SOME PERSPECTIVES OF SMOOTH AND LOCALLY
SPARSE ESTIMATORS

by

Zhenhua Lin
B.Sc., Fudan University, 2008
M.Sc., Simon Fraser University, 2010

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

© Zhenhua Lin 2013
SIMON FRASER UNIVERSITY
Summer 2013

All rights reserved.

However, in accordance with the Copyright Act of Canada, this work may be reproduced without authorization under the conditions for “Fair Dealing.” Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.
Name: Zhenhua Lin

Degree: Master of Science

Title of Project: Some Perspectives of Smooth and Locally Sparse Estimators

Examining Committee:

Chair: Dr. Tim Swartz, Professor

Dr. Jiguo Cao
Senior Supervisor, Associate Professor

Dr. Richard Lockhart
Supervisor, Professor

Dr. Qian Zhou
Internal Examiner, Assistant Professor

Date Approved:
Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the “Institutional Repository” link of the SFU Library website (www.lib.sfu.ca) at http://summit/sfu.ca and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2011
Abstract

In this thesis we develop some new techniques for computing smooth and meanwhile locally sparse (i.e. zero on some sub-regions) estimators of functional principal components (FPCs) in functional principal component analysis (FPCA) and coefficient functions in functional linear regression (FLR). Like sparse models in ordinary data analysis, locally sparse estimators in functional data analysis enjoy less variability and better interpretability.

In the first part of the thesis, we develop smooth and locally sparse estimators of FPCs. For an FPC, the sub-regions on which it has significant magnitude are interpreted as where sample curves have major variations. The non-null sub-regions of our estimated FPCs coincide with the sub-regions where the corresponding FPC has significant magnitude. This makes our derived FPCs easier to interpret: those non-null sub-regions are where sample curves have major variations. An efficient algorithm is designed to compute our estimators using projection deflation. Our estimators are strongly consistent and asymptotically normal under mild conditions. Simulation studies also show that FPCs estimated by our method explain similar variations of sample curves as FPCs estimated from other methods.

In the second part of the thesis, we develop a new regularization technique called “functional SCAD” (fSCAD), which is the functional generalization of the well-known SCAD (smoothly clipped absolute deviation) regularization, and then apply it to derive a smooth and locally sparse estimator of the coefficient function in FLR. The fSCAD enables us to identify the null sub-regions of the coefficient function without over shrinking the non-zero values. The smoothness of our estimator is regularized by a roughness penalty. We also develop an efficient algorithm to compute the estimator in practice via B-Splines expansion. An asymptotic analysis shows that our estimator enjoys the oracle property, i.e. it performs as well as if we knew the true null sub-regions of the coefficient function in advance. The simulation studies show that our estimator has superior numerical performance.
To my family.
“Everything should be made as simple as possible, but not simpler.”

— Albert Einstein (1879 - 1955)
Acknowledgements

It is difficult to overemphasize my gratitude to my advisor Dr. Jiguo Cao. His constant encouragement, insightful suggestions and great support make this thesis possible. I have been benefiting so much from working closely with him in the last two years. Jiguo is very generous and patient to me. He is not just a professional advisor in research, but also a supportive friend for life. It is extremely lucky for me to have Jiguo as my thesis advisor.

I am deeply grateful to my supervisor, Dr. Richard Lockhart, for his encouragement and guidance on my study of mathematical statistics, probability theory and some related advanced mathematics that are of importance to this thesis. I learnt a lot from the books he recommended and lent to me and the discussions with him on analysis and even topology. I also thank Dr. Qian Zhou and Dr. Tim Swartz for serving on my examining committee and spending their precious time on reviewing my work.

I wish to express my special appreciation to Dr. Wenyuan Li. With him as my mentor, I enjoyed learning, using and exploring statistics when I worked in BC Hydro. I learned a lot of statistical skills from him, particularly the skill of applying statistical methods to solve practical problems. This pleasant experience and his strong encouragement played a very important role in my decision to pursue professional training and then a career in the area of statistics.

I would like to acknowledge many people in our department, support staff and faculty, for always being helpful over the years. I thank my friends at Simon Fraser University for their help. A particular acknowledgement goes to Tianyu Guan, Dilinuer Kuerban, Jinhee Lim, Jie Liu, Lu Wang, Qian Wang, Tingting Wen, Yuanyu Yang and Kasra Yousefi. I also would like to express my sincere thanks to my dear friends Jiyi Chen, Wei Gao, Xiaoyuan Guo, Dr. Ming Hua, Dr. Bin Jiang, Jiarong Jiang, Yuanbin Shen, Dr. Zhengzheng Xing, Dr. Xiaojin Xu and Dr. Bin Zhou.
Finally, I would like to dedicate this thesis to my wife Dr. Liu Yang, my parents and parents-in-law for their boundless love, support, encouragement and company through all these years. I hope my achievement makes them proud of me, as I am proud of them.
Contents

Approval ii
Partial Copyright License iii
Abstract iv
Dedication v
Quotation vi
Acknowledgements vii
List of Tables xii
List of Figures xiii
List of Algorithms xiv

1 Introduction 1

2 Interpretable FPCA 4
  2.1 Introduction and Motivation ................................. 4
  2.2 Preliminaries .................................................. 6
  2.3 Methodology .................................................... 8
    2.3.1 Formulation ................................................. 8
    2.3.2 A Practical Approach ...................................... 9
    2.3.3 The iFPCA Algorithm .................................... 12
    2.3.4 Efficient Update of $G_{\alpha}^{-1}$ and $G_{\alpha}^{-1}Q_{\alpha}$ .......... 14
2.3.5 Projection Deflation .................................. 16
2.3.6 Tuning Parameter Selection ............................... 17
2.4 Asymptotic Properties .................................... 18
2.5 Simulation Studies ........................................ 20
  2.5.1 Simulation 1 ........................................ 20
  2.5.2 Simulation 2 ........................................ 23
2.6 Applications .............................................. 24
  2.6.1 Application 1: Canadian Weather Data ................. 24
  2.6.2 Application 2: Pinch Force Data ....................... 27
2.7 Summary .................................................. 28
2.8 Proofs ..................................................... 29
  2.8.1 Proof of Theorem 2.1 ................................ 30
  2.8.2 Proof Strategy for Other Theorems .................... 31
  2.8.3 A Key Lemma ......................................... 32
  2.8.4 Proofs of Theorem 2.2 and 2.3 ......................... 36
  2.8.5 Proof of Theorem 2.4 ................................ 42

3 Functional SCAD and SLoS Estimator 47
  3.1 Introduction ............................................. 47
  3.2 Methodology ............................................. 50
    3.2.1 Functional SCAD ................................... 50
    3.2.2 Penalized B-spline Expansion ....................... 52
    3.2.3 The SLoS Estimator ................................ 53
    3.2.4 Choice of Tuning Parameters ....................... 56
    3.2.5 Extension to Multiple Regressors ................... 57
  3.3 Oracle Property ......................................... 57
  3.4 Simulation Studies ....................................... 59
    3.4.1 Study 1 ........................................... 60
    3.4.2 Study 2 ........................................... 65
  3.5 Application: Canadian Weather Data ..................... 69
  3.6 Concluding Remarks ..................................... 70
  3.7 Proofs .................................................. 71

4 Summary of Thesis 81
List of Tables

3.1 Integrated MSEs of estimators and MSEs on test data in Study 1 . . . . . . 61
3.2 Integrated MSEs of estimators and MSEs on test data in Study 2 . . . . . . 66
List of Figures

2.1 The first two FPCs of the pinch force data ................................. 6
2.2 An example of B-spline basis functions .................................... 10
2.3 Estimated FPCs and their pointwise MSEs in Simulation 1 .............. 22
2.4 Comparison of explained variance in Simulation 1 .......................... 23
2.5 Estimated FPCs and their pointwise MSEs in Simulation 2 .............. 24
2.6 Comparison of the explained variance in Simulation 2 ..................... 25
2.7 Daily temperature curves in Canadian weather data ........................ 26
2.8 The estimated FPCs for Canadian weather data .............................. 26
2.9 The recorded force curves in pinch force data .............................. 27
2.10 The estimated FPCs for pinch force data .................................... 28
2.11 Comparison of the explained variance in Application 2 ................... 29

3.1 The true coefficient function $\beta(t)$ in Study 2 ............................ 61
3.2 Pointwise bias of $\hat{\beta}(t)$ in Study 1 ..................................... 63
3.3 Pointwise standard deviation $\hat{\beta}(t)$ in Study 1 .......................... 63
3.4 Pointwise RMSE of $\hat{\beta}(t)$ in Study 1 .................................... 64
3.5 Boxplot and histogram of estimated intercepts in Study 1 ............... 64
3.6 True coefficient functions in Study 2 ....................................... 65
3.7 Pointwise bias of $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ in Study 2 .................. 67
3.8 Pointwise standard deviation of $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ in Study 2 ...... 67
3.9 Pointwise bias of $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ in Study 2 .................. 68
3.10 Boxplot and histogram of estimated intercepts in Study 2 ............... 68
3.11 Estimates of the coefficient function in Canadian weather data ........ 69
List of Algorithms

1  The Greedy Backward Elimination Algorithm .......................... 15
2  The iFPCA Algorithm ..................................................... 16
Chapter 1

Introduction

In the last two decades, functional data analysis has emerged into the statistical society as an important tool for analyzing data where the observations can be viewed as curves defined on a continuum. A non-extensive list of classic topics in functional data analysis includes spline smoothing, functional principal component analysis (FPCA), functional linear regression, etc. For more practical and theoretical treatments on the subject of functional data analysis, readers are referred to an excellent introductory text written by Ramsay and Silverman (2005).

A basic task in functional data analysis is to estimate some unknown curves from a set of sample curves. For example, in FPCA, the fundamental problem is to estimate the first several principal component curves. A popular approach is to transform estimating a functional object into estimating a finite-dimension vector through basis expansion, such as Fourier basis expansion and B-spline expansion. A common issue of basis expansion approaches is the excessive roughness exhibited by estimators. The issue comes from the fact that in practice we only have a finite number of observations, and each sample curve is usually only observed at some discrete design points and meanwhile contaminated by random errors. Additional smoothing regularization is required to eliminate the undesired roughness. In fact, smoothness is of fundamental importance to a good estimator of a curve, and the development of smoothing methods is one of the main streams in the area of functional data analysis. Among many other methods, regularization via roughness penalty is a very popular smoothing technique.

In ordinary data analysis, statistical models with a sparse representation are of practical merit, particularly when the number of predictors is large. Sparse models are usually
less variable and more interpretable. Locally sparse models in functional data analysis also possess these two desired qualities, where “locally sparse” means a curve is identically zero on some sub-regions of the domain. However, in the literature, there are very few works concerning the interpretability and the locally sparse representation of functional objects. Also, existing methods either use an indirect regularization technique or have multiple stages, which makes them complicated and difficult to implement in practice. In this thesis, we advance locally sparse modeling in functional data analysis by developing some new regularization techniques that are simpler but yet possess nice statistical properties, and applying them to derive smooth and locally sparse estimators in various tasks of functional data analysis.

In Chapter 2, we concerns locally sparse estimators of functional principal components. FPCA is a common tool to perform dimensionality reduction when a sample of random curves is observed. It is also a tool to explore the major source of variability in the observed curves around the overall mean curve. For example, FPCs are often interpreted by looking at the change of the FPCs over the entire time range, i.e., the sub-regions with relatively large magnitude of FPCs are interpreted as where the curves have major variations. We propose a novel penalty-based method and combine it with regularization on roughness to derive smooth and locally sparse FPCs that are nonzero on the sub-regions where curves have major variations and strictly zero in others. Our estimated FPCs are easier to interpret in the sense that, the sub-regions on which our estimated FPCs are nonzero indicate where sample curves have major variations. To compute locally sparse FPCs in practice, we design an efficient algorithm based on projection deflation. We also establish the consistency and asymptotic normality of our estimators.

In Chapter 3, we develop a new regularization method called “functional SCAD” (fSCAD for short) for locally sparse modeling and then apply it to derive a smooth and locally sparse estimator of the coefficient function in functional linear regression. The fSCAD can be viewed as a functional generalization of the ordinary SCAD regularization proposed by Fan and Li (2001). When combined with penalized B-spline expansion in practice, it can also be seen as a generalized SCAD with diverging number of parameters. The combination of the proposed fSCAD technique and penalized B-spline expansion yields smooth and locally sparse estimates of coefficient functions in functional linear regression models. Comparing to existing methods in the literature, our method for smooth and locally sparse modeling has two distinct features. First, unlike the FLiRTI method proposed by James et al. (2009),
in which local sparsity is indirectly regularized by imposing sparsity on transformed basis coefficients, we use fSCAD to directly regularize the estimated coefficient function for local sparsity. Second, unlike the two-stage procedure in Zhou et al. (2013), our method combines the fSCAD and the roughness penalty together in a single optimization criterion to produce an estimate of the null sub-regions and a smooth estimate of $\beta(t)$ on its non-null sub-regions simultaneously in a single stage. These two features make our estimating procedure conceptually more elegant and computationally simpler. Moreover, our smooth and locally sparse estimator also enjoys the oracle property, which means it performs as well as if we knew the true null region of the coefficient function in advance.

In Chapter 4, we summarize the thesis.
Chapter 2

Interpretable FPCA

2.1 Introduction and Motivation

Functional principal component analysis (FPCA) is a common tool to perform dimensionality reduction when a sample of random curves is observed. It aims to discover the major source of variability in the observed curves around the overall mean curve. Since introduced by Rao (1958) when he studied a comparison of growth curves, it has attracted many researchers’ attention. For example, Castro et al. (1986) investigated some fundamentals of FPCA, such as relating FPCA to Karhunen-Loève theorem and the best \( m \)-dimensional functional linear model. Dauxois et al. (1982) studied asymptotic properties of empirical eigenfunctions and eigenvalues when the sample curves are fully observable. Benko et al. (2009) extended this work to a more practical setting where sample curves themselves have to be reconstructed from noisy observations at finite design points. Hall and Hosseini-Nasab (2006, 2009) studied some statistical properties, particularly the high-order estimation errors of empirical eigenfunctions and eigenvalues. To overcome excessive variability of empirical eigenfunctions, Rice and Silverman (1991) proposed smoothed estimators of eigenfunctions via penalizing the sample variance for the roughness of empirical eigenfunctions. The consistency of the estimators was established by Pezzulli and Silverman (1993). Subsequently, Silverman (1996) proposed an alternative way to obtain smoothed estimators of eigenfunctions through modifying the norm structure, and established the consistency under some regularity conditions. Qi and Zhao (2011) established the asymptotic normality of the estimators of Silverman (1996). A kernel-based method for smoothing eigenfunctions was proposed by Boente and Fraiman (2000). Classical FPCA assumes either that sample curves
are fully observable or that a set of dense and regularly spaced points of sample curves is available. The extension of FPCA to sparse data such as longitudinal data was studied by James et al. (2000) and Yao et al. (2005). An introductory exposition of FPCA can be found in Chapters 8 and 9 of Ramsay and Silverman (2005).

FPCs are often interpreted by looking at the change of the FPCs over the entire time range, i.e. the sub-regions with relatively significant magnitude of FPCs are interpreted as where the curves have major variations. If FPCs are nonzero on these sub-regions while strictly zero on others, then it is easier to interpret these FPCs. For instance, these FPCs can be interpreted as that major variations of the curves exist only in those non-null sub-regions. However, these sub-regions are often hard to be identified by na"ive users. For example, in Figure 2.1, we show the first two FPCs of the pinch force data (see Section 2.6 for details), which are estimated by the classic regularized FPCA method (Silverman, 1996). The first FPC is nonzero roughly on the interval $[7, 65]$ and the second FPC is nonzero roughly on the interval $[5, 50]$, which can be interpreted as that the group of curves has major variations on these two intervals. To identify these intervals, a straightforward simple hard-thresholding method would set an FPC curve to zero on the intervals where the absolute value of the FPC is below a prespecified threshold. However, this approach suffers from two problems. First, the obtained new curve is discontinuous and hence nonsmooth. The discontinuity will make further analysis and inference based on FPCs difficult. Second, an FPC should be produced so that it points to a direction explaining most or at least a large portion of variance that has not been explained by other FPCs. But the hard-thresholding method does not guarantee that the new FPC it produces from the original one points to a new direction that accounts for a large portion of the unexplained variance. To the best of our knowledge, the problem of identifying sub-regions with major variations has not been studied in the literature.

In this chapter, we propose a novel penalty-based method to derive smooth FPCs that are nonzero on the sub-regions where curves have major variations and strictly zero on others. In comparison with the classic regularized FPCA method (Silverman, 1996), our estimated FPCs are more interpretable in the sense that, the sub-regions on which our estimated FPCs are nonzero indicate where sample curves have major variations. Therefore, we call our method as the interpretable functional principal component analysis, or in short, iFPCA. Our estimated functional principal component is called iFPC. The main idea of iFPCA is to penalize the supports of smooth FPCs. In practice, these iFPCs can be estimated via the
basis approach where each basis function has a compact support. The main contributions of this work include developing an efficient method to estimate iFPCs and showing that the estimated iFPCs are asymptotically normal and strongly consistent under some regularity conditions.

The rest of the chapter is organized as follows. In Section 2.2, we give a background on regularized FPCA. We then introduce our iFPCA method in Section 2.3. The asymptotic normality and consistency of the estimated iFPCs are stated in Section 2.4. Section 2.5 includes two simulation studies to illustrate the finite sample performance of our iFPCA method. Section 2.6 demonstrates our iFPCA method in two real applications. We summarize the chapter in Section 2.7. Proof of theorems are collected in Section 2.8.

2.2 Preliminaries

Let $X(t)$ be a square integrable stochastic process on a compact interval $\mathcal{I}$. That is to say $X \in L^2(\mathcal{I})$ almost surely, where $L^2(\mathcal{I})$ is the Hilbert space of square integrable functions on $\mathcal{I}$ equipped with the usual inner product $\langle f, g \rangle = \int_{\mathcal{I}} f(t)g(t) \, dt$ and the corresponding norm $\|f\| = \sqrt{\langle f, f \rangle}$ for $f, g \in L^2(\mathcal{I})$. Since our focus is on estimating eigenfunctions which are invariant to the overall mean function of $X$, without loss of generality, we assume $\mathbb{E}X(t) = 0$ for all $t \in \mathcal{I}$. Then the covariance function $C$ of $X$ is $C(s, t) = \mathbb{E}[X(s)X(t)]$ for $s, t \in \mathcal{I}$. Its corresponding covariance operator $\mathcal{C}$ is defined by the mapping $(\mathcal{C}f)(s) = \int_{\mathcal{I}} C(s, t)f(t) \, dt$. Assume $X$ and $C$ satisfy the conditions of Karhunen-Loève theorem so that $X$ admits a
Karhunen-Loève expansion

\[ X(t) = \sum_{k=1}^{\infty} Z_k \xi_k(t), \]

and \( C \) admits a decomposition

\[ C(s, t) = \sum_{k=1}^{\infty} \lambda_k \xi_k(s) \xi_k(t), \]

where \( Z_1, Z_2, \ldots \) are uncorrelated random variables, functions \( \{ \xi_k : k = 1, 2, \ldots \} \) form an orthonormal basis of \( L^2(\mathcal{I}) \), and each \( \xi_k(t) \) is an eigenfunction of \( C \) corresponding to the eigenvalue \( \lambda_k \). Furthermore, each \( Z_k \) has mean \( \mathbb{E}Z_k = 0 \) and variance \( \mathbb{V}Z_k = \lambda_k \).

**Notation Conventions.** Throughout this chapter, we write \( \xi(t) \) when we emphasize that \( \xi \) is a function of \( t \in \mathcal{I} \), and write \( \xi \) when we treat it as an element in the function space \( L^2(\mathcal{I}) \). Operators on \( L^2(\mathcal{I}) \) are written by calligraphic fonts. Vectors and matrices are written in boldface. Random variables are written in capital letters. A hat or tilde on a symbol is used to denote an estimated value. For instance, \( \hat{\xi} \) or \( \tilde{\xi} \) stands for an estimated value of \( \xi \). The symbol \( \mathbb{E} \) is reserved for expectation while \( \mathbb{V} \) is for variance.

The task of functional principal component analysis is to estimate the first \( m \) eigenfunctions of the unknown operator \( C \) based on a sample of \( n \) observed curves \( X_1(t), X_2(t), \ldots, X_n(t) \). It is well known that each function \( \xi_k(t) \) is the solution of the following successive optimization problems

\[
\max_{\|\xi\| = 1, \xi \perp H_{k-1}} \frac{\langle \xi, C\xi \rangle}{\|\xi\|^2} \tag{2.1}
\]

where \( H_{k-1} \) denotes the space spanned by \( \xi_1, \xi_2, \ldots, \xi_{k-1} \). Accordingly, the ordinary FPCA is formulated as to find the first \( m \) FPCs \( \hat{\xi}_k \) \( (k = 1, 2, \ldots, m) \) that solve the successive optimization problems of (2.1) but with \( C \) replaced by the empirical covariance operator \( \hat{C} \) corresponding to the empirical covariance function \( \hat{C}(s, t) = \frac{1}{n} \sum_{i=1}^{n} X_i(s)X_i(t) \), and \( H_k \) replaced by \( \hat{H}_k \overset{\text{def}}{=} \text{span}\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k\} \). Here, \( m \) is a positive integer determined by users.

However, eigenfunctions that are estimated based on the empirical covariance function, \( \hat{C}(s, t) \), exhibit excessive variability in general, so that smoothing is desirable. Regularization through a roughness penalty is a popular method to smooth estimators of functions. Among others, the estimation method proposed by Silverman (1996) incorporates a roughness penalty on estimated eigenfunctions through modifying the default norm \( \| \cdot \| \). Below we briefly describe this methodology when a penalty is made on the second-order derivative.
However, we shall point out that depending on the application, one can also penalize the derivatives of other orders. For example, to pursue more smoothness, one can penalize the derivative of a higher order.

Let $W_2^2$ be the Sobolev space

$$W_2^2 \overset{\text{def}}{=} \{ f : f, f' \text{ are absolutely continuous on } \mathcal{I} \text{ and } f'' \in L^2(\mathcal{I}) \}.$$ 

Define linear operator $D^2$ on $W_2^2$ by $D^2 f = f''$, i.e. $D^2$ is the second-order derivative operator on $W_2^2$. For a given scalar parameter $\gamma \geq 0$, define a new inner product

$$\langle f, g \rangle_\gamma = \langle f, g \rangle + \gamma \langle D^2 f, D^2 g \rangle$$

and its corresponding norm $\|f\|_\gamma = \sqrt{\|f\|^2 + \gamma \|D^2 f\|^2}$ on $W_2^2$. Then, the roughness penalty is enforced by using the new norm $\|\cdot\|_\gamma$, rather than the default one. That is, the $k$th smooth FPC $\tilde{\xi}_k$ is found to maximize

$$\langle \xi, \hat{C}\xi \rangle \frac{\|\xi\|^2}{\|\xi\|^2_\gamma}$$

subject to $\|\xi\|_\gamma = 1$ and $\langle \xi, \tilde{\xi}_j \rangle_\gamma = 0$ for $j < k$. The advantage of this formulation has been discussed by Silverman (1996).

### 2.3 Methodology

#### 2.3.1 Formulation

Generally, a smooth FPC tends to be nonzero over the entire domain. To obtain an FPC which is strictly zero except on the sub-regions where sample curves have major variations, we propose to place a penalty on the support of the estimated FPC. More specifically, if $S(\xi) \overset{\text{def}}{=} \int_{\mathcal{I}} 1\{\xi(t) \neq 0\} \, dt$ denotes the length of the support of $\xi(t)$, then the $k$th iFPC $\hat{\xi}_k(t)$ is chosen from $W_2^2$ to maximize

$$\frac{\langle \xi, \hat{C}\xi \rangle}{\|\xi\|^2 + \gamma \|D^2 \xi\|^2} - \rho_k S(\xi)$$

subject to $\|\xi\|_\gamma = 1$ and $\langle \xi, \hat{\xi}_j \rangle_\gamma = 0$ for $j < k$, where each $\rho_k > 0$ is a prespecified parameter. The intuition is that, a large $\rho_k$ prefers an iFPC $\hat{\xi}(t)$ with a larger region on which $\hat{\xi}(t)$ is
zero. When $\rho_k = 0$, the iFPC $\hat{\xi}_k(t)$ is identical to the smooth FPC proposed by Silverman (1996), while as $\rho_k$ approaches to the infinity, $\hat{\xi}_k(t)$ tends to be zero everywhere. Therefore, the parameter $\rho_k$ controls the trade-off between fidelity and interpretability.

### 2.3.2 A Practical Approach

In this section we develop a practical approach to estimate the iFPCs defined by (2.2) via a basis approach. We choose a set of basis functions $\{\phi_j(t) : 1 \leq j \leq p\}$ each with a small support of roughly equal length. An example of such a basis is the B-spline basis. To define the B-spline basis, we divide the compact interval $\mathcal{I}$ into subintervals separated by a sequence of knots. Over each subinterval, a B-spline basis function is a polynomial of specified order at most $M$. An order $M$ B-spline basis function is nonzero over no more than $M$ consecutive subintervals. This property is called the compact support property, which is very important for efficient computation and makes B-spline basis functions popularly used (Ramsay and Silverman, 2005). When a dense sequence of knots is used to define the B-spline basis, each B-spline basis function is only nonzero in a very short subinterval. Figure 2.2 shows one of 23 basis functions in a cubic B-spline basis system with 21 equally spaced knots. This basis function is only nonzero on the interval $[0.1,0.3]$. For more treatments on B-splines, we recommend the text written by de Boor (1978). In this chapter, we assume that all basis functions are only nonzero on short subintervals with the same length except for a few basis functions on the boundaries. This assumption is well satisfied by the B-spline basis, which will be used in our simulation studies and applications.

Suppose each curve $X_i(t)$, $i = 1, \ldots, n$, is measured or observed at some discrete design points in the presence of random errors. We estimate $X_i(t)$ by $\hat{X}_i(t) = \sum_{j=1}^p s_{ij} \phi_j(t)$ from the noisy data using the B-spline smoothing method (Ramsay and Silverman, 2005). If we use $\phi$ to denote the column vector $(\phi_1, \phi_2, \ldots, \phi_p)^T$, then $(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)^T = S\phi$, where $S$ is an $n \times p$ matrix with elements $S(i, j) = s_{ij}$.

Suppose each function $\xi(t)$ in (2.2) has an expansion $\xi(t) = \sum_{j=1}^p a_j \phi_j(t)$. Then the number of nonzero coefficients $a_j$ can serve as a surrogate of $S(\xi)$ because of the compact support property of the basis functions $\phi_j(t)$. Similarly, if $a$ denotes the vector $(a_1, a_2, \ldots, a_p)^T$, then $\xi = a^T \phi$. The covariance function is then estimated by

$$
\hat{C}(s, t) = \frac{1}{n} \sum_{i=1}^n \hat{X}_i(s)\hat{X}_i(t) = \frac{1}{n} \phi^T(s)S^T S \phi(t)
$$
and the variance explained by $\xi$ is estimated by
\[
\langle \xi, \hat{\xi} \rangle = \frac{1}{n} \int a^T \phi(t) \phi^T(t) a \, ds \, dt = \frac{1}{n} a^T W S^T S W a,
\]
where $W(i, j) = \langle \phi_i, \phi_j \rangle$. Also, $\|\xi\|^2 = \langle a^T \phi, a^T \phi \rangle = a^T W a$. Similarly, if $R$ denotes a matrix such that $R(i, j) = \langle D^2 \phi_i, D^2 \phi_j \rangle$, then $\|\nabla^2 \xi\|^2 = a^T R a$. If we define $G = W + \gamma R$ and $Q = \frac{1}{n} W S^T S W$, then the first term of (2.2) becomes
\[
\frac{\langle \xi, \hat{\xi} \rangle}{\|\xi\|^2 + \gamma \|\nabla^2 \xi\|^2} = \frac{1}{n} a^T W S^T S W a}{a^T (W + \gamma R) a} = \frac{a^T Q a}{a^T G a}. \tag{2.3}
\]
We also use the number $\|a\|_0$ of nonzero loadings of $a$ as an approximate surrogate of $S(\xi)$. Thus, the optimization problem (2.2) is approximated as finding the vector $\hat{a}_k$ that maximizes
\[
\frac{a^T Q a}{a^T G a} - \rho_k \|a\|_0
\]
subject to $a^T G a = 1$ and $a^T G \hat{a}_j = 0$ for $j < k$, which respectively correspond to the two constraints of (2.2).

In practice, due to the orthogonality constraint, except the coefficient vector $\hat{a}_1$ of the first iFPC $\hat{\xi}_1$, other $\hat{a}_k$'s are not free to choose its zero loadings and hence may be forced to have no zero loading. Therefore, in the case that we want to relax the orthogonality.
constraint for pursuing more interpretability, the objective function (2.4) is difficult to work with. To circumvent the difficulty, we employ a “deflation” technique analogous to the matrix deflation detailed by White (1958) for computing eigenvalues and eigenvectors of a matrix.

We first describe our deflation technique in functional settings. The basic idea is to remove the effect of the first \( k - 1 \) principal components from the data when we estimate the \( k \)th principal component. More specifically, suppose we already have \( \hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1} \). The effect of these components are removed by projecting each sample curve \( X_i \) into the subspace perpendicular to all of \( \hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1} \). The projection of \( X_i \) in this way is called the residual of \( X_i \) perpendicular to the subspace \( \hat{H}_{k-1} = \text{span}\{\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}\} \), denoted by \( \hat{r}_k^i \) (note that \( \hat{r}_1^i = X_i \)). Then the empirical covariance function \( \hat{C}_k \) after the effect of the first \( k - 1 \) components is removed is estimated by

\[
\hat{C}_k(s,t) = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_k^i(s)\hat{r}_k^i(t). \tag{2.5}
\]

Based on \( \hat{C}_k \), the \( k \)th iFPC is found to maximize

\[
\langle \xi, \hat{C}_k \xi \rangle - \rho_k S(\xi) \tag{2.6}
\]

subject to \( \|\xi\|_\gamma = 1 \). Analogous to the projection deflation of a matrix described in Mackey (2009), we call our deflation “projection deflation”, and \( \hat{C}_k \) is the deflated covariance function. It is worth noting that estimators \( \hat{\xi}_1, \hat{\xi}_2, \ldots \) with the use of deflation technique or without, are close to be orthogonal but not strictly orthogonal in the \( L^2(\mathcal{F}) \) space.

We now show how to compute iFPCs using projection deflation based on the chosen B-spline basis system that we previously introduced. First of all, each residual can be written as a linear combination of basis functions, i.e., \( \hat{r}_k^i = \sum_{j=1}^{p} S_{ij}^{(k)} \phi_j \). Let \( S_k \) be the matrix with elements \( S_k(i,j) = S_{ij}^{(k)} \). An efficient method to compute \( S_k \) is provided in Section 2.3.5. Now define \( Q_k = \frac{1}{n} WS_k^T S_k W \) (note that \( S_1 = S \) and \( Q_1 = Q \)). Also, as aforementioned, it is assumed \( \xi = a^T \phi \). Then

\[
\langle \xi, \hat{C}_k \xi \rangle = \frac{1}{n} \int a^T \phi(s) \phi^T(s) S_k^T S_k \phi(t) \phi^T(t) a \, ds \, dt = \frac{1}{n} a^T W S_k^T S_k W a = a^T Q_k a.
\]
Again, by using $\|a\|_0$ to serve as surrogate of $S(\xi)$, the optimization problem (2.6) is transformed into finding the coefficient vector $\hat{a}_k$ of the $k$th iFPC $\hat{\xi}_k$ to maximize

$$a^T Q_k a - \rho_k \|a\|_0$$

subject to $a^T G a = 1$. Or alternatively we find the vector $\hat{a}_k$ to

$$\begin{align*}
&\text{maximize} & a^T Q_k a \\
&\text{subject to} & a^T G a = 1 \text{ and } \|a\|_0 \leq \tau_k,
\end{align*}$$

where the constant $\tau_k$ is a prespecified positive integer. By formulation (2.8), we can relax the orthogonality of estimated FPCs in exchange to interpretability. Also, the optimization problem (2.8) is equivalent to the optimization problem (2.7) in the sense that if $\hat{a}_k$ is the optimal solution of (2.7), then $\hat{a}_k$ is also the optimal solution of (2.8) with $\tau_k = \|\hat{a}_k\|_0$ (d’Aspremont et al., 2008). The advantage of (2.8) is that it is relatively easier to choose a value of $\tau_k$ than $\rho_k$, as $\tau_k$ is directly related to the penalty on the support of $\hat{\xi}_k$. Unfortunately, the optimization problem (2.8) is NP-hard (Wang and Wu, 2012), which means that no efficient algorithm is known to solve it. In the next section, we develop a greedy backward elimination method to approximately solve the optimization problem (2.8).

### 2.3.3 The iFPCA Algorithm

We first introduce some notations. Let $\alpha$ denote a subset of $\{1, 2, \ldots, p\}$, $M_\alpha$ denote the submatrix of $M$ obtained by keeping only those rows and columns whose indices are in $\alpha$, and $a_\alpha$ denote the subvector of $a$ by keeping the elements whose indices are in $\alpha$. To reduce the notational burden, the subscript $k$, used for the $k$-th FPC in the previous sections, is suppressed in this section. Now consider in (2.8) fixing the $i_1, i_2, \ldots, i_j$-th elements of $a$ to zero. This is equivalent to zeroing all the $i_1, i_2, \ldots, i_j$-th columns and rows of the matrices $G$ and $Q$, and hence the optimization problem (2.8) reduces to finding the vector $\hat{a}_\alpha$ that maximizes

$$a_\alpha^T Q_\alpha a_\alpha$$

subject to $a_\alpha^T G_\alpha a_\alpha = 1$ where $\alpha = \{1, 2, \ldots, p\} - \{i_1, i_2, \ldots, i_j\}$. It is easy to show that $\hat{a}_\alpha$ is an eigenvector corresponding to the largest eigenvalue $\lambda_1(Q_\alpha, G_\alpha)$ in the generalized eigenvalue problem (GEP) $Q_\alpha a_\alpha = \lambda G_\alpha a_\alpha$. Therefore, to solve the optimization problem
(2.8), it is crucial to find out a subset $\alpha$ of indices $\{1,2,\ldots,p\}$ such that $|\alpha| = \tau$ and $\lambda_1(Q_{\alpha}, G_{\alpha})$ is maximized.

The idea of backward elimination to find such $\alpha$ is to remove a nonzero loading of $a$ iteratively until there are only $\tau$ nonzero loadings left. In each iteration, the loading with the minimum impact on explaining variance will be chosen in a greedy fashion. More precisely, if $\alpha_j$ denotes the value of $\alpha$ when the $j$-th iteration is completed, and by convention, $\alpha^0 = \{1,2,\ldots,p\}$, then the $i_j$th loading of $a$ such that

$$i_j = \arg\min_{i \in \alpha_{j-1}} \left( \lambda_1(Q_{\alpha_{j-1}}, G_{\alpha_{j-1}}) - \lambda_1(Q_{\alpha_{j-1}-\{i\}}, G_{\alpha_{j-1}-\{i\}}) \right)$$

(2.9)

will be nullified in the $j$th iteration, and $\alpha^j = \alpha^{j-1} - \{i_j\}$. Note that in (2.9), $\lambda_1(Q_{\alpha_{j-1}}, G_{\alpha_{j-1}}) - \lambda_1(Q_{\alpha_{j-1}-\{i\}}, G_{\alpha_{j-1}-\{i\}})$ is the loss of explained variance due to elimination of the $i$th loading.

It requires $O(p^3)$ FLOPS (FLoating-point Operations Per Second) to solve the optimization (2.9), because it takes at least $O(p^2)$ FLOPS to compute the leading eigenvalue of a generalized eigenvalue problem with two $p \times p$ matrices. When $p$ is large, it is still quite computationally inefficient. Instead of computing the exact value of $\lambda_1(Q_{\alpha_{j-1}}, G_{\alpha_{j-1}}) - \lambda_1(Q_{\alpha_{j-1}-\{i\}}, G_{\alpha_{j-1}-\{i\}})$, we borrow the idea of the approximate minimum variance loss (AMVL) criterion by Wang and Wu (2012), originally proposed in the setting of ordinary principal component analysis: choose $i_j$ with the smallest upper bound of the loss of variance due to elimination of the $i_j$th loading, provided that the upper bound is easy to compute. We now derive an upper bound of the loss of variance due to elimination of the $i$th loading. The inequality below generalizes the one in Wang and Wu (2012).

**Theorem 2.1.** Suppose $A$ and $B$ are $p \times p$ symmetric matrices, and $B$ is invertible. Denote $\lambda_1(A, B)$ as the largest eigenvalue of the generalized eigenvalue problem $Ax = \lambda Bz$, and $v$ as the corresponding eigenvector. Let $v_i$ be the $i$-th element of the vector $v$, $\alpha = \{1,2,\ldots,p\} - \{i\}$, $v_\alpha$ be the subvector of $v$ by keeping the elements whose indices are in $\alpha$, and

$$B^{-1} = \begin{pmatrix} F & f_i \\ f_i^T & y \end{pmatrix}, \quad B^{-1} A = \begin{pmatrix} P_1 & P_2 \\ P_3^T & c \end{pmatrix}.$$
\[ \lambda_1(A, B) - \lambda_1(A_\alpha, B_\alpha) \leq \frac{v_1 y^{-1} f^T_i v_\alpha [\lambda_1(A, B) - c] + v_i v_\alpha^T d_2}{1 - v_i^2}. \] (2.10)

Provided \( \lambda_1(A, B), B^{-1} \) and \( B^{-1}A \) have been computed, the upper bound in (2.10) can be computed in \( O(p) \) extra FLOPS. One can check that the matrix \( G_\alpha \) is positive-definite, symmetric and hence invertible. Therefore, Theorem 2.1 applies to the matrix pair \((Q_\alpha, G_\alpha)\). Based on the inequality (2.10), by using the AMVL criterion, the optimization (2.9) can be approximately solved in \( O(p^2) \) FLOPS provided that \( \lambda_1(Q_\alpha, G_\alpha) \) has been computed. When \( p \) is relatively small, the Cholesky decomposition can be used to compute all eigenvalues and eigenvectors of the GEP \( Q_\alpha a_\alpha = \lambda G_\alpha a_\alpha \). When \( p \) is large, the Lanczos algorithm is preferred since it can take only \( O(qp^2 \log p) \) FLOPS (Kuczyński and Woźniakowski, 1992; Leyk and Woźniakowski, 1996) to accurately estimate the leading eigenvalue and its eigenvector of the GEP \( Q_\alpha a_\alpha = \lambda G_\alpha a_\alpha \), where \( q \) is a constant factor depending on the tolerance level. When \( p \) is large, \( q \) is typically much smaller than \( p \) (i.e. typically, \( q \ll p \)). Therefore, in total it takes \( O(p^2 \log p) \) FLOPS for the AMVL approach to approximately solve the problem (2.9).

The greedy backward elimination algorithm based on AVML criterion to compute the first iFPC is outlined in Algorithm 1. Note that in Algorithm 1, by using the technique described in Section 2.3.4, the update to \( G_\alpha^{-1} \) and \( G_\alpha^{-1} Q_\alpha \) in each iteration can be done in \( O(p^2) \) FLOPS. Therefore, the computational complexity of the greedy backward elimination algorithm is \( O(p^3 \log p) \) FLOPS by noting that \( \tau \leq p \). The iFPCA algorithm is outlined in Algorithm 2, which uses the projection deflation described in Section 2.3.5 to iteratively compute \( \hat{a}_k \) by applying the greedy backward elimination algorithm (Algorithm 1) on matrices \( Q_k \) and \( G \) until the first \( m \) iFPCs are obtained. In total, the iFPCA algorithm takes \( O(nmp^2 + m^2 p^2 + mp^3 \log p) \) FLOPS to complete. When \( n = O(p) \) and \( m \) is fixed, the computational complexity of the iFPCA algorithm grows in the order of \( O(p^3 \log p) \). This complexity is almost the same as the regularized FPCA described in Ramsay and Silverman (2005), which requires at least \( O(p^3) \) FLOPS.

### 2.3.4 Efficient Update of \( G_\alpha^{-1} \) and \( G_\alpha^{-1} Q_\alpha \)

Generally, it takes \( O(p^3) \) to compute \( G_\alpha^{-1} \) and \( G_\alpha^{-1} Q_\alpha \) from the scratch. However, in each iteration of Algorithm 1, we can take the advantage of the fact that \( G^{-1}_{\alpha \cup \{i\}} \) and \( G^{-1}_{\alpha \cup \{i\}} Q_{\alpha \cup \{i\}} \) have been computed in the previous iteration or in the initialization stage,
Algorithm 1 The Greedy Backward Elimination Algorithm

1: $\beta \leftarrow \emptyset$
2: $\alpha \leftarrow \{1, 2, \ldots, p\}$
3: $s \leftarrow p$
4: compute $G^{-1}_\alpha$ and $G^{-1}_\alpha Q_\alpha$
5: compute $\lambda_1(Q_\alpha, G_\alpha)$ and its corresponding normalized eigenvector $v$.
6: while $s > \tau$ do
7:     for all $i \in \alpha$ do
8:         $h_i \leftarrow \frac{v_i y^T v_i (\lambda(Q_\alpha, G_\alpha) - c) + v_i v_i^T p^2}{1 - v_i^T}$
9:     end for
10:     $i_s \leftarrow \arg \min_i h_i$
11:     $\beta \leftarrow \beta \cup \{i_s\}$
12:     $\alpha \leftarrow \alpha - \{i_s\}$
13:     update $G^{-1}_\alpha$ and $G^{-1}_\alpha Q_\alpha$ according to Appendix A
14:     update $\lambda_1(Q_\alpha, G_\alpha)$ and its corresponding normalized eigenvector $v$.
15:     $s \leftarrow s - 1$
16: end while
17: let $a$ be a vector such that $a(\beta) = 0$ and $a(\alpha) = v$
18: normalize $a$ so that $a^T G a = 1$
19: output $\lambda_1(Q_\alpha, G_\alpha)$ and the vector $a$. The corresponding estimated iFPC is $\hat{\xi}(t) = a^T \phi(t)$.

To efficiently update $G^{-1}_\alpha$ and $G^{-1}_\alpha Q_\alpha$. Below we present a technique which can bring the computational cost of updating $G^{-1}_\alpha$ and $G^{-1}_\alpha Q_\alpha$ down to $O(p^2)$ FLOPS in each iteration of Algorithm 1.

Suppose we write the following matrices in a block form:

$$Q_{\alpha \cup \{i\}} = \begin{pmatrix} Q_\alpha & q_i \\ q_i^T & w \end{pmatrix}, \quad G^{-1}_{\alpha \cup \{i\}} = \begin{pmatrix} F & f_i \\ f_i^T & y \end{pmatrix}, \quad G^{-1}_{\alpha \cup \{i\}} Q_{\alpha \cup \{i\}} = \begin{pmatrix} P_1 & P_2 \\ P_3^T & c \end{pmatrix}.$$

Then it is easy to check that

$$G^{-1}_\alpha = F - \frac{1}{y} f_i f_i^T.$$

Therefore, if $G^{-1}_{\alpha \cup \{i\}}$ is already computed, then only $O(p^2)$ additional FLOPS are required to compute $G^{-1}_\alpha$. Also,

$$G^{-1}_\alpha Q_\alpha = F Q_\alpha - \frac{1}{y} f_i f_i^T Q_\alpha = (F Q_\alpha + f_i q_i^T) - y^{-1} f_i (f_i^T Q_\alpha + y q_i^T) = P_1 - y^{-1} f_i p_3^T.$$
CHAPTER 2. INTERPRETABLE FPCA

Algorithm 2 The iFPCA Algorithm

1: compute matrices $\mathbf{S}, \mathbf{W}, \mathbf{R}$ based on the chosen basis and the given data
2: $\mathbf{G} \leftarrow \mathbf{W} + \gamma \mathbf{R}$
3: for all $k = 1, 2, \ldots, m$ do
4: $\mathbf{Q} \leftarrow \mathbf{W}^T \mathbf{S} \mathbf{W}$
5: compute $\hat{a}_k$ and $\hat{\lambda}_k$ using Algorithm 1 on $\mathbf{G}$ and $\mathbf{Q}$
6: Update $\mathbf{S}$ according to Appendix B
7: end for
8: the total variance of data is estimated by $\lambda_{\text{tot}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}(i,:) \mathbf{W}(i,:)^T$
9: output the estimated first $m$ iFPCs: $\hat{\xi}_1(t) = \hat{a}_1^T \phi(t)$, $\hat{\xi}_2(t) = \hat{a}_2^T \phi(t)$, \ldots, $\hat{\xi}_m(t) = \hat{a}_m^T \phi(t)$. The estimated percentages of variance that each iFPC takes account of are calculated as $\hat{\lambda}_1/\lambda_{\text{tot}}, \hat{\lambda}_2/\lambda_{\text{tot}}, \ldots, \hat{\lambda}_m/\lambda_{\text{tot}}$, respectively.

$$
\mathbf{G}_{\alpha \cup \{i\}}^{-1} \mathbf{Q}_{\alpha \cup \{i\}} = \begin{pmatrix}
\mathbf{F} & \mathbf{f}_i \\
\mathbf{f}_i^T & y
\end{pmatrix}
\begin{pmatrix}
\mathbf{Q}_\alpha & \mathbf{q}_i \\
\mathbf{q}_i^T & w
\end{pmatrix}
= \begin{pmatrix}
\mathbf{FQ}_\alpha + \mathbf{f}_i \mathbf{q}_i^T & \mathbf{Fq}_i + w \mathbf{f}_i \\
\mathbf{f}_i^T \mathbf{Q}_\alpha + y \mathbf{q}_i^T & \mathbf{f}_i^T \mathbf{q}_i + wy
\end{pmatrix}
= \begin{pmatrix}
\mathbf{P}_1 & \mathbf{p}_2 \\
\mathbf{p}_2^T & c
\end{pmatrix}.
$$

Therefore, $\mathbf{G}_{\alpha \cup \{i\}}^{-1} \mathbf{Q}_{\alpha \cup \{i\}}$ can also be computed in $O(p^2)$ FLOPS provided that $\mathbf{G}_{\alpha \cup \{i\}}^{-1} \mathbf{Q}_{\alpha \cup \{i\}}$ is computed already.

2.3.5 Projection Deflation

In this section we described the projection deflation technique in detail. We also present an efficient method to compute the matrix $\mathbf{S}_k$ introduced in Section 2.3.2.

First of all, recall that $\hat{r}_i^k$ denotes the residual of $\hat{X}_i$ perpendicular to the space spanned by the first $k - 1$ iFPCs $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$, i.e. $\hat{r}_i^k = \hat{X}_i - \sum_{j=1}^{k-1} b_{ij} \hat{\xi}_j$ for some scalars $b_{ij}$. However, as the iFPCs $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$ may not be orthogonal in the inner product $\langle \cdot, \cdot \rangle$, the coefficients $b_{ij}$‘s are dependent on the index $k$. In other words, we have to recompute $b_{ij}$‘s for each $\hat{r}_i^k$, which is computationally inefficient. To address this issue, we use the Gram-Schmidt process to orthogonalize $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$ as we progressively compute iFPCs.

Suppose $\eta_1, \eta_2, \ldots, \eta_{k-1}$ are orthogonal functions produced by Gram-Schmidt process on $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$. Instead of representing the residual $\hat{r}_i^k$ by $\hat{r}_i^k = \hat{X}_i - \sum_{j=1}^{k-1} b_{ij} \hat{\xi}_j$, we write $\hat{r}_i^k = \hat{X}_i - \sum_{j=1}^{k-1} b_{ij} \eta_j$ and each $\eta_\nu = \sum_{j=1}^{k-1} h_{\nu j} \phi_j$ for $\nu = 1, 2, \ldots, k - 1$. Recall that the element $s_{ij}^{(k)} = \mathbf{S}_k(i,j)$ of the matrix $\mathbf{S}_k$ is the coefficient of $\hat{r}_i^k$ with respect to the $j$th basis function $\phi_j$. Suppose we already have $\mathbf{S}_k, \eta_1, \eta_2, \ldots, \eta_{k-1}, \hat{\xi}_k$ and we want to compute $\mathbf{S}_{k+1}$
(note that $S_1 = S$). We first continue the Gram-Schmidt process on $\eta_1, \eta_2, \ldots, \eta_{k-1}$ and $\hat{\xi}_k = \hat{a}_k^T \phi$ to obtain $\eta_k$. It is equivalent to computing the residual $\vartheta$ of $\hat{\xi}_k$ with respect to the space spanned by $\eta_1, \eta_2, \ldots, \eta_{k-1}$ and then normalize $\vartheta$. Let $c_\nu \eta_\nu$ be the projection of $\hat{\xi}_k$ on $\eta_\nu = h_\nu^T \phi$. Then $\langle \hat{\xi}_k - c_\nu \eta_\nu, \eta_\nu \rangle = 0$, which implies $\hat{a}_k \text{Wh}_\nu = c_\nu h_\nu^T \text{Wh}_\nu$ and hence

$$c_\nu = \frac{\hat{a}_k \text{Wh}_\nu}{h_\nu^T \text{Wh}_\nu}.$$ 

Let $c$ denote the vector $(c_1, c_2, \ldots, c_k)^T$. Then $\vartheta = \hat{\xi}_k - \sum_{\nu=1}^k c_\nu \eta_\nu = (\hat{a}_k^T - c^T H) \phi$ where $H(\nu, j) = h_{\nu j}$. By normalizing $\vartheta$ to have unit length, we obtain $\eta_k$. This process takes $O(kp^2)$ FLOPS to complete.

To compute the residual of $\hat{X}_i$ perpendicular to the subspace spanned by the first $k$ iFPCs, we only need to compute the residual of $\hat{r}_i^k$ with respect to $\eta_k$ since by assumption $\hat{r}_i^k \perp \eta_j$ and $\eta_k \perp \eta_j$ for all $j < k$. Thus, the residual $\hat{r}_i^{k+1}$ is found to make $\hat{r}_i^k = b_{ik} \eta_k + \hat{r}_i^{k+1}$ hold for some scalar $b_{ik}$ and $(\hat{r}_i^{k+1}, \eta_k) = 0$. In matrix form, it becomes $S_k^k \phi = bh_k^T \phi + S_{k+1}^{k+1} \phi$ and $S_{k+1} \text{Wh}_k = 0$, where $h_k = (h_{k1}, h_{k2}, \ldots, h_{kp})^T$. They together give

$$b = \frac{S_k \text{Wh}_k}{h_k^T \text{Wh}_k}, \quad S_{k+1} = S_k - \frac{S_k \text{Wh}_k h_k^T}{h_k^T \text{Wh}_k}.$$ 

This process takes $O(np^2)$ FLOPS to complete. Thus, in total it takes $O(np^2 + kp^2)$ FLOPS to compute $S_{k+1}$.

### 2.3.6 Tuning Parameter Selection

The optimization criterion (2.8) has two tuning parameters: $\gamma$ controls the roughness of iFPCs, and $\tau_k$ controls the sparsity or interpretability of the $k$-th iFPC, $k = 1, \ldots, m$. A two-dimensional cross-validation procedure can be employed to search good values of $\gamma$ and $\tau_k$ simultaneously. Alternatively, we propose a two-step procedure, in which $\gamma$ is chosen in the first step, and $\tau$ is determined in the second step. Our computational experiments show that the two-step procedure produces reasonably good results and is computationally efficient.

In the first step, $\gamma$ can be chosen based on prior knowledge, required degree of smoothness, or cross-validation (CV) as described in Ramsay and Silverman (2005). Once $\gamma$ is chosen, $\tau_k$ can be chosen by CV in the second step, which is introduced as follows. Suppose the first $(k - 1)$ iFPCs, $\hat{\xi}_1(t), \ldots, \hat{\xi}_{k-1}(t)$, have been obtained, and we are now estimating
the $k$-th iFPC $\hat{\xi}_k(t)$. To determine $\tau_k$, for each positive integer $\tau_k = 1, 2, \ldots, p$ (Recall that $p$ is the number of basis functions, and is the maximum possible value of $\tau_k$), we compute the corresponding CV score. For example, if a $K$-fold CV procedure is employed, then the CV score for $\tau_k$ is

$$ CV(\tau_k) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i \in P_j} \|r_i^{(-j)}(\tau_k)\|^2, $$

where $P_j$ is the $j$-th partition used in the $K$-fold CV procedure, and $r_i^{(-j)}(\tau_k)$ is the residual of $X_i(t)$ orthogonal to the span of the iFPCs $\hat{\xi}_1(t), \ldots, \hat{\xi}_{k-1}(t)$, and $\hat{\xi}_k(t)$. Here $\hat{\xi}_k^{(-j)}(t)$ is the $k$-th iFPC estimated from all sample curves excluding those in the $j$-th partition.

In principle, the tuning parameter $\tau_k$ is chosen by minimizing the CV score. On the other hand, as $\tau_k$ increases, the decrease of CV score is sometimes very small. Then choosing $\tau_k$ which minimizes the CV score may result in an unnecessarily large value of $\tau_k$. In this case, the well known “+1 SE” rule or its variants can be used (Breiman et al., 1984). The “+1 SE” rule selects the smallest $\tau$ whose CV score is below the smallest CV score plus one standard error of the smallest CV score.

Alternatively, if we want to control the loss of variance explained due to the pursuit of interpretability, then we can choose the smallest $\tau_k$ so that the variance explained by iFPCA is at least $\delta\%$ of the variance explained by the regularized FPCA method (Silverman, 1996), where $\delta$ is a prespecified number between 0 and 100.

### 2.4 Asymptotic Properties

In the special case that the tuning parameter used in (2.2) is set as $\rho_k = 0$ for all $k = 1, 2, \ldots, m$, our iFPCA estimators, $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$, and $\hat{\xi}_1(t), \hat{\xi}_2(t), \ldots, \hat{\xi}_m(t)$, are reduced to the regularized FPCA estimators, $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m$, and $\tilde{\xi}_1(t), \tilde{\xi}_2(t), \ldots, \tilde{\xi}_m(t)$, proposed by Silverman (1996). The regularized FPCA estimators have been shown to be strongly consistent by Silverman (1996) and have the asymptotic normality property by Qi and Zhao (2011). In this section, we will show that our iFPCA estimators also enjoy these nice properties under more general conditions on the $\rho_k$’s. We present the main results in this section while provide their proofs in Section 2.8. Except where otherwise stated, the notation is understood as defined in the previous sections.

In our setup, the sample functions, $X_1(t), X_2(t), \ldots, X_n(t)$, are assumed to be independently and identically distributed realizations of the stochastic process $X(t)$ that is
CHAPTER 2. INTERPRETABLE FPCA

introduced in Section 2.2. We shall further assume the following technical conditions, which are identical to those made in Silverman (1996):

(A1) The covariance function $C$ is strictly positive-definite and the trace $\int_{\mathcal{I}} C(t,t) \, dt$ is finite. This assumption ensures that the covariance operator $C$ is a Hilbert-Schmidt operator and hence a compact operator. Thus, the operator $C$ has a complete sequence of eigenfunctions $\xi_1(t), \xi_2(t), \ldots$, and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Furthermore, it implies $E(\|X\|^4) < \infty$.

(A2) Each eigenfunction $\xi_k(t)$ falls in the space $W_2^2$. When roughness penalty is made on derivatives of other orders, the space $W_2^2$ is replaced by the corresponding suitable space of “smooth” functions on $\mathcal{I}$.

(A3) The first $m$ eigenvalues have multiplicity 1 and hence $\lambda_1 > \lambda_2 > \cdots > \lambda_m > \lambda_{m+1} \geq \cdots > 0$. This condition is assumed for simplicity. The result can be extended to the general case.

Note that even when an eigenvalue $\lambda_k$ has multiplicity 1, both $\xi_k(t)$ and $-\xi_k(t)$ are its corresponding eigenfunctions, although they have opposite directions. In the sequel, when we mention eigenfunctions, their directions are assumed to be given. Also, we assume that the directions of iFPCs $\hat{\xi}_k(t)$ and Silverman’s FPCs $\tilde{\xi}_k(t)$ are chosen so that $\langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma \geq 0$ and $\langle \tilde{\xi}_k, \xi_k \rangle \geq 0$ for all $k = 1, 2, \ldots, m$.

Theorem 2.2 (Asymptotic Normality). Let $\hat{\xi}_k(t)$ be the estimated iFPC that maximizes (2.2), and $\hat{\lambda}_k = \langle \hat{\xi}_k, \hat{\xi}_k \rangle$. Assume conditions A1-A3 hold. As $n \to \infty$,

(a) the joint distribution of

\[ \{\sqrt{n}(\hat{\lambda}_1 - \lambda_1), \sqrt{n}(\hat{\lambda}_2 - \lambda_2), \ldots, \sqrt{n}(\hat{\lambda}_m - \lambda_m)\} \]  

(2.11)

converges to a Gaussian distribution with mean zero, if $\gamma = o(n^{-1/2})$ and $\rho_k = o(n^{-1/2})$ for all $k = 1, 2, \ldots, m$;

(b) the vector

\[ \{\sqrt{n}(\hat{\xi}_1 - \xi_1), \sqrt{n}(\hat{\xi}_2 - \xi_2), \ldots, \sqrt{n}(\hat{\xi}_m - \xi_m)\} \]  

(2.12)

converges to a Gaussian random element with mean zero, if $\gamma = o(n^{-1})$ and $\rho_k = o(n^{-1})$ for all $k = 1, 2, \ldots, m$. 
Theorem 2.2 already implies that our iFPCA estimators are consistent. Moreover, we will show that they are strongly consistent in Theorem 2.3.

**Theorem 2.3** (Strong Consistency). Let \( \hat{\xi}_k(t) \) be the estimated iFPC that maximizes (2.2), and \( \hat{\lambda}_k = \langle \hat{\xi}_k, \hat{C}\hat{\xi}_k \rangle \). Under the conditions A1-A3, if \( \gamma \to 0 \) and \( \rho_k \to 0 \) for all \( k = 1, 2, \ldots, m \), then \( \hat{\lambda}_k \to \lambda_k \) and \( \hat{\xi}_k \to \xi_k \) almost surely as \( n \to \infty \) for all \( k = 1, 2, \ldots, m \).

In Section 2.3, we use projection deflation technique to relax the orthogonality constraint. Even in this case, our iFPCA estimators are still strongly consistent, as shown in the following theorem.

**Theorem 2.4** (Strong Consistency using Deflation Technique). Let \( \hat{\xi}_k(t) \) be the estimated iFPC that maximizes (2.6), and \( \hat{\lambda}_k = \langle \hat{\xi}_k, \hat{C}\hat{\xi}_k \rangle \). Under the conditions A1-A3, if \( \gamma \to 0 \) and \( \rho_k \to 0 \) for all \( k = 1, 2, \ldots, m \), then \( \hat{\lambda}_k \to \lambda_k \) and \( \hat{\xi}_k \to \xi_k \) almost surely as \( n \to \infty \) for all \( k = 1, 2, \ldots, m \).

### 2.5 Simulation Studies

We evaluate the finite sample performance of our iFPCA method and compare it with the classic regularized FPCA method (Silverman, 1996) in two simulation studies. In the first simulation study, the true FPCs are nonzero in some sub-regions and strictly zero elsewhere. This scenario is in favour of our iFPCA method. In this scenario, we will show that our iFPCA method, in comparison with the classic regularized FPCA method, produces more accurate FPC estimators, and the loss of variance explained by the estimated iFPPCs is negligible. In the second simulation study, the true FPCs are almost nonzero over the entire domain. This scenario is in favour of the classic regularized FPCA method, but we will show that our iFPCA method is still competitive to the classic regularized FPCA method in this scenario.

#### 2.5.1 Simulation 1

In this simulation study, 500 curves are simulated by \( X_i(t) = b_{i1}\xi_1(t) + b_{i2}\xi_2(t) + b_{i3}\xi_3(t) \) where \( b_{ik} \sim N(0, \lambda_k) \), \( \lambda_1 = 3^2 \), \( \lambda_2 = 2.5^2 \) and \( \lambda_3 = 2^2 \), \( t \in [0, 1] \). The three true FPCs, \( \xi_k(t), k = 1, 2, 3 \), are simulated as \( \xi_k(t) = \sum_{\ell=1}^{53} a_{k\ell}\phi_\ell(t) \), where each \( \phi_\ell(t) \) is one of the 53 basis functions of the cubic B-spline basis system with 51 knots equally spaced in \([0,1]\).
For each FPC $\xi_k$ ($k = 1, 2, 3$), all of its basis coefficients $a_{k\ell}$, except $2k + 2$ of them, are zero. Figure 2.3 displays the three true FPCs in the top three panels. The simulated data are generated as pairs $(t_j, y_{ij})$ for $i = 1, 2, \ldots, 500$ and $j = 1, 2, \ldots, 51$, where $t_j$ is the $j$th design point equally-spaced on $[0,1]$, and $y_{ij} = X_i(t_j) + \epsilon_{ij}$ is the value of $X_i$ at $t_j$ with a random measurement error $\epsilon_{ij} \sim N(0,1)$.

We estimate the three FPCs using our iFPCA method and the classic regularized FPCA method from the simulated data. Both methods first estimate $X_i(t)$, $i = 1, \ldots, 500$, from the simulated data by penalized spline smoothing method (Ramsay and Silverman, 2005). The spline basis system we use is a cubic B-spline basis system with 51 equally-spaced knots in $[0,1]$. Both our iFPCA method and the classic regularized FPCA method express each of the three FPCs as a linear combination of the same basis functions. Both methods set the tuning parameter, $\gamma$, which controls the roughness of the estimated FPCs, as $\gamma = 10^{-7}$.

We choose $\tau_k$, $k = 1, 2, 3$, which controls the interpretability of the $k$-th iFPC, by 10-fold cross-validation in the iFPCA method. The simulation is implemented with 100 simulation replicates.

The top three panels in Figure 2.3 display the estimated FPCs using our iFPCA method and the classic regularized FPCA method in one random simulation replicate. They show that the estimated FPCs using the classic regularized FPCA method are nonzero on almost the entire interval $[0, 1]$, and all of them have three modes instead of the true one mode. In contrast, each estimated FPC using the iFPCA method has only one mode, and is only nonzero on almost the same interval as the corresponding true FPC. The above results from the two FPCA methods are also seen in most of the other 99 simulation replicates. In fact, the estimated FPCs using our iFPCA method have the same nonzero intervals as the corresponding true FPCs in 97%, 98%, and 100% of the total 100 simulation replicates for the first, second and third FPCs, respectively. The bottom three panels in Figure 2.3 show the pointwise mean squared errors (MSE) of the estimated FPCs using the two FPCA methods. All three estimated FPCs using our iFPCA method have smaller pointwise MSEs than those using the classic regularized FPCA method on almost the entire interval. Our iFPCA method also seems to obtain more robust estimates of FPCs than the classic regularized FPCA method.

In comparison with the classic regularized FPCA method, the iFPCA method has the additional constraint that FPCs have to be zero in most intervals to increase their interpretability. Therefore, the iFPCA method has the risk of having the estimated FPCs
explaining less variance of the sample curves. Figure 2.4 summarizes the average cumulative percentages of variance explained by the FPCs estimated using our iFPCA method and the classic regularized FPCA method. It shows that the FPCs estimated using our iFPCA method explain almost the same cumulative percentages of variance of the sample curves, less than 3.2% on average. Figure 2.4 also displays the boxplot of the differences of cumulative percentages of variance explained by FPCs using the two methods in 100 simulation replicates. The differences of cumulative percentages of explained variance are all less than 4.5% in 100 simulation replicates. These differences are partly due to the fact that the classic regularized FPCA method undesirably catches more random measurement errors, as it is observed that FPCs estimated by this method are nonzero almost on the entire domain and each of them has three modes, rather than just one true mode.
Figure 2.4: The left panel shows the average cumulative percentages of variance explained by the estimated FPCs using our iFPCA method in comparison with the classic regularized FPCA method in 100 simulation replicates in Simulation 1. The right panel displays the boxplot of the loss of cumulative percentages of variance explained by the estimated FPCs using our iFPCA method compared with the classic regularized FPCA method in 100 simulation replicates.

2.5.2 Simulation 2

In this simulation study, 500 curves are simulated with the same settings as in the previous simulation study except that the three true FPCs, $\xi_k(t)$, $k = 1, 2, 3$, are simulated as $\xi_1(t) = \sin 2\pi t$, $\xi_2(t) = \sin 4\pi t$, and $\xi_3(t) = \sin 6\pi t$, where $t \in [0, 1]$. All three of these FPCs are designed such that they are nonzero in almost the entire domain. We estimate the three FPCs using our iFPCA method and the classic regularized FPCA method. We use the same cubic B-spline basis functions and the same method to choose tuning parameters as in Simulation 1. The simulation is implemented with 100 simulation replicates. In this scenario, the iFPCA method is expected to not perform as well as the classic regularized FPCA method. But we will see that the iFPCA method is still quite competitive with the classic regularized FPCA method.

Figure 2.5 displays the estimated FPCs using our iFPCA method and the classic regularized FPCA method in one random simulation replicate. It shows that the estimated FPCs using the iFPCA method are very close to the true FPCs. The bottom three panels in Figure 2.5 also show the pointwise mean squared errors (MSE) of the estimated FPCs using the two methods. The pointwise MSE of the estimated FPCs using the iFPCA method is
CHAPTER 2. INTERPRETABLE FPCA

Figure 2.5: The top three panels display the estimated FPCs using our iFPCA method and the classic regularized FPCA method in one random simulation replicate in Simulation 2. The bottom three panels show the pointwise mean squared errors (MSE) of the estimated FPCs using the two methods.

only slightly larger than those using the classic regularized FPCA method.

Figure 2.6 shows that the average cumulative percentages of variance explained by the estimated FPCs using our iFPCA method are almost the same as those estimated by the classic regularized FPCA method. The differences of the cumulative percentages of explained variance using the two methods in 100 simulation replicates are no larger than 6%.

2.6 Applications

2.6.1 Application 1: Canadian Weather Data

Figure 2.7 displays the daily temperature recorded at 35 Canadian cities in one year. These data have been analyzed in Chapter 8 of Ramsay and Silverman (2005) to use FPCA to find the major sources of variability of the 35 temperature curves. We will use our
proposed iFPCA method to analyze these 35 curves, and compare this method with the classic regularized FPCA method.

Both the iFPCA method and the classic regularized FPCA method first estimate the underlying continuous temperature curves from the discrete temperature measurements by penalized spline smoothing method (Ramsay and Silverman, 2005). The spline basis system we use is a cubic B-spline basis system with 365 equally-spaced knots in [1,365]. Both the iFPCA method and the classic regularized FPCA method express each of the three FPCs as a linear combination of the same basis functions. Both methods set the tuning parameter, $\gamma$, which controls the roughness of the estimated FPCs, as $\gamma = 3000$. The iFPCA method chooses the other tuning parameters $\tau_k, k = 1, 2, 3$, which control the interpretability of iFPCs, by 10-fold cross-validation.

Figure 2.8 displays the first three FPCs estimated by our iFPCA method and the classic regularized FPCA method. These three estimated FPCs using the two methods have the same trend, so their interpretations are consistent. On the other hand, the three FPCs estimated by the classic regularized FPCA method are nonzero in the entire time interval, while the estimated three iFPCs are only nonzero in a short time interval. Hence, the
estimated three iFPCs can help us identify the precise time intervals on which the 35 curves have major variations.

The estimated first iFPC is only nonzero in $[1, 150] \cup [247, 365]$, which indicates that the first major variation of the 35 temperature curves happens in the spring, fall and winter. The first iFPC explains 82.7% variance of the 35 temperature curves, which is only 6.2% less than the percentage of variance explained by the first FPC using the classic regularized FPCA. The estimated second iFPC is only nonzero in $[132, 257]$, which represents the second major variation of the 35 curves that happens in the summer. The second iFPC explains 8.9% variance of the 35 curves, which is only 0.35% less than the percentage of variance
explained by the second FPC estimated by the classic regularized FPCA method. The third iFPC is only positive in $[77, 143]$ and negative in $[1, 20]$, which indicates that the third major variation of the 35 curves is the temperature change between spring and winter.

### 2.6.2 Application 2: Pinch Force Data

Figure 2.9 displays the recorded force exerted by the thumb and forefinger during 20 brief pinches. Silverman (1996) analyzed the pinch force data using the classic regularized FPCA method. We will use the iFPCA method to analyze the same data, and compare the iFPCA method with the classic regularized FPCA method.

![Figure 2.9: The recorded force exerted by the thumb and forefinger during 20 brief pinches.](image)

Both methods first estimate the underlying continuous functions from the discrete data by penalized spline smoothing method (Ramsay and Silverman, 2005). The spline basis system we use is a cubic B-spline basis system with 151 equally-spaced knots in $[0,100]$. In both the iFPCA method and the classic regularized FPCA method, each of the three FPCs is expressed as a linear combination of the same basis functions. Both methods set the tuning parameter, $\gamma$, which controls the roughness of the estimated FPCs, as $\gamma = 37$, as advised by Silverman (1996). The second tuning parameter $\tau_k$, $k = 1, 2, 3$, which controls the interpretability of the $k$-th iFPC, is chosen by 10-fold cross-validation in the iFPCA method.

Figure 2.10 shows the estimated three FPCs using our iFPCA method and the classic
regularized FPCA method for the pinch force data. All three FPCs estimated by both methods are very close, but the first two estimated FPCs using the iFPCA method are strictly zero in some time intervals, which increases the interpretability of the FPCs. For example, the first FPC estimated by the iFPCA method is only nonzero in [8.7, 59.1], and it accounts for 67.7% of the total variance of the 20 pinch force curves. It may be interpreted as that at least 67.7% variation of the 20 pinch force curves is concentrated in this time interval. The third FPC is far away from zero on the entire time interval because of its nature of explaining variation of sample curves, and our iFPCA method is able to produce a very similar estimate for the third FPC as the classic regularized FPCA method. Beside the improved interpretability of the estimated three FPCs using our iFPCA method, they also explain almost the same percentages of total variance of the 20 pinch force curves, as shown in Figure 2.11.

![Figure 2.10: The estimated three FPCs using our iFPCA method (solid lines) and the classic regularized FPCA method (dashed lines) for the pinch force data.](image)

2.7 Summary

Functional principal component analysis (FPCA) has been widely used to identify the major source of variations of a sample of random curves. These variation sources are represented by estimated functional principal components (FPCs). The sub-regions on which FPCs are significantly far away from zero are interpreted as the parts of the sample curves that have major variations. However, the significance threshold is hard to be determined by naïve users.

In this chapter, we propose a novel interpretable FPCA (iFPCA) method, which can
estimate FPCs that are only nonzero on some sub-regions and strictly zero on the rest intervals. The sub-regions on which FPCs are nonzero can be interpreted as the parts of the sample curves that have major variations. We develop an efficient algorithm to estimate the interpretable FPCs using projection deflation. We also show that the estimated interpretable FPCs are strongly consistent and asymptotically normal under mild conditions. The simulation studies show that our iFPCA method can obtain FPCs of more interpretability than other FPCA methods, while at the same time these interpretable FPCs account for similar variations of the sample curves as the FPCs that are estimated by other FPCA methods. Our iFPCA method is also demonstrated by analyzing real Canadian temperature data and pinch force data as given in Ramsay and Silverman (2005).

2.8 Proofs

The proofs for Theorem 2.1, Theorem 2.2, 2.3 and 2.4 are organized as follows. The proof of Theorem 2.1 is given in Section 2.8.1. In Section 2.8.2, we outline our proof strategy for Theorem 2.2 and then briefly introduce some results important to our proofs from Qi and Zhao (2011). After that, we start our proofs by first establishing Lemma 2.5 in Section 2.8.3, which is a key result in our proofs. This lemma not only serves as a technical tool in proving our theorems, but also is of independent interest. Roughly speaking, it asserts two
finite-dimensional Hilbert spaces are close to each other if their bases are close to each other. In Subsection 2.8.4, we proceed to prove Theorem 2.2 and 2.3 by first establishing Lemma 2.7 which contains the essence of the proofs of the two theorems. Some technical details of the proof of this lemma are moved to Lemma 2.8 and Lemma 2.9 so that the mainstream of the proof is manifest. The two theorems are proved in the end of the subsection. Section 2.8.5 is dedicated to the proof of Theorem 2.4. We first establish three auxiliary lemmas, namely, Lemma 2.10, Lemma 2.11 and Lemma 2.12. At last, we prove Theorem 2.4.

2.8.1 Proof of Theorem 2.1

Proof of Theorem 2.1. First of all, note that when $B$ is invertible, the generalized eigenproblem $Az = \lambda Bz$ is equivalent to the ordinary eigenproblem $B^{-1}Az = \lambda z$. Write the matrices $A$, $B$, $B^{-1}A$ and the vector $v$ in a block form so that

$$A = \begin{pmatrix} A_\alpha & a_i \\ a_i^T & w \end{pmatrix}, \quad B = \begin{pmatrix} B_\alpha & b_i \\ b_i^T & z \end{pmatrix}, \quad B^{-1} = \begin{pmatrix} F & f_i \\ f_i^T & y \end{pmatrix}, \quad v = \begin{pmatrix} v_\alpha \\ v_i \end{pmatrix}. $$

By nothing that $B_\alpha^{-1} = F - \frac{1}{y} f_i f_i^T$, we have

$$v_\alpha^T B_\alpha^{-1} A_\alpha v_\alpha = v_\alpha^T (FA_\alpha - y^{-1} f_i f_i^T A_\alpha) v_\alpha \leq \lambda(A_\alpha, B_\alpha) \|v_\alpha\|^2 = \lambda(A_\alpha, B_\alpha) (1 - v_i^2),$$

or equivalently,

$$v_\alpha^T (FA_\alpha + f_i a_i^T) v_\alpha - y^{-1} v_\alpha^T f_i (f_i^T A_\alpha + ya_i^T) v_\alpha \leq \lambda(A_\alpha, B_\alpha) (1 - v_i^2). \quad (2.13)$$

Now write $B^{-1}A$ in a block form so that

$$B^{-1}A = \begin{pmatrix} P_1 & p_2 \\ p_3^T & c \end{pmatrix}.$$

Then $FA_\alpha + f_i a_i^T = P_1$ and $f_i^T A_\alpha + ya_i^T = p_3^T$ since

$$B^{-1}A = \begin{pmatrix} F & f_i \\ f_i^T & y \end{pmatrix} \begin{pmatrix} A_\alpha & a_i \\ a_i^T & w \end{pmatrix} = \begin{pmatrix} FA_\alpha + f_i a_i^T & Fa_i + wf_i \\ f_i^T A_\alpha + ya_i^T & f_i^T a_i + wy \end{pmatrix} = \begin{pmatrix} P_1 & p_2 \\ p_3^T & c \end{pmatrix}. \quad (2.14)$$
From (2.14) and $v^T B^{-1} A v = \lambda(A, B)$, we have
\[ v_\alpha^T P_1 v_\alpha + v_i p_3^T v_\alpha + v_i v_\alpha^T p_2 + v_i^2 c = \lambda(A, B). \]
Thus,
\[ v_\alpha^T P_1 v_\alpha = \lambda(A, B) - (v_i v_\alpha^T (p_2 + p_3) + v_i^2 c). \]
With (2.13), this gives
\[ \lambda(A, B, B_\alpha) \geq \frac{\lambda(A, B) - v_i v_\alpha^T (p_2 + p_3) - v_i^2 c - y^{-1} v_\alpha^T \xi f p_3^T v_\alpha}{1 - h^2}, \]
or
\[ \lambda(A, B) - \lambda(A, B_\alpha) \leq \frac{y^{-1} v_\alpha^T \xi f p_3^T v_\alpha + v_i v_\alpha^T (p_2 + p_3) + v_i^2 c - v_i^2 \lambda(A, B)}{1 - v_i^2} \]
Noting that $v_\alpha^T p_3 + v_i c = v_i \lambda(A, B)$, we have further
\[ \lambda(A, B) - \lambda(A, B_\alpha) \leq \frac{v_i y^{-1} f_i^T v_\alpha [\lambda(A, B) - c] + v_i v_\alpha^T p_2}{1 - v_i^2}. \] (2.15)
This proves Theorem 1.

2.8.2 Proof Strategy for Other Theorems

In Qi and Zhao (2011), $\sqrt{n}(\hat{\xi}_n - \xi_k)$ has been shown to converge to a Gaussian random element. As $\sqrt{n}(\hat{\xi}_n - \xi_k) = \sqrt{n}(\hat{\xi}_k - \xi_k) + \sqrt{n}(\hat{\xi}_k - \xi_k)$, if we can show
\[ \|\hat{\xi}_k - \xi_k\| = o_p(n^{-1/2}), \] (2.16)
then $\sqrt{n}(\hat{\xi}_k - \xi_k)$ converges to 0 in probability and hence by Slutsky’s theorem, $\sqrt{n}(\hat{\xi}_k - \xi_k)$ and $\sqrt{n}(\hat{\xi}_k - \xi_k)$ converge to the same random element in distribution. Instead of directly proving (2.16), we consider a more general problem. Take a scalar sequence $\{\nu_n\}$ convergent to zero and then choose $\rho_k = o(\nu_n)$ for all $k = 1, 2, \ldots, m$. When $\nu_n = n^{-1/2}$ we recover the assumption on $\rho_k$’s in (i) of Theorem 2.2, and when $\nu_n = n^{-1}$, we recover the assumption on $\rho_k$’s in (ii) of Theorem 2.2. In other words, we will show that $|\hat{\lambda}_k - \lambda_k|$ and $\|\hat{\xi}_k - \xi_k\|^2$ have a convergence rate at least as fast as $\rho_k$. That is, $|\hat{\lambda}_k - \lambda_k| = O_p(\rho_k)$ and $\|\hat{\xi}_k - \xi_k\|^2 = O_p(\rho_k)$.

In our proof, we need the “half-smoothing” operator that is mentioned in Silverman
(1996) and studied more thoroughly in Qi and Zhao (2011). Here we cite some results crucial to our proofs from Qi and Zhao (2011). Recall that $D^2$ is the second-derivative operator on $W^2_2$. Let $D^2_*$ be the Hilbert adjoint of $D^2$. Define $S^2 = (I + \gamma D^2_* D^2)^{-1}$. The inverse $S^{-1}$ of $S$ exists and is a self-adjoint operator. By Lemma 6 of Qi and Zhao (2011), \{$(\tilde{\lambda}_k, S^{-1}\tilde{\xi}_k) : k \in \mathbb{N}$\} are eigenvalues and eigenfunctions of the compact positive operator $\hat{S}C\hat{S}$, and moreover, there are no other eigenvalues for $\hat{S}C\hat{S}$. Furthermore, under the assumption that the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ of the operator $C$ have multiplicity one, $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m$ have multiplicity one almost surely. Therefore, we can assume $\tilde{\lambda}_1 > \tilde{\lambda}_2 > \cdots > \tilde{\lambda}_m$ without affecting the asymptotic properties. As $\hat{S}C\hat{S}$ is a compact self-adjoint positive operator, its eigenfunctions \{${S^{-1}\tilde{\xi}_k} : k \in \mathbb{N}$\} form an orthogonal basis of $L^2(I)$ and equipped with the inner product $\langle \cdot, \cdot \rangle_\gamma$, as the operator $S : L^2(I) \rightarrow W^2_2$ is one-to-one. Also, for each $\eta \in W^2_2$, we have $\langle S^{-1}\eta, S^{-1}\eta \rangle_\gamma = \langle \eta, \eta \rangle_\gamma$. Therefore, we can derive the following useful equation

$$\langle \hat{\xi}_i, \hat{\xi}_j \rangle = \langle S^{-1}\hat{\xi}_i, (S^2\hat{S}S^{-1}\hat{\xi}_j) \rangle = \tilde{\lambda}_j \langle S^{-1}\hat{\xi}_i, S^{-1}\hat{\xi}_j \rangle = \begin{cases} \tilde{\lambda}_j & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

### 2.8.3 A Key Lemma

Let $P^\perp$ be the projection onto $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$, with respect to the inner product $\langle \cdot, \cdot \rangle_\gamma$. Let $\hat{P}^\perp$ be the projection onto $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$ with respect to the same inner product structure. Note that we are dealing with two inner products in $W^2_2$: the one from $L^2(I)$ and the modified one $\langle \cdot, \cdot \rangle_\gamma$. These two projections are both defined in terms of the inner product $\langle \cdot, \cdot \rangle_\gamma$. Intuitively, when each $\hat{\xi}_j$ is close to $\tilde{\xi}_j$, the projections $\hat{P}$ and $\hat{P}$ are expected to be close to each other in the operator norm $\| \cdot \|_\gamma$ defined as

$$\| U \|_\gamma = \sup_{\eta \in W^2_2, \| \eta \|_\gamma = 1} \| U\eta \|_\gamma.$$

The following lemma gives a more precise statement of this intuition. In the sequel the notation $o(1)$ is understood as some quantity converging to zero almost surely (with probability one).

**Lemma 2.5.** Fix any $\gamma \geq 0$. 
(a) If \( \langle \hat{\xi}_j, \tilde{\xi}_j \rangle_\gamma^2 = 1 + o_p(t_n) \) for all \( j < k \), then \( \| \hat{P}^\perp - \tilde{P}^\perp \|_\gamma = o_p(\sqrt{t_n}) \) and \( \| \hat{P} - \tilde{P} \|_\gamma = o_p(\sqrt{t_n}) \).

(b) If \( \langle \hat{\xi}_j, \tilde{\xi}_j \rangle_\gamma^2 = 1 + o(1) \) for all \( j < k \), then \( \| \hat{P}^\perp - \tilde{P}^\perp \|_\gamma = o(1) \) and \( \| \hat{P} - \tilde{P} \|_\gamma = o(1) \).

These conclusions hold even when \( \hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1} \) are not orthogonal with respect to the inner product \( \langle \cdot, \cdot \rangle_\gamma \).

Before beginning the proof, we shall point out that in the above lemma claim (a) does not imply claim (b), as the convergence in (a) is “in probability” while the one in (b) is “almost surely”.

Proof. We first prove (a). For each \( j < k \), expanding \( \hat{\xi}_j \) by the basis \( \tilde{\xi}_1, \tilde{\xi}_2, \ldots \) we assume \( \hat{\xi}_j = \sum_{i=1}^{\infty} b_{ji} \tilde{\xi}_i \) where \( b_{ji} = \langle \hat{\xi}_j, \tilde{\xi}_i \rangle_\gamma \). First, note that

\[
\sum_{i \neq j} b_{ji}^2 = \sum_{i=1}^{\infty} b_{ji}^2 - b_{jj}^2 = \| \hat{\xi}_j \|_\gamma^2 - b_{jj}^2 = 1 - b_{jj}^2 = 1 - \langle \hat{\xi}_j, \tilde{\xi}_j \rangle_\gamma^2.
\]

This implies

\[
\sum_{i \neq j} b_{ji}^2 = o_p(t_n)
\]

because by assumption \( \langle \hat{\xi}_j, \tilde{\xi}_j \rangle_\gamma^2 = 1 + o_p(t_n) \). Hence,

\[
\| \hat{P}^\perp \hat{\xi}_j - \tilde{\xi}_j \|_\gamma = \sqrt{\sum_{i \geq k} b_{ji}^2} = o_p(\sqrt{t_n})
\]

for each \( j < k \). Let \( \zeta \) be any element in \( W_2^2 \) such that \( \| \zeta \|_\gamma = 1 \). The projection of \( \zeta \) into the subspace spanned by \( \hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1} \) can be written as \( \hat{P}^\perp \zeta = \sum_{j=1}^{k-1} c_j \tilde{\xi}_j \) (if \( \hat{\xi}_i = g \hat{\xi}_j \) for some scalar \( g \) and \( i < j \), then we set \( c_i = 0 \)). Let \( \eta = \zeta - \hat{P}^\perp \zeta \). Then for each \( j < k \),

\[
0 = \langle \eta, \hat{\xi}_j \rangle_\gamma = \sum_{i=1}^{\infty} b_{ji} \langle \eta, \tilde{\xi}_i \rangle_\gamma = b_{jj} \langle \eta, \tilde{\xi}_j \rangle_\gamma + \sum_{i \neq j} b_{ji} \langle \eta, \tilde{\xi}_i \rangle_\gamma.
\]

Thus,

\[
|b_{jj} \langle \eta, \tilde{\xi}_j \rangle_\gamma| = \left| \sum_{i \neq j} b_{ji} \langle \eta, \tilde{\xi}_i \rangle_\gamma \right|.
\]
On the other hand, by Cauchy-Schwarz inequality
\[
\left| \sum_{i \neq j} b_{ji} \langle \eta, \bar{\xi}_i \rangle \gamma \right|^2 \leq \left( \sum_{i \neq j} b_{ji}^2 \right) \left( \sum_{i \neq j} \langle \eta, \bar{\xi}_i \rangle^2 \gamma \right) \leq \left( \sum_{i \neq j} b_{ji}^2 \right) \| \eta \|^2_\gamma \leq \left( \sum_{i \neq j} b_{ji}^2 \right). \tag{2.22}
\]

Combining (2.21) and (2.22) we have
\[
0 \leq |b_{jj} \langle \eta, \bar{\xi}_j \rangle \gamma|^2 \leq \left( \sum_{i \neq j} b_{ji}^2 \right).
\]

In light of (2.19), this implies
\[
|\langle \eta, \bar{\xi}_j \rangle \gamma|^2 \leq |b_{jj}|^{-2} \left( \sum_{i \neq j} b_{ji}^2 \right) = o_p (\iota_n)
\]

because \( b_{jj}^2 = 1 + o(\iota_n) \). Thus,
\[
\| \bar{\mathcal{P}} \perp \eta \|_\gamma = \sqrt{\sum_{j<k} \langle \eta, \bar{\xi}_j \rangle^2_\gamma} = o_p (\sqrt{\iota_n}).
\]

This result, combined with (2.20), gives
\[
\| \bar{\mathcal{P}} \perp \zeta - \bar{\mathcal{P}} \perp \hat{\zeta} \|_\gamma = \left\| \sum_{j<k} c_j \hat{\xi}_j - \sum_{j<k} c_j \bar{\mathcal{P}} \perp \hat{\xi}_j - \bar{\mathcal{P}} \perp \eta \right\|_\gamma \\
\leq \sum_{j=1}^{k-1} |c_j| \| \bar{\mathcal{P}} \perp \hat{\xi}_j - \hat{\xi}_j \|_\gamma + \| \bar{\mathcal{P}} \perp \eta \|_\gamma \\
= \sum_{j=1}^{k-1} |c_j| o_p (\sqrt{\iota_n}) + o_p (\sqrt{\iota_n}). \tag{2.23}
\]

When \( \hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1} \) are orthogonal, \( c_j = \langle \zeta, \hat{\xi}_j \rangle \gamma \) and hence \( |c_j| \leq 1 \) for each \( j < k \). In this case,
\[
\sum_{j=1}^{k-1} |c_j| o_p (\sqrt{\iota_n}) + o_p (\sqrt{\iota_n}) = o_p (\sqrt{\iota_n}). \tag{2.24}
\]
Then this means we can find a subset for all $h$ is fixed. Then $1 + \sum_{i} b_{ji} \xi_{i} = \sum_{i} (\hat{P}^\perp \zeta, \xi_{i}) \gamma \xi_{i}$, so that $\sum_{j<i} c_{ji} b_{ji} = (\theta, \hat{\xi}_{i})_{\gamma}$ for all $i \geq 1$. Let $c = (c_{1}, c_{2}, \ldots, c_{k-1})^T$, $B$ be a $(k-1) \times (k-1)$ matrix such that $B(i, j) = b_{ji}$ for $1 \leq i, j \leq k-1$, and $h = ((\theta, \hat{\xi}_{1})_{\gamma}, (\theta, \hat{\xi}_{2})_{\gamma}, \ldots, (\theta, \hat{\xi}_{k-1})_{\gamma})^T$. Then by a matrix form we have $Bc = h$. Because $\sum_{i \neq j} b_{ji} = o_{p}(\epsilon_{n})$ and $b_{jj} = 1 + o_{p}(\epsilon_{n})$, $B(j, i) = b_{ji} = o_{p}(1)$ for $i \neq j$ and and $B(j, j) = 1 + o_{p}(1)$. By Lemma 2.6 (see below), $c = h + o_{p}(1)1$. Therefore, we still have

$$\sum_{j=1}^{k-1} |c_{j}| o_{p}(\sqrt{\epsilon_{n}}) + o_{p}(\sqrt{\epsilon_{n}}) = o_{p}(\sqrt{\epsilon_{n}}).$$

(2.25) Combining (2.23), (2.24) and (2.25), we have

$$\|\hat{P}^\perp \zeta - \hat{P}^\perp \zeta\|_{\gamma} = o_{p}(\sqrt{\epsilon_{n}}).$$

Since this holds for every $\zeta \in W_{2}$ such that $\|\zeta\|_{\gamma} = 1$, we have $\|\hat{P}^\perp - \hat{P}^\perp\|_{\gamma} = o_{p}(\sqrt{\epsilon_{n}})$. Since $I = \hat{P} + \hat{P}^\perp = \hat{P} + \hat{P}^\perp$, we also have $\|\hat{P} - \hat{P}\|_{\gamma} = o_{p}(\sqrt{\epsilon_{n}})$.

The above proof is still valid if we replace $o_{p}(\cdot)$ by $o(1)$. Thus, (b) is also true. □

**Lemma 2.6.** Suppose $B_{n}$ is a $q \times q$ matrix such that $B_{n}^{ij} = o_{p}(1)$ for $i \neq j$ and $B_{n}^{ii} = 1 + o_{p}(1)$. If $c_{n}$ is a sequence of vectors in $\mathbb{R}^{q}$ and $h$ is a vector such that $B_{n}c_{n} = h$, where $h$ is fixed. Then $c_{n} = h + o_{p}(1)1$.

**Proof.** By assumption, for any fixed $\epsilon > 0$, for sufficiently large $n$, we have $\Pr\{|B_{n}^{ij}| > \epsilon\} < \epsilon$ for all $i \neq j$ and $\Pr\{|B_{n}^{ii}| > 1 + \epsilon\} < \epsilon$ for all $i = 1, 2, \ldots, q$. Suppose $\Omega$ is the sample space. Then this means we can find a subset $\Omega_{1}$ of $\Omega$ such that $\Pr(\Omega_{1}) \geq 1 - q^{2}\epsilon$ and for $\omega \in \Omega_{1}$ and sufficiently large $n$, $|B_{n}^{ij}(\omega)| \leq \epsilon$ for all $i \neq j$ and $1 - \epsilon \leq |B_{n}^{ii}(\omega)| \leq 1 + \epsilon$ for all $i = 1, 2, \ldots, q$. When $\epsilon$ is taken to be sufficiently small, $B_{n}$ is invertible for sufficiently large $n$, since the determinant of $B_{n}$ is bounded away from 0. In this case, since $B_{n}(\omega) \rightarrow I$, we have $B_{n}^{-1}(\omega) \rightarrow I$, since the set of invertible $q \times q$ matrices is a Banach algebra, where the inverse operator $B \rightarrow B^{-1}$ is continuous. Thus, $c_{n}(\omega) = B_{n}^{-1}h \rightarrow h$ for all $\omega \in \Omega_{1}$, or in other words, $c_{n} = h + o_{p}(1)1$. □
2.8.4 Proofs of Theorem 2.2 and 2.3

The essence of the proofs for Theorem 2.2 and 2.3 is contained in the following lemma.

Lemma 2.7. Suppose \( \rho_k = o(\iota_n) \) where \( \iota_n \to 0 \). For each \( k = 1, 2, \ldots, m \), we have

\[
\hat{\lambda}_k \overset{a.s.}{\to} \hat{\lambda}_k, \quad \hat{\lambda}_k - \tilde{\lambda}_k = o_p(\iota_n);
\]

and

\[
\langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma^2 \overset{a.s.}{\to} 1, \quad \langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma^2 = 1 + o_p(\iota_n) .
\]

Proof. We adopt the same strategy in the proof by Silverman (1996) to use mathematical induction on the statement \( \mathcal{H}_k \) of two convergences (2.26) and (2.27). In what follows we will prove \( \mathcal{H}_k \) holds by assuming \( \mathcal{H}_j \) holds for each \( j < k \).

On the one hand,

\[
\hat{\lambda}_k = \langle \hat{P}\tilde{\xi}_k, \hat{C}\tilde{P}\tilde{\xi}_k \rangle \geq \langle \hat{P}\hat{\xi}_k, \hat{C}\hat{P}\hat{\xi}_k \rangle - \rho_k S(\hat{\xi}_k) + \rho_k S(\tilde{\xi}_k)
\]

\[
= \langle \hat{P}\hat{\xi}_k, \hat{C}\hat{P}\hat{\xi}_k \rangle + o_p(\iota_n)
\]

\[
= \hat{\lambda}_k + o_p(\iota_n),
\]

where the (2.29) is due to Lemma 2.8 (see below). On the other hand,

\[
\hat{\lambda}_k = \langle \hat{P}\tilde{\xi}_k, \hat{C}\tilde{P}\tilde{\xi}_k \rangle - \rho_k S(\hat{\xi}_k) + \rho_k S(\tilde{\xi}_k)
\]

\[
\geq \langle \hat{P}\hat{\xi}_k, \hat{C}\hat{P}\hat{\xi}_k \rangle - \rho_k S(\hat{\xi}_k) + \rho_k S(\tilde{\xi}_k)
\]

\[
= \langle \hat{P}\tilde{\xi}_k, \hat{C}\tilde{P}\tilde{\xi}_k \rangle + o_p(\iota_n)
\]

\[
= \hat{\lambda}_k + o_p(\iota_n),
\]

where the inequality (2.31) comes from the fact that \( \hat{\xi}_k \) maximizes \( \langle \hat{P}\xi, \hat{C}\hat{P}\xi \rangle - \rho_k S(\xi) \) among all \( \xi \in W_2^2 \) such that \( \|\xi\|_\gamma = 1 \) and the equality (2.32) is due to Lemma 2.9 (see below) and \( \rho_k = o(\iota_n) \). From (2.30) and (2.33) we conclude that \( |\hat{\lambda}_k - \tilde{\lambda}_k| = o_p(\iota_n) \). If in the above we replace all \( o_p(\iota_n) \) with \( o(1) \), the above proof still holds. Thus, \( \hat{\lambda}_k \overset{a.s.}{\to} \hat{\lambda}_k \).
We now proceed to prove (2.27). First note that

\[
\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \sum_{j \geq k} (\hat{\xi}_k, \hat{\xi}_j)_\gamma \hat{\xi}_j, \hat{\mathcal{C}} \sum_{j \geq k} (\hat{\xi}_k, \hat{\xi}_j)_\gamma \hat{\xi}_j \rangle = \sum_{i,j \geq k} \langle \hat{\xi}_k, \hat{\xi}_i \rangle \gamma \langle \hat{\xi}_k, \hat{\xi}_j \rangle \gamma \langle \hat{\xi}_i, \hat{\xi}_j \rangle = \sum_{j \geq k} \tilde{\lambda}_j (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2,
\]

where the last equality is again due to (2.17). Also, since \( \hat{\mathcal{P}} \) is a projection, \( \| \hat{\mathcal{P}} \hat{\xi}_k \|_\gamma^2 \leq \| \hat{\xi}_k \|_\gamma = 1 \). Then

\[
\tilde{\lambda}_k - \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle \geq \tilde{\lambda}_k \| \hat{\mathcal{P}} \hat{\xi}_k \|_\gamma^2 - \sum_{j \geq k} \tilde{\lambda}_j (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 = \sum_{j \geq k} (\tilde{\lambda}_k - \tilde{\lambda}_j) (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 \geq 0 \tag{2.34}
\]

as \( \tilde{\lambda}_k > \tilde{\lambda}_j \) for all \( j > k \). On the other hand, \( \hat{\lambda}_k = \tilde{\lambda}_k + o(\tau_n) \) and by Lemma 2.8,

\[
\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle + o_p(\tau_n) = \hat{\lambda}_k + o_p(\tau_n) = \tilde{\lambda}_k + o_p(\tau_n). \tag{2.35}
\]

With (2.34), this implies

\[
\sum_{j \geq k} (\hat{\lambda}_k - \hat{\lambda}_j) (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 = o_p(\tau_n).
\]

Thus,

\[
\sum_{j \geq k} (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 = \frac{1}{\hat{\lambda}_k - \hat{\lambda}_{k+1}} \sum_{j \geq k} (\hat{\lambda}_k - \hat{\lambda}_{k+1}) (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 \\
\leq \frac{1}{\hat{\lambda}_k - \hat{\lambda}_{k+1}} \sum_{j \geq k} (\hat{\lambda}_k - \hat{\lambda}_j) (\hat{\xi}_k, \hat{\xi}_j)_\gamma^2 \\
= o_p(\tau_n) \tag{2.36}
\]
since $\tilde{\lambda}_k - \tilde{\lambda}_j > 0$ almost surely for all $j > k$. Also, we have
\[
\sqrt{\sum_{j<k} (\hat{\xi}_k, \tilde{\xi}_j)^2} = \sqrt{\sum_{j<k} (\hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_j)^2} \\
\leq \sum_{j<k} |\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_j \rangle_\gamma| \\
= \sum_{j<k} \left( |\langle \hat{P} \hat{\xi}_k, \hat{P} \hat{\xi}_j \rangle_\gamma| + o_p(\sqrt{\tau_n}) \right) \quad (\hat{P} \hat{\xi}_k = 0) \\
= o_p(\sqrt{\tau_n}).
\]

Thus,
\[
\sum_{j<k} (\hat{\xi}_k, \tilde{\xi}_j)^2 \leq o_p(\tau_n).
\]

This result, with (2.36) together, gives the desired conclusion (2.27):
\[
(\hat{\xi}_k, \hat{\xi}_k)_\gamma = 1 - \sum_{j<k} (\hat{\xi}_k, \tilde{\xi}_j)_\gamma^2 - \sum_{j>k} (\hat{\xi}_k, \xi_j)_\gamma^2 = 1 + o_p(\tau_n).
\]

The above proof is still valid if we replace $o_p(\cdot)$ by $o(1)$. Thus, we complete the proof. \qed

**Lemma 2.8.** Suppose $\mathcal{H}_j$ holds for each $j < k$. Then $\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle + o_p(\tau_n)$ and $\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle + o(1)$.

**Proof.** Clearly, $\hat{\mathcal{P}} \hat{\xi}_k = \hat{\xi}_k - \hat{\mathcal{P}} \hat{\xi}_k$. Thus
\[
\langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\xi}_k, \hat{\mathcal{C}} \hat{\xi}_k \rangle - \langle \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle - \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\xi}_k \rangle + \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle. \quad (2.37)
\]

First, note that
\[
\langle \hat{\xi}_k, \hat{\mathcal{C}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{C}} \hat{\mathcal{P}} \hat{\xi}_k \rangle \quad (2.38)
\]
as $\hat{\xi}_k = \hat{\mathcal{P}} \hat{\xi}_k$. Next, expanding $\hat{\xi}_k$ by the basis $\{\hat{\xi}_j : j \in \mathbb{N}\}$ we get $\hat{\xi}_k = \sum_{j=1}^{\infty} \langle \hat{\xi}_k, \hat{\xi}_j \rangle_\gamma \hat{\xi}_j$. Thus, $\hat{\mathcal{P}} \hat{\xi}_k = \sum_{j<k} \langle \hat{\xi}_k, \hat{\xi}_j \rangle_\gamma \hat{\xi}_j$, and
Thus, combining (2.37), (2.38), (2.46), (2.47) and (2.48) together, we prove that \( \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_k \rangle + o_p(t_n). \) Replacing \( o_p(\cdot) \) by \( o(1) \) we prove \( \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_k \rangle = \langle \hat{\mathcal{P}} \hat{\xi}_k, \hat{\mathcal{P}} \hat{\xi}_k \rangle + o(1). \) \( \square \)
Lemma 2.9. Suppose \( \mathcal{H}_j \) holds for all \( j < k \). Then \( \langle \hat{\mathcal{P}} \xi_k, \hat{C} \hat{\mathcal{P}} \xi_k \rangle = \langle \xi_k, \hat{C} \xi_k \rangle + o_p(\tau_n) \) and \( \langle \hat{\mathcal{P}} \xi_k, \hat{C} \hat{\mathcal{P}} \xi_k \rangle = \langle \xi_k, \hat{C} \xi_k \rangle + o(1) \).

Proof. Clearly \( \hat{\mathcal{P}} \xi_k = \xi_k - \hat{\mathcal{P}} \perp \xi_k = \xi_k - \sum_{j<k} \langle \hat{\xi}_j, \hat{\xi}_k \rangle \hat{\xi}_j \). Thus,

\[
\langle \hat{\mathcal{P}} \xi_k, \hat{C} \hat{\mathcal{P}} \xi_k \rangle = \langle \xi_k, \hat{C} \xi_k \rangle - \langle \xi_k, \hat{C} \hat{\mathcal{P}} \perp \xi_k \rangle - \langle \hat{\mathcal{P}} \perp \xi_k, \hat{C} \xi_k \rangle + \langle \hat{\mathcal{P}} \perp \xi_k, \hat{C} \hat{\mathcal{P}} \perp \xi_k \rangle,
\]

based on which we again use the divide-and-conquer strategy. First of all,

\[
\langle \xi_k, \hat{C} \hat{\mathcal{P}} \perp \xi_k \rangle = \langle \xi_k, \hat{C} \sum_{j<k} \langle \hat{\xi}_j, \hat{\xi}_k \rangle \gamma \hat{\xi}_j \rangle = \sum_{j<k} \langle \hat{\xi}_j, \hat{\xi}_k \rangle \gamma \langle \hat{\xi}_k, \hat{C} \hat{\xi}_j \rangle.
\]

For each \( j < k \), we have

\[
\langle \hat{\xi}_k, \hat{C} \hat{\mathcal{P}} \perp \hat{\xi}_k \rangle = \langle \hat{\xi}_k, \hat{\mathcal{C}} \sum_{i \geq 1} \langle \hat{\xi}_j, \hat{\xi}_i \rangle \gamma \hat{\xi}_i \rangle
= \sum_{i \geq 1} \langle \hat{\xi}_j, \hat{\xi}_i \rangle \gamma \langle \hat{\xi}_i, \hat{\mathcal{C}} \hat{\xi}_i \rangle
= \sum_{i \geq 1} \langle \hat{\xi}_j, \hat{\xi}_i \rangle \gamma \langle \hat{\xi}_i, \hat{\mathcal{C}} \hat{\xi}_i \rangle
= \lambda_k \langle \hat{\xi}_j, \hat{\xi}_k \rangle \gamma,
\]

where the last equality is due to (2.17). Thus,

\[
\langle \hat{\xi}_k, \hat{C} \hat{\mathcal{P}} \perp \hat{\xi}_k \rangle = \lambda_k \sum_{j<k} \langle \hat{\xi}_j, \hat{\xi}_k \rangle^2 \gamma.
\]

By Lemma 2.5, we also have

\[
\langle \hat{\xi}_j, \hat{\xi}_k \rangle \gamma = \langle \hat{\mathcal{P}} \perp \hat{\xi}_j, \hat{\mathcal{P}} \hat{\xi}_k \rangle \gamma = \langle \hat{\mathcal{P}} \perp \hat{\xi}_j, \hat{\mathcal{P}} \hat{\xi}_k \rangle \gamma + o_p(\sqrt{\tau_n}) = o(\sqrt{\tau_n}).
\]

Therefore,

\[
\langle \hat{\xi}_k, \hat{C} \hat{\mathcal{P}} \perp \hat{\xi}_k \rangle = \lambda_k \sum_{j<k} \langle \hat{\xi}_j, \hat{\xi}_k \rangle^2 \gamma = o(\tau_n).
\]

Similarly,

\[
\langle \hat{\mathcal{P}} \perp \hat{\xi}_k, \hat{C} \hat{\xi}_k \rangle = o_p(\tau_n).
\]
Also, we have

\[ |\langle \hat{P}^\perp \xi_k, \hat{C} \hat{P}^\perp \xi_k \rangle| = \sum_{i,j<k} \langle \hat{\xi}_i, \hat{\xi}_k \rangle_\gamma \langle \hat{\xi}_j, \hat{\xi}_k \rangle_\gamma \leq \|\hat{C}\| \sum_{i,j<k} \langle \hat{\xi}_i, \hat{\xi}_k \rangle_\gamma \langle \hat{\xi}_j, \hat{\xi}_k \rangle_\gamma = o_p(t_n). \]

by (2.50) and \( \|\hat{C} - C\| < \infty \) almost surely. Combining (2.49), (2.51), (2.52) and (2.53), we prove the \( \langle \hat{P}\xi_k, \hat{C} \hat{P}\xi_k \rangle = \langle \hat{\xi}_k, \hat{C}\hat{\xi}_k \rangle + o_p(t_n). \) Replacing \( o_p(\cdot) \) by \( o(1) \) we then prove \( \langle \hat{P}\xi_k, \hat{C} \hat{P}\xi_k \rangle = \langle \hat{\xi}_k, \hat{C}\hat{\xi}_k \rangle + o(1). \)

We are now ready to prove Theorem 2.2 and Theorem 2.3.

**Proof of Theorem 2.2.** As \( \sqrt{n} (\hat{\xi}_k - \xi_k) = \sqrt{n} (\hat{\xi}_k - \tilde{\xi}_k) + \sqrt{n} (\tilde{\xi}_k - \xi_k) \), by Lemma 2.7, when \( \rho_k = o(n^{-1}) \) for all \( k = 1, 2, \ldots, m \), we have

\[ \|\hat{\xi}_k - \tilde{\xi}_k\|_\gamma^2 \leq \|\hat{\xi}_k - \hat{\xi}_k\|_\gamma^2 + \|\hat{\xi}_k - \tilde{\xi}_k\|_\gamma^2 - 2 \langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma = 2 - 2 \langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma = \left( \langle \hat{\xi}_k, \hat{\xi}_k \rangle_\gamma - 1 \right)^2 + 1 - \langle \hat{\xi}_k, \tilde{\xi}_k \rangle_\gamma^2 = o_p(n^{-1}) + o_p(n^{-1}) = o_p(n^{-1}) \]

by noting that \( \langle \hat{\xi}_k, \hat{\xi}_k \rangle_\gamma = 1 + o_p(n^{-1/2}) \). In other words, \( \|\hat{\xi}_k - \tilde{\xi}_k\| \leq o_p(n^{-1/2}) \). Thus \( \sqrt{n} (\hat{\xi}_k - \tilde{\xi}_k) \) converges to 0 in probability for each \( k = 1, 2, \ldots, m \). So does the joint vector \( \{\sqrt{n} (\hat{\xi}_k - \tilde{\xi}_k) : k = 1, 2, \ldots, m \} \). By Slutsky’s theorem, \( \{\sqrt{n} (\hat{\xi}_k - \xi_k) : k = 1, 2, \ldots, m \} \) and \( \{\sqrt{n} (\tilde{\xi}_k - \xi_k) : k = 1, 2, \ldots, m \} \) converge to the same random element in distribution. As Qi and Zhao (2011) have shown that the latter converges to a Gaussian random element when \( \gamma = o_p(n^{-1}) \), \( \{\sqrt{n} (\hat{\xi}_k - \xi_k) : k = 1, 2, \ldots, m \} \) converges to the same Gaussian random element as well.

By a similar argument, we can show that when \( \gamma = o_p(n^{-1/2}) \) and \( \rho_k = o_p(n^{-1/2}) \) for \( k = 1, 2, \ldots, m \), the vector \( \{\sqrt{n} (\hat{\lambda}_k - \lambda_k) : k = 1, 2, \ldots, m \} \) converges to a Gaussian distribution with mean 0.

**Proof of Theorem 2.3.** When \( \rho_k = o(1) \), then by Lemma 2.7, \( \|\hat{\xi}_k - \tilde{\xi}_k\| \leq \|\hat{\xi}_k - \tilde{\xi}_k\|_\gamma \overset{a.s.}{\to} 0 \). In Silverman (1996), it has been shown that \( \|\hat{\xi}_k - \xi_k\|_\gamma \overset{a.s.}{\to} 0 \) under the condition that \( \gamma \to 0 \).
as $n \to \infty$. Thus,

$$\|\hat{\xi}_k - \xi_k\| \leq \|\hat{\xi}_k - \hat{\xi}_k\| + \|\hat{\xi}_k - \xi_k\| \leq \|\hat{\xi}_k - \hat{\xi}_k\| \gamma + \|\hat{\xi}_k - \xi_k\| \xrightarrow{a.s.} 0.$$ 

By a similar argument, we can show that $\hat{\lambda}_k \xrightarrow{a.s.} \lambda_k$. 

### 2.8.5 Proof of Theorem 2.4

When using projection deflation, the estimated $\hat{\xi}_k$'s are not orthogonal in $\langle \cdot, \cdot \rangle_\gamma$. Also, note that the residuals of curves $X_i$'s are defined in terms of inner product $\langle \cdot, \cdot \rangle$ (we cannot project $X_i$'s in terms of the modified inner product since $X_i$ might not be in the domain of $D^2$). To reduce the notational burden, we reuse $\hat{\mathcal{P}}$ to denote the projection perpendicular to $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{k-1}$ in the space $L^2(\mathcal{I})$. In this case, $\hat{\mathcal{P}}$ is a self-adjoint operator on $L^2(\mathcal{I})$. Recall that $\xi_1, \xi_2, \ldots$ denote the eigenfunctions of $\mathcal{C}$. Let $\mathcal{P}$ denote the projection perpendicular to $\xi_1, \xi_2, \ldots, \xi_{k-1}$.

We start with establishing three auxiliary results and then prove Theorem 2.4 in the end.

**Lemma 2.10.** If $\langle \hat{\xi}_j, \xi_j \rangle = o(1)$ and $\|\hat{\xi}_j\| = 1 + o(1)$ for all $j < k$, then

$$\|\hat{\mathcal{P}} - \mathcal{P}\| = o(1)$$

and

$$\|\hat{\mathcal{P}} - \mathcal{P}\| = o(1).$$

**Proof.** The proof is almost along the lines of proof of Lemma 2.5, with $\hat{\xi}_j$ replaced by $\xi_j$, $\langle \cdot, \cdot \rangle_\gamma$ replaced by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_\gamma$ replaced by $\|\cdot\|$. 

Recall that we use $H_{k-1}$ to denote the subspace spanned by the first $k-1$ eigenfunctions $\xi_1, \xi_2, \ldots, \xi_{k-1}$ of the covariance operator $\mathcal{C}$. It is easy to see that the operator $\mathcal{C}_k = \mathcal{P}\mathcal{C}\mathcal{P}$ has eigenfunctions $\xi_k, \xi_{k+1}, \ldots$ and eigenvalues $\lambda_k > \lambda_{k+1} > \cdots \geq 0$ in the subspace $H_{k-1}^\perp$. Its null space is exactly $H_{k-1}$. The following lemma shows that if an operator $\mathcal{U}$ is sufficiently close to $\mathcal{C}_k$ and $\langle \xi, \mathcal{U}\xi \rangle$ is close to $\lambda_k$ enough, then projection of the function $\xi$ of unit length on the null space of $\mathcal{C}_k$ must be close to zero.
Lemma 2.11. Let \( \{ \mathcal{U}_n \} \) be a sequence of operators on \( L^2(\mathcal{S}) \) such that \( \sup_{\| \xi \| = 1} |\langle \xi, \mathcal{U}_n \xi \rangle - \langle \xi, \mathcal{C}_k \xi \rangle| = o(1) \). If \( \xi_n^* \) is a function such that \( \langle \xi_n^*, \mathcal{U}_n \xi_n^* \rangle = \lambda_k + o(1) \) and \( \| \xi^* \| = \| \xi^* \| + o(1) \), then \( \| \mathcal{P}^\perp \xi_n^* \| = o(1) \).

Proof. First of all, we have \( \xi_n^* = \mathcal{P}^\perp \xi_n^* + \mathcal{P} \xi_n^* \). Then by Cauchy-Schwarz inequality,

\[
\langle \mathcal{P}^\perp \xi_n^*, \mathcal{U}_n \mathcal{P}^\perp \xi_n^* \rangle = \langle \mathcal{P}^\perp \xi_n^*, \mathcal{U}_n \mathcal{P}^\perp \xi_n^* \rangle - \langle \mathcal{P}^\perp \xi_n^*, \mathcal{C}_k \mathcal{P}^\perp \xi_n^* \rangle = \langle \mathcal{P}^\perp \xi_n^*, (\mathcal{U}_n - \mathcal{C}_k) \mathcal{P}^\perp \xi_n^* \rangle = o(1)
\]
as \( \sup_{\| \xi \| = 1} |\langle \xi, \mathcal{U}_n \xi \rangle - \langle \xi, \mathcal{C}_k \xi \rangle| = o(1) \). Similarly, \( \langle \mathcal{P}^\perp \xi_n^*, \mathcal{U}_n \mathcal{P} \xi_n^* \rangle = o(1) \), \( \langle \mathcal{P} \xi_n^*, \mathcal{U}_n \mathcal{P}^\perp \xi_n^* \rangle = o(1) \), and \( \langle \mathcal{P} \xi_n^*, \mathcal{U}_n \mathcal{P} \xi_n^* \rangle = \langle \mathcal{P} \xi_n^*, \mathcal{C}_k \mathcal{P} \xi_n^* \rangle + o(1) \). Thus,

\[
\langle \xi_n^*, \mathcal{U}_n \xi_n^* \rangle = \langle \mathcal{P} \xi_n^*, \mathcal{U}_n \mathcal{P} \xi_n^* \rangle + \langle \mathcal{P} \xi_n^*, \mathcal{U}_n \mathcal{P}^\perp \xi_n^* \rangle + \langle \mathcal{P}^\perp \xi_n^*, \mathcal{U}_n \mathcal{P} \xi_n^* \rangle + \langle \mathcal{P}^\perp \xi_n^*, \mathcal{U}_n \mathcal{P}^\perp \xi_n^* \rangle
\]

\[
= \langle \mathcal{P} \xi_n^*, \mathcal{C}_k \mathcal{P} \xi_n^* \rangle + o(1).
\]

On the other hand, by assumption, \( \langle \xi_n^*, \mathcal{U}_n \xi_n^* \rangle = \lambda_k + o(1) \). Combining this with (2.54) gives \( \langle \mathcal{P} \xi_n^*, \mathcal{C}_k \mathcal{P} \xi_n^* \rangle = \lambda_k + o(1) \). Since

\[
\left( \frac{\mathcal{P} \xi_n^*}{\| \mathcal{P} \xi_n^* \|}, \frac{\mathcal{C}_k \mathcal{P} \xi_n^*}{\| \mathcal{P} \xi_n^* \|} \right) \leq \lambda_k,
\]
we have \( \lambda_k + o(1) = \langle \mathcal{P} \xi_n^*, \mathcal{C}_k \mathcal{P} \xi_n^* \rangle \leq \lambda_k \| \mathcal{P} \xi_n^* \|^2 \). Since \( \| \mathcal{P} \xi_n^* \|^2 \leq \| \xi_n^* \|^2 \leq 1 \), we must have \( \| \mathcal{P} \xi_n^* \| = 1 + o(1) \), which implies that \( \| \mathcal{P}^\perp \xi_n^* \| = o(1) \) since \( \| \xi_n^* \| = 1 + o(1) \).

\[ \square \]

Lemma 2.12. If \( \mathbb{E}(\| X \|^2) < \infty \) and \( \langle \hat{\xi}_j, \xi_j \rangle^2 = o(1) \) for all \( j < k \), then

\[
\sup_{\| \xi \| = 1} |\langle \xi, \mathcal{C}_k \xi \rangle - \langle \xi, \mathcal{C}_k \xi \rangle| = o(1).
\]
Proof. For each $\xi \in W^2_2$,

$$\langle \xi, \hat{C}_k \xi \rangle = \int_{\mathcal{D}} \xi(s) \left( \int_{\mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i^k(s) \hat{X}_i^k(t) \xi(t) \, dt \right) \, ds$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left( \int_{\mathcal{D}} \xi(s) \hat{P}_i(s) \, ds \right) \left( \int_{\mathcal{D}} \xi(s) \hat{P}_i(s) \, dt \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle \xi, \hat{P}_i \rangle^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle \hat{P} \xi, X_i \rangle^2.$$ 

On the other hand,

$$\langle \xi, \hat{C}_k \xi \rangle - \langle \xi, \hat{P} \xi \rangle = \int_{\mathcal{D}} \xi(s) \left( \int_{\mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i^k(s) X_i(t) \xi(t) \, dt \right) \, ds$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle \xi, \hat{P}_i \rangle \langle \xi, X_i \rangle$$

By Lemma 2.10,

$$\sup_{\|\xi\|_\gamma = 1} |\langle \hat{P} \xi, X_i \rangle - \langle \xi, X_i \rangle| \leq \sup_{\|\xi\|_\gamma = 1} \|X_i\| (\|P - \hat{P}\| \|\xi\|) = o(1) \|X_i\|.$$ 

Thus,

$$\sup_{\|\xi\|_\gamma = 1} |\langle \hat{P} \xi, X_i \rangle^2 - \langle \xi, X_i \rangle^2| = o(1) \|X_i\|^2.$$ 

Therefore,

$$\langle \xi, \hat{C}_k \xi \rangle - \langle \xi, \hat{P} \xi \rangle = o(1) \frac{1}{n} \sum_{i=1}^{n} \|X_i\|^2 = o(1),$$

since by the strong law of large number, we have $\frac{1}{n} \sum \|X_i\|^2 \to E\|X\|^2 < \infty$. Also, since $\|C - \hat{C}\| = o(1)$ by Dauxois et al. (1982), we have

$$\langle \xi, \hat{C}_k \xi \rangle - \langle \xi, CP \xi \rangle = \langle \xi, \hat{C}_k \xi \rangle - \langle \xi, \hat{C}_k \xi \rangle = o(1)$$

for all $\xi \in W^2_2$ such that $\|\xi\|_\gamma = 1$. 

Proof of Theorem 2.4. The paradigm of our proof is almost identical to the one in Silverman
(1996). Again, we use mathematical induction on the statement $\mathcal{H}_k^j$ of statements:

\[
\hat{\lambda}_k \overset{a.s.}{\to} \lambda_k, \quad (2.55)
\]

\[
\|\hat{\xi}_k\| \to 1; \quad (2.56)
\]

and

\[
\langle \hat{\xi}_k, \xi_k \rangle^2 \overset{a.s.}{\to} 1. \quad (2.57)
\]

Assume $\mathcal{H}_j^j$ holds for $j = 1, 2, \ldots, k - 1$ and we are going to prove $\mathcal{H}_k^j$.

We first establish (2.55). On the one hand,

\[
\lambda_k = \langle \xi_k, C_k \xi_k \rangle \geq \langle \xi_k^*, C_k \xi_k^* \rangle \geq \langle \hat{\xi}_k, C_k \hat{\xi}_k \rangle \quad (2.58)
\]

\[
= \langle \hat{\xi}_k, \hat{C}_k \hat{\xi}_k \rangle + o(1) = \hat{\lambda}_k + o(1) \quad (2.59)
\]

because of Lemma and 2.12. On the other hand,

\[
\hat{\lambda}_k = \langle \hat{\xi}_k, \hat{C}_k \hat{\xi}_k \rangle - \rho_k S(\hat{\xi}_k) + \rho_k S(\hat{\xi}_k)
\]

\[
\geq \left( \frac{\xi_k}{\|\xi_k\|_\gamma}, \frac{\hat{\xi}_k \hat{\xi}_k}{\|\xi_k\|_\gamma} \right) - \rho_k S(\xi_k) + \rho_k S(\hat{\xi}_k) \quad (2.60)
\]

\[
= \left( \frac{\xi_k}{\|\xi_k\|_\gamma}, \frac{C_k \xi_k}{\|\xi_k\|_\gamma} \right) + o(1) - \rho_k S(\xi_k) + \rho_k S(\hat{\xi}_k) \quad (2.61)
\]

\[
= \frac{\lambda_k}{\|\xi_k\|_\gamma^2} + o(1) = \lambda_k + o(1), \quad (2.62)
\]

where the inequality (2.60) comes from the fact that $\hat{\xi}_k$ maximizes $\langle \xi, \hat{C}_k \xi \rangle - \rho_k S(\xi)$ among all $\xi \in W_2^\gamma$ such that $\|\xi\|_\gamma = 1$ and the equality (2.61) is due to Lemma 2.12. From (2.59) and (2.62) we conclude that $\hat{\lambda}_k - \lambda_k = o(1)$.

Second, we prove (2.56). The proof is actually identical to the one by Silverman (1996). We repeat it here for the sake of completeness. As all inequalities in (2.58) tend to equality, it follows

\[
\|\hat{\xi}_k\|^2 = \frac{\langle \hat{\xi}_k, C_k \hat{\xi}_k \rangle}{\langle \hat{\xi}_k, C_k \hat{\xi}_k \rangle} \overset{a.s.}{\to} 1,
\]

so that $\|\hat{\xi}_k\| \overset{a.s.}{\to} 1$. 


Finally, we proceed to prove (2.57). First note that

$$\langle \xi_k^*, C_k \xi_k^* \rangle = \sum_{j \geq k} \lambda_j \langle \xi_k^*, \xi_j \rangle^2.$$  

Also, note that $\|P \xi_k^*\| \leq \|\xi_k^*\| \leq 1$. Then

$$\lambda_k - \langle \xi_k^*, C_k \xi_k^* \rangle \geq \lambda_k \|P \xi_k^*\|^2 - \sum_{j \geq k} \lambda_j \langle \xi_k^*, \xi_j \rangle^2 = \sum_{j > k} (\lambda_k - \lambda_j) \langle \xi_k^*, \xi_j \rangle^2 \geq 0 \quad (2.63)$$

as $\lambda_k > \lambda_j$ for all $j > k$. On the other hand, by (2.58), (2.59) and (2.55),

$$\lambda_k - \langle \xi_k^*, C_k \xi_k^* \rangle = o(1). \quad (2.64)$$

Thus,

$$\sum_{j > k} \langle \xi_k^*, \xi_j \rangle^2 = o(1) \quad (2.65)$$

since $\lambda_k - \lambda_j > 0$ for all $j > k$. By Lemma 2.11 we obtain $\|P_{-\xi_k^*}\| \xrightarrow{a.s.} 0$. This result, with (2.65) together, gives the desired conclusion (2.57):

$$\langle \xi_k^*, \xi_k \rangle^2 = 1 - \|P_{-\xi_k^*}\|^2 - \sum_{j > k} \langle \xi_k^*, \xi_j \rangle^2 \xrightarrow{a.s.} 1.$$  

\[\square\]
Chapter 3

Functional SCAD and SLoS Estimator

3.1 Introduction

In this chapter, we study the problem of estimating the continuous coefficient function $\beta(t)$ in the functional linear regression (FLR) model

$$Y_i = \mu + \int_D X_i(t) \beta(t) \, dt + \varepsilon_i,$$

where each $Y_i$ is a scalar response, $\mu$ is the grand mean, $X_1, X_2, \ldots, X_n$ are independent realizations of an unknown centered stochastic process $X$ defined on the domain $D = [0, T]$, $\varepsilon_i \sim i.i.d. \, N(0, \sigma^2_\varepsilon)$, and $\mathbb{E}(X(t)\varepsilon) = 0$ for $t \in D$ almost everywhere. The coefficient function $\beta(t)$ is also called “regression weight function”, because the cumulative contribution of a functional predictor $X_i$ to its corresponding response $Y_i$ is weighted by $\beta(t)$. It is of practical interest to know where $X$ has no effect on $Y$, and where $X$ contributes to $Y$ and in what weight. Statistically, this corresponds to the problem of identifying the null sub-regions of $\beta(t)$ and meanwhile estimating $\beta(t)$ on its non-null sub-regions, which is the focus of this chapter.

Historically, functional linear regression originated from ordinary linear regression with a large number of predictors. When the number of predictors is very large, estimators produced by ordinary least squares exhibit excessive variability, and hence have poor performance on prediction. Researchers developed many approaches to rectify the issue, such
as partial least squares (PLS), principal components regression (PCR), and ridge regression, etc. When the predictors were actually discrete observations of some continuous process, Hastie and Mallows (1993) pointed out that none of these methods directly made use of the spatial information, i.e. the correlation and the order between predictors. They suggested that it would be more natural to regularize the variability by directly restricting the coefficient vector to be a smooth function. They also conceived the ideas of penalized least squares and smooth basis expansions. These ideas were then explored more thoroughly in Ramsay and Silverman (1997).

For more theoretical and practical treatments of functional linear regression, readers are referred to Ramsay and Silverman (2005) and Ferraty and Vieu (2006). Below we give a brief and non-extensive review of recent works. Cardot et al. (2003) studied some theoretical aspects of FLR, such as identifiability, existence and unicity of the estimator of $\beta(t)$. They also proposed an estimator based on penalized B-spline expansion with a roughness penalty. The method developed by Crambes et al. (2009) was also based on penalized B-spline expansion, but it included an unusual term in the roughness penalty to ensure the existence of their estimator without additional assumptions on sample curves. In both works, convergence rates of estimators were given in terms of a norm induced by the covariance operator of the process $X$. Instead of B-spline basis, Li and Hsing (2007) used a penalized Fourier basis to estimate $\beta(t)$. With the nicer properties of the Fourier basis, they were able to derive an $L^2$-norm convergence rate for their estimator. The above three works all followed the idea of penalized least squares proposed by Hastie and Mallows (1993). In contrast, some researchers utilized the spectral decomposition of the covariance function of $X$. For instance, Cardot et al. (2003) also proposed to obtain an initial estimate through principal components regression and then smooth it by means of B-spline approximation. In Cai and Hall (2006) and Hall and Horowitz (2007), $\beta(t)$ was estimated based on the spectral decomposition as well.

Besides the classic task of estimating $\beta(t)$, researchers also studied some other aspects of FLR. For example, Cardot et al. (2007) investigated the convergence rates of some estimators in the weak topology of $L^2(D)$. Preda (2007) and Yuan and Cai (2010) explored FLR from the point of view of reproducing kernel Hilbert space. Müller and Stadtmüller (2005) studied generalized functional linear models. Yao et al. (2005) extended the scope of FLR to the sparse longitudinal data. Fan and Zhang (2000) and Lian (2012) studied FLR with a functional response. Recently, James et al. (2009) recognized the importance of the
interpretability of estimated $\beta(t)$ and proposed a so-called “FLiRTI” approach that intended to produce a more interpretable estimator. The FLiRTI estimate of $\beta(t)$ is zero on some sub-regions and has a simple structure on the remaining sub-regions. It is also computed via basis expansion approach. The sparsity is indirectly achieved by using a difference operator to transform the basis coefficients and then using Dantzig selector (Candes and Tao, 2007) on the transformed coefficients. Zhou et al. (2013) proposed a method for simultaneously identifying the null sub-regions of $\beta(t)$ and estimating $\beta(t)$ on its non-null sub-regions. The method has two stages: in the first stage, Dantzig selector is used to obtain an initial estimate of the null sub-regions, and in the second stage, the estimated null sub-regions are iteratively refined and meanwhile the estimate of $\beta(t)$ on the non-null sub-regions is produced via a group SCAD penalty proposed by Wang et al. (2007).

In this chapter, we propose a one-stage procedure, called SLoS (Smooth and Locally Sparse), to simultaneously identify the null sub-regions of $\beta(t)$ and produce a smooth estimator of $\beta(t)$ on the non-null sub-regions, where by “locally sparse” we means a curve is zero on some sub-regions of its domain. To estimate the null sub-regions, we develop a new regularization technique, called “functional SCAD” (fSCAD for short), which can be viewed as a functional generalization of the ordinary SCAD proposed in Fan and Li (2001). The fSCAD enables our method to identify null sub-regions without over shrinking the non-zero values. Practically, we compute our estimator via penalized B-spline expansions. The smoothness of the estimator is regularized by a roughness penalty. Comparing to existing methods in the literature, our SLoS method has two distinct features. First, unlike the FLiRTI method, our fSCAD directly regularizes the estimated coefficient function. Second, unlike the two-stage procedure in Zhou et al. (2013), our SLoS method combines the fSCAD and the roughness penalty together in a single optimization criterion to produce an estimate of the null sub-regions and a smooth estimate of $\beta(t)$ on its non-null sub-regions simultaneously in a single stage. These two features make our estimating procedure conceptually more elegant and computationally simpler. Moreover, our SLoS estimator enjoys the oracle property, which means it performs as well as if we knew the true null sub-regions of $\beta(t)$ in advance.

The rest of the chapter is organized as follows. In Section 3.2 we present the fSCAD regularization technique with a discussion of its property, and then apply it to derive the SLoS estimator of $\beta(t)$. The oracle property of our estimator is discussed in Section 3.3. In Section 3.4 we evaluate the numerical performance of our estimator via simulation studies.
In Section 3.5 we apply our method to study the Canadian weather data. The chapter is concluded in Section 3.6. Proofs of theorems are collected in Section 3.7.

3.2 Methodology

3.2.1 Functional SCAD

To motivate our functional SCAD, we first review the SCAD penalty from a slightly different angle. The SCAD was proposed by Fan and Li (2001) in the setting of regression with multiple scalar predictors. Here, we use the ordinary linear regression as an example to illustrate the SCAD technique. Consider a linear model with $\kappa$ variables $b_1, b_2, \ldots, b_\kappa$ and covariates $z_1, z_2, \ldots, z_\kappa$. The model is typically written as $y_i = \mu + \sum_{j=1}^{\kappa} b_j z_{ij} + \varepsilon_i$, where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. For the purpose of introducing our fSCAD, we write it as $y_i = z_i \cdot b + \varepsilon_i$.

Using the least-squares as loss function, the SCAD estimator of $b_1, b_2, \ldots, b_\kappa$ is

$$ (\hat{b}^{\text{scad}}, \hat{\mu}) = \arg \min_{a \in \mathbb{R}^\kappa, \mu \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu - z_i \cdot a)^2 + \sum_{j=1}^{\kappa} p_\lambda(|a_j|) \right\}, \quad (3.2) $$

where the penalty function $p_\lambda(\cdot)$ is the SCAD penalty function of Fan and Li (2001), defined on $[0, +\infty]$ as

$$ p_\lambda(u) = \begin{cases} 
\lambda u & \text{if } 0 \leq u \leq \lambda \\
-\frac{u^2 - 2a\lambda u + \lambda^2}{2(a-1)} & \text{if } \lambda < u < a\lambda \\
\frac{(a+1)\lambda^2}{2} & \text{if } u \geq a\lambda,
\end{cases} $$

where a suggested value for $a$ is 3.7 according to Fan and Li (2001), and $\lambda$ is a tuning parameter varying with sample size. According to Fan and Li (2001), when $\kappa$ is a fixed value, the SCAD penalty enjoys the so-called oracle property: it is able to identify the true sub-model with probability tending to one and produce an asymptotically normal estimate for each non-zero variable. In other words, the estimator $\hat{b}^{\text{scad}}$ performs as well as if we knew the true sub-model, i.e. which true variables $b_j$’s are zero. Even when the number $\kappa$ of variables tends to $\infty$ as sample size $n$ grows, under certain regularity conditions, $\hat{b}^{\text{scad}}$ still enjoys the oracle property, as shown in Fan and Peng (2004).

If we view $z$ as a discrete process over the index set $\{1, 2, \ldots, \kappa\}$, when we replace $\{1, 2, \ldots, \kappa\}$ by a continuum $D = [0, T]$, then the dot product is analogously replaced by
Theorem 3.1. Let \( \alpha \) be a equispaced sequence in the domain \([0, T]\), and \( \alpha_j(t) \) denote the restriction of \( \alpha(t) \) on the subinterval \([t_{j-1}, t_j]\), i.e. \( \alpha_j(t) = \alpha(t) \) for \( t \in [t_{j-1}, t_j] \) and zero elsewhere. If \( \alpha(t) \) is continuous, then for any \( q > 0 \),
\[
\frac{1}{T} \int_D p_\lambda(|\alpha(t)|) \, dt = \lim_{M \to \infty} \frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( M^{\frac{1}{q}} T^{-\frac{1}{q}} \| \alpha_j \|_q \right), \tag{3.3}
\]
where \( \| \alpha_j \|_q \overset{def}{=} \left( \int_D |\alpha_j(t)|^q \, dt \right)^{1/q} \).

The identity (3.3) shows that, the fSCAD of \( \alpha(t) \) is the limit of the average of SCAD penalty on the normalized \( L^q \) norm of \( \alpha(t) \) on each small subinterval. Now we discuss the shrinkage nature of fSCAD according to this identity. First of all, since
\[
\left( \min_{t \in [t_{j-1}, t_j]} |\alpha_j|^q \right) \frac{T}{M} \leq \int_{t_{j-1}}^{t_j} |\alpha_j(t)|^q \, dt = \| \alpha_j \|_q^q \leq \| \alpha_j \|_\infty^q \frac{T}{M},
\]
immediately, we have
\[
\min_{t \in [t_{j-1}, t_j]} |\alpha_j(t)| \leq M^{\frac{1}{2}} T^{-\frac{1}{q}} \| \alpha_j \|_q \leq \| \alpha_j \|_\infty \leq \| \alpha \|_\infty. \tag{3.4}
\]

Let \( H(\alpha_j) \) denote \( M^{\frac{1}{2}} T^{-\frac{1}{q}} \| \alpha_j \|_q \). For any fixed \( t \in D \), if \( \beta(t) \neq 0 \), then by the continuity of \( \beta(t) \) we have \( |\beta(t)| > \epsilon \) on a small neighborhood \( \mathcal{N}_\epsilon(t) \) of \( t \) for some \( \epsilon > 0 \). When

the inner product \( \langle f, g \rangle \overset{def}{=} \int_D f(t)g(t) \, dt \), and we actually end up with the functional linear model \( Y_i = \mu + \langle X_i, \beta \rangle + \varepsilon_i \). For the penalty term \( \sum_{j=1}^{N} p_\lambda(|\alpha_j|) \) in (3.2), if we normalize it by \( \frac{1}{T} \), then it becomes the integral \( \frac{1}{T} \int_D p_\lambda(|\alpha(t)|) \, dt \). We call the functional \( L(\alpha) \overset{def}{=} \frac{1}{T} \int_D p_\lambda(|\alpha(t)|) \, dt \) the functional SCAD (fSCAD for short) penalty. Since the ordinary SCAD estimator has the oracle property, we might expect that in a functional linear model, an estimator \( \hat{\beta}(t) \) that is regularized by fSCAD also has such a nice property, i.e. with probability tending to one, the estimator \( \hat{\beta}(t) \) is identically zero on the sub-regions where \( \beta(t) \) is zero, and meanwhile has a pointwise asymptotic normal distribution on the non-null sub-regions of \( \beta(t) \). Under some mild regularity conditions, this is true when B-spline expansion is employed to estimate \( \beta(t) \) and the fSCAD penalty is properly scaled.

In the above, we presented the fSCAD by an analogy to the ordinary SCAD. However, it can also be naturally introduced in the setting of locally sparse modeling, due to the following remarkable result that provides an insight into the fSCAD.
CHAPTER 3. FUNCTIONAL SCAD AND SLOS ESTIMATOR

52

If \( M \) is sufficiently large, the subinterval \( D_j = [t_{j-1}, t_j] \) containing \( t \) is inside \( \mathcal{N}_\epsilon(t) \). Then for any consistent continuous estimator \( \hat{\beta}(t) \) of \( \beta(t) \), with probability tending to one, we have \(|\hat{\beta}_j(s)| \geq \epsilon/2\) for all \( s \in \mathcal{N}_\epsilon(t) \) and hence \( H(\hat{\beta}_j) \geq \epsilon/2\). As \( \lambda_n \to 0 \), \( H(\hat{\beta}_j) > a\lambda_n \) with probability tending to one, which indicates that the consistent estimator \( \hat{\beta}(t) \) does not get penalized for its values at \( t \). On the other hand, if \( \beta(t) = 0 \) on a small interval \( \mathcal{N}_0 \), then \(|\hat{\beta}(t)| \to 0\) on \( \mathcal{N}_0 \) with probability tending to one. When \( M \) is sufficiently large, one or more subintervals \( D_j = [t_{j-1}, t_j] \) are inside \( \mathcal{N}_0 \). Then \( H(\hat{\beta}_j) \to 0 \) in probability. By choosing appropriate \( \lambda_n \), the penalty \( p_{\lambda_n}(H(\hat{\beta}_j)) \) has a growing rate \( \lambda_n \) on \( H(\hat{\beta}_j) \) and hence force \( \beta(t) \) to become identically zero on \( D_j \). In the light of the identity (3.3), the above argument shows that, intuitively, the fSCAD shrinks the values of an estimator \( \hat{\beta}(t) \) towards zero on the null sub-regions of \( \beta(t) \), but does not over shrink the estimate on the non-null sub-regions. This shrinkage feature of fSCAD brings it numerous potential applications in locally sparse modeling. For example, in Section 3.2.3, we use it to derive our SLoS estimator of the coefficient function in functional linear regression.

Theorem 3.1 also provides a practical way to approximately compute the fSCAD penalty. For example, we can choose a large \( M \) and then approximately compute \( \frac{1}{T} \int_D p_{\lambda_n}(||\alpha(t)||) \) by \( \frac{1}{M} \sum_{j=1}^M p_{\lambda_n}(M^{d+1}T^{-\frac{1}{2}} ||\alpha_j||_q) \). When B-spline expansion (introduced in the next section) is used to approximate \( \alpha \), we might take \( q = 2 \) because it is relatively easy to compute the \( L^2 \) norm of a B-spline function. Moreover, for B-spline basis, each \( ||\alpha_j||_2 \) only involves a few basis coefficients due to the compact support property of B-spline basis functions. In this particular setting, the fSCAD can be viewed as a generalized SCAD with a diverging number of parameters.

3.2.2 Penalized B-spline Expansion

To prepare for the introduction of our SLoS estimator, in this section we briefly review the penalized B-spline expansion approach for estimating \( \beta(t) \). Recall that a B-spline basis is defined by an order and a sequence of knots. Suppose we use an order \( (d+1) \) and place \( M + 1 \) equispaced knots \( 0 = t_0 < t_1 < \cdots < t_M = T \) in the domain \([0, T]\) to define a B-spline basis. As we have introduced in 2.3.2, over each subinterval, each B-spline basis function \( B_j(t) \) is a polynomial of the degree at most \( d \). Moreover, each B-spline basis function is nonzero over no more than \( d + 1 \) consecutive subintervals. When \( M \) is large, each B-spline basis function is only nonzero on a very short sub-region. This is the compact support property of B-spline basis functions, which is very important for us to practically compute
our SLoS estimator. In Section 2.3.2 we showed an example of B-spline basis functions. Readers are referred to the section for more information.

Let \( S_{dM} \) denote the function space spanned by the B-spline basis functions \( \{ B_j(t) : j = 1, 2, \ldots, M + d \} \) we defined above. A classical method of estimating coefficient function \( \beta(t) \) in (3.1) based on B-spline expansion is to choose \( \alpha \in S_{dM} \) such that \( \alpha \) minimizes the sum of squared errors. However, this method usually produces an estimator that exhibits excessive variability, particularly when \( M \) is relatively large. A popular approach to obtain a smooth estimator is regularization through roughness penalty. For example, a smooth estimator of \( \beta(t) \), proposed in Cardot et al. (2003), is defined as

\[
\hat{\beta}_{\text{smooth}} = \arg \min_{\alpha \in S_{dM}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - \int_D X_i(t) \alpha(t) \, dt \right]^2 + \gamma \| D^m \alpha \|^2 \right\},
\]

where \( \mu \) is assumed to be zero, \( D^m \) is the \( m \)th order derivative operator with \( m \leq d \), and the parameter \( \gamma \geq 0 \) is a tuning parameter varying with the sample size \( n \). Note that by setting \( \gamma = 0 \), the above smooth estimator is reduced to the classical least squares estimator, while as \( \gamma \rightarrow \infty \), the estimator \( \hat{\beta}_{\text{smooth}} \) tends to become a constant function. Thus, \( \gamma \) serves as a smoothing parameter which controls the smoothness of \( \hat{\beta}_{\text{smooth}} \).

Although regularization via roughness penalty is able to produce a smooth estimator, in general, it does not yield a locally sparse estimator. In other words, it is not able to identify the null sub-regions of \( \beta \) due to the presence of random errors \( \varepsilon \) in the model.

### 3.2.3 The SLoS Estimator

We now combine fSCAD and penalized B-spline expansions together to define our SLoS estimator which is simultaneously smooth and locally sparse:

\[
(\hat{\beta}, \hat{\mu}) = \arg \min_{\substack{\alpha \in S_{dM} \\ \mu \in \mathbb{R}}} Q(\alpha), \tag{3.5}
\]

where the penalized least squares \( Q(\alpha) \) is defined as

\[
Q(\alpha) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - \mu - \int_D X_i(t) \alpha(t) \, dt \right]^2 + \gamma \| D^m \alpha \|^2 + \frac{M}{T} \int_D p_\lambda(\| \alpha(t) \|) \, dt. \tag{3.6}
\]
It is worthy of note that in the above penalized least squares, we combine two penalty terms in a single optimization criterion: the roughness penalty \( \gamma \| \mathcal{D}^m \alpha \| \) regularizes the smoothness of the estimator, and the fSCAD penalty \( \frac{M}{T} \int_D p_\lambda(|\alpha(t)|) \, dt \) directly regularizes the local sparseness. We also scale up the fSCAD penalty by the factor \( M \) to make \( \hat{\beta}(t) \) enjoy the oracle property.

Below we show how to compute the SLoS estimator \( \hat{\beta}(t) \) in practice. To simplify the notation and convey the gist of the computation, we assume \( \mu = 0 \). The case that \( \mu \neq 0 \) is considered at the end of the section. We use \( B(t) \) to denote the vector \( (B_1(t), B_2(t), \ldots, B_{M+d}(t))^T \).

Assume \( \alpha(t) = B^T(t) b \) where \( b = (b_1, b_2, \ldots, b_{M+d})^T \) is the coefficient vector of \( \alpha(t) \) with respect to the chosen B-spline basis \( \{B_1(t), B_2(t), \ldots, B_{M+d}(t)\} \). Let \( U \) be an \( n \times (M + d) \) matrix with entries \( u_{ij} = \int_D X_i(t) B_j(t) \, dt \) for \( i = 1, \ldots, n, j = 1, \ldots, (M + d) \). Let \( y = (Y_1, Y_2, \ldots, Y_n)^T \). Then the first term of \( Q(\alpha) \) in (3.6) is written as

\[
\frac{1}{n} \sum_{i=1}^n \left[ Y_i - \int_D X_i(t) \alpha(t) \, dt \right]^2 = \frac{1}{n} (y - Ub)^T (y -Ub).
\] (3.7)

Let \( V \) be an \( (M + d) \times (M + d) \) matrix with entries \( v_{ij} = \langle \mathcal{D}^m B_i, \mathcal{D}^m B_j \rangle \) for \( 1 \leq i, j \leq M + d \). Then the second term of \( Q(\alpha) \) in (3.6) is

\[
\gamma \| \mathcal{D}^m \alpha \|^2 = \gamma b^T V b.
\] (3.8)

According to Theorem 3.1, we approximate the penalty term \( \frac{M}{T} \int_D p_\lambda(|\alpha(t)|) \, dt \) by

\[
\frac{M}{T} \int_D p_\lambda(|\alpha(t)|) \, dt \approx \sum_{j=1}^M p_\lambda \left( \frac{\| \alpha_j \|_2}{\sqrt{T/M}} \right).
\]

It is not hard to show that

\[
\| \alpha_j \|_2 = \left( \int_{t_{j-1}}^{t_j} \alpha^2(t) \, dt \right)^{1/2} = \left( b^T W_j b \right)^{1/2}
\]

where \( W_j \) is an \( (M + d) \times (M + d) \) matrix with entries \( w_{uv} = \int_{t_{j-1}}^{t_j} B_u(t) B_v(t) \, dt \) if \( j \leq u, v \leq j + d \) and zero if otherwise. The SCAD penalty function \( p_\lambda(\cdot) \) might be approximated by the local quadratic approximation (LQA) proposed in Fan and Li (2001) as follows.
When \( u \approx u(0) \), the LQA of SCAD function \( p_\lambda(u) \) is

\[
p_\lambda(|u|) \approx p_\lambda(|u(0)|) + \frac{1}{2} \frac{p'_\lambda(|u(0)|)}{|u(0)|} (u^2 - u(0)^2).
\]

Therefore, to use the LQA, we first employ penalized B-spline expansion to obtain an initial estimate \( \alpha(0)(t) \) of \( \beta(t) \). Then for \( \alpha \approx \alpha(0) \), we have

\[
M \sum_{j=1}^{M} p_\lambda \left( \frac{\|\alpha_j\|_2}{\sqrt{T/M}} \right) \approx \frac{1}{2} \sum_{j=1}^{M} \frac{p'_\lambda \left( \frac{\|\alpha_{j}^{(0)}\|_2}{\sqrt{T/M}} \right)}{\|\alpha_{j}^{(0)}\|_2} \frac{\|\alpha_j\|_2}{T/M} + G(\alpha(0)),
\]

where \( G(\alpha(0)) \) is a term that does not contain \( \alpha \) and is defined as

\[
G(\alpha(0)) \overset{\text{def}}{=} \sum_{j=1}^{M} p_\lambda \left( \frac{\|\alpha_{j}^{(0)}\|_2}{\sqrt{T/M}} \right) - \frac{1}{2} \sum_{j=1}^{M} \frac{p'_\lambda \left( \frac{\|\alpha_{j}^{(0)}\|_2}{\sqrt{T/M}} \right)}{\|\alpha_{j}^{(0)}\|_2} \frac{\|\alpha_j\|_2}{\sqrt{T/M}}.
\]

Let

\[
W^{(0)}(0) = \frac{1}{2} \sum_{j=1}^{M} \left( \frac{p'_\lambda(\|\alpha_{j}^{(0)}\|_2 \sqrt{M/T})}{\|\alpha_{j}^{(0)}\|_2} \frac{\sqrt{M/T}}{\sqrt{T/M}} W_j \right).
\]

Then we have

\[
\frac{M}{T} \int_{D} p_\lambda(|\alpha(t)|) \, dt \approx b^T W^{(0)} b + G(\alpha(0)).
\]

Now, we put (3.7), (3.8) and (3.11) together to write \( Q(\alpha) \) as

\[
Q(b) = \frac{1}{n} (y - Ub)^T (y - Ub) + \gamma b^T V b + b^T W^{(0)} b + G(\alpha(0)).
\]

Thus, optimizing \( Q(\alpha) \) with respect to \( \alpha \) is equivalent to optimizing \( Q(b) \) with respect to \( b \) in (3.12). Also note that the term \( G(\alpha(0)) \) has no impact on optimizing \( Q(b) \). Differentiating \( Q(b) \) with respect to \( b \) and setting it to zero, we have the following equation:

\[
U^T Ub + n\gamma V b + nW^{(0)} b = U^T y
\]

with the solution

\[
\hat{b} = \left( U^T U + n\gamma V + nW^{(0)} \right)^{-1} U^T y.
\]
As in Fan and Li (2001), we repeat the above computational steps until \( \hat{b} \) converges. In summary, we have the following algorithm to compute \( \hat{b} \) and obtain the estimator \( \hat{\beta}(t) = B^T(t)\hat{b} \):

**Step 1:** Compute the initial estimate \( \hat{b}^{(0)} = \left(U^T U + n\gamma V\right)^{-1}U^T y \);

**Step 2:** Given \( \hat{b}^{(i)} \), compute \( W^{(i)} \) and \( \hat{b}^{(i+1)} = \left(U^T U + n\gamma V + nW^{(i)}\right)^{-1}U^T y \);

**Step 3:** Repeat Step 2 until the convergence of \( \hat{b} \) is reached.

When \( \mu \neq 0 \), we modify the matrices \( U \), \( V \) and \( W^{(i)} \) as follows to produce an estimate of \( \hat{\mu} \) simultaneously: add one more parameter \( \mu \) to the topmost of the column vector \( b \); add one more column of all ones to the leftmost of \( U \); add one column of all zeros to the leftmost of both \( V \) and \( W^{(i)} \), and then add one more row of all zeros to the topmost of both \( V \) and \( W^{(i)} \). With these changes, the algorithm above can be carried out as usual.

### 3.2.4 Choice of Tuning Parameters

In practice, we choose a large \( M \) so that the fSCAD can be well approximated. Once \( M \) is chosen, we can use a two-dimension cross-validation procedure to choose \( \gamma \) and \( \lambda \) simultaneously. Note that both \( \gamma \) and \( \lambda \) have effect on shrinking variables. Based on this observation, we propose an iterative search procedure as follows: 1) initially, we set \( \lambda = 0 \) and find the best \( \gamma_0 \) among a set of candidate values. Then find the best \( \lambda \) with the \( \gamma \) fixed at current value \( \gamma_0 \); 2) choose a new \( \gamma \) by searching around the old value while keeping \( \lambda \) fixed at its current value; 3) search the best \( \lambda \) around the old value while keep \( \gamma \) fixed at its current value; 4) repeat Step 2) and 3) several times. The selection criteria in the above procedure can be cross-validation, generalized cross-validation, BIC, AIC or RIC. In general, the \( \gamma_0 \) in the first step is larger than those found in the second step since a positive value of \( \lambda \) shrinks some variables towards zero. The decrease of \( \gamma \) will then result in an increase of \( \lambda \) because a smaller value of \( \gamma \) has less shrinkage effect. Thus, in the above iterative procedure, in general we might expect a decreasing sequence of \( \gamma \) and an increasing sequence of \( \lambda \). As the iteration continues, we expect the pair \( (\gamma, \lambda) \) converges to the optimal one. This procedure is more efficient than the two-dimension cross-validation. We also demonstrate its effectiveness in simulation studies.
3.2.5 Extension to Multiple Regressors

The extension of our method to the case of multiple regressors $X^{(1)}, X^{(2)}, \ldots, X^{(K)}$ and coefficient functions $\beta_1(t), \beta_2(t), \ldots, \beta_K(t)$ is quite easy and straightforward, as we shall see below. When there are multiple coefficient functions, for each of them there corresponds an fSCAD penalty. The computational steps described in Section (3.2.3) are modified as follows to account for this situation. Corresponding to each regressor $X^{(k)}$, we compute the matrix $U_k$. Let $U = (U_1, U_2, \ldots, U_K)$ be the column concatenation of $U_1, U_2, \ldots, U_K$, and correspondingly set $V = \text{diag}(V_1, V_2, \ldots, V_K)$. For each $k$, we also have a matrix $W^{(0,k)}$ which corresponds to the matrix $W^{(0)}$ in (3.10). Then the $W^{(0)}$ in (3.12) is replaced by $W^{(0)} = \text{diag}(W^{(0,1)}, W^{(0,2)}, \ldots, W^{(0,K)})$. After these modifications, the iterative algorithm described in Section (3.2.3) can be carried out without any change.

3.3 Oracle Property

Let $N(\beta)$ denote the null region of $\beta(t)$ and $S(\beta)$ denote the non-null region of $\beta(t)$, i.e. $N(\beta) = \{t \in D : \beta(t) = 0\}$ and $S(\beta) = \{t \in D : \beta(t) \neq 0\}$. We show that, under some regularity conditions, our SLoS estimator $\hat{\beta}(t)$ enjoys the oracle property for identifying the null region $N(\beta)$ and at the same time estimating $\beta(t)$ on $S(\beta)$ with an optimal pointwise convergence rate. In other words, our estimator performs as well as if the true null region $N(\beta)$ was known. We state the conditions and the theorems in this section while defer the proofs to Section 3.7.

The assumed regularity conditions are:

(C1) $\|X\|_2$ is almost surely bounded, i.e. $\|X\|_2 < \infty$ a.s.

(C2) $|\beta(p')(u_1) - \beta(p')(u_2)| \leq c|u_1 - u_2|^\nu$ for some constant $c > 0$, integer $p'$ and $\nu \in [0, 1]$. Let $p \overset{\text{def}}{=} p' + \nu$. Assume $p \leq d$, where $d$ is the degree of the B-spline basis. This condition says $\beta(t)$ is in the Hölder space $C^{p',\nu}(D)$. We assume $p > \frac{1}{2}$.

(C3) Define $a_n \overset{\text{def}}{=} \frac{1}{\sqrt{S(\beta)}} \int_{\beta} |\beta(t)|^2 dt$ and we assume $a_n = O(n^{-\frac{1}{2}} M^{-\frac{1}{2}})$; we also define $h_n \overset{\text{def}}{=} \sqrt{S(\beta)} \int_{\beta} |\beta(t)|^2 dt$ and assume $h_n = o_P(M^{-\frac{1}{2}})$.

In the above, (C1) and (C2) are the same as (H1) and (H3) of Cardot et al. (2003). Assumption (C3) is analogous to regularity conditions (B') and (C') in Fan and Peng (2004)
to ensure the unbiasedness and guarantee that the penalty does not dominate the least squares.

Recall that $\gamma$ and $\lambda$ are tuning parameter varying with $n$. To emphasize their dependency on the sample size $n$, we denote them by $\gamma_n$ and $\lambda_n$, respectively. For the oracle property to hold, we assume the following conditions on choosing values of $M$, $\gamma_n$ and $\lambda_n$:

(C4) $M = o(n)$, $M^{-p} = o(n^{-\frac{1}{2}})$, $\gamma_n n^{\frac{1}{2}} M^{-\frac{1}{2}} = o(1)$, $\gamma_n n M^{2(m-p)} = o(1)$, $\lambda_n = o(1)$ and $\lambda_n \sqrt{n}/\sqrt{M} \to \infty$.

According to Theorem XII.1 of de Boor (1978), there is some $s \defeq \sum_{j=1}^{M+d} b_{sj} B_j \in S_{dM}$ such that $\|s - \beta\|_\infty \leq C_1 M^{-p}$ for some constant $C_1 \geq 0$. Moreover, it is clear that the coefficient $b_{sj}$ can be chosen to be zero if the support of the corresponding basis function $B_j$ is contained in the null region $N(\beta)$. The following theorem assures the existence of a local solution of (3.5) with probability tending to one and gives an upper bound of the convergence rate with respect to $(\hat{b}_s, \hat{\mu})$.

**Theorem 3.2 (Existence of SLoS Estimator).** Under conditions (C1)-(C4), with probability tending to one, there exists a local minimizer $(\hat{\beta}, \hat{\mu})$ of (3.5) such that $\hat{\beta}(t) = B^T(t) \hat{b}$, $\|\hat{b} - b_s\| = O_P(n^{-1/2}M)$ and $|\hat{\mu} - \mu| = O_P(n^{-1/2})$.

According to the above theorem, we have a pointwise root-$n/M$ consistent local minimizer $\hat{\beta}(t)$ and a root-$n$ consistent local minimizer $\hat{\mu}$. This local minimizer $(\hat{\beta}, \hat{\mu})$ is taken as our SLoS estimator for $\beta(t)$ and the intercept $\mu$. The estimator $\hat{\beta}(t)$ has an asymptotic property which corresponds to the oracle property reported in Zhou et al. (2013).

We now prepare some notations for stating the oracle property. First, recall that the support of $\beta(t)$ is defined as the closure of $S(\beta)$. For any $\epsilon > 0$ and a subset of the real line $A$, the $\epsilon$-neighborhood of $A$, defined by $\{t \in D : \inf_{u \in A} |t - u| < \epsilon\}$, is denoted by $A^\epsilon$. Also, when we say a sequence of subsets $A_n$ of the real line converges to a subset $F$, we mean the Lebesgue measure of the symmetric difference of $A_n$ and $F$ converges to zero. Given a $\lambda_n$ and $M$, we divide the domain $D$ into two parts: $D_1 = \{t \in D : |\beta(t)| \geq aC_4(\lambda_n + M^{-p})\}$ for some constant $C_4 > 0$ that is determined in Section 3.7, and $D - D_1$. Let $B_1$ be the sub-vector of $B$ such that each $B_j(t)$ in $B_1$ has a support inside $D_1$. Correspondingly, let $U_1$ be the columns of $U$ associated to $B_1$, where $U$ is defined in Section 3.2.3.

**Theorem 3.3 (Oracle Property).** Under conditions (C1)-(C4), as $n \to \infty$,
(a) For every $t$ not in the support of $\beta(t)$, we have $\hat{\beta}(t) = 0$ with probability tending to one. Moreover, for every $\epsilon > 0$, we have $N(\beta) \subset N^*(\hat{\beta})$ with probability tending to one. Thus, $N(\hat{\beta}) \to N(\beta)$ and $S(\hat{\beta}) \to S(\beta)$ in probability.

(b) For every $t$ such that $\beta(t) \neq 0$, we have

$$\left(\frac{n}{M}\right)^{1/2} \left[\hat{\beta}(t) - \beta(t)\right] \xrightarrow{d} N(0, \sigma^2(t)),$$

where

$$\sigma^2(t) = \lim_{n \to \infty} \frac{n}{M} B_1(t) \left(U_1^T U_1\right)^{-1} B_1(t).$$

(c) For the estimated intercept $\hat{\mu}$, we have

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

In the above theorem, the first statement of (a) says that pointwise $\hat{\beta}(t)$ is identically zero on the null region except those points on the boundaries. The second statement of (a) effectively states that the null region identified by our SLoS estimator is consistent with the true null region $N(\beta)$. This conclusion is much stronger than the pointwise consistency in the first statement. (b) says that the SLoS estimator $\hat{\beta}(t)$ on the non-null region has a pointwise asymptotic normal distribution. Thus, it achieves the optimal pointwise convergence rate, which is $M^{1/2}n^{-1/2}$. Given (b), asymptotic pointwise confidence bands for $\beta(t)$ can be constructed. (c) says that $\hat{\mu}$ is asymptotically normally distributed and thus it also achieves the optimal convergence rate $n^{-1/2}$. The result can be used to construct confidence intervals for $\mu$ and perform hypothesis test on $\mu$.

### 3.4 Simulation Studies

We conducted two simulation studies to investigate the numerical performance of our SLoS estimator. We compared it to the following estimators: the ordinary least squares (OLS for short) estimator which corresponded to $\gamma = 0$ and $\lambda = 0$, the smooth estimator in Cardot et al. (2003) which corresponded to $\lambda = 0$, and the oracle estimator. The oracle estimator was computed by placing equispaced knots only on the non-null region $S(\beta)$ and regularized
by roughness penalty. For the comparison, we reported two quantities

\[ A_0 = \frac{1}{\ell_0} \int_{N(\beta)} (\hat{\beta}(t) - \beta(t))^2 \, dt \]

and

\[ A_1 = \frac{1}{\ell_1} \int_{S(\beta)} (\hat{\beta}(t) - \beta(t))^2 \, dt, \]

where \( \ell_0 \) was the length of the null region and \( \ell_1 \) was the length of the non-null region of \( \beta(t) \). \( A_0 \) and \( A_1 \) measure the integrated mean squared error (MSE) between an estimated coefficient function \( \hat{\beta}(t) \) and the true function \( \beta(t) \) on the null region and the non-null region, respectively. We also reported the pointwise bias, standard deviation and RMSE at 100 points \( 0, 0.1, \ldots, 9.9, 10 \). Moreover, we reported the prediction MSE of each method on independent test datasets. For the intercept, we computed the boxplot of the estimated intercepts by these four methods. We also computed the histogram of \( \sqrt{n}(\hat{\mu} - \mu)/\sigma_\varepsilon \) for our SLoS estimator. All tuning parameters were chosen by a 10-fold cross-validation procedure.

### 3.4.1 Study 1

In this simulation study, we considered the model

\[ Y_i = 1 + \int_0^{10} X_i(t) \beta(t) \, dt + \varepsilon_i, \]

where

\[ \beta(t) = \max\{0, t \sin(\pi(t - 2)/3)\} \]

and \( \varepsilon \sim N(0, 0.25) \) The null-region of \( \beta(t) \) was \( N(\beta) = [0, 2] \cup [5, 8] \), as shown in Figure 3.1. The covariate functions \( X_i(t) \) were generated in the following way: each \( X_i(t) \) was a cubic B-spline function on \([0, 10]\) with 51 equispaced knots and the corresponding basis coefficients were generated uniformly from \([-5, 5]\). We took \( m = 51 \) observations within \([0, 10]\) from each true covariate function as the observed data and reconstructed each \( X_i(t) \) by B-spline approximations. We generated 300 datasets with sample size \( n = 500 \). For each dataset, we also generated a separate test dataset with the same sample size. To estimate the coefficient function \( \beta(t) \), we used the cubic B-spline basis defined by 51 equispaced knots \( 0 = t_0 < t_1 < \cdots < t_{50} = 10 \) on \([0, 10]\).
Figure 3.1: The true coefficient function $\beta(t)$ in Study 2.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$A_0$ ($\times 10^{-3}$)</th>
<th>$A_1$ ($\times 10^{-3}$)</th>
<th>MSE ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>-</td>
<td>2.39 (1.29)</td>
<td>1.18 (0.50)</td>
</tr>
<tr>
<td>OLS</td>
<td>3.99 (0.94)</td>
<td>3.98 (1.13)</td>
<td>3.08 (0.78)</td>
</tr>
<tr>
<td>Smooth</td>
<td>3.63 (1.02)</td>
<td>3.85 (1.21)</td>
<td>2.97 (0.82)</td>
</tr>
<tr>
<td>SLoS</td>
<td>0.72 (0.29)</td>
<td>3.22 (0.93)</td>
<td>1.56 (0.52)</td>
</tr>
</tbody>
</table>

Table 3.1: The second and the third columns report the integrated MSEs $A_0$ and $A_1$ of estimators on the null sub-regions and the non-null sub-regions, respectively. The last column reports the MSEs on the test data in Study 1. Each entry is the Monte Carlo average of 300 runs. The corresponding Monte Carlo standard deviation is included in parentheses.

Table 3.1 clearly shows SLoS as the winner except the oracle procedure in each column. On the null sub-regions, our method performs much better than the smooth and the OLS estimators in terms of the integrated MSE $A_0$. On the non-null sub-regions, in terms of the integrated MSE $A_1$, they all have similar performances, with our estimator slightly better than the smooth and the OLS estimators. Our method also has a much better performance on the test data than the smooth and the OLS estimators in terms of the MSE on the test data. The performance of the oracle procedure on the test data is only slightly better than the SLoS method.
Figure 3.2, 3.3 and 4.4 show the pointwise bias, standard deviation and the RMSE of estimated $\beta(t)$, respectively. On the null sub-regions of $\beta(t)$, our SLoS estimator has a similar performance of the oracle procedure at most parts and is consistently better than other methods. On the non-null sub-regions of $\beta(t)$, the performance of the SLoS estimator is fairly similar to the performances of the smooth and the oracle estimators. We also observe that there are some spikes around the boundaries of null and non-null sub-regions for all methods except the oracle procedure. These spikes are the major reason why our SLoS estimator has slightly larger integrated MSEs than the oracle procedure. It might be surprising that OLS has a relatively smaller pointwise standard deviation and RMSE on the non-null sub-regions. This might be because the number of knots (about 21 in most simulation replicates) is relatively small and the sample size $n = 500$ is relatively large. They together produces an OLS estimator with a smaller pointwise variance and hence RMSE. In Study 2, we use a much smaller sample size, in which case OLS does not have a pointwise performance better than other methods. Also, in terms of MSE on the test data, OLS does not work better than other methods, as shown in Table 3.1. The left panel of Figure 3.5 shows a boxplot of estimated intercepts. We can see that the performance of all methods is fairly close. The right panel of Figure 3.5 shows a histogram of $\Delta \mu = \sqrt{n}(\hat{\mu} - \mu)/\sigma$, where $\hat{\mu}$ is estimated by our method. We can see that it matches the probability density function of the standard Gaussian distribution very well. This confirms the part (c) of Theorem 3.3 that asymptotically the estimated intercept has a normal distribution.
CHAPTER 3. FUNCTIONAL SCAD AND SLO S ESTIMATOR

Figure 3.2: The pointwise bias of estimated $\beta(t)$ in Study 1.

Figure 3.3: The pointwise standard deviation of estimated $\beta(t)$ in Study 1.
Figure 3.4: The pointwise RMSE of estimated $\beta(t)$ in Study 1.

Figure 3.5: The left panel displays the boxplot of estimated intercepts by SLoS, smooth, OLS and oracle estimators in Study 1. The right panel shows a histogram of estimated intercepts by SLoS method in Study 1. The curve over the bins is the probability density function of the standard normal distribution.
3.4.2 Study 2

In this simulation study, we consider the model

\[
Y_i = 2 + \int_0^{10} X_{1i}(t)\beta_1(t) \, dt + \int_0^{10} X_{2i}(t)\beta_2(t) \, dt + \varepsilon_i,
\]

with two coefficient functions

\[
\beta_1(t) = \begin{cases} 
3 \cos \frac{\pi t}{6} & \text{if } t \leq 3 \\
0 & \text{otherwise}
\end{cases}
\quad \beta_2(t) = \begin{cases} 
4 \sin \frac{\pi (t+4)}{6} & \text{if } t \geq 6 \\
0 & \text{otherwise}
\end{cases}
\]

and \( \varepsilon \sim N(0, 1) \). The null sub-regions of \( \beta_1(t) \) and \( \beta_2(t) \) are \( N(\beta_1) = [3, 10] \) and \( N(\beta_2) = [0, 6] \), respectively, as shown in Figure 3.6. The covariate functions \( X_{1i}(t) \) and \( X_{2i}(t) \) were generated by using the same method in Study 1. We generated 300 datasets with sample size \( n = 150 \). For each dataset, we also generated a separate test dataset with the same sample size. The cubic B-spline basis used to compute estimates of \( \beta_1(t) \) and \( \beta_2(t) \) was defined by 41 equispaced knots \( 0 = t_0 < t_1 < \cdots < t_{40} = 10 \) on \( [0, 10] \).

Figure 3.6: The top panel shows the true coefficient function \( \beta_1(t) \) and the bottom panel displays the true coefficient function \( \beta_2(t) \) in Study 2.

Again, Table 3.2 clearly shows that except the oracle procedure, our method is the winner of each column. From the results, we can see that on the null sub-regions, the performance of SLoS estimator is far better than the smooth estimator and the OLS estimator. On the
Table 3.2: The second and fourth columns show the integrated MSEs of estimators on null regions for $\beta_1(t)$ and $\beta_2(t)$, respectively. The third and fifth columns report the integrated MSEs of estimators on non-null regions for $\beta_1(t)$ and $\beta_2(t)$, respectively. The last column reports the MSEs on the test data in Study 2. Each entry is the Monte Carlo average of 300 runs. The corresponding Monte Carlo standard deviation is included in parentheses.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\beta_1(t)$</th>
<th>$\beta_2(t)$</th>
<th>MSE ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>$A_0$ ($\times 10^{-3}$)</td>
<td>$A_1$ ($\times 10^{-3}$)</td>
<td>$A_0$ ($\times 10^{-3}$)</td>
</tr>
<tr>
<td>OLS</td>
<td>7.47 (3.57)</td>
<td>10.17 (5.97)</td>
<td>9.61 (4.36)</td>
</tr>
<tr>
<td>Smooth</td>
<td>5.63 (2.93)</td>
<td>7.32 (4.52)</td>
<td>6.73 (2.96)</td>
</tr>
<tr>
<td>SLoS</td>
<td>0.29 (0.35)</td>
<td>6.31 (4.32)</td>
<td>0.60 (0.35)</td>
</tr>
</tbody>
</table>

non-null sub-regions, it still slightly outperforms the smooth estimator and much better than the OLS estimator. The last column of Table 3.2 is the MSE of each estimator on the test data. It shows that our SLoS estimator performs much better than the smooth and the OLS estimators. In all measures, the oracle estimator is only slightly better than our SLoS estimator. However, we should keep in mind that in practice we generally do not know the location of the true null sub-regions and thus the oracle estimator does not exist.

Figure 3.7, 3.8 and 3.9 show the pointwise bias, standard deviation and RMSE of each method on estimating $\beta_1(t)$ and $\beta_2(t)$, respectively. Again, they show that on the null sub-regions of coefficient functions, our SLoS method has a performance comparable to the oracle procedure and works consistently much better than the smooth and OLS estimators. On the non-null sub-regions, the oracle, the SLoS and the smooth estimators have similar performances. We also observe that unlike the Study 1, the OLS acts consistently worse than other method, which might be due to the small sample size $n = 150$ in this simulation study. In the left panel of Figure 3.10 we show a boxplot of estimated intercepts. From the figure, we can see that the oracle estimator and the SLoS estimator produce almost identical results. The estimated intercept by the smooth estimator seems to have a slightly larger variance, and the OLS procedure produces the estimate with an even larger variance. The right panel of Figure 3.10 shows a histogram of $\Delta \mu = \sqrt{n}(\hat{\mu} - \mu)/\sigma_\varepsilon$, where $\hat{\mu}$ is estimated by SLoS procedure. The histogram matches the probability density function of the standard normal distribution very well, which supports our conclusion that $\hat{\mu}$ estimated by SLoS has an asymptotic normal distribution.
Figure 3.7: The pointwise bias of estimated $\hat{\beta}_1(t)$ (on the left panel) and $\hat{\beta}_2(t)$ (on the right panel) in Study 2.

Figure 3.8: The pointwise standard deviation of estimated $\hat{\beta}_1(t)$ (on the left panel) and $\hat{\beta}_2(t)$ (on the right panel) in Study 2.
Figure 3.9: The pointwise RMSE of estimated $\hat{\beta}_1(t)$ (on the left panel) and $\hat{\beta}_2(t)$ (on the right panel) in Study 2.

Figure 3.10: The left panel displays the boxplot of estimated intercepts by SLoS, smooth, OLS and oracle estimators in Study 1. The right panel shows a histogram of estimated intercepts by SLoS method in Study 2. The curve over the bins is the probability density function of the standard normal distribution.
3.5 Application: Canadian Weather Data

We applied our SLoS approach to study the classic Canadian weather data. In the data, daily temperature was recorded at 35 Canadian cities in one year. Figure 2.7 shows the temperature curves in the data. The total annual rainfall in each city was also recorded on a log scale. Besides studying the variations of daily temperature (see Chapter 2 for more information), researchers were also interested in the relationship between the daily temperature curves and the annual rainfall. Previous studies such as James et al. (2009) suggested that the temperature in summer time had limited or no impact on annual rainfall. This suggestion was supported by our study using the SLoS method.

In the study, the relationship between daily temperature and the annual rainfall was modeled by the functional linear regression model (3.1), where daily temperature was the covariate and annual rainfall was the response. We used the SLoS method to determine the coefficient function. In the computation, we used a cubic B-spline basis defined on \([0, 365]\) with 43 basis functions. We also restricted the start and the end of the coefficient function \(\beta(t)\) to be equal to respect the periodic nature of temperature cycles. The tuning parameters were chosen by cross-validation using the proposed two-stage procedure. Figure 3.11 shows the temperature coefficient function estimated by our SLoS method and the
classic penalized B-spline expansion method (Cardot et al., 2003). From the figure, we can see that the coefficient function is zero roughly in January and summer months (June, July and August). This suggests temperature in summer had no effect on annual rainfall, which supports the discovery of previous research on this data. Our method also produced a smooth estimate of $\beta(t)$ on the non-null sub-regions. It indicates that temperature in the early spring, fall and late winter time had significant contribution to annual rainfall, which is supported by the smooth estimator.

3.6 Concluding Remarks

When a large number of predictors are involved in a regression problem, parsimonious models via SCAD or other regularization methods have been proved to enjoy less variability and better interpretability. We propose the functional SCAD regularization method which extends the ordinary SCAD to the functional setting. The functional SCAD, when combined with the penalized B-spline expansion method, yields a smooth and locally sparse (SLoS) estimator of the coefficient function in a functional linear model. The SLoS procedure is a combination of three techniques: the functional SCAD that is responsible for identifying the null sub-regions of the coefficient function without over shrinking the non-zero values, the B-spline basis expansion that is used to practically compute the SLoS estimator efficiently thanks to its compact support property, and the roughness regularization that assures the smoothness of our estimator even when a large number of knots are used to define the B-spline basis. Therefore, our method is able to accurately identify the null region and simultaneously produce a smooth estimator on the non-null region. We also show that our estimator enjoys the oracle property, which means asymptotically it performs as well as if we knew the true null sub-regions of the coefficient function. Comparing to existing methods in the literature, our estimation procedure is conceptually more elegant and computationally simpler. The simulation studies show that our estimator has superior numerical performance and finite-sample properties. The application on Canadian weather data demonstrates the practical merit of our method.

Finally, we shall point out that, the functional SCAD is a very general regularization technique. Its nice shrinkage property entitles it to numerous applications in locally sparse modeling beyond functional linear regression. For example, it can be used in a spline smoothing problem to obtain a smooth and locally sparse estimator of the unknown curve, and in
function principal component analysis to produce smooth and locally sparse functional principal components.

3.7 Proofs

Proof of Theorem 3.1. From (3.4) we already have

\[ \min_{t \in [t_{j-1}, t_j]} |\alpha_j(t)| \leq M^{\frac{1}{q}} T^{-\frac{1}{q}} \|\alpha_j\|_q \leq \|\alpha_j\|_\infty. \]

Since \( p_\lambda \) is continuous and monotone, we then have

\[
\frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( M^{\frac{1}{q}} T^{-\frac{1}{q}} \|\alpha_j\|_q \right) \leq \frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( \max_{t \in [t_{j-1}, t_j]} |\alpha(t)| \right) \\
= \frac{1}{T} \sum_{j=1}^{M} \max_{t \in [t_{j-1}, t_j]} p_\lambda (|\alpha(t)|) \frac{T}{M} \\n\rightarrow \frac{1}{T} \int_{D} p_\lambda (|\alpha(t)|) \, dt.
\]

where (3.13) is the upper Darboux sum of \( p_\lambda \circ |\alpha| \) with respect to the partition defined by the equispaced points \( 0 = t_0 < t_1 < \cdots < t_M = T \). Note that the mesh size of this partition converges to 0 as \( M \to \infty \). Also note that the continuity of \( \alpha \) implies the continuity of \( |\alpha| \) and then the continuity of \( p_\lambda \) implies the continuity of \( p_\lambda \circ |\alpha| \). Therefore, the upper Darboux sum (3.13) exists and equals to its Riemann integral which exists since \( p_\lambda \) is bounded. That is why we have (3.14). Similarly,

\[
\frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( M^{\frac{1}{q}} T^{-\frac{1}{q}} \|\alpha_j\|_q \right) \geq \frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( \min_{t \in [t_{j-1}, t_j]} |\alpha(t)| \right) \\
= \frac{1}{T} \sum_{j=1}^{M} \min_{t \in [t_{j-1}, t_j]} p_\lambda (|\alpha(t)|) \frac{T}{M} \\
\rightarrow \frac{1}{T} \int_{D} p_\lambda (|\alpha(t)|) \, dt.
\]

Thus, by the standard sandwich theorem we have

\[
\lim_{M \to \infty} \frac{1}{M} \sum_{j=1}^{M} p_\lambda \left( M^{\frac{1}{q}} T^{-\frac{1}{q}} \|\alpha_j\|_q \right) = \frac{1}{T} \int_{D} p_\lambda (|\alpha(t)|) \, dt.
\]
Now we proceed to prove Theorem 3.2. The proof is in the same spirit of the one in Fan and Li (2001). We first write

\[ Q(b) = \ell(b) + \int_D p_{\lambda_n}(|\alpha(t)|) \, dt \]

with

\[ \ell(b) \overset{def}{=} \frac{1}{n} \sum_{i=1}^n \left( \langle \beta, X_i \rangle - \langle B^T b, X_i \rangle + \varepsilon_i \right)^2 + \gamma b^T V b \]

\[ \ell(b) = \frac{1}{n} \sum_{i=1}^n \left( \langle \beta - B^T b, X_i \rangle + \varepsilon_i \right)^2 + \gamma b^T V b \]

\[ \ell(b) = \frac{1}{n} \sum_{i=1}^n \left( \langle \beta - B^T b, X_i \rangle^2 + 2\varepsilon_i \langle \beta - B^T b, X_i \rangle + \varepsilon_i^2 \right) + \gamma b^T V b \]

\[ \ell(b) = (\Gamma_n(\beta - B^T b), (\beta - B^T b) + 2 \frac{n}{n} \sum_{i=1}^n \varepsilon_i \langle \beta - B^T b, X_i \rangle + \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 + \gamma b^T V b, \]

where \( \Gamma_n \) is the empirical version of the covariance operator \( \Gamma \) of \( X \), defined by

\[ (\Gamma_n x)(v) \overset{def}{=} \frac{1}{n} \sum_{i=1}^n \int_D X_i(v) X_i(u) x(u) \, du. \]

Therefore,

\[ \nabla \ell(b) = 2(\Gamma_n B, B^T) b - 2(\Gamma_n \beta, B) - \frac{2}{n} \sum_{i=1}^n \varepsilon_i \langle X_i, B \rangle + 2\gamma V b \]

and

\[ \Delta \ell(b) = 2(\Gamma_n B, B^T) + 2\gamma V. \]

**Proof of Theorem 3.2.** To simplify the notation, we shall assume the model does not contain an intercept, i.e. \( \mu = 0 \). At the point \( b = b_s \), we have

\[ \nabla \ell(b_s) = 2(\Gamma_n B, B^T) b_s - 2(\Gamma_n \beta, B) - \frac{2}{n} \sum_{i=1}^n \varepsilon_i \langle X_i, B \rangle + 2\gamma V b_s. \]

\[ = 2(\Gamma_n s - \beta, B) - 2(\Gamma_n \beta, B) - \frac{2}{n} \sum_{i=1}^n \varepsilon_i \langle X_i, B \rangle + 2\gamma V b_s \]

\[ = 2(\Gamma_n (s - \beta), B) - 2(\frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i, B) + 2\gamma V b_s. \]  \( (3.15) \)
CHAPTER 3. FUNCTIONAL SCAD AND SLOS ESTIMATOR

We now develop bounds for $\nabla \ell(b_s)$ and $\Delta \ell(b_s)$. We begin by bounding each term in (3.15). For the first term, according to Cardot et al. (2003), $||\Gamma_n(s - \beta)||^2/2 + \gamma_n||D^2s||^2 < C_5(M^{-2p} + \gamma_nM^{2(m-p)})$ a.s. for some constant $C_5 > 0$, and $\sup_{1 \leq i \leq M+d} \sum_{j=1}^{M+d} |(B_i, B_j)| = O(M^{-1})$. For the second term, according to assumption (C1) and by CLT (see Aldous (1976)), $\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right)$ converges to a Gaussian random element in $L^2(D)$ in distribution, whose mean is 0. This implies

$$\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i, B \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right| ||B|| = O_P \left( n^{-\frac{1}{2}} M^{-\frac{1}{2}} \right) 1 \quad (3.16)$$

where $||B||$ denotes the vector $(||B_1||, ||B_2||, \ldots, ||B_{M+d}||)^T$. For the third term, note that each entry in $V$ is also of order $M^{-1}$, and the number of non-zero elements in each row of $V$ is at most $d + 1$, where $d$ is the degree of the B-spline basis. Given these bounds, and noting that $\sup_{M \geq 1} ||b_s||_\infty < \infty$ by the local property of B-spline bases, we have

$$|\nabla \ell(b_s)| \leq 2||\Gamma_n(s - \beta)|| ||B|| + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i, B \right| + 2\gamma nVb_s$$

$$= O_P \left( M^{-p-\frac{1}{2}} + \gamma_n^2 M^{(m-p)-\frac{1}{2}} + M^{-\frac{3}{2}} n^{-\frac{1}{2}} + \gamma_n M^{-1} \right) 1.$$  

$$= O_P \left( M^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \quad (\text{by (C4)})$$

By Zhou et al. (1998), $0 < \lambda_{\min}(\Gamma_n B, B) = \Theta_P(M^{-1})$ and $\lambda_{\max}(\Gamma_n B, B) = \Theta_P(M^{-1})$, where $\lambda_{\min}(\Gamma_n B, B)$ and $\lambda_{\max}(\Gamma_n B, B)$ denote the smallest and the largest eigenvalues of the matrix $(\Gamma_n B, B)$, and the notation $u = \Theta_P(v)$ denotes $u/v$ is bounded away from zero and infinity in probability. Since $0 < \lambda_{\min}(\gamma_n V) = O(\gamma_n M^{-1}) = \lambda_{\max}(\gamma_n V)$ and $\gamma_n = o(1)$ according (C4), we then have that $\Delta \ell(b) = \Theta_P(M^{-1})$ and is positive definite.

Let $S(s)$ denote the set $\{t \in D : s(t) \neq 0\}$. For bounds of the first-order and the
CHAPTER 3. FUNCTIONAL SCAD AND SLOS ESTIMATOR

second-order derivatives of \( \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt \) with respect to \( b_s \), we have

\[
\left| \nabla \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt \right| = \left| \int_{S(s)} \nabla p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt \right|
\]

\[
= \int_{S(s)} p'_{\lambda_n} \left( |B^T(t)b_s| \right) B(t) \, dt
\]

\[
\leq \sqrt{\int_{S(s)} p'_{\lambda_n} \left( |B^T(t)b_s| \right)^2 \, dt} \|B\|
\]

\[
= O_P(M^{-\frac{1}{2}}a_n)1
\]

and

\[
\Delta \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt = \int_{S(s)} \Delta p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt
\]

\[
= \int_{S(s)} p''_{\lambda_n} \left( |B^T(t)b_s| \right)|B|B^T \ast (\text{sgn}(b_s)\text{sgn}(b_s)^T) \, dt
\]

\[
\leq h_n \sqrt{\int_D \left[ \text{sgn}(b_s) \ast B(t) \right] \left[ \text{sgn}(b_s)B(t)^T \right] \, dt}
\]

\[
= o_P(M^{-2})Z,
\]

where \( \ast \) denotes the entrywise multiplication of two vectors or matrices, and \( Z \) is a sparse matrix with 1 in the location \( (i, j) \) such that \( 0 \leq |i - j| \leq 4 \).

Now, let \( \delta_n = n^{-1/2}M^{1/2} + a_n \). Then

\[
D(u) \stackrel{def}{=} Q(b_s + \delta_n u) - Q(b_s)
\]

\[
\geq \ell(b_s + \delta_n u) - \ell(b_s) + \frac{M}{T} \int_{S(s)} p_{\lambda_n} \left( |B^T(t)(b_s + \delta_n u)| \right) \, dt
\]

\[
- \frac{M}{T} \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt
\]

\[
= \delta_n \nabla^T \ell(b_s) + \frac{1}{2} \delta_n^2 u^T \Delta \ell(b_s) u
\]

\[
+ \frac{M}{T} \left[ \delta_n \nabla^T \left( \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt \right) u + \frac{1}{2} \delta_n^2 u^T \Delta \left( \int_{S(s)} p_{\lambda_n} \left( |B^T(t)b_s| \right) \, dt \right) u \right]
\]

\[
= O_P(n^{-1})\|u\| + \Theta_P(n^{-1})\|u\|^2 + O_P(n^{-1})\|u\| + \delta_n^2 o_P(M^{-1})\|u\|^2 > 0
\]
for sufficiently large $C_3$ and $u$ such that $\|u\| = C_3$. In other words, for any given $\epsilon > 0$, there exists a large constant $C_3$ such that

$$\Pr \left\{ \inf_{\|u\|=C_3} Q(b_s + \delta_n u) > Q(b_s) \right\} \geq 1 - \epsilon.$$ 

This mean, there is a local minimizer $\hat{b}$ such that $\|\hat{b} - b_s\| = O_P(n^{-1/2}M^\frac{1}{4} + a_n) = O_P(n^{-1/2}M^\frac{1}{2}).$

When $\mu \neq 0$, a straightforward extension of the above argument can show that besides $\|\hat{b} - b_s\| = O_P(n^{-1/2}M^\frac{1}{4})$, there exists an estimate $\hat{\mu}$ for $\mu$ such that $|\hat{\mu} - \mu| = O_P(n^{-1/2})$. $\square$

Now we proceed to prove Theorem 3.3. In Section 3.3 we have divided $D$ into $D_1$ and $D - D_1$. We now further divide $D - D_1$ into two groups: $D_2 = \{t \in D : \beta(t) = 0\} = N(\beta)$ and $D_3 = D - D_1 - D_2$. Let $S_j$ denote the support of $B_j(t)$. Given $M$ and $\lambda_n$, besides $B_1$ in Section 3.3, we also further divide the rest of $B$ into two groups: $B_2$ that consists of the those $B_j$ such that $S_j \subset D_2$, and $B_3$ that consists of the rest. Correspondingly, we divide $b$ into three parts: $b_1, b_2$ and $b_3$. Thus, in matrix form, for $\alpha$ in (3.6) we have

$$\alpha = B_1^T b_1 + B_2^T b_2 + B_3^T b_3.$$ 

Also, we have a decomposition of $U = (U_1, U_2, U_3)$. Let $F_j$ denote the union of supports of those basis functions in $B_j$ for $j = 1, 2, 3$. By the compact support property of B-spline basis functions, it is easy to see that $F_j$ converges to $D_j$ for each $j = 1, 2, 3$ as $M \to \infty$.

**Proof of (a) of Theorem 3.3.** We shall first show that $\hat{b}_2 = 0$ with probability tending one. Fix $b_j$ at $b_{sj}$ except a $b_k$ in $b_2$:

$$\frac{\partial Q(b)}{\partial b_k} = \frac{\partial \ell(b)}{\partial b_k} + M \frac{T}{T} \int_D p_{\lambda_n}^t \left( |B^T(t)b_s| \right) B_k(t) \text{sgn}(b_k) \, dt$$

$$= \frac{\partial \ell(b)}{\partial b_k} + \sum_{i=1}^{M+d} \frac{\partial^2 \ell(b)}{\partial b_i \partial b_k} (b_i - b_{si}) + \text{sgn}(b_k) \lambda_n M \frac{T}{T} \int_D p_{\lambda_n}^t \left( |B^T(t)b_s| \right) \lambda^{-1} B_k(t) \, dt$$

$$= \lambda_n \left[ O_P(n^{-1/2}M^{-\frac{1}{2}}/\lambda_n) + O_P(n^{-1/2}M^{-\frac{1}{2}}/\lambda_n) \right.$$ 

$$\left. + \text{sgn}(b_k)T^{-1}\lambda_n^{-1} M \int_{S_k} p_{\lambda_n}^t (|b_k B_k(t)|) B_k(t) \, dt \right].$$
CHAPTER 3. FUNCTIONAL SCAD AND SLOS ESTIMATOR

76

Since by condition (C4) \( n^{-1/2}M^{-\frac{1}{2}}/\lambda_n \to 0 \) and \( \liminf_{n \to \infty} \liminf_{\theta \to \theta^+} p'_{\lambda_n}(\theta)\lambda_n^{-1} > 0 \), we have

\[
\lambda_n^{-1}M \int_{S_k} p'_{\lambda_n}(|b_k B_k(t)|) B_k(t) \, dt \geq \Theta P(1)
\]

and hence the sign of \( \frac{\partial Q(b)}{\partial b_k} \) is determined by \( b_k \). Therefore, \( \hat{b}_k = 0 \) as \( \hat{b} \) minimizes \( Q(b) \). This shows that \( \hat{b}_2 = 0 \) with probability tending to one. Thus, the union \( \hat{F}_2 \subseteq D_2 \) in probability and converges to \( N(\beta) \) in probability since \( F_2 \subseteq D_2 = N(\beta) \) and \( F_2 \to D_2 \).

Since \( \|\hat{b} - b_s\| = O_P(n^{-1/2}M^{\frac{1}{2}}) \) and \( \|s - \beta\|_\infty = O(M^{-p}) \), we have \( \|\hat{\beta} - \beta\|_\infty = O_P(n^{-1/2}M^{\frac{1}{2}} + M^{-p}) = O_P(\lambda_n + M^{-p}) \) as \( n^{-1/2}M^{\frac{1}{2}} \) is dominated by \( \lambda_n \) given the condition (C4). That means, there is some \( C > 1 \) such that \( \|\hat{\beta} - \beta\|_\infty \leq C(\lambda_n + M^{-p}) \) in probability. We set \( C_4 = 2C \). Then, for all \( t \in D_1 \) (and hence \( |\beta(t)| \geq 2Ca(\lambda_n + M^{-p}) \)) we have \( |\hat{\beta}(t)| \geq Ca(\lambda_n + M^{-p}) > a\lambda_n \) collectively in probability. Thus, \( D_1 \subseteq S(\hat{\beta}) \) in probability. This also implies that \( N(\hat{\beta}) \subseteq D_2 \cup D_3 \) in probability as \( n \to \infty \). Since \( D_3 \) shrinks to \( \emptyset \) and \( \hat{F}_2 \to N(\beta) \), for any \( \epsilon > 0 \), we then have \( N(\beta) \subseteq N'(\hat{\beta}) \) in probability. This proves that \( N(\hat{\beta}) \to N(\beta) \) and \( S(\hat{\beta}) \to S(\beta) \) in probability. Now suppose \( t \) is not in the support of \( \beta \). Then it must be in the interior of \( N(\beta) \). Therefore, for sufficiently large \( M \), there is always some \( j \) such that \( B_{j-d-1}, B_{j-d}, \ldots, B_{j+d+1} \) whose supports are all inside \( N(\beta) \). Since \( \hat{b}_2 = 0 \) with probability tending to one, \( t \in N(\hat{\beta}) \) with probability tending to one as well. This completes the proof of part (a). As a by-product, we have \( \nabla \int_{D_1} p(B^T(t)\hat{b}) = 0 \) with probability tending to 1. We will need this result in the proof of part (b). \( \Box \)

Proof of (b) of Theorem 3.3. Fix any \( t \in D \) such that \( \beta(t) \neq 0 \). Choose \( n \) and \( M \) that are sufficiently large so that \( |\beta(v)| > C_4a(\lambda_n + M^{-p}) \) over more than \( 2(d+1) \) consecutive \( M^{-1} \) subintervals around \( t \) (this is possible given the continuity of \( \beta \)). Then, we have

\[
\sqrt{\frac{n}{M}} \left( \hat{\beta}(t) - \beta(t) \right) = n^{\frac{1}{2}}M^{-\frac{1}{2}}B_1^T(t) \left( \hat{b}_1 - b_{s1} \right) + n^{\frac{1}{2}}M^{-\frac{1}{2}} \left( B_1^T(t)b_{s1} - \beta(t) \right). \hspace{1cm} (3.17)
\]

The term \( B_1^T(t)b_{s1} - \beta(t) \) in the RHS of (3.17) is the B-spline approximation error, for which we have

\[
n^{\frac{1}{2}}M^{-\frac{1}{2}} \left( B_1^T(t)b_{s1} - \beta(t) \right) = n^{\frac{1}{2}}M^{-\frac{1}{2}} \left( s(t) - \beta(t) \right) = O(n^{\frac{1}{2}}M^{-\frac{1}{2}}p) t \hspace{1cm} (3.18)
\]
since \( \| s - \beta \|_\infty = O(M^{-p}) \). For the first term in the RHS of (3.17), we shall first study the estimate \( \hat{b}_1 \). From part (a), with probability tending to one, \( \hat{b}_2 = 0 \). Since \( \nabla \int_{D_1} p_{\lambda_n}(B^T(t)\hat{b}) = 0 \) and all basis functions \( B_j(t) \) associated to \( \hat{b}_1 \) have a support entirely inside \( D_1 \), with probability tending to 1, we have

\[
\hat{b}_1 = \left( U_1^T U_1 + \gamma n V_{11} \right)^{-1} U_1^T (y - U_3 \hat{b}_3) - \left( U_1^T U_1 + \gamma n V_{11} \right)^{-1} \frac{\gamma n}{2} (V_{31}^T + V_{13}) \hat{b}_3,
\]

where \( V \) is written as a block form

\[
V = \begin{pmatrix}
V_{11} & V_{12} & V_{13} \\
V_{21} & V_{22} & V_{23} \\
V_{31} & V_{32} & V_{33}
\end{pmatrix}.
\]

Given that

\[
y = \varepsilon + \langle X, B^T b_s + e \rangle = \varepsilon + U b_s + \zeta
\]

where \( \zeta_i = \langle X_i, e \rangle \) and \( e(t) = \beta(t) - B^T (t) b_s \), and by noting \( b_2 = 0 \), we have

\[
\hat{b}_1 = \left( U_1^T U_1 + \gamma n n V \right)^{-1} U_1^T (\varepsilon + U_1 b_{s1} + U_3 (b_{s3} - \hat{b}_3) + \zeta) - \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} \frac{\gamma n}{2} (V_{31}^T + V_{13}) \hat{b}_3.
\]

Therefore,

\[
\hat{b}_1 - b_{s1} = \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} U_1^T (\varepsilon + U_1 b_{s1} + \zeta + U_3 (b_{s3} - \hat{b}_3)) - b_{s1}
\]

\[
- \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} \frac{\gamma n}{2} (V_{31}^T + V_{13}) \hat{b}_3
\]

\[
= \left( U_1^T U_1 \right)^{-1} U_1^T \varepsilon + \left[ \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} - \left( U_1^T U_1 \right)^{-1} \right] U_1^T \varepsilon
\]

\[
- \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} \frac{\gamma n}{2} n V_{11} b_{s1}
\]

\[
+ \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} U_1^T (\zeta + U_3 (b_3 - \hat{b}_3))
\]

\[
- \left( U_1^T U_1 + \gamma n n V_{11} \right)^{-1} \frac{\gamma n}{2} (V_{31}^T + V_{13}) \hat{b}_3
\]

(3.19)
Below we derive the bounds of (3.19), (3.20), (3.21) and (3.22).

First of all, we have $\lambda_{\text{min}}(U_1^T U_1) = \Theta_P(nM^{-1})$ according to Zhou et al. (1998), and also $\lambda_{\text{min}}(\gamma_n n V) = O_P(\gamma_n n M^{-1})$. Thus, $\lambda_{\text{max}}[(U_1^T U_1 + \gamma_n n V)^{-1}] = O_P(n^{-1}M)$. Let $A \overset{\text{def}}{=} U_1^T U$ and $\Delta A \overset{\text{def}}{=} \gamma_n n V$. Then according to Horn and Johnson (2012), we have

$$
\| (A + \Delta A)^{-1} - A^{-1} \| \leq \frac{\| A^{-1} \| \| \Delta A \|}{1 - \| A^{-1} \Delta A \|} = \frac{O(M^2 n^{-2} \gamma_n n M^{-1})}{1 - O(\gamma_n n M^{-1})O(M^{-1})} = O(\gamma_n M n^{-1}).
$$

Also, by CLT (see the proof of Theorem 3.2), we have

$$
(U_1^T \varepsilon)_k = \sum_{i=1}^{n} \varepsilon_i (X_i, B_k) = O_P(n^{\frac{3}{2}} M^{-\frac{1}{2}}). \quad (3.23)
$$

Thus, the bound for (3.19) is

$$
\left[ (U_1^T U_1 + \gamma_n n V)^{-1} - (U_1^T U_1)^{-1} \right] U_1^T \varepsilon = O_P(\gamma_n M^{\frac{1}{2}} n^{-\frac{1}{2}}). \quad (3.24)
$$

For (3.20), we have

$$
(U_1^T U_1 + \gamma_n n V)^{-1} \gamma_n n V b_1 = O_P(\gamma_n) 1. \quad (3.25)
$$

By $\| \hat{b}_3 - b_{s3} \|_\infty = O_P(n^{-\frac{1}{2}} M^{\frac{1}{2}})$, we have the bound

$$
U_3 (b_{s3} - \hat{b}_3) = O_P(n^{-\frac{1}{2}} M^{\frac{1}{2}}) 1. \quad (3.26)
$$

Since

$$
(U_1^T \zeta)_k = \sum_{i=1}^{n} (X_i, e) (X_i, B_k) 
\leq \sum_{i=1}^{n} |(X_i, e) \cdot (X_i, B_k)| 
\leq \sum_{i=1}^{n} \| X_i \|_1 \| e \|_\infty \| X_i \|_2 \| B_k \|_2 
= O_P(M^{-p}) \sum_{i=1}^{n} \| X_i \|_2^2 \| B_k \|_2 
= O_P(M^{-p-\frac{1}{2}} n),
$$

Hence, we have the desired bounds.
CHAPTER 3. FUNCTIONAL SCAD AND SLOS ESTIMATOR

with (3.26) and \( U_1^T 1 = O_p(M^{-\frac{1}{2}} n^{\frac{1}{2}}) \) (again, by CLT), for (3.21) we have

\[
(U_1^T U_1 + \gamma n V)^{-1} U_1^T (\zeta + U_3 (b_{s3} - \hat{b}_3)) = O_p(M^{-p+\frac{1}{2}} + n^{-1} M) 1. \tag{3.27}
\]

For (3.22), since it is easy to show that \( \hat{b}_3 \) is bounded in probability, we have

\[
(U_1^T U_1 + \gamma n V_{11})^{-1} \frac{\gamma n}{2} (V_{31} + V_{13}) \hat{b}_3 = O_p(n^{-1} M \gamma_n n M^{-1}) = O_p(\gamma_n). \tag{3.28}
\]

Putting (3.24), (3.25), (3.27) and (3.28) together, we have

\[
\hat{b}_1 - b_{s1} = (U_1^T U_1)^{-1} U_1^T \varepsilon + O_p(\gamma_n M^{\frac{1}{2}} n^{-\frac{1}{2}} + \gamma_n + M^{-p+\frac{1}{2}} + n^{-1} M). \tag{3.29}
\]

With (3.17) and (3.18), this gives

\[
\sqrt{\frac{n}{M}} \left( \hat{\beta}(t) - \beta(t) \right) = n^{\frac{1}{2}} M^{-\frac{1}{2}} B(t) \left( \hat{b}_1 - b_{s1} \right) + n^{\frac{1}{2}} M^{-\frac{1}{2}} \left( B(t) b_{s1} - \beta(t) \right)
= n^{\frac{1}{2}} M^{-\frac{1}{2}} B(t) \left( U_1^T U_1 \right)^{-1} U_1^T \varepsilon
+ O_p(\gamma_n + n^{\frac{1}{2}} M^{-\frac{1}{2}} \gamma_n + n^{\frac{1}{2}} M^{-p} + M^{\frac{1}{2}} n^{-\frac{1}{2}}) t
+ O(n^{\frac{1}{2}} M^{-\frac{1}{2}} \gamma_n) t
= n^{\frac{1}{2}} M^{-\frac{1}{2}} B(t) \left( U_1^T U_1 \right)^{-1} U_1^T \varepsilon + o_P(1) t
\xrightarrow{d} N(0, \sigma^2(t))
\]

where according to Zhou et al. (2013), the asymptotic variance \( \sigma^2(t) \) can be determined by

\[
\sigma^2(t) = \lim_{n \to \infty} \frac{M}{n} B(t) \left( U_1^T U_1 \right)^{-1} B(t).
\]

Proof of (c) of Theorem 3.3. An easy calculation shows that

\[
\hat{\mu} = \frac{1}{n} 1^T (y - U \hat{b}).
\]
Thus,
\[
\sqrt{n}(\hat{\mu} - \mu) = \sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle X_i, \hat{\beta} \rangle) - \mu \right]
\]
\[
= \sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle X_i, \hat{\beta} \rangle + \varepsilon_i) - \mu \right]
\]
\[
= \sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \beta \rangle - \langle X_i, \hat{\beta} \rangle + \varepsilon_i \right]
\]
\[
= \sqrt{n} \bar{\varepsilon} + \sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \beta - \hat{\beta} \rangle \right]
\]
\[
= \sqrt{n} \bar{\varepsilon} + \sqrt{n} \langle \bar{X}, \beta - \hat{\beta} \rangle,
\]
where \( \bar{\varepsilon} = n^{-1} \sum_{i=1}^{n} \varepsilon_i \) and \( \bar{X}(t) = n^{-1} \sum_{i=1}^{n} X_i(t) \). Also, since \( \|b_s - \hat{b}\|_2 = O_P(M^{1/2}n^{-1/2}) \) by Theorem 3.2 and \( \|\bar{X}\|_2 = O_P(n^{-1/2}) \) by CLT, with \( e = \beta - B^T b_s \) we have
\[
|\langle \bar{X}, \beta - \hat{\beta} \rangle| = |\langle \bar{X}, B^T b_s + e - B^T \hat{b} \rangle|
\]
\[
\leq |\langle \bar{X}, e \rangle| + |\langle \bar{X}, B^T (b_s - \hat{b}) \rangle|
\]
\[
\leq \|\bar{X}\|_2 \|e\|_2 + \|\bar{X}\|_2 \|B^T (b_s - \hat{b})\|_2
\]
\[
= O_P(n^{-1/2})[O_P(M^{-p}) + O_P(M^{1/2}n^{-1/2})]
\]
\[
= O_P(n^{-1/2}M^{-p} + M^{1/2}n^{-1}).
\]
Thus, given Condition (C4), we have
\[
\sqrt{n}(\hat{\mu} - \mu) = \sqrt{n} \bar{\varepsilon} + O_P(M^{-p} + M^{1/2}n^{-1/2}) = \sqrt{n} \bar{\varepsilon} + o_P(1).
\]
Note that \( \sqrt{n} \bar{\varepsilon} \xrightarrow{d} N(0, \sigma^2_\varepsilon) \). By Slutsky’s Theorem, \( \sqrt{n}(\hat{\mu} - \mu) \) and \( \sqrt{n} \bar{\varepsilon} \) converges to the same limit \( N(0, \sigma^2_\varepsilon) \) in distribution. \( \square \)
Chapter 4

Summary of Thesis

In Chapter 2, we propose an interpretable FPCA (iFPCA) method, which can produce smooth and locally sparse FPCs that are only nonzero on sub-regions where sample curves have significant variations and strictly zero on the other sub-regions. This feature makes our derived FPCs more interpretable. In practice we develop an efficient algorithm to compute the interpretable FPCs using projection deflation, and in theory we show that the estimated interpretable FPCs are strongly consistent and asymptotically normal under mild conditions. The simulation studies show that FPCs produced by the iFPCA method possess more interpretability than FPCs obtained by other methods, while at the same time account for similar variations of the sample curves as the FPCs that are estimated by other FPCA methods. Our iFPCA method is also demonstrated by analyzing the Canadian weather data and the pinch force data.

In Chapter 3, we propose the fSCAD regularization technique and apply it to derive a smooth and locally sparse (SLoS) estimator of the coefficient function in functional linear regression. The SLoS estimation procedure is a beautiful combination of the following techniques: the fSCAD regularization that is able to identify the null sub-regions of the coefficient function without over shrinking the non-zero values, the B-spline expansion that enables us to practically compute the SLoS estimator thanks to the compact support property of B-spline basis functions, and the regularization via roughness penalty that assures the smoothness of our estimator even when a large number of knots are used to define the B-spline basis. Therefore, our method is able to produce a precise estimate of the null region of the coefficient function and simultaneously yield a smooth estimator on its non-null region. Asymptotic analysis shows that our estimator enjoys the oracle property, which means
asymptotically it performs as well as if we knew the true null sub-regions of the coefficient function. Comparing to existing methods in the literature, our estimation procedure is conceptually more elegant and computationally simpler. The simulation studies show that it has superior numerical performance and finite-sample properties. We also demonstrate our method in a study on the Canadian weather data. In the study, our method suggests that summer time had limited or no impact on annual rainfall, which is supported by previous studies.

The fSCAD is a very general regularization technique that can be applied to applications beyond functional linear regression. In the future, we plan to investigate its applications in other tasks of functional data analysis, such as spline smoothing and functional principal component analysis. We are also interested in estimating the degree of freedom of SLoS, because an estimate of degree of freedom enables us to use various information criteria such as BIC or AIC in selecting tuning parameters.
Bibliography


