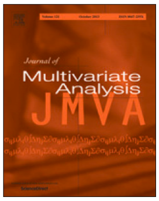
Journal of Multivariate Analysis 185 (2021) 104769

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Fast implementation of partial least squares for function-on-function regression

Zhiyang Zhou

Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, United States

ARTICLE INFO

Article history:
Received 17 December 2020
Received in revised form 5 May 2021
Accepted 6 May 2021
Available online 12 May 2021

AMS 2020 subject classifications:
primary 62R10
secondary 62G08

Keywords:
Functional data analysis
Functional linear model
Krylov subspace
Principal component analysis

ABSTRACT

People employ the function-on-function regression to model the relationship between two stochastic processes. Fitting this model, widely used strategies include functional partial least squares algorithms which typically require iterative eigen-decomposition. Here we introduce a route of functional partial least squares based upon Krylov subspace. Our route can be expressed in two forms equivalent to each other in exact arithmetic: One is non-iterative with explicit expressions of the estimator and prediction, facilitating the theoretical derivation and potential extensions; the other one stabilizes numerical outputs. The consistency of estimation and prediction is established under regularity conditions. It is highlighted that our proposal is competitive in terms of both estimation and prediction accuracy but consumes much less execution time.

1. Introduction

In recent years, the functional data analysis has enjoyed a rapid development, due to the growing demands of digging out the rich information from complicated data structures like trajectories and images. As a fundamental model in functional data analysis, the function-on-function regression (FoFR, arguably first proposed by Ramsay and Dalzell [31]) is a generalization of multivariate regression with the coefficient vector evolving into a bivariate coefficient function. It helps to model the relationship between two stochastic processes.

Excellent contributions have been made to the investigation into FoFR. Among others, Cuevas et al. [8] estimated the coefficient function by interpolation. [11,12] utilized the Nadaraya–Watson estimator in predicting the conditional expectation of response. A more prevailing stream of fitting FoFR, as documented in monographs like ([20], Section 8.3; [32], Chapter 16), is to reduce the intrinsically infinite dimension of coefficient function by focusing on a pre-determined lower dimensional space. There are various candidates for such a space: It can be a linear space spanned by wavelets, orthogonal polynomials, penalized splines, etc.; a more recent option is the reproducing kernel Hilbert space employed by [25,35]. Nevertheless, one may prefer a data-driven strategy named the functional principal component regression (FPCR): It constructs the lower dimensional space from leading eigenfunctions of auto-covariance operators. Least-squares-type projections onto such a space were proposed by [5,6,43] as estimators for coefficient function; Wang [37] made the projection through random effect models. Incorporating a regularization into [6], Benatia et al. [3] enabled FPCR even for FoFR with ill-posed auto-covariance operators. It is known that FPCR fails to involve the correlation between response and predictor in truncating the Karhunen–Loève series, resulting in a possible loss in accuracy. A remedy for this point is the functional partial least squares (FPLS). FPLS is a terminology shared by a series of algorithms. At least three

E-mail address: zhiyang.zhou@northwestern.edu.

https://doi.org/10.1016/j.jmva.2021.104769
0047-259X/© 2021 Elsevier Inc. All rights reserved.

of them are applicable to FoFR: [4] included respective extensions of the nonlinear iterative PLS (NIPALS, [39]) and the statistically inspired modification of PLS (SIMPLS, [22]), both initially designed for the multivariate context. SigComp [26] sequentially maximizes penalized Rayleigh quotients subject to constraints on normalization and orthogonality. Although these three FPLS algorithms have shown their own numerical advantages, they all have to solve iterative eigen-problems which may take time.

Our work chooses to constrain the estimator to a subspace named after (Alexei) Krylov. This idea expands the alternative partial least squares (APLS, [9], an FPLS algorithm designed for the scalar-on-function regression) and is abbreviated as fAPLS with initial “f” emphasizing the application to functional response. In the context of scalar-on-function regression, APLS is equivalent to NIPALS and SIMPLS. Though this equivalency is unlikely to hold for FoFR, our fAPLS is still expected to have little difference with NIPALS or SIMPLS in terms of accuracy. Meanwhile, involving no eigen-problem, fAPLS would inherit multiple features, e.g., closed-form estimators and less running time, from APLS.

The remaining portion of this paper is organized as below. After clarifying the model settings, Section 2 presents two equivalent expressions of fAPLS estimator. They facilitate the empirical implementation and theoretical derivation, respectively. In Section 3 fAPLS is compared with competitors under distinct simulation scenarios. Its performance is evaluated in terms of both accuracy and execution time. We apply fAPLS to two real-world datasets in Section 4. fAPLS is adaptable to more complex settings, e.g., sparsely observed predictors, correlated subjects or non-linear modeling; we include corresponding discussions in Section 5. More theoretical and computational details, e.g., assumptions, proofs and code trunks, are relegated to the Appendix for conciseness.

2. Model and method

In what follows, we formalize FoFR before sketching existing FPLS algorithms, viz. NIPALS [4], SIMPLS [4] and SigComp [26]. We then jump to our method which is motivated by Proposition 1.

2.1. Model

Let $X = X(s)$ and $Y = Y(t)$ be two L_2-processes respectively defined on closed intervals $\mathbb{I}_X$ and $\mathbb{I}_Y \subset \mathbb{R}$. FoFR is formulated as

$$Y(t) = \mu_Y(t) + \int_{\mathbb{I}_X} \{X(s) - \mu_X(s)\}\beta(s,t)\mathrm{d}s + \varepsilon(t), \tag{1}$$

where $\beta \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)$ is the coefficient function to be estimated; $\mu_X \in L_2(\mathbb{I}_X)$ and $\mu_Y \in L_2(\mathbb{I}_Y)$ denote unknown expectations of X and Y, respectively. The zero-mean Gaussian process $\varepsilon(t)$ has a covariance function continuous on $\mathbb{I}_Y \times \mathbb{I}_Y$ and is uncorrelated with $X(s)$, i.e., $\mathrm{cov}\{X(s), \varepsilon(t)\} = 0$ for all $(s,t) \in \mathbb{I}_X \times \mathbb{I}_Y$. The model (1) becomes

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

defining a random integral operator $\mathcal{L}_X : L_2(\mathbb{I}_X \times \mathbb{I}_Y) \to L_2(\mathbb{I}_Y)$ such that, for each $f \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)$,

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

Write auto-covariance functions $r_{XX} = r_{XX}(s,t) = \mathrm{cov}\{X(s), X(t)\}$ and $r_{YY} = r_{YY}(s,t) = \mathrm{cov}\{Y(s), Y(t)\}$, continuous respectively on $\mathbb{I}_X \times \mathbb{I}_X$ and $\mathbb{I}_Y \times \mathbb{I}_Y$. As well, define a cross-covariance function $r_{XY} = r_{XY}(s,t) = \mathrm{cov}\{X(s), Y(t)\}$ continuous with respect to $(s,t) \in \mathbb{I}_X \times \mathbb{I}_Y$. Correspondingly, an auto-covariance operator $R_{XX} : L_2(\mathbb{I}_X) \to L_2(\mathbb{I}_X)$ is given by, for each $f \in L_2(\mathbb{I}_X)$, $R_{XX}(f)(\cdot) = \int_{\mathbb{I}_X} r_{XX}(s,\cdot)f(s)\mathrm{d}s$. One more auto-covariance operator $R_{YY} : L_2(\mathbb{I}_Y) \to L_2(\mathbb{I}_Y)$ is defined in complete analogy to R_{XX}. Let $(\lambda_{j,X}, \phi_{j,X})$ (resp. $(\lambda_{j,Y}, \phi_{j,Y})$) be the two-tuple consisting of the jth leading eigenvalue and eigenfunction of R_{XX} (resp. R_{YY}). It is standard for functional data analysis to assume that $\sum_{j=1}^{\infty} \lambda_{j,X} < \infty$ and $\sum_{j=1}^{\infty} \lambda_{j,Y} < \infty$, with positive $\lambda_{j,X}$ and $\lambda_{j,Y}$.

Define a linear integral operator $\Gamma_{XX} : L_2(\mathbb{I}_X \times \mathbb{I}_Y) \to L_2(\mathbb{I}_X \times \mathbb{I}_Y)$ such that, for each $f \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)$,

$$\Gamma_{XX}(f)(s,t) = \int_{\mathbb{I}_X} r_{XX}(s,s')f(s',t)\mathrm{d}s', \quad (s,t) \in \mathbb{I}_X \times \mathbb{I}_Y.$$

Made of Γ_{XX} and β, a p-dimensional Krylov subspace is denoted by

$$\mathrm{KS}_p(\Gamma_{XX}, \beta) = \mathrm{span}\{\Gamma_{XX}(\beta), \ldots, \Gamma_{XX}^p(\beta)\}$$

in which $\mathrm{span}\{\cdot\}$ denotes the linear space spanned by elements inside braces. Let Γ_{XX}^0 be the identity operator. Thus, $\Gamma_{XX}^j : L_2(\mathbb{I}_X \times \mathbb{I}_Y) \to L_2(\mathbb{I}_X \times \mathbb{I}_Y)$, $j \in \mathbb{Z}^+$, are recursively defined as, for each $f \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)$ and each $(s,t) \in \mathbb{I}_X \times \mathbb{I}_Y$,

$$\Gamma_{XX}^j(f)(s,t) = (\Gamma_{XX} \circ \Gamma_{XX}^{j-1})(f)(s,t) = \Gamma_{XX}\{\Gamma_{XX}^{i-1}(f)\}(s,t) = \int_{\mathbb{I}_X} r_{XX}(s,s')\{\Gamma_{XX}^{i-1}(f)(s',t)\}\mathrm{d}s'.$$

Noting that $\Gamma_{XX}^j(\beta) = \Gamma_{XX}^{j-1}(r_{XY})$ for all positive j, we indeed incorporate the correlation between X and Y into $\mathrm{KS}_p(\Gamma_{XX}, \beta)$.

Proposition 1. *Under (C1) in the Appendix, true parameter* $\beta \in \overline{\mathrm{KS}_\infty(\Gamma_{XX}, \beta)} = \overline{\mathrm{span}\{\Gamma_{XX}^j(\beta) \mid j \geq 1\}}$*, with the overline representing the closure.*

Remark 1. Proposition 1 is not a corollary of [9, Theorem 3.2]. The latter one merely implies an identity weaker than Proposition 1: Fixing arbitrary $t_0 \in \mathbb{I}_Y$, univariate function $\beta(\cdot, t_0) \in \mathrm{span}\{\Gamma_{XX}^j(\beta)(\cdot, t_0) \mid j \geq 1\}$.

2.2. Background

We first sketch FPLS algorithms NIPALS, SIMPLS and SigComp. For FoFR, NIPALS seeks to approximate β within $\mathrm{span}\{u_{k,\mathrm{N}}(s)v_{k',\mathrm{N}}(t) \mid k, k' \in \{1, \ldots, K\}\}$. The two-tuple $(u_{1,\mathrm{N}}, v_{1,\mathrm{N}})$ maximizes $\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} u(s)r_{XY}(s,t)v(t)\mathrm{d}s\mathrm{d}t\}^2$ with respect to $(u, v) \in L_2(\mathbb{I}_X) \times L_2(\mathbb{I}_Y)$ subject to $\|u\|_2 = \|v\|_2 = 1$, denoting the L_2-norm by $\|\cdot\|_2$. NIPALS proceeds by updating $X - \mu_X$ and $Y - \mu_Y$ with their respective projections onto the orthogonal complement of $\mathrm{span}\{\int_{\mathbb{I}_X}\{X(s) - \mu_X(s)\}u_{1,\mathrm{N}}(s)\mathrm{d}s\}$. r_{XY} is updated accordingly. These steps are repeated with all indices eventually raised to K. Similar to NIPALS, SIMPLS begins with

$$(u_{1,\mathrm{S}}, v_{1,\mathrm{S}}) = \mathop{\arg\max}_{\substack{u \in L_2(\mathbb{I}_X), v \in L_2(\mathbb{I}_Y) \\ \|u\|_2 = \|v\|_2 = 1}} \left\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} u(s)r_{XY}(s,t)v(t)\mathrm{d}s\mathrm{d}t\right\}^2,$$

but it then approximates β by $\sum_{k=1}^K \sum_{k'=1}^K u_{k,\mathrm{S}}(s)\int_{\mathbb{I}_X} u_{k',\mathrm{S}}(s')r_{XY}(s',t)\mathrm{d}s'$, where $(u_{k,\mathrm{S}}, v_{k,\mathrm{S}})$, $k \in \{2, \ldots, K\}$, is the solution to

$$\max_{\substack{u \in L_2(\mathbb{I}_X), v \in L_2(\mathbb{I}_Y) \\ \|u\|_2 = \|v\|_2 = 1}} \left\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} u(s)r_{XY}(s,t)v(t)\mathrm{d}s\mathrm{d}t\right\}^2$$

subject to $\int_{\mathbb{I}_X} u_{k',\mathrm{S}}(s)\{R_{XX}(u)(s)\}\mathrm{d}s = 0$, $k' \in \{1, \ldots, k-1\}$. Assuming that $\beta \approx \sum_{k=1}^K u_{k,\mathrm{SC}}(s)v_{k,\mathrm{SC}}(t)$, SigComp first obtains

$$u_{1,\mathrm{SC}} = \mathop{\arg\max}_{u \in L_2(\mathbb{I}_X)} (1 + \mathrm{PEN}_1)^{-1}\int_{\mathbb{I}_X}\int_{\mathbb{I}_X}\int_{\mathbb{I}_Y} u(s)r_{XY}(s,t)r_{XY}(s',t)u(s')\mathrm{d}t\mathrm{d}s\mathrm{d}s'$$

subject to $\int_{\mathbb{I}_X} u(s)\{R_{XX}(u)(s)\}\mathrm{d}s = 1$, with PEN_1 penalizing the smoothness. The subsequent $u_{k,\mathrm{SC}}$, $k \in \{2, \ldots, K\}$, are sequentially constructed following orthonormality constraints. Given $u_{1,\mathrm{SC}}, \ldots, u_{K,\mathrm{SC}}$, desired functions $v_{1,\mathrm{SC}}, \ldots, v_{K,\mathrm{SC}}$ are exactly the last K elements of the solution to

$$\max_{b_0, b_1, \ldots, b_K \in L_2(\mathbb{I}_Y)} \mathrm{E}\left[\int_{\mathbb{I}_Y}\left\{Y(t) - b_0(t) - \sum_{k=1}^K b_k(t)\int_{\mathbb{I}_X} X(s)u_{k,\mathrm{SC}}(s)\mathrm{d}s\right\}^2 \mathrm{d}t\right] + \mathrm{PEN}_2$$

with penalty term PEN_2.

Inspired by Proposition 1, we propose to approximate β by the least-squares solution

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

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

$$h_{jj'} = \int_{\mathbb{I}_Y}\left[\int_{\mathbb{I}_X}\int_{\mathbb{I}_X} r_{XX}(s,s')\{\Gamma_{XX}^j(\beta)(s,t)\}\{\Gamma_{XX}^{j'}(\beta)(s',t)\}\mathrm{d}s\mathrm{d}s'\right]\mathrm{d}t = \int_{\mathbb{I}_Y}\int_{\mathbb{I}_X}\{\Gamma_{XX}^j(\beta)(s,t)\}\{\Gamma_{XX}^{j'+1}(\beta)(s,t)\}\mathrm{d}s\mathrm{d}t,$$

$$\alpha_i = \int_{\mathbb{I}_Y}\left[\int_{\mathbb{I}_X}\int_{\mathbb{I}_X} r_{XX}(s,s')\{\Gamma_{XX}^j(\beta)(s,t)\}\beta(s',t)\mathrm{d}s\mathrm{d}s'\right]\mathrm{d}t = \int_{\mathbb{I}_Y}\int_{\mathbb{I}_X}\{\Gamma_{XX}(\beta)(s,t)\}\{\Gamma_{XX}^j(\beta)(s,t)\}\mathrm{d}s\mathrm{d}t.$$

Proposition 1 justifies (2) by entailing that $\lim_{p\to\infty}\|\beta_{p,\mathrm{fAPLS}} - \beta\|_2 = 0$, which is crucial to our theoretical results delivered later in Section 2.4.

2.3. Estimation and prediction

Suppose n two-tuples (X_i, Y_i), $i \in \{1, \ldots, n\}$, are all independent realizations of (X, Y). We understand that, in practice, trajectories X_i and Y_i are recorded discretely. As long as observation points for each curve are sufficiently dense (see Section 5 if this denseness assumption is not satisfied), one may presmooth the curves through interpolation or smoothing techniques, e.g., the penalized B-spline smoothing [32, Chapter 5]; see [40] for error rates associated with penalized splines. For convenience, we keep using X_i and Y_i for smoothed curves.

It is natural to estimate $r_{XX}(s,s')$ and $r_{XY}(s,t)$ $(= \Gamma_{XX}(\beta)(s,t))$, $(s,s',t) \in \mathbb{I}_X \times \mathbb{I}_X \times \mathbb{I}_Y$, respectively, by

$$\hat{r}_{XX}(s,s') = \frac{1}{n}\sum_{i=1}^n X_i^{\mathrm{cent}}(s)X_i^{\mathrm{cent}}(s'), \quad \hat{r}_{XY}(s,t) = \widehat{\Gamma}_{XX}(\beta)(s,t) = \frac{1}{n}\sum_{i=1}^n X_i^{\mathrm{cent}}(s)Y_i^{\mathrm{cent}}(t), \tag{3}$$

where $X_i^{\mathrm{cent}} = X_i - \bar{X} = X_i - n^{-1}\sum_{i=1}^{n} X_i$ and $Y_i^{\mathrm{cent}} = Y_i - \bar{Y} = Y_i - n^{-1}\sum_{i=1}^{n} Y_i$. Given $\widehat{\Gamma}_{XX}^{j}(\beta)$, one can estimate $\Gamma_{XX}^{j+1}(\beta)(s,t)$ by

$$\widehat{\Gamma}_{XX}^{i+1}(\beta)(s,t) = \int_{\mathbb{I}_X} \hat{r}_{XX}(s,s')\{\widehat{\Gamma}_{XX}^{j}(\beta)(s',t)\}\mathrm{d}s' \tag{4}$$

in which the integral sign here represents the numerical integral through the trapezoidal rule; Tasaki [36] bounded the corresponding approximation error.

Plugging (3) and (4) into (2), an estimator for both $\beta_{p,\mathrm{fAPLS}}$ and β comes:

$$\hat{\beta}_{p,\mathrm{fAPLS}} = [\widehat{\Gamma}_{XX}(\beta), \ldots, \widehat{\Gamma}_{XX}^{p}(\beta)]\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p, \tag{5}$$

where $\widehat{\boldsymbol{H}}_p = [\hat{h}_{jj'}]_{1\le j,j'\le p}$ and $\widehat{\boldsymbol{\alpha}}_p = [\hat{\alpha}_1, \ldots, \hat{\alpha}_p]^\top$ are respectively consisting of

$$\hat{h}_{jj'} = \int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} \{\widehat{\Gamma}_{XX}^{j}(\beta)(s,t)\}\{\widehat{\Gamma}_{XX}^{j'+1}(\beta)(s,t)\}\mathrm{d}s\mathrm{d}t, \quad \hat{\alpha}_j = \int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} \{\widehat{\Gamma}_{XX}(\beta)(s,t)\}\{\widehat{\Gamma}_{XX}^{j}(\beta)(s,t)\}\mathrm{d}s\mathrm{d}t.$$

Finally, given $X_0 \sim X$ and $t \in \mathbb{I}_Y$,

$$g(X_0)(t) = \mathrm{E}\{Y(t) \mid X = X_0\} = \mu_Y(t) + \mathcal{L}_{X_0}(\beta)(t) \tag{6}$$

is predicted by

$$\hat{g}_{p,\mathrm{fAPLS}}(X_0)(t) = \bar{Y}(t) + \int_{\mathbb{I}_X} X_0^{\mathrm{cent}}(s)\hat{\beta}_{p,\mathrm{fAPLS}}(s,t)\mathrm{d}s. \tag{7}$$

Matrix $\widehat{\boldsymbol{H}}$ at (5) is always invertible if we were able to work in exact arithmetic. But it is not the case for finite precision arithmetic: As p increases, the linear system from $\widehat{\Gamma}_{XX}(\beta), \ldots, \widehat{\Gamma}_{XX}^{p}(\beta)$ may be close to singular. As suggested by Delaigle and Hall [9, Section 4.2], orthonormalizing $\widehat{\Gamma}_{XX}(\beta), \ldots, \widehat{\Gamma}_{XX}^{p}(\beta)$ (with respect to $\hat{r}_{XX}$) into $\hat{\psi}_1, \ldots, \hat{\psi}_p$ (see Algorithm 1 or [23, pp. 102]), we reformulate the optimization problem at (2) into the empirical version:

$$\max_{c_1,\ldots,c_p\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{I}_Y}\left\{Y_i^{\mathrm{cent}}(t) - \sum_{j=1}^{p} c_j \int_{\mathbb{I}_X} X_i^{\mathrm{cent}}(s)\hat{\psi}_j(s,t)\mathrm{d}s\right\}^2 \mathrm{d}t. \tag{8}$$

We then reach a numerically stabilized estimator for β:

$$\tilde{\beta}_{p,\mathrm{fAPLS}} = [\hat{\psi}_1, \ldots, \hat{\psi}_p][\hat{\gamma}_1, \ldots, \hat{\gamma}_p]^\top = \sum_{j=1}^{p}\hat{\gamma}_j\hat{\psi}_j, \tag{9}$$

where the p-tuple $(\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ is the maximizer of (8), with

$$\hat{\gamma}_j = \int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} \hat{r}_{XY}(s,t)\hat{\psi}_j(s,t)\mathrm{d}s\mathrm{d}t.$$

A prediction for $g(X_0)$ at (6), alternative to $\hat{g}_{p,\mathrm{fAPLS}}(X_0)$ at (7), is thus given by

$$\tilde{g}_{p,\mathrm{fAPLS}}(X_0)(t) = \bar{Y}(t) + \int_{\mathbb{I}_X} X_0^{\mathrm{cent}}(s)\tilde{\beta}_{p,\mathrm{fAPLS}}(s,t)\mathrm{d}s. \tag{10}$$

It is worth emphasizing that, in exact arithmetic, $\hat{\beta}_{p,\mathrm{fAPLS}}$ at (5) (resp. $\hat{g}_{p,\mathrm{fAPLS}}$ at (7)) is identical to $\tilde{\beta}_{p,\mathrm{fAPLS}}$ at (9) (resp. $\tilde{g}_{p,\mathrm{fAPLS}}$ at (10)), because $\{\widehat{\Gamma}_{XX}(\beta), \ldots, \widehat{\Gamma}_{XX}^{p}(\beta)\}$ and $\{\hat{\psi}_1, \ldots, \hat{\psi}_p\}$ literally span the same space. Nevertheless, in practice $\tilde{\beta}_{p,\mathrm{fAPLS}}$ and $\tilde{g}_{p,\mathrm{fAPLS}}$ stand out due to their numerical stability for finite precision arithmetic, whereas the more explicit expressions of $\hat{\beta}_{p,\mathrm{fAPLS}}$ and $\hat{g}_{p,\mathrm{fAPLS}}$ make themselves preferred in theoretical derivations.

2.4. Asymptotic properties

Under regularity conditions, Proposition 2 (resp. Proposition 3) demonstrates the consistency in L_2 and/or supremum metric in probability of $\hat{\beta}_{p,\mathrm{fAPLS}}$ (resp. $\hat{g}_{p,\mathrm{fAPLS}}(X_0)$). In these results, we allow p to diverge as a function of n, but its rate is capped to be at most $O(n^{1/2})$ if $\|r_{XX}\|_2 < 1$ and even slower otherwise. More discussions on technical assumptions may be found in the Appendix.

Proposition 2. *Holding (C1)–(C5), as n diverges,* $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_2 = o_p(1)$*. If upgrade (C5) to (C6), then the convergence becomes uniform, i.e.,* $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_\infty = o_p(1)$*, with* $\|\cdot\|_\infty$ *denoting the supremum metric.*

Proposition 3. *Given* $X_0 \sim X$*, conditions (C1)–(C5) suffice for the zero-convergence (in probability) of* $\|\hat{g}_{p,\mathrm{fAPLS}}(X_0) - g(X_0)\|_2$ *(i.e.,* $\|\hat{g}_{p,\mathrm{fAPLS}}(X_0) - g(X_0)\|_2 = o_p(1)$*), while the uniform version (viz.* $\|\hat{g}_{p,\mathrm{fAPLS}}(X_0) - g(X_0)\|_\infty = o_p(1)$*) is entailed jointly by (C1)–(C4) and (C6)–(C7).*

Algorithm 1 Modified Gram–Schmidt orthonormalization with respect to $\hat{r}_{XX}$

```
for $j$ in $1, \ldots, p$ do
    $\hat{\psi}_j^{[1]} \leftarrow \widehat{\Gamma}_{XX}^{j}(\beta)$.
    if $j \geq 2$ then
        for $j'$ in $1, \ldots, j-1$ do
            $\hat{\psi}_j^{[j'+1]} \leftarrow \hat{\psi}_j^{[j']} - \left\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X}\int_{\mathbb{I}_X} \hat{r}_{XX}(s,s')\hat{\psi}_j^{[j']}(s,t)\hat{\psi}_{j'}(s',t)\mathrm{d}s\mathrm{d}s'\mathrm{d}t\right\}\hat{\psi}_{j'}$.
        end for
    end if
    $\hat{\psi}_j \leftarrow \left\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X}\int_{\mathbb{I}_X} \hat{r}_{XX}(s,s')\hat{\psi}_j^{[j]}(s,t)\hat{\psi}_j^{[j]}(s',t)\mathrm{d}s\mathrm{d}s'\mathrm{d}t\right\}^{-1/2}\hat{\psi}_j^{[j]}$.
end for
```

2.5. Tuning parameter

Our theoretical results in Section 2.4 are established with diverging but capped p. As well, the divergence rate of p varies with covariance functions r_{XX} and r_{XY}. It implies that the optimal p must be adaptive to data, neither too small nor too large. So, we have to tune the value of p. Except for the cross-validation, commonly-adopted tuning schemes like the generalized cross validation [7] and various information criteria demand an estimator for the degree of freedom which is absent in our context due to the intrinsic complexity of modeling. We hence take use of the five-fold cross-validation which was employed too by SigComp. In particular, p is chosen as the minimizer of

$$\mathrm{CV}(p) = \frac{1}{5}\sum_{k=1}^{5}\frac{\sum_{i\in I_k}\|Y_i - \tilde{g}_{p,\mathrm{fAPLS}}^{(-k)}(X_i)\|_2^2}{\sum_{i\in I_k}\|Y_i - \sum_{i\in I_{\mathrm{test}}\setminus I_k} Y_i/(\#I_{\mathrm{test}} - \#I_k)\|_2^2},$$

where $\{I_1, \ldots, I_5\}$ is a partition of index set for testing, say I_{test}; # represents the cardinality; $\tilde{g}_{p,\mathrm{fAPLS}}^{(-k)}(X_i)$ predicts $g(X_i)$ and is constructed from data points corresponding to $I_{\mathrm{test}} \setminus I_k$. Define the fraction of variance explained (FVE) as $\mathrm{FVE}(p) = \sum_{j=1}^{p}\lambda_{j,X}/\sum_{j=1}^{\infty}\lambda_{j,X}$. Then the search for the value of p is limited to $[1, p_{\max}]$, where $p_{\max}$ is set to be the smallest integer such that $\mathrm{FVE}(p_{\max})$ exceeds a pre-determined close-to-one threshold, e.g., 99%. This FVE criterion is commonly exploited by FPCR to determine the truncation point of Karhunen–Loève series. Since FPLS algorithms are typically more parsimonious than FPCR in terms of number of basis functions, $p_{\max}$ formed in this way tends to be large enough.

3. Simulation

In total we went through three simulation scenarios. They varied from one another in μ_Y, X, and β (as specified later) but shared the analogous setup of error term $\varepsilon = \varepsilon(t)$ which was a zero-mean Gaussian process with covariance function $\mathrm{E}\{\varepsilon(t), \varepsilon(t')\} = \sigma_\varepsilon^2\rho^{|t-t'|}$, $t, t' \in [0, 1]$ $(= \mathbb{I}_X = \mathbb{I}_Y$ in simulation). Given μ_Y, X, and β, parameters ρ and σ_ε^2 determined the signal-noise-ratio (SNR), viz. the ratio of $[\int_{\mathbb{I}_Y}\mathrm{var}\{\mathcal{L}_X(\beta)(t)\}\mathrm{d}t]^{1/2}$ to $[\int_{\mathbb{I}_Y}\mathrm{var}\{\varepsilon(t)\}\mathrm{d}t]^{1/2}$. ρ took either 0.1 (low autocorrelation of error term) or 0.9 (high autocorrelation of error term), while two levels of σ_ε^2 were set up so that SNR was moderate and fell between roughly 1 and 10; see Tables 1 and 2 for specific settings of ρ and σ_ε^2. In each scenario, we generated $n = 300$ independent and identically distributed (i.i.d.) pairs of trajectories with 80% kept for training and 20% for testing. Each curve was recorded at 101 equally spaced points $\{0, 1/101, \ldots, 100/101, 1\}$. We repeated this procedure 50 times and hence created 50 datasets, for each combination of μ_Y, X, β, ρ, and σ_ε^2. For every artificial dataset, fAPLS was compared with competitors in terms of the relative integrated squared estimation error (ReISEE) and/or the relative integrated squared prediction error (ReISPE):

$$\mathrm{ReISEE} = \frac{\|\beta - \hat{\beta}\|_2^2}{\|\beta\|_2^2}, \quad \mathrm{ReISPE} = \frac{\sum_{i\in I_{\mathrm{test}}}\|Y_i - \hat{Y}_i\|_2^2}{\sum_{i\in I_{\mathrm{test}}}\|Y_i - \sum_{i\in I_{\mathrm{train}}} Y_i/\#I_{\mathrm{train}}\|_2^2},$$

where $\hat{\beta}$ estimates β and $\hat{Y}_i$ predicts Y_i, $i \in \{1, \ldots, n\}$; I_{train} is the index set for training. We summarize ReISEEs and ReISPEs in Table 1; included in Table 2 are average values of p and total execution times.

3.1. Simulation I

Assume $\mu_Y = 0$. We took 100, 10, and 1 as the top three eigenvalues of Γ_{XX}, whereas $\lambda_{j,X} = 0$ for all $j \geq 4$. Correspondingly, the first three eigenfunctions of Γ_{XX} were respectively set to be (normalized) shifted Legendre polynomials [17, pp. 773–774] of orders 2, 3, and 4, say P_2, P_3, and P_4, viz.

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

Table 1
The averages ×100 (and standard deviations ×100) of ReISPEs and ReISEEs in numerical experiments. Values of ρ and σ_ε^2 were designed, whereas SNR was computed accordingly. Row minimums are underlined. fAPLS stood out because of its favorable mean estimation errors. As for the prediction accuracy, the outputs of all the four FPLS routes were fairly close.

	ρ	σ_ε^2	SNR	fAPLS	SigComp	NIPALS	SIMPLS
Estimation error: mean ReISEE ×100 (standard deviation ×100)							
Simulation I	0.1	1	10	33.04 (14.53)	72.63 (25.77)	73.02 (7.12)	42.18 (18.53)
		100	1	37.84 (8.80)	84.12 (21.39)	73.71 (5.96)	45.41 (14.27)
	0.9	1	11	33.20 (14.64)	72.60 (27.28)	71.82 (8.49)	42.21 (18.57)
		100	1	37.84 (10.22)	86.67 (23.26)	73.78 (6.53)	44.44 (16.77)
Simulation II	0.1	1	7	0.89 (0.62)	1.35 (0.48)	7.44 (1.33)	0.95 (0.49)
		80	1	13.01 (3.23)	13.05 (5.50)	21.00 (6.69)	12.75 (3.79)
	0.9	1	7	1.42 (2.02)	1.68 (0.75)	7.90 (1.63)	1.64 (1.72)
		80	1	17.46 (15.45)	21.03 (16.43)	26.55 (12.73)	23.46 (20.67)
Simulation III	0.1	0.05	7	2.33 (5.92)	8.98 (21.21)	4.28 (8.54)	1.99 (5.75)
		1	2	7.04 (4.25)	19.32 (26.67)	13.66 (20.08)	8.70 (15.92)
	0.9	0.05	7	6.27 (14.55)	6.62 (18.34)	6.22 (11.46)	5.33 (10.82)
		1	2	23.53 (30.15)	20.36 (27.07)	27.88 (32.14)	16.68 (26.25)
Prediction error: mean ReISPE ×100 (standard deviation ×100)							
Simulation I	0.1	1	10	1.65 (0.39)	1.74 (0.43)	1.94 (0.43)	1.80 (0.45)
		100	1	47.69 (5.37)	47.69 (5.40)	47.70 (5.37)	47.62 (5.37)
	0.9	1	11	1.56 (0.36)	1.62 (0.43)	1.85 (0.48)	1.71 (0.42)
		100	1	46.57 (6.10)	46.57 (6.09)	46.71 (6.08)	46.61 (6.07)
Simulation II	0.1	1	7	2.69 (0.50)	2.57 (0.49)	2.69 (0.50)	2.69 (0.49)
		80	1	68.88 (5.33)	68.80 (5.64)	68.89 (5.33)	68.90 (5.40)
	0.9	1	7	2.68 (0.63)	2.56 (0.62)	2.70 (0.63)	2.70 (0.63)
		80	1	68.90 (6.35)	69.06 (6.48)	68.88 (6.39)	69.14 (6.43)
Simulation III	0.1	0.05	7	28.68 (4.32)	28.62 (4.43)	28.61 (4.39)	28.56 (4.42)
		1	2	90.18 (3.48)	90.11 (3.38)	90.21 (3.59)	90.04 (3.50)
	0.9	0.05	7	28.37 (5.41)	28.11 (5.57)	28.33 (5.36)	28.37 (5.38)
		1	2	89.36 (4.40)	89.16 (4.92)	89.36 (4.48)	89.22 (4.70)
FA	–	–	–	81.10 (3.51)	81.97 (3.94)	80.99 (3.69)	80.75 (3.59)
Gait	–	–	–	62.65 (12.68)	71.23 (15.64)	69.41 (16.12)	65.04 (15.83)

As is well known, they are of unit norm and mutually orthogonal on [0, 1]. The slope function and realizations of predictor were respectively given by

$$\beta(s,t) = P_2(s)P_2(t) + P_3(s)P_3(t) + P_4(s)P_4(t), \quad X_i(s) = \zeta_{i1}P_2(s) + \zeta_{i2}P_3(s) + \zeta_{i3}P_4(s),$$

with ζ_{ij} independently distributed as $\mathcal{N}(0, \lambda_{j,X})$, $j \in \{1, \dots, 3\}$.

Simulation I was equipped with a true coefficient belonging to $\mathrm{KS}_3(\Gamma_{XX}, \beta)$ and hence was in favor of our proposal. As expected, fAPLS enjoyed lower estimation errors for this scenario; see Table 1. Nevertheless, as for prediction errors, the outputs from all the four methods were fairly comparable. We speculated that their extra estimation bias fell outside the range of Γ_{XX}, viz., $\{\Gamma_{XX}(f) \mid f \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)\}$, and hence impacted little on prediction after taking integrals. ReISEEs of all methods changed little with ρ or σ_ε^2, while their prediction accuracy was sensitive to σ_ε^2: As σ_ε^2 became smaller, prediction errors were all lowered. Meanwhile, four FPLS routes all chose around two components. The biggest advantage of fAPLS was on execution time: It ran ten times, one hundred times, and fifty times as fast as SigComp, NIPALS, and SIMPLS, respectively; see Table 2. This phenomenon was not surprising, because, compared with others, fAPLS involves fewer tuning parameters and no eigen-decomposition.

3.2. Simulation II

Define two covariance functions as follows:

$$\Sigma_1 = \Sigma_1(s,s') = \exp\{-(10|s-s'|)^2\}, \quad \Sigma_2 = \Sigma_2(s,s') = \{1 + 20|s-s'| + (20|s-s'|)^2/3\}\exp(-20|s-s'|).$$

Generating $\zeta_1, \dots, \zeta_7$ as i.i.d. realizations of the zero-mean Gaussian process with covariance function Σ_2, we constructed

$$\mu_Y(t) = \zeta_1(t), \quad \beta(s,t) = \zeta_2(s)\zeta_3(t) + \zeta_4(s)\zeta_5(t) + \zeta_6(s)\zeta_7(t).$$

Our setup was finished by sampling X_i, $i \in \{1, \dots, 300\}$, from the zero-mean Gaussian process with covariance function Σ_1. This setting appeared too in [26, Section 4.1.1].

The performance of four approaches was analogous to that in Simulation I: fAPLS stood out again in terms of estimation accuracy. Noting that numbers recorded in Table 1 were magnified 100 times, prediction errors from all routes were pretty close. Though the four methods shared the identical search scope for number of components, models from fAPLS and SigComp were typically more parsimonious (viz. of fewer numbers of components) than the remaining two; see Table 2. Especially, when there was more noise, viz. $\sigma_\varepsilon^2 = 80$, fAPLS led to the simplest model.

Table 2
Averages (and standard deviations) of component numbers and total running time (in seconds) for numerical experiments. Values of ρ and σ_ε^2 were designed, whereas SNR was computed accordingly. Row minimums are underlined. SigComp was more likely to build up concise models because it was accompanied by the smallest average number of components in most cases, while fAPLS typically took the second place. In terms of the execution time, the advantage of fAPLS was clear, except in the case of Gait data whose sample size is small.

	ρ	σ_ε^2	SNR	fAPLS	SigComp	NIPALS	SIMPLS
Number of components: average number (standard deviation)							
Simulation I	0.1	1	10	2.2 (0.4)	2.2 (0.4)	$\underline{2.1}$ (0.3)	2.2 (0.4)
		100	1	$\underline{2.1}$ (0.3)	$\underline{2.1}$ (0.4)	$\underline{2.1}$ (0.3)	$\underline{2.1}$ (0.4)
	0.9	1	11	$\underline{2.2}$ (0.4)	$\underline{2.2}$ (0.4)	$\underline{2.2}$ (0.4)	$\underline{2.2}$ (0.4)
		100	1	$\underline{2.1}$ (0.3)	$\underline{2.1}$ (0.4)	$\underline{2.1}$ (0.3)	2.2 (0.4)
Simulation II	0.1	1	7	6.2 (1.0)	$\underline{3.1}$ (0.2)	10.3 (1.1)	9.9 (1.2)
		80	1	$\underline{2.0}$ (0.3)	3.5 (1.1)	4.3 (1.3)	4.3 (1.3)
	0.9	1	7	6.5 (1.8)	$\underline{3.0}$ (0.0)	10.2 (1.4)	10.0 (1.7)
		80	1	$\underline{2.6}$ (1.9)	4.7 (1.3)	4.6 (1.8)	5.2 (2.7)
Simulation III	0.1	0.05	7	1.9 (0.5)	$\underline{1.4}$ (0.6)	2.2 (0.6)	1.8 (0.7)
		1	2	$\underline{1.3}$ (0.5)	$\underline{1.3}$ (0.5)	2.1 (0.4)	1.4 (0.6)
	0.9	0.05	7	2.1 (0.7)	$\underline{1.5}$ (0.8)	2.3 (0.5)	2.0 (0.9)
		1	2	1.7 (0.9)	$\underline{1.4}$ (0.7)	2.3 (0.6)	1.5 (0.8)
FA	–	–	–	4.1 (0.8)	$\underline{3.3}$ (0.7)	4.6 (0.5)	4.8 (0.4)
Gait	–	–	–	$\underline{2.7}$ (0.9)	4.1 (1.4)	5.0 (1.6)	4.9 (1.5)
Total running time in seconds for all replicates/splits							
Simulation I	0.1	1	10	$\underline{3.0}$	41.8	350.3	150.2
		100	1	$\underline{3.4}$	41.4	358.5	149.6
	0.9	1	11	$\underline{3.4}$	41.2	327.2	148.7
		100	1	$\underline{3.4}$	40.1	327.3	147.5
Simulation II	0.1	1	7	$\underline{36.1}$	48.6	398.5	241.7
		80	1	$\underline{35.8}$	51.4	416.3	242.2
	0.9	1	7	$\underline{35.7}$	50.1	375.0	242.3
		80	1	$\underline{36.4}$	49.3	377.7	242.9
Simulation III	0.1	0.05	7	$\underline{6.9}$	41.6	327.9	163.2
		1	2	$\underline{6.4}$	44.0	336.8	164.4
	0.9	0.05	7	$\underline{6.4}$	41.0	272.0	162.8
		1	2	$\underline{6.4}$	41.6	275.7	163.6
FA	–	–	–	$\underline{5.4}$	71.8	266.4	101.6
Gait	–	–	–	$\underline{1.6}$	1.7	29.2	23.5

3.3. Simulation III

We considered a setting similar to the ones in multiple works [13, Section 4.1; 21, Section 4.1; 26, Section 4.1.2]:

$$\mu_Y(t) = 2\exp\{-(t-1)^2\}, \quad \beta(s,t) = \sin(\pi s)\cos(2\pi t), \quad X_i(s) = \sum_{m=1}^{10} \frac{1}{m^2}\{\zeta_{i1m}\sin(m\pi s) + \zeta_{i2m}\cos(m\pi s)\},$$

where ζ_{ijm}, $i \in \{1,\ldots,300\}$, $j \in \{1,2\}$, $m \in \{1,\ldots,10\}$, are all i.i.d. standard normal.

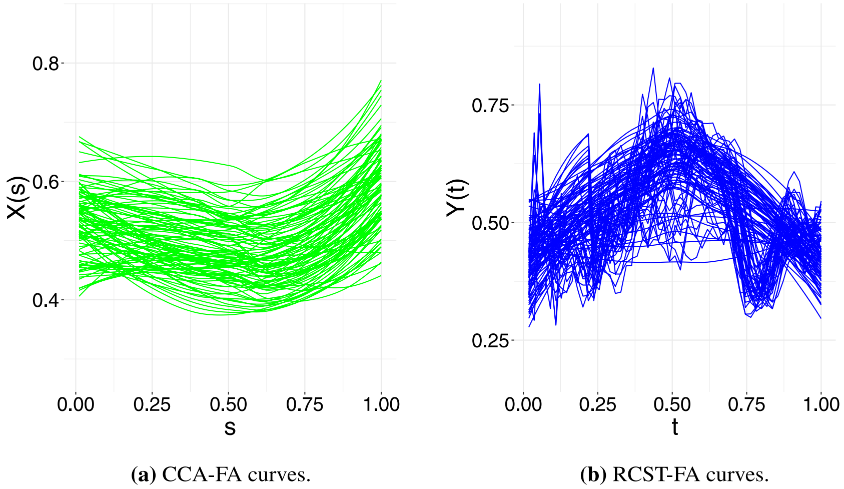
This was a scenario where SIMPLS seemed to work generally better than others in estimation. Noticing again that errors in Table 1 were 100 times as great as original numbers, the estimation performance of fAPLS was still comparable, except in the case with a large autocorrelated error term ($\rho = 0.9$) and a small SNR ($= 1$). As for prediction, though fAPLS did not correspond to the minimum error for any setting, its outputs remained close to corresponding minimums; see Table 1. Number of components picked up by fAPLS was in average about 0.5 more than those from SigComp but was of the same level of those from NIPALS and SIMPLS; see Table 2.

4. Application

We applied the four approaches, viz. fAPLS, SigComp, NIPALS, and SIMPLS, to two real-world datasets. Their predictive performance was evaluated again by ReISPE. We generated 50 ReISPE values for each approach and each dataset, after repeating the following random split 50 times: Around 20% of all the pairs of trajectories were retained for testing and the remaining for training.

4.1. Fractional anisotropy (FA)

As a magnetic resonance imaging technique, the diffusion tensor imaging (DTI) tractography may measure the diffusivity of water. In the brain, water diffuses anisotropically along white matter tracts but isotropically elsewhere.



(a) CCA-FA curves.

(b) RCST-FA curves.

Fig. 1. 100 pairs of CCA- and RCST-FA curves collected in a clinical trial. Each CCA- (resp. RCST-) FA curve consists of FA values measured along CCA (resp. RCST). Small FA values imply the demyelination of white matter tracts and become indicators of damages on the central nervous system. An investigation into the spatial association of paired CCA- and RCST-FA trajectories may contribute to the diagnosis of corresponding diseases.

One widely-used diffusivity measure given by DTI is the fractional anisotropy (FA) which ranges from zero to one: FA $= 1$ means a diffusion occurring only in the direction of white matter tracts, while the zero FA corresponds to the isotropic diffusion. Small FA values accordingly may imply a loss of myelin (viz. demyelination) of white matter tracts. In this way, DTI is powerful in characterizing microstructural changes for neuropathology [2], e.g., diagnosing the multiple sclerosis (MS). MS is an immune disorder jeopardizing the central nervous system; the immune system of an MS patient attacks myelin in the brain, leading to demyelination. Symptoms of MS vary a lot on a case-by-case basis and depend on which nerves are affected. Severe MS can even cause lifelong disability.

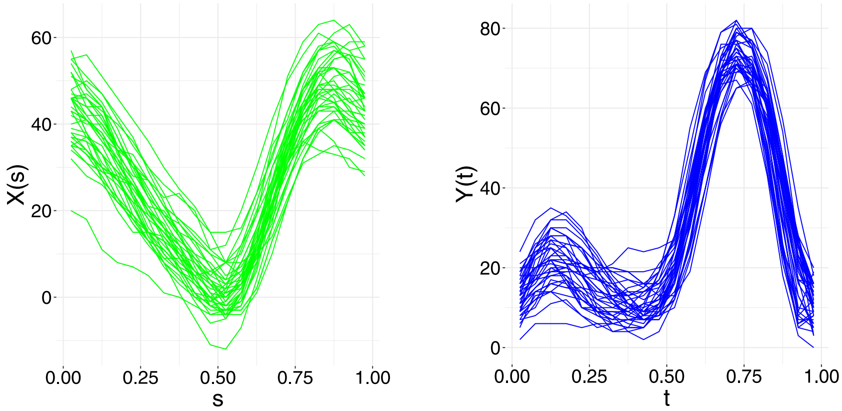
Initially collected by the Johns Hopkins University and Kennedy-Krieger Institute, dataset `DTI` in `R` package `refund` [14] contains FA values along two sorts of white matter tracts, viz. the corpus callosum (CCA) and right corticospinal tract (RCST) for participants of a clinical trial on MS. At each visit, 93 (resp. 55) FA values along CCA (resp. RCST) were measured for each participant. These FA values formed the so-called CCA- (resp. RCST-) FA curves. There were altogether 382 pairs of CCA- and RCST-FA curves whose spatial association was previously studied by, e.g., [21,26], through FoFR. We took CCA-FA curves (Fig. 1a) as predictors and RCST-FA curves (Fig. 1b) as responses, imputing missingness through local polynomial regression (with the help of `R` package `spatialEco` [10]). As illustrated at the second line from the bottom of Table 1, SIMPLS took the minimum mean ReISPE value, whereas ReISPEs from fAPLS were of the lowest variation. But, in fact, the difference in ReISPE was extremely close. fAPLS stood out again in its time consumption.

4.2. Boys' gait

Human motion has been a research topic for over two thousands years, dating back to the period of ancient Greece. It is believed that the gait of an individual is an indicator of his/her neuromuscular development or impairment [29]. A sound model on people's gait would help to define the normal and abnormal walking and, as well, to discover causes and deviations of abnormality. Dataset `gait` in `R` package `fda` [33] was collected at the Motion Analysis Laboratory at Children's Hospital, San Diego, recording hip and knee angles over one gait cycle for 39 boys. Observed at 20 time points, one gait cycle began and ended at the time point when the heel touched the ground. Lian [25] tried to summarize the gait by depicting how these two joints interacted; they applied FoFR and regressed the knee angle curves on hip angle curves. Analogously, we took the hip angle curves (Fig. 2a) as predictors and knee angle curves (Fig. 2b) as responses. In terms of both average and variance of ReISPE values, fAPLS was more accurate than competitors (see Table 1) and also built up the most parsimonious model (see Table 2). For this dataset, it was not apparent to see the advantage of fAPLS in execution time: SigComp ran almost as fast as fAPLS; see the last row of Table 2. We guessed the small sample size ($= 39$) reduced the computational burden of SigComp.

5. Conclusion and discussion

Fitting FoFR, we suggest fAPLS, an FPLS route via Krylov subspace. fAPLS estimator (5) owns a concise and explicit expression. Meanwhile, we introduce an alternative but equivalent version (9), stabilizing numerical outputs. Resulting in a competitive accuracy in both estimation and prediction, fAPLS consumes less running time than other FPLS routes. It is out of question that our proposal is far from perfect. One major concern lies at the value of p, viz. the dimension



(a) Hip angle curves.

(b) Knee angle curves.

Fig. 2. Curves of 39 boys' hip and knee angles within one gait cycle that begins and ends when the heel touched the ground. For each boy, his hip (resp. knee) angle curve is drawn by connecting values of hip (resp. knee) angles observed at 20 time points. A sound model between paired hip and knee angle trajectories would reflect boys' physical growth.

of Krylov subspace: When p is small, e.g., not larger than 10, everything should work out well; otherwise, the space spanned by orthonormalized basis functions, viz. $\text{span}\{\hat{\psi}_1, \ldots, \hat{\psi}_p\}$, is possible to be distant from its theoretical target $\text{span}\{\mathit{\Gamma}_{XX}(\beta), \ldots, \mathit{\Gamma}_{XX}^p(\beta)\}$. The source of this numerical error is twofold: the accumulated error in the recursive estimation for $\mathit{\Gamma}_{XX}^j(\beta)$, $j \in \{1, \ldots, p\}$, and the bias induced by modified Gram–Schmidt orthonormalization. Accordingly, one needs to reconsider the value of p, in case the five-fold cross-validation in Section 2.5 recommends a large one.

Curves X_i and Y_i are expected to be observed so densely that we may enjoy small errors in presmoothing. However, fAPLS estimation is still doable without fulfilling this denseness assumption: As long as covariance functions r_{XX} and r_{XY} are estimated, it suffices to obtain fAPLS estimators at (5) and (9). Even for sparsely observed trajectories, one may estimate both r_{XX} and r_{XY} following the local linear smoothing (e.g., [24,42]) or spline smoothing (e.g., [41]). These techniques are helpful too in accommodating measurement errors; refer to [44] for the scalar-on-function regression with contaminated observations. In the case of geographic data, the spatial correlation (i.e., X_i and $X_{i'}$, $i \neq i'$, no longer mutually independent) leads to a potential inconsistency of FPLS estimators; see [34, Theorem 1] for this issue in the multivariate context. A naive correction, transplanted from [34, Section 4.1], is to instead implement the regression on transformed observations (X_i^*, Y_i^*), $i \in \{1, \ldots, n\}$, such that, for all $(s, t) \in \mathbb{I}_X \times \mathbb{I}_Y$, $[X_1^*(s), \ldots, X_n^*(s)]^\top = \boldsymbol{V}_{XX}^{-1/2}(s)[X_1(s), \ldots, X_n(s)]^\top$ and $[Y_1^*(t), \ldots, Y_n^*(t)]^\top = \boldsymbol{V}_{YY}^{-1/2}(t)[Y_1(t), \ldots, Y_n(t)]^\top$, with matrices $\boldsymbol{V}_{XX}(s) = [\text{cov}\{X_i(s), X_{i'}(s)\}]_{n\times n}$ and $\boldsymbol{V}_{YY}(t) = [\text{cov}\{Y_i(t), Y_{i'}(t)\}]_{n\times n}$. But it is even challenging to recover $\boldsymbol{V}_{XX}$ and $\boldsymbol{V}_{YY}$ sufficiently accurately without specifying the dependence structure, since there is only one observation for each i. Alternatively and more practically, one can target at correcting naive $\hat{r}_{XX}$ and $\hat{r}_{XY}$ for dependent subjects: Paul and Peng [30] offered a solution to this point.

Though all X_i's (resp. Y_i's) are assumed to share the identical time domain $\mathbb{I}_X$ (resp. $\mathbb{I}_Y$), one may encounter the phase variation (also known as time variation, misalignment, etc.), i.e., observed curves suffer the lateral displacement and/or deformation. For instance, replications of handwriting are likely to take different lengths of time; i.e., while each time the handwriting is initiated at time 0, the end point differ from replication to replication [28, Section 1.2]. In such cases, it is heuristic to register curves first, i.e., transform the arguments of X_i and Y_i in pre-processing. Specifically, for curves X_i (resp. Y_i), introduce smooth and strictly increasing warping functions $wp_{X,i}$ (resp. $wp_{Y,i}$). The phase variation is anticipated to be removed from the further analysis by taking use of registered cures $\widetilde{X}_i(s) = X_i\{wp_{X,i}(s)\}$ and $\widetilde{Y}_i(t) = Y_i\{wp_{X,i}(t)\}$ rather than original ones. If there exist landmarks of interest, e.g., certain peaks and/or troughs, throughout the data, this curve registration may be carried out so that these landmarks occur roughly at the same time. Details on registration are available at, e.g., [32, Chapter 7] and more recent [28].

fAPLS has got a naive extension to multiple functional covariates, i.e., associated with each realization $Y_i \sim Y$, there are $m > 1$ functional covariates, say $X_{ij} \sim X_{\cdot j}$, $j \in \{1, \ldots, m\}$, and correspondingly m coefficient functions $\beta^{(j)}$, $j \in \{1, \ldots, m\}$. 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 (2), an ad hoc estimator for m-tuple $(\beta^{(1)}, \ldots, \beta^{(m)})$ is thus

$$(\hat{\beta}^{(1)}_{\mathrm{fAPLS}}, \ldots, \hat{\beta}^{(m)}_{\mathrm{fAPLS}}) = \mathop{\arg\min}_{\theta^{(j)} \in \mathrm{KS}_p(\widehat{\Gamma}_{X_{\cdot j}X_{\cdot j}}, \beta^{(j)}),\ 1 \le j \le m} \frac{1}{m} \sum_{i=1}^{m} \int_{\mathbb{I}_Y} \left[Y_i(t) - \bar{Y}_i(t) - \sum_{j=1}^{m} \int_{\mathbb{I}_{X_{\cdot j}}} \{X_{ij}(s) - \bar{X}_{\cdot j}(s)\} \theta^{(j)}(s,t) \mathrm{d}s \right]^2 \mathrm{d}t,$$

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

It appears that fAPLS merely works for linear models, because, without the linear model assumption (1), neither Proposition 1 nor the consistency holds. Fortunately, there is a promising way of applying fAPLS (and other FPLS algorithms) to nonlinear modeling: Recursively linearize the procedure of maximizing likelihood and then embed FPLS algorithms into the iteratively reweighted least squares [15]. This idea has been successfully applied by [1,38].

CRediT authorship contribution statement

Zhiyang Zhou: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization.

Acknowledgments

The author is grateful to Editor-in-Chief, Executive Editor, Associate Editor, and anonymous reviewers. Their insightful comments dramatically improved our original manuscript. Special thanks go to Professor Richard A. Lockhart at the Simon Fraser University for his constructive suggestions. This work is financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix

We refer to R package FRegSigCom [27] for SigComp and GitHub (https://github.com/hanshang/FPLSR; accessed on May 17, 2021) for (functional) NIAPLS and SIMPLS. Our R codes for fAPLS are publicly available too at GitHub (https://github.com/ZhiyangGeeZhou/fAPLS; accessed on May 17, 2021). We run the code trunks on a laptop with AMD® Ryzen™ 5 4500U @6 × 2.38 GHz and 16 GB RAM.

Our technical assumptions are summarized as below.

(C1) $\sum_{j,j'=1}^{\infty} \lambda_{j,X}^{-2} \left\{ \int_{\mathbb{I}_Y} \int_{\mathbb{I}_X} \phi_{j,X}(s) r_{XY}(s,t) \phi_{j',Y}(t) \mathrm{d}s\mathrm{d}t \right\}^2 < \infty$. Moreover, β belongs to $\mathrm{range}(\Gamma_{XX}) = \{\Gamma_{XX}(f) \mid f \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)\}$.

(C2) $\mathrm{E}(\|X\|_2^4) < \infty$ for all $t \in \mathbb{I}_Y$.

(C3) As $n \to \infty$, $p = p(n) = O(n^{1/2})$.

(C4) Let $\mathbb{I}_X = [0, 1]$. Both $\|\xi_{XX}\|_{\infty,2}$ and $\|\eta_{XX}\|_{\infty,2}$ are of order $O_p(1)$ as $n \to \infty$, with ξ_{XX} and η_{XX} defined as in Lemma A.1 and $\|\cdot\|_{\infty,2}$ defined such that $\|f\|_{\infty,2} = \sup_{s \in \mathbb{I}_X} \left\{ \int_{\mathbb{I}_X} f^2(s,t) \mathrm{d}t \right\}^{1/2}$ for $f \in L_2(\mathbb{I}_X \times \mathbb{I}_X)$.

(C5) Additional requirements on p vary with the magnitude of $\|r_{XX}\|_2$; they also depend on τ_p, the smallest eigenvalue of $\boldsymbol{H}_p$.

- If $\|r_{XX}\|_2 \ge 1$, then, as $n \to \infty$, $n^{-1}\tau_p^{-2} p^4 \|r_{XX}\|_2^{4p} \max(1, \tau_p^{-2} p^2 \|r_{XX}\|_2^{4p})$ and $n^{-1}\tau_p^{-3} p^5 \|r_{XX}\|_2^{6p}$ are both of order $o(1)$;
- if $\|r_{XX}\|_2 < 1$, then $n^{-1}\tau_p^{-4} = o(1)$ as n diverges.

(C6) Keep everything in (C5) but substitute $\|r_{XX}\|_\infty$ for $\|r_{XX}\|_2$. Meanwhile, require that $\|\beta_{p,\mathrm{fAPLS}} - \beta\|_\infty = o(1)$ as p diverges, viz. an enhanced version of Proposition 1.

(C7) Stochastic process Y is “eventually totally bounded in mean” (as defined by Hoffmann-Jørgensen [18, (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 $\mathrm{Cover}(\mathbb{T})$, for each set $\mathbb{A} \in \mathrm{Cover}(\mathbb{T})$, such that $\inf_{n \in \mathbb{Z}^+} n^{-1} \mathrm{E} \{\sup_{t,t' \in \mathbb{A}} |Y(t) - Y(t')|\} < \epsilon$.

Introducing (C1), He et al. [16, Theorem 2.3] confirmed the identifiability of β and derived its closed form (A.2). (C1) was also the foundation of [43]. Assumptions (C2)–(C4) are prerequisites for the convergence of $\widehat{\Gamma}_{XX}^{j}(\beta)$ $(= \widehat{\Gamma}_{XX}^{j-1}(\hat{r}_{XY}))$ which is uniform in $j \ge 1$. One may feel unclear about the technical conditions stated in (C5) for the scenario of $\|r_{XX}\|_2 \ge 1$: virtually a special case 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 $\|r_{XX}\|_2 \ge 1$ than in the case of $\|r_{XX}\|_2 < 1$ (for the latter case p is allowed to diverge at the rate of $O(n^{1/2})$); that is why Delaigle and Hall [9] suggested changing the scale on which X is measured. (C6) is stronger than (C5), enabling us to consider the L_∞-convergence. At last, we add (C7) as a prerequisite for the uniform law of large numbers for $\{Y_i \mid i \ge 1\}$.

Lemma A.1. *For each $(s, s', t) \in \mathbb{I}_X \times \mathbb{I}_X \times \mathbb{I}_Y$,*

$$\hat{r}_{XX}(s, s') = r_{XX}(s, s') + n^{-1/2}\xi_{XX}(s, s') + n^{-1}\eta_{XX}(s, s'), \quad \hat{r}_{XY}(s, t) = r_{XY}(s, t) + n^{-1/2}\xi_{XY}(s, t) + n^{-1}\eta_{XY}(s, t),$$

where, with identity operator $I : \mathbb{R} \to \mathbb{R}$,

$$\xi_{XX}(s, s') = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I - \mathrm{E})[\{X_i(s) - \mu_X(s)\}\{X_i(s') - \mu_X(s')\}], \quad \eta_{XX}(s, s') = -n\{\bar{X}(s) - \mu_X(s)\}\{\bar{X}(s') - \mu_X(s')\},$$

$$\xi_{XY}(s, t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I - \mathrm{E})[\{X_i(s) - \mu_X(s)\}\{Y_i(t) - \mu_Y(t)\}], \quad \eta_{XY}(s, t) = -n\{\bar{X}(s) - \mu_X(s)\}\{\bar{Y}(t) - \mu_Y(t)\},$$

and $\|\xi_{XX}\|_2$, $\|\eta_{XX}\|_2$, $\|\xi_{XY}\|_2$, and $\|\eta_{XY}\|_2$ all equal $O_p(1)$ as n diverges.

Proof of Lemma A.1. It is an immediate implication of Delaigle and Hall [9, (5.1)]. □

Lemma A.2. *Assume (C1) and (C2) and that there is $C > 0$ such that, for all n, we have $p \le Cn^{-1/2}$ (i.e., condition (C3)). Then, for each $\epsilon > 0$, there are positive C_1, C_2, and n_0 such that, for each $n > n_0$,*

$$\Pr\left[\bigcap_{j=1}^{p}\left\{\|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 \le n^{-1/2}\|r_{XX}\|_2^{j-1}\{C_1 + C_2(j-1)\}\right\}\right] \ge 1 - \epsilon.$$

Assuming one more condition (C4),

$$\Pr\left[\bigcap_{j=1}^{p}\left\{\|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_\infty \le n^{-1/2}\|r_{XX}\|_\infty^{i-1}\{C_1 + C_2(j-1)\}\right\}\right] \ge 1 - \epsilon.$$

Proof of Lemma A.2. Since $\Gamma_{XX}(\beta) = r_{XY}$ and $\widehat{\Gamma}_{XX}(\beta) = \hat{r}_{XY}$, Lemma A.2 is simply implied by Lemma A.1 when $p = 1$. For integer $j \ge 2$ and each $(s, s', t) \in \mathbb{I}_X \times \mathbb{I}_X \times \mathbb{I}_Y$,

$$\begin{aligned}\left|\widehat{\Gamma}_{XX}^{j}(\beta)(s,t) - \Gamma_{XX}^{j}(\beta)(s,t)\right| &= \left|\widehat{\Gamma}_{XX}\{\widehat{\Gamma}_{XX}^{j-1}(\beta) - \Gamma_{XX}^{j-1}(\beta)\}(s,t) + \{(\widehat{\Gamma}_{XX} - \Gamma_{XX})\Gamma_{XX}^{j-1}(\beta)\}(s,t)\right| \\ &\le \left\{\int_{\mathbb{I}_X}\hat{r}_{XX}^2(s,s')\mathrm{d}w\right\}^{1/2}\left[\int_{\mathbb{I}_X}\{\widehat{\Gamma}_{XX}^{j-1}(\beta) - \Gamma_{XX}^{j-1}(\beta)\}(s',t)\mathrm{d}w\right]^{1/2} \\ &\quad + \left[\int_{\mathbb{I}_X}\{\hat{r}_{XX}(s,s') - r_{XX}(s,s')\}^2\mathrm{d}s'\right]^{1/2}\left\{\int_{\mathbb{I}_X}\Gamma_{XX}^{j-1}(\beta)(s',t)\mathrm{d}s'\right\}^{1/2}.\end{aligned}$$

It implies that, by the triangle inequality,

$$\|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 \le \|\hat{r}_{XX}\|_2\|\widehat{\Gamma}_{XX}^{j-1}(\beta) - \Gamma_{XX}^{j-1}(\beta)\|_2 + \|\hat{r}_{XX} - r_{XX}\|_2\|\Gamma_{XX}^{j-1}(\beta)\|_2.$$

On iteration it gives that

$$\|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 \le \|\hat{r}_{XX}\|_2^{j-1}\|\widehat{\Gamma}_{XX}(\beta) - \Gamma_{XX}(\beta)\|_2 + \|\hat{r}_{XX} - r_{XX}\|_2\sum_{j'=1}^{j-1}\|\hat{r}_{XX}\|_2^{j-j'-1}\|\Gamma_{XX}^{j'}(\beta)\|_2. \tag{A.1}$$

For each $\epsilon > 0$, there is $n_0 > 0$ such that, for all $n > n_0$, we have

$$\begin{aligned}1 - \epsilon/2 &\le \Pr(\|\hat{r}_{XX} - r_{XX}\|_2 \le C_0 n^{-1/2}) \le \Pr(\|\hat{r}_{XX}\|_2 \le \|r_{XX}\|_2 + C_0 n^{-1/2}), \\ 1 - \epsilon/2 &\le \Pr(\|\hat{r}_{XY} - r_{XY}\|_2 \le C_0 n^{-1/2}),\end{aligned}$$

with constant $C_0 > 0$, by Lemma A.1. It follows (A.1) that

$$\begin{aligned}
&1-\epsilon\\
&\le \Pr\left(\bigcap_{j=1}^{p}\left[\|(\widehat{\Gamma}^j_{XX}-\Gamma^j_{XX})(\beta)\|_2 \le C_0 n^{-1/2}\left\{(\|r_{XX}\|_2 + C_0 n^{-1/2})^{j-1} + \sum_{j'=1}^{j-1}\|r_{XX}\|_2^{j'}\|\beta\|_2(\|r_{XX}\|_2 + C_0 n^{-1/2})^{j-j'-1}\right\}\right]\right)\\
&\le \Pr\left(\bigcap_{j=1}^{p}\left[\|(\widehat{\Gamma}^j_{XX}-\Gamma^j_{XX})(\beta)\|_2 \le C_0 n^{-1/2}\|r_{XX}\|_2^{j-1}\left\{\left(1+\frac{C_0 n^{-1/2}}{\|r_{XX}\|_2}\right)^{j-1} + \|\beta\|_2\sum_{j'=1}^{j-1}\left(1+\frac{C_0 n^{-1/2}}{\|r_{XX}\|_2}\right)^{j-j'-1}\right\}\right]\right)\\
&\le \Pr\left(\bigcap_{j=1}^{p}\left[\|\widehat{\Gamma}^j_{XX}(\beta)-\Gamma^j_{XX}(\beta)\|_2 \le n^{-1/2}\|r_{XX}\|_2^{j-1}\{C_1 + C_2(j-1)\}\right]\right), \quad (\text{since } p \le Cn^{1/2})
\end{aligned}$$

where $C_1 = C_0\exp(CC_0/\|r_{XX}\|_2)$ and $C_2 = \|\beta\|_2 C_1$.

Suppose (C4) holds. Similar to (A.1),

$$\begin{aligned}
\|\widehat{\Gamma}^j_{XX}(\beta)-\Gamma^j_{XX}(\beta)\|_\infty &\le \|\hat{r}_{XX}\|_\infty^{j-1}\|\widehat{\Gamma}_{XX}(\beta)-\Gamma_{XX}(\beta)\|_\infty + \|\hat{r}_{XX}-r_{XX}\|_\infty\sum_{j'=1}^{j-1}\|\hat{r}_{XX}\|_\infty^{j-j'-1}\|\Gamma^{j'}_{XX}(\beta)\|_\infty\\
&\le \|\hat{r}_{XX}\|_\infty^{j-1}\|\widehat{\Gamma}_{XX}(\beta)-\Gamma_{XX}(\beta)\|_\infty + \|\hat{r}_{XX}-r_{XX}\|_\infty\sum_{j'=1}^{j-1}\|\hat{r}_{XX}\|_\infty^{j-j'-1}\|r_{XX}\|_\infty^{j'}\|\beta\|_\infty.
\end{aligned}$$

Mimicking the argument above for the L_2 sense, one obtains that

$$\Pr\left(\bigcap_{j=1}^{p}\left[\|\widehat{\Gamma}^j_{XX}(\beta)-\Gamma^j_{XX}(\beta)\|_\infty \le n^{-1/2}\|r_{XX}\|_\infty^{j-1}\{C_1 + C_2(j-1)\}\right]\right) \ge 1-\epsilon,$$

with, at this time, $C_1 = C_0\exp(CC_0/\|r_{XX}\|_\infty)$ and $C_2 = \|\beta\|_\infty C_1$. The identity that $\|\beta\|_\infty < \infty$ originates from the continuity of eigenfunctions $\phi_{i,X}$'s and $\phi_{i,Y}$'s (refer to the Mercer's theorem). □

Proof of Proposition 1. From condition (C1), He et al. [16, Theorem 2.3] derived the unique closed-form of β. In particular, for each $(s,t) \in \mathbb{I}_X \times \mathbb{I}_Y$,

$$\beta(s,t) = \Gamma_{XX}^{-1}(r_{XY})(s,t) = \sum_{j,j'=1}^{\infty}\frac{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X}\phi_{j,X}(s)r_{XY}(s,t)\phi_{j',Y}(t)\mathrm{d}s\mathrm{d}t}{\lambda_{j,X}}\phi_{j,X}(s)\phi_{j',Y}(t). \tag{A.2}$$

Introduce $\beta_p \in L_2(\mathbb{I}_X \times \mathbb{I}_Y)$ such that

$$\beta_p(s,t) = \sum_{j=1}^{p}\frac{\phi_{j,X}(s)}{\lambda_{j,X}}\int_{\mathbb{I}_X}\phi_{j,X}(s')r_{XY}(s',t)\mathrm{d}s'.$$

It follows that

$$\Gamma_{XX}(\beta_p)(s,t) = \sum_{j=1}^{p}\phi_{j,X}(s)\int_{\mathbb{I}_X}\phi_{j,X}(s')r_{XY}(s',t)\mathrm{d}w.$$

Now

$$[(\lambda_{1,X}I - \Gamma_{XX})\circ\cdots\circ(\lambda_{p,X}I - \Gamma_{XX})](\beta_p) = 0$$

in which the left-hand side equals $\sum_{i=j}^{p} a_j\Gamma^j_{XX}(\beta_p)$ with $a_0 = \prod_{j=1}^{p}\lambda_{j,X} > 0$. Therefore,

$$\beta_p = -\sum_{j=1}^{p}\frac{a_j}{a_0}\Gamma^j_{XX}(\beta_p).$$

Denote by P_p : range$(\Gamma_{XX}) \to$ range(Γ_{XX}) the operator that projects elements in range(Γ_{XX}) to span$\{f_{jj'} \in L_2(\mathbb{I}_X \times \mathbb{I}_Y) \mid f_{jj'}(s,t) = \phi_{j,X}(s)\phi_{j',Y}(t), 1 \le j \le p, j' \ge 1\}$. Thus $\beta_p = P_p(\beta)$. Since $\Gamma^j_{XX}(\beta_p) = P_p[\Gamma^j_{XX}(\beta)]$, one has

$$P_p\left[\beta + \sum_{j=1}^{p}\frac{a_j}{a_0}\Gamma^j_{XX}(\beta)\right] = 0,$$

implying that, for all p,

$$P_p(\beta) \in \{P_p(f) \mid f \in \overline{\mathrm{KS}_\infty(\Gamma_{XX}, \beta)}\}.$$

Taking limits as $p \to \infty$ on both sides of the above formula, we obtain $\beta \in \overline{\mathrm{KS}_\infty(\Gamma_{XX}, \beta)}$ and accomplish the proof. □

Proof of Proposition 2. Recall $\beta_{p,\mathrm{fAPLS}}$ (2) and $\hat{\beta}_{p,\mathrm{fAPLS}}$ (5) and notations in defining them. The Cauchy–Schwarz inequality implies that

$$\begin{aligned}|\hat{h}_{jj'} - h_{jj'}| &\le \|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 \|\widehat{\Gamma}_{XX}^{j'+1}(\beta)\|_2 + \|\widehat{\Gamma}_{XX}^{j'+1}(\beta) - \Gamma_{XX}^{j'+1}(\beta)\|_2 \|\Gamma_{XX}^{j}(\beta)\|_2 \\ &\le \|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 \|\hat{r}_{XX}\|_2^{j+1} \|\beta\|_2 + \|\widehat{\Gamma}_{XX}^{j'+1}(\beta) - \Gamma_{XX}^{j'+1}(\beta)\|_2 \|r_{XX}\|_2^{j} \|\beta\|_2.\end{aligned}$$

By Lemmas A.1 and A.2, for each $\epsilon > 0$ and $p \le Cn^{1/2}$, there are positive n_0, C_3 and C_4 such that, for all $n > n_0$,

$$\begin{aligned}&1-\epsilon \\ &\le \Pr\left[\bigcap_{j,j'=1}^{p} \left\{|\hat{h}_{jj'} - h_{jj'}| \le \|\widehat{\Gamma}_{XX}^{j}(\beta) - \Gamma_{XX}^{j}(\beta)\|_2 (\|r_{XX}\|_2 + C_0 n^{-1/2})^{j'+1} \|\beta\|_2 + \|\widehat{\Gamma}_{XX}^{j'+1}(\beta) - \Gamma_{XX}^{j'+1}(\beta)\|_2 \|r_{XX}\|_2^{j} \|\beta\|_2\right\}\right] \\ &\le \Pr\left(\bigcap_{j,j'=1}^{p} \left[|\hat{h}_{jj'} - h_{jj'}| \le \ n^{-1/2} \|r_{XX}\|_2^{i+j'} \{C_3 \max(j, j') + C_4\}\right]\right).\end{aligned}$$

Thus

$$\begin{aligned}\|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2^2 &\le \sum_{j,j'=1}^{p} |\hat{h}_{jj'} - h_{jj'}|^2 \\ &= O_p\left(\frac{1}{n} \sum_{j,j'=1}^{p} \|r_{XX}\|_2^{2j+2j'}\right) + O_p\left\{\frac{1}{n} \sum_{j,j'=1}^{p} \max(j^2, j'^2) \|r_{XX}\|_2^{2j+2j'}\right\} \\ &= \begin{cases} O_p(n^{-1} p^4 \|r_{XX}\|_2^{4p}), & \text{if } \|r_{XX}\|_2 \ge 1, \\ O_p(n^{-1}), & \text{if } \|r_{XX}\|_2 < 1. \end{cases}\end{aligned} \tag{A.3}$$

Here $\|\cdot\|_2$ is abused for the matrix norm induced by the Euclidean norm, i.e., for arbitrary $\boldsymbol{A} \in \mathbb{R}^{p \times p'}$ and $\boldsymbol{b} \in \mathbb{R}^{p' \times 1}$, $\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{b}:\|\boldsymbol{b}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{b}\|_2$ is actually the largest eigenvalue of $\boldsymbol{A}$. It reduces to the Euclidean norm for vectors. It is analogous to (A.3) to deduce that

$$\|\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2^2 = \sum_{j=1}^{p} |\hat{\alpha}_j - \alpha_j|^2 = \begin{cases} O_p(n^{-1} p^3 \|r_{XX}\|_2^{2p}), & \text{if } \|r_{XX}\|_2 \ge 1, \\ O_p(n^{-1}), & \text{if } \|r_{XX}\|_2 < 1. \end{cases} \tag{A.4}$$

Denote by τ_p the smallest eigenvalue of $\boldsymbol{H}_p$. Noting that $\|\boldsymbol{H}_p^{-1}\|_2 = \tau_p^{-1}$, for $p \le Cn^{1/2}$,

$$\|(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\boldsymbol{H}_p^{-1}\|_2 \le \tau_p^{-1} \|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2 = \begin{cases} O_p(n^{-1/2} \tau_p^{-1} p^2 \|r_{XX}\|_2^{2p}), & \text{if } \|r_{XX}\|_2 \ge 1, \\ O_p(n^{-1/2} \tau_p^{-1}), & \text{if } \|r_{XX}\|_2 < 1. \end{cases}$$

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,

$$\|\boldsymbol{M}_p\|_2 \le \|\boldsymbol{I} + \boldsymbol{H}^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2^{-1} \|\boldsymbol{H}^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2^2 \le (1-\rho)^{-1} \tau_p^{-2} \|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2^2,$$

provided that $\tau_p^{-1} \|\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p\|_2 \le \rho < 1$ (refer to Delaigle and Hall [9, (7.18)]). Revealed by the identity that $\widehat{\boldsymbol{H}}_p^{-1} = \{\boldsymbol{I} + \boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\}^{-1} \boldsymbol{H}_p^{-1}$,

$$\begin{aligned}\|\widehat{\boldsymbol{H}}_p^{-1} - \boldsymbol{H}_p^{-1}\|_2 &\le \{\|\boldsymbol{H}_p^{-1}(\widehat{\boldsymbol{H}}_p - \boldsymbol{H}_p)\|_2 + \|\boldsymbol{M}_p\|_2\} \|\boldsymbol{H}_p^{-1}\|_2 \\ &= \begin{cases} O_p(n^{-1/2} \tau_p^{-2} p^2 \|r_{XX}\|_2^{2p}) + O_p(n^{-1} \tau_p^{-3} p^4 \|r_{XX}\|_2^{4p}), & \text{if } \|r_{XX}\|_2 \ge 1, \\ O_p(n^{-1/2} \tau_p^{-2}) + O_p(n^{-1} \tau_p^{-3}), & \text{if } \|r_{XX}\|_2 < 1. \end{cases}\end{aligned} \tag{A.5}$$

Combining (A.4), (A.5) and the identity that

$$\begin{aligned}\|\boldsymbol{\alpha}_p\|_2 &= \left[\sum_{j=1}^{p}\left\{\int_{\mathbb{I}_Y}\int_{\mathbb{I}_X} r_{XY}(s,t)\Gamma^j_{XX}(\beta)(s,s')\mathrm{d}s\mathrm{d}s'\right\}^2\right]^{1/2} \leq \left\{\sum_{j=1}^{p}\|r_{XY}\|_2^2\|\Gamma^j_{XX}(\beta)\|_2^2\right\}^{1/2}\\ &= \begin{cases} O(p^{1/2}\|r_{XX}\|_2^p), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O(1), & \text{if } \|r_{XX}\|_2 < 1,\end{cases}\end{aligned} \tag{A.6}$$

we reach that

$$\begin{aligned}&\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2 \leq \|\widehat{\boldsymbol{H}}_p^{-1}\|_2\|\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{\alpha}_p\|_2 + \|\widehat{\boldsymbol{H}}_p^{-1} - \boldsymbol{H}_p^{-1}\|_2\|\boldsymbol{\alpha}_p\|_2\\ &\quad = \begin{cases} O_p(n^{-1/2}\tau_p^{-1}p^{3/2}\|r_{XX}\|_2^{p}) + \ O_p(n^{-1/2}\tau_p^{-2}p^{5/2}\|r_{XX}\|_2^{3p}) + O_p(n^{-1}\tau_p^{-3}p^{9/2}\|r_{XX}\|_2^{5p}), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O_p(n^{-1/2}\tau_p^{-1}) + O_p(n^{-1/2}\tau_p^{-2}) + O_p(n^{-1}\tau_p^{-3}), & \text{if } \|r_{XX}\|_2 < 1,\end{cases}\\ &\quad = \begin{cases} O_p(n^{-1/2}\tau_p^{-1}p^{3/2}\|r_{XX}\|_2^{p}) + \ O_p(n^{-1/2}\tau_p^{-2}p^{5/2}\|r_{XX}\|_2^{3p}) + O_p(n^{-1}\tau_p^{-3}p^{9/2}\|r_{XX}\|_2^{5p}), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O_p(n^{-1/2}\tau_p^{-2}) + O_p(n^{-1}\tau_p^{-3}), \quad (\text{since } \tau_p \leq h_{jj} = O(1)) & \text{if } \|r_{XX}\|_2 < 1.\end{cases}\end{aligned} \tag{A.7}$$

For each $(s,t) \in \mathbb{I}_X \times \mathbb{I}_Y$,

$$\begin{aligned}&|\hat{\beta}_{p,\mathrm{fAPLS}}(s,t) - \beta_{p,\mathrm{fAPLS}}(s,t)|^2\\ &\quad = \left|[\widehat{\Gamma}_{XX}(\beta)(s,s'),\ldots,\widehat{\Gamma}^p_{XX}(\beta)(s,s')]\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - [\Gamma_{XX}(\beta)(s,s'),\ldots,\Gamma^p_{XX}(\beta)(s,s')]\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\right|^2\\ &\quad \leq \left|\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2\left[\sum_{j=1}^{p}\{\widehat{\Gamma}^j_{XX}(\beta)(s,s')\}^2\right]^{1/2} + \|\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2\left(\sum_{j=1}^{p}[\{\widehat{\Gamma}^j_{XX} - \Gamma^j_{XX}\}(\beta)(s,s')]^2\right)^{1/2}\right|^2\\ &\quad \leq 2\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\left[\sum_{j=1}^{p}\{\widehat{\Gamma}^j_{XX}(\beta)(s,s')\}^2\right] + 2\|\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\left[\sum_{j=1}^{p}\{\widehat{\Gamma}^j_{XX}(\beta)(s,s') - \Gamma^j_{XX}(\beta)(s,s')\}^2\right].\end{aligned}$$

Thus, $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2$ is bounded as below:

$$\begin{aligned}\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2^2 &\leq 2\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^{p}\|\Gamma^j_{XX}(\beta)\|_2^2 + 2\|\boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^{p}\|\Gamma^j_{XX}(\beta) - \widehat{\Gamma}^j_{XX}(\beta)\|_2^2\\ &\leq 2\|\widehat{\boldsymbol{H}}_p^{-1}\widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1}\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^{p}\|\Gamma^j_{XX}(\beta)\|_2^2 + 2\tau_p^{-2}\|\boldsymbol{\alpha}_p\|_2^2\sum_{j=1}^{p}\|\widehat{\Gamma}^j_{XX}(\beta) - \Gamma^j_{XX}(\beta)\|_2^2,\end{aligned} \tag{A.8}$$

where, owing to (A.7),

$$\text{the first term of (A.8)} = \begin{cases} O_p(n^{-1}\tau_p^{-2}p^4\|r_{XX}\|_2^{4p}) + \ O_p(n^{-1}\tau_p^{-4}p^6\|r_{XX}\|_2^{8p}) + O_p(n^{-2}\tau_p^{-6}p^{10}\|r_{XX}\|_2^{12p}), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}), & \text{if } \|r_{XX}\|_2 < 1;\end{cases}$$

the order of the second term of (A.8) is given by (A.6) and Lemma A.2, i.e.,

$$\text{the second term of (A.8)} = \begin{cases} O(n^{-1}\tau_p^{-2}p^4\|r_{XX}\|_2^{4p}), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O_p(n^{-1}\tau_p^{-2}), & \text{if } \|r_{XX}\|_2 < 1.\end{cases}$$

In this way we deduce

$$\begin{aligned}&\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_2^2\\ &\quad = \begin{cases} O_p(n^{-1}\tau_p^{-2}p^4\|r_{XX}\|_2^{4p}) + \ O_p(n^{-1}\tau_p^{-4}p^6\|r_{XX}\|_2^{8p}) + O_p(n^{-2}\tau_p^{-6}p^{10}\|r_{XX}\|_2^{12p}), & \text{if } \|r_{XX}\|_2 \geq 1,\\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}), & \text{if } \|r_{XX}\|_2 < 1.\end{cases}\end{aligned} \tag{A.9}$$

A set of necessary conditions for the zero-convergence (in probability) of (A.9) is contained in (C5). Once they are fulfilled, we conclude the L_2 convergence (in probability) of $\hat{\beta}_{p,\mathrm{fAPLS}}$ to β, following Proposition 1.

We complete the proof by bounding the estimating error in the supremum metric:

$$
\begin{aligned}
\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_\infty^2 &= \left\| [\widehat{\Gamma}_{XX}(\beta), \ldots, \widehat{\Gamma}_{XX}^p(\beta)] \widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - [\Gamma_{XX}(\beta), \ldots, \Gamma_{XX}^p(\beta)] \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p \right\|_\infty^2 \\
&\leq 2\|\widehat{\boldsymbol{H}}_p^{-1} \widehat{\boldsymbol{\alpha}}_p - \boldsymbol{H}_p^{-1} \boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\Gamma_{XX}^j(\beta)\|_\infty^2 + 2\tau_p^{-2} \|\boldsymbol{\alpha}_p\|_2^2 \sum_{j=1}^p \|\widehat{\Gamma}_{XX}^i(\beta) - \Gamma_{XX}^j(\beta)\|_\infty^2 \qquad \text{(A.10)} \\
&= \begin{cases} O_p(n^{-1}\tau_p^{-2} p^4 \|r_{XX}\|_\infty^{4p}) + O_p(n^{-1}\tau_p^{-4} p^6 \|r_{XX}\|_\infty^{8p}) + O_p(n^{-2}\tau_p^{-6} p^{10} \|r_{XX}\|_\infty^{12p}), & \text{if } \|r_{XX}\|_\infty \geq 1, \\ O_p(n^{-1}\tau_p^{-4}) + O_p(n^{-2}\tau_p^{-6}), & \text{if } \|r_{XX}\|_\infty < 1, \end{cases}
\end{aligned}
$$

where (A.10) is the counterpart of (A.8). $\|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta_{p,\mathrm{fAPLS}}\|_\infty$ converges to zero (in probability), once (C6) is satisfied. 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 3. Notice that

$$
\begin{aligned}
\|\hat{g}_{p,\mathrm{fAPLS}}(X_0) - g(X_0)\|_2 &\leq \|\bar{Y} - \mu_Y\|_2 + \|\bar{X} - \mu_X\|_2 \|\beta\|_2 + \|X_0 - \bar{X}\|_2 \|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_2, \\
\|\hat{g}_{p,\mathrm{fAPLS}}(X_0) - g(X_0)\|_\infty &\leq \|\bar{Y} - \mu_Y\|_\infty + \|\bar{X} - \mu_X\|_2 \|\beta\|_\infty + \|X_0 - \bar{X}\|_2 \|\hat{\beta}_{p,\mathrm{fAPLS}} - \beta\|_\infty.
\end{aligned}
$$

The finite trace of R_{XX} (resp. R_{YY}), viz. $\sum_{j=1}^\infty \lambda_{j,X} = \mathrm{E}(\|X - \mu_X\|_2^2) < \infty$ (resp. $\sum_{j=1}^\infty \lambda_{j,Y} = \mathrm{E}(\|Y - \mu_Y\|_2^2) < \infty$), entails that $\|\bar{X} - \mu_X\|_2 = o_{\mathrm{a.s.}}(1)$ (resp. $\|\bar{Y} - \mu_Y\|_2 = o_{\mathrm{a.s.}}(1)$); see [19, (2.1.3)]. The proof is complete once we verify the zero-convergence (in probability and under (C7)) of $\|\bar{Y} - \mu_Y\|_\infty$ following Hoffmann-Jørgensen [18, Theorem 2]. □

References

[1] A.M.H. Albaqshi, Generalized Partial Least Squares Approach for Nominal Multinomial Logit Regression Models with a Functional Covariate (Ph.D. thesis), University of Northern Colorado, 2017.

[2] A.L. Alexander, J.E. Lee, M. Lazar, A.S. Field, Diffusion tensor imaging of the brain, Neurotherapeutics 4 (2007) 316–329.

[3] D. Benatia, M. Carrasco, J.-P. Florens, Functional linear regression with functional response, J. Econ. 201 (2017) 269–291.

[4] U. Beyaztas, H.L. Shang, On function-on-function regression: partial least squares approach, Environ. Ecol. Stat. 27 (2020) 95–114.

[5] J.-M. Chiou, Y.-F. Yang, Y.-T. Chen, Multivariate functional linear regression and prediction, J. Multivariate Anal. 146 (2016) 301–312.

[6] C. Crambes, A. Mas, Asymptotics of prediction in functional linear regression with functional outputs, Bernoulli 19 (2013) 2627–2651.

[7] P. Craven, G. Wahba, Smoothing noisy data with spline functions, Numer. Math. 31 (1979) 377–403.

[8] A. Cuevas, M. Febrero, R. Fraiman, Linear functional regression: The case of fixed design and functional response, Canad. J. Statist. 30 (2002) 285–300.

[9] A. Delaigle, P. Hall, Methodology and theory for partial least squares applied to functional data, Ann. Statist. 40 (2012) 322–352.

[10] J.S. Evans, SpatialEco, 2020, R package version 1.3-3.

[11] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Kernel regression with functional response, Electron. J. Stat. 5 (2011) 159–171.

[12] F. Ferraty, I. Van Keilegom, P. Vieu, Regression when both response and predictor are functions, J. Multivariate Anal. 109 (2012) 10–28.

[13] J. Goldsmith, J. Bob, C.M. Crainiceanu, B. Caffo, D. Reich, Penalized functional regression, J. Comput. Graph. Statist. 20 (2011) 830–851.

[14] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, J. Gellar, J. Harezlak, M.W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, P.T. Reiss, Refund: Regression with functional data, 2019, R package version 0.1-21.

[15] 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 (1984) 149–192.

[16] G. He, H.-G. Müller, J.-L. Wang, W. Yang, Functional linear regression via canonical analysis, Bernoulli 16 (2010) 705–729.

[17] U.W. Hochstrasser, Orthogonal polynomials, in: M. Abramowitz, I.A. Stegun (Eds.), Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, in: Applied Mathematics Series, vol. 55, Dover Publications, Inc., New York, 1972, pp. 773–802, Tenth original printing with corrections.

[18] J. Hoffmann-Jørgensen, Necessary and sufficient condition for the uniform law of large numbers, in: A. Beck, R. Dudley, M. Hahn, J. Kuelbs, M. Marcus (Eds.), Probability in Banach Spaces V, Springer, Berlin, 1985, pp. 258–272.

[19] J. Hoffmann-Jørgensen, G. Pisier, The law of large numbers and the central limit theorem in banach spaces, Ann. Probab. 4 (1976) 587–599.

[20] L. Horváth, P. Kokoszka, Functional Data Analysis, in: Springer Series in Statistics, Springer, New York, 2012.

[21] A.E. Ivanescu, A.-M. Staicu, F. Scheipl, S. Greven, Penalized function-on-function regression, Comput. Statist. 30 (2015) 539–568.

[22] S. de Jong, SIMPLS: An alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (1993) 251–263.

[23] K. Lange, Numerical Analysis for Statisticians, second ed., Springer, New York, 2010.

[24] Y. Li, T. Hsing, Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data, Ann. Statist. 38 (2010) 3321–3351.

[25] H. Lian, Minimax prediction for functional linear regression with functional responses in reproducing kernel Hilbert spaces, J. Multivariate Anal. 140 (2015) 395–402.

[26] R. Luo, X. Qi, Function-on-function linear regression by signal compression, J. Amer. Statist. Assoc. 112 (2017) 690–705.

[27] R. Luo, X. Qi, FRegSigCom: Functional regression using signal compression approach, 2018, R package version 0.3.0.

[28] J.S. Marron, J.O. Ramsay, L.M. Sangalli, A. Srivastava, Functional data analysis of amplitude and phase variation, Statist. Sci. 30 (2015) 468–484.

[29] R.A. Olshen, E.N. Biden, M.P. Wyatt, D.H. Sutherland, Gait analysis and the bootstrap, Ann. Statist. 17 (1989) 1419–1440.

[30] D. Paul, J. Peng, Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach, Electron. J. Stat. 5 (2011) 1960–2003.

[31] J.O. Ramsay, C.J. Dalzell, Some tools for functional data analysis, J. R. Stat. Soc. Ser. B Stat. Methodol. 53 (1991) 539–572.

[32] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, second ed., in: Springer Series in Statistics, Springer, New York, 2005.

[33] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, Fda: Functional data analysis, 2020, R package version 5.1.4.

[34] M. Singer, T. Krivobokova, A. Munk, B. de Groot, Partial least squares for dependent data, Biometrika 103 (2016) 351–362.

[35] X. Sun, P. Du, X. Wang, P. Ma, Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework, J. Amer. Statist. Assoc. 113 (2018) 1601–1611.

[36] H. Tasaki, Convergence rates of approximate sums of Riemann integrals, J. Approx. Theory 161 (2009) 477–490.
[37] W. Wang, Linear mixed function-on-function regression models, Biometrics 70 (2014) 794–801.
[38] Y. Wang, J.G. Ibrahim, H. Zhu, Partial least squares for functional joint models with applications to the Alzheimer's disease neuroimaging initiative study, Biometrics 76 (2020) 1109–1119.
[39] H. Wold, Path models with latent variables: the NIPALS approach, in: H. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon, V. Capecchi (Eds.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building, Academic Press, New York, 1975, pp. 307–335.
[40] L. Xiao, Asymptotic theory of penalized splines, Electron. J. Stat. 13 (2019) 747–794.
[41] L. Xiao, C. Li, W. Checkley, C. Crainiceanu, Fast covariance estimation for sparse functional data, Stat. Comput. 28 (2018) 511–522.
[42] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, J. Amer. Statist. Assoc. 100 (2005) 577–590.
[43] F. Yao, H.-G. Müller, J.-L. Wang, Functional linear regression analysis for longitudinal data, Ann. Statist. 33 (2005) 2873–2903.
[44] Z. Zhou, R.A. Lockhart, Partial least squares for sparsely observed curves with measurement errors, 2020, arXiv:2003.11542.