifically, FPC scores obtained from FPCA are used as covariates in these regression and classification methods. Selecting the number of FPC scores is a critical problem when applying these methods. We demonstrate the effectiveness of our method by showcasing its performance in FPC-based methods for functional linear regression and functional data classification.

3.5.1 Capital Bike-share Data

We analyse a dataset from the Capital Bike-share System (CBS) in Washington, D.C., to investigate the relationship between the hourly rental profile and the total rental time (TRT) in hours on the same day. This relationship serves as a potential commercial indicator for assessing whether bicycle demand exceeds supply in a given region. The CBS dataset records the date and time of each rental, offering valuable insights into public transportation usage and environmental factors.

For this study, we consider rental transactions for the year 2017, restricting our analysis to users with memberships. Given that rental patterns differ significantly between weekdays and weekends, we focus solely on weekends, including holidays, resulting in a total of 116 days of data. Additionally, for each day, we remove records where rental durations exceed five hours, as these extremely long rental times are likely due to users forgetting to return bikes.

Our objective is to formally assess how the hourly rental profile, denoted as $X(\cdot)$, affect the TRT, denoted as Y, for each day. Figure 3.3 displays the hourly rental profiles across all 116 days. Figure 3.5 presents the histogram of rental durations, which is available in 3.A.III. We consider a scalar-on-function regression model to quantify the relationship between X and Y, which is defined as follows:

$$Y_i = \alpha + \int \beta(t) X_i(t) dt + e_i, \quad i \in \{1, 2, \ldots, 249\}, \tag{3.7}$$

where the coefficient function $\beta(\cdot)$ quantifies the impact of hourly rental numbers on TRT.

We employ the FPC-based method developed by Hall and Horowitz (2007) to fit model (3.7). To compare the performance of the proposed method with other IC-based methods, we randomly split the dataset into 90% training data and 10% test data. Since X is observed on a common grid, we estimate the mean function and covariance function using sample averages and the empirical covariance matrix on the training data. Let $\hat{d}$ denote the number of FPCs selected by a selection algorithm. Applying FPCA to X, model (3.7)

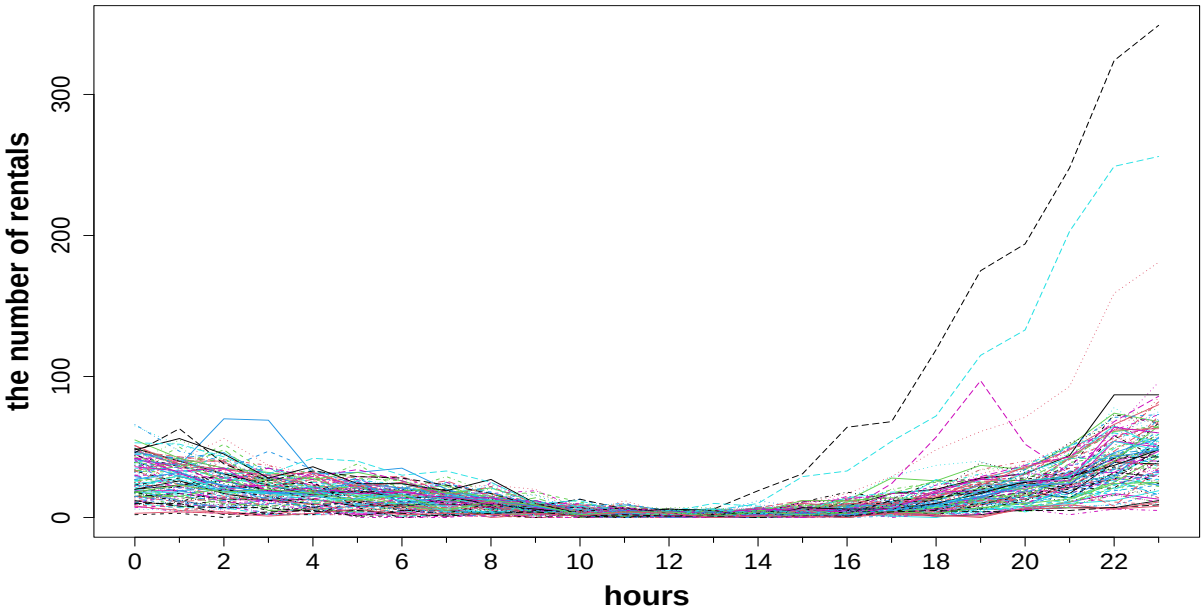


Figure 3.3: The number of rentals over 24 hours for all days in the training set.

reduces to a linear regression model:

$$Y_i = \beta_0 + \sum_{\nu=1}^{\hat{d}} \beta_\nu \xi_{i\nu} + e_i,$$

where $\beta_0 = \alpha + \int_t \beta(t)\mu(t)dt$ and $\beta_\nu = \int_t \beta(t)\phi_\nu(t)dt$ for $\nu \geq 1$. After obtaining the estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{\hat{d}})^\top$ by ordinary least squares, the estimate of β in model (3.7) is given by $\hat{\beta}(t) = \sum_{\nu=1}^{\hat{d}} \hat{\beta}_\nu \hat{\phi}_\nu(t)$, where ϕ_ν denote the estimated eigenfunctions of X.

Using the estimated slope function $\hat{\beta}(t)$, we predict TRT for days in the test set, and compute the prediction error, defined as $\sum_{i\in\mathcal{I}}(y_i - \hat{y}_i)^2/|\mathcal{I}|$, where $\mathcal{I}$ is the index set for test days and $|\mathcal{I}|$ denotes the total number of test days. Table 3.3 summarizes the estimated rank and the corresponding prediction errors for the proposed method and IC-based methods across 500 independent splits.

The results demonstrate the superiority of our method in determining the order of functional data, as indicated by lower prediction errors. The estimated slope function based

Table 3.3: The estimated rank of the bike sharing data with averaged prediction errors under 500 runs. The estimated rank is the mode of estimated ranks.

method	FLE	AIC_{Yao}	BIC_{PACE}	AIC_{m}	BIC_{m}
$\hat{d}$	4	22	22	22	21
prediction error	7584.569	7822.712	7822.712	7822.712	7903.395

on the first four eigenfunctions, the first three estimated eigenfunctions, and additional estimated eigenfunctions, are displayed in Figures 3.6 - 3.8 in 3.A.III.

3.5.2 Beijing Air Pollutants Data

Air pollution has become a major environmental concern in many cities across China due to rapid industrialization and urbanization. Fine particulate matter (PM) with an aerodynamic diameter less than $2.5\,\mu$m, often referred to as $\text{PM}_{2.5}$, is one of the main pollutants in Beijing (Zhang et al., 2023). Exposure to $\text{PM}_{2.5}$ has been associated with cardiovascular and respiratory diseases and even lung cancer (Pope III et al., 2002; Hoek et al., 2013; Lelieveld et al., 2015).

As noted in Liang et al. (2015) and Zhang et al. (2023), $\text{PM}_{2.5}$ levels are highly influenced by meteorological conditions, particularly wind and humidity conditions. In this example, we focus on the effect of dew point temperature (DEW), which serves as a proxy for both relative humidity and air temperature (Alduchov and Eskridge, 1996). Moreover, DEW has been recognized as an important predictor for the levels of $\text{PM}_{2.5}$. For example, the first branching rule of a tree model proposed in Zhang et al. (2017) is decided by DEW, and its selection frequency is 100% in the selection frequency chart of Liang et al. (2015).

To demonstrate the effectiveness of the proposed method, we use DEW to predict the daily $\text{PM}_{2.5}$ level, denoted as Y, whose label is based on the daily average readings of $\text{PM}_{2.5}$. Specifically, daily average readings are divided into four categories: $Y = 0$ if $\text{PM}_{2.5} \leq 35\,\mu\text{gm}^{-3}$, $Y = 1$ if $35\,\mu\text{gm}^{-3} < \text{PM}_{2.5} \leq 75\,\mu\text{gm}^{-3}$, $Y = 2$ if $75\,\mu\text{gm}^{-3} < \text{PM}_{2.5} \leq 150\,\mu\text{gm}^{-3}$, $Y = 3$, if $150\,\mu\text{gm}^{-3} < \text{PM}_{2.5}$. This partition adapts the current

national ambient air quality standard in China, which is also adopted in Zhang et al. (2017).

To predict Y with DEW, we consider the following generalized scalar-on-function regression:

$$\log \frac{\Pr(Y = k \mid X)}{\Pr(Y = 3 \mid X)} = \alpha_k + \int_t \beta_k(t) X(t) dt + e \quad \text{for } k = 0, 1, 2,$$

where $X(t)$ denotes the hourly DEW readings. Such models with a general link function were studied in Müller and Stadtmüller (2005). We analyse hourly air pollution data collected from Huairou, an urban district in the northern part of Beijing, spanning March 2013 to February 2017. To improve prediction accuracy, we stratify the data based on human activity patterns and seasonal effects. First, we separate workdays from weekends and holidays, ensuring that each subset exhibits homogeneous human activity patterns. Given the strong seasonal variability of air pollution levels in Beijing (Zhang et al., 2017), we partition 12 months into four seasons. In this study, we consider winter months (December–February of the following year) as winter heating is one of the most important factors directly impacting $PM_{2.5}$ levels (Zhang et al., 2023). This partitioning strategy is widely used in air pollution studies (Liang et al., 2015). Finally, to evaluate the prediction accuracy, we randomly select 10% of the data as a test set while using the remaining 90% for training.

In the training set, we estimate the mean function and covariance function of X using the same procedures as in the Capital Bike-share study since DEW is recorded on a common grid. The dimension $\hat{d}$ is determined by the proposed method and other IC-based methods. Using the estimated slope function and the FPC scores, we predict Y in the test set and evaluate classification accuracy on the test set to compare the effectiveness of different methods in estimating d. Table 3.4 summarizes the estimated rank and the classification accuracy for winter workdays over 500 independent runs.

The results indicate that our method achieves the highest classification accuracy, demonstrating its superior predictive performance in classifying functional data, and thus its great capability to estimate the order of functional data. Furthermore, our analysis suggests that DEW is a strong predictor of $PM_{2.5}$ pollution levels, consistent with prior studies (Zhang et al., 2017).

Table 3.4: The estimated rank of the Beijing air pollutants data and the classification accuracy for different methods in winter workdays under 500 runs. The estimated rank is the mode of estimated ranks.

method	FLE	AIC_{Yao}	BIC_{PACE}	AIC_{Li}	BIC_{Li}
$\hat{d}$	4	11	11	4	13
accuracy	0.542	0.522	0.523	0.542	0.509

3.A Additional Details and Numerical Results

3.A.I Estimated Eigenfunctions

In this section, we provide the data generation process and include an additional figure illustrating the variability of the estimated eigenfunctions.

We generate the data $\mathcal{D} = (Y_{ij}, T_{ij})$ based on the model (3.3) and the Karhunen–Loève expansion in equation (3.2), where $\mu(t) = t + 10\exp\{-(t-5)^2\}$ for $t \in [0, 10]$ and $\sigma_\varepsilon^2 = 0.01$. The eigenfunctions are given by

$$\phi_1(t) = 5^{-1/2}\cos(\pi t/5),\ \ \phi_2(t) = -5^{-1/2}\sin(\pi t/5),$$

and $\lambda_1 = 25, \lambda_2 = 4$ and $\lambda_\nu = 0$ for $\nu \geq 3$. Thus, the dimension of X is 2. Figure 3.4 illustrates the variability of estimated eigenfunctions based on one simulation run. From the figure, we can see that the first two estimated eigenfunctions from each subset appear similar in shape, whereas the third estimated eigenfunction deviates notably.

3.A.II Additional Simulation Results

This section provides extra simulation studies results, focusing on scenarios where $d = 6$ or the underlying processes are non-Gaussian. Tables 3.5 and 3.6 show the results when $d = 6$ and the underlying processes are Gaussian. Table 3.7 - Table 3.10 illustrate the results for non-Gaussian processes, where $d = 3$ for first two tables and $d = 6$ for the last two tables.

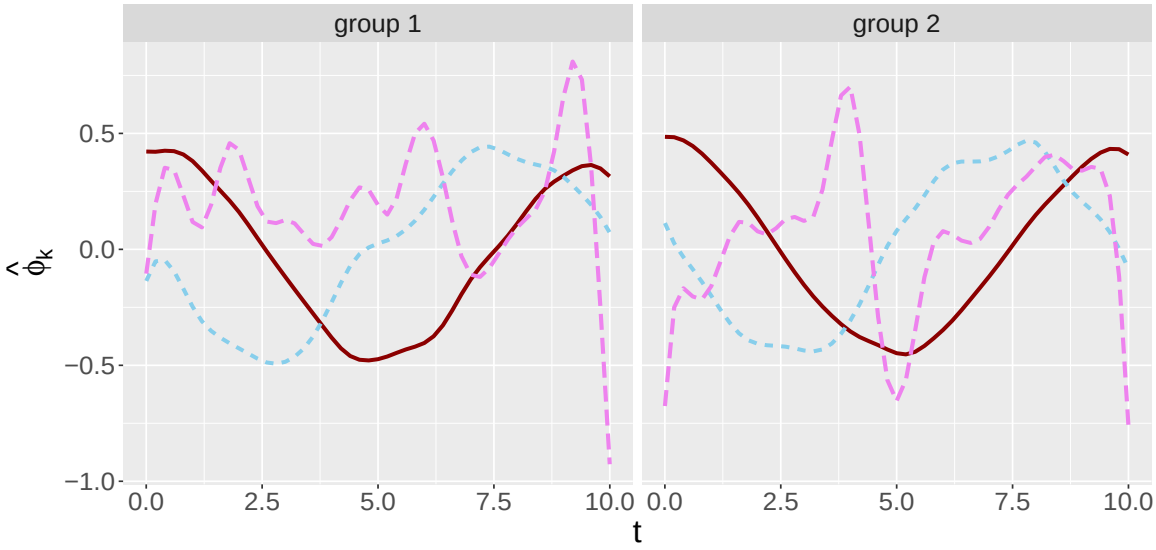


Figure 3.4: The left panel displays the first three estimated eigenfunctions derived from $\mathcal{D}_1$, while the right panel presents those from $\mathcal{D}_2$. Here the solid line represents $\hat{\phi}_1$, the dash line is $\hat{\phi}_2$, and the long dash line is $\hat{\phi}_3$, for both panels.

Table 3.5: Comparison of order determination for Gaussian random processes under irregular design with $d = 6$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	200				300			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 11$	**FLE**		0.112	0.120	0.124	0.096	0.158	0.148	0.164	0.152
	AIC_{Li}		0.126	0.102	0.084	0.016	0.256	0.192	0.176	0.050
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.198	0.218	0.218	0.230	0.064	0.104	0.094	0.158
	BIC_{PACE}		0.550	0.496	0.522	0.468	0.328	0.320	0.274	0.320
$m = 26$	**FLE**		0.200	0.204	0.170	0.162	0.220	0.242	0.208	0.172
	AIC_{Li}		0.258	0.370	0.468	0.734	0.206	0.348	0.466	0.780
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.000	0.008	0.006	0.002	0.000	0.000	0.000	0.000
	BIC_{PACE}		0.012	0.026	0.014	0.016	0.014	0.020	0.012	0.000

Table 3.6: Comparison of order determination for Gaussian random processes with $d = 6$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	200				300			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 51$, irregular	**FLE**		0.632	0.622	0.626	0.590	0.744	0.762	0.742	0.734
	AIC_{Li}		0.572	0.670	0.828	0.956	0.390	0.676	0.798	0.982
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	BIC_{PACE}		0.152	0.106	0.062	0.028	0.032	0.010	0.010	0.000
$m = 51$, regular	**FLE**		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AIC_{Li}		1.000	0.960	0.624	0.082	1.000	1.000	0.986	0.462
	BIC_{Li}		0.000	0.216	0.344	0.366	0.000	0.410	0.590	0.680
	AIC_{Yao}		0.606	0.074	0.004	0.000	0.566	0.100	0.012	0.000
	BIC_{PACE}		0.608	0.080	0.004	0.000	0.568	0.100	0.012	0.000

Table 3.7: Comparison of order determination for non-Gaussian random processes under irregular design with $d = 3$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	100				200			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 11$	**FLE**		0.848	0.814	0.818	0.770	0.888	0.890	0.912	0.882
	AIC_{Li}		0.822	0.868	0.872	0.596	0.772	0.888	0.952	0.818
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.346	0.362	0.296	0.372	0.132	0.120	0.112	0.118
	BIC_{PACE}		0.548	0.504	0.488	0.594	0.290	0.260	0.260	0.336
$m = 26$	**FLE**		0.946	0.934	0.950	0.940	0.972	0.956	0.964	0.946
	AIC_{Li}		0.278	0.604	0.756	0.976	0.218	0.642	0.832	0.984
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.014	0.014	0.010	0.022	0.000	0.002	0.000	0.006
	BIC_{PACE}		0.056	0.042	0.034	0.038	0.008	0.006	0.002	0.012

Table 3.8: Comparison of order determination for non-Gaussian random processes with $d = 3$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	100				200			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 51$, irregular	**FLE**		0.978	0.974	0.964	0.962	0.984	0.978	0.972	0.978
	AIC_{Li}		0.094	0.448	0.684	0.972	0.086	0.528	0.774	0.994
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000
	BIC_{PACE}		0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000
$m = 51$, regular	**FLE**		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AIC_{Li}		0.338	0.010	0.000	0.000	0.750	0.206	0.058	0.000
	BIC_{Li}		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AIC_{Yao}		0.040	0.000	0.000	0.000	0.078	0.000	0.000	0.000
	BIC_{PACE}		0.042	0.000	0.000	0.000	0.080	0.000	0.000	0.000

Table 3.9: Comparison of order determination for non-Gaussian random processes under irregular design with $d = 6$ and $m = 11, 26$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	200				300			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 11$	**FLE**		0.066	0.058	0.036	0.050	0.096	0.082	0.058	0.096
	AIC_{Li}		0.046	0.022	0.024	0.010	0.060	0.036	0.040	0.010
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.550	0.568	0.552	0.550	0.458	0.498	0.502	0.470
	BIC_{PACE}		0.428	0.382	0.384	0.418	0.610	0.528	0.580	0.592
$m = 26$	**FLE**		0.174	0.162	0.142	0.152	0.220	0.214	0.192	0.196
	AIC_{Li}		0.570	0.570	0.596	0.520	0.642	0.652	0.654	0.666
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.046	0.044	0.048	0.026	0.004	0.006	0.000	0.006
	BIC_{PACE}		0.284	0.230	0.244	0.222	0.122	0.102	0.072	0.064

Table 3.10: Comparison of order determination for Gaussian random processes with $d = 6$ and $m = 51$: entries in Columns 4–11 are the percentage of accurate order determination across 500 iterations.

Number of measurements	Methods	n	200				300			
		σ_ε^2	0.1	0.5	1	4	0.1	0.5	1	4
$m = 51$, irregular	**FLE**		0.426	0.398	0.440	0.406	0.532	0.474	0.514	0.516
	AIC_{Li}		0.396	0.542	0.668	0.850	0.252	0.478	0.614	0.870
	BIC_{Li}		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AIC_{Yao}		0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	BIC_{PACE}		0.100	0.052	0.064	0.034	0.014	0.004	0.006	0.000
$m = 51$, regular	**FLE**		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AIC_{Li}		1.000	0.952	0.600	0.070	1.000	1.000	0.978	0.522
	BIC_{Li}		0.000	0.214	0.324	0.378	0.000	0.400	0.570	0.706
	AIC_{Yao}		0.500	0.088	0.004	0.000	0.540	0.094	0.016	0.000
	BIC_{PACE}		0.504	0.094	0.004	0.000	0.544	0.094	0.016	0.000

3.A.III Additional Real Data Analyses

This section provides additional figures to demonstrate the dataset and the performance of the proposed method. Figure 3.5 illustrates the histogram of the total rental time for all days in the training set. Based on the estimated order for the capital bike data, we plot the estimated slope function $\hat{\beta}$ when $\hat{d} = 4$ in Figure 3.6. Furthermore, Figure 3.7 shows the first three estimated eigenfunctions, while Figure 3.8 illustrates the 5th and the 6th estimated eigenfunctions.

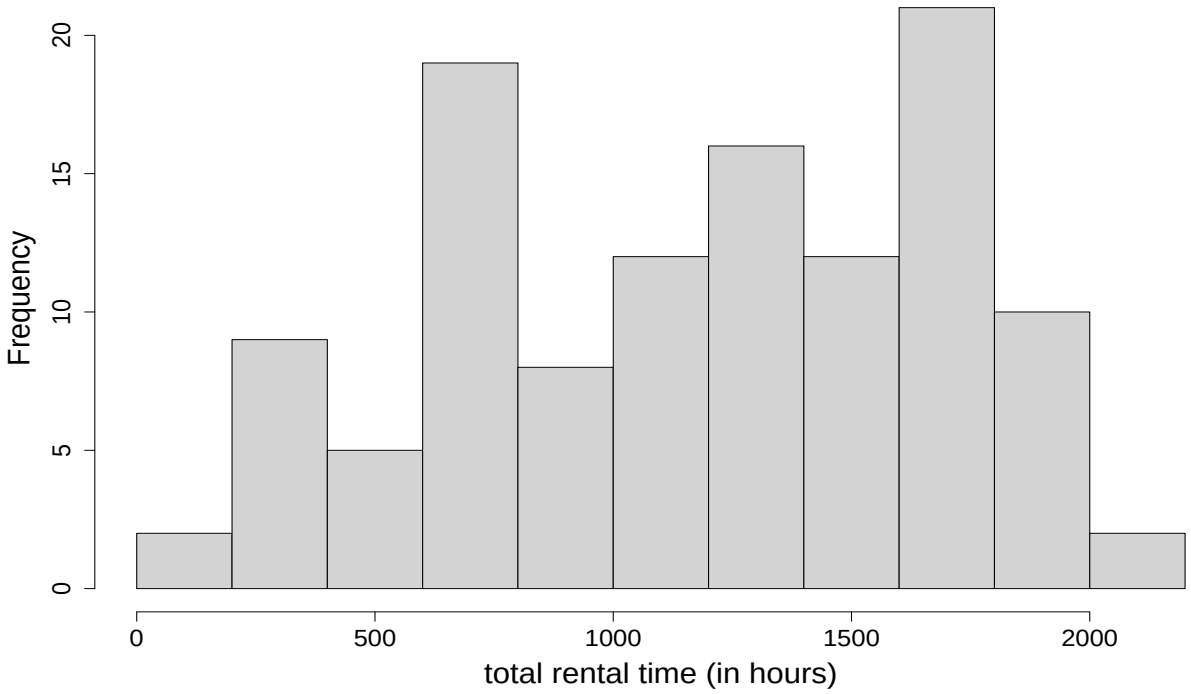


Figure 3.5: The histogram of total rental time (in hours)

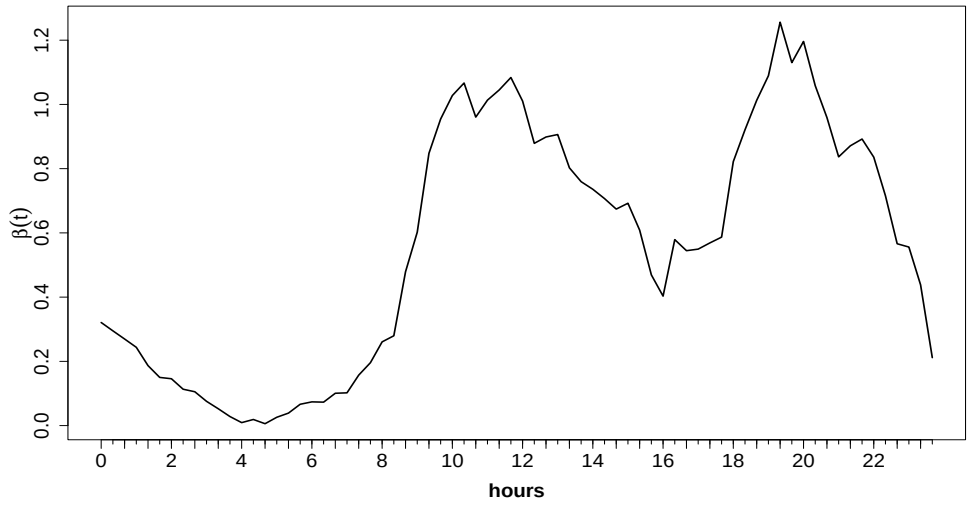


Figure 3.6: The estimated slope function, $\hat{\beta}(t)$, for the bike-sharing dataset, using the first four estimated eigenfunctions.

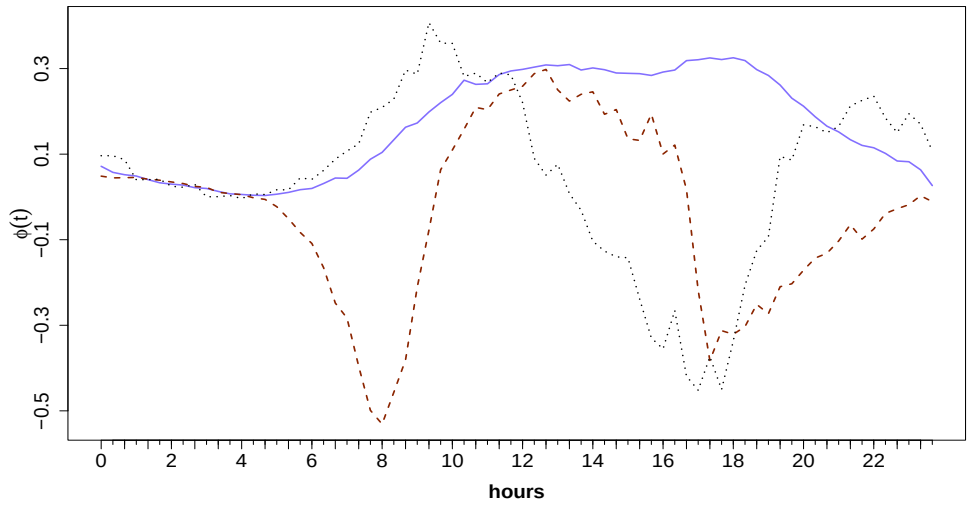


Figure 3.7: The plot shows the top three estimated eigenfunctions for the bike-sharing data, where the first eigenfunction is presented in solid line, the second eigenfunction is in dash line, and the third eigenfunction is the dotted line.

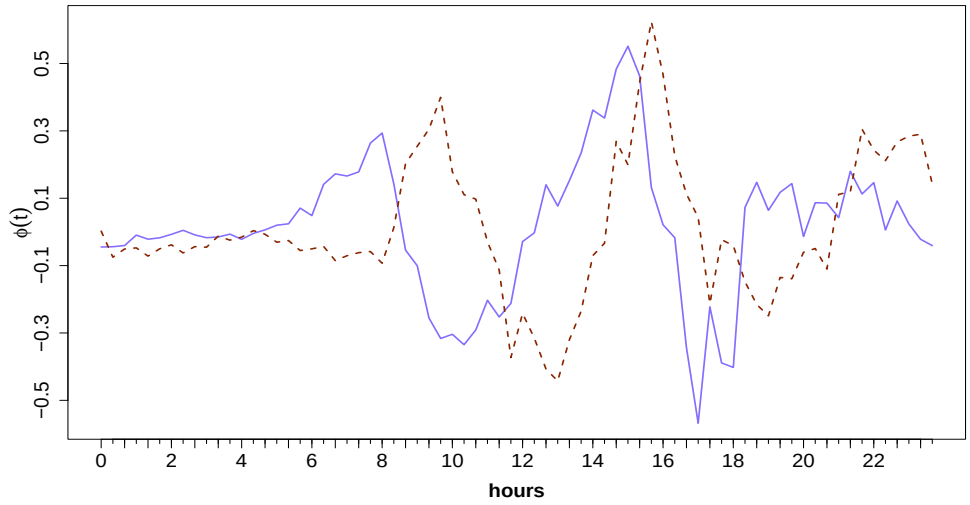


Figure 3.8: The 5th estimated eigenfunction (solid line) and the estimated 6th eigenfunction (dash line) for the bike-sharing data.

Chapter 4

Learning Shared and Source-specific Subspaces across Multiple Data Sources for Functional Data

4.1 Introduction

The theoretical foundations of FPCA have been extensively developed over the past two decades (Yao et al., 2005b; Hall and Hosseini-Nasab, 2006; Zhou et al., 2024). FPCA is also commonly adopted as a regularization tool when inverting the covariance operator is required. For example, the inverse of the covariance operator is unbounded without regularization in functional linear regression (Hall and Horowitz, 2007; Zhou et al., 2023) and functional generalized linear models (Dou et al., 2012). While FPCA methods are well established both in theory and applications, comparatively less attention has been devoted to settings involving multiple functional data sources, where one seeks to aggregate information and uncover shared structure. Yet, such scenarios are increasingly common in modern applications, where data collected at different sites often present heterogeneity. For instance, in the air pollutants study discussed in Section 4.5, measurements are obtained from five cities with distinct geographical locations and climate conditions.

When functional data are collected from multiple sources, heterogeneity across sources poses additional challenges for analysis. A naive extension of single-source FPCA is to compute source-specific sample covariance operators and then perform FPCA on their average. However, this averaging procedure implicitly assumes homogeneity across sources, and dominant patterns present only in particular sources may introduce severe bias in identifying the true shared structure. To illustrate, we construct a toy example in Example 4.1 with two sources. The covariance operator of the second source is generated by applying a simple rotation to the basis functions of the first source, while retaining the same eigenvalues. Even though the two sources have identical sample sizes and eigenvalues, the eigenfunction estimated from the averaged covariance operator deviates substantially from the truth, as shown in Figure 4.1.

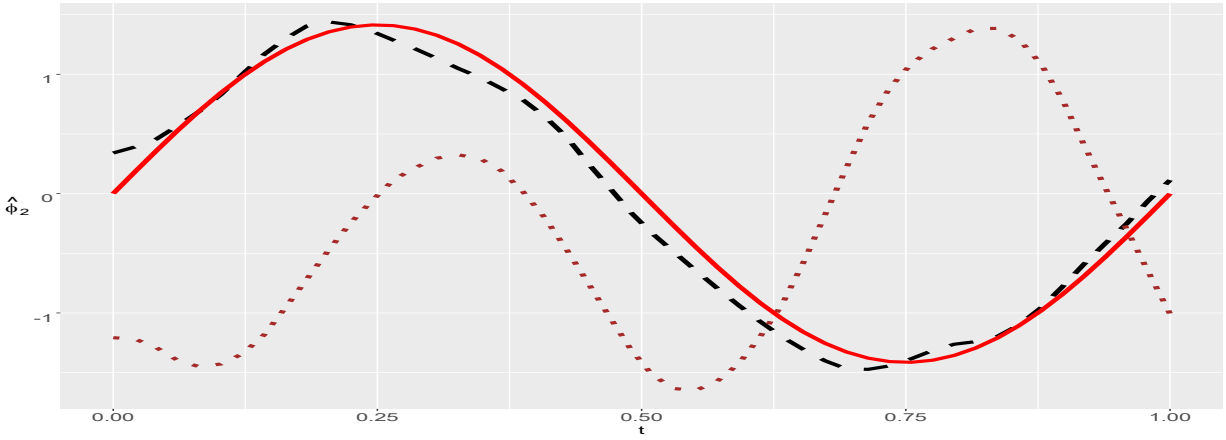


Figure 4.1: Data are generated from model (4.4) and Example 4.1. The solid line represents the true shared eigenfunction across two sites. The dashed line shows the estimated shared eigenfunction obtained using the proposed method in Section 4.2.3. The dotted line corresponds to the eigenfunction estimated from the averaged covariance operator.

Beyond naive averaging, much of the existing literature on multi-source functional data has focused on mean estimation under the framework of transfer learning. For example, Cai et al. (2024) studied phase transition phenomena in mean function estimation based on discretely sampled functional data and established the corresponding minimax rates. Another line of work concerns common FPCA (CFPCA) and partially common FPCA (PCFPCA). CFPCA assumes that the leading eigenfunctions for all sources are the same,

with differences arising only in the eigenvalues (Benko et al., 2009; Boente et al., 2010), whereas PCFPCA allows some eigenfunctions to be shared and others to remain source-specific (Jiao et al., 2024). For instance, Jiao et al. (2024) considered a forest-structured hierarchy of communities, where sources in each layer were partitioned into communities such that the data sources within the same community partially shared a subset of basis functions.

However, existing CFPCA and PCFPCA methods are designed for fully or densely observed functional data, which may be too restrictive in many real-world applications. Moreover, although the parameter of interest is often a low-dimensional shared subspace, most existing approaches rely on linear combinations of covariance functions estimated from each source. Such strategies remain prone to bias when the eigenvalues associated with source-specific directions dominate those of the shared subspace across sources, regardless of how the averaging weights are chosen. A more natural perspective is to focus directly on the subspace itself: if the shared subspace is the primary object of interest, one can safely discard the eigenvalues during the data integration process and retain only the eigenfunctions. This perspective motivates the projection-based approach we develop in this chapter.

To our knowledge, no existing methods directly pursue this subspace-focused strategy in the functional data literature to extract the shared subspace. The idea is inspired by Li et al. (2024), which aims to enhance the estimation efficiency of the shared structure under the framework of transfer learning for multivariate data. They identified the shared space by solving an optimization problem on the Grassmann manifold, where the manifold is often defined for finite-dimensional Euclidean space.

We generalize the method to functional data, where the underlying space is infinite-dimensional. Specifically, we first estimate the sample covariance function for each source using a local linear smoother, thereby accommodating the nonparametric nature of functional data. Aggregating observations within each source allows us to borrow strength across subjects, enabling estimation even for sparse functional data. Discretizing the smoothed covariance estimates yields estimated eigenpairs and corresponding sample projection operators for each source. We then construct a weighted average projection estimator, $\widehat{\boldsymbol{P}}_w$, from the source-wise sample projection operators. By utilizing the spectral

properties of $\widehat{\boldsymbol{P}}_w$, we identify the shared subspace, which is then used to estimate the source-specific subspace by re-projecting the estimated covariance operator of the specific source onto the orthogonal complement of the shared subspace.

Furthermore, we establish theoretical convergence guarantees for both the shared and source-specific estimators. The convergence rate of the estimator of the shared projection operator highlights the gain in estimation efficiency achieved by effectively leveraging information from multiple data sources. These theoretical insights align with the numerical results from simulation studies under various settings, where the proposed method consistently outperforms single-source estimators.

The remainder of this chapter is organized as follows. In Section 4.2.1, we introduce the data generation mechanism and standard FPCA for one source. Then, we rigorously define the shared structure for functional data in Section 4.2.2. The proposed method is detailed in Section 4.2.3. We develop asymptotic results for the proposed estimator in Section 4.3. Detailed proofs are relegated to the online supplement. We showcase the finite-sample performance of the proposed method under various simulation settings in Section 4.4. Finally, we apply the proposed method to a real world example in Section 4.5.

4.2 Estimation via Data Integration

4.2.1 Standard FPCA

Before presenting the proposed method for multiple data sources, we begin by reviewing the standard FPCA for a single source and introducing the notation. Let $X(t)$ be a continuous and square-integrable stochastic process defined on a compact interval $\mathcal{T} = [0, 1]$ with $\mathbb{E}[X(t)] = 0$ for $t \in \mathcal{T}$. Denote $\mathcal{H} = \mathcal{L}^2(\mathcal{T})$, the Hilbert space of square-integrable functions on $\mathcal{T}$, equipped with inner product $\langle f, g\rangle_2 = \int_{t\in\mathcal{T}} f(t)g(t)\,\mathrm{d}t$. Moreover, $\|f\|_2^2 = \langle f, f\rangle_2$. As presented in Section 1.2, FPCA seeks to find a projection operator $\widetilde{\boldsymbol{P}}$ of rank m that minimizes the expected mean-squared distance between X and its projection, which is formalized as follows:

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P}\in\mathcal{P}_m}{\arg\min}\, \mathbb{E}\|X - \boldsymbol{P}X\|_2^2. \tag{4.1}$$

Here, $\mathcal{P}_m = \{\boldsymbol{P} \in \mathcal{B}(\mathcal{H}) \mid \boldsymbol{P}^2 = \boldsymbol{P}, \operatorname{rank}(\boldsymbol{P}) = m, \boldsymbol{P} \text{ is self-adjoint}\}$, and $\mathcal{B}(\mathcal{H})$ denotes the set of bounded linear operators on $\mathcal{H}$. In addition, any $\boldsymbol{P} \in \mathcal{P}_m$ can be written as $\boldsymbol{P} = \sum_{\nu=1}^{m} e_\nu \otimes e_\nu$, where $f \otimes g(\cdot) = \langle f, \cdot \rangle\, g$ is a rank-one operator on $\mathcal{H}$ for $f, g \in \mathcal{H}$.

An equivalent problem to the optimization problem in (4.1) is

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P} \in \mathcal{P}_m}{\arg\min}\, \mathbb{E}\|X - \boldsymbol{P}X\|_2^2 - \mathbb{E}\|X\|_2^2 = \underset{\boldsymbol{P} \in \mathcal{P}_m}{\arg\max}\, \langle \boldsymbol{G}, \boldsymbol{P} \rangle_{\mathrm{HS}}\,, \tag{4.2}$$

where $\boldsymbol{G}$ is the covariance operator induced by the covariance function of X. Specifically, for any $f \in \mathcal{H}$,

$$(\boldsymbol{G}f)(t) = \int_{s \in \mathcal{T}} G(s,t) f(s)\, \mathrm{d}s$$

with $G(s,t) = \mathbb{E}[X(s)X(t)]$ for $s, t \in \mathcal{T}$. Since G is continuous, symmetric, and positive semi-definite. By Mercer's theorem,

$$G(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s) \phi_\nu(t), \quad t, s \in \mathcal{T}, \tag{4.3}$$

where (λ_ν, ϕ_ν) is the νth eigenvalue-eigenfunction pair of $\boldsymbol{G}$ satisfying $\boldsymbol{G}\phi_\nu = \lambda_\nu \phi_\nu$, and $\{\phi_\nu\}_{\nu \geq 1}$ forms an orthonormal basis of $\mathcal{H}$. It is easy to see that $e_i = \phi_i$ for $i = 1, \ldots, m$ by combining equations (4.2) and (4.3) Therefore, the Karhunen–Loève expansion given in equation (1.6) implies that we can use the leading m eigenfunctions of $\boldsymbol{G}$ to approximate X.

As aforementioned in Section 1.1, functional data are often recorded at discrete points with noise. Suppose that there are n i.i.d. sample paths of X, with N_i observations for subject i, observed at times T_{ij} with additive measurement error:

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, N_i,$$

where we assume there are n i.i.d. sample paths of X, with N_i observations for subject i, observed at times T_{ij} with additive measurement error. Here N_i are i.i.d. copies of a random variable N; T_{ij} are subject-specific observation times in $\mathcal{T}$; and ε_{ij} are independent with mean zero and variance σ_ε^2. Moreover, N, T, ε, and X are assumed to be independent.

We first estimate the covariance function $G(t, s)$ by pooling observations across subjects, following the idea in Yao et al. (2005b), where a local linear smoother is employed. Next, we estimate the eigenvalue–eigenfunction pairs $(\lambda_\nu, \phi_\nu), \nu = 1, 2, \ldots, m$, from the discretization of the smoothed covariance function; see Rice and Silverman (1991) and Ramsay and Silverman (2005) for more details. The choice of m can be determined by the fraction of variance explained (FVE), information-criterion-based methods (Yao et al., 2005b; Li et al., 2013), or methods integrating information from both eigenvalues and eigenfunctions (Zhang et al., 2025a).

4.2.2 Multi-source Model

Suppose that there are K sources, and let $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the index set. For any $k \in \mathcal{K}$, we assume that the observations $Y_{ij}^{[k]}$ collected at $T_{ij}^{[k]}$ are generated from the following:

$$Y_{ij}^{[k]} = X_i^{[k]}\left(T_{ij}^{[k]}\right) + \varepsilon_{ij}^{[k]}, \quad i = 1, 2, \ldots, n_k, \ j = 1, 2, \ldots, N_i^{[k]}. \tag{4.4}$$

Analogous to the single-source model, we assume $N_i^{[k]}$ are i.i.d. copies of $N^{[k]}$, and $T_{ij}^{[k]}$'s are observation times in $\mathcal{T}$ for source k. Measurement errors $\varepsilon_{ij}^{[k]}$ are i.i.d with mean zero and variance $(\sigma_\varepsilon^{[k]})^2$. Moreover, $N^{[k]}$, $T^{[k]}$, $\varepsilon^{[k]}$ and $X^{[k]}$ are independent.

The Karhunen–Loève expansion of $X^{[k]}(t)$ for $k = 1, \ldots, K$, is

$$X^{[k]}(t) = \sum_{\nu=1}^{\infty} \xi_\nu^{[k]} \phi_\nu^{[k]}(t), \quad t \in \mathcal{T}, \tag{4.5}$$

where $\xi_\nu^{[k]} = \int_{\mathcal{T}} X^{[k]}(t)\phi_\nu^{[k]}(t)\,\mathrm{d}t$ are uncorrelated mean-zero random variables with variance $\lambda_\nu^{[k]}$. It is common to assume that the underlying process $X^{[k]}$ is well-approximated by the leading m_k eigenfunctions (Yao et al., 2005b). Moreover, the number of eigenfunctions that can be consistently estimated depends on the sample size and the number of observations per subject (Zhou et al., 2024). Throughout this chapter, we focus on the leading m_k functional principal components for source k. More precisely, we assume $\operatorname{rank}(\boldsymbol{P}_k) = m_k$ for $k \in \mathcal{K}$. This assumption is standard in the CFPCA/PCFPCA literature, where m_k is typically fixed and independent of n_k; see, for example, Benko et al. (2009) and Jiao et al.

(2024). To better characterize the inherent infinite-dimensional nature of functional data, we instead allow m_k to diverge as n_k grows. Finally, let $\mathcal{H}_0^{[k]}$ denote the subspace spanned by the leading m_k eigenfunctions of source k.

As shown in (4.3), the covariance structure is determined by its eigenfunctions, which capture variability patterns, and the corresponding eigenvalues, which quantify their magnitudes. Ideally, if the sample sizes and the number of observations per subject for different sources are roughly the same, the number of selected eigenfunctions is about the same. Furthermore, if $\phi_\nu^{[k]} = \phi_\nu$ for all $k \in \mathcal{K}$ and $\nu \in \mathbb{N}_+$, the shared subspace can be obtained from the leading eigenfunctions of the averaged sample covariance operators. In practice, however, these assumptions are often too restrictive. Real-world multi-source datasets may exhibit heterogeneity in covariance structures or imbalanced sample sizes, which introduces additional challenges for data integration in the following ways:

- The number of eigenfunctions that can be consistently estimated depends on the sample size and the average of the number of observations per subject. Consequently, m_k may vary across sources when n_k or the average $N_i^{[k]}$ differ substantially.
- For some source k, the space spanned by its leading m_k eigenfunctions can be decomposed into a shared subspace and a source-specific subspace. If the latter is not properly isolated, it may introduce bias into the recovery of the shared subspace.
- In some sources, the eigenfunctions associated with the shared subspace may correspond to relatively small eigenvalues, whereas source-specific eigenfunctions may be associated with larger eigenvalues. This imbalance can substantially degrade the performance of naive averaging approaches.

To integrate information across sources and address these challenges, we assume that there is a shared subspace $\mathcal{H}_s$ with $\dim(\mathcal{H}_s) = m_s$, spanned by the same eigenfunctions across all sources, where m_s diverges as the sample sizes increase. Thus, the shared structure can be formalized as follows: Let $\mathcal{I}_s^{[k]}$ contain the indices of the eigenfunctions that span $\mathcal{H}_s$ for source $k \in \mathcal{K}$. Then by (4.5), we rewrite $X^{[k]}(t)$ as

$$X^{[k]}(t) = \sum_{\nu \in \mathcal{I}_s^{[k]}} \xi_\nu^{[k]} \phi_\nu^{[k]}(t) + \sum_{\nu \notin \mathcal{I}_s^{[k]}} \xi_\nu^{[k]} \phi_\nu^{[k]}(t), \quad t \in \mathcal{T}.$$

It is important to note that while the shared eigenfunctions are identical across sources, their associated indices within each source may differ; that is, $\mathcal{I}_s^{[k]}$ is not necessarily equal to $\mathcal{I}_s^{[k']}$ for $k \neq k'$. However, the space spanned by these functions is invariant across $k \in \mathcal{K}$. Specifically, for each $k \in \mathcal{K}$,

$$\mathcal{H}_s^{[k]} \triangleq \left\{ f = \sum_{\nu=1}^{m_s} \alpha_\nu \phi_{\pi_k(\nu)} \mid \alpha_1, \alpha_2, \ldots, \alpha_{m_s} \in \mathbb{R} \right\} = \mathcal{H}_s,$$

where $\pi_k : \{1, 2, \ldots, m_s\} \to \mathcal{I}_s^{[k]}$ is a bijective function. This shared structure is similar to that defined in Jiao et al. (2024). Accordingly, define the projection operators

$$\boldsymbol{P}_s^{[k]} = \sum_{\nu \in \mathcal{I}_s^{[k]}} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}, \quad \boldsymbol{P}_p^{[k]} = \sum_{\nu \notin \mathcal{I}_s^{[k]},\ \nu \leq m_k} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}, \tag{4.6}$$

where $\boldsymbol{P}_s^{[k]}$ projects onto the shared subspace and $\boldsymbol{P}_p^{[k]}$ onto the source-specific subspace.

The following toy example illustrates the shared structure.

Example 4.1. *Consider the Fourier basis functions:*

$$\phi_1(t) = 1, \quad \phi_\nu(t) = \begin{cases} \sqrt{2}\sin(\nu\pi t) & \text{if } \nu \geq 2 \text{ and } \nu \text{ is even,} \\ \sqrt{2}\cos\{(\nu-1)\pi t\} & \text{if } \nu \geq 2 \text{ and } \nu \text{ is odd.} \end{cases}$$

Consider two sources with covariance operators $\boldsymbol{G}^{[k]} = \sum_{\nu=1}^{\infty} \lambda_\nu^{[k]} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}$ *for* $k = 1, 2$. *For source 1, set* $\lambda_\nu^{[1]} = \nu^{-2}$ *and* $\phi_\nu^{[1]} = \phi_\nu$ *for* $\nu \in \mathbb{N}_+$. *For source 2, we assume* $\lambda_2^{[2]} = \lambda_3^{[1]}$, $\lambda_3^{[2]} = \lambda_2^{[1]}$, *and* $\lambda_\nu^{[2]} = \lambda_\nu^{[1]}$ *for all other* $\nu \in \mathbb{N}_+$. *Define the eigenfunctions as*

$$\phi_1^{[2]} = 2^{-1/2}(\phi_4 + \phi_5), \quad \phi_2^{[2]} = \phi_2, \quad \phi_3^{[2]} = \phi_3, \quad \phi_4^{[2]} = \phi_1,$$

and, for $\nu \geq 5$,

$$\phi_\nu^{[2]} = 2^{-1/2}(\phi_{2\nu-4} + \phi_{2\nu-3}).$$

If $m_1 = m_2 = 3$, *then the shared dimension is* $m_s = 2$. *In this case,* $\pi_1(\nu) = \nu$ *for* $\nu \leq m_s$, *while* $\pi_2(1) = 3$ *and* $\pi_2(2) = 2$.

Remark 4.1. *In this construction, the eigenfunction associated with the largest eigenvalue always corresponds to the source-specific variation, whereas the shared subspace lies in the tail and is associated with relatively small eigenvalues. Consequently, weighting sample covariance operators and extracting leading eigenfunctions to identify the shared subspace may still perform poorly regardless of the choice of weights.*

4.2.3 Multi-source FPCA

To identify the shared structure across multiple sources and generalize the standard FPCA, we consider focusing exclusively on the eigenfunctions while discarding the eigenvalues during the integration step. If source-specific variation is of further interest, analysis can be refined within the source of interest based on the estimated shared structure.

Motivated by the above idea and Example 4.1, we propose the following projection-based estimator:

$$\widehat{\boldsymbol{P}}_w = n_t^{-1} \sum_{k \in \mathcal{K}} n_k \widetilde{\boldsymbol{P}}^{[k]}, \tag{4.7}$$

where $n_t = \sum_{k \in \mathcal{K}} n_k$ and $\widetilde{\boldsymbol{P}}^{[k]} = \sum_{\nu=1}^{m_k} \tilde{\phi}_\nu^{[k]} \otimes \tilde{\phi}_\nu^{[k]}$. Here $\tilde{\phi}_\nu^{[k]}$ is the νth eigenfunction estimated from the smoothed covariance function of the kth source. Note that, consistency of $\tilde{\phi}_\nu^{[k]}$ implies the consistency of $\widetilde{\boldsymbol{P}}^{[k]}$.

The target of interest is the leading m_s eigenfunctions of $\widehat{\boldsymbol{P}}_w$, or equivalently, the subspace they span. Let $\widehat{\boldsymbol{P}}_s$ denote the projection operator onto this estimated shared subspace. Then, it follows from (4.2) that

$$\begin{aligned}
\widehat{\boldsymbol{P}}_s &= \underset{\boldsymbol{P} \in \mathcal{P}_{m_s}}{\arg\max} \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \left\langle \widetilde{\boldsymbol{P}}^{[k]}, \boldsymbol{P} \right\rangle_{\mathrm{HS}} \\
&= \underset{\boldsymbol{P} \in \mathcal{P}_{m_s}}{\arg\min} \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \|\widetilde{\boldsymbol{P}}^{[k]} - \boldsymbol{P}\|_{\mathrm{HS}}^2,
\end{aligned}$$

where $\mathcal{P}_{m_s}$ denotes the collection of rank-m_s projection operators in $\mathcal{H}$. Therefore, the projection operator of the shared structure minimizes the weighted average of Hilbert–Schmidt distances between $\widetilde{\boldsymbol{P}}^{[k]}$ and $\boldsymbol{P}$ across all sources.

Note that at the population level,

$$\boldsymbol{P}_w = n_t^{-1} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}^{[k]} = \boldsymbol{P}_s + \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]}, \tag{4.8}$$

which implies that the eigenvalues corresponding to the shared subspace are always equal to one. While the largest eigenvalue of $\sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} / n_t$ is strictly smaller than one. Since $\widetilde{\boldsymbol{P}}^{[k]}$ is consistent, $\widehat{\boldsymbol{P}}_w$ is also consistent. Consequently, the rank of the shared subspace can be identified from the scree plot of $\widehat{\boldsymbol{P}}_w$; see Figure 4.2 for illustration. Further discussions

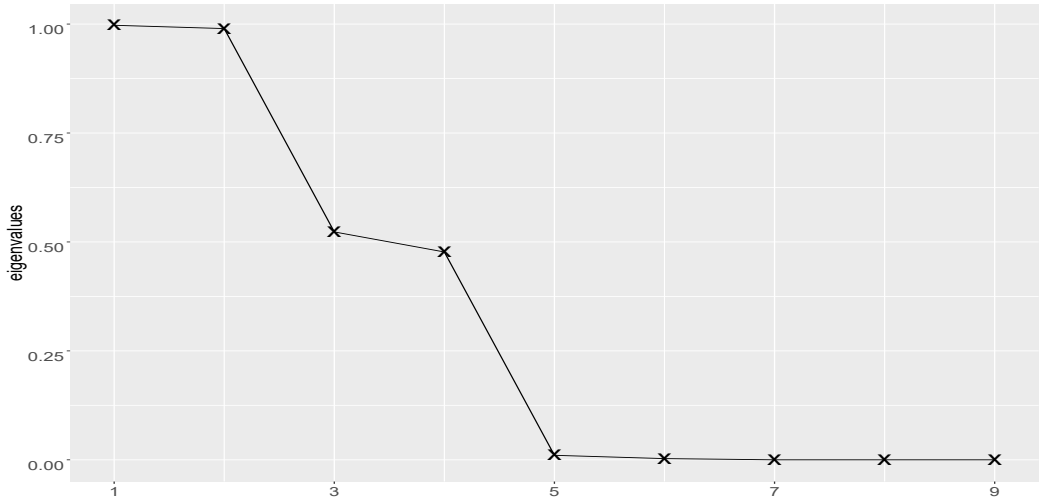


Figure 4.2: Scree plot of $\widehat{\boldsymbol{P}}_w$ based on Example 4.1 with $m_1 = m_2 = 3$. Data are generated from two sources, each with 200 subjects and 25 observations per subject.

of the eigengap between the shared subspace and $\sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} / n_t$ are provided in Remark 4.2 and Example 4.2. Let $(\hat{\theta}_\nu, \hat{\psi}_\nu)$ denote the νth eigen-pair of $\widehat{\boldsymbol{P}}_w$. Therefore, we have

$$\widehat{\boldsymbol{P}}_s = \sum_{\nu=1}^{m_s} \hat{\psi}_\nu \otimes \hat{\psi}_\nu. \tag{4.9}$$

If source-specific variation for source k is of interest and $m_k > m_s$, the corresponding projector $\boldsymbol{P}_p^{[k]}$, which is defined in (4.6), can be estimated using only the kth source data after obtaining $\widehat{\boldsymbol{P}}_s$. The space associated with $\widehat{\boldsymbol{P}}_s$ can be viewed as an average of the

shared subspace across K sources. Consequently, the basis functions generating $\widehat{\boldsymbol{P}}_s$ are not necessarily orthogonal to those generating the source-specific variation. To preserve orthogonality, we need to define a projection operator $\widehat{\boldsymbol{P}}_d$ that projects functions onto $\widehat{\mathcal{H}}_d^{[k]} = \widetilde{\mathcal{H}}_0^{[k]} \ominus \widehat{\mathcal{H}}_s$, where $\widetilde{\mathcal{H}}_0^{[k]}$ and $\widehat{\mathcal{H}}_s$ are two subspaces associated with $\widetilde{\boldsymbol{P}}^{[k]}$ and $\widehat{\boldsymbol{P}}_s$, respectively. Define $\widehat{\boldsymbol{T}}^{[k]} = \widetilde{\boldsymbol{P}}^{[k]}(\mathrm{id} - \widehat{\boldsymbol{P}}_s)\widetilde{\boldsymbol{P}}^{[k]}$, where id is the identity operator that maps any functions in $\mathcal{H}$ to itself. Denote by $\widehat{\boldsymbol{P}}_d$ the projection operator generated by the eigenfunctions of $\widehat{\boldsymbol{T}}^{[k]}$ associated with the eigenvalue one. Note that $\widehat{\boldsymbol{T}}^{[k]}(f) = f$ for any $f \in \widehat{\mathcal{H}}_d^{[k]}$, which implies $\widehat{\boldsymbol{P}}_d(f) = f$. The construction of $\widehat{\boldsymbol{P}}_d$ implicitly forces any function not in $\widetilde{\mathcal{H}}_0^{[k]} \cap \widehat{\mathcal{H}}_s^{\perp}$ to be zero as we set these eigenvalues to be 0. Thus, $\widehat{\boldsymbol{P}}_d(f) = 0$ for any $f \in \mathcal{H} \ominus \widehat{\mathcal{H}}_d^{[k]}$. To ensure the eigenfunctions in the source-specific subspace are identifiable, we then project $\widetilde{\boldsymbol{G}}^{[k]}$, the estimated covariance function for the kth source, onto $\widehat{\boldsymbol{P}}_d^{[k]}$. The source-specific subspace $\widehat{\boldsymbol{P}}_p^{[k]}$ is then represented by the leading $m_k - m_s$ eigenfunctions of

$$\widehat{\boldsymbol{G}}_p^{[k]} = \widehat{\boldsymbol{P}}_d^{[k]} \widetilde{\boldsymbol{G}}^{[k]} \widehat{\boldsymbol{P}}_d^{[k]}.$$

The refined estimator of projection operator for the kth source is given by $\widehat{\boldsymbol{P}}^{[k]} = \widehat{\boldsymbol{P}}_p^{[k]} + \widehat{\boldsymbol{P}}_s$. The entire procedure is summarized in Algorithm 3.

4.3 Statistical Theory

This section is devoted to the theoretical justification of the proposed method. The shared structure is formalized in Section 4.2.2. In what follows, let $\|\boldsymbol{A}\|_{\mathrm{op}} = \sup_{f \in \mathcal{H}} \|\boldsymbol{A}(f)\|_2 / \|f\|_2$ denote the operator norm of $\boldsymbol{A}$. For a compact operator, the operator norm equals its largest eigenvalue. For any continuous function f defined on $[0,1]$, $\|f\|_\infty = \sup_{t \in [0,1]} |f(t)|$. Let $a_n \lesssim b_n$ denote $a_n \leq C b_n$ for all $n \in \mathbb{N}^+$ with some positive constant C, and $\gtrsim$ is defined similarly. Lastly, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We need the following assumptions to develop theoretical properties for the proposed estimator.

Assumption 4.1. *The shared space $\mathcal{H}_s$ is spanned by the same basis functions across all sources. Furthermore, we assume $m_s \leq \min_{k \in \mathcal{K}} m_k$.*

Algorithm 3 Multi-source FPCA

1. For each source k, estimate the covariance function $\widetilde{G}^{[k]}(t,s)$ and eigenpairs $\{(\tilde{\lambda}_\nu^{[k]}, \tilde{\phi}_\nu^{[k]}) \mid \nu = 1, 2, \ldots, m_k\}$ using the method detailed in Section 4.2.1. Moreover, compute $\widetilde{\boldsymbol{P}}^{[k]} = \sum_{\nu=1}^{m_k} \tilde{\phi}_\nu^{[k]} \otimes \tilde{\phi}_\nu^{[k]}$.

2. Construct the weighted average projection operator $\widehat{\boldsymbol{P}}_w$ as in equation (4.7), and perform eigendecomposition of its discretized version to obtain eigenpairs $(\hat{\theta}_\nu, \hat{\psi}_\nu)$.

3. Identify the rank m_s of $\mathcal{H}_s$ based on the eigenvalues of $\widehat{\boldsymbol{P}}_w$. The projection operator of the shared subspace is determined by using equation (4.9).

4.1 (optional) If source-specific variation for source k is of interest and $m_k > m_s$, obtain $\widehat{\boldsymbol{P}}_d^{[k]}$ by acquiring the eigenspace of $\widetilde{\boldsymbol{P}}^{[k]}(\mathrm{id} - \widehat{\boldsymbol{P}}_s)\widetilde{\boldsymbol{P}}^{[k]}$ associated with the eigenvalue equals one. Then, we re-project $\widetilde{\boldsymbol{G}}^{[k]}$ onto the space associated with $\widehat{\boldsymbol{P}}_d^{[k]}$ to obtain $\widehat{\boldsymbol{G}}_p^{[k]} = \widehat{\boldsymbol{P}}_d^{[k]} \widetilde{\boldsymbol{G}}^{[k]} \widehat{\boldsymbol{P}}_d^{[k]}$.

4.2 Extract the leading $m_k - m_s$ eigenfunctions of $\widehat{\boldsymbol{G}}_p^{[k]}$ to estimate the source-specific subspace and its associated projection operator $\widehat{\boldsymbol{P}}_p^{[k]}$.

Assumption 4.2. *For the source-specific subspaces, we assume*

$$\left\| \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} \right\|_{op} \leq 1 - d,$$

where $d > 0$ is a constant that is independent of sample sizes.

Remark 4.2. *Note that the largest eigenvalue of $\boldsymbol{P}_w$ defined in (4.8) satisfies $\lambda_1(P_w) = 1$. Assumption 4.2 essentially provides a lower bound on the discrepancy between the largest two eigenvalues of P_w. The magnitude of $1-d$ characterizes the degree of similarity among the source-specific subspaces across sources. A smaller d indicates greater challenges in estimating the shared space. In particular, if $d = 0$, then an eigenfunction is shared by all sources, and by Assumption 4.1, this eigenfunction must belong to $\mathcal{H}_s$.*

Example 4.2. *Let us revisit Example 4.1 and further assume $n_1 = n_2$. It is easy to see that $\left\| 2^{-1} \sum_{k \in \mathcal{K}} \boldsymbol{P}_p^{[k]} \right\|_{op} = 1/2$, and $d = 1/2$. If instead, we let $\phi_1^{[2]} = 2^{-1/2}(\phi_1 + \phi_4)$ and $\phi_4^{[2]} = \phi_5$ while keeping $m_1 = m_2 = 3$, then $\| 2^{-1} \sum_{k \in K} \boldsymbol{P}_p^{[k]} \|_{op} = 1/2 + 2^{1/2}/4$, and $d = 1/2 - 2^{1/2}/4$. Thus, d is smaller in second case as $\phi_1^{[2]}$ is partially aligned with $\phi_1^{[1]} = \phi_1$, indicating an overlap between two source-specific subspaces. In contrast, if $\phi_1^{[2]}$ is completely aligned with $\phi_1^{[1]}$, then $m_s = 3$ and this eigenfunction belongs to the shared space.*

Assumptions 4.3 - 4.8 should be understood as they hold for any $k \in \mathcal{K}$. For brevity, we omit the superscript $[k]$ or k. These assumptions ensure that the covariance function and eigenpairs are well estimated for each source.

Assumption 4.3. *There exists a positive constant c_0 such that $\mathbb{E}\xi_\nu^4 \leq c_0(\lambda_\nu)^2$.*

Assumption 4.4. *The second order partial derivatives of $G(s,t)$, $\partial G(s,t)/\partial t \partial s$, $\partial G(s,t)/\partial s^2$, and $\partial G(s,t)/\partial t^2$, are all bounded on $\mathcal{T} \times \mathcal{T}$.*

Assumption 4.5. *The eigenvalues of $\boldsymbol{G}$ follow a polynomial decay rate. That is, there exists a positive constant $a > 1$ such that, for any $\nu \in \mathbb{N}_+$,*

$$\nu^{-a} \gtrsim \lambda_\nu \gtrsim \lambda_{\nu+1} \gtrsim \nu^{-(a+1)}$$

Assumption 4.6. *Suppose* $\|\phi_\nu\|_\infty = O(1)$ *for* $\nu \in \mathbb{N}_+$, *and there exists a positive constant* c

$$\|\phi_\nu^{(\ell)}\|_\infty \lesssim \nu^{c/2}\|\phi_\nu^{(\ell-1)}\|_\infty \quad \text{for } \ell = 1, 2,$$

where $f^{(\ell)}$ *denotes the* ℓ*th derivative of* f.

Assumption 4.7. $\mathbb{E}(\|X\|_\infty^{2\alpha})$ *and* $\mathbb{E}(|\varepsilon|^{2\alpha})$ *are bounded for some* $\alpha > 3$.

Assumption 4.8. *The observation times* T_{ij} *are independent of* X *and* ε. *Moreover,* T_{ij}*'s are iid with a density function bounded away from zero and infinity on* $\mathcal{T}$.

Assumptions 4.3 and 4.4 are widely adopted in the functional data literature when smoothing methods are employed; see Yao et al. (2005b) and Cai and Yuan (2011) for example. Assumption 4.5 considers polynomial decay rates for eigenvalues, where the resulting lower bounds of eigengaps facilitate a theoretical framework for the estimated functional principal components (Hall and Horowitz, 2007; Zhou et al., 2023). Assumption 4.6 was considered in Zhou et al. (2024) to control the smoothness of eigenfunctions and their derivatives. In particular, it is easy to see that $\|\phi_\nu^{(2)}\|_\infty \lesssim \nu^c$, which implies that higher order derivatives can grow up to a polynomial rate with respect to the index ν of the eigenfunction. The moments condition in Assumption 4.7 ensures the uniform convergences of the estimated covariance function, which can be found, for example, in Li and Hsing (2010) and Zhang and Wang (2016).

We first present the convergence rate of $\widehat{\boldsymbol{P}}_s$ in terms of the operator norm. Note that the projection operator is the aggregation of the projection operator for each source. Thus, Theorems 4.1 and 4.2 are based on the FPCA theory for a single source.

Theorem 4.1. *Suppose Assumptions 4.1 and 4.2 hold. For each source* k, *suppose Assumptions 4.3 - 4.8 hold with respective constants, and denote*

$$\bar{N}_2^{[k]} = \left\{ n_k^{-1} \sum_{i=1}^{n_k} (N_i^{[k]})^{-2} \right\}^{-1/2}.$$

For any $k \in \mathcal{K}$, we further assume $m_k^{2a_k+2}/n_k = o(1)$, $\bar{N}_2^{[k]} \gtrsim m_k$, and for $\nu \leq m_k$,

$$\frac{m_k^{2a_k+2}}{n_k(\bar{N}_2^{[k]})^2 h_k^2(\nu)} = o(1), \ h_k^4(\nu)\max\{m_k^{2a_k+2c_k}, m_k^{4a_k}\log(n_k)\} = o(1),$$

where

$$h_k(\nu) = (n_k\bar{N}_2^{[k]})^{-1/5}\nu^{(a_k-2c_k-2)/5}(1+\nu^{a_k}/\bar{N}_2^{[k]})^{1/5}.$$

Under these assumptions, we have

$$\|\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s\|_{op} = O_p(\kappa_{n_t}).$$

Here,

- *When $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$, we further assume $c_k + 2 < 4a_k$ for all $k \in \mathcal{K}$. Then,*

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}} + \frac{\left(\sum_{k\in\mathcal{K}} m_k^{c_k/2+3}\right)^{2/5}}{n_t^{2/5}}.$$

  If $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$ for all $k \in \mathcal{K}$, then

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}}.$$

- *When $\bar{N}_2^{[k]} \asymp m_k$ for all $k \in \mathcal{K}$, we further assume $\max_{k\in\mathcal{K}} 2(a_k+1)/(a_k+c_k/2+1) > 1$. Then,*

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^{2a_k+2}\right)^{1/2}}{n_t^{1/2}} + \frac{\left(\sum_{k\in\mathcal{K}} m_k^{2a_k+c_k/2+1}\right)^{2/5}}{n_t^{2/5}}.$$

Remark 4.3. *The additional conditions on a_k and c_k arise from aggregating estimated eigenfunctions within the shared space, which are very mild. For example, standard bases used in FDA, including Fourier, Legendre, and wavelet, satisfy Assumption 4.6 with $c_k = 2$. The polynomial decay rate for eigenvalues in Assumption 4.5 normally assumes $a_k > 1$. Consequently, both $c_k + 2 < a_k$ and $\max_{k\in\mathcal{K}} 2(a_k+1)/(a_k+c_k/2+1) > 1$ hold.*

Remark 4.4. *To illustrate the efficiency gain from data integration, consider the following example. For simplicity, we focus on the dense case, where* $\kappa_{n_t} = \left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}/n_t^{1/2}$. *Suppose there exists a source* k_0 *such that* $n_{k_0} \asymp n_t$ *and* $n_k/n_{k_0} = o(1)$ *for all* $k \neq k_0$. *When* $m_k \asymp m$ *for all* k, *we have* $\kappa_{n_t} \asymp m^2/n_{k_0}^{1/2}$. *In contrast, if only source* k *is available for some* $k \neq k_0$, *the convergence rate for estimating the projection operator of the shared subspace is* $O_p[(m_s m)/n_k^{1/2}]$; *see Proposition 4.1 in the supplementary material. This rate is based on the oracle setting where the true indices of the shared eigenfunctions are known for source* k. *The ratio between the two convergence rates is* $(mn_k)^{1/2}/(m_s n_{k_0})$. *For example, if* $m_s \asymp m$, *the ratio is* $o(1)$. *This demonstrates the improvement in estimation efficiency achieved through data integration. This theoretical result aligns with the numerical findings reported in Table 4.1 in Section 4.4.*

Theorem 4.2. *Under Assumptions for Theorem 4.1, if* $m_k > m_s$, *then the estimator of the* k*th source-specific subspace satisfies*

$$\left\|\widehat{\boldsymbol{P}}_p^{[k]} - \boldsymbol{P}_p^{[k]}\right\|_{op} = O_p(\zeta_{n_k}).$$

Here,

- *When* $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$,

$$\zeta_{n_k} = \frac{(m_k - m_s)^2}{n_k^{1/2}} + \left\{\frac{(m_k - m_s)^{c_k/2+3}}{n_k}\left(\frac{m_k - m_s}{m_k}\right)^{a_k}\right\}^{2/5}.$$

  If we further assume $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k + c_k/2 - 2}$, *then*

$$\zeta_{n_t} = \frac{(m_k - m_s)^2}{n_k^{1/2}}.$$

- *When* $\bar{N}_2^{[k]} \asymp m_k$,

$$\zeta_{n_k} = \frac{(m_k - m_s)^{a_k+1}}{n_k^{1/2}} + \frac{(m_k - m_s)^{(4a_k + c_k + 2)/5}}{n_k^{2/5}}.$$

Remark 4.5. *Let the source-specific subspace for source k_0 be denoted by $\mathcal{H}_d^{[k_0]} = \mathcal{H}_0^{[k_0]} \ominus \mathcal{H}_s$. When $m_{k_0} > m_s$, it is not surprising that the convergence rate in Theorem 4.2 is of the same order as that of a single-source estimator, unless $\dim(\mathcal{H}_d^{[k_0]}) \ll \dim(\mathcal{H}_s)$. In particular, we focus on the dense case discussed in Remark 4.4, and suppose $m_k \asymp m$ for all $k \in \mathcal{K}$. In the oracle setting where the true indices of the eigenfunctions spanning $\mathcal{H}_d^{[k_0]}$ are known, the convergence rate for estimating the projection operator of the source-specific subspace is $O_p[(m - m_s)m/n_{k_0}^{1/2}]$ based on Proposition 4.1 in the supplementary material, while Theorem 4.2 gives $\zeta_{n_{k_0}} = (m - m_s)^2/n_{k_0}^{1/2}$. Thus, the ratio between the two convergence rates is $1/(1 + r)$, where $r = \dim(\mathcal{H}_s)/\dim(\mathcal{H}_d^{[k_0]})$. This result implies that if $\dim(\mathcal{H}_d^{[k_0]}) \asymp \dim(\mathcal{H}_s)$, the efficiency gain is asymptotically constant; however, when $\dim(\mathcal{H}_d^{[k_0]}) \ll \dim(\mathcal{H}_s)$, the gain becomes substantial.*

For example, in the extreme case where $\dim(\mathcal{H}_d^{[k_0]}) = 1$ and $m_s \asymp m \to \infty$ as $n_k \to \infty$ for all $k \in \mathcal{K}$, the estimated subspace $\widehat{\mathcal{H}}_d^{[k_0]}$ must be orthogonal to the estimated shared subspace, and accurate estimation of $\mathcal{H}_s$ markedly improves the recovery of $\mathcal{H}_d^{[k_0]}$. Conversely, if $\dim(\mathcal{H}_s) = 1$ but $\dim(\mathcal{H}_d^{[k_0]}) \to \infty$ as $n_k \to \infty$, improvements in estimating $\mathcal{H}_s$ have negligible effect on the estimation of $\mathcal{H}_d^{[k_0]}$. In summary, although the information for estimating the source-specific subspace resides solely within source k_0, the orthogonality constraint between $\mathcal{H}_s$ and $\mathcal{H}_0^{[k_0]} \ominus \mathcal{H}_s$ can help enhance estimation efficiency.

4.4 Simulation

In this section, we perform simulation studies to investigate the finite sample performance of the proposed method, denoted as Multi-FPCA. We consider three sources in total. To showcase the superior performance of the proposed method, we consider three different sample size settings: 1.) $n_1 = 50$ and $n_2 = n_3 = 200$; 2.) $n_1 = 100, n_2 = n_3 = 200$; 3.) $n_1 = 100, n_2 = n_3 = 400$. In addition, we consider the number of observations per subject $N_i = N \in \{10, 25, 50\}$, where $N = 10$ and 50 are referred to as sparse and dense functional data, respectively. The intermediate case, $N = 25$, represents a transition state between sparse and dense, hereby referred to as neither.

We generate data based on model (4.4) and the Karhunen–Loève expansion in (4.5).

Observation times $T_{ij}^{[k]}$ are uniformly generated in $[0,1]$. For brevity, we assume the true mean function is zero. In terms of the covariance structure, we first denote the ordinary Fourier basis as follows:

$$\phi_1(t) = 1, \quad \phi_\nu(t) = \begin{cases} \sqrt{2}\sin(\nu\pi t) \text{ if } \nu \geq 2 \text{ and } \nu \text{ is even}, \\ \sqrt{2}\cos\{(\nu-1)\pi t\} \text{ if } \nu \geq 2 \text{ and } \nu \text{ is odd}. \end{cases}$$

The eigen-systems for the three sources are specified as follows:

Source 1: Set $\lambda_1^{[1]} = 24, \lambda_2^{[1]} = 12, \lambda_3^{[1]} = 6$, and $\lambda_\nu = 0$ for $\nu \geq 4$. The eigenfunctions are $\phi_\nu^{[1]} = \phi_\nu$ for $\nu \in \mathbb{N}_+$. The variance of the measurement error, σ_ε^2, is 0.49.

Source 2: Set $\lambda_1^{[2]} = 12, \lambda_2^{[2]} = 10, \lambda_3^{[2]} = 8$, $\lambda_4^{[2]} = 4$, and $\lambda_\nu = 0$ for $\nu \geq 5$. As for the eigenfunctions, $\phi_1^{[2]} = \phi_3$, $\phi_2^{[2]} = (\sin\theta)\phi_4 + (\cos\theta)\phi_5$, $\phi_3^{[2]} = \phi_2$, $\phi_4^{[2]} = (\sin\theta)\phi_6 + (\cos\theta)\phi_7$ with $\theta = \pi/4$. The variance of the measurement error, σ_ε^2, is 0.16.

Source 3: Set $\lambda_1^{[3]} = 20, \lambda_2^{[3]} = 10, \lambda_3^{[3]} = 5$, and $\lambda_\nu = 0$ for $\nu \geq 4$. The eigenfunctions are $\phi_1^{[3]} = (\sin\omega)\phi_4 + (\cos\omega)\phi_5$, $\phi_2^{[3]} = \phi_2$, $\phi_3^{[3]} = \phi_3$ with $\omega = \pi/3$. The variance of the error, σ_ε^2, is 0.16.

The above settings imply $m_s = 2$.

To illustrate the benefit of integrating information, we compare Multi-FPCA with an estimator using only source 1 data. For the latter, we first estimate $\widetilde{\boldsymbol{P}}^{[1]}$ with the method in Section 4.2.1. We then compute $\widetilde{\boldsymbol{P}}_s^{[1]}$ and $\widetilde{\boldsymbol{P}}_p^{[1]}$ by providing the true indices of the shared eigenfunctions in the first source. In particular, $\widetilde{\boldsymbol{P}}_s^{[1]} = \tilde{\phi}_2^{[1]} \otimes \tilde{\phi}_2^{[1]} + \tilde{\phi}_3^{[1]} \otimes \tilde{\phi}_3^{[1]}$. It is worth noting that the true indices are provided in computing $\widetilde{\boldsymbol{P}}_s^{[1]}$ and $\widetilde{\boldsymbol{P}}_p^{[1]}$, while the indices of the shared eigenfunctions are always estimated from the data for Multi-FPCA. The estimation procedure for the Multi-FPCA is detailed in Section 4.2.3 and summarized in Algorithm 3. 500 independent simulation trials were run for each setting mentioned above.

To assess the performance of these methods, we consider the following metric:

$$\frac{1}{M}\sum_{i=1}^{M} \|\widehat{\boldsymbol{P}}_i - \boldsymbol{P}\|,$$

where $M = 500$ and $\|\cdot\|$ denotes either the operator norm or Hilbert–Schmidt norm. Here $\widehat{\boldsymbol{P}}_i$ denotes the generic estimate of the target (sub)space of interest, i.e., the entire space, the shared subspace or the source-specific subspace, based on Multi-FPCA or using the source 1 data only from the ith simulation run, and $\boldsymbol{P}$ is the corresponding true (sub)space.

We report both the averaged norm differences and the associated standard error. The results of estimating the shared subspace and the source-specific subspace for source 1 are displayed in Table 4.1 and Table 4.2, respectively. Table 4.3 summarizes the results for estimating the entire space for source 1.

Table 4.1: Mean and standard error (in parenthesis) comparison of the norm differences for the *shared* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.188(0.06)	0.294(0.09)	0.306(0.10)	0.475(0.15)
	25	0.117(0.03)	0.188(0.04)	0.251(0.09)	0.388(0.13)
	50	0.089(0.02)	0.144(0.03)	0.247(0.11)	0.377(0.14)
$(100, 200, 200)$	10	0.160(0.04)	0.254(0.06)	0.238(0.07)	0.368(0.11)
	25	0.101(0.02)	0.166(0.04)	0.196(0.07)	0.306(0.10)
	50	0.077(0.01)	0.126(0.02)	0.181(0.07)	0.279(0.09)
$(100, 400, 400)$	10	0.142(0.05)	0.224(0.06)	0.237(0.07)	0.368(0.10)
	25	0.088(0.02)	0.142(0.03)	0.189(0.06)	0.296(0.09)
	50	0.067(0.02)	0.108(0.02)	0.176(0.06)	0.272(0.08)

Generally, the estimation accuracy improves for both methods as sample sizes or the number of observations per subject increases in terms of the norm differences. Tables 4.1 - 4.3 evidently illustrate the superior performance of Multi-FPCA. From Table 4.1, we find Multi-FPCA always outperforms the method using only single source data even if the true indices of the shared eigenfunctions are provided for the latter case. This confirms that Multi-FPCA effectively recovers the shared structure by leveraging information from additional sources, which also helps to reduce the estimation uncertainty. The improved

Table 4.2: Mean and standard error (in parenthesis) comparison of the norm differences for the *source-specific* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.122(0.07)	0.167(0.08)	0.208(0.11)	0.287(0.15)
	25	0.091(0.03)	0.126(0.04)	0.201(0.10)	0.279(0.15)
	50	0.085(0.05)	0.117(0.05)	0.216(0.12)	0.299(0.16)
$(100, 200, 200)$	10	0.103(0.07)	0.142(0.08)	0.152(0.09)	0.209(0.11)
	25	0.078(0.02)	0.109(0.03)	0.153(0.07)	0.212(0.1)
	50	0.067(0.02)	0.093(0.02)	0.153(0.08)	0.212(0.11)
$(100, 400, 400)$	10	0.094(0.08)	0.128(0.08)	0.158(0.10)	0.216(0.11)
	25	0.073(0.02)	0.102(0.03)	0.150(0.07)	0.207(0.10)
	50	0.065(0.02)	0.090(0.02)	0.148(0.07)	0.205(0.09)

Table 4.3: Mean and standard error (in parenthesis) comparison of the norm differences for the *entire* space across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.191(0.08)	0.307(0.10)	0.261(0.11)	0.403(0.14)
	25	0.126(0.03)	0.213(0.05)	0.182(0.05)	0.296(0.08)
	50	0.103(0.05)	0.176(0.05)	0.152(0.06)	0.257(0.07)
$(100, 200, 200)$	10	0.159(0.07)	0.256(0.08)	0.207(0.09)	0.320(0.11)
	25	0.105(0.02)	0.179(0.04)	0.145(0.04)	0.237(0.06)
	50	0.084(0.01)	0.144(0.02)	0.115(0.03)	0.196(0.04)
$(100, 400, 400)$	10	0.149(0.08)	0.239(0.09)	0.207(0.09)	0.320(0.11)
	25	0.097(0.02)	0.162(0.03)	0.139(0.04)	0.229(0.06)
	50	0.078(0.02)	0.132(0.02)	0.116(0.03)	0.196(0.04)

shared-space estimation further enhances the recovery of the source-specific subspace as shown in Table 4.2. These two improved estimators, $\widehat{\boldsymbol{P}}_s$ and $\widehat{\boldsymbol{P}}_p^{[1]}$, together lead to a better estimate for the entire subspace, as shown in Table 4.3.

4.5 Real Data Illustration

Air pollution has become one of the major environmental concerns in China due to its rapid industrial expansion and urbanization over the past four decades (Guan et al., 2016). Fine particulate matter (PM) with an aerodynamic diameter less than 2.5 μm, often referred to as $PM_{2.5}$, is one of the main pollutants in multiple cities in China (Huang et al., 2014). Cardiovascular and respiratory diseases, and even lung cancer, are associated with long-term exposure to $PM_{2.5}$ (Pope III et al., 2002; Hoek et al., 2013; Lelieveld et al., 2015).

In this study, our objective is to identify the shared variation pattern of the daily $PM_{2.5}$ levels of five Chinese major cities: Beijing, Chengdu, Guangzhou, Shanghai, and Shenyang. These cities are located in economically developed regions, collectively contributing over 10% of China's Gross Domestic Product and approximately 6% of its total population. The data were collected by U.S. diplomatic posts located in the five cities. The U.S. Embassy in Beijing started releasing hourly $PM_{2.5}$ readings in April 2008, followed by the consulates in Guangzhou, Shanghai, Chengdu, and Shenyang in November and December 2011, June 2012, and April 2013, respectively.

To begin, we stratify the data by human activity patterns and seasonal effects. Holidays and weekends are excluded to ensure that the selected days reflect homogeneous human activity patterns. Given the substantial influence of meteorological conditions on daily $PM_{2.5}$ levels, we analyze the data separately for winter months (November to February of the following year) and summer months (May to August). This partitioning accounts for the winter heating regulated by local governments in northern cities such as Beijing and Shenyang, where heating typically begins in November, following the approach of Liang et al. (2015). Finally, daily measurements are aggregated into weekly averages to mitigate temporal dependence.

The numbers of selected FPCs are 4, 4, 6, 4, 5 for Beijing, Chengdu, Guangzhou,

Shanghai, and Shenyang, respectively, in winter, and 5, 6, 4, 5, 5 in summer. Figure 4.3 displays the eigenvalues obtained from the weighted average of the projection operator for the winter and summer months. In general, the alignment of variability patterns across five cities is stronger in summer than in winter, suggesting human activities patterns are much more similar in summer.

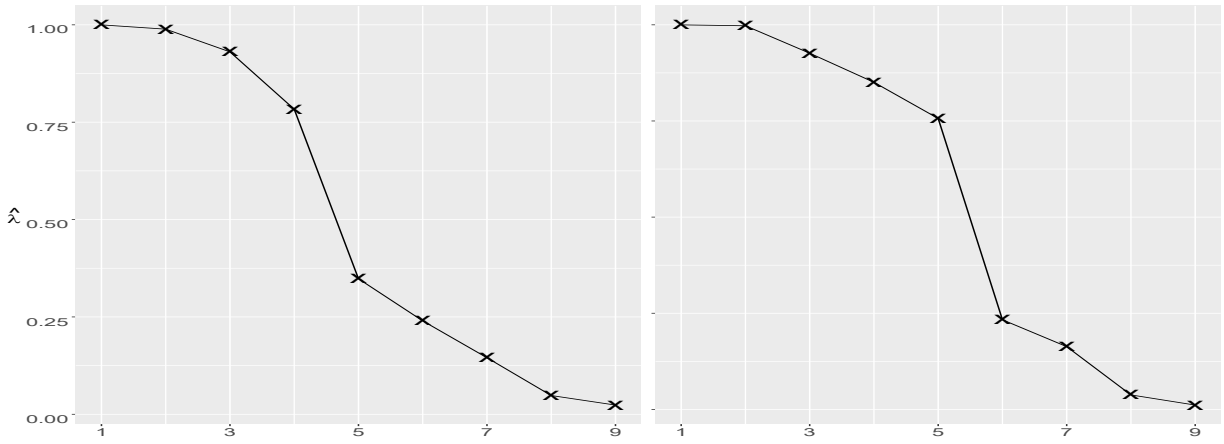


Figure 4.3: The left panel shows the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{winter}}$ for five cities in the winter months, and the right panel demonstrates the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{summer}}$ in the summer months. Both projected operators are calculated based on equation (4.7).

The top two eigenvalues are close to one for both seasons, implying two dominant shared components. The associated eigenfunctions are shown in Figure 4.4. In both seasons, $\hat{\psi}_2$ increases after the morning rush hour, peaks at approximately 4:30 PM, and declines slightly thereafter, suggesting that it captures afternoon traffic emissions. The shape of $\hat{\psi}_1$ differs between winter and summer, likely reflecting meteorological effects. Overall, these findings indicate the existence of a shared structure in the variation pattern of daily $PM_{2.5}$ levels across the five cities, despite their substantial geographical and climatic diversity.

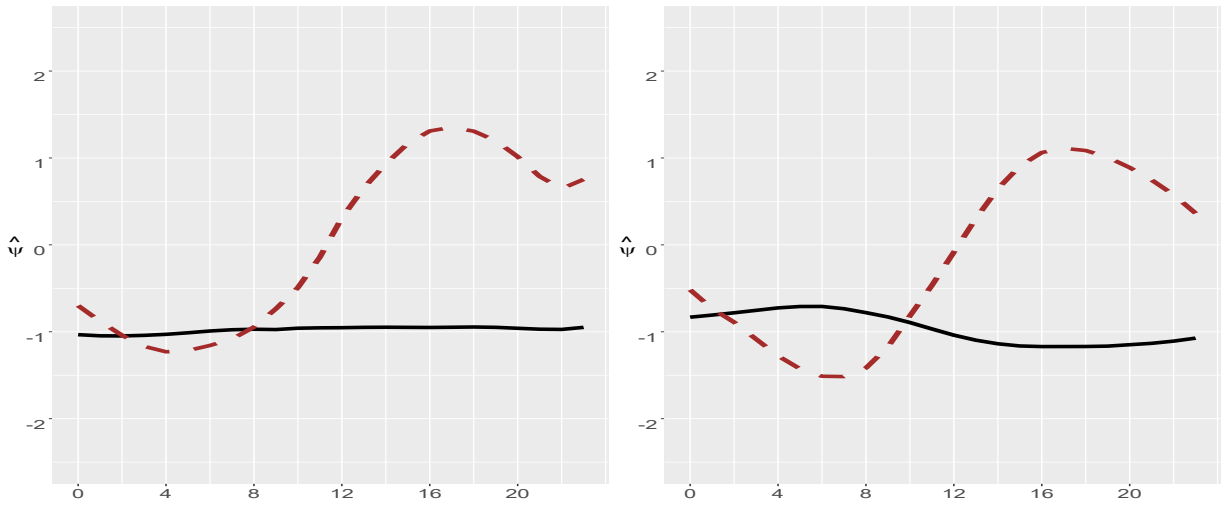


Figure 4.4: The left panel displays the leading two estimated eigenfunctions of $\widehat{\boldsymbol{P}}_{\text{winter}}$, and the right panel presents those of $\widehat{\boldsymbol{P}}_{\text{summer}}$. In both panels, the solid and dashed lines represent $\hat{\psi}_1$ and $\hat{\psi}_2$, respectively, obtained from the corresponding $\widehat{\boldsymbol{P}}_s$.

4.A Proof of the Theorems in Section 4.3

This section presents the proof for the two main theorems in Section 4.3.

4.A.I Proof of Theorem 4.1

Proof. Note that

$$\boldsymbol{P}_s^{[k]} = \sum_{\nu=1}^{m_s} \phi_{\pi_k(\nu)} \otimes \phi_{\pi_k(nu)}$$

for all $k \in K$, while these $\phi_{\pi_k(\nu)}$ spanned the same subspace $\mathcal{H}_s$ based on Assumption 4.1. Therefore,

$$\begin{aligned}\boldsymbol{P}_w &= \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \boldsymbol{P}_s^{[k]} + \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \boldsymbol{P}_p^{[k]} \\ &= \boldsymbol{P}_s + \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \boldsymbol{P}_p^{[k]},\end{aligned}$$

where $\dim(\boldsymbol{P}_s) = m_s$. Note that, each $\tilde{\phi}_\nu^{[k]}$ is a consistent estimator of $\phi_\nu^{[k]}$ based on Lemma 4.1 under the conditions of the theorem. We can treat $\widehat{\boldsymbol{P}}_w$ is a perturbed version of $\boldsymbol{P}_w$. Let us denote $\delta_j(\boldsymbol{A}, \widetilde{\boldsymbol{A}}) = \|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_{\text{op}}/\eta_j(A)$, where $\eta_j(\boldsymbol{A}) = \min_{i\neq j}|\lambda_i(\boldsymbol{A}) - \lambda_j(\boldsymbol{A})|/2$. Here, $\lambda_i(\boldsymbol{A})$ denote the ith eigenvalue of $\boldsymbol{A}$. According to the proof of Corollary 5.1.5 in Hsing and Eubank (2015), the estimated projection operator cannot deviated from the true value a lot in terms of the operator norm. That is,

$$\|\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s\|_{\text{op}} \leq \frac{\delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w)}{1 - \delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w)} \tag{4.10}$$

where

$$\delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w) = \frac{\|\boldsymbol{P}_w - \widehat{\boldsymbol{P}}_w\|_{\text{op}}}{\eta_1(\boldsymbol{P}_w)}. \tag{4.11}$$

It is easy to see that $\min_{i\neq 1}|\lambda_1(\boldsymbol{P}_w) - \lambda_i(\boldsymbol{P}_w)| \geq d$ from Assumption 4.2. From Lemma 4.2, we know that

$$\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}_w\|_{\text{op}} \leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k} \big\|\tilde{\phi}_\nu^{[k]} - \phi_\nu^{[k]}\big\|_2.$$

The proof is concluded based on Lemma 4.1, depending on whether $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$ or $\bar{N}_2^{[k]} \asymp m_k$. Note that, if $\delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w) = o_p(1)$, then $\|\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s\|_{\text{op}} \lesssim 2\delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w) = \kappa_{n_t}$. Thus, Let us first consider the order of $\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}_w\|_{\text{op}}$.

Case 1: $\bar{N}_2^{[k]} \gtrsim r_k^{a_k}$ for all $k \in \mathcal{K}$.

From Lemma 4.1, we know

$$\big\|\tilde{\phi}_\nu^{[k]} - \phi_\nu^{[k]}\big\|_2^2 \lesssim \frac{\nu^2}{n_k} + \frac{\nu^{(4a_k+2c_k+2)/5}}{(n_k\bar{N}_2^{[k]})^{4/5}}, \quad i.p.$$

for $\nu \leq m_k$. Thus,

$$\begin{aligned}
&\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}_w\|_{\text{op}} \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k}\left\{\frac{\nu^2}{n_k} + \frac{\nu^{(4a_k+2c_k+2)/5}}{(n_k\bar{N}_2^{[k]})^{4/5}}\right\}^{1/2} \quad i.p. \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k}\left\{\frac{\nu}{n_k^{1/2}} + \frac{\nu^{(2a_k+c_k+1)/5}}{(n_k\bar{N}_2^{[k]})^{2/5}}\right\} \quad i.p. \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{1/2} m_k^2 + \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k} \frac{\nu^{(2a_k+c_k+1)/5}}{(n_k\bar{N}_2^{[k]})^{2/5}}, \quad i.p.
\end{aligned} \tag{4.12}$$

If we further have $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$ for all $k \in \mathcal{K}$, the first term of equation (4.12) dominates the second term, and

$$(4.12) \lesssim \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}}. \tag{4.13}$$

It is easy to see that $m_k^4 < m_k^{2a_k+2} = o(n_k)$ by Assumption 4.5 as $a_k > 1$. Thus, equation (4.13) $= o_p(1)$. If we only have $\bar{N}_2^{[k]} \gtrsim r_k^{a_k}$ for all $k \in \mathcal{K}$, then the second term in equation (4.12) is bounded as follows:

$$\begin{aligned}
&\frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k} \frac{\nu^{(2a_k+c_k+1)/5}}{(n_k\bar{N}_2^{[k]})^{2/5}} \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{3/5} \left\{\frac{m_k^{a_k+c_k/2+3}}{\bar{N}_2^{[k]}}\right\}^{2/5} \\
&\lesssim \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{3/5} m_k^{c_k/5+6/5} \\
&\leq \frac{1}{n_t}\left\{\sum_{k\in\mathcal{K}} n_k\right\}^{3/5}\left\{\sum_{k\in\mathcal{K}} m_k^{c_k/2+3}\right\}^{2/5} \\
&\lesssim \frac{1}{n_t^{2/5}}\left\{\sum_{k\in\mathcal{K}} m_k^{c_k/2+3}\right\}^{2/5}
\end{aligned} \tag{4.14}$$

Note that, $c_k/2 + 3 < 2a_k + 2$ by the assumptions of the theorem. Therefore, equation (4.14) $= o_p(1)$.

Combining equations (4.12) - (4.14), we have

$$\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}\|_{\text{op}} \lesssim \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}} + \frac{\left\{\sum_{k\in\mathcal{K}} m_k^{c_k/2+3}\right\}^{2/5}}{n_t^{2/5}} = o_p(1).$$

Case 2: $\bar{N}_2^{[k]} \asymp r_k$ for all $k \in \mathcal{K}$.

Under this scenario, Lemma 4.1 implies

$$\|\tilde{\phi}_\nu^{[k]} - \phi_\nu^{[k]}\|_2^2 \lesssim \frac{\nu^{2a_k+2}}{n_k\left(\bar{N}_2^{[k]}\right)^2} + \frac{\nu^{(8a_k+2c_k+2)/5}}{n_k^{4/5}\left(\bar{N}_2^{[k]}\right)^{8/5}}.$$

Thus,

$$
\begin{aligned}
&\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}_w\|_{\text{op}} \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k}\left\{\frac{\nu^{2a_k+2}}{n_k\left(\bar{N}_2^{[k]}\right)^2} + \frac{\nu^{(8a_k+2c_k+2)/5}}{n_k^{4/5}\left(\bar{N}_2^{[k]}\right)^{8/5}}\right\}^{1/2} \qquad i.p. \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \sum_{\nu=1}^{m_k}\left\{\frac{\nu^{a_k+1}}{n_k^{1/2}\bar{N}_2^{[k]}} + \frac{\nu^{(4a_k+c_k+1)/5}}{n_k^{2/5}\left(\bar{N}_2^{[k]}\right)^{4/5}}\right\} \qquad i.p. \\
&\leq \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{1/2}\frac{m_k^{a_k+2}}{\bar{N}_2^{[k]}} + \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \frac{m_k^{(4a_k+c_k+6)/5}}{n_k^{2/5}\left(\bar{N}_2^{[k]}\right)^{4/5}} \qquad i.p. \qquad (4.15)
\end{aligned}
$$

The firth term of equation (4.15) can be simplified as

$$
\begin{aligned}
&\frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{1/2}\frac{m_k^{a_k+2}}{\bar{N}_2^{[k]}} \\
&\asymp \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{1/2} m_k^{a_k+1} \\
&\leq n_t^{-1/2}\Big(\sum_{k\in\mathcal{K}} m_k^{2a_k+2}\Big)^{1/2} = o_p(1),
\end{aligned}
$$

where the last line due to $m_k^{2a_k+2}/n_k = o(1)$. The second term of equation (4.15) can be simplified as

$$
\begin{aligned}
&\frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k \frac{m_k^{(4a_k+c_k+6)/5}}{n_k^{2/5}\left(\bar{N}_2^{[k]}\right)^{4/5}} \\
&\asymp \frac{1}{n_t}\sum_{k\in\mathcal{K}} n_k^{3/5} m_k^{(4a_k+c_k+2)/5} \\
&\leq n_t^{-2/5}\Big(\sum_{k\in\mathcal{K}} m_k^{2a_k+c_k/2+1}\Big)^{2/5} = o_p(1)
\end{aligned}
$$

where the last line holds if $(2a_k + c_k/2 + 1)/2(a_k + 1) < 1$. Therefore,

$$\delta_1(\widehat{\boldsymbol{P}}_w, \boldsymbol{P}_w) \lesssim 2\|\widehat{\boldsymbol{P}}_w - \boldsymbol{P}\|_{\text{op}}$$

□

4.A.II Proof of Theorem 4.2

Proof. Recall $\widehat{\boldsymbol{P}}_p^{[k]} = \sum_{\nu=1}^{m_k - m_s} \hat{\psi}_\nu^{[k]} \otimes \hat{\psi}_\nu^{[k]}$, where $\hat{\psi}_\nu^{[k]}$ is the νth leading eigenfunction of $\widehat{\boldsymbol{G}}_p^{[k]} = \widehat{\boldsymbol{P}}_d \widetilde{\boldsymbol{G}}^{[k]} \widehat{\boldsymbol{P}}_d$. Let denote $\mathcal{I}_p^{[k]} = \{\nu \in \mathbb{N}_+ \mid \nu \notin \mathcal{I}_s^{[s]}, \nu \leq m_k\}$ be the index set of the eigenfunctions belong to the kth source-specific variation, and let $\Phi_\nu = \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}$ for $\nu \in \mathcal{I}_p^{[k]}$. Clearly, $|\mathcal{I}_p^{[k]}| = m_k - m_s$. Moreover, let $\widehat{\Psi}_\nu^{[k]} = \hat{\psi}_\nu^{[k]} \otimes \hat{\psi}_\nu^{[k]}$. First, consider $\boldsymbol{G}_p^{[k]} = \boldsymbol{P}_d \boldsymbol{G}^{[k]} \boldsymbol{P}_d$, where $\boldsymbol{P}_d$ is the population level counterpart of $\widehat{\boldsymbol{P}}_d$. Since $\widehat{\boldsymbol{P}}_d$ is driven from $\widehat{\boldsymbol{T}}$, we consider $\boldsymbol{T} = \boldsymbol{P}^{[k]}(\text{id} - \boldsymbol{P}_s)\boldsymbol{P}^{[k]} = \boldsymbol{P}^{[k]} - \boldsymbol{P}_s$. Clearly, the eigenvalues of $\boldsymbol{T}$ are either 0 or 1 as $\mathcal{H}_s \subset \mathcal{H}_0^{[k]}$. Moreover, $\widehat{\boldsymbol{T}}$ is a consistent estimator of $\boldsymbol{T}$, which implies $\widehat{\boldsymbol{G}}_p^{[k]}$ is a consistent estimator of $\boldsymbol{G}_p^{[k]}$. Let us denote the set of eigenfunctions ψ_ν of $\boldsymbol{G}_p^{[k]}$, and we can define $\Psi_\nu^{[k]} = \psi_\nu^{[k]} \otimes \psi_\nu^{[k]}$. Clearly the projection space associated with $\dot{\boldsymbol{P}}_p^{[k]} = \sum_{\nu=1}^{m_k - m_s} \Psi_\nu^{[k]}$ equals to the projection space associated with $\boldsymbol{P}^{[k]} - \boldsymbol{P}_s^{[k]}$, that is

$$\dot{\boldsymbol{P}}_p^{[k]}(\mathcal{H}) = (\boldsymbol{P}^{[k]} - \boldsymbol{P}_s^{[k]})(\mathcal{H}).$$

Therefore, we have $\psi_\nu^{[k]} = \phi_{\pi_k(\nu)}$ for some permutation function $\pi_k : \{1, 2, \ldots, m_k - m_s\} \to \mathcal{I}_p^{[k]}$.

The above arguments implies

$$\|\widehat{\boldsymbol{P}}_p^{[k]} - \boldsymbol{P}_p^{[k]}\|_{\text{op}} = \|\widehat{\boldsymbol{P}}_p^{[k]} - \dot{\boldsymbol{P}}_p^{[k]}\|_{\text{op}}.$$

In addition, Lemma 4.2 implies that

$$\|\widehat{\boldsymbol{P}}_p^{[k]} - \dot{\boldsymbol{P}}_p^{[k]}\|_{\text{op}} \leq \sum_{\nu=1}^{m_k - m_s} \|\widehat{\Psi}_\nu^{[k]} - \Psi_\nu^{[k]}\|_{\text{op}} \leq \sum_{\nu=1}^{m_k - m_s} \|\hat{\psi}_\nu^{[k]} - \psi_\nu^{[k]}\|_2.$$

Note that, $\boldsymbol{G}_p^{[k]}$ is a finite truncation version of $\boldsymbol{G}^{[k]}$. It can be viewed as taking the first m_k eigenpairs of $\boldsymbol{G}^{[k]}$, then setting some eigenvalues to be zero if the associated eigenfunctions is in the shared space. Therefore, $\boldsymbol{G}_p^{[k]}$ satisfied Assumptions 4.3 - 4.7 as long as the kth source satisfied these assumptions. Lastly, the consistency of $\widehat{\boldsymbol{G}}_p^{[k]}$ implies that the eigenpairs of $\boldsymbol{G}_p^{[k]}$ should be consistently estimated by the eigenpairs of $\widehat{\boldsymbol{G}}_p^{[k]}$, where the convergence rates depend on the eigenstructure of $\boldsymbol{G}_p^{[k]}$.

The proof is concluded by applying the same argument in proving Theorem 4.1.

Case 1: $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$.

$$
\begin{aligned}
&\|\widehat{\boldsymbol{P}}_p^{[k]} - \boldsymbol{P}_p^{[k]}\|_{\text{op}} \\
\leq &\sum_{\nu=1}^{m_k - m_s} \|\hat{\psi}_\nu^{[k]} - \psi_\nu^{[k]}\|_2 \\
\leq &\sum_{\nu=1}^{m_k - m_s} \left(\frac{\nu}{n_k^{1/2}} + \frac{\nu^{(2a_k + c_k + 1)/5}}{(n_k \bar{N}_2^{[k]})^{2/5}} \right) \\
\leq &\frac{(m_k - m_s)^2}{n_k^{1/2}} + \left\{ \frac{(m_k - m_s)^{c_k/2+3}}{n_k} \left(\frac{m_k - m_s}{m_k} \right)^{a_k} \right\}^{2/5} = \zeta_{n_k}.
\end{aligned}
$$

If we further have $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k + c_k/2 - 2}$, then

$$
\zeta_{n_k} = \frac{(m_k - m_s)^2}{n_k^{1/2}}
$$

Case 2: $\bar{N}_2^{[k]} \asymp m_k$.

$$
\|\tilde{\psi}_\nu^{[k]} - \psi_\nu^{[k]}\|_2^2 \lesssim \frac{\nu^{2a_k+2}}{n_k \left(\bar{N}_2^{[k]} \right)^2} + \frac{\nu^{(8a_k + 2c_k + 2)/5}}{n_k^{4/5} \left(\bar{N}_2^{[k]} \right)^{8/5}}.
$$

Thus,

$$
\begin{aligned}
&\|\widehat{P}_p^{[k]} - P_p^{[k]}\|_{\mathrm{op}} \\
&\leq \sum_{\nu=1}^{m_k - m_s} \left\{ \frac{\nu^{a_k+1}}{n_k^{1/2} \bar{N}_2^{[k]}} + \frac{\nu^{(4a_k+c_k+1)/5}}{n_k^{2/5} \left(\bar{N}_2^{[k]}\right)^{4/5}} \right\} \\
&\leq \frac{(m_k - m_s)^{a_k+2}}{n_k^{1/2} \bar{N}_2^{[k]}} + \frac{(m_k - m_s)^{(4a_k+c_k+6)/5}}{n_k^{2/5} \left(\bar{N}_2^{[k]}\right)^{4/5}} \\
&\leq \frac{(m_k - m_s)^{a_k+1}}{n_k^{1/2}} + \frac{(m_k - m_s)^{(4a_k+c_k+2)/5}}{n_k^{2/5}} = \zeta_{n_k}
\end{aligned}
$$

□

4.B Proposition 4.1 and its Proof

Proposition 4.1. *Suppose* $\boldsymbol{P}_{\mathcal{I}^{[k]}} = \sum_{\nu\in\mathcal{I}^{[k]}} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}$, *and* $\widetilde{\boldsymbol{P}}_{\mathcal{I}^{[k]}} = \sum_{\nu\in\mathcal{I}^{[k]}} \widetilde{\phi}_\nu^{[k]} \otimes \widetilde{\phi}_\nu^{[k]}$, *where* $\mathcal{I}^{[k]}$ *is a known index set for source* k. *Let* $|\mathcal{I}^{[k]}|$ *denote the number of elements in* $\mathcal{I}^{[k]}$. *Then, under the conditions of Theorem 4.1,*

$$\|\widetilde{\boldsymbol{P}}_{\mathcal{I}^{[k]}} - \boldsymbol{P}_{\mathcal{I}^{[k]}}\|_{op} = O_p(\tau_{n_k}),$$

where

- *When* $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$,

$$\tau_{n_k} = |\mathcal{I}^{[k]}| \left(\frac{m_k}{n_k^{1/2}} + \frac{m_k^{1-2a_k/5}}{n_k^{2/5}} \right).$$

  If $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$, *then*

$$\tau_{n_k} = |\mathcal{I}^{[k]}| \frac{m_k}{n_k^{1/2}}.$$

- *When* $\bar{N}_2^{[k]} \asymp m_k$, *then,*

$$\kappa_{n_t} = |\mathcal{I}^{[k]}| \left\{ \frac{m_k^{a_k}}{n_k^{1/2}} + \frac{m_k^{[4(a_k-1)+c_k+1]/5}}{n_k^{2/5}} \right\}.$$

Proof. The results are the direct calculation based on Lemmas 4.1 and 4.2. For example, we consider the case when $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$. Then,

$$\|\widetilde{\boldsymbol{P}}_{\mathcal{I}^{[k]}} - \boldsymbol{P}_{\mathcal{I}^{[k]}}\|_{\text{op}} \le \sum_{\nu\in\mathcal{I}^{[k]}} \|\tilde{\phi}_\nu^{[k]} - \tilde{\phi}_\nu^{[k]}\|_2 \le \sum_{\nu\in\mathcal{I}^{[k]}} \frac{\nu}{n_k^{1/2}} \le |\mathcal{I}^{[k]}| \frac{m_k}{n_k^{1/2}}.$$

The calculation for other cases are the same. □

4.C Technical Lemmas

The entire section is about FPCA results for single source. For brevity, we omit the superscript $[k]$ or k.

We present the following Lemma 4.1 concerning the estimation error of FPCs using single data source, which is a re-statement of Corollary 1 of Zhou et al. (2024). Let

$$\bar{N}_2 = \left\{ n^{-1} \sum_{i=1}^{n} (N_i)^{-2} \right\}^{-1/2}.$$

Lemma 4.1. *Assume Assumptions 4.3 - 4.8 hold. For all $\nu \leq m \in \mathbb{N}_+$, where $m^{2a+2}/n = o(1)$. Suppose conditions*

$$(m^{2a+2})/\{n\bar{N}_2^2 h^2(\nu)\} = o(1), \text{ and } h^4(\nu)\max\{m^{2a+2c}, m^{4a}\log(n)\} = o(1)$$

are satisfied. Let

$$h(\nu) = (n\bar{N}_2)^{-1/5}\nu^{(a-2c-2)/5}(1+\nu^a/\bar{N}_2)^{1/5},$$

then we have

- *If $\bar{N}_2 \gtrsim \nu^a$,*

$$\|\tilde{\phi}_\nu - \phi_\nu\|_2^2 = O_p\left(\frac{\nu^2}{n} + \frac{\nu^{(4a+2c+2)/5}}{(n\bar{N}_2)^{4/5}}\right).$$

  If we further have $\bar{N}_2 \gtrsim n^{1/4}\nu^{a+c/2-2}$, then

$$\|\tilde{\phi}_\nu - \phi_\nu\|_2^2 = O_p\left(\frac{\nu^2}{n}\right).$$

- *If $\bar{N}_2 = o(\nu^a)$,*

$$\|\tilde{\phi}_\nu - \phi_\nu\|_2^2 = O_p\left(\frac{\nu^{2a+2}}{n\bar{N}_2^2} + \frac{\nu^{(8a+2c+2)/5}}{(n\bar{N}_2^2)^{4/5}}\right).$$

The following lemma connects the convergence rate of the projection operator in terms of the operator norm with the convergence rate of FPCs. Note that the estimated projection

is constructed based on estimated eigenfunction. That is, we have the index set $\mathcal{I}$ and the associated $\tilde{\phi}_\nu, \nu \in \mathcal{I}$ first, then $\widetilde{P}$ is based on $\tilde{\phi}_\nu, \nu \in \mathcal{I}$.

Lemma 4.2. *Suppose* $\boldsymbol{P} = \sum_{\nu=1}^{m} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}$, *and* $\widetilde{\boldsymbol{P}}^{[k]} = \sum_{\nu=1}^{m} \widetilde{\phi}_\nu^{[k]} \otimes \widetilde{\phi}_\nu^{[k]}$. *Then* $\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_{op} \leq \sum_{\nu=1}^{m} \|\tilde{\phi}_\nu - \phi_\nu\|_2$.

Proof. Let $\boldsymbol{\Phi}_\nu = \phi_\nu \otimes \phi_\nu$ and $\widetilde{\boldsymbol{\Phi}}_\nu = \tilde{\phi}_\nu \otimes \tilde{\phi}_\nu$. The triangle inequality implies,

$$\begin{aligned}
\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_{\text{op}} &\leq \sum_{\nu=1}^{m} \|\boldsymbol{\Phi}_\nu - \widetilde{\boldsymbol{\Phi}}_\nu\|_{\text{op}} \\
&= \sum_{\nu=1}^{m} \left[1 - \left\{ 1 - \frac{1}{2} \left\| \tilde{\phi}_\nu^{[k]} - \phi_\nu^{[k]} \right\|_2^2 \right\}^2 \right]^{1/2},
\end{aligned}$$

where the last line comes from the part(3) of Lemma 5.1.7 in Hsing and Eubank (2015). Therefore,

$$\begin{aligned}
&\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}^{[k]}\|_{\text{op}} \\
&\leq \sum_{\nu=1}^{m} \left[1 - \left\{ 1 - \frac{1}{2} \left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2^2 \right\}^2 \right]^{1/2} \\
&= \sum_{\nu=1}^{m} \left[1 - \left\{ 1 - \left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2^2 + \frac{1}{4} \left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2^4 \right\} \right]^{1/2} \\
&= \sum_{\nu=1}^{m} \left[\left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2^2 - \frac{1}{4} \left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2^4 \right]^{1/2} \\
&\leq \sum_{\nu=1}^{m} \left\| \tilde{\phi}_\nu - \phi_\nu \right\|_2.
\end{aligned}$$

□

The final technical lemma establishes the convergence rate for $\widetilde{\boldsymbol{G}}$ in terms of the operator norm.

Lemma 4.3. *Under Assumptions 4.3, 4.4, 4.7, 4.8, we have*

- *When* $\bar{N}_2/(n/\log n)^{1/4} \to 0$ *and* $h \asymp \left[\left\{n(\bar{N}_2)^2\right\}/\log n\right]^{-1/6}$,

$$\left\|\widetilde{\boldsymbol{G}} - \boldsymbol{G}\right\|_{op} \lesssim \left\{\frac{\log n}{n(\bar{N}_2)^2}\right\}^{1/3} \quad a.s.$$

- *When* $\bar{N}_2/\left\{n/\log n\right\}^{1/4} \to C_1$ *for some positive constant* C_1 *and* $h \asymp (n/\log n)^{1/4}$,

$$\left\|\widetilde{\boldsymbol{G}} - \boldsymbol{G}\right\|_{op} \lesssim \left(\frac{\log n}{n}\right)^{1/2} \quad a.s.$$

- *When* $\bar{N}_2/(n/\log n)^{1/4} \to \infty$, $h^4 = o(n/\log n)$, *and* $\bar{N}_2 h \to \infty$

$$\left\|\widetilde{\boldsymbol{G}} - \boldsymbol{G}\right\|_{op} \lesssim \left\{\frac{\log n}{n}\right\}^{1/2} \quad a.s.$$

Proof. It is easy to see that, for any $f \in \mathcal{L}^2(\mathcal{T})$, $\int_{t\in\mathcal{T}}|f(t)|dt \le \|f\|_2$ by Jensen's inequality. Thus, for any $f \in \mathcal{L}^2(\mathcal{T})$,

$$\begin{aligned}
&\|\widetilde{\boldsymbol{G}} - \boldsymbol{G}\|_{\text{op}}^2 \\
&= \sup_{\|f\|_2=1}\left\|\int_{s\in\mathcal{T}}\left\{\widetilde{G}(t,s) - G(t,s)\right\}f(s)ds\right\|_2^2 \\
&= \sup_{\|f\|_2=1}\int_{t\in\mathcal{T}}\left[\int_{s\in\mathcal{T}}\left\{\widetilde{G}(t,s) - G(t,s)\right\}f(s)ds\right]^2 dt \\
&\le \sup_{t,s\in\mathcal{T}}\left|\widetilde{G}(t,s) - G(t,s)\right|^2 \sup_{\|f\|_2=1}\left\|\int_{t\in\mathcal{T}}\left\{\int_{s\in\mathcal{T}}f(s)ds\right\}^2 dt\right\| \\
&\le \sup_{t,s\in\mathcal{T}}\left|\widetilde{G}(t,s) - G(t,s)\right|^2
\end{aligned}$$

The rest of the proof follows from part(b) of Theorem 3 in Zhou et al. (2024). □

Chapter 5

Concluding Remarks and Future Work

In this dissertation, we have studied several important problems in functional data analysis: (1) testing mean differences between two groups of functional data when heterogeneity in the covariance structure is present; (2) determining the number of functional principal components (FPCs) to retain in functional principal component analysis (FPCA); and (3) extracting shared and source-specific subspaces from multiple functional data sources. In this chapter, we summarize the main contributions of this dissertation and discuss several potential directions for future research.

In Chapter 2, we introduce a statistical framework designed to identify mean differences between two groups of functional data, particularly in situations where each subject has only a limited number of irregularly spaced observations. Unlike typical FPCA-based approaches, our method does not require the assumption of a common covariance structure across two groups. Specifically, under the Reproducing kernel Hilbert space framework, we build a functional Bahadur representation for the mean function estimator. This estimator is constructed by applying smoothing splines to the aggregated data within each group. The functional Bahadur representation facilitates the development of pointwise limiting distributions and the weak convergence of the mean function estimator for each group. Subsequently, we present a testing procedure for the mean difference, as the functional

data from these two groups are independent.

It is worth emphasizing that the theoretical results obtained in this chapter are derived in a distribution-wise manner. Specifically, the Type I error is controlled for sufficiently large sample sizes under a fixed probability distribution. However, these asymptotic results generally fail to guarantee that the Type I error is well controlled across a family of distributions, even when the sample size is large; see Romano (2004) and Waudby-Smith et al. (2024) for more details. Currently, many distribution-free methods focus on testing conditional independence (Lundborg et al., 2022; Shah and Peters, 2020). In the context of FDA, the development of uniform inference procedures remains underexplored. It is still unclear what proper assumptions should be imposed to restrict the family of distributions to enable uniform inference for functional data. Most current work in FDA on statistical inference focuses on asymptotic properties for a fixed probability distribution, as seen in Wang (2021) and Pomann et al. (2016).

In Chapter 3, we developed the functional ladle estimator (FLE) to determine the number of FPCs to retain in FPCA. When the covariance operator has finite rank, eigenfunction estimates exhibit substantial variability once the index exceeds the true rank, whereas the associated eigenvalue estimates tend to be close to zero. Leveraging this observation, we constructed the FLE, which demonstrated superior performance in selecting the optimal number of FPCs across a range of simulation settings. Classical approaches, such as AIC-based methods, assume that the true model lies within the set of candidate models. However, when smoothing techniques are employed, bias is inevitable if the average number of observations per subject is not sufficiently large relative to the sample size. Our method, however, relies on the variability of the estimated eigenfunctions, making it more flexible and broadly applicable. Numerical studies demonstrate strong performance across diverse settings, while information-criteria-based methods often show sensitivity to specific simulation settings.

Despite these promising results, several theoretical aspects remain to be explored. In particular, analyzing contaminated discrete observations from infinite-dimensional processes presents significant challenges. The smoothing step introduces additional tuning parameters, such as bandwidths, which complicate the theoretical analysis. The asymptotic behaviour of the estimated eigenfunctions in the null space of the covariance operator

remains unknown. In future work, we aim to investigate these theoretical properties in greater depth to further enhance the rigour and applicability of our approach.

In Chapter 4, we introduced a novel procedure to integrate multiple functional data sources and extract the shared and source-specific subspaces for these data sources. Specifically, we focus exclusively on subspaces by retaining the sample eigenfunction estimates while discarding the estimated eigenvalues. The shared space is isolated by examining the scree plot of the proposed projection estimator. With the help of this estimator, we can unveil the variation pattern across sources even if the eigenvalues of the eigenfunctions spanned this subspace are relatively small within individual sources. Furthermore, the source-specific subspace is obtained by reprojecting the sample covariance operator onto the orthogonal complement of the shared space. Simulation studies and real data applications demonstrate the effectiveness of the proposed method in capturing the shared features among multiple functional datasets.

One potential direction for future research is to consider that the similarity between different data sources may not be absolute. For example, $\|\boldsymbol{P}_s^{[1]} - \boldsymbol{P}_s^{[2]}\|_{\mathrm{HS}} \leq \eta$ for some small constant $\eta > 0$, where $\boldsymbol{P}_s^{[1]}$ and $\boldsymbol{P}_s^{[2]}$ denote the projection operators onto the shared subspaces of the first and second functional data sources, respectively. In this scenario, the eigenvalues of $\boldsymbol{P}_w$ are no longer equal to one, and we need to further develop methods to isolate the shared subspace and the source-specific subspaces.

References

Alduchov, O. A. and Eskridge, R. E. (1996). Improved Magnus form approximation of saturation vapor pressure. *Journal of Applied Meteorology and Climatology*, 35(4):601–609.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159.

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, 37(1):1 –34.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, Inc., Chichester, 1st edition.

Boente, G., Rodriguez, D., and Sued, M. (2010). Inference under functional proportional and common principal component models. *Journal of Multivariate Analysis*, 101(2):464–475.

Bücher, A. and Kojadinovic, I. (2019). A note on conditional versus joint unconditional weak convergence in bootstrap consistency results. *Journal of Theoretical Probability*, 32(3):1145–1165.

Cai, T. T., Kim, D., and Pu, H. (2024). Transfer learning for functional mean estimation: Phase transition and adaptive algorithms. *The Annals of Statistics*, 52(2):654–678.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5):2330–2355.

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538.

Charkaborty, A. and Panaretos, V. M. (2022). Testing for the rank of a covariance operator. *The Annals of Statistics*, 50(6):3510–3537.

Coffey, N., Harrison, A., Donoghue, O., and Hayes, K. (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Human Movement Science*, 30(6):1144–1166.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Dai, X., Müller, H.-G., and Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560.

Dauxois, J., Pousse, A., and Romain, Y. (1982). 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.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(2):267–286.

Dou, W. W., Pollard, D., and Zhou, H. H. (2012). Estimation in functional regression for general exponential families. *The Annals of Statistics*, 40(5):2421–2451.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2023). *refund: Regression with Functional Data*. R package version 0.1-32.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. Springer, New York, 2nd edition.

Guan, W.-J., Zheng, X.-Y., Chung, K. F., and Zhong, N.-S. (2016). Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. *The Lancet*, 388(10054):1939–1951.

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

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):109–126.

Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(4):689–705.

Hao, M., Liu, K.-Y., Xu, W., and Zhao, X. (2021). Semiparametric inference for the functional Cox model. *Journal of the American Statistical Association*, 116(535):1319–1329.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: A review. *Environmental Health*, 12(1):43.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Ltd, Chichester.

Huang, R.-J., Zhang, Y., Bozzetti, C., Ho, K.-F., Cao, J.-J., Han, Y., Daellenbach, K. R., Slowik, J. G., Platt, S. M., Canonaco, F., Zotter, P., Wolf, R., Pieber, S. M., Bruns, E. A., Crippa, M., Ciarelli, G., Piazzalunga, A., Schwikowski, M., Abbaszade, G., Schnelle-Kreis, J., Zimmermann, R., An, Z., Szidat, S., Baltensperger, U., Haddad, I. E., and Prévôt, A. S. H. (2014). High secondary aerosol contribution to particulate pollution during haze events in China. *Nature*, 514(7521):218–222.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):271–293.

Jiao, S., Frostig, R., and Ombao, H. (2024). Filtrated common functional principal component analysis of multigroup functional data. *The Annals of Applied Statistics*, 18(2):1160–1177.

Kley, T., Volgushev, S., Dette, H., and Hallin, M. (2016). Quantile spectral processes: Asymptotic analysis and inference. *Bernoulli*, 22(3):1770–1807.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, New York.

Kokoszka, P., Rice, G., and Shang, H. L. (2017). Inference for the autocovariance of a functional time series under conditional heteroscedasticity. *Journal of Multivariate Analysis*, 162:32–50.

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

Lelieveld, J., Evans, J., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525:367–71.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504):1284–1294.

Li, Z., Qin, K., He, Y., Zhou, W., and Zhang, X. (2024). Knowledge transfer across multiple principal component analysis studies. DOI: 10.48550/arXiv.2403.07431.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Lundborg, A. R., Shah, R. D., and Peters, J. (2022). Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805.

Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P., and Reich, D. (2010). MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis Journal*, 16(2):166–177.

Pigoli, D., Aston, J. A. D., Dryden, I. L., and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, 101(2):409–422.

Pomann, G.-M., Staicu, A.-M., and Ghosh, S. (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(3):395–414.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9):1132–1141.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2nd edition.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 53(1):233–243.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584.

Rudin, W. (1986). *Principles of Mathematical Analysis*. McGraw - Hill, New York, 3rd edition.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Shang, Z. and Cheng, G. (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638.

Shang, Z. and Cheng, G. (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics*, 43(4):1742–1773.

Staicu, A.-M., Li, Y., Crainiceanu, C. M., and Ruppert, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, 41(4):932–949.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

Wang, Q. (2021). Two-sample inference for sparse functional data. *Electronic Journal of Statistics*, 15(1):1395–1423.

Wang, Y., Wang, G., Klinedinst, B., Willette, A., and Wang, L. (2025). Statistical inference for mean functions of complex 3D objects. *Statistica Sinica*.

Waudby-Smith, I., Kennedy, E. H., and Ramdas, A. (2024). Distribution-uniform anytime-valid sequential inference. DOI: https://arxiv.org/abs/2311.03343.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903.

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

Zhang, C., Sang, P., and Qin, Y. (2025a). Order determination for functional data. DOI: 10.48550/arXiv.2503.03000.

Zhang, C., Sang, P., and Qin, Y. (2025b). Two-sample inference for sparse functional data. *Electronic Journal of Statistics*, 19(1):792–864.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079.

Zhang, J.-T. and Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics*, 41(1):51–71.

Zhang, J.-T., Liang, X., and Xiao, S. (2010). On the two-sample Behrens-Fisher problem for functional data. *Journal of Statistical Theory and Practice*, 4(4):571–587.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.

Zhang, Y., Chen, S. X., and Bao, L. (2023). Air pollution estimation under air stagnation–A case study of Beijing. *Environmetrics*, 34(6):e2819.

Zhou, H., Wei, D., and Yao, F. (2024). Theory of functional principal component analysis for noisy and discretely observed data. DOI: 10.48550/arXiv.2209.08768, accepted by the Annals of Statistics.

Zhou, H., Yao, F., and Zhang, H. (2023). Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393.

Zhou, Y., Bhattacharjee, S., Carroll, C., Chen, Y., Dai, X., Fan, J., Gajardo, A., Hadjipantelis, P. Z., Han, K., Ji, H., Zhu, C., Müller, H.-G., and Wang, J.-L. (2022). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5.9.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(3):581–603.