Statistical Learning and Inference For Functional Predictor Models via Reproducing Kernel Hilbert Space

by

Meichen Liu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Functional regression is a cornerstone for understanding complex relationships where predictors or responses (or both) are functions. A particularly powerful framework within this domain is the Reproducing Kernel Hilbert Space (RKHS), which facilitates the handling of infinite-dimensional data through a finite set of parameters.

This thesis delves into three specific topics within functional regression using RKHS, showcasing innovative methodologies and their applications to real-world data. The first topic explores functional linear expectile regression, a method that offers a nuanced view of conditional response distributions, particularly beneficial for asymmetric distributions or when tail behaviour is of interest. The second topic ventures into functional smoothed score (SS) classification. The study investigates the functional classifier's generalization ability and convergence property. The last chapter addresses the challenge of estimation and inference for the slope function in logistic regression under case-control designs, which exert influence over rare disease research.

The three topics contribute to functional regression, offering robust and theoretically sound methodologies for analyzing complex data structures. Through the representation theorem of RKHS, this thesis advances statistical modeling techniques and provides practical tools for tackling real-world problems in diverse scientific domains.

Preface

The research conducted in this thesis was under the supervision of Dr. Linglong Kong and Dr. Bei Jiang at the University of Alberta.

Chapter 3 of the thesis has been published as Liu, M., Pietrosanu, M., Liu, P., Jiang, B., Zhou, X., Kong, L., & Initiative, the A. D. N. (2022). Reproducing kernel-based functional linear expectile regression. Canadian Journal of Statistics, 50(1), 241–266. I was responsible for the development of methodology, theoretical proof, data analysis, manuscript composition, and revision. This work was completed under the supervision of Dr. Linglong Kong and Dr. Bei Jiang. Dr. Linglong Kong is the corresponding author.

Chapter 4 of the thesis has been submitted to the Journal of Machine Learning Research. I was responsible for the development of methodology, most of the theoretical derivation, numerical studies as well as manuscript composition. This work was conducted under the supervision of Dr. Linglong Kong and Dr. Bei Jiang. Dr. Linglong Kong is the corresponding author.

Chapter 5 is ready to be submitted soon. I was responsible for the development of methodology, theoretical derivation, numerical studies, and manuscript composition. This work was conducted under the supervision of Dr. Linglong Kong and Dr. Bei Jiang. Dr. Linglong Kong will be the corresponding author.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Linglong Kong and Dr. Bei Jiang, who have supported me throughout my Ph.D. study and research with their excellent guidance, patience, and understanding.

I would also like to express my sincere thanks to Dr. Linglong Kong, Dr. Bei Jiang, Dr. Ivan Mizera, and Dr. Xingyu Li for serving on my dissertation examination committee. Additionally, I appreciate my external examiner: Dr. Haiying Wang for his thorough review and insightful comments.

My sincere thanks are extended to Dr. Xingcai Zhou, Dr. Jinhan Xie, Dr. Peijun Sang, Mattehew Pietrosanu, and Dr. Peng Liu for their valuable comments, help, and suggestions on my research.

I would also thank all the professors who taught me for their mentorship and encouragement at the University of Alberta and the family of the Department of Mathematical and Statistical Sciences for their help and support with administrative matters. Last but not least, I would like to thank my parents, my families, and dear friends for their unconditional companionship and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Outline of the Thesis

Innovative data manipulating technologies and modern computing have propelled researchers to collect, store, and study data of complex structures. Among these, functional data represents a fundamental type. Functional data analysis (FDA) encompasses the study and theoretical framework for data represented as functions, images, shapes, or other general objects [182]. This approach finds applications in various fields, including medical science, biology, and signal processing, where data naturally manifest as functional forms [143]. Instances abound in gene expression microarray data, single nucleotide polymorphism (SNP) data, magnetic resonance imaging (MRI) data, high-frequency financial data, etc [67, 96, 143, 197].

While intrinsically infinite-dimensional, functional data is always observed discretely over a grid, where observations at nearby grid points are highly correlated due to their spatial or temporal nature. The number of grid points could be much larger than the number of observations. The high dimensionality of observed data poses challenges both for theory and computation.

Research on functional data or other structured data can lead to more nuanced and accurate understandings of complex phenomena and thus offer profound implications for a wide range of applications. Some merits include:

1. Improved modeling: FDA allows for more complex and flexible modeling of

data that would be difficult or impossible to model using traditional methods. For example, it can account for non-linear trends, temporal dependencies, and other patterns that may be missed by simpler models.

2. Enhanced prediction: FDA often leads to more accurate predictions by incorporating information from the entire curve or structure, rather than relying on summary statistics or discrete data points.

3. Improved decision-making: Analyzing functional data can provide deeper insights into patterns and relationships within the data, leading to more informed decision-making.

In this thesis, we aim at a class of functional predictor models where the response Y, either continuous or dichotomous, is related to a square-integrable random function $X(\cdot)$ through the linear form $\eta(X) := \alpha_0 + \int_{\mathcal{I}} X(t)\beta_0(t)dt$, where α_0 is the intercept, $\mathcal{I}$ is a compact subset of an Euclidean space, $\beta_0(\cdot)$ denotes an unknown slope function. The domain $\mathcal{I}$ is based on a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ consisting of n independent copies of $(X(t), Y)$.

Some tools for functional linear regression are among functional principal components analysis (FPCA) [61], basis spline, smoothing spline, etc. FPCA has taken off to be the most prevalent and effective tool in FDA since invented. This is partly because it facilitates the conversion of inherently infinite-dimensional functional data to a finite-dimensional vector of random scores. In a functional linear model, FPCA ultimately relies on an efficient representation of β_0 in terms of the leading functional principal components of X [16]. However, these FPCs might fail to form an appropriate basis to express β_0 or have little predictive power. Consequently, FPCA-based methods might not excel the other alternatives. In practice, this phenomenon has been observed for functional data in the Canadian weather dataset [16, 143], as well as the Alzheimer's Disease Neuroimaging Initiative (ADNI) data.

Another efficient, also commonly used method is the reproducing kernel Hilbert

space (RKHS). Due to the unbounded inverse of a compact operator, regularization is routinely adopted for any procedure that involves the inverse of a compact operator, as it does in the RKHS method. The smoothness regularized estimator enjoys superiority when FPCA couldn't demonstrate desirable prediction power and circumvents additional assumptions on the spacing of the eigenvalues of the covariance operator for X as well as Fourier coefficients of β_0 concerning the eigenfunctions, which are required by the FPCA-based approaches.

The general regularization method to estimate a function of interest η on a generic domain $\mathcal{I}$ using stochastic data takes the form of

$$L(\eta|\text{data}) + \lambda J(\eta) \tag{1.1}$$

where $L(\eta|\text{data})$ is usually taken as the average loss or the minus log-likelihood of the data and $J(\eta)$ is a quadratic roughness functional with a null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ of low dimension, and the smoothing parameter λ in (1.1) is the Lagrange multiplier to balance the fidelity to the data and the plausibility.

By adding a roughness penalty $J(\eta)$ to $L(\eta)$, one considers only smooth functions in the space $\{\eta : J(\eta) < \infty\}$ or a subspace therein. To assist analysis and computation, one needs a metric and a geometry in the space, and the score $L(\eta) + \lambda J(\eta)$ to be continuous in η under the metric. The reproducing kernel Hilbert space (RKHS), of which a brief introduction is presented in the next chapter, is adequately equipped for the purpose.

1.1 Thesis Outline

The thesis focuses on several unexplored aspects in recent functional regression literature and provides a brief review in terms of functional predictor models and RKHS in Chapter 2.

- **Chapter 3:** Functional Linear Expectile Regression in RKHS

We propose a scalar-on-function linear expectile regression model under the RKHS framework. The proposed estimator is shown to achieve optimal convergence rates, supported by both theoretical bounds and practical implementation via the alternating direction method of multipliers (ADMM) algorithm. Empirical validation through simulation studies and a neuroimaging data analysis underscores its advantages over FPCA-based methods.

- **Chapter 4:** Functional Data Classification with Smoothed Score Classifier

  In this chapter, we investigated the functional smoothed score classifier's generalization ability and Fisher consistency. Theoretical and numerical analyses demonstrate the high-accuracy slope function estimation and reveal the trade-off between tuning parameter selection and candidate function class size. Computationally, we tackle the nonconvex optimization by bringing an efficient proximal gradient algorithm. Extensive numerical studies demonstrate the favorable performance of the proposed method compared with some popular classifiers.

- **Chapter 5:** RKHS-Based Functional Linear Logistic Regression under Case Control designs

  We explored the estimation and inference of the slope function in logistic regression under the case-control designs. By embedding the slope function in RKHS and applying roughness regularization, this methodology enhances estimation precision. Theoretical contributions include establishing the convergence rate of the slope coefficient function and the asymptotic normality of the test statistic. Simulation studies and two real-data applications demonstrate the superiority of this estimator compared to traditional methods.

Chapter 2

Background

2.1 Functional Predictor Models

Most work in functional predictor regression is based on a variant of the functional linear model (FLM), introduced by Ramsay and Dalzell [142] and first written in its commonly encountered form by Hastie and Mallows [64]:

$$Y_i = \alpha + \int X_i(t)\beta(t)dt + \varepsilon_i, \tag{2.1}$$

where $Y_i, i = 1, \ldots, N$ is a continuous response, $X_i(t)$ a functional predictor, $\beta(t)$ functional coefficient, α intercept, and $\varepsilon_i \sim N(0, \sigma^2)$ residual errors. For non-Gaussian responses, many have worked with the generalized FLM (GFLM), first introduced by Marx and Eilers [113] for exponential family responses as

$$g\{E(Y_i)\} = \alpha + \int X_i(t)\beta(t)dt \tag{2.2}$$

for some link function $g(\cdot)$. **ramsay_silverman_2005** discussed using truncated basis function expansions for $X_i(t) = \sum_{k=1}^{K_X} X_{ik}^* \phi_k(t)$ and $\beta(t) = \sum_{k=1}^{K_B} B_k^* \psi_k(t)$. If we let $\mathbf{X}_i^* = \left[X_{i1}^*, \ldots, X_{iK_X}^*\right]^\top, \mathbf{B}^* = \left[B_1^*, \ldots, B_{K_B}^*\right]^\top, \phi(t) = [\phi_1(t), \ldots, \phi_{K_X}(t)]^\top$, and $\psi(t) = [\psi_1(t), \ldots, \psi_{K_B}(t)]^\top$, then the functional regression terms can be written

$$\int X_i(t)\beta(t)dt = \mathbf{X}_i^* J_{\phi,\psi} \mathbf{B}^* \tag{2.3}$$

$$= \mathbf{X}_i^{**} \mathbf{B}^* \tag{2.4}$$

where $\mathbf{X}_i^{**} = \mathbf{X}^* J_{\phi,\psi}$ and $J_{\phi,\psi} = \int \phi(t)\psi(t)^\top dt$, and thus the FLM simplifies into a linear model of dimension K_B. If the same orthogonal basis functions are used for ϕ and ψ, then $J_{\phi,\psi} = I$ and $\mathbf{X}_i^{**} = \mathbf{X}_i^*$. What's more, If all $X_i(t)$ are observed on the same regular grid $\boldsymbol{t}$ of size T, then one can write $J_{\phi,\psi} = \boldsymbol{\Phi}\boldsymbol{\Psi}^\top$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}(t_1), \ldots, \boldsymbol{\phi}(t_T)]$ and $\boldsymbol{\Psi} = [\boldsymbol{\psi}(t_1), \ldots, \boldsymbol{\psi}(t_T)]$.

Regularization techniques in functional predictor regression involve methods such as truncation, roughness penalties, or sparsity applied to functional coefficients $\beta(t)$ and/or predictors $X_i(t)$. Regularizing $\beta(t)$ serves to (a) mitigate collinearity, (b) enhance coefficient interpretability, and (c) potentially improve estimation and prediction efficiency by leveraging the functional nature of the data across time points t. The optimal choice of basis functions and regularization strategy varies across datasets based on function characteristics and true functional coefficients.

For $X_i(t)$, regularization aims to (a) diminish measurement errors, enhancing estimation efficiency for Gaussian responses and reducing bias for nonlinear models with non-Gaussian outcomes, and (b) accommodate functional predictors with sparse and irregular grids.

Methodological developments in functional predictor regression often align with the **ramsay_silverman_2005** strategy, employing diverse basis functions and regularization approaches. Each method is related to Model (2.3) above. Additional advancements introduced by various researchers extend models to handle non-Gaussian or correlated responses, incorporate nonfunctional fixed or random effect terms, handle multiple functional predictors, accommodate functions on irregular grids, extend to nonlinear functional predictor models, and introduce specific approaches for variable selection or inference.

2.1.1 Principal component analysis (PCA), spline and their combinations

The original work on FLM was done using multivariate analysis techniques for functions sampled on a common grid, with principal component regression (PCR), partial least squares (PLS), or ridge regression used to reduce collinearity in the resulting high-dimensional multiple regression. In PCR, orthogonal empirically determined PC bases estimated from the decomposition of $X^\top X$ are used for both Φ and Ψ. According to Jolliffe [77], the original work on PC, for instance, Kendall [82] suggested a variable selection approach to regularization, supposing that the PCs explaining the largest amount of variability in $X_i(t)$ may not necessarily be the most important in predicting Y_i. However, because in many cases the first few PCs appear most important, over time many researchers have regularized by truncating at the first few PCs explaining a fixed proportion of total variability. Both strategies have been used in practice. Other methods extending PCR have been developed for function predictor regression.

Pascal [132] fit an FLM with PCR through truncation and discussed some theoretical results. Müller and Stadtmüller [123] introduced functional PCR (fPCR) for GFLM, using smoothed fPC bases for both $X_i(t)$ and $\beta(t)$ with regularization done by PC truncation. The basis functions were estimated by decomposing a kernel-smoothed estimate of the covariance function after subtracting off a kernel-smoothed estimate of the overall mean, while removing white noise errors. When the functions are irregularly and sparsely sampled, the PC scores Y_i^* are computed using PACE [195], which borrows strength across sparse curves to estimate the covariance matrix and estimates individual PC scores as best linear unbiased predictors (BLUPs) of a linear mixed effect model (LMM).

In a multilevel functional data setting with multiple curves per subject, Crainiceanu et al. [29] used ML-fPCA [138] to estimate the multilevel PC scores, using a truncated set of subject-level scores as the predictors X_i^{**} in Model (2.4), effectively using

subject-level PCs as bases for both ϕ and Ψ and regularizing by truncation. They used a restricted likelihood ratio test (RLRT) [28] to select the truncation parameter, and also discussed a fully Bayesian Markov chain Monte Carlo (MCMC)-based version to obtain estimates and inference. Crainiceanu and Goldsmith [27] developed a Bayesian modeling approach for PCR using WinBUGS, modeling the PC scores and eigenvalues stochastically (but still conditioning on estimated eigenvectors as fixed), which they showed leads to more accurate inference.

Holan et al. [68] developed Bayesian methods to classify subjects based on a functional predictor that is a nonstationary time series. They computed a time-varying Fourier spectrum to transform the time series to the time-frequency dual space, and then constructed a logistic GFLM with the spectrogram image as a predictor, using PC for the basis functions and stochastic search variable selection (SSVS) for regularization. Randolph et al. [144] introduced a method for PCR for which regularization is done by a structured L_2 penalty that incorporates presumed structure directly into the estimation process through a linear penalty operator, leading to a weighted ridge regression in the PC space.

Hastie and Mallows [64] pointed out that standard multivariate approaches such as PCR do not take the ordering inherent to the functional data into account. They introduced the FLM (2.1) and used penalized splines for $\beta(t)$ while assuming all functions were sampled on a common grid $\boldsymbol{t}$. Others have also incorporated spline bases for regularization in the FLM or GFLM settings.

Marx and Eilers [113] introduced penalized signal regression (PSR) for the GFLM that uses B-splines for $\beta(t)$ that are penalized by the P-spline first difference L2 penalty introduced in [41]. They subsequently extended this work in various ways, adding multiple linear and additive scalar fixed effects and linear random effects to the model [42], accommodating multidimensional signals such as images [114], and allowing prespecified weights for adaptive smoothing in Li and Marx [95].

James [73] fit a GFLM using natural cubic spline bases to represent the predictor

functions $X_i(t)$ with regularization done by the truncation inherent in the knot selection. They specified a measurement error model $X_i = X_i^* \mathbf{\Phi}_i + \epsilon_i$, where X_i and $\mathbf{\Phi}_i$ are vectors with the functional predictors and natural cubic spline basis functions evaluated on the varying sampling grid $\boldsymbol{t}_i$ across functions, and X_i^* a vector containing the spline coefficients for a common set of knots.

Zhang et al. [200] presented a spline-based FLM with periodic spline bases used for both $\mathbf{\Phi}$ and Ψ, and regularization by roughness penalties. Similar to James [73] and Müller and Stadtmüller [123], their approach accommodated irregularly sampled functions and adjusted for measurement error in $X_i(t)$. Crambes et al. [30] presented a smoothing spline-based FLM, and proved some theoretical results; Yuan and Cai [198] introduced a general reproducing kernel Hilbert space approach to functional linear regression regularized by roughness penalties, implemented using smoothing spline-like kernels.

There are numerous further studies combining the FPCA and splines to perform regularization. One method involves representing the functional predictor $X_i(t)$ using splines and conducting a PC decomposition of the spline coefficients. James [73] applied this method by fitting unpenalized cubic splines to the functional predictors and then performing a FPCA of the matrix of spline coefficients $X^* = [X_1^*, X_2^*, \ldots, X_N^*]$. This basis was then used to further transform the $X(t)$. Natural spline bases for $\mathbf{\Phi}$ and eigenvectors of the spline bases for $\mathbf{\Psi}$ were fit using the expectation maximization (EM) algorithm and weighted least squares (WLS), with the spline coefficient eigenvalues as weights.

Reiss and Ogden [149] adopted a similar strategy in a way to bring together the ideas of PCR and penalized spline regression (PSR) [113]. For Gaussian outcomes, their model transforms $X_i(t)$ using a $T \times K_x$ B-spline design matrix $\mathbf{\Phi}$ and then performs a FPCA on $\mathbf{X\Phi}$ to estimate a truncated set of eigenfunctions Ψ, with $X^{**} = X\Phi\Psi$ in Model (2.4). Regularization is achieved by applying roughness penalties (or P-spline penalties) to either the fPC(fPCR$_C$) or the regression coefficients (fPCR$_R$).

A similar strategy was used in Reiss et al. [148] but they replaced PC bases with partial least squares (PLS) bases (fPLS $_C$, fPLS$_R$), empirical bases to maximize the correlation between $X_i(t)$ and Y_i. They also introduced statistical inferences including bootstrap-based simultaneous confidence bands and RLRT [28] to test the significance of the functional coefficient alternative.

Another approach is to transform the functional predictors $X_i(t)$ using PCA while parameterizing the coefficient function $\beta(t)$ using splines. Cardot et al. [18] introduced a two-step method that first performs PCR and then uses B-splines with a roughness penalty to smooth the resulting $\beta(t)$. Crainiceanu and Goldsmith [27]presented a Bayesian generalized functional linear model (GFLM) with PCs for $X_i(t)$ and B-splines for $\beta(t)$, which was regularized using the random walk prior described by [91] on the spline coefficients. The PC scores and eigenvalues were modelled stochastically. Goldsmith et al. [54] developed a variational Bayes approximation for this model, which was sufficient for estimation but not necessarily for inference.

The frequentist adaptation proposed by Goldsmith et al. [51] is known as Penalized Functional Regression (PFR). PFR utilizes PCs for the functional predictor Φ and truncated power series for Ψ, while applying L_2 penalties to regularize the coefficients of the truncated power series spline, as discussed by Ruppert et al. [155]. PFR can handle sparse, irregular, or multi-level functional data by incorporating PC scores obtained through PACE or ML-fPCA into the Generalized Functional Linear Model (GFLM). Non-functional linear predictors are also considered, with a recommendation to retain a substantial number of PCs for irregularly sampled functions. Functional hypothesis testing procedures based on the Randomization-based Likelihood Ratio Test (RLRT) for comparing functional vs constant effects and selecting among multiple functional predictors were proposed by Swihart et al. [169].

This methodology was extended by Goldsmith et al. [52] to handle data with repeated measurements in both the response and the functional predictor. Scalar random effects were introduced to the GFLM to account for repeated measurements.

Although the model includes i.i.d. random effects, it could be extended to more complex random effect structures. However, this extension does not account for the correlation among repeated functions when calculating PC scores. An alternative approach considering this correlation was presented by Gertheiss et al. [48]. They utilized an extension of ML-fPCA [138] for Longitudinal Functional Principal Component Analysis [56] to compute separate eigenvectors for random intercepts and slopes. A shared PC score between them was estimated using a regression approach, and this score was incorporated into the GFLM.

2.1.2 Wavelets and general bases for $X_i(t)$ and/or $\beta(t)$

Brown et al. [14] introduced a wavelet-based, Bayesian approach for fitting FLMs to data on a common equally spaced grid $\boldsymbol{t}$. They used orthogonal wavelet bases to represent both $X_i(t)$ and $\beta(t)$, and used stochastic search variable selection (SSVS) to select a subset of important wavelet coefficients. Their approach effectively used the orthonormal wavelet transform matrix for both Φ and Ψ, but coefficients were actually calculated using the pyramid-based discrete wavelet transform algorithm, allowing the approach to scale up to functional data on very large grids. As mentioned above, the use of wavelets and sparsity leads to adaptive regularization, making it well-suited for modeling spatially heterogeneous functions with local features such as spikes. Subsequent work involving wavelet representations for $X_i(t)$ and $\beta(t)$ includes that of Heo [67], Malloy et al. [109], Wang et al. [184], and Zhao et al. [203].

A common basis transform Φ for $X_i(t)$ and $\beta(t)$ was employed by Ratcliffe et al. [146] for FLM, where $X_i^{**} = X_i^* \Phi \Phi^\top$ and regularization was posed by truncation. They used Fourier basis coefficients but presented their method for any general basis, This approach was extended to a logistic FLM by Ratcliffe et al. [145]. Transforming 2D image data into 1D variance functions, Ogden et al. [130] fitted an FLM using Fourier bases for both $X_i(t)$ and $\beta(t)$, applying regularization through a roughness penalty.

Sparsity in basis coefficients was leveraged for regularization by Zhu et al. [205] and Lee et al. [94]. Zhu et al. [205] modeled dichotomous outcomes using a Bayesian probit model with general orthogonal basis functions for $X_i(t)$ and $\beta(t)$, employing sparsity-inducing SSVS. Lee and Park [93] introduced an FLM for Gaussian responses with a common general basis transform for $X_i(t)$ and $\beta(t)$, utilizing sparsity penalties such as LASSO, adaptive LASSO, or SCAD. Similar applications in Fan et al. [44] used group-LASSO for variable selection across multiple functional predictors.

James et al. [75] introduced an interpretable FLR, encouraging sparsity in $\beta(t)$ for interpretability through L_1 penalization. Piecewise constant basis functions for $\beta(t)$ were employed, extendable to other high-dimensional basis expansions like splines, Fourier, and wavelets. A Bayesian method for predicting a Gaussian response from an imaging predictor was proposed by Goldsmith et al. [53], using an Ising prior and Markov random field prior to encouraging clustering and borrowing strength among significant regions of the coefficient surface.

2.1.3 Nonlinear functional predictor models

Besides assuming the functional linear structures, several approaches have been proposed to handle nonlinear functional predictor regression.

Li and Marx [95] extended the PSR method of Marx and Eilers [113] to include general additive polynomial terms such as $\int \{X_i(t)\}^2 B_2(t)dt$. Yao and Müller [194] further allowed for full quadratic functional predictors, represented by the equation:

$$Y_i = \alpha + \int X_i(t)B_1(t)dt + \iint X_i(t)X_i(s)B_2(t,s)dtds + \varepsilon_i.$$

This provides greater flexibility compared to the PSR method described by Li and Marx [95], which only allows for diagonal cross products in the second term. They used PC decompositions to estimate empirical basis functions for $X_i(t)$ and used these as basis functions for $X_i(t), B_1(t)$, and $B_2(t,s)$, with regularization through truncation. Their approach can be generalized to full polynomials of any order.

Yang et al. [193] extended the spectrogram-based method of Holan et al. [68] to the penalized quadratic regression setting, using a Bayesian modeling approach with SSVS for regularization across the PC dimensions.

In 2005, James and Silverman introduced the functional adaptive model estimation (FAME) method, which is the first nonparametric extension of the FLM. The model, represented by

$$g\left\{E\left(Y_i\right)\right\} = \alpha + \sum_{k=1}^{K} f_k \left\{ \int X_i(t) B_k(t) dt \right\},$$

extends projection pursuit regression to functional predictors and exponential family outcomes, here f_k is a smooth function of unspecified form. They used natural cubic splines to represent $X_i(t), B_k(t)$, and $f_k(\cdot)$ with the roughness penalties. FAME was later extended to incorporate multiple functional predictors.

There is fruitful literature on nonlinearities in functional predictor regression using single-index models, which involve the function $f\left(\int X_i(t)\beta(t)dt\right)$ for some smooth $f(\cdot)$. Li et al. [99] used single-index models to model interactions between scalar and functional predictors, and further employed complete orthogonal basis to represent $X_i(t)$. Marx et al. [115] used single-index models to extend the PSR of Marx and Eilers [113] in the setting of 2D functional predictors, using tensor B-splines for the predictor surface and 1D B-splines for the index function, with regularization through P-spline penalties. Fan et al. [44] allowed separate additive single-index terms for p functional predictors, using a common orthogonal basis for $X_{ij}(t)$ and $B_j(t)$, and another basis for $g_j(\cdot)$ for $j = 1, \ldots, p$. Truncation for penalization and group-LASSO were conducted across the functional predictors for variable selection.

Müller and Yao [124] introduced functional additive models (FAMs), which extend the fPC regression approach of Yao et al. [196] to nonlinear models that involve additive nonparametric functions of the PC scores, using the model

$$Y_i = \alpha + \sum_{k=1}^{K_x} f_k(Y_{ik}^*) + \varepsilon_i,$$

where Y_{ik}^* represents the scores for a set of truncated principal components (PCs) and

f_k is a smoothed function using a kernel approach. Zhang [201] proposed an alternative approach that utilized COSSO regularization, which is a form of L1 penalization, instead of truncating the PCs. This method allowed dimensions that are important for predicting Y_i even though they may explain less variability in $X_i(t)$.

McLean et al. [117] introduced a functional generalized additive model (FGAM) that extends the traditional generalized additive models (GAMs) to the generalized functional predictor regression $g\{E(Y_i)\} = \alpha + \int f\{X_i(t), t\}\, dt + \varepsilon_i(t)$ for noise-free functions observed on a common grid t. g is a general link function for exponential families and $f(x, t)$ is a smooth functional additive regression surface. The surface was parameterized using tensor B-splines and regularized using P-spline-type penalization. The model was fit using penalized iteratively WLS with generalized cross validation (GCV) to estimate the smoothing parameters. McLean et al. [116] also proposed a Bayesian FGAM for sparsely observed functions on irregular grids. They handled the measurement error and unevenly spaced $X_i(t)$ by using PC decompositions, updating the PC scores within the MCMC, and using tensor product B-splines with random walk penalties on the coefficients.

2.2 Reproducing Kernel Hilbert Space (RKHS)

The Reproducing Kernel Hilbert Space (RKHS) is a particular type of Hilbert space that consists of functions with reproducing kernels, as described in Berlinet and Thomas-Agnan [8]. Its characteristics and theories were developed by Aizerman et al. [3] and Aronszajn [4] successively. Although the RKHS was initially studied in pure mathematics, it gained significance in machine learning with the introduction of kernel Support Vector Machine (SVM) by Boser et al. [11] and Vapnik [174]. Eigenfunctions were further developed for the eigenvalue problem of operators and functions, as discussed in Williams and Seeger [186], and were utilized in machine learning Bengio et al. [7], quantum mechanics, and other domains in physics Kusse and Westwig [90]. The connection between these developments and RKHS resides in the utilization of

the weighted inner product in Hilbert space, which is a characteristic of both the RKHS and the development of eigenfunctions as described by Williams and Seeger [186].

Definition 2.1 ([147]) *A Hilbert space $\mathcal{H}$ is an inner product space that is a complete metric space with respect to the norm or distance function induced by the inner product.*

Remark 2.1 *The Hilbert space, often high dimensional, generalizes the Euclidean space to a finite or infinite dimensional space. It is a special case of Banach space equipped with a norm defined using an inner product notion. All Hilbert spaces are Banach spaces but the converse is not true.*

Definition 2.2 ([4, 8]) *A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ with a reproducing kernel $k : \mathcal{X}^2 \to \mathbb{R}$ where $k(\boldsymbol{x}, .) \in \mathcal{H}$ and $f(x) = \langle k(x,), f \rangle$.*

To illustrate, consider the kernel function $k(\boldsymbol{x}, \boldsymbol{y})$ which is a function of two variables. Suppose, for n points, we fix one of the variables to have $k(\boldsymbol{x}_1, \boldsymbol{y}), k(\boldsymbol{x}_2, \boldsymbol{y}), \ldots, k(\boldsymbol{x}_n, \boldsymbol{y})$. These are all functions of the variable y. RKHS is a function space which is the set of all possible linear combinations of these functions [3, 84, 119]:

$$\mathcal{H} := \left\{ f(.) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, .) \right\} = \left\{ f(.) = \sum_{i=1}^{n} \alpha_i k_{\boldsymbol{x}_i}(.) \right\} \tag{2.5}$$

where we let $k_{\boldsymbol{x}}(.) := k(\boldsymbol{x}, .)$. According to Eq.(2.5), every function in the RKHS can be written as a linear combination of kernel functions, thus they form the bases of an RKHS. Consider two functions in this space represented as $f = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{y})$ and $g = \sum_{j=1}^{n} \beta_j k(\boldsymbol{x}, \boldsymbol{y}_j)$. Hence, the inner product in RKHS is calculated as:

$$\langle f, g\rangle_k \overset{(2.5)}{=} \left\langle \sum_{i=1}^{n} \alpha_i k\left(\boldsymbol{x}_i, .\right), \sum_{j=1}^{n} \beta_j k\left(\boldsymbol{y}_j, .\right) \right\rangle_k$$
$$\overset{(a)}{=} \left\langle \sum_{i=1}^{n} \alpha_i k\left(\boldsymbol{x}_i, .\right), \sum_{j=1}^{n} \beta_j k\left(., \boldsymbol{y}_j\right) \right\rangle_k$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right) \tag{2.6}$$

since the kernel is symmetric. Hence, the norm in RKHS is calculated as: $\|f\|_k := \sqrt{\langle f, f\rangle_k}$.

The subscript of norm and inner product in RKHS has various notations in the research papers. Some most famous notations are $\langle f, g\rangle_k, \langle f, g\rangle_{\mathcal{H}}, \langle f, g\rangle_{\mathcal{H}_k}$ where $\mathcal{H}_k$ denotes the Hilbert space associated with kernel k.

Definition 2.3 (L_p Space) *Consider a function f with domain $[a, b]$. For $p > 0$, let the L_p norm be defined as:*

$$\|f\|_p := \left(\int |f(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{\frac{1}{p}}.$$

The L_p space is defined as the set of functions with bounded L_p norm:

$$L_p(a, b) := \{f : [a, b] \to \mathbb{R} \mid \|f\|_p < \infty\}.$$

Definition 2.4 (Sobolev Space [150]) *A Sobolev space is a vector space of functions equipped with L_p norms and derivatives:*

$$\mathcal{W}_{m,p} := \{f \in L_p(0, 1) \mid D^m f \in L_p(0, 1)\},$$

where $D^m f$ denotes the m-th order derivative.

The Sobolev spaces are RKHS with some specific kernels [129].

Remark 2.2 *Given a kernel, the corresponding RKHS is unique (up to isometric isomorphisms). Given an RKHS, the corresponding kernel is unique. In other words, each kernel generates a new RKHS.*

RKHS is a space of functions and not a space of vectors. Namely, the basis vectors of RKHS are basis functions named eigenfunctions. Because the RKHS is a space of functions rather than a space of vectors, we usually do not know the exact location of pulled points to the RKHS but we know the relation of them as a function.

2.2.1 Reproducing property

Consider only one component in Eq. (2.6) for g to have $g(\boldsymbol{x}) = \sum_{j=1}^{n} \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}) = \beta k(\boldsymbol{x}, \boldsymbol{x})$ where we take $\beta = 1$ to have $g(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) = k_{\boldsymbol{x}}(.)$. That is to say, assume the function $g(x)$ is a kernel in the RKHS space. Also consider the function $f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$ in the space. According to Eq. (2.6), the inner product of these functions is:

$$\begin{aligned} \langle f(\boldsymbol{x}), g(\boldsymbol{x}) \rangle_k &= \langle f, k_{\boldsymbol{x}}(.) \rangle_k \\ &= \left\langle \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}) \right\rangle_k \\ &= \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) \overset{(a)}{=} f(\boldsymbol{x}), \end{aligned} \tag{2.7}$$

As Eq. (2.7) indicates, the function f can be reproduced from the inner product of that function with one of the kernels in the space. This shows the reproducing property of the RKHS space. A special case of Eq. (2.7) is $\langle k_{\boldsymbol{x}}, k_{\boldsymbol{x}} \rangle_k = k(\boldsymbol{x}, \boldsymbol{x})$.

2.2.2 Representation in RKHS

We provide a proof for Eq. (2.6) and explain why that equation defines the RKHS.

Theorem 2.1 (Representer Theorem [85, 153, 198]) *For a set of data $\mathcal{X} = \{x_i\}_{i=1}^{n}$, consider a RKHS $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ with kernel function k. For any function $\ell : \mathbb{R}^2 \to \mathbb{R}$ (usually called the loss function), consider the optimization problem:*

$$f^* \in \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell\left(f(\boldsymbol{x}_i), \boldsymbol{y}_i\right) + \eta \Omega\left(\|f\|_k\right),$$

where $\eta \geq 0$ is the regularization parameter and $\Omega\left(\|f\|_k\right)$ is a penalty term such as $\|f\|_k^2$. The solution of this optimization can be expressed as:

$$f^* = \sum_{i=1}^{n} \alpha_i k\left(\boldsymbol{x}_i, .\right) = \sum_{i=1}^{n} \alpha_i k_{\boldsymbol{x}_i}(.).$$

A proof of Theorem 2.1 can be found in Ghojogh et al. [49] and Rudin [153].

RKHS provides valid estimations to functions, which can be extended to FLR (2.1). Assume that the slope function $\beta_0(\cdot)$ resides in an RKHS $\mathcal{H} = \mathcal{H}(k)$, a subspace of square-integrable functions with the domain $\mathcal{I}$, equipped with a reproducing kernel k. Typically, we assume $\mathcal{H}(k)$ to be the second-order Sobolev space. Let without loss of generality that $\mathcal{I} = [0, 1]$, one squared norm [13] that makes $\mathcal{W}_{r,2}$ an RKHS is

$$\sum_{j=0}^{r-1} \left(\int_{\mathcal{I}} \beta^{(j)}(t)\ \mathrm{d}t \right)^2 + \int_{\mathcal{I}} \left(\beta^{(r)}(t)\right)^2\ \mathrm{d}t. \tag{2.8}$$

In the context of FLR, the penalty functional J on the slope function β can be conveniently defined as the squared norm or semi-norm associated with $\mathcal{H}$, we adopt the typical penalty given by $J(\beta) = \int_0^1 [\beta^{(r)}(t)]^2\ \mathrm{d}t$. Another setting of particular interest is $\mathcal{I} = [0, 1]^2$ which naturally occurs when X represents an image [17]. A popular choice in this setting is the thin plate spline where J is defined as

$$J(\beta) = \int_0^1 \int_0^1 \left[\left(\frac{\partial^2 \beta}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 \beta}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 \beta}{\partial x_2^2} \right)^2 \right] dx_1 dx_2$$

and (x_1, x_2) are the arguments of bivariate function β. Other examples of $\mathcal{I}$ include $\mathcal{I} = \{1, 2, \ldots, p\}$ for some positive integer p, and unit sphere in an Euclidean space among others. The readers are referred to Wahba [175] for common choices of $\mathcal{H}$ and J in these as well as other contexts.

The null space of functional J, defined as $\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$, forms a finite-dimensional linear subspace of $\mathcal{H}$ with some orthonormal basis $(\xi_1, \xi_2, \ldots, \xi_M)$, where $M = \dim(\mathcal{H}_0)$. The orthogonal complement $\mathcal{H}_1$ of the null space $\mathcal{H}_0$ is such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. It can be shown that $\mathcal{H}_1$ also forms an RKHS with the same inner product as $\mathcal{H}$, but restricted to $\mathcal{H}_1$. More generally, for any $\beta \in \mathcal{H}$, there exists

$\beta_1 \in \mathcal{H}_0$ and $\beta_2 \in \mathcal{H}_1$ such that the decomposition $\beta = \beta_1 + \beta_2$ is unique [57, 127]. Let k be the reproducing kernel of $\mathcal{H}_1$ such that $J(\beta_2) = \|\beta_2\|_{\mathcal{H}}^2 = \|\beta\|_k^2$, defined as the RKHS norm of β. Consequently, as we demonstrate in lemma 2.2, the functional slope coefficient β_0 can be represented by a finite-dimensional form through basis $(\xi_1, \xi_2, \ldots, \xi_M)$ as well as reproducing kernel k.

Lemma 2.2 *Let $(\xi_1, \ldots, \xi_M)$ be a basis of $\mathcal{H}_0$. There exist vectors $e = (e_1, \ldots, e_M)^\top$ and $c = (c_1, \ldots, c_n)^\top$ allowing the solution $\widehat{\beta}_n$ to the problem in Eq. 2.1 to be expressed as*

$$\widehat{\beta}_n(t) = \sum_{i=1}^{M} e_i \xi_i(t) + \sum_{j=1}^{n} c_j \int_{\mathcal{I}} k(s,t) X_j(t) \, \mathrm{d}s. \tag{2.9}$$

Lemma 2.2 is a generalization of the renowned representer theorem 2.1 for smoothing splines [175]. Although the minimization over $\widehat{\beta}_n$ in Equation (1.1) is taken over an infinite-dimensional space $\mathcal{H}$, the above result implies that the solution lies in a finite-dimensional subspace. Thus, it suffices to estimate the coefficients e and c in Equation (2.9).

2.2.3 Common kernel functions

There exist many different kernel functions that are widely used in machine learning [49, 152]. In the following, we list some of the most well-known kernels.

- Linear Kernel: Linear kernel is the simplest kernel to calculate the inner product of points: $k(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \boldsymbol{y}$, and data are not pulled to any other space in linear kernel but in the input space, using a linear kernel may or may not be equivalent to non-kernelized method.

- Radial Basis Function (RBF) or Gaussian Kernel:

$$k(\boldsymbol{x}, \boldsymbol{y}) := \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2}{\sigma^2}\right)$$

  where $\gamma := 1/\sigma^2$ and σ^2 is the variance of kernel. A proper value for this parameter is $\gamma = 1/d$ where d is the dimensionality of data. RBF kernel has

a scaled Gaussian (or normal) distribution where the normalization factor of distribution is usually ignored. Hence, it is also called the Gaussian kernel. It has also been widely used in RBF networks [131] and kernel density estimation [162].

- Laplacian Kernel: The Laplacian kernel (the Laplace kernel) is similar to the RBF kernel but with L_1 norm rather than squared L_2 norm. The Laplacian kernel is:

$$k(\boldsymbol{x}, \boldsymbol{y}) := \exp\left(-\gamma \|x - y\|_1\right) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_1}{\tau^2}\right)$$

  where $\|\boldsymbol{x} - \boldsymbol{y}\|_1$ is also called the Manhattan distance. Under some specific situations, the Laplacian kernel has been found to perform better than Gaussian kernel [154]. This makes sense thanks to the sparsity principal of L_1 norm [65]. However, the computation and derivative of L_1 norm is more difficult than L_2 norm.

- Polynomial Kernel: Polynomial kernel applies a polynomial function with degree δ (a positive integer) on inner product of points:

$$k(\boldsymbol{x}, \boldsymbol{y}) := \left(\gamma \boldsymbol{x}^\top \boldsymbol{y} + c\right)^d,$$

  where $\gamma > 0$ is the slope and c is the intercept.

Chapter 3

Reproducing Kernel Based Functional Linear Expectile Regression

3.1 Introduction

Functional data have grown ubiquitous in medical data analysis, biology, and image and signal processing, among many other fields [67, 96, 143, 197]. While intrinsically functional, this type of data is almost always observed discretely over a grid, where the number of grid points is often larger than the number of observations. Due to the spatial or temporal nature of this grid, observations at nearby grid points are often highly correlated. Specialized techniques are consequently crucial in the proper analysis of functional data.

Traditional analytic approaches typically assume that errors are independent and identically distributed (i.i.d.) with a symmetric and homoscedastic density [59]. These assumptions cannot be guaranteed in practice, particularly in high-dimensional settings [101]. As typical examples, consider modelling meteorological outcomes (e.g., from the Canadian weather dataset, as in [143] or [16]) or clinical outcomes (e.g., Mini Mental State Examination (MMSE) scores, a clinical survey-based measure used to quantify Alzheimer's disease severity, as in [72]). In these and many other settings, there is no guarantee that the conditional response distribution will be symmetric, much less Gaussian. It is more often the case that stochastic error terms are heteroscedastic and that the conditional response distribution is highly skewed or heavy-

tailed. As noted in [126], heteroscedastic errors lead to inefficient or inconsistent parameter and covariance estimation.

From a practical standpoint, in regression settings where error is heteroscedastic or asymmetric, several estimators may be required for a satisfactory picture of the relationship between the response variable and model predictors. Each of these estimators may speak to a different notion of the location of the conditional response distribution, such as its different quantiles levels. Neuroimaging data analysis is one such setting where responses at multiple extreme levels, representing outlying or abnormal cases, are of more practical interest than, say, a single conditional mean. [134] further emphasizes the particular need for functional tools not focused solely on conditional mean estimation in neuroimaging data analysis and more general fields of application.

Motivated by the dependence of traditional coefficient estimators on error homoscedasticity and symmetry assumptions, [126] first introduced expectile regression, also called asymmetric least squares regression [59, 176]. Expectiles are analogous to quantiles and can similarly be computed for a random variable Y at any level $\tau \in (0, 1)$, but are determined by the tail expectations rather than the tail probabilities of a distribution. While quantile regression has a strong intuitive appeal, well-studied robustness properties, and broad applications in a variety of research fields [86], [126] motivates expectiles by pointing out three major drawbacks to quantile estimators: their nondifferentiability, their relative inefficiency for near-Gaussian error distributions, and the difficulty inherent in computing their covariance.

It is then a natural development to consider expectile regression with functional predictors (i.e., in a "scalar-on-function" framework). In this chapter, we are concerned with the model

$$Y = \int_{\mathcal{T}} X(t)\beta_0(t)\ \mathrm{d}t + \varepsilon, \tag{3.1}$$

where Y is a scalar response, $X : \mathcal{T} \to \mathbb{R}$ is a square-integrable stochastic process,

and $\beta_0 : \mathcal{T} \to \mathbb{R}$ is the slope function. We assume that the domain $\mathcal{T}$ is a compact subset of a Euclidean space.

Most recent approaches to functional linear regression are based on functional principal components analysis (FPCA) [61]. FPCA ultimately relies on an efficient representation of β_0 in terms of the leading functional principal components of X [16]. However, these functional principal components might not form an appropriate basis to express β_0 or might have little predictive power. Consequently, FPCA-based methods might not perform well. In practice, this phenomenon has been observed for functional data in the Canadian weather dataset [16, 143], in the Alzheimer's Disease Neuroimaging Initiative (ADNI) data analyzed later, and more generally, in principal components regression and singular value decomposition methods for linear inverse problems [37]. Numerous other works have considered the model in Equation (3.1) using FPCA-based approaches [15, 31, 60, 61, 76, 102, 159].

In this chapter, we study instead the functional linear expectile regression model from the perspective of a reproducing kernel Hilbert space (RKHS): we assume that the slope function β_0 resides in an RKHS $\mathcal{H}(K)$. In this more general framework, the functional covariance operator C and the reproducing kernel K of the RKHS are not required to be related. This assumption differs from the implicit requirements in FPCA-based frameworks that the ordered eigenfunctions of K and C perfectly coincide. FPCA-based approaches further assume that the slope function β_0 can be efficiently represented in terms of leading functional principal components [16, 198]. RKHS-based estimators, such as those proposed in this chapter, circumvent this restriction.

As illustrated in [198], the eigenstructure of the RKHS plays an important role in estimation and prediction, making RKHS-based methods more difficult to implement. To our knowledge, the literature on RKHS-based approaches to functional data analysis is limited. [23] considered a joint asymptotic framework for studying semi-nonparametric regression models where (finite-dimensional) Euclidean parame-

ters and (infinite-dimensional) functional parameters are both of interest: the authors derived convergence rates for estimators of both. [141] studied functional Cox models with right-censored data in the presence of both functional and scalar covariates in an RKHS framework. Notably, the authors proved that their functional coefficient estimator achieves the minimax optimal rate of convergence in penalized log partial likelihood settings. [100] derived various asymptotic results regarding kernel quantile regression (KQR) and proposed an efficient algorithm to compute entire KQR solution paths.

In this chapter, we propose a regularized estimator for the functional linear expectile regression model in an RKHS framework. Specifically, unlike existing FPCA-based approaches to expectile regression, we use the reproducing kernel to approximate functional effects and capture local features. Theoretically (when the eigenfunctions of K and C agree) and empirically (regardless of this agreement) we find that our estimators exhibit stronger convergence rates relative to FPCA-based estimators. We further incorporate shrinkage penalties as a means to improve estimate interpretability and generalizability for prediction. We derive upper and lower bounds for minimax convergence in prediction error and establish minimax convergence rate optimality for our proposed estimator. We demonstrate that RKHS-based methods simplify functional coefficient estimate regularization (e.g., via smoothness, sparsity, or Tikhonov penalties) and allow model estimation to be formulated as a convex optimization problem. Our RKHS-based estimator can thus be efficiently computed: the alternating direction method of multipliers (ADMM) algorithm we apply makes our procedure simple to implement and allows us to incorporate existing computational techniques for smoothing splines.

The remainder of the article is organized as follows. In Section 2, we discuss expectile regression and RKHSs and establish the minimax optimality of our proposed estimator. In Section 3, we reformulate model estimation as a convex optimization problem and derive an ADMM iterative update scheme using a finite-dimensional

representation of the slope function obtained via the representer theorem. Section 4 investigates finite-sample performance through simulation studies and a real-world data analysis, the latter using data from the ADNI [72]. A subsequent appendix contains technical proofs of this article's main results.

As notation to be used throughout this article, let $\|\cdot\|_2$ denote the Euclidean L_2 norm. For two positive real sequences $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$, we write $a_k \asymp b_k$ to indicate that the sequence of ratios $(a_k/b_k)_{k\in\mathbb{N}}$ is bounded away from zero and infinity.

3.2 Theoretical Properties

We first introduce functional linear expectile regression, our proposed estimator, and the setting where $\beta_0 \in \mathcal{H}(K)$. Following this, we derive upper and lower bounds for the minimax rate of convergence in prediction error and establish the minimax optimality of our proposed estimator.

3.2.1 Expectiles and functional linear expectile regression

Let Y be a random variable with a distribution function F and a finite mean. The τth expectile $\mu_\tau = \mu_\tau(F)$ of Y, as defined by [126], is

$$\mu_\tau(F) = \arg\min_{\eta\in\mathbb{R}} E_Y\ r_\tau(y-\eta),$$

for $\tau \in (0,1)$, where $r_\tau(y-\eta) = |\tau - \mathbb{1}(y<\eta)|(y-\eta)^2$.

Expectiles share many desirable characteristics of quantiles and various additional computational advantages [126]. [78] showed that the expectiles of a distribution F are the quantiles of a distribution G defined explicitly as

$$G(y) = \frac{P(y) - yF(y)}{2(P(y) - yF(y)) + (y-\mu)},$$

where $P(y) = \int_{-\infty}^{y} x\ \mathrm{d}F(x)$ and $\mu = \int_{-\infty}^{\infty} x\ \mathrm{d}F(x)$.

As a generalization of ordinary mean regression, expectile regression is known to be statistically more efficient than quantile regression when standard assumptions such

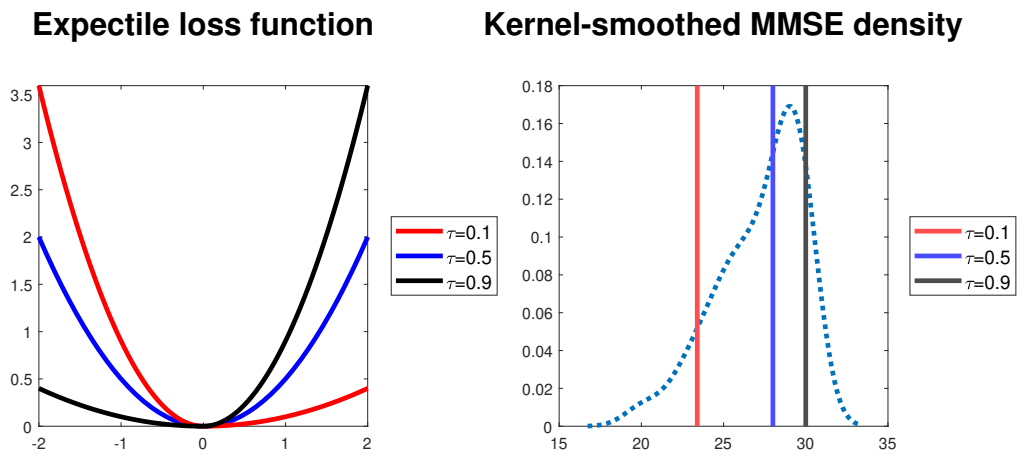


Figure 3.1: (Left) The expectile loss function for $\tau = 0.1, 0.5, 0.9$ in red, blue, and black, respectively. (Right) Kernel-smoothed estimate of the MMSE score density function (dotted blue) from the ADNI dataset. The corresponding expectiles at $\tau = 0.1, 0.5, 0.9$ are indicated in solid red, blue, and black, respectively.

as error homoscedasticity are not severely violated [102]. Unlike quantile regression, expectile regression uses a smooth loss function which, in terms of general computation, is considerably easier to optimize [59]. [69] and [89] explored the asymptotic properties of sample expectiles and established their uniform consistency under the assumption of a finite mean. Unlike quantiles, expectiles are also guaranteed to be unique under this assumption. The asymptotic normality of the sample expectile estimator follows directly with the additional assumption of a finite second moment. Similar to quantiles, expectiles characterize and give more insight into a distribution of interest.

Figure 3.1 illustrates the expectile loss function at $\tau = 0.1, 0.5, 0.9$. When $\tau < 0.5$, the cost of a positive error is lower than that of a negative one, encouraging a smaller expectile μ_τ. Larger expectiles are correspondingly encouraged when $\tau > 0.5$. At $\tau = 0.5$, the loss $r_{0.5}$ is equivalent to the least squares loss and recovers the mean of the distribution. In settings where the distribution of Y is highly skewed rather than symmetric, τ can be chosen to obtain a more desirable location estimate. This is illustrated in Figure 3.1 using MMSE data from the ADNI.

In this part, we are primarily interested in establishing the convergence properties of our proposed regularized sample estimator for β_0 in the functional linear expectile regression model of Equation 3.1,

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} r_\tau \left(y_i - \int_{\mathcal{T}} x_i(t)\beta(t)\ \mathrm{d}t \right) + \lambda J(\beta), \tag{3.2}$$

where $\{(x_i, y_i) : i = 1, \ldots, n\}$ is a set of observed training data, J is a penalty function assessing the "plausibility" of a candidate β, and $\lambda \geq 0$ is a tuning parameter controlling the strength of the penalty J. For convenience, we suppress notation indicating implicit dependence on τ.

3.2.2 RKHS

We assume that the slope function β_0 resides in an RKHS $\mathcal{H} = \mathcal{H}(K)$, a subspace of square-integrable functions with the domain $\mathcal{T}$, equipped with a reproducing kernel K. The canonical example of $\mathcal{H}(K)$ is a Sobolev space: assuming, without loss of generality, that $\mathcal{T} = [0, 1]$, the Sobolev space of order r [55] can be defined as

$$\begin{aligned} \mathcal{W}_2^r = \mathcal{W}_2^r([0,1]) = \{\beta : [0,1] \to \mathbb{R} : \beta, \beta^{(1)}, \ldots, \beta^{(r-1)} \\ \text{are absolutely continuous and } \beta^{(r)} \in \mathcal{L}_2\}. \end{aligned}$$

One squared norm that will make $\mathcal{W}_2^r$ an RKHS [13] is $\sum_{j=0}^{r-1} \left\{ \int_{\mathcal{T}} \beta^{(j)}(t)\ \mathrm{d}t \right\}^2 + \int_{\mathcal{T}} \{\beta^{(r)}(t)\}^2\ \mathrm{d}t$. The penalty functional J on the slope function β can be conveniently defined as the squared norm or semi-norm associated with $\mathcal{H}$ [17]: one possible choice is $J(\beta) = \int_0^1 [\beta^{(r)}(t)]^2\ \mathrm{d}t$. The null space of J, defined as $\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$, forms a finite-dimensional linear subspace of $\mathcal{H}$ with some orthonormal basis $(\xi_1, \xi_2, \ldots, \xi_M)$, where $M = \dim(\mathcal{H}_0)$. The orthogonal complement $\mathcal{H}_1$ of the null space $\mathcal{H}_0$ is such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. It can be shown that $\mathcal{H}_1$ also forms an RKHS with the same inner product as $\mathcal{H}$, but restricted to $\mathcal{H}_1$. More generally, for any $\beta \in \mathcal{H}$, there exists $\beta_1 \in \mathcal{H}_0$ and $\beta_2 \in \mathcal{H}_1$ such that the decomposition $\beta = \beta_1 + \beta_2$ is unique [57, 127]. Let K be the reproducing kernel of $\mathcal{H}_1$ such that

$J(\beta_2) = \|\beta_2\|_{\mathcal{H}}^2 = \|\beta\|_K^2$, defined as the RKHS norm of β. Consequently, as we will demonstrate, we can find a finite-dimensional representation for the functional slope coefficient β_0.

We next consider the two kernels crucial to the estimation process. First, recalling that $\mathcal{T} \subset \mathbb{R}$ is a compact set, a reproducing kernel $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a real, symmetric, square-integrable, nonnegative-definite function. There is a one-to-one correspondence between a reproducing kernel K and an RKHS $\mathcal{H}(K)$. Mercer's theorem implies that K admits the spectral decomposition $K(s,t) = \sum_{k=1}^{\infty} \varrho_k \varphi_k(s)\varphi_k(t)$, where the eigenvalues $(\varrho_k)_{k\in\mathbb{N}}$ are in nonincreasing order and $(\varphi_k)_{k\in\mathbb{N}}$ are the corresponding eigenfunctions.

For any real, square-integrable, semidefinite function R, define $L_R : \mathcal{L}_2 \to \mathcal{L}_2$ as the linear integral operator $L_R(f)(\cdot) = \langle R(s,t), f\rangle_{\mathcal{L}^2(\mathcal{T})} = \int_{\mathcal{T}} R(s,\cdot)f(s)\,\mathrm{d}s$. By the spectral theorem, there exists a sequence of orthonormal eigenfunctions $(\psi_k^R)_{k\in\mathbb{N}}$ and a corresponding sequence of nonincreasing eigenvalues $(\theta_k^R)_{k\in\mathbb{N}}$ such that $R(s,t) = \sum_{k\in\mathbb{N}} \theta_k^R \psi_k^R(s)\psi_k^R(t)$ for all $s,t \in \mathcal{T}$, and $L_R(\psi_k^R) = \theta_k^R \psi_k^R$ for $k \in \mathbb{N}$. Additionally, for all $s,t \in \mathcal{T}$, $R^{1/2}(s,t) = \sum_{k\in\mathbb{N}} \sqrt{\theta_k^R}\psi_k^R(s)\psi_k^R(t)$. We say that two linear operators are aligned if they share the same ordered (i.e., with corresponding eigenvalues in nonincreasing order) sequence of eigenfunctions.

Let $L_{R^{1/2}}$ be the linear operator defined by $L_{R^{1/2}}(\psi_k^R) = \sqrt{\theta_k^R}\psi_k^R$. It is clear that $L_{R^{1/2}} = (L_R)^{1/2}$. Further defining $(R_1R_2)(s,t) = \int_{\mathcal{T}} R_1(s,u)R_2(u,t)\,\mathrm{d}u$, it follows that $L_{R_1R_2} = L_{R_1} \circ L_{R_2} = L_{R_2} \circ L_{R_1}$.

With the previous results in mind, consider the covariance kernel $C : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ for X, defined as $C(s,t) = E\left([X(s) - EX(s)]\,[X(t) - EX(t)]\right)$. Of course, we require that the covariance kernel C be continuous and square-integrable over $\mathcal{T}\times\mathcal{T}$. Similar to K, C admits the spectral decomposition $C(s,t) = \sum_{k=1}^{\infty} \mu_k \phi_k(s)\phi_k(t)$. The two eigenfunction sequences $(\varphi_k)_{k\in\mathbb{N}}$ and $(\phi_k)_{k\in\mathbb{N}}$ are different in general. However, under certain conditions, K and C can be simultaneously diagonalized [25].

Using the eigenstructures of the reproducing and covariance kernels K and C, we

can define the linear operator $L_{K^{1/2}CK^{1/2}}$ in a compositional fashion as $L_{K^{1/2}CK^{1/2}} = L_{K^{1/2}} \circ L_C \circ L_{K^{1/2}}$. By the spectral theorem, $K^{1/2}CK^{1/2}$ has the spectral decomposition $K^{1/2}CK^{1/2}(s,t) = \sum_{k=1}^{\infty} \nu_k \zeta_k(s)\zeta_k(t)$, where the sequence of eigenvalues $(\nu_k)_{k\in\mathbb{N}}$ is arranged in nonincreasing order and $(\zeta_k)_{k\in\mathbb{N}}$ is the corresponding sequence of orthonormal eigenfunctions. Obviously, the eigenvalues $(\nu_k)_{k\in\mathbb{N}}$ are determined by the eigenvalues of both K and C and the alignment of their respective eigenfunctions. We will eventually show that the convergence rate of our proposed estimator is related to the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$.

Before discussing estimation of the functional coefficient β_0 over $\mathcal{H}(K)$, we impose two basic assumptions on the reproducing and covariance kernels, whose eigenstructures determine the optimal convergence rate.

(A1) The eigenvalues of $K^{1/2}CK^{1/2}$ satisfy $\nu_k \asymp k^{-2r}$ for some $r > 0$.

(A2) For any square-integrable function f,

$$E\left[\int_{\mathcal{T}} [X(t) - EX(t)]\, f(t)\, \mathrm{d}t\right]^4 \leq c\left(E\left[\int_{\mathcal{T}} [X(t) - EX(t)]\, f(t)\, \mathrm{d}t\right]^2\right)^2$$

for some constant $c > 0$.

Assumption A1 pertains to the decay rate of ν_k. As already discussed, this rate is determined by the eigenstructures of the kernels K and C, specifically, their individual eigenvalue decay rates and the alignment between their eigenfunctions. The eigenvalues of the covariance kernel C obey $\mu_k \asymp k^{-2r^C}$ if the Sacks–Ylvisaker condition of order $r^C - 1$ is satisfied for some integer $r^C \geq 1$ [151, 198]. As an example, the Ornstein-Uhlenbeck covariance kernel $C(s,t) = \exp(-|s-t|)$ has $r^C = 1$. For Sobolev spaces, various covariance functions are known to satisfy the Sacks–Ylvisaker condition [151]. Concerning the eigenvalue decay rate of the kernel K, if $\mathcal{H}$ is the r^Kth order Sobolev space $\mathcal{W}_2^{r^K}$, it is known that $\varrho_k \asymp k^{-2r^K}$ [120].

When K and C are aligned, i.e., when they share a common ordered eigenfunction set so that $\phi_k = \varphi_k$ for $k \in \mathbb{N}$ [16], it follows that $r = r^C + r^K$ in Assumption A1.

However, if K and C are not aligned, then the eigenvalues of the two operators alone cannot determine the order r. For example, the eigenvalues for the Sobolev class $\mathcal{W}_2^r$ for $r > 1/2$ follow a polynomial decay rate.

Assumption A2 restricts the fourth moment of the linear functional $\int_{\mathcal{T}} X(t)f(t)\ \mathrm{d}t$, ensuring bounded kurtosis. When X is a Gaussian process, for example, Assumption A2 is satisfied with $c = 3$.

3.2.3 Minimax convergence properties

We take $\mathcal{H}(K) = \mathcal{W}_2^2$ and define the penalty function as $J(\beta) = \int_{\mathcal{T}} [\beta''(t)]^2\ \mathrm{d}t = \|\beta\|_K^2$. Consequently, $\mathcal{H}_0$ is the linear space spanned by $\xi_1(t) = 1$ and $\xi_2(t) = t$.

The accuracy of $\hat{\beta}_n$ can be measured via the squared RKHS norm associated with the covariance kernel C [198], as

$$\left\|\hat{\beta}_n - \beta_0\right\|_C^2 = E_{X^*}\left(\int X^*(t)\hat{\beta}_n(t)\ \mathrm{d}t - \int X^*(t)\beta_0(t)\ \mathrm{d}t\right)^2,$$

where X^* is an independent copy of X and the expectation on the right-hand side is taken over X^*. The above quantity measures the mean squared prediction error for a random, future observation of X.

Theorem 3.1 (Minimax lower bound) *Under Assumption A1,*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\hat{\beta}_n}\sup_{\beta_0\in\mathcal{H}(K)}\mathbb{P}_{\beta_0}\left\{\|\hat{\beta}_n - \beta_0\|_C \geq an^{-\frac{2r}{2r+1}}\right\} = 1, \tag{3.3}$$

where the infimum is taken over all possible estimators $\hat{\beta}_n$ computed from the training data.

Theorem 3.2 (Minimax upper bound) *Under Assumptions A1 and A2,*

$$\lim_{A\to 0}\limsup_{n\to\infty}\sup_{\beta_0\in\mathcal{H}(K)}\mathbb{P}_{\beta_0}\left\{\|\hat{\beta}_n - \beta_0\|_C \geq An^{-\frac{2r}{2r+1}}\right\} = 0. \tag{3.4}$$

provided that the tuning parameter satisfies $\lambda \asymp n^{-2r/(2r+1)}$.

By Theorems 2 and 3, the regularized estimator $\hat{\beta}_n$ is minimax rate optimal: the minimax rate of convergence for the prediction error is $n^{-2r/(2r+1)}$. As discussed previously, this optimal rate of convergence depends jointly on the eigenvalue decay rate of the operator $L_{K^{1/2}CK^{1/2}}$ (i.e., of the eigenvalues of K and C) through r and, more importantly, on alignment between the eigenfunctions of K and C.

3.3 Computation

In this section, we propose an efficient computational approach for model estimation using the ADMM algorithm. We begin with an application of the representer theorem to establish that the proposed estimator lies in a finite-dimensional subspace. We subsequently discuss hyperparameter tuning and propose our estimation algorithm.

3.3.1 Representer theorem

Theorem 3.3 (Representer theorem) *Let $(\xi_1, \ldots, \xi_M)$ be a basis of $\mathcal{H}_0$. There exist vectors $e = (e_1, \ldots, e_M)^\top$ and $c = (c_1, \ldots, c_n)^\top$ allowing the solution $\hat{\beta}_n$ to the problem in Equation (3.2) to be expressed as*

$$\hat{\beta}_n(t) = \sum_{i=1}^{M} e_i \xi_i(t) + \sum_{k=1}^{n} c_k \int_{\mathcal{T}} K(s,t) X_k(t) \, \mathrm{d}s. \tag{3.5}$$

Theorem 3 is a generalization of the well-known representer lemma for smoothing splines [175]. Although the minimization over $\hat{\beta}_n$ in Equation (3.2) is taken over an infinite-dimensional space $\mathcal{H}(K)$, the above result implies that the solution lies in a finite-dimensional subspace. Thus, it suffices to estimate the coefficients e and c in Equation (3.5). By Theorem 3, we can conclude that

$$\int_{\mathcal{T}} X(t)\beta(t)dt = \sum_{i=1}^{M} e_i \int_{\mathcal{T}} X(t)\xi_i(t) \, \mathrm{d}t + \sum_{k=1}^{n} c_k \int_{\mathcal{T}} \int_{\mathcal{T}} X(t)K(s,t)X_k(s) \, \mathrm{d}s \, \mathrm{d}t.$$

Let $Y = (Y_1, Y_2, \ldots, Y_n)^\top$ and let T represent the $n \times M$ matrix with the (i,j)th entry $T_{ij} = \int_{\mathcal{T}} X_i(t)\xi_j(t) \, \mathrm{d}t$ for $i = 1, \ldots, n$ and $j = 1, \ldots, M$. Similarly, let Σ

be the $n \times n$ matrix with the (i,j)th entry $\Sigma_{ij} = \int_{\mathcal{T}} \int_{\mathcal{T}} X_i(t) K(s,t) X_j(s)\ \mathrm{d}s\ \mathrm{d}t$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. It follows from the reproducing property that

$$J(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \int_{\mathcal{T}} \int_{\mathcal{T}} X_i(t) K(s,t) X_j(s)\ \mathrm{d}s\ \mathrm{d}t = c^\top \Sigma c.$$

We make use of this representation in the following subsections for model estimation.

c

3.3.2 Hyperparameter tuning

As with most smoothing methods, the selection of the tuning parameter λ influences the performance of the regularized estimator $\hat{\beta}_n$. There are various tools available for this task, such as K-fold cross-validation [87], the Bayesian information criterion (BIC), generalized maximum likelihood [175], and generalized cross-validation (GCV) [55].

In this chapter, unless otherwise noted, we employ GCV as a practical criterion for choosing the optimal tuning parameter value. Because the regularized estimator is a linear estimator and can be written as $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) = H(\lambda) y = Te + \Sigma c$, where $H(\lambda)$ is the "hat matrix" for a particular value of λ, we may select the the value of λ that minimizes [175]

$$GCV(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} r_\tau(\hat{y}_i - y_i)}{(1 - \mathrm{Tr}(H(\lambda))/n)^2}.$$

3.3.3 ADMM algorithm

We next apply the ADMM algorithm to estimate the functional linear expectile model. Pseudocode for the proposed estimation procedure is provided in Algorithm 1.

Developed in the 1970s and further summarized in [12], the ADMM algorithm is a simple and efficient approach for solving convex optimization problems. It has found renewed popularity in large-scale computing through its ability to decentralize large, global problems into small, local ones. The ADMM algorithm has been employed in

quantile regression [58], two-way functional hazard models [96], and Gaussian graphical models [107], to name only a few applications.

Using the results and notation of Section 3.3.1, the optimization problem in Equation (3.1) can be reformulated as a convex optimization problem with respect to e, c, and the auxiliary variable u as

$$\begin{array}{ll} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} r_\tau(y_i - u_i) + \lambda c^\top \Sigma c \\ \text{subject to} & u_i = T_i e + \Sigma_i c, \quad i = 1, \ldots, n, \end{array} \tag{3.6}$$

where T_i and Σ_i denote the ith rows of T and Σ, respectively. The scaled ADMM algorithm [12] uses an objective function defined by the augmented Lagrangian form of the above problem,

$$\begin{aligned} L_\sigma(u, e, c, h) =& \frac{1}{n}\sum_{i=1}^{n} r_\tau(Y_i - u_i) + \lambda c^\top \Sigma c + \\ & \frac{\sigma}{2}\sum_{i=1}^{n}(u_i - T_i e - \Sigma_i c + h_i)^2 - \frac{\sigma}{2}\sum_{i=1}^{n} h_i^2, \end{aligned}$$

which we aim to minimize over $u = (u_1, \ldots, u_n)^\top$, e, c, and $h = (h_1, \ldots, h_n)^\top$ without restriction.

The scaled ADMM update scheme for the $(k+1)$th iteration is straightforward to derive:

$$\begin{aligned} u_i^{k+1} &= \arg\min_{u_i} \left\{ \frac{1}{n}\sum_{i=1}^{n} r_\tau(Y_i - u_i) + \frac{\sigma}{2}(u_i - T_i e^k - \Sigma_i c^k + h_i^k)^2 \right\} \\ &= \begin{cases} \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2\tau y_i}{\sigma + 2\tau}, & y_i \geq u_i \\ \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2(1-\tau) y_i}{\sigma + 2(1-\tau)}, & y_i < u_i \end{cases} \\ (e^{k+1},\ c^{k+1}) &= \arg\min_{e_i, c_i} \left\{ \lambda c^\top \Sigma c + \frac{\sigma}{2}\left(u_i^{k+1} - T_i e - \Sigma_i c + h_i^k\right)^2 \right\} \\ h_i^{k+1} &= h_i^k + u_i^{k+1} - T_i e^{k+1} - \Sigma_i c^{k+1}. \end{aligned}$$

The update step for (e, c) above can be explicitly solved using the sub-iterations

$$e^{k+1} = (T^\top T)^{-1} \left[\sum_{i=1}^{n} T_i^\top (u_i^{k+1} - \Sigma_i c^k + h_i^k) \right],$$

$$c^{k+1} = (2\lambda\Sigma/\sigma + \Sigma^\top \Sigma)^{-1} \left[\sum_{i=1}^{n} \Sigma_i^\top (u_i^{k+1} - T_i e^{k+1} + h_i^k) \right].$$

Stopping conditions for the proposed scheme can be defined in terms of the size of the problem's primal and dual residuals: we terminate the algorithm when $r^k = \|u - Te - \Sigma c\| \leq \epsilon_{\text{dual}}$ and $s^k = \sigma \left(T(e^{k+1} - e^k) + \Sigma(c^{k+1} - c^k) \right) \leq \epsilon_{\text{pri}}$. Here, $\epsilon_{pri} = \sqrt{n}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max(\|u\|_2, \|Te + \Sigma c\|_2) > 0$ and $\epsilon_{\text{dual}} = \sqrt{n}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \|h\|_2 > 0$ are feasibility tolerances for the primal and dual feasibility conditions, where $\epsilon_{\text{abs}} > 0$ and $\epsilon_{\text{rel}} > 0$ are absolute and relative tolerances, respectively. In all of the numerical studies presented in the current section, we follow the suggestion in [12] by fixing $\epsilon_{\text{rel}} = 10^{-4}$, $\epsilon_{\text{abs}} = 10^{-2}$, and $\sigma = 2$.

Algorithm 1 ADMM algorithm for functional linear expectile regression.

Input: u^0, e^0, c^0, h^0 (initial estimates); σ (step size parameter); λ (tuning parameter)

repeat
  for $i = 1, \ldots, n$ **do**
    if $u_i^k \leq y_i^k$ **then**
      $u_i^{k+1} \leftarrow \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2\tau y_i}{\sigma + 2\tau}$
    else
      $u_i^{k+1} \leftarrow \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2(1-\tau) y_i}{\sigma + 2(1-\tau)}$
    end if
  end for
  $e^{k+1} \leftarrow (T^\top T)^{-1} \left[\sum_{i=1}^{n} T_i^\top (u_i^{k+1} - \Sigma_i c^k + h_i^k) \right]$
  $c^{k+1} \leftarrow (2\lambda\Sigma/\sigma + \Sigma^\top \Sigma)^{-1} \left[\sum_{i=1}^{n} \Sigma_i^\top (u_i^{k+1} - T_i e^{k+1} + h_i^k) \right]$
  $h_i^{k+1} \leftarrow h_i^k + u_i^{k+1} - T_i e^{k+1} - \Sigma_i c^{k+1}$
until stopping criteria are met
compute estimated slope function $\hat{\beta}$ from the optimal e, c

Output: $L_\sigma(u, e, c, h)$, $\hat{\beta}$

3.3.4 Convergence of the ADMM algorithm

We next apply a general result of [12] to verify the convergence of our proposed ADMM-based approach for estimating the functional linear expectile regression model.

For convenience, we return to a more-general formulation of the ADMM algorithm:

$$\begin{aligned} &\text{minimize} && F(x,z) = f(x) + g(z) \\ &\text{subject to} && G(x,z) = Ax + Bz - c = 0, \end{aligned} \tag{3.7}$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, where $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$, and $c \in \mathbb{R}^p$ [12]. For our setting, x and z correspond to u and $(e^\top, c^\top)^\top$; f and g to the empirical expectile loss $\frac{1}{n}\sum_{i=1}^n r_\tau(Y_i - u_i)$ and $\lambda c^\top \Sigma c$; and A, B, and c to the identity matrix I, $[T_i, \Sigma_i]$, and 0, respectively. To guarantee convergence, we verify two additional conditions, referring to Assumptions 1 and 2 of [12].

First, we require that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, proper, and convex. This requirement is naturally satisfied for our formulation in (3.6).

Second, we require that the nonaugmented Lagrangian $L_0(x, z, y) = f(x) + g(z) + y^\top(Ax + Bz - c)$ has a saddle point, i.e., that there exists a (not necessarily unique) $(x^\star, z^\star, y^\star)$ satisfying $L_0(x^\star, z^\star, y) \le L_0(x^\star, z^\star, y^\star) \le L_0(x, z, y^\star)$ for all (x, z, y). The existence of a saddle point follows immediately from the saddle point theorem [121, Theorem B.29] and the fact that we are optimizing over a real space (specifically, with a nonempty interior), that $G(x, z)$ is affine, and that $G(0, 0) = 0$ (since $c = 0$).

Consequently, we can guarantee that the estimate and objective function iterates in our ADMM-based implementation will converge to the solution and optimal value, respectively, of the original problem. The particular benefit of an ADMM-based approach is that each update has a closed form, which speeds up numerical computation relative to traditional interior point methods or other generic algorithms. Indeed, empirical results in existing literature have illustrated the clear superiority that ADMM-based algorithms have in a variety of settings [19, 133].

3.4 Numerical Experiments

We now investigate the finite-sample performance of our proposed estimators. We are specifically interested in comparisons between our proposed estimator and one

using an FPCA-based approach that uses the first four leading eigenfunctions [80, 143, 198]. Three sets of simulations in Section 3.4.1 examine the effects of eigenvalue decay, kernel alignment, and various error distributions on the convergence of both estimators.

3.4.1 Simulation studies

In the following sets of simulation studies, we consider $\mathcal{T} = [0, 1]$ and let $\mathcal{H} = \mathcal{H}(K)$ be the set of functions in the linear span of the cosine basis [16], i.e., $\mathcal{H}(K) = \{g(t) = \sqrt{2}\sum_{k\in\mathbb{N}} g_k \cos(k\pi t) : g_k \in \mathbb{R}, k \in \mathbb{N}\} \subset \mathcal{W}_2^2$. When endowed with the squared norm

$$\|f\|^2_{\mathcal{H}(K)} = \int_{\mathcal{T}} (f'')^2 = \int_0^1 \left(\sqrt{2}\sum_{k\in\mathbb{N}} (k\pi)^2 g_k \cos k\pi t\right)^2 = \sum_{k\in\mathbb{N}} (k\pi)^4 g_k^2,$$

$\mathcal{H}$ is an RKHS with the reproducing kernel

$$\begin{aligned} K(s,t) &= \sum_{k\in\mathbb{N}} 2(k\pi)^{-4} \cos(k\pi s)\cos(k\pi t) \\ &= -\frac{1}{3}\left(B_4(|s-t|/2) + B_4((s+t)/2)\right), \end{aligned}$$

where B_k is the kth Bernoulli polynomial

$$B_{2m}(x) = (-1)^{m-1} 2(2m)! \sum_{k\in\mathbb{N}} \frac{\cos(2\pi k x)}{(2\pi k)^{2m}},$$

for $x \in [0, 1]$. Additionally, we choose $(\xi_1(t) = 1,\ \xi_2(t) = t)$ as the basis for the null space $\mathcal{H}_0$.

To quantify the behaviour of varying coefficient estimates, we calculate prediction error (PE) on a test dataset $\{(x_i^*, y_i^*) : i = 1, \ldots, n^*\}$, given by

$$\mathrm{PE}_\tau = \left(\frac{1}{n^*}\sum_{j=1}^{n^*} \left\| \int_{\mathcal{T}} x_j^*(t)\hat{\beta}_n(t)\ \mathrm{d}t - \int_{\mathcal{T}} x_j^*(t)\beta_0(t)\ \mathrm{d}t \right\|_2^2\right)^{1/2}.$$

As a more direct comparison between the RKHS- and FPCA-based estimators, we also report relative prediction error, defined by $\mathrm{PE}_\tau^{\mathrm{FPCA}}/\mathrm{PE}_\tau^{\mathrm{RKHS}}$, where $\mathrm{PE}_\tau^{\mathrm{FPCA}}$ and $\mathrm{PE}_\tau^{\mathrm{RKHS}}$ represent prediction errors for the two methods. In all simulation studies, results are averaged over 100 simulated training and test datasets.

In the first simulation study, we focus primarily on the effect of eigenvalue decay rate. We define the covariance operator as

$$C(s,t) = \sum_{k=1}^{50} 2k^{-2r_2} \cos(k\pi s)\cos(k\pi t),$$

where $r_2 = 1, 2, 3$ imposes different decay rates on the eigenvalues of C: a larger value of r_2 yields stronger eigenvalue decay. In this setting, the two kernels, K and C, share the same ordered set of eigenfunctions.

We follow the data generation procedure in [61] and [16]. The response is generated as $Y = \int_0^1 X(t)\beta_0(t)\,\mathrm{d}t + \varepsilon$, with $\beta_0(t) = \sum_{k=1}^{50} \beta_k \phi_k(t)$; $\beta_k = 4(-1)^{k+1}k^{-2}$ and $\phi_k(t) = \sqrt{2}\cos(k\pi t)$ for $k = 1, \ldots, 50$; and $\varepsilon \sim N(0, 0.5)$. The functional covariate is generated as $X(t) = \sum_{k=1}^{50} \gamma_k U_k \phi_k(t)$, where $\gamma_k = (-1)^{k+1}k^{-r_2}$ and $U_k \overset{\text{i.i.d.}}{\sim} U[-\sqrt{3}, \sqrt{3}]$. The U_ks have a mean of zero and unit variance and each X is observed at 101 equally spaced grid points on $[0, 1]$. We emphasize that the data generation process is ultimately driven by the choice of the covariance operator.

Results for the first simulation are presented in Figure 3.2. First, the generally positive performance of the FPCA-based estimator is not surprising, as β_0 is a linear combination of the leading eigenfunctions of the functional covariate X. Nonetheless, our RKHS-based estimator demonstrates higher relative predictive performance except in certain settings with $r_2 = 1$, where the eigenvalue decay rate is small. In these settings, the standard errors of the PE and relative PE measures across simulations are typically small. Together, these results suggest a systematically lower PE for the proposed method. The PE of both estimators generally decreases as r_2 increases, as expected. Both methods appear to converge at similar rates as the sample size increases, although the RKHS-based estimator again outperforms the FPCA-based one.

In the second simulation study, we are primarily interested in how alignment between the reproducing kernel K and the covariance kernel C influences the performance of the RKHS- and FPCA-based estimators. We define the covariance kernel

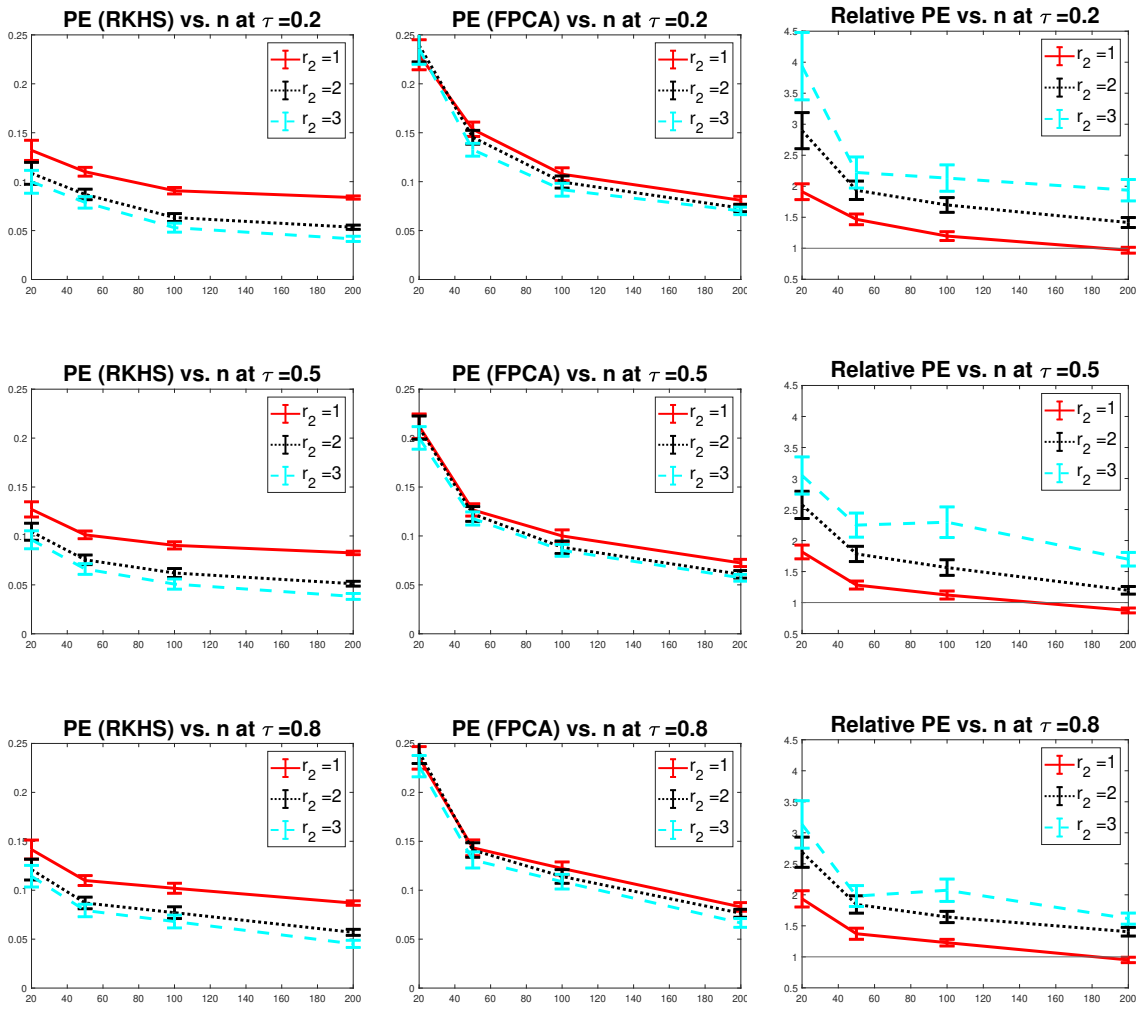


Figure 3.2: Effect of covariance kernel eigenvalue decay rate on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the first simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$.

in this setting as

$$C(s,t) = \sum_{k=1}^{50} 2(|k-k_0|+1)^{-2}\cos(k\pi s)\cos(k\pi t).$$

To control the extent of the alignment between K and C, the leading eigenfunctions of C are located around the k_0th eigenfunction of the reproducing kernel K: we consider $k_0 = 5, 10, 20$, with larger values of k_0 corresponding to worse alignment [16]. In all other aspects, the data generation process matches that of the first simulation study.

Figure 3.3 presents results for the second simulation study. As expected, the FPCA-based estimator generally shows worse PE relative to the RKHS-based estimator. We observe that relative PE increases with worsened alignment, most notably when $k_0 = 20$. Furthermore, with increasing k_0, poor alignment between K and C seems to have a significant impact on the FPCA-based estimator but little effect on the proposed RKHS-based one. The standard error for relative PE is large in some settings, but still leads us to conclude that the proposed method gives systematically better PE. These empirical results are consistent with our theoretical expectations and illustrate the merit of our RKHS-based perspective.

In the third simulation study, we investigate the ability of our proposed approach to cope with different types of error distributions. Specifically, we consider distributions that are either heteroscedastic or asymmetric.

We use the same setup as the first simulation study (excepting the distribution of ε), with $r_2 = 2$. As asymmetric error distributions, we take $\varepsilon \sim \text{Gamma}(2, 0.2)$ and $\varepsilon \sim \text{Beta}(5, 1)$ for left- and right-skewed errors, respectively. Heteroscedastic errors are sampled as a mixture of $N(0, 0.25)$, $N(0, 0.375)$, and $N(0, 0.5)$ distributions, representing a simple case with three heteroscedastic groups.

Results are presented in Figure 3.4 for the third simulation study. As a general trend, our proposed RKHS-based estimator shows better performance than the FPCA-based estimator, with relative PE typically falling between one and four. Standard error for relative PE is moderate across the different settings but is again sug-

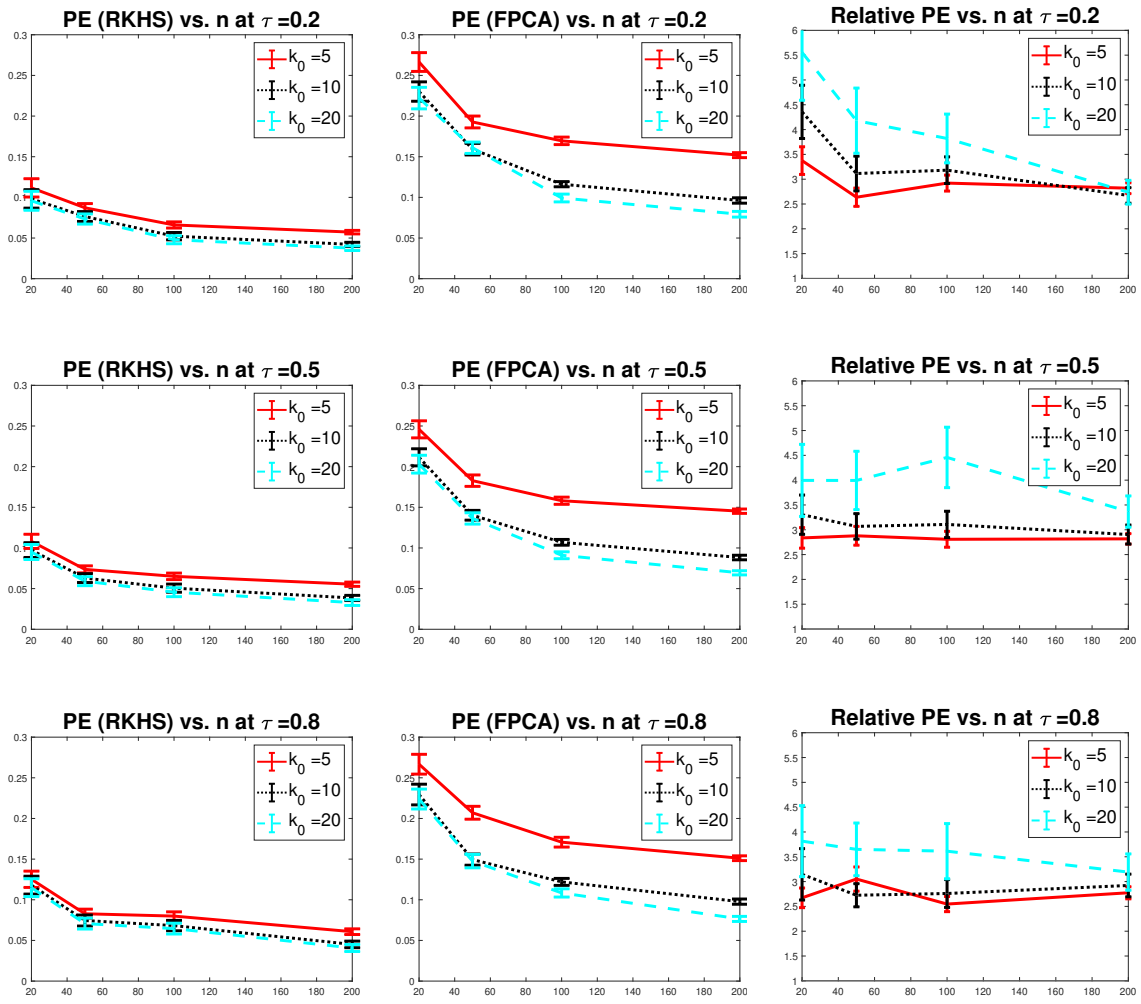


Figure 3.3: Effect of reproducing and covariance kernel alignment on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the second simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$.

gestive of a systematically lower PE for the proposed RKHS-based estimator. In the setting with right-skewed errors, PE for both estimators is relatively smaller when $\tau = 0.2$ than when $\tau = 0.8$: this result is reversed for left-skewed errors. These results, for both asymmetric and heteroscedastic error distributions, demonstrate the power of expectile regression in dealing with various error distributions, relative to methods that focus on conditional mean estimation. This simulation study highlights the versatility of our expectile model in cases of model error misspecification.

3.4.2 Application to ADNI data

We next apply the proposed RKHS-based estimator in an analysis of MMSE scores from 199 patients in the ADNI dataset. In the functional linear model, the response Y is MMSE score while the functional predictor X is fractional anisotropy (FA) as a function of distance along the midsagittal corpus callosum skeleton (scaled to $\mathcal{T} = [0, 1]$). The corresponding functional linear model is

$$MMSE = \int_0^1 \beta_0(t) FA(t) \, \mathrm{d}t + \varepsilon.$$

Figure 3.5 plots FA values, observed at 83 grid points, for all 199 patients. For tuning and evaluating both estimators, approximately 80% of the data is used for four-fold cross validation while the remaining 20% is held out as a test set. Context and the visualization of the neuroimaging data in Figure 3.5 suggest that the functional predictor $X = \text{FA}$ may be periodic on $[0, 1]$.

We let $\mathcal{H}(K) = \mathcal{W}_2^{\text{per}}$ be the second-order Sobolev space of periodic functions on $[0, 1]$, endowed with the norm $\|b\|_{\mathcal{H}}^2 = \left[\int_0^1 b(t)dt\right]^2 + \int_0^1 [b''(t)]^2 \, dt$ and the reproducing kernel $K(s, t) = 1 - \frac{1}{24} B_4(|s - t|)$, where B_4 is the fourth Bernoulli polynomial [175].

Estimates obtained using our proposed method at the expectile levels $\tau = 0.1, \ldots, 0.9$ are shown in Figure 3.6. As expected, for any fixed t, $\hat{\beta}_\tau(t)$ increases with τ. For the sake of practical interpretation, it is useful that these functional estimates do not cross each other.

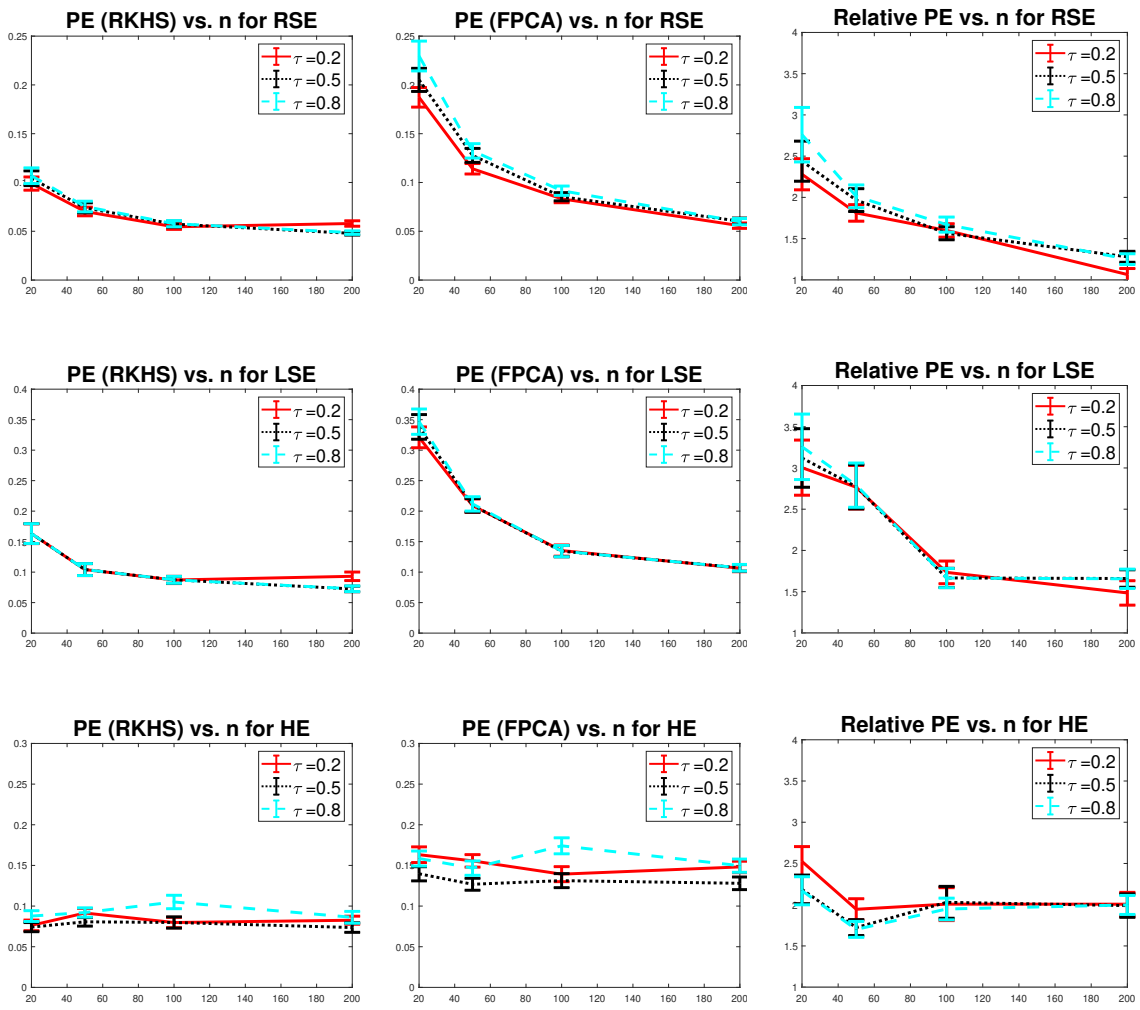


Figure 3.4: Effect of abnormal errors on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the third simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$. RSE, LSE, and HE indicate left-skewed, right-skewed, and heteroscedastic error distributions, respectively.

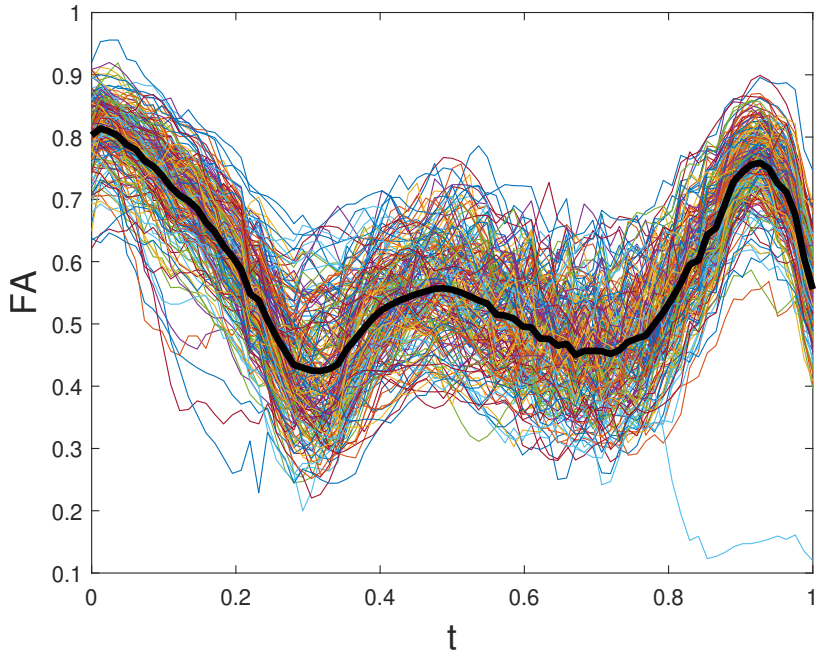


Figure 3.5: FA curves evaluated along the corpus callosum skeleton. The solid black line is the mean FA curve.

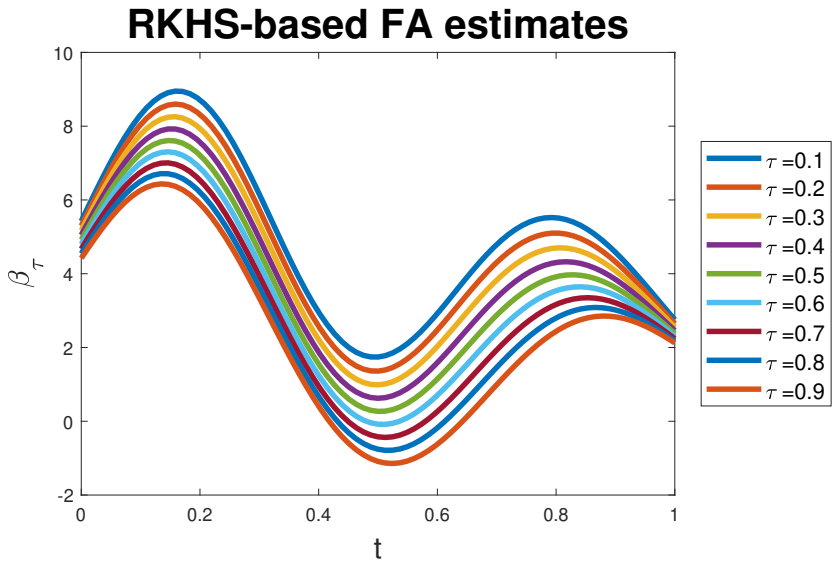


Figure 3.6: RKHS-based estimates $\hat{\beta}_\tau$ at $\tau = 0.1, \ldots, 0.9$ in the ADNI data analysis, describing the functional effect of FA on MMSE score.

Table 3.1: Test set prediction error in the ADNI analysis for the RKHS- and FPCA-based predictors at $\tau = 0.1, \ldots, 0.9$.

Expectile level τ	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
RKHS PE (SE)	0.9954	0.9936	0.9922	0.9913	0.9907	0.9905	0.9907	0.9911	0.9918
	(0.0153)	(0.0152)	(0.0151)	(0.015)	(0.0150)	(0.0149)	(0.0149)	(0.0148)	(0.0148)
FPCA PE (SE)	1.0164	1.0161	1.0058	1.0006	0.9999	0.9982	1.0000	1.0004	1.0008
	(0.0174)	(0.0174)	(0.0163)	(0.0164)	(0.0161)	(0.016)	(0.0160)	(0.0161)	(0.0157)

We also considered FPCA-based estimates obtained using 4, 6, 8, and 10 functional principal components. These estimates, illustrated in Figure 3.7, are clearly not ideal for at least a couple reasons. First, the FPCA-based estimates cross each other, unlike the RKHS-based estimates in Figure 3.6. This "crossing problem" is further discussed in [66] in the context of quantile regression. Second, the FPCA-based estimates are sensitive to the user-specified number of principal components. The discrete nature of this hyperparameter makes it difficult to tune finely, unlike the continuous hyperparameter λ in our RKHS-based approach.

Table 3.1 moreover shows that, at each expectile level considered, the proposed RKHS-based estimator outperforms the FPCA-based one in predicting MMSE. These results emphasize the practical importance and advantages of our RKHS-based approach in functional linear expectile regression.

As an informal aside (due to the computation time involved), we also compared the computational efficiency of different implementations of our proposed RKHS-based estimator. Our first implementation is as presented in Section 3.3.3 using the ADMM algorithm while the second uses an interior point (IP) algorithm [118]. The latter is popularly applied to constrained optimization problems. We found our ADMM implementation to be far superior to the IP implementation: the latter typically requires at least 100 times more computation time than the former until convergence. These results can be made available on request.

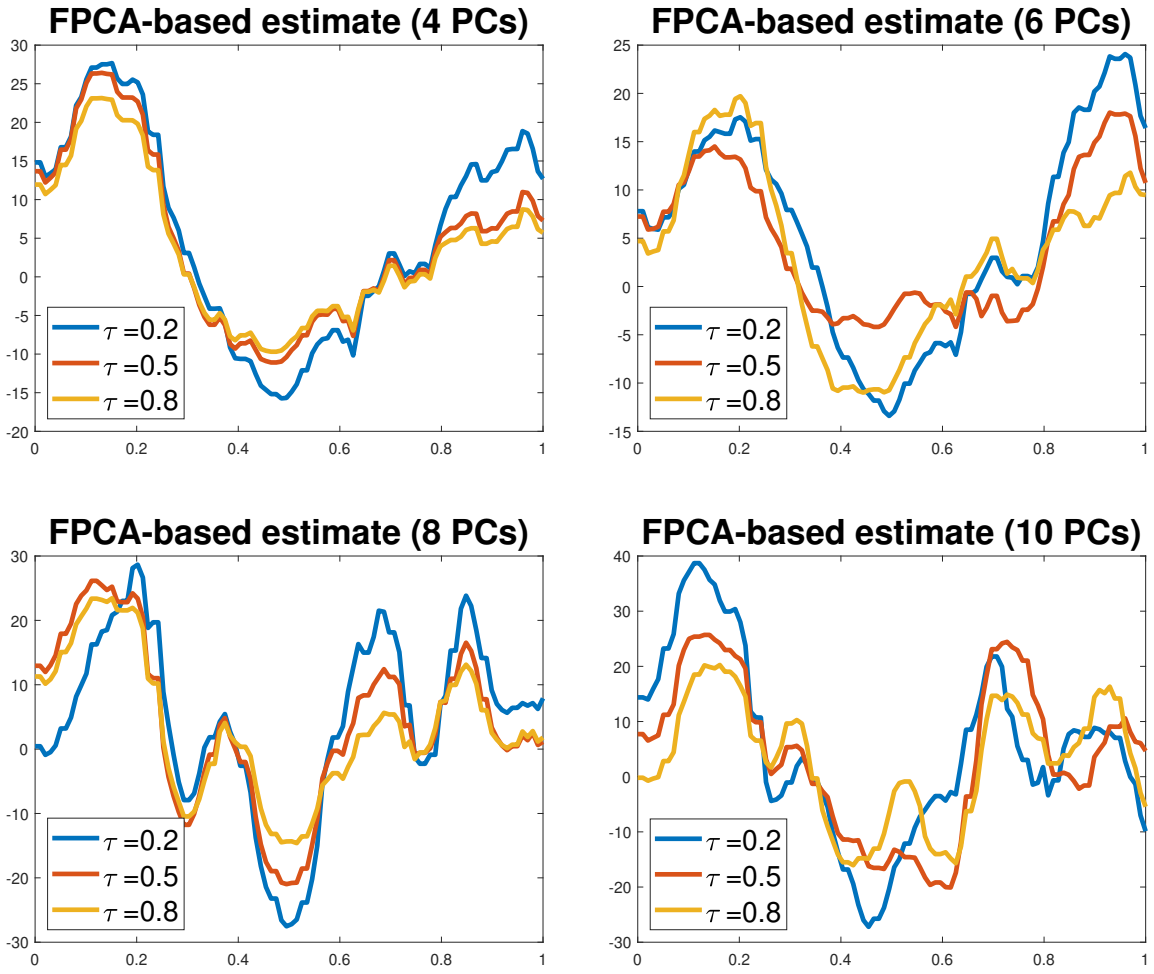


Figure 3.7: FPCA-based estimates $\hat{\beta}_\tau$ at $\tau = 0.2, 0.5, 0.8$ in the ADNI data analysis, describing the functional effect of FA on MMSE score. The number of functional principal components (PCs) used is indicated in each subplot: 4, 6, 8, and 10 PCs explain 79.9%, 86.0%, 89.3%, and 91.5%, respectively, of the observed variance in functional FA.

3.5 Discussion

In this chapter, we proposed a regularized estimator for the functional linear expectile regression model under an RKHS framework. We derived upper and lower bounds for the minimax rate of convergence of prediction error and established the minimax optimality of our proposed estimator. While most existing approaches to functional linear expectile regression rely on FPCA, we argue that these approaches are too restrictive in their assumption regarding eigenvalue spacing. Additionally, FPCA-based methods rely on the assumption that leading principal components (which are determined by only the functional predictor X and not the response Y) are predictive of the response: in practice, this assumption is typically not valid.

We demonstrated the general superiority of our proposed RKHS-based approach in three sets of simulation studies and an application to an ADNI neuroimaging dataset. In particular, we illustrated the degradation of FPCA-based estimators when its implicit assumptions regarding the eigenstructures of the reproducing and covariance kernels are violated. Our results showed that both eigenfunction alignment and eigenvalue decay rates between the reproducing and covariance kernels have an important impact on estimator performance.

For the sake of illustration, we focused on a univariate functional predictor X with a domain $\mathcal{T}$ that is a compact subset of $\mathbb{R}$. We took $\mathcal{T} = [0, 1]$ and used the corresponding canonical Sobolev space as a working example. Our theoretical results apply nonetheless to more general RKHSs, provided that $\mathcal{T}$ remains a compact subset of an arbitrary Euclidean space. For example, the derived optimal convergence rate still holds for Sobolev spaces on $\mathcal{T} = [0, 1]^2$, e.g., for imaging data, with the decay rate r determined by the corresponding reproducing and covariance kernels. The developments in this chapter thus have wide applications in spatial statistics, 2D and 3D image analysis, and longitudinal data analysis.

Settings where the reproducing and covariance kernels are not well aligned (i.e., in

the sense of their eigenfunctions) are interesting topics for future work. As suggested by our ADNI analysis, another natural generalization of our approach is the inclusion of scalar predictors, e.g., age, gender, and diagnosis status, for a partial functional expectile regression model. While it is straightforward to accommodate scalar covariate effect estimation from an algorithmic perspective, the optimality of the corresponding estimators requires more work to establish. Informally (and with results available on request), PE for the RKHS- and FPCA-based estimators are comparable when scalar age, gender, and diagnosis status effects are included in the model. We suspect that this decrease in relative PE can be attributed to the relative complexity of the two models and possibly the overwhelming usefulness of these scalar covariates as predictors. We feel that the full impact of scalar predictors on empirical performance, such as in high-dimensional settings, should be investigated in future work.

3.6 Proof of Main Results

Proof of Theorem 1. Recall the functional model $Y = \int_{\mathcal{T}} X(t)\beta_0(t)\,\mathrm{d}t + \varepsilon$ specified in the main text. Fix an expectile level $\tau \in (0,1)$ and assume that ε follows an asymmetric normal distribution with the density function

$$f(\varepsilon) = \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}}\frac{1}{\sqrt{\pi\sigma^2}}\exp\{-r_\tau(\varepsilon/\sigma)\}, \tag{3.8}$$

where $r_\tau(u) = |\tau - I(u<0)|u^2$. Further assume that β_0 belongs to an RKHS $\mathcal{H}(K)$.

Consider the functional space

$$\mathcal{H}^* = \left\{ \beta = \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}}\zeta_k : (b_{M+1},\ldots,b_{2M}) \in \{0,1\}^M \right\},$$

where $(\zeta_k)_{k\in\mathbb{N}}$ is a sequence of orthonormal eigenfunctions of $K^{1/2}CK^{1/2}$. The function $\|\cdot\|_K$ is a semi-norm on $\mathcal{H}(K)$ and M is some large number to be discussed later.

For any $\beta \in \mathcal{H}^*$, observe that

$$\begin{aligned} J(\beta) = \|\beta\|_K^2 &= \left\| \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}} \zeta_k \right\|_K^2 \\ &= \sum_{k=M+1}^{2M} b_k^2 M^{-1} \left\| L_{K^{1/2}} \zeta_k \right\|_K^2 \\ &\leq \sum_{k=M+1}^{2M} M^{-1} \left\| L_{K^{1/2}} \zeta_k \right\|_K^2 \\ &= 1, \end{aligned}$$

which follows from the fact that $\langle L_{K^{1/2}} \zeta_k, L_{K^{1/2}} \zeta_l \rangle_K = \langle L_K \zeta_k, \zeta_l \rangle_K = \langle \zeta_k, \zeta_l \rangle_{\mathcal{L}_2} = \delta_{kl}$. Therefore, $\mathcal{H}^* \subset \mathcal{H}(K) = \{\beta : \|\beta\|_K < \infty\}$.

The Gilbert-Varshamov bound [172, Lemma 2.9] establishes that, for any $M \geq 8$, there exists a set $\{b^{(0)}, b^{(1)}, \ldots, b^{(N)}\} \subset \{0,1\}^M$ such that

(i) $b^{(0)} = (0, \ldots, 0)^\top$;

(ii) $H(b^{(i)}, b^{(j)}) \geq M/8$ for any distinct $b^{(i)}, b^{(j)} \in \mathcal{B}$, where $H(\cdot, \cdot)$ denotes Hamming distance; and

(iii) $N \geq 2^{M/8}$.

Define the subset

$$\mathcal{B} = \left\{ \beta^{(0)}, \ldots, \beta^{(N)} : \beta^{(i)} = \sum_{k=M+1}^{2M} b_{k-M}^{(i)} M^{-1/2} L_{K^{1/2}} \zeta_k, i = 1, \ldots, N \right\} \subset \mathcal{H}^*$$

and let M be the smallest integer greater than $c_0 n^{1/(2r+1)}$ for some constant $c_0 > 0$.

Then for i and j satisfying $0 \le i \le j \le N$,

$$\begin{aligned}
\left\|\beta^{(i)} - \beta^{(j)}\right\|_C^2 &= \left\| L_{C^{1/2}} \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)}) M^{-1/2} L_{K^{1/2}} \zeta_k \right\|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \| L_{C^{1/2}} L_{K^{1/2}} \zeta_k \|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \nu_k. \\
&\ge \nu_{2M} M^{-1} \sum_{k=1}^{M} (b_k^{(i)} - b_k^{(j)})^2 \\
&= 4\nu_{2M} M^{-1} H(b^{(i)}, b^{(j)}) \\
&\ge \nu_{2M}/2 \\
&\ge c_1 2^{-(2r+1)} M^{-2r} \\
&\ge 2c\alpha^{2r/(2r+1)} n^{-2r/(2r+1)},
\end{aligned}$$

where $c > 0$ is some constant.

We apply the results of [172] to establish a lower bound based on multiple hypothesis testing. Under the assumption that the slope function β_0 belongs to the subset $\mathcal{B}$, we construct a subset $\{\beta^{(0)}, ..., \beta^{(N)}\} \subset \mathcal{H}^*$ with N increasing in n such that, for some positive constant c and for i and j such that $0 \le i \le j \le N$,

$$\|\beta^{(i)} - \beta^{(j)}\|_C^2 \ge c\alpha^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}} \tag{3.9}$$

and

$$\frac{1}{N} \sum_{j=1}^{N} \mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) \le \alpha \log N, \tag{3.10}$$

where P_β denotes the joint conditional distribution of Y given X and KL represents Kullback-Leibler divergence. By Theorem 2.5 of [172], it follows that

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathcal{H}^*} \mathbb{P}(\|\beta^{(i)} - \beta^{(j)}\|_C^2 \ge c\alpha^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}}) \ge \frac{\sqrt{N}}{\sqrt{N}+1} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log N}}\right). \tag{3.11}$$

Note that $M, N \to \infty$ as $n \to \infty$. This implies that the right-hand side of (3.11) can

be made arbitrarily close to 1 as $n \to \infty$ and $\alpha \to 0$. We conclude that

$$\lim_{\alpha \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}} \sup_{\beta_0 \in \mathcal{H}^*} \mathbb{P}(\|\beta^{(i)} - \beta^{(j)}\|_C^2 \geq a n^{-\frac{2r}{2r+1}}) = 1. \tag{3.12}$$

This lower bound for the asymmetric normal distribution yields a lower bound for general error distributions. Let P_j, for $j = 1, \ldots, N$, represent the joint distribution of the observed sample $\{(x_k, y_k) : k = 1, \ldots, n\}$ under the assumption that $\beta_0 = \beta^{(j)}$. It follows that

$$P_j = \prod_{k=1}^{n} \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\tau} + \sqrt{1-\tau}} \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{ -r_\tau \left(\frac{y_k - \int_{\mathcal{T}} x_k(t)^\top \beta^{(j)}(t)}{\sigma} \right) \right\}. \tag{3.13}$$

The Kullback-Leibler divergence between $P_{\beta^{(i)}}$ and $P_{\beta^{(j)}}$ is

$$\begin{aligned}
\mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) &= E_{\beta^{(i)}} \log(P_{\beta^{(i)}} / P_{\beta^{(j)}}) \\
&= n E_{\beta^{(i)}} \left[r_\tau \left(\frac{Y - \int_{\mathcal{T}} X(t)\beta^{(j)}(t)\, \mathrm{d}t}{\sigma} \right) - r_\tau \left(\frac{Y - \int_{\mathcal{T}} X(t)^\top \beta^{(i)}(t)\, \mathrm{d}t}{\sigma} \right) \right] \\
&\leq n \max(\tau, 1-\tau) \left(\int_{\mathcal{T}} X(t)^\top (\beta^{(j)}(t) - \beta^{(i)}(t))\, \mathrm{d}t \right)^2.
\end{aligned}$$

The inequality above holds since, defining $\mu^i = \int_{\mathcal{T}} X(t)^\top \beta^{(i)}(t)\, \mathrm{d}t$,

$$\begin{aligned}
&E_{\beta^{(i)}} \left[r_\tau \left(\frac{Y - \mu^j}{\sigma} \right) - r_\tau \left(\frac{Y - \mu^i}{\sigma} \right) \right] \\
&\quad = \int_{\mu^i}^{\infty} \tau \left[\left(\frac{y - \mu^j}{\sigma} \right)^2 - \left(\frac{y - \mu^i}{\sigma} \right)^2 \right] f(y - \mu^i)\, \mathrm{d}y \\
&\qquad + \int_{-\infty}^{\mu^i} (1-\tau) \left[\left(\frac{y - \mu^j}{\sigma} \right)^2 - \left(\frac{y - \mu^i}{\sigma} \right)^2 \right] f(y - \mu^i)\, \mathrm{d}y \\
&\qquad + \int_{\mu^j}^{\mu^i} (2\tau - 1) \left(\frac{y - \mu^j}{\sigma} \right)^2 f(y - \mu^i)\, \mathrm{d}y,
\end{aligned}$$

where

$$\int_{\mu^j}^{\mu^i} (2\tau - 1) \left(\frac{y - \mu^j}{\sigma} \right)^2 f(y - \mu^i)\, \mathrm{d}y \leq |1 - 2\tau| \left(\frac{\mu^i - \mu^j}{\sigma} \right)^2 \int_{\mu^j}^{\mu^i} f(y - \mu^i)\, \mathrm{d}y.$$

Thus,

$$\begin{aligned}
\mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) &\leq n \max(\tau, 1-\tau) \left(\int_{\mathcal{T}} X_k(t)^\top (\beta^{(j)}(t) - \beta^{(i)}(t)) \, \mathrm{d}t \right)^2 \\
&= n \max(\tau, 1-\tau) \| L_{c^{1/2}} (\beta^{(j)}(t) - \beta^{(i)}(t)) \|_{\mathcal{L}_2}^2 \\
&= n \max(\tau, 1-\tau) \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \nu_k \\
&\leq n \max(\tau, 1-\tau) \nu_M M^{-1} \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 \\
&= 4n \max(\tau, 1-\tau) \nu_M M^{-1} H(b^{(i)}, b^{(j)}) \\
&\leq 4n \max(\tau, 1-\tau) \nu_M \\
&\leq 4c_2 n \max(\tau, 1-\tau) M^{-2r}.
\end{aligned}$$

Consequently, when $0 < \alpha < 1/8$,

$$\frac{1}{N} \sum_{j=1}^{N} KL(P_j \mid P_0) \leq 4c_2 n \max(\tau, 1-\tau) M^{-2r} \leq \alpha \log 2^{M/8} \leq \alpha \log N.$$

By taking M to be the smallest integer greater than $c_2 \alpha^{-1/(2r+1)} n^{1/(2r+1)}$ with $c_2 = (8c_1 \log 2)^{1/(2r+1)}$, the desired result follows. ■

Proof of Theorem 2. Recall that $L_{K^{1/2}}(\mathcal{L}_2) = \mathcal{H}(K)$. Therefore, there exist $f_0, \hat{f} \in \mathcal{L}_2$ such that $\beta_0 = L_{K^{1/2}} f_0$ and $\hat{\beta}_\lambda = L_{K^{1/2}} \hat{f}_\lambda$. For brevity, we assume that $\mathcal{H}(K)$ is dense in $\mathcal{L}_2$, which ensures that f_0 and $\hat{f}_\lambda$ are uniquely defined. The proof in the general case proceeds in exactly the same fashion by restricting consideration to $\mathcal{L}_2 / \ker(L_{K^{1/2}})$.

For brevity, define $T = L_{K^{1/2} C K^{1/2}}$. Let T^ν denote a linear operator from $\mathcal{L}_2$ to $\mathcal{L}_2$ such that $T^\nu \varphi_k = s_k^\nu \varphi_k$. Prediction error can then be written as

$$\|\hat{\beta} - \beta_0\|_C^2 = \left\| T^{1/2} \left(\hat{f}_\lambda - f_0 \right) \right\|_{\mathcal{L}_2}^{2.}$$

and, furthermore,

$$\hat{f}_\lambda = \operatorname*{arg\,min}_{f \in \mathcal{L}_2} \left[\frac{1}{n} \sum_{i=1}^{n} r_\tau \left(y_i - \langle x_i, L_{K^{1/2}} f \rangle_{\mathcal{L}_2} \right)^2 + \lambda \|f\|_{\mathcal{L}_2}^2 \right].$$

Recalling that $y_i = \langle x_i, L_{K^{1/2}} f_0 \rangle_{\mathcal{L}_2} + \varepsilon_i$,

$$C_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} e_i x_i(s) x_i(t),$$

where $e_i = \tau$ if $y_i \geq \left\langle x_i, L_{K^{1/2}} \hat{f}_\lambda \right\rangle_{\mathcal{L}_2}$ and $e_i = 1 - \tau$ otherwise. Define $T_n = L_{K^{1/2}} L_{C_n} L_{K^{1/2}}$, where L_{C_n} is an integral operator such that, for any $h \in \mathcal{L}_2$,

$$L_{C_n} h(\cdot) = \int_{\mathcal{T}} C_n(s, \cdot) h(s) \, \mathrm{d}s.$$

Consequently, $\hat{f}_\lambda = (T_n + \lambda \mathbf{1})^{-1} (T_n f_0 + g_n)$, where $\mathbf{1}$ is the identity operator and $g_n = \frac{1}{n} \sum_{i=1}^{n} e_i \epsilon_i L_{K^{1/2}} x_i$.

Next, define $f_\lambda = (T + \lambda \mathbf{1})^{-1} T f_0$. By the triangle inequality,

$$\left\| T^{1/2} \left(\hat{f}_\lambda - f_0 \right) \right\|_{\mathcal{L}_2} = \left\| T^{1/2} (f_\lambda - f_0) \right\|_{\mathcal{L}_2} + \left\| T^{1/2} \left(\hat{f}_\lambda - f_\lambda \right) \right\|_{\mathcal{L}_2}. \tag{3.14}$$

The first term on the right-hand side can be easily bounded. To proceed, we appeal to the following lemma.

Lemma 3.4 *Lemma A1. For* $0 < \nu < 1$, $\left\| T^v (f_\lambda - f_0) \right\|_{\mathcal{L}_2} \leq (1-\nu)^{1-\nu} \nu^\nu \lambda^\nu \left\| f_0 \right\|_{\mathcal{L}_2}$.

Taking $\nu = 1/2$ in Lemma A1 establishes that $\left\| T^{1/2} (f_\lambda - f_0) \right\|_{\mathcal{L}_2}^2 \leq \frac{1}{4} \lambda \left\| f_0 \right\|_{\mathcal{L}_2}^2$.

We now turn to the second term on the right-hand side of Equation (3.14). Observe that

$$f_\lambda - \hat{f}_\lambda = (T + \lambda \mathbf{1})^{-1} (T_n + \lambda \mathbf{1}) \left(f_\lambda - \hat{f}_\lambda \right) + (T + \lambda \mathbf{1})^{-1} (T - T_n) \left(f_\lambda - \hat{f}_\lambda \right)$$

and that $(T_n + \lambda \mathbf{1}) \hat{f}_\lambda = T_n f_0 - g_n$. Therefore,

$$\begin{aligned} f_\lambda - \hat{f}_\lambda &= (T + \lambda \mathbf{1})^{-1} T_n (f_\lambda - f_0) + \lambda (T + \lambda \mathbf{1})^{-1} f_\lambda + (T + \lambda \mathbf{1})^{-1} g_n \\ &\quad + (T + \lambda \mathbf{1})^{-1} (T - T_n) \left(f_\lambda - \hat{f}_\lambda \right) \\ &= (T + \lambda \mathbf{1})^{-1} T (f_\lambda - f_0) + (T + \lambda \mathbf{1})^{-1} (T_n - T) (f_\lambda - f_0) + \lambda (T + \lambda \mathbf{1})^{-1} f_\lambda \\ &\quad + (T + \lambda \mathbf{1})^{-1} g_n + (T + \lambda \mathbf{1})^{-1} (T - T_n) \left(f_\lambda - \hat{f}_\lambda \right) \end{aligned}$$

We first consider bounding $\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2}$ for some $\nu \in (0, 1/2-1/(4r))$. By the triangle inequality,

$$\begin{aligned}\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} \leq & \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2} \\ & + \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n-T\right)\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2} \\ & + \lambda\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}f_{\lambda}\right\|_{\mathcal{L}_2} + \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_n\right\|_{\mathcal{L}_2} \\ & + \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T-T_n\right)\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2}.\end{aligned}$$

Lemma 3.5 *Lemma A2. Assume that there exists a constant $c_3 > 0$ such that, for any $f \in \mathcal{L}_2$, $E\langle X, f\rangle_{\mathcal{L}_2}^4 \leq c_3(E\langle X, f\rangle_{\mathcal{L}_2}^2)^2$. Then for any $\nu > 0$ such that $2r(1-2\nu) > 1$,*

$$\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n-T\right)T^{-\nu}\right\|_{op} = O_p\left(\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right),$$

where $\|\cdot\|_{op}$ denotes the usual operator norm, i.e., $\|U\|_{op} = \sup_{\{h:\|h\|_{\mathcal{L}_2}=1\}}\|Uh\|_{\mathcal{L}_2}$ for an operator $U : \mathcal{L}_2 \to \mathcal{L}_2$.

By an application of Lemma A2,

$$\begin{aligned}&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T-T_n\right)\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} \\ &\quad \leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T-T_n\right)T^{-\nu}\right\|_{\mathrm{op}}\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} \\ &\quad \leq o_p(1)\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2}\end{aligned}$$

whenever $\lambda \geq cn^{-2r/(2r+1)}$ for some constant $c > 0$. Similarly,

$$\begin{aligned}&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n-T\right)\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2} \\ &\quad \leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n-T\right)T^{-\nu}\right\|_{\mathrm{op}}\left\|T^{\nu}\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2} \\ &\quad \leq o_p(1)\left\|T^{\nu}\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} = O_p\Big(&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T\left(f_{\lambda}-f_0\right)\right\|_{\mathcal{L}_2} \\ &+ \lambda\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}f_{\lambda}\right\|_{\mathcal{L}_2} + \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_n\right\|_{\mathcal{L}_2}\Big).\end{aligned}$$

By Lemma A1,

$$\begin{aligned}\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2} &\leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T^{1-\nu}\right\|_{\mathrm{op}}\left\|T^{\nu}\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2}\\ &\leq \left\|T^{\nu}\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2}\\ &\leq (1-\nu)^{1-\nu}\nu^{\nu}\lambda^{\nu}\left\|f_{0}\right\|_{\mathcal{L}_2}.\end{aligned}$$

Lemma 3.6 *Lemma A3. When* $0 \leq \nu \leq 1/2$,

$$\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_{n}\right\|_{L_2} = O_{p}\left(\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right).$$

Lemma A3 and the preceding result imply that

$$\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} = O_{p}\left(\lambda^{\nu}+\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right) = O_{p}\left(\lambda^{\nu}\right),$$

provided that $c_1 n^{-2r/(2r+1)} \leq \lambda \leq c_2 n^{-2r/(2r+1)}$ for some constants c_1 and c_2 satisfying $0 < c_1 < c_2 < \infty$.

Recall that

$$\begin{aligned}\left\|T^{1/2}\big(f_{\lambda}-\hat{f}_{\lambda}\big)\right\|_{\mathcal{L}_2} =& \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T\big(f_{\lambda}-f_{0}\big)\right\|_{\mathcal{L}_2}\\ &+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_{n}-T\big)\big(f_{\lambda}-f_{0}\big)\right\|_{\mathcal{L}_2}\\ &+\lambda\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}f_{\lambda}\right\|_{\mathcal{L}_2}+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}g_{n}\right\|_{\mathcal{L}_2}\\ &+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T-T_{n}\big)\big(f_{\lambda}-\hat{f}_{\lambda}\big)\right\|_{\mathcal{L}_2},\end{aligned}$$

so we can bound $\left\|T^{1/2}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|$ by bounding the five terms on the right-hand side of the above equation. By Lemma A1,

$$\begin{aligned}\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2} &\leq \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T^{1/2}\right\|_{\mathrm{op}}\left\|T^{1/2}\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2}\\ &\leq \frac{1}{2}\lambda^{1/2}\left\|f_{0}\right\|_{\mathcal{L}_2}.\end{aligned}$$

Lemma 3.7 *Lemma A4. Under the conditions of Lemma A2,*

$$\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\left(T_{n}-T\right)T^{-\nu}\right\|_{op} = O_{p}\left(\left(n\lambda^{1/(2r)}\right)^{-1/2}\right).$$

By Lemmas A1 and A4,

$$\begin{aligned}
&\big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n-T\big)\big(f_\lambda-f_0\big)\big\|_{\mathcal{L}_2}\\
&\qquad\le \big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n-T\big)T^{-\nu}\big\|_{\mathrm{op}}\big\|T^{\nu}\big(f_\lambda-f_0\big)\big\|_{\mathcal{L}_2}\\
&\qquad\le O_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\lambda^{\nu}\big)\\
&\qquad= o_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\big).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n-T\big)\big(f_\lambda-\hat{f}_\lambda\big)\big\|_{\mathcal{L}_2}\\
&\qquad\le \big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n-T\big)T^{-\nu}\big\|_{\mathrm{op}}\big\|T^{\nu}\big(f_\lambda-\hat{f}_\lambda\big)\big\|_{\mathcal{L}_2}\\
&\qquad\le O_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\lambda^{\nu}\big)\\
&\qquad= o_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\big).
\end{aligned}$$

By Lemma A3, $\big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}g_n\big\|_{\mathcal{L}_2}=O_p\left(\big(n\lambda^{1/(2r)}\big)^{-1/2}\right)$.

Finally, together with the fact that $\lambda\big\|T^{1/2}(T+\lambda\mathbf{1})^{-1}f_\lambda\big\|_{\mathcal{L}_2}=O(\lambda)$, we conclude that $\|T^{1/2}(f_\lambda-\hat{f}_\lambda)\|_{\mathcal{L}_2}=O_p(n^{-\frac{2r}{2r+1}})$, as desired. ■

Chapter 4

Semi-functional Smoothed Score Classification in Reproducing Kernel Hilbert Space

This chapter considers the challenge of estimating the smoothed score (SS) classifier, integrating both functional and scalar covariates in order to make predictions for binary responses. The general binary response model has the form [70]

$$Y = \begin{cases} 1, & \text{if } f(\boldsymbol{R}, \boldsymbol{\gamma}) \geq 0, \\ -1, & \text{otherwise}, \end{cases} \tag{4.1}$$

where $Y \in \{-1, 1\}$ is a binary dependent variable, $\boldsymbol{R}$ is a vector of explanatory variables, f is a function that may or may not be known a priori, and $\boldsymbol{\gamma}$ is a vector of parameters whose values must be estimated from observations. In plain settings, model (4.1) is assumed to be linear, i.e., $f(\boldsymbol{R}, \boldsymbol{\gamma}) = \boldsymbol{R}^\top \boldsymbol{\gamma}$, where $\boldsymbol{R}$, $\boldsymbol{\gamma} \in \mathbb{R}^p$.

In recent decades, the accessibility of more intricate and structured datasets has significantly expanded, owing to booming technological innovations. A substantial portion of these datasets falls under the category of functional data. While functional data are frequently observed along a one-dimensional continuum, this framework also encompasses observations in higher-dimensional domains, such as 2D or 3D image data, time-space data, as well as observations conducted on manifolds and other non-Euclidean domains. Typical functional datasets include the Canadian daily temperature over one year, stock price, fMRI medical records, etc. As the high- or

infinite-dimensional structured data contain rich information, leveraging them can greatly enhance the prediction accuracy of a model.

A natural generalization is to incorporate an additional functional predictor $X(t)$ into model (4.1) such that $f(X, \boldsymbol{R}; \boldsymbol{\theta}) = \int X(t)\beta(t)dt + \boldsymbol{R}^\top \boldsymbol{\gamma}$. The partial functional binary response model is therefore presented as:

$$Y = \text{sign}\left(\int X(t)\beta(t)dt + \boldsymbol{R}^\top \boldsymbol{\gamma}\right), \tag{4.2}$$

where $\text{sign}(u) = 1$, if $u \geq 0$, and -1 otherwise. Our goal is to estimate the coefficient $\boldsymbol{\theta} := [\beta(t), \boldsymbol{\gamma}]$ that minimizes the disagreement between Y and the right-hand side of (4.2). Formally, suppose that $\{x_i, \boldsymbol{r}_i, y_i\}_{i=1}^n$ are n i.i.d copies of $(X(\cdot), \boldsymbol{R}, Y)$ that follows an unknown distribution $P = P(X(\cdot), \boldsymbol{R}, Y)$. Then the true value of $\boldsymbol{\theta}$ is defined as

$$\begin{aligned}\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \big\{ & w(-1)P(f(X, \boldsymbol{R}; \boldsymbol{\theta}) \geq 0, Y = -1) \\ & + w(1)P(f(X, \boldsymbol{R}; \boldsymbol{\theta}) < 0, Y = 1)\big\},\end{aligned} \tag{4.3}$$

where $w(\cdot)$ is a prespecified misclassification cost. We estimate $\boldsymbol{\theta}^*$ through minimizing an empirical risk, as (4.3) can be rewritten as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} E\left\{w(Y)L_{01}\left(Yf\left(X, \boldsymbol{R}; \boldsymbol{\theta}\right)\right)\right\}, \tag{4.4}$$

where $L_{01}(u) = \frac{1}{2}\{1 - \text{sign}(u)\}$ is the 0-1 loss.

Our work is primarily driven by the urgent necessity to develop statistical methodologies tailored to handle electronic health records (EHR) data in clinical research, which exhibit diversity, high dimensionality, and inherent structures. Specifically, our focus centers on the analysis of data originating from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a valuable repository of genetic, functional neuroimaging, and clinical data. This data resource enables the development of robust diagnostic tools, facilitating early detection and intervention in Alzheimer's disease (AD). AD is a debilitating and progressive neurological disorder that afflicts millions of individuals worldwide, compounded by the lack of effective treatment options.

With the global population aging, the prevalence of AD is projected to surge, rendering it a significant public health concern.

The dataset under investigation comprises functional observations collected via diffusion tensor imaging (DTI), in conjunction with AD status, mini-mental state examination (MMSE) scores, and other scalar demographic attributes such as age and gender. By harnessing the MMSE scores acquired directly from patients, we can ascertain the minimal clinically important difference for AD, as previously investigated by [9] and [45].

Another significant application of the SS classifier resides in the realm of personalized medicine. Clinical evidence has underscored the considerable heterogeneity in patient responses to identical treatments. Consequently, the task of designing the most effective treatment for each individual has garnered substantial attention and has evolved into a foundational issue in personalized medicine. In the context of estimating optimal individualized treatment regimes (ITR), let's denote a continuous outcome variable as S, where higher values signify improved conditions, and $Y \in \{-1, 1\}$ stands for the assigned treatment. An ITR $D(\cdot)$ is a map into treatment Y so that a patient presenting with $(X, \boldsymbol{R}) = (x, \boldsymbol{r})$ is recommended to receive a specific treatment $D(x, \boldsymbol{r})$. The optimal ITR D^* [204] is defined as the minimizer of

$$E\left[\frac{S}{\pi(Y|X,\boldsymbol{R})}\mathbb{1}(Y \neq D(X,\boldsymbol{R}))\right],$$

where $\mathbb{1}$ is the indicator function, and $\pi(Y \mid X, \boldsymbol{R}) = P(Y = 1|X, \boldsymbol{R})$ is known as the propensity score. Assuming $D(X, \boldsymbol{R})$ takes the form of $\text{sign}(f(X, \boldsymbol{R}; \boldsymbol{\theta}))$ in model (4.2), the optimal ITR is reduced to the problem (4.3) with the weight $w(Y) = S/\pi(Y|X, \boldsymbol{R})$.

As indicated above, estimating $\boldsymbol{\theta}^*$ defined in (4.3) can be accomplished through minimizing an empirical risk function. In light of the formulation in (4.4), the problem we want to address is closely related to classification problems [5], but functional covariates are taken into consideration in our setting. There exists wide literature on

functional data classification. For regression-based functional classification models, popular approaches are among functional generalized linear models [74, 122, 123], for instance, the logistic regression dealing with a binary response and a functional covariate. An alternative to regression-based functional classification stands on linear or quadratic discriminant analysis methods. [34, 35] established the theoretical support for asymptotically vanishing misclassification rates for them. [191] further extended linear discriminant analysis to high-dimensional functional covariates through an appropriate regularization. Interested readers may refer to [182] for a more detailed review.

Another class of classifiers is constructed through minimizing an empirical risk function, for instance, large margin classifiers, which are more relevant to our work. Among these classifiers, the support vector machines (SVM) [32, 174], have been extensively studied. The non-asymptotic error bound was established in Blanchard et al. [10]. Multiple variants of SVM have been proposed, including the weighted SVM [40, 103], robust truncated SVM [139, 188], and distance-weighted discrimination (DWD) classifier [112, 177]. Recently Sang et al. [158] extended the DWD classifier to functional data and developed its theoretical properties.

Despite their wide success, it is essential to remain aware of potential limitations of the SVM. Firstly, SVM classifiers exhibit sensitivity to noisy training data [165, 188]. When outliers are significantly distant from their respective class boundaries in the training data, SVM classifiers are highly influenced by these points due to the unbounded hinge loss. Another noticeable drawback is that the number of support vectors can be quite large, especially in problems with numerous scalar covariates, not to mention the infinite-dimensional functional data in our case. Fitting an SVM classifier with a substantial number of support vectors can be time-consuming.

We propose the smoothed score approach to address the aforementioned issues associated with SVM classifiers, by utilizing a bounded smooth loss function inspired by [46] and [71]. Estimating the effect of the functional covariate, which is represented

as $\beta(t)$ in model (4.2), entails dimension reduction, as $\beta(t)$ is an infinite-dimensional parameter. In the literature, functional principal component analysis (FPCA) is a popular dimension reduction method. Achieving a good estimate of the true FPCs of X is essential to ensure desirable theoretical properties of an estimator of β. However, this entails stringent assumptions on the covariance function of X [61]. In contrast, using the reproducing kernel Hilbert space (RKHS) framework can circumvent such assumptions; see [198] and [16] for more detailed discussions. Moreover, FPCA may introduce artificial irregularities in the estimate due to the discontinuous roughness control through the choice of the number of retained FPCs. Under the RKHS framework, we impose a penalty on the roughness of β, which leads to a continuous control with a smoothing parameter.

We would like to highlight that model (4.2), is closely related to, but fundamentally different from, the standard classification task whose main objective is to accurately predict the class label Y based on $f(X, R; \boldsymbol{\theta})$. This distinction can be explained by the fact that popular classification techniques, for instance, SVM and AdaBoost, often aim to minimize an empirical risk using a convex upper bound for the L_{01} loss considered in (4.4) [5]. Under certain conditions, the minimizer of these surrogate losses may agree with the Bayes rule, which is defined as the minimizer of the L_{01} loss. However, as argued in [46], using such surrogate loss functions may lead to an unreliable estimate of $\boldsymbol{\theta}^*$. Besides the prediction accuracy, we are also interested in constructing a consistent estimator, which entails an appropriate surrogate loss function whose minimizer coincides with $\boldsymbol{\theta}^*$.

In this chapter, we introduce a class of loss functions designed to meet the aforementioned needs. The estimator of $\boldsymbol{\theta}^*$ is obtained from minimizing a regularized empirical risk with this class of loss functions. We develop an efficient algorithm to solve the minimization problem.

The Fisher consistency and the generalization capacity of our SS estimator are established. A notable aspect of our work is its generality compared to [46], as our

model encompasses both functional and scalar covariates, thereby enhancing prediction accuracy and interpretability. The generalization theory quantifies the optimal classification performance concerning the training sample size and the class of candidate decision functions. It not only provides insights into why the SS classifier is expected to yield high-accuracy performance but also elucidates the trade-off between the choice of the tuning parameter and the size of the candidate function class.

As the other significant contribution of this work, we also establish the near optimal minimax rate in prediction, aligning the results in [16] and [202]. The rate, depending on a bandwidth parameter, is established by using the Rademacher complexity, which in turn is characterized by the eigenvalues of a certain compact operator defined through the RKHS and the covariance operator of the functional covariate.

On the computational front, we develop a proximal gradient algorithm to tackle the non-convexity of the SS loss function, ensuring benign estimation within a convex constraint set. Extensive simulations and a real-world data example demonstrate superior prediction and estimation performance compared to some widely used classifiers.

The remainder of the paper is organized as follows. In Section 4.1, we introduce the RKHS-based smoothed score classifier for the functional partial linear model. Theoretical properties of the proposed classifier are established in Section 4.2. We carry out extended simulation studies and a real data application in Section 4.3 to investigate the finite sample performance of the SS functional classifier. We conclude this article in Section 4.4. All technical proofs are provided in the Supplementary Material.

4.1 Methodology

4.1.1 Smoothed score loss

Let $\mathcal{K}(t)$ be a symmetric kernel function satisfying $\int \mathcal{K}(t)dt = 1$, and $h > 0$ be a bandwidth hyperparameter. We consider a class of smoothed score (SS) loss functions defined as

$$G_h(u) = G(u/h) = \int_{u/h}^{\infty} \mathcal{K}(t)dt = \int \mathbb{1}\{u \leq th\}\mathcal{K}(t)dt. \tag{4.5}$$

The function G_h serves as a differentiable surrogate of the L_{01} loss to make the minimization problem in (4.4) tractable [46], as the L_{01} loss is known to be NP-hard to optimize. It converges to $\mathbb{1}(u < 0)$ for any fixed $u \neq 0$ as the bandwidth parameter h shrinks towards 0.

As illustrated in [46], $G_h(\cdot)$ falls into a quite broad and inclusive class of surrogate loss functions. It can be generalized to the ramp loss, ψ-loss [165], or other variants, for example, the truncated hinge loss [188]. The key to its success resides in incorporating the bandwidth parameter h in the objective function and shrinking it towards 0 with a proper rate.

Given the surrogate loss G_h, we then define the surrogate risk function

$$S_h(\boldsymbol{\theta}) = E[w(Y)G_h(Yf(X, \boldsymbol{R}; \boldsymbol{\theta}))] \tag{4.6}$$

and its empirical counterpart

$$S_{nh}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} w(Y_i)G_h(Y_i f(X_i, \boldsymbol{R}_i; \boldsymbol{\theta})).$$

Our goal is to estimate the optimal slope function $\beta_0(t)$ and the coefficient vector $\boldsymbol{\gamma}_0$ that minimize the smoothed score function S_h through minimizing S_{nh}.

The SS loss borrows the idea from the smoothed maximum score estimator, which was introduced by Horowitz [71] for the linear binary response model (4.1). Under the assumption that the median of noise random variable U conditional on predictors is 0, it is the binary-response analog of the least-absolute-deviation estimator of a linear

median regression model. Since the heteroskedasticity of U with an unknown form is accommodated in the estimation, one does not need to know the form of relation between predictors X and the distribution of U. Therefore, the smoothed maximum score estimator is in a "model-free" regime.

4.1.2 RKHS and the penalized estimator

We assume that $\beta_0(\cdot)$ resides in an RKHS $\mathcal{H}$, a subspace of the collection of square-integrable functions on $\mathcal{I}$. To achieve desirable smoothness for the estimator, we incorporate a penalty term $J(\beta)$ as a regularization on the complexity of β, which is commonly defined through a squared norm or semi-norm associated with $\mathcal{H}$.

Without loss of generality we assume $\mathcal{I} = [0, 1]$ and take $\mathcal{H}$ as the Sobolev space of order m, which is defined as

$$\mathcal{W}_2^m([0,1]) = \{f : [0,1] \to \mathbb{R}, f, f^{(1)}, \ldots, f^{(m-1)} \text{are absolutely continuous and } f^{(m)} \in L_2[0,1]\}.$$

According to Chapter 2.3 in [57], $\mathcal{W}_2^m([0,1])$ is an RKHS when endowed with the (squared) norm

$$\|f\|^2_{\mathcal{W}_2^m} = \sum_{l=0}^{m-1} \left\{ \int_0^1 f^{(l)}(t)dt \right\}^2 + \int_0^1 \left\{ f^{(m)}(t) \right\}^2 dt.$$

Write $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{\beta : J(\beta) = 0\}$ is the null space of $J(\beta)$. Then $\mathcal{H}_0$ is a finite dimensional space with basis functions $\{\phi_1, \phi_2, \cdots \phi_m\}$, with $\dim(\mathcal{H}_0) = m$. Accordingly, we find the orthogonal complement of $\mathcal{H}_0$ in $\mathcal{H}$, denoted by $\mathcal{H}_1$, which also forms an RKHS with reproducing kernel $K(\cdot, \cdot) : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$. The reproducing kernel satisfies $J(\beta) = \|\beta\|_K^2 = \|\beta\|_{\mathcal{H}}^2$ for any $\beta \in \mathcal{H}_1$, where the subscript K indicates the correspondence between the inner product and the reproducing kernel.

We estimate $\theta = (\beta_0(t), \boldsymbol{\gamma}_0)$ by minimizing the following penalized empirical smoothed score function:

$$\begin{aligned} S_{nh,\lambda}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} w(y_i) G\left(y_i \left\{ \int x_i(t)\beta(t)dt + \boldsymbol{r}_i \boldsymbol{\gamma} \right\} / h \right) + \lambda J(\beta) \\ &= S_{nh}(\theta) + \lambda J(\beta). \end{aligned} \tag{4.7}$$

The first term of the objective function (4.7) measures how well the classifier fits the data whereas the penalty functional $J(\beta)$ assesses the plausibility of β. In the second term, $\lambda > 0$ is a tuning parameter balancing the fidelity to the data and the plausibility of β. We choose the roughness penalty as $J(\beta) = \int_0^1 \left\{\beta^{(m)}(t)\right\}^2 dt$ in $\mathcal{W}_2^m([0,1])$. To ease the notation, we write $S_\lambda(\theta) = S_{nh,\lambda}(\theta)$ unless confusion may occur in the rest of the chapter.

4.1.3 Estimation and proximal gradient algorithm

As shown in Chapter 2.3 in [57], the reproducing kernel K, which is a nonnegative definite operator on $L_2(\mathcal{I})$, is continuous and square-integrable. Denote $\int_{\mathcal{I}} K(\cdot, s)x(s)ds$ by $(Kx)(\cdot)$ for any $x \in L_2(\mathcal{I})$. The estimation of $\beta_0(t)$ relies on the following theorem.

Theorem 4.1 *Assume $\hat{\beta}_n(t)$ is the solution to (4.7). Then $\hat{\beta}_n(t)$ is a linear combination of the basis functions $\phi_1, \phi_2 \cdots \phi_m$ and representers $Kx_1, Kx_2 \cdots Kx_n$:*

$$\hat{\beta}_n(t) = \sum_{i=1}^{m} d_i \phi_i(t) + \sum_{j=1}^{n} c_j (Kx_j)(t), \tag{4.8}$$

where $\mathbf{d} = (d_1, d_2, \cdots d_m)^\top \in \mathbb{R}^m$, and $\mathbf{c} = (c_1, c_2, \cdots c_n)^\top \in \mathbb{R}^n$.

Theorem 4.1 is a generalization of the well-known representer lemma for smoothing splines [175]. It indicates that the solution to the minimization problem (4.7) has a finite expansion based on $\{K(\cdot, x_j) : j = 1, \ldots, n\}$ and $\{\phi_i : i = 1, \ldots, m\}$. Therefore, the estimation of infinite dimensional $\hat{\beta}_n(t)$ is converted to the estimation of finite-dimensional coefficients $\boldsymbol{d}$ and $\boldsymbol{c}$. We have by Theorem 4.1 that

$$\int_{\mathcal{I}} X(t)\hat{\beta}_n(t)dt = \sum_{i=1}^{m} d_i \int_{\mathcal{I}} X(t)\phi_i(t)\, dt + \sum_{j=1}^{n} c_j \int_{\mathcal{I}}\int_{\mathcal{I}} X(t)K(s,t)X_j(s)\, ds\, dt.$$

Let $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^\top$, T denote the $n \times m$ matrix with the (i,j)th entry $T_{ij} = \int_{\mathcal{I}} X_i(t)\phi_j(t)\, dt$ for $i = 1, \ldots, n, j = 1, \ldots, m$, and $\boldsymbol{T}_i$ denotes the ith row of matrix T. Similarly, let Σ be the $n \times n$ matrix with the (i,j)th entry $\Sigma_{ij} = \int_{\mathcal{I}}\int_{\mathcal{I}} X_i(t)K(s,t)X_j(s)\, ds\, dt$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$, and $\boldsymbol{\Sigma}_i$ be the ith row

of matrix Σ. It follows from the reproducing property that

$$J(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \int_{\mathcal{I}}\int_{\mathcal{I}} X_i(t)K(s,t)X_j(s)\,ds\,dt = \boldsymbol{c}^\top \Sigma \boldsymbol{c}.$$

With a slight abuse of the notation, we aim to minimize the following empirical smoothed score function:

$$Q(\boldsymbol{\zeta}) = \frac{1}{n}\sum_{i=1}^{n} w(y_i) G\left(y_i\left(\mathbf{T}_i\boldsymbol{d} + \boldsymbol{\Sigma}_i\boldsymbol{c} + \boldsymbol{r}_i\boldsymbol{\gamma}\right)/h\right) + \lambda \mathbf{c}^\top \Sigma \mathbf{c} \tag{4.9}$$

to estimate $\boldsymbol{\zeta} := [\boldsymbol{\gamma}, \boldsymbol{d}, \boldsymbol{c}]$.

Though the nonsmoothness of L_{01} loss is tackled by utilizing the differentiable G_h loss, $Q(\boldsymbol{\zeta})$ is inherently nonconvex. Thus retrieving the global minimizer of $Q(\boldsymbol{\zeta})$ is yet intractable. We apply the proximal algorithm [125] to iteratively estimate $\boldsymbol{\zeta}$ by minimizing a sequence of quadratic approximation of $Q(\boldsymbol{\zeta})$ over a convex constraint set[1] Ω. The iterating steps are presented as follows until a stopping criterion is met:

$$\begin{aligned}\boldsymbol{\zeta}^{k+1} &= \arg\min_{\boldsymbol{\zeta}\in\Omega}\left\{S_{nh}\left(\boldsymbol{\zeta}^k\right) + \left\langle \nabla S_{nh}\left(\boldsymbol{\zeta}^k\right), \left(\boldsymbol{\zeta} - \boldsymbol{\zeta}^k\right)\right\rangle + \frac{1}{2\eta}\left\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^k\right\|_2^2 + \lambda \mathbf{c}^\top \Sigma \mathbf{c}\right\} \\ &= \arg\min_{\boldsymbol{\theta}\in\Omega}\left\{\frac{1}{2\eta}\left\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^k + \eta\nabla S_{nh}\left(\boldsymbol{\zeta}^k\right)\right\|_2^2 + \lambda \mathbf{c}^\top \Sigma \mathbf{c}\right\},\end{aligned} \tag{4.10}$$

where η is the step size to be specified later.

Due to the nonconvexity of $S_{nh}(\boldsymbol{\zeta})$, there may exist multiple local minimizers of $Q(\boldsymbol{\zeta})$. To further regularize the sequence of estimators, we force $\boldsymbol{\zeta}^{k+1}$ to stay in the set Ω in each iteration in (4.10), where we assume $Q(\boldsymbol{\zeta})$ is well behaved in the sense of [185].

Remark 4.1 *To ensure appealing performance of the SS classifier, we require the following restricted strong convexity (RSC) and restricted smoothness (RSM) [46] hold over Ω. Particularly, there exists a set $\Omega = \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\|_2 \le B\}$ for some B which*

[1]This Ω will further align with $\bar{\Omega}$ in section 4.2 to guarantee its theoretical properties

may increases with respect to n*, such that* $\boldsymbol{\zeta}^* \in \Omega$*, and for any* $\boldsymbol{\zeta}, \boldsymbol{\zeta}' \in \Omega$*,*

$$S_{nh}(\boldsymbol{\zeta}') \geq S_{nh}(\boldsymbol{\zeta}) + \nabla S_{nh}(\boldsymbol{\zeta})^\top (\boldsymbol{\zeta}' - \boldsymbol{\zeta}) + \frac{1}{2}\rho_- \|\boldsymbol{\zeta}' - \boldsymbol{\zeta}\|_2^2$$

and

$$S_{nh}(\boldsymbol{\zeta}') \leq S_{nh}(\boldsymbol{\zeta}) + \nabla S_{nh}(\boldsymbol{\zeta})^\top (\boldsymbol{\zeta}' - \boldsymbol{\zeta}) + \frac{1}{2}\rho_+ \|\boldsymbol{\zeta}' - \boldsymbol{\zeta}\|_2^2,$$

where $0 < \rho_- \leq \rho_+ < +\infty$ *are two constants.*

This condition is crucial to the high statistical and computational efficiency of the proximal gradient algorithm [46, 185]. The ρ_--strongly convexity and ρ_+-smoothness of the empirical risk is assumed in the set Ω. In general, as $n \to \infty$, the empirical risk $S_{nh}(\boldsymbol{\zeta})$ is strongly convex and smooth locally in Ω if the corresponding population risk $S_h(\boldsymbol{\zeta})$ is strongly convex and smooth in the same region under mild conditions. If we choose a proper smoothed score function such that $S_h(\boldsymbol{\zeta})$ is twice differentiable, it amounts to saying that the minimal and maximal eigenvalues of the population Hessian $\nabla^2 S_h(\boldsymbol{\zeta})$ are bounded away from zero and infinity for $\boldsymbol{\zeta} \in \Omega$ [46].

Remark 4.2 *For our smoothed score classifier, taking* $\Omega = \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\|_2 \leq B\}$ *where* $\|\boldsymbol{\zeta}^*\|_2 \leq B$*, satisfies our need. Moreover, the smoothed score function* G *satisfies (i)* G' *has bounded support on* $[-1, 1]$*, (ii)* $\|G'\|_\infty, \|G''\|_\infty$ *and* $-\int G''(v)vdv > 0$ *are bounded from above by some universal constants. Such functions can be constructed by orthogonal polynomial basis functions [172].*

4.2 Theoretical Properties

In the current section, we develop theoretical results for the statistical performance of the estimated SS classifier. We first build a non-asymptotic upper bound on the generalization ability of the estimated classifier compared with the Bayes classifier, and furthermore, derive the nonstandard statistical convergence rate for the estimator.

Let $\mathcal{X}$ be a subspace of $L^2(\mathcal{I}) \times \mathbb{R}^p$. The tuple $Z = (X, \boldsymbol{R}, Y)$ denotes a random object taking values in $\mathcal{X} \times \{-1, 1\}$, and $\{z_i = (x_i, r_i, y_i)\}_{i=1}^n$ are n i.i.d copies of Z. We shall use $\widetilde{X}$ and $\tilde{x}$ to denote $(X, \boldsymbol{R})$ and (x, r), respectively. Without loss of generality,

we assume that the weight function $w(y)$ is known, and choose $w(y) = 1/P(Y = y)$ in the rest of this section.

4.2.1 Generalization error of the SS classifier

Condition 4.1 *There exists a constant $c \in (0, 1/2)$ such that $c \leq P(Y = 1) \leq 1 - c$.*

Condition 4.1 ensures that the weight function $w(y)$ is bounded away from 0 and infinity. In our learning problem, we examine classification accuracy on inputs outside the training sample via an error function that measures the generalization ability [165]. The error function, denoted as generalization error (GE), is defined as

$$\mathrm{Err}(f) = P(w(Y)Yf(X, \boldsymbol{R}) < 0) = \frac{1}{2}E[w(Y)\{1 - \mathrm{sign}(Yf(\widetilde{X}))\}]. \tag{4.11}$$

Its empirical version, denoted as the empirical GE (EGE), is given by

$$(2n)^{-1}\sum_{i=1}^{n} w(Y_i)\{1 - \mathrm{sign}(Y_i f(\widetilde{X}_i))\}.$$

If we knew the function $p(\tilde{x}) = P(Y = 1|\widetilde{X} = \tilde{x})$, we would be able to identify the optimal classification hyperplane that minimizes equation (4.11):

$$f_B(\tilde{x}) = p(\tilde{x}) - \frac{w(1)}{w(1) + w(-1)}.$$

We measure the learning accuracy of a generic classifier f by the difference between the actual and ideal performances, denoted as $Er(f, f_B) = \mathrm{Err}(f) - \mathrm{Err}(f_B)$. For an arbitrary loss function ℓ other than $L_{01}(u) = \{1 - \mathrm{sign}(u)\}/2$, the associated risk function for a classifier f is defined correspondingly as

$$L(f, f_B) := E[\ell(f) - \ell(f_B)]. \tag{4.12}$$

We next build a non-asymptotic upper bound on the $Er(\hat{f}, f_B)$ as a function of the tuning parameter λ and sample size n, in terms of the bandwidth parameter h and the size of the candidate classification sets $G(\mathcal{F}) = \{G_f = \{\tilde{x} : f(\tilde{x}) \geq 0\} : f \in \mathcal{F}\}$, generated by the function class $\mathcal{F}$ comprising candidate decision functions. The class

$\mathcal{F}$ can depend on sample size n [166]. For our theory, the ideal optimal classification set $G_{f_B} = \{\tilde{x} \in \mathcal{X} : f_B(\tilde{x}) \geq 0\}$ is not confined to $G(\mathcal{F})$. Rather, we assume $\bar{f} = \text{sign}(f_B)$ can be perfectly approximated by $\mathcal{F}$. The theoretical results are built upon the size of $G(\mathcal{F})$, measured by the metric entropy to be defined.

In particular, we take ℓ as S_h defined in equation (4.12), and $L(f, f_B) = \{S_h(f) - S_h(f_B)\}$. We consider the following conditions to study theoretical properties for the estimated SS estimator:

Condition 4.2 *For some positive sequence $s_n \to 0$ as $n \to \infty$, there exists $f_0 \in \mathcal{F}$ such that $L(f_0, \bar{f}) \leq s_n/2$, i.e., $\inf_{f \in \mathcal{F}} L(f, \bar{f}) \leq s_n/2$.*

Condition 4.3 *There exist some constants $0 < \alpha \leq +\infty$ and $c_1 > 0$ such that $P(\tilde{x} \in \mathcal{X} : |f_B(\tilde{x})| \leq \delta) \leq c_1 \delta^\alpha$ for any sufficiently small $\delta \geq 0$.*

Condition 4.3 essentially imposes an Hölder type of regularity that describe the performance of the Bayes hyperplane f_B near the decision boundary $\{\tilde{x} : f_B(\tilde{x}) = 0\}$. To characterize the size of $G(\mathcal{F})$, we provide a formal definition of the metric entropy for any measurable sets. Let $d(\cdot, \cdot)$ denote a distance in $\mathcal{X}$, satisfying $d(G_1, G_2) = \int_{\{G_1 \Delta G_2\}} dP = P(G_1 \Delta G_2)$ for any $G_1, G_2 \in \mathcal{X}$, where $G_1 \Delta G_2 = (G_1 \backslash G_2) \cup (G_2 \backslash G_1)$ denotes the set difference between G_1 and G_2. For a given class $\mathcal{B}$ of subsets of $\mathcal{X}$ and any $\varepsilon > 0$, define an ε-bracketing set of $\mathcal{B}$, denoted as $\{(G_1^l, G_1^u), \ldots, (G_m^l, G_m^u)\}$, by any set that satisfies for any $G \in \mathcal{B}$ there is a j such that $G_j^l \subset G \subset G_j^u$ and $\max_{1 \leq j \leq m} d(G_j^u, G_j^l) \leq \varepsilon$. Then the metric entropy $H(\varepsilon, \mathcal{B})$ of $\mathcal{B}$ with bracketing is defined as the logarithm of the cardinality of an ε-bracketing set of $\mathcal{B}$ of the smallest size.

Define

$$\begin{aligned} \mathcal{G}(k) &= \{G_f = \{\tilde{x} : f(\tilde{x}) \geq 0\} : f \in \mathcal{F}, J(f) \leq k\} \\ &\subset \mathcal{G}(\mathcal{F}) = \{G_f = \{\tilde{x} : f(\tilde{x}) \geq 0\} : f \in \mathcal{F}, J(f) < +\infty\}, \end{aligned}$$

where $J(f)$ is the penalty term $\|\beta\|_{\mathcal{H}}^2$ as defined previously and $J_0 = \max(J(\beta_0), 1)$.

Condition 4.4 *For some positive constants c_2, c_3 and c_4, define*

$$\phi(\varepsilon_n, k) = \int_{c_4 L}^{c_3^{1/2} L^{\frac{\alpha}{2(\alpha+1)}}} H^{1/2}\left(u^2/2, \mathcal{G}(k)\right) du/L,$$

where $L = L(\varepsilon_n, \lambda, k) = \min(\varepsilon_n^2 + \lambda J_0(k/2-1), 1)$. There exists some $\varepsilon_n > 0$ such that

$$\sup_{k \geq 1} \phi(\varepsilon_n, k) \leq c_2 n^{1/2}.$$

In our case, $c_2 = 2^{-23/2}$, $c_3 = \max\{2^{(1+2\alpha)/\alpha}[4(4c_1)^{\frac{1}{\alpha+1}}+1]+2^{(1+\alpha)/\alpha}, 8\}$, and $c_4 = 2^{-6}$.

The following lemma indicates that our SS classifier approximates the Bayes classifier $\bar{f}(\tilde{x}) = \text{sign}(f_B(\tilde{x}))$. This constitutes a crucial feature of the SS loss. It demonstrates the feasibility of achieving optimal classifier f_B by replacing the L_{01} loss with the S_h function.

Lemma 4.2 *For any G defined in equation (4.5), the Bayes classifier $\bar{f}(\tilde{x}) = \text{sign}(f_B(\tilde{x}))$ minimizes both $ES_h(f(\widetilde{X}))$ and $Err(f(\widetilde{X}))$. That is, for any f, we have*

$$ES_h(f(\widetilde{X})) \geq ES_h(\bar{f}(\widetilde{X})), \quad \textit{and}$$
$$E\left[\frac{1}{2}w(Y)\{1 - \text{sign}(Y\bar{f}(\widetilde{X}))\}\right] \leq E\left[\frac{1}{2}w(Y)\{1 - \text{sign}(Yf(\widetilde{X}))\}\right], \quad \textit{as } h \to 0.$$

Remark 4.3 *The minimizers for $ES_h(f(\widetilde{X}))$ and $E[\frac{1}{2}w(Y)(1 - \text{sign}(Yf(\widetilde{X})))]$ are not necessarily unique. For example, cf_B is also a minimizer for both quantities for any constant $c \geq 1$. By Lemma 4.2 and condition 4.2, we could choose sufficiently small h_n, such that $Er(f_0, f_B) \leq L(f_0, \bar{f}) + s_n/2 \leq s_n$.*

Theorem 4.3 *Suppose that Conditions 4.1 - 4.4 are met. There exists a constant $c_5 > 0$, such that the estimated SS classifier $\text{sign}(\hat{f})$ satisfies*

$$P\left(Er(\hat{f}, f_B) \geq \delta_n^2\right) \leq 3.5 \exp\left(-c_5 n \lambda^{\frac{\alpha+2}{\alpha+1}} J_0^{\frac{\alpha+2}{\alpha+1}}\right),$$

provided that $h \to 0$ and $\lambda^{-1} \geq 2\delta_n^{-2} J_0$, where $\delta_n^2 = \min(\max(\varepsilon_n^2, s_n), 1)$.

Corollary 4.4 *Under the conditions of Theorem 4.3 ,*

$$\left|Er(\hat{f}, f_B)\right| = O_p\left(\delta_n^2\right),$$
$$E\left|L(\hat{f}, f_B)\right| = O\left(\delta_n^2\right),$$

provided that $n\left(\lambda J_0\right)^{\frac{\alpha+2}{\alpha+1}}$ *is bounded away from 0.*

Theorem 4.3 and Corollary 4.4 illustrate the trade-off between the magnitude of λ and the risk bounds on $Er(\hat{f}, f_B)$; the best performance is achieved when λ gives the best balance between the size of $G(\mathcal{F})$ and n. In applications, we need to verify that the assumptions are satisfied for $s_n \to 0$ and $\varepsilon_n \to 0$, and then choose the optimal δ_n. The optimal λ that yields the best rate δ_n^2 for the SS classifier is determined by two constrains: one is $\lambda \leq (2J_0)^{-1}\delta_n^2$ and the other is $J_0\lambda = O\left(n^{-\frac{\alpha+1}{\alpha+2}}\right)$. To entertain bothe of them, we choose λ of order of $(J_0)^{-1}\delta_n^2$. Consequently, if $\alpha = \infty$, then the bound in Theorem 4.3 becomes $3.5\exp\left(-c_5 n\delta_n^2\right)$; if $\alpha \to 0$, then the bound reduces to $3.5\exp\left(-c_5 n\delta_n^4\right)$.

4.2.2 Convergence rate of SS estimator

We will establish convergence rate of SS estimator based on surrogate loss $G_h(\cdot)$. Specifically, the following result highlight that the estimated coefficient will coincide with $\boldsymbol{\theta}^*$ in (4.3) when $h \to 0^+$.

Proposition 4.5 *For any coefficient* $\boldsymbol{\theta}$*, if* $\tilde{\boldsymbol{\theta}}$ *minimizes* $S_h(\boldsymbol{\theta})$ *in (4.6), then* $\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}}$ *when* $h \to 0^+$.

With a little abuse of notation, we define β^* and $\boldsymbol{\gamma}^*$ to minimize the smoothed score risk (4.6), recall that β^* resides in a RKHS $\mathcal{H}$ and $\boldsymbol{\gamma}^* \in \mathbb{R}^p$. By Proposition 4.5, we have

$$(\beta^*, \boldsymbol{\gamma}^*) = \arg\min_{\boldsymbol{\theta} \in \mathcal{H}\times\mathbb{R}^p} E[w(Y)G_h(Yf(X, \boldsymbol{R}))].$$

Our penalized SS estimator can be written as

$$(\hat{\beta}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\beta\in\mathcal{H}, \boldsymbol{\gamma}} S_{nh}(\beta, \boldsymbol{\gamma}) + \lambda J(\beta), \tag{4.13}$$

where $J(\beta) = \|\beta\|_K^2$, and

$$S_{nh}(\beta, \boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} w(y_i) G\left(y_i\left(\int x_i(t)\beta(t)dt + \boldsymbol{r}_i\boldsymbol{\gamma}\right)/h\right).$$

Note that $\|\cdot\|_K$ is a norm of $\mathcal{H}_K$ associated with a positive definite kernel K in compact set $\mathcal{I}\times\mathcal{I}$, which is deduced by RKHS inner product $\langle\cdot,\cdot\rangle_K$. Since $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, we write $\beta = \beta_0 + \beta_1$ with $\beta_0 \in \mathcal{H}_0$ and $\beta_1 \in \mathcal{H}_1$, for any $\beta \in \mathcal{H}$. So, we have

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\gamma}}) = \arg\min_{\beta_0\in\mathcal{H}_0, \beta_1\in\mathcal{H}_1, \boldsymbol{\gamma}} S_{nh}(\beta_0, \beta_1, \boldsymbol{\gamma}) + \lambda\|\beta_1\|_K^2, \tag{4.14}$$

and

$$S_{nh}(\beta_0, \beta_1, \boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} w(y_i) G\left(y_i\left(\langle x_i, \beta_0+\beta_1\rangle_{L_2} + \boldsymbol{r}_i\boldsymbol{\gamma}\right)/h_n\right).$$

Although the SS loss function avoids the discontinuity of the 0-1 loss, it is generally nonconvex. Therefore, the global solution to (4.13) or (4.14) is still intractable. To address this issue, we restrict true value $(\beta_0^*, \beta_1^*, \boldsymbol{\gamma}^*)$ and the local minimizer to a convex set $\bar{\Omega} = \left\{(\beta_0, \beta_1, \boldsymbol{\gamma}) : \|\beta_0\|_{L_2} \le R, \|\beta_1\|_{L_2} \le R, \|\boldsymbol{\gamma}\|_2 \le R\right\}\}$ for some $R > 0$.

One key feature of RKHS is the reproducing property, which implies that $\beta_1(t) = \langle\beta_1, K_t\rangle_K$ for any $\beta_1 \in \mathcal{H}_k$ and $t \in \mathcal{I}$, where $K_t(\cdot) = K(t,\cdot)$. With abuse of notation, denote also by K the linear operator, that is, $K : L_2 \to \mathcal{H}_1$ satisfying $K\beta = \int K(\cdot, s)\beta(s)ds \in \mathcal{H}_1$ for $\beta \in L_2$. Since $\mathcal{H}_1$ is identical to the range of $K^{1/2}$, the square-root operator of K. Define the covariance operator of X as $\Gamma = E(X\otimes X)$, which is a linear operator, where $f\otimes g : L_2 \to L_2$, $(f\otimes g)h = \langle g, h\rangle f$ for $f, g \in L_2$, and any $h \in L_2$.

Define $T = K^{1/2}\Gamma K^{1/2}$, which will play a critical role in the theoretical analysis. Assuming $E\|X\|^4 < \infty$, then Γ and T are compact operators. By the Mercer's Theorem, we have a spectral expansion for T:

$$T = \sum_{j=1}^{\infty} u_j\psi_j \otimes \psi_j,$$

where $u_1 \ge u_2 \ge \cdots \ge 0$ are the eigenvalues with $u_j \to 0$ and $\{\psi_j\}$ are their orthonormalized eigenfunctions. In addition, we consider the following Rademacher

complexity

$$E\left[\sup_{(\beta_0,\beta_1,\boldsymbol{\gamma})\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\langle x_i,\beta_0+\beta_1\rangle_{L_2}+\boldsymbol{r}_i\boldsymbol{\gamma}\right)\right],$$

where $\mathcal{F}=\{(\beta_0,\beta_1,\boldsymbol{\gamma}):\left\|\Gamma^{1/2}\beta_0\right\|_{L_2}\leq v_1,\left\|\Gamma^{1/2}\beta_1\right\|_{L_2}\leq v_1,\|\beta_1\|_{\mathcal{H}_K}\leq 1,\|\boldsymbol{\gamma}\|_2\leq v_2\}$, and $\sigma_i\in\{-1,1\}$ are i.i.d. Rademacher variables. In the Proof section, we will show the bound on the above Rademacher complexity depends on

$$\mathcal{R}(v)=\left\{\frac{1}{n}\sum_{j=1}^{n}\min(u_j,v^2)\right\}^{1/2}.$$

To establish the convergence rate of the estimator, we further impose the following conditions.

Condition 4.5 *Both X and components of $\boldsymbol{R}$ are sub-Gaussian.*

Condition 4.6 *$E\left\{[\langle X,\beta\rangle+\boldsymbol{R}\boldsymbol{\gamma}]^2\right\}\asymp E\left[\langle X,\beta\rangle^2+(R\boldsymbol{\gamma})^2\right]$ for all $\beta\in\mathcal{H}$, and $\boldsymbol{\gamma}\in\mathbb{R}^p$, where $a\asymp b$ means $0<C_1\leq a/b\leq C_2<\infty$. Additionally, $E(\boldsymbol{R}^T\boldsymbol{R})$ has eigenvalues bounded away from zero and infinity.*

Condition 4.7 *The eigenvalues $u_j\leq Cj^{-\mu}$ for some $\mu>1$.*

Condition 4.8 *There exists a set $\widetilde{\Omega}=\left\{(\beta,\boldsymbol{\gamma}):\|\beta\|_{\mathcal{H}}^2+\|\boldsymbol{\gamma}\|_2^2\leq B\right\}$ for some B, such that $(\beta^*,\boldsymbol{\gamma}^*)\in\widetilde{\Omega}$, and for any $(\beta,\boldsymbol{\gamma})\in\widetilde{\Omega}$,*

$$\bar{S}_h\left([\langle X,\beta\rangle+\boldsymbol{R}\boldsymbol{\gamma}]\right)-\bar{S}_h([\langle X,\beta^*\rangle+\boldsymbol{R}\boldsymbol{\gamma}^*])\geq\frac{1}{2}\rho_-^*E\left[\langle X,\beta-\beta^*\rangle+\boldsymbol{R}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right]^2,$$

where $0<\rho_-^*<+\infty$ is a constant, and

$$\bar{S}_h\left([\langle X,\beta\rangle+\boldsymbol{R}\boldsymbol{\gamma}]\right):=E\left[w(Y)G_h(Y\left([\langle X,\beta\rangle+\boldsymbol{R}\boldsymbol{\gamma}]\right))\right].$$

Condition 4.9 *The bandwidth h_n, which is dependent on n, satisfies $h_n\to 0$ and $h_n^{(1-\nu)}n^{\frac{\mu}{2(\mu+1)}}\to\infty$ as $n\to\infty$ for some $0<\nu<1$.*

Remark 4.4 *The sub-Gaussian assumption in Condition 4.5 is frequently adopted in high-dimensional analysis for applying concentration inequalities, and this condition*

holds if X is a Gaussian process and $\boldsymbol{R}$ follows a multivariate normal distribution. For Condition 4.6, the first part indicates that the correlation between X and $\boldsymbol{R}$ is weak. Further, under Condition 4.6, we have

$$E\left\{[\langle X,\beta\rangle+\boldsymbol{R}\boldsymbol{\gamma}]^2\right\}\asymp E\left[\langle X,\beta_0\rangle_{L_2}^2+\langle X,\beta_1\rangle_{L_2}^2+\|\boldsymbol{\gamma}\|_2^2\right], \tag{4.15}$$

for all $\beta=\beta_0+\beta_1\in\mathcal{H}_0\oplus\mathcal{H}_1$, and $\boldsymbol{\gamma}\in\mathbb{R}^p$.

Condition 4.7 is also considered in [198], which specifies the smoothness of the linear operator T. The eigenvalues of the reproducing kernel in some Sobolev satisfy this condition; interested readers may refer to [198] and [163]. Condition 4.8 essentially serves as a qualification of the identifiability condition of the objective function at local minimum $(\beta^*,\boldsymbol{\gamma}^*)$. Moreover, Condition 4.8 renders $\bar{S}_h$ to be restricted strong convex at the local minimum $(\beta^*,\boldsymbol{\gamma}^*)$, as $\nabla\bar{S}_h([\langle X,\beta^*\rangle+\boldsymbol{R}\boldsymbol{\gamma}^*])=0$. Condition 4.9 requires that the bandwidth used in G_h for defining the SS loss function cannot decay too fast with respect to n.

Theorem 4.6 *Assume that Conditions 4.1 and 4.5-4.9 hold. If $\lambda\asymp h_n^{-(1-\nu)}n^{-\frac{\mu}{\mu+1}}$ is met for some $0<\nu<1$, we have*

$$\left\|\Gamma^{1/2}(\hat{\beta}_0-\beta_0^*)\right\|_{L_2}+\left\|\Gamma^{1/2}(\hat{\beta}_1-\beta_1^*)\right\|_{L_2}+\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2=O_p\left(h_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}\right).$$

Further,

$$\left\|\Gamma^{1/2}(\hat{\beta}-\beta^*)\right\|_{L_2}+\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2=O_p\left(h_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}\right).$$

Remark 4.5 *The rate $h_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}$ is typically slower than the optimal convergence rate in kernel smoothing. For example, the rate is $n^{-\frac{\mu}{2(\mu+1)}}$ for functional linear mean regression in [16], $n^{-\frac{\mu}{2(\mu+1)}}\log n$ for partially linear functional quantile regression in [192], and $n^{-\frac{\mu(1+2r)}{2(\mu(1+2r)+1)}}$ with some $r\in[0,1/2]$ for functional partially linear support vector machine in [202]. Letting $\nu=1/2$, $h_n\asymp n^{-\frac{1}{2(\mu+1)}}$, we have $h_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}=n^{-\frac{2\mu-1}{4(\mu+1)}}$, which is close to the rate in [16] and [202].*

Remark 4.6 *From Theorem 4.6 and the second part of Condition 4.6, the convergence rate of the prediction risk can be established as*

$$E^\star\left[\left(\langle X^\star, \hat{\beta}-\beta^*\rangle + \boldsymbol{R}^\star(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\right)^2\right] = O_p\left(n^{-\frac{2\mu-1}{2(\mu+1)}}\right),$$

by taking $\nu = 1/2$ *and* $h_n \asymp n^{-\frac{1}{2(\mu+1)}}$*, where* $(X^\star, \boldsymbol{R}^\star)$ *is an independent copy of* $(X, \boldsymbol{R})$*, and* $E^\star$ *is the expectation over* $(X^\star, \boldsymbol{R}^\star)$*. It measures the mean squared prediction error for a random future observation on* $(X, \boldsymbol{R})$*. This convergence rate is slightly slower than the optimal rate* $n^{-\frac{2\mu}{2\mu+1}}$ *in [16]. When the linear operator* T *is sufficiently smooth, that is,* μ *is sufficiently large, they asymptotically converge to* $O_p(n^{-1})$*, which attains the parametric rate.*

4.3 Numerical Studies

4.3.1 Simulations

This section showcases the finite-sample performance of the proposed functional SS classifier under various data-generating processes. Each experiment is concerned with the binary response model (4.2) where the discriminant function is defined as

$$f\left(X_i, \boldsymbol{R}_i\right) = \int_0^1 X_i(t)\beta(t)dt + \boldsymbol{R}_i^\top \boldsymbol{\gamma}, \tag{4.16}$$

and $Y = \text{sign}\left(f\left(X_i, \boldsymbol{R}_i\right)\right)$ for $i = 1, \ldots, n$. The functional covariate X is generated in the following way [198]:

$$X_i(t) = \sum_{j=1}^{50} \xi_{ij}\zeta_j\phi_j(t),$$
$$\xi_{ij} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(-\sqrt{3}, \sqrt{3}),$$

where $\zeta_j = (-1)^{j+1}j^{-1}$for $j = 1, \ldots, 50$, and $\phi_1(t) = 1$ and $\phi_j(t) = \sqrt{2}\cos((j-1)\pi t), j \geq 2$ for $t \in [0, 1]$. Observations at 200 equally spaced times points in the interval $[0, 1]$ are made on each sample path of $X(t)$. In (4.16), $\boldsymbol{R} = (R_0, R_1, R_2)^\top$ denotes three scalar covariates with $R_0 = (1, \ldots, 1)^\top$, and (R_1, R_2) independently generated from a truncated normal distribution within the interval $[-2, 2]$.

We consider two choices for the slope function $\beta(t)$. In the first scenario, the slope function is generated as a linear combination of the FPCs of $X(t)$. Particularly, $\beta(t) = \sum_{j=1}^{50} 4(-1)^{j+1} j^{-2} \phi_j(t)$. The coefficient vector of the scalar covariates $\boldsymbol{R}$ takes the values $\boldsymbol{\gamma} = (\alpha, -2, 3)^\top$ or $(\alpha, 0, 0)^\top$ to ensure the discriminant function f depends on the scalar covariates or not. The intercept α is set as 0.01 and 7 to make balanced and imbalanced samples, respectively. In particular, we set $\alpha = 0.01$ to generate the samples where the proportion of $y = -1$, denoted by ρ in Table 4.1, stays around 0.5, whereas $\alpha = 7$ is chosen to generate samples with $\rho = 0.05$. In the second scenario, we have $\beta(t) = e^{-t}$, and $\boldsymbol{\gamma} = (\alpha, -0.5, 1)^\top$ or $(\alpha, 0, 0)^\top$. The intercept α is set to be 0.01 or 3 to differ the value of ρ in a similar manner as in the first scenario.

Besides the proposed functional SS classifiers (SS in short), we also consider other commonly used functional classifiers for comparison, including the functional logistic regression (Logi in short) [38], the support vector machine (SVM in short) [202], and the functional DWD classifier (DWD in short) [158]. For the proposed SS approach, we choose the cumulative distribution function of the standard normal distribution (Gaussian kernel) as the smoothed score function G that satisfies the requirements in Remark 4.2. The optimal tuning parameters λ and bandwidth h_n are selected through five-fold cross-validation.

In each simulation trial, we randomly generate a training set of size $n = 200, 500$, and 700 to fit these classifiers and then evaluate their estimation and prediction accuracy on a test sample of size 500. Additionally, we compare the estimation errors of $\beta(t)$ and $\boldsymbol{\gamma}$ for these functional classifiers. To assess the uncertainty in the estimation and prediction accuracy of each classifier, 200 independent simulation trials are conducted in each scenario.

Tables 4.1 and 4.2 summarize the means and the standard errors of the misclassification rates for each classifier under these two designs of β. Figures 4.1 and 4.2 depict the L_2 estimation errors of $\hat{\beta}$, and l_2 estimation errors of $\hat{\boldsymbol{\gamma}}$ regarding the aforementioned functional classifiers over 200 simulation trials for various sample sizes,

Table 4.1: The mean misclassification errors on the test sample across 200 simulations with the standard errors in parentheses in Scenario 1. The columns ρ and $\boldsymbol{\gamma}$ indicate the approximate proportion between two categories, whether or not the true discriminant function depends on the scalar covariates.

n	ρ	$\boldsymbol{\gamma}$	SS	Logi	SVM	DWD	$\boldsymbol{\gamma}$	SS	Logi	SVM	DWD
200	0.5	(0,0)	0.021	0.026	0.022	0.097	(-2,3)	0.021	0.023	0.020	0.116
			(0.010)	(0.014)	(0.012)	(0.042)		(0.009)	(0.011)	(0.010)	(0.042)
500	0.5	(0,0)	0.014	0.017	0.015	0.092	(-2,3)	0.013	0.016	0.014	0.106
			(0.007)	(0.009)	(0.008)	(0.038)		(0.007)	(0.008)	(0.007)	(0.043)
700	0.5	(0,0)	0.009	0.012	0.012	0.092	(-2,3)	0.008	0.011	0.011	0.116
			(0.005)	(0.005)	(0.006)	(0.041)		(0.004)	(0.005)	(0.005)	(0.043)
200	0.05	(0,0)	0.031	0.028	0.210	0.070	(-2,3)	0.017	0.022	0.169	0.079
			(0.016)	(0.015)	(0.024)	(0.016)		(0.010)	(0.014)	(0.022)	(0.018)
500	0.05	(0,0)	0.018	0.016	0.217	0.069	(-2,3)	0.008	0.012	0.017	0.075
			(0.008)	(0.007)	(0.019)	(0.014)		(0.005)	(0.007)	(0.019)	(0.019)
700	0.05	(0,0)	0.015	0.012	0.217	0.069	(-2,3)	0.007	0.006	0.172	0.078
			(0.006)	(0.005)	(0.017)	(0.015)		(0.004)	(0.004)	(0.018)	(0.019)

respectively.

In general, the value of the sample unbalance agent in the binary response model, denoted by ρ under this context, has a great impact on the performance of the classifiers. Compared with balanced cases ($\rho = 0.5$), the misclassification rates and estimation errors are relatively higher in almost all the unbalanced cases ($\rho = 0.05$). However, the proposed SS classifier addresses this issue quite well through imposing larger class weights for the small class $Y = -1$. In contrast, the alternative classifiers without an appropriate adjustment cannot effectively tackle class imbalance. In particular, fluctuations appear in the SVM and DWD approaches. The functional DWD classifier produces favourable results with a mean misclassification rate around 7% compared with 18% given by the SVM with unbalanced samples. A possible reason might be that the DWD classifier makes use of all observations in the training set, rather than just the support vectors in SVM, to determine the decision boundary

Table 4.2: The mean misclassification errors on the test sample across 200 simulations with the standard errors in parentheses in Scenario 2. The columns ρ and $\boldsymbol{\gamma}$ indicate the approximate proportion between two categories, whether or not the true discriminant function depends on the scalar covariates.

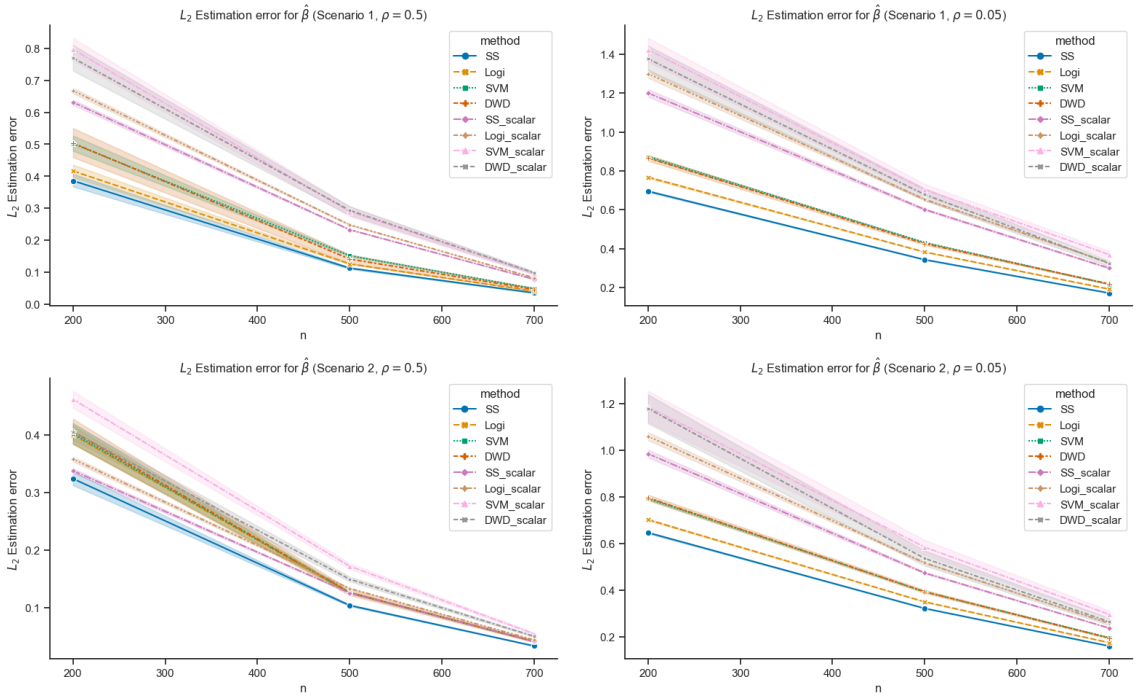
n	ρ	$\boldsymbol{\gamma}$	SS	Logi	SVM	DWD	$\boldsymbol{\gamma}$	SS	Logi	SVM	DWD
200	0.5	(0,0)	0.020	0.021	0.016	0.095	(-.5,1)	0.020	0.019	0.015	0.116
			(0.010)	(0.013)	(0.011)	(0.042)		(0.011)	(0.011)	(0.010)	(0.041)
500	0.5	(0,0)	0.009	0.010	0.009	0.092	(-.5,1)	0.009	0.011	0.009	0.106
			(0.007)	(0.007)	(0.006)	(0.037)		(0.007)	(0.008)	(0.007)	(0.040)
700	0.5	(0,0)	0.004	0.005	0.005	0.091	(-.5,1)	0.003	0.005	0.004	0.110
			(0.003)	(0.003)	(0.003)	(0.039)		(0.003)	(0.004)	(0.003)	(0.040)
200	0.05	(0,0)	0.029	0.023	0.208	0.080	(-.5,1)	0.018	0.020	0.178	0.073
			(0.017)	(0.014)	(0.020)	(0.017)		(0.012)	(0.013)	(0.023)	(0.019)
500	0.05	(0,0)	0.012	0.010	0.209	0.079	(-.5,1)	0.009	0.010	0.180	0.070
			(0.008)	(0.007)	(0.018)	(0.016)		(0.006)	(0.007)	(0.019)	(0.019)
700	0.05	(0,0)	0.008	0.005	0.201	0.078	(-.5,1)	0.006	0.004	0.172	0.071
			(0.005)	(0.004)	(0.018)	(0.017)		(0.004)	(0.003)	(0.017)	(0.016)

[112, 158]; thus it is more resilient in unbalanced training samples. When the simulated data are balanced with $\rho = 0.5$, the SS classifier still yields better results in comparison with alternative methods.

Regarding the effect of scalar covariates, when $\boldsymbol{\gamma} = (\alpha, -2, 3)^\top$ in scenario one or $(\alpha, -0.5, 1)^\top$ in scenario two, the proposed functional SS classifier with scalar covariates is superior to its competitors without incorporating the scalar variables in terms of prediction. This fact demonstrates the importance of accounting for scalar covariates when the true discriminant function indeed depends on them.

Finally, as sample size increases, the prediction and estimation errors will decrease. The simulation results of the SS classifier align with Theorems 4.3 and 4.6, demonstrating the desirable generalization ability and the consistency of the proposed estimator. The functional SS classifier generally outperforms the other three classifiers especially when the sample size is relatively small and when there is a small

Figure 4.1: The mean L_2 estimation errors of $\hat{\beta}$ over 200 simulation samples under four scenarios, with the shade determined by standard deviations. Balanced/Unbalanced represent the class proportion $\rho = 0.5/0.05$, respectively.

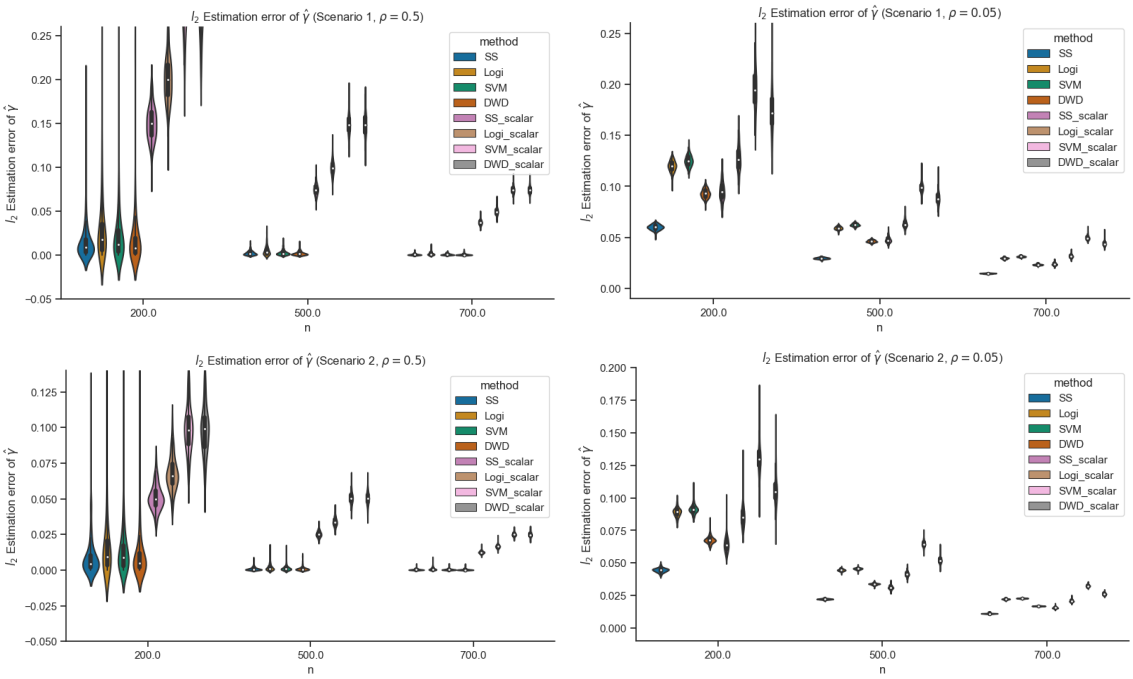


proportion of -1s. This suggests that the functional SS classifier is more efficient in utilizing limited data and less sensitive to class imbalance, rendering it a robust choice under various conditions. Additionally, in terms of accuracy in parameter estimation, we see that the SS approach yields lower estimation errors compared to logistic regression, SVM and DWD under both scenarios of $\beta(t)$. Therefore, the SS functional approach outperforms other methods that do not have statistical guarantees for estimation consistency, and tends to be more reliable than its competitors when parameter estimation and model interpretation become the priority.

4.3.2 Application to ADNI data classification

We apply the proposed SS classifier as well as the alternatives considered in Section 4.3.1 to the ADNI dataset, which consists of $N = 199$ subjects after removing subjects

Figure 4.2: Violin plots of the l_2 estimation errors for estimating $\hat{\gamma}$ over 200 simulation samples under two scenarios. Balanced/Unbalanced represent the class proportion $\rho = 0.5/0.05$, respectively.



with missing values. The data were obtained from the ongoing ADNI study, where researchers are interested in identifying biomarkers of the AD from genetic, structural, functional neuroimaging, and clinical data.

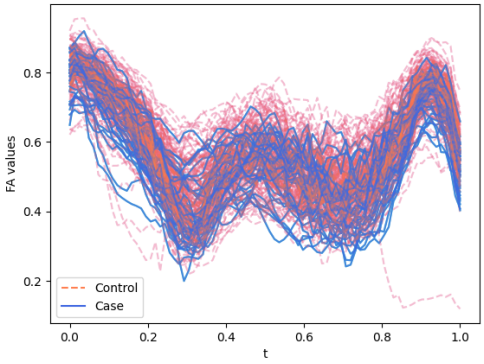


Figure 4.3: FA profiles of the two categories

The ADNI dataset mainly consists of two parts. The first part is the neuroimaging data collected by DTI. Specifically, fractional anisotropy (FA) values were measured

at 83 locations along the corpus callosum (CC) fibre tract for each subject. An illustrative plot of all FA curves, which are treated as $X(t)$ in our analysis, is presented in Figure 4.3. It is shown that there exist differences in FA curves between the control group and the AD group. The CC is the largest fibre tract in the human brain and is a topographically organized structure. It is responsible for a large proportion of communications between the two hemispheres and connects homologous areas in the two cerebral hemispheres. Hence the transferring of visual, motoric, somatosensory, and auditory information is largely ensured by CC. These facts explain why FA values along the CC differ between these two groups. The other part consists of demographic features like gender (a categorical variable), handedness (left hand or right hand, a categorical variable), age, education level, mini-mental state examination (MMSE) scores and AD status.

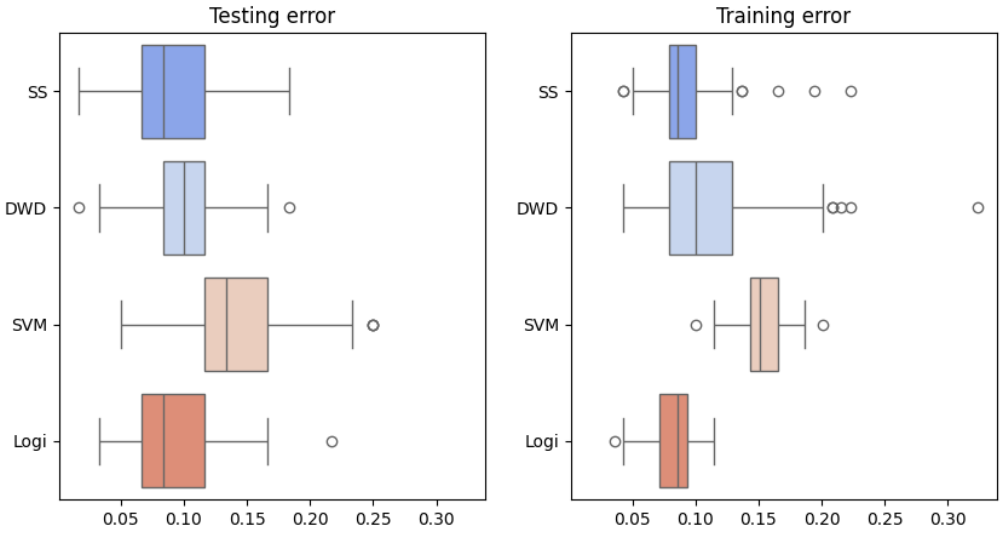


Figure 4.4: Boxplots of the misclassification rates on the test set and training set of the classifiers where 70% of the data is used for training.

The AD status is a categorical variable with three levels: normal control (NC), mild cognitive impairment (MCI), and AD. In our study, we combine the NC and MCI categories into one level for classification, and then this status variable is treated as

a binary response variable in our following analysis. There are $N_1 = 38$ AD patients in our analysis, accounting for 20% of the whole sample.

One can find a more detailed description of the ADNI data at http://adni.loni.usc.edu/. There has been extensive research on this dataset; see [106, 158, 170] and [97] for example. Our main objective is to use the FA trajectories and demographic features to predict the status of AD to investigate the relationship between the progression of AD status and the patient's clinical measurements. The 199 subjects are randomly divided into a training set with $n = 0.7N$ subjects and a testing set with the other $N - n$ subjects. We randomly split the whole dataset into training and testing sets 200 times.

The testing and training errors are summarized in Figure 4.4. The proposed SS classifier shares comparable results with the functional logistic regression in prediction. Moreover, our SS classifier performs better than the SVM and DWD approaches. To sum up, the SS classifier can be a good, even better, alternative to the off-the-shelf classifiers like the SVM [46].

4.4 Discussion

In this chapter, we introduce a regularized Smoothed Score (SS) classifier within the framework of RKHS to handle the classification of functional data. The SS loss is quite versatile through adapting to distinct kernel functions. The RKHS framework allows us to flexibly control the smoothness of the estimated projection direction, resulting in enhanced prediction accuracy. Additionally, in contrast to some popular off-the-shelf classifiers for functional data, we establish Fisher consistency for the estimation of the slope function within the RKHS framework.

Our numerical studies underscore that the SS method exhibits favorable generalization and estimation properties, surpassing the predictive performance of the logistic regression, SVM and DWD classifiers. Notably, we develop a proximal gradient descent algorithm to address the issue of non-convexity of the objective function in

optimizations. The scalar covariates are also accounted for in our classifier in a linear manner to achieve a good trade-off between flexibility and interpretability.

There are several potential directions for extending our work. First, to deal with the case where the covariates Z are high-dimensional, imposing sparsity on the coefficient vector $\boldsymbol{\gamma}$ is one possible solution to avoid overfitting. Second, in many practical situations in clinical diagnostics, it is demanding to quantify the uncertainty of the estimated slope function and conduct valid statistical inference. These goals entail statistically sound tools, which we leave for future work.

4.5 Proof of the Main Theorems

4.5.1 The proof for generalization error of SS classifier

Proof of lemma 4.2:

Assume $w(1) + w(-1) = 1$, otherwise we replace $w(1), w(-1)$ by $w(1)/(w(1) + w(-1))$ and $w(-1)/(w(1) + w(-1))$ respectively. We first verify that $\bar{f}$ minimizes $Err(f) = \frac{1}{2}E(1 - \text{sign}(Yf))$. Notice that

$$\begin{aligned}
&E[\frac{1}{2}w(Y)(1 - \text{sign}(Yf))|\tilde{X} = \tilde{x}] \\
=&\frac{1}{2}[w(1)(1 - \text{sign}(f))p(\tilde{x}) + w(-1)(1 - \text{sign}(-f))(1 - p(\tilde{x}))] \\
=&\frac{1}{2}\{w(1)p(\tilde{x}) + w(-1)(1 - p(\tilde{x})) - [w(1)p(\tilde{x}) - w(-1)(1 - p(\tilde{x}))]\,\text{sign}(f)\}.
\end{aligned}$$

which is minimized when $f = f_B$. Of course, the solution is not unique, $\bar{f} = \text{sign}(f_B)$ is also a minimizer. For the smoothed score loss S_h, we could similarly write

$$\begin{aligned}
ES_h(f) &= E\{[w(Y)G(Yf/h)]|\tilde{X} = \tilde{x}\} \\
&= p(\tilde{x})w(1)G(f/h) + G(-f/h)(1 - p(\tilde{x})) := A(f).
\end{aligned}$$

The minimizer of $A(f)$ is also given by f_B, moreover, $S_h(f(\tilde{x})) \to \frac{1}{2}(1 - Y\,\text{sign}(f(\tilde{x}))$ as $h \to \infty$ for any decision rule f, which conclude the result of Lemma 4.2.

Proof of Theorem 4.3: As prerequisites for the main proof, we introduce the L_2-metric entropy with bracketing for a function class $\mathcal{F}$. For any $\varepsilon > 0$, denote

$\left\{l_1^l, l_1^u, \ldots, l_m^l, l_m^u\right\}$ an ε-bracketing function if for any $f \in \mathcal{F}$ there is a j such that $l_j^l \leq l(f,\cdot) \leq l_j^u$ and $\max_{\{1\leq j\leq m\}} \left\|l_j^u - l_j^l\right\|_2 \leq \varepsilon$, $\|\cdot\|_2$ is the usual L_2-norm, where $\|l\|_2^2 = \int l^2 dP$. Then the L_2 metric entropy of $\mathcal{F}$ with bracketing $H_B(\varepsilon, \mathcal{F})$ is defined as a logarithm of the cardinality of the ε-bracketing of the smallest size.

Let $\widetilde{S}_h(f) = S_h(f) + \lambda J(f)$ be the smoothed score function to be minimized, where $S_h(f) = w(Y_i)G(Y_i f(\tilde{X}_i)/h)$, h depends on n, we will ignore the subscript for simplicity. Let $\widetilde{\ell}(f, Z_i) = \ell(f, Z_i) + \lambda J(f)$ be the corresponding sign-cost function, where $\ell(f, Z_i) = \frac{1}{2}w(Y_i)(1 - \text{sign}(Y_i f(\tilde{X}_i)))$and $Z_i = (X_i, \boldsymbol{R}_i, Y_i)$.

Define the scaled empirical process $E_n\left(\widetilde{\ell}(f, Z) - \widetilde{S}_h(f_0, Z)\right)$, as

$$\begin{aligned} & E_n\left(\widetilde{\ell}(f, Z) - \widetilde{S}_h(f_0, Z)\right) \\ =& \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\ell}(f, Z_i) - \widetilde{S}_h(f_0, Z_i) - E\left(\widetilde{\ell}(f, Z_i) - \widetilde{S}_h(f_0, Z_i)\right)\right) \\ =& E_n\left(l(f, Z) - S_h(f_0, Z)\right). \end{aligned} \tag{4.17}$$

where $Z = (X, \boldsymbol{R}, Y)$ is the tuple. We also define the sieved function classes for the candidate classifiers for $j = 1, 2, \ldots$, and $i = 1, 2, \ldots$,

$$\begin{aligned} A_{i,j} =& \{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq Er(f, \bar{f}) < 2^i \delta_n^2, \\ & 2^{j-1}\max(J(\beta_0), 1) \leq J(f) < 2^j \max(J(\beta_0), 1)\} \end{aligned}$$

and

$$A_{i,0} = \left\{f \in \mathcal{F} : 2^{i-1}\delta_n^2 \leq Er(f, \bar{f}) < 2^i\delta_n^2,\ J(\beta) < \max(J(\beta_0), 1)\right\},$$

We assume without loss of generality that $J(\beta_0) \geq 1$ in the sequel. By applying the sieve method in Shen et al. [165] and Shen and Wong [166]. The idea is to split the problem of bounding $P\left(Er(\hat{f}, \bar{f}) \geq \delta_n^2\right)$ to bounding a sequence of empirical processes that are induced by the cost function $\widetilde{\ell}$. In particular, for bounding $P(A_{ij}), i, j = 1, \ldots, n$, we employ the large deviation inequality for empirical processes presented in theorem 3 [166].Controlling the mean and variance defined by $l(f, Z_i)$ and penalty $\lambda J(f)$ yields an inequality for the sequence of empirical processes and therefore, for $Er(\hat{f}, \bar{f})$.

Next, we establish the connection between $Er(\hat{f}, \bar{f})$ and the empirical processes (4.17). By the definition of $S_h(f, Z)$, we have

$$S_h\left(f, Z_i\right) \to \ell\left(f, Z_i\right), \text{as } h \to 0 \tag{4.18}$$

Since $\hat{f}$ is the minimizer of $n^{-1}\sum_{i=1}^{n} \widetilde{S}_h\left(f, Z_i\right)$, and $L\left(f_0, \bar{f}\right) \xrightarrow{h\to 0} Er\left(f_0, \bar{f}\right)$, along with (4.18), we have for h sufficiently small,

$$\begin{aligned}\left\{Er(\hat{f}, \bar{f}) \geq \delta_n^2\right\} &\subset \left\{ \sup_{\left\{f \in \mathcal{F}: Er(f,\bar{f}) \geq \delta_n^2\right\}} n^{-1} \sum_{i=1}^{n} \left(\widetilde{S}_h\left(f_0, Z_i\right) - \widetilde{S}_h\left(f, Z_i\right)\right) \geq \delta_n^2/2 \right\} \\ &\subset \left\{ \sup_{\left\{f \in \mathcal{F}: Er(f,\bar{f}) \geq \delta_n^2\right\}} n^{-1} \sum_{i=1}^{n} \left(\widetilde{S}_h\left(f_0, Z_i\right) - \widetilde{\ell}\left(f, Z_i\right)\right) \geq 0 \right\}.\end{aligned}$$

Hence

$$\begin{aligned}P\left(Er(\hat{f}, \bar{f}) \geq \delta_n^2\right) \leq P^* \Bigg(&\sup_{\left\{f \in \mathcal{F}: Er(f,\bar{f}) \geq \delta_n^2\right\}} n^{-1} \\ &\times \sum_{i=1}^{n} \left(\widetilde{S}_h\left(f_0, Z_i\right) - \widetilde{\ell}\left(f, Z_i\right)\right) \geq 0 \Bigg) := I\end{aligned}$$

where P^* denotes the outer probability measure. It is sufficient to bound the corresponding probability over A_{ij} in order to bound I, for each $i, j = 1, \ldots$. To this end, we need to build some inequalities regarding the first and second moments of the scaled empirical process $\widetilde{\ell}(f, Z) - \widetilde{S}_h\left(f_0, Z\right)$ for $f \in A_{ij}$.

For the first moment, following the result that

$$\begin{aligned}&E\left(\ell(f, Z) - S_h\left(f_0, Z\right)\right) \\ =&E\left(\ell(f, Z) - S_h(\bar{f}, Z)\right) - E\left(S_h\left(f_0, Z\right) - S_h(\bar{f}, Z)\right),\end{aligned}$$

According to condition 4.2, $L\left(f_0, \bar{f}\right) \leq s_n/2 \leq \delta_n^2/2$.

Next, assume that $\lambda \cdot \max\left(J(\beta_0), 1\right) \leq \delta_n^2/2$, and choose a sufficiently small h, we have, for any integers $i, j \geq 1$, $|S_h(\bar{f}, Z) - \ell(\bar{f}, Z)| < \delta_n^2/2$,

$$\begin{aligned}\inf_{A_{i,j}} E\left(\widetilde{\ell}(f, Z) - \widetilde{S}_h\left(f_0, Z\right)\right) &\geq M(i, j) \\ &= \left(2^{i-1}\delta_n^2\right) - \delta_n^2/2 + \lambda\left(2^{j-1} - 1\right) J(\beta_0)\end{aligned} \tag{4.19}$$

and

$$\inf_{A_{i,0}} E\left(\widetilde{\ell}(f,Z)-\widetilde{S}_h\left(f_0,Z\right)\right)\geq\left(2^{i-1}-1\right)\delta_n^2\geq M(i,0)=2^{i-2}\delta_n^2, \tag{4.20}$$

here the inequality $2^i-1\geq 2^{i-1}$ has been used.

For the second moment, it follows from the definition [165] and condition 4.3 that for any $f\in\mathcal{F}$,

$$\begin{aligned}
Er(f,\bar{f}) &= \ell(f)-\ell(\bar{f}) \\
&= \frac{1}{2}E\left|f_B(\tilde{X})\right|\left|\operatorname{sign}(\bar{f}(\tilde{X}))-\operatorname{sign}(f(\tilde{X}))\right| \\
&\geq \frac{1}{2}\delta E\left(\left|\operatorname{sign}(\bar{f}(\tilde{X}))-\operatorname{sign}(f(\tilde{X}))\right| I\left(\left|f_B(\tilde{X})\right|\geq\delta\right)\right) \\
&\geq \frac{1}{2}\delta\left(E\left|\operatorname{sign}(\bar{f}(\tilde{X}))-\operatorname{sign}(f(\tilde{X}))\right|-2c_1\delta^{\alpha}\right) \\
&\geq 4^{-1}\left(4c_1\right)^{-1/\alpha}\left(E\left|\operatorname{sign}(\bar{f}(\tilde{X}))-\operatorname{sign}(f(\tilde{X}))\right|\right)^{(\alpha+1)/\alpha}
\end{aligned}$$

with a choice of $\delta=\left(E\left|\operatorname{sign}(\bar{f}(\tilde{X}))-\operatorname{sign}(f(\tilde{X}))\right|/4c_1\right)^{1/\alpha}$.

Then we could establish a connection between the first and second moments. By Lemma 4.2, for any $f\in\mathcal{F}$, $E(S_h(f)-\ell(f))\xrightarrow{h\to 0}0$, then it follows from the triangular inequality that for any $\tilde{x}\in\mathcal{X}$, and h sufficiently small,

$$\begin{aligned}
E\left(\ell(f,Z)-S_h\left(f_0,Z\right)\right)^2 \leq & E\left|\frac{1}{2}w(Y)(1-\operatorname{sign}(Yf(\tilde{X})))-w(Y)G\left(Yf_0(\tilde{X})/h\right)\right| \\
\leq & \left(\frac{1}{2}Ew(Y)\left|\operatorname{sign}(Y\bar{f}(\tilde{X}))-\operatorname{sign}(Yf(\tilde{X}))\right|\right. \\
& +\frac{1}{2}Ew(Y)\left|\operatorname{sign}(Y\bar{f}(\tilde{X}))-\operatorname{sign}\left(Yf_0(\tilde{X})\right)\right| \\
& \left.+E\left|w(Y)G\left(Yf_0(\tilde{X})/h\right)-\frac{1}{2}w(Y)\left(1-\operatorname{sign}\left(Yf_0(\tilde{X})\right)\right)\right|\right).
\end{aligned}$$

For any $f\in A_{i,j}, Er(f,\bar{f})^{\frac{\alpha}{\alpha+1}}\geq(2^{-1}\delta_n^2)^{\frac{\alpha}{\alpha+1}}\geq 2^{-1}\delta_n^2\geq L\left(f_0,\bar{f}\right)$ and $L(f_0,\bar{f})+s_n/2\geq Er(f_0,\bar{f})$, indicating that

$$\begin{aligned}
& E\left(\ell(f,Z)-S_h\left(f_0,Z\right)\right)^2 \\
& \leq\left(4\left(4c_1\right)^{\frac{1}{\alpha+1}}\left(Er(f,\bar{f})^{\frac{\alpha}{\alpha+1}}+Er\left(f_0,\bar{f}\right)^{\frac{\alpha}{\alpha+1}}\right)+\left(ES_h(f_0)-E\ell(f_0)\right)\right) \\
& \leq c_3(Er(f,\bar{f})/2)^{\frac{\alpha}{\alpha+1}},
\end{aligned}$$

where $c_3 = \max\{2^{(1+2\alpha)/\alpha}[4(4c_1)^{\frac{1}{\alpha+1}}+1]+2^{(1+\alpha)/\alpha}, 8\}$. As a result,

$$\sup_{A_{i,j}} E\left(\ell(f,Z) - S_h\left(f_0,Z\right)\right)^2 \leq v^2(i,j) = c_3 M(i,j)^{\frac{\alpha}{\alpha+1}}; i = 1,\ldots, j = 0,\ldots.$$

Using the assumption that $\max \lambda\left(J\left(f_0\right),1\right) \leq \delta_n^2/2$, inequality (4.19), and (4.20), we have

$$\begin{aligned} I \leq & \sum_{i,j} P^* \left(\sup_{A_{i,j}} E_n \left(S_h\left(f_0,Z\right) - l(f,Z)\right) \geq M(i,j) \right) \\ & + \sum_{i} P^* \left(\sup_{A_{i,0}} E_n \left(S_h\left(f_0,Z\right) - l(f,Z)\right) \geq M(i,0) \right) := I_1 + I_2, \end{aligned}$$

Next we proceed to bound I_i separately. For I_1, we verify the required conditions (4.5)-(4.7) in theorem 3 of Shen and Wong [166]. To compute the metric entropy in (4.7), we now define a bracketing function for $S_h\left(f_0,Z\right) - \ell(f,Z)$. Denote an ε-bracketing set for $\{G_f : G_f = \{\tilde{x} \in \mathcal{X} : f(\tilde{x}) \geq 0\}, f \in A_{i,j}\}$ to be $\left\{\left(G_1^l, G_1^u\right), \ldots, \left(G_m^l, G_m^u\right)\right\}$. Let $s_j^l(\tilde{x})$ be -1 if $\tilde{x} \in G_j^u$ and 1 otherwise, and $s_j^u(x)$ be -1 if $\tilde{x} \in G_j^l$ and 1 otherwise; $j = 1,\ldots,m$. Then, $\left\{\left(s_1^l, s_1^u\right), \ldots, \left(s_m^l, s_m^u\right)\right\}$ forms an ε-bracketing function of $-\operatorname{sign}(f)$ for $f \in A_{i,j}$. Define

$$\begin{aligned} l_j^u(z) &= 1 + \left(s_j^u(\tilde{x})(1+y)/2 - s_j^l(\tilde{x})(1-y)/2\right) - S_h\left(f_0,z\right) \\ l_j^l(z) &= 1 + \left(s_j^l(\tilde{x})(1+y)/2 - s_j^u(\tilde{x})(1-y)/2\right) - S_h\left(f_0,z\right) \end{aligned}$$

Then for any $\varepsilon \geq M(i,j)$ and $f \in A_{i,j}$, there exists a $j, 1 \leq j \leq m$ such that $l_j^l(z) \leq \ell(f,z) - S_h\left(f_0,z\right) \leq l_j^u(z)$ for any $z = (y,\tilde{x})$, and $\left(E\left(l_j^u - l_j^l\right)^2\right)^{1/2} = \left(E\left(s_j^u(x) - s_j^l(x)\right)^2\right)^{1/2} \leq 2^{1/2}\varepsilon^{1/2}$. Thus $\left(E\left(l_j^u - l_j^l\right)^2\right)^{1/2} \leq \min\left((2\varepsilon)^{1/2}, 2^{1/2}\right)$. Consequently, $H_B\left(\varepsilon, \mathcal{F}\left(2^j\right)\right) \leq H\left(\varepsilon^2/2, \mathcal{G}\left(2^j\right)\right)$ for any $\varepsilon > 0$ and $j = 0,\ldots,$ where $\mathcal{F}\left(2^j\right) = \{\ell(f,z) - S_h(f,z) : f \in \mathcal{F}, J(f) \leq 2^j\}$. Using the fact that $\int_{aM(i,j)}^{v(i,j)} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(i,j)$ is nonincreasing in i and $M(i,j)$; $i = 1,\ldots.$ we have

$$\begin{aligned} & \int_{aM(i,j)}^{v(i,j)} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(i,j) \\ & \leq \int_{aM(1,j)}^{c_3^{1/2} M(1,j)^{\alpha/2(\alpha+1)}} H^{1/2}\left(u^2/2, \mathcal{G}\left(2^j\right)\right) du/M(1,j) \leq \phi\left(\varepsilon_n, 2^j\right), \end{aligned}$$

where $a = \varepsilon/32$. Then (4.7) is derived by Assumption 4.4 with a choice of $\varepsilon = 1/2$, c_3, and c_4. Finally, it is easy to see that (4.5)-(4.6) are satisfied with $\varepsilon = 1/2$ with the choice of $M(i,j), v(i,j)$, and $T = 1$.

By $0 < \delta_n \leq 1$ and $\lambda \max\left(J\left(f_0\right), 1\right) \leq \delta_n^2/2$ as well as theorem 3 of Shen and Wong [166] with $M = n^{1/2}M(i,j), v = v^2(i,j)$, $\varepsilon = 1/2$, and $T = 1$, we have

$$\begin{aligned}
I_1 &\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-\frac{(1-\varepsilon)nM(i,j)^2}{2\left(4v^2(i,j) + M(i,j)T/3\right)}\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 nM(i,j)^{\frac{\alpha+2}{\alpha+1}}\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 n\left[2^{i-1}\delta_n^2 - 2^{-1}\delta_n^2 + \left(2^{j-1}-1\right)\lambda J\left(f_0\right)\right]^{\frac{\alpha+2}{\alpha+1}}\right) \\
&\leq \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp\left(-c_5 n\left[\left(2^{i-2}\delta_n^2\right)^{\frac{\alpha+2}{\alpha+1}} + \left(\left(2^{j-1}-1\right)\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right]\right) \\
&\leq 3\exp\left(-c_5 n\left(\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right) / \left[\left(1 - \exp\left(c_5 n\left(\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right)\right)\right]^2
\end{aligned}$$

c_5 here is a positive constant. I_2 can be bounded in the same manner with the same upper.

Finally,

$$I \leq 6\exp\left(-c_5 n\left(\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right) / \left[\left(1 - \exp\left(-c_5 n\left(\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right)\right)\right]^2$$

This implies that $I^{1/2} \leq \left(5/2 + I^{1/2}\right)\exp\left(-c_5 n\left(\lambda J\left(f_0\right)\right)^{\frac{\alpha+2}{\alpha+1}}\right)$. The non-asymptotic upper bound is then followed from the fact $I \leq I^{1/2} \leq 1$.

Proof of Corollary 4.4: The rate with respect to the associated risk is a direct result from the exponential inequality in Theorem 4.3.

4.5.2 The proof for convergence rate of SS classifier

Proof of Proposition 4.5: First, we consider the case of $f(x, \boldsymbol{r}; \boldsymbol{\theta}) \neq 0$. For given $x, \boldsymbol{r}$, we note

$$\begin{aligned}
\boldsymbol{\theta}^*(x, \boldsymbol{r}) &= \arg\min_{\boldsymbol{\theta}} E\left\{w(Y) L_{01}\left(Y f\left(X, \boldsymbol{R}; \boldsymbol{\theta}\right)\right) | X = x, \boldsymbol{R} = \boldsymbol{r}\right\} \\
&= \arg\min_{\boldsymbol{\theta}} \{w(1) L_{01}(f(x, \boldsymbol{r}; \boldsymbol{\theta})) E\left[Y = 1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] \\
&\qquad + w(-1) L_{01}(-f(x, \boldsymbol{r}; \boldsymbol{\theta})) E\left[Y = -1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right]\} \\
&= \arg\min_{\boldsymbol{\theta}} \{w(1) E\left[Y = 1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] I(f(x, \boldsymbol{r}; \boldsymbol{\theta}) < 0) \\
&\qquad + w(-1) E\left[Y = -1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] I(f(x, \boldsymbol{r}; \boldsymbol{\theta}) > 0)\} \qquad (4.21)
\end{aligned}$$

Next, for each $x, \boldsymbol{r}$, we have

$$\begin{aligned}
& E[w(Y) G_h(Y f(X, \boldsymbol{R}; \boldsymbol{\theta}) | X = x, \boldsymbol{R} = \boldsymbol{r}] \\
= \; & w(1) G_h(f(x, \boldsymbol{r}; \boldsymbol{\theta})) E\left[Y = 1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] \\
& + w(-1) G_h(-f(x, \boldsymbol{r}; \boldsymbol{\theta})) E\left[Y = -1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right]
\end{aligned}$$

Further, when $h \to 0^+$, one gets

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}(x, \boldsymbol{r}) &= \arg\min_{\boldsymbol{\theta}} \lim_{h \to 0^+} E[w(Y) G_h(Y f(X, \boldsymbol{R}; \boldsymbol{\theta}) | X = x, \boldsymbol{R} = \boldsymbol{r}] \\
&= \arg\min_{\boldsymbol{\theta}} \{w(1) E\left[Y = 1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] I(f(x, \boldsymbol{r}; \boldsymbol{\theta}) < 0) \\
&\qquad + w(-1)\; E\left[Y = -1 | X = x, \boldsymbol{R} = \boldsymbol{r}\right] I(f(x, \boldsymbol{r}; \boldsymbol{\theta}) > 0)\} \qquad (4.22)
\end{aligned}$$

From (4.21) and (4.22), we have $\boldsymbol{\theta}^*(x, \boldsymbol{r}) = \tilde{\boldsymbol{\theta}}(x, \boldsymbol{r})$ for any given $x, \boldsymbol{r}$. In addition, if $f(x, \boldsymbol{r}; \boldsymbol{\theta}) = 0$, obviously $L_{01}(0) = G_h(0) = 1/2$ for any $h > 0$; Thus, $\boldsymbol{\theta}^*(x, \boldsymbol{r}) = \tilde{\boldsymbol{\theta}}(x, \boldsymbol{r})$ again. It leads $\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}}$ under the distribution $P(X, \boldsymbol{R})$.The result holds.

To derive the convergence rate of the SS estimator, we first present a bound for a partial functional binary classifier of Redemacher complexity.

Lemma 4.7 *For any $v_1, v_2 > 0$ and $0 < \nu < 1$,*

$$E\left[\sup_{\beta, \boldsymbol{\gamma}} \left| \frac{1/n \sum_{i=1}^n \sigma_i w(y_i) \left[\langle x_i, \beta \rangle_{L_2} + \boldsymbol{r}_i \boldsymbol{\gamma}\right]}{v_1^{-1} \left(\left\|\Gamma^{1/2} \beta_0\right\|_{L_2} + \left\|\Gamma^{1/2} \beta_1\right\|_{L_2}\right) + \left\|\beta_1\right\|_{\mathcal{H}_K} + v_2^{-1} \left\|\boldsymbol{\gamma}\right\|_2} \right| \right] \leq C\left(\mathcal{R}(v_1) + \frac{v_1}{\sqrt{n}} + \frac{v_2}{\sqrt{n}}\right),$$

where $\sigma_i \in \{-1, +1\}$ are i.i.d. Rademacher variables. And we have

$$E\left[\sup_{\beta,\boldsymbol{\gamma}}\left|(P-P_n)\frac{w(y)\left[G\left(y\left(\langle x,\beta\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right)-G\left(y\left(\langle x,\beta^*\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{v_1^{-1}\left(\left\|\Gamma^{1/2}(\beta_0-\beta_0^*)\right\|_{L_2}+\left\|\Gamma^{1/2}(\beta_1-\beta_1^*)\right\|_{L_2}\right)+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}+v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right]$$
$$\leq Ch_n^{-(1-\nu)}\left(\mathcal{R}(v_1)+\frac{v_1}{\sqrt{n}}+\frac{v_2}{\sqrt{n}}\right).$$

Proof: Recall that $G(u)=\int_u^\infty \mathcal{K}(t)dt$. For any $u_1, u_1 \in \mathbb{R}$ and $0<\nu<1$, when $h_n>0$ small enough, we have

$$\left|G\left(\frac{u_1}{h_n}\right)-G\left(\frac{u_2}{h_n}\right)\right|=\int_{(u_1\wedge u_2)/h_n}^{(u_1\vee u_2)/h_n}\mathcal{K}(t)dt\leq h_n^{\nu}\left|\frac{u_1}{h_n}-\frac{u_2}{h_n}\right|\leq\left|\frac{u_1}{h_n}-\frac{u_2}{h_n}\right|, \quad (4.23)$$

which implies that $G(\cdot)$ is Lipschitz continuous. By the symmetrization argument [135], and the contraction inequality for the Rademacher complexity (see, for example, Theorem 2.2 in [88]), we have

$$E\left[\sup_{\beta,\boldsymbol{\gamma}}\left|(P-P_n)\frac{w(y)\left[G\left(y\left(\langle x,\beta\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right)-G\left(y\left(\langle x,\beta^*\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{v_1^{-1}\left(\left\|\Gamma^{1/2}(\beta_0-\beta_0^*)\right\|_{L_2}+\left\|\Gamma^{1/2}(\beta_1-\beta_1^*)\right\|_{L_2}\right)+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}+v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right]$$
$$\leq CE\left[\sup_{\beta,\boldsymbol{\gamma}}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)\left[G\left(y_i\left(\langle x_i,\beta\rangle_{L_2}+\boldsymbol{r}_i\boldsymbol{\gamma}\right)/h_n\right)-G\left(y_i\left(\langle x_i,\beta^*\rangle_{L_2}+\boldsymbol{r}_i\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{v_1^{-1}\left(\|\Gamma^{1/2}(\beta_0-\beta_0^*)\|_{L_2}+\|\Gamma^{1/2}(\beta_1-\beta_1^*)\|_{L_2}\right)+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}+v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right]$$
$$\leq CE\left[\sup_{\beta,\boldsymbol{\gamma}}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)y_i\left[\langle x_i,\beta-\beta^*\rangle_{L_2}+\boldsymbol{r}_i(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right]/h_n^{1-\nu}}{v_1^{-1}\left(\|\Gamma^{1/2}(\beta_0-\beta_0^*)\|_{L_2}+\|\Gamma^{1/2}(\beta_1-\beta_1^*)\|_{L_2}\right)+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}+v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right]$$
$$= CE\left[\sup_{\beta,\boldsymbol{\gamma}}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)\left[\langle x_i,\beta-\beta^*\rangle_{L_2}+\boldsymbol{r}_i(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right]/h_n^{1/\nu}}{v_1^{-1}\left(\|\Gamma^{1/2}(\beta_0-\beta_0^*)\|_{L_2}+\|\Gamma^{1/2}(\beta_1-\beta_1^*)\|_{L_2}\right)+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}+v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right]$$
$$\leq Ch_n^{-(1-\nu)}E\left[\sup_{\beta_0}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)\left[\langle x_i,\beta_0-\beta_0^*\rangle_{L_2}\right]}{v_1^{-1}\|\Gamma^{1/2}(\beta_0-\beta_0^*)\|_{L_2}}\right|\right]$$
$$+Ch_n^{-(1-\nu)}E\left[\sup_{\beta_1}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)\left[\langle x_i,\beta_1-\beta_1^*\rangle_{L_2}\right]}{v_1^{-1}\|\Gamma^{1/2}(\beta_1-\beta_1^*)\|_{L_2}+\|\beta_1-\beta_1^*\|_{\mathcal{H}_K}}\right|\right]$$
$$+Ch_n^{-(1-\nu)}E\left[\sup_{\boldsymbol{\gamma}}\left|\frac{1/n\sum_{i=1}^n\sigma_i w(y_i)\left[\boldsymbol{r}_i(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right]}{v_2^{-1}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^*\|_2}\right|\right], \quad (4.24)$$

where $\sigma_i (i=1,\cdots,n)$ are i.i.d. Redemacher variables, the first inequality follows from the symmetrization inequality (Theorem 2.1 in [88]), the second inequality uses the contraction inequality for the Redamacher complexity (Theorem 2.2 in [88]), and the third step holds because of $\sigma_i y_i \in \{-1,+1\}$.

Similar to the proofs of Lemma 1 in [202] and of Lemma 3.1 in [192], we have

$$E\left[\sup_{\beta_1:\|\beta_1\|_{\mathcal{H}_K}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(y_i)\left[\langle x_i,\beta_1\rangle_{L_2}\right]\right|\right]\leq C\mathcal{R}(v_1), \tag{4.25}$$

$$E\left[\sup_{\beta_0:\left\|\Gamma^{1/2}\beta_0\right\|_{L_2}\leq v_1}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(y_i)\left[\langle x_i,\beta_0\rangle_{L_2}\right]\right|\right]\leq C\frac{v_1}{\sqrt{n}}. \tag{4.26}$$

For the linear term, we have

$$E\left[\sup_{\|\boldsymbol{\gamma}\|_2\leq v_2}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(y_i)\boldsymbol{r}_i\boldsymbol{\gamma}\right|\right]\leq E\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(y_i)\boldsymbol{r}_i\right\|_2\sup_{\|\boldsymbol{\gamma}\|_2\leq v_2}\|\boldsymbol{\gamma}\|_2\right]\leq C\frac{v_2}{\sqrt{n}}, \tag{4.27}$$

using the sub-Gaussianity of the components of $\boldsymbol{r}_i$. So, for any $\beta_0\in\mathcal{H}_\prime$, $\beta_1\in\mathcal{H}_1$, and $\boldsymbol{\gamma}\in\mathbb{R}^p$, standardize them as $\beta_0':=\dfrac{\beta_0}{v_1^{-1}\left\|\Gamma^{1/2}\beta_0\right\|_{L_2}}$, $\beta_1':=\dfrac{\beta_1}{v_1^{-1}\left\|\Gamma^{1/2}\beta_1\right\|_{L_2}+\|\beta_1\|_{\mathcal{H}_K}}$, and $\boldsymbol{\gamma}':=\dfrac{\boldsymbol{\gamma}}{v_2^{-1}\|\boldsymbol{\gamma}\|_2}$, which satisfy that $\left\|\Gamma^{1/2}\beta_0'\right\|_{L_2}\leq v_1$, $\left\|\Gamma^{1/2}\beta_1'\right\|_{L_2}\leq v_1$, $\|\beta_1'\|_{\mathcal{H}_K}\leq 1$, and $\|\boldsymbol{\gamma}\|_2\leq v_2$. Based on (4.25)-(4.27), we have

$$\begin{aligned}
&E\left[\sup_{\beta_1\in\mathcal{H}_1}\left|\frac{n^{-1}\sum_{i=1}^{n}\sigma_i w(y_i)\left[\langle x_i,\beta_1\rangle_{L_2}\right]}{v_1^{-1}\left\|\Gamma^{1/2}\beta_1\right\|_{L_2}+\|\beta_1\|_{\mathcal{H}_K}}\right|\right]\\
&\quad\leq E\left[\sup_{\beta_1:\left\|\Gamma^{1/2}\beta_1\right\|_{L_2}\leq v_1,\|\beta_1\|_{\mathcal{H}_K}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w(y_i)\left[\langle x_i,\beta_1\rangle_{L_2}\right]\right|\right]\leq C\mathcal{R}(v_1),\\
&E\left[\sup_{\beta_0\in\mathcal{H}_0}\left|\frac{n^{-1}\sum_{i=1}^{n}\sigma_i w(y_i)\left[\langle x_i,\beta_0\rangle_{L_2}\right]}{v_1^{-1}\left\|\Gamma^{1/2}\beta_0\right\|_{L_2}}\right|\right]\leq C\frac{v_1}{\sqrt{n}},\\
&E\left[\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p}\left|\frac{n^{-1}\sum_{i=1}^{n}\sigma_i w(y_i)\boldsymbol{r}_i\boldsymbol{\gamma}}{v_2^{-1}\|\boldsymbol{\gamma}\|_2}\right|\right]\leq C\frac{v_2}{\sqrt{n}}.
\end{aligned}$$

Therefore, for (4.24), we have

$$(4.24)\leq Ch_n^{-(1-\nu)}\left(\mathcal{R}(v_1)+\frac{v_1}{\sqrt{n}}+\frac{v_2}{\sqrt{n}}\right).$$

This completes the proof.

Now, we transfer the bound in expectation in Lemma 4.7 to the following bound in probability.

Lemma 4.8 *With probability at least* $1-\exp\left(-\min\left(\dfrac{n\mathcal{R}^2(v_1,v_2)}{(v_1+v_2)^2},\dfrac{n\mathcal{R}(v_1,v_2)}{v_2\log(n)}\right)\right)$, *we have*

$$\sup_{\beta,\boldsymbol{\gamma}}\left|(P-P_n)\frac{w(y)\left[G\left(y\left(\langle x,\beta\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right)-G\left(y\left(\langle x,\beta^*\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{D(\beta_0-\beta^*,\beta_1-\beta_1^*,\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)}\right|\leq Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2),$$

where $\mathcal{R}(v_1, v_2) = \mathcal{R}(v_1) + \frac{v_1}{\sqrt{n}} + \frac{v_2}{\sqrt{n}}$, *and* $D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*) = \|\beta_0 - \beta_0^*\|_{L_2} + v_1^{-1}\left(\left\|\Gamma^{1/2}(\beta_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\beta_1 - \beta_1^*)\right\|_{L_2}\right) + \|\beta_1 - \beta_1^*\|_{\mathcal{H}_K} + v_2^{-1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2$,

Proof: By (4.23), we have

$$\begin{aligned}
& \left|\frac{w(y)\left[G\left(y\left(\langle x, \beta\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right) - G\left(y\left(\langle x, \beta^*\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right| \\
\leq & \left|\frac{h_n^{-(1-\nu)} w(y) y\left(\langle x, \beta - \beta^*\rangle_{L_2} + \boldsymbol{r}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right)}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right| \\
\leq & C\left|\frac{h_n^{-(1-\nu)}\left(\langle x, \beta_0 - \beta_0^*\rangle_{L_2} + \langle x, \beta_1 - \beta_1^*\rangle_{L_2} + \boldsymbol{r}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right)}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right| \\
\leq & C\left|\frac{h_n^{-(1-\nu)}\left(\langle x, \beta_0 - \beta_0^*\rangle_{L_2} + \langle Kx, \beta_1 - \beta_1^*\rangle_{\mathcal{H}_K} + \boldsymbol{r}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right)}{\|\beta_0 - \beta_0^*\|_{L_2} + \|\beta_1 - \beta_1^*\|_{\mathcal{H}_K} + v_2^{-1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2}\right| \\
\leq & C\left|\frac{h_n^{-(1-\nu)}\left(\|x\|_{L_2}\|\beta_0 - \beta_0^*\|_{L_2} + \|Kx\|_{\mathcal{H}_K}\|\beta_1 - \beta_1^*\|_{\mathcal{H}_K} + \|\boldsymbol{r}\|_2\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2\right)}{\|\beta_0 - \beta_0^*\|_{L_2} + \|\beta_1 - \beta_1^*\|_{\mathcal{H}_K} + v_2^{-1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2}\right| \\
\leq & C h_n^{-(1-\nu)}\left(\|x\|_{L_2} + v_2\|\boldsymbol{r}\|_2\right), \qquad (4.28)
\end{aligned}$$

and

$$\begin{aligned}
& Var\left(\frac{w(y)\left[G\left(y\left(\langle x, \beta\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right) - G\left(y\left(\langle x, \beta^*\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right) \\
\leq & CVar\left(\frac{h_n^{-(1-\nu)}\left(\langle x, \beta_0 - \beta_0^*\rangle_{L_2} + \langle x, \beta_1 - \beta_1^*\rangle_{L_2} + \boldsymbol{r}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\right)}{v_1^{-1}\left(\|\Gamma^{1/2}(\beta_0 - \beta_0^*)\|_{L_2} + \|\Gamma^{1/2}(\beta_1 - \beta_1^*)\|_{L_2}\right) + v_2^{-1}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2}\right) \\
& Ch_n^{-2(1-\nu)}(v_1^2 + v_2^2). \qquad (4.29)
\end{aligned}$$

By the Adamczak bound on pages 24-25 in [88], and (4.28)-(4.29), we have

$$\begin{aligned}
& \sup_{\beta, \boldsymbol{\gamma}}\left|(P - P_n)\frac{w(y)\left[G\left(y\left(\langle x, \beta\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right) - G\left(y\left(\langle x, \beta^*\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right| \\
\leq & CE\left[\sup_{\beta, \boldsymbol{\gamma}}\left|(P - P_n)\frac{w(y)\left[G\left(y\left(\langle x, \beta\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right) - G\left(y\left(\langle x, \beta^*\rangle_{L_2} + \boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]}{D(\beta_0 - \beta^*, \beta_1 - \beta_1^*, \boldsymbol{\gamma} - \boldsymbol{\gamma}^*)}\right|\right] \\
& + Ch_n^{-(1-\nu)}\left((v_1 + v_2)\sqrt{\frac{t}{n}} + C\left(\left\|\max_i \|X\|_{L_2}\right\|_{\psi_1} + v_2\left\|\max_i \|\boldsymbol{r}_i\|_2\right\|_{\psi_1}\right)\frac{t}{n}\right), \qquad (4.30)
\end{aligned}$$

with probability at least $1 - e^{-t}$, where ψ_1 is the Orlicz norm associated with the function $\psi_1(z) = e^x - 1$. By Lemma 2.2.2 of [173], one gets $\left\|\max_i \|X\|_{L_2}\right\|_{\psi_1} \leq$

$C\log(n)$ and $\|\max_i \|\boldsymbol{r}_i\|_2\|_{\psi_1} \leq C\log(n)$ because of Condition 4.5. By Lemma 4.7, we have

$$(4.30) \leq Ch^{-(1-\nu)}\Big(\mathcal{R}(v_1)+\frac{v_1}{\sqrt{n}}+\frac{v_2}{\sqrt{n}}+ \left[(v_1+v_2)\sqrt{\frac{t}{n}}+[\log(n)+v_2\log(n)]\,\frac{t}{n}\right]\Big).$$

Let $t=\min\left(\frac{n\mathcal{R}^2(v_1,v_2)}{(v_1+v_2)^2},\frac{n\mathcal{R}(v_1,v_2)}{(1+v_2)\log(n)}\right)$, we finish the proof of the lemma.

Proof of Theorem 4.6: By the definition of $(\hat{\beta},\hat{\boldsymbol{\gamma}})$ in (4.13), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w(y_i)G\left(y_i\left(\langle x_i,\hat{\beta}\rangle_{L_2}+\boldsymbol{r}_i\hat{\boldsymbol{\gamma}}\right)/h_n\right)+\lambda\left\|\hat{\beta}_1\right\|_{\mathcal{H}_K}^2 \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} w(y_i)G\left(y_i\left(\langle x_i,\beta^*\rangle_{L_2}+\boldsymbol{r}_i\boldsymbol{\gamma}^*\right)/h_n\right)+\lambda\|\beta_1^*\|_{\mathcal{H}_K}^2. \qquad (4.31)
\end{aligned}$$

By Lemma 4.8, with probability at least $1-\exp\left(-\min\left(\frac{n\mathcal{R}^2(v_1,v_2)}{(v_1+v_2)^2},\frac{n\mathcal{R}(v_1,v_2)}{v_2\log(n)}\right)\right)$, we have

$$\begin{aligned}
&E\left\{w(y)\left[G\left(y\left(\langle x,\beta\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}\right)/h_n\right)-G\left(y\left(\langle x,\beta^*\rangle_{L_2}+\boldsymbol{r}\boldsymbol{\gamma}^*\right)/h_n\right)\right]\right\} \\
\leq\ & \lambda\|\beta_1^*\|_{\mathcal{H}_K}^2-\lambda\left\|\hat{\beta}_1\right\|_{\mathcal{H}_K}^2+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)D(\hat{\beta}_0-\beta^*,\hat{\beta}_1-\beta_1^*,\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) \\
=\ & -2\lambda\langle\beta_1^*,\hat{\beta}_1-\beta_1^*\rangle-\lambda\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K}^2+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)D(\hat{\beta}_0-\beta^*,\hat{\beta}_1-\beta_1^*,\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) \\
\leq\ & 2\lambda\|\beta_1^*\|_{\mathcal{H}_K}\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K}-\lambda\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K}^2+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)D(\hat{\beta}_0-\beta^*,\hat{\beta}_1-\beta_1^*,\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) \\
\leq\ & 2\lambda\|\beta_1^*\|_{\mathcal{H}_K}^2+\frac{\lambda}{2}\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K}^2-\lambda\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K}^2+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)\left\|\hat{\beta}_1-\beta_1^*\right\|_{\mathcal{H}_K} \\
&+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)\left[\left\|\hat{\beta}_0-\beta_0^*\right\|_{L_2}+v_2^{-1}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\right] \\
&+Ch_n^{-(1-\nu)}\mathcal{R}(v_1,v_2)\left[v_1^{-1}\left(\left\|\Gamma^{1/2}(\hat{\beta}_0-\beta_0^*)\right\|_{L_2}+\left\|\Gamma^{1/2}(\hat{\beta}_1-\beta_1^*)\right\|_{L_2}\right)\right] \qquad (4.32)
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\lambda \left\|\beta_1^*\right\|_{\mathcal{H}_K}^2 + \frac{\lambda}{2}\left\|\hat{\beta}_1 - \beta_1^*\right\|_{\mathcal{H}_K}^2 - \lambda\left\|\hat{\beta}_1 - \beta_1^*\right\|_{\mathcal{H}_K}^2 + \frac{Ch_n^{-2(1-\nu)}}{\lambda}\mathcal{R}^2(v_1, v_2) \\
&\quad + \frac{\lambda}{2}\left\|\hat{\beta}_1 - \beta_1^*\right\|_{\mathcal{H}_K}^2 + Ch_n^{-(1-\nu)}\mathcal{R}(v_1, v_2)\left[\left\|\hat{\beta}_0 - \beta_0^*\right\|_{L_2}\right] \\
&\quad + Ch_n^{-(1-\nu)}\mathcal{R}(v_1, v_2)\left[v_1^{-1}\left(\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2}\right) + v_2^{-1}\left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2\right] \\
&\leq 2\lambda \left\|\beta_1^*\right\|_{\mathcal{H}_K}^2 + \frac{Ch_n^{-2(1-\nu)}}{\lambda}\mathcal{R}^2(v_1, v_2) + Ch_n^{-(1-\nu)}\mathcal{R}(v_1, v_2)\left[\left\|\hat{\beta}_0 - \beta_0^*\right\|_{L_2}\right] + Ch_n^{-(1-\nu)} \cdot \\
&\quad \mathcal{R}(v_1, v_2)\left[v_1^{-1}\left(\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2}\right) + v_2^{-1}\left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2\right]. \qquad (4.33)
\end{aligned}
$$

Let $v_1 \asymp v_2 \asymp n^{-\frac{\mu}{2(\mu+1)}}$, and Condition 4.6 , then $\mathcal{R}(v_1) \leq Cn^{-\frac{\mu}{\mu+1}}$ by Lemma 3.1 of [192]. Thus, $\frac{v_1}{\sqrt{n}} \asymp n^{-\frac{2\mu+1}{2(\mu+1)}}$. Since $\frac{2\mu+1}{2(\mu+1)} > \frac{\mu}{\mu+1}$, $\mathcal{R}(v_1, v_2) = \mathcal{R}(v_1) + \frac{v_1}{\sqrt{n}} + \frac{v_2}{\sqrt{n}} \leq Cn^{-\frac{\mu}{\mu+1}}$. Taking $\lambda \asymp h_n^{-(1-\nu)}n^{-\frac{\mu}{\mu+1}}$, we have

$$
(4.32) \leq Ch_n^{-(1-\nu)}n^{-\frac{\mu}{\mu+1}} + Ch_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}\left[\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2} + \left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2\right]. \qquad (4.34)
$$

By (4.15) in Remark 4.4, (4.32) and(4.34), either

$$
\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2} + \left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2 \leq Ch_n^{-(1-\nu)}n^{-\frac{\mu}{\mu+1}}, \qquad (4.35)
$$

or

$$
\begin{aligned}
&\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2}^2 + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2}^2 + \left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2^2 \\
&\leq Ch_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}\left[\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2} + \left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2\right] \qquad (4.36)
\end{aligned}
$$

Clearly, (4.35) implies . For (4.36), by the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we have

$$
\left\|\Gamma^{1/2}(\hat{\beta}_0 - \beta_0^*)\right\|_{L_2} + \left\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1^*)\right\|_{L_2} + \left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_2 \leq Ch_n^{-(1-\nu)}n^{-\frac{\mu}{2(\mu+1)}}.
$$

This completes the proof of the main result.

Chapter 5

Inference for Functional Logistic Regression under Case Control Designs

5.1 Introduction

One of the pivotal challenges during the big-data era becomes to tackle data of multi-modality and massive volume. Despite the rapid development of computational resources, fitting a model using massive data often exceeds available computational power due to the limitation of memory. One antidote is to develop complex distributed computing systems that can directly handle big data. However, this approach brings forth unnecessary model complexity and inefficient data deployment. Furthermore, computational time required for processing the full dataset can be prohibitively extensive, necessitating the use of high-performance computing resources which are in practice hard to obtain.

Another commonly used strategy involves selecting random subsamples from the extensive dataset to serve as a surrogate. However, subsampling can result in loss of statistical accuracy especially for highly unbalanced large datasets, meaning the estimates may have high variability. Thus it's crucial to design an effective sampling method that can minimize loss of accuracy while addressing the unbalanced data problem.

Additionally, the multi-modality of large datasets poses another layer of obstacles. In applications such as clinical, epidemiological sciences, and meteorology where the stakes are exceptionally high, many variables are measured or observed at multiple time points or spatial locations. These kinds of variables gave birth to the generation of functional data analysis (FDA) [**rramsay_silverman_2006**, 15, 123, 196, 198] since these variables can be viewed as functions of time or spatial locations.

In some applications, we may encounter massive functional data. One example is the kidney transplant data acquired from the Organ Procurement Transplant Network/United Network for Organ Sharing (Optn/UNOS). This extensive dataset encompasses information on about 1 million recipients as well as over 4 million records monitored during the post-transplantation follow-up period to assess the success of kidney transplants. We use the Glomerular filtration rate (GFR) trajectories as the predictor and patient demographics as scalar covariates to measure kidney function. Similarly in environmental sciences, extreme events, though less frequent, hold significant importance. The air pollution dataset provided by the Environmental Protection Agency (EPA) meticulously logs daily concentrations of pollutants such as PM2.5, ozone, and nitrogen dioxide across thousands of monitoring stations nationwide. This dataset, instrumental for evaluating air quality trends and formulating environmental policies, spans multiple decades and again includes millions of records, the enormity of such datasets often stretches into hundreds of gigabytes. The horizontal and vertical dimensions of such functional massive yet unbalanced data make it almost impossible to utilize the full dataset for classification.

The existing literature on subsampling schemes mainly revolves around models with both scalar responses and predictors. For linear regression, a leverage-score-based random subsampling and its asymptotic estimator properties were explored by [108]. [180] introduced an information-based optimal subdata selection (IBOSS) where subsample are selected deterministically without random sampling. For Logistic regression and generalized linear models, the A-optimality, L-optimality criterion

Poisson subsamplings are proposed in [2, 24, 178, 181]. Recently, [179], [43] and [1] used the subsampling method for quantile regressions. For generalized additive models, [187] developed a scalable estimation method via marginal predictor discretization. Recently, [105] developed the L-Optimality subsampling for functional linear and generalized linear models based on the penalized B-splines.

In the more refined context of classification in imbalanced datasets, the case-control (CC) study, a traditional yet more efficient response-biased sampling, is widely used for investigating the relationship between existing factors and rare disease incidence in epidemiology, where samples are taken separately according to the value of the response, e.g, through disease registry records [22, 92, 137]. This also closely relates to choice-based sampling in econometrics [110]. Especially, people are interested in the relationship between the predictors and individuals' choices. It would be more interesting to take samples of individuals based on one or some specific choices made by individuals than to take a single sample from the entire population [20, 21, 47, 104, 137, 160, 161]. Other popular sampling designs such as the length-biased, case-cohort, and general biased sampling schemes have been studied by [83, 136, 167, 190, 199]. Generally speaking, case-control or other biased samples are more likely to contain useful information relevant to one's interest [140].

As steps forward, local case-control (LCC) and local uncertainty sampling (LUS) are proposed on the ground of CC scheme by [47, 62] respectively. They attempt to remedy response imbalance locally throughout the feature space and extend to multi-class logistic regression. Both strategically select near-boundary samples using a pilot estimator, enhancing the model's sensitivity to the minority class.

This paper adapts the CC and LCC sampling scheme for functional and semi-functional linear logistic regression to ensure a balanced and informative training set. LCC is indispensable in unbalanced functional datasets where the goal is to achieve high predictive performance and reliable inference in applications where predictions can have profound implications. The proposed LCC sampling procedure includes:

(1)it assigns an acceptance probability for each data point and selects observations according to the assigned probabilities through a flexible pilot estimation; (2) it fits a logistic model with the subsampled observations to obtain an estimate of the unknown model parameter.

The coefficient function is assumed to reside in a Reproducing Kernel Hilbert Space (RKHS) and can be estimated using roughness regularization. By ensuring a balanced and informative training set, To the best of our knowledge, there is scarce work on subsampling for functional data, our framework is distinct from [105] in multiple aspects, first we target the imbalance issues in massive functional data classification and are more flexible to incorporate additional scalar terms. Additionally, the utilization of RKHS in semi-functional logistic regression models brings forth several advantages.

(1) The RKHS approaches offer greater flexibility in choosing function spaces, allowing for the modeling of more complex relationships. They are not restricted to a fixed basis, unlike penalized B-splines, and can adapt to various types of data smoothly and effectively [128]. (2) RKHS inherently includes a regularization parameter within its framework, which offers a systematic approach to control model complexity. This built-in regularization is more nuanced and can be more effective compared to the ad-hoc penalty terms used in penalized B-splines [171].

The roughness regularization method assuming smooth structures in a RKHS was widely employed nowadays. In particular, functional linear regression (FLR) have been investigated in[16, 17]. [50] considered fast-fitting approaches for generalized functional linear models through penalized spline regression. [39] adopted a penalized likelihood approach to study the generalized FLR. [164] proposed a roughness penalty approach and conducted nonparametric inferences for generalized functional linear models. [81] designed a functional logistic model based on elastic net regularization to identify the genes. A class of generalized scalar-on-image regression models was developed by [183] via the total variation penalty enforced on the coefficient function estimator. [98] studied the method for inferences in generalized partial functional

linear models based on RKHS. These methods often assume a balanced dataset and may not perform well in the presence of class imbalance. The challenge becomes more pronounced when the functional data are accompanied by scalar covariates, necessitating a semi-functional approach that can accommodate both data types.

Our contributions are multifold and can be summarized as follows:

- We introduce the CC and LCC approach for semi-functional logistic regression models, bridging the gap in the literature by addressing both functional and scalar covariates in an imbalanced dataset scenario. The procedure generates a consistent estimator within the original model family when the model is correctly specified without the need of post-estimation corrections.

- Our method capitalizes on the strengths of RKHS to handle the infinite-dimensional functional data, allowing us to capture intricate data structures through a kernelized representation. Theoretically, we establish the functional Bahadur representation [98, 164] for the functional slope and further explore the joint definition to incorporate both scalar and functional variables through a predefined weighted inner product, the asymptotic joint distribution of the maximum partial likelihood estimators are derived under CC and LCC sampling schemes. To detect the significance of the slope function, an inferential tool is developed based on the penalized likelihood ratio test under the CC scheme along with the confidence bands presented in the last section of this chapter.

We provide extensive empirical and real data experiments demonstrating that our methods outperform traditional random sampling and yield superior model performance, making it more accessible for large-scale applications without compromising on computational integrity or predictive accuracy.

The rest of this article is organized as follows. In Section 5.2, we present the functional linear logistic model under CC design, introduce the norms under RKHS, and the computation algorithm. In Section 5.3, we study the Frećhet derivatives of

the penalized likelihood, the Bahadur representation for the function slope, and its asymptotic properties. A penalized likelihood ratio test based on the CC sampling is also presented. We further explore the LCC sampling for semi-functional logistic regression and its joint asymptotic properties in Section 5.4 and 5.5. The performance of the proposed CC and LCC estimator is demonstrated through extensive simulation studies in Section 5.6.1,5.6.2. Finally, applications on two real datasets in 5.7 illustrate the superior performance of the proposed approaches. A brief discussion is given in Section 5.8. Technical details are deferred to the supplementary materials.

5.2 Functional Logistic Regression for the Case Control Design

Logistic regression [26] is one of the most fundamental statistical tools to model categorical outcomes. Let $Y \in \mathcal{Y}$ be a categorical response variable. Without loss of generality, we focus on binary response coded as $0/1$. Consider a functional linear logistic regression

$$P(Y = 1 \mid X = x) = \frac{\exp\{\alpha + \int x(t)\beta(t)\mathrm{d}t\}}{1 + \exp\{\alpha + \int x(t)\beta(t)\mathrm{d}t\}} \equiv p_1(x; \theta), \tag{5.1}$$

where $X(t)$ is a square-integrable function recorded on the interval $\mathcal{I} = [0, 1]$, $\alpha \in \mathbb{R}^1$ is an intercept term, $\beta(t) : \mathcal{I} \to \mathbb{R}$ is the slope function, and $\theta = (\alpha, \beta(t))^\top$. Throughout this paper, we shall assume $\theta_0^* = (\alpha_0^*, \beta_0(t))^\top$ represents the true value of θ when the logistic regression (5.1) is correctly specified.

Let $X^0(t), X^1(t)$ represent two real-valued random predictor processes over $\mathcal{I}$, a compact subspace of $\mathbb{R}$, where the superscripts 0 and 1 of X are mnemonic reminders that the samples correspond to $Y = 0$ and $Y = 1$. Under the case-control or retrospective sampling scheme, we observe two independent samples with total sample size $n = n_0 + n_1$ from each of these processes: $X_1^0(t) \ldots, X_{n_0}^0(t); X_1^1(t) \ldots, X_{n_1}^1(t)$, where n_1 and n_0 are pre-specified in case-control studies. The "case-control design" in this chapter refers to the classical method where separate random samples are taken from

the case and control populations, for instance, through disease registry records [22].

Remark 5.1 *According to Lin et al. [104], Prentice and Pyke [137], and Scott and Wild [160], the prospective estimating equation derived from the maximum likelihood estimation is still valid for a consistent estimator of the slope function $\beta(t)$, except for the intercept term α under case-control design. In other words, the case-control design only leads to biased estimation of the intercept α_0^*, and we specify the biased true value as α_0. With this view, we are able to conduct valid inference analysis for $\beta(t)$, forget the sampling scheme and treat the data as if they are drawn by random sampling from the whole population.*

5.2.1 RKHS

Under the functional linear logistic regression model (5.1), the slope function $\beta_0(\cdot)$ is a real-valued function, and thus it's generally to do dimension reduction or to pose additional constraints due to the infinite dimensionality of $\beta_0(t)$. A possible way is to write $\beta_0(t)$ as a truncated expansion of certain basis functions; see for example FPCA, B-splines, or Fourier basis functions [39, 98, 164]. Nevertheless, as pointed out in [143, 198], the truncation parameter changes in a discrete manner, which may yield imprecise control over the model complexity, hence result in inaccurate and unutterable functional estimates with "artificial" bumps. Instead, we adopt the roughness penalty approach to prevent the aforementioned problems.

Let $\beta \in H^m(\mathbb{I})$, the m-order Sobolev space is defined by

$$\begin{aligned} H^m(\mathbb{I}) =& \{\beta : \mathbb{I} \mapsto \mathbb{R} \mid \beta^{(j)}, j = 0, \ldots, m-1, \\ & \text{are absolutely continuous, and } \beta^{(m)} \in L^2(\mathbb{I})\}. \end{aligned}$$

Therefore, the unknown parameter $\theta = (\alpha, \beta)$ belongs to $\mathcal{H} \equiv \mathbb{R}^1 \times H^m(\mathbb{I})$. We further assume $m > 1/2$ to ensure that $H^m(\mathbb{I})$ is a reproducing kernel Hilbert space [98, 164, 198]. Under case-control design, we propose to find the estimates by minimizing the

following regularized problems:

$$\begin{aligned}\left(\hat{\alpha}_{n,\lambda}, \hat{\beta}_{n,\lambda}\right) &= \underset{(\alpha,\beta)\in\mathcal{H}}{\arg\inf} L_{n,\lambda}(\theta) = \underset{(\alpha,\beta)\in\mathcal{H}}{\arg\inf} \left[L_n(\theta) + \lambda/2 J(\beta,\beta)\right] \\ &\equiv \underset{(\alpha,\beta)\in\mathcal{H}}{\arg\inf} -\frac{1}{n}\left[\sum_{i=1}^{n_1} \log\{p_1(X_i^1;\theta)\} + \sum_{i=1}^{n_0} \log\{p_0(X_i^0;\theta)\}\right] + (\lambda/2) J(\beta,\beta)\end{aligned} \tag{5.2}$$

where $p_0(\cdot) = 1 - p_1(\cdot)$, $J(\beta_1,\beta_2) = \int_0^1 \beta_1^{(m)}(t)\beta_2^{(m)}(t)\mathrm{d}t$ is a roughness penalty function, and λ is a smoothing parameter which controls the balance between the bias and the smoothness of the parameter. Here we use $\lambda/2$ for simplifying expressions when calculating Fréchet derivatives.

Before introducing an inner product and norms of $H^m(\mathbb{I})$ and $\mathcal{H}$, we first give some notation. Let $X_1,\ldots,X_{n_0}$ denote $X_1^0,\ldots,X_{n_0}^0$, and let $X_{n_0+1},\ldots,X_n$ denote $X_1^1,\ldots,X_{n_1}^1$, then $X = (X_1^0,\ldots,X_{n_0}^0, X_1^1,\ldots,X_{n_1}^1)^\top$ represent the design matrix without intercept. Let $B(X) = p_0(X;\theta_0)p_1(X;\theta_0)$ and $E_l(\cdot)$ denote the conditional expectation of X given $Y = l$, $l = 0, 1$. The inner product for any $\beta_1, \beta_2 \in H^m(\mathbb{I})$ is then defined by

$$\langle \beta_1, \beta_2 \rangle_1 = V(\beta_1,\beta_2) + \lambda J(\beta_1,\beta_2), \tag{5.3}$$

where $V(\beta_1,\beta_2) = \int_0^1\int_0^1 C(s,t)\beta_1(s)\beta_2(t)\mathrm{d}s\mathrm{d}t$ and $C(s,t) \equiv \rho_0 E_0\{B(X^0)X^0(t)X^0(s)\} + \rho_1 E_1\{B(X^1)X^1(t)X^1(s)\}$ which can be viewed as a weighted covariance operator of X. Denote the corresponding norm as $\|\cdot\|_1$. As for the full parameter space $\mathcal{H}$, we also define the inner product, for any $\theta_1 = (\alpha_1,\beta_1)$ and $\theta_2 = (\alpha_2,\beta_2) \in \mathcal{H}$,

$$\begin{aligned}&\langle (\alpha_1,\beta_1),\ (\alpha_2,\beta_2)\rangle \\ &\quad\equiv \sum_{l=0}^{1} \rho_l E_l\left\{B(X^l)\left(\alpha_1 + \int_0^1 X^l(t)\beta_1(t)\mathrm{d}t\right)\left(\alpha_2 + \int_0^1 X^l(t)\beta_2(t)\mathrm{d}t\right)\right\} \\ &\quad\quad + \lambda J(\beta_1,\beta_2).\end{aligned} \tag{5.4}$$

We note that the corresponding norm $\|\theta\|^2 = \langle\theta,\theta\rangle$ is well defined under some conditions.

5.2.2 Representer theorem

In this subsection, we introduce the representer theorem, which guarantees a finite-dimensional representation of the estimate $\hat{\beta}_{n,\lambda} \in H^m(\mathbb{I})$. Consider the (squared) norm

$$\|\beta\|^2_{\mathcal{W}_2^m} = \sum_{l=0}^{m-1} \left\{ \int_0^1 \beta^{(l)}(t) \mathrm{d}t \right\}^2 + \int_0^1 \left\{ \beta^{(m)}(t) \right\}^2 \mathrm{d}t$$

for any $\beta \in H^m(\mathbb{I})$ makes $H^m(\mathbb{I})$ an RKHS. Let $H_0(\mathbb{I}) = \{\beta \in \mathcal{H} : J(\beta, \beta) = 0\}$ denote the null space of J, which is a finite-dimensional linear subspace of $H^m(\mathbb{I})$ with the basis functions $\psi_1, \ldots, \psi_m$. Denote by $H_1(\mathbb{I})$ its orthogonal complement in $H^m(\mathbb{I})$ such that $H^m(\mathbb{I}) = H_0(\mathbb{I}) \oplus H_1(\mathbb{I})$. Similarly, for any $\beta \in H^m(\mathbb{I})$, there exists a unique decomposition $\beta = \beta_0 + \beta_1$ such that $\beta_0 \in H_0(\mathbb{I})$ and $\beta_1 \in H_1(\mathbb{I})$. Notice that $H_1(\mathbb{I})$ is also an RKHS with the inner product of $H^m(\mathbb{I})$ restricted to $H_1(\mathbb{I})$.

Let K and $K_1 : \mathbb{I} \times \mathbb{I} \to \mathbb{R}$ denote the reproducing kernels of $H^m(\mathbb{I})$ and $H_1(\mathbb{I})$, respectively. We then have $J(\beta_1, \beta_1) = \|\beta_1\|_K^2 = \|\beta_1\|^2_{\mathcal{W}_2^m}$ for any $\beta_1 \in H_1(\mathbb{I})$. According to the well-known Representer lemma [33, 168], the solution to equation (5.2) can be expressed as

$$\hat{\beta}_{n\lambda}(t) = \sum_{l=1}^{m} d_l \psi_l(t) + \sum_{i=1}^{n} c_i \xi_i(t), \tag{5.5}$$

where $\xi_i(t) = \int_0^1 K_1(t,s) X_i(s) ds$, and $(KX_i)(\cdot) \equiv \int_0^1 K(\cdot, s) X_i(s) ds \in H^m(\mathbb{I})$ for $i = 1, \ldots, n$. It demonstrates that the solution of problem (5.2) can be found in a finite-dimensional subspace based on ψ_i as well as ξ_i. Let $\boldsymbol{d} = (d_1, \ldots, d_m)^\top, \boldsymbol{c} = (c_1, \ldots, c_n)^\top$ and $\Xi = (J(\xi_i, \xi_j))_{ij} \in \mathbb{R}^{n \times n}$. Define the functions $\ell_l(a) = al - \log\{1 + \exp(a)\}$, for $l = 0, 1$. As a result, the estimation of infinite dimensional $\beta(t)$ reduces to the estimation of finite scalar coefficients $\boldsymbol{d}$ and $\boldsymbol{c}$ by minimizing the following problems

$$\begin{aligned} \ell_{n,\lambda}(\alpha, \boldsymbol{d}, \boldsymbol{c}) = -\frac{1}{n} \Bigg[& \sum_{i=1}^{n_0} \ell_0 \left\{ \alpha + \sum_{l=1}^{m} d_l \int_{\mathcal{I}} X_i^0(t) \phi_j(t)\, dt + \sum_{j=1}^{n} c_j \langle \xi_i, \xi_j \rangle_{H^m(\mathbb{I})} \right\} \\ & + \sum_{i=1}^{n_1} \ell_1 \left\{ \alpha + \sum_{l=1}^{m} d_l \int_{\mathcal{I}} X_i^1(t) \phi_j(t)\, dt + \sum_{j=1}^{n} c_j \langle \xi_i, \xi_j \rangle_{H^m(\mathbb{I})} \right\} \Bigg] + \frac{\lambda}{2} \boldsymbol{c}^\top \Xi \boldsymbol{c}. \end{aligned} \tag{5.6}$$

The objective function (5.6) is strictly convex in $\alpha + \int_0^1 X(t)\beta(t)\mathrm{d}t$, which is a linear transformation about β. Hence, we can apply the Newton-Raphson algorithm to compute the minimizer $(\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda})$ of (5.6) for a fixed smoothing parameter λ. Let $a(X_i^l;\theta) = \alpha + \sum_{l=1}^m d_l \int_{\mathcal{I}} X_i^l(t)\phi_j(t)\,\mathrm{d}t + \sum_{j=1}^n c_j \langle \xi_i, \xi_j \rangle_{H^m(\mathbb{I})}$. Given an initial estimate $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$, define $\tilde{B}_i^l = p_0\{a(X_i^l;\tilde{\theta})\}p_1\{a(X_i^l;\tilde{\theta})\}$ and $\tilde{Y}_i^l = a(X_i^l;\tilde{\theta}) - [p_1\{a(X_i^l;\tilde{\theta})\} - l]/\tilde{B}_i^l$, then the quadratic approximation of $\ell_l\{a(X_i^l;\theta)\}$ at $\tilde{\theta}$ is $\tilde{B}_i^l\{\tilde{Y}_i^l - a(X_i^l;\tilde{\theta})\}^2/2 + C_i^l$, where C_i^l is independent of θ. We then update $(\tilde{\alpha}, \tilde{\beta})$ by the minimizer of the penalized weighted least squares form as follows

$$\frac{1}{n}\sum_{l=0}^{1}\sum_{i=1}^{n_l} \tilde{B}_i^l \left\{\tilde{Y}_i - a\left(X_i^l;\theta\right)\right\}^2 + \frac{\lambda}{2}\boldsymbol{c}^\top \Xi \boldsymbol{c}.$$

To select the smoothing parameter λ, we adopt the generalized approximate cross-validation (GACV) score developed in [57, 189], which provides a cross-validation approximation to the Kullback-Leibler distance between an estimate and the true.

5.3 Asymptotic Properties

We will next introduce the Bahadur representation for the penalized estimators, which greatly facilitates the following asymptotic analysis. We start with several notations. Define $\rho_l = \lim_{n\to\infty} n_l/n$, for $l = 0, 1$. Let $\|\cdot\|_{L^2}$ denote the Euclidean L^2 norm. For two positive real sequences $(a_k)_{k\geq 1}$ and $(b_k)_{k\geq 1}$, we write $a_k \asymp b_k$ to indicate that the sequence of ratios $(a_k/b_k)_{k\geq 1}$ is bounded away from zero and infinity. $\overset{a}{\sim}$ denotes approximately distributed. To establish the theoretical properties, we require the following regularity conditions.

Condition 5.1 *There exist two positive constants c_1 and c_2 such that $0 < c_1 \leq \min_{l=0,1}\rho_l \leq \max_{l=0,1}\rho_l \leq c_2 < 1$, moreover, $|\rho_0 - n_0/n| = O_P(n^{-1/2})$.*

This condition requires that the limiting proportion of cases ($Y = 1$) and controls ($Y = 0$) is not close to either 0 or 1. Meanwhile, the convergence rate n_0/n cannot be too slow. Actually, under case-control design, n_0, n, ρ_0 and ρ_1 are user-specified

quantities, which can be easily satisfied. Similar condition can be found in Lin et al. [104].

Define a linear bounded operator $C(\cdot)$ from $L^2(\mathbb{I})$ to $L^2(\mathbb{I})$: $(C\beta)(t) = \int_0^1 C(s,t)\beta(s)ds$ where $C(s,t) = \sum_{l=0}^1 \rho_l E_l\{B(X^l)X^l(t)X^l(s)\}$. $C\beta$ is further regulated as follows.

Condition 5.2 *$C(s,t)$ is continuous on $\mathbb{I} \times \mathbb{I}$. Furthermore, for any $\beta \in L^2(\mathbb{I})$ satisfying $C\beta = 0$, we have $\beta = 0$.*

Condition 5.2 indicates that V is positive definite, allowing the inner product (5.3) to be well defined. Furthermore, $H_m(\mathbb{I})$ is a reproducing kernel Hilbert space (RKHS) under $\langle\cdot,\cdot\rangle_1$, and we denote its reproducing kernel function as $K(s,t)$. As discussed in [164], C admits the spectral decomposition $C(s,t) = \sum_{j=1}^\infty \zeta_j\psi_j(s)\psi_j(t)$ by Mercer's theorem, where $\{\psi_j(\cdot), \zeta_j \geq 0\}_{j\geq 1}$ forms an orthonormal basis in $L^2(\mathbb{I})$ under the usual L^2-norm.

Condition 5.3 *There exists a sequence of basis functions $\{\varphi_j\}_{j\geq 1} \subsetneq H^m(\mathbb{I})$ such that $\|\varphi_j\|_{L^2} \leq C_\varphi j^a$ holds uniformly over j for some constants $a \geq 0, C_\varphi > 0$, and that*

$$V(\varphi_\mu, \varphi_j) = \delta_{\mu j}, \quad J(\varphi_\mu, \varphi_j) = \rho_j\delta_{\mu j} \quad \text{for any} \quad \mu, j \geq 1.$$

where $\delta_{\mu j} = 1$ if $\mu = j$ and 0 otherwise, and ρ_j is a nondecreasing nonnegative sequence satisfying $\rho_j \asymp j^{2k}$ for some constant $k > a + 1/2$. Furthermore, any $\beta \in H^m(\mathbb{I})$ admits the Fourier expansion $\beta = \sum_{j=1}^\infty V(\beta, \varphi_j)\varphi_j$ with convergence in $H^m(\mathbb{I})$ under $\langle\cdot,\cdot\rangle_1$.

Condition 5.3 assumes the existence of a sequence of basis functions in $H^m(\mathbb{I})$ that can simultaneously diagonalize V and J in the inner product (5.3).

We define $K_t(\cdot) \equiv K(t,\cdot) \in H^m(\mathbb{I})$ for any $t \in \mathbb{I}$. According to [63, 164], the relationship among (ϕ_j, ρ_j), $K_t(\cdot)$ and W_λ are as follows

$$K_t(\cdot) = \sum_{j=1}^\infty \frac{\varphi_j(t)}{1+\lambda\rho_j}\varphi_j(\cdot),$$

and

$$(W_\lambda\varphi_j)(\cdot) = \frac{\lambda\rho_j}{1+\lambda\rho_j}\varphi_j(\cdot).$$

Let $\tau(x) = \sum_{j=1}^{\infty} \frac{x_j}{1+\lambda\rho_j}\varphi_j$, where $x_j = \int_0^1 x(t)\varphi_j(t)dt$, for any $x \in L^2(\mathbb{I})$. According to Proposition 2.4 of Shang and Cheng [164], we know that $R_x \equiv ([E\{B(X)\}]^{-1}, \tau(x)) \in \mathcal{H}$, satisfies $\langle R_x, \theta\rangle = \alpha + \int_0^1 x(t)\beta(t)dt$ for any $\theta = (\alpha, \beta) \in \mathcal{H}$. Note that R_x depends on λ through the definition of $\tau(x)$. Let $P_\lambda\theta_1 = (0, W_\lambda\beta_1)$ for any $\theta_1 = (\alpha_1, \beta_1) \in \mathcal{H}$. Then $P_\lambda\theta_1 \in \mathcal{H}$ and $\langle P_\lambda\theta_1, \theta_2\rangle = \langle W_\lambda\beta_1, \beta_2\rangle_1$ for any $\theta_2 = (\alpha_2, \beta_2) \in \mathcal{H}$.

The following regularity conditions on X^l are required theoretically to pose smoothness and tail conditions on the logistic loss $\ell_l(a)$ for $l = 0, 1$.

Condition 5.4 *Assume that X^l is almost surely bounded with respect to L^2-norm, and there exists a constant $M_0 > 0$ such that for any $\beta \in H^m(\mathbb{I})$,*

$$E\left\{\left|\int_0^1 X^l(t)\beta(t)dt\right|^4\right\} \le M_0\left[E\left\{\left|\int_0^1 X^l(t)\beta(t)dt\right|^2\right\}\right]^2. \tag{5.7}$$

Given that $\|X^l\|_{L^2} \le c$ *a.s.* for some constant $c > 0$, there exists some constant $s \in (0, 1)$ such that

$$E\left\{\exp\left(s\|X\|_{L^2}\right)\right\} < \infty, \tag{5.8}$$

which implies that $B(X^l)$ is bounded away from zero. In particular, there exists some positive constant $C_2 \le 1$ such that $C_2^{-1} \le B(X^l) \le C_2$ a.s.

The Fréchet derivative of $\ell_{n,\lambda}(\theta)$ w.r.t. θ is then given by

$$\begin{aligned} S_{n,\lambda}(\theta)\Delta\theta \equiv & D\ell_{n,\lambda}(\theta)\Delta\theta = \frac{1}{n}\sum_{i=1}^{n_0}\left[\dot{\ell}_0\left(\left\langle R_{X_i^0}, \theta\right\rangle\right)\left\langle R_{X_i^0}, \Delta\theta\right\rangle\right. \\ & \left. + \sum_{i=1}^{n_1}\dot{\ell}_1\left(\left\langle R_{X_i^1}, \theta\right\rangle\right)\left\langle R_{X_i^1}, \Delta\theta\right\rangle\right] - \langle P_\lambda\theta, \Delta\theta\rangle \\ \equiv & S_{n,\lambda}^0(\theta)\Delta\theta + S_{n,\lambda}^1(\theta)\Delta\theta \end{aligned}$$

where $S_{n,\lambda}^l(\theta)\Delta\theta = \frac{1}{n}\sum_{i=1}^{n_l}[\dot{\ell}_l(\langle R_{X_i^l}, \theta\rangle)\langle R_{X_i^l}, \Delta\theta\rangle - \langle P_\lambda\theta, \Delta\theta\rangle]$, $\Delta\theta = (\Delta\alpha, \Delta\beta)$, and $\Delta\theta_j = (\Delta\alpha_j, \Delta\beta_j)$ for $j = 1, 2, 3$. Recall that $X = \left(X_1^0, \ldots, X_{n_0}^0, X_1^1, \ldots, X_{n_1}^1\right)^\top$ represents the design matrix without intercept. The second- and third-order Fréchet derivatives of $\ell_{n,\lambda}(\theta)$ can be compactly written as: $DS_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\ddot{\ell}\left(\langle R_{X_i}, \theta\rangle\right)\langle R_{X_i}, \Delta\theta_1\rangle$ $\langle R_{X_i}, \Delta\theta_2\rangle - \langle P_\lambda\Delta\theta_1, \Delta\theta_2\rangle$ and $D^2S_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2\Delta\theta_3 = \frac{1}{n}\sum_{i=1}^{n}\ell'''\left(\langle R_{X_i}, \theta\rangle\right)\langle R_{X_i}, \Delta\theta_1\rangle$

$\langle R_{X_i}, \Delta\theta_2 \rangle \langle R_{X_i}, \Delta\theta_3 \rangle$. where $\ddot{\ell}(a) = -p_0(a)p_1(a), p_0 = 1/(1+\exp(a)), p_1(a) = 1 - p_0(a)$ and ℓ''' are the derivative of $\ddot{\ell}(a)$ w.r.t. a. We further define $S_n(\theta) = [\sum_{i=1}^{n_0} \dot{\ell}_0(\langle R_{X_i^0}, \theta\rangle) R_{X_i^0} + \sum_{i=1}^{n_1} \dot{\ell}_1(\langle R_{X_i^1}, \theta\rangle) R_{X_i^1}]/n$,

$$\begin{aligned} S(\theta) &= \rho_0 E_0 \left\{ \dot{\ell}_0(\langle R_{X^0}, \theta\rangle) R_{X^0} \right\} + \rho_1 E_1 \left\{ \dot{\ell}_1(\langle R_{X^1}, \theta\rangle) R_{X^1} \right\} \\ &\equiv S_0(\theta) + S_1(\theta), \end{aligned}$$

and $S_\lambda(\theta) = S(\theta) - P_\lambda \theta = \rho_0 E_0 \left\{ \dot{\ell}_0(\langle R_{X^0}, \theta\rangle) R_{X^0} \right\} + \rho_1 E_1 \left\{ \dot{\ell}_1(\langle R_{X^1}, \theta\rangle) R_{X^1} \right\} - P_\lambda \theta$.

We now introduce the functional Bahadur representation for the penalized estimators under the case-control functional logistic model, which provides a unified treatment for the upcoming inference problems on $\beta_0(\cdot)$. Note that for any $\theta = (\alpha, \beta) \in \mathcal{H}$, we have $\|\theta\|_2 = |\alpha| + \|\beta\|_{L^2}$. It has been demonstrated in [164] that there is a specific relationship between the norms $|\cdot|_2$ and $|\cdot|$, which is summarized in the following lemma.

Lemma 5.1 *There exists a constant $\kappa > 0$ such that for any $\theta \in \mathcal{H}$, $\|\theta\|_2 \leq \kappa h^{-(2a+1)/2} \|\theta\|$, where $h \equiv \lambda^{1/(2k)}$ for some constant k, a defined in condition 5.3.*

To obtain the Bahadur representation, we first establish the concentration inequality through lemma 5.2. Let $E_l\{\cdot\}$ denote the expectation over control or case populations $l = 0, 1$. We define

$$\begin{aligned} H_n(\theta) &= \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n_0} \left[\psi_n\left(X_i^0; \theta\right) R_{X_i^0} - E_0\left\{ \psi_n(X^0; \theta) R_{X^0} \right\} \right] \right. \\ &\quad \left. + \sum_{i=1}^{n_1} \left[\psi_n\left(X_i^1; \theta\right) R_{X_i^1} - E_1\left\{ \psi_n(X^1; \theta) R_{X^1} \right\} \right] \right] \\ &\equiv \frac{1}{\sqrt{n}} \left[H_n^0(\theta) + H_n^1(\theta) \right], \end{aligned}$$

$$\mathcal{F}_{p_n} = \{\theta = (\alpha, \beta) \in \mathcal{H} : |\alpha| \leq 1, \|\beta\|_{L^2} \leq 1, J(\beta, \beta) \leq p_n\},$$

where $X^l \in \mathcal{X}$, $\psi_n(X; \theta)$ is a function over $\mathcal{X} \times \mathcal{H}$ might depend on n, and $p_n \geq 1$.

Lemma 5.2 *Assume Conditions 5.1 - 5.4 hold.* $\psi_n\left(X_i^l;0\right)=0$ *a.s.. Additionally, there exists a constant* $C_\psi>0$ *s.t. the following Lipschitz continuity holds:*

$$|\psi_n(X;\theta_1)-\psi_n(X;\theta_2)|\leq C_\psi\|\theta_1-\theta_2\|_2 \quad \textit{for any } \theta_1,\theta_2\in\mathcal{F}_{p_n}. \tag{5.9}$$

Then let $\gamma=1-1/(2m)$*, as* $n\to\infty$*,*

$$\sup_{\theta\in\mathcal{F}_{p_n}}\frac{\|H_n(\theta)\|}{p_n^{1/(4m)}\|\theta\|_2^\gamma+n^{-1/2}}=O_P\left(\left(h^{-1}\log\log n\right)^{1/2}\right).$$

Condition 5.5

$$\left\|\widehat{\theta}_{n,\lambda}-\theta_0\right\|=O_P((nh)^{-1/2}+h^k).$$

for k *specified in condition 5.3.*

This assumption is concerned with the convergence rate of $\widehat{\theta}_{n,\lambda}$.

Proposition 5.3 *Assume Conditions 5.1 to 5.4 hold, and the following rate conditions on* h *(or equivalently,* λ*) are satisfied:*

$$\begin{gathered} h=o(1),\quad n^{-1/2}h^{-1}=o(1),\\ n^{-1/2}h^{-(a+1)-((2k-2a-1)/(4m))}(\log n)(\log\log n)^{1/2}=o(1).\end{gathered} \tag{5.10}$$

The condition 5.5 is then satisfied.

This proposition shows that the convergence rate of $\widehat{\theta}_{n,\lambda}$ stated earlier can be achieved if conditions 5.2 to 5.4are satisfied and the smoothing parameter λ is properly chosen. No estimation consistency is required in Proposition 5.3.

We are now able to establish the Bahadur representation for the proposed estimators.

Theorem 5.4 (Bahadur representation for functional data) *Assume Conditions 5.1 to 5.5 hold.* $h=o(1)$ *and* $\log\left(h^{-1}\right)=O(\log n)$ *as* $n\to\infty$*. Furthermore, equation* (5.7) *holds. Then, as* $n\to\infty$*,*

$$\left\|\widehat{\theta}_{n,\lambda}-\theta_0-S_{n,\lambda}\left(\theta_0\right)\right\|=O_P\left(a_n\right), \tag{5.11}$$

where $a_n=n^{-1/2}h^{-(4ma+6m-1)/(4m)}r_n(\log n)(\log\log n)^{1/2}+h^{-1/2}r_n^2$*, and* $r_n\equiv(nh)^{-1/2}+h^k$*.*

The Bahadur representation greatly facilitates the derivation of the joint limit distribution for the functional slope estimate. Using this representation, we then derive the following pointwise limit distribution of the slope function estimate.

Theorem 5.5 *Suppose that the conditions of Theorem 5.4 are satisfied,* $\sup_{j\geq 1}\|\varphi_j\|_{\sup} \leq C_\varphi j^a$ *for* $j \geq 1$. *As* $n \to \infty, nh^{2a+1}(\log(1/h))^{-4} \to \infty$, $n^{1/2}a_n = o(1)$ *and* $\sum_{j=1}^{\infty} \frac{|\varphi_j(t)|^2}{(1+\lambda\rho_j)^2} \asymp h^{-(2a+1)}$. *Then we have for any* $t \in \mathbb{I}$,

$$\frac{\sqrt{n}\left(\widehat{\beta}_{n,\lambda}(t) - \beta_0(t) + (W_\lambda\beta_0)(t)\right)}{\sqrt{\sum_{j=1}^{\infty}\left(|\varphi_j(t)|^2/(1+\lambda\rho_j)^2\right)}} \xrightarrow{d} N(0,1).$$

Additionally, if $\sqrt{n}(W_\lambda\beta_0)(t)/\sqrt{\sum_{j=1}^{\infty}\frac{|\varphi_j(t)|^2}{(1+\lambda\rho_j)^2}} = o(1)$, *then*

$$\frac{\sqrt{n}\left(\widehat{\beta}_{n,\lambda}(t) - \beta_0(t)\right)}{\sqrt{\sum_{j=1}^{\infty}|\varphi_j(t)|^2/(1+\lambda\rho_j)^2}} \xrightarrow{d} N(0,1).$$

Theorem 5.5 provides the convergence rate of the local estimate $\widehat{\beta}_{n,\lambda}(t)$ as $\sqrt{nh^{2a+1}}$. The factor a (defined in Condition 5.3) generically reflects the impact of the covariance operator on the convergence rate. The condition $\sqrt{n}(W_\lambda\beta_0)(t)/\sqrt{\sum_{j=1}^{\infty}\frac{|\varphi_j(t)|^2}{(1+\lambda\rho_j)^2}} = o(1)$ holds if $nh^{4k} = o(1)$ and the true slope function $\beta_0 = \sum_j b_j\varphi_j$ satisfies $\sum_j b_j^2\rho_j^2 < \infty$ [164]. Under the condition of Theorem 5.5, the asymptotic bias for the estimation of $\beta_0(t_0)$ vanishes at any fixed point $t_0 \in \mathbb{I}$. The pointwise confidence interval for $\beta_0(t_0)$ is given by

$$P\left(\beta_0(t_0) \in \left[\widehat{\beta}_{n,\lambda}(t_0) \pm z_{\xi/2}\frac{\sqrt{\sum_{j=1}^{\infty}\left(|\varphi_j(t)|^2/(1+\lambda\rho_j)^2\right)}}{\sqrt{n}}\right]\right) \to 1-\xi$$

as $n \to \infty$, where z_ξ is the lower ξth quantile of $\Phi(\cdot)$, the standard normal cumulative distribution function, that is $\Phi(z_\xi) = 1-\xi$.

5.4 Local Case Control Sampling for Semiparametric Functional Linear Models

Although CC subsampling, thanks to its efficiency than uniform subsampling when the datasets are marginally imbalanced, has been widely used in practice in epidemi-

ology and social science studies [111], the distribution of subsampled data is skewed by the sample selection process since the acceptance probability relies on the response variable [47, 62, 110]. It follows that correction methods are needed to adjust for the selection bias. Another method to remove bias in CC subsampling is to weigh each sampled data point by the inverse of its acceptance probability. This is known as the weighted case-control (WCC) method, which is consistent and unbiased [47, 110], but may sacrifice the variance of the resulting estimator.

As a step forward, we will introduce the estimation framework with the local case-control (LCC) sampling scheme [47, 62] for the unbalanced functional data through a pilot experiment in this section. LCC has proven consistent in parametric logistic regression [47] and excels the CC and WCC scheme for its relative efficiency. Our extension is even versatile to incorporate scalar parameters. Unlike correction-based methods that are specialized for certain models such as linear model, the maximum likelihood estimation (MLE) is proposed that integrates the correction into the MLE formulation, and this approach allows us to deal with arbitrary sampling probability and produces a consistent estimator within the original model family as long as the underlying logistic model is correctly specified.

Suppose that $T_i = (Y_i, Z_i, X_i)$, $i = 1, \ldots, n$, are i.i.d. copies of $T = (Y, Z, X)$, where $Y \in \mathcal{Y}$, $U = (X, Z) \in \mathcal{U}$ is the dichotomous response and covariate variables, $Z \in \mathbb{R}^p$ and $X(\cdot)$ are the scalar and functional covariates, respectively. With a little abuse of the notation, assume $\theta = (\alpha, \beta)$ where $\alpha \in \mathbb{R}^p$, $\beta(\cdot)$ are the unknown scalar parameters and coefficient function. The full parameter space for θ is $\mathbb{R}^p \times H^m(\mathbb{I}) \equiv \mathcal{H}$. The semi-functional linear logistic regression is posed as

$$P(Y = 1 \mid U = u) = \frac{\exp\{Z^\top \alpha + \int x(t)\beta(t)\mathrm{d}t\}}{1 + \exp\{Z^\top \alpha + \int x(t)\beta(t)\mathrm{d}t\}} \equiv p_1(u; \theta), \tag{5.12}$$

Recall that the true value of θ is $\theta_0^* = (\alpha_0^*, \beta_0(t))^\top$.

Case-control sampling typically involves selecting all cases and a fixed multiple c of controls (We choose $c = 1$ in subsection 5.6.1). A nearly equivalent, simpler

method is accept-reject sampling with acceptance probability $f(y)$ and log-selection bias $b = \log \frac{f(1)}{f(0)}$. The debiasing algorithm leading to all consistent estimators is:

(a) Generate $s_i \sim \text{Bernoulli}(f(y_i))$ independently.

(b) Fit logistic regression to $S = \{T_i : s_i = 1\}$ to get unadjusted $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$.

(c) Adjust $\hat{\alpha} = \hat{\alpha}_S - b$; keep $\hat{\beta} = \hat{\beta}_S$.

S yields an i.i.d. sample from an augmented population $\mathcal{S}$ where

$$\mathbb{P}_{\mathcal{S}}(T) = \mathbb{P}(T|S=1) = \frac{f(Y)\mathbb{P}(T)}{\bar{f}} \tag{5.13}$$

with $\bar{f} = f(1)\mathbb{P}(Y=1) + f(0)\mathbb{P}(Y=0)$ being $S=1$'s marginal probability. We have the log-odds function $\eta(u)$ for the biased population $\mathbb{P}_S$ as

$$\begin{aligned} \eta(u) &= \log \frac{\mathbb{P}(Y=1|U=u,S=1)}{\mathbb{P}(Y=0|U=u,S=1)} \\ &= \log \frac{\mathbb{P}(Y=1|U=u)}{\mathbb{P}(Y=0|U=u)} + \log \frac{\mathbb{P}(S=1|Y=1,U=u)}{\mathbb{P}(S=1|Y=0,U=u)} \\ &= a(u) + b. \end{aligned}$$

Here, $\eta(u)$ is shifted by b from the original population's log-odds $a(u)$. If the model is correct, logistic regression on the subsample consistently estimates $\eta(u)$, and thus $a(u)$. If b depends on u, we have $\eta(u) = a(u) + b(u)$.

If a is correctly specified, then the following function family

$$\eta(\boldsymbol{u};\theta) = a(\boldsymbol{u},\theta) + \log \frac{f(\boldsymbol{u},1)}{f(\boldsymbol{u},0)} \tag{5.14}$$

for $\mathcal{S}$, i.e. the true parameter θ_0^* in (5.12) also satisfies $\eta(u;\theta_0^*) = \log \frac{\mathbb{P}(Y=1|U=u,S=1)}{\mathbb{P}(Y=0|U=u,S=1)}$. The LCC sampling uses this identity and can be summarized as follows [47, 62]:

(a) For each pair of observation $(\boldsymbol{u}_i, y_i), i = 1, \ldots, n$, generate a random binary variable $s_i \in \{0,1\}$, drawn from the $\{0,1\}$-valued Bernoulli distribution $\mathcal{B}(\boldsymbol{u}_i, y_i)$ with acceptance probability

$$\mathbb{P}_{\mathcal{B}(\boldsymbol{u}_i,y_i)}(s_i = 1) = f(\boldsymbol{u}_i, y_i);$$

where $f(\boldsymbol{u}, y) \in [0,1]$ is arbitrary given sampling probability function.

(b) Keep the samples with $s_i = 1$ for $i \in \{1, \ldots, n\}$. Fit a semi-functional logistic regression based on the selected samples by solving the optimization problem

$$\arg\sup_{(\alpha,\beta)\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} s_i \left[y_i \cdot a(\boldsymbol{u}_i;\theta) - \log\left(1 + \frac{f(\boldsymbol{u}_i,1)}{f(\boldsymbol{u}_i,0)} e^{a(\boldsymbol{u}_i;\theta)}\right)\right] - (\lambda/2)J(\beta,\beta).$$

Consequently, the computational cost in the second step is greatly reduced to fitting the model with subsamples with sample size $\sum_{i=1}^{n} s_i$.

It follows that given arbitrary sampling probability function $f(\boldsymbol{u}, y) \in [0,1]$, θ_0^* can be obtained by using MLE to the new population $\mathcal{S}$:

$$\arg\sup_{(\alpha,\beta)\in\mathcal{H}} F(\theta) := \mathbb{E}_{\boldsymbol{u},y,s\sim\mathcal{S}}\ s\left[y \cdot \eta(\boldsymbol{u};\theta) - \log\left(1 + e^{\eta(\boldsymbol{u};\theta)}\right)\right] - (\lambda/2)J(\beta,\beta). \quad (5.15)$$

In practice, the model parameter can be estimated by the modified empirical conditional MLE to the sampled data $\{(\boldsymbol{u}_i, y_i, s_i) : i = 1, \ldots, n\}$:

$$\arg\sup_{(\alpha,\beta)\in\mathcal{H}} \hat{F}_{n,\lambda}(\theta) := \frac{1}{n}\sum_{i=1}^{n} s_i \left[y_i \cdot \eta(\boldsymbol{u}_i;\theta) - \log\left(1 + e^{\eta(\boldsymbol{u}_i;\theta)}\right)\right] - (\lambda/2)J(\beta,\beta), \quad (5.16)$$

Let $\widehat{\theta}_{Sub} = \arg\sup_{(\alpha,\beta)\in\mathcal{H}} \hat{F}_{n,\lambda}(\theta)$ be the LCC based estimator. In the next section, we will show that $\widehat{\theta}_{Sub}$ is a consistent estimator of θ_0^* when the model is correctly specified.

We describe local case-control subsampling, a generalization of standard case-control sampling that both improves on its efficiency and resolves its problem of inconsistency. To achieve these benefits, we require a pilot estimate, that is, a good guess $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$ for the population-optimal θ_0^*.

Local case-control sampling differs from case-control sampling only in that the acceptance probability f is allowed to depend on u as well as y [47, 62]. Our criterion for selection will be the degree of "surprise" we experience upon observing y_i given u_i:

$$f(u,y) = \left|y - \tilde{p}(u)\right| = \begin{cases} 1 - \tilde{p}(u), & y = 1, \\ \tilde{p}(u), & y = 0, \end{cases} \quad (5.17)$$

where $\tilde{p}(u) = \frac{\exp\{Z^\top\tilde{\alpha} + \int x(t)\tilde{\beta}(t)\mathrm{d}t\}}{1+\exp\{Z^\top\tilde{\alpha} + \int x(t)\tilde{\beta}(t)\mathrm{d}t\}}$ is the pilot estimate of $\mathbb{P}(Y = 1|U = u)$. If $a(u)$ is well approximated by the pilot estimate, then $\eta(u) \approx 0$ throughout feature space. That is, conditional on selection into $\mathcal{S}$, y_i given u_i is nearly a fair coin toss.

To motivate this choice heuristically, recall that the Fisher information for the log odds of a Bernoulli random variable is maximized when the probability is $\frac{1}{2}$: fair coin tosses are more informative than heavily biased ones.

In marginally imbalanced data sets where $\mathbb{P}(Y = 1|U = u)$ is small everywhere in the predictor space, a good pilot has $\tilde{p}(u) \approx 0$ for all u, and the number of cases discarded by this algorithm will be quite small. If we wish to avoid discarding any cases, we can always modify the algorithm so that instead of keeping $(u, 1)$ with probability $f(u, 1)$, we keep it with probability 1 and assign weight $f(u, 1)$.

5.5 Asymptotic Properties of the LCC Estimator

5.5.1 A partially linear extension of RKHS theory

In this subsection, we extend the inner product and corresponding norms defined in (5.4) to our semi-nonparametric setup and examine the asymptotic behavior of the proposed functional LCC method described in Section 5.4.

Denote $\dot{F}(y;\eta), \ddot{F}(y;\eta)$ and $F'''(y;\eta)$ as the first-, second-, and third-order derivatives of $F(y;\eta)$ with respect to η, respectively. Define

$$\begin{aligned} I(U) &\equiv -E\left(\ddot{F}_\eta\left(Y; Z^\top\alpha_0 + \int_0^1 X(t)\beta_0(t)dt\right) \mid U\right) \\ &= E_{S,Y}\left(s \cdot \frac{e^{\eta(U;\theta)}}{(1+e^{\eta(U;\theta)})^2} \mid U\right) \\ &= E_Y\left(f(U,y) \cdot \frac{(f_1 p_1)(f_0 p_0)}{(f_1 p_1 + f_0 p_0)^2} \mid U\right) \\ &= E\left(\frac{f_1 p_1 \cdot f_0 p_0}{f_1 p_1 + f_0 p_0}|U\right) \end{aligned} \tag{5.18}$$

where $f_1 = f(u, 1)$ and $p_1(u;\theta)$ are the arbitrary sampling probability and consistent class probability estimated by the pilot estimation.

The rectified inner product for $\mathcal{H}$ is thus, for any $\theta_1 = (\alpha_1, \beta_1), \theta_2 = (\alpha_2, \beta_2) \in \mathcal{H}$,

$$\begin{aligned} \left\langle \theta_1, \theta_2 \right\rangle \equiv E_U\left\{ I(U)\left(Z^\top\alpha_1 + \int_0^1 X(t)\beta_1(t)\mathrm{d}t\right)\left(Z^\top\alpha_2 + \int_0^1 X(t)\beta_2(t)\mathrm{d}t\right)\right\} \\ +\lambda J\left(\beta_1, \beta_2\right). \end{aligned} \tag{5.19}$$

Under the norm $\|\theta\|^2 = \langle\theta,\theta\rangle$, we shall construct two linear operators, $R_u \in \mathcal{H}$, for any $u \in \mathcal{U}$, and $P_\lambda : \mathcal{H} \mapsto \mathcal{H}$ satisfying

$$\langle R_u, \theta\rangle = Z^\top\alpha + \int_0^1 X(t)\beta(t)\mathrm{d}t \qquad \text{for any } u \in \mathcal{U} \text{ and } \theta \in \mathcal{H} \tag{5.20}$$

$$\langle P_\lambda\theta_1, \theta_2\rangle = \lambda J(\beta_1,\beta_2) \qquad \text{for any } \theta_1, \theta_2 \in \mathcal{H}. \tag{5.21}$$

Define $B(X) = E\{I(U)|X\}$ the newly weighting function to form the weighted covariance operator of X and $K(\beta_1,\beta_2)$ be a (symmetric) reproducing kernel of $H^m(\mathbb{I})$ endowed with the inner product $\langle\beta_1,\beta_2\rangle_1 = V(\beta_1,\beta_2) + \lambda J(\beta_1,\beta_2)$ where $V(\beta_1,\beta_2) \equiv \int_0^1\int_0^1 E_X\{B(X)X(t)X(s)\}\beta_1(t)\beta_2(s)\mathrm{d}t\mathrm{d}s$ in a similar manner as [63, 98].

Define a p-dimensional functional-valued vector $G(X) \equiv (G_1(X),\ldots,G_p(X))^\top$ projecting the covariate vector Z on $X(\cdot)$, which satisfies $E\{I(U)(Z-G(X))X\} = 0$, where $G(X) = E\{I(U)Z|X\}/B(X)$.

We next state a regularity condition guaranteeing the linearity of Z projection on X and positive definiteness of the matrix Ω.

Condition 5.6 *G_k has a finite second moment for $k = 1,\ldots,p$, there exists $\tilde\beta_k$, such that $G_k(X)$ can be represented as $G_k(X) = \int_0^1 X(t)\tilde\beta_k(t)dt$ with $V(\tilde\beta_k,\tilde\beta_k) < \infty$ and the $p\times p$ matrix $\Omega \equiv E\{I(U)(Z-G(X))(Z-G(X))^\top\}$ is positive definite.*

5.5.2 Joint limit distribution

In this subsection, the joint asymptotic normality of both parametric and functional parts is demonstrated when the model is correctly specified, we start with some further assumptions.

Condition 5.7 (a) *$F(y;\eta)$ is three times continuously differentiable and concave w.r.t. η. There exists a bounded open interval $\mathcal{I} \supset \mathcal{I}_0$ and positive constants C_0 and C_1 s.t.*

$$E\Big\{\exp\Big(\sup_{\eta\in\mathcal{I}}\big|\ddot F_\eta(Y;\eta)\big|/C_0\Big)\big|U\Big\} \le C_1 \qquad a.s. \tag{5.22}$$

$$E\Big\{\exp\Big(\sup_{\eta\in\mathcal{I}}\big|F'''_\eta(Y;\eta)\big|/C_0\Big)\big|U\Big\} \le C_1 \qquad a.s. \tag{5.23}$$

(b) *There exists a positive constant* C_2 *s.t.* $C_2^{-1} \leq I(U) \leq C_2$, *a.s.*

(c) $\epsilon \equiv \dot{F}_\eta(Y;\eta)$ *satisfies* $E(\epsilon|U) = 0$, $E(\epsilon^2|U) = I(U)$, *a.s., and* $E\{\epsilon^4\} < \infty$.

Following analogous eigen-system construction in Condition 5.3, all $\beta \in H^m(\mathbb{I})$ have the Fourier representation $\beta = \sum_{v=1}^{\infty} V(\beta, \varphi_v)\varphi_v$. Define $A = (A_1, \ldots, A_p)^\top$ with $A_k(t) = \sum_v V\left(\tilde{\beta}_k, \varphi_v\right)\varphi_v(t)/\left(1 + \lambda\rho_v\right)$.

Condition 5.8 *There exists a constant* $s_2 \in (0,1)$, *such that*

$$E\left\{\exp\left(s_2\left(Z^\top Z\right)^{1/2}\right)\right\} < \infty \tag{5.24}$$

$$E\left\{\exp\left(s_2\left(\left(Z - \langle A, \tau(X)\rangle_1\right)^\top\left(Z - \langle A, \tau(X)\rangle_1\right)\right)^{1/2}\right)\right\} < \infty \tag{5.25}$$

where $\tau(X) = \sum_v X_v\varphi_v(t)/\left(1 + \lambda\rho_v\right)$ *with* $X_v = \int_0^1 X(t)\varphi_v(t)dt$. *Moreover, for any* $\alpha \in \mathbb{R}^p$, *there exists a constant* M_2 *satisfying*

$$E\left(\left|Z^\top\alpha\right|^4\right) \leq M_2\left[E\left(\left|Z^\top\alpha\right|^2\right)\right]^2$$

A similar discussion of condition 5.6 can be found in [63, 98]. Recall that k is specified in condition 5.3, and let $h = \lambda^{1/(2k)}$. The following theorem derives the asymptotics of $\widehat{\theta}_{Sub}$. Define for any $x_0 \in L^2(\mathbb{I})$, $\tilde{x}_0 = x_0 \cdot \sigma_{x_0}^{-1}$ where $\sigma_{x_0}^2 = \sum_{v=1}^{\infty}|x_v^0|^2/(1+\lambda\rho_v)^2$ and $x_v^0 = \int_0^1 x_0(t)\phi_v(t)dt$.

Theorem 5.6 (Joint limit distribution) *Assume Conditions 5.2-5.3,5.6-5.7 holds, and* $\|R_{\tilde{u}}\| = O(1)$ *for any* $\tilde{u} = (\tilde{z}, \tilde{x}_0)$, *and* $E\{exp(s^*|\epsilon|)\} < \infty$ *for some* $s^* > 0$. *Suppose there exists* $b \in ((2a+1)/2k, a/k+1]$ *such that* $\tilde{\beta}_j$ *satisfies*

$$\sum_\nu \left|V(\tilde{\beta}_j, \varphi_\nu)\right|^2\rho_\nu^b < \infty \qquad \textit{for any } j = 1, \ldots, p. \tag{5.26}$$

Furthermore, if $na_n^2 = o(1), nh^{4k} = o(1)$ *and* $nh^{2a+1}(\log n)^{-4} \to \infty$ *hold, and* $\beta_0 = \sum_v b_v\phi_v$ *satisfies the condition* $\sum_v b_v^2\rho_v^2 \leq \infty$, *then as* $n \to \infty$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\alpha}_{Sub} - \alpha_0^*) \\ \frac{\sqrt{n}}{\sigma_{x_0}}\left(\int_0^1 x_0(t)\widehat{\beta}_{Sub}(t)dt - \int_0^1 x_0(t)\beta_0(t)dt\right) \end{pmatrix} \xrightarrow{d} N(0, \Psi), \tag{5.27}$$

where $\widehat{\theta}_{Sub} \equiv (\widehat{\alpha}_{Sub}, \widehat{\beta}_{Sub})$ *and*

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & 1 \end{pmatrix}$$

Condition (5.26) is essential to obtaining the asymptotic independence between the scalar estimators and the estimator of the functional part [98]. It is also vital to guaranteeing the $\sqrt{n}$-consistency of the parametric estimators since it controls the decay rates of the coefficients for the projection $G(X)$. Theorem 5.6 helps to construct the joint asymptotic under LCC sampling for the scalar as well as functional estimates and simplifies the construction of the prediction interval for a new response with given new covariates.

5.6 Simulation Studies

5.6.1 Simulation I: CC scheme

In this section, we evaluate the performance of the proposed methods in estimating the functional and scalar coefficients as well as constructing the pointwise confidence intervals via the following settings.

Estimation and confidence bands of $\beta(\cdot)$

In this subsection, we examine the numerical performance of the proposed procedures for estimation and inference of the slope function $\beta_0(t)$ as well as α. Simulated data are generated from the following functional logistic regression model (FLRM):

$$P(Y = 1 \mid X = x) = \frac{\exp\left\{\alpha_0 + \int x(t)\beta_0(t)\right\}}{1 + \exp\left\{\alpha_0 + \int x(t)\beta_0(t)\right\}}, \tag{5.28}$$

where $Y \in \{0, 1\}$ is the binary response variable, the predictor $X_i(t)$ is observed at 1000 time points over $[0, 1]$. Three different cases are considered below. For each case, the intercept α_0 is set, respectively, to achieve the population percentage of cases $\Pr(Y = 1) = 0.05$, indicating a rare incidence rate. For the proposed method, we collect case-control samples by taking a random of n_1 cases from the case population

and a random sample of n_0 controls from the control population, separately. We present the experimental results for the combinations of $n \in \{300, 500\}$ with three different $\beta(t)$ settings. For simplicity, we set $n_0 = n_1$ for the case-control sampling. For comparison, we also collect prospective samples with sample size $n = n_0 + n_1$ and then compute the corresponding criteria. We take $m = 2$ for the choice of $H^m(\mathbb{I})$ and adopt the aforementioned GACV criterion to select the roughness penalty parameter λ. The results are based on 500 data replications for each case. The nominal significance level is chosen to be 5%.

Case 1: In the first setting, we consider the same one as Setting 4 in [164]. The predictor process X_i is simulated as $X_i(t) = \sum_{j=1}^{100} \sqrt{\lambda_j}\eta_{ij}V_j(t)$, where $\lambda_j = (j - 0.5)^{-2}\pi^{-2}$, $V_j(t) = \sqrt{2}\sin((j-0.5)\pi t), t \in [0,1]$, and $\eta_{ij} = \xi_{ij}I_{\{|\xi_{ij}|\leq 0.5\}} + 0.5I_{\{\xi_{ij}>0.5\}} - 0.5I_{\{\xi_{ij}<-0.5\}}$, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, 100$ with ξ_{ij} being a standard normal random variable. In addition, the true slope function β_0 is set to be $B \cdot 3 \cdot 10^5\{t^{11}(1 - t)^6\}$ for $t \in [0, 1]$, here B controls signal strengths in PLRT test.

Case 2: In the second setting, each $X_i(t)$ is the same Brownian motion simulated as $X_i(t) = \sum_{j=1}^{100} \sqrt{\lambda_j}\eta_{ij}V_j(t)$, where η_{ij} 's are independent standard normal for $i = 1, \ldots, n$ and $j = 1, \ldots, 100$. Whereas the true slope function is chosen as

$$\beta_0^B(t) = \frac{B}{\sqrt{\sum_{k=1}^{\infty} k^{-2\xi-1}}} \sum_{i=1}^{100} j^{-\xi-0.5}V_j(t), \quad \xi = 0.1.$$

Case 3: We follow the setting in [15]. The functional covariate is generated as $X_i(t) = \sum_{k=1}^{50} \zeta_k\xi_{ik}\psi_k$ for $i = 1, \ldots, n$, where $\zeta_k = (-1)^{k+1}k^{-1}$, and $\psi_1 = 1$ and $\psi_k = \sqrt{2}\cos(k\pi t)$ for $k \geq 2$ are the eigenfunctions of $X(t)$. The random coefficients ξ_{ik} 's are i.i.d generated from uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. The true coefficient function is $\beta_0(t) = B \cdot \sum_{k=1}^{50} \beta_k\psi_k(t)$ with $\beta_k = 4(-1)^{k+1}k^{-2}$.

Table 5.1 summarizes the estimation errors of the estimated slope function $\widehat{\beta}_{n,\lambda}(\cdot)$ using the proposed and prospective approaches, where estimation error is defined as

the integrated mean squared error (IMSE) of the estimate.

$$\text{IMSE} = \frac{1}{500}\sum_{r=1}^{500}\int(\hat{\beta}^{(r)}(t) - \beta(t))^2 dt \tag{5.29}$$

We observe that the proposed method performs well in all three simulation cases. For almost all combinations, the estimation errors of random design are reduced by nearly or more than 50% when applying the case-control design. The performances become superior to the prospective method, as the functional signal (controlled by B) becomes stronger. In particular, for case 3, the large estimation error of the prospective method is no surprise, considering the large mean estimate deviation in Figure 5.1.

Figures 5.1 depict the average of the estimated coefficient functions and the corresponding pointwise confidence interval compared with the true functions. It can be seen that for cases 2-3, the estimates from our proposed case-control design match well with the true function except for the deviation at the left end for our estimate since the true slope function oscillates vigorously at that location with insufficient data basis. In Case 1, our approach also tracks the true function better than the prospective counterpart. Additionally, the confidence bands from all case-control designs are much narrower than that of random sampling.

Table 5.1: Estimation errors with the standard errors in parentheses for the three simulated cases with three signal strengths $B \in \{0.1, 0.5, 1\}$.

Setting	n	Method	$B = 0.1$	$B = 0.5$	$B = 1$
Case 1	300	proposed	0.030(2.382)	0.210(2.855)	0.746(3.135)
		prospective	0.067(5.728)	0.722(6.149)	2.125(8.929)
	500	proposed	0.013(2.926)	0.214(2.363)	0.722(2.622)
		prospective	0.018(4.366)	0.367(4.601)	1.262(4.716)
Case 2	300	proposed	0.003(1.516)	0.025(1.467)	0.104(1.445)
		prospective	0.012(2.750)	0.048(2.917)	0.110(3.122)
	500	proposed	0.001(1.304)	0.021(1.192)	0.088(1.251)
		prospective	0.005(2.239)	0.042(2.335)	0.120(2.424)
Case 3	300	proposed	.0006(0.807)	0.018(0.910)	0.076(1.139)
		prospective	.0056(1.527)	0.748(12.24)	12.54(28.05)
	500	proposed	.0007(0.692)	0.013(0.718)	0.045(0.935)
		prospective	.0038(1.214)	0.085(1.566)	1.252(3.731)

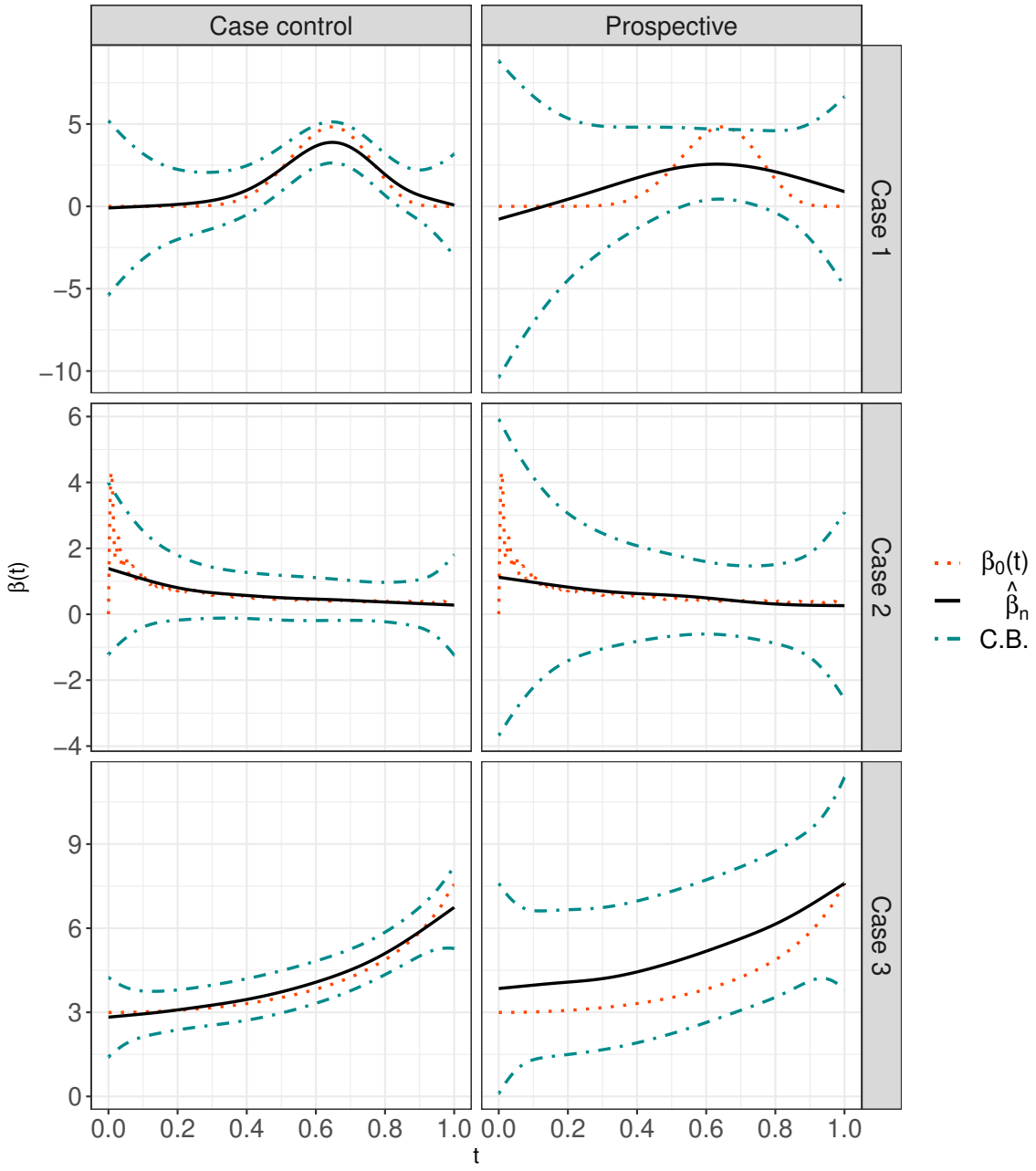


Figure 5.1: True slope function $\beta_0(t)$, estimates $\hat{\beta}$, and 95% confidence band (C.B.) from the proposed CC v.s. prospective methods under three cases with $n_0 = 250, n = 500$. In each panel, the dotted orange line represents the true slope function, the solid black and dot-dash blue lines depict the mean estimates of the slope function and the estimated 95% pointwise confidence bands, respectively.

Inference on $\beta(\cdot)$ via PLRT under CC scheme

We investigate the performance of the PLRT test under the null hypothesis H_0 : $\beta = 0$. Tables 5.2-5.4 summarize the sizes and powers under different cases. The parameter is set to $B \in \{0, 0.1, 0.5, 1\}$. Note that we obtain the size of the tests when $B = 0$ and the power when $B = 0.1, 0.5, 1$. For $B = 0$, the sizes of the proposed test are closer to the nominal level of 5% than the prospective counterpart with highly imbalanced samples, and the performance gets better as the sample size increases. We also observe that the powers of all tests increase as the sample size and the signal strength increase. In particular, the proposed method generally performs better than the prospective test since the proposed method incorporates more information from the case samples. In contrast, the powers of the prospective method increase much slower than the proposed method. In addition, it can be seen that for stronger signals $B = 1$, especially in cases 1 and 3, the powers of the proposed PLRT approach one as the sample size gets larger, demonstrating the efficiency and asymptotic property of the test, while in Case 2, the PLRT power exceeds 80% for $n_0 = 250$.

Table 5.2: The empirical sizes and powers for testing $H_0 : \beta_0(t) = 0$.

n	Method	$B = 0$	$B = 0.1$	$B = 0.5$	$B = 1$
300	CC	0.054(±0.226)	0.089(±0.279)	0.369(±0.483)	0.915(±0.420)
	US	0.073(±0.260)	0.068(±0.285)	0.098(±0.297)	0.229(±0.252)
500	CC	0.057(±0.232)	0.082(±0.275)	0.538(±0.499)	0.993(±0.083)
	US	0.040(±0.196)	0.048(±0.214)	0.113(±0.317)	0.301(±0.459)

5.6.2 Simulation II: LCC scheme

In this section, we compare the proposed functional LCC approach to all its consistent ancestors, the standard weighted (WCC), unweighted case-control sampling (CC) with bias correction and uniform sampling (US), concerning the following semi-

Table 5.3: Case 2: the empirical sizes and powers for testing $H_0 : \beta_0(t) = 0$.

n	Method	$B = 0$	$B = 0.1$	$B = 0.5$	$B = 1$
300	CC	0.086(±0.281)	0.093(±0.291)	0.253(±0.435)	0.623(±0.485)
	US	0.071(±0.257)	0.079(±0.270)	0.091(±0.288)	0.147(±0.354)
500	CC	0.092(±0.289)	0.106(±0.308)	0.340(±0.474)	0.824(±0.381)
	US	0.071(±0.257)	0.063(±0.243)	0.093(±0.291)	0.182(±0.386)

Table 5.4: Case 3: the empirical sizes and powers for testing $H_0 : \beta_0(t) = 0$.

n	Method	$B = 0$	$B = 0.1$	$B = 0.5$	$B = 1$
300	CC	0.121(±0.326)	0.781(±0.414)	1.000(±0.000)	1.000(±0.000)
	US	0.093(±0.291)	0.199(±0.399)	0.950(±0.218)	0.951(±0.216)
500	CC	0.132(±0.339)	0.934(±0.248)	1.000(±0.000)	1.000(±0.000)
	US	0.091(±0.288)	0.291(±0.454)	0.998(±0.032)	0.997(±0.055)

functional logistic regression model (sFLRM)

$$P(Y = 1 \mid X = x) = \frac{\exp\left\{Z^\top \alpha_0 + \int x(t)\beta_0(t)\right\}}{1 + \exp\left\{Z^\top \alpha_0 + \int x(t)\beta_0(t)\right\}} \tag{5.30}$$

We first generate a large ($n = 10^5$) sample from the population described in cases 1-3 of Simulation I in section 5.6.1. The additional $Z_i \in \mathbb{R}^2$ is generated from the standard normal distribution $\mathcal{N}(0, 1)$. Second, a pilot model using the unweighted case-control method with bias correction on $n_{pilot} = 500$ data points. Next, we conduct the functional LCC sampling according to that pilot model. Moreover, we generate another $n_{test} = 100,\ 000$ data points to test the prediction accuracy of different methods.

For fair comparisons, we obtain standard case-control (CC), weighted case-control (WCC), and uniform sampling (US) estimates using the total number of observations $n_{LCC} + n_{pilot}$ seen by the LCC model as well as the pilot model, so the LCC estimate must pay for its pilot sample. The entire procedure is again replicated 500 times for

each case.

Table 5.5: Estimation errors and standard errors for the three simulated cases

	Method	Bias($\widehat{\beta}_{n,\lambda}$)	s.e.	Bias($\widehat{\alpha}_{n,\lambda}$)	s.e	Bias($\widehat{\theta}_{n,\lambda}$)	s.e.
Case 1	LCC	3.4073	2.608	0.0059	0.0055	3.4132	2.6084
	CC	3.0778	1.4989	0.0045	0.0041	3.0823	1.4984
	WCC	3.4801	1.2154	0.0931	0.0426	3.5732	1.2198
	US	5.0993	3.9819	0.0152	0.0167	5.1145	3.980
Case 2	LCC	0.215	0.2294	0.0877	0.0394	0.3027	0.2342
	CC	0.3604	0.5445	0.0038	0.0033	0.3642	0.5445
	WCC	0.4449	0.6842	0.0051	0.0047	0.45	0.6843
	US	0.8382	1.2498	0.013	0.0139	0.8512	1.2497
Case 3	LCC	0.4042	0.2964	0.0442	0.0584	0.4483	0.3356
	CC	0.4227	0.2425	0.0327	0.0375	0.4554	0.2558
	WCC	0.6702	0.2991	0.1027	0.1329	0.7729	0.3441
	US	1.044	0.8721	0.1913	0.2921	1.2354	1.1132

Table 5.5 presents the squared bias and variance of the estimators $\hat{\alpha}_{Sub}$, $\hat{\beta}_{Sub}$, and $\hat{\theta}_{Sub}$, across 500 simulations for each of the four methods: LCC, CC, WCC, and US. The LCC approach demonstrates superior performance with relatively smaller estimation errors and lower standard errors than CC, WCC, and US sampling techniques.

The LCC method exhibits a consistently smaller bias compared to WCC and a reduced variance relative to CC. This dual advantage positions LCC as a more robust approach than its counterparts. Specifically, the bias of LCC is reduced by up to 30% compared to WCC, and its variance is decreased by approximately 25% concerning CC. Noticeably, there are substantial enhancements in case 3 concerning the precision of the estimates.

The variance improvement seen with LCC over CC can be attributed to the mitigation of conditional imbalance in the subsample selection. On the other hand, the bias

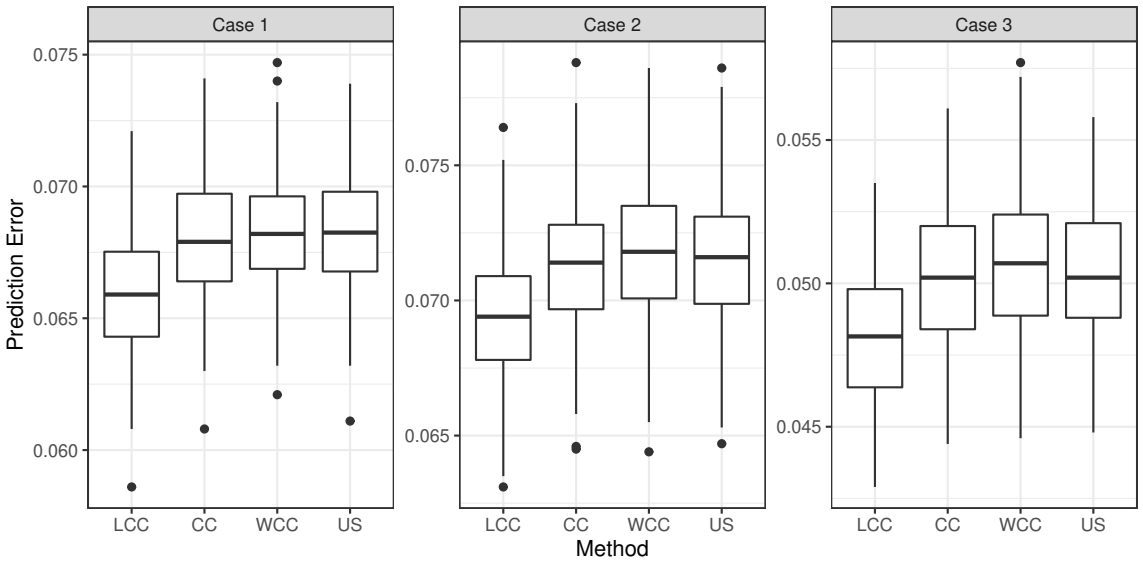


Figure 5.2: Prediction errors on test data across the four approaches in 3 cases

reduction compared to WCC suggests that while these methods are unbiased in theory as sample sizes approach infinity, LCC is more closely aligned with its asymptotic behavior even in finite samples.

Figure 5.2 elucidates prediction errors on test data across the four approaches in 3 cases. There is a uniform pattern resembling the results shown in estimation results. The LCC method always obtains the most accurate and robust predictions. The performance hierarchy generally observed is CC outperforming WCC, which in turn typically surpasses US. LCC consistently outperforms all three alternative methods. When considering the relative efficiency, LCC often exhibits an efficiency gain of over 40% against CC and even higher against WCC and US.

The LCC subsamping not only reduces bias and standard error more effectively than debiased CC but also outperforms WCC and US in most scenarios. With its ability to maintain high accuracy with reduced data usage, LCC stands out as a highly efficient approach for statistical estimation, especially in situations where computational resources are at a premium or when dealing with large datasets that make full-sample analysis impractical. This efficiency gain translates into tangible improvements in statistical inference and predictive modeling.

5.7 Real Data Application

5.7.1 Multiple Sclerosis (MS) Data

In this subsection, we apply the CC method to classify corpus callosum tracts as originating from either MS patients or control patients, as described in [39, 50]. The corpus callosum (CCA) is the largest white matter structure in the brain, connecting the left and right cerebral hemispheres. Its tracts are composed of axons surrounded by fatty myelin sheaths, which are crucial for signal transmission. In the case of multiple sclerosis, the myelin sheaths surrounding the axons are damaged, resulting in severe disability. By using diffusion tensor imaging (DTI) tractography [6], a magnetic resonance imaging (MRI) technique that measures water diffusivity within white-matter tracts, it is possible to diagnose and study multiple sclerosis (MS).

Neuroimaging-based medical examinations are crucial for diagnosing Multiple Sclerosis (MS) and other conditions such as demyelinating diseases, Alzheimer's disease, and epilepsy, etc [156, 157]. In particular, they enable the identification of non-MS related nerve fiber damage, as well as informative biomarkers that enhance the diagnosis of MS. Several studies have established correlations between MS progression and quantitative measures obtained through DTI imaging. The dataset, which is part of the `R refund` package, consists of $n_1 = 334$ cases and $n_0 = 42$ controls. The response variable is coded as 1 (MS) and 0 (nondisease). The functional predictor used to classify a tract is the diffusivity profile of the corpus callosum (CC) fiber tract, which is measured at 93 locations for each subject.

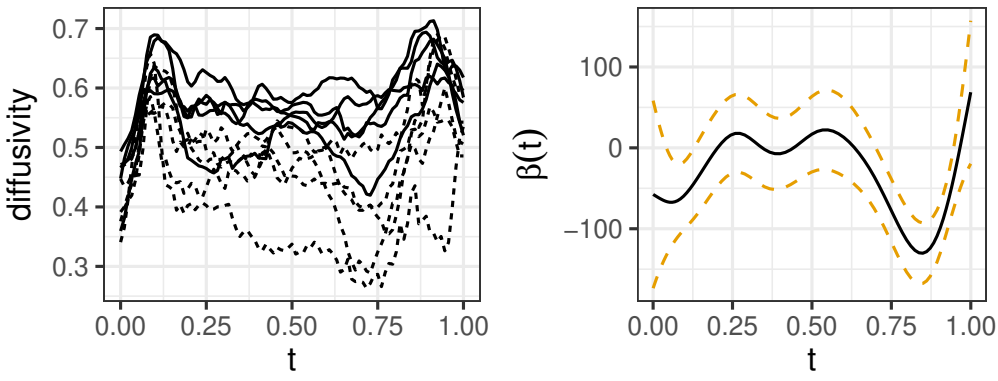


Figure 5.3: Left: Diffusivity profiles for a random sample of five case tracts (dashed) and five control tracts (solid). Right: Coefficient function estimates for the DTI tractography example, the coefficient estimate (solid), and the connected point-wise 95% confidence limits (dashed) from the proposed approach.

The left panel of Figure 5.3 displays the diffusivity profiles of a random sample of five case tracts and five control tracts. The right panel of figure 5.3 depicts the estimated slope functions derived from minimizing objective (5.2). The corresponding 95% pointwise confidence intervals for the estimated slope function are also displayed. To obtain pointwise confidence intervals, we employ the leave one fold out of five, which are depicted in the right panel of Figure 5.3. The estimated slope function with its confidence interval suggests that the effect of the CC fiber tract on the occurrence of MS varies over time, with the effect becoming more significant at the marginal region of the CC fiber tract.

We next investigate the functional testing problem: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ using the proposed PLRT method via RKHS estimator along with the FPCA estimator [15, 17, 98]. The tuning parameter is determined by the GACV criterion. The resulting p-values are 0.000, 0.0012, respectively, which indicates a significant association between the diffusivity profile of the CC tract and the occurrence of MS. This finding is consistent with the coefficient estimation, confidence interval, and previous research reported in [39].

5.7.2 Kidney transplant data

We apply the four consistent subsampling methods to estimate the semi-functional logistic model from the kidney transplant data set.

The kidney transplant data is acquired from the Organ Procurement Transplant Network/United Network for Organ Sharing (Optn/UNOS) as of December 2023, where there are basic descriptions (e.g. age, race, gender, and height) of the kidney transplant recipients at the time of transplant and the information (e.g. serum creatinine, recipient status and the follow-up time) during the followed-up period. This data is available at https://optn.transplant.hrsa.gov/ with the permission of OPTN/UNOS.

The kidneys are a pair of organs in the human body that help maintain a healthy body whose main functions are to remove waste and regulate the chemical (electrolyte) composition of the blood. When renal failure occurs, the kidneys can no longer perform these vital roles, threatening a patient's life. Chronic renal failure, in contrast to acute renal failure, can be treated through kidney transplants. A successful kidney transplant can restore normal renal function and extend the patients' survival time.

However, even after transplant, recipients confront high risks of losing graft function or needing re-transplant. Accurately predicting long-term outcomes post-transplant is therefore crucial. Being able to determine if and how long a recipient will survive would not only help clinicians optimize follow-up care, adjust immunosuppression, and manage complications.

Glomerular filtration rate (GFR) serves as a key predictor to measure kidney function, which considers the blood creatinine level and several associated factors simultaneously, like age, race, and body size. As described in [36, 105], GFR values are calculated through distinct equations respectively for adults (Age$\geq$18) and children (Age$\leq$18). Our goal is to analyze GFR trajectories in the first 6 years post-transplant to predict the recipients' 10-year survival. Optimizing predictive models using GFR

and other factors could help clinicians intervene earlier and recipients make informed choices to maximize their outcomes.

As of December 2023, the UNOS dataset includes over 4M follow-up records, and over 1M kidney recipients. After matching, deleting missing data and feature engineering for the GFR trajactories, there are around 202,831 patients followed for at least 6 years post-transplant. The recipients who die or need to be re-transplanted during the sixth to tenth year ($Y = 0$) accounts for 19.77% (44521/225144) of the whole sample. Fig. 5.4 display the mean GFR trajectories for the two categories.

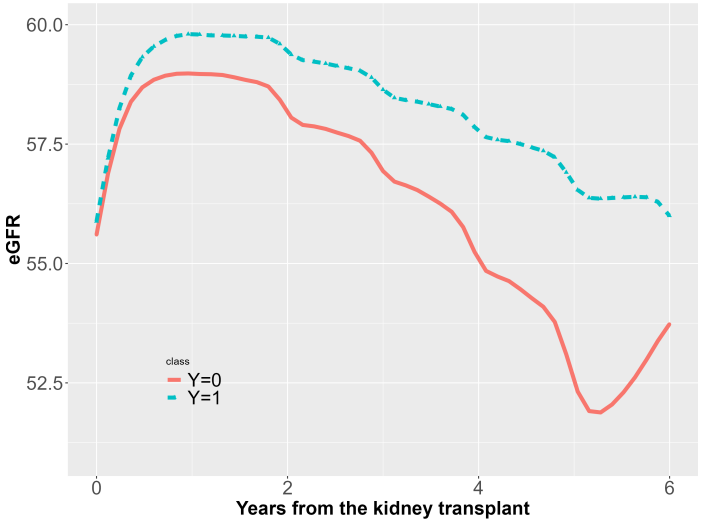


Figure 5.4: The mean GFR curves for the group of recipients who die or need to be retransplanted during the sixth to tenth year after the transplant ($Y = 0$) and the group of recipients who have lived for at least ten years after transplant ($Y = 1$).

We consider fitting a semi-functional logistic regression model:

$$\mathbb{P}\left(Y_i \mid \text{eGFR}_i, Z_i\right) = \frac{\exp\left(Z_i^\top \alpha + \int_0^6 \text{eGFR}_i(t) \cdot \beta(t) dt\right)}{1 + \exp\left(Z_i^\top \alpha + \int_0^6 \text{eGFR}_i(t) \cdot \beta(t) dt\right)}.$$

where Z_i includes the demographic characteristics of the patients: Age of Transplantation, gender and mean height over the 6 years.

We present the prediction error and computational time in Table 6 for the four sampling methods. To evaluate the computational efficiency of the subsampling strategies, we record the CPU times (in seconds) of the four strategies. We use the R programming language (enhanced R distribution Microsoft R 4.1.2) to parallel implement each method. All computations are carried on one HPC node running Linux CentOS 7 with 20 Intel Skylake cores (2.4GHz, AVX512) and around 4GiB RAM per core. The magnitude of our improvement over standard CC, WCC, and US sampling is substantial here but could be much larger in a data set with an even stronger signal. The key point is that standard case-control methods have no way to exploit conditional imbalance, so the more there is, the more local case-control dominates the other methods.

	LCC	CC	WCC	US
Prediction error	0.0734	0.1173	0.1203	0.0990
Computational time(s)	88.356	36.309	40.952	35.261

Table 5.6: prediction error and computational time using kidney dataset for the four sampling methods

5.8 Discussion

The exploration of functional logistic regression under CC as well as LCC design leverages the strengths of RKHS and roughness regularization to address the estimation and inference issues for the coefficients.

The empirical studies, including the application to real-world datasets such as the UNOS kidney transplant data, underscore the practical utility of the proposed methods. These applications highlight the method's ability to handle massive functional data effectively, providing insights into critical phenomena like kidney function post-transplantation. The superior performance of the LCC sampling, in terms of consistency and retaining estimation robustness, is particularly compelling. This effi-

ciency is crucial for statistical estimation and predictive modeling in scenarios where computational resources are limited or full-sample analysis is impractical.

While the article provides a solid foundation for functional logistic regression under two general case-control designs, several avenues for future research emerge from this work. One potential extension involves exploring nonlinear relationships between functional predictors and binary outcomes that could offer further improvements in the estimation of the slope coefficient function. Moreover, incorporating adaptive regularization methods might enhance the model's ability to capture complex system patterns.

Furthermore, extending the framework to accommodate multi-class outcomes could significantly broaden the applicability of functional logistic regression. Many real-world problems involve categorizing observations into more than two groups, and adapting the current methodology to handle such scenarios would be a valuable contribution to the field.

5.9 Proof of the Main Theorems

Recall that for any function $\psi_n(X^l;\theta)$ function over $\mathcal{X}\times\mathcal{H}$, which might depend on n, $E_l\{\cdot\}$ means the expectation over control or case populations $l=0,1$.

5.9.1 Proof of Proposition 5.3

Proof. we first present the proof of Lemma 5.2 that will be useful for the proof of Proposition 5.3.

Proof of lemma 5.2.

For any $\theta_1,\theta_2\in\mathcal{F}_{p_n}$, using the lipschitz continuity (5.9), we have $\|(\psi_n(X_i;\theta)-\psi_n(X_i;\tilde{\theta}))R_{X_i}\|\leq C_\psi\|R_{X_i}\|\cdot\|\theta-\tilde{\theta}\|_2$. Let $\mathcal{R}_n=\{\|R_{X_i}\|\}_{i=1}^n$, where $X=(X_1^0,\ldots,X_{n_0}^0,X_1^1,\ldots,X_{n_1}^1)^\top$. We could apply the same argument as proving Lemma 3.4. in [164], thus as $n\to\infty$,

$$\sup_{\theta\in\mathcal{F}_{pn}} \frac{\|H_n(\theta)\|}{p_n^{1/(4m)}\|\theta\|_2^{\gamma} + n^{-1/2}} = O_P\left(\left(h^{-1}\log\log n\right)^{1/2}\right).$$

Lemma 5.2 holds. ■

The proof of this proposition then proceeds in two parts in analogy to [164]. First, we show that there exists an unique element $\theta_\lambda \in \mathcal{H}$ satisfying $S_\lambda(\theta_\lambda) = 0$ and $\|\theta_\lambda - \theta_0\| = O\left(h^k\right)$. Second, there exists uniquely an element $\widehat{\theta}_{n,\lambda} \in \mathcal{H}$ satisfying $S_{n,\lambda}(\widehat{\theta}_{n,\lambda}) = 0$ and $\|\widehat{\theta}_{n,\lambda} - \theta_0\| = O_P(r_n)$, where $r_n = (nh)^{-1/2} + h^k$. Note such $\widehat{\theta}_{n,\lambda}$ is exactly the smoothing spline estimator since it is the zero of the first-order Fréchet derivative of the penalized likelihood function $\ell_{n,\lambda}(\theta)$.

Lemma 5.7 *For any* $\theta \in \mathcal{H}, \sum_0^1 \rho_l E_l\left\{|\langle R_{X^l}, \theta\rangle|^4\right\} \leq 8\tilde{M}\|\theta\|^4$, *where* $\tilde{M} = \max\{1, M_0\}$.

(I). Define the operator

$$T_{1h}(\theta) = \theta + S_\lambda(\theta_0 + \theta), \theta \in \mathcal{H}.$$

It is easy to see that

$$\|T_{1h}(\theta)\| = \|\theta + S_\lambda(\theta_0 + \theta)\| \leq \|\theta + S_\lambda(\theta_0 + \theta) - S_\lambda(\theta_0)\| + \|S_\lambda(\theta_0)\|.$$

Let $\mathbb{B}(\varepsilon) = \{\theta \in \mathcal{H} : \|\theta\| \leq \varepsilon\}$ be the ball in $\mathcal{H}$ of radius ε. Note $S(\theta_0) = 0$ which implies $S_\lambda(\theta_0) = -P_\lambda\theta_0$. It follows that

$$\|P_\lambda\theta_0\| = \sup_{\|\tilde{\theta}\|=1} \langle P_\lambda\theta_0, \tilde{\theta}\rangle = \sup|\lambda J(\theta_0, \tilde{\theta})| \leq \sqrt{\lambda J(\theta_0, \theta_0)} \sup_{\|\tilde{\theta}\|=1} \sqrt{\lambda J(\tilde{\theta}, \tilde{\theta})}$$

and

$$\|S_\lambda(\theta_0)\| = \|P_\lambda\theta_0\| \leq \sqrt{\lambda J(\beta_0, \beta_0)} \leq (J(\beta_0, \beta_0) + 1)^{1/2} h^k \equiv r_{1n}/2.$$

By the expression of the Fréchet derivative $DS_\lambda(\theta_0)$, we have for any θ, θ', $\langle DS_\lambda(\theta_0)\theta, \theta'\rangle = DS_\lambda(\theta_0)\theta\theta' = -\langle\theta, \theta'\rangle$, which implies $DS_\lambda(\theta_0) = -id$. It then follows by the bound-

edness of $E_l\{\exp(\ddot{\ell}(a))\}$, $E_l\{\exp(\ell'''(a))\}$ that

$$
\begin{aligned}
&\|\theta + S_\lambda(\theta_0+\theta) - S_\lambda(\theta_0)\| \\
=&\left\|\theta + DS_\lambda(\theta_0)\theta + \int_0^1\int_0^1 sD^2S_\lambda(\theta_0+ss'\theta)\theta\theta \mathrm{d}s\mathrm{d}s'\right\| \\
=&\left\|\int_0^1\int_0^1 sD^2S_\lambda(\theta_0+ss'\theta)\theta\theta \mathrm{d}s\mathrm{d}s'\right\| \\
=&\left\|\int_0^1\int_0^1 s\rho_0 E_0\left\{\ell'''\left(\langle R_{X^0},\theta_0+ss'\theta\rangle\right)\left(\langle R_{X^0},\theta\rangle\right)^2 R_{X^0}\right\}\right. \\
&\left.+\int_0^1\int_0^1 s\rho_1 E_1\left\{\ell'''\left(\langle R_{X^1},\theta_0+ss'\theta\rangle\right)\left(\langle R_{X^1},\theta\rangle\right)^2 R_{X^1}\right\}\right\| \\
\leq&\int_0^1\int_0^1 sE\left\{\sum_{l=0}^1 \rho_l E_l\left\{\left|\ell'''\left(\langle R_{X^l},\theta_0+ss'\theta\rangle\right)\right|\right\}\left(\langle R_{X^l},\theta\rangle\right)^2\|R_{X^l}\|\right\} \\
\leq&(1/2)\sum_{l=0}^1 \rho_l E_l\left\{\left(\langle R_{X^l},\theta\rangle\right)^2\|R_{X^l}\|\right\} \\
\leq&(1/2)\sum_{l=0}^1 \rho_l\left\{E_l\left\{|\langle R_{X^l},\theta\rangle|^4\right\}^{1/2} E_l\left\{\|R_{X^l}\|^2\right\}^{1/2}\right\} \\
\leq&(1/2)\sqrt{8\tilde{M}}C_2\|\theta\|^2 C_R^{1/2}h^{-1/2} = C_3\|\theta\|^2h^{-1/2},
\end{aligned}
$$

where $C_3 = C_2\sqrt{2\tilde{M}C_R}$ is an absolute constant, $\tilde{M}$ is specified in lemma S.3. of [164]. Since $C_3r_{1n}h^{-1/2} = O(h^{k-1/2}) = o(1)$, here $k > a+1/2 \geq 1/2$ and $h = o(1)$ lead to that $h^{k-1/2} = o(1)$, as $n\to\infty$, for any $\theta\in\mathbb{B}(r_{1n})$, $\|T_{1h}(\theta)\| \leq C_3h^{-1/2}r_{1n}^2 + r_{1n}/2 < r_{1n}$. So $T_{1h}(\mathbb{B}(r_{1n})) \subset \mathbb{B}(r_{1n})$.

Next, we show that T_{1h} is a contraction mapping. For any $\theta_j = (\alpha_j,\beta_j)\in\mathbb{B}(r_{1n})$ for $j=1,2$. Taylor's expansion leads to that

$$
\begin{aligned}
&T_{1h}(\theta_1) - T_{1h}(\theta_2) \\
=&\theta_1-\theta_2 + S_\lambda(\theta_0+\theta_1) - S_\lambda(\theta_0+\theta_2) \\
=&\int_0^1\left[DS_\lambda(\theta_0+\theta_2+s(\theta_1-\theta_2)) - DS_\lambda(\theta_0)\right](\theta_1-\theta_2)\,\mathrm{d}s \\
=&\int_0^1\int_0^1 D^2S_\lambda(\theta_0+s'(\theta_2+s(\theta_1-\theta_2)))(\theta_1-\theta_2)(\theta_2+s(\theta_1-\theta_2))\,\mathrm{d}s\mathrm{d}s'.
\end{aligned}
$$

Using the similar arguments of $\|T_{1h}(\theta)\|$, we then get

$$
\begin{aligned}
&\|T_{1h}(\theta_1) - T_{1h}(\theta_2)\| \\
\leq &\int_0^1 \int_0^1 E\Big\{ \sum_{l=0}^{1} \rho_l E_l \left\{ |\ell'''(\langle R_{X^l}, \theta_0 + s'(\theta_2 + s(\theta_1 - \theta_2))\rangle)| \mid X^l \right\} \\
&|\langle R_{X^l}, \theta_1 - \theta_2\rangle \langle R_{X^l}, \theta_2 + s(\theta_1 - \theta_2)\rangle| \cdot \|R_{X^l}\| \Big\} \mathrm{d}s\mathrm{d}s' \\
\leq &\int_0^1 \int_0^1 \sum_{l=0}^{1} \rho_l E_l \left\{ |\langle R_{X^l}, \theta_1 - \theta_2\rangle| \cdot |\langle R_{X^l}, \theta_2 + s(\theta_1 - \theta_2)\rangle| \cdot \|R_{X^l}\| \right\} \mathrm{d}s\mathrm{d}s' \\
\leq &\sum_{l=0}^{1} \rho_l C_R^{1/2} h^{-1/2} \int_0^1 \int_0^1 \Big\{ E_l \left\{ |\langle R_{X^l}, \theta_1 - \theta_2\rangle|^4 \right\}^{1/4} \\
&\times E_l \left\{ |\langle R_{X^l}, \theta_2 + s(\theta_1 - \theta_2)\rangle|^4 \right\}^{1/4} \Big\} \mathrm{d}s\mathrm{d}s' \\
\leq &C_R^{1/2} \sqrt{8\tilde{M}} C_2 h^{-1/2} \left(\|\theta_2\| + \|\theta_1 - \theta_2\| \right) \|\theta_1 - \theta_2\| \\
\leq &C_3' h^{-1/2} r_{1n} \|\theta_1 - \theta_2\| < 1/2 \|\theta_1 - \theta_2\|,
\end{aligned}
$$

where $C_3' = 3C_2\sqrt{8\tilde{M}C_R}$, since as $n \to \infty$, $h^{-1/2} r_{1n} = o(1)$. Therefore, T_{1h} is a contraction mapping on $\mathbb{B}(r_{1n})$. By contraction mapping theorem, there exists uniquely an element $\theta_\lambda' \in \mathbb{B}(r_{1n})$ such that $T_{1h}(\theta_\lambda') = \theta_\lambda'$. Define $\theta_\lambda = \theta_0 + \theta_\lambda'$, then we have $S_\lambda(\theta_\lambda) = 0$ and $\|\theta_\lambda - \theta_0\| \leq r_{1n}$.

(II). Define the operator

$$
T_{2h}(\theta) = \theta - \left[DS_\lambda(\theta_\lambda) \right]^{-1} S_{n,\lambda}(\theta_\lambda + \theta).
$$

Here, the invertibility of $DS_\lambda(\theta_\lambda)$ follows the same manner as in [164], where $C_0 = 1, C_1 = e$. The norm of the inverse operator $DS_\lambda(\theta_\lambda)^{-1}$ falls between $(1/2, 3/2)$. Rewrite T_{2h} as

$$
\begin{aligned}
T_{2h}(\theta) = &- DS_\lambda(\theta_\lambda)^{-1} \left[DS_{n,\lambda}(\theta_\lambda)\theta - DS_\lambda(\theta_\lambda)\theta \right] \\
&- DS_\lambda(\theta_\lambda)^{-1} \left[S_{n,\lambda}(\theta_\lambda + \theta) - S_{n,\lambda}(\theta_\lambda) - DS_{n,\lambda}(\theta_\lambda)\theta \right] \\
&- DS_\lambda(\theta_\lambda)^{-1} S_{n,\lambda}(\theta_\lambda) \\
\equiv \quad &I_1 + I_2 + I_3.
\end{aligned}
$$

We will bound the three terms respectively and conclude the proof.

For I_3, recall that $\dot{\ell}_l(a) = l - \exp(a)/(1+\exp(a))$, for $l = 0, 1, i = 1, \ldots, n_l$. Define $\tilde{O}_i^l = \dot{\ell}_l(\langle R_{X_i^l}, \theta_\lambda \rangle) R_{X_i^l}$ and $O_i^l = \tilde{O}_i^l - E\{\tilde{O}_i^l\}$. Notice that $\epsilon_i^l = \dot{\ell}_l(\langle R_{X_i^l}, \theta_0 \rangle)$. It then follows that

$$\begin{aligned} S_\lambda(\theta_\lambda) =& \rho_0 \left\{ E_0 \left\{ \dot{\ell}_0(\langle R_{X^0}, \theta_\lambda \rangle) R_{X^0} \right\} \right\} \\ &+ \rho_1 \left\{ E_1 \left\{ \dot{\ell}_1(\langle R_{X^1}, \theta_\lambda \rangle) R_{X^1} \right\} \right\} - P_\lambda \theta_\lambda \\ =& S_0(\theta_\lambda) + S_1(\theta_\lambda) - P_\lambda \theta_\lambda. \end{aligned}$$

By Lemmas S.3 and S.4 in [164], $n^{-1} r_{1n}^2 h^{-2a-1} \asymp n^{-1} h^{2k-2a-1} = o\left((nh)^{-1}\right)$ and $S_\lambda(\theta_\lambda) = 0$, we have

$$\begin{aligned} &\left\| S_{n,\lambda}(\theta_\lambda) \right\|^2 = \left\| S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda) \right\|^2 \\ &\leq 2 \left\| S_{n,\lambda}^0(\theta_\lambda) - S_\lambda^0(\theta_\lambda) \right\|^2 + 2 \left\| S_{n,\lambda}^1(\theta_\lambda) - S_\lambda^1(\theta_\lambda) \right\|^2 \end{aligned}$$

for $l = 0, 1$, then

$$\begin{aligned} &E_l \left\{ \left\| S_{n,\lambda}^l(\theta_\lambda) - S_\lambda^l(\theta_\lambda) \right\|^2 \right\} \\ \leq& E_l \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n_l} O_i^l \right\|^2 \right\} + (\rho_l - n_l/n)^2 E_l \left\{ \left| \dot{\ell}_l\left(\langle R_{X^l}, \theta_\lambda \rangle\right) \right|^2 \left\| R_{X^l} \right\|^2 \right\} \\ =& n^{-1}(n_l/n) E_l \left\{ \left\| O_i^l \right\|^2 \right\} + (\rho_l - n_l/n)^2 E_l \left\{ \left| \dot{\ell}_l\left(\langle R_{X^l}, \theta_\lambda \rangle\right) \right|^2 \left\| R_{X^l} \right\|^2 \right\} \\ \leq& \left(n^{-1}(n_l/n) + (\rho_l - n_l/n)^2 \right) E_l \left\{ \left| \dot{\ell}_l\left(\langle R_{X^l}, \theta_\lambda \rangle\right) \right|^2 \left\| R_{X^l} \right\|^2 \right\} \\ \leq& 2\left(n^{-1}(n_l/n) + (\rho_l - n_l/n)^2 \right) E_l \left\{ \left\| R_{X^l} \right\|^2 \right\} \\ =& O\left((nh)^{-1} \right), \end{aligned}$$

the last equation follows from lemma S.4 [164], and $\rho_l - n_l/n = O_P(n^{-1/2})$. Thus $\left\| S_{n,\lambda}(\theta_\lambda) \right\| = O_P\left((nh)^{-1/2}\right)$. Let C_4 be large constant such that, with probability approaching one, $\left\| S_{n,\lambda}(\theta_\lambda) \right\| \leq C_4 (nh)^{-1/2}$. Let $r_{2n} = 2C_4(nh)^{-1/2}$, and restrict θ to be an element in $\mathbb{B}(r_{2n}) \equiv \{\theta \in \mathcal{H} : \|\theta\| \leq r_{2n}\}$.

Define $\mathcal{E}_n = \cap_{i=1}^n A_i^l$ for $X = \left(X_1^0, \ldots, X_{n_0}^0, X_1^1, \ldots, X_{n_1}^1\right)^\top$, where $A_i^l = \left\{ \left\| X_i^l \right\|_{L^2} \leq C \log n \right\}$, where C is a positive constant such that $C \log n > 1$. Since X^l is sub-Gaussian for $l = 0, 1$, we can choose C to be large enough so that $P(\mathcal{E}_n)$ approaches one as $n \to \infty$.

By the boundedness of logistic objective function, we may choose the above C to be large so that $\cap_i A_i^l$ has probability approaching one, and $P\left(A_i^{lc}\right) = O\left(n^{-1}\right)$.

Let $d_n = \kappa h^{-(2a+1)/2}$ and $p_n = d_n^{-2}\lambda^{-1}$. To handle I_2, it follows by Taylor's expansion that

$$S_{n,\lambda}\left(\theta_\lambda + \theta\right) - S_{n,\lambda}\left(\theta_\lambda\right) - DS_{n,\lambda}\left(\theta_\lambda\right)\theta = \int_0^1 \int_0^1 sD^2 S_{n,\lambda}\left(\theta_\lambda + s's\theta\right)\theta\theta \mathrm{d}s\mathrm{d}s'.$$

Assume $\mathcal{E} = \cap_i A_i^l$ in the rest of the proof. For any $\theta \in \mathcal{H}\backslash\{0\}$, let $\bar{\theta} = \theta/\left(d_n\|\theta\|\right)$, then $\bar{\theta} \in \mathcal{F}_{p_n}$, where $d_n = \kappa h^{-(2a+1)/2}$ and $p_n = d_n^{-2}\lambda^{-1} = \kappa^{-2}h^{-(2k-2a-1)}$. Let

$$\psi_n\left(X_i^l; \bar{\theta}\right) = \frac{\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right| \cdot \left\|R_{X_i^l}\right\| \cdot \left\langle R_{X_i^l}, \bar{\theta}\right\rangle}{\sqrt{2C_R}(C\log n)^2 h^{-(2a+1)/2}} I_{A_i^l}, i = 1, 2, \ldots, n.$$

Using $\|R_{X_i^l}\|^2 \leq C_R(1 + (C\log n)^2 h^{-(2a+1)})$, we get that ψ_n satisfies Lipschitz continuity (5.9). It follows by Lemma 5.2 that, with probability approaching one, for any $\theta \in \mathcal{H}\backslash\{0\}$,

$$\frac{1}{\sqrt{n}}\left\|\sum_{l=0}^{l}\sum_{i=1}^{n_l}\left[\psi_n\left(X_i^l;\bar{\theta}\right)R_{X_i^l} - E_l\left\{\psi_n\left(X_i^l;\bar{\theta}\right)R_{X_i^l}\right\}\right]\right\| \leq C'' p_n^{1/(4m)}\left(h^{-1}\log\log n\right)^{1/2}.$$

for some large $C'' > 0$, leading to

$$\begin{aligned}
&\left\|\sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right| \cdot \left\|R_{X_i^l}\right\| \cdot \left\langle R_{X_i^l}, \theta\right\rangle I_{A_i^l} R_{X_i^l}\right.\right. \\
&\quad \left.\left. - E_l\left\{\sup_{a\in\mathbb{R}}\left|\ell_a'''\left(a\right)\right| \cdot \left\|R_{X_i^l}\right\| \cdot \left\langle R_{X_i^l}, \theta\right\rangle I_{A_i^l} R_{X_i^l}\right\}\right]\right\| \\
\leq &\sqrt{2C_R}C''C^3\kappa^\gamma\sqrt{n}h^{-\left(2a+\frac{3}{2}\right)-\frac{2k-2a-1}{4m}}(\log n)^2(\log\log n)^{1/2}\|\theta\|.
\end{aligned}$$

The above inequality also holds for $\theta = 0$. Meanwhile, by the logistic objective function,

$$\begin{aligned}
&E_l\left\{\sup_{a\in\mathbb{R}}\left|\ell'''(a)\right| \cdot \|R_{X^l}\| \cdot \left|\langle R_{X^l}, \theta\rangle\right|^2\right\} \\
\leq &E_l\left\{\|R_{X^l}\| \cdot \left|\langle R_{X^l}, \theta\rangle\right|^2\right\} \\
\leq &C_R^{1/2}h^{-1/2}C_2\left(8\tilde{M}\right)^{1/2}\|\theta\|^2.
\end{aligned}$$

Therefore, we assume on $\cap_{i=1}^{n} A_i^l$, for any $\theta \in \mathbb{B}(r_{2n})$, we have, for some $s, s' \in [0,1]$,

$$
\begin{aligned}
&\left\| D^2 S_{n,\lambda}\left(\theta_\lambda + s' s\theta\right)\theta\theta \right\| \\
=&n^{-1}\left\| \sum_{l=0}^{1}\sum_{i=1}^{n_l} \ell'''\left(\left\langle R_{X_i^l}, \tilde{\theta}\right\rangle\right)\left|\left\langle R_{X_i^l}, \theta\right\rangle\right|^2 R_{X_i^l} \right\| \\
\leq&n^{-1}\sum_{l=0}^{1}\sum_{i=1}^{n_l}\sup_{a\in\mathbb{R}}\left|\ell'''(a)\right|\cdot\left\|R_{X_i^l}\right\|\cdot\left|\left\langle R_{X_i^l}, \theta\right\rangle\right|^2 \\
=&n^{-1}\left\langle \sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\sup_{a\in\mathbb{R}}\left|\ell'''(a)\right|\cdot\left\|R_{X_i^l}\right\|\cdot\left\langle R_{X_i^l}, \theta\right\rangle I_{A_i^l} R_{X_i^l}\right.\right. \\
&\left.\left. -E_l\left\{\sup_{a\in\mathbb{R}}\left|\ell'''(a)\right|\cdot\left\|R_{X_i^l}\right\|\cdot\left\langle R_{X_i^l}, \theta\right\rangle I_{A_i^l} R_{X_i^l}\right\}\right], \theta \right\rangle \\
&+\sum_{l=0}^{1} n_l/n E_l\left\{\sup_{a\in\mathbb{R}}\left|\ell'''(a)\right|\cdot\left\|R_{X_i^l}\right\|\cdot\left|\left\langle R_{X_i^l}, \theta\right\rangle\right|^2 I_{A_i^l}\right\} \\
\leq&C'''\left[n^{-1}h^{-2(a+1)-\frac{2k-2a-1}{4m}}(\log n)^2(\log\log n)^{1/2} + n^{-1/2}h^{-1}\right]\times\|\theta\| \\
\leq&\|\theta\|/18,
\end{aligned}
$$

the last equality follows by the assumption (5.10).

We address I_1 using similar arguments in the analysis of I_2. Define

$$
\psi\left(X_i^l;\theta\right) = \ddot{\ell}\left(\left\langle R_{X_i^l}, \theta_\lambda\right\rangle\right)\left\langle R_{X_i^l}, \theta\right\rangle I_{A_i^l}.
$$

Then for any $\theta_j = (\alpha_j, \beta_j) \in \mathcal{H}, j = 1, 2$,

$$
\begin{aligned}
&\left|\psi\left(X_i^l;\theta_1\right) - \psi\left(X_i^l;\theta_2\right)\right| \\
=&\left|\ddot{\ell}\left(\left\langle R_{X_i^l}, \theta_\lambda\right\rangle\right)\right|\left|\left\langle R_{X_i^l}, \theta_1 - \theta_2\right\rangle\right| I_{A_i^l} \\
\leq&I_{A_i^l}\left(|\alpha_1 - \alpha_2| + \|X^l\|_{L^2}\|\beta_1 - \beta_2\|_{L^2}\right) \\
\leq&(C\log n)\|\theta_1 - \theta_2\|_2 .
\end{aligned}
$$

Let $\psi_n\left(X_i^l;\theta\right) = (C\log n)^{-1}\psi\left(X_i^l;\theta\right)$, then ψ_n satisfies equation (5.9). For any $\theta \in \mathcal{H}\backslash\{0\}$, define $\bar{\theta} = (\bar{\alpha}, \bar{\beta}) \equiv \theta/\left(d_n\|\theta\|\right)$. It then follows by Lemma 5.1 that

$$
\|\bar{\theta}\|_2 \leq d_n\|\bar{\theta}\| = 1,
$$

which implies that $|\bar{\alpha}| + \|\bar{\beta}\|_{L^2} \leq 1$. Meanwhile, $\lambda J(\bar{\beta}, \bar{\beta}) \leq \|\bar{\theta}\|^2 = d_n^{-2}$, implying $J(\bar{\beta}, \bar{\beta}) \leq d_n^{-2}\lambda^{-1} = p_n$. Therefore, $\bar{\beta} \in \mathcal{F}_{p_n}$. By Lemma 5.2, for some constant

$C' > 0$ and with probability approaching one, for any $\theta \in \mathbb{B}(r_{2n})$,

$$\begin{aligned}
&\left\| \sum_{i=1}^{n_0} \left[\psi_n\left(X_i^0; \bar{\theta}\right) R_{X_i^0} - E_0\left\{ \psi_n(X^0; \bar{\theta}) R_X^0 \right\} \right] \right. \\
&\left. + \sum_{i=1}^{n_1} \left[\psi_n\left(X_i^1; \bar{\theta}\right) R_{X_i^1} - E_1\left\{ \psi_n(X^1; \bar{\theta}) R_X^1 \right\} \right] \right\| \\
&\leq C' \left(n^{1/2} p_n^{1/(4m)} + 1 \right) \left(h^{-1} \log\log n \right)^{1/2},
\end{aligned}$$

which implies

$$\begin{aligned}
&\left\| \sum_{i=1}^{n_0} \left[\psi_n\left(X_i^0; \theta\right) R_{X_i^0} - E_0\left\{ \psi_n(X^0; \theta) R_X^0 \right\} \right] \right. \\
&\left. + \sum_{i=1}^{n_1} \left[\psi_n\left(X_i^1; \theta\right) R_{X_i^1} - E_1\left\{ \psi_n(X^1; \theta) R_X^1 \right\} \right] \right\| \\
\leq & C' \kappa h^{-(a+1)} \left(n^{1/2} p_n^{1/(4m)} + 1 \right) (C \log n) (\log\log n)^{1/2} \|\theta\|.
\end{aligned}$$

On the other side, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left\| E_l \left\{ \ddot{\ell}\left(\left\langle R_{X_i^l}, \theta_\lambda \right\rangle \right) \left\langle R_{X_i^l}, \theta \right\rangle R_{X_i^l} I_{A_i^{lc}} \right\} \right\| \\
= & \sup_{\|\theta'\|=1} \left| E_l \left\{ \ddot{\ell}\left(\left\langle R_{X_i^l}, \theta_\lambda \right\rangle \right) \left\langle R_{X_i^l}, \theta \right\rangle \left\langle R_{X_i^l}, \theta' \right\rangle I_{A_i^{lc}} \right\} \right| \\
\leq & \sup_{\|\theta'\|=1} E_l \left\{ \left| \left\langle R_{X_i^l}, \theta \right\rangle \right| \cdot \left| \left\langle R_{X_i^l}, \theta' \right\rangle \right| I_{A_i^{lc}} \right\} \\
\leq & E_l \left\{ \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^4 \right\}^{1/4} \sup_{\|\theta'\|=1} E_l \left\{ \left| \left\langle R_{X_i^l}, \theta' \right\rangle \right|^4 \right\}^{1/4} P\left(A_i^{lc}\right)^{1/2} \\
\leq & C_2 \sqrt{8\tilde{M}} \|\theta\| P\left(A_i^{lc}\right)^{1/2} = o(1)\|\theta\|.
\end{aligned}$$

Thus, with probability approaching one, for any $\theta \in \mathbb{B}(r_{2n})$, we get

$$
\begin{aligned}
&\|DS_{n,\lambda}(\theta_\lambda)\theta - DS_\lambda(\theta_\lambda)\theta\| \\
\leq & n^{-1}\sum_{l=0}^{1}\left\|\sum_{i=1}^{n_l}\left[\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right.\right. \\
& \left.\left. -E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right\}\right]\right\| \\
& + \sum_{l=0}^{1}(n_l/n-\rho_l)\left\|E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right\}\right\| \\
& + \sum_{l=0}^{1}\rho_l\left\|E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^{lc}}\right\}\right\| \\
= & n^{-1}\sum_{l=0}^{1}\left\|\sum_{i=1}^{n_l}\left[\psi\left(X_i^l;\theta\right)R_{X_i^l} - E_l\left\{\psi\left(X_i^l;\theta\right)R_{X_i^l}\right\}\right]\right\| \\
& + \sum_{l=0}^{1}(n_l/n-\rho_l)\left\|E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right\}\right\| \\
& + \sum_{l=0}^{1}\rho_l\left\|E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_\lambda\right\rangle\right)\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^{lc}}\right\}\right\| \\
= & O\left(n^{-1/2}h^{-(a+1)-\frac{2k-2a-1}{4m}}(\log n)(\log\log n)^{1/2}\right)\|\theta\| + o(1)\|\theta\| + o(1)\|\theta\| \\
= & o(1)\|\theta\|,
\end{aligned}
$$

where the last inequality follows by the assumption (5.10). Hence, with probability approaching one, for any $\theta \in \mathbb{B}(r_{2n})$,

$$\|I_1\| \leq \left\|DS_\lambda(\theta_\lambda)^{-1}\right\|\|DS_{n,\lambda}(\theta_\lambda)\theta - DS_\lambda(\theta_\lambda)\theta\| \leq r_{2n}/18.$$

By the above analysis of the terms I_1, I_2, I_3, we have that, for any $\theta \in \mathbb{B}(r_{2n})$, with probability approaching one,

$$
\begin{aligned}
&\|T_{2h}(\theta)\| \\
\leq & \|I_1\| + \|I_2\| + \|I_3\| \\
\leq & (3/2)\left(r_{2n}/18 + r_{2n}/18 + r_{2n}/2\right) = 11r_{2n}/12
\end{aligned}
$$

namely, $T_{2n}(\mathbb{B}(r_{2n})) \subset \mathbb{B}(r_{2n})$. Next, we show that T_{2h} is a contraction mapping.

For any $\theta_1, \theta_2 \in \mathbb{B}(r_{2n})$, we have, by Taylor's expansion, that

$$\begin{aligned}
&T_{2h}(\theta_1) - T_{2h}(\theta_2) \\
=&\theta_1 - \theta_2 - DS_\lambda(\theta_\lambda)^{-1}\left[S_{n,\lambda}(\theta_\lambda + \theta_1) - S_{n,\lambda}(\theta_\lambda + \theta_2)\right] \\
=&-DS_\lambda(\theta_\lambda)^{-1}\int_0^1\int_0^1 D^2 S_{n,\lambda}(\theta_\lambda + s'(\theta_2 + s(\theta_1 - \theta_2)))\cdot \\
&(\theta_2 + s(\theta_1 - \theta_2))(\theta_1 - \theta_2)\,\mathrm{d}s\mathrm{d}s' \\
&-DS_\lambda(\theta_\lambda)^{-1}\left[DS_{n,\lambda}(\theta_\lambda) - DS_\lambda(\theta_\lambda)\right](\theta_1 - \theta_2) \\
\equiv&-I_4 - I_5.
\end{aligned}$$

Using exactly the same arguments as the analysis of the terms I_1 and I_2, it can be shown that, with probability approaching one, for any $\theta_1, \theta_2 \in \mathbb{B}(r_{2n})$,

$$\begin{aligned}
&\|I_4\| \\
=&O\left(n^{-1}h^{-2(a+1)-\frac{2k-2a-1}{4m}}(\log n)^2(\log\log n)^{1/2} + n^{-1/2}h^{-1}\right)\|\theta_1 - \theta_2\| \\
\leq&\|\theta_1 - \theta_2\|/3, \\
&\|I_5\| \\
=&O\left(n^{-1/2}h^{-(a+1)-\frac{2k-2a-1}{4m}}(\log n)(\log\log n)^{1/2}\right)\|\theta_1 - \theta_2\| + o(1)\|\theta_1 - \theta_2\| \\
\leq&\|\theta_1 - \theta_2\|/3.
\end{aligned}$$

implying that $\|T_{2h}(\theta_1) - T_{2h}(\theta_2)\| \leq 2\|\theta_1 - \theta_2\|/3$. Therefore, T_{2h} is a contraction mapping from $\mathbb{B}(r_{2n})$ to itself. By contraction mapping theorem, there exists uniquely an element $\theta' \in \mathbb{B}(r_{2n})$ such that $T_{2h}(\theta') = \theta'$, implying $S_{n,\lambda}(\theta_\lambda + \theta') = 0$. Let $\widehat{\theta}_{n,\lambda} = \theta_\lambda + \theta'$, then $S_{n,\lambda}(\widehat{\theta}_{n,\lambda}) = 0$, that is, $\widehat{\theta}_{n,\lambda}$ is the smoothing spline estimator. Furthermore, with probability approaching one,

$$\begin{aligned}
&\left\|\widehat{\theta}_{n,\lambda} - \theta_0\right\| \\
\leq&\left\|\widehat{\theta}_{n,\lambda} - \theta_\lambda\right\| + \|\theta_\lambda - \theta_0\| \leq r_{2n} + r_{1n} \\
=&O\left((nh)^{-1/2} + h^k\right) = O(r_n).
\end{aligned}$$

■

5.9.2 Proof of Theorem 5.4

Proof.

The proof of Theorem 5.4 replies on concentration inequality established in Lemma 5.2. By the rate concluded in Condition 5.5, we can find a large constant $M > 0$ such that, with probability approaching one, $\|\widehat{\theta}_{n,\lambda} - \theta_0\| \leq Mr_n$. Denote $\theta = \widehat{\theta}_{n,\lambda} - \theta_0$. Assume $\|\theta\| \leq Mr_n$ since its complement is negligible in terms of probability. Let $d_n = \kappa M h^{-(2a+1)/2} r_n$, $\tilde{\theta} = d_n^{-1}\theta$, and $p_n = \kappa^{-2} h^{1-2m}$, where κ is the constant given in Lemma 5.1. Clearly $p_n \geq 1$ since h approaches zero and $1 - 2m < 0$. In fact, $\|\theta\| \leq Mr_n$ implies $\tilde{\theta} \in \mathcal{F}_{p_n}$. To see this, write $\tilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$. By Lemma 5.1, $\|\tilde{\theta}\|_2 = d_n^{-1}\|\theta\|_2 \leq d_n^{-1}\kappa h^{-(2a+1)/2}\|\theta\| \leq d_n^{-1}\kappa h^{-(2a+1)/2} Mr_n = 1$. Therefore, $|\widetilde{\alpha}| \leq 1$ and $\|\widetilde{\beta}\|_{L^2} \leq 1$. Thus, $\tilde{\theta} \in \mathcal{F}_{p_n}$ follows from

$$\begin{aligned}
&J(\widetilde{\beta}, \widetilde{\beta}) \\
=&d_n^{-2}\lambda^{-1}(\lambda J(\beta, \beta)) \\
\leq&d_n^{-2}\lambda^{-1}\|\theta\|^2 \leq d_n^{-2}\lambda^{-1}\left(Mr_n\right)^2 = \kappa^{-2}h^{1-2m} = p_n.
\end{aligned}$$

For $i = 1, \ldots, n$, define $A_i^l = \left\{\left\|X_i^l\right\|_{L^2} \leq C\log n\right\}$. By Condition 5.4, we can fix $C > 1$ as large enough such that

$$n^{1/2}h^{-1/2}P\left(I_{A_i^{lc}}\right)^{1/4} = o\left(p_n^{1/(4m)}\left(h^{-1}\log\log n\right)^{1/2}\right),$$

and $P\left(\mathcal{E}_n\right)$ approaches one, as $n \to \infty$, where $\mathcal{E}_n = \cap_{i=1}^n A_i^l$. Let $D_n = (C\log n)^{-1} d_n^{-1}$. Define $\psi(X_i^l; \theta) = [\dot{\ell}_l(\langle R_{X_i^l}, \theta + \theta_0\rangle) - \dot{\ell}_l(\langle R_{X_i^l}, \theta_0\rangle)]$, and $\psi_n(X_i^l; \tilde{\theta}) = D_n\psi(X_i^l; d_n\tilde{\theta})I_{A_i^l}$. Then for any $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathcal{F}_{p_n}$,

$$\begin{aligned}
&\left|\psi_n\left(X_i^l; \tilde{\theta}_1\right) - \psi_n\left(X_i^l; \tilde{\theta}_2\right)\right| \\
=&D_n\left|\psi\left(X_i^l; d_n\tilde{\theta}_1\right) - \psi\left(X_i^l; d_n\tilde{\theta}_2\right)\right| I_{A_i^l} \\
=&D_n\left|\dot{\ell}_l\left(\left\langle R_{X_i^l}, \theta_0 + d_n\tilde{\theta}_1\right\rangle\right) - \dot{\ell}_l\left(\left\langle R_{X_i^l}, \theta_0 + d_n\tilde{\theta}_2\right\rangle\right)\right| I_{A_i^l} \\
\leq&D_n \sup_{a\in\mathbb{R}}\left|\ddot{\ell}\left(a\right)\right| \cdot d_n \cdot \left|\left\langle R_{X_i^l}, \tilde{\theta}_1 - \tilde{\theta}_2\right\rangle\right| I_{A_i^l} \\
\leq&D_n(C\log n)d_n\left\|\tilde{\theta}_1 - \tilde{\theta}_2\right\|_2 \\
=&\left\|\tilde{\theta}_1 - \tilde{\theta}_2\right\|_2.
\end{aligned}$$

Since $\|\theta\| \leq Mr_n$ implies $\tilde{\theta} \in \mathcal{F}_{p_n}$, it follows by Lemma 5.2 that there exists $C' > 0$

such that for large n, with probability approaching one

$$\begin{aligned}
&\left\| 1/\sqrt{n} \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi_n \left(X_i^l; d_n \tilde{\theta} \right) R_{X_i^l} - E_l \left\{ \psi_n \left(X^l; d_n \tilde{\theta} \right) R_{X^l} \right\} \right] \right\| \\
\leq & C' \left(p_n^{1/(4m)} \|\tilde{\theta}\|_2^{\gamma} + n^{-1/2} \right) \left(h^{-1} \log\log n \right)^{1/2} \\
\leq & C' \left(p_n^{1/(4m)} + n^{-1/2} \right) \left(h^{-1} \log\log n \right)^{1/2},
\end{aligned}$$

where $\gamma = 1 - 1/(2m)$. We denote the term $(1/\sqrt{n}) \sum_{l=0}^{1} \sum_{i=1}^{n_l} \Big[\psi_n(X_i^l; d_n \tilde{\theta}) R_{X_i^l} - E_l \{ \psi_n(X^l; d_n \tilde{\theta}) R_{X^l} \}]$ by $H_n(\tilde{\theta})$.

On the other hand, using Cauchy-Schwarz inequality, we have for $l = 0, 1$,

$$\begin{aligned}
&\left\| E_l \left\{ \psi \left(X_i^l; d_n \tilde{\theta} \right) R_{X_i^l} I_{A_i^{lc}} \right\} \right\| \\
\leq & E_l \left\{ \left| \psi \left(X_i^l; d_n \tilde{\theta} \right) \right| \cdot \left\| R_{X_i^l} \right\| I_{A_i^{lc}} \right\} \\
\leq & E_l \left\{ \sup_{a \in \mathbb{R}} \left| \ddot{\ell}(a) \right| d_n \left| \left\langle R_{X_i^l}, \tilde{\theta} \right\rangle \right| \cdot \left\| R_{X_i^l} \right\| I_{A_i^{lc}} \right\} \\
\leq & d_n E_l \left\{ \sup_{a \in \mathbb{R}} \left| \ddot{\ell}(a) \right| \left(1 + \left\| X_i^l \right\|_{L^2} \right) \left\| R_{X_i^l} \right\| I_{A_i^{lc}} \right\} \\
\leq & d_n E_l \left\{ \left(\sup_{a \in \mathbb{R}} \left| \ddot{\ell}(a) \right| \right)^4 \left(1 + \left\| X_i^l \right\|_{L^2} \right)^4 \right\}^{1/4} P \left(I_{A_i^{lc}} \right)^{1/4} E \left\{ \left\| R_{X_i^l} \right\|^2 \right\}^{1/2} \\
\leq & d_n E_l \left\{ \left(1 + \left\| X_i^l \right\|_{L^2} \right)^4 \right\}^{1/4} P \left(I_{A_i^{lc}} \right)^{1/4} C_R^{1/2} h^{-1/2}.
\end{aligned}$$

Consequently, by choice of C and direct examinations, we have

$$\begin{aligned}
& D_n n^{1/2} \left\| E_l \left\{ \psi \left(X_i^l; d_n \tilde{\theta} \right) R_{X_i^l} I_{A_i^{lc}} \right\} \right\| \\
\leq & C_R^{1/2} E_l \left\{ \left(1 + \left\| X_i^l \right\|_{L^2} \right)^4 \right\}^{1/4} n^{1/2} h^{-1/2} P \left(I_{A_i^{lc}} \right)^{1/4} \\
\leq & C' p_n^{1/(4m)} \left(h^{-1} \log\log n \right)^{1/2}.
\end{aligned}$$

Since on $\mathcal{E}_n, I_{A_i^l} = 1$ for $i = 1, \ldots, n$, then with probability approaching one,

$$\begin{aligned}
& n^{-1/2} D_n \left\| \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi\left(X_i^l; \theta\right) R_{X_i^l} - E_l \left\{ \psi(X^l; \theta) R_{X^l} \right\} \right] \right\| \\
=& n^{-1/2} \left\| \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi_n\left(X_i^l; \tilde{\theta}\right) R_{X_i^l} - E_l \left\{ \psi_n\left(X_i^l; \tilde{\theta}\right) R_{X_i^l} \right\} \right] \right. \\
& \left. - D_n \sum_{l=0}^{1} n_l \, E_l \left\{ \psi\left(X_i^l; d_n \tilde{\theta}\right) R_{X_i^l} I_{A_i^{lc}} \right\} \right\| \\
\leq & \left\| H_n(\tilde{\theta}) \right\| + D_n n^{-1/2} \sum_{l=0}^{1} n_l \left\| E_l \left\{ \psi\left(X_i^l; d_n \tilde{\theta}\right) R_{X_i^l} I_{A_i^{lc}} \right\} \right\| \\
\leq & 2C' p_n^{1/(4m)} \left(h^{-1} \log\log n\right)^{1/2}.
\end{aligned}$$

Thus, we can show that by direct calculations, with probability approaching one,

$$\begin{aligned}
& \left\| S_{n,\lambda}\left(\theta + \theta_0\right) - S_{n,\lambda}\left(\theta_0\right) - \sum_{l=0}^{1} \rho_l E_l \left\{ S_{n,\lambda}^l\left(\theta + \theta_0\right) - S_{n,\lambda}^l\left(\theta_0\right) \right\} \right\| \\
= & \left\| S_n\left(\theta + \theta_0\right) - S_n\left(\theta_0\right) - \sum_{l=0}^{1} \rho_l E_l \left\{ S_n^l\left(\theta + \theta_0\right) - S_n^l\left(\theta_0\right) \right\} \right\| \\
= & n^{-1} \left\| \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi\left(X_i^l; \theta\right) R_{X_i^l} - E_l \left\{ \psi\left(X_i^l; \theta\right) R_{X_i^l} \right\} \right] \right\| \\
& + \sum_{l=0}^{1} (\rho_l - n_l/n) \left\| E_l \left\{ \psi\left(X_i^l; \theta\right) R_{X_i^l} \right\} \right\| \\
\leq & 2C' n^{-1/2} p_n^{1/(4m)} D_n^{-1} \left(h^{-1} \log\log n\right)^{1/2} \\
= & 2C' \kappa^{\gamma} C^2 M n^{-1/2} h^{-\frac{4ma+6m-1}{4m}} r_n (\log n)^2 (\log\log n)^{1/2}.
\end{aligned}$$

Note that $S_{n,\lambda}\left(\theta + \theta_0\right) = 0$. Define $DS_\lambda(\theta) = \sum_{l=0}^{1} \rho_l E_l S_{n,\lambda}^l(\theta)$ for any θ. For any $\theta_1, \theta_2 \in \mathcal{H}, \langle DS_\lambda(\theta)\theta_1, \theta_2 \rangle = -\langle \theta_1, \theta_2 \rangle$, we have $DS_\lambda(\theta_0) = -id$, where id denotes the identity operator on $\mathcal{H}$. By Taylor's expansion,

$$\begin{aligned}
& \sum_{l=0}^{1} \rho_l E_l \left\{ S_{n,\lambda}^l\left(\theta + \theta_0\right) - S_{n,\lambda}^l\left(\theta_0\right) \right\} \\
= & DS_\lambda\left(\theta_0\right)\theta + \int_0^1 \int_0^1 s \sum_{l=0}^{1} \rho_l E_l \left\{ D^2 S_{n,\lambda}^l\left(\theta_0 + s s' \theta\right) \theta\theta \right\} \mathrm{d}s\mathrm{d}s' \\
= & -\theta + \int_0^1 \int_0^1 s \sum_{l=0}^{1} \rho_l E_l \left\{ \ell'''\left(\langle R_{X^l}, \theta_0 + s' s \theta \rangle\right) \left|\langle R_{X^l}, \theta \rangle\right|^2 R_{X^l} \right\} \mathrm{d}s\mathrm{d}s'.
\end{aligned}$$

Since for $l = 0, 1$,

$$\begin{aligned}
&\left\| \int_0^1 \int_0^1 s \sum_{l=0}^{1} \rho_l E_l \left\{ \ell'''\left(Y; \langle R_{X^l}, \theta_0 + s' s\theta \rangle\right) |\langle R_{X^l}, \theta \rangle|^2 R_{X^l} \right\} \mathrm{d}s\mathrm{d}s' \right\| \\
\leq & \int_0^1 \int_0^1 s \sum_{l=0}^{1} \rho_l E_l \left\{ \left|\ell'''\left(\langle R_{X^l}, \theta_0 + s' s\theta \rangle\right)\right| \cdot |\langle R_{X^l}, \theta \rangle|^2 \|R_{X^l}\| \right\} \mathrm{d}s\mathrm{d}s' \\
\leq & (1/2) C_\ell \sum_{l=0}^{1} \rho_l E_l \left\{ |\langle R_{X^l}, \theta \rangle|^2 \|R_{X^l}\| \right\} \\
\leq & (1/2) C_\ell \sum_{l=0}^{1} \rho_l \left[E_l \left\{ |\langle R_{X^l}, \theta \rangle|^4 \right\}^{1/2} E_l \left\{ \|R_{X^l}\|^2 \right\}^{1/2} \right] \\
\leq & (1/2) C_\ell \left(8\tilde{M} C_2^2 C_R \right)^{1/2} h^{-1/2} \|\theta\|^2 \\
\leq & (1/2) C_\ell C_2 \left(8\tilde{M} C_R \right)^{1/2} M^2 h^{-1/2} r_n^2.
\end{aligned}$$

Thus, with probability approaching one, we get

$$\begin{aligned}
&\left\| \widehat{\theta}_{n,\lambda} - \theta_0 - S_{n,\lambda}\left(\theta_0\right) \right\| \\
\leq & 2C' \kappa^\gamma C^2 M n^{-1/2} h^{-\frac{4ma+6m-1}{4m}} r_n (\log n)(\log\log n)^{1/2} \\
& + (1/2) C_\ell C_2 \left(8\tilde{M} C_R \right)^{1/2} M^2 h^{-1/2} r_n^2.
\end{aligned}$$

■

Proof of Theorem 5.5.

The proof of theorem 5.5 follows directly from the Lindeberg Central limit theorem and the discussion of corollary 3.7 in [164]. ■

5.9.3 Likelihood ratio test under case control design

Under the case-control design, we conduct a penalized likelihood ratio test for the functional slope $\beta(t)$. Consider the following hypothesis:

$$H_0 : \beta = \beta_0 \quad \text{versus } H_1 : \beta \in \mathcal{H} - \{\beta_0\},$$

where $\beta_0 \in H^m(\mathbb{I})$. The penalized likelihood ratio test is implemented via the following steps. First, we define a test statistic PLRT concerning the full parameter space $\mathcal{H}$ as PLRT $= \ell_{n,\lambda}\left(\theta_0\right) - \ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right)$, where $\widehat{\theta}_{n,\lambda}$ is the maximizer of $\ell_{n,\lambda}(\theta)$ over $\mathcal{H}$.

Then, the test statistic for β is defined as

$$T_P \equiv \text{PLRT}_1 - \text{PLRT}_2 = \ell_{n,\lambda}\left(\widehat{\theta}_0\right) - \ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right).$$

where $\widehat{\theta}_0$ is the maximizer under the null hypothesis, i.e., the parameter space contains only the intercept α. Theorem 5.8 below derives the null limiting distribution of PLRT.

Theorem 5.8 (Likelihood ratio testing) *Assume $H_0 : \theta = \theta_0$ holds, Conditions 5.1 to 5.5 are satisfied for the hypothesized value θ_0. As $n \to \infty$, h satisfy the following rate conditions:*

$$n^{1/2}a_n = o(1), \quad nr_n^3 = o(1),$$
$$nh^{2k+1} = O(1), \quad (nh)^{-1} = o(1),$$
$$n^{1/2}h^{-(a+1/2+(2k-2a-1)/(4m))}r_n^2(\log n) \times (\log\log n)^{1/2} = o(1),$$
and $$n^{1/2}h^{-(2a+1+(2k-2a-1)/(4m))} \times r_n^3(\log n)^2 \times (\log\log n)^{1/2} = o(1).$$

Let $u_n = h^{-1}\sigma_1^4/\sigma_2^2, \sigma^2 = \sigma_1^2/\sigma_2^2$ and $\sigma_l^2 = h\sum_j (1+\lambda\rho_j)^{-l}$ for $l = 1, 2$. Then under H_0, as $n \to \infty$, we have

$$-(2u_n)^{-1/2}\left(2n\sigma^2 \cdot \text{PLRT} + u_n + n\sigma^2 \left\|W_\lambda\beta_0\right\|_1^2\right) \xrightarrow{d} N(0,1).$$

The PLRT relies on the Bahadur representation and the inner products defined in (5.3). It can be shown that $n\left\|W_\lambda\beta_0\right\|_1^2 = o(n\lambda) = o\left(u_n\right)$. Therefore, as $n \to \infty$, $-2n\sigma^2 \cdot \text{PLRT}$ is asymptotically $N\left(u_n, 2u_n\right)$. According to the parametric likelihood ratio testing, we have $-2n\sigma^2 \cdot \text{PLRT}_1 = O_P(1)$. Meanwhile, theorem 5.8 shows that $-2n\sigma^2 \cdot \text{PLRT}_2 \overset{a}{\sim} \chi^2_{u_n}$. Thus, the null limit distribution for testing the significance of the slope function β is also $\chi^2_{u_n}$, specifically, $-2n\sigma^2 T_P \overset{a}{\sim} \chi^2_{u_n}$.

Proof of Theorem 5.8.

The proof is finished in two parts. First, we show that the PLRT is asymptotically equivalent to a quadratic form. The proof of this part relies on the concentration

inequality established in Lemma 5.9. Second, using the functional Bahadur representation established in Theorem 5.4 and a central limit theorem for generalized quadratic forms by [79], we likewise show that the quadratic form derived in the first part has Gaussian limit.

Under $H_0 : \theta = \theta_0$, the hypothesized value θ_0 is deemed as the "true" parameter of the model. By consistency, for some $M > 0$ and with probability approaching one, $\|\widehat{\theta}_{n,\lambda} - \theta_0\| \leq Mr_n$. Thus, define $\theta = \widehat{\theta}_{n,\lambda} - \theta_0$, we may as well assume $\|\theta\| \leq Mr_n$, where , since the complement is trivial in terms of probability measure. For $l = 0, 1, i = 1, \ldots, n_l$, define similarly $A_i^l = \{\|X_i\|_{L^2} \leq C \log n\}$.

By the boundedness of logistic objective and Condition 5.4, it is not hard to choose a suitable $C > 1$ to be large so that, as $n \to \infty, P(A_i^{lc})^{1/2} = o(n^{-1/2} h^{-(a+1+\frac{2k-2a-1}{4m})} \times (\log n)(\log\log n)^{1/2})$. Let $\mathcal{E}_n = \cap_{l=0}^{1} \cap_{i=1}^{n_l} A_i^l$, clearly $\mathcal{E}_n$ has probability approaching one, and we may assume that $\mathcal{E}_n$ holds in the rest of the proof. By Taylor's expansion,

$$
\begin{aligned}
&\ell_{n,\lambda}(\theta_0) - \ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right) \\
=&\ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda} - \theta\right) - \ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right) \\
=&-S_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right)\theta + \int_0^1\int_0^1 sDS_{n,\lambda}\left(\widehat{\theta}_{n,\lambda} - s's\theta\right)\theta\theta \mathrm{d}s\mathrm{d}s' \\
=&\int_0^1\int_0^1 s\left[DS_{n,\lambda}\left(\widehat{\theta}_{n,\lambda} - s's\theta\right) - DS_{n,\lambda}(\theta_0)\right]\theta\theta \mathrm{d}s\mathrm{d}s' \\
&+ (1/2)\left[DS_{n,\lambda}(\theta_0) - \sum_{l=0}^{1}\rho_l E_l\left\{DS_{n,\lambda}^l(\theta_0)\right\}\right]\theta\theta \\
&+ (1/2)\sum_{l=0}^{1}\rho_l E_l\left\{DS_{n,\lambda}^l(\theta_0)\right\}\theta\theta \\
\equiv& I_1 + I_2 + I_3.
\end{aligned}
$$

Note that $\sum_{l=0}^{1}\rho_l E_l\left\{DS_{n,\lambda}^l(\theta_0)\right\} = -id$ on $\mathcal{H}$, then $\sum_{l=0}^{1}\rho_l E_l\{DS_{n,\lambda}^l(\theta_0)\}\theta\theta = -\|\theta\|^2$. The remainder of the proof proceeds by approximating the orders of I_1 and I_2, and finding the null limit distribution of the PLRT. To approximate I_1, denote $\theta' = \widehat{\theta}_{n,\lambda} - s's\theta - \theta_0$, which by definition is equal to $(1 - s's)\theta$, where $0 \leq s', s \leq 1$.

Direct calculations lead to

$$\begin{aligned}
&\left[DS_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}-s's\theta\right)-DS_{n,\lambda}\left(\theta_0\right)\right]\theta\theta\\
=&\left[DS_{n,\lambda}\left(\theta'+\theta_0\right)-DS_{n,\lambda}\left(\theta_0\right)\right]\theta\theta\\
=&\frac{1}{n}\sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\ddot{\ell}\left(\left\langle R_{X_i^l},\theta'+\theta_0\right\rangle\right)-\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_0\right\rangle\right)\right]\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2\\
\leq&\frac{1}{n}\sum_{l=0}^{1}\sum_{i=1}^{n_l}\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right|\cdot\left|\left\langle R_{X_i^l},\theta'\right\rangle\right|\cdot\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2\\
\leq&\frac{Mr_n}{n}\sum_{l=0}^{1}\sum_{i=1}^{n_l}\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right|\cdot\left\|R_{X_i^l}\right\|\cdot\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2\\
\leq&\frac{Mr_n}{n}\left\langle\sum_{l=0}^{1}\sum_{i=1}^{n_l}\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right|\cdot\left\|R_{X_i^l}\right\|\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l},\theta\right\rangle
\end{aligned}$$

Let $d_n=\kappa Mh^{-(2a+1)/2}r_n$ and $\tilde{\theta}=d_n^{-1}\theta$, where κ is the constant given in Lemma 5.1. Note that $\|\theta\|\leq Mr_n$ implies $\tilde{\theta}\in\mathcal{F}_{p_n}$, where $p_n=\kappa^{-2}h^{2a+1-2k}$. Let

$$\psi_n\left(X_i;\tilde{\theta}\right)=\frac{\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right|\cdot\left\|R_{X_i}\right\|\left\langle R_{X_i},\tilde{\theta}\right\rangle}{\sqrt{2C_R}(C\log n)^2h^{-(2a+1)/2}}I_{A_i}.$$

where $A_1,\ldots,A_{n_0}$ denote the control sets $A_1^0,\ldots,A_{n_0}^0$, and $A_{n_0+1},\ldots,A_n$ denote the case sets $A_1^1,\ldots,A_{n_1}^1$, by Lemma S.4 in [164], for any $\tilde{\theta}_1,\tilde{\theta}_2\in\mathcal{F}_{p_n}$,

$$\begin{aligned}
&\left|\psi_n\left(X_i;\tilde{\theta}_1\right)-\psi_n\left(X_i;\tilde{\theta}_2\right)\right|\\
=&\frac{\sup_{a\in\mathbb{R}}\left|\ell'''\left(a\right)\right|\cdot\left\|R_{X_i}\right\|\cdot\left|\left\langle R_{X_i},\tilde{\theta}_1-\tilde{\theta}_2\right\rangle\right|}{\sqrt{2C_R}(C\log n)^2h^{-(2a+1)/2}}I_{A_i}\\
\leq&\left\|\tilde{\theta}_1-\tilde{\theta}_2\right\|_2.
\end{aligned}$$

Then by Lemma 5.2, with probability approaching one, we know that $\|H_n(\tilde{\theta})\|\leq C'p_n^{1/(4m)}\left(h^{-1}\log\log n\right)^{1/2}$, for some large $C'>0$, where

$$H_n(\tilde{\theta})=n^{-1/2}\sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\psi_n\left(X_i^l;\tilde{\theta}\right)R_{X_i^l}-E_l\left\{\psi_n(X^l;\tilde{\theta})R_{X^l}\right\}\right].$$

Thus, on $\mathcal{E}_n$,

$$\begin{aligned}
&\left\|\sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\left\|R_{X_i^l}\right\|\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}\right.\right.\\
&\left.\left.-E_l\left\{\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\left\|R_{X_i^l}\right\|\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right\}\right]\right\|\\
\leq& n^{1/2}d_n\sqrt{2C_R}(C\log n)^2h^{-(2a+1)/2}C'p_n^{1/(4m)}\left(h^{-1}\log\log n\right)^{1/2}\\
=&\sqrt{2C_R}C'C^2\kappa^\gamma Mn^{1/2}r_nh^{-\left(2a+\frac{3}{2}+\frac{2k-2a-1}{4m}\right)}(\log n)^2(\log\log n)^{1/2}.
\end{aligned}$$

In the meantime, by Lemmas 5.7 and S.4 in [164], for $l=0,1$,

$$\begin{aligned}
&E_l\left\{\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\|R_{X^l}\|\cdot|\langle R_{X^l},\theta\rangle|^2\right\}\\
\leq& E\left\{\|R_{X^l}\|\cdot|\langle R_{X^l},\theta\rangle|^2\right\}\\
\leq& C_R^{1/2}h^{-1/2}\left(8\tilde{M}C_2^2\right)^{1/2}\|\theta\|^2\\
\leq& C_R^{1/2}\left(8\tilde{M}C_2^2\right)^{1/2}M^2h^{-1/2}r_n^2
\end{aligned}$$

So for some large C'', with probability approaching one,

$$\begin{aligned}
&|I_1|\\
\leq&\frac{Mr_n}{n}\left\|\sum_{l=0}^{1}\sum_{i=1}^{n_l}\left[\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\left\|R_{X_i^l}\right\|\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}\right.\right.\\
&\left.\left.-E_l\left\{\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\left\|R_{X_i^l}\right\|\left\langle R_{X_i^l},\theta\right\rangle R_{X_i^l}I_{A_i^l}\right\}\right]\right\|\|\theta\|\\
&+Mr_n\sum_{l=0}^{1}n_l/nE_l\left\{\sup_{a\in\mathbb{R}}|\ell'''(a)|\cdot\|R_{X^l}\|\cdot|\langle R_{X^l},\theta\rangle|^2\right\}\\
\leq& C''\left(n^{-1/2}h^{-\left(2a+\frac{3}{2}+\frac{2k-2a-1}{4m}\right)}r_n^3(\log n)^2(\log\log n)^{1/2}+h^{-1/2}r_n^3\right).
\end{aligned}$$

Therefore, $I_1=O_P(n^{-1/2}h^{-\left(2a+\frac{3}{2}+\frac{2k-2a-1}{4m}\right)}r_n^3(\log n)^2(\log\log n)^{1/2}+h^{-1/2}r_n^3)=o_P\left(n^{-1}h^{-1/2}\right)$. Similar techniques can be applied for I_2. Notice that on $\mathcal{E}_n, I_{A_i^l}=1$ for $l=0,1, i=$

$1, \ldots, n_l$, thus,

$$
\begin{aligned}
& 2I_2 \\
=& \frac{1}{n} \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^2 - E_l \left\{ \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^2 \right\} \right] \\
& + \sum_{l=0}^{1} [n_l/n - \rho_l] E_l \left\{ \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^2 \right\} \\
=& n^{-1} \left\langle \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi \left(X_i^l; \theta \right) R_{X_i^l} - E_l \left\{ \psi \left(X_i^l; \theta \right) R_{X_i^l} \right\} \right], \theta \right\rangle \\
& - E_l \left\{ \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^2 I_{A_i^{lc}} \right\} \\
& + \sum_{l=0}^{1} [n_l/n - \rho_l] E_l \left\{ \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left| \left\langle R_{X_i^l}, \theta \right\rangle \right|^2 \right\},
\end{aligned}
$$

where $\psi \left(X_i^l; \theta \right) = \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \left\langle R_{X_i^l}, \theta \right\rangle I_{A_i^l}$. Let $d_n = \kappa M h^{-(2a+1)/2} r_n, \tilde{\theta} = d_n^{-1} \theta$ and $\psi_n \left(X_i^l; \tilde{\theta} \right) = (C \log n)^{-1} d_n^{-1} \psi \left(X_i^l; d_n \tilde{\theta} \right)$. For any $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathcal{F}_{p_n}$, where $p_n = \kappa^{-2} h^{2a+1-2k}$, it can be shown that

$$
\begin{aligned}
& \left| \psi_n \left(X_i^l; \tilde{\theta}_1 \right) - \psi_n \left(X_i^l; \tilde{\theta}_2 \right) \right| \\
=& (C \log n)^{-1} d_n^{-1} \left| \ddot{\ell} \left(\left\langle R_{X_i^l}, \theta_0 \right\rangle \right) \right| \cdot \left| \left\langle R_{X_i^l}, d_n \left(\tilde{\theta}_1 - \tilde{\theta}_2 \right) \right\rangle \right| I_{A_i^l} \\
\leq & \left\| \tilde{\theta}_1 - \tilde{\theta}_2 \right\|_2 .
\end{aligned}
$$

Again by Lemma 5.2, for some large C', with probability approaching one,

$$
\left\| H_n(\tilde{\theta}) \right\| \leq C' p_n^{1/(4m)} \left(h^{-1} \log \log n \right)^{1/2} .
$$

Thus

$$
\begin{aligned}
& \left\| \sum_{l=0}^{1} \sum_{i=1}^{n_l} \left[\psi \left(X_i^l; \theta \right) R_{X_i^l} - E_l \left\{ \psi \left(X_i^l; \theta \right) R_{X_i^l} \right\} \right] \right\| \\
\leq & C' (C \log n) d_n h^{-\frac{2k-2a-1}{4m}} n^{1/2} \left(h^{-1} \log \log n \right)^{1/2} \\
= & C' C \kappa M n^{1/2} h^{-\left(a+1+\frac{2-2a-1}{4m} \right)} r_n (\log n) (\log \log n)^{1/2}.
\end{aligned}
$$

By choosing an approapriate C, and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
& E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_0\right\rangle\right)\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2 I_{A_i^{lc}}\right\} \\
\leq & C_2 E_l\left\{\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2 I_{A_i^{lc}}\right\} \\
\leq & C_2 E_l\left\{\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^4\right\}^{1/2} P\left(A_i^{lc}\right)^{1/2} \\
\leq & C_2\left(8\tilde{M}C_2^2\right)^{1/2}(Mr_n)^2 P\left(A_i^{lc}\right)^{1/2} \\
= & o\left(n^{-1/2}h^{-\left(a+1+\frac{2k-2a-1}{4m}\right)}r_n^2(\log n)(\log\log n)^{1/2}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
& \sum_{l=0}^{1}[n_l/n-\rho_l]E_l\left\{\ddot{\ell}\left(\left\langle R_{X_i^l},\theta_0\right\rangle\right)\left|\left\langle R_{X_i^l},\theta\right\rangle\right|^2\right\} \\
\leq & \sum_{l=0}^{1}[n_l/n-\rho_l]C_2\left(8\tilde{M}C_2^2\right)^{1/2}(Mr_n)^2 = O_P(n^{-1/2}r_n^2).
\end{aligned}
$$

This implies that

$$
I_2 = O_P\left(n^{-1/2}h^{-\left(a+1+\frac{2k-2a-1}{4m}\right)}r_n^2(\log n)(\log\log n)^{1/2}\right) = o_P\left(n^{-1}h^{-1/2}\right).
$$

By the assumptions of the theorem, we then get

$$
\begin{aligned}
-2n\cdot\mathrm{PLRT} &= -2n\cdot\left(\ell_{n,\lambda}\left(\theta_0\right)-\ell_{n,\lambda}\left(\widehat{\theta}_{n,\lambda}\right)\right) \\
&= n\left\|\widehat{\theta}_{n,\lambda}-\theta_0\right\|^2 - 2n\left(I_1+I_2\right) \\
&= n\left\|\widehat{\theta}_{n,\lambda}-\theta_0\right\|^2 + o_P\left(h^{-1/2}\right).
\end{aligned}
$$

We next focus on the leading term $\|\widehat{\theta}_{n,\lambda}-\theta_0\|^2$ to derive the distribution of the LRT test. By Theorem 5.4 and the conditions of this theorem, $n^{1/2}\|\widehat{\theta}_{n,\lambda}-\theta_0-S_{n,\lambda}\left(\theta_0\right)\| = O_P\left(n^{1/2}a_n\right) = o_P(1)$. So

$$
\begin{aligned}
& -2n\cdot\mathrm{PLRT} \\
= & n\left\|\widehat{\theta}_{n,\lambda}-\theta_0\right\|^2 + o_P\left(h^{-1/2}\right) \\
= & \left(n^{1/2}\left\|S_{n,\lambda}\left(\theta_0\right)\right\| + o_P(1)\right)^2 + o_P\left(h^{-1/2}\right) \\
= & n\left\|S_{n,\lambda}\left(\theta_0\right)\right\|^2 + n^{1/2}\left\|S_{n,\lambda}\left(\theta_0\right)\right\|\cdot o_P(1) + o_P\left(h^{-1/2}\right). \qquad (5.31)
\end{aligned}
$$

We then study the asymptotic property of the term $\|S_{n,\lambda}(\theta_0)\|^2$. Let $\epsilon_1, \ldots, \epsilon_{n_0}$, $\epsilon_{n_0+1}, \ldots, \epsilon_n$ represent $\epsilon_1^0, \ldots, \epsilon_{n_0}^0, \epsilon_1^1, \ldots, \epsilon_{n_0}^1$. By direct calculations, we have

$$n\|S_{n,\lambda}(\theta_0)\|^2 = n^{-1}\left\|\sum_{i=1}^n \epsilon_i R_{X_i}\right\|^2 - 2\left\langle \sum_{i=1}^n \epsilon_i R_{X_i}, P_\lambda\theta_0 \right\rangle + n\|P_\lambda\theta_0\|^2.$$

Write $\beta_0 = \sum_\nu b_\nu^0 \varphi_\nu$. We know that $\sum_\nu |b_\nu^0|^2 \rho_\nu < \infty$. Then

$$\begin{aligned}
&E\left\{\left|\left\langle \sum_{i=1}^n \epsilon_i R_{X_i}, P_\lambda\theta_0 \right\rangle\right|^2\right\} \\
=&\sum_{l=0}^1 n_l E_l\left\{(\epsilon^l)^2 |\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\} \\
=&\sum_{l=0}^1 (n_l - n\rho_l) E_l\left\{(\epsilon^l)^2 |\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\} + \sum_{l=0}^1 n\rho_l E_l\left\{(\epsilon^l)^2 |\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\}.
\end{aligned}$$

We have for $l = 0, 1$,

$$\begin{aligned}
&(n_l - n\rho_l) E_l\left\{(\epsilon^l)^2 |\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\} \\
\le&(n_l - n\rho_l) E_l\left\{|\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\} \\
=&(n_l - n\rho_l) E_l\left\{\left|\int_0^1 X^l(t)\left(W_\lambda\beta_0\right)(t)dt\right|^2\right\} \\
\le&(n_l - n\rho_l) E_l\left\{\|R_{X^l}\|^2\right\}\|W_\lambda\beta_0\|_1^2 = o(n\lambda).
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\sum_{l=0}^1 n\rho_l E_l\left\{(\epsilon^l)^2 |\langle R_{X^l}, P_\lambda\theta_0\rangle|^2\right\} \\
=&n\rho_l E_l\left\{B(X^l)\left|\int_0^1 X^l(t)\left(W_\lambda\beta_0\right)(t)dt\right|^2\right\} \\
=&nV\left(W_\lambda\beta_0, W_\lambda\beta_0\right) \\
=&nV\left(\sum_\nu b_\nu^0 \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}, \sum_\nu b_\nu^0 \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\right) \\
=&n\lambda\sum_\nu \left|b_\nu^0\right|^2 \rho_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} = o(n\lambda),
\end{aligned}$$

where the last equality follows from $\sum_\nu |b_\nu^0|^2 \rho_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} = o(1)$, as $\lambda \to 0$ that can be shown by dominated convergence theorem. Therefore, $\langle \sum_{i=1}^n \epsilon_i R_{X_i}, P_\lambda\theta_0\rangle =$

$o_P\left((n\lambda)^{1/2}\right)$. It can be shown by direct calculations and dominated convergence theorem that

$$\begin{aligned}
& n\left\|P_\lambda\theta_0\right\|^2 \\
=& n\left\|W_\lambda\beta_0\right\|_1^2 \\
=& n\left\langle \sum_\nu b_\nu^0 \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\varphi_\nu, \sum_\nu b_\nu^0 \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\varphi_\nu \right\rangle_1 \\
=& n\sum_\nu \left|b_\nu^0\right|^2 \frac{\left(\lambda\rho_\nu\right)^2}{1+\lambda\rho_\nu} = o(n\lambda)
\end{aligned}$$

Thus, by the condition $n\lambda = O\left(h^{-1}\right)$ we get that $n\left\|S_{n,\lambda}\left(\theta_0\right)\right\|^2 = n^{-1}\left\|\sum_{i=1}^n \epsilon_i R_{X_i}\right\|^2 + o_P\left(h^{-1/2}+n\lambda\right)$. Therefore, it follows by (5.31) that

$$\begin{aligned}
& -2n\cdot \mathrm{PLRT} \\
=& n^{-1}\left\|\sum_{i=1}^n \epsilon_i R_{X_i}\right\|^2 + n\left\|W_\lambda\beta_0\right\|_1^2 + n^{1/2}\left\|S_{n,\lambda}\left(\theta_0\right)\right\|\cdot o_P(1) \\
& + o_P\left(h^{-1/2}\right).
\end{aligned} \tag{5.32}$$

Finally, we examine the term $n^{-1}\left\|\sum_{i=1}^n \epsilon_i R_{X_i}\right\|^2$ which could be split to

$$\begin{aligned}
\left\|\sum_{i=1}^n \epsilon_i R_{X_i}\right\|^2 &= \sum_{i=1}^n \epsilon_i^2 \left\langle R_{X_i}, R_{X_i}\right\rangle + \sum_{1\le i<j\le n} W_{ij} \\
&\equiv \sum_{i=1}^n \epsilon_i^2 \left\langle R_{X_i}, R_{X_i}\right\rangle + W(n)
\end{aligned}$$

where $W_{ij} = 2\epsilon_i\epsilon_j\left\langle R_{X_i}, R_{X_j}\right\rangle$. By the same arguments in [79, 164], the term $W(n)/\sqrt{2n^2h^{-1}\sigma_2^2} \xrightarrow{d} N(0,1)$, as $n\to\infty$. Notice that

$$\begin{aligned}
& E\left\{\left|\left\langle R_X, R_X\right\rangle\right|^2\right\} \\
\le& 2C_2^2 + 2E\left\{\left|\left\langle \tau(X), \tau(X)\right\rangle_1\right|^2\right\} \\
=& 2C_2^2 + 2E\left\{\left|\sum_\nu \frac{X_\nu^2}{1+\lambda\rho_\nu}\right|^2\right\} \\
\le& 2C_2^2 + 2E\left\{\sum_\nu \frac{X_\nu^4}{1+\lambda\rho_\nu}\right\}\sum_\nu \frac{1}{1+\lambda\rho_\nu} = O\left(h^{-2}\right).
\end{aligned} \tag{5.33}$$

Thus

$$
\begin{aligned}
& E\left\{\left|\sum_{i=1}^{n}\left[\epsilon_i^2\left\|R_{X_i}\right\|^2-E\left\{\epsilon_i^2\left\|R_{X_i}\right\|^2\right\}\right]\right|^2\right\} \\
\leq & 2\sum_{l=0}^{1} n_l E_l\left\{(\epsilon^l)^4\left|\langle R_{X^l}, R_{X^l}\rangle\right|^2\right\}+ \\
& 2\sum_{l=0}^{1}(n_l-n\rho_l) E_l\left\{(\epsilon^l)^4\left|\langle R_{X^l}, R_{X^l}\rangle\right|^2\right\} \\
= & O\left(nh^{-2}\right), \\
& n^{-1}\sum_{i=1}^{n}\epsilon_i^2\left\|R_{X_i}\right\|^2 \\
= & E\left\{\epsilon_i^2\left\|R_{X_i}\right\|^2\right\}+O_P\left(n^{-1/2}h^{-1}\right) \\
= & 1+h^{-1}\sigma_1^2+O_P\left(n^{-1/2}h^{-1}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
& n\left\|S_{n,\lambda}\left(\theta_0\right)\right\|^2 \\
= & n^{-1}\sum_{i=1}^{n}\epsilon_i^2\left\|R_{X_i}\right\|^2+n^{-1}W(n)+O_P\left(h^{-1/2}+n\lambda\right) \\
= & O_P\left(h^{-1}+n^{-1/2}h^{-1}+h^{-1/2}+n\lambda\right)=O_P\left(h^{-1}\right).
\end{aligned}
$$

Therefore, it follows by (5.32) that

$$
\begin{aligned}
& -2n\cdot\mathrm{PLRT} \\
= & n^{-1}\sum_{i=1}^{n}\epsilon_i^2\left\|R_{X_i}\right\|^2+n^{-1}W(n)+n\left\|W_\lambda\beta_0\right\|_1^2+o_P\left(h^{-1/2}\right) \\
= & h^{-1}\sigma_1^2+n^{-1}W(n)+n\left\|W_\lambda\beta_0\right\|_1^2+o_P\left(h^{-1/2}\right).
\end{aligned}
$$

This implies that, as $n\to\infty$,

$$
-\frac{2n\cdot\mathrm{PLRT}+h^{-1}\sigma_1^2+n\left\|W_\lambda\beta_0\right\|_1^2}{\sqrt{2\sigma_2^2h^{-1}}}\xrightarrow{d}N(0,1).
$$

We have completed the proof of Theorem 5.8. ■

Proof of Theorem 5.6. The proof of Theorem 5.6 follows from the proof of Theorem 3 of [98] with the weighting operator $I(U)$ defined in (5.18), and is omitted here. ■

Chapter 6

Concluding Remarks and Future Research

The realm of functional predictor regression has witnessed significant advancements, particularly over the last decade. While a subset of these methodologies incorporates nonlinear dynamics, the majority adhere to variations of the Gaussian model or Generalized Functional Linear Models, designated as Models (1.1) and (2.2). The primary distinction across these methods lies in their selection of basis functions for portraying the predictors $X_i(t)$ and the functional coefficient $\beta(t)$. Principal components, splines, and wavelets, individually or in combination, are among the favored choices, coupled with distinct regularization strategies. This thesis rotates around these models and the evaluation of effective regularization techniques remains a crucial need and potential future directions. Such advancements would greatly benefit practitioners in selecting the most appropriate methodological approach for their specific dataset.

The field of Topological Data Analysis (TDA) has emerged with myriad ideas, yet its practical impact on data analysis remains modest. This could be attributed to the novelty of the techniques, their complexity, or a possible mismatch between the methods and practical applications. The effectiveness of TDA as a specialized tool appears promising for specific challenges, such as analyzing data related to cosmic structures.

The integration of deep learning techniques with FDA presents another intriguing

avenue for future research. Its ability to handle large datasets and uncover complex patterns is opening new frontiers in data analysis. This synergy could be particularly beneficial in fields where both the geometric structure and high-dimensional data play crucial roles, such as genomics, neuroimaging, and complex system analysis. Moreover, there is a pressing need for comprehensive comparison and collaboration between deep learning approaches and traditional statistical methods. Such studies across a variety of scientific disciplines would be invaluable.

Bibliography

[1] Mingyao Ai, Fei Wang, Jun Yu, and Huiming Zhang. "Optimal subsampling for large-scale quantile regression". In: *Journal of Complexity* 62 (2021), p. 101512.

[2] Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. "Optimal subsampling algorithms for big data regressions". In: *Statistica Sinica* 31 (2021).

[3] Mark A. Aizerman, E. M. Braverman, and L. I. Rozonoer. "Theoretical foundations of the potential function method in pattern recognition learning". In: *Automation and remote control* 25 (1964), pp. 821–837.

[4] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[5] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds". In: *Journal of the American Statistical Association* 101.473 (2006), pp. 138–156.

[6] Peter J Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. "In vivo fiber tractography using DT-MRI data". In: *Magnetic resonance in medicine* 44.4 (2000), pp. 625–632.

[7] Yoshua Bengio, Pascal Vincent, Jean-François Paiement, Olivier Delalleau, Marie Ouimet, and Nicolas Le Roux. *Spectral clustering and kernel PCA are learning eigenfunctions*. Tech. rep. Departement d'Informatique et Recherche Operationnelle, Technical Report 1239, 2003.

[8] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[9] Leslie J. Bisson, Melissa A. Kluczynski, William M Wind, Marc S Fineberg, Geoffrey A. Bernas, Michael A. Rauh, John M. Marzo, and Robert J. Smolinski. "Design of a randomized controlled trial to compare debridement to observation of chondral lesions encountered during partial meniscectomy: The ChAMP (Chondral Lesions And Meniscus Procedures) Trial". In: *Contemporary Clinical Trials* 45 (2015), pp. 281–286.

[10] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. "Statistical performance of support vector machines". In: *The Annals of Statistics* 36.2 (2008), pp. 489–531.

[11] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[12] Stephen Boyd. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends in Machine Learning* 3.1 (2010), pp. 1–122.

[13] H. Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY: Springer, 2011.

[14] Philip J Brown, Tom Fearn, and Marina Vannucci. "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem". In: *Journal of the American Statistical Association* 96.454 (2001), pp. 398–408.

[15] T Tony Cai and Peter Hall. "Prediction in functional linear regression". In: *The Annals of Statistics* 34.5 (2006), pp. 2159–2179.

[16] T Tony Cai and Ming Yuan. "Minimax and adaptive prediction for functional linear regression". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1201–1216.

[17] T Tony Cai and Ming Yuan. "Optimal estimation of the mean function based on discretely sampled functional data: Phase transition". In: *The Annals of Statistics* 39.5 (2011), pp. 2330–2355.

[18] Hervé Cardot, Frédéric Ferraty, André Mas, and Pascal Sarda. "Testing hypotheses in the functional linear model". In: *Scandinavian Journal of Statistics* 30.1 (2003), pp. 241–255.

[19] Colin Chen and Ying Wei. "Computational issues for quantile regression". In: *Sankhyā: The Indian Journal of Statistics* 67.2 (2005), pp. 399–417.

[20] Kani Chen. "Parametric models for response-biased sampling". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.4 (2001), pp. 775–789.

[21] Kani Chen, Yuanyuan Lin, Yuan Yao, and Chaoxu Zhou. "Regression analysis with response-selective sampling". In: *Statistica Sinica* 27 (2017), pp. 1699–1714.

[22] Kani Chen and Shaw-Hwa Lo. "Case-cohort and case-control analysis with Cox's model". In: *Biometrika* 86.4 (1999), pp. 755–764.

[23] Guang Cheng and Zuofeng Shang. "Joint asymptotics for semi-nonparametric regression models with partially linear structure". In: *The Annals of Statistics* 43.3 (2015), pp. 1351–1390.

[24] Qianshun Cheng, Haiying Wang, and Min Yang. "Information-based optimal subdata selection for big data logistic regression". In: *Journal of Statistical Planning and Inference* 209 (2020), pp. 112–122.

[25] Keith Conrad. *The minimal polynomial and some applications*. Tech. rep. Department of Mathematics, University of Connecticut, 2014.

[26] David R Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232.

[27] Ciprian M. Crainiceanu and A. Jeffrey Goldsmith. "Bayesian Functional Data Analysis Using WinBUGS." In: *Journal of statistical software* 32.11 (2010), pp. 1–33.

[28] Ciprian M. Crainiceanu and David Ruppert. "Likelihood ratio tests in linear mixed models with one variance component". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1 (2004), pp. 165–185.

[29] Ciprian M. Crainiceanu, Ana-Maria Staicu, and Chongzhi Di. "Generalized multilevel functional regression". In: *Journal of the American Statistical Association* 104.488 (2009), pp. 1550–1561.

[30] Christophe Crambes, Alois Kneip, and Pascal Sarda. "Smoothing splines estimators for functional linear regression". In: *The Annals of Statistics* 37.1 (2009), pp. 35–72.

[31] Christophe Crambes, Alois Kneip, and Pascal Sarda. "Smoothing Splines Estimators for Functional Linear Regression". In: *The Annals of Statistics* 37.1 (2009), pp. 35–72.

[32] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[33] Felipe Cucker and Steve Smale. "On the mathematical foundations of learning". In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49.

[34] Aurore Delaigle and Peter Hall. "Achieving near perfect classification for functional data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2 (2012), pp. 267–286.

[35] Aurore Delaigle and Peter Hall. "Classification using censored functional data". In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1269–1283.

[36] Jianghu J Dong, Liangliang Wang, Jagbir Gill, and Jiguo Cao. "Functional principal component analysis of glomerular filtration rate curves after kidney transplant". In: *Statistical Methods in Medical Research* 27.12 (2018), 3785–3796.

[37] David L. Donoho and Iain M. Johnstone. "Adapting to Unknown Smoothness via Wavelet Shrinkage". In: *Journal of the American Statistical Association* 90.432 (1995), pp. 1200–1224.

[38] Pang Du and Xiao Wang. "Penalized likelihood functional regression". In: *Statistica Sinica* 24.2 (2014), pp. 1017–1041.

[39] Pang Du and Xiao Wang. "Penalized likelihood functional regression". In: *Statistica Sinica* 24.2 (2014), pp. 1017–1041.

[40] Shu-Xin Du and Sheng-Tan Chen. "Weighted support vector machine for classification". In: *2005 IEEE International Conference on Systems, Man and Cybernetics*. Vol. 4. IEEE. 2005, pp. 3866–3871.

[41] Paul H.C. Eilers and Brian D Marx. "Flexible smoothing with B-splines and penalties". In: *Statistical science* 11.2 (1996), pp. 89–121.

[42] Paul H.C. Eilers and Brian D. Marx. "Generalized linear additive smooth structures". In: *Journal of Computational and Graphical Statistics* 11.4 (2002), pp. 758–783.

[43] Yan Fan, Yukun Liu, and Lixing Zhu. "Optimal subsampling for linear quantile regression models". In: *Canadian Journal of Statistics* 49.4 (2021), pp. 1039–1057.

[44] Yingying Fan, Gareth M James, and Peter Radchenko. "Functional additive regression". In: *The Annals of Statistics* (2015), pp. 2296–2325.

[45] Huijie Feng, Jingyi Duan, Yang Ning, and Jiwei Zhao. "Test of Significance for High-dimensional Thresholds with Application to Individualized Minimal Clinically Important Difference". In: *Journal of the American Statistical Association* (2023), pp. 1–13.

[46] Huijie Feng, Yang Ning, and Jiwei Zhao. "Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses". In: *The Annals of Statistics* 50.4 (2022), pp. 2284–2305.

[47] William Fithian and Trevor Hastie. "Local case-control sampling: Efficient subsampling in imbalanced data sets". In: *Annals of statistics* 42.5 (2014), p. 1693.

[48] Jan Gertheiss, Jeff Goldsmith, Ciprian M. Crainiceanu, and Sonja Greven. "Longitudinal scalar-on-functions regression with application to tractography data." In: *Biostatistics (Oxford, England)* 14.3 (2013), pp. 447–461.

[49] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. *Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nystr\" om Method, and Use of Kernels in Machine Learning: Tutorial and Survey*. 2021.

[50] Jeff Goldsmith, Jennifer Bobb, Ciprian M. Crainiceanu, Brian Caffo, and Daniel Reich. "Penalized Functional Regression". In: *Journal of Computational and Graphical Statistics* 20.4 (2011), pp. 830–851.

[51] Jeff Goldsmith, Jennifer F. Bobb, Ciprian M. Crainiceanu, Brian Caffo, and Daniel S. Reich. "Penalized functional regression". In: *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 20.4 (2011), pp. 830–851.

[52] Jeff Goldsmith, Ciprian M. Crainiceanu, Brian Caffo, and Daniel Reich. "Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.3 (2012), pp. 453–469.

[53] Jeff Goldsmith, Lei Huang, and Ciprian M. Crainiceanu. "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection". In: *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 23.1 (2014), pp. 46–64.

[54] Jeff Goldsmith, Matt P Wand, and Ciprian Crainiceanu. "Functional regression via variational Bayes". In: *Electronic journal of statistics* 5 (2011), p. 572.

[55] Gene H. Golub, Michael Heath, and Grace Wahba. "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* 21.2 (1979), pp. 215–223.

[56] Sonja Greven, Ciprian Crainiceanu, Brian Caffo, and Daniel Reich. "Longitudinal functional principal component analysis". In: *Recent Advances in Functional Data Analysis and Related Topics*. Springer. 2011, pp. 149–154.

[57] Chong Gu. *Smoothing Spline ANOVA Models*. New York: Springer, 2013.

[58] Yuwen Gu, Jun Fan, Lingchen Kong, Shiqian Ma, and Hui Zou. "ADMM for high-dimensional sparse penalized quantile regression". In: *Technometrics* 60.3 (2018), pp. 319–331.

[59] Yuwen Gu and Hui Zou. "High-dimensional generalizations of asymmetric least squares regression and their applications". In: *The Annals of Statistics* 44.6 (2016), pp. 2661–2694.

[60] Mengmeng Guo, Lan Zhou, Jianhua Z Huang, and Wolfgang Karl Härdle. "Functional data analysis of generalized regression quantiles". In: *Statistics and Computing* 25.2 (2015), pp. 189–202.

[61] Peter Hall and Joel L. Horowitz. "Methodology and convergence rates for functional linear regression". In: *The Annals of Statistics* 35.1 (Feb. 2007), pp. 70–91.

[62] Lei Han, Kean Ming Tan, Ting Yang, and Tong Zhang. "Local uncertainty sampling for large-scale multiclass logistic regression". In: *The Annals of Statistics* 48.3 (2020), pp. 1770–1788.

[63] Meiling Hao, Kin-yat Liu, Wei Xu, and Xingqiu Zhao. "Semiparametric inference for the functional Cox model". In: *Journal of the American Statistical Association* 116.535 (2021), pp. 1319–1329.

[64] Trevor Hastie and Colin Mallows. "[A statistical view of some chemometrics regression tools]: Discussion". In: *Technometrics* 35.2 (1993), pp. 140–143.

[65] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[66] Xuming He. "Quantile curves without crossing". In: *The American Statistician* 51.2 (1997), pp. 186–192.

[67] Yafei Wang; Linglong Kong; Bei Jiang; Xingcai Zhou; Shimei Yu; Li Zhang; Giseon Heo. "Wavelet-based LASSO in functional linear quantile regression". In: *Journal of Statistical Computation and Simulation* 89.6 (2019), pp. 1111–1130.

[68] Scott H Holan, Christopher K Wikle, Laura E Sullivan-Beckers, and Reginald B Cocroft. "Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals". In: *Biometrics* 66.3 (2010), pp. 914–924.

[69] Hajo Holzmann and Bernhard Klar. "Expectile asymptotics". In: *Electronic Journal of Statistics* 10.2 (2016), pp. 2355–2371.

[70] Joel L. Horowitz. "A Smoothed Maximum Score Estimator for the Binary Response Model". In: *Econometrica* 60.3 (1992), pp. 505–531.

[71] Joel L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer, 2009.

[72] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging* 27.4 (2008), pp. 685–691.

[73] Gareth M James. "Generalized linear models with functional predictors". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 411–432.

[74] Gareth M James. "Generalized linear models with functional predictors". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 411–432.

[75] Gareth M. James, Jing Wang, and Ji Zhu. "Functional linear regression that's interpretable". In: *The Annals of Statistics* 37.5 (2009), pp. 2083–2108.

[76] Gareth M. James, Jing Wang, and Ji Zhu. "Functional linear regression that's interpretable". In: *The Annals of Statistics* 37.5A (Oct. 2009), pp. 2083–2108.

[77] Ian T Jolliffe. "A note on the use of principal components in regression". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 31.3 (1982), pp. 300–303.

[78] M. C. Jones. "Expectiles and M-quantiles are quantiles". In: *Statistics & Probability Letters* 20.2 (1994), pp. 149–153.

[79] Peter de Jong. "A central limit theorem for generalized quadratic forms". In: *Probability Theory and Related Fields* 75.2 (1987), 261–277.

[80] Kengo Kato. "Estimation in functional linear quantile regression". In: *The Annals of Statistics* 40.6 (2012), pp. 3108–3136.

[81] Mitsunori Kayano, Hidetoshi Matsui, Rui Yamaguchi, Seiya Imoto, and Satoru Miyano. "Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection". In: *Biostatistics* 17.2 (2016), pp. 235–248.

[82] Maurice G Kendall. "A course in multivariate analysis: London". In: *Charles Griffin & Co* (1957).

[83] Jane Paik Kim, Wenbin Lu, Tony Sit, and Zhiliang Ying. "A unified approach to semiparametric transformation models under general biased sampling schemes". In: *Journal of the American Statistical Association* 108.501 (2013), pp. 217–227.

[84] George Kimeldorf and Grace Wahba. "Some results on Tchebycheffian spline functions". In: *Journal of mathematical analysis and applications* 33.1 (1971), pp. 82–95.

[85] George Kimeldorf and Grace Wahba. "Some results on Tchebycheffian spline functions". In: *Journal of Mathematical Analysis and Applications* 33.1 (1971), pp. 82–95.

[86] Roger Koenker. "Quantile regression: 40 years on". In: *Annual Review of Economics* 9 (2017), pp. 155–176.

[87] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[88] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. New York: Springer, 2011.

[89] Volker Krätschmer and Henryk Zähle. "Statistical Inference for Expectile-based Risk Measures". In: *Scandinavian Journal of Statistics* 44.2 (2017), pp. 425–454.

[90] Bruce R Kusse and Erik A Westwig. *Mathematical Physics: Applied Mathematics For Scientists and Engineers*. 2nd ed. Wiley-VCH, 2006.

[91] Stefan Lang and Andreas Brezger. "Bayesian P-splines". In: *Journal of computational and graphical statistics* 13.1 (2004), pp. 183–212.

[92] J. F. Lawless. "Likelihood and pseudo likelihood estimation based on response-biased observation". In: *Lecture Notes-Monograph Series* (1997), pp. 43–55.

[93] Eun Ryung Lee and Byeong U. Park. "Sparse estimation in functional linear regression". In: *Journal of Multivariate Analysis* 105.1 (2012), pp. 1–17.

[94] Seunggeun Lee, Fei Zou, and Fred A. Wright. "Convergence and prediction of principal component scores in high-dimensional settings". In: *Annals of statistics* 38.6 (2010), pp. 3605–3629.

[95] Bin Li and Brian D Marx. "Sharpening P-spline signal regression". In: *Statistical Modelling* 8.4 (2008), pp. 367–383.

[96] Gen Li, Jianhua Z Huang, and Haipeng Shen. "To wait or not to wait: Two-way functional hazards model for understanding waiting in call centers". In: *Journal of the American Statistical Association* 113.524 (2018), pp. 1503–1514.

[97] Jialiang Li, Chao Huang, and Zhub Hongtu. "A Functional Varying-Coefficient Single-Index Model for Functional Response Data". In: *Journal of the American Statistical Association* 112.519 (2017), 1169–1181.

[98] Ting Li and Zhongyi Zhu. "Inference for generalized partial functional linear regression". In: *Statistica Sinica* 30 (2020).

[99] Yehua Li, Naisyin Wang, and Raymond J Carroll. "Generalized functional linear models with semiparametric single-index interactions". In: *Journal of the American Statistical Association* 105.490 (2010), pp. 621–633.

[100] Youjuan Li, Yufeng Liu, and Ji Zhu. "Quantile Regression in Reproducing Kernel Hilbert Spaces". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 255–268.

[101] Zhaoyuan Li and Jianfeng Yao. "Testing for heteroscedasticity in high-dimensional regressions". In: *Econometrics and Statistics* 9 (2019), pp. 122–139.

[102] Lina Liao, Cheolwoo Park, and Hosik Choi. "Penalized expectile regression: An alternative to penalized quantile regression". In: *Annals of the Institute of Statistical Mathematics* 71.2 (2019), pp. 409–438.

[103] Chun-Fu Lin and Sheng-De Wang. "Fuzzy support vector machines". In: *IEEE Transactions on Neural Networks* 13.2 (2002), pp. 464–471.

[104] Yuanyuan Lin, Jinhan Xie, Ruijian Han, and Niansheng Tang. "Post-selection Inference of High-dimensional Logistic Regression Under Case–Control Design". In: *Journal of Business & Economic Statistics* 41.2 (2023), pp. 624–635.

[105] Hua Liu, Jinhong You, and Jiguo Cao. "Functional L-Optimality Subsampling for Functional Generalized Linear Models with Massive Data". In: *Journal of Machine Learning Research* 24.219 (2023).

[106] Meichen Liu, Matthew Pietrosanu, Peng Liu, Bei Jiang, Xingcai Zhou, Linglong Kong, and Alzheimer's Disease Neuroimaging Initiative. "Reproducing kernel-based functional linear expectile regression". In: *Canadian Journal of Statistics* 50.1 (2022), pp. 241–266.

[107] Cong Ma, Junwei Lu, and Han Liu. "Inter-Subject Analysis: A Partial Gaussian Graphical Model Approach". In: *Journal of the American Statistical Association* 116 (2020), pp. 1–57.

[108] Ping Ma, Michael W Mahoney, and Bin Yu. "A statistical perspective on algorithmic leveraging". In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 861–911.

[109] Elizabeth J. Malloy, Jeffrey S. Morris, Sara D. Adar, Helen Suh, Diane R. Gold, and Brent A. Coull. "Wavelet-based functional linear mixed models: an application to measurement error–corrected distributed lag models". In: *Biostatistics (Oxford, England)* 11.3 (2010), pp. 432–452.

[110] Charles F Manski, Daniel McFadden, et al. *Structural analysis of discrete data with econometric applications*. MIT press Cambridge, MA, 1981.

[111] Nathan Mantel and William Haenszel. "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease". In: *JNCI: Journal of the National Cancer Institute* 22.4 (Apr. 1959), pp. 719–748.

[112] James Stephen Marron, Michael J Todd, and Jeongyoun Ahn. "Distance-weighted discrimination". In: *Journal of the American Statistical Association* 102.480 (2007), pp. 1267–1271.

[113] Brian D Marx and Paul H.C. Eilers. "Generalized linear regression on sampled signals and curves: a P-spline approach". In: *Technometrics* 41.1 (1999).

[114] Brian D Marx and Paul H.C. Eilers. "Multidimensional penalized signal regression". In: *Technometrics* 47.1 (2005), pp. 13–22.

[115] Brian D. Marx, Paul H.C. Eilers, and Bin Li. "Multidimensional single-index signal regression". In: *Chemometrics and Intelligent Laboratory Systems* 109.2 (2011), pp. 120–130.

[116] Mathew W McLean, Fabian Scheipl, Giles Hooker, Sonja Greven, and David Ruppert. *Bayesian functional generalized additive models with sparsely observed covariates*. 2013.

[117] Matthew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. "Functional generalized additive models". In: *Journal of computational and graphical statistics* 23.1 (2014), pp. 249–269.

[118] Sanjay Mehrotra. "On the Implementation of a Primal-Dual Interior Point Method". In: *SIAM Journal on Optimization* 2.4 (1992), pp. 575–601.

[119] James Mercer. "Functions of positive and negative type and their connection with the theory of integral equations". In: *Philosophical Transactions of the Royal Society* A.209 (1909), pp. 415–446.

[120] Charles A Micchelli and Grace Wahba. *Design Problems for Optimal Surface Interpolation*. Tech. rep. ADA070012. Department of Statistics, University of Wisconsin–Madison, 1979.

[121] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Cambridge, MA: MIT press, 2018.

[122] Hans-Georg Müller. "Functional modelling and classification of longitudinal data". In: *Scandinavian Journal of Statistics* 32.2 (2005), pp. 223–240.

[123] Hans-Georg Müller and Ulrich Stadtmüller. "Generalized functional linear models". In: *the Annals of Statistics* 33.2 (2005), pp. 774–805.

[124] Hans-Georg Müller and Fang Yao. "Functional additive models". In: *Journal of the American Statistical Association* 103.484 (2008), pp. 1534–1544.

[125] Yu. Nesterov. "Gradient methods for minimizing composite functions". In: *Mathematical Programming* 140.1 (2012), 125–161.

[126] Whitney Newey and James L. Powell. "Asymmetric Least Squares Estimation and Testing". In: *Econometrica* 55.4 (1987), pp. 819–47.

[127] Alvaro Nosedal-Sanchez, Curtis B. Storlie, Thomas C.M. Lee, and Ronald Christensen. "Reproducing Kernel Hilbert Spaces for Penalized Regression: A Tutorial". In: *The American Statistician* 66.1 (2012), pp. 50–60.

[128] Alvaro Nosedal-Sanchez, Curtis B Storlie, Thomas CM Lee, and Ronald Christensen. "Reproducing kernel Hilbert spaces for penalized regression: A tutorial". In: *The American Statistician* 66.1 (2012), pp. 50–60.

[129] Erich Novak, Mario Ullrich, Henryk Woźniakowski, and Shun Zhang. "Reproducing kernels of Sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability". In: *Analysis and Applications* 16.05 (2018), pp. 693–715.

[130] R. Todd Ogden, Carl E. Miller, Kunio Takezawa, and Seishi Ninomiya. "Functional regression in crop lodging assessment with digital images". In: *Journal of Agricultural, Biological, and Environmental Statistics* 7 (2002), pp. 389–402.

[131] Mark J. L. Orr. *Introduction to radial basis function networks*. Tech. rep. Center for Cognitive Science, University of Edinburgh, 1996.

[132] Cardot Hervé; Ferraty Frédéric; Sarda Pascal. "Functional linear model". In: *Statistics & Probability Letters* 45.1 (1999), pp. 11–22.

[133] Matthew Pietrosanu, Jueyu Gao, Linglong Kong, Bei Jiang, and Di Niu. "Advanced algorithms for penalized quantile and composite quantile regression". In: *Computational Statistics* 36 (2020), pp. 333–346.

[134] Matthew Pietrosanu, Haoxu Shu, Bei Jiang, Linglong Kong, Giseon Heo, Qianchuan He, John Gilmore, and Hongtu Zhu. "Estimation for the bivariate quantile varying coefficient model with application to diffusion tensor imaging data analysis". In: *Biostatistics* 24.2 (2023), pp. 465–480.

[135] David Pollard. *Convergence of Stochastic Processes*. New York: Springer, 2012.

[136] Ross L Prentice. "A case-cohort design for epidemiologic cohort studies and disease prevention trials". In: *Biometrika* 73.1 (1986), pp. 1–11.

[137] Ross L Prentice and Ronald Pyke. "Logistic disease incidence models and case-control studies". In: *Biometrika* 66.3 (1979), pp. 403–411.

[138] Chongzhi Di; Ciprian M. Crainiceanu; Brian Caffo; Naresh M. Punjabi. "Multilevel functional principal component analysis". In: *The annals of applied statistics* 3.1 (2009), pp. 458–488.

[139] Xingye Qiao and Yufeng Liu. "Adaptive weighted learning for unbalanced multicategory classification". In: *Biometrics* 65.1 (2009), pp. 159–168.

[140] Jing Qin. *Biased Sampling, Over-identified Parameter Problems and Beyond*. Vol. 5. Springer, 2017.

[141] Simeng Qu, Jane-Ling Wang, and Xiao Wang. "Optimal estimation for the functional Cox model". In: *The Annals of Statistics* 44.4 (Aug. 2016), pp. 1708–1738.

[142] James Ramsay and CJ Dalzell. "Some tools for functional data analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 539–561.

[143] James Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

[144] Timothy W Randolph, Jaroslaw Harezlak, and Ziding Feng. "Structured penalties for functional linear models—partially empirical eigenvectors for regression". In: *Electronic journal of statistics* 6 (2012), p. 323.

[145] Sarah J Ratcliffe, Gillian Z Heller, and Leo R Leader. "Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression". In: *Statistics in medicine* 21.8 (2002), pp. 1115–1127.

[146] Sarah J Ratcliffe, Leo R Leader, and Gillian Z Heller. "Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression". In: *Statistics in Medicine* 21.8 (2002), pp. 1103–1114.

[147] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Functional Analysis*. Academic Press, 1972.

[148] Philip T. Reiss, Lei Huang, and Maarten Mennes. "Fast function-on-scalar regression with penalized basis expansions." In: *The international journal of biostatistics* 6.1 (2010), 28–NA.

[149] Philip T. Reiss and R. Todd Ogden. "Functional Principal Component Regression and Functional Partial Least Squares". In: *Journal of the American Statistical Association* 102.479 (2007), pp. 984–996.

[150] Michael Renardy and Robert C Rogers. *An Introduction to Partial Differential Equations*. Vol. 13. Springer Science & Business Media, 2006.

[151] Klaus Ritter, Grzegorz W Wasilkowski, and Henryk Wozniakowski. "Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions". In: *The Annals of Applied Probability* 5.2 (1995), pp. 518–540.

[152] José Luis Rojo-Álvarez, Manel Martínez-Ramón, Jordi Muñoz Marí, and Gustavo Camps-Valls. *Digital Signal Processing with Kernel Methods*. Wiley Online Library, 2018.

[153] Cynthia Rudin. *Prediction: Machine Learning and Statistics (MIT 15.097), Lecture on Kernels*. Tech. rep. Massachusetts Institute of Technology, 2012.

[154] Matthias Rupp. "Machine learning for quantum mechanics in a nutshell". In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1058–1073.

[155] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric Regression*. 12. Cambridge university press, 2003.

[156] Rowayda A Sadek. "An improved MRI segmentation for atrophy assessment". In: *International Journal of Computer Science Issues (IJCSI)* 9.3 (2012), p. 569.

[157] Rowayda A Sadek. "Regional atrophy analysis of MRI for early detection of Alzheimer's disease". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6.1 (2013), pp. 49–58.

[158] Peijun Sang, Adam B Kashlak, and Linglong Kong. "A reproducing kernel Hilbert space framework for functional classification". In: *Journal of Computational and Graphical Statistics* 32.3 (2023), pp. 1000–1008.

[159] Sabine K. Schnabel and Paul H.C. Eilers. "Optimal expectile smoothing". In: *Computational Statistics & Data Analysis* 53.12 (2009), pp. 4168–4177.

[160] Alastair J Scott and Chris J Wild. "Fitting logistic models under case-control or choice based sampling". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 48.2 (1986), pp. 170–182.

[161] Alastair J Scott and Chris J Wild. "Fitting regression models to case-control data by maximum likelihood". In: *Biometrika* 84.1 (1997), pp. 57–71.

[162] David W Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 1992.

[163] Zuofeng Shang and Guang Cheng. "Local and global asymptotic inference in smoothing spline models". In: *The Annals of Statistics* 41.5 (2013), pp. 2608–2638.

[164] Zuofeng Shang and Guang Cheng. "Nonparametric inference in generalized functional linear models". In: *The Annals of Statistics* 43.4 (2015), pp. 1742–1773.

[165] Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. "On ψ-Learning". In: *Journal of the American Statistical Association* 98.463 (2003), 724–734.

[166] Xiaotong Shen and Wing Hung Wong. "Convergence rate of sieve estimates". In: *The Annals of Statistics* 22.2 (1994), pp. 580–615.

[167] Yu Shen, Jing Ning, and Jing Qin. "Analyzing length-biased data with semiparametric transformation and accelerated failure time models". In: *Journal of the American Statistical Association* 104.487 (2009), pp. 1192–1202.

[168] Hyejin Shin and Seokho Lee. "An RKHS approach to robust functional linear regression". In: *Statistica Sinica* (2016), pp. 255–272.

[169] Bruce J Swihart, Jeff Goldsmith, and Ciprian M. Crainiceanu. "Restricted likelihood ratio tests for functional effects in the functional linear model". In: *Technometrics* 56.4 (2014), pp. 483–493.

[170] Qingguo Tang, Linglong Kong, David Ruppert, and Rohana J Karunamuni. "Partial functional partially linear single-index models". In: *Statistica Sinica* 31.1 (2021), pp. 107–133.

[171] Ryan Tibshirani. *Nonparametric Regression: Splines and RKHS Methods*. 2023.

[172] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer Science & Business Media, 2008.

[173] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. New York: Springer Verlag, 1996.

[174] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 1995.

[175] Grace Wahba. *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1990.

[176] Linda Schulze Waltrup, Fabian Sobotka, Thomas Kneib, and Göran Kauermann. "Expectile and quantile regression—David and Goliath?" In: *Statistical Modelling* 15.5 (2015), pp. 433–456.

[177] Boxiang Wang and Hui Zou. "Another look at distance-weighted discrimination". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 80.1 (2018), pp. 177–198.

[178] HaiYing Wang. "More Efficient Estimation for Logistic Regression with Optimal Subsamples". In: *Journal of Machine Learning Research* 20.132 (2019), pp. 1–59.

[179] Haiying Wang and Yanyuan Ma. "Optimal subsampling for quantile regression in big data". In: *Biometrika* 108.1 (2021), pp. 99–112.

[180] HaiYing Wang, Min Yang, and John Stufken. "Information-based optimal subdata selection for big data linear regression". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 393–405.

[181] HaiYing Wang, Rong Zhu, and Ping Ma. "Optimal subsampling for large sample logistic regression". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 829–844.

[182] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. "Functional data analysis". In: *Annual Review of Statistics and Its Application* 3.1 (2016), pp. 257–295.

[183] Xiao Wang, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. "Generalized scalar-on-image regression models via total variation". In: *Journal of the American Statistical Association* 112.519 (2017), pp. 1156–1168.

[184] Xiaohui Wang, Shubhankar Ray, and Bani K Mallick. "Bayesian curve classification using wavelets". In: *Journal of the American Statistical Association* 102.479 (2007), pp. 962–973.

[185] Zhaoran Wang, Han Liu, and Tong Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems". In: *The Annals of Statistics* 42.6 (2014), p. 2164.

[186] Christopher Williams and Matthias Seeger. "The effect of the input density distribution on kernel-based classifiers". In: *Proceedings of the 17th international conference on machine learning*. 2000.

[187] Simon N Wood, Zheyuan Li, Gavin Shaddick, and Nicole H Augustin. "Generalized additive models for gigadata: modeling the UK black smoke network daily data". In: *Journal of the American Statistical Association* 112.519 (2017), pp. 1199–1210.

[188] Yichao Wu and Yufeng Liu. "Robust Truncated Hinge Loss Support Vector Machines". In: *Journal of the American Statistical Association* 102.479 (2007), 974–983.

[189] Dong Xiang and Grace Wahba. "A generalized approximate cross validation for smoothing splines with non-Gaussian data". In: *Statistica Sinica* 6 (1996), pp. 675–692.

[190] Gongjun Xu, Tony Sit, Lan Wang, and Chiung-Yu Huang. "Estimation and inference of quantile regression for survival data under biased sampling". In: *Journal of the American Statistical Association* 112.520 (2017), pp. 1571–1586.

[191] Kaijie Xue, Jin Yang, and Fang Yao. "Optimal Linear Discriminant Analysis for High-Dimensional Functional Data". In: *Journal of the American Statistical Association* (2023), pp. 1–10.

[192] Hongmei Lin Yan Zhou Weiping Zhang and Heng Lian. "Partially linear functional quantile regression in a reproducing kernel Hilbert space". In: *Journal of Nonparametric Statistics* 34.4 (2022), pp. 789–803.

[193] Wen-Hsi Yang, Christopher K Wikle, Scott H Holan, and Mark L Wildhaber. "Ecological prediction with nonlinear multivariate time-frequency functional data models". In: *Journal of agricultural, biological, and environmental statistics* 18 (2013), pp. 450–474.

[194] Fang Yao and Hans-Georg Müller. "Functional quadratic regression". In: *Biometrika* 97.1 (2010), pp. 49–64.

[195] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional data analysis for sparse longitudinal data". In: *Journal of the American statistical association* 100.470 (2005), pp. 577–590.

[196] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional linear regression analysis for longitudinal data". In: *The Annals of Statistics* 33.6 (2005), pp. 2873–2903.

[197] Dengdeng Yu, Li Zhang, Ivan Mizera, Bei Jiang, and Linglong Kong. "Sparse wavelet estimation in quantile regression with multiple functional predictors". In: *Computational Statistics & Data Analysis* 136 (2019), pp. 12–29.

[198] Ming Yuan and T Tony Cai. "A reproducing kernel Hilbert space approach to functional linear regression". In: *The Annals of Statistics* 38.6 (2010), pp. 3412–3444.

[199] Donglin Zeng and D. Y. Lin. "Efficient estimation of semiparametric transformation models for two-phase cohort studies". In: *Journal of the American Statistical Association* 109.505 (2014), pp. 371–383.

[200] Daowen Zhang, Xihong Lin, and MaryFran Sowers. "Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome." In: *Biometrics* 63.2 (2007), pp. 351–362.

[201] Hongxiao Zhu; Fang Yao; Hao Helen Zhang. "Structured functional additive regression in reproducing kernel Hilbert spaces". In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 76.3 (2013), pp. 581–603.

[202] Yingying Zhang, Yan-Yong Zhao, and Heng Lian. "Statistical rates of convergence for functional partially linear support vector machines for classifications". In: *Journal of Machine Learning Research* 23 (2022), pp. 1–24.

[203] Yihong Zhao, R. Todd Ogden, and Philip T. Reiss. "Wavelet-based LASSO in functional linear regression". In: *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 21.3 (2012), pp. 600–617.

[204] Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". In: *Journal of the American Statistical Association* 107.499 (2012), 1106–1118.

[205] Hongxiao Zhu, Marina Vannucci, and Dennis D. Cox. "A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors". In: *Biometrics* 66.2 (2009), pp. 463–473.