Supervised Basis Functions Applied to Functional Regression and Classification

by

Zhiyang Zhou

M.Sc., Nankai University, 2012
B.Sc., Beijing Normal University, 2009

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

in the
Department of Statistics & Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2020

Approval

Name:	**Zhiyang Zhou**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	**Supervised Basis Functions Applied to Functional Regression and Classification**
Examining Committee:	**Chair:** Jinko Graham Professor
	Richard A. Lockhart Senior Supervisor Professor
	Derek Bingham Supervisor Professor
	Jiguo Cao Internal Examiner Associate Professor
	Nancy E. Heckman External Examiner Professor Department of Statistics University of British Columbia
Date Defended:	**July 29, 2020**

Abstract

In fitting functional linear models, including scalar-on-function regression (SoFR) and function-on-function regression (FoFR), the intrinsically infinite dimension of the problem often demands an limitation to a subspace spanned by a finite number of basis functions. In this sense, the choice and construction of basis functions matters. We discuss herein certain supervised choices of basis functions for regression/classification with densely/sparsely observed curves, and give both numerical and theoretical perspectives.

For SoFR, the functional principal component (FPC) regression may fail to provide good estimation or prediction if the response is highly correlated with some excluded FPCs. This is not rare since the construction of FPCs never involves the response. We hence develop regression on functional continuum (FC) basis functions whose framework includes, as special cases, both FPCs and functional partial least squares (FPLS) basis functions.

Aiming at the binary classification of functional data, we then propose the continuum centroid classifier (CCC) built upon projections of functional data onto the direction parallel to FC regression coefficient. One of the two subtypes of CCC (asymptotically) enjoys no misclassification.

Implementation of FPLS traditionally demands that each predictor curve be recorded as densely as possible over the entire time span. This prerequisite is sometimes violated by, e.g., longitudinal studies and missing data problems. We accommodate FPLS for SoFR to scenarios where curves are sparsely observed. We establish the consistency of proposed estimators and give confidence intervals for responses.

FPLS is widely used to fit FoFR. Its implementation is far from unique but typically involves iterative eigen decomposition. We introduce an new route for FoFR based upon Krylov subspaces. The method can be expressed in two equivalent forms: one of them is non-iterative with explicit forms of estimators and predictions, facilitating the theoretical derivation; the other one stabilizes numerical outputs. Our route turns out to be less time-consuming than other methods with competitive accuracy.

Keywords: Functional continuum regression; function-on-function regression; Krylov subspace; functional partial least squares; functional principal component; scalar-on-function regression

Dedication

To my family.

Acknowledgements

I eventually arrive at this ending phase of my journey to the highest degree in Statistics (though it takes a little bit longer than my initial expectation). I could not have made this far without the help and support from my family, professors, and friends.

My family is unconditionally and persistently supportive for all my feasible ideas. Its infinite love and endless encouragement lit up my pathway ahead, especially during the most frustrating period that I experienced. A particular family member was a female red-eared slider turtle (*Trachemys scripta elegans*) who had accompanied me for over twenty-three years but passed away a few days ago. May she rest in peace!

My superb supervisor, Professor Richard A. Lockhart, is gifted in mathematical theories and is versatile in various branches of statistics. He is enthusiastic in acquiring and absorbing up-to-date knowledge. He is humble and easy-going and is always willing to offer assistance to everyone else. He taught me how to write with clarity and how to give suggestions with politeness. He backed me up as much as possible, not only financially but also academically; I was spoiled so much that I could go both far and deep on any topics interesting me. Once I got stuck somewhere in the exploration, a discussion with him could dispel the mist and pull me out immediately. This journey has been so intriguing and rewarding because of him!

Professor Derek Bingham drew my attention to applications to scientific discoveries. Professors Jiguo Cao and Nancy E. Heckman offered constructive suggestions on technical issues. I appreciate their willingness of joining in the examining committee.

Fortunately I had opportunities to work as a teaching assistant instructed by, respectively, Professors Joan Hu, Thomas M. Loughin, Brad McNeney, Tim Swartz, Steve K. Thompson, and Liangliang Wang. They might differ in their pedagogical styles but were equally devoted in sharing knowledge with students. Special thanks go to Professors Joan Hu, Thomas M. Loughin, and Peijun (Perry) Sang for their endorsement in my job hunting.

This journey would be lonely without the accompany of my cohort including but not limited to Anqi (Angela) Chen, Siyuan (Jensen) Chen, Tian Chen, JinCheol Choi, Jingxue (Grace) Feng, Shufei Ge, Michael Grosskopf, Tianyu Guan, Lulu Guo, Botao (Bobby) Han, Grace Hsu, Boyi Hu, Haiyang (Jason) Jiang, Dongdong Li, Tian Li, Chuyuan (Cherlane) Lin, Luyao Lin, Yan (Lillian) Lin, Dongmeng Liu, Sichen (Coco) Liu, Yanjun Liu, Yue Ma, Jacob Mortensen, Payman Nickchi, Yunlong (Ben) Nie, Derek Qiu, Pulindu Ratnasekera, Haoyao Ruan, Will Ruth, Nate Sandholtz, Chenlu Shi, Cheng-Yu Sun, Meng (Maggie) Sun,

Trevor Thomson, Jiahao Tian, Yunwei Tu, Haixu (Alex) Wang, Jie (John) Wang, Shijia Wang, Yueren Wang, Qi (Emma) Wen, Jiying Wen, Sidi Wu, Yifan (Lucas) Wu, Yi Xiong, Mengxiao Xu, Yuping Yang, Faezeh Yazdi, Lu Yi, Ying (Daisy) Yu, Jiarui (Erin) Zhang, Yang (Maple) Bai, Ying Zhang, and Haoxuan (Charlie) Zhou. I wish them all the joy and success wherever they go.

Responsible staff members, Charlene Bradbury, Kelly Jay, Sadika Jungic, Carlye Vroom, and Jay Young, saved me from massive paper work. Their contribution, combined with efforts from all faculty members and students, created and maintained the delightful and inclusive environment in our department.

Last but not least, I would like to express my gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for the financial support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the development of technology, the demand for functional data analysis (FDA) is increasing. It is frequent to encounter data that are recorded continuously within a non-degenerate and compact domain of "time", e.g., $[0, 1]$. Formally, consider an L^2 process X whose argument takes values from a "time" domain $\mathbb{T}_X$. Write $L^2(\mathbb{T}_X)$ (resp. $L^2(\mathbb{T}_X^2)$) as the L^2-space on $\mathbb{T}_X$ (resp. $\mathbb{T}_X^2$) with respect to (w.r.t.) the Lebesgue measure. The auto-covariance operator of X, say $\mathcal{V}_X : L^2(\mathbb{T}_X) \to L^2(\mathbb{T}_X)$, is then given, for all $f \in L^2(\mathbb{T}_X)$, by

$$\mathcal{V}_X(f)(\cdot) = \int_{\mathbb{T}_X} f(s) v_X(s, \cdot) \mathrm{d}s, \tag{1.1}$$

where

$$v_X = v_X(s, t) = \mathrm{cov}\{X(s), X(t)\}. \tag{1.2}$$

As a standard assumption in FDA research, $v_X \in L^2(\mathbb{T}_X^2)$ implies countably many nonnegative eigenvalues of $\mathcal{V}_X$ sorted in a decreasing order, say $\lambda_{1,X} \geq \lambda_{2,X} \geq \cdots$, and corresponding eigenfunctions $\phi_{1,X}, \phi_{2,X}, \ldots$. We further require $\sum_{j=1}^{\infty} \lambda_{j,X} < \infty$ and abuse the notation $\|\cdot\|_2$ for the L^2-norm of each L^2-space involved.

The (linear) scalar-on-function regression (SoFR) is an elementary model in FDA, bridging scalar response Y to functional predictor X. To be specific,

$$Y = \mu_Y + \int_{\mathbb{T}_X} \beta(X - \mu_X) + \varepsilon, \tag{1.3}$$

where μ_X (resp. μ_Y) is the expectation of X (resp. Y) and white noise ε has mean zero and variance σ_ε^2. The notation $\int_{\mathbb{T}_X} f$ is short for $\int_{\mathbb{T}_X} f(t)\mathrm{d}t$. To assure identifiability of β, assume that β belongs to $\mathrm{span}(\phi_1, \phi_2, \ldots)$, where $\mathrm{span}(\cdot)$ is the linear space spanned by functions in the parentheses. People have applied SoFR to several domains including chemometrics (e.g., predicting scalars according to near infra-red (NIR) spectroscopy [40]), food manufacturing (e.g., controlling biscuit quality [1]), geoscience (e.g., investigating climate data from the United States [7]), medical imaging (e.g., analyzing diffusion tensor imaging (DTI) tractog-

raphy [37]) and many others. In practice, the interpretability of β is the largest advantage of SoFR over competitors; refer to [79].

The infinite-dimensional structure of L^2-spaces makes data analysis challenging: the dimension of the parameter space exceeds the number of observed subjects, and hence dimension-reduction techniques are indispensable in model fitting. To estimate β and to predict the conditional expectation

$$\eta(X^*) = \mathrm{E}(Y \mid X = X^*) = \mu_Y + \int_{\mathbb{T}_X} \beta(X^* - \mu_X) \tag{1.4}$$

for X^* distributed as X, the standard approach is to express β in terms of a linear combination of functions $w_1, \ldots, w_p$ truncated from countably many basis functions $w_1, w_2, \ldots$ in $L^2(\mathbb{T}_X)$. This inspires people to approximate β and $\eta(X^*)$, respectively, by:

$$\beta_p = \underset{\theta \in \mathrm{span}(w_1,\ldots,w_p)}{\arg\min} \mathrm{E}\left\{ Y - \mu_Y - \int_{\mathbb{T}_X} \theta(X - \mu_X) \right\}^2, \tag{1.5}$$

$$\eta_p(X^*) = \mu_Y + \int_{\mathbb{T}_X} \beta_p(X^* - \mu_X), \tag{1.6}$$

where β_p is the slope of the best approximation (within $\mathrm{span}(w_1, \ldots, w_p)$ and in the L^2 sense) to Y by a linear function of X; for $p = 0$, we put $\beta_0 = 0$ for later convenience, completing the definition.

Suppose n two-tuples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently and identically distributed (iid) as (X, Y). It is understood that the trajectories in the observed data will have no analytical expression and hence there is no way to compute corresponding integrals exactly. Nevertheless, numerical techniques are available, e.g., quadrature rules, as long as the set of points at which each curve is observed is sufficiently dense. Errors would be introduced in these approximations though they would be bounded (see, e.g., [91] bounding errors associated with the trapezoidal rule). Such bounds depend upon the smoothness of underlying trajectories. Accordingly, interpolations (e.g., various splines) are often involved; refer to, e.g., [101] for theoretical results on certain penalized splines. Especially in Chapters 2, 3 and 5, we assume curves are observed densely enough and, for convenience, abuse integral signs for corresponding empirical approximations throughout the entire thesis. Then, plug-in estimates for (1.5) and (1.6) are respectively expressed as

$$\hat{\beta}_p = \underset{\theta \in \mathrm{span}(\hat{w}_1,\ldots,\hat{w}_p)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - \int_{\mathbb{T}_X} \theta(X_i - \bar{X}) \right\}^2, \tag{1.7}$$

$$\hat{\eta}_p(X^*) = \bar{Y} + \int_{\mathbb{T}_X} \hat{\beta}_p(X^* - \bar{X}), \tag{1.8}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Obviously, these estimates vary with the choice of w_j as well as the quality of $\hat{w}_j$. Although this framework is compatible with a

basis independent of the data (e.g., polynomial basis, Fourier basis, wavelets, splines, etc.), it is more reasonable to force it to adapt to data (e.g., the functional principal component (FPC) and functional partial least squares (FPLS) bases).

When the response Y is changed from scalar to an L^2 process defined on another "time" domain $\mathbb{T}_Y$, one may resort to (linear) function-on-function regression (FoFR, first proposed by [75]):

$$Y(t) = \mu_Y(t) + \int_{\mathbb{T}_X} \beta(s,t)\{X(s) - \mu_X(s)\}\mathrm{d}s + \varepsilon(t)$$

where coefficient β is now bivariate, defined on $\mathbb{T}_X \times \mathbb{T}_Y$ and μ_Y becomes functional. We assume the zero-mean Gaussian process ε is uncorrelated with X (i.e., $\mathrm{E}\{X(s), \varepsilon(t)\} = 0$ for all $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$) with a covariance function v_ε which is continuous on $\mathbb{T}_Y^2$. We rewrite FoFR in the form

$$Y(t) = \mu_Y(t) + \mathcal{L}_X(\beta)(t) + \varepsilon(t), \tag{1.9}$$

defining a random integral operator $\mathcal{L}_X : L^2(\mathbb{T}_X \times \mathbb{T}_Y) \to L^2(\mathbb{T}_Y)$ such that, for each $f \in L^2(\mathbb{T}_X \times \mathbb{T}_Y)$,

$$\mathcal{L}_X(f)(\cdot) = \int_{\mathbb{T}_X} \{X(s) - \mu_X(s)\} f(s,\cdot)\mathrm{d}s.$$

Assuming (C.A.4.1) in Section A.4, [44, Theorem 2.3] shows that the true parameter value β can be defined uniquely through least squares, viz. $\beta = \arg\min_{\theta \in L^2(\mathbb{T}_X \times \mathbb{T}_Y)} \mathrm{E}\,\|Y - \mu_Y - \mathcal{L}_X(\theta)\|_2^2$; in detail, for each $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$, we have

$$\beta(s,t) = \sum_{i,j=1}^{\infty} \frac{\mathrm{cov}(\int_{\mathbb{T}_X} X\phi_{i,X}, \int_{\mathbb{T}_Y} Y\phi_{j,Y})}{\lambda_{i,X}} \phi_{i,X}(s)\phi_{j,Y}(t), \tag{1.10}$$

where $\lambda_{j,Y}$ (resp. $\phi_{j,Y}$) is the jth top eigenvalue (resp. eigenfunction) of $\mathcal{V}_Y$ (defined in complete analogy to $\mathcal{V}_X$ at (1.1)). Estimation and prediction for FoFR can still be implemented through projection. Indeed, FPC regression (FPCR) for FoFR approximates β by its orthogonal projection on $\mathrm{span}\{f_{ij} \in L^2(\mathbb{I}_X \times \mathbb{I}_Y) \mid f_{ij}(s,t) = \phi_{i,X}(s)\phi_{j,Y}(t), 1 \le i \le p, 1 \le j \le q\}$, or equivalently, drops the tail of the series on the farthest right-hand side of (1.10).

The main body of this thesis comprises our previous works [109, 110, 111, 112] which are presented independently from each other. Chapter 2 offers a supervised option for w_j for SoFR which results in better accuracy in both estimation and prediction. Applying proposals in Chapter 2 to binary classification, Chapter 3 reveals the possible improvement of error rate associated with this strategy. Chapter 4 implements FPLS in the more challenging context of sparsely observed functional data with measurement errors. Fitting FoFR, a new route for FPLS is provided by Chapter 5. Possible further work is described in Chapter 6. For the conciseness, technical details are consigned to appendices.

Chapter 2

Functional continuum regression

2.1 Introduction

2.1.1 Functional principal component and functional partial least squares bases

Recall the framework detailed in Chapter 1. Among all the bases exploited in FDA, the most prevailing one is the FPC basis i.e., $\{\phi_{1,X}, \phi_{2,X}, \ldots\}$, where $\phi_{j,X}$ is the jth eigenfunction of $\mathcal{V}_X$ (1.1), or equivalently, given $\phi_{k,X}$ for all $k \in \{1, \ldots, j-1\}$, one has

$$\phi_j = \underset{w:\|w\|_2=1}{\arg\max} \int_{\mathbb{T}_X} w\mathcal{V}_X(w) \tag{2.1}$$

subject to

$$\int_{\mathbb{T}_X} w\phi_{1,X} = \cdots = \int_{\mathbb{T}_X} w\phi_{j-1,X} = 0.$$

Function $\phi_{j,X}$ (2.1) is estimated by $\hat{\phi}_j$, the jth eigenfunction of operator $\widehat{\mathcal{V}}_X$ defined by substituting

$$\hat{v}_X(s,t) = \widehat{\mathrm{cov}}\{X(s), X(t)\} = \frac{1}{n}\sum_{i=1}^{n}\{X_i(s) - \bar{X}(s)\}\{X_i(t) - \bar{X}(t)\}$$

for $v_X(s,t)$ (1.2) in the definition of $\mathcal{V}_X$ (1.1), i.e., for all $f \in L^2(\mathbb{T}_X)$,

$$\widehat{\mathcal{V}}_X(f)(\cdot) = \int_{\mathbb{T}_X} f(s)\hat{v}_X(s,\cdot)\mathrm{d}s. \tag{2.2}$$

During the past few decades, extensive work has focused on FPC; more details can be found in a number of monographs (e.g., [48, 76]) and review papers (e.g., [31, 97]). As defined in (2.1), the construction of the functional principal component basis is "unsupervised"; this basis does not involve the response Y; the first few elements of this basis seek to explain as much of the variation of X as possible, whereas they are not necessarily important in

representing β. That is, it is possible for one or more members in the abandoned part $\{\phi_{p+1}, \phi_{p+2}, \ldots\}$ to be highly correlated with the response.

Some efforts have already been made to target this well-known defect, including [73], in which (multivariate) partial least squares (PLS) is extended to the functional domain, i.e., FPLS. This technique relies on a basis which is defined in a sequential manner. Namely, given $w_{k,\mathrm{FPLS}}$ for all $k \in \{1, \ldots, j-1\}$,

$$w_{j,\mathrm{FPLS}} = \underset{w:\|w\|_2=1}{\arg\max}\ \mathrm{cov}^2\left\{Y - \eta_{j-1,\mathrm{FPLS}}(X), \int_{\mathbb{T}_X} Xw\right\} \tag{2.3}$$

subject to

$$\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{1,\mathrm{FPLS}}) = \cdots = \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j-1,\mathrm{FPLS}}) = 0.$$

The empirical counterpart of this parameter is the estimator

$$\hat{w}_{j,\mathrm{FPLS}} = \underset{w:\|w\|_2=1}{\arg\max} \left[\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\eta}_{j-1,\mathrm{FPLS}}(X_i)\}\int_{\mathbb{T}_X} w(X_i - \bar{X})\right]^2$$

subject to

$$\int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_X(\hat{w}_{1,\mathrm{FPLS}}) = \cdots = \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_X(\hat{w}_{j-1,\mathrm{FPLS}}) = 0,$$

where $\eta_{j-1,\mathrm{FPLS}}$ and $\hat{\eta}_{j-1,\mathrm{FPLS}}$ are respective counterparts of (1.6) and (1.8).

FPLS has since been investigated and developed, e.g., in [2, 28, 78]. PLS and its derivatives are referred to as "fully supervised" and may suffer the "double-dipping" problem: they employ the covariance between Y and X both for the construction of basis functions and for further prediction. The resulting findings are possibly vulnerable and sensitive to small signals; see [52]. By contrast, [69] suggested a linear combination of FPC and FPLS bases; their proposal lies between unsupervised and fully supervised techniques. Different from these authors, we borrow the idea of (multivariate) continuum regression [88] and extend it to learning for functional data.

2.1.2 Continuum regression

In the context of multivariate analysis, continuum regression works for the linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}$ with response $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ and design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$; we assume these are both column-mean-centered. The method projects $\boldsymbol{y}$ onto the linear space spanned by mutually orthogonal vectors $\mathbf{X}\boldsymbol{w}_{1,\alpha}, \ldots, \mathbf{X}\boldsymbol{w}_{p,\alpha}$, after successively computing the d-vectors given by

$$\boldsymbol{w}_{j,\alpha} = \underset{\boldsymbol{w}:\boldsymbol{w}^\top\boldsymbol{w}=1}{\arg\max}\ (\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})^2(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w})^{\alpha/(1-\alpha)-1}, \tag{2.4}$$

with the constraint that $\boldsymbol{w}_{k,\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} = 0$ for all $k \in \{1, \ldots, j-1\}$. Here $\alpha \in [0, 1)$ and p ($\leq d$) are both to be tuned. The most appealing property of continuum regression, as

proved by [88], is that the framework of continuum regression encompasses ordinary least squares (OLS) ($\alpha = 0$), PLS ($\alpha = 1/2$), and (multivariate) principal component regression ($\alpha \to 1$). The tuning parameter α actually controls the degree of supervision (i.e., the extent of involvement of the response), giving more flexibility to the resulting estimator and prediction.

There have been some further developments of continuum regression. [90] connected it to the ridge regression. [13] revealed the analytical form of (2.4). [57] combined continuum regression with the kernel learning to accommodate nonlinear regression. [17] proved the possible inconsistency of estimators produced by continuum regression, while [18] showed the consistency of continuum regression in estimating the central (dimensional-reduction) subspace defined by [20] and [21, pp. 105].

The remainder of this chapter develops our functional approach. Section 2.2 introduces functional continuum (FC) regression and some special cases. Our consistency results are presented in Section 2.3, based on which Section 2.4 derives an effective algorithm. Empirical evidence appears in Section 2.5, where our method is compared with existing ones in terms of both estimation and prediction. Section 2.6 discusses the pros and cons of FC regression as well as possible future work. For the sake of brevity, technical details are relegated to Appendix A.1.

2.2 Functional continuum regression

2.2.1 Functional continuum basis

We begin by defining the (truncated) FC basis denoted by $\{w_{1,\alpha}, \ldots, w_{p,\alpha}\}$. For a predetermined $\alpha \in [0, 1)$, we construct the basis in a sequential way. Given $w_{1,\alpha}, \ldots, w_{j-1,\alpha}$, define

$$w_{j,\alpha} = \underset{w:\|w\|_2=1}{\arg\max}\ T_\alpha(w) \tag{2.5}$$

subject to

$$\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{1,\alpha}) = \cdots = \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j-1,\alpha}) = 0, \tag{2.6}$$

where

$$T_\alpha = T_\alpha(w) = \left\{\int_{\mathbb{T}_X} w\mathcal{V}_X(w)\right\}^{\alpha/(1-\alpha)-1} \operatorname{cov}^2\left(Y, \int_{\mathbb{T}_X} Xw\right). \tag{2.7}$$

The optimization problem (2.5) constrained by (2.6) is exactly a functional counterpart of (2.4). By controlling α, the degree of supervision, one can recover some well-known special cases including FPC and FPLS; see Section 2.2.2. Analogous to (1.5) and (1.6) respectively,

we define the parameters,

$$\begin{aligned}\beta_{p,\alpha} &= \underset{\theta\in\mathrm{span}(w_{1,\alpha},\ldots,w_{p,\alpha})}{\arg\min} \mathrm{E}\left\{\int_{\mathbb{T}_X}(\beta-\theta)(X^*-\mu_X)\right\}^2 \\ &= \sum_{j=1}^{p}\left\{\int_{\mathbb{T}_X}\beta\mathcal{V}_X(w_{j,\alpha})\right\}\left\{\int_{\mathbb{T}_X} w_{j,\alpha}\mathcal{V}_X(w_{j,\alpha})\right\}^{-1/2} w_{j,\alpha}\end{aligned} \tag{2.8}$$

and the ideal predictor

$$\begin{aligned}\eta_{p,\alpha}(X^*) &= \mu_Y + \int_{\mathbb{T}_X}\beta_{p,\alpha}(X^*-\mu_X) \\ &= \mu_Y + \sum_{j=1}^{p}\left\{\int_{\mathbb{T}_X}\beta\mathcal{V}_X(w_{j,\alpha})\right\}\left\{\int_{\mathbb{T}_X} w_{j,\alpha}\mathcal{V}_X(w_{j,\alpha})\right\}^{-1/2}\int_{\mathbb{T}_X} w_{j,\alpha}(X^*-\mu_X);\end{aligned} \tag{2.9}$$

these give approximations to β in (1.3) and $\eta(X^*)$ in (1.4).

Having defined $w_{j,\alpha}$ in (2.5), we now give its empirical counterpart $\hat{w}_{j,\alpha}$ which is also defined recursively. Once the first $j-1$ empirical components are determined, the next $\hat{w}_{j,\alpha}$ is taken as the maximizer of the following optimization problem:

$$\begin{aligned}&\underset{w}{\text{maximize}} \quad \widehat{T}_\alpha(w) = \left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})\int_{\mathbb{T}_X} w(X_i-\bar{X})\right\}^2\left\{\int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_X(w)\right\}^{\alpha/(1-\alpha)-1} \\ &\text{subject to} \quad \|w\|_2 = 1 \quad \text{and} \quad \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_X(\hat{w}_{1,\alpha}) = \cdots = \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_X(\hat{w}_{j-1,\alpha}) = 0,\end{aligned} \tag{2.10}$$

where operator $\widehat{\mathcal{V}}_X$ is defined as in (2.2). Further, $\beta_{p,\alpha}$ from (2.8) and $\eta_{p,\alpha}(X^*)$ from (2.9) are respectively estimated by

$$\begin{aligned}\hat{\beta}_{p,\alpha} &= \underset{\theta\in\mathrm{span}(\hat{w}_{1,\alpha},\ldots,\hat{w}_{p,\alpha})}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\bar{Y}-\int_{\mathbb{T}_X}\theta(X_i-\bar{X})\right\}^2 \\ &= \sum_{j=1}^{p}\left\{\int_{\mathbb{T}_X}\beta\widehat{\mathcal{V}}_X(\hat{w}_{j,\alpha})\right\}\left\{\int_{\mathbb{T}_X}\hat{w}_{j,\alpha}\widehat{\mathcal{V}}_X(\hat{w}_{j,\alpha})\right\}^{-1/2}\hat{w}_{j,\alpha} \\ &= \sum_{j=1}^{p}\widehat{\mathrm{cov}}\left(Y,\int_{\mathbb{T}_X}X\hat{w}_{j,\alpha}\right)\widehat{\mathrm{var}}^{-1/2}\left(\int_{\mathbb{T}_X}X\hat{w}_{j,\alpha}\right)\hat{w}_{j,\alpha}\end{aligned} \tag{2.11}$$

and

$$\hat{\eta}_{p,\alpha}(X^*) = \bar{Y} + \int_{\mathbb{T}_X}\hat{\beta}_{p,\alpha}(X^*-\bar{X}). \tag{2.12}$$

Return to the definition of $w_{j,\alpha}$ in (2.5). Though it looks like a natural extension of (2.4), at least two concerns arise with the non-concavity of objective functions $T_\alpha(w)$ (2.7) and $\widehat{T}_\alpha(w)$ in (2.10) and the infinite dimension of $L^2(\mathbb{T}_X)$: one is the existence of $w_{j,\alpha}$ and $\hat{w}_{j,\alpha}$ which is not trivial at all since neither the unit sphere nor the unit ball in $L^2(\mathbb{T}_X)$ is

compact; the other is whether or not, for arbitrary $\alpha \in [0, 1)$, β can be fully expressed in terms of a linear combination of members of the FC basis $\{w_{1,\alpha}, w_{2,\alpha}, \ldots\}$.

Proposition 2.1. *Given $w_{1,\alpha}, \ldots, w_{j-1,\alpha}$, the objective function T_α (2.7), subject to conditions (2.6), has a maximizer. So does $\widehat{T}_\alpha$ in (2.10) with fixed $\hat{w}_{1,\alpha}, \ldots, \hat{w}_{j-1,\alpha}$.*

Proposition 2.2. *For arbitrary $\alpha \in [0, 1)$, β belongs to $\overline{\text{span}(w_{1,\alpha}, w_{2,\alpha}, \ldots)}$, the closure of $\text{span}(w_{1,\alpha}, w_{2,\alpha}, \ldots)$.*

2.2.2 Special cases

FC regression inherits the inclusion property of continuum regression; i.e., for certain α, FC regression reduces to some existing methods. First, as $\alpha \to 1$, the variance term $\int_{\mathbb{T}_X} w\mathcal{V}_X(w)$ dominates the objective function T_α (2.7) and the role of $\text{cov}(Y, \int_{\mathbb{T}_X} Xw)$ is negligible. We assert that, in this scenario, FC basis is identical to FPC basis.

Proposition 2.3. *If $\text{cov}(Y, \int_{\mathbb{T}_X} X\phi_{j,X}) \neq 0$ for all $j \in \{1, \ldots, p\}$, then, with fixed $j \in \{1, \ldots, p\}$, $\|w_{j,\alpha} - \phi_{j,X}\|_2 \to 0$ as $\alpha \to 1$.*

At the other extreme ($\alpha = 0$), note that

$$w_{1,0} = \underset{w:\|w\|=1}{\arg\max} \frac{\text{cov}^2(Y, \int_{\mathbb{T}_X} Xw)}{\int_{\mathbb{T}_X} w\mathcal{V}_X(w)} = \underset{w:\|w\|=1}{\arg\max} \frac{\text{cov}^2(Y, \int_{\mathbb{T}_X} Xw)}{\text{var}(Y)\,\text{var}(\int_{\mathbb{T}_X} Xw)}.$$

Geometrically, $w_{1,\alpha}$ maximizes the squared cosine of the angle between $\int_{\mathbb{T}_X} Xw$ and Y. Therefore, $\int_{\mathbb{T}_X} Xw_{1,0}$ is parallel to the orthogonal projection of Y onto X, meaning that $\text{cov}(Y, \int_{\mathbb{T}_X} Xw)$ must be zero for all w such that $\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{1,0}) = 0$. That is to say, the sequential construction terminates at $w_{1,\text{FC}}$ and no subsequent element exists. Obviously, in this situation, FC regression is equivalent to a functional version of OLS regression.

Another special case lies midway between these two extremes, i.e., $\alpha = 1/2$. Under constraints (2.6), we then have

$$\text{cov}\left\{Y - \eta_{j-1,1/2}(X), \int_{\mathbb{T}_X} Xw\right\} = \text{cov}\left(Y, \int_{\mathbb{T}_X} Xw\right).$$

One can see that this case is identical to functional PLS introduced in Section 2.1.1.

2.3 Theoretical properties

2.3.1 Equivalent forms of the functional continuum basis

Considering residuals of X and Y after the first $j-1$ steps, we merge the $j-1$ side-conditions (2.6) and objective T_α (2.7) together. This reformulation simplifies forthcoming proofs and facilitates the implementation in Section 2.4 as well.

Proposition 2.4. *Let $X^{(1,\alpha)} = X - \mu_X$ and $Y^{(1,\alpha)} = Y - \mu_Y$. For every integer $j \geq 2$, given $w_{k,\alpha}$ satisfying $\int_{\mathbb{T}_X} w_{k,\alpha}\mathcal{V}_X(w_{k,\alpha}) > 0$ for all $k \in \{1, \ldots, j-1\}$, write*

$$X^{(j,\alpha)} = X - \mu_X - \sum_{k=1}^{j-1} \int_{\mathbb{T}_X} w_{k,\alpha}(X - \mu_X) \left\{ \int_{\mathbb{T}_X} w_{k,\alpha}\mathcal{V}_X(w_{k,\alpha}) \right\}^{-1/2} \mathcal{V}_X(w_{k,\alpha})$$

and

$$Y^{(j,\alpha)} = Y - \eta_{j-1,\alpha}(X) = \int_{\mathbb{T}_X} \beta X^{(j,\alpha)}.$$

Then, $w_{j,\alpha}$ (2.5) can be found by maximizing $T^*_{j,\alpha}$ on the unit sphere, i.e.,

$$w_{j,\alpha} = \underset{w:\|w\|=1}{\arg\max}\ T^*_{j,\alpha}(w),$$

where

$$\begin{aligned} T^*_{j,\alpha}(w) &= \mathrm{cov}^2 \left\{ Y^{(j,\alpha)}, \int_{\mathbb{T}_X} X^{(j,\alpha)} w \right\} \left\{ \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(w) \right\}^{\alpha/(1-\alpha)-1} \\ &= \left\{ \int_{\mathbb{T}_X} \beta\mathcal{V}_{X^{(j,\alpha)}}(w) \right\}^2 \left\{ \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(w) \right\}^{\alpha/(1-\alpha)-1}. \end{aligned} \tag{2.13}$$

An empirical counterpart of Proposition 2.4 naturally follows.

Proposition 2.5. *Fix an integer $i \in \{1, \ldots, n\}$. Let $\widehat{X}_i^{(1,\alpha)} = X_i - \bar{X}$ and $\widehat{Y}_i^{(1,\alpha)} = Y_i - \bar{Y}$. For every integer $j \geq 2$, given $\hat{w}_{k,\alpha}$ with $\int_{\mathbb{T}_X} \hat{w}_{k,\alpha}\widehat{\mathcal{V}}_X(\hat{w}_{k,\alpha}) > 0$ for all $k \in \{1, \ldots, j-1\}$, write*

$$\widehat{X}_i^{(j,\alpha)} = X_i - \bar{X} - \sum_{k=1}^{j-1} \int_{\mathbb{T}_X} \hat{w}_{k,\alpha}(X_i - \bar{X}) \left\{ \int_{\mathbb{T}_X} \hat{w}_{k,\alpha}\widehat{\mathcal{V}}_X(\hat{w}_{k,\alpha}) \right\}^{-1/2} \widehat{\mathcal{V}}_X(\hat{w}_{k,\alpha})$$

and

$$\widehat{Y}_i^{(j,\alpha)} = Y_i - \hat{\eta}_{j-1,\alpha}(X_i) = \int_{\mathbb{T}_X} \beta\widehat{X}_i^{(j,\alpha)}.$$

Then,

$$\hat{w}_{j,\alpha} = \underset{w:\|w\|=1}{\arg\max}\ \widehat{T}^*_{j,\alpha}(w), \tag{2.14}$$

where

$$\begin{aligned} \widehat{T}^*_{j,\alpha}(w) &= \widehat{\mathrm{cov}}^2 \left\{ \widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} w \right\} \left\{ \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(w) \right\}^{\alpha/(1-\alpha)-1} \\ &= \left\{ \int_{\mathbb{T}_X} \beta\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(w) \right\}^2 \left\{ \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(w) \right\}^{\alpha/(1-\alpha)-1} \end{aligned} \tag{2.15}$$

with $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}} = \widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(s,t) = \sum_{i=1}^n \widehat{X}_i^{(j,\alpha)}(s)\widehat{X}_i^{(j,\alpha)}(t)/n$.

Previously, the FC basis has been defined as a set of maximizers of sequential optimization problems. Proposition 2.6 below derives an alternative but more explicit form of these desired solutions: they are constructed by adjusting the projection of function β on some directions.

Proposition 2.6. *Given $\alpha \in [0,1)$ and $w_{1,\alpha}, \ldots, w_{j-1,\alpha}$. Let $\lambda_{k,X^{(j,\alpha)}}$ denote the kth top eigenvalue of $\mathcal{V}_{X^{(j,\alpha)}}$ with corresponding eigenfunction $\phi_k^{(j,\alpha)}$. Suppose $\lambda_{1,X^{(j,\alpha)}}$ has multiplicity $m \geq 1$, i.e., $\lambda_{1,X^{(j,\alpha)}} = \cdots = \lambda_{m,X^{(j,\alpha)}} > \lambda_{m+1,X}^{(j,\alpha)}$. If $\mathcal{V}_{X^{(j,\alpha)}}(\beta)$ is not orthogonal to $\operatorname{span}\{\phi_{1,X}^{(j,\alpha)}, \ldots, \phi_{m,X}^{(j,\alpha)}\}$, then there exists $\delta^{(j,\alpha)} \in (-1,0) \cup (0,\infty)$ such that $w_{j,\alpha}$ is of unit L^2-norm and*

$$w_{j,\alpha} \propto \sum_{k=1}^{\infty} \frac{\lambda_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta \phi_k^{(j,\alpha)}}{\lambda_{k,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}} \phi_k^{(j,\alpha)},$$

where the three boundary values of $\delta^{(j,\alpha)}$, i.e., -1, 0 and ∞, correspond to FPC ($\delta^{(j,\alpha)} \to -1$), FPLS ($\delta^{(j,\alpha)} \to 0$) and functional OLS ($\delta^{(j,\alpha)} \to \infty$), respectively.

2.3.2 Consistency of the empirical functional continuum basis and corresponding estimators

We need one more condition, as follows.

(C.2.1) For each $j \in \{1, \ldots, p\}$, $T_{j,\alpha}^*(w)$ in (2.13) has a unique maximizer (up to sign change) on the unit sphere $\{w \in L^2(\mathbb{T}_X) : \|w\|_2 = 1\}$.

Our main result, Theorem 2.1, demonstrates the consistency of $\hat{w}_{j,\alpha}$ defined in (2.14), $\hat{\beta}_{p,\alpha}$ from (2.11) and $\hat{\eta}_{p,\alpha}(X^*)$ from (2.12) in the case of "fixed p and infinite n".

Theorem 2.1. *Fix $\alpha \in [0,1)$ and integer p. Under (C.2.1), we have, for all $j \in \{1, \ldots, p\}$, $\|\hat{w}_{j,\alpha} - w_{j,\alpha}\|_2 \to_p 0$ as $n \to \infty$. It follows that $\|\hat{\beta}_{p,\alpha} - \beta_{p,\alpha}\|_2$ and $|\hat{\eta}_{p,\alpha}(X^*) - \eta_{p,\alpha}(X^*)|$ both converge to zero in probability as $n \to \infty$, where X^* is a realization of X and independent from $X_1, \ldots, X_n$.*

Remark 2.1. We do not have to impose uniqueness on the maximizer of $\widehat{T}_{j,\alpha}^*(w)$ in (2.15); if $\arg\max_{\|w\|=1} \widehat{T}_{j,\alpha}^*(w)$ is not unique, the proof of Theorem 2.1 is still valid as long as the resulting $\hat{w}_{j,\alpha}$ is measurable. [51, Lemma 2] provided a route to construct such a measurable $\hat{w}_{j,\alpha}$.

2.4 Implementation

We understand that in practice each curve can only be observed at finitely many spots; that is why the integrals involved generally have to be approximated in numerically, e.g., by various finite sums. Alternatively, people may choose to recover (or pre-smooth) unknown curves through penalized splines [96, pp. 98]; when the observation time points are sufficiently dense in $\mathbb{T}_X$, the resulting curves are expected to be consistent approximations to

true underlying ones; see, e.g., [101]. These approximation procedures definitely affect the accuracy of $\hat{w}_{j,\alpha}$ (2.14) and $\hat{\beta}_{p,\alpha}$ (2.11), but corresponding discussions are out of the scope of this work. For convenience, we will keep the integral notations, even in the description of implementation.

It is feasible to duplicate the idea in [16, 57] to tackle the maximization problem (2.10). Nevertheless, implementation is more natural and straightforward if we apply the following identity, an empirical version of Proposition 2.6.

Proposition 2.7. *Fix $\hat{w}_{1,\alpha}, \ldots, \hat{w}_{j-1,\alpha}$. Let $\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}}$ be the kth largest eigenvalue of $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}$ with corresponding eigenfunction $\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}$. Suppose $\hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}} = \cdots = \hat{\lambda}_{m,\widehat{X}^{(j,\alpha)}} > \hat{\lambda}_{m+1,\widehat{X}^{(j,\alpha)}}$. If $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta) = \sum_{i=1}^{n} \widehat{X}_i^{(j,\alpha)} \widehat{Y}_i^{(j,\alpha)}/n$ is not orthogonal to $\operatorname{span}\{\hat{\phi}_{1,\widehat{X}^{(j,\alpha)}}, \ldots, \hat{\phi}_{m,\widehat{X}^{(j,\alpha)}}\}$, then there exists $\hat{\delta}^{(j,\alpha)} \in (-1,0) \cup (0,\infty)$ such that*

$$\hat{w}_{j,\alpha} = \left[\sum_{k=1}^{\infty} \frac{\widehat{\operatorname{cov}}^2\{\widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}}{\{\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}}/\hat{\delta}^{(j,\alpha)}\}^2} \right]^{-1/2} \sum_{k=1}^{\infty} \frac{\widehat{\operatorname{cov}}\{\widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}}{\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}}/\hat{\delta}^{(j,\alpha)}.} \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}. \tag{2.16}$$

Remark 2.2. The infinite series (2.16) reduces to a product of matrices as in [13] if curves are approximated by linear combinations of splines (or other known functions).

Remark 2.3. When the m top eigenvalues of $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}$ are equal, we must assume that $\widehat{\mathcal{V}}_{\widehat{X}^{(p,\alpha)}}(\beta)$ is not orthogonal to $\operatorname{span}\{\hat{\phi}_{1,\widehat{X}^{(j,\alpha)}}, \ldots, \hat{\phi}_{m,\widehat{X}^{(j,\alpha)}}\}$; otherwise, the ridge-type solution (2.16) may be not a global maximizer. Corresponding examples are artificially constructible, yet they are rare in practice (see [13, 52]) especially when ε and X in (1.3) are both continuously distributed. Actually, if the assumption is not fulfilled, one can always project $\widehat{X}_i^{(j,\alpha)}$ onto the complement of $\operatorname{span}\{\hat{\phi}_{1,\widehat{X}^{(j,\alpha)}}, \ldots, \hat{\phi}_{m,\widehat{X}^{(j,\alpha)}}\}$ and update $\widehat{X}_i^{(j,\alpha)}$ with the projection.

Proposition 2.7 suggests merely considering w of a ridge-type. It helps to narrow down the search scope for $\hat{w}_{j,\alpha}$ by reformulating (2.5) as a univariate maximization problem. The only unknown item in (2.16), $\hat{\delta}^{(j,\alpha)}$, is taken as

$$\hat{\delta}^{(j,\alpha)} = \operatorname*{arg\,max}_{\delta \in (-1,0)\cup(0,\infty)} Q_{j,\alpha}(\delta) = \operatorname*{arg\,min}_{\delta \in (-1,0)\cup(0,\infty)} -\ln Q_{j,\alpha}(\delta),$$

where

$$
\begin{aligned}
Q_{j,\alpha}(\delta) = & \left[\sum_{k=1}^{\infty} \frac{\widehat{\mathrm{cov}}^2\{\widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)}\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}}{\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}}/\delta}\right]^2 \\
& \times \left[\sum_{k=1}^{\infty} \frac{\widehat{\mathrm{cov}}^2\{\widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)}\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}}{\{\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}}/\delta\}^2}\right]^{\alpha/(1-\alpha)} \\
& \times \left[\sum_{k=1}^{\infty} \frac{\widehat{\mathrm{cov}}^2\{\widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)}\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}\widehat{\mathrm{var}}\{\int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)}\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}}\}}{\{\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(j,\alpha)}}/\delta\}^2}\right]^{\alpha/(1-\alpha)-1}
\end{aligned}
\tag{2.17}
$$

is obtained by substituting the right-hand side of (2.16) for w in (2.15). The univariate function $\ln Q_{j,\alpha}$ depends not only on j and α but also on the observations, which makes it inconvenient to theoretically investigate this function's behavior. However, for the specific datasets to be investigated in Section 2.5, there seems to be no more than one local maximum within either $(-1, 0)$ or $(0, \infty)$; see Figure 2.1. As a result, the maximization in each piece can be handled by a symbolic computation program.

To reduce computational burden and increase the efficiency of Algorithm 2.1, we compute $\widehat{X}_i^{(j,\alpha)}$ and $\hat{\beta}_{j,\alpha}$ in a recursive way, viz.

$$
\widehat{X}_i^{(j,\alpha)} = \widehat{X}_i^{(j-1,\alpha)} - \widehat{\mathrm{var}}^{-1/2}\left(\int_{\mathbb{T}_X} X\hat{w}_{j-1,\alpha}\right)\left\{\int_{\mathbb{T}_X} \hat{w}_{j-1,\alpha}(X_i - \bar{X})\right\}\widehat{\mathcal{V}}_X(\hat{w}_{j-1,\alpha}),
$$

and

$$
\hat{\beta}_{j,\alpha} = \hat{\beta}_{j-1,\alpha} + \widehat{\mathrm{cov}}\left(Y, \int_{\mathbb{T}_X} X\hat{w}_{j,\alpha}\right)\widehat{\mathrm{var}}^{-1/2}\left(\int_{\mathbb{T}_X} X\hat{w}_{j,\alpha}\right)\hat{w}_{j,\alpha},
$$

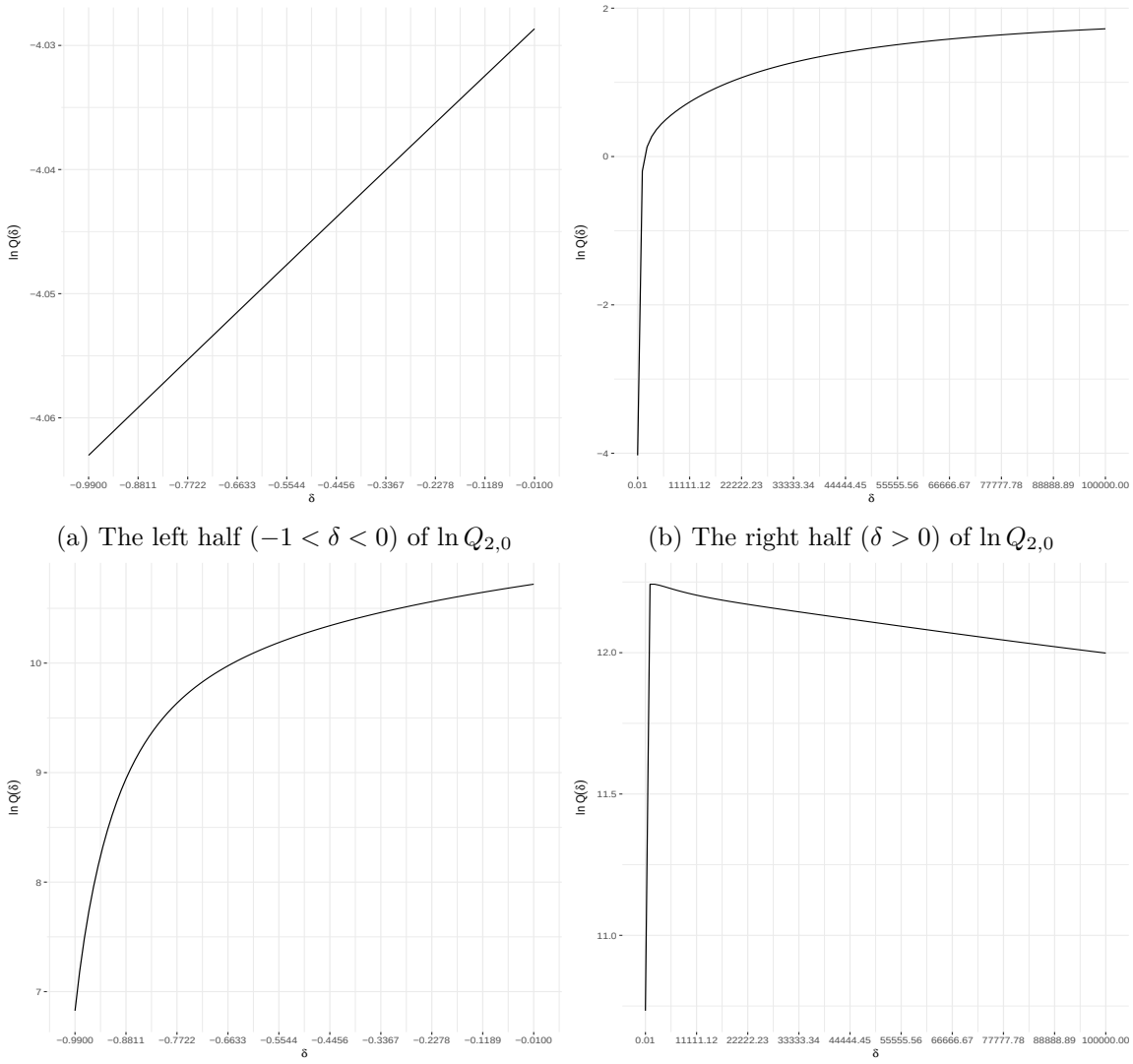
starting with $\widehat{X}_i^{(1,\alpha)} = X_i - \bar{X}$, $\widehat{Y}_i^{(1,\alpha)} = Y_i - \bar{Y}$ and $\hat{\beta}_{0,\alpha} = 0$.

2.4.1 Tuning parameters

The result of FC regression relies on the choice of two parameters: α, the continuum parameter, and p, the number of basis functions included in the model. Favoring a much lower expense in computation, we tune them through the generalized cross-validation (GCV, [23]): specifically, for each possible pair (p, α), we define

$$
\mathrm{GCV}(p, \alpha) = \sum_{i=1}^{n} \frac{\{Y_i - \hat{\eta}_{p,\alpha}(X_i)\}^2}{(n - p - 1)^2},
$$

i.e., the GCV criterion with the degrees of freedom (DoF) heuristically taken as p. This is a compromise when we have no idea on how to estimate DoF associated with FC regression (due to the intrinsic complexity). This tuning scheme is less time-consuming than cross-validation.



(a) The left half ($-1 < \delta < 0$) of $\ln Q_{2,0}$

(b) The right half ($\delta > 0$) of $\ln Q_{2,0}$

(c) The left half ($-1 < \delta < 0$) of $\ln Q_{5,0.4}$

(d) The right half ($\delta > 0$) of $\ln Q_{5,0.4}$

Figure 2.1: Plots of $\ln Q_{2,0}$ and $\ln Q_{5,0.4}$ for Tecator™ data (spectra vs. fat). Each pair of curves (i.e., the top two or bottom two) applies to Tecator™ data (spectra vs. fat) discussed in Section 2.5.2, with two different sets of values for (p, α). Neither pair of graphs shows more than one maximizer.

Define the fraction of variance explained (FVE) by

$$\mathrm{FVE}(j_0) = \sum_{j=1}^{j_0} \lambda_{j,X} \Big/ \sum_{j=1}^{\infty} \lambda_{j,X}, \tag{2.18}$$

where $\lambda_{j,X}$ is replaced in practice with its empirical counterpart. Classically the optimal 2-tuple (p, α) is chosen after a fixed search (FS), i.e., minimizing $\mathrm{GCV}(p, \alpha)$ over a pre-configured rectangle mesh grid, say $\{1, \ldots, p_{\max}\} \times \{\alpha_1, \ldots, \alpha_J\}$, where J is based on the computational capacity and $p_{\max}$ is assigned such that the first $p_{\max}$ components explain

most of the variation, e.g.,

$$p_{\max} = \min\{j \in \mathbb{Z}^+ : \mathrm{FVE}(j_0) \geq 95\%\}. \tag{2.19}$$

This strategy of determining $p_{\max}$ is commonly adopted for FPCR; it is still applicable to FC regression, since supervised methods tend to include fewer basis functions than FPCR; see [28, Section 6].

Our recursive implementation implies that the results corresponding to $\{1, \ldots, p_{\max} - 1\} \times \{\alpha_1, \ldots, \alpha_J\}$ are interim ones needed only in pursuing the outputs for $\{p_{\max}\} \times \{\alpha_1, \ldots, \alpha_J\}$, i.e., FS is actually carried out over $\{p_{\max}\} \times \{\alpha_1, \ldots, \alpha_J\}$. [10] argued that FS is less efficient (in terms of the computational burden) than a random search (RS) which allows the upper bound of p to vary with the value of α. Specifically, sample J iid two-tuples uniformly from $\{1, \ldots, p_{\max}\} \times [0, 1)$, say $\{(p_{1,\max}, \alpha_1), \ldots, (p_{J,\max}, \alpha_J)\}$, and form a non-rectangular search grid $\{(p, \alpha_j) : 1 \leq p \leq p_{j,\max}, 1 \leq j \leq J\}$. The cardinality of this non-rectangular grid is smaller than that of the rectangular one, while (most likely) their projections on the first (resp. second) dimension are of the identical cardinality $p_{\max}$ (resp. J); in this case, [10] illustrated that RS would save some time without an enormous sacrifice in accuracy.

Pseudocodes for our implementation are summarized in Algorithm 2.1. In the following section, we will employ both FS and RS in tuning parameters for FC regression.

2.5 Numerical illustration

To illustrate the performance of FC regression, the results given by our method (with both FS and RS) were compared with those from supFPC [69], pFPLS [2], FPLS_R- and FPCR_R-REML (both recommended by [78] after a series of comparisons) and smoothed FPC (smFPC, [76, Section 9.3]). Among these the first four are supervised, while the other two are categorized as unsupervised.

With the aid of R [74], RStudio™ [81] and the R-package `fda` [77], we coded all the methods mentioned in the preceding paragraph except for FPCR_R-REML which was implemented using R-function `fpcr` in [39]. Our source codes are accessible at `https://github.com/ZhiyangGeeZhou/Functional-continuum-regression`.

2.5.1 Simulation study

The dataset `CanadianWeather` in [77] contains the (base 10 logarithm of) precipitation at 35 different locations in Canada averaged over 1960 to 1994. Taking these curves as iid, we estimated the mean function, top three eigenvalues and corresponding eigenfunctions. Then we exploited them as true values for, respectively, $\mu_X, \lambda_{1,X}, \lambda_{2,X}, \lambda_{3,X}, \phi_{1,X}, \phi_{2,X}$, and $\phi_{3,X}$ in our simulation. Analogous to [69], each sample in our simulation consisted of 100

Algorithm 2.1 FC regression tuned by GCV

for (p, α) in a finite set **do**
  for i from 1 to n **do**
    if $p = 1$ **then**
      $\widehat{X}_i^{(p,\alpha)} \leftarrow X_i - \bar{X}$.
      $\widehat{Y}_i^{(p,\alpha)} \leftarrow Y_i - \bar{Y}$.
      $\hat{\beta}_{p-1,\alpha} \leftarrow 0$.
    else
      $\widehat{X}_i^{(p,\alpha)} \leftarrow \widehat{X}_i^{(p-1,\alpha)} - c_2 \cdot c_3 \cdot \widehat{\mathcal{V}}_X(\hat{w}_{p-1,\alpha})$.
      $\widehat{Y}_i^{(p,\alpha)} \leftarrow \widehat{Y}_i^{(p-1,\alpha)} - \hat{\eta}_{p-1,\alpha}(X_i)$.
    end if
  end for
  $\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}}, \hat{\phi}_{j,\widehat{X}^{(p,\alpha)}} \leftarrow$ the jth eigenvalue and eigenfunction of $\widehat{\mathcal{V}}_{\widehat{X}^{(p,\alpha)}}$.
  $a_j \leftarrow \widehat{\mathrm{cov}}\{\widehat{Y}^{(p,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(p,\alpha)} \hat{\phi}_{j,\widehat{X}^{(p,\alpha)}}\}$.
  $b_j \leftarrow \widehat{\mathrm{var}}\{\int_{\mathbb{T}_X} \widehat{X}^{(p,\alpha)} \hat{\phi}_{j,\widehat{X}^{(p,\alpha)}}\}$.
  $$Q_{p,\alpha}(\delta) \leftarrow \left\{\sum_{j=1}^{\infty} \frac{a_j^2}{\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(p,\alpha)}}/\delta}\right\}^2 \left[\sum_{j=1}^{\infty} \frac{a_j^2}{\{\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(p,\alpha)}}/\delta\}^2}\right]^{\alpha/(1-\alpha)} \times \left[\sum_{j=1}^{\infty} \frac{a_j^2 b_j}{\{\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(p,\alpha)}}/\delta\}^2}\right]^{\alpha/(1-\alpha)-1}.$$
  $\hat{\delta}^{(p,\alpha)} \leftarrow \arg\min_{\delta \in (-1,0)\cup(0,\infty)} -\ln Q_{p,\alpha}(\delta)$.
  $$\hat{w}_{p,\alpha} \leftarrow \left[\sum_{j=1}^{\infty} \frac{a_j^2}{\{\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(p,\alpha)}}/\hat{\delta}^{(p,\alpha)}\}^2}\right]^{-1/2} \sum_{j=1}^{\infty} \frac{a_j}{\hat{\lambda}_{j,\widehat{X}^{(p,\alpha)}} + \hat{\lambda}_{1,\widehat{X}^{(p,\alpha)}}/\hat{\delta}^{(p,\alpha)}} \hat{\phi}_{j,\widehat{X}^{(p,\alpha)}}.$$
  $c_1 \leftarrow \widehat{\mathrm{cov}}(Y, \int_{\mathbb{T}_X} X\hat{w}_{p,\alpha})$.
  $c_2 \leftarrow \widehat{\mathrm{var}}^{-1/2}(\int_{\mathbb{T}_X} X\hat{w}_{p,\alpha})$.
  $c_3 \leftarrow \int_{\mathbb{T}_X} \widehat{X}_i^{(1,\alpha)} \hat{w}_{p,\alpha}$.
  $\hat{\beta}_{p,\alpha} \leftarrow \hat{\beta}_{p-1,\alpha} + c_1 c_2 \hat{w}_{p,\alpha}$.
  for i from 1 to n **do**
    $\hat{\eta}_{p,\alpha}(X_i) \leftarrow \bar{Y} + \int_{\mathbb{T}_X} \widehat{X}_i^{(1,\alpha)} \hat{\beta}_{p,\alpha}$.
  end for
  $\mathrm{GCV}(p,\alpha) \leftarrow (n-p-1)^{-2} \sum_{i=1}^{n} \{Y_i - \hat{\eta}_{p,\alpha}(X_i)\}^2$.
end for
optimal $(p,\alpha) \leftarrow \arg\min_{(p,\alpha)} \mathrm{GCV}(p,\alpha)$.

iid functional predictors X_i such that

$$X_i = \mu_X + \sum_{j=1}^{3} \xi_{ij} \phi_{j,X}$$

with responses Y_i generated as

$$Y_i = \int_{\mathbb{T}_X} \beta(X_i - \mu_X) + \sigma\varepsilon_i,$$

where $\xi_{ij}\lambda_{j,X}^{-1/2}$ and ε_i were all assumed to be iid as $\mathcal{N}(0,1)$. We used two levels (2 and 20) of signal-to-noise-ratio (SNR)

$$\mathrm{SNR} = \sigma^{-1}\mathrm{var}^{1/2}\left(\int_{\mathbb{T}_X}\beta X_i\right) = \sigma^{-1}\left\{\sum_{j=1}^{3}\lambda_{j,X}\left(\int_{\mathbb{T}_X}\beta\phi_{j,X}\right)^2\right\}^{1/2}.$$

The choice of coefficient function β must be limited to $\mathrm{span}(\phi_{1,X}, \phi_{2,X}, \phi_{3,X})$; see Remark 2.4 below. Specifically, we considered two sorts of coefficient function:

$$\beta = \phi_{1,X} \tag{2.20}$$

and

$$\beta = \phi_{3,X}. \tag{2.21}$$

No matter how supervised they are, all the methods were expected to be favored by the scenario where $\beta = \phi_{1,X}$; the other scenario was intended to imitate the target our proposal is designed for: the true coefficient function is orthogonal to the top few eigenfunctions of $\mathcal{V}_X$.

Remark 2.4. Decompose β to be $\beta = \beta^{(1)} + \beta^{(2)}$ in which $\beta^{(1)}$ (resp. $\beta^{(2)}$) is the projection of β onto $\mathrm{span}(\phi_{1,X}, \phi_{2,X}, \phi_{3,X})$ (resp. its complement in $L^2(\mathbb{T}_X)$). Then

$$\int_{\mathbb{T}_X}\beta(X_i - \mu_X) = \int_{\mathbb{T}_X}\beta^{(1)}(X_i - \mu_X)$$

since $X_i - \mu_X \in \mathrm{span}(\phi_{1,X}, \phi_{2,X}, \phi_{3,X})$. In other words, $\beta^{(2)}$ vanishes when we take the inner product between β and $X_i - \mu_X$, making $\beta^{(2)}$ unidentifiable in the regression model.

For each of the four combinations of β and SNR, we generated 200 samples and applied all the techniques to each sample. The estimation quality was directly evaluated via the (point-wise) root mean squared error (RMSE) defined as

$$\mathrm{RMSE}(t) = \left[\frac{1}{200}\sum_{r=1}^{200}\{\beta(t) - \hat{\beta}_r(t)\}^2\right]^{1/2},$$

where $\hat{\beta}_r(t)$ was the estimated coefficient function for the rth sample.

As described in Section 2.4.1, there were two strategies in tuning FC regression. To accomplish RS, we randomly generated a brand new search grid following Section 2.4.1 for each sample, taking $J = 10$ and $p_{\max} = 2$, because of $(\lambda_{1,X} + \lambda_{2,X})/(\lambda_{1,X} + \lambda_{2,X} + \lambda_{3,X}) \approx 97\%$. FS for FC regression searched over a 2×11 grid, $\{1,2\} \times (\{0 \times 10^{-1}, \ldots, 9 \times 10^{-1}\} \cup \{.999\})$, where the same scope for p, i.e., $\{1,2\}$, was used for all five other methods. In

the implementation of smFPC, supFPC and pFPLS, smoothing penalty parameters were chosen from $\{0\} \cup \{10^0, \ldots, 10^5\}$. Moreover, as suggested by [69], candidate values of the "weight" parameter needed by supFPC were taken from $\{0 \times 10^{-1}, \ldots, 10 \times 10^{-1}\}$.

When $\beta = \phi_{1,X}$ and SNR = 20 (Figure 2.2a), all the techniques performed close to each other. Curves corresponding to FC regression were not the most outstanding ones: they were slightly worse than pFPLS and smFPC but better than FPLS_R-REML and FPCR_R-REML whose RMSE values became dramatically high at both ends of the domain. For any method, RMSE values were enlarged as the noise increased (or equivalently the SNR decreased); see Figure 2.2b. Compared with pFPLS and smFPC, FC regression was more sensitive to the change of SNR, possibly because the tuning of one more parameter introduced more variability and/or FC regression did not penalize the smoothness of estimated basis functions or coefficient functions. Another interesting phenomenon was that FC regression with RS outperformed that with FS under both levels of SNR: the setup of search points was also a source of bias which was likely to be alleviated by the randomization.

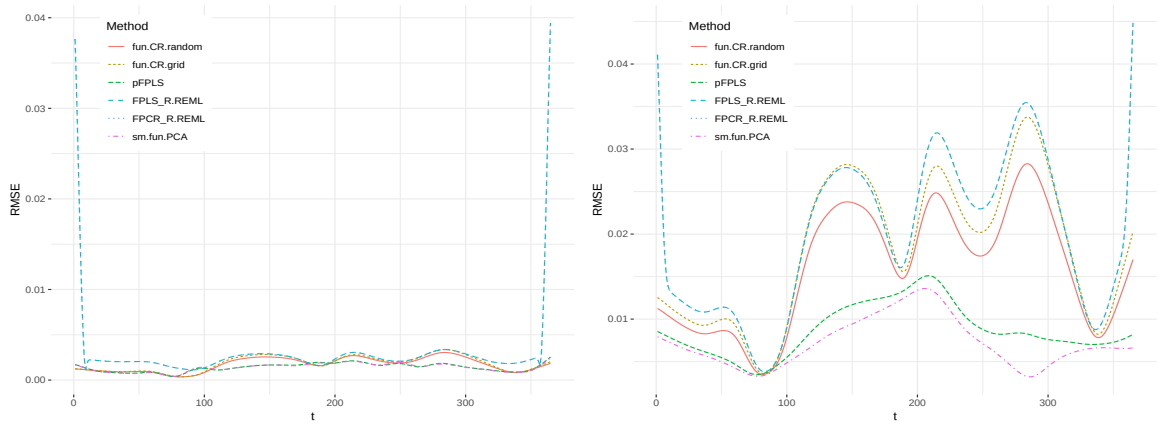
Unsurprisingly, as shown in Figure 2.3, the scenario of $\beta = \phi_{3,X}$ did not favor the smFPC which was unlikely to involve the third eigenfunction. FC regression outperformed competitors regardless of SNR; it returned the lowest RMSE uniformly in the whole domain. When encountering noisier settings, RMSE curves of FC regression were almost overlapped by those from FPLS_R-REML and FPCR_R-REML except at the ends of $\mathbb{T}_X$; see Figure 2.3b. Note that curves for supFPC were not included in either Figure 2.2 or 2.3, as RMSE values from supFPC were much larger than those from other approaches; it seemed that either the estimators from [69] were not consistent or they needed a larger sample size to reach a more satisfying accuracy.

2.5.2 Application to real datasets

For each of following two datasets, we randomly reserved roughly 10% of all the samples of each dataset for testing and used the remainder for training. We repeated the random split 200 times. To mitigate impacts from different testing sets and facilitate the comparison in prediction, we defined the relative mean squared prediction error (ReMSPE), which is a ratio of the prediction error from a competitor to the one from the mean training response:

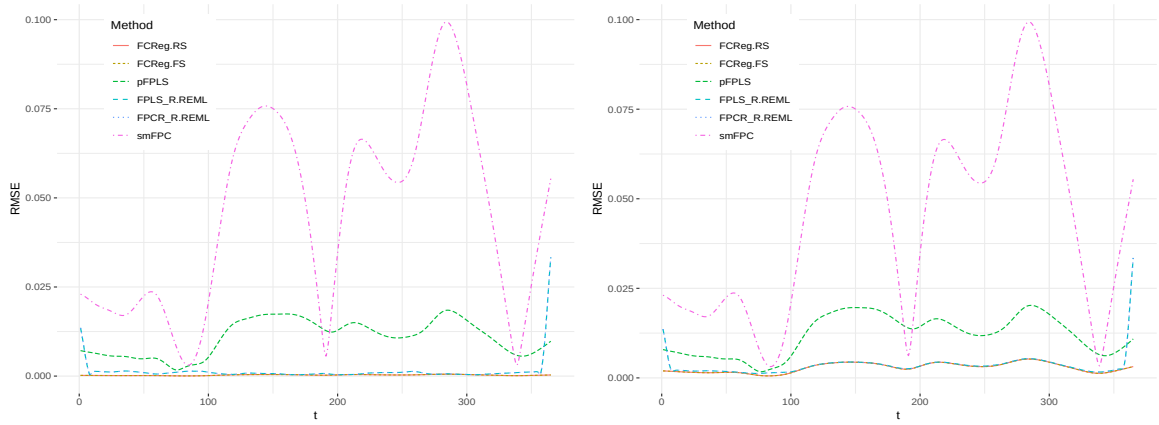
$$\text{ReMSPE} = \frac{\sum_{i \in \text{ID}_{\text{test}}} (Y_i - \widehat{Y}_i)^2}{\sum_{i \in \text{ID}_{\text{test}}} (Y_i - \bar{Y}_{\text{train}})^2}, \tag{2.22}$$

where ID_{test} was the index set for testing data, and $\widehat{Y}_i$ was the prediction corresponding to Y_i. For each approach, we generated a boxplot of the 200 values of ReMSPE. As for the candidate pool for tuning parameters, we kept all the settings in Section 2.5.1 except the one for p; we raised its upper bound from 2 to 5 to accommodate the new datasets.



(a) SNR = 20 (b) SNR = 2

Figure 2.2: RMSE curves of estimated coefficient functions when β = (2.20). Subfigures are displayed with identical scales. In the legend of each subfigure, the six linetypes (or colors), from top to bottom, correspond to FC regression (tuned by RS), FC regression (tuned by FS), pFPLS, FPLS$_R$-REML, FPCR$_R$-REML and smFPC, respectively. Curves corresponding to FPLS$_R$-REML and FPCR$_R$-REML almost overlap each other in each subfigure. supFPC does not perform well in estimation for this case and hence its RMSE curve is not shown.



(a) SNR=20 (b) SNR=2

Figure 2.3: RMSE curves of estimated coefficient functions when β = (2.21). Subfigures are displayed with identical scales. In the legend of each subfigure, the six linetypes (or colors), from top to bottom, correspond to FC regression (tuned by RS), FC regression (tuned by FS), pFPLS, FPLS$_R$-REML, FPCR$_R$-REML and smFPC, respectively. Curves corresponding to FPLS$_R$-REML and FPCR$_R$-REML almost overlap each other in each subfigure. supFPC does not perform well in estimation for this case and hence its RMSE curve is not shown.

Medfly data

Investigated in substantial literature (see, e.g., [66, 83]), the Mediterranean fruit fly, or medfly for short, has become a popular object of study, partly owing to its short lifespan.

The medfly data (`http://faculty.bscb.cornell.edu/~hooker/FDA2008/medfly.Rdata`; accessed 27-Feb-2019) records lifespans of 50 female flies as well as numbers of eggs laid by each of them in each of the 26 days. People would like to uncover how lifespan is influenced by fecundity as time goes on by relating the curves of egg count to lifespans.

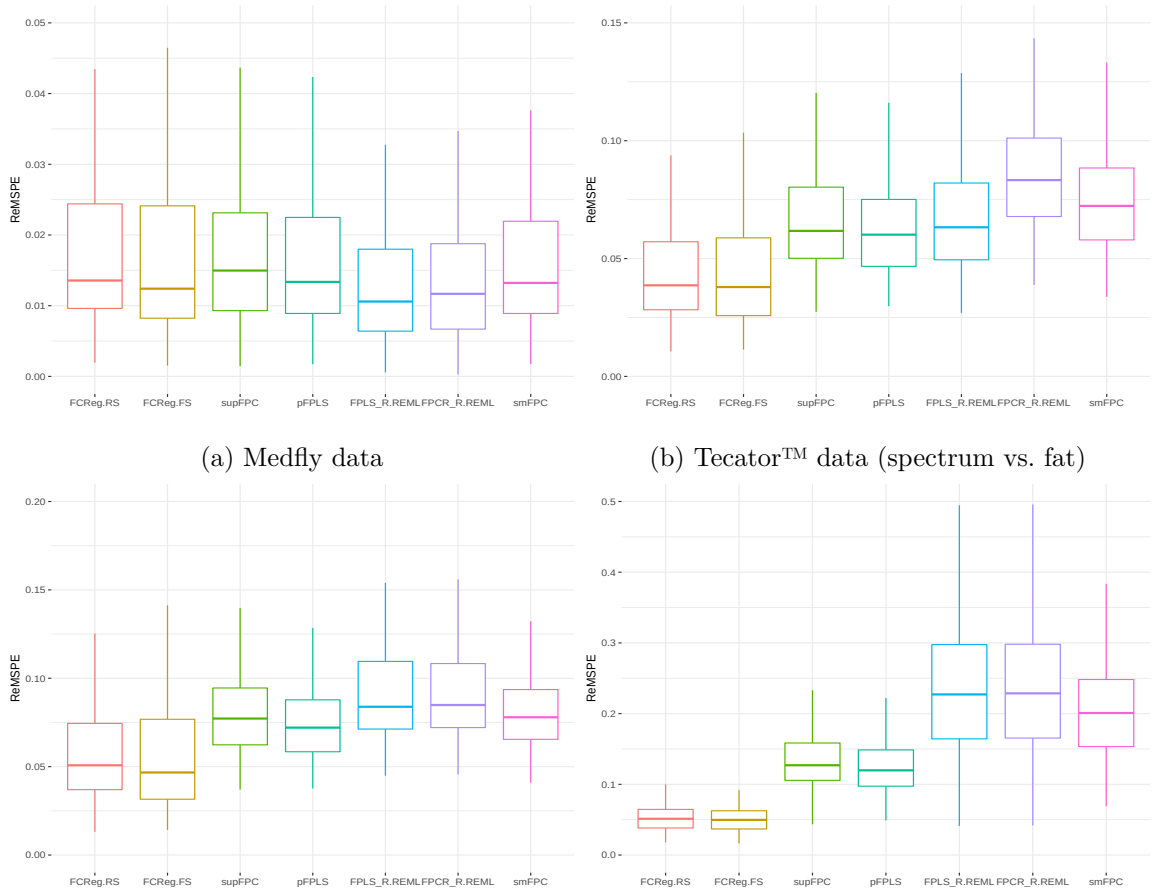
Taking the egg count and lifespan as predictor and response respectively, all the seven methods, no matter whether supervised or unsupervised, performed fairly close to each other, though FPCR_R- and FPLS_R-REML appeared slightly better than the other five in terms of both of the mean and dispersion of ReMSPE values; see Figure 2.4a. More formally, we carried out (two-sided) paired t-tests for the comparison between RS-tuning and FS-tuning FC regression as well as the ones between each FC regression and each of other existing methods, involving 11 comparisons in total. At a significance level of 0.05 with the Bonferroni correction, there were significant differences in comparisons of FPLS_R-REML vs. FC regression (with RS), FPLS_R-REML vs. FC regression (with FS) and FPCR_R-REML vs. FC regression (with FS) (with respective p-values 2.0×10^{-4}, 3.2×10^{-10} and 3.8×10^{-7}).

Tecator™ data

A Tecator™ Infratec Food and Feed Analyzer recorded NIR absorbance spectra (ranging from 850 to 1050 nm and divided into 100 channels) of 240 finely chopped pure meat samples with different fat, moisture and protein contents. The dataset (`http://lib.stat.cmu.edu/datasets/tecator`; accessed 04-Aug-2019) contains the absorbance spectra (i.e., the base 10 logarithm of transmittance at each wavelength) and the three contents measured in percent by analytic chemistry.

We regressed the fat, moisture and protein contents, respectively, on the absorbance spectra. Roughly, from a graphical viewpoint, Figure 2.4b categorized the seven approaches into three groups according to their performance: FC regression on the left end, the three in the middle (including supFPC, pFPLS and FPLS_R-REML), and another two on the very right (i.e., FPCR_R-REML and smFPC); supervised strategies were apparently preferred. This phenomenon did not hold in the cases of moisture (Figure 2.4c) or protein (Figure 2.4d), but FC regression, especially the one tuned by FS, still took the lead in terms of ReMSPE. Again, focusing on the 11 comparisons mentioned in Section 2.5.2, paired t-tests could reach a relatively formal conclusion. At a significance level of 0.05 with the Bonferroni correction, when the response was fat (resp. moisture), the only insignificantly different comparison was FC regression (with RS) vs. FC regression (with FS) corresponding to p-value 0.13 (resp. 0.82). In the application to the protein data, even the difference between the two FC regression techniques became significant (with p-value 2.0×10^{-3}), implying a loss in accuracy when adopting RS instead of FS.

Despite the better performance of FS in tuning FC regression, it cost over 50% more than the time consumed by RS when applied to Tecator™ data; see the last three columns of Table 2.1. This trade-off made RS preferred when one was not too sensitive to the accuracy.



(a) Medfly data

(b) Tecator™ data (spectrum vs. fat)

(c) Tecator™ data (spectrum vs. moisture)

(d) Tecator™ data (spectrum vs. protein)

Figure 2.4: Boxplots of ReMSPE values for different methods for real applications. In each subfigure, the seven boxes, from left to right, correspond to FC regression (tuned by RS), FC regression (tuned by FS), supFPC, pFPLS, FPLS$_R$-REML, FPCR$_R$-REML and smFPC, respectively.

2.6 Concluding remarks

Specially designed for scalar-on-function regression models, the framework of FC regression encompasses the well-known FPCR and FPLS, etc.. We gave various equivalent forms of FC basis functions which lower the difficulty of optimization in numerical implementation. Consistency of the estimators was demonstrated for the case of fixed p. Verified in numerical studies and compared with several existing methods, our strategy was overall competitive in terms of both estimation and prediction.

The core of our algorithm is to locate the constrained global maximizer of the logarithm of $Q_{j,\alpha}(\delta)$ in (2.17). In Section 2.5, thanks to the simplicity of the curves of $\ln Q_{j,\alpha}$, we did not have to initiate the maximization with multiple start points. Even so, our implementation was still more involved than competitors when the number of curves becomes larger; see

Table 2.1: Time consumed (in seconds) in Section 2.5 by different approaches after 200 repeats (running on a desktop with Intel® Core™ i5-7500 CPU @ 2 × 3.40 GHz and 8 GB RAM)

	Simulation				Medfly	Tecator™		
	$\beta = \phi_{1,X}$		$\beta = \phi_{3,X}$			Fat	Moisture	Protein
SNR	20	2	20	2				
Number of curves	100				50	240		
FC regression (with RS)	120	123	125	120	66	814	817	743
FC regression (with FS)	129	124	140	139	101	1363	1340	1238
supFPC	232	240	239	235	31	327	336	328
pFPLS	1030	1016	1017	1009	229	1191	1193	1193
FPLS_R-REML	76	72	75	79	21	80	71	74
FPCR_R-REML	75	74	73	77	19	69	66	71
smFPC	104	101	101	108	19	110	114	110

Table 2.1. It can always be worse: curves of $\ln Q_{j,\alpha}$ may be more complex in some real datasets. In such cases, we have to avoid being trapped in some local maxima by trying multiple initial values. But this strategy would definitely slow down the implementation of FC regression. For instance, under the same computing environment, if we try 100 initial points in each maximization, the time used by FC regression with FS for the Tecator™ data would be over 30 times as much as the corresponding number posted in Table 2.1. Concerned about this disadvantage, we introduced RS, which turned out to be an effective way to improve the most time-consuming cases of the numerical study with little loss in accuracy. Under certain circumstances (see Figure 2.2), FC regression with RS even achieved a better performance in estimation because the randomization reduced the bias caused by the pinned discrete search grid.

Last but not least, FC regression possesses the potential to be further improved and extended. As mentioned in Section 2.5.1, the accuracy of FC regression could suffer from no penalty on smoothness. It would be helpful if we introduce one more tuning parameter to force the resulting estimator to be smoother. With the assistance of RS, the time consumption should be comparable with that of the version without penalty. In addition, with a generalization analogous to that in [15], it may be possible to handle multiple responses simultaneously and even a functional response. Another possible direction of research is to enhance the robustness by replacing variance and covariance terms with robust counterparts, just as [84] did for continuum regression.

Chapter 3

Continuum centroid classifier for functional data

3.1 Introduction

Recall two-tuples $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid as (X, Y), with scalar Y and functional X defined on, without loss of generality, $\mathbb{T}_X = [0, 1]$. When Y takes values from $\{0, 1\}$, the problem becomes a binary functional classification. Potential applications include disease diagnosis using medical imaging. There has been extensive work on functional data classification. [32, 85] defined distances among curves without projecting curves to specific directions, while many other researchers suggested reducing the instrinsically infinite dimension at the first step. Typical strategies of the latter form apply multivariate classification techniques to FPC scores $\int_{\mathbb{T}_X} X_i \phi_j$ with ϕ_j as in (2.1), including but not limited to linear and quadratic Bayes classifiers on FPC scores [35], logistic regression on FPC scores [58] and support vector machines on FPC scores [80].

Our work is mainly motivated by the centroid classifier proposed by [28] (later detailed in Section 3.1.2) who projected functional data to the direction of either $\beta_{p,\mathrm{WFPC}}$ (3.3) or $\beta_{p,\mathrm{FPLS}}$ (3.4). In this way, the authors converted the classification problem to the estimation of the slope function of SoFR (1.3). As mentioned in Chapter 2, $\beta_{p,\mathrm{WFPC}}$ (3.3) and $\beta_{p,\mathrm{FPLS}}$ (3.4) are either unsupervised or too supervised; our proposal $\beta_{p,\alpha}$ (2.8) is hence expected to improve the performance of the previous two directions.

3.1.1 Formalization of the problem

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid copies of (X, Y), where X is a random function defined on $\mathbb{T}_X = [0, 1]$, and Y is the label of X taking values from $\{0, 1\}$. In other words, each X_i is sampled from a mixture of two populations Π_0 and Π_1 with the indicator $Y_i = \mathbb{1}(X_i \in \Pi_1)$. Of interest is the binary classification for a newly observed X^* distributed as X but independent of $X_1, \ldots, X_n$. Denote by $\mu_X^{[k]}$ and $v_X^{[k]}$ respectively the sub-mean and

sub-covariance functions for Π_k, $k = 0, 1$, i.e.,

$$\mu_X^{[k]} = \mu_X^{[k]}(\cdot) = \mathrm{E}\{X(\cdot) \mid Y = k\}$$

and

$$v_X^{[k]} = v_X^{[k]}(s,t) = \mathrm{cov}\{X(s), X(t) \mid Y = k\}.$$

Corresponding to $v_X^{[k]}$, sub-covariance operator $\mathcal{V}_X^{[k]} : L^2(\mathbb{T}_X) \to L^2(\mathbb{T}_X)$ is defined such that, for $f \in L^2(\mathbb{T}_X)$,

$$\mathcal{V}_X^{[k]}(f)(\cdot) = \int_{\mathbb{T}_X} f(s) v_X^{[k]}(s,\cdot)\mathrm{d}s.$$

$\mathcal{V}_X^{[k]}$ possesses a countable number of nonnegative eigenvalues, say $\lambda_{k,1} \geq \lambda_{k,2} \geq \cdots > 0$, and corresponding eigenfunctions, say $\phi_{k,1}, \phi_{k,2}, \ldots$

Let $\pi_0 = \Pr(X \in \Pi_0) = \Pr(Y = 0) \in (0,1)$. Decomposing functions μ_X and v_X (1.2), we have $\mu_X = \pi_0 \mu_X^{[0]} + (1-\pi_0)\mu_X^{[1]}$ and, in view of the law of total covariance,

$$v_X(s,t) = v_X^W(s,t) + v_X^B(s,t),$$

where

$$v_X^W(s,t) = \pi_0 v_X^{[0]}(s,t) + (1-\pi_0) v_X^{[1]}(s,t) \tag{3.1}$$

and

$$v_X^B(s,t) = \pi_0(1-\pi_0)\{\mu_X^{[1]}(s) - \mu_X^{[0]}(s)\}\{\mu_X^{[1]}(t) - \mu_X^{[0]}(t)\}$$

are respectively the within- and between-group covariance functions.

3.1.2 Review of centroid classifier

For the binary classification of X^*, projecting curves onto the one-dimensional space spanned by pre-determined $\omega \in L^2(\mathbb{T}_X)$, i.e., $\mathrm{span}(\omega)$, [28] exploited the resulting projection in constructing classifiers. Specifically, they defined a classifier

$$\mathcal{D}(X^* \mid \omega) = \left\{ \int_{\mathbb{T}_X} \frac{\omega}{\|\omega\|_2}(X^* - \mu_X^{[1]}) \right\}^2 - \left\{ \int_{\mathbb{T}_X} \frac{\omega}{\|\omega\|_2}(X^* - \mu_X^{[0]}) \right\}^2 + 2\ln\frac{\pi_0}{1-\pi_0}, \tag{3.2}$$

where $\|\omega\|_2^{-1} |\int_{\mathbb{T}_X} \omega(X^* - \mu_X^{[k]})|$, the magnitude of the projection of $X^* - \mu_X^{[k]}$ onto $\mathrm{span}(\omega)$, can be regarded as the distance from X^* to $\mu_X^{[k]}$. When $\mathcal{D}(X^* \mid \omega)$ is positive, X^* is thought to be closer to $\mu_X^{[0]}$ and hence assigned to Π_0 and otherwise to Π_1. Given ω, this principle is identical to the linear discriminant analysis (LDA) assuming $\int_{\mathbb{T}_X} X\omega$ to be normally distributed conditional on $X \in \Pi_k$ with $\mathrm{var}(\int_{\mathbb{T}_X} X\omega \mid X \in \Pi_k) = \|\omega\|_2^2$, i.e., $\int_{\mathbb{T}_X} X\omega \mid X \in \Pi_k \sim \mathcal{N}(\int_{\mathbb{T}_X} \mu_X^{[k]}\omega, \|\omega\|_2^2)$.

It remains to determine the direction $\omega \in L^2(\mathbb{T}_X)$ so as to optimize the misclassification rate

$$\begin{aligned}&\mathrm{err}\{\mathcal{D}(X^* \mid \omega)\}\\ &\quad = \pi_0 \Pr\{\mathcal{D}(X^* \mid \omega) < 0 \mid X^* \in \Pi_0\} + (1-\pi_0)\Pr\{\mathcal{D}(X^* \mid \omega) > 0 \mid X^* \in \Pi_1\}.\end{aligned}$$

[28] succeeded in bridging the binary classification problem to SoFR (1.3) by taking $\omega = \beta_{p,\mathrm{WFPC}}$ or $\beta_{p,\mathrm{FPLS}}$, both of which are derived from (1.5). More specifically,

$$\beta_{p,\mathrm{WFPC}} = \underset{\theta \in \mathrm{span}(\phi_1^W, \ldots, \phi_p^W)}{\arg\min} \mathrm{E}\left\{Y - \mu_Y - \int_{\mathbb{T}_X} \theta(X - \mu_X)\right\}^2 \tag{3.3}$$

and

$$\beta_{p,\mathrm{FPLS}} = \underset{\theta \in \mathrm{span}(w_{1,\mathrm{FPLS}}, \ldots, w_{p,\mathrm{FPLS}})}{\arg\min} \mathrm{E}\left\{Y - \mu_Y - \int_{\mathbb{T}_X} \theta(X - \mu_X)\right\}^2, \tag{3.4}$$

where ϕ_j^W is the top jth eigenfunction of v_X^W (3.1) and $w_{j,\mathrm{FPLS}}$ is defined as (2.3). Integer p is tuned through cross-validation. The resulting classifier $\mathcal{D}(X^* \mid \beta_{p,\mathrm{WFPC}})$ (resp. $\mathcal{D}(X^* \mid \beta_{p,\mathrm{FPLS}})$) is abbreviated here to be PCC (resp. PLCC).

The rest of this chapter is organized as follows. In Section 3.2, we introduce the two subtypes of our classifier, including both the population and empirical versions, and then establish the property of (asymptotically) perfect classification. The numerical illustration in Section 3.3 investigates the performance of our proposal, highlighting the settings in favor of it. Section 3.4 gives concluding remarks as well as discussions. Technical details are relegated to Appendix A.2.

3.2 Continuum centroid classifier

The continuum centroid classifier (CCC) is defined by substituting $\beta_{p,\alpha}$ (2.8) for ω in (3.2) and, simultaneously, dropping the assumption $\mathrm{var}(\int_{\mathbb{T}_X} X\omega \mid X \in \Pi_k) = \|\omega\|_2^2$. In detail, CCC assigns trajectory X^* by applying LDA or the quadratic discriminant analysis (QDA) to the projection of X^*. The method hence has two subtypes, say CCC-L and CCC-Q. The latter is more general and given by

$$\begin{aligned}\mathcal{D}_Q(X^* \mid \beta_{p,\alpha}) = \sigma_{[1]}^{-2}(\beta_{p,\alpha})\left\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(X^* - \mu_X^{[1]})\right\}^2 - \sigma_{[0]}^{-2}(\beta_{p,\alpha})\left\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(X^* - \mu_X^{[0]})\right\}^2\\ + 2\ln\frac{\pi_0 \sigma_{[1]}(\beta_{p,\alpha})}{(1-\pi_0)\sigma_{[0]}(\beta_{p,\alpha})}\end{aligned} \tag{3.5}$$

with

$$\sigma^2_{[k]}(\omega) = \mathrm{var}\left(\int_{\mathbb{T}_X} X\omega \mid X \in \Pi_k\right) \tag{3.6}$$

for each k; it reduces to the former one

$$\mathcal{D}_L(X^* \mid \beta_{p,\alpha}) = \left\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(X^* - \mu_X^{[1]})\right\}^2 - \left\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(X^* - \mu_X^{[0]})\right\}^2 + 2\sigma^2(\beta_{p,\alpha}) \ln \frac{\pi_0}{1-\pi_0} \tag{3.7}$$

if one believes $\sigma^2(\omega) = \sigma^2_{[0]}(\omega) = \sigma^2_{[1]}(\omega)$. Analogous to (3.2), positive $\mathcal{D}_L(X^* \mid \beta_{p,\alpha})$ (or $\mathcal{D}_Q(X^* \mid \beta_{p,\alpha})$) suggests classifying X^* to Π_0.

Under conditions (C.A.2.1) in Appendix A.2, Proposition 2.2 implies that $\beta_{p,\alpha}$ and $\beta_{p,\mathrm{FPLS}}$ have identical limiting behaviours as p diverges. In that case, when $\pi_0 = 1/2$, CCC-L is equivalent to PLCC asymptotically. Even if (C.A.2.1) is violated, in theory one can still expect an asymptotically perfect classification (without misclassification) given by CCC-L, as long as (C.A.2.2) in Appendix A.2 stands; see Proposition 3.1. It is worth emphasizing that Proposition 3.1 does not require normality or specific variance structure of the two subpopulations.

Proposition 3.1. *Under condition (C.A.2.2) in Appendix A.2, CCC-L asymptotically leads to no misclassification as $p \to \infty$.*

3.2.1 Empirical implementation

In general it is impossible to observe entire trajectories. In this sense, the procedure of estimating $\beta_{p,\alpha}$ (2.8) in Chapter 2 is not detailed enough since Algorithm 2.1 is not described in the matrix form. We improve the specification for discretely observed trajectories in this section.

For brevity, the X_i are all assumed to be densely digitized on $M+1$ equispaced time points $t_m = (m-1)\Delta t$, $m = 1, \ldots, M+1$, with $\Delta t = 1/M$. Reformulating the infinite-dimensional optimization problem (2.5) as a finite-dimensional one, we employ (penalized) cubic spline smoothing [76, Sections 5.2.4–5.2.5] on each trajectory, i.e., we seek a surrogate of X_i in the N ($= M+3$, as recommended by [76, pp. 86]) dimensional linear space

$$BS_N = \mathrm{span}(\psi_1, \ldots, \psi_N) \tag{3.8}$$

spanned by cubic B-splines $\psi_1, \ldots, \psi_N$; refer to, e.g., [25, Chapter 4], for more detail on B-splines. Specifically, the estimator for the ith trajectory is

$$\widehat{X}_i = \hat{\boldsymbol{c}}_i^\top \boldsymbol{\psi}, \tag{3.9}$$

where

$$\boldsymbol{\psi} = \boldsymbol{\psi}(\cdot) = [\psi_1(\cdot), \ldots, \psi_N(\cdot)]^\top \tag{3.10}$$

and

$$\hat{\boldsymbol{c}}_i = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \theta_i \mathbf{Pen})^{-1} \boldsymbol{\Psi}^\top \boldsymbol{X}_i, \tag{3.11}$$

with matrices

$$\boldsymbol{\Psi} = [\psi_k(t_m)]_{(M+1)\times N} = [\boldsymbol{\psi}(t_1), \ldots, \boldsymbol{\psi}(t_{M+1})]^\top, \tag{3.12}$$

$$\mathbf{Pen} = \left[\int_{\mathbb{T}_X} \psi''_{l_1} \psi''_{l_2}\right]_{1\le l_1, l_2 \le N}, \tag{3.13}$$

$$\boldsymbol{X}_i = [X_i(t_1), \ldots, X_i(t_{M+1})]^\top,$$

and smoothing parameter $\theta_i > 0$. Thanks to the denseness in observation, under regularity conditions, the smoothing technique is able to recover underlying curves accurately (in the L^2 sense).

Proposition 3.2. *Assuming (C.A.2.3) and (C.A.2.4) in the appendix, for each i, $\|\widehat{X}_i - X_i\|_2 \to 0$ in probability as $M \to \infty$.*

Proposition 3.3. *Suppose*

$$\tilde{w}_{j,\alpha} = \max_{w:\|w\|_2=1} \widetilde{T}^*_{j,\alpha}(w)$$

is an estimator of $w_{j,\alpha}$ (2.5), where $\widetilde{T}^*_{j,\alpha}$ is obtained by substituting $\widehat{X}_i$ (3.9) for X_j in the expression of $\widetilde{T}^*_{j,\alpha}$ (2.15). Then $\tilde{w}_{j,\alpha}$ must lie in BS_N (3.8).

Start with optimizing $\widetilde{T}^*_{1,\alpha}(w)$. Proposition 3.3 narrows our search from $\{w : \|w\|_2 = 1, w \in L^2(\mathbb{T}_X)\}$ to

$$\{w : w = \boldsymbol{b}^\top \boldsymbol{\psi}, \boldsymbol{b}^\top \mathbf{W} \boldsymbol{b} = 1, \boldsymbol{b} \in \mathbb{R}^{N\times 1}\} = \{w : w = \boldsymbol{b}^\top \mathbf{W}^{-1/2} \boldsymbol{\psi}, \boldsymbol{b}^\top \boldsymbol{b} = 1, \boldsymbol{b} \in \mathbb{R}^{N\times 1}\}$$

with invertible and symmetric matrix

$$\mathbf{W} = \left[\int_{\mathbb{T}_X} \psi_{l_1} \psi_{l_2}\right]_{1\le l_1, l_2 \le N}. \tag{3.14}$$

Maximization of $\widetilde{T}^*_{1,\alpha}(w)$ (subject to $\|w\|_2 = 1$) is reformulated as the N-dimensional optimization problem

$$\max_{\boldsymbol{b}\in\mathbb{R}^{N\times 1}} (\boldsymbol{b}^\top \mathbf{W}^{1/2} \widehat{\mathbf{C}}_{\mathrm{c}}^\top \mathbf{Y}_{\mathrm{c}})^2 (\boldsymbol{b}^\top \mathbf{W}^{1/2} \widehat{\mathbf{C}}_{\mathrm{c}}^\top \widehat{\mathbf{C}}_{\mathrm{c}} \mathbf{W}^{1/2} \boldsymbol{b})^{\alpha/(1-\alpha)-1} \tag{3.15}$$

subject to $\boldsymbol{b}^\top \boldsymbol{b} = 1$, where, with $\hat{\boldsymbol{c}}_i$ (3.11),

$$\widehat{\mathbf{C}}_{\mathrm{c}} = \left[\hat{\boldsymbol{c}}_1 - \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{c}}_i, \ldots, \hat{\boldsymbol{c}}_n - \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{c}}_i \right]^\top$$

and

$$\boldsymbol{Y}_{\mathrm{c}} = \left[Y_1 - \frac{1}{n}\sum_{i=1}^{n} Y_i, \ldots, Y_n - \frac{1}{n}\sum_{i=1}^{n} Y_i \right]^\top .$$

Note that the solution to (3.15) is necessarily located in the row space of $\widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}$, i.e., the search region is further restricted to $\{w : w = \boldsymbol{b}^\top \mathbf{V}^\top \mathbf{W}^{-1/2}\boldsymbol{\psi}, \boldsymbol{b}^\top \boldsymbol{b} = 1, \boldsymbol{b} \in \mathbb{R}^{r\times 1}\}$, where $r = \mathrm{rank}(\widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}) \leq \min\{N, n\}$ and $N \times r$ matrix $\mathbf{V}$ comes from the thin singular value decomposition (thin SVD) of $\widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}$: $\widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}$ is decomposed into $\mathbf{U}\mathbf{R}\mathbf{V}^\top$ with diagonal invertible $r \times r$ square matrix $\mathbf{R}$ and semi-orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$. In this way the N-dimensional optimization (3.15) may be reduced to be of r dimensions. Write $\mathbf{G}_1 = \mathbf{U}\mathbf{R}$. The estimator for the first FC basis function then takes the form

$$\tilde{w}_{1,\alpha} = \boldsymbol{b}_{1,\alpha}^\top \mathbf{V}^\top \mathbf{W}^{-1/2}\boldsymbol{\psi}$$

in which

$$\boldsymbol{b}_{1,\alpha} = \underset{\boldsymbol{b}\in\mathbb{R}^{r\times 1}:\boldsymbol{b}^\top\boldsymbol{b}=1}{\arg\max} (\boldsymbol{b}^\top \mathbf{G}_1^\top \boldsymbol{Y}_{\mathrm{c}})^2 (\boldsymbol{b}^\top \mathbf{G}_1^\top \mathbf{G}_1 \boldsymbol{b})^{\alpha/(1-\alpha)-1}.$$

Subsequently and successively, for $j \geq 2$, given r vectors $b_{1,\alpha}, \ldots, b_{j-1,\alpha}$, we just have to replace previous $\mathbf{G}_1$ with deflated $\mathbf{G}_j = \mathbf{P}_{j-1}\mathbf{G}_1$, where $\mathbf{P}_0 = \mathbf{I}_n$, and $\mathbf{P}_{j-1} = \mathbf{I}_n - \mathbf{H}_{j-1}(\mathbf{H}_{j-1}^\top \mathbf{H}_{j-1})^{-1}\mathbf{H}_{j-1}^\top$ is the projection matrix associated with the orthogonal complement of column space of

$$\mathbf{H}_{j-1} = \widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}[\mathbf{V}\boldsymbol{b}_{1,\alpha}, \ldots, \mathbf{V}\boldsymbol{b}_{j-1,\alpha}] = \mathbf{U}\mathbf{R}[\boldsymbol{b}_{1,\alpha}, \ldots, \boldsymbol{b}_{j-1,\alpha}].$$

Namely, for all j,

$$\tilde{w}_{j,\alpha} = \boldsymbol{b}_{j,\alpha}^\top \mathbf{V}^\top \mathbf{W}^{-1/2}\boldsymbol{\psi}, \tag{3.16}$$

where

$$\begin{aligned}\boldsymbol{b}_{j,\alpha} &= \underset{\boldsymbol{b}\in\mathbb{R}^{r\times 1}:\boldsymbol{b}^\top\boldsymbol{b}=1}{\arg\max} (\boldsymbol{b}^\top \mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}})^2 (\boldsymbol{b}^\top \mathbf{G}_j^\top \mathbf{G}_j \boldsymbol{b})^{\alpha/(1-\alpha)-1} \\ &= \left\{ \boldsymbol{Y}_{\mathrm{c}}^\top \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j + \delta_{j,\alpha}^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-2} \mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}} \right\}^{-1/2} (\mathbf{G}_j^\top \mathbf{G}_j + \delta_{j,\alpha}^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-1} \mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}} \end{aligned} \tag{3.17}$$

is deduced from [13, Proposition 2.1]; and $\zeta_{j,\alpha}$ is the largest eigenvalue of $\mathbf{G}_j^\top \mathbf{G}_j$. The only unknown item in (3.17), viz. $\delta_{j,\alpha}$, is the maximizer of the univariate function

$$\begin{aligned}\widetilde{Q}_{j,\alpha}(\delta) = &\left\{\boldsymbol{Y}_{\mathrm{c}}^\top \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j + \delta^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-1}\mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}}\right\}^2 \\ &\times \left\{\boldsymbol{Y}_{\mathrm{c}}^\top \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j + \delta^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-2}\mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}}\right\}^{\alpha/(\alpha-1)} \\ &\times \left\{\boldsymbol{Y}_{\mathrm{c}}^\top \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j + \delta^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-1}\mathbf{G}_j^\top \mathbf{G}_j (\mathbf{G}_j^\top \mathbf{G}_j + \delta^{-1}\zeta_{j,\alpha}\mathbf{I}_r)^{-1}\mathbf{G}_j^\top \boldsymbol{Y}_{\mathrm{c}}\right\}^{\alpha/(1-\alpha)-1}.\end{aligned}$$

confined within $(-1,\infty)\setminus\{0\}$. Here $\widetilde{Q}_{j,\alpha}(\delta)$ is an empirical counterpart of (2.17). We expect to implement this maximization using an arbitrary computer algebra system.

Fixing p, we proceed to derive an estimator for $\beta_{p,\alpha}$ (2.8),

$$\tilde{\beta}_{p,\alpha} = [\tilde{w}_{1,\alpha},\ldots,\tilde{w}_{p,\alpha}](\mathbf{H}_p^\top \mathbf{H}_p)^{-1}\mathbf{H}_p^\top \boldsymbol{Y}_{\mathrm{c}} = \boldsymbol{\psi}^\top \mathbf{W}^{-1/2}\mathbf{V}[\boldsymbol{b}_{1,\alpha},\ldots,\boldsymbol{b}_{p,\alpha}](\mathbf{H}_p^\top \mathbf{H}_p)^{-1}\mathbf{H}_p^\top \boldsymbol{Y}_{\mathrm{c}}.$$

Remark 3.1. Despite the possible ambiguity in representing $\boldsymbol{b}_{j,\alpha}$ in (3.17), the consistency of $\tilde{w}_{j,\alpha}$ (3.16) is not affected, as long as (C.A.2.6) in Appendix A.2 is fulfilled; refer to Remark 2.1.

Like the observed training trajectories X_i's, the new trajectory to be classified, X^*, is discretely observed. The complete trajectory has to be estimated by

$$\widehat{X}^* = \hat{\boldsymbol{c}}^{*\top}\boldsymbol{\psi},$$

where $\hat{\boldsymbol{c}}^*$ comes from applying B-spline smoothing to $X^*(t_1),\ldots,X^*(t_M)$. Let n_0 (resp. n_1) denote the number of training trajectories belonging to Π_0 (resp. Π_1). Estimating mean functions $\mu_X^{[k]}$ by

$$\hat{\mu}_X^{[k]} = \frac{1}{n_k}\sum_{i=1}^n \widehat{X}_i \mathbb{1}(X_i \in \Pi_k) = \frac{1}{n_k}\sum_{i=1}^n \hat{\boldsymbol{c}}_i^\top \boldsymbol{\psi}\mathbb{1}(X_i \in \Pi_k), \tag{3.18}$$

the empirical CCC-Q and -L are then given by, respectively,

$$\begin{aligned}\widehat{\mathcal{D}}_Q(\widehat{X}^* \mid \tilde{\beta}_{p,\alpha}) = &\hat{\sigma}_{[1]}^{-2}(\tilde{\beta}_{p,\alpha})\left\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}(\widehat{X}^* - \hat{\mu}_X^{[1]})\right\}^2 \\ &- \hat{\sigma}_{[0]}^{-2}(\tilde{\beta}_{p,\alpha})\left\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}(\widehat{X}^* - \hat{\mu}_X^{[0]})\right\}^2 + 2\ln\frac{n_0\hat{\sigma}_{[1]}(\tilde{\beta}_{p,\alpha})}{n_1\hat{\sigma}_{[0]}(\tilde{\beta}_{p,\alpha})},\end{aligned} \tag{3.19}$$

and

$$\begin{aligned}\widehat{\mathcal{D}}_L(\widehat{X}^* \mid \tilde{\beta}_{p,\alpha}) = &\hat{\sigma}_{\mathrm{pool}}^{-2}(\tilde{\beta}_{p,\alpha})\left\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}(\widehat{X}^* - \hat{\mu}_X^{[1]})\right\}^2 \\ &- \hat{\sigma}_{\mathrm{pool}}^{-2}(\tilde{\beta}_{p,\alpha})\left\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}(\widehat{X}^* - \hat{\mu}_X^{[0]})\right\}^2 + 2\ln\frac{n_0}{n_1},\end{aligned} \tag{3.20}$$

where

$$\hat{\sigma}^2_{\text{pool}}(\omega) = (n-2)^{-1} \sum_{k=0}^{1} \sum_{i=1}^{n} \left\{ \int_{\mathbb{T}_X} \omega(\widehat{X}_i - \hat{\mu}_X^{[k]}) \right\}^2 \mathbb{1}(X_i \in \Pi_k), \tag{3.21}$$

and

$$\hat{\sigma}^2_{[k]}(\omega) = (n_k - 1)^{-1} \sum_{i=1}^{n} \left\{ \int_{\mathbb{T}_X} \omega(\widehat{X}_i - \hat{\mu}_X^{[k]}) \right\}^2 \mathbb{1}(X_i \in \Pi_k), \quad k = 0, 1. \tag{3.22}$$

Proposition 3.4. *Fix $p \in \mathbb{Z}^+$ and $\alpha \in [0, 1)$ and assume (C.A.2.3)–(C.A.2.6) in Appendix A.2. Empirical classifier $\widehat{\mathcal{D}}_Q(\widehat{X}^* \mid \tilde{\beta}_{p,\alpha})$ (3.19) (resp. $\widehat{\mathcal{D}}_L(\widehat{X}^* \mid \tilde{\beta}_{p,\alpha})$ (3.20)) converges to its population version $\mathcal{D}_Q(X^* \mid \beta_{p,\alpha})$ (3.5) (resp. $\mathcal{D}_L(X^* \mid \beta_{p,\alpha})$ (3.7)) in probability as n diverges. Further, if (C.A.2.2) holds too, then*

$$\lim_{p\to\infty} \lim_{n\to\infty} \text{err}\{\widehat{\mathcal{D}}_L(\widehat{X}^* \mid \tilde{\beta}_{p,\alpha})\} = 0.$$

3.2.2 Tuning parameters

Analogous to Section 2.4.1, the generalized cross-validation (GCV) tuning scheme is employed here, i.e., minimize (w.r.t. (p, α))

$$\text{GCV}(p, \alpha) = \frac{\sum_{i=1}^{n} \left[Y_i - \mathbb{1}\{\widehat{\mathcal{D}}_Q(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) < 0\} \right]^2}{(n - p - 2)^2}$$

for CCC-Q or

$$\text{GCV}(p, \alpha) = \frac{\sum_{i=1}^{n} \left[Y_i - \mathbb{1}\{\widehat{\mathcal{D}}_L(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) < 0\} \right]^2}{(n - p - 2)^2}$$

for CCC-L, where the digit 2 in parenthesis in the denominator corresponds to the number of populations. Algorithm 3.1 details the implementation. As illustrated in Section 2.4.1, using a random nonrectangular grid is accompanied with little loss in prediction accuracy compared to using a fixed regular search grid. We adjust $p_{\max}$ at (2.19) by concentrating it on within-group covariance v_X^W (3.1), viz.

$$p_{\max} = \min \left\{ j_0 \in \mathbb{Z}^+ : \sum_{j=1}^{j_0} \hat{\lambda}_j^W \Big/ \sum_{j=1}^{\infty} \hat{\lambda}_j^W \geq 99\% \right\},$$

with $\hat{\lambda}_j^W$ estimating the jth top eigenvalue of v_X^W at (3.1).

Algorithm 3.1 CCC tuned via GCV

$p_{\max} \leftarrow$ upper bound of number of FC basis functions.
for α in a finite set **do**
  for p from 1 to $p_{\max,\alpha}$ **do**
    if $p = 1$ **then**
      $\mathbf{P} \leftarrow \mathbf{I}_n$.
      thin SVD of $\widehat{\mathbf{C}}_{\mathrm{c}}\mathbf{W}^{1/2}$: $\mathbf{URV}^\top$.
      $\mathbf{G}_1 \leftarrow \mathbf{UR}$.
    else
      $\mathbf{P} \leftarrow \mathbf{P}\{\mathbf{I}_n - \mathbf{G}_{p-1}\boldsymbol{b}_{p-1,\alpha}(\boldsymbol{b}_{p-1,\alpha}^\top\mathbf{G}_{p-1}^\top\mathbf{G}_{p-1}\boldsymbol{b}_{p-1,\alpha})^{-1}\boldsymbol{b}_{p-1,\alpha}^\top\mathbf{G}_{p-1}^\top\}$.
    end if
    $\mathbf{G}_p \leftarrow \mathbf{PG}_1$.
    $\zeta \leftarrow$ largest eigenvalue of $\mathbf{G}_p^\top\mathbf{G}_p$.
    $\mathbf{L}(\delta) \leftarrow (\mathbf{G}_p^\top\mathbf{G}_p + \delta^{-1}\zeta\mathbf{I}_r)^{-1}$.
    $Q(\delta) \leftarrow \{\boldsymbol{Y}_{\mathrm{c}}^\top\mathbf{G}_p\mathbf{L}(\delta)\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}}\}^2\{\boldsymbol{Y}_{\mathrm{c}}^\top\mathbf{G}_p\mathbf{L}^2(\delta)\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}}\}^{\alpha/(1-\alpha)}$
      $\times\{\boldsymbol{Y}_{\mathrm{c}}^\top\mathbf{G}_p\mathbf{L}(\delta)\mathbf{G}_p^\top\mathbf{G}_p\mathbf{L}(\delta)\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}}\}^{\alpha/(1-\alpha)-1}$.
    $\delta_{p,\alpha} \leftarrow \arg\min_{\delta\in(-1,0)\cup(0,\infty)} -\ln Q(\delta)$.
    $\boldsymbol{b}_{p,\alpha} \leftarrow \mathbf{L}(\delta_{p,\alpha})\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}}/\{\boldsymbol{Y}_{\mathrm{c}}^\top\mathbf{G}_p\mathbf{L}^2(\delta_{p,\alpha})\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}}\}^{1/2}$.
    $\hat{w}_{p,\alpha} \leftarrow \boldsymbol{b}_{p,\alpha}^\top\mathbf{V}^\top\mathbf{W}^{-1/2}\boldsymbol{\psi}$.
    if $p = 1$ **then**
      $\tilde{\beta}_{p,\alpha} \leftarrow n^{-1/2}(\boldsymbol{b}_{p,\alpha}^\top\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}})(\boldsymbol{b}_{p,\alpha}^\top\mathbf{G}_p^\top\mathbf{G}_p\boldsymbol{b}_{p,\alpha})^{-1/2}\hat{w}_{p,\alpha}$.
    else
      $\tilde{\beta}_{p,\alpha} \leftarrow \tilde{\beta}_{p-1,\alpha} + n^{-1/2}(\boldsymbol{b}_{p,\alpha}^\top\mathbf{G}_p^\top\boldsymbol{Y}_{\mathrm{c}})(\boldsymbol{b}_{p,\alpha}^\top\mathbf{G}_p^\top\mathbf{G}_p\boldsymbol{b}_{p,\alpha})^{-1/2}\hat{w}_{p,\alpha}$.
    end if
    for i from 1 to n **do**
      $\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i \leftarrow n^{-1/2}\left\{\sum_{j=1}^p(\boldsymbol{b}_{j,\alpha}^\top\mathbf{G}_j\boldsymbol{Y}_{\mathrm{c}})(\boldsymbol{b}_{j,\alpha}^\top\mathbf{G}_j^\top\mathbf{G}_j\boldsymbol{b}_{j,\alpha})^{-1/2}\boldsymbol{b}_{j,\alpha}^\top\mathbf{V}_j^\top\right\}\mathbf{W}^{1/2}\hat{c}_i$.
    end for
    $\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\hat{\mu}_X^{[k]} \leftarrow \mathrm{mean}\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i\mathbb{1}(X_i \in \Pi_k) : i = 1,\ldots,n\}, \quad k = 0, 1.$
    $\hat{\sigma}_{[k]}^2(\tilde{\beta}_{p,\alpha}) \leftarrow \mathrm{var}\{\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i\mathbb{1}(X_i \in \Pi_k) : i = 1,\ldots,n\}, \quad k = 0, 1.$
    $\hat{\sigma}_{\mathrm{pool}}^2(\tilde{\beta}_{p,\alpha}) \leftarrow (n-2)^{-1}\{(n_0-1)\hat{\sigma}_{[0]}^2(\tilde{\beta}_{p,\alpha}) + (n_1-1)\hat{\sigma}_{[1]}^2(\tilde{\beta}_{p,\alpha})\}$.
    for i from 1 to n **do**
      $\widehat{\mathcal{D}}_Q(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) \leftarrow \hat{\sigma}_{[1]}^{-2}(\tilde{\beta}_{p,\alpha})(\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i - \int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\hat{\mu}_X^{[1]})^2$
        $-\hat{\sigma}_{[0]}^{-2}(\tilde{\beta}_{p,\alpha})(\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i - \int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\hat{\mu}_X^{[0]})^2 + 2\ln\frac{n_0\hat{\sigma}_{[1]}(\tilde{\beta}_{p,\alpha})}{n_1\hat{\sigma}_{[0]}(\tilde{\beta}_{p,\alpha})}$.
      $\widehat{\mathcal{D}}_L(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) \leftarrow \hat{\sigma}_{\mathrm{pool}}^{-2}(\tilde{\beta}_{p,\alpha})(\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i - \int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\hat{\mu}_X^{[1]})^2$
        $-\hat{\sigma}_{\mathrm{pool}}^{-2}(\tilde{\beta}_{p,\alpha})(\int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\widehat{X}_i - \int_{\mathbb{T}_X}\tilde{\beta}_{p,\alpha}\hat{\mu}_X^{[0]})^2 + 2\ln(n_0/n_1)$.
    end for
  end for
  $\mathrm{GCV}(p,\alpha) \leftarrow \sum_{i=1}^n[Y_i - \mathbb{1}\{\widehat{\mathcal{D}}_Q(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) < 0\}]^2/(n-p-2)^2$
    or $\sum_{i=1}^n[Y_i - \mathbb{1}\{\widehat{\mathcal{D}}_L(\widehat{X}_i \mid \tilde{\beta}_{p,\alpha}) < 0\}]^2/(n-p-2)^2$.
end for
$(p_{\mathrm{opt}}, \alpha_{\mathrm{opt}}) \leftarrow \arg\min_{(p,\alpha)}\mathrm{GCV}(p,\alpha)$.

3.3 Numerical illustration

In spite of theoretical arguments illustrating the asymptotically perfect classification of CCC-L in specific cases, we were still in need of more evidences to support our proposals, especially CCC-Q. Therefore, we resorted to numerical studies so as to compare the performance of PCC, PLCC and the two CCC classifiers in finite-sample applications. Candidate pools for tuning parameters were set up as follows. For CCC classifiers, combinations of tuning parameters came from $\{(p, \alpha_j) : 1 \leq p \leq p_{j,\max}, 1 \leq j \leq 13\}$ with $\{\alpha_1, \ldots, \alpha_{13}\} = \{0 \times 10^{-1}, \ldots, 9 \times 10^{-1}, 1 - 10^{-2}, \ldots, 1 - 10^{-4}\}$ and $p_{j,\max}$ as constructed in Section 2.4.1. PLCC and PCC both took $\{1, \ldots, p_{\max}\}$ as the candidate pool for their p. Boxplots were all created via R-package `ggplot2` [99]. There were also tables summarizing statistics of the box plots. Formal comparisons among means were carried out via the paired t-test (resp. the corrected resampled t-test proposed by [67]) for independent (resp. resampled) samples involved in the upcoming studies with simulated data (resp. real-world data).

3.3.1 Simulation study

We generated $R = 200$ samples, each containing $n = 200$ curves X_i, $i = 1, \ldots, 200$. In each sample, we randomly preserved 80% of the curves for training and used the remaining 20% for testing. Each curve was observed at 101 equally spaced points in $\mathbb{T}_X = [0, 1]$, i.e., $\{i/100 : i = 0, \ldots, 100\}$; curves were generated in an iid way as

$$X_i = \sum_{k=0}^{1} \left(\sum_{j=1}^{5} \lambda_{k,j}^{1/2} Z_{ij} \phi_{kj} + \mu_X^{[k]} \right) \mathbb{1}(X_i \in \Pi_k).$$

Without loss of generality, the difference of two mean functions was set to be exactly $\mu_X^{[1]}$, i.e., $\mu_X^{[0]}(\cdot) \equiv 0$. Instead of a 50/50 mixture of Gaussian processes, we considered a challenging setup: $Z_{ij} \sim \exp(1) - 1$ and $\pi_0 = \Pr(X_i \in \Pi_0) = 80\%$. Although $v_X^{[0]}$ and $v_X^{[1]}$ shared the identical nonzero eigenvalues $(200, 100, 1, 0.2, 0.1)$, they might differ in eigenfunctions; specifically, we took the jth-order shifted Legendre polynomial (scaled to have norm one) [54] as the jth eigenfunction of $v_X^{[0]}$, $j = 1, \ldots, 5$, viz.

$$\begin{aligned}
\phi_{0,1}(t) &= \sqrt{3}(2t - 1), \\
\phi_{0,2}(t) &= \sqrt{5}(6t^2 - 6t + 1), \\
\phi_{0,3}(t) &= \sqrt{7}(20t^3 - 30t^2 + 12t - 1), \\
\phi_{0,4}(t) &= 3(70t^4 - 140t^3 + 90t^2 - 20t + 1),
\end{aligned}$$

and

$$\phi_{0,5}(t) = \sqrt{11}(252t^5 - 630t^4 + 560t^3 - 210t^2 + 30t - 1).$$

Meanwhile we accounted for two sorts of combinations of $\mu_X^{[1]}$ and $\phi_{1,j}$:

$$\mu_X^{[1]} = \rho\lambda_{1,1}^{1/2}\phi_{1,1} \quad \text{with} \quad \phi_{1,j} = \phi_{0,j} \tag{3.23}$$

and

$$\mu_X^{[1]} = \rho\lambda_{1,3}^{1/2}\phi_{1,3} \quad \text{with} \quad \phi_{1,j} = \phi_{0,5-j}. \tag{3.24}$$

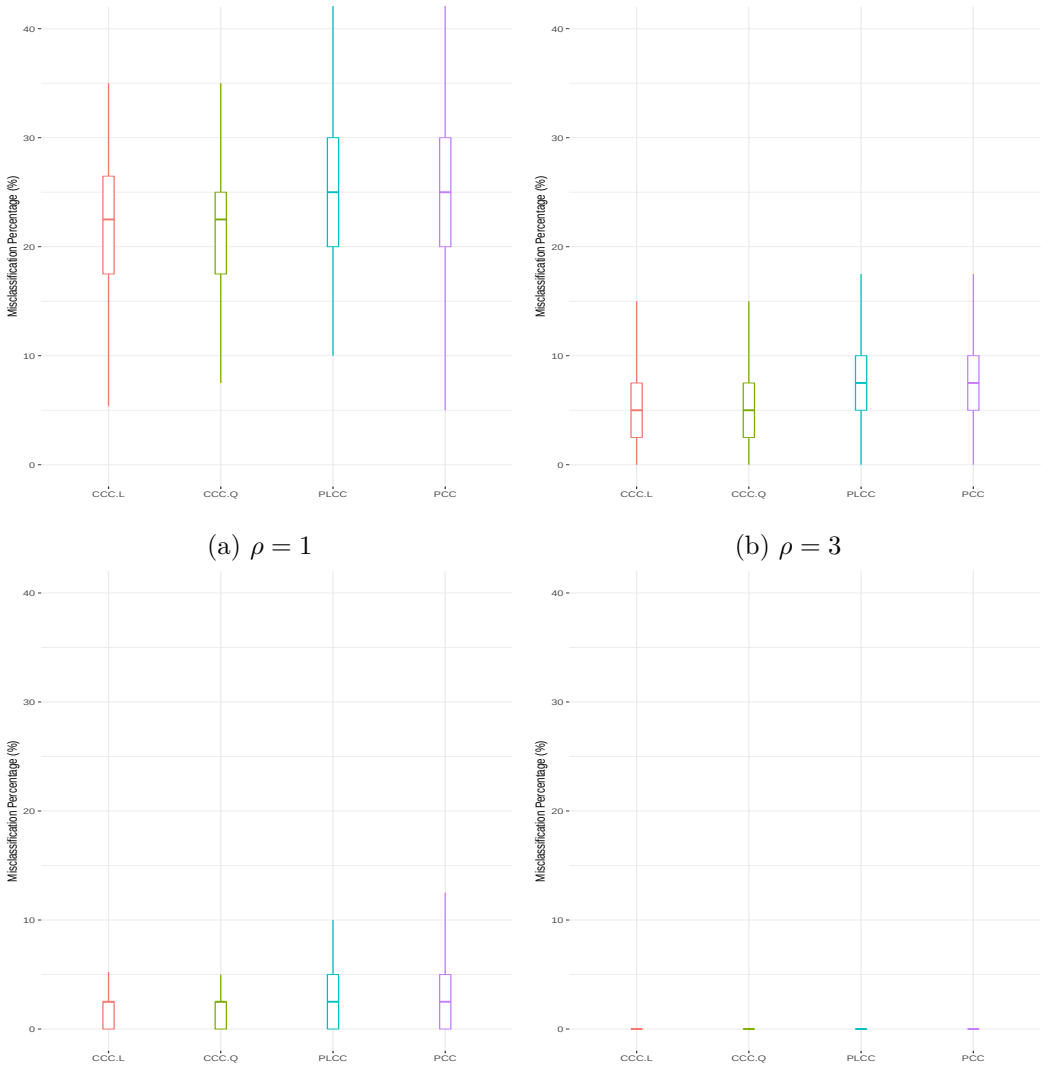
In both scenarios, ρ $(= 1, 3, 5, 10)$ controlled the magnitude of gap between $\mu_X^{[1]}$ and $\mu_X^{[0]}$ relative to $\lambda_{1,j}^{1/2}$ and also the ratio of the between-group variation to the within-group one. These eight combinations in total of settings would help us clarify the joint impacts of the direction and magnitude of $\mu_X^{[1]} - \mu_X^{[0]}$ on misclassification.

Design (3.23) favored the identification of the direction of $\mu_X^{[1]} - \mu_X^{[0]}$ as it was parallel to the first eigenfunction not only of v_X (1.2) but also of v_X^W (3.1); for each value of ρ, all the four classifiers performed fairly close to each other. As indicated by Figure 3.1 and the first four rows of Table 3.1, larger ρ typically meant more separable subpopulations and hence the overall error rate became lower and lower with increasing ρ. In the case of $\rho = 1$, the two data clouds were likely to have substantial overlap, resulting in average error rates of over 20%; all the classifiers achieved perfect performance when $\rho = 5$ or 10.

It was a different story in design (3.24) which restricted $\mu_X^{[1]} - \mu_X^{[0]}$ to be parallel to ϕ_{03}, the least important eigenfunction of

$$\begin{aligned} v_X^W(s,t) = 160.02(&\phi_{0,1}(s)\phi_{0,1}(t) + 80.04\phi_{0,2}(s)\phi_{0,2}(t) \\ &+ 40.08\phi_{0,5}(s)\phi_{0,5}(t) + 20.16\phi_{0,4}(s)\phi_{0,4}(t) + \phi_{0,3}(s)\phi_{0,3}(t). \end{aligned}$$

In this case, focused only on decomposing v_X^W, PCC probably failed to extract the correct direction of $\mu_X^{[1]} - \mu_X^{[0]}$ and naturally yielded more misclassification regardless of ρ. Moreover, $v_X^{[1]}$ shared the same eigenfunctions with $v_X^{[0]}$ but in a reversed order; they were no longer equal at all. This setting violated the assumption of CCC-L and PLCC, through these two classifiers had little problem in recovering $\phi_{0,3}$. Consequently, when $\rho = 1$, CCC-Q significantly outperformed the other three classifiers; see Figure 3.2a. When ρ grew to 3, the difference in sub-covariance did not matter as much as in the case of $\rho = 1$. That was why the performance of CCC-L and PLCC was improved. When we further enlarged the value of ρ, the two groups became clearly identifiable for CCC-L, -Q and PLCC and few errors were committed by these three classifiers.



(a) $\rho = 1$ (b) $\rho = 3$

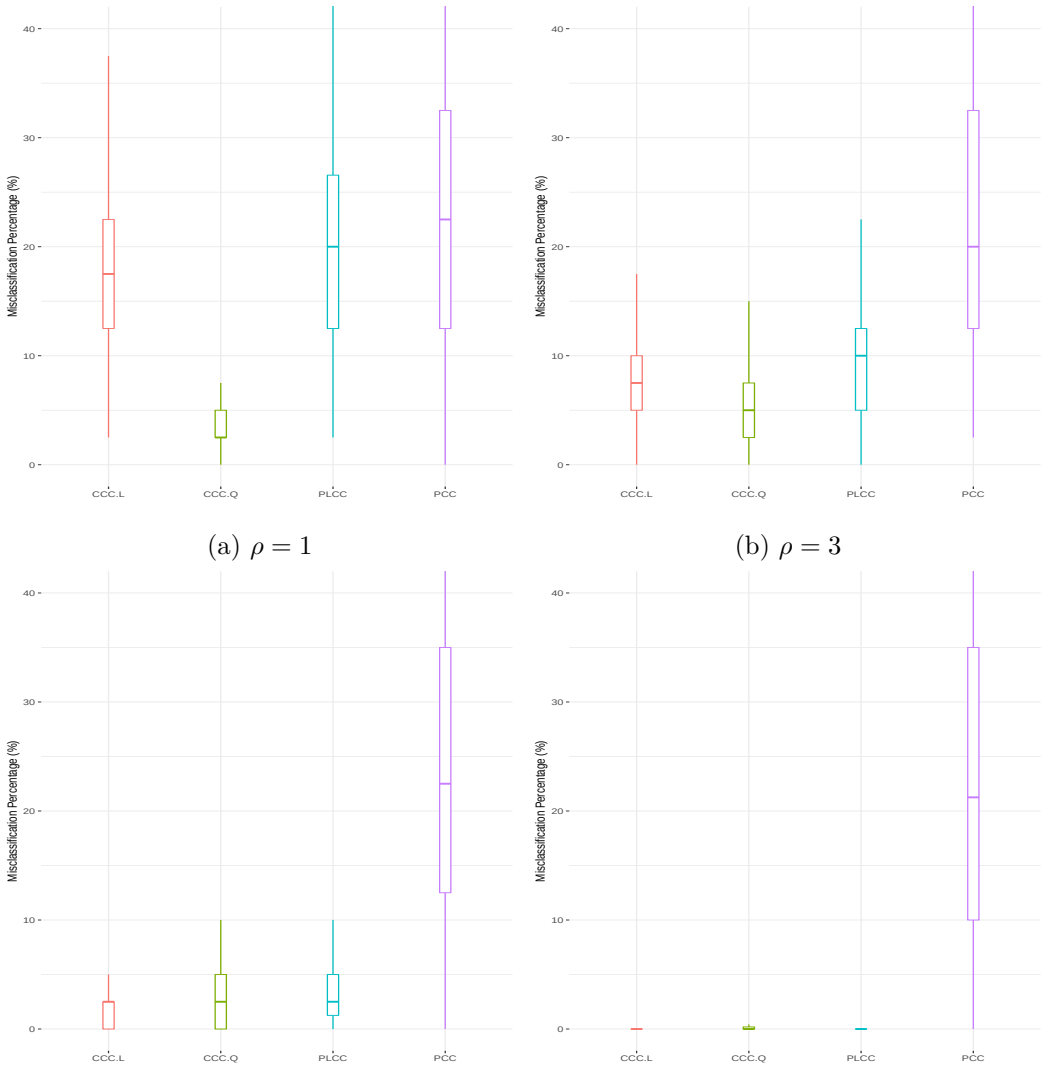
(c) $\rho = 5$ (d) $\rho = 10$

Figure 3.1: Boxplots of misclassification percentage (MP, %) for simulation (3.23). In each panel, the four boxes, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC and PCC, respectively and use identical scales.

In summary, supervised options, CCC-L, -Q and PLCC, were more likely to capture the direction of $\mu_X^{[1]} - \mu_X^{[0]}$. Classifiers holding the equal variance assumption might fail if the two groups were close to each other and equipped with different sub-covariance functions (e.g., design (3.24) with $\rho = 1$); in contrast, under this circumstance, CCC-Q stood out.

3.3.2 Real data application

For each dataset analyzed below, we repeat 200 times a random split with the ratio 8 : 2, i.e., at each repetition, we train the classifiers with 80% of the data points and test them on the remaining 20%. Table 3.2 summarizes the means and standard deviations of MP.



(a) $\rho = 1$ (b) $\rho = 3$

(c) $\rho = 5$ (d) $\rho = 10$

Figure 3.2: Boxplots of MP (%) for simulation (3.24). In each panel, the four boxes, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC and PCC, respectively. The four subfigures come with identical scales.

First we revisit the Tecator™ data in Section 2.5.2 by categorizing all meat samples into two groups: Π_1 consisted of meat samples with protein content less than 16% and the rest constituted Π_0. The 240 spectrum curves (or their second order derivative curves as recommended by [33, Section 7.2.2] and [40]) were regarded as functional covariates in the study. For both sorts of curves, the two CCC subtypes, especially CCC-Q, showed error rates considerably lower than those from PLCC and PCC. We speculate that this phenomenon for PLCC (resp. PCC) was caused by the poor performance of $\|\beta_{p,\mathrm{FPLS}}\|_2^2$ (resp. $\|\beta_{p,\mathrm{WFPC}}\|_2^2$) in replacing $\mathrm{var}(\int_{\mathbb{T}_X} X\beta_{p,\mathrm{FPLS}} \mid X \in \Pi_k)$ (resp. $\mathrm{var}(\int_{\mathbb{T}_X} X\beta_{p,\mathrm{WFPC}} \mid X \in \Pi_k)$).

Table 3.1: Average MP (%) for simulation studies (with standard deviations in parentheses) corresponding to different classifiers. Row minimums are underlined.

Design	ρ	CCC-L	CCC-Q	PLCC	PCC
(3.23)	1	21.71 (6.662)	21.35 (6.351)	25.68 (7.423)	25.79 (7.550)
(3.23)	3	5.906 (4.321)	6.009 (4.211)	6.831 (4.227)	6.812 (4.156)
(3.23)	5	2.159 (2.523)	2.125 (2.486)	2.650 (2.932)	2.688 (2.996)
(3.23)	10	.1495 (.5873)	.1492 (.5872)	.2188 (.7026)	.2125 (.6990)
(3.24)	1	17.30 (6.663)	3.969 (2.937)	19.70 (9.255)	23.98 (14.52)
(3.24)	3	7.487 (4.459)	4.877 (3.749)	9.238 (4.682)	23.61 (14.55)
(3.24)	5	2.337 (2.215)	2.611 (2.675)	3.323 (2.849)	25.41 (15.79)
(3.24)	10	.3496 (.8291)	.3684 (.8126)	.4250 (.9823)	23.38 (14.78)

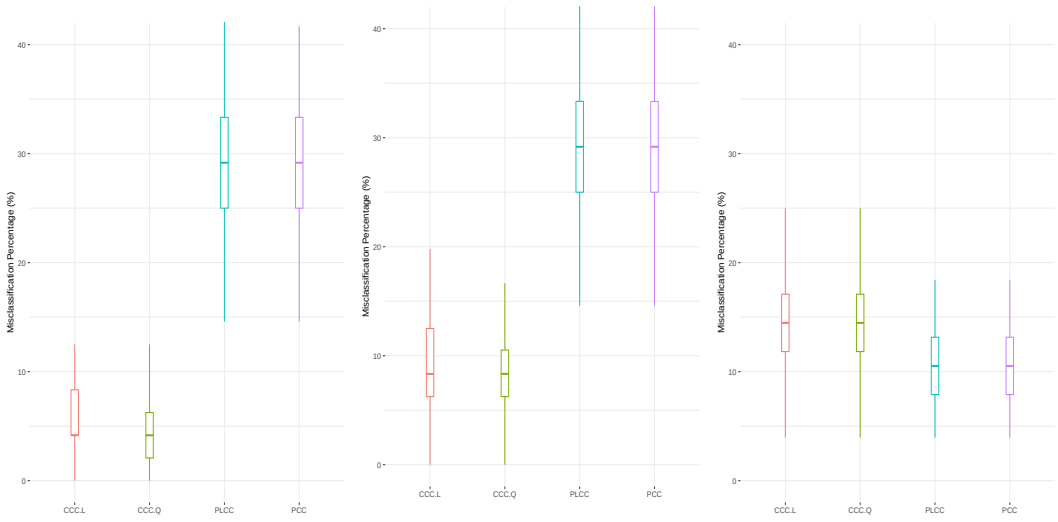
Table 3.2: Average MP (%) for real applications (with standard deviations in parentheses) corresponding to different classifiers. Row minimums are underlined.

	CCC-L	CCC-Q	PLCC	PCC
Tecator™ (original)	5.521 (3.254)	4.562 (2.766)	29.57 (6.447)	28.95 (6.756)
Tecator™ (2nd order derivative)	9.288 (3.888)	8.897 (3.838)	29.79 (5.841)	29.79 (5.841)
DTI	14.32 (3.987)	14.33 (4.014)	10.69 (3.276)	10.69 (3.276)

We considered a diffusion tensor imaging (DTI) study in the second example. DTI is a modern tool in mapping white matter tractography in brains. In tractography, the fractional anisotropy (FA), a scalar ranging from 0 to 1 and reflecting the fiber density, axonal diameter and myelination, is measured at a specific spot in the white matter. Along a tract of interest, FAs form a tract FA profile. A study on these profiles may help people quantify pathological changes resulted from multiple sclerosis (MS) [38], an immune disorder affecting the central nervous system [9]. Dataset `DTI` from R-package `classiFunc` [64] contains tract FA profiles for corpus callosum of 382 subjects comprised of healthy people (Π_0) and MS patients (Π_1). By classifying these profiles, we tried to judge the status of each subject: healthy or suffering from MS. As displayed in Table 3.2 and Figure 3.3c, PLCC and PCC appeared to perform slightly better than CCC subtypes. However, we could not formally reject the null hypothesis that mean error rates of all the four classifiers reached a tie (according to the corrected resampled t-test); that is, the four classifiers had misclassification rates of a similar level.

3.4 Conclusion and discussion

We propose two subtypes of CCC classifiers, CCC-L and -Q, for binary classification of curves. Theoretically, under certain circumstances, CCC-L enjoys the (asymptotic) zero misclassification regardless of the distribution assumption, while, in empirical studies, CCC-Q seems preferred due to a generally more competitive output. Once regularity conditions are met, our implementation results in empirical classifiers which are consistent for their theoretical counterparts, for the case of "fixed p and infinite n".



(a) Tecator™ (original)

(b) Tecator™ (2nd order derivative)

(c) DTI

Figure 3.3: Boxplots of MP (%) for real applications. The top two panels reflect respective results with original and second order derivative curves. In each panel, the four boxes, from left to right, correspond to classifiers CCC-L, CCC-Q, PLCC and PCC, respectively. The three subfigures are displayed with identical scales.

PLCC is slightly less time-consuming than PCC since the former one does not carry out SVD. By contrast, the time-consumption of CCC classifiers grows rapidly as the size of training set increases, which agrees with the discussion in [109, Section 6]. When applied to simulated datasets (with $n = 200$), and despite one more hyper-parameter being involved, CCC classifiers are very competitive in computing time. Unfortunately they become far more time-consuming when n is around 400 (e.g., in the case of analyzing DTI data); see Table 3.3.

In the numerical experiments in Section 3.3, we do gain some benefits from the introduction of the supervision controller α. Nevertheless, one cannot be too optimistic; actually, tuning one more parameter may yield more variation and even bias. Evidence for this possibility is offered by the unsatisfactory performance of CCC classifiers in the application to DTI data; more evidence is encountered when we employ these four classifiers to analyze simulated example 3 in [28], where PLCC overwhelms the other three.

Regarding a further extension to the functional classification with K (≥ 3) classes, one possible strategy is to carry out binary classifiers repeatedly. To be explicit, for a newcomer X^*, each time we only consider two distinct labels and then assign either label to it, or equivalently, throw a vote for either of the two labels. After all $\binom{K}{2}$ binary classifications, the label that wins the most votes is eventually assigned to X^*.

Table 3.3: Time consumed (in seconds) by each method in each numerical study in Section 3.3 (running on a desktop with R [74], Rstudio [81], Intel® Core™ i5-7500 CPU @ 2×3.40 GHz and 8 GB RAM). Two CCC subtypes share only one column because they differ in only a few steps and consume very similar amounts of time. Row minimums are underlined.

Scenario	CCC-L (or -Q)	PLCC	PCC
Design (3.23) & $\rho = 1$	<u>205.9</u>	590.6	714.5
Design (3.23) & $\rho = 3$	<u>204.4</u>	584.5	705.1
Design (3.23) & $\rho = 5$	<u>205.0</u>	584.6	702.6
Design (3.23) & $\rho = 10$	<u>204.4</u>	577.9	703.8
Design (3.24) & $\rho = 1$	<u>252.0</u>	602.1	721.7
Design (3.24) & $\rho = 3$	<u>252.7</u>	602.1	718.6
Design (3.24) & $\rho = 5$	<u>251.8</u>	601.1	717.7
Design (3.24) & $\rho = 10$	<u>253.8</u>	600.4	714.9
Tecator™ (original)	<u>293.7</u>	524.6	631.9
Tecator™ (2nd order derivative)	<u>479.8</u>	557.6	662.8
DTI	1723	<u>663.1</u>	729.6

Chapter 4

Partial least squares for sparsely observed curves with measurement errors

4.1 Introduction

During past decades, numerous efforts have been put on FPC and FPLS; we refer readers to, e.g., [31], for a general review of both techniques. Nevertheless, most of these works are designed only for dense settings, i.e., realizations of X are supposed to be (almost) completely observed, a condition which is not expected to be fulfilled all the time. For example, in typical clinical trials, medical indications cannot be monitored 24/7; instead, participants are required to visit the clinic repeatedly on specific dates. Due to costs and convenience, for each subject, the scheduled visiting frequency is doomed to be sparse. What is even worse is that subjects are more willing to show up on their own basis with frequencies lower and more irregular than scheduled. Another apparent instance exists with missing data problems where a number of recordings may be erased naturally or manually. Meanwhile, in many practical applications, observed curves and responses are often contaminated by measurement errors.

Admitting that sparsity in observation as well as errors in measurement are commonplace, we now suppose $X_i \overset{\text{iid}}{\sim} X$ are unobservable. Instead the ith trajectory is merely measured at L_i time points $T_{i1}, \ldots, T_{iL_i} \in \mathbb{T}_X = [0, 1]$ in a noisy form such that,

$$\widetilde{X}_i(T_{i\ell}) = X_i(T_{i\ell}) + \sigma_e e_{i\ell}, \tag{4.1}$$

where $\sigma_e > 0$ and white noises $e_{i\ell}$ have zero mean and unit variance. We assume that all the time points and error terms are independent across subjects and from each other. More precise description is detailed by Appendix A.3. This joint setup of sparsity and error-in-variable is considered too by existing literature, e.g., [104], [105], [102], and [82].

Remark 4.1. For each i, $T_{i1}, \ldots, T_{iL_i}$ do not need to be ordered increasingly or descendingly. Additionally we suggest not taking $\widetilde{X}_i$ as the sum of X_i and a white noise process, otherwise more mathematical efforts are needed in definition to ensure rigor; instead (4.1) suffices for our proposal which utilizes only random variables $\widetilde{X}_i(T_{i1}), \ldots, \widetilde{X}_i(T_{iL_i})$ and never attempts to approximate integrals involving the entire trajectory $\widetilde{X}_i$.

We now slightly modify the definition of the FPLS basis from (2.3). For brevity in notation, the modified FPLS basis in this chapter remains denoted by $\{w_1, w_2, \ldots\}$. Introducing the cross-covariance function, v_{XY}, such that

$$v_{XY} = v_{XY}(\cdot) = \operatorname{cov}\{Y, X(\cdot)\}, \tag{4.2}$$

an implementation of FPLS ([28, Appendix A.2]) starts with $X^{[1]} = X - \mu_X$, $Y^{[1]} = Y - \mu_Y$ and

$$w_1 = \underset{\|w\|_2=1}{\arg\max} \operatorname{cov}^2\left(Y^{[1]}, \int_{\mathbb{T}_X} X^{[1]} w\right) = \|v_{XY}\|_2^{-1} v_{XY}. \tag{4.3}$$

It then constructs subsequent basis functions in a successive way: given the first $j-1$ basis functions $w_1, \ldots, w_{j-1}$, assuming $v_{XY}^{[j]} \not\equiv 0$, we define

$$w_j = \underset{\|w\|_2=1}{\arg\max} \operatorname{cov}^2\left(Y^{[j]}, \int_{\mathbb{T}_X} X^{[j]} w\right) = \left\|v_{XY}^{[j]}\right\|_2^{-1} v_{XY}^{[j]}, \tag{4.4}$$

where superscript $[j]$ corresponds to the jth least-squares deflation. That is to say,

$$\begin{aligned} X^{[j]} &= X^{[j]}(\cdot) = X^{[j-1]}(\cdot) \\ &\quad - \left(\int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right) \operatorname{var}^{-1}\left(\int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right) \operatorname{cov}\left\{X^{[j-1]}(\cdot), \int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right\}, \quad (4.5) \\ Y^{[j]} &= Y^{[j-1]} \\ &\quad - \left(\int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right) \operatorname{var}^{-1}\left(\int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right) \operatorname{cov}\left(Y^{[j-1]}, \int_{\mathbb{T}_X} X^{[j-1]} w_{j-1}\right), \quad (4.6) \end{aligned}$$

and

$$v_{XY}^{[j]} = v_{XY}^{[j]}(\cdot) = \operatorname{cov}\left\{Y^{[j]}, X^{[j]}(\cdot)\right\}. \tag{4.7}$$

Remark 4.2. As a functional counterpart of Proposition 1.1 in [14], the orthogonality of basis functions, or equivalently, the identity $\int_{\mathbb{T}_X} w_{j_1} w_{j_2} = 0$ if $j_1 \geq 2$ and $j_1 > j_2$, follows from the above algorithm. This property is verified by noting that, from (4.5) and (4.6) (both defined through least squares), the random variable

$$\begin{aligned} \int_{\mathbb{T}_X} X^{[j_1]} w_{j_2} &= \int_{\mathbb{T}_X} X^{[j_2]} w_{j_2} \\ &\quad - \sum_{j=j_2}^{j_1-1} \left(\int_{\mathbb{T}_X} X^{[j]} w_j\right) \operatorname{var}^{-1}\left(\int_{\mathbb{T}_X} X^{[j]} w_j\right) \operatorname{cov}\left(\int_{\mathbb{T}_X} X^{[j]} w_{j_2}, \int_{\mathbb{T}_X} X^{[j]} w_j\right) \end{aligned}$$

is located in $\operatorname{span}(\int_{\mathbb{T}_X} X^{[j_2]} w_{j_2}, \ldots, \int_{\mathbb{T}_X} X^{[j_1-1]} w_{j_1-1})$ whose orthogonal complement contains $Y^{[j_1]}$. In this way, (4.4) and (4.7) jointly imply that, as long as $v_{XY}^{[j_1]} \not\equiv 0$,

$$\int_{\mathbb{T}_X} w_{j_1} w_{j_2} = \left\| v_{XY}^{[j_1]} \right\|_2^{-1} \int_{\mathbb{T}_X} v_{XY}^{[j_1]} w_{j_2} = \left\| v_{XY}^{[j_1]} \right\|_2^{-1} \operatorname{cov}\left(Y^{[j_1]}, \int_{\mathbb{T}_X} X^{[j_1]} w_{j_2} \right) = 0.$$

Remark 4.3. Alternatively, w_j (4.4) is the maximizer of $\operatorname{cov}^2\left(Y, \int_{\mathbb{T}_X} Xw\right)$ subject to $\|w\|_2 = 1$ and $\int_{\mathbb{T}_X} w w_{1,\mathrm{FPLS}} = \cdots = \int_{\mathbb{T}_X} w w_{j-1,\mathrm{FPLS}} = 0$. That is, it differs from $w_{j,\mathrm{FPLS}}$ (2.3) only in the constraints: the former one is constrained by the ordinary orthogonality while the latter one requires orthogonality w.r.t. v_X (1.2).

[50, 104, 105] succeeded in extending (classical) FPC (with dense observation) to the challenging setting in this chapter. Among their proposals, the particular proposal from [104, 105] is abbreviated as PACE. Compared with FPC, FPLS is more adaptive to data, leading to a more parsimonious basis as well as more interpretability [79]; however, to the best of our knowledge, FPLS has few extension applicable to the sparsity setting. In this work, we attempt to fill in this blank by developing a new technique named Partial LEAst Square for Sparsity (PLEASS), aiming at handling both sparse observations and measurement errors simultaneously.

The remainder of this chapter is organized as follows. Section 4.2 describes the implementation procedure for PLEASS. In Section 4.3 we present asymptotic results including not only the consistency of estimators but also confidence intervals for predictions. Section 4.4 applies PACE and PLEASS to both simulated and real datasets and compares their resulting performances. This is followed by concluding remarks in Section 4.5.

4.2 Methodology

4.2.1 Estimation and prediction

As is guaranteed by [28, Theorem 3.2], the true slope β must be located in the closure of $\operatorname{span}(w_1, w_2, \ldots)$; it is hence the limit (as p diverges and in the L^2 sense) of β_p (1.5). With w_j defined in (4.4), β_p in (1.5) and $\eta_p(X^*)$ in (1.6) become

$$\beta_p = \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} [w_1, \ldots, w_p]^\top \tag{4.8}$$

and

$$\eta_p(X^*) = \mu_Y + \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} [\xi_1^*, \ldots, \xi_p^*]^\top, \tag{4.9}$$

respectively, where

$$\boldsymbol{c}_p = \left[\int_{\mathbb{T}_X} w_1 \mathcal{V}_X(\beta), \ldots, \int_{\mathbb{T}_X} w_p \mathcal{V}_X(\beta) \right]^\top = [\|v_{XY}\|_2, 0, \ldots, 0]^\top, \tag{4.10}$$

$$\mathbf{\Lambda}_p = \left[\int_{\mathbb{T}_X} w_{j_1} \mathcal{V}_X(w_{j_2})\right]_{1\le j_1, j_2\le p},$$

and the so-called jth FPLS score

$$\xi_j^* = \int_{\mathbb{T}_X} w_j(X^* - \mu_X). \tag{4.11}$$

The farthest right-hand side of (4.10) is derived from identity (4.3) and Remark 4.2.

The first phase of PLEASS is to find estimators for μ_X at (1.3), v_A at (1.2), v_C at (4.2), and σ_e^2 at (4.1), respectively, say, $\hat{\mu}_X$, $\hat{v}_A$, $\hat{v}_C$ and $\hat{\sigma}_e^2$. Existing methods for the reconstruction of variance and covariance structure from sparse observations roughly fall into three categories: a) spline smoothing (e.g., a fast covariance estimation (FACE) by [102]), b) kernel smoothing (e.g., LLS in [104, 105, 59]; and modified kernel smoothing in [71]), and c) maximum likelihood (ML, e.g., restricted ML in [50, 72]; and quasi-ML in [108]). Typically, the third category requires initial values obtained from some method in the first two categories and hence is more time-consuming. In the numerical study (Section 4.4), we adopt both LLS (with details relegated to Appendix A.3.1) and FACE. Our reason for this choice is two-fold: LLS (of nice asymptotic properties [42]) is also exploited by PACE and hence leads to a more fair comparison between PLEASS and PACE; FACE runs faster and outputs competitive accuracy.

Remark 4.4. In theory, the framework of PLEASS is flexible as to how to estimate μ_X, v_A, v_C, and σ_e^2, as long as $\|\hat{\mu}_X - \mu_X\|_\infty$, $\|\hat{v}_A - v_A\|_\infty$, $\|\hat{v}_C - v_C\|_\infty$, and $|\hat{\sigma}_e^2 - \sigma_e^2|$ all converge to zero as n diverges (with $\|\cdot\|_\infty$ denoting the L^∞-norm). It is even more flexible in practice and permits whatever way of recovery preferred by users. Technical conditions in Appendix A.3.2 vary with your final choice. Also, theoretical results in upcoming Section 4.3 are merely demos corresponding to LLS but are adaptable to other approaches.

Implied by [28, Property (3.4)], the first p FPLS basis functions actually span the functional Krylov subspace of $L^2(\mathbb{T}_X)$, i.e.,

$$\operatorname{span}(w_1, \ldots, w_p) = \operatorname{span}\{\mathcal{V}_X(\beta), \ldots, \mathcal{V}_X^p(\beta)\} = \operatorname{span}\{v_{XY}, \mathcal{V}_X(v_{XY}), \ldots, \mathcal{V}_X^{p-1}(v_{XY})\}, \tag{4.12}$$

where $\mathcal{V}_X^j$ is the jth power of operator $\mathcal{V}_X$ (1.2) and estimated by $\widehat{\mathcal{V}}_X^j$ such that, for all $f \in L^2(\mathbb{T}_X)$,

$$\widehat{\mathcal{V}}_X^j(f)(\cdot) = \int_{\mathbb{T}_X} \widehat{\mathcal{V}}_X^{j-1}(f)(\cdot)\hat{v}_X(t, \cdot)\mathrm{d}t. \tag{4.13}$$

Combined with the orthogonality in (Remark 4.2), identity (4.12) inspires us to estimate w_j (4.4) by sequentially orthonormalizating j functions $\hat{v}_{XY}, \widehat{\mathcal{V}}_X(\hat{v}_{XY}), \ldots, \widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY})$, i.e., subtracting projections onto previous functions and then scaling the remaining part to one in terms of L^2-norm. In particular, $\hat{w}_1 = \hat{v}_{XY}/\|\hat{v}_{XY}\|_2$ and, for $j \ge 2$, $\hat{w}_j$ is succesively

Algorithm 4.1 Modified Gram-Schmidt orthonormalization in estimating w_j

for j in $1, \ldots, p$ **do**
  $\hat{w}_j^{[1]} \leftarrow \widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY})$
  if $j \geq 2$ **then**
    for i in $1, \ldots, j-1$ **do**
      $\hat{w}_j^{[i+1]} \leftarrow \hat{w}_j^{[i]} - \hat{w}_i \int_{\mathbb{T}_X} \hat{w}_j^{[i]} \hat{w}_i$
    end for
  end if
  $\hat{w}_j \leftarrow \hat{w}_j^{[j]} / \|\hat{w}_j^{[j]}\|_2$
end for

given by

$$\hat{w}_j = \frac{\widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY}) - \sum_{k=1}^{j-1} \hat{w}_k \int_{\mathbb{T}_X} \hat{w}_k \widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY})}{\left\| \widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY}) - \sum_{k=1}^{j-1} \hat{w}_k \int_{\mathbb{T}_X} \hat{w}_k \widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY}) \right\|_2}. \tag{4.14}$$

Alternatively, the modified Gram-Schmidit procedure (Algorithm 4.1) gives mathematically equivalent but numerically more stable estimators for w_j; see, e.g., [56, pp. 102]. Plugging both

$$\hat{\boldsymbol{c}}_p = [\|\hat{v}_{XY}\|_2, 0, \ldots, 0]^\top \tag{4.15}$$

and

$$\widehat{\boldsymbol{\Lambda}}_p = \left[\int_{\mathbb{T}_X} \hat{w}_{j_1} \widehat{\mathcal{V}}_X(\hat{w}_{j_2}) \right]_{1 \leq j_1, j_2 \leq p}, \tag{4.16}$$

into β_p (4.8), we then have

$$\hat{\beta}_p = \hat{\boldsymbol{c}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1} [\hat{w}_1, \ldots, \hat{w}_p]^\top. \tag{4.17}$$

This estimator converges to the true β as n and p diverge at specific rates (as described in Theorem 4.1).

Denote by $\widetilde{X}^*$ the noisy counterpart of X^*. Predicting $\eta(X^*)$ (1.4) is separate from estimating β; since $\widetilde{X}^*$ is tainted by noise terms and only measured at $L^* \sim L$ (as restricted by (C.A.3.1) in Section A.3.2) time points, for now it is not practical to numerically compute the integral $\int_{\mathbb{T}_X} \hat{\beta}_p \widetilde{X}^*$. Taking a detour through conditional expectation, we target predicting not $\eta(X^*)$ but instead $\tilde{\eta}_\infty(X^*)$ (4.20), a surrogate for $\eta(X^*)$ (1.4). Specifically, write

$$\begin{aligned} \widetilde{\boldsymbol{X}}^* &= \left[\widetilde{X}^*(T_1^*), \ldots, \widetilde{X}^*(T_{L^*}^*) \right]^\top, \\ \boldsymbol{\mu}_X^* &= \mathrm{E}(\widetilde{\boldsymbol{X}}^*) = [\mu_X(T_1^*), \ldots, \mu_X(T_{L^*}^*)]^\top, \end{aligned}$$

$$\boldsymbol{\Sigma}_{\widetilde{X}^*} = \left[v_X(T_{l_1}^*, T_{l_2}^*)\right]_{1\le l_1,l_2\le L^*} + \sigma_e^2 \mathbf{I}_{L^*},$$

and, for $j \in \{1,\dots,p\}$,

$$\boldsymbol{h}_j^* = \left[\mathcal{V}_X(w_j)(T_1^*),\dots,\mathcal{V}_X(w_j)(T_{L^*}^*)\right]^\top,$$

with $L^* \times L^*$ identity matrix $\mathbf{I}_{L^*}$. Given $L^*, T_1^*, \dots, T_{L^*}^*$, and using

$$\operatorname{cov}\begin{bmatrix} \widetilde{\boldsymbol{X}}^* \\ \xi_1^* \\ \vdots \\ \xi_p^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\widetilde{X}^*} & \boldsymbol{h}_1^* & \cdots & \boldsymbol{h}_p^* \\ \boldsymbol{h}_1^{*\top} & & & \\ \vdots & & \boldsymbol{\Lambda}_p & \\ \boldsymbol{h}_p^{*\top} & & & \end{bmatrix},$$

it is well known that, for ξ_j^* given in (4.11), the best linear unbiased predictor is

$$\tilde{\xi}_j^* = \mathrm{E}(\xi_j^* \mid \widetilde{\boldsymbol{X}}^*, L^*, T_1^*, \dots, T_{L^*}^*) = \boldsymbol{h}_j^{*\top} \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1} (\widetilde{\boldsymbol{X}}^* - \boldsymbol{\mu}_X^*) \tag{4.18}$$

It is best see, e.g., [43, Theorem 1]. Indeed, if $\xi_1^*, \dots, \xi_p^*$ and $\widetilde{\boldsymbol{X}}^*$ are jointly Gaussian then (4.18) is the best predictor, linear or otherwise, in terms of minimizing $\mathrm{E}\{\xi_j^* - f(\widetilde{\boldsymbol{X}}^*)\}^2$ w.r.t. Lebesgue measurable functions f on $\mathbb{R}^{L^*}$. Thus, if we introduce the $L^* \times p$ matrix $\mathbf{H}_p = [\boldsymbol{h}_1^*, \dots, \boldsymbol{h}_p^*]$, then a reasonable surrogate for $\eta_p(X^*)$ (1.6) is

$$\tilde{\eta}_p(X^*) = \mu_Y + \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} [\tilde{\xi}_1^*, \dots, \tilde{\xi}_p^*]^\top = \mu_Y + \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} \mathbf{H}_p^\top \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1} (\widetilde{\boldsymbol{X}}^* - \boldsymbol{\mu}_X^*). \tag{4.19}$$

Conditioning on $L^*, T_1^*, \dots, T_{L^*}^*$, $\tilde{\xi}_j^*$ (4.18) is the (orthogonal) projection of ξ_j^* (4.11) onto $\operatorname{span}\{\widetilde{X}^*(T_1^*), \dots, \widetilde{X}^*(T_{L^*}^*)\}$. Accordingly, $\lim_{p\to\infty} \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} \mathbf{H}_p^\top \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1} (\widetilde{\boldsymbol{X}}^* - \boldsymbol{\mu}_X^*)$ exists as a projection of $\eta(X^*) - \mu_X$ and so does

$$\tilde{\eta}_\infty(X^*) := \lim_{p\to\infty} \tilde{\eta}_p(X^*). \tag{4.20}$$

For $\eta(X^*)$ (1.4), a plug-in prediction

$$\hat{\eta}_p(X^*) = \frac{1}{n}\sum_{i=1}^n Y_i + \hat{\boldsymbol{c}}_p^\top \hat{\boldsymbol{\Lambda}}_p^{-1} \widehat{\mathbf{H}}_p^\top \widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*}^{-1} (\widetilde{\boldsymbol{X}}^* - \hat{\boldsymbol{\mu}}_X^*) \tag{4.21}$$

follows by replacing $\boldsymbol{\Sigma}_{\widetilde{X}^*}$, $\boldsymbol{\mu}_X^*$, $\mathbf{H}_j$ and $\boldsymbol{\Lambda}_p$ involved in (4.19) with their respective empirical counterparts. These counterparts are

$$\widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*} = \left[\hat{v}_X(T_{l_1}^*, T_{l_2}^*)\right]_{1\le l_1,l_2\le L^*} + \hat{\sigma}_e^2 \mathbf{I}_{L^*}, \tag{4.22}$$

$$\hat{\boldsymbol{\mu}}_X^* = \left[\hat{\mu}_X(T_1^*), \dots, \hat{\mu}_X(T_{L^*}^*)\right]^\top, \tag{4.23}$$

$$\widehat{\mathbf{H}}_p = \left[\widehat{\mathcal{V}}_X(\hat{w}_j)(T_l^*)\right]_{1\le l\le L^*, 1\le j\le p}, \tag{4.24}$$

and $\widehat{\mathbf{\Lambda}}_p$ from (4.16).

It remains to construct a confidence interval (CI) for $\eta(X^*)$ (1.4). From the viewpoint of projection again,

$$\begin{aligned}
&\operatorname{cov}([\xi_1^* - \tilde{\xi}_1^*, \ldots, \xi_p^* - \tilde{\xi}_p^*]^\top \mid L^*, T_1^*, \ldots, T_{L^*}^*) \\
&\quad = \operatorname{cov}([\xi_1^*, \ldots, \xi_p^*]^\top \mid L^*, T_1^*, \ldots, T_{L^*}^*) - \operatorname{cov}([\tilde{\xi}_1^*, \ldots, \tilde{\xi}_p^*]^\top \mid L^*, T_1^*, \ldots, T_{L^*}^*) \\
&\quad = \mathbf{\Lambda}_p - \mathbf{H}_p^\top \mathbf{\Sigma}_{\widetilde{X}^*}^{-1} \mathbf{H}_p.
\end{aligned}$$

Under the Gaussian assumption (specifically condition (C.4.1) in Corollary 4.2.1), $\hat{\eta}_p(X^*) - \eta(X^*)$ turns out to be asymptotically (conditionally) normal, as long as $\boldsymbol{c}_p^\top \mathbf{\Lambda}_p^{-1}(\mathbf{\Lambda}_p - \mathbf{H}_p^\top \mathbf{\Sigma}_{\widetilde{X}^*}^{-1} \mathbf{H}_p)\mathbf{\Lambda}_p^{-1}\boldsymbol{c}_p$ converges to a positive number as p goes to infinity (see the detailed condition (C.4.2) in Corollary 4.2.1). An asymptotic $(1-\alpha)$ CI for $\eta(X^*)$ is then

$$\hat{\eta}_p(X^*) \pm \Phi_{\alpha/2}^{-1} \left\{ \hat{\boldsymbol{c}}_p^\top \widehat{\mathbf{\Lambda}}_p^{-1} (\widehat{\mathbf{\Lambda}}_p - \widehat{\mathbf{H}}_p^\top \widehat{\mathbf{\Sigma}}_{\widetilde{X}^*}^{-1} \widehat{\mathbf{H}}_p) \widehat{\mathbf{\Lambda}}_p^{-1} \hat{\boldsymbol{c}}_p \right\}^{1/2},$$

with the $(1-\alpha/2)$ standard normal quantile $\Phi_{\alpha/2}^{-1}$.

4.2.2 Selection of number of basis functions

As for the tuning target, it is doable to adapt the generalized cross-validation (GCV) at Section 2.4.1 to the context over here. But the (leave-one-out) cross-validation (CV) sounds more reasonable: we are unclear on how to estimate the degrees of freedom (DoF) associated with PLEASS prediction at (4.21); its intrinsic complexity results in no natural extension of DoF computation for (multivariate) PLS [55]. Specifically, we here choose $p \in [0, p_{\max}]$ by minimizing

$$\mathrm{CV}(p) = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{\eta}_p^{(-i)}(X_i)\}^2$$

in which $\hat{\eta}_p^{(-i)}(X_i)$ predicts the ith response with all the other subjects kept for training. As for $p_{\max}$, the definition at (2.19) works here, because it is one default rule in truncating the Karhunen-Loève series and, as numerically illustrated by [28, Section 6], FPLS needs fewer terms than FPC to reach the same accuracy. Another candidate for $p_{\max}$ is provided by [27, Section 3], i.e., $p_{\max} = n/2$, which is acceptable for a small or moderate n.

4.3 Asymptotic properties

Our theoretical results are derived from Conditions (C.A.3.1)–(C.A.3.14) in Appendix A.3.2. The first six of these detail the assumptions of model (4.1) while and the remaining assumptions are set up for consistency of LLS in Appendix A.3.1. For arbitrarily fixed integer p, the consistency of $\hat{\beta}_p$ (1.7) is deduced immediately from Lemmas A.3 and A.4 in Appendix A.3.2 and follow the line of proof of [109, Theorem 1]. Unfortunately, this argument does not ap-

Algorithm 4.2 PLEASS tuned through GCV

Obtain $\hat{\mu}_X$, $\hat{v}_X$, $\hat{v}_{XY}$ and $\hat{\sigma}_e^2$ following A.3.1.
for j in $1, \ldots, p_{\max} - 1$ **do**
$\widehat{\mathcal{V}}_X^j(\hat{v}_{XY})(\cdot) \leftarrow \int_{\mathbb{T}_X} \hat{v}_X(\cdot, t)\widehat{\mathcal{V}}_X^{j-1}(\hat{v}_{XY})(t)\mathrm{d}t$.
end for
Extract $\hat{w}_j$ from $\hat{v}_{XY}, \widehat{\mathcal{V}}_X(\hat{v}_{XY}), \ldots, \widehat{\mathcal{V}}_X^{p_{\max}-1}(\hat{v}_{XY})$ following Algorithm 4.1.
$\hat{\beta}_0 \leftarrow 0$.
$\hat{\eta}_0(X^*) \leftarrow n^{-1}\sum_{i=1}^n Y_i$.
for p in $1, \ldots, p_{\max}$ **do**
$\hat{\beta}_p \leftarrow \hat{\boldsymbol{c}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1}[\hat{w}_1, \ldots, \hat{w}_p]^\top$ with $\hat{\boldsymbol{c}}_p$ (4.15) and $\widehat{\boldsymbol{\Lambda}}_p$ (4.16).
$\hat{\eta}_p(X^*) \leftarrow n^{-1}\sum_{i=1}^n Y_i + \hat{\boldsymbol{c}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1}\widehat{\mathbf{H}}_p^\top \widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*}^{-1}(\widetilde{\boldsymbol{X}}^* - \hat{\boldsymbol{\mu}}_X^*)$
with $\widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*}$ (4.22), $\hat{\boldsymbol{\mu}}_X^*$ (4.23) and $\widehat{\mathbf{H}}_p$ (4.24).
end for
$p_{\mathrm{opt}} \leftarrow \arg\min_{0 \le p \le p_{\max}} \mathrm{GCV}(p)$.
CI for $\eta(X^*) \leftarrow \hat{\eta}_{p_{\mathrm{opt}}}(X^*) \pm \Phi_{\alpha/2}^{-1}\left\{\hat{\boldsymbol{c}}_{p_{\mathrm{opt}}}^\top \widehat{\boldsymbol{\Lambda}}_{p_{\mathrm{opt}}}^{-1}(\widehat{\boldsymbol{\Lambda}}_{p_{\mathrm{opt}}} - \widehat{\mathbf{H}}_{p_{\mathrm{opt}}}^\top \widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*}^{-1}\widehat{\mathbf{H}}_{p_{\mathrm{opt}}})\widehat{\boldsymbol{\Lambda}}_{p_{\mathrm{opt}}}^{-1}\hat{\boldsymbol{c}}_{p_{\mathrm{opt}}}\right\}^{1/2}$.

ply to the scenario with p diverging with n, since the sequential construction (4.14) tends to induce a bias accumulated with the increase of p. As a consequence, it is indispensable to impose a sufficiently slow divergence rate on p, e.g., as stated in (C.A.3.15).

Theorem 4.1. *Assume that (C.A.3.1)–(C.A.3.15) in Appendix A.3.1 all hold. As n goes to infinity, $\|\hat{\beta}_p - \beta\|_2 \to_p 0$. If $\|\beta_p - \beta\|_\infty \to_p 0$ and condition (C.A.3.15) is replaced with the stronger assumption (C.A.3.16) then this L^2 convergence result can be strengthened to a uniform version, namely, $\|\hat{\beta}_p - \beta\|_\infty \to_p 0$.*

Analogous to PACE, given X^*, our PLEASS does not result in a consistent prediction: $\mathrm{E}\{\hat{\eta}_p(X^*) - \eta(X^*)\}$ converges to zero but $\hat{\eta}_p(X^*) - \eta(X^*)$ does not; the limit of $\hat{\eta}_p(X^*)$ in (4.21) is $\tilde{\eta}_\infty(X^*)$ in (4.20) instead.

Theorem 4.2. *Under assumptions (C.A.3.1)–(C.A.3.15), as n goes to infinity, $\hat{\eta}_p(X^*) - \tilde{\eta}_\infty(X^*)$ converges to zero (unconditionally) in probability.*

In spite of the fact that PLEASS does not yield the desired consistent prediction for $\eta(X^*)$, Theorem 4.2 implies that $\hat{\eta}_p(X^*) - \eta(X^*)$ is asymptotically distributed as $\tilde{\eta}_\infty(X^*) - \eta(X^*)$. An asymptotic $(1-\alpha)$ CI for $\eta(X^*)$ is therefore available. In particular, the result for Gaussian cases is presented in Corollary 4.2.1.

Corollary 4.2.1. *In addition to assumptions (C.A.3.1)–(C.A.3.15), we assume two more:*
(C.4.1) FPLS scores $\int_{\mathbb{T}_X} w_j(X - \mu_X)$ and measurement errors $e_{i\ell}$ are jointly Gaussian.
(C.4.2) $\boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1}(\boldsymbol{\Lambda}_p - \mathbf{H}_p^\top \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1}\mathbf{H}_p)\boldsymbol{\Lambda}_p^{-1}\boldsymbol{c}_p \to \omega > 0$ as p goes to infinity.
Then,

$$\frac{\hat{\eta}_p(X^*) - \eta(X^*)}{\sqrt{\hat{\boldsymbol{c}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1}(\widehat{\boldsymbol{\Lambda}}_p - \widehat{\mathbf{H}}_p^\top \widehat{\boldsymbol{\Sigma}}_{\widetilde{X}^*}^{-1}\widehat{\mathbf{H}}_p)\widehat{\boldsymbol{\Lambda}}_p^{-1}\hat{\boldsymbol{c}}_p}} \to_d \mathcal{N}(0,1).$$

4.4 Numerical illustration

PLEASS is compared here with PACE in terms of finite-sample numerical performance. As mentioned in Section 4.2.1, both LLS and FACE (implemented respectively via `R` packages `fdapace` [24] and `face` [103]) were utilized to estimate population quantities μ_X at (1.3), v_X at (1.2), v_C at (4.2), and σ_e^2 at (4.1). Resulting combinations, viz. PLEASS+LLS, PACE+LLS, PLEASS+FACE and PACE+FACE, are abbreviated as PLEASS.L, PACE.L, PLEASS.F and PACE.F, respectively. Corresponding code trunks are accessible at `https://github.com/ZhiyangGeeZhou/PLEASS`.

4.4.1 Simulation study

Each sample consisted of $n = 300$ iid paired realizations of (X, Y) with X and Y both of zero mean. The X was set up as a Gaussian process, i.e., $\lambda_{j,X}^{-1/2}\rho_j$ were all iid as standard normal with the jth FPC score $\rho_j = \int_{\mathbb{T}_X} \phi_{j,X}(X - \mu_X)$. Error terms $e_{i\ell}$ were also standard normal. We took 100, 90, 80, 10, 9, 8, 1, 0.9, and 0.8 as the top nine eigenvalues of operator $\mathcal{V}_X$ at (1.1); all the rest were 0. Correspondingly, the top nine eigenfunctions were taken to be (normalized) shifted Legendre polynomials [45, pp. 773–774] of order 1 to 9, say $P_1, \ldots, P_9$; unit-normed and mutually orthogonal on $[0, 1]$, they were generated through `R`-package `orthopolynom` [70]. The slope function β was given by one of the following cases:

$$\beta = P_1 + P_2 + P_3, \tag{4.25}$$

$$\beta = P_4 + P_5 + P_6, \tag{4.26}$$

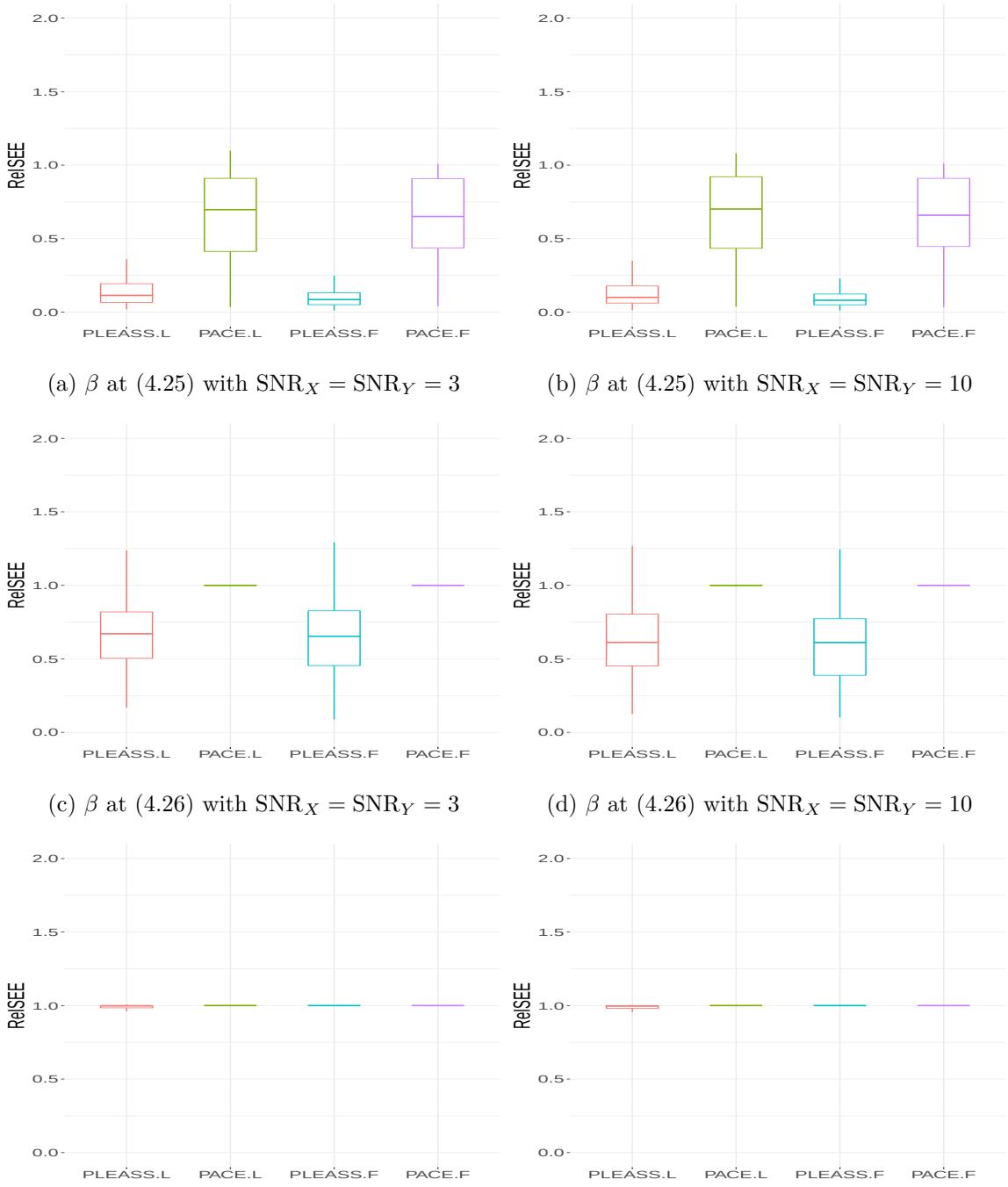
$$\beta = P_7 + P_8 + P_9. \tag{4.27}$$

Two sorts of signal-to-noise-ratio (SNR) were defined, i.e., $\mathrm{SNR}_X = (\sum_{j=1}^{\infty} \lambda_{j,X})^{1/2}/\sigma_e$ and $\mathrm{SNR}_Y = \mathrm{sd}(\int_{\mathbb{T}_X} \beta X)/\sigma_\varepsilon$. For simplicity, we took $\mathrm{SNR}_X = \mathrm{SNR}_Y$ ($= 3$ or 10). To embody the sparsity assumptions, in each sample, X_i was observed only at L_i ($\overset{\text{iid}}{\sim}$ Unif$\{3, 4, 5, 6\}$) points uniformly selected from $[0, 1]$. In total there were six combinations of settings. 200 iid samples were generated for each of them. We randomly reserved 20% of the subjects in each sample for testing and used the remainder for training. After running through all samples, we computed 200 values of relative integrated squared estimation error (ReISEE)

$$\mathrm{ReISEE} = \|\beta\|_2^{-2}\|\beta - \hat{\beta}_p\|_2^2. \tag{4.28}$$

Since neither PACE nor PLEASS leads to consistent predictions, it is better to evaluate the prediction quality via the coverage percentage (CP) of CIs constructed for testing subjects, viz.

$$\mathrm{CP} = \sum_{i \in I_{\text{test}}} \mathbb{1}\left\{\eta(X_i) \in \widehat{\mathrm{CI}}_i\right\} \Big/ \#\mathrm{ID}_{\text{test}},$$



(a) β at (4.25) with $\text{SNR}_X = \text{SNR}_Y = 3$

(b) β at (4.25) with $\text{SNR}_X = \text{SNR}_Y = 10$

(c) β at (4.26) with $\text{SNR}_X = \text{SNR}_Y = 3$

(d) β at (4.26) with $\text{SNR}_X = \text{SNR}_Y = 10$

(e) β at (4.27) with $\text{SNR}_X = \text{SNR}_Y = 3$

(f) β at (4.27) with $\text{SNR}_X = \text{SNR}_Y = 10$

Figure 4.1: Boxplots of ReISEE values under different simulated settings: SNR value varies with column, while rows differ in β. In each subfigure, from left to right, the four boxes respectively correspond to PLEASS.L, PLEASS.F, PACE.L, and PACE.F.

where $\widehat{\text{CI}}_i$ is the asymptotic (95%) CI for $\eta(X_i)$, and I_{test} is the index set for testing portion with cardinality $\#\text{ID}_{\text{test}}$.

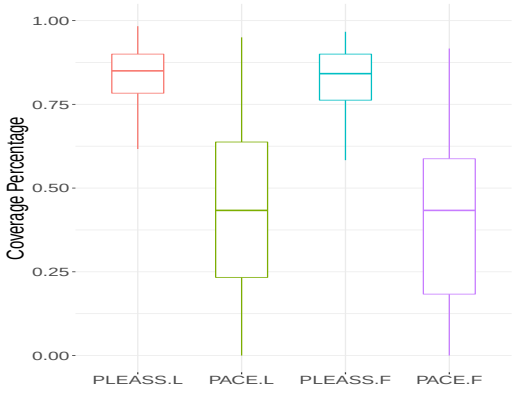
When β was constructed from eigenfunctions corresponding to large or moderate eigenvalues (viz. β at (4.25) or (4.26)), PLEASS performed better in term of ReISEE; see the first two rows of Figure 4.1. Particularly, at the second row of Figure 4.1, ReISEE values of PLEASS were mostly lower than one, while PACE boxes was trapped at one. An ReISEE box sticking around one implied estimates concentrated around the most trivial $\hat{\beta} = 0$, i.e., the corresponding method failed to output non-trivial estimates. This failure was caused by zero inner products between $\hat{v}_{XY}$ and estimated basis functions; this happened frequently if β was mainly associated with a small portion of total variation (of $\mathcal{V}_A$) that was likely to be smoothed out in recovering v_X and v_{XY}. Such was exactly the case for PACE in the scenario (4.26) and for both PACE and PLEASS with β at (4.27).

As seen in Figure 4.2, CP boxes belonging to PACE stayed at a low level, especially for scenarios (4.26) and (4.27). This phenomenon was consistent with the performance of PACE in estimating β under corresponding settings. In contrast, PLEASS was more likely to output CP values closer to the stated level (95%), though we must admit that their coverage was still far from satisfactory especially with β at (4.26) and (4.27). Looking into those $\eta(X_i)$ not covered by $\widehat{\mathrm{CI}}_i$, we noticed that the majority of missed $\eta(X_i)$ fell at the right-hand side of $\widehat{\mathrm{CI}}_i$. A possible cause of miss-covering lay in the bias of estimates for means of X and Y; a larger size of training set might be helpful. Moreover, although SNR had little impact on estimation (compare the two columns of Figure 4.1), CP values appeared to be higher with a smaller SNR (compare the two columns of Figure 4.2): $\eta(X_i)$ did not vary with SNR, while larger $\hat{\sigma}_e^2$ (resulting from smaller SNR) widened $\widehat{\mathrm{CI}}_i$ and enhanced the coverage of $\widehat{\mathrm{CI}}_i$.

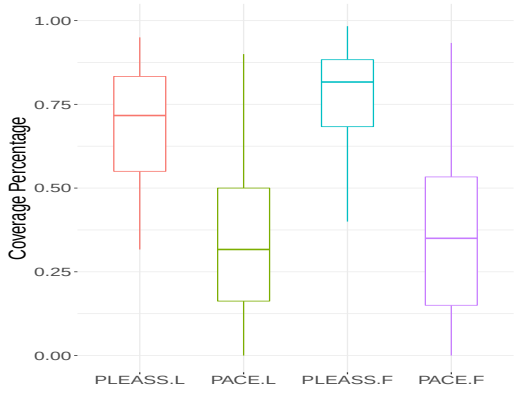
4.4.2 Application to real data

We then applied PLEASS to two real datasets. The first came from a clinical trial, whereas the second was densely observed but recorded with numerous missing values.

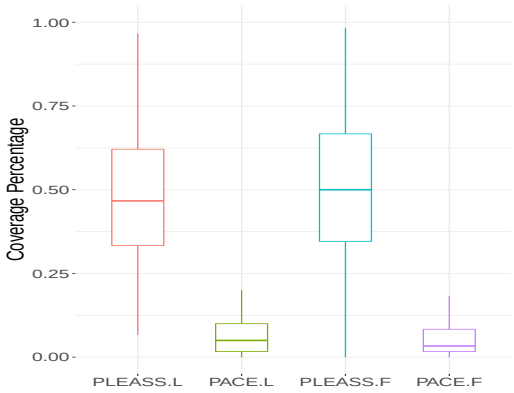
Primary Biliary Cholangitis data. Initially shared by [93], dataset `pbcseq` (accessible in R-package `survival` created by [92]) was collected in a randomized placebo controlled trial of D-penicillamine, a drug designed for primary biliary cholangitis (PBC, also known as primary biliary cirrhosis). PBC is a chronic disease in which bile ducts in the liver are slowly destroyed; it can cause more serious problems including liver cancer. All the participants of the clinical trial were supposed to revisit the Mayo Clinic at six months, one year, and annually after their initial diagnoses. However, participants' visiting frequencies, with an average of 6, varied among patients, ranging from 1 to 16 and leading to sparse and irregular recordings. Although the clinical trial lasted from January 1974 through May 1984, to satisfy the prerequisites of LLS, we included only measurements within the first 3000 days and removed subjects with fewer than two visits. At each visit, several body indices were measured and recorded,



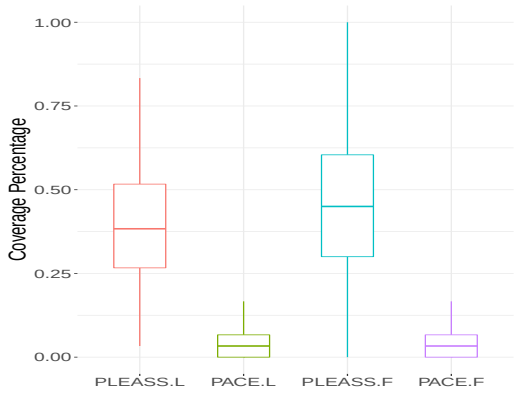
(a) β at (4.25) with $\text{SNR}_X = \text{SNR}_Y = 3$



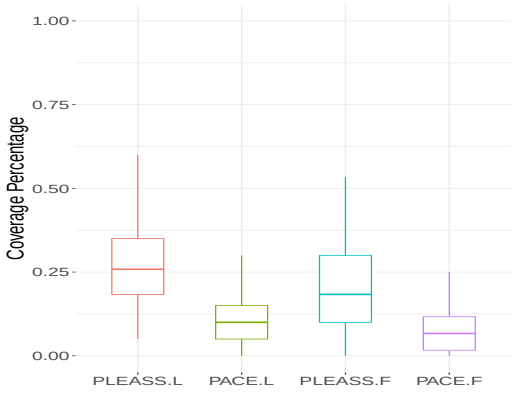
(b) β at (4.25) with $\text{SNR}_X = \text{SNR}_Y = 10$



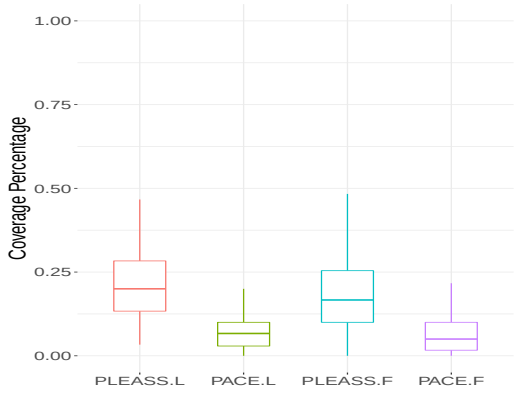
(c) β at (4.26) with $\text{SNR}_X = \text{SNR}_Y = 3$



(d) β at (4.26) with $\text{SNR}_X = \text{SNR}_Y = 10$



(e) β at (4.27) with $\text{SNR}_X = \text{SNR}_Y = 3$



(f) β at (4.27) with $\text{SNR}_X = \text{SNR}_Y = 10$

Figure 4.2: Boxplots of coverage percentage under different simulated settings: SNR value varies with column, while rows differ in β. In each subfigure, from left to right, the four boxes respectively correspond to PLEASS.L, PLEASS.F, PACE.L, and PACE.F.

including serum albumin (SerAlbu, in mg/dL) and prothrombin time (ProTime, in seconds) whose relationship was studied by [105]. We also focused on this pair of indi-

cators and attempted to model a linear connection between participants' last ProTime measurements (response) and their SerAlbu profiles (predictor).

Diffusion tensor imaging (DTI) data. As mentioned in Section 3.3.2, the fractional anisotropy (FA) is measured at a specific spot in the white matter, ranging from 0 to 1 and reflecting the fiber density, axonal diameter and myelination. Along a tract of interest, these values form an FA tract profile. Collected at Johns Hopkins University and the Kennedy-Krieger Institute, dataset `DTI` (in R-package `refund` [39]) contains FA tract profiles for corpus callosum measured via DTI. Though these trajectories were not sparsely measured, a few of them had missing records which we were able to handle using PACE and PLEASS without presmoothing or interpolation. We linked the FA tract profile (our predictor) to the corresponding Paced Auditory Serial Addition Test (PASAT, see [94] for more on PASAT) score (our response). The PASAt score is a standard assessment of the capacity for and rate of information processing. This study explored a potential linkage between a modern medical imaging technique and a classical diagnosis on brain function. Moreover, to ensure independence among trajectories, we kept only the latest profile for each participant and excluded those curves with no PASAT score.

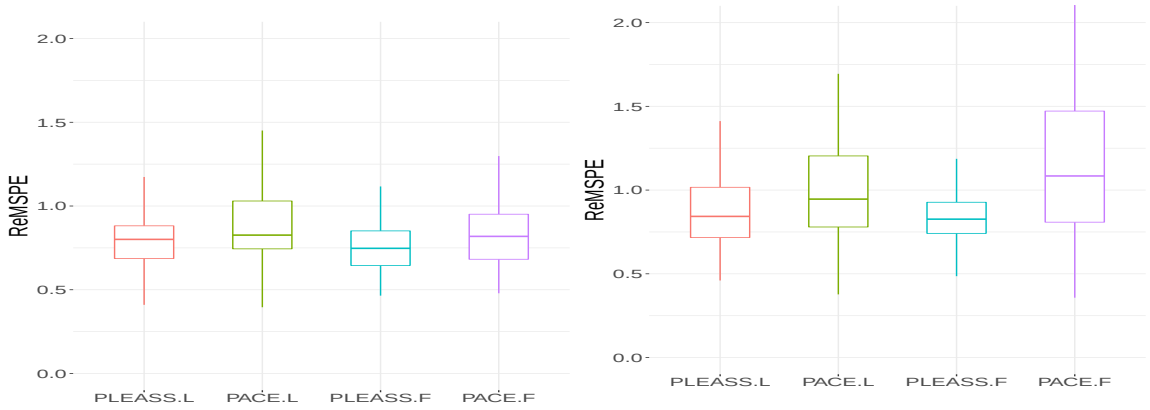
For each real dataset, 200 (independent) random splits were carried out. As we did in the simulation, (roughly) 80% of subjects in each split were put into the training set and the remainder was kept for testing. In these cases there was no way to calculate ReISEE or CP values; the comparison between PACE and PLEASS was hence carried out in terms of ReMSPE at (2.22), viz. for each method and each split,

$$\text{ReMSPE} = \frac{\sum_{i\in \text{ID}_{\text{test}}}(Y_i - \widehat{Y}_i)^2}{\sum_{i\in \text{ID}_{\text{test}}}(Y_i - \bar{Y}_{\text{train}})^2},$$

where $\widehat{Y}_i$ is the prediction for the ith subject, and $\bar{Y}_{\text{train}}$ is the sample mean of testing responses. ReMSPE values for PBC and DTI cases were collected and summarized into boxes; see Figure 4.3. In both applications, PLEASS was demonstrated to be more competitive than PACE, enjoying lower medians and smaller dispersion of ReMSPE values. Analogous to the previous simulation study, Figure 4.3 shows that FACE performs close to LLS when used with PLEASS. As a result, PLEASS.F might be preferred if a low time consumption were particularly appreciated.

4.5 Concluding remarks

The main contributions of our work are summarized as follows. First, we propose PLEASS, a variant of FPLS modified for scenarios in which functional predictors are only observed at sparse time points. Secondly, we show that PLEASS is applicable to SoFR coupled



(a) ALP (predictor) vs. latest AST (response)

(b) FA tract profile (predictor) vs. PASAT (response)

Figure 4.3: Boxplots of ReMSPE values for real data analysis. In each subfigure, from left to right, the four boxes respectively correspond to PLEASS.L, PLEASS.F, PACE.L, and PACE.F.

with measurement errors, a setting more complex than the one in Delaigle and Hall [28]. Third, not only do we give estimators and predictions via PLEASS, but also we construct CIs for mean responses. Assuming that p diverges as a function of n, consistency of our estimator is among the few asymptotic results available for FPLS and its variants. Fourth, we numerically illustrate the advantage of PLEASS in specific scenarios.

Estimators for the variance and covariance structure could be further revised. When estimating the value at a specific point, LLS borrows strength from a neighbourhood whose bandwidth turns out to impact PLEASS to a certain extent. Competitors of LLS are more or less haunted too by the challenging bandwidth selection problem whose solution remains an open question. The data-driven selection adopted by Appendix A.3.1 may not be optimal in practice or perhaps not even close to optimal. If trajectories are no longer independent of each other, the proposal of [71] is more competitive. One more limitation concerns the nature of the missingness: sparse observations (and missing values) are assumed independent of trajectories and measurement errors, as in (C.A.3.3). Once the missingness is permitted to be correlated with unobserved time points, we speculate that, after necessary modifications, estimates of the ML type would be still promising in estimating components of covariance structure.

In contrast with PLEASS, which is for now concentrated on SoFR only, PACE is more versatile: aside from handling even FoFR, PACE is capable as well of recovering predictor trajectories. By merging PLEASS into the framework of [110], we are working to adapt it to FoFR. We believe it would offer ancillary information when curves correlated to to-be-reconstructed ones are observed.

Chapter 5

Partial least squares for function-on-function regression via Krylov subspaces

5.1 Introduction

Sometimes one would like to model the relationship between two stochastic curves. To exemplify this type of interest, two instances are listed as below. As a fundamental model in FDA, FoFR at (1.9) may be helpful to corresponding scientific explorations.

Diffusion tensor imaging (DTI) data (dataset `DTI` in `R` package `classiFunc` [64], involved too in Section 3.3.2). DTI is powerful for characterizing microstructural changes for neuropathology [4]. One widely used DTI measure is called the fractional anisotropy (FA). An FA tract profile consists of FA values (ranging between zero and one) along a tract of interest in the brain. Originally collected at the Johns Hopkins University and the Kennedy-Krieger Institute, FA tract profiles for the corpus callosum (CCA) and the right corticospinal tract (RCST) for 142 individuals are included in dataset `DTI` in `R` package `refund` [39] (involved too in Section 4.4.2). Imputing missing values among them, [64] created `DTI` in `classiFunc`. There are already investigations on associations between these CCA and RCST trajectories available in the literature; see, e.g., [49].

Boys' gait (BG) data (dataset `gait` in `R` package `fda` [77]). This dataset records hip and knee angles in degrees for 39 walking boys. For each individual, through a 20-point movement cycle, these angles form two curves. Then BG may be partially reflected by the relationship between hip and knee curves.

Excellent contributions have been made to the investigation of FoFR. In general, due to the intrinsically infinite dimension, people have to consider an approximation to β within

certain subspaces of $L^2(\mathbb{T}_X \times \mathbb{T}_Y)$. Traditionally, these subspaces are constructed from predetermined functions, e.g., splines and Fourier basis functions. But a more prevailing option may be data-driven: FPCR approximates β by

$$\beta_{p,q,\mathrm{FPCR}}(s,t) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\mathrm{cov}(\int_{\mathbb{T}_X} X\phi_{i,X}, \int_{\mathbb{T}_Y} Y\phi_{j,Y})}{\lambda_{i,X}} \phi_{i,X}(s)\phi_{j,Y}(t); \tag{5.1}$$

see, e.g., [105, Eq. 5]. Accompanied with a penalized estimation, [60] and [89] limit their discussions of coefficient estimators to reproducing kernel Hilbert spaces. The Tikhonov (viz. ridge-type) regularization in [8] yields a remedy for ill-posed β when not all $\lambda_{i,X}$ are nonzero. Distinct from these works, our consideration is based on a subspace of $L^2(\mathbb{T}_X \times \mathbb{T}_Y)$ named after (Alexei) Krylov, viz.

$$\mathrm{KS}_p(\mathcal{V}_X, \beta) = \mathrm{span}\{\mathcal{V}_X^i(\beta) \mid 1 \le i \le p\}, \tag{5.2}$$

where $\mathcal{V}_X^1$ (resp. $\mathcal{V}_X^0$) is indeed operator $\mathcal{V}_X$ (resp. identity operator I), while operator $\mathcal{V}_X^i : L_2(\mathbb{T}_X \times \mathbb{T}_Y) \to L_2(\mathbb{T}_X \times \mathbb{T}_Y)$, $i \ge 1$, is defined recursively as, for each $f \in L_2(\mathbb{T}_X \times \mathbb{T}_Y)$ and each $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$,

$$\begin{aligned}
\mathcal{V}_X^i(f)(s,t) &= (\mathcal{V}_X \circ \mathcal{V}_X^{i-1})(f)(s,t) \\
&= \mathcal{V}_X\{\mathcal{V}_X^{i-1}(f)\}(s,t) \\
&= \int_{\mathbb{T}_X} v_X(s,u)\{\mathcal{V}_X^{i-1}(f)(u,t)\}\mathrm{d}u.
\end{aligned}$$

Noting that $\mathcal{V}_X^i(\beta) = \mathcal{V}_X^{i-1}(v_{XY})$ (with $v_{XY}(s,t) = \mathrm{cov}\{X(s), Y(t)\}$) for all $i \in \mathbb{Z}^+$, the (p-dimensional) Krylov subspace at (5.2) incorporates both X and Y and hence overcomes the lack of supervision of the truncated eigenspaces used in FPCR.

Definition (5.2) is a natural generalization of (4.12); it also expands the Krylov subspace method previously defined for (multivariate) PLS. In the multivariate context, PLS is a terminology shared by a series of algorithms yielding supervised (i.e., related-to-response) basis functions; [12, Section 2.2] briefs several well-known examples of them, including the nonlinear iterative PLS (NIPALS, [100]) and the statistically inspired modification of PLS (SIMPLS, [26]). For single-vector-response, these two lead to outputs identical to that from the Krylov subspace method; but they are known to yield different results when the response is of more than one vectors; see [22, Section 7.2]. Likewise, their respective functional counterparts are equivalent to each other for scalar-response but become diverse again for FoFR. We refer readers to [11] for a straightforward extension of NIPALS and SIMPLS for FoFR. Shooting at the same model, SigComp [61] embeds penalties into NIPALS. It is Proposition 5.1 that drives us to pick up the Krylov subspace method as our route.

Proposition 5.1. *Under (C.A.4.1), β belongs to $\overline{\mathrm{KS}_\infty(\mathcal{V}_X, \beta)} = \overline{\mathrm{span}\{\mathcal{V}_X^i(\beta) \mid i \geq 1\}}$, with the overline representing closure.*

Remark 5.1. It is worth noting that Proposition 5.1 is not a corollary of [28, Theorem 3.2]; the latter implies only an identity weaker than Proposition 5.1: namely, fixing an arbitrary $t_0 \in \mathbb{T}_Y$, the univariate function $\beta(\cdot, t_0)$ belongs to $\overline{\mathrm{span}\{\mathcal{V}_X^i(\beta)(\cdot, t_0) \mid i \geq 1\}}$.

As an extension of the alternative PLS (APLS, [28], designed for the scalar-on-function regression), our proposal is abbreviated as fAPLS, with letter "f" emphasizing its application to FoFR. The remaining portion of this chapter is organized as follows. Section 4.2 details two equivalent expressions of fAPLS estimators, facilitating the empirical implementation and theoretical derivation, respectively. In Section 5.3 fAPLS is compared with competitors in applications to both simulated and authentic datasets. The framework of fAPLS has the potential to be extended to more complex settings, e.g., correlated subjects and non-linear modelling; we include promising directions in Chapter 6. More assumptions and proofs are relegated to Appendix A.4 for conciseness.

5.2 Method

We propose to project β to (5.2) and to utilize the least squares solution

$$\beta_{p,\mathrm{fAPLS}} = \underset{\theta \in \mathrm{KS}_p(\mathcal{V}_X, \beta)}{\arg\min} \, \mathrm{E} \, \|Y - \mu_Y - \mathcal{L}_X(\theta)\|_2^2 = [\mathcal{V}_X(\beta), \ldots, \mathcal{V}_X^p(\beta)] \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p, \tag{5.3}$$

where $\boldsymbol{H}_p = [h_{ij}]_{1 \leq i,j \leq p}$ and $\boldsymbol{\alpha}_p = [\alpha_1, \ldots, \alpha_p]^\top$ denote $p \times p$ and $p \times 1$ matrices, respectively, with

$$\begin{aligned} h_{ij} &= \int_{\mathbb{T}_Y} \left\{ \int_{\mathbb{T}_X} \int_{\mathbb{T}_X} v_X(s,u) \mathcal{V}_X^i(\beta)(s,t) \mathcal{V}_X^j(\beta)(u,t) \mathrm{d}s\mathrm{d}u \right\} \mathrm{d}t \\ &= \int_{\mathbb{T}_Y} \int_{\mathbb{T}_X} \mathcal{V}_X^i(\beta)(s,t) \mathcal{V}_X^{j+1}(\beta)(s,t) \mathrm{d}s\mathrm{d}t, \\ \alpha_i &= \int_{\mathbb{T}_Y} \left\{ \int_{\mathbb{T}_X} \int_{\mathbb{T}_X} v_X(s,u) \mathcal{V}_X^i(\beta)(s,t) \beta(u,t) \mathrm{d}s\mathrm{d}u \right\} \mathrm{d}t \\ &= \int_{\mathbb{T}_Y} \int_{\mathbb{T}_X} \mathcal{V}_X(\beta)(s,t) \mathcal{V}_X^i(\beta)(s,t) \mathrm{d}s\mathrm{d}t. \end{aligned} \tag{5.4}$$

Proposition 5.1 justifies (5.3) by entailing that $\lim_{p \to \infty} \|\beta_{p,\mathrm{fAPLS}} - \beta\|_2 = 0$, which is crucial for the consistency of our estimators described later.

It is natural to estimate $v_X(s,t)$ and $v_{XY}(s,t)$ $(= \mathcal{V}_X(\beta)(\mathrm{s,t}))$, respectively, by

$$\hat{v}_X(s,t) = \frac{1}{n} \sum_{i=1}^n X_i^{\mathrm{cent}}(s) X_i^{\mathrm{cent}}(t) \tag{5.5}$$

$$\widehat{\mathcal{V}}_X(\beta)(s,t) = \hat{v}_{XY}(s,t) = \frac{1}{n} \sum_{i=1}^n X_i^{\mathrm{cent}}(s) Y_i^{\mathrm{cent}}(t) \tag{5.6}$$

in which $X_i^{\text{cent}} = X_i - \bar{X}$ and $Y_i^{\text{cent}} = Y_i - \bar{Y}$, with $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$. Given $\widehat{\mathcal{V}}_X^i(\beta)$, one can estimate $\mathcal{V}_X^{i+1}(\beta)(s,t)$ by

$$\widehat{\mathcal{V}}_X^{i+1}(\beta)(s,t) = \int_{\mathbb{T}_X} \hat{v}_X(s,u)\widehat{\mathcal{V}}_X^i(\beta)(u,t)\mathrm{d}u. \tag{5.7}$$

Plugging (5.5), (5.6) and (5.7) all into (5.3), an estimator for β results:

$$\hat{\beta}_{p,\text{fAPLS}} = [\widehat{\mathcal{V}}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta)]\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p, \tag{5.8}$$

where $\widehat{\boldsymbol{H}}_p = [\hat{h}_{ij}]_{1\leq i,j\leq p}$ and $\widehat{\boldsymbol{\alpha}}_p = [\hat{\alpha}_1, \ldots, \hat{\alpha}_p]^\top$ have respective components given by

$$\begin{aligned} \hat{h}_{ij} &= \int_{\mathbb{T}_Y}\int_{\mathbb{T}_X} \widehat{\mathcal{V}}_X^i(\beta)(s,t)\widehat{\mathcal{V}}_X^{j+1}(\beta)(s,t)\mathrm{d}s\mathrm{d}t, \\ \hat{\alpha}_i &= \int_{\mathbb{T}_Y}\int_{\mathbb{T}_X} \widehat{\mathcal{V}}_X(\beta)(s,t)\widehat{\mathcal{V}}_X^i(\beta)(s,t)\mathrm{d}s\mathrm{d}t. \end{aligned} \tag{5.9}$$

Finally, given a new trajectory $X_0 \sim X$ and aa point $t \in \mathbb{T}_Y$,

$$\eta(X_0)(t) = \mathrm{E}\{Y(t) \mid X = X_0\} = \mu_Y(t) + \mathcal{L}_{X_0}(\beta)(t) \tag{5.10}$$

is predicted by

$$\hat{\eta}_{p,\text{fAPLS}}(X_0)(t) = \bar{Y}(t) + \int_{\mathbb{T}_X} X_0^{\text{cent}}(s)\hat{\beta}_{p,\text{fAPLS}}(s,t)\mathrm{d}s. \tag{5.11}$$

The matrix $\widehat{\boldsymbol{H}}$ at (5.8) would be invertible if we were able to work in exact arithmetic. But this is not the case for finite precision arithmetic: as p increases, the linear system of $\widehat{\mathcal{V}}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta)$ may become close to singular. To overcome this numerical difficulty, as suggested by [28, Section 4.2], we orthonormalize $\widehat{\mathcal{V}}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta)$ (w.r.t. $\hat{v}_X$) into $\hat{w}_1, \ldots, \hat{w}_p$ (see Algorithm 5.1 or [56, pp. 102]) and reformulate the optimization problem at (5.3) into the empirical version:

$$\max_{[c_1,\ldots,c_p]^\top \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{T}_Y} \left\{ Y_i(t) - \bar{Y}(t) - \sum_{j=1}^p c_j \int_{\mathbb{T}_X} X_i^{\text{cent}}(s)\hat{w}_j(s,t)\mathrm{d}s \right\}^2 \mathrm{d}t. \tag{5.12}$$

We then reach a numerically stabilized estimator for β:

$$\tilde{\beta}_{p,\text{fAPLS}} = [\hat{w}_1, \ldots, \hat{w}_p][\hat{\gamma}_1, \ldots, \hat{\gamma}_p]^\top = \sum_{i=1}^p \hat{\gamma}_i \hat{w}_i, \tag{5.13}$$

where $[\hat{\gamma}_1, \ldots, \hat{\gamma}_p]^\top$ is the maximizer of (5.12), with

$$\hat{\gamma}_i = \int_{\mathbb{T}_Y}\int_{\mathbb{T}_X} \hat{v}_{XY}(s,t)\hat{w}_i(s,t)\mathrm{d}s\mathrm{d}t.$$

Algorithm 5.1 Modified Gram-Schmidt orthonormalization w.r.t. $\hat{v}_X$

for i in $1, \ldots, p$ **do**
  $\hat{w}_i^{[1]} \leftarrow \widehat{\mathcal{V}}_X^i(\beta)$.
  if $i \geq 2$ **then**
    for j in $1, \ldots, i-1$ **do**
      $\hat{w}_i^{[j+1]} \leftarrow \hat{w}_i^{[j]} - \left\{\int_{\mathbb{T}_Y}\int_{\mathbb{T}_X}\int_{\mathbb{T}_X} \hat{v}_X(s,u)\hat{w}_i^{[j]}(s,t)\hat{w}_j(u,t)\mathrm{d}s\mathrm{d}u\mathrm{d}t\right\}\hat{w}_j$.
    end for
  end if
  $\hat{w}_i \leftarrow \left\{\int_{\mathbb{T}_Y}\int_{\mathbb{T}_X}\int_{\mathbb{T}_X} \hat{v}_X(s,u)\hat{w}_i^{[i]}(s,t)\hat{w}_i^{[i]}(u,t)\mathrm{d}s\mathrm{d}u\mathrm{d}t\right\}^{-1/2}\hat{w}_i^{[i]}$.
end for

A prediction for $\eta(X_0)$ at (5.10), alternative to $\hat{\eta}_{p,\mathrm{fAPLS}}(X_0)$ at (5.11), is thus given by

$$\tilde{\eta}_{p,\mathrm{fAPLS}}(X_0)(t) = \bar{Y}(t) + \int_{\mathbb{T}_X} X_0^{\mathrm{cent}}(s)\tilde{\beta}_{p,\mathrm{fAPLS}}(s,t)\mathrm{d}s. \tag{5.14}$$

It is worth emphasizing that, in exact arithmetic, $\hat{\beta}_{p,\mathrm{fAPLS}}$ at (5.8) (resp. $\hat{\eta}_{p,\mathrm{fAPLS}}$ at (5.11)) is identical to $\tilde{\beta}_{p,\mathrm{fAPLS}}$ at (5.13) (resp. $\tilde{\eta}_{p,\mathrm{fAPLS}}$ at (5.14)), because $\{\widehat{\mathcal{V}}_X^i(\beta) \mid 1 \leq i \leq p\}$ and $\{\hat{w}_i \mid 1 \leq i \leq p\}$ literally span the same space. Nevertheless, in practice $\tilde{\beta}_{p,\mathrm{fAPLS}}$ and $\tilde{\eta}_{p,\mathrm{fAPLS}}$ stand out due to their numerical stability for finite precision arithmetic, whereas the more explicit expressions of $\hat{\beta}_{p,\mathrm{fAPLS}}$ and $\hat{\eta}_{p,\mathrm{fAPLS}}$ make themselves preferred in theoretical derivations.

There is one hyper-parameter to tune. We use the generalized cross-validation (GCV) again; in particular, p is chosen within $[1, p_{\max}]$ as the minimizer of

$$\mathrm{GCV}(p) = (n-p-1)^{-2}\sum_{i=1}^{n}\int_{\mathbb{T}_Y}\{Y_i(t) - \tilde{\eta}_{p,\mathrm{fAPLS}}(X_i)(t)\}^2\mathrm{d}t,$$

where $p_{\max}$ depends upon FVE at (2.18) such that $\mathrm{FVE}(p_{\max})$ barely exceeds a pre-determined close-to-one threshold, e.g., 99% in Section 5.3.

5.2.1 Asymptotic properties

Under regularity conditions, Proposition 5.2 (resp. Proposition 5.3) verifies the consistency in L_2 and/or supremum metric (in probability) of $\hat{\beta}_{p,\mathrm{fAPLS}}$ (resp. $\hat{\eta}_{p,\mathrm{fAPLS}}(X_0)$). In these results, we allow p to diverge as a function of n, but its rate is capped to be at most $O(\sqrt{n})$ if $\|v_X\|_2 < 1$ and even slower otherwise. More discussion of the technical assumptions may be found at the beginning of Appendix A.4.

Proposition 5.2. *Assuming (C.A.4.1)–(C.A.4.5), as n diverges, $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_2 = o_p(1)$. If we strengthen (C.A.4.5) to (C.A.4.6), then the convergence becomes uniform, i.e., $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_\infty = o_p(1)$, with $\|\cdot\|_\infty$ denoting the supremum norm.*

Proposition 5.3. *Given $X_0 \sim X$, conditions (C.A.4.1)–(C.A.4.5) suffice to imply convergence to 0 (in probability) of $\|\hat{\eta}_{p,\mathrm{fAPLS}}(X_0)-\eta(X_0)\|_2$ (i.e., $\|\hat{\eta}_{p,\mathrm{fAPLS}}(X_0)-\eta(X_0)\|_2 = o_p(1)$), while the uniform version (viz. $\|\hat{\eta}_{p,\mathrm{fAPLS}}(X_0) - \eta(X_0)\|_\infty = o_p(1)$) is entailed jointly by (C.A.4.1)–(C.A.4.4) and (C.A.4.6)–(C.A.4.7).*

5.3 Numerical study

Our proposal fAPLS was compared with competitors in terms of ReISEE at (4.28), viz.

$$\mathrm{ReISEE} = \frac{\|\beta - \hat{\beta}\|_2^2}{\|\beta\|_2^2},$$

and/or relative integrated squared prediction error (ReISPE)

$$\mathrm{ReISPE} = \frac{\sum_{i\in \mathrm{ID}_{\mathrm{test}}} \|Y_i - \hat{Y}_i\|_2^2}{\sum_{i\in \mathrm{ID}_{\mathrm{test}}} \|Y_i - \bar{Y}_{\mathrm{train}}\|_2^2},$$

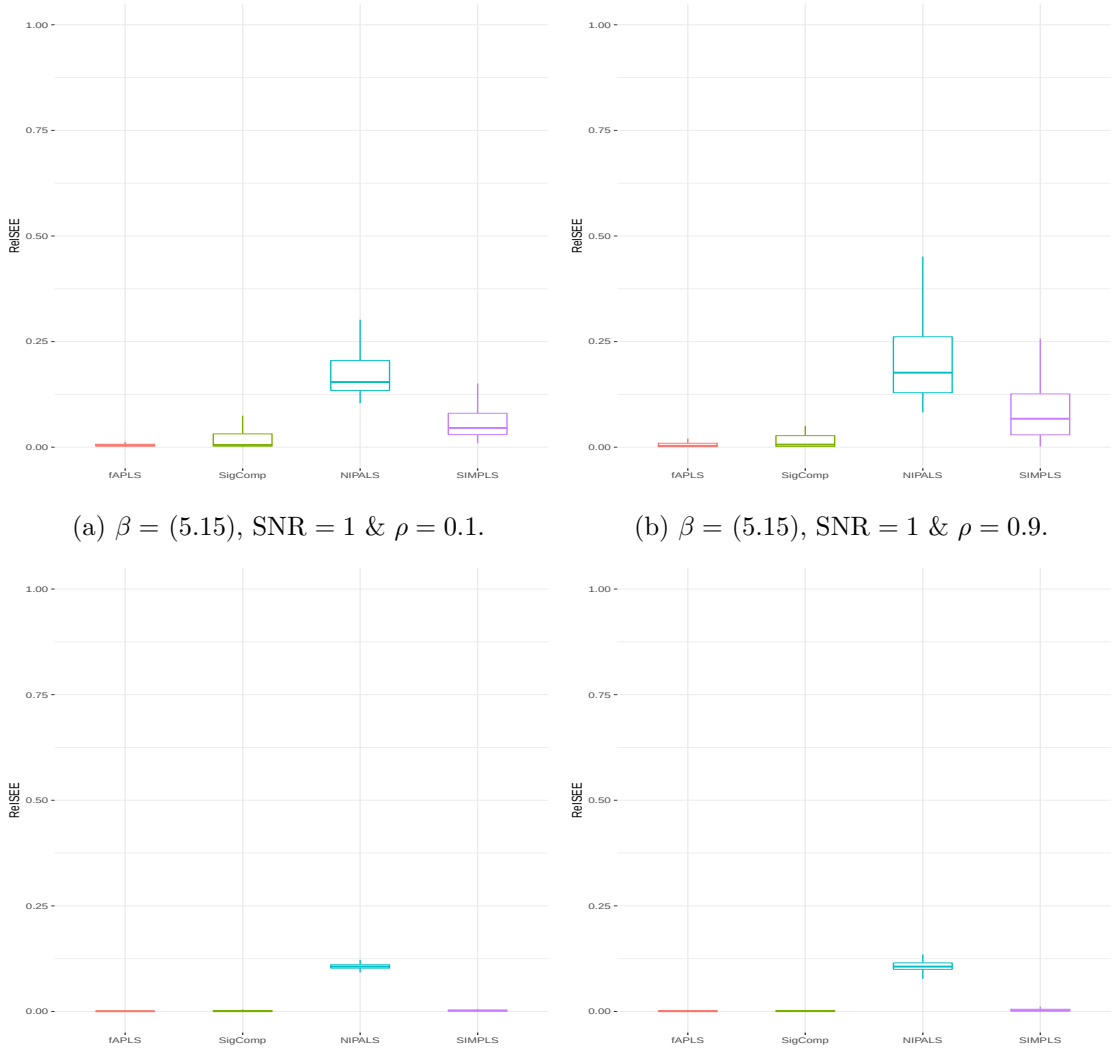
where $\hat{\beta}$ estimates β and $\hat{Y}_i$ predicts Y_i, $1 \le i \le n$; here we denote by $\mathrm{ID}_{\mathrm{test}}$ the index set for testing. Subsequent comparisons involved other FPLS routes for FoFR, including SigComp [61] and (functional) NIPALS and SIMPLS [11]. We referred to their original source codes posted respectively at `R` package `FRegSigCom` [62] and GitHub (`https://github.com/hanshang/FPLSR`; accessed on June 12, 2020). Code trunks for our implementation are currently available at GitHub too (`https://github.com/ZhiyangGeeZhou/fAPLS`; accessed on June 12, 2020).

5.3.1 Simulation

Each of the 200 toy samples consisted of n (= 300) independent and identically distributed (iid) pairs of trajectories (with 80% used for training). For simplicity, assume $\mu_X = \mu_Y = 0$. We took 100, 10 and 1 as the top three eigenvalues of $\mathcal{V}_X$, whereas $\lambda_{i,X} = 0$ for all $i \ge 4$. Correspondingly, the first three eigenfunctions of $\mathcal{V}_X$ were respectively set to be (normalized) shifted Legendre polynomials of order 2 to 4 [45, pp. 773–774] (these were also involved in Section 3.3.1), say P_2, P_3 and P_4, viz.

$$\begin{aligned}
\phi_{1,X}(t) &= P_2(t) = \sqrt{5}(6t^2 - 6t + 1),\\
\phi_{2,X}(t) &= P_3(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1),\\
\phi_{3,X}(t) &= P_4(t) = 3(70t^4 - 140t^3 + 90t^2 - 20t + 1).
\end{aligned}$$

As is known, these polynomials are of unit norm and mutually orthogonal on $[0, 1]$ (this set is both $\mathbb{T}_X$ and $\mathbb{T}_Y$ in our simulation). Two sorts of slope functions were respectively given



(a) $\beta = (5.15)$, SNR $= 1$ & $\rho = 0.1$.

(b) $\beta = (5.15)$, SNR $= 1$ & $\rho = 0.9$.

(c) $\beta = (5.15)$, SNR $= 5$ & $\rho = 0.1$.

(d) $\beta = (5.15)$, SNR $= 5$ & $\rho = 0.9$.

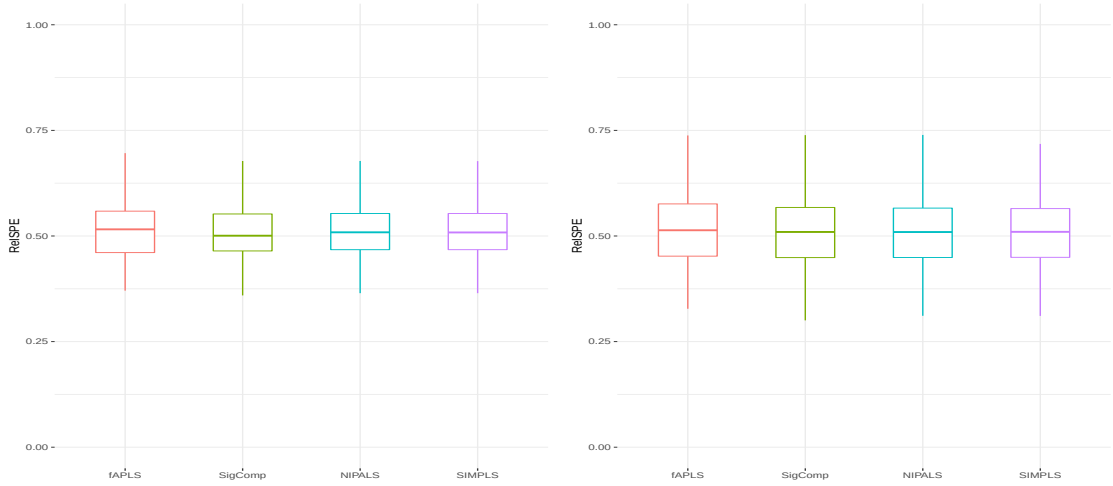
Figure 5.1: Boxplots of ReISEE values for simulation with β at (5.15). The four boxes in each subfigure, from left to right, correspond to fAPLS, SigComp, NIPALS and SIMPLS, respectively. All the plots come with the identical scale.

by

$$\beta(s,t) = P_2(s)P_2(t), \tag{5.15}$$

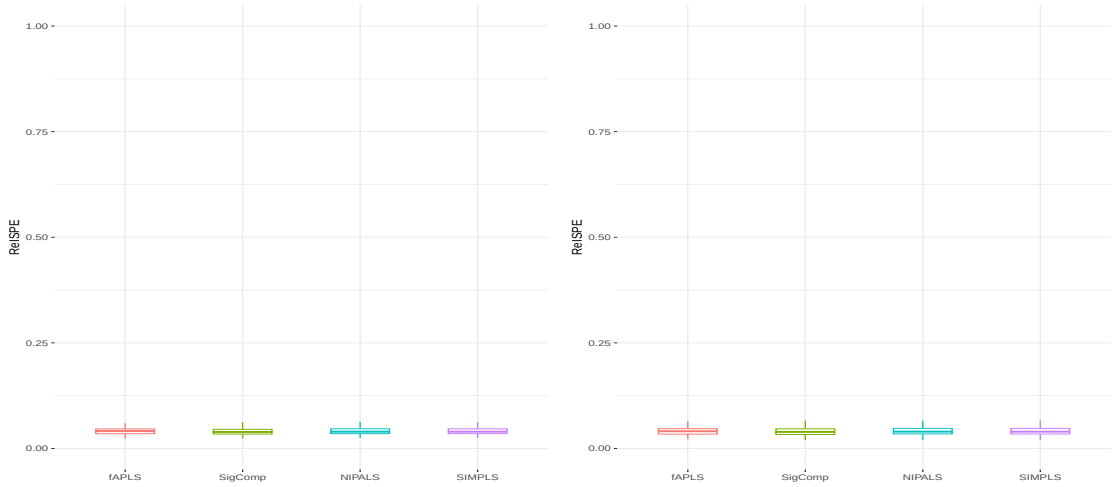
$$\beta(s,t) = P_4(s)P_4(t). \tag{5.16}$$

For our zero-mean Gaussian process ε, we chose the covariance function $r_\varepsilon = r_\varepsilon(s,t) = \sigma^2\rho^{|s-t|}$, with ρ controlling the autocorrelation of ε and σ determined by the value of signal-to-noise ratio (SNR $= \sigma^{-1}\sqrt{\text{var}(\|Y\|_2^2)}$). Different values of ρ (resp. SNR) were involved: 0.1 and 0.9 (resp. 1 and 5). In total there were eight combinations of $(\beta, \text{SNR}, \rho)$.



(a) $\beta = (5.15)$, SNR $= 1$ & $\rho = 0.1$.

(b) $\beta = (5.15)$, SNR $= 1$ & $\rho = 0.9$.

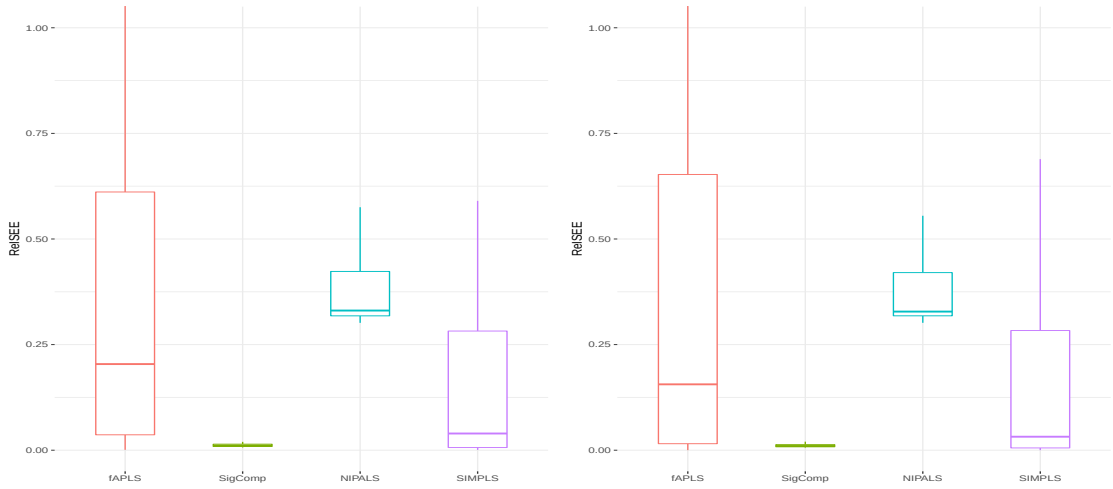


(c) $\beta = (5.15)$, SNR $= 5$ & $\rho = 0.1$.

(d) $\beta = (5.15)$, SNR $= 5$ & $\rho = 0.9$.

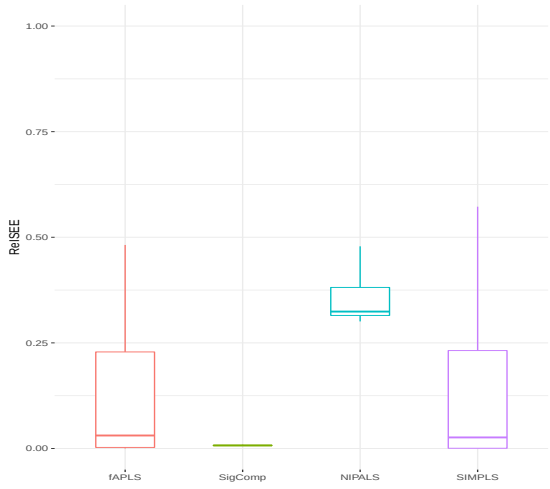
Figure 5.2: Boxplots of ReISPE values for simulation with β at (5.15). The four boxes in each subfigure, from left to right, correspond to fAPLS, SigComp, NIPALS and SIMPLS, respectively. All the plots come with the identical scale.

A common point shared by Figures 5.1–5.4 was that the two plots of the same row differ little. That is, ρ, the degree of autocorrelation of error process, had little impact on estimation or prediction. This phenomenon was consistent with observations in the multivariate context. Fixing levels of β and ρ, as SNR became larger, each approach led to relatively higher accuracy (or equivalently, lower values of ReISEE and ReISPE). Profiting from the smoothness penalty, SigComp was the most accurate strategy under almost all the settings; in general the prediction and estimation accuracy of fAPLS was comparable to that of NIPALS and SIMPLS. In particular, when the signal was absolutely strong (viz. β at (5.15)), fAPLS produced satisfactory estimators (see Figure 5.1) and was fully competitive in terms

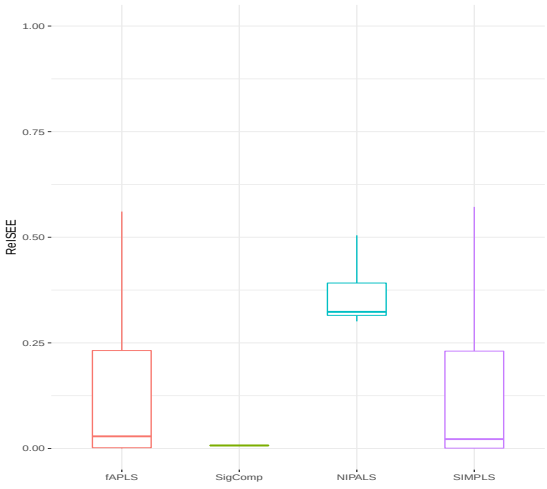


(a) $\beta = (5.16)$, SNR $= 1$ & $\rho = 0.1$.

(b) $\beta = (5.16)$, SNR $= 1$ & $\rho = 0.9$.



(c) $\beta = (5.16)$, SNR $= 5$ & $\rho = 0.1$.

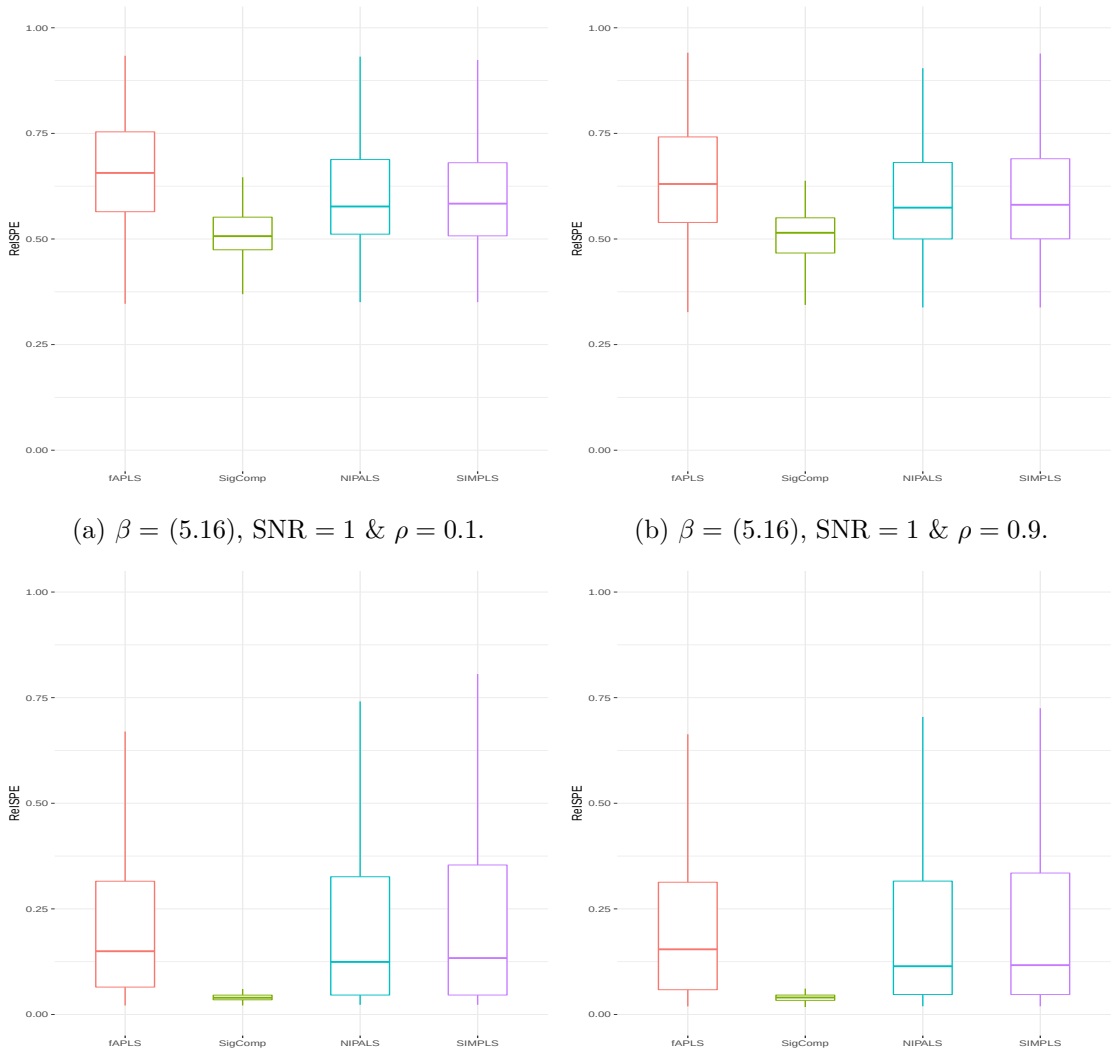


(d) $\beta = (5.16)$, SNR $= 5$ & $\rho = 0.9$.

Figure 5.3: Boxplots of ReISEE values for simulation with β at (5.16). The four boxes in each subfigure, from left to right, correspond to fAPLS, SigComp, NIPALS and SIMPLS, respectively. All the plots come with the identical scale.

of prediction (see Figure 5.2). Encountering the weakest (both absolutely and relatively) signal (viz. β at (5.16) and SNR $= 1$), fAPLS performed the worst: its estimation error was the most fluctuating (see Figures 5.3a and 5.3b), though in this case fAPLS prediction errors were still comparable with those given by NIPALS and SIMPLS (see Figures 5.4a and 5.4b).

The biggest advantage of fAPLS was on the running time: under all the eight simulation settings, it ran much faster than the other three (see Table 5.1). This phenomenon was not surprising, because, compared with the other three competing routes, fAPLS involves neither eigendecomposition nor choosing a tuning parameter for a penalty.



(a) $\beta = (5.16)$, SNR $= 1$ & $\rho = 0.1$.

(b) $\beta = (5.16)$, SNR $= 1$ & $\rho = 0.9$.

(c) $\beta = (5.16)$, SNR $= 5$ & $\rho = 0.1$.

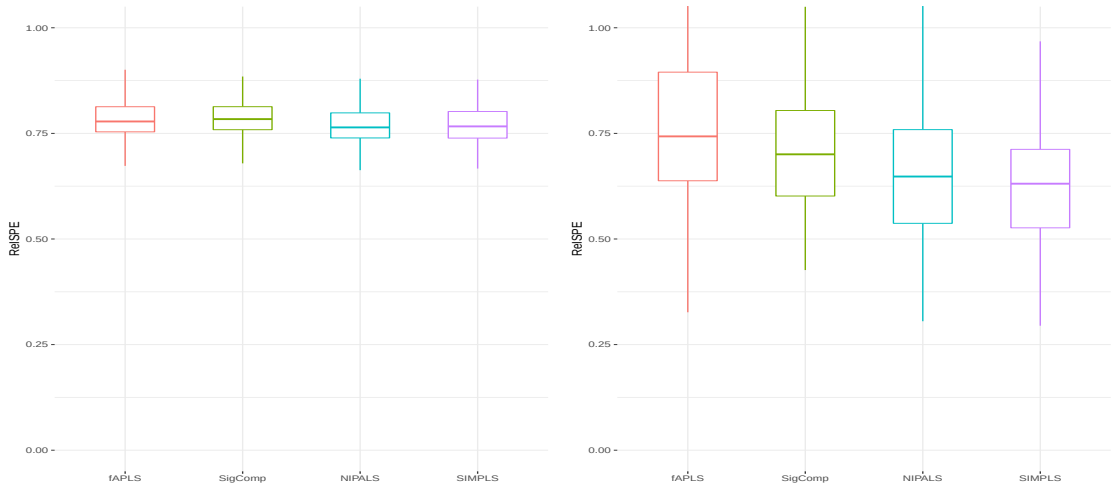
(d) $\beta = (5.16)$, SNR $= 5$ & $\rho = 0.9$.

Figure 5.4: Boxplots of ReISPE values for simulation with β at (5.16). The four boxes in each subfigure, from left to right, correspond to fAPLS, SigComp, NIPALS and SIMPLS, respectively. All the plots come with the identical scale.

5.3.2 Application

We now revisit the two datasets described in Section 5.1. For DTI (resp. BG) data, we took CCA FA tract profiles (resp. hip angle curves) as predictors and RCST FA tract profiles (resp. knee angle curves) as responses. For each dataset, we repeated the random split for 200 times: taking roughly 20% of all the data points for testing and using the remainder for training. After analyzing these training subsets, corresponding to each approach, we generated 200 ReISPE values.

Outputs for DTI data from the four approaches were fairly close to each other in terms of ReISPE (see Figure 5.5a), while BG data seemed in favor of SIMPLS (see Figure 5.5b).



(a) DTI (b) BG

Figure 5.5: Boxplots of ReISPE values for two applications. The plots have identical scales.

We guess the relatively small sample size (= 39) of BG data was a cause deteriorating fAPLS predictions.

There was a lack of dominant eigenvalues of $\mathcal{V}_X$ for DTI data. As a consequence, $p_{\max}$ became as high as 23, slowing down the implementation of fAPLS. That is, as is seen in Table 5.1 that, compared with other cases, DTI dataset consumed much more time in running fAPLS.

5.4 Concluding remarks

When fitting FoFR, we suggest fAPLS, a route of FPLS via Krylov subspaces. The fAPLS estimator has a concise and explicit expression. Meanwhile, we introduce an alternative and equivalent form of it, which stabilizes numerical outputs. fAPLS is competitive with existing FPLS routes in terms of estimation and prediction errors and is less computationally involved.

Up to this point we have avoided applications to geodata. Spatial correlation (i.e., if X_i and X_j, $i \neq j$ are no longer mutually independent) can lead to inconsistency of PLS estimators; see Singer et al. [86, Theorem 1] for the multivariate context with single-vector-response. A naive correction, transplanted from Singer et al. [86, Section 4.1], is to instead implement the regression on transformed observations (X_i^*, Y_i^*), $i = 1, \ldots, n$, such that, for all $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$, $[X_1^*(s), \ldots, X_n^*(s)]^\top = \boldsymbol{V}_X^{-1/2}(s)[X_1(s), \ldots, X_n(s)]^\top$ and $[Y_1^*(t), \ldots, Y_n^*(t)]^\top = \boldsymbol{V}_Y^{-1/2}(t)[Y_1(t), \ldots, Y_n(t)]^\top$, with $n \times n$ matrices $\boldsymbol{V}_X(s) = [\mathrm{cov}\{X_i(s), X_j(s)\}]$ and $\boldsymbol{V}_Y(t) = [\mathrm{cov}\{Y_i(t), Y_j(t)\}]$. But it is even challenging to recover $\boldsymbol{V}_X$ and $\boldsymbol{V}_Y$ sufficiently accurately without specifying the dependence structure, since there is only one observation

Table 5.1: Time consumed (in seconds) by 200 repeats in numerical studies (running on a laptop with Intel® Core™ i5-5200U CPU @2×2.20 GHz and 8 GB RAM)

	$\beta^*(s,t) = P_2(s)P_2(t)$			
	SNR = 1		SNR = 5	
	$\rho = 0.1$	$\rho = 0.9$	$\rho = 0.1$	$\rho = 0.9$
fAPLS	6.0	6.1	6.1	6.2
SigComp	363.4	372.7	365.2	373.2
NIPALS	197.6	199.1	195.0	197.6
SIMPLS	183.2	187.9	181.2	188.2

	$\beta^*(s,t) = P_4(s)P_4(t)$			
	SNR = 1		SNR = 5	
	$\rho = 0.1$	$\rho = 0.9$	$\rho = 0.1$	$\rho = 0.9$
fAPLS	6.0	6.1	6.1	6.2
SigComp	363.4	372.7	365.2	373.2
NIPALS	197.6	199.1	195.0	197.6
SIMPLS	183.2	187.9	181.2	188.2

	Application	
	DTI	BG
fAPLS	94.1	2.5
SigComp	837.2	14.1
NIPALS	835.2	35.3
SIMPLS	125.5	20.0

for each i. Alternatively and more practically, one can target at correcting naive $\widehat{\mathcal{V}}_X$ and $\hat{r}_{XY}$ for dependent subjects; Paul and Peng [71] offers a solution along these lines.

fAPLS has a heuristic extension to multiple functional covariates, i.e., associated with each realization $Y_i \sim Y$, there are $m > 1$ functional covariates, say $X_{ij} \sim X_{\cdot j}$, $1 \le j \le m$, and correspondingly m coefficient functions $\beta^{*(j)}$, $1 \le j \le m$. In particular,

$$Y_i(t) = \mu_Y(t) + \sum_{i=1}^{m} \mathcal{L}_{X_{ij}}(\beta^{*(j)}) + \varepsilon_i(t),$$

where Y_i and X_{ij} are assumed to be independent across all i. Following the idea of (5.3), an ad hoc estimator for true $(\beta^{*(1)}, \ldots, \beta^{*(m)})$ is thus

$$(\hat{\beta}^{(1)}_{\text{fAPLS}}, \ldots, \hat{\beta}^{(m)}_{\text{fAPLS}}) = \underset{\beta^{(j)} \in \text{KS}_p(\widehat{\mathcal{V}}_{X_{\cdot j} X_{\cdot j}}, \beta^{*(j)}),\ 1 \le j \le m}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \int_{\mathbb{T}_Y} \left\{ Y_i(t) - \bar{Y}_i(t) - \sum_{j=1}^{m} \int_{\mathbb{T}_{X_{\cdot j}}} (X_{ij} - \bar{X}_{\cdot j})(s) \beta^{(j)}(s,t) \mathrm{d}s \right\}^2 \mathrm{d}t,$$

with $\bar{X}_{\cdot j} = m^{-1} \sum_{j=1}^{m} X_{ij}$ and domains $\mathbb{T}_{X_{\cdot j}}$ varying with j. Of course, it becomes necessary to introduce penalties once the above minimizer is not uniquely defined.

Chapter 6

Future perspectives

In this dissertation, approximations to β in fitting SoFR and FoFR are always taken from linear spaces spanned by certain basis functions. It is also feasible to assume that the slope function resides in other sorts of spaces, e.g., as in [106, 60, 89], a reproducing kernel Hilbert space (RKHS) induced by a positive semi-definite bivariate kernel, say $\kappa(\cdot, \cdot)$. Typically $\kappa(\cdot, \cdot)$ is made of Bernoulli polynomials [29]. A more data-driven option for $\kappa(\cdot, \cdot)$ may lead to better performance.

Indeed it is promising to consider a model slightly more general than linear ones, say the single index model [87]:

$$Y = f\left(\int_{\mathbb{T}} X\beta\right) + \text{error}, \tag{6.1}$$

where underlying $f(\cdot)$ is allowed to be any smooth real-valued function defined on $\mathbb{R}$. Pioneer works like [68] could be of great help in extending FPLS for (6.1). [33, Eq. (1.1)] further generalizes (6.1) to the nonparametric scalar-on-function regression:

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

with unspecified non-linear operator $m : L^2(\mathbb{T}_X) \to \mathbb{R}$. The functional Nadaraya-Watson (FNW) estimator [79] for $m(\cdot)$ involves a pre-defined semi-metric, say $d(\cdot, \cdot)$, virtually defining the similarity among paired predictor curves. Existing candidates for $d(\cdot, \cdot)$ include the ones based on FPC, FPLS basis functions or the ensemble of FPC and FPLS via stacking [36]. Another strategy for estimating $m(\cdot)$ is to restrict it to an RKHS, resulting in estimators with closed forms [79]. Both FNW and RKHS approaches have the potential to be improved by manipulating the extent of supervision.

Our proposals apply to more complex models including the (functional) generalized linear models and proportional hazard (PH) models. Inherited from Marx [65], the basic idea is to maximize likelihood via iteratively reweighted LS (IRLS, Green [41]) and then to embed methods for linear models into each step of IRLS. Successful recent applications of this strategy include [3, 98]: [3] classified curves by fitting logistic regression models; the

joint modeling in [98] consisted of a functional linear mixed-effects model and a PH model, incorporating APLS with IRLS.

Bibliography

[1] A. Aguilera, M. Escabias, C. Preda, and G. Saporta. Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemometrics Intell. Lab. Syst.*, 104:289–305, 2010. doi: 10.1016/j.chemolab.2010.09.007.

[2] A. Aguilera, M. Aguilera-Morillo, and C. Preda. Penalized versions of functional PLS regression. *Chemometrics Intell. Lab. Syst.*, 154:80–92, 2016. doi: 10.1016/j.chemolab.2016.03.013.

[3] A. M. H. Albaqshi. *Generalized Partial Least Squares Approach for Nominal Multinomial Logit Regression Models with a Functional Covariate*. PhD thesis, University of Northern Colorado, 2017.

[4] A. L. Alexander, J. E. Lee, M. Lazar, and A. S. Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4:316–329, 2007. doi: 10.1016/j.nurt.2007.05.011.

[5] T. Amemiya. *Advanced Econometrics*. Harvard University Press, Cambridge, 1985.

[6] T. S. Angell and A. Kirsch. *Optimization Methods in Electromagnetic Radiation*. Springer Monographs in Mathematics. Springer, New York, 2004. doi: 10.1007/b97629.

[7] A. Baíllo. A note on functional linear regression. *J. Stat. Comput. Simul.*, 79:657–669, 2009. doi: 10.1080/00949650701836765.

[8] D. Benatia, M. Carrasco, and J.-P. Florens. Functional linear regression with functional response. *J. Econom.*, 201:269–291, 2017. doi: 10.1016/j.jeconom.2017.08.008.

[9] K. Berer and G. Krishnamoorthy. Microbial view of central nervous system autoimmunity. *FEBS Lett.*, 588:4207–4213, 2014. doi: 10.1016/j.febslet.2014.04.007.

[10] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012.

[11] U. Beyaztas and H. L. Shang. On function-on-function regression: partial least squares approach. *Environ. Ecol. Stat.*, 27:95–114, 2020. doi: 10.1007/s10651-019-00436-1.

[12] A. C. Bissett. *Improvements to PLS Methodology*. PhD thesis, University of Manchester, 2015.

[13] A. Björkström and R. Sundberg. A generalized view on continuum regression. *Scand. J. Stat.*, 26:17–30, 1999. doi: 10.1111/1467-9469.00134.

[14] R. Bro and L. Eldén. Pls works. *J. Chemometrics*, 23:69–71, 2009. doi: 10.1002/cem.1177.

[15] R. Brooks and M. Stone. Joint continuum regression for multiple predictands. *J. Am. Stat. Assoc.*, 89:1374–1377, 1994. doi: 10.1080/01621459.1994.10476876.

[16] F. Y. Chan and T. K. Mak. Discussion on "continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression". *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 52:264–265, 1990.

[17] X. Chen and R. D. Cook. Some insights into continuum regression and its asymptotic properties. *Biometrika*, 97:985–989, 2010. doi: 10.1093/biomet/asq024.

[18] X. Chen and L.-P. Zhu. Connecting continuum regression with sufficient dimension reduction. *Stat. Probab. Lett.*, 98:44–49, 2015. doi: 10.1016/j.spl.2014.12.007.

[19] C. K. Chui. Concerning rates of convergence of Riemann sums. *J. Approx. Theory*, 4:279–287, 1971. doi: 10.1016/0021-9045(71)90016-5.

[20] R. D. Cook. Graphics for regressions with a binary response. *J. Am. Stat. Assoc.*, 91:983–992, 1996. doi: 10.2307/2291717.

[21] R. D. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York, 1998. doi: 10.1002/9780470316931.

[22] R. D. Cook and L. Forzani. Partial least squares prediction in high-dimensional regression. *Ann. Stat.*, 47:884–908, 2019.

[23] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numer. Math.*, 31:377–403, 1978. doi: 10.1007/BF01404567.

[24] X. Dai, P. Z. Hadjipantelis, K. Han, and H. Ji. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2018. URL `https://CRAN.R-project.org/package=fdapace`. R package version 0.4.0.

[25] C. de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, New York, revised edition, 2001.

[26] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst.*, 18:251–263, 1993. doi: 10.1016/0169-7439(93)85002-X.

[27] A. Delaigle and P. Hall. Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 74:267–286, 2012. doi: 10.1111/j.1467-9868.2011.01003.x.

[28] A. Delaigle and P. Hall. Methodology and theory for partial least squares applied to functional data. *Ann. Stat.*, 40:322–352, 2012. doi: 10.1214/11-AOS958.

[29] K. Dilcher. Bernoulli and euler polynomials. In F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors, *NIST Handbook of Mathematical Functions*, pages 587–599. Cambridge University Press, New York, 2010.

[30] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, 1996.

[31] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *Int. Stat. Rev.*, 85:61–83, 2017. doi: 10.1111/insr.12116.

[32] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Comput. Stat. Data Anal.*, 44:161–173, 2003. doi: 10.1016/S0167-9473(03)00032-X.

[33] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer, New York, 2006. doi: 10.1007/0-387-36620-2.

[34] E. I. Fredholm. Sur une classe d'équations fonctionnelles. *Acta Math.*, 27:365–390, 1903. doi: 10.1007/BF02421317.

[35] P. Galeano, E. Joseph, and R. E. Lillo. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57:281–291, 2015. doi: 10.1080/00401706.2014.902774.

[36] J. Goldsmith and F. Scheipl. Estimator selection and combination in scalar-on-function regression. *Comput. Stat. Data Anal.*, 70:362–372, 2014. doi: 10.1016/j.csda.2013.10.009.

[37] J. Goldsmith, J. Bob, C. M. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *J. Comput. Graph. Stat.*, 20:830–851, 2011. doi: 10.1198/jcgs.2010.10007.

[38] J. Goldsmith, C. M. Crainiceanu, B. Caffo, and D. Reich. Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, 61:453–469, 2012. doi: 10.1111/j.1467-9876.2011.01031.x.

[39] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss. *refund: Regression with Functional Data*, 2016. R package version 0.1-16.

[40] C. Goutis. Second-derivative functional regression with applications to near infrared spectroscopy. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 60:103–114, 1998. doi: 10.1111/1467-9868.00111.

[41] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 46:149–192, 1984.

[42] P. Hall, H.-G. MÃijller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Stat.*, 34:1493–1517, 2006. doi: 10.1214/009053606000000272.

[43] D. Harville. Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Stat.*, 4:384–395, 1976. doi: 10.1214/aos/1176343414.

[44] G. He, H.-G. Müller, J.-L. Wang, and W. Yang. Functional linear regression via canonical analysis. *Bernoulli*, 16:705–729, 2010. doi: 10.3150/09-BEJ228.

[45] W. Hochstrasser. Orthogonal polynomials. In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, pages 771–802. National Bureau of Standards, Washington D.C., 1972. 10th printing.

[46] J. Hoffmann-Jørgensen. Necessary and sufficient condition for the uniform law of large numbers. In A. Beck, R. Dudley, M. Hahn, J. Kuelbs, and M. Marcus, editors, *Probability in Banach Spaces V*, volume 1153 of *Lecture Notes in Mathematics*, pages 258–272. Springer, Berlin, 1985.

[47] J. Hoffmann-Jørgensen and G. Pisier. The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probab.*, 4:587–599, 1976.

[48] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, New York, 2012. doi: 10.1007/978-1-4614-3655-3.

[49] A. E. Ivanescu, A.-M. Staicu, F. Scheipl, and S. Greven. Penalized function-on-function regression. *Comput. Stat.*, 30:539–568, 2015.

[50] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000. doi: 10.1093/biomet/87.3.587.

[51] R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Stat.*, 40:633–643, 1969. doi: 10.1214/aoms/1177697731.

[52] S. Jung. Continuum directions for supervised dimension reduction. *Comput. Stat. Data Anal.*, 125:27–43, 2018. doi: 10.1016/j.csda.2018.03.015.

[53] B. V. Khvedelidze. Fredholm theorems. In *Encyclopedia of Mathematics*. Springer and the European Mathematical Society, 2011. URL `https://www.encyclopediaofmath.org/index.php/Fredholm_theorems#References`. [Online; accessed 21-Sep-2019].

[54] T. H. Koornwinder, R. Wong, R. Koekoek, and R. F. Swarttouw. Orthogonal polynomials. In F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors, *NIST Handbook of Mathematical Functions*, pages 435–484. Cambridge University Press, New York, 2010.

[55] N. Krämer and M. Sugiyama. The degrees of freedom of partial least squares regression. *J. Am. Stat. Assoc.*, 106:697–705, 2011. doi: 10.1198/jasa.2011.tm10107.

[56] K. Lange. *Numerical Analysis for Statisticians*. Springer, New York, 2nd edition, 2010. doi: 10.1007/978-1-4419-5945-4.

[57] M. H. Lee and Y. Liu. Kernel continuum regression. *Comput. Stat. Data Anal.*, 68: 190–201, 2013. doi: 10.1016/j.csda.2013.06.016.

[58] X. Leng and H.-G. Müller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22:68–76, 2005. doi: 10.1093/bioinformatics/bti742.

[59] Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Stat.*, 38:3321–3351, 2010. doi: 10.1214/10-AOS813.

[60] H. Lian. Minimax prediction for functional linear regression with functional responses in reproducing kernel hilbert spaces. *J. Multivar. Anal.*, 140:395–402, 2015. doi: 10.1016/j.jmva.2015.06.005.

[61] R. Luo and X. Qi. Function-on-function linear regression by signal compression. *J. Am. Stat. Assoc.*, 112:690–705, 2017. doi: 10.1080/01621459.2016.1164053.

[62] R. Luo and X. Qi. *FRegSigCom: Functional Regression using Signal Compression Approach*, 2018. R package version 0.3.0.

[63] T. Lyche, C. Manni, and H. Speleers. Foundations of spline theory: B-splines, spline approximation, and hierarchical refinement. In A. Kunoth, T. Lyche, G. Sangalli, and S. Serra-Capizzano, editors, *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra*, pages 1–76. Springer, Cham, Switzerland, 2018. doi: 10.1007/978-3-319-94911-6.

[64] T. Maierhofer and F. Pfisterer. *classiFunc: Classification of Functional Data*, 2018. R package version 0.1.1.

[65] B. D. Marx. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38:374–381, 1996.

[66] H.-G. Müller, J.-L. Wang, W. B. Capra, P. Liedo, and J. R. Carey. Early mortality surge in protein-deprived females causes reversal of sex differential of life expectancy in Mediterranean fruit flies. *Proc. Natl. Acad. Sci. U. S. A.*, 94:2762–2765, 1997.

[67] C. Nadeau and Y. Bengio. Inference for the generalization error. *Mach. Learn.*, 52:239–281, 2003. doi: 10.1023/A:1024068626366.

[68] P. Naik and C. Tsai. Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 62:763–771, 2000. doi: 10.1111/1467-9868.00262.

[69] Y. Nie, L. Wang, B. Liu, and J. Cao. Supervised functional principal component analysis. *Stat. Comput.*, 28:713–723, 2018. doi: 10.1007/s11222-017-9758-2.

[70] F. Novomestky. *orthopolynom: Collection of functions for orthogonal and orthonormal polynomials*, 2013. URL `https://CRAN.R-project.org/package=orthopolynom`. R package version 1.0-5.

[71] D. Paul and J. Peng. Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach. *Electron. J. Statist.*, 5:1960–2003, 2011. doi: 10.1214/11-EJS662.

[72] J. Peng and D. Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Stat.*, 18:995–1015, 2009. doi: 10.1198/jcgs.2009.08011.

[73] C. Preda and G. Saporta. PLS regression on a stochastic process. *Comput. Stat. Data Anal.*, 48:149–158, 2005. doi: 10.1016/j.csda.2003.10.003.

[74] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. ver.3.6.1 "Action of the Toes".

[75] J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 53:539–572, 1991.

[76] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2005. doi: 10.1007/b98888.

[77] J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2017. R package version 2.4.7.

[78] P. T. Reiss and R. T. Ogden. Functional principal component regression and functional partial least squares. *J. Am. Stat. Assoc.*, 102:984–996, 2007. doi: 10.1198/016214507000000527.

[79] P. T. Reiss, J. Goldsmith, H. L. Shang, and R. T. Ogden. Methods for scalar-on-function regression. *Int. Stat. Rev.*, 85:228–249, 2017. doi: 10.1111/insr.12163.

[80] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69:730–742, 2006. doi: 10.1016/j.neucom.2005.12.010.

[81] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. RStudio version 1.1.383.

[82] T. Rubín and V. M. Panaretos. Sparsely observed functional time series: estimation and prediction. *Electron. J. Statist.*, 14:1137–1210, 2020. doi: 10.1214/20-EJS1690.

[83] P. Sang, L. Wang, and J. Cao. Parametric functional principal component analysis. *Biometrics*, 73:802–810, 2017. doi: 10.1111/biom.12641.

[84] S. Serneels, P. Filzmoser, C. Croux, and P. J. V. Espen. Robust continuum regression. *Chemometrics Intell. Lab. Syst.*, 76:197–204, 2005. doi: 10.1016/j.chemolab.2004.11.002.

[85] H. Shin. An extension of Fisher's discriminant analysis for stochastic processes. *J. Multivar. Anal.*, 99:1191–1216, 2008. doi: 10.1016/j.jmva.2007.08.001.

[86] M. Singer, T. Krivobokova, A. Munk, and B. de Groot. Partial least squares for dependent data. *Biometrika*, 103:351–362, 2016. doi: 10.1093/biomet/asw010.

[87] T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54:1461–1481, 1986. doi: 10.2307/1914309.

[88] M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 52:237–269, 1990.

[89] X. Sun, P. Du, X. Wang, and P. Ma. Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *J. Am. Stat. Assoc.*, 113:1601–1611, 2018. doi: 10.1080/01621459.2017.1356320.

[90] R. Sundberg. Continuum regression and ridge regression. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 55:653–659, 1993.

[91] H. Tasaki. Convergence rates of approximate sums of riemann integrals. *J. Approx. Theory*, 161:477–490, 2009. doi: 10.1016/j.jat.2008.10.005.

[92] T. M. Therneau. *A Package for Survival Analysis in S*, 2015. URL `https://CRAN.R-project.org/package=survival`. version 2.38.

[93] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. doi: 10.1007/978-1-4757-3294-8.

[94] T. N. Tombaugh. A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Arch. Clin. Neuropsychol.*, 21:53–76, 2006. doi: https://doi.org/10.1016/j.acn.2005.07.006.

[95] F. Utreras. Natural spline functions, their associated eigenvalue problem. *Numer. Math.*, 42:107–117, 1983. doi: 10.1007/BF01400921.

[96] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1990. doi: 10.1137/1.9781611970128.

[97] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annu. Rev. Stat. Appl.*, 3:257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624.

[98] Y. Wang, J. G. Ibrahim, and H. Zhu. Partial least squares for functional joint models with applications to the alzheimer's disease neuroimaging initiative study. *Biometrics*, 2020. doi: 10.1111/biom.13219. in press.

[99] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. doi: 10.1007/978-0-387-98141-3.

[100] H. Wold. Path models with latent variables: the NIPALS approach. In H. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi, editors, *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, pages 307–335. Academic Press, New York, 1975.

[101] L. Xiao. Asymptotic theory of penalized splines. *Electron. J. Statist.*, 13:747–794, 2019. doi: 10.1214/19-EJS1541.

[102] L. Xiao, C. Li, W. Checkley, and C. Crainiceanu. Fast covariance estimation for sparse functional data. *Stat. Comput.*, 28:511–522, 2018. doi: 10.1007/s11222-017-9744-8.

[103] L. Xiao, C. Li, W. Checkley, and C. Crainiceanu. *face: Fast Covariance Estimation for Sparse Functional Data*, 2019. URL `https://CRAN.R-project.org/package=face`. R package version 0.1-5.

[104] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.*, 100:577–590, 2005. doi: 10.1198/016214504000001745.

[105] F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *Ann. Stat.*, 33:2873–2903, 2005. doi: 10.1214/009053605000000660.

[106] M. Yuan and T. T. Cai. A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Stat.*, 38:3412–3444, 2010. doi: 10.1214/09-AOS772.

[107] E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications*. Applied Mathematical Sciences. Springer, New York, 1995. doi: 10.1007/978-1-4612-0821-1.

[108] L. Zhou, H. Lin, and H. Liang. Efficient estimation of the nonparametric mean and covariance functions for longitudinal and sparse functional data. *J. Am. Stat. Assoc.*, 113:1550–1564, 2018. doi: 10.1080/01621459.2017.1356317.

[109] Z. Zhou. Functional continuum regression. *J. Multivar. Anal.*, 173:328–346, 2019. doi: 10.1016/j.jmva.2019.03.006.

[110] Z. Zhou. Partial least squares for function-on-function regression via Krylov subspaces. arXiv:2005.04798, 2020.

[111] Z. Zhou and R. Lockhart. Partial least squares for sparsely observed curves with measurement errors. arXiv:2003.11542, 2020.

[112] Z. Zhou and P. Sang. Continuum centroid classifier for functional data. Under review, 2019.

Appendix A

Technical details

A.1 Technical details for Chapter 2

Definition A.1. Before moving further, we recall a few mathematical terms.

(D.A.1.1) Weak convergence. A sequence $\{x_n\}$ in $L^2(\mathbb{T}_X)$ is said to weakly converge to $x^* \in L^2(\mathbb{T}_X)$ if $\lim_{n\to\infty} \int_{\mathbb{T}_X} x_n u = \int_{\mathbb{T}_X} x^* u$ holding for each $u \in L^2(\mathbb{T}_X)$.

(D.A.1.2) Weak sequential closedness. A subset $C \subseteq L^2(\mathbb{T}_X)$ is said to be weakly sequentially closed if each weakly convergent sequence in C converges weakly to an element in C.

(D.A.1.3) Weak sequential upper semi-continuity. A real-valued function f defined on $L^2(\mathbb{T}_X)$ is weakly sequentially upper semi-continuous if $f(x^*) \geq \overline{\lim}_{n\to\infty} f(x_n)$ holds for every sequence $\{x_n\}$ converging weakly to x^*.

(D.A.1.4) Weak sequential compactness. A subset $C \subseteq L^2(\mathbb{T}_X)$ is weakly sequentially compact in $L^2(\mathbb{T}_X)$ if C is weakly sequentially closed and each sequence $\{x_n\}$ in C has a weakly convergent subsequence.

Lemma A.1 is the cornerstone of our proof of the existence of $w_{j,\alpha}$ and $\hat{w}_{j,\alpha}$. Lemma A.2, essential in proving Theorem 2.1, establishes the convergence of empirical $\widehat{T}^*_{j,\alpha}$ in (2.15) to its theoretical counterpart $T^*_{j,\alpha}$ in (2.13).

Lemma A.1. *Suppose $C \subseteq L^2(\mathbb{T}_X)$ is a bounded and weakly sequentially closed set (D.A.1.2). Suppose $f : C \to \mathbb{R}$ is weakly sequentially upper semi-continuous (D.A.1.3). Then f has a maximizer on C.*

Proof of Lemma A.1. Firstly prove that $f_0 = \sup_{x\in C} f(x) < \infty$. To the contrary, suppose that $f_0 = \infty$. Then there is a sequence x_n in C such that $f(x_n) \geq n$ for each $n \in \{1, 2, \ldots\}$. Since C is bounded and weakly sequentially closed (D.A.1.2), we may apply Alaoglu's theorem (see, e.g., [6, Theorem A.56]) to see that C is weakly sequentially compact (D.A.1.4), i.e., x_n must have a subsequence x_{n_k} weakly converging (D.A.1.1) to $x^* \in C$. Due to the

weakly sequential upper semi-continuity (D.A.1.3) of f, we have

$$f(x^*) \geq \overline{\lim_{k\to\infty}} f(x_{n_k}) \geq \overline{\lim_{k\to\infty}} n_k = \infty.$$

This contradicts the assumption that f is real-valued.

Next, there always exists a sequence x_n such that $\lim_{n\to\infty} f(x_n) = f_0$. Find a weakly convergent (D.A.1.1) subsequence x_{n_k} with limit $x^* \in C$. Thus,

$$f_0 = \sup_{x\in C} f(x) \geq f(x^*) \geq \overline{\lim_{k\to\infty}} f(x_{n_k}) = \lim_{n\to\infty} f(x_n) = f_0.$$

The sandwich rule indicates that $x^* \in C$ is a maximizer of f on C and completes this proof. □

Remark A.1. Although the assumption of Lemma A.1 can be further relaxed, Lemma A.1 suffices for our needs in this paper. For a more general version, please refer to Theorem 5.3 and Remark 5.4 in an unpublished 2013 technical report by Prof. Alen Alexanderian (`https://aalexan3.math.ncsu.edu/articles/hilbert.pdf`; accessed 21-Sep-2019).

Lemma A.2. *Recall $T^*_{j,\alpha}(w)$ in (2.13) and $\widehat{T}^*_{j,\alpha}(w)$ in (2.15). If $\|\hat{w}_{k,\alpha} - w_{k,\alpha}\|_2$ converges to zero in probability as n diverges for all $k \in \{1,\ldots,j-1\}$, then $\widehat{T}^*_{j,\alpha}(w)$ converges to $T^*_{j,\alpha}(w)$ in probability uniformly over the unit ball, i.e., for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{ \sup_{w:\|w\|_2\leq 1} |\widehat{T}^*_{j,\alpha}(w) - T^*_{j,\alpha}(w)| < \epsilon \right\} = 1.$$

Proof of Lemma A.2. The proof consists of three phases. First follow Eq. (5.1) in [28] to conclude that, as $n \to \infty$,

$$\widehat{\mathcal{V}}_{X^{(j,\alpha)}}(\beta) \to_p \mathcal{V}_{X^{(j,\alpha)}}(\beta) \quad \text{and} \quad \hat{v}_{X^{(j,\alpha)}} \to_p v_{X^{(j,\alpha)}},$$

both in the L^2 sense. Moreover, for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left\{ \int_{\mathbb{T}_X} \int_{\mathbb{T}_X} \{\hat{v}_{X^{(j,\alpha)}} - \hat{v}_{\widehat{X}^{(j,\alpha)}}\}^2 > \epsilon \right\} \subseteq \left\{ \int_{\mathbb{T}_X} \{X^{(j,\alpha)} - \widehat{X}^{(j,\alpha)}\}^2 > \delta \right\}$$

and

$$\left\{ \int_{\mathbb{T}_X} \{\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta) - \widehat{\mathcal{V}}_{X^{(j,\alpha)}}(\beta)\}^2 > \epsilon \right\} \subseteq \left\{ \int_{\mathbb{T}_X} \{X^{(j,\alpha)} - \widehat{X}^{(j,\alpha)}\}^2 > \delta \right\}.$$

The Continuous Mapping Theorem guarantees the convergence to zero in probability of $\|\widehat{X}^{(j,\alpha)} - X^{(j,\alpha)}\|_2$ and further yields that, in the L^2 sense,

$$\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta) \to_p \mathcal{V}_{X^{(j,\alpha)}}(\beta) \quad \text{and} \quad \hat{v}_{\widehat{X}^{(j,\alpha)}} \to_p v_{X^{(j,\alpha)}}.$$

Recall $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(s,t) = n^{-1}\sum_{i=1}^n \widehat{X}_i^{(j,\alpha)}(s)\widehat{X}_i^{(j,\alpha)}(t)$ and $\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta) = n^{-1}\sum_{i=1}^n \widehat{X}_i^{(j,\alpha)}\widehat{Y}_i^{(j,\alpha)}$. For convenience, write

$$f_{j,\alpha} = f_{j,\alpha}(w) = \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(\beta) \quad \text{and} \quad g_{j,\alpha} = g_{j,\alpha}(w) = \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(w)$$

and their empirical counterparts

$$\hat{f}_{j,\alpha} = \hat{f}_{j,\alpha}(w) = \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta) \quad \text{and} \quad \hat{g}_{j,\alpha} = \hat{g}_{j,\alpha}(w) = \int_{\mathbb{T}_X} w\widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(w).$$

By the Cauchy–Schwarz inequality, as $n \to \infty$,

$$\begin{aligned} \sup_{w:\|w\|_2\leq 1} |f_{j,\alpha}(w) - \hat{f}_{j,\alpha}(w)| &= \sup_{w:\|w\|_2\leq 1} \left| \int_{\mathbb{T}_X} w\{\mathcal{V}_{X^{(j,\alpha)}}(\beta) - \widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta)\} \right| \\ &\leq \|\mathcal{V}_{X^{(j,\alpha)}}(\beta) - \widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(\beta)\|_2 \to_p 0, \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \sup_{w:\|w\|_2\leq 1} |g_{j,\alpha}(w) - \hat{g}_{j,\alpha}(w)| &= \sup_{w:\|w\|_2\leq 1} \left| \int_{\mathbb{T}_X} w\{\mathcal{V}_{X^{(j,\alpha)}}(w) - \widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}}(w)\} \right| \\ &\leq \|v_{X^{(j,\alpha)}} - \hat{v}_{\widehat{X}^{(j,\alpha)}}\|_2 \to_p 0. \end{aligned} \tag{A.2}$$

Next we deduce a continuous mapping theorem specific for uniform convergence in probability. Suppose m is a continuous $\mathbb{R}^2 \to \mathbb{R}$ function. For arbitrary $\epsilon > 0$, there are $w_{n,\epsilon} \in \{w : \|w\|_2 \leq 1\}$ and $\delta > 0$ such that

$$\begin{aligned} &\left\{ \sup_{w:\|w\|_2\leq 1} |m\{f_{j,\alpha}(w), g_{j,\alpha}(w)\} - m\{\hat{f}_{j,\alpha}(w), \hat{g}_{j,\alpha}(w)\}| > \epsilon \right\} \\ &\quad \subseteq \left\{ |m\{f_{j,\alpha}(w_{n,\epsilon}), g_{j,\alpha}(w_{n,\epsilon})\} - m\{\hat{f}_{j,\alpha}(w_{n,\epsilon}), \hat{g}_{j,\alpha}(w_{n,\epsilon})\}| > \epsilon \right\} \\ &\quad \subseteq \left\{ |f_{j,\alpha}(w_{n,\epsilon}) - \hat{f}_{j,\alpha}(w_{n,\epsilon})|^2 + |g_{j,\alpha}(w_{n,\epsilon}) - \hat{g}_{j,\alpha}(w_{n,\epsilon})|^2 > \delta^2/2 \right\} \\ &\quad \subseteq \left\{ |f_{j,\alpha}(w_{n,\epsilon}) - \hat{f}_{j,\alpha}(w_{n,\epsilon})| > \delta/2 \right\} \cup \{|g_{j,\alpha}(w_{n,\epsilon}) - \hat{g}_{j,\alpha}(w_{n,\epsilon})| > \delta/2\} \\ &\quad \subseteq \left\{ \sup_{w:\|w\|_2\leq 1} |f_{j,\alpha}(w) - \hat{f}_{j,\alpha}(w)| > \delta/2 \right\} \bigcup \left\{ \sup_{w:\|w\|_2\leq 1} |g_{j,\alpha}(w) - \hat{g}_{j,\alpha}(w)| > \delta/2 \right\}, \end{aligned}$$

which further indicates that

$$\lim_{n\to\infty} \Pr \left\{ \sup_{w:\|w\|_2\leq 1} |m\{f_j(w), g_{j,\alpha}(w)\} - m\{\hat{f}_j(w), \hat{g}_{j,\alpha}(w)\}| > \epsilon \right\} = 0.$$

Lemma A.2 follows from the identities $\widehat{T}^*_{j,\alpha} = \hat{f}^2_{j,\alpha}\hat{g}^{\alpha/(1-\alpha)-1}_{j,\alpha}$ and $T^*_{j,\alpha} = f^2_{j,\alpha}g^{\alpha/(1-\alpha)-1}_{j,\alpha}$. □

Proof of Proposition 2.1. Denote the unit sphere and unit ball in $L^2(\mathbb{T}_X)$ by

$$S = \{w \in L^2(\mathbb{T}_X) : \|w\|_2 = 1\} \quad \text{and} \quad B = \{w \in L^2(\mathbb{T}_X) : \|w\|_2 \leq 1\},$$

respectively. Write

$$W^{\perp}_{j-1,\alpha} = \left\{ w \in L^2(\mathbb{T}_X) : \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{1,\alpha}) = \cdots = \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j-1,\alpha}) = 0 \right\}$$

and

$$\widehat{W}_{j-1,\alpha}^{\perp} = \left\{ w \in L^2(\mathbb{T}_X) : \int_{\mathbb{T}_X} w\mathcal{V}_X(\hat{w}_{1,\alpha}) = \cdots = \int_{\mathbb{T}_X} w\mathcal{V}_X(\hat{w}_{j-1,\alpha}) = 0 \right\}.$$

Clearly, $W_{j-1,\alpha}^{\perp} \cap B$ is weakly sequentially closed and bounded and $T_\alpha(w)$ is weakly sequentially upper semi-continuous when restricted to $W_{j-1,\alpha}^{\perp} \cap B$. According to Lemma A.1, $T_\alpha(w)$ has a maximizer within $W_{j-1,\alpha}^{\perp} \cap B$. This maximizer, say w^*, must locate in $W_{j-1,\alpha}^{\perp} \cap S$, otherwise we can construct $w' = w^*/\|w^*\|_2$ with $T_\alpha(w') = \|w^*\|_2^{2\alpha/(\alpha-1)} T_\alpha(w^*) > T_\alpha(w^*)$. Likewise, $\widehat{T}_\alpha(w)$ has a maximizer in $\widehat{W}_{j-1,\alpha}^{\perp} \cap S$, too. □

Proof of Proposition 2.2. Consider two special cases: when $\alpha = 0$, as stated in Section 2.2.2, we have $\beta \propto w_{1,0}$; for $\alpha = 1/2$, combine Eqs. (3.4) and (3.11) in [28].

For any integer $j \geq 2$ and $\alpha \in (0, 1/2) \cup (1/2, 1)$, let

$$\begin{aligned} f = f(w) &= \operatorname{cov}\left(Y, \int_{\mathbb{T}_X} Xw\right) = \int_{\mathbb{T}_X} w\mathcal{V}_X(\beta), \\ g = g(w) &= \int_{\mathbb{T}_X} w\mathcal{V}_X(w), \\ h = h(w) &= \|w\|_2^2, \end{aligned}$$

and, for all $k \in \{1, \ldots, j-1\}$,

$$e_k = e_k(w) = 2\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{k,\alpha}).$$

Then $T_\alpha = f^2 g^{\alpha/(1-\alpha)-1}$. The Lagrange multiplier rule for Banach spaces, as stated, e.g., in [107, pp. 270–271], ensures that there are real numbers $\delta_1, \ldots, \delta_j$, for each $w \in L^2(\mathbb{T}_X)$,

$$\begin{aligned} f(w_{j,\alpha}) g^{\alpha/(1-\alpha)-2}(w_{j,\alpha}) \left[2g(w_{j,\alpha})\, \mathrm{D}f(w_{j,\alpha})(w) + \{\alpha/(1-\alpha) - 1\} f(w_{j,\alpha})\, \mathrm{D}g(w_{j,\alpha})(w)\right] \\ = \delta_j\, \mathrm{D}h(w_{j,\alpha})(w) + \sum_{k=1}^{j-1} \delta_k\, \mathrm{D}e_k(w_{k,\alpha})(w), \quad \text{(A.3)} \end{aligned}$$

where $\mathrm{D}f(w_{j,\alpha})$, $\mathrm{D}g(w_{j,\alpha})$, $\mathrm{D}h(w_{j,\alpha})$, and $\mathrm{D}e_k(w_{k,\alpha})$, all surjections from $L^2(\mathbb{T}_X)$ to $\mathbb{R}$, are the first-order (Fréchet) derivatives of f, g, h, and e_k evaluated at $w_{k,\alpha}$, respectively; in particular, for $w \in L^2(\mathbb{T}_X)$,

$$\begin{aligned} \mathrm{D}f(w_{j,\alpha})(w) &= \int_{\mathbb{T}_X} w\mathcal{V}_X(\beta), \\ \mathrm{D}g(w_{j,\alpha})(w) &= 2\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j,\alpha}), \\ \mathrm{D}h(w_{j,\alpha})(w) &= 2\int_{\mathbb{T}_X} w w_{j,\alpha}, \end{aligned}$$

and, for all $k \in \{1, \ldots, j-1\}$,

$$\mathrm{D}e_k(w_{k,\alpha})(w) = 2\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{k,\alpha}).$$

Since w at (A.3) is arbitrary, we see that

$$f(w_{j,\alpha})g^{\alpha/(1-\alpha)-2}(w_{j,\alpha})\left[2g(w_{j,\alpha})\mathcal{V}_X(\beta)+\{\alpha/(1-\alpha)-1\}f(w_{j,\alpha})\mathcal{V}_X(w_{j,\alpha})\right] \\ =\delta_j w_{j,\alpha}+\sum_{k=1}^{j-1}\delta_k\mathcal{V}_X(w_{k,\alpha}). \quad \text{(A.4)}$$

Cases with $\{\alpha/(1-\alpha)-1\}f^2(w_{j,\alpha})g^{\alpha/(1-\alpha)-1}(w_{j,\alpha})=0$ and $\gamma_j=0$ are both eliminated: the former one corresponds to the uninteresting minimum of T_α, while the latter one leads to the unconstrained maximizer of T_α which actually never falls on the unit sphere. By Fredholm's theorems (see, e.g., [34, 53]), solve the integral equation (A.4) to get

$$w_{j,\alpha}=U_{j,\alpha}\left(\gamma_j\beta+\sum_{k=1}^{j-1}\gamma_k w_{k,\alpha}\right),$$

where $U_{j,\alpha}:L^2(\mathbb{T}_X)\to L^2(\mathbb{T}_X)$ takes w to $\{(\mathcal{V}_X+\gamma_0 I)^{-1}\circ\mathcal{V}_X\}(w)$, with $\gamma_0=\gamma_0(j,\alpha)\in\mathbb{R}$ and identity operator I, and where $\gamma_1,\ldots,\gamma_j$ accommodate j side-conditions (2.6). It follows that, with $K_{j,\alpha}=U_{j,\alpha}\circ\cdots\circ U_{1,\alpha}$,

$$\mathrm{span}(w_{1,\alpha},\ldots,w_{j,\alpha})=\mathrm{span}\{K_{1,\alpha}(\beta),\ldots,K_{j,\alpha}(\beta)\},$$

because, for each $k\in\{1,\ldots,j\}$, $w_{k,\alpha}$ (resp. $K_{k,\alpha}(\beta)$) belongs to $\mathrm{span}\{K_{1,\alpha}(\beta),\ldots,K_{j,\alpha}(\beta)\}$ (resp. $\mathrm{span}(w_{1,\alpha},\ldots,w_{j,\alpha})$).

Finally, we verify that $\beta\in\overline{\mathrm{span}\{K_{1,\alpha}(\beta),K_{2,\alpha}(\beta),\ldots\}}$. Introduce the orthogonal projection operator P_p that takes $w\in L^2(\mathbb{T}_X)$ to $\sum_{j=1}^p\phi_{j,X}\int_{\mathbb{T}_X}w\phi_{j,X}$. Write $\beta_{p,\mathrm{FPC}}=P_p(\beta)$. Now, the identity

$$\left[\left\{\frac{\lambda_{1,X}}{\lambda_{1,X}+\gamma_0(1,\alpha)}I-(P_p\circ U_{1,\alpha})\right\}\circ\cdots\circ\left\{\frac{\lambda_{p,X}}{\lambda_{p,X}+\gamma_0(p,\alpha)}I-(P_p\circ U_{p,\alpha})\right\}\right](\beta_{p,\mathrm{FPC}})=0$$

implies that

$$\beta_{p,\mathrm{FPC}}\in\mathrm{span}\{(P_p\circ K_{1,\alpha})(\beta_{p,\mathrm{FPC}}),\ldots,(P_p\circ K_{p,\alpha})(\beta_{p,\mathrm{FPC}})\}.$$

In view of $(P_p\circ K_{j,\alpha})(\beta_{p,\mathrm{FPC}})=(P_p\circ K_{j,\alpha})(\beta)$ for all $j\in\{1,\ldots,p\}$, after taking limits in the L^2 sense as $p\to\infty$ on both sides of the following formula

$$\beta_{p,\mathrm{FPC}}\in\left\{P_p(w):w\in\overline{\mathrm{span}\{K_{1,\alpha}(\beta),K_{2,\alpha}(\beta),\ldots\}}\right\},$$

we conclude the proof. □

Proof of Proposition 2.3. For simplicity, we assume no tie among eigenvalues of operator $\mathcal{V}_X$. Then

$$\int_{\mathbb{T}_X}\phi_{j,X}\mathcal{V}_X(\phi_{j'})=\begin{cases}\lambda_{j,X} & \text{if } j=j',\\ 0 & \text{if } j\neq j'.\end{cases}$$

We now prove the proposition by mathematical induction.

For any w ($\neq \phi_{1,X}$) on S with $\int_{\mathbb{T}_X} w\mathcal{V}_X(w) > 0$, there exists $\alpha_0 > 2/3$ such that, for all $\alpha \in (\alpha_0, 1)$,

$$0 < \left\{ \int_{\mathbb{T}_X} w\mathcal{V}_X(w)/\lambda_{1,X} \right\}^{\alpha/(1-\alpha)-1} < \frac{\mathrm{cov}^2\{Y - \mu_Y, \int_{\mathbb{T}_X} X\phi_{1,X}\}}{\mathrm{cov}^2\{Y - \mu_Y, \int_{\mathbb{T}_X} Xw\}},$$

because $0 < \int_{\mathbb{T}_X} w\mathcal{V}_X(w)/\lambda_{1,X} < 1$ and $\mathrm{cov}^2\{Y - \mu_Y, \int_{\mathbb{T}_X} X\phi_{1,X}\} > 0$. It follows that $T_\alpha(\phi_{1,X})/T_\alpha(w) > 1$ for all $\alpha \in (\alpha_0, 1)$ and hence $\|w_{1,\alpha} - \phi_{1,X}\|_2 \to 0$ as $\alpha \to 1$.

Suppose we have $w_{k,\alpha} = \phi_k$ for all $k \in \{1, \ldots, j-1\}$ and $j \geq 2$. For w ($\neq \phi_{p,X}$) satisfying constraints (2.6) and $\int_{\mathbb{T}_X} w\mathcal{V}_X(w) > 0$, along with sufficiently large α, the inequality

$$0 < \left\{ \int_{\mathbb{T}_X} w\mathcal{V}_X(w)/\lambda_{j,X} \right\}^{\alpha/(1-\alpha)-1} < \frac{\mathrm{cov}^2(Y, \int_{\mathbb{T}_X} X)}{\mathrm{cov}^2(Y, \int_{\mathbb{T}_X} Xw)},$$

always holds. Thus, as $\alpha \to 1$, $\phi_{j,X} = \arg\max_w T_{j,\alpha}(w)$ subject to (2.6) and hence $w_{j,\alpha} = \phi_{j,X}$. □

Proof of Proposition 2.4. Define S and $W_{j-1,\alpha}^{\perp}$ as in the proof of Proposition 2.1. Apparently, $T_\alpha(w) = T_{j,\alpha}^*(w)$ for all $w \in W_{j-1,\alpha}^{\perp}$. That is, $w_{j,\alpha}$ is also the solution to

$$\begin{aligned} &\underset{w}{\text{maximize}} \quad T_{j,\alpha}^*(w) \\ &\text{subjectto} \quad \|w\|_2 = 1 \quad \text{and} \quad \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{1,\alpha}) = \cdots = \int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j-1,\alpha}) = 0. \end{aligned} \tag{A.5}$$

For any $w \in S$, construct $w^* \in S$ proportional to

$$w - \sum_{k=1}^{j-1} \frac{\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{k,\alpha})}{\int_{\mathbb{T}_X} w_{k,\alpha}\mathcal{V}_X(w_{k,\alpha})} w_{k,\alpha}.$$

Due to

$$\left\| w - \sum_{k=1}^{j-1} \frac{\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{k,\alpha})}{\int_{\mathbb{T}_X} w_{k,\alpha}\mathcal{V}_X(w_{k,\alpha})} w_{k,\alpha} \right\|_2 \leq 1$$

and $\alpha/(\alpha-1) < 0$ (excluding the trivial case $\alpha = 0$), it is easy to verify that $w^* \in W_{j-1,\alpha}^{\perp}$ and

$$T_{j,\alpha}^*(w^*) = T_{j,\alpha}^*(w) \left\| w - \sum_{j=1}^{p-1} \frac{\int_{\mathbb{T}_X} w\mathcal{V}_X(w_{j,\alpha})}{\int_{\mathbb{T}_X} w_{j,\alpha}\mathcal{V}_X(w_{j,\alpha})} w_{j,\alpha} \right\|_2^{2\alpha/(\alpha-1)} \geq T_{j,\alpha}^*(w).$$

This inequality becomes an equality only when $w \in W_{j-1,\alpha}^{\perp}$; in other words, it suffices to drop side-conditions (A.5) when maximizing $T_{j,\alpha}^*(w)$ subject to $\|w\|_2 = 1$. □

Proof of Proposition 2.5. Replace population values in the proof of Proposition 2.4 with their empirical counterparts. □

Proof of Proposition 2.6. Let $h = h(w) = \|w\|_2^2$, $g_{j,\alpha} = g_{j,\alpha}(w) = \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(w)$, and

$$f_{j,\alpha} = f_{j,\alpha}(w) = \operatorname{cov}\left\{Y^{(j,\alpha)}, \int_{\mathbb{T}_X} X^{(j,\alpha)} w\right\} = \int_{\mathbb{T}_X} w\mathcal{V}_{X^{(j,\alpha)}}(\beta).$$

Then $T_{j,\alpha} = f_{j,\alpha}^2 g_{j,\alpha}^{\alpha/(1-\alpha)-1}$ and $w_{j,\alpha}$ (2.5) must be a solution to the constrained optimization problem

$$\begin{aligned} &\underset{w}{\text{maximize}} \quad f_{j,\alpha}^2(w) \\ &\text{subject to} \quad g_{j,\alpha}(w) = g_0 \quad \text{and} \quad h(w) = 1 \end{aligned}$$

for certain $g_0 \in (0, \lambda_{1,X^{(j,\alpha)}}]$, where $\lambda_{k,X^{(j,\alpha)}}$ is the kth largest eigenvalue of operator $\mathcal{V}_{X^{(j,\alpha)}}$ with corresponding eigenfunction $\phi_{k,X^{(j,\alpha)}}$.

Check the case with $g_0 = \lambda_{1,X^{(j,\alpha)}} > 0$ (i.e., the functional principal component basis). Provided that $\lambda_{1,X^{(j,\alpha)}}$ has multiplicity $= m \geq 1$, we can write

$$w_{j,\alpha} = a_1\phi_{1,X^{(j,\alpha)}} + \cdots + a_m\phi_{m,X^{(j,\alpha)}},$$

where $a_1, \ldots, a_m \in [-1, 1]$ and $a_1^2 + \cdots + a_m^2 = 1$. The Cauchy–Schwarz inequality implies that the maximum of

$$f_{j,\alpha}^2(w) = \left\{\sum_{k=1}^m a_k \int_{\mathbb{T}_X} \phi_{k,X^{(j,\alpha)}} \mathcal{V}_{X^{(j,\alpha)}}(\beta)\right\}^2 = \left(\sum_{k=1}^m a_j \lambda_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta\phi_{k,X^{(j,\alpha)}}\right)^2$$

is achieved if and only if m-vector

$$(a_1, \ldots, a_m) \propto \left(\lambda_{1,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta\phi_{1,X^{(j,\alpha)}}, \ldots, \lambda_{m,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta\phi_{m,X^{(j,\alpha)}}\right).$$

Therefore, as $\delta^{(j,\alpha)} \to -1$,

$$w_{j,\alpha} \propto \sum_{k=1}^{\infty} \frac{\lambda_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta\phi_{k,X^{(j,\alpha)}}}{\lambda_{k,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}} \phi_{k,X^{(j,\alpha)}}.$$

Unless $g_0 = \lambda_{1,X^{(j,\alpha)}} > 0$, apply the Lagrange multiplier rule for Banach spaces as in the proof of Proposition 2.2 and arrive at identity

$$f_{j,\alpha}(w_{j,\alpha})\mathcal{V}_{X^{(j,\alpha)}}(\beta) = \delta_1\mathcal{V}_{X^{(j,\alpha)}}(w_{j,\alpha}) + \delta_2 w_{j,\alpha},$$

with $\delta_1, \delta_2 \in \mathbb{R}$, where δ_2 must be nonzero as the the maximizer of $T_{j,\alpha}^*$ never falls on the unit sphere. Also, we rule out the case of $f_{j,\alpha}(w_{j,\alpha}) = 0$ corresponding to the uninteresting minimum of $T_{j,\alpha}^*$.

If $\delta_1 = 0$, the functional continuum basis reduces to functional PLS basis and $w_{j,\alpha} \propto \mathcal{V}_{X^{(j,\alpha)}}(\beta)$. When $\delta^{(j,\alpha)}$ is close enough to 0, $\lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}$ becomes dominant over $\lambda_{k,X^{(j,\alpha)}}$ for all $k \in \mathbb{Z}^+$, i.e., $\lambda_{k,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}$ and $\lambda_{k',X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}$ approach each

other for $k \neq k'$. Accordingly, as $\delta^{(j,\alpha)} \to 0$,

$$w_{j,\alpha} \propto \sum_{k=1}^{\infty} \lambda_{k,X^{(j,\alpha)}} \phi_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta \phi_{k,X^{(j,\alpha)}} \propto \sum_{k=1}^{\infty} \frac{\lambda_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta \phi_{k,X^{(j,\alpha)}}}{\lambda_{k,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}} \phi_{k,X^{(j,\alpha)}}.$$

In the case with nonzero δ_1, solving the following inhomogeneous Fredholm integral equation w.r.t. $w_{j,\alpha}$,

$$f_{j,\alpha}(w_{j,\alpha}) \mathcal{V}_{X^{(j,\alpha)}}(\beta)/\delta_1 = \delta_2 w_{j,\alpha}/\delta_1 + \mathcal{V}_{X^{(j,\alpha)}}(w_{j,\alpha}),$$

we also obtain the solution

$$w_{j,\alpha} \propto \left[\left\{ \mathcal{V}_{X^{(j,\alpha)}} + \frac{\lambda_{1,X^{(j,\alpha)}}}{\delta^{(j,\alpha)}} I \right\}^{-1} \circ \mathcal{V}_{X^{(j,\alpha)}} \right] (\beta) = \sum_{k=1}^{\infty} \frac{\lambda_{k,X^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta \phi_{k,X^{(j,\alpha)}}}{\lambda_{k,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}} \phi_{k,X^{(j,\alpha)}},$$

where $\delta^{(j,\alpha)} = \delta_1 \lambda_{1,X^{(j,\alpha)}}/\delta_2$. The existence and uniqueness of this solution is guaranteed by Fredholm's theorems [53] which hold here because $v_{X^{(j,\alpha)}} \in L^2(\mathbb{T}_X^2)$.

The last phase of this proof is to ascertain that $\delta^{(j,\alpha)} \notin (-\infty, -1)$. Without loss of generality, assume that $\int_{\mathbb{T}_X} \phi_{k,X^{(j,\alpha)}} \mathcal{V}_{X^{(j,\alpha)}}(\beta) \geq 0$ for all k, otherwise we can use $-\phi_{k,X^{(j,\alpha)}}$ instead. The identity $\sum_{k=1}^{\infty} \lambda_{k,X^{(j,\alpha)}} < \infty$ further indicates that, if $\delta^{(j,\alpha)} \in (-\infty, -1)$, then there must exist k_0 such that $\lambda_{k_0,X^{(j,\alpha)}} + \lambda_{1,X^{(j,\alpha)}}/\delta^{(j,\alpha)}$ is negative. Under this circumstance, changing the sign of it will increase $f_{j,\alpha}^2(w_{j,\alpha})$ without altering $g_{j,\alpha}(w_{j,\alpha})$ or violating the unit norm constraint. This contradicts the definition of $w_{j,\alpha}$ and hence completes the proof. □

Proof of Theorem 2.1. We resort to an argument similar to the proof adopted by [5, Theorem 4.1.1] and extend it from the finite-dimensional setting to the functional context. The unit ball B is as defined in the proof of Proposition 2.1. Start with $j = 1$. Let $N_{1,\delta}$ be a neighborhood in $L^2(\mathbb{T}_X)$ containing $w_{1,\alpha}$, namely, for $\delta \in (0, 2)$,

$$N_{1,\delta} = \{w \in L^2(\mathbb{T}_X) : \|w - w_{1,\alpha}\|_2 < \delta\}.$$

Verify that $B \setminus N_{1,\delta}$ is weakly sequentially closed and bounded and $T_{1,\alpha}^*(w)$ is weakly sequentially upper semi-continuous within $B \setminus N_{1,\delta}$. Then Lemma A.1 guarantees the existence of $\max_{w \in B \setminus N_{1,\delta}} T_{1,\alpha}^*(w)$.

Write

$$\epsilon = T_{1,\alpha}^*(w_{j,\alpha}) - \max_{w \in B \setminus N_{1,\delta}} T_{1,\alpha}^*(w) > 0$$

and observe that

$$\begin{aligned}
&\left\{ \sup_{w:\|w\|_2=1} |\widehat{T}_{1,\alpha}^*(w) - T_{1,\alpha}^*(w)| < \frac{\epsilon}{2} \right\} \\
&\subseteq \left\{ T_{1,\alpha}^*(\hat{w}_{p,\alpha}) > \widehat{T}_{1,\alpha}^*(\hat{w}_{1,\alpha}) - \frac{\epsilon}{2} \right\} \cup \left\{ \widehat{T}_{1,\alpha}^*(w_{p,\alpha}) > T_{1,\alpha}^*(w_{1,\alpha}) - \frac{\epsilon}{2} \right\} \\
&\subseteq \left\{ T_{1,\alpha}^*(\hat{w}_{1,\alpha}) > \widehat{T}_{1,\alpha}^*(w_{1,\alpha}) - \frac{\epsilon}{2} \right\} \cup \left\{ \widehat{T}_{1,\alpha}^*(w_{1,\alpha}) > T_{1,\alpha}^*(w_{1,\alpha}) - \frac{\epsilon}{2} \right\} \\
&\subseteq \left\{ T_{1,\alpha}^*(\hat{w}_{1,\alpha}) > T_{1,\alpha}^*(w_{1,\alpha}) - \epsilon \right\}
\end{aligned}$$

$$\subseteq \{\hat{w}_{1,\alpha} \in N_{1,\delta}\}.$$

By Lemma A.2, $\lim_{n\to\infty} \Pr(\hat{w}_{1,\alpha} \in N_{1,\delta}) = 1$. Considering the arbitrariness of δ, we conclude that $\|\hat{w}_{1,\alpha} - w_{1,\alpha}\|_2$ converges to zero in probability as n diverges. If the convergence of $\hat{w}_{1,\alpha}, \ldots, \hat{w}_{j-1,\alpha}$ holds, the prerequisite of Lemma A.2 is fulfilled. Mimicking the argument for $j = 1$, we deduce the convergence to zero in probability of $\|\hat{w}_{j,\alpha} - w_{j,\alpha}\|_2$ as $n \to \infty$ for arbitrarily given j.

As for $\hat{\beta}_{p,\alpha}$ in (2.11) and $\hat{\eta}_{p,\alpha}(X_0)$ in (2.12), their convergence can be proved after we combine identities (A.1) and (A.2) with the convergence of $\hat{w}_{j,\alpha}$ and employ the continuous mapping theorem for convergence in probability. □

Proof of Proposition 2.7. Follow the same argument as in the proof for Proposition 2.6 but substitute empirical items for the population counterparts. Meanwhile, take the following identity into consideration:

$$\begin{aligned}
\hat{\lambda}_{k,\widehat{X}^{(j,\alpha)}} \int_{\mathbb{T}_X} \beta \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}} &= \int_{\mathbb{T}_X} \beta \widehat{\mathcal{V}}_{\widehat{X}^{(j,\alpha)}} \left(\hat{\phi}_{k,\widehat{X}^{(j,\alpha)}} \right) \\
&= \widehat{\mathrm{cov}} \left\{ \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} \beta, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}} \right\} \\
&= \widehat{\mathrm{cov}} \left\{ \widehat{Y}^{(j,\alpha)}, \int_{\mathbb{T}_X} \widehat{X}^{(j,\alpha)} \hat{\phi}_{k,\widehat{X}^{(j,\alpha)}} \right\}.
\end{aligned}$$

This completes the argument. □

A.2 Technical details for Chapter 3

We use the following conditions in the theoretical part of Chapter 3.

(C.A.2.1) The true relationship between $Y = \mathbb{1}(X \in \Pi_1)$ and X is linear, i.e., there is $\beta \in L^2(\mathbb{T}_X)$ such that SoFR (1.3) holds.

(C.A.2.2) $\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(\mu_X^{[1]} - \mu_X^{[0]})\}^2 / \operatorname{var}\{\int_{\mathbb{T}_X} \beta_{p,\alpha}(X - \mu_X^{[k]}) \mid X \in \Pi_k\}$ diverges as $p \to \infty$ for each α and k.

(C.A.2.3) Realizations of X are twice continuously differentiable and $\|X'\|_2$ is bounded almost surely.

(C.A.2.4) $\tau_1, \ldots, \tau_N$ are eigenvalues of $\mathbf{Pen}^{-1/2}\mathbf{WPen}^{-1/2}$ such that $\tau_1 = \tau_2 = 0$, $\tau_3 \geq \cdots \geq \tau_N$, and $C_1(l-2)^{-4} \leq \tau_l \leq C_2(l-2)^{-4}$ for $l \geq 3$, with neither C_1 nor C_2 depending on l or N.

(C.A.2.5) $M \to \infty$ and $M^{-1}\max(\theta^*, \theta_1, \ldots, \theta_n) \to 0$ as $n \to \infty$, where $\theta^* > 0$ is the smoothing parameter involved in recovering X^*.

(C.A.2.6) For all j ($\leq p \leq \operatorname{rank}(\widehat{\mathbf{C}}_c\mathbf{W}^{1/2})$), $T_{j,\alpha}(w)$ has a unique maximizer (up to sign) in $\{w \in L^2(\mathbb{T}_X) : \|w\|_2 = 1\}$.

Under (C.A.2.1) with $\pi_0 = 1/2$, CCC-L and PLCC are equivalent to each other, because $\beta_{p,\mathrm{FPLS}}$ (3.4) and $\beta_{p,\alpha}$ (2.8) share the same limit as $p \to \infty$. It is not necessary to hold (C.A.2.1) for the (asymptotically) perfect classification for CCC-L. Condition (C.A.2.2) implies that, after projected to the direction of $\beta_{p,\alpha}$, as p diverges, the within-group covariance becomes more and more ignorable when compared with the between-group one, i.e., the two groups become more and more separable. It is analogous to assumption (4.4)(d) in [27] and assures us of the (asymptotic) perfect classification of CCC-L. Assumptions (C.A.2.3) and (C.A.2.4) jointly guarantee that the smoothed curves converge to the true ones as observations become denser and denser; although the latter one has been proved by [95, Eq. 4] for natural splines, we have little knowledge on whether it still holds for B-splines and hence have to assume it following [23, Eq. A4.3.1]. If we have extra regularity conditions (C.A.2.5) and (C.A.2.6) (identical to (C.2.1)), the proposed empirical implementation in Section 3.2.1 turns out to be consistent in probability.

Proof of Proposition 3.1. Write $\gamma_{p,\alpha} = \int_{\mathbb{T}_X} \beta_{p,\alpha}(\mu_X^{[1]} - \mu_X^{[0]})$ and $R_{p,\alpha}^{[k]} = \int_{\mathbb{T}_X} \beta_{p,\alpha}(X^* - \mu_X^{[k]})$. Recalling (3.6), $\sigma^2_{[k]}(\beta_{p,\alpha}) = \mathrm{var}(R_{p,\alpha}^{[k]} \mid X \in \Pi_k)$, $k = 0, 1$. Thus,

$$\begin{aligned}
&\Pr\{\mathcal{D}_L(X^* \mid \beta_{p,\alpha}) < 0 \mid X^* \in \Pi_0\} \\
&\quad = \Pr\left\{ (R_{p,\alpha}^{[0]} - \gamma_{p,\alpha})^2 - (R_{p,\alpha}^{[0]})^2 < 2\sigma^2_{[0]}(\beta_{p,\alpha}) \ln \frac{1-\pi_0}{\pi_0} \,\middle|\, X^* \in \Pi_0 \right\} \\
&\quad = \Pr\left[\frac{R_{p,\alpha}^{[0]}}{\sigma_{[0]}(\beta_{p,\alpha})} > \frac{\gamma^2_{p,\alpha} + 2\sigma^2_{[0]}(\beta_{p,\alpha}) \ln\{\pi_0/(1-\pi_0)\}}{2\gamma\sigma_{[0]}(\beta_{p,\alpha})} \,\middle|\, X^* \in \Pi_0 \right] \\
&\quad \le \frac{4\sigma^2_{[0]}(\beta_{p,\alpha})/\gamma^2_{p,\alpha}}{\left[1 + 2\gamma^{-2}_{p,\alpha}\sigma^2_{[0]}(\beta_{p,\alpha}) \ln\{\pi_0/(1-\pi_0)\}\right]^2},
\end{aligned}$$

where the upper bound is derived from Chebyshev's inequality and the identity that random variable $R_{p,\alpha}^{[0]}/\sigma_{[0]}(\beta_{p,\alpha})$ (conditional on the event $X^* \in \Pi_0$) is of zero mean and unit variance. Similarly, we deduce that

$$\Pr\{\mathcal{D}_L(X^* \mid \beta_{p,\alpha}) > 0 \mid X^* \in \Pi_1\} \le \frac{4\sigma^2_{[1]}(\beta_{p,\alpha})/\gamma^2_{p,\alpha}}{\left[1 + 2\gamma^{-2}_{p,\alpha}\sigma^2_{[0]}(\beta_{p,\alpha}) \ln\{(1-\pi_0)/\pi_0\}\right]^2}.$$

Eventually, as p diverges, the zero-convergence of

$$\begin{aligned}
&\mathrm{err}\{\mathcal{D}_L(X^* \mid \beta_{p,\alpha})\} \\
&\quad = \pi_0 \Pr\{\mathcal{D}_L(X^* \mid \beta_{p,\alpha}) < 0 \mid X^* \in \Pi_0\} + (1-\pi_0) \Pr\{\mathcal{D}_L(X^* \mid \beta_{p,\alpha}) > 0 \mid X^* \in \Pi_1\}
\end{aligned}$$

results from (C.A.2.2) (i.e., $\sigma^2_{[k]}(\beta_{p,\alpha})/\gamma^2_{p,\alpha} \to 0$ as p diverges for each α and k). □

Proof of Proposition 3.2. Recall $\Delta t = |\mathbb{T}_X|/M$ and matrices $\boldsymbol{\psi}$ (3.10), $\hat{\boldsymbol{c}}_i$ (3.11), $\boldsymbol{\Psi}$ (3.12), $\mathbf{W}$ (3.14) and $\mathbf{Pen}$ (3.13), all defined in Section 3.2.1. Introduce operator $\mathcal{P}_{BS_N}$ such that, for each $f \in L^2(\mathbb{T}_X)$, $\mathcal{P}_{BS_N} f$ is the orthogonal projection of f onto BS_N (3.8), i.e.,

$$\mathcal{P}_{BS_N} f = \left[\int_{\mathbb{T}_X} f\psi_1, \ldots, \int_{\mathbb{T}_X} f\psi_N \right] \mathbf{W}^{-1} \boldsymbol{\psi}.$$

For each i, specifically, $\mathcal{P}_{BS_N} X_i = c_i^\top \boldsymbol{\psi}$ with

$$c_i = \mathbf{W}^{-1} \left[\int_{\mathbb{T}_X} X_i \psi_1, \ldots, \int_{\mathbb{T}_X} X_i \psi_N \right]^\top.$$

We chop $\|\widehat{X}_i - X_i\|_2$ into two segments: $\|\mathcal{P}_{BS_N} X_i - X_i\|_2$ and $\|\widehat{X}_i - \mathcal{P}_{BS_N} X_i\|_2$. Combined with [63, Theorem 16], condition (C.A.2.3) implies

$$\|\mathcal{P}_{BS_N} X_i - X_i\|_2 = O_p(\Delta t) = O_p(M^{-1}) \quad \text{as} \quad M \to \infty.$$

Further, condition (C.A.2.3) allows us to follow [19, Theorem 5] to verify that, as $M \to \infty$, $\|\mathbf{W} - \Delta t \boldsymbol{\Psi}^\top \boldsymbol{\Psi}\|_F^2 = O(M^{-2})$, $\|\mathbf{W} - \Delta t \boldsymbol{\Psi}^\top \boldsymbol{\Psi} - \Delta t \theta_i \mathbf{Pen}\|_F^2 = O(1)$, and

$$\left\| \left[\int_{\mathbb{T}_X} X_i \psi_1, \ldots, \int_{\mathbb{T}_X} X_i \psi_N \right]^\top - \Delta t \boldsymbol{X}_i^\top \boldsymbol{\Psi} \right\|_F^2 = O_p(M^{-3}),$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Let $\tau_1, \ldots, \tau_N$ be eigenvalues of

$$\mathbf{Z} = \mathbf{Pen}^{-1/2} \mathbf{W} \mathbf{Pen}^{-1/2}$$

with corresponding eigenvectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N$. Noting that $\lim_{N\to\infty} \max_{l_1,l_2} |\int_{\mathbb{T}_X} \psi_{l_1} \psi_{l_2}|$ and $\lim_{N\to\infty} \max_{l_1,l_2} |\int_{\mathbb{T}_X} \psi''_{l_1} \psi''_{l_2}|$ are both finite, the squared second trunk

$$\begin{aligned}
&\|\widehat{X}_i - \mathcal{P}_{BS_N} X_i\|_2^2 \\
&= (\hat{\boldsymbol{c}}_i^\top - \boldsymbol{c}_i^\top) \mathbf{W} (\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i) \\
&= \left\{ \Delta t \boldsymbol{X}_i^\top \boldsymbol{\Psi} (\Delta t \boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \Delta t \theta_i \mathbf{Pen})^{-1} - \boldsymbol{c}_i^\top \right\} \mathbf{W} \left\{ (\Delta t \boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \Delta t \theta_i \mathbf{Pen})^{-1} \Delta t \boldsymbol{\Psi}^\top \boldsymbol{X}_i - \boldsymbol{c}_i \right\} \\
&= \boldsymbol{c}_i^\top \mathbf{Pen}^{1/2} \left\{ \mathbf{Z} (\mathbf{Z} + \Delta t \theta_i \mathbf{I}_N)^{-1} - \mathbf{I}_N \right\} \mathbf{Z} \left\{ (\mathbf{Z} + \Delta t \theta_i \mathbf{I}_N)^{-1} \mathbf{Z} - \mathbf{I}_N \right\} \mathbf{Pen}^{1/2} \boldsymbol{c}_i + o_p(1) \\
&= (\Delta t)^2 \theta_i^2 \boldsymbol{c}_i^\top \mathbf{Pen}^{1/2} (\mathbf{Z} + \Delta t \theta_i \mathbf{I}_N)^{-1} \mathbf{Z} (\mathbf{Z} + \Delta t \theta_i \mathbf{I}_N)^{-1} \mathbf{Pen}^{1/2} \boldsymbol{c}_i + o_p(1) \\
&= \sum_{l=1}^{N} \frac{\tau_l}{(\theta_i^{-1} \Delta t^{-1} \tau_l + 1)^2} (\boldsymbol{c}_i^\top \mathbf{Pen}^{1/2} \boldsymbol{e}_l)^2 + o_p(1) \\
&\le \Delta t \theta_i \sum_{l=1}^{N_1 - 1} \tau_l (\boldsymbol{c}_i^\top \mathbf{Pen}^{1/2} \boldsymbol{e}_l)^2 + \sum_{l=N_1}^{N} \tau_l (\boldsymbol{c}_i^\top \mathbf{Pen}^{1/2} \boldsymbol{e}_l)^2 + o_p(1) \\
&= o_p(1) \quad \text{as} \quad M \to \infty,
\end{aligned}$$

where $N_1 \in \mathbb{Z}^+$ is so defined that $\tau_{N_1} = \max\{\tau_N, (\Delta t \theta_i)^{1/2}\}$ and diverges as $M \to \infty$ owing to (C.A.2.4). □

Proof of Proposition 3.3. Recall $\mathcal{P}_{BS_N}$ defined in the proof of Proposition 3.2. Writing $\tilde{w}_{j,\alpha} = \mathcal{P}_{BS_N} \tilde{w}_{j,\alpha} + (\mathcal{I} - \mathcal{P}_{BS_N}) \tilde{w}_{j,\alpha}$, with identity operator $\mathcal{I}$, one has $0 < \|\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}\|_2 \le 1$ and $\int_{\mathbb{T}_X} \widehat{X}_i \mathcal{P}_{BS_N} \tilde{w}_{j,\alpha} = \int_{\mathbb{T}_X} \widehat{X}_i \tilde{w}_{j,\alpha}$ since $\widehat{X}_i \in BS_N$ for all i. If $0 < \|\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}\|_2 < 1$ (i.e., $(\mathcal{I} - \mathcal{P}_{BS_N}) \tilde{w}_{j,\alpha} > 0$), then $\|\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}\|_2^{-1} \mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}$ satisfies that

$$\widetilde{T}_{j,\alpha}^* \left(\frac{\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}}{\|\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}\|_2} \right) = \|\mathcal{P}_{BS_N} \tilde{w}_{j,\alpha}\|_2^{2\alpha/(\alpha-1)} \widetilde{T}_{j,\alpha}^*(\tilde{w}_{j,\alpha}) > \widetilde{T}_{j,\alpha}^*(\tilde{w}_{j,\alpha}),$$

which violates the definition of $\tilde{w}_{j,\alpha}$. and reaches a contradiction. This contradiction implies that $(\mathcal{I} - \mathcal{P}_{BS_N})\tilde{w}_{j,\alpha}$ must be 0. □

Proof of Proposition 3.4. Under conditions (C.A.2.3)–(C.A.2.6), fixing p, Proposition 3.2 and Theorem 2.1 assure us of the zero-convergence (in probability) of $\|\tilde{\beta}_{p,\alpha} - \beta_{p,\alpha}\|_2$ as n diverges. As Proposition 3.2 applies to X^*, the convergence of empirical classifiers follows. □

A.3 Technical details for Chapter 4

A.3.1 A glance at the local linear smoother

Let $\kappa = \kappa(\cdot)$ be a function on $\mathbb{R}$ satisfying (C.A.3.8)–(C.A.3.10) in A.3.2; examples include the symmetric Beta family [30, Eq. 2.5] that takes the Epanechnikov kernel $\kappa(t) = .75(1 - t^2)\mathbb{1}(|t| \leq 1)$ as a special case. LLS actually falls into the framework of weighted least squares (WLS, [30, pp. 58–59]). Given integers M and m (with values specified in the following cases (A.3.i)–(A.3.iv)), matrices $\mathbf{1}_M$ (viz. the M-vector of ones), $\boldsymbol{u}$ (viz. an M-vector), $\mathbf{T}$ (viz. an $M \times m$ matrix) and $\mathbf{W}$ (viz. an $M \times M$ non-negative definite matrix), one solves

$$\min_{a_0, \boldsymbol{a}} (\boldsymbol{u} - a_0\mathbf{1}_M - \mathbf{T}\boldsymbol{a})^\top \mathbf{W} (\boldsymbol{u} - a_0\mathbf{1}_M - \mathbf{T}\boldsymbol{a})$$

for $a_0 \in \mathbb{R}$ and m-vector $\boldsymbol{a} = [a_1, \ldots, a_m]^\top$. The actual estimate uses only the WLS solution for a_0 given by

$$\begin{aligned} \hat{a}_0 &= (\mathbf{1}_M^\top \mathbf{W}^{1/2} \mathbf{P}^\perp_{\mathbf{W}^{1/2}\mathbf{T}} \mathbf{W}^{1/2} \mathbf{1}_M)^+ \mathbf{1}_M^\top \mathbf{W}^{1/2} \mathbf{P}^\perp_{\mathbf{W}^{1/2}\mathbf{T}} \mathbf{W}^{1/2} \boldsymbol{u} \\ &= [\mathbf{1}_M^\top \{\mathbf{W} - \mathbf{W}\mathbf{T}(\mathbf{T}^\top \mathbf{W}\mathbf{T})^+ \mathbf{T}^\top \mathbf{W}\} \mathbf{1}_M]^+ \mathbf{1}_M^\top \{\mathbf{W} - \mathbf{W}\mathbf{T}(\mathbf{T}^\top \mathbf{W}\mathbf{T})^+ \mathbf{T}^\top \mathbf{W}\} \boldsymbol{u} \end{aligned} \tag{A.6}$$

in which the superscript "+" denotes the Moore-Penrose generalized inverse and

$$\mathbf{P}^\perp_{\mathbf{W}^{1/2}\mathbf{T}} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{T}(\mathbf{T}^\top \mathbf{W}\mathbf{T})^+ \mathbf{T}^\top \mathbf{W}^{1/2}.$$

In particular, four different combinations of $\boldsymbol{u}$, $\mathbf{T}$ and $\mathbf{W}$ yield estimates of the four targets of interest, respectively:

(A.3.i) Given $t \in \mathbb{T}_X$, estimate $\mu_X(t)$ by $\hat{\mu}_X(t) = \hat{a}_0$ (A.6) with $\sum_{1 \leq i \leq n} L_i$-vectors

$$\boldsymbol{u} = \left[\widetilde{X}_1(T_{11}), \ldots, \widetilde{X}_1(T_{1L_1}), \ldots, \widetilde{X}_n(T_{n1}), \ldots, \widetilde{X}_n(T_{nL_n})\right]^\top$$

and

$$\mathbf{T} = [t - T_{11}, \ldots, t - T_{1L_1}, \ldots, t - T_{n1}, \ldots, t - T_{nL_n}]^\top$$

and $\sum_{1 \leq i \leq n} L_i \times \sum_{1 \leq i \leq n} L_i$ matrix

$$\mathbf{W} = \operatorname{diag}\left\{\kappa\left(\frac{t - T_{11}}{h_{\mu_X}}\right), \ldots, \kappa\left(\frac{t - T_{1L_1}}{h_{\mu_X}}\right), \ldots, \kappa\left(\frac{t - T_{n1}}{h_{\mu_X}}\right), \ldots, \kappa\left(\frac{t - T_{nL_n}}{h_{\mu_X}}\right)\right\}.$$

(A.3.ii) Write $\bar{Y} = n^{-1}\sum_{1\le i\le n} Y_i$. For arbitrary $t \in \mathbb{T}$, $\hat{v}_C(t) = \hat{a}_0 - \bar{Y}\cdot\hat{\mu}_X(t)$, where $\hat{a}_0$ follows (A.6) with $\sum_{1\le i\le n} L_i$-vectors

$$\begin{aligned}\boldsymbol{u} = \Big[&\{\widetilde{X}_1(T_{11}) - \hat{\mu}_X(T_{11})\}(Y_1 - \bar{Y}), \ldots, \{\widetilde{X}_1(T_{1L_1}) - \hat{\mu}_X(T_{1L_1})\}(Y_1 - \bar{Y}), \ldots, \\ &\{\widetilde{X}_n(T_{n1}) - \hat{\mu}_X(T_{n1})\}(Y_n - \bar{Y}), \ldots, \{\widetilde{X}_n(T_{nL_n}) - \hat{\mu}_X(T_{nL_n})\}(Y_n - \bar{Y})\Big]^\top\end{aligned}$$

and $\mathbf{T} = [t - T_{11}, \ldots, t - T_{1L_1}, \ldots, t - T_{n1}, \ldots, t - T_{nL_n}]^\top$ as well as $\sum_{1\le i\le n} L_i \times \sum_{1\le i\le n} L_i$ matrix

$$\mathbf{W} = \mathrm{diag}\left\{\kappa\left(\frac{t - T_{11}}{h_{v_{XY}}}\right), \ldots, \kappa\left(\frac{t - T_{1L_1}}{h_{v_{XY}}}\right), \ldots, \kappa\left(\frac{t - T_{n1}}{h_{v_{XY}}}\right), \ldots, \kappa\left(\frac{t - T_{nL_n}}{h_{v_{XY}}}\right)\right\}.$$

(A.3.iii) Fix $s, t \in \mathbb{T}_X$. Then $\hat{v}_A(s,t) = \hat{a}_0 - \hat{\mu}_X(s)\hat{\mu}_X(t)$, where $\hat{a}_0$ is fitted as (A.6) with $\sum_{1\le i\le n} L_i(L_i - 1)$-vector

$$\begin{aligned}\boldsymbol{u} = \Big[&\ldots, \widetilde{X}_i(T_{i\ell})\widetilde{X}_i(T_{i1}), \ldots, \widetilde{X}_i(T_{i\ell})\widetilde{X}_i(T_{i,\ell-1}), \\ &\widetilde{X}_i(T_{i\ell})\widetilde{X}_i(T_{i,\ell+1}), \ldots, \widetilde{X}_i(T_{i\ell})\widetilde{X}_i(T_{iL_i})\ldots\Big]^\top,\end{aligned}$$

$\sum_{1\le i\le n} L_i(L_i - 1) \times 2$ matrix

$$\mathbf{T} = \begin{bmatrix} \cdots & s - T_{i\ell} & \cdots & s - T_{i\ell} & s - T_{i\ell} & \cdots & s - T_{i\ell} & \cdots \\ \cdots & t - T_{i1} & \cdots & t - T_{i,\ell-1} & t - T_{i,\ell+1} & \cdots & t - T_{iL_i} & \cdots \end{bmatrix}^\top$$

and $\sum_{1\le i\le n} L_i(L_i - 1) \times \sum_{1\le i\le n} L_i(L_i - 1)$ matrix

$$\begin{aligned}\mathbf{W} = \mathrm{diag}\Bigg\{&\ldots, \kappa\left(\frac{s - T_{i\ell}}{h_{v_X}}\right)\kappa\left(\frac{t - T_{i1}}{h_{v_X}}\right), \ldots, \kappa\left(\frac{s - T_{i\ell}}{h_{v_X}}\right)\kappa\left(\frac{t - T_{i,\ell-1}}{h_{v_X}}\right), \\ &\kappa\left(\frac{s - T_{i\ell}}{h_{v_X}}\right)\kappa\left(\frac{t - T_{i,\ell+1}}{h_{v_X}}\right), \ldots, \kappa\left(\frac{s - T_{i\ell}}{h_{v_X}}\right)\kappa\left(\frac{t - T_{iL_i}}{h_{v_X}}\right), \ldots\Bigg\}.\end{aligned}$$

(A.3.iv) Rotate the two tuple $(T_{i\ell_1}, T_{i\ell_2})$ to become

$$\begin{bmatrix} T^{\#}_{i\ell_1} \\ T^{\#}_{i\ell_2} \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}\begin{bmatrix} T_{i\ell_1} \\ T_{i\ell_2} \end{bmatrix}.$$

For arbitrarily fixed $t \in \mathbb{T}_X$, $\tilde{v}(t) = \hat{a}_0 - \hat{\mu}^2_X(t)$, where $\hat{a}_0$ follows (A.6) with $\sum_{1\le i\le n} L_i$-vector

$$\boldsymbol{u} = \left[\widetilde{X}^2_1(T^{\#}_{11})\ldots\widetilde{X}^2_1(T^{\#}_{1L_1})\ldots\widetilde{X}^2_n(T^{\#}_{n1})\ldots\widetilde{X}^2_n(T^{\#}_{nL_n})\right]^\top,$$

$\sum_{1\le i\le n} L_i \times 2$ matrix

$$\mathbf{T} = \begin{bmatrix} -T_{11}^{\#} & \cdots & -T_{1L_i}^{\#} & \cdots & -T_{n1}^{\#} & \cdots & -T_{nL_i}^{\#} \\ t/\sqrt{2} - T_{11}^{\#} & \cdots & t/\sqrt{2} - T_{1L_1}^{\#} & \cdots & t/\sqrt{2} - T_{n1}^{\#} & \cdots & t/\sqrt{2} - T_{nL_n}^{\#} \end{bmatrix}^\top$$

and $\sum_{1\le i\le n} L_i \times \sum_{1\le i\le n} L_i$ matrix

$$\mathbf{W} = \operatorname{diag}\left\{ \kappa\left(\frac{t/\sqrt{2} - T_{11}^{\#}}{h_{\sigma_e}}\right), \ldots, \kappa\left(\frac{t/\sqrt{2} - T_{1L_1}^{\#}}{h_{\sigma_e}}\right), \ldots, \right.$$
$$\left. \kappa\left(\frac{t/\sqrt{2} - T_{n1}^{\#}}{h_{\sigma_e}}\right), \ldots, \kappa\left(\frac{t/\sqrt{2} - T_{nL_n}^{\#}}{h_{\sigma_e}}\right) \right\}.$$

Then, as suggested in [104, 105], σ_e^2 is estimated by averaging $\tilde{v}(t) - \hat{v}_X(t,t)$ over a truncated subinterval of $\mathbb{T}_X = [0,1]$, say $\mathbb{T}_X^* = [1/4, 3/4]$, i.e.,

$$\hat{\sigma}_e^2 = 2\int_{\mathbb{T}_X^*} \{\tilde{v}(t) - \hat{v}_X(t,t)\}\mathrm{d}t.$$

Bandwidths h_{μ_X}, $h_{v_{XY}}$, h_{v_X} and h_{σ_e} are all tuned through GCV, i.e., they are chosen to minimize

$$\frac{\boldsymbol{u}^\top \mathbf{W}^{1/2} \mathbf{P}^{\perp}_{\mathbf{W}^{1/2}[\mathbf{1}_M, \mathbf{T}]} \mathbf{W}^{1/2} \boldsymbol{u}}{\{\sum_{i=1}^n L_i - \operatorname{tr}(\mathbf{P}_{\mathbf{W}^{1/2}[\mathbf{1}_M, \mathbf{T}]})\}^2}$$
$$= \frac{\boldsymbol{u}^\top \{\mathbf{W} - \mathbf{W}[\mathbf{1}_M, \mathbf{T}]([\mathbf{1}_M, \mathbf{T}]^\top \mathbf{W}[\mathbf{1}_M, \mathbf{T}])^+ [\mathbf{1}_M, \mathbf{T}]^\top \mathbf{W}\} \boldsymbol{u}}{\{\sum_{i=1}^n L_i - \operatorname{tr}(\mathbf{P}_{\mathbf{W}^{1/2}[\mathbf{1}_M, \mathbf{T}]})\}^2}$$

with their respective corresponding values of $\boldsymbol{u}$, $\mathbf{T}$ and $\mathbf{W}$. [30, Eq. 4.3] suggested a rule of thumb for the bandwidth, i.e., a crude guess of bandwidth.

A.3.2 Assumptions, lemmas, and proofs

Recall that trajectories are observed at time points $T_{i\ell} \overset{\text{iid}}{\sim} T$, the numbers of observations are $L_i \overset{\text{iid}}{\sim} L$, predictor trajectories are $X_i \overset{\text{iid}}{\sim} X$, scalar responses are $Y_i \overset{\text{iid}}{\sim} Y$ and measurement errors are $e_{i\ell} \overset{\text{iid}}{\sim} e$ and $\varepsilon_{i\ell} \overset{\text{iid}}{\sim} \varepsilon$. Write f_1, f_2 and f_3 as respective pdfs of $T_{i\ell}$, $(T_{i\ell}, \widetilde{X}_i(T_{i\ell}))$ and $(T_{i\ell_1}, T_{i\ell_2}, \widetilde{X}_i(T_{i\ell_1}), \widetilde{X}_i(T_{i\ell_2}))$ in which $\widetilde{X}_i(T_{i\ell}) = X_i(T_{i\ell}) + \sigma_e e_{i\ell}$ are noisy observations. Regularity conditions (C.A.3.1)–(C.A.3.7) are imposed on the above random variables and pdfs. Some other restrictions are necessary for hyper-parameters in LLS: specifically, kernel function κ is assumed to fulfill (C.A.3.8)–(C.A.3.10); taking bandwidths h_{μ_X}, h_{v_X}, h_{σ_e} and $h_{v_{XY}}$ as functions of n, (C.A.3.11)–(C.A.3.14) require proper convergence rates of them as n diverges. Condition (C.A.3.15) (resp. (C.A.3.16)) ensures the consistency of PLEASS estimators in the L^2 (resp. L^∞) sense. These conditions limit the highest divergence rate of $p = p(n)$: as n diverges, p is chosen to diverge not faster than $O(\zeta_2^{-1}) = O(n^{1/2} h_{v_X}^2)$, where ζ_2 is defined as in (C.A.3.12). This restriction on p is fairly close to the setting in

[28, Theorem 5.3], where the discussion is limited to the case of $\|v_X\|_2 < 1$ (reachable by changing the scale on which X_i is measured). In detail we suppose:

(C.A.3.1) $\mathrm{E}(L) < \infty$ and $\Pr(L \geq 2) > 0$.

(C.A.3.2) μ_X and v_{XY} are both continuous on $\mathbb{T}_X$. v_X is continuous on $\mathbb{T}_X^2$. Hence $\|\mu_X\|_\infty$, $\|v_{XY}\|_\infty$ and $\|v_X\|_\infty$ are all finite.

(C.A.3.3) $X_i, T_{i1}, \ldots, T_{iL_i}$ and $e_{i1}, \ldots, e_{iL_i}$ are all independent of L_i in the sense that, given $L_i = \ell$, $X_i, T_{i1}, \ldots, T_{i\ell}$ and $e_{i1}, \ldots, e_{i\ell}$ are all independent and the conditional laws are those of X, T, and e.

(C.A.3.4) $\mathrm{E}\{X(T) - \mu_X(T) + e\sigma_e\}^4 < \infty$.

(C.A.3.5) $(\mathrm{d}^2/\mathrm{d}t^2)f_1$ exists and is continuous on $\mathbb{T}_X$ with pdf f_1 supported on $\mathbb{T}_X$.

(C.A.3.6) $(\mathrm{d}^2/\mathrm{d}t^2)f_2$ exists and is uniformly continuous on $\mathbb{T}_X \times \mathbb{R}$.

(C.A.3.7) $\{\mathrm{d}^2/(\mathrm{d}t_1\mathrm{d}t_2)\}f_3$, $(\mathrm{d}^2/\mathrm{d}t_1^2)f_3$ and $(\mathrm{d}^2/\mathrm{d}t_2^2)f_3$ all exist and are uniformly continuous on $\mathbb{T}_X^2 \times \mathbb{R}^2$.

(C.A.3.8) Kernel function κ in Appendix A.3.1 is symmetric (w.r.t. the y axis) and non-negative on $\mathbb{R}$ such that $\int_{\mathbb{R}} \kappa(t)\mathrm{d}t = 1$, $\int_{\mathbb{R}} t^2\kappa(t)\mathrm{d}t < \infty$, and $\int_{\mathbb{R}} \kappa^2(t)\mathrm{d}t < \infty$.

(C.A.3.9) Kernel function κ is compactly supported, i.e., it has a bounded support.

(C.A.3.10) The Fourier transform of κ is absolutely integrable, i.e., $\int_{\mathbb{R}} |\int_{\mathbb{R}} e^{-ist}\kappa(s)\mathrm{d}s|\mathrm{d}t < \infty$.

(C.A.3.11) $h_{\mu_X} \to 0$, $nh_{\mu_X}^4 \to \infty$, $nh_{\mu_X}^6 = O(1)$, and $\zeta_1 = n^{-1/2}h_{\mu_X}^{-1} = o(1)$, as $n \to \infty$.

(C.A.3.12) $h_{v_X} \to 0$, $nh_{v_X}^6 \to \infty$, $nh_{v_X}^8 = O(1)$, and $\zeta_2 = n^{-1/2}h_{v_X}^{-2} = o(1)$, as $n \to \infty$.

(C.A.3.13) $h_{\sigma_e} \to 0$, $nh_{\sigma_e}^4 \to \infty$, $nh_{\sigma_e}^6 = O(1)$, and $\zeta_3 = n^{-1/2}(h_{v_X}^{-2} + h_{\sigma_e}^{-1}) = o(1)$, as $n \to \infty$.

(C.A.3.14) $h_{v_{XY}} \to 0$, $nh_{v_{XY}}^4 \to \infty$, $nh_{v_{XY}}^6 = O(1)$, and $\zeta_4 = n^{-1/2}(h_{\mu_X}^{-1} + h_{v_{XY}}^{-1}) = o(1)$, as $n \to \infty$.

(C.A.3.15) As $n \to \infty$, $p = p(n) = O(\zeta_2^{-1})$. Additional requirements on p vary with the magnitude of $\|v_X\|_2$; they also depend on τ_p, the smallest eigenvalue of $\mathbf{D}_p$ at (A.11).

- If $\|v_X\|_2 \geq 1$, then the two terms $\tau_p^{-1}p\|v_X\|_2^{2p}\zeta_4 \max(1, \tau_p^{-1}p\|v_X\|_2^{2p})$ and $\tau_p^{-1}p^2\|v_X\|_2^{2p}\zeta_2 \max(1, \tau_p^{-1}p\|v_X\|_2^{2p})$ are both of order $o(1)$;
- if $\|v_X\|_2 < 1$, then $\tau_p^{-2}\max(\zeta_2, \zeta_4)$ and $\tau_p^{-1}\max(\zeta_2, \zeta_4)$ are both of order $o(1)$.

(C.A.3.16) Condition (C.A.3.15) holds with the L^2-norm $\|\cdot\|_2$ replaced by the infinity norm $\|\cdot\|_\infty$.

Among the above conditions, the first fourteen are inherited from [104, 105]. So is Lemma A.3 that states the convergence rate of estimators through LLS in Appendix A.3.1.

Lemma A.3. *Under assumptions (C.A.3.1)–(C.A.3.14), as $n \to \infty$,*

$$\begin{aligned}\|\hat{\mu}_X - \mu_X\|_\infty &= O_p(\zeta_1) = o_p(1),\\ \|\hat{v}_X - v_X\|_\infty &= O_p(\zeta_2) = o_p(1),\\ |\hat{\sigma}_e^2 - \sigma_e^2| &= O_p(\zeta_3) = o_p(1),\end{aligned}$$

and

$$\|\hat{v}_{XY} - v_{XY}\|_\infty = O_p(\zeta_4) = o_p(1), \tag{A.7}$$

where ζ_1, ζ_2, ζ_3 and ζ_4 are respectively defined in (C.A.3.11)–(C.A.3.14).

Proof of Lemma A.3. Synthesize [104, Theorem 1 and Corollary 1] and [105, Lemma A.1]. □

Lemma A.4. *Assume (C.A.3.1)–(C.A.3.14) and that there is a $C > 0$ such that for all n we have $p \in [1, C\zeta_2^{-1}]$. Then, for each $\epsilon > 0$, there are positive constants C_1 and C_2 and an integer $n_0 > 0$ such that, for each $n > n_0$,*

$$\Pr\left[\bigcap_{j=1}^{p}\{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2 \le C_1\|v_X\|_2^{j-1}\zeta_4 + C_2(j-1)\|v_X\|_2^{j-1}\zeta_2\}\right] \ge 1-\epsilon,$$

and

$$\Pr\left[\bigcap_{j=1}^{p}\{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_\infty \le C_1\|v_X\|_\infty^{j-1}\zeta_4 + C_2(j-1)\|v_X\|_\infty^{j-1}\zeta_2\}\right] \ge 1-\epsilon.$$

Proof of Lemma A.4. Recall definitions of $\mathcal{V}_X$ at (1.1) and $\widehat{\mathcal{V}}_X$ at (2.2). Since $\mathcal{V}_X(\beta) = v_{XY}$ and $\widehat{\mathcal{V}}_X(\beta) = \hat{v}_{XY}$, Lemma A.4 reduces to (A.7) when $j = 1$. For integer $j \ge 2$ and each $t \in \mathbb{T}_X$, the inequality that

$$\begin{aligned}&|\widehat{\mathcal{V}}_X^j(\beta)(t) - \mathcal{V}_X^j(\beta)(t)|\\ &\quad= |\widehat{\mathcal{V}}_X\{\widehat{\mathcal{V}}_X^{j-1}(\beta) - \mathcal{V}_X^{j-1}(\beta)\}(t) + (\widehat{\mathcal{V}}_X - \mathcal{V}_X)\{\mathcal{V}_X^{j-1}(\beta)\}(t)|\\ &\quad\le \|\widehat{\mathcal{V}}_X^{j-1}(\beta) - \mathcal{V}_X^{j-1}(\beta)\|_2\left\{\int_{\mathbb{T}_X}\hat{v}_X^2(s,t)\mathrm{d}s\right\}^{1/2}\\ &\qquad+ \|\mathcal{V}_X^{j-1}(\beta)\|_2\left[\int_{\mathbb{T}_X}\{\hat{v}_X(s,t) - v_X(s,t)\}^2\mathrm{d}s\right]^{1/2} \quad \text{(Cauchy-Schwarz)}\end{aligned}$$

implies that

$$\begin{aligned}\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2 &\le \|\hat{v}_X\|_2\|\mathcal{V}_X^{j-1}(\beta) - \widehat{\mathcal{V}}_X^{j-1}(\beta)\|_2 + \|\mathcal{V}_X^{j-1}(\beta)\|_2\|v_X - \hat{v}_X\|_2,\\ \|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_\infty &\le \|\hat{v}_X\|_\infty\|\mathcal{V}_X^{j-1}(\beta) - \widehat{\mathcal{V}}_X^{j-1}(\beta)\|_\infty + \|\mathcal{V}_X^{j-1}(\beta)\|_2\|v_X - \hat{v}_X\|_\infty.\end{aligned}$$

On iteration these two inequalities give that, respectively,

$$\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2$$

$$\leq \|\hat{v}_X\|_2^{j-1}\|\mathcal{V}_X(\beta) - \widehat{\mathcal{V}}_X(\beta)\|_2 + \|v_X - \hat{v}_X\|_2 \sum_{k=1}^{j-1} \|\mathcal{V}_X^k(\beta)\|_2 \|\hat{v}_X\|_2^{j-k-1}, \tag{A.8}$$

$$\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_\infty$$
$$\leq \|\hat{v}_X\|_\infty^{j-1}\|\mathcal{V}_X(\beta) - \widehat{\mathcal{V}}_X(\beta)\|_\infty + \|v_X - \hat{v}_X\|_\infty \sum_{k=1}^{j-1} \|\mathcal{V}_X^k(\beta)\|_2 \|\hat{v}_X\|_\infty^{j-k-1}. \tag{A.9}$$

For each $\epsilon > 0$, there is $n_0 > 0$ such that, for all $n > n_0$, we have

$$\begin{aligned}
1 - \epsilon/2 \leq\ & \Pr(\|\hat{v}_X - v_X\|_2 \leq C_0\zeta_2) \leq \Pr(\|\hat{v}_X\|_2 \leq \|v_X\|_2 + C_0\zeta_2), \\
1 - \epsilon/2 \leq\ & \Pr(\|\hat{v}_X - v_X\|_\infty \leq C_0\zeta_2) \leq \Pr(\|\hat{v}_X\|_\infty \leq \|v_X\|_\infty + C_0\zeta_2), \\
1 - \epsilon/2 \leq\ & \Pr(\|\hat{v}_{XY} - v_{XY}\|_2 \leq C_0\zeta_4), \\
1 - \epsilon/2 \leq\ & \Pr(\|\hat{v}_{XY} - v_{XY}\|_\infty \leq C_0\zeta_4),
\end{aligned}$$

with constant $C_0 > 0$, by Lemma A.3. It follows from (A.8) that

$$\begin{aligned}
1 - \epsilon \leq\ & \Pr\left[\bigcap_{j=1}^{p} \Big\{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2 \leq (\|v_X\|_2 + C_0\zeta_2)^{j-1}C_0\zeta_4 \right. \\
& \qquad \left. + C_0\zeta_2 \sum_{k=1}^{j-1} \|v_X\|_2^k \|\beta\|_2 (\|v_X\|_2 + C_0\zeta_2)^{j-k-1}\Big\}\right] \\
\leq\ & \Pr\left[\bigcap_{j=1}^{p} \Big\{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2 \leq C_0(1 + C_0\zeta_2/\|v_X\|_2)^{j-1}\|v_X\|_2^{j-1}\zeta_4 \right. \\
& \qquad \left. + C_0\|\beta\|_2\zeta_2\|v_X\|_2^{j-1} \sum_{k=1}^{j-1} (1 + C_0\zeta_2/\|v_X\|_2)^{j-k-1}\Big\}\right] \\
\leq\ & \Pr\left[\bigcap_{j=1}^{p} \Big\{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2 \leq C_1\|v_X\|_2^{j-1}\zeta_4 + C_2(j-1)\|v_X\|_2^{j-1}\zeta_2\Big\}\right], \\
& \qquad \text{(if } p \leq C\zeta_2^{-1} \text{ with arbitrarily fixed } C > 0)
\end{aligned}$$

where $C_1 = C_0 \exp(CC_0/\|v_X\|_2) \geq C_0 \exp(CC_0/\|v_X\|_\infty)$ and $C_2 = \|\beta\|_2 C_1$. It is worth noting that we have assumed that the range of p is constrained in $[1, C\zeta_2^{-1}]$; the quantity $(1 + C_0\zeta_2/\|v_X\|_2)^p$ may not be bounded if p diverges too fast. Similarly, inequality (A.9) implies that, for $1 \leq p \leq C\zeta_2^{-1}$,

$$\Pr\left[\bigcap_{j=1}^{p} \{\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_\infty \leq C_1\|v_X\|_\infty^{j-1}\zeta_4 + C_2(j-1)\|v_X\|_\infty^{j-1}\zeta_2\}\right] \geq 1 - \epsilon.$$

□

Proof of Theorem 4.1. Ascribed to [28, Eq. (3.6)], the following alternative expression of β_p at (4.9) dramatically facilitates our further moves:

$$\beta_p = \beta_p(\cdot) = [\mathcal{V}_X(\beta)(\cdot), \ldots, \mathcal{V}_X^p(\beta)(\cdot)]\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p, \tag{A.10}$$

where

$$\mathbf{D}_p = [d_{j_1,j_2}]_{1\le j_1,j_2\le p}, \tag{A.11}$$

$$\boldsymbol{\alpha}_p = [\alpha_1,\ldots,\alpha_p]^\top, \tag{A.12}$$

with

$$d_{j_1,j_2} = \int_{\mathbb{T}_X} \mathcal{V}_X^{j_1+1}(\beta)\mathcal{V}_X^{j_2}(\beta) = \int_{\mathbb{T}_X} \mathcal{V}_X^{j_1}(\beta)\mathcal{V}_X^{j_2+1}(\beta),$$
$$\alpha_j = \int_{\mathbb{T}_X} \mathcal{V}_X(\beta)\mathcal{V}_X^{j}(\beta) = \int_{\mathbb{T}_X} v_{XY}\mathcal{V}_X^{j}(\beta).$$

As is known, $\mathbf{D}_p^{-1}$ and $\boldsymbol{\alpha}_p$ are bounded, respectively, as

$$\|\mathbf{D}_p^{-1}\|_2 = \tau_p^{-1} \tag{A.13}$$

and

$$\begin{aligned}\|\boldsymbol{\alpha}_p\|_2 = \left[\sum_{j=1}^{p}\left\{\int_{\mathcal{T}} v_{XY}\mathcal{V}_X^{j}(\beta)\right\}^2\right]^{1/2} &\le \left[\sum_{j=1}^{p}\|v_{XY}\|_2^2\|\mathcal{V}_X^{j}(\beta)\|_2^2\right]^{1/2} \quad \text{(Cauchy-Schwarz)}\\ &= \begin{cases} O(p^{1/2}\|v_X\|_2^p) & \text{if } \|v_X\|_2 \ge 1\\ O(1) & \text{if } \|v_X\|_2 < 1.\end{cases}\end{aligned} \tag{A.14}$$

Accordingly, rewrite $\hat{\beta}_p$ at (4.17) as

$$\hat{\beta}_p = \hat{\beta}_p(\cdot) = [\widehat{\mathcal{V}}_X(\beta)(\cdot),\ldots,\widehat{\mathcal{V}}_X^{p}(\beta)(\cdot)]\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p \tag{A.15}$$

in which $\widehat{\mathbf{D}}_p = [\hat{d}_{j_1,j_2}]_{1\le j_1,j_2\le p}$ and $\hat{\boldsymbol{\alpha}}_p = [\hat{\alpha}_1,\ldots,\hat{\alpha}_p]^\top$ are respective empirical counterparts of $\mathbf{D}_p$ at (A.11) and $\boldsymbol{\alpha}_p$ at (A.12), with

$$\hat{d}_{j_1,j_2} = \int_{\mathbb{T}_X} \widehat{\mathcal{V}}_X^{j_1+1}(\beta)\widehat{\mathcal{V}}_X^{j_2}(\beta),$$
$$\hat{\alpha}_j = \int_{\mathbb{T}_X} \widehat{\mathcal{V}}_X(\beta)\widehat{\mathcal{V}}_X^{j}(\beta) = \int_{\mathbb{T}_X} \hat{v}_{XY}\widehat{\mathcal{V}}_X^{j}(\beta).$$

Observe that, by the Cauchy-Schwarz inequality,

$$\begin{aligned}|\alpha_j - \hat{\alpha}_j| &= \left|\int_{\mathbb{T}_X}(v_{XY}-\hat{v}_{XY})\mathcal{V}_X^{j}(\beta)\right| + \left|\int_{\mathbb{T}_X}\hat{v}_{XY}\{\widehat{\mathcal{V}}_X^{j}(\beta)-\mathcal{V}_X^{j}(\beta)\}\right|\\ &\le \|\beta\|_2\|v_X\|_2^j\|\hat{v}_{XY}-v_{XY}\|_2 + \|\hat{v}_{XY}\|_2\|\widehat{\mathcal{V}}_X^{j}(\beta)-\mathcal{V}_X^{j}(\beta)\|_2.\end{aligned}$$

For every $\epsilon > 0$ and $1 \le p \le C\zeta_2^{-1}$, there is $n_0 > 0$ such that, $\forall n > n_0$,

$$1-\epsilon \le \Pr\left[\bigcap_{j=1}^{p}\{|\alpha_j-\hat{\alpha}_j| \le C_3\|v_X\|_2^{j-1}\zeta_4 + C_4(j-1)\|v_X\|_2^{j-1}\zeta_2\}\right], \quad \text{(Lemmas A.3–A.4)}$$

with constants $C_3, C_4 > 0$. Analogously, writing $\Delta_{jk} = \hat{d}_{jk} - d_{jk}$, the Cauchy-Schwarz inequality implies that

$$\begin{aligned}|\Delta_{jk}| &\le \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 \|\widehat{\mathcal{V}}_X^k(\beta)\|_2 + \|\widehat{\mathcal{V}}_X^k(\beta) - \mathcal{V}_X^k(\beta)\|_2 \|\mathcal{V}_X^{j+1}(\beta)\|_2 \\ &\le \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 \|\hat{v}_X\|_2^k \|\beta\|_2 + \|\widehat{\mathcal{V}}_X^k(\beta) - \mathcal{V}_X^k(\beta)\|_2 \|v_X\|_2^{j+1} \|\beta\|_2,\end{aligned}$$

and further, by Lemmas A.3 and A.4, as long as $1 \le p \le C\zeta_2^{-1}$,

$$\begin{aligned}1 - \epsilon &\le \Pr\left[\bigcap_{j,k=1}^{p} \{|\Delta_{jk}| \le \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 (\|v_X\|_2 + C_0\zeta_2^{-1})^k \|\beta\|_2 \right. \\ &\qquad \left. + \|\widehat{\mathcal{V}}_X^k(\beta) - \mathcal{V}_X^k(\beta)\|_2 \|v_X\|_2^{j+1} \|\beta\|_2\}\right] \\ &\le \Pr\left[\bigcap_{j,k=1}^{p} \{|\Delta_{jk}| \le C_5 \|v_X\|_2^{j+k} \zeta_4 + C_6 \max(j, k-1) \|v_X\|_2^{j+k} \zeta_2\}\right],\end{aligned}$$

where C_5 and C_6 are positive constants. Thus, if $\boldsymbol{\Delta}_p = [\Delta_{jk}]_{p\times p} = \widehat{\mathbf{D}}_p - \mathbf{D}_p$, then

$$\begin{aligned}\|\boldsymbol{\Delta}_p\|_2^2 &\le \sum_{1\le j,k\le p} \Delta_{jk}^2 \\ &= O_p\left(\zeta_4^2 \sum_{1\le j,k\le p} \|v_X\|_2^{2j+2k}\right) + O_p\left[\zeta_2^2 \sum_{1\le j,k\le p} \max\{j^2, (k-1)^2\} \|v_X\|_2^{2j+2k}\right] \\ &= \begin{cases} O_p(p^2\|v_X\|_2^{4p}\zeta_4^2) + O_p(p^4\|v_X\|_2^{4p}\zeta_2^2) & \text{if } \|v_X\|_2 \ge 1 \\ O_p(\zeta_4^2) + O_p(\zeta_2^2) & \text{if } \|v_X\|_2 < 1. \end{cases}\end{aligned} \tag{A.16}$$

In a similar manner, one proves that

$$\begin{aligned}\|\hat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2^2 &= \sum_{1\le j\le p} |\hat{\alpha}_j - \alpha_j|^2 \\ &= O_p\left(C_1\zeta_4^2 \sum_{1\le j\le p} \|v_X\|_2^{2j-2}\right) + O_p\left\{\zeta_2^2 \sum_{1\le j\le p} (j-1)^2 \|v_X\|_2^{2j-2}\right\} \\ &= \begin{cases} O_p(p\|v_X\|_2^{2p}\zeta_4^2) + O_p(p^3\|v_X\|_2^{2p}\zeta_2^2) & \text{if } \|v_X\|_2 \ge 1 \\ O_p(\zeta_4^2) + O_p(\zeta_2^2) & \text{if } \|v_X\|_2 < 1. \end{cases}\end{aligned} \tag{A.17}$$

Denote by τ_p the smallest eigenvalue of $\mathbf{D}_p$. Notice that, for $p = p(n) = O(\zeta_2^{-1})$,

$$\begin{aligned}&\|\mathbf{D}_p^{-1}\boldsymbol{\Delta}_p\|_2 \\ &\quad \le \tau_p^{-1}\|\boldsymbol{\Delta}_p\|_2 \\ &\quad = \begin{cases} O_p(\tau_p^{-1}p\|v_X\|_2^{2p}\zeta_4) + O_p(\tau_p^{-1}p^2\|v_X\|_2^{2p}\zeta_2) & \text{if } \|v_X\|_2 \ge 1 \\ O_p(\tau_p^{-1}\zeta_4) + O_p(\tau_p^{-1}\zeta_2) & \text{if } \|v_X\|_2 < 1. \end{cases} \quad \text{(by (A.13) and (A.16))}\end{aligned}$$

Provided that (C.A.3.15) holds, for sufficiently large n, one has $\tau_p^{-1}\|\boldsymbol{\Delta}_p\|_2 < \gamma$, for some $\gamma \in (0,1)$. In this case, [28, Eq. (7.18)] argues that, as n goes to infinity,

$$\widehat{\mathbf{D}}_p^{-1} = \{\mathbf{I} - \mathbf{D}_p^{-1}\boldsymbol{\Delta}_p + O_p(\tau_p^{-2}\|\boldsymbol{\Delta}_p\|_2^2)\}\mathbf{D}_p^{-1},$$

which can be rewritten as

$$\begin{aligned}
&\|\widehat{\mathbf{D}}_p^{-1} - \mathbf{D}_p^{-1}\|_2 \\
&\quad = \|\{O_p(\tau_p^{-2}\|\boldsymbol{\Delta}_p\|_2^2) - \mathbf{D}_p^{-1}\boldsymbol{\Delta}_p\}\mathbf{D}_p^{-1}\|_2 \\
&\quad = \begin{cases} \tau_p^{-1}\|O_p(\tau_p^{-2}p^2\|v_X\|_2^{4p}\zeta_4^2) + O_p(\tau_p^{-2}p^4\|v_X\|_2^{4p}\zeta_2^2) - \mathbf{D}_p^{-1}\boldsymbol{\Delta}_p\|_2 & \text{if } \|v_X\|_2 \geq 1 \\ \tau_p^{-1}\|O_p(\tau_p^{-2}\zeta_4^2) + O_p(\tau_p^{-2}\zeta_2^2) - \mathbf{D}_p^{-1}\boldsymbol{\Delta}_p\|_2 & \text{if } \|v_X\|_2 < 1 \end{cases} \quad \text{(by (A.16))} \\
&\quad = \begin{cases} O_p(\tau_p^{-2}p\|v_X\|_2^{2p}\zeta_4) + O_p(\tau_p^{-2}p^2\|v_X\|_2^{2p}\zeta_2) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(\tau_p^{-2}\zeta_4) + O_p(\tau_p^{-2}\zeta_2) & \text{if } \|v_X\|_2 < 1. \end{cases} \quad \text{(by (C.A.3.15))}
\end{aligned} \tag{A.18}$$

Combining (A.13), (A.14), (A.17) and (A.18), one obtains

$$\begin{aligned}
&\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2 \\
&\qquad \leq \|\widehat{\mathbf{D}}_p^{-1} - \mathbf{D}_p^{-1}\|_2\|\boldsymbol{\alpha}_p\|_2 + \|\widehat{\mathbf{D}}_p^{-1}\|_2\|\hat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2 \\
&\qquad = \begin{cases} O_p(\tau_p^{-2}p^{3/2}\|v_X\|_2^{3p}\zeta_4) + \ O_p(\tau_p^{-2}p^{5/2}\|v_X\|_2^{3p}\zeta_2) \\ \quad + O_p(\tau_p^{-1}p^{1/2}\|v_X\|_2^{p}\zeta_4) + O_p(\tau_p^{-1}p^{3/2}\|v_X\|_2^{p}\zeta_2) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(\tau_p^{-2}\zeta_4) + O_p(\tau_p^{-2}\zeta_2) + O_p(\tau_p^{-1}\zeta_4) + O_p(\tau_p^{-1}\zeta_2) & \text{if } \|v_X\|_2 < 1. \end{cases}
\end{aligned} \tag{A.19}$$

Next, for each $t \in \mathbb{T}_X$, we have

$$\begin{aligned}
&|\hat{\beta}_p(t) - \beta_p(t)|^2 \\
&= \left|[\widehat{\mathcal{V}}_X(\beta)(t), \ldots, \widehat{\mathcal{V}}_X^p(\beta)(t)]\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - [\mathcal{V}_X(\beta)(t), \ldots, \mathcal{V}_X^p(\beta)(t)]\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\right|^2 \\
&\leq \left|\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2\left[\sum_{j=1}^p\{\widehat{\mathcal{V}}_X^j(\beta)(t)\}^2\right]^{1/2} + \|\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2\left[\sum_{j=1}^p\{\widehat{\mathcal{V}}_X^j(\beta)(t) - \mathcal{V}_X^j(\beta)(t)\}^2\right]^{1/2}\right|^2 \\
&\leq 2\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\left[\sum_{j=1}^p\{\widehat{\mathcal{V}}_X^j(\beta)(t)\}^2\right] + 2\|\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\left[\sum_{j=1}^p\{\widehat{\mathcal{V}}_X^j(\beta)(t) - \mathcal{V}_X^j(\beta)(t)\}^2\right].
\end{aligned}$$

Thus $\|\hat{\beta}_p - \beta_p\|_2$ is bounded as below:

$$\begin{aligned}
\|\hat{\beta}_p - \beta_p\|_2^2 &\leq 2\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^p\|\mathcal{V}_X^j(\beta)\|_2^2 + 2\|\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^p\|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_2^2 \\
&\leq 2\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^p\|\mathcal{V}_X^j(\beta)\|_2^2 \qquad \text{(A.20)} \\
&\quad + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^p\|\widehat{\mathcal{V}}_X^j(\beta) - \mathcal{V}_X^j(\beta)\|_2^2. \qquad \text{(A.21)}
\end{aligned}$$

Owing to (A.19),

$$(\text{A.20}) = \begin{cases} O_p(\tau_p^{-4}p^4\|v_X\|_2^{8p}\zeta_4^2) + O_p(\tau_p^{-4}p^6\|v_X\|_2^{8p}\zeta_2^2) \\ \quad + O_p(\tau_p^{-2}p^2\|v_X\|_2^{4p}\zeta_4^2) + O_p(\tau_p^{-2}p^4\|v_X\|_2^{4p}\zeta_2^2) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(\tau_p^{-4}\zeta_4^2) + O_p(\tau_p^{-4}\zeta_2^2) + O_p(\tau_p^{-2}\zeta_4^2) + O_p(\tau_p^{-2}\zeta_2^2) & \text{if } \|v_X\|_2 < 1; \end{cases}$$

the rate of (A.21) is given by (A.14) and Lemma A.4 jointly, i.e.,

$$(\text{A.21}) = \begin{cases} O_p(\tau_p^{-2}p^2\|v_X\|_2^{4p}\zeta_4^2) + O_p(\tau_p^{-2}p^4\|v_X\|_2^{4p}\zeta_2^2) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(\tau_p^{-2}\zeta_4^2) + O_p(\tau_p^{-2}\zeta_2^2) & \text{if } \|v_X\|_2 < 1. \end{cases}$$

In this way we deduce

$$\|\hat{\beta}_p - \beta_p\|_2^2 = \begin{cases} O_p(\tau_p^{-4}p^4\|v_X\|_2^{8p}\zeta_4^2) + O_p(\tau_p^{-4}p^6\|v_X\|_2^{8p}\zeta_2^2) \\ \quad + O_p(\tau_p^{-2}p^2\|v_X\|_2^{4p}\zeta_4^2) + O_p(\tau_p^{-2}p^4\|v_X\|_2^{4p}\zeta_2^2) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(\tau_p^{-4}\zeta_4^2) + O_p(\tau_p^{-4}\zeta_2^2) + O_p(\tau_p^{-2}\zeta_4^2) + O_p(\tau_p^{-2}\zeta_2^2) & \text{if } \|v_X\|_2 < 1. \end{cases} \quad (\text{A.22})$$

Condition (C.A.3.15) then implies that both (A.20) and (A.21) converge to 0 in probability. The consistency of PLEASS estimators in the L^2 sense follows, from the L^2 convergence of β_p to β [28, Theorem 3.2].

Finally, we bound the estimation error in the supremum metric:

$$\begin{aligned} &\|\hat{\beta}_p - \beta_p\|_\infty^2 \\ &= \left\|[\widehat{\mathcal{V}}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta)](\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p) + [\widehat{\mathcal{V}}_X(\beta) - \mathcal{V}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta) - \mathcal{V}_X^p(\beta)]\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\right\|_\infty \\ &\leq \left[\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2 \left\{\sum_{j=1}^p \|\widehat{\mathcal{V}}_X^j(\beta)\|_\infty^2\right\}^{1/2} + \|\mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2 \left\{\sum_{j=1}^p \|\widehat{\mathcal{V}}_X^j(\beta) - \mathcal{V}_X^j(\beta)\|_\infty^2\right\}^{1/2}\right]^2 \\ &\leq 2\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\mathcal{V}_X^j(\beta)\|_\infty^2 + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\mathcal{V}_X^j(\beta) - \widehat{\mathcal{V}}_X^j(\beta)\|_\infty^2 \\ &\leq 2\|\widehat{\mathbf{D}}_p^{-1}\hat{\boldsymbol{\alpha}}_p - \mathbf{D}_p^{-1}\boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\mathcal{V}_X^j(\beta)\|_\infty^2 \quad \text{(different from (A.20) only in the metric)} \\ &\quad + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\widehat{\mathcal{V}}_X^j(\beta) - \mathcal{V}_X^j(\beta)\|_\infty^2 \quad \text{(different from (A.21) only in the metric)} \\ &= \begin{cases} O_p(\tau_p^{-4}p^4\|v_X\|_\infty^{8p}\zeta_4^2) + O_p(\tau_p^{-4}p^6\|v_X\|_\infty^{8p}\zeta_2^2) \\ \quad + O_p(\tau_p^{-2}p^2\|v_X\|_\infty^{4p}\zeta_4^2) + O_p(\tau_p^{-2}p^4\|v_X\|_\infty^{4p}\zeta_2^2) & \text{if } \|v_X\|_\infty \geq 1 \\ O_p(\tau_p^{-4}\zeta_4^2) + O_p(\tau_p^{-4}\zeta_2^2) + O_p(\tau_p^{-2}\zeta_4^2) + O_p(\tau_p^{-2}\zeta_2^2) & \text{if } \|v_X\|_\infty < 1. \end{cases} \end{aligned}$$

That is, the upper bound for $\|\hat{\beta}_p - \beta_p\|_\infty$ can be obtained from (A.22) by replacing $\|v_X\|_2$ with $\|v_X\|_\infty$. Condition (C.A.3.16) completes the proof for the zero-convergence of $\|\hat{\beta}_p - \beta\|_\infty$, as long as we assume $\|\beta_p - \beta\|_\infty \to 0$ as $p \to \infty$. □

Proof of Theorem 4.2. Recall β_p (at (4.8) and (A.10)) and $\hat{\beta}_p$ (at (4.17) and (A.15)). Introduce $\boldsymbol{s}_p = [\mathcal{V}_X(w_1), \ldots, \mathcal{V}_X(w_p)]^\top$ and its empirical counterpart $\hat{\boldsymbol{s}}_p = [\widehat{\mathcal{V}}_X(\hat{w}_1), \ldots, \widehat{\mathcal{V}}_X(\hat{w}_p)]^\top$. Note the identities $\boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} \mathbf{H}_p^\top = \mathcal{V}_X(\beta_p)$ and $\hat{\boldsymbol{c}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1} \widehat{\mathbf{H}}_p^\top = \widehat{\mathcal{V}}_X(\hat{\beta}_p)$. Thus, (C.A.3.1)–(C.A.3.15) jointly ensure that, for arbitrarily given $L^*, T_1^*, \ldots, T_{L^*}$,

$$
\begin{aligned}
&\|\mathbf{H}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{c}_p - \widehat{\mathbf{H}}_p \widehat{\boldsymbol{\Lambda}}_p^{-1} \hat{\boldsymbol{c}}_p\|_2^2 \\
&\quad \le L^* \|\boldsymbol{s}_p^\top \boldsymbol{\Lambda}_p^{-1} \boldsymbol{c}_p - \hat{\boldsymbol{s}}_p^\top \widehat{\boldsymbol{\Lambda}}_p^{-1} \hat{\boldsymbol{c}}_p\|_\infty^2 \\
&\quad = L^* \sup_{t \in \mathbb{T}_X} \left| \int_{\mathbb{T}_X} v_X(s,t) \{\beta_p(s) - \hat{\beta}_p(s)\} \mathrm{d}s + \int_{\mathbb{T}_X} (v_X - \hat{v}_X)(s,t) \hat{\beta}_p(s) \mathrm{d}s \right|^2 \\
&\quad \le L^* \sup_{t \in \mathbb{T}_X} \left| \left\{ \int_{\mathbb{T}_X} v_X^2(s,t) \mathrm{d}s \right\}^{1/2} \|\beta_p - \hat{\beta}_p\|_2 + \left\{ \int_{\mathbb{T}_X} (v_X - \hat{v}_X)^2(s,t) \mathrm{d}s \right\}^{1/2} \|\hat{\beta}_p\|_2 \right|^2 \\
&\quad \le L^* (\|v_X\|_\infty \|\beta_p - \hat{\beta}_p\|_2 + \|v_X - \hat{v}_X\|_\infty \|\hat{\beta}_p\|_2)^2 \\
&\quad \to_p 0. \quad \text{(by Lemma A.3 and Theorem 4.1)}
\end{aligned}
$$

The convergence to 0 (in probability and conditional on L^* and $T_1^*, \ldots, T_{L^*}$) of $\hat{\eta}_p(X^*) - \tilde{\eta}_\infty(X^*)$ (with $\hat{\eta}_p(X^*)$ at (1.8) and $\tilde{\eta}_\infty(X^*)$ at (4.20)) follows from Lemma A.3 and the continuous mapping and Slutsky's theorems. Since L^* and $T_1^*, \ldots, T_{L^*}$ are arbitrary, the dominated convergence theorem enables us to drop the conditioning. This completes the proof of Theorem 4.2. □

Proof of Corollary 4.2.1. Recall $\eta_p(X^*)$ at (4.9), $\tilde{\eta}_p(X^*)$ at (4.19) and $\tilde{\eta}_\infty(X^*)$ at (4.20). As discussed in the last paragraph of Section 4.2.1, $[\tilde{\xi}_1^* - \xi_1^*, \ldots, \tilde{\xi}_p^* - \xi_p^*]^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_p - \mathbf{H}_p^\top \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1} \mathbf{H}_p)$. It follows that $\tilde{\eta}_p(X^*) - \eta_p(X^*) \sim \mathcal{N}\{0, \boldsymbol{c}_p^\top \boldsymbol{\Lambda}_p^{-1} (\boldsymbol{\Lambda}_p - \mathbf{H}_p^\top \boldsymbol{\Sigma}_{\widetilde{X}^*}^{-1} \mathbf{H}_p) \boldsymbol{\Lambda}_p^{-1} \boldsymbol{c}_p\}$ and further $\hat{\eta}_p(X^*) - \eta_p(X^*)$ converges (in distribution) to $\mathcal{N}(0, \omega)$ as $n \to \infty$, by Theorem 4.2. The Slutsky's theorem completes the proof. □

A.4 Technical details for Chapter 5

(C.A.4.1) $\sum_{i,j=1}^\infty \lambda_{i,X}^{-2} \left\{ \int_{\mathbb{T}_Y} \int_{\mathbb{T}_X} \phi_{i,X}(s) v_{XY}(s,t) \phi_{j,Y}(t) \mathrm{d}s \mathrm{d}t \right\}^2 < \infty$ and β belongs to the range of $\mathcal{V}_X$, say range($\mathcal{V}_X$).

(C.A.4.2) $\mathrm{E}(\|X\|_2^4) < \infty$ for all $t \in \mathbb{T}_Y$.

(C.A.4.3) Let $\mathbb{T}_X = [0,1]$. Both $\|\xi_X\|_{\infty,2}$ and $\|\psi_X\|_{\infty,2}$ are of order $O_p(1)$ as $n \to \infty$, with ξ_X and ψ_X defined in the statement of Lemma A.5 and $\|\cdot\|_{\infty,2}$ defined such that $\|f\|_{\infty,2} = \sup_{s \in \mathbb{T}_X} \{\int_{\mathbb{T}_X} f^2(s,t) \mathrm{d}t\}^{1/2}$ for $f \in L^2(\mathbb{T}_X^2)$.

(C.A.4.4) As $n \to \infty$, $p = p(n) = O(n^{1/2})$. Meanwhile, $\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2 / \tau_p \le \rho$ for certain $\rho \in (0,1)$ when n is sufficiently large. (Here τ_p is the smallest eigenvalue of $\boldsymbol{H}_p$. Here $\|\cdot\|_2$ is abused for the matrix norm induced by the Euclidean norm, i.e., for arbitrary $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ and $\boldsymbol{b} \in \mathbb{R}^{q \times 1}$ $\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{b}: \|\boldsymbol{b}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{b}\|_2$ is actually the largest eigenvalue of $\boldsymbol{A}$. It reduces to the Euclidean norm for vectors.)

(C.A.4.5) Additional requirements on p vary with the magnitude of $\|v_X\|_2$; they also depend on τ_p, the smallest eigenvalue of $\boldsymbol{H}_p$.

- If $\|v_X\|_2 \geq 1$, then, as $n \to \infty$, $n^{-1}\tau_p^{-2}p^4\|v_X\|_2^{4p}\max(1, \tau_p^{-2}p^2\|v_X\|_2^{4p})$ and $n^{-1}\tau_p^{-3}p^5\|v_X\|_2^{6p}$ are both of order $o(1)$;
- if $\|v_X\|_2 < 1$, then $(n\tau_p^4)^{-1} = o(1)$ as n diverges.

(C.A.4.6) Keep everything in (C.A.4.5) but substitute $\|v_X\|_\infty$ for $\|v_X\|_2$. Meanwhile, require that $\|\beta_{p,\mathrm{fAPLS}} - \beta\|_\infty = o(1)$ as p diverges, viz. an enhanced version of Proposition 5.1.

(C.A.4.7) Stochastic process Y is "eventually totally bounded in mean" (defined as [46, Eq. 5–7]); i.e., in our context,

- $\mathrm{E}(\|Y\|_\infty) < \infty$;
- for each $\epsilon > 0$, there is a finite cover of $\mathbb{T}$, say $\mathrm{Co}(\mathbb{T})$, for each set $\mathbb{A} \in \mathrm{Co}(\mathbb{T})$, such that $\inf_{n\in\mathbb{Z}^+} n^{-1}\,\mathrm{E}\{\sup_{s,t\in\mathbb{A}}|Y(s) - Y(t)|\} < \epsilon$.

Introduced by [44], (C.A.4.1) is set up to guarantee the uniqueness and identifiability of β in FoFR (1.9). It is also adopted by [105]. Assumptions (C.A.4.2)–(C.A.4.4) are prerequisites for L^2-convergence results in [28]. One may feel unclear about the technical conditions stated in (C.A.4.5) for the scenario of $\|v_X\|_2 \geq 1$: virtually a special case for is that $n^{-1}\max(\tau_p^{-4}, \tau_p^{-6}, \tau_p^{-8}) = o(1)$ and $p = O(\ln\ln n)$. Apparently, p is more restricted when $\|v_X\|_2 \geq 1$ than in the case of $\|v_X\|_2 < 1$ (for the latter case p is allowed to diverge at the rate of $O(n^{1/2})$); that is why [28] suggested changing the scale on which X is measured. (C.A.4.6) is an upgrade of (C.A.4.5), handling the uniform convergence (in probability). At last, we add (C.A.4.7) as a prerequisite of the uniform law of large numbers for $\{Y_i \mid i \geq 1\}$.

Lemma A.5. *For each $(s, u, t) \in \mathbb{T}_X^2 \times \mathbb{T}_Y$,*

$$\begin{aligned}\hat{v}_X(s,t) &= v_X(s,t) + n^{-1/2}\xi_X(s,t) + n^{-1}\psi_X(s,t),\\ \hat{v}_{XY}(s,t) &= v_{XY}(s,t) + n^{-1/2}\xi_{XY}(s,t) + n^{-1}\psi_{XY}(s,t)\end{aligned} \tag{A.23}$$

where, with identity operator $I : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}\xi_X(s,t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I - \mathrm{E})[\{X_i(s) - \mu_X(s)\}\{X_i(t) - \mu_X(t)\}],\\ \psi_X(s,t) &= -n\{\bar{X}(s) - \mu_X(s)\}\{\bar{X}(t) - \mu_X(t)\},\\ \xi_{XY}(s,t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I - \mathrm{E})[\{X_i(s) - \mu_X(s)\}\{Y_i(t) - \mu_Y(t)\}],\\ \psi_{XY}(s,t) &= -n\{\bar{X}(s) - \mu_X(s)\}\{\bar{Y}(t) - \mu_Y(t)\},\end{aligned}$$

and $\|\xi_X\|_2$, $\|\psi_X\|_2$, $\|\xi_{XY}\|_2$ and $\|\psi_{XY}\|_2$ all equal $O_p(1)$ as n diverges.

Proof of Lemma A.5. It is an immediate implication of [28, Eq. 5.1]. □

Lemma A.6. *Assume (C.A.4.1) and (C.A.4.2) and that there is $C > 0$ such that, for all n, we have $p \le Cn^{-1/2}$. Then, for each $\epsilon > 0$, there are positive C_1, C_2 and n_0 such that, for each $n > n_0$,*

$$\Pr\left[\bigcap_{i=1}^{p}\left\{\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 \le n^{-1/2}\|v_X\|_2^{i-1}\{C_1 + C_2(i-1)\}\right\}\right] \ge 1 - \epsilon.$$

Assuming one more condition (C.A.4.3),

$$\Pr\left[\bigcap_{i=1}^{p}\left\{\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_\infty \le n^{-1/2}\|v_X\|_\infty^{i-1}\{C_1 + C_2(i-1)\}\right\}\right] \ge 1 - \epsilon.$$

Proof of Lemma A.6. Since $\mathcal{V}_X(\beta) = v_{XY}$ and $\widehat{\mathcal{V}}_X(\beta) = \hat{v}_{XY}$, Lemma A.6 is simply implied by Lemma A.5 when $p = 1$. For integer $i \ge 2$ and each $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$,

$$\begin{aligned}
&|\widehat{\mathcal{V}}_X^i(\beta)(s,t) - \mathcal{V}_X^i(\beta)(s,t)| \\
&\quad = |\widehat{\mathcal{V}}_X\{\widehat{\mathcal{V}}_X^{i-1}(\beta) - \mathcal{V}_X^{i-1}(\beta)\}(s,t) + (\widehat{\mathcal{V}}_X - \mathcal{V}_X)\{\mathcal{V}_X^{i-1}(\beta)\}(s,t)| \\
&\quad \le \left\{\int_{\mathbb{T}_X} \hat{v}_X^2(s,u)\mathrm{d}u\right\}^{1/2}\left[\int_{\mathbb{T}_X}\{\widehat{\mathcal{V}}_X^{i-1}(\beta) - \mathcal{V}_X^{i-1}(\beta)\}(u,t)\mathrm{d}u\right]^{1/2} \\
&\qquad + \left[\int_{\mathbb{T}_X}\{\hat{v}_X(s,u) - v_X(s,u)\}^2\mathrm{d}u\right]^{1/2}\left\{\int_{\mathbb{T}_X}\mathcal{V}_X^{i-1}(\beta)(u,t)\mathrm{d}u\right\}^{1/2}.
\end{aligned}$$

It implies that, by the triangle inequality,

$$\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 \le \|\hat{v}_X\|_2\|\widehat{\mathcal{V}}_X^{i-1}(\beta) - \mathcal{V}_X^{i-1}(\beta)\|_2 + \|\hat{v}_X - v_X\|_2\|\mathcal{V}_X^{i-1}(\beta)\|_2.$$

On iteration it gives that

$$\begin{aligned}
&\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 \\
&\qquad \le \|\hat{v}_X\|_2^{i-1}\|\widehat{\mathcal{V}}_X(\beta) - \mathcal{V}_X(\beta)\|_2 + \|\hat{v}_X - v_X\|_2\sum_{j=1}^{i-1}\|\hat{v}_X\|_2^{i-j-1}\|\mathcal{V}_X^j(\beta)\|_2. \quad \text{(A.24)}
\end{aligned}$$

For each $\epsilon > 0$, there is $n_0 > 0$ such that, for all $n > n_0$, we have

$$1 - \epsilon/2 \le \Pr(\|\hat{v}_X - v_X\|_2 \le C_0 n^{-1/2}) \le \Pr(\|\hat{v}_X\|_2 \le \|v_X\|_2 + C_0 n^{-1/2})$$

and

$$1 - \epsilon/2 \le \Pr(\|\hat{v}_{XY} - v_{XY}\|_2 \le C_0 n^{-1/2}),$$

with constant $C_0 > 0$, by Lemma A.5. It follows (A.24) that

$$\begin{aligned}
1 - \epsilon \le \Pr\Bigg[\bigcap_{i=1}^{p}\Bigg[\|(\widehat{\mathcal{V}}_X^i - \mathcal{V}_X^i)(\beta)\|_2 \le C_0 n^{-1/2}\Bigg\{&(\|v_X\|_2 + C_0 n^{-1/2})^{i-1} \\
&+ \sum_{j=1}^{i-1}\|v_X\|_2^j\|\beta\|_2(\|v_X\|_2 + C_0 n^{-1/2})^{i-j-1}\Bigg\}\Bigg]\Bigg]
\end{aligned}$$

$$\leq \Pr\left[\bigcap_{i=1}^{p}\left[\|(\widehat{\mathcal{V}}_X^i - \mathcal{V}_X^i)(\beta)\|_2 \leq C_0 n^{-1/2}\|v_X\|_2^{i-1}\left\{(1+C_0n^{-1/2}/\|v_X\|_2)^{i-1}\right.\right.\right.$$
$$\left.\left.\left. + \|\beta\|_2\sum_{j=1}^{i-1}(1+C_0n^{-1/2}/\|v_X\|_2)^{i-j-1}\right\}\right]\right]$$
$$\leq \Pr\left[\bigcap_{i=1}^{p}\left\{\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2\right.\right.$$
$$\left.\left. \leq n^{-1/2}\|v_X\|_2^{i-1}\{C_1 + C_2(i-1)\}\right\}\right], \quad (\text{since } p \leq Cn^{1/2})$$

where $C_1 = C_0\exp(CC_0/\|v_X\|_2)$ and $C_2 = \|\beta\|_2 C_1$.

Suppose (C.A.4.3) holds. Similar to (A.24),

$$\begin{aligned}\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_\infty &\leq \|\hat{v}_X\|_\infty^{i-1}\|\widehat{\mathcal{V}}_X(\beta) - \mathcal{V}_X(\beta)\|_\infty \\ &\quad + \|\hat{v}_X - v_X\|_\infty\sum_{j=1}^{i-1}\|\hat{v}_X\|_\infty^{i-j-1}\|\mathcal{V}_X^j(\beta)\|_\infty \\ &\leq \|\hat{v}_X\|_\infty^{i-1}\|\widehat{\mathcal{V}}_X(\beta) - \mathcal{V}_X(\beta)\|_\infty \\ &\quad + \|\hat{v}_X - v_X\|_\infty\sum_{j=1}^{i-1}\|\hat{v}_X\|_\infty^{i-j-1}\|v_X\|_\infty^j\|\beta\|_\infty.\end{aligned}$$

Mimicking the argument above for the L^2 sense, one obtains that

$$\Pr\left[\bigcap_{i=1}^{p}\left\{\|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_\infty \leq n^{-1/2}\|v_X\|_\infty^{i-1}\{C_1 + C_2(i-1)\}\right\}\right] \geq 1-\epsilon,$$

with, at this time, $C_1 = C_0\exp(CC_0/\|v_X\|_\infty)$ and $C_2 = \|\beta\|_\infty C_1$. The finiteness of $\|\beta\|_\infty$ originates from the continuity of eigenfunctions $\phi_{i,X}$'s and $\phi_{i,Y}$'s (refer to the Mercer's theorem). □

Proof of Proposition 5.1. Recall $\beta_{p,q,\text{FPCR}}$ at (5.1) and introduce $\beta_{p,\infty,\text{FPCR}} \in L^2(\mathbb{T}_X \times \mathbb{T}_Y)$ such that

$$\beta_{p,\infty,\text{FPCR}}(s,t) = \lim_{q\to\infty}\beta_{p,q,\text{FPCR}}(s,t) = \sum_{i=1}^{p}\frac{\phi_{i,X}(s)}{\lambda_{i,X}}\int_{\mathbb{T}_X}\phi_{i,X}(u)v_{XY}(u,t)\mathrm{d}u.$$

It follows that

$$\mathcal{V}_X(\beta_{p,\infty,\text{FPCR}})(s,t) = \sum_{i=1}^{p}\phi_{i,X}(s)\int_{\mathbb{T}_X}\phi_{i,X}(u)v_{XY}(u,t)\mathrm{d}u.$$

Now

$$[(\lambda_{1,X}I - \mathcal{V}_X)\circ\cdots\circ(\lambda_{p,X}I - \mathcal{V}_X)](\beta_{p,\infty,\text{FPCR}}) = 0$$

in which the left-hand side equals $\sum_{i=0}^{p} a_i \mathcal{V}_X^i(\beta_{p,\infty,\mathrm{FPCR}})$ with $a_0 = \prod_{i=1}^{p} \lambda_{i,X} > 0$. Therefore,

$$\beta_{p,\infty,\mathrm{FPCR}} = -\sum_{i=1}^{p} \frac{a_i}{a_0} \mathcal{V}_X^i(\beta_{p,\infty,\mathrm{FPCR}}).$$

Denote by $P_p : \mathrm{range}(\mathcal{V}_X) \to \mathrm{range}(\mathcal{V}_X)$ the operator that projects elements in $\mathrm{range}(\mathcal{V}_X)$ to $\mathrm{span}\{f_{ij} \in L^2(\mathbb{T}_X \times \mathbb{T}_Y) \mid f_{ij}(s,t) = \phi_{i,X}(s)\phi_{j,Y}(t), 1 \le i \le p, j \ge 1\}$. Thus $\beta_{p,\infty,\mathrm{FPCR}} = P_p(\beta)$. Since $\mathcal{V}_X^i(\beta_{p,\infty,\mathrm{FPCR}}) = P_p[\mathcal{V}_X^i(\beta)]$, one has

$$P_p\left[\beta + \sum_{i=1}^{p} \frac{a_i}{a_0} \mathcal{V}_X^i(\beta)\right] = 0,$$

implying that, for all p,

$$P_p(\beta) \in \{P_p(f) \mid f \in \overline{\mathrm{KS}_\infty(\mathcal{V}_X, \beta)}\}.$$

Taking limits as $p \to \infty$ on both sides of the above formula, we obtain $\beta \in \overline{\mathrm{KS}_\infty(\mathcal{V}_X, \beta)}$ and accomplish the proof. □

Proof of Proposition 5.2. Recall $\beta_{p,\mathrm{fAPLS}}$ (5.3) and $\hat{\beta}_{p,\mathrm{fAPLS}}$ (5.8) and notations in defining them. The Cauchy-Schwarz inequality implies that

$$\begin{aligned}
|\hat{h}_{ij} - h_{ij}| &\le \|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 \|\widehat{\mathcal{V}}_X^{j+1}(\beta)\|_2 + \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 \|\mathcal{V}_X^i(\beta)\|_2 \\
&\le \|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 \|\hat{v}_X\|_2^{j+1} \|\beta\|_2 + \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 \|v_X\|_2^i \|\beta\|_2.
\end{aligned}$$

By Lemmas A.5 and A.6, for each $\epsilon > 0$ and $p \le Cn^{1/2}$, there are positive n_0, C_3 and C_4 such that, for all $n > n_0$,

$$\begin{aligned}
1 - \epsilon \le \Pr\Bigg[\bigcap_{i,j=1}^{p} \Big\{ |\hat{h}_{ij} - h_{ij}| \le & \|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2 (\|v_X\|_2 + C_0 n^{-1/2})^{j+1} \|\beta\|_2 \\
& + \|\widehat{\mathcal{V}}_X^{j+1}(\beta) - \mathcal{V}_X^{j+1}(\beta)\|_2 \|v_X\|_2^i \|\beta\|_2 \Big\} \Bigg] \\
\le \Pr\Bigg[\bigcap_{i,j=1}^{p} \Big\{ |\hat{h}_{ij} - h_{ij}| \le n^{-1/2} \|v_X\|_2^{i+j} \{C_3 \max(i,j) + C_4\} \Big\} \Bigg].
\end{aligned}$$

Thus

$$\begin{aligned}
\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2^2 &\le \sum_{j,k=1}^{p} |\hat{h}_{ij} - h_{ij}|^2 \\
&= O_p\left(n^{-1} \sum_{i,j=1}^{p} \|v_X\|_2^{2i+2j} \right) + O_p\left\{ n^{-1} \sum_{i,j=1}^{p} \max(i^2, j^2) \|v_X\|_2^{2i+2j} \right\} \\
&= \begin{cases} O_p(n^{-1}p^2 \|v_X\|_2^{4p}) + O_p(n^{-1}p^4 \|v_X\|_2^{4p}) & \text{if } \|v_X\|_2 \ge 1 \\ O_p(n^{-1}) & \text{if } \|v_X\|_2 < 1 \end{cases}
\end{aligned}$$

$$= \begin{cases} O_p(n^{-1}p^4\|v_X\|_2^{4p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1}) & \text{if } \|v_X\|_2 < 1. \end{cases} \tag{A.25}$$

It is analogous to (A.25) to deduce that

$$\|\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2^2 = \sum_{j=1}^{p} |\hat{\alpha}_j - \alpha_j|^2 = \begin{cases} O_p(n^{-1}p^3\|v_X\|_2^{2p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1}) & \text{if } \|v_X\|_2 < 1. \end{cases} \tag{A.26}$$

Denote by τ_p the smallest eigenvalue of $\boldsymbol{H}_p$. Noting that $\|\boldsymbol{H}_p^{-1}\|_2 = \tau_p^{-1}$, for $p \leq Cn^{1/2}$,

$$\|(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\boldsymbol{H}_p^{-1}\|_2 \leq \tau_p^{-1}\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2 = \begin{cases} O_p(n^{-1/2}\tau_p^{-1}p^2\|v_X\|_2^{2p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1/2}\tau_p^{-1}) & \text{if } \|v_X\|_2 < 1. \end{cases}$$

Introduce random matrix $\boldsymbol{M}_p \in \mathbb{R}^{p\times p}$ such that

$$\boldsymbol{I} - \boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p) + \boldsymbol{M}_p = \{\boldsymbol{I} + \boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\}^{-1},$$

i.e.,

$$\boldsymbol{M}_p = \{\boldsymbol{I} + \boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\}^{-1}\boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p).$$

Therefore,

$$\|\boldsymbol{M}_p\|_2 \leq \|\boldsymbol{I} + \boldsymbol{H}^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2^{-1}\|\boldsymbol{H}^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2^2 \leq (1-\rho)^{-1}\tau_p^{-2}\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2^2,$$

provided that $\tau_p^{-1}\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2 \leq \rho < 1$ (refer to (C.A.4.4)). Revealed by the identity that $\widehat{\boldsymbol{H}}_p^{-1} = \{\boldsymbol{I} + \boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\}^{-1}\boldsymbol{H}_p^{-1}$,

$$\begin{aligned} &\|\widehat{\boldsymbol{H}}_p^{-1} - \boldsymbol{H}_p^{-1}\|_2 \\ &\quad \leq \{\|\boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2 + \|\boldsymbol{M}_p\|_2\}\|\boldsymbol{H}_p^{-1}\|_2 \\ &\quad = \begin{cases} O_p(n^{-1/2}\tau_p^{-2}p^2\|v_X\|_2^{2p}) + O_p(n^{-1}\tau_p^{-3}p^4\|v_X\|_2^{4p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1/2}\tau_p^{-2}) + O_p(n^{-1}\tau_p^{-3}) & \text{if } \|v_X\|_2 < 1. \end{cases} \end{aligned} \tag{A.27}$$

Combining (A.26), (A.27) and the identity that

$$\begin{aligned} \|\boldsymbol{\alpha}_p\|_2 &= \left[\sum_{i=1}^{p}\left\{\int_{\mathbb{T}_Y}\int_{\mathbb{T}_X} v_{XY}(s,t)\mathcal{V}_X^i(\beta)(s,t)\mathrm{d}s\mathrm{d}t\right\}^2\right]^{1/2} \\ &\leq \left[\sum_{i=1}^{p}\|v_{XY}\|_2^2\|\mathcal{V}_X^i(\beta)\|_2^2\right]^{1/2} \\ &= \begin{cases} O(p^{1/2}\|v_X\|_2^p) & \text{if } \|v_X\|_2 \geq 1 \\ O(1) & \text{if } \|v_X\|_2 < 1, \end{cases} \end{aligned} \tag{A.28}$$

we reach that

$$\begin{aligned} &\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2 \\ &\quad \leq \|\widehat{\boldsymbol{H}}_p^{-1}\|_2\|\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2 + \|\widehat{\boldsymbol{H}}_p^{-1} - \boldsymbol{H}_p^{-1}\|_2\|\boldsymbol{\alpha}_p\|_2 \end{aligned}$$

$$
\begin{aligned}
&= \begin{cases} O_p(n^{-1/2}\tau_p^{-1}p^{3/2}\|v_X\|_2^p) \\ \qquad + O_p(n^{-1/2}\tau_p^{-2}p^{5/2}\|v_X\|_2^{3p}) + O_p(n^{-1}\tau_p^{-3}p^{9/2}\|v_X\|_2^{5p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1/2}\tau_p^{-1}) + O_p(n^{-1/2}\tau_p^{-2}) + O_p(n^{-1}\tau_p^{-3}) & \text{if } \|v_X\|_2 < 1 \end{cases} \\
&= \begin{cases} O_p(n^{-1/2}\tau_p^{-1}p^{3/2}\|v_X\|_2^p) \\ \qquad + O_p(n^{-1/2}\tau_p^{-2}p^{5/2}\|v_X\|_2^{3p}) + O_p(n^{-1}\tau_p^{-3}p^{9/2}\|v_X\|_2^{5p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1/2}\tau_p^{-2}) + O_p(n^{-1}\tau_p^{-3}) \quad (\text{since } \tau_p \leq h_{ii} = O(1)) & \text{if } \|v_X\|_2 < 1. \end{cases}
\end{aligned} \tag{A.29}
$$

For each $(s,t) \in \mathbb{T}_X \times \mathbb{T}_Y$,

$$
\begin{aligned}
&|\hat{\beta}_{p,\mathrm{fAPLS}}(s,t) - \beta_{p,\mathrm{fAPLS}}(s,t)|^2 \\
&\quad = \Bigg| [\widehat{\mathcal{V}}_X(\beta)(s,t), \ldots, \widehat{\mathcal{V}}_X^p(\beta)(s,t)] \widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p \\
&\qquad - [\mathcal{V}_X(\beta)(s,t), \ldots, \mathcal{V}_X^p(\beta)(s,t)] \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p \Bigg|^2 \\
&\quad \leq \Bigg| \|\widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2 \left[\sum_{i=1}^p \{\widehat{\mathcal{V}}_X^i(\beta)(s,t)\}^2 \right]^{1/2} \\
&\qquad + \|\boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2 \left[\sum_{i=1}^p [\{\widehat{\mathcal{V}}_X^i - \mathcal{V}_X^i\}(\beta)(s,t)]^2 \right]^{1/2} \Bigg|^2 \\
&\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \left[\sum_{i=1}^p \{\widehat{\mathcal{V}}_X^i(\beta)(s,t)\}^2 \right] \\
&\qquad + 2\|\boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \left[\sum_{i=1}^p \{\widehat{\mathcal{V}}_X^i(\beta)(s,t) - \mathcal{V}_X^i(\beta)(s,t)\}^2 \right].
\end{aligned}
$$

Thus $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2$ is bounded as below:

$$
\begin{aligned}
&\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2^2 \\
&\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta)\|_2^2 + 2\|\boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta) - \widehat{\mathcal{V}}_X^i(\beta)\|_2^2 \\
&\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta)\|_2^2 \qquad \text{(A.30)} \\
&\qquad + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_2^2, \qquad \text{(A.31)}
\end{aligned}
$$

where, owing to (A.29),

$$
\text{(A.30)} = \begin{cases} O_p(n^{-1}\tau_p^{-2}p^4\|v_X\|_2^{4p}) \\ \qquad + O_p(n^{-1}\tau_p^{-4}p^6\|v_X\|_2^{8p}) + O_p(n^{-2}\tau_p^{-6}p^{10}\|v_X\|_2^{12p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}) & \text{if } \|v_X\|_2 < 1; \end{cases}
$$

the order of (A.31) is jointly given by (A.28) and Lemma A.6, i.e.,

$$(\text{A.31}) = \begin{cases} O(n^{-1}\tau_p^{-2}p^4\|v_X\|_2^{4p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1}\tau_p^{-2}) & \text{if } \|v_X\|_2 < 1. \end{cases}$$

In this way we deduce

$$\begin{aligned} &\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2^2 \\ &= \begin{cases} O_p(n^{-1}\tau_p^{-2}p^4\|v_X\|_2^{4p}) \\ \qquad + O_p(n^{-1}\tau_p^{-4}p^6\|v_X\|_2^{8p}) + O_p(n^{-2}\tau_p^{-6}p^{10}\|v_X\|_2^{12p}) & \text{if } \|v_X\|_2 \geq 1 \\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}) & \text{if } \|v_X\|_2 < 1. \end{cases} \end{aligned} \tag{A.32}$$

A set of necessary conditions for the zero-convergence (in probability) of (A.32) is contained in (C.A.4.5). Once they are fulfilled, we conclude the L^2 convergence (in probability) of $\hat{\beta}_{p,\mathrm{fAPLS}}$ to β following Proposition 5.1.

We complete the proof by bounding the estimating error in the supremum metric:

$$\begin{aligned} &\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_\infty^2 \\ &\quad = \left\| [\widehat{\mathcal{V}}_X(\beta), \ldots, \widehat{\mathcal{V}}_X^p(\beta)]\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - [\mathcal{V}_X(\beta), \ldots, \mathcal{V}_X^p(\beta)]\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p \right\|_\infty^2 \\ &\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta)\|_\infty^2 + 2\|\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta) - \widehat{\mathcal{V}}_X^i(\beta)\|_\infty^2 \\ &\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\mathcal{V}_X^i(\beta)\|_\infty^2 \quad \text{(compare (A.30))} \\ &\qquad + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2 \sum_{i=1}^p \|\widehat{\mathcal{V}}_X^i(\beta) - \mathcal{V}_X^i(\beta)\|_\infty^2, \quad \text{(compare (A.31))} \\ &\quad = \begin{cases} O_p(n^{-1}\tau_p^{-2}p^4\|v_X\|_\infty^{4p}) \\ \qquad + O_p(n^{-1}\tau_p^{-4}p^6\|v_X\|_\infty^{8p}) + O_p(n^{-2}\tau_p^{-6}p^{10}\|v_X\|_\infty^{12p}) & \text{if } \|v_X\|_\infty \geq 1 \\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}) & \text{if } \|v_X\|_\infty < 1, \end{cases} \end{aligned}$$

converging to zero (in probability) with the satisfaction of (C.A.4.6). The zero-convergence (in probability) of $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_\infty$ follows if we assume that $\|\beta_{p,\mathrm{fAPLS}} - \beta\|_\infty \to 0$ as p diverges. □

Proof of Proposition 5.3. Notice that

$$\begin{aligned} &\|\hat{\eta}_{p,\mathrm{fAPLS}}(X_0) - \eta(X_0)\|_2 \leq \|\bar{Y} - \mu_Y\|_2 + \|\bar{X} - \mu_X\|_2\|\beta\|_2 + \|X_0 - \bar{X}\|_2\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_2, \\ &\|\hat{\eta}_{p,\mathrm{fAPLS}}(X_0) - \eta(X_0)\|_\infty \leq \|\bar{Y} - \mu_Y\|_\infty + \|\bar{X} - \mu_X\|_2\|\beta\|_\infty + \|X_0 - \bar{X}\|_2\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_\infty. \end{aligned}$$

The finite trace of $\mathcal{V}_X$ (resp. $\mathcal{V}_Y$), viz. $\sum_{i=1}^\infty \lambda_{i,X} = \mathrm{E}(\|X - \mu_X\|_2^2) < \infty$ (resp. $\sum_{i=1}^\infty \lambda_{i,Y} = \mathrm{E}(\|Y - \mu_Y\|_2^2) < \infty$), entails that $\|\bar{X} - \mu_X\|_2 = o_{\mathrm{a.s.}}(1)$ (resp. $\|\bar{Y} - \mu_Y\|_2 = o_{\mathrm{a.s.}}(1)$); see [47, Eq 2.1.3]. The proof is completed once we verify the zero-convergence (in probability and under (C.A.4.7)) of $\|\bar{Y} - \mu_Y\|_\infty$ following [46, Theorem 2]. □