Partial Least Squares Regression with Multiple Functional and Scalar Predictors

Jongmin Mun and Jeong Hoon Jang

July 9, 2025

1 Introduction

Functional regression settings involving a large number of predictors, $X_{i1}(t), X_{i2}(t), \ldots, X_{iK}(t)$, are becoming increasingly common. For example, Storey et al. (2005) analyzes two gene expression datasets measured over time, which involve only a small number of patients but tens of thousands of functional predictors.

Infinite dimensional beta estimation problem (ill-posed problem)

Multicollinearity and high-dimensionality in scalar predictors are not handled.

Also, there may be high correlation between functional and scalar predictors.

Many approaches have been proposed 1. Roughness Penalty 2. basis approach (power-series, B-splines, wavelets) - Power Series splines (Goldsmith et al., 2011, Journal of Computational Statistics) - B-splines (H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statistica Sinica 13 (2003) 571–591. T.T.Cai,P.Hall,Prediction in functional linear regression, Ann.Stat. 34(5) (2006) 2159–2179.) - Wavelet (Y.Zhao,R.T.Ogden,P.T.Reiss,Wavel based lasso in functional linear regression, J.Comput.Graph.Stat.21(3)(2012)600–61.)7

3. FPCA approach - Hall, P. and Horowitz, J. L. (2007), 'Methodology and convergence rates for functional linear regression', The Annals of Statistics 35(1), 70–91. - Reiss P. T., Ogden R. T.. Functional principal component regression and functional partial least squares, Journal of the American Statistical Association, 2007, vol. 102 (pg. 984-996) - Febrero-Bande et al., 2017

4. FPLS Approach - Preda and Saporta (2005) - Reiss and Odgen, 2007 - Aguilera AM, Escabias M, Preda C, Saporta G. Using basis expansions for estimating functional PLS regression applications withchemometric data.Chemom Intell Lab Syst. 2010;104:289-305. - Delaigle and Hall, 2012 - Febrero-Bande et al., 2017 - Beyaztas and Shang, (2022) A Robust Functional Partial Least Squares for Scalar-on-Multiple-Function Regression. Journal of Chemometrics. - Saricam et al., On partial least-squares estimation in scalar-on-function regression models. Journal of Chemometrics. - Mutis et al. (2025) On function-on-function linear quantile regression

2 Data Objects and Model Formulation

We assume, without loss of generality, that the functional predictors $X_{i1}(t), X_{i2}(t), \ldots, X_{iK}(t)$ are observed over the domain $0 \leq t \leq 1$. We also assume that each functional predictor belongs to the space $L^2([0,1])$, the Hilbert space of square-integrable real-valued functions with respect to the Lebesgue measure dt on $[0,1]$.

Remark 1 *Our method is applicable to settings where each functional predictor belongs to a different Hilbert space, possibly defined over a distinct compact domain in* $\mathbb{R}^d$*, for arbitrary* d*, and observed at different time points. However, for notational simplicity, our discussion assumes a common Hilbert space over domain* $[0,1]$ *for all functional predictors.*

Write $X = (X^{(1)}, \ldots, X^{(K)})$ as a multivariate functional object that belongs to $\mathcal{F} =$

$L^2(\mathcal{T}_1) \times \cdots \times L^2(\mathcal{T}_K)$—a cartesian product of individual $L^2(\mathcal{T}_k)$ spaces. Note that if $K = 1$, X reduces to a univariate functional object. We can also express the functional object X evaluated on the multi-dimensional argument $\mathbf{t} = (t_1, \ldots, t_k)^\top \in \mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_K$ as a K-dimensional vector $X(\mathbf{t}) = (X^{(1)}(t_1), \ldots, X^{(K)}(t_K))^\top$. The inner product of $f_1 = (f_1^{(1)}, \ldots, f_1^{(K)})$ and $f_2 = (f_2^{(1)}, \ldots, f_2^{(K)})$ in $\mathcal{F}$ is defined as $\langle f_1, f_2 \rangle_{\mathcal{F}} = \sum_{k=1}^K \langle f_1^{(k)}, f_2^{(k)} \rangle_{L^2} = \sum_{k=1}^K \int_{\mathcal{T}_k} f_1^{(k)}(t_k) f_2^{(k)}(t_k) dt_k$ with norm $\|f_1\|_{\mathcal{F}} = \langle f_1, f_1 \rangle_{\mathcal{F}}^{1/2} = \{\sum_{k=1}^K \int_{\mathcal{T}_k} f_1^{(k)}(t_k)^2 dt_k\}^{1/2}$.

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ denote a p-dimensional multivariate scalar data in $\mathbb{R}^p$. We assume that $\mathbf{Z}$ is a random vector with finite first two moments and equipped with the Euclidean inner product and norm; i.e., for $\mathbf{v_1} = (v_{11}, \ldots, v_{1p})^\top$ and $\mathbf{v_2} = (v_{21}, \ldots, v_{2p})^\top$ in $\mathbb{R}^p$, $\langle \mathbf{v_1}, \mathbf{v_2} \rangle = \mathbf{v_1}^\top \mathbf{v_2} = \sum_{r=1}^p v_{1r} v_{2r}$, and $\|\mathbf{v}_1\| = \langle \mathbf{v}_1, \mathbf{v}_1 \rangle^{1/2} = (\sum_{r=1}^p v_{1r}^2)^{1/2}$.

Without loss of generality, assume that both multivariate functional and scalar data are centered; that is, $E(X) = 0$ and $E(\mathbf{Z}) = \mathbf{0}$. Our goal is to predict a real outcome Y based on X and $\mathbf{Z}$ via the following functional regression model:

$$Y = \sum_{r=1}^p \alpha_r Z_r + \sum_{k=1}^K \int_{\mathcal{T}_k} \beta^{(k)}(t_k) X^{(k)}(t_k) dt_k + \epsilon = \langle \boldsymbol{\alpha}, \mathbf{Z} \rangle + \langle \beta, X \rangle_{\mathcal{F}} + \epsilon \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top \in \mathbb{R}^p$ and $\beta = (\beta^{(1)}, \ldots, \beta^{(K)}) \in \mathcal{F}$ are respectively the regression coefficient vector and function that characterize the effect of scalar and functional predictors on the outcome.

Our strategy is to formulate a hybrid random object, by combining X and $\mathbf{Z}$ into an ordered pair:

$$\mathbf{W} = (X, \mathbf{Z}) \in \mathcal{H}, \text{ where } \mathcal{H} := \mathcal{F} \times \mathbb{R}^p \tag{2}$$

An alternative notation for the hybrid object can be obtained by evaluating its functional part on $\mathbf{t}$ and expressing it as a $(K + p)$-dimensional vector: $\mathbf{W}[\mathbf{t}] = (X(\mathbf{t}), \mathbf{Z})^\top$, with $X(\mathbf{t}) = (X^{(1)}(t_1), \ldots, X^{(K)}(t_K))^\top \in \mathbb{R}^K$ and $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top \in \mathbb{R}^p$.

Definition 2 (Hilbert space of hybrid objects) *We define the inner product between any two hybrid objects,* $\mathbf{h}_1 = (f_1, \mathbf{v}_1)$ *and* $\mathbf{h}_2 = (f_2, \mathbf{v}_2)$*, as*

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle_{\mathcal{H}} := \langle f_1, f_2 \rangle_{\mathcal{F}} + \omega \langle \mathbf{v_1}, \mathbf{v_2} \rangle = \sum_{k=1}^{K} \int_{\mathcal{T}_k} f_1^{(k)}(t_k) f_2^{(k)}(t_k) dt_k + \omega \sum_{r=1}^{p} v_{1r} v_{2r}, \tag{3}$$

with norm $\| \cdot \|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$.

In (3), ω is a positive weight that needs to be pre-specified or estimated. It is mainly used to take into account heterogeneity between functional and scalar parts in terms of measurement scale and/or amount of variation (see Section 4.4). Without loss of generality and for the clarity of illustration, all the following theoretical results will be derived for $\omega = 1$. The results remain valid for any positive weights. Then the functional regression model (1) with multiple functional and scalar predictors can be re-expressed in terms of the scalar-on-hybrid regression model as follows

$$Y = \langle \boldsymbol{\eta}, \mathbf{W} \rangle_{\mathcal{H}} + \epsilon, \tag{4}$$

where $\boldsymbol{\eta} = (\beta, \boldsymbol{\alpha}) \in \mathcal{H}$ is a hybrid regression coefficient characterizing the association between the hybrid predictor $\mathbf{W}$ and the outcome Y.

The basic idea of partial least squares (PLS) is to simultaneously decompose the hybrid predictor $\mathbf{W}$ and the real outcome Y in terms of zero mean uncorrelated PLS scores $(\rho_l)_{l \in \mathbb{N}}$ with maximum predictive performance as follows

$$\mathbf{W}[\mathbf{t}] = \sum_{l=1}^{\infty} \rho_l \boldsymbol{\delta}_l[\mathbf{t}] \quad \text{and} \quad Y = \sum_{l=1}^{\infty} \rho_l \nu_l + \epsilon, \tag{5}$$

where $(\boldsymbol{\delta}_l)_{l \in \mathbb{N}} \in \mathcal{H}$ and $(\nu_l)_{l \in \mathbb{N}} \in \mathbb{R}$ are appropriate bases. The PLS scores are obtained as $\rho_l = \langle \mathbf{W}^{[l]}, \boldsymbol{\xi}_l \rangle_{\mathcal{H}}$, where $\mathbf{W}^{[l]}$ denotes the residualized predictor sequentially derived as the residual of the regression of $\mathbf{W}^{[l-1]}$ on ρ_{l-1} with $\mathbf{W}^{[1]} = \mathbf{W}$, and $\boldsymbol{\xi}_l = (\psi_l, \boldsymbol{\theta}_l) \in \mathcal{H}$ is the hybrid

PLS component, with $\psi_l = (\psi_l^{(1)}, \ldots, \psi_l^{(K)}) \in \mathcal{F}$ and $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lp})^\top \in \mathbb{R}^p$, sequentially chosen to maximize the squared covariance between ρ_l and the residualized outcome $Y^{[l]}$—i.e., $\text{Cov}^2(\rho_l, Y^{[l]}) = \text{Cov}^2(\langle \mathbf{W}^{[l]}, \boldsymbol{\xi}_l \rangle_{\mathcal{H}}, Y^{[l]})$. Here, $Y^{[l]}$ is also sequentially obtained as the residual of the regression of $Y^{[l-1]}$ on ρ_{l-1} with $Y^{[1]} = Y$. The sequentially derived hybrid PLS components $\{\boldsymbol{\xi}_l\}_{l=1}^{\infty}$, are orthonormal in a sense that $\langle \boldsymbol{\xi}_l, \boldsymbol{\xi}_j \rangle_{\mathcal{H}} = \mathbb{1}(l = j)$.

Challenges

1. Independent variables consist of multiple highly structured images and scalar predictors.

2. Our sample size is small compared to the dimension and number of functional and scalar predictors.

3. Existing partial least squares (PLS) methods can only accommodate (i) univariate or multivariate functional predictors without any scalar predictors (Preda and Saporta, 2005, Delaigle and Hall, 2012, Febrero-Bande et al., 2017, Beyaztas and Shang, 2020); or (2) a univariate functional predictor with other scalar predictors (Wang, 2018).

3 Naive PLS Algorithm

This section introduces a naive pointwise hybrid PLS algorithm and highlights its limitations. The algorithm treats each functional predictor as a long vector of discretized points, derives PLS components independently at each observed time point $\mathbf{t}$, and aggregates them into a functional object to compute the PLS score. Since the regression coefficient computation follows directly from the PLS components and scores, our discussion here focuses on the latter and omits the former. Denote $\{(Y_1, \mathbf{W}_1), \ldots, (Y_n, \mathbf{W}_n)\}$ as n independent data pairs (observed sample) distributed as $(Y, \mathbf{W})$. The goal of the PLS algorithm is to decompose

Algorithm 1 Naive Hybrid PLS (regression coefficient omitted)

Require: Hybrid predictors $\{\mathbf{W}_1, \ldots, \mathbf{W}_n\} = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$, responses $\{Y_1, \ldots, Y_n\}$, functional predictor evaluation points $\mathbf{t}_1, \ldots, \mathbf{t}_m \in [0,1]^K$

for i = 1, 2, ... n **do** ▷ Centering (6)

  $\{(Y_i^{[1]}, X_i^{[1]}, Z_i^{[1]})\} \leftarrow \{(Y_i - \bar{Y}, X_i - \bar{X}, Z_i - \bar{Z})\}$

for l = 1, 2, ... L **do**

  PLS component computation step:

  for $\mathbf{t} = \mathbf{t}_1, \ldots \mathbf{t}_m$ **do** ▷ Compute PLS loading

    $\widehat{\boldsymbol{\xi}}_l[\mathbf{t}] \leftarrow \dfrac{\sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]}[\mathbf{t}]}{\| \sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]} \|_{\mathcal{H}}}$ ▷ (8)

  $\{\widehat{\rho}_{1l}, \ldots, \widehat{\rho}_{nl}\} \leftarrow \left\{ \langle \mathbf{W}_1^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}}, \ldots, \langle \mathbf{W}_n^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} \right\}$ ▷ Compute PLS scores (7)

  Orthogonalization step:

  for $\mathbf{t} = \mathbf{t}_1, \ldots \mathbf{t}_m$ **do**

    $\widehat{\boldsymbol{\delta}}_l[\mathbf{t}] \leftarrow \dfrac{\sum_{i=1}^n \mathbf{W}_i^{[l]}[\mathbf{t}] \widehat{\rho}_{il}}{\sum_{i=1}^n \widehat{\rho}_{il}^2}.$

  $\mathbf{W}_i^{[l+1]}[\mathbf{t}] \leftarrow \mathbf{W}_i^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\delta}}_l[\mathbf{t}] \widehat{\rho}_{il}$

  $\widehat{\nu}_l \leftarrow \dfrac{\sum_{i=1}^n Y_i^{[l]} \widehat{\rho}_{il}}{\sum_{i=1}^n \widehat{\rho}_{il}^2}$

  $Y_i^{[l+1]} \leftarrow Y_i^{[l]} - \widehat{\nu}_l \widehat{\rho}_{il}$

Output PLS scores $\left\{\widehat{\rho}_{il}\right\}_{i \in [n], l \in [L]}$

the predictor $\mathbf{W}_i$ $(i = 1, \ldots, n)$ and the real response Y_i in terms of zero mean uncorrelated PLS scores $(\rho_{il})_{l \in \mathbb{N}}$ with maximum predictive performance. Here and below, a superscript in square brackets denotes the iteration index of the algorithm. We begin by denoting the centered data as

$$Y_i^{[1]} = Y_i - \bar{Y} \text{ and } \mathbf{W}_i^{[1]} = (X_i^{[1]}, \mathbf{Z}_i^{[1]}) = (X_i - \bar{X}, \mathbf{Z}_i - \bar{\mathbf{Z}}) = \mathbf{W}_i - \bar{\mathbf{W}}. \tag{6}$$

We describe the l-th step of the algorithm for $l = 1, 2, \ldots, L$, which consists of the following substeps. The complete procedure is summarized in Algorithm 1.

Step 1. We compute the lth PLS score of the ith subject as

$$\widehat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} = \langle X_i^{[l]}, \widehat{\psi}_l \rangle_{\mathcal{F}} + \langle \mathbf{Z}_i^{[l]}, \widehat{\boldsymbol{\theta}}_l \rangle, \tag{7}$$

where the l-th PLS loading $\widehat{\boldsymbol{\xi}}_l = (\widehat{\psi}_l, \widehat{\boldsymbol{\theta}}_l) \in \mathcal{H}$ maximizes $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}_i^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}}, Y_i^{[l]}) = \widehat{\mathrm{Cov}}^2(\widehat{\rho_{il}}, Y_i^{[l]})$. Specifically, we define the l-th PLS loadings pointwise as follows:

$$\widehat{\boldsymbol{\xi}}_l[\mathbf{t}] = \frac{\sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]}[\mathbf{t}]}{\| \sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]} \|_{\mathcal{H}}} = \left[\frac{\sum_{i=1}^n Y_i^{[l]} X_i^{[l]}(\mathbf{t})}{\| \sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]} \|_{\mathcal{H}}}, \frac{\sum_{i=1}^n Y_i^{[l]} \mathbf{Z}_i^{[l]}}{\| \sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]} \|_{\mathcal{H}}} \right]^\top, \tag{8}$$

where the normalizing factor is computed as

$$\| \sum_{i=1}^n Y_i^{[l]} \mathbf{W}_i^{[l]} \|_{\mathcal{H}} = [\sum_{i=1}^n \sum_{j=1}^n Y_i^{[l]} Y_j^{[l]} \{ \langle X_i^{[l]}, X_j^{[l]} \rangle_{\mathcal{F}} + \langle \mathbf{Z}_i^{[l]}, \mathbf{Z}_j^{[l]} \rangle \}]^{1/2}. \tag{9}$$

Step 2. Obtain the subsequent residualized outcomes and predictors as $Y_i^{[l+1]} = Y_i^{[l]} - \widehat{\nu}_l \widehat{\rho}_{il}$ and $\mathbf{W}_i^{[l+1]}[\mathbf{t}] = \mathbf{W}_i^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\delta}}_l[\mathbf{t}] \widehat{\rho}_{il}$, where

$$\widehat{\nu}_l = \frac{\sum_{i=1}^n Y_i^{[l]} \widehat{\rho}_{il}}{\sum_{i=1}^n \widehat{\rho}_{il}^2} \quad \text{and} \quad \widehat{\boldsymbol{\delta}}_l[\mathbf{t}] = \frac{\sum_{i=1}^n \mathbf{W}_i^{[l]}[\mathbf{t}] \widehat{\rho}_{il}}{\sum_{i=1}^n \widehat{\rho}_{il}^2}.$$

Note, $\widehat{\nu}_l$ is the least squares estimate from linear regression of $Y_i^{[l]}$ on $\widehat{\rho}_{il}$, and $\widehat{\boldsymbol{\delta}}_l[\mathbf{t}]$ is the least squares estimate from a linear regression of $\mathbf{W}_i^{[l]}[\mathbf{t}]$ on $\widehat{\rho}_{il}$.

Step 3. Let $l = l + 1$ and back to Step 1.

This naive pointwise PLS algorithm, however, can be computationally expensive for multiple dense functional data and is not feasible for irregular functional data. Moreover, the pointwise estimates only use information from data on the particular argument value, and thus can show substantial variability across the domain, resulting in overfitting and unstable predictions.

4 Proposed PLS Algorithm

To address the limitations outlined in Section 3, we propose novel strategies for implementing the steps of the hybrid PLS algorithm. Our approach provides an efficient and robust means of producing PLS components and scores in the presence of multiple dense and/or irregular functional predictors and scalar predictors. It also incorporates a regularization scheme that enables the algorithm to borrow strength and exploit structural relationships within and between the functions to avoid overfitting of the PLS components and to improve the generalizability and interpretability of the predictive model. Each iteration of our approach consists of two subroutines: regularized estimation of smoothed PLS components and orthogonalization, detailed in Sections 4.1 and 4.2, respectively. After a suitable number of iterations, the hybrid regression coefficient is estimated, as described in Section 4.3. For notational simplicity, we omit the iteration index l in the following discussion, with the understanding that the subroutines apply to any iteration. The complete algorithm is summarized in Algorithm 2.

4.1 Step 1: Computing smoothed PLS components

The function part ψ of the hybrid PLS component $\boldsymbol{\xi}$ is inherently infinite-dimensional, rendering its estimation impossible with finite sample size. As such, we first approximate each functional predictor (observed version when $l = 1$, or residualized version when $l \geq 2$) as a linear combination of M basis functions and coefficients. Due to the hybrid nature of the predictors, this approximation requires careful explanation, provided in Section 4.1.1. This representation allows us to work with coefficient vectors, simplifying notation. Section 4.1.2 outlines the core strategy for computing the PLS components in this reduced space. Building on this, we then introduce our regularized and smoothed estimation procedure, which

constitutes the first step of the iterative algorithm.

4.1.1 Finite-basis approximation of the predictors and regression coefficients

In practice, the functional predictors $X^{(1)}, \ldots, X^{(K)}$ are not observed as complete curves, but only through their values at a finite set of time points, which may differ across the K predictors. Thus, each observation is represented by a finite-dimensional vector $X_i^{(k)}(\mathbf{t}_k) \in \mathbb{R}^{|\mathbf{t}_k|}$, where $\mathbf{t}_k \subset \mathcal{T}_k$. The first step in functional data analysis typically involves reconstructing the underlying functional form from these discrete measurements, commonly by assuming that sample paths lie in a finite-dimensional space spanned by basis functions.

Given an orthonormal basis $\{b_l(t)\}$, the functional predictors and the corresponding regression coefficients can be decomposed into

$$X_{ij}(t) = \sum_{l=1}^{\infty} \theta_{ijl}\, b_l(t), \quad \beta_j(t) = \sum_{l=1}^{\infty} \eta_{0,jl}\, b_l(t),$$

where θ_{ijl} and $\eta_{0,jl}$ are the coefficients of $X_{ij}(t)$ and $\beta_j(t)$ corresponding to the lth basis function $b_l(t)$, respectively. While it is possible to use a different basis for each functional predictor, we adopt a common basis system for all predictors for notational simplicity.

For $k = 1, \ldots, K$, let $\Phi^{(k)} := \{\phi_1^{(k)}, \ldots, \phi_M^{(k)}\}$, with $M \geq 1$, be an M-dimensional basis in $L^2(\mathcal{T}_k)$. In our proposed hybrid PLS regression algorithm, this basis system is used to approximate both the predictors and the PLS components. Additionally, it serves as a regularization mechanism, borrowing strength across components and helping to prevent overfitting. The number of basis functions M for each functional predictor can be set to a sufficiently large value (e.g., 15 or 20) to incorporate a rich variety of patterns that functional predictors may exhibit. Note that there is no need to carefully choose different M values for different functional predictors, as an appropriate penalization scheme will be introduced in

the later section to regularize each functional predictor to achieve a desired smoothness (see Section 4.1.3).

For $k = 1, \ldots, K$, we project the kth functional predictor $X_i^{(k)}$ onto span($\Phi^{(k)}$):

$$X_i^{(k)} \approx \sum_{m=1}^{M} c_{im}^{(k)} \phi_m^{(k)}, \quad i = 1, \ldots, n,$$

where the coefficients $c_{im}^{(k)}$ satisfy

$$X_i^{(k)}(t_k) = \sum_{m=1}^{M} c_{im}^{(k)} \phi_m^{(k)}(t_k), \quad \text{for all } t_k \in \mathbf{t}_k \subset \mathcal{T}_k.$$

Thus, the n observations of the kth functional predictor can be represented by an $n \times M$ matrix of coefficients, where rows index samples and columns index basis functions:

$$\widetilde{\mathbf{C}}^{(k)} := (c_{im}^{(k)}).$$

Importantly, the core computation of our algorithm involves the inner product between n hybrid predictor observations and a single hybrid object (the PLS direction), defined via the inner product in Definition 2. With the basis expansion, this reduces to simple matrix operations. For each $k = 1, \ldots, K$, let $\mathbf{\Phi}^{(k)} := (\phi_{m\ell}^{(k)})$ denote the $M \times M$ symmetric matrix with entries the inner products between basis functions

$$\phi_{m\ell}^{(k)} := \langle \phi_m^{(k)}, \phi_\ell^{(k)} \rangle = \int_{\mathcal{T}_k} \phi_m^{(k)}(t_k)\, \phi_\ell^{(k)}(t_k)\, dt_k. \tag{10}$$

4.1.2 Smoothed estimation of PLS components

Building on the basis function approximation from Section 4.1.1, we now introduce an intermediate strategy for implementing the first step of the PLS algorithm. This approach

directly obtains a hybrid PLS direction, sharing the same structure as the hybrid predictors presented in (4), by simply solving a linear equation.

We now describe the strategy for computing the PLS direction at the l-th iteration. For simplicity, we omit the iteration index and denote the residualized response by Y, with finite-sample observations given by $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Let $\widetilde{W} \in \mathcal{H}$ denote the residualized, basis-approximated hybrid predictor, and let $\widetilde{\mathbf{C}}^{(1)}, \ldots, \widetilde{\mathbf{C}}^{(k)} \in \mathbb{R}^{n \times M}$ denote the basis coefficient matrices of its functional part in the finite-sample observations. Our strategy seeks the direction $\boldsymbol{\xi} \in \mathcal{H}$, whose functional parts lying in $\text{span}(\Phi^{(1)}), \ldots, \text{span}(\Phi^{(K)})$, that maximizes the empirical covariance between the resulting PLS component $\langle \widetilde{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}$ and the response.

For the computation of the estimated PLS direction $\hat{\boldsymbol{\xi}}$, we obtain the basis coefficients of its functional part, denoted by $\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(K)} \in \mathbb{R}^M$, and the scalar part $\mathbf{d} \in \mathbb{R}^p$, by solving linear systems

$$\boldsymbol{\Phi}^{(1)}\mathbf{d}^{(1)} = (\widetilde{\mathbf{C}}^{(1)}\boldsymbol{\Phi}^{(1)})^\top \mathbf{y}, \ldots, \boldsymbol{\Phi}^{(K)}\mathbf{d}^{(K)} = (\widetilde{\mathbf{C}}^{(K)}\boldsymbol{\Phi}^{(K)})^\top \mathbf{y}, \text{ and } \mathbf{d} := \mathbf{Z}^\top \mathbf{y}.$$

The estimated PLS direction is the tuple $\hat{\boldsymbol{\xi}} \in \mathcal{H}$, defined with these coefficients and normalized with respect to the hybrid Hilbert norm (Definition 2) as follows:

$$\hat{\boldsymbol{\xi}} := \frac{1}{\left(\mathbf{d}^\top \mathbf{d} + \sum_{k=1}^K (\mathbf{d}^{(k)})^\top \boldsymbol{\Phi}^{(K)} \mathbf{d}^{(k)}\right)^{1/2}} \left(\sum_{m=1}^M d_m^{(1)} \phi_m^{(1)}, \ldots, \sum_{m=1}^M d_m^{(K)} \phi_m^{(K)}, \mathbf{d} \right). \tag{11}$$

The details of this hybrid norm computation are provided in Appendix D.2. Finally, the estimated lth PLS scores for the n observations, denoted $\hat{\boldsymbol{\rho}} \in \mathbb{R}^n$ is computed via hybrid inner product. In a matrix multiplication form, it is computed as:

$$\hat{\boldsymbol{\rho}} := \left(\langle \widetilde{\mathbf{W}}_1, \hat{\boldsymbol{\xi}} \rangle_{\mathcal{H}}, \ldots, \langle \widetilde{\mathbf{W}}_n, \hat{\boldsymbol{\xi}} \rangle_{\mathcal{H}} \right)^\top$$

$$= \frac{1}{\left(\mathbf{d}^\top \mathbf{d} + \sum_{k=1}^{K} (\mathbf{d}^{(k)})^\top \mathbf{\Phi}^{(K)} \mathbf{d}^{(k)}\right)^{1/2}} \left(\mathbf{Z}\mathbf{d} + \sum_{k=1}^{K} \widetilde{\mathbf{C}}^{(k)} \mathbf{\Phi}^{(k)} \mathbf{d}^{(k)} \right).$$

The details of this computation are provided in Appendix D.2. Although the unnormalized coefficients are computed separately for the functional and scalar part of the PLS direction, the normalization step couples them, allowing the components to influence each other. This enables the procedure to account for the correlation between functional and scalar components.

Proposition 1 *Let $\boldsymbol{\xi} \in \mathcal{H}$ be any unit-norm direction such that its 1st through Kth functional components lie in* $\mathrm{span}(\Phi^{(1)}), \ldots, \mathrm{span}(\Phi^{(K)})$*, respectively. Then, the empirical covariance between the projected component $\langle \widetilde{\mathbf{W}}, \boldsymbol{\xi} \rangle_{\mathcal{H}}$ and the response Y, approximated in terms of the basis, is given by*

$$\widehat{\mathrm{Cov}}\left(\langle \widetilde{\mathbf{W}}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y \right) := \frac{1}{n} \sum_{i=1}^{n} y_i \langle \widetilde{\mathbf{W}}_i, \boldsymbol{\xi} \rangle_{\mathcal{H}}.$$

Then the PLS direction $\hat{\boldsymbol{\xi}}$ computed in (11) *maximizes* $\widehat{\mathrm{Cov}}(\langle \widetilde{\mathbf{W}}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y)$.

The proof of Proposition 1 is given in the Appendix D.2.

4.1.3 Proposed Approach: Regularized and Smoothed Computation

Although Proposition 1 provides an efficient and principled way to compute the PLS components $\boldsymbol{\xi}$, their functional part $\psi = (\psi^{(1)}, \ldots, \psi^{(K)})$ may not be smooth, which can lead to overfitting and unstable predictions. Thus, in this section, we provide a modification of Proposition 1 that yields regularized (smoothed) PLS components.

A widely-used approach to smoothing a function $\psi^{(k)}$ is to penalize its roughness that

can be quantified by the integrated squared second derivative:

$$\text{PEN}\left(\psi^{(k)}\right) = \int_{\mathcal{T}_k} \ddot{\psi}^{(k)}(t)^2 dt = \mathbf{a}^{(k)T} \ddot{J}^{(k)} \mathbf{a}^{(k)}, \tag{12}$$

where the last term is obtained using the basis expansion $\psi^{(k)}(t) = \mathbf{a}^{(k)T}\mathbf{b}^{(k)}(t)$.

Now, maximizing $\widehat{\text{Cov}}^2\left(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y\right)$ is not the only aim; we also want to prevent the roughness of each $\psi^{(k)}$—that is, $\text{PEN}(\psi^{(k)})$—from being too large. One possible way to achieve this is to extend Rice and Silverman (1991)'s approach of obtaining smoothed functional PCA components, and find $\boldsymbol{\xi}^*$ that maximizes:

$$\widehat{\text{Cov}}^2\left(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y\right) - \sum_{k=1}^{K} \lambda_k \text{PEN}\left(\psi^{(k)}\right), \tag{13}$$

subject to ${\boldsymbol{\xi}^*}^\top J^* \boldsymbol{\xi}^* = 1$. The equation (13) quantifies the trade-off between fidelity to the data (in this case the sample covariance between the $\rho = \langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}$ and Y in the direction of $\boldsymbol{\xi}$) and roughness as measured by the sum of the individual penalty terms in (12). The smoothing parameters $\{\lambda_k\}_{k=1}^K$ control the relative importance between the two objectives. Note that the penalty term can be written concisely using matrices as:

$$\sum_{k=1}^{K} \lambda_k \text{PEN}\left(\psi^{(k)}\right) = \sum_{k=1}^{K} \lambda_k \mathbf{a}^{(k)T} \ddot{J}^{(k)} \mathbf{a}^{(k)} = {\boldsymbol{\xi}^*}^\top \Lambda \ddot{J}^* \boldsymbol{\xi}^*,$$

where $\Lambda \in \mathbb{R}^{(MK+p)\times(MK+p)}$ is defined as

$$\Lambda = \text{blkdiag}(\lambda_1 I_M, \cdots, \lambda_K I_M, 0_{p\times p}), \tag{14}$$

$0_{p\times p}$ denotes a $p \times p$ matrix with all elements equal to zero.

In this paper, we take an alternative strategy that extends the idea of Silverman (1996).

This approach is known to produce appropriately smoothed functional PCA components under milder conditions than the Rice and Silverman (1991)'s approach by replacing the usual L^2-orthonormality constraint with the one that takes account the roughness of functions via the modified inner product. Specifically, the following proposition provides a generalized Rayleigh quotient problem which modifies the $\mathcal{H}$-orthonormality constraint of the hybrid PLS components to incorporate roughness penalty, and thus whose solution corresponds to a smoothed PLS component maximizing $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y)$ over the class of all functions satisfying sufficient smoothness.

Proposition 2 *At each l-th iteration of the first step of the PLS algorithm involving the residualized response and predictors $Y_i \equiv Y_i^{[l]}$ and $\mathbf{W}_i \equiv \mathbf{W}_i^{[l]}$, consider the basis function expansions of the functional predictors $X_i^{(k)}(t_k) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t_k)$, $i = 1, \ldots, n$, and those of the functional parts of the PLS component $\psi^{(k)}(t_k) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t_k)$, $k = 1, \ldots, K$. Recall from* (26) *that V^* is defined as*

$$\mathbf{V}^* = n^{-2} \begin{bmatrix} J\widetilde{C}^\top \mathbf{y}\mathbf{y}^\top \widetilde{C} J & J\widetilde{C}^\top \widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top \mathbf{Z} \\ \mathbf{Z}^\top \widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top \widetilde{C} J & \mathbf{Z}^\top \widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top \mathbf{Z} \end{bmatrix} = n^{-2} J^* \widetilde{C}^{*\top} \widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top \widetilde{C}^* J^* \in \mathbb{R}^{(MK+p)\times(MK+p)}.$$

and $\boldsymbol{\xi}^* = (\mathbf{a}^\top, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{MK+p}$. Then, the l-th smoothed PLS component, $\widehat{\boldsymbol{\xi}} \equiv \widehat{\boldsymbol{\xi}}_l$ can be obtained in the following way. We first compute the basis coefficients by solving the optimization problem

$$\begin{aligned} &\underset{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}}{\arg\max} \ \frac{\boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*\top} (J^* + \Lambda \ddot{J}^*) \boldsymbol{\xi}^*}, \quad \textit{or equivalently,} \\ &\underset{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}}{\arg\max} \ \boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^* \quad \text{subject to} \quad \boldsymbol{\xi}^{*\top} (J^* + \Lambda \ddot{J}^*) \boldsymbol{\xi}^* = 1. \end{aligned} \tag{15}$$

We then express the smoothed component in terms of the basis functions as

$$\widehat{\boldsymbol{\xi}}[\mathbf{t}] = [\widehat{\psi}(\mathbf{t})^\top, \widehat{\boldsymbol{\theta}}^\top]^\top = B^*[\mathbf{t}]^\top \widehat{\boldsymbol{\xi}}^*.$$

The l-th PLS score $\widehat{\rho}_i \equiv \widehat{\rho}_{il}$ can be obtained as

$$\widehat{\rho}_i = \langle \mathbf{W}_i, \widehat{\boldsymbol{\xi}} \rangle_{\mathcal{H}} \tag{16}$$

The proof of Proposition 2 is provided in Appendix D.2. Here, λ_k $(k = 1, \ldots, K)$ is a positive smoothing parameter that controls the trade-off between the goodness of fit and the amount of smoothness in $\psi^{(k)}$. Smaller λ_k produces PLS components that fit more closely to the observed data, but selecting too small value may cause overfitting. The estimated PLS component reverts back to the unregularized version of Proposition 1 when $\lambda_k = 0$ for all k. On the other hand, larger λ_k places more emphasis on smoothing $\psi^{(k)}$. In the limit $\lambda_k \to \infty$, each of the functions is forced to be of the form $\psi^{(k)}(t) = a + bt$ for some constants a and b. In practice, $\{\lambda_k\}_{k=1}^K$ as well as the number of PLS components L to be estimated can be chosen by cross-validation based on a certain predictive performance criterion (e.g., mean squared error).

The constraint $\boldsymbol{\xi}^{*\top}(J^* + \Lambda \ddot{J}^*)\boldsymbol{\xi}^* = 1$ enforces the following modified orthonormality of the estimated PLS components $(\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \ldots)$ based on the modified inner product $\langle\langle h_l, h_j \rangle\rangle = \langle h_l, h_j \rangle_{\mathcal{H}} + \sum_{k=1}^K \lambda_k \langle \ddot{f}_l^{(k)}, \ddot{f}_j^{(k)} \rangle_{L^2}$ which now takes into account the roughness of the functions:

$$\begin{aligned}
\langle\langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j \rangle\rangle &= \langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j \rangle_{\mathcal{H}} + \sum_{k=1}^K \lambda_k \left\langle \ddot{\widehat{\psi}}_l^{(k)}, \ddot{\widehat{\psi}}_j^{(k)} \right\rangle_{L^2} \\
&= \sum_{k=1}^K \int_{\mathcal{T}_k} \widehat{\psi}_l^{(k)}(t_k) \widehat{\psi}_j^{(k)}(t_k) dt_k + \sum_{r=1}^p \widehat{\theta}_{lr} \widehat{\theta}_{jr} + \sum_{k=1}^K \lambda_k \int_{\mathcal{T}_k} \ddot{\widehat{\psi}}_l^{(k)}(t_k) \ddot{\widehat{\psi}}_j^{(k)}(t_k) dt_k \\
&= \mathbb{1}(l = j).
\end{aligned}$$

This modified orthonormality condition is one of the main geometric properties of the proposed PLS algorithm and is formally stated with proof in Section 5 (see Proposition 5), along with other properties.

To solve (15) in practice, we first perform a Choleski factorization: $LL^\top = J^* + \Lambda \ddot{J}^*$, where L is a lower triangular matrix, and then reduce it to maximizing the classical Rayleigh quotient problem:

$$\max_{\mathbf{e}\in\mathbb{R}^{MK+p}} \frac{\mathbf{e}^\top E\mathbf{e}}{\mathbf{e}^\top\mathbf{e}} \iff \max_{\mathbf{e}\in\mathbb{R}^{MK+p}} \mathbf{e}^\top E\mathbf{e} \quad \text{subject to} \quad \|\mathbf{e}\|^2 = 1$$

through the transformations $\mathbf{e} = L^\top\boldsymbol{\xi}^*$ and $E = L^{-1}\mathbf{V}^*L^{\top-1}$. Once $\mathbf{e}$ is obtained as the first eigenvector of E, we set $\boldsymbol{\xi}^* = L^{\top-1}\mathbf{e}$ and $\widehat{\boldsymbol{\xi}}[\mathbf{t}] = B^*[\mathbf{t}]^\top\boldsymbol{\xi}^*$.

4.2 Step 2: Computing Residualized Predictors via Hybrid-on-Scalar Regression

The second step of the PLS algorithm involves iteratively obtaining the $(l+1)$-th residualized predictors as $\mathbf{W}_i^{[l+1]}[\mathbf{t}] = \mathbf{W}_i^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\delta}}_l[\mathbf{t}]\widehat{\rho}_{il}$, where $\widehat{\boldsymbol{\delta}}_l[\mathbf{t}]$ is the least squares estimate from regressing the previous (l-th) residualized outcome $\mathbf{W}_i^{[l]}[\mathbf{t}]$ on the corresponding estimated PLS score $\widehat{\rho}_{il}$. The pointwise least squares estimate, $\widehat{\boldsymbol{\delta}}_l[\mathbf{t}] = (\sum_{i=1}^n \mathbf{W}_i^{[l]}[\mathbf{t}]\widehat{\rho}_{il})/(\sum_{i=1}^n \widehat{\rho}_{il}^2)$, presented in Section 3, however, is highly inefficient and may not be even feasible for multiple dense and/or irregular functional data. Moreover, the pointwise least squares estimator can show substantial variability across the domain, which becomes problematic for the entire algorithm as its instability passes on to subsequent residualized outcomes and PLS components. Therefore, in this section, we introduce a novel strategy based on a hybrid-on-scalar regression model for obtaining $\widehat{\boldsymbol{\delta}}_l[\mathbf{t}]$ and implementing the second step.

Again, we omit l in the notations. Let $\widehat{\boldsymbol{\rho}} = (\widehat{\rho}_1, \ldots, \widehat{\rho}_n)^\top$ denote an n-dimensional

vector of PLS scores estimated from the proposed algorithm, and let the $(K + P)$-vector $\boldsymbol{\delta}[\mathbf{t}] = [\pi(\mathbf{t})^\top, \boldsymbol{\gamma}^\top]^\top$ denote the regression coefficient consisting of the functional part $\pi(\mathbf{t}) = (\pi^{(1)}(t_1), \ldots, \pi^{(K)}(t_K))^\top \in \mathbb{R}^K$ and scalar part $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \in \mathbb{R}^p$. The hybrid-on-scalar regression model we are interested in fitting is then:

$$\widetilde{W}[\mathbf{t}] = \widehat{\boldsymbol{\rho}}\, \boldsymbol{\delta}[\mathbf{t}]^\top + \mathbf{e}[\mathbf{t}], \tag{17}$$

where $\mathbf{e}[\mathbf{t}]$ is an $n \times (K + p)$ matrix of errors.

As in Section 4.1, the hybrid response can be expanded using basis functions as $\widetilde{W}[\mathbf{t}] = \widetilde{C}^* B^*[\mathbf{t}]$. Similarly, we expand each functional regression coefficient using the basis functions as $\pi^{(k)}(t_k) = \sum_{m=1}^{M} d_m^{(k)} b_m^{(k)}(t_k) = \mathbf{d}^{(k)\top}\mathbf{b}^{(k)}(t_k)$, where $\mathbf{d}^{(k)} = (d_1^{(k)}, \ldots, d_M^{(k)})\top \in \mathbb{R}^M$ are basis coefficients. Here, a different basis system can be flexibly chosen to expand the regression coefficient, but without loss of generality we use the same system as in Section 4.1. With the $(MK + p)$-dimensional row vector $\mathbf{d}^* = (\mathbf{d}^{(1)\top}, \ldots, \mathbf{d}^{(K)\top}, \gamma_1, \ldots, \gamma_p)$ that stacks the basis coefficients and scalar regression coefficients, we can express the linear predictor of the regression as: $\widehat{\boldsymbol{\rho}}\, \boldsymbol{\delta}[\mathbf{t}]^\top = \widehat{\boldsymbol{\rho}}\, \mathbf{d}^* B^*[\mathbf{t}]$. Finally, these basis expansions of the hybrid response and the regression coefficient enable us to re-write the hybrid-on-scalar regression model (17) in a finite-dimensional form:

$$\widetilde{C}^* B^*[\mathbf{t}] = \widehat{\boldsymbol{\rho}}\, \mathbf{d}^* B^*[\mathbf{t}] + \mathbf{e}[\mathbf{t}]. \tag{18}$$

Based on the model formulation (18), we can now formulate the following least squares criterion whose minimizing solution gives the estimate of the regression coefficient $\boldsymbol{\delta}[\mathbf{t}]$ in the

hybrid-on-scalar regression model (17):

$$\begin{aligned} \mathrm{SSE}(\delta) &= \int_{\mathcal{T}} \left\| \widetilde{C}^* B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}\, \mathbf{d}^* B^*[\mathbf{t}] \right\|_F^2 d\mathbf{t} \\ &= tr\left(\widetilde{C}^{*T}\widetilde{C}^* J^*\right) - 2\, tr\left(\mathbf{d}^* J^* \widetilde{C}^{*T}\widehat{\boldsymbol{\rho}}\right) + tr\left(\widehat{\boldsymbol{\rho}}^\top \widehat{\boldsymbol{\rho}}\, \mathbf{d}^* J^* \mathbf{d}^{*T}\right), \end{aligned} \tag{19}$$

where tr denotes the trace of a matrix, and $\|\cdot\|_F$ is a Frobenius norm. Now taking the derivative of (19) with respect to $\mathbf{d}^*$ and setting the result to zero, we find that $\mathbf{d}^*$ satisfies the system of linear equations

$$\widehat{\boldsymbol{\rho}}^\top \widehat{\boldsymbol{\rho}}\, \mathbf{d}^* J^* = \widehat{\boldsymbol{\rho}}^\top C^* J^*,$$

whose solutions lead to the following estimates of $\mathbf{d}^*$ and $\boldsymbol{\delta}[\mathbf{t}]$:

$$\widehat{\mathbf{d}}^* = \widehat{\boldsymbol{\rho}}^\top C^* J^* \left(\widehat{\boldsymbol{\rho}}^\top \widehat{\boldsymbol{\rho}} J^*\right)^{-1} \quad \text{and} \quad \widehat{\boldsymbol{\delta}}[\mathbf{t}] = B^*[\mathbf{t}]^\top \widehat{\mathbf{d}}^{*\top}.$$

Then, we can use $\widehat{\boldsymbol{\delta}}[\mathbf{t}]$ to obtain the subsequent residualized hybrid predictor as $\mathbf{W}_i^{[l+1]}[\mathbf{t}] = \mathbf{W}_i^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\delta}}_l[\mathbf{t}]\widehat{\rho}_{il}$.

The step for obtaining the subsequent residualized response is identical to that of the naive PLS algorithm in in Section 3. Specifically, the $(l-1)$-th residualized response can be obtained as $Y_i^{[l+1]} = Y_i^{[l]} - \widehat{\nu}_l \widehat{\rho}_{il}$, where $\widehat{\nu}_l = \sum_{i=1}^n Y_i^{[l]}\widehat{\rho}_{il} / \sum_{i=1}^n \widehat{\rho}_{il}^2$ is the least square estimate of the simple linear model regressing $Y_i^{[l]}$ on $\widehat{\rho}_{il}$.

4.3 Obtaining the Estimated Hybrid Regression Coefficient

The regression coefficient $\boldsymbol{\eta}$ in model (4) can be written in terms of the estimated PLS components and scores. First, note that we can show $\widehat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} = \langle \mathbf{W}_i, \widehat{\boldsymbol{\zeta}}_l \rangle_{\mathcal{H}}$, where $\widehat{\boldsymbol{\zeta}}_l = \widehat{\boldsymbol{\xi}}_l - \sum_{u=1}^{l-1} \langle \widehat{\boldsymbol{\delta}}_u, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} \widehat{\boldsymbol{\zeta}}_u$ for $l \geq 2$, with $\widehat{\boldsymbol{\zeta}}_1 = \widehat{\boldsymbol{\xi}}_1$. Then, the decomposition of Y_i in (5)

leads to

$$Y_i = \sum_{l=1}^{L} \widehat{\nu}_l \langle \mathbf{W}_i^{[l]}, \widehat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} + \epsilon_i = \left\langle \mathbf{W}_i, \sum_{l=1}^{L} \widehat{\nu}_l \widehat{\boldsymbol{\zeta}}_l \right\rangle_{\mathcal{H}} + \epsilon_i,$$

which, given the uniqueness of $\boldsymbol{\eta}$, leads to

$$\widehat{\boldsymbol{\eta}} = \sum_{l=1}^{L} \widehat{\nu}_l \widehat{\boldsymbol{\zeta}}_l.$$

Here, the number of PLS components L to be estimated can be chosen by cross-validation.

4.4 Data Preprocessing

Functional and scalar elements of the hybrid predictors often have incompatible units and/or exhibit different amounts of variation. This can be problematic for our PLS framework which is not scale invariant as: i) each predictor has different chance of contributing to the predictor/response structure; and ii) a predictor with high correlation to Y but relatively low variance may be overlooked.

To obtain PLS components that have a meaningful interpretation, we standardize the predictor data via the following steps. The first step is to account for discrepancies *within* respective functional and scalar parts, if needed. If elements of multivariate functional data

Algorithm 2 Hybrid PLS (under construction)

Require: Response $\mathbf{y} \in \mathbb{R}^n$, hybrid predictors approximation coefficient $\widetilde{C}^* \in \mathbb{R}^{n \times (MK+p)}$ (22), Gram matrix with respect to basis functions $J^* \in \mathbb{R}^{(MK+p)\times(MK+p)}$ (23), Gram matrix with respect to second derivatives of the basis functions $\ddot{J}^* \in \mathbb{R}^{(MK+p)\times(MK+p)}$ (25), regularization matrix $\Lambda \in \mathbb{R}^{(MK+p)\times(MK+p)}$ (14)

1: $\widetilde{C}^{*[1]} \leftarrow \widetilde{C}^*$
2: **for** l = 1, 2, ...T **do**
3: $\quad \mathbf{V}^* \leftarrow n^{-2} J^* \widetilde{C}^{*\top} \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \widetilde{C}^* J^*$ ▷ (26)
4: $\quad \widehat{\boldsymbol{\xi}}_l^* \leftarrow \arg\max_{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}} \boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^* \quad \text{subject to} \quad \boldsymbol{\xi}^{*\top}(J^* + \Lambda \ddot{J}^*)\boldsymbol{\xi}^* = 1$ ▷ (15)
5: $\quad \widehat{\boldsymbol{\rho}}_l \leftarrow \widetilde{C}^* J^* \widehat{\boldsymbol{\xi}}_l^*$ ▷ $\in \mathbb{R}^n$; (16)
6: Output a

$X_i = (X_i^{(1)}, \ldots, X_i^{(K)})$ are measured in different units or have quite different domains, one can standardize them to have mean zero and integrated variance of one. If multivariate scalar predictors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ exhibit different amounts of variation, one can standardize them to have mean zero and unit variance. The second step is to eliminate the discrepancies *between* functional and scalar parts. To accomplish this aim, we choose an appropriate weight ω in the hybrid inner product (3) that ensures functional and vector parts have comparable variance. A sensible data-driven approach to choosing an appropriate weight is to set

$$\omega = \frac{\sum_{i=1}^n \|X_i\|_{\mathcal{F}}^2}{\sum_{i=1}^n \|\mathbf{Z}_i\|^2},$$

In practice, this weighting scheme can be implemented by formulating the hybrid object as $\mathbf{W} = (X, \omega^{1/2}\mathbf{Z})$, whose vector part has been scaled by a factor of $\omega^{1/2}$.

5 Properties of the Proposed PLS Framework

In this section, we derive some of the theoretical and geometric properties of the proposed PLS framework.

5.1 Tucker's Criterion

We derive that the PLS components obtained from the first step of the proposed algorithm satisfy the Tucker's Criterion (Tucker, 1938) extended to our scalar-on-hybrid regression model setting (4). We omit l in the notations and first define the cross-covariance terms between the response and predictors (observed versions if $l = 1$ and residualized versions if

$l \geq 2$):

$$\sigma_{YX} = (\sigma_{YX}^{(1)}, \ldots, \sigma_{YX}^{(K)}) = (E(YX^{(1)}), \ldots, E(YX^{(K)})) = E(YX) \in \mathcal{F}$$

$$\sigma_{YZ} = (\sigma_{YZ,1}, \ldots, \sigma_{YZ,p})^\top = (E(YZ_1), \ldots, E(YZ_p))^\top = E(Y\mathbf{Z}) \in \mathbb{R}^p$$

$$\Sigma_{YW} = (\sigma_{YX}, \sigma_{YZ}) = (E(YX), E(Y\mathbf{Z})) = E(Y\mathbf{W}) \in \mathcal{H}$$

We also introduce two cross-covariance operators, $\mathcal{C}_{YW} = E(\mathbf{W} \otimes_{\mathcal{H}} Y) : \mathcal{H} \to \mathbb{R}$ and $\mathcal{C}_{WY} = E(Y \otimes \mathbf{W}) : \mathbb{R} \to \mathcal{H}$, which respectively map $h = (f, \mathbf{v}) \in \mathcal{H}$ to $\mathbb{R}$ and $d \in \mathbb{R}$ to $\mathcal{H}$ as follows:

$$\mathcal{C}_{YW}\mathbf{h} = E\{\langle \mathbf{W}, \mathbf{h} \rangle_{\mathcal{H}} Y\} = \langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} = \sum_{k=1}^{K} \int_{\mathcal{T}_k} \sigma_{YX}^{(k)}(t_k) f^{(k)}(t_k) dt_k + E(Y\mathbf{Z}^\top)\mathbf{v}$$

$$\mathcal{C}_{WY} d = E(\langle Y, d \rangle \mathbf{W}) = E(Y\mathbf{W})d = \Sigma_{YW}\, d.$$

Now define a new operator $\mathcal{U} = \mathcal{C}_{WY} \circ \mathcal{C}_{YW} : \mathcal{H} \to \mathcal{H}$, which performs the following mapping:

$$\mathcal{U}\mathbf{h} = \mathcal{C}_{WY}(\mathcal{C}_{YW}\mathbf{h}) = \mathcal{C}_{YW}(\langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}}) = \Sigma_{YW} \langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} = (\Sigma_{YW} \otimes \Sigma_{YW})\mathbf{h}.$$

In other words, $\mathcal{U} = \Sigma_{YW} \otimes \Sigma_{YW}$. The following proposition presents several important properties of $\mathcal{U}$.

Proposition 3 *The operator $\mathcal{U} = \mathcal{C}_{WY} \circ \mathcal{C}_{YW} = \Sigma_{YW} \otimes \Sigma_{YW} : \mathcal{H} \to \mathcal{H}$ is positive and self-adjoint. Furthermore, if there exist finite constants Q_1 and Q_2 such that*

$$\max_{k=1,\ldots,K} \sup_{t_k \in \mathcal{T}_k} \sigma_{YX}^{(k)2}(t_k) < Q_1 \quad \text{and} \quad \max_{r=1,\ldots,p} \sigma_{YZ,r}^2 < Q_2, \tag{20}$$

and if each $\sigma_{YX}^{(k)}$ $(k = 1, \ldots, K)$ is uniformly continuous in a sense that for any $\epsilon > 0$, there

exists $\delta_k > 0$ *such that for every* $t_k, t_k^* \in \mathcal{T}_k$ *with* $|t_k - t_k^*| < \delta_k$*, we have that*

$$|\sigma_{yx}^{(k)}(t_k) - \sigma_{yx}^{(k)}(t_k^*)| < \epsilon,$$

then $\mathcal{U}$ *is a compact operator.*

The proof of Proposition 3 is provided in Appendix E. By the Hilbert-Schmidt theorem (e.g., Theorem 4.2.4 in Hsing and Eubank, 2015), Proposition 3 guarantees the existence of a complete orthonormal system of eigenfunctions $\{\boldsymbol{\xi}_{(u)}\}_{u\in\mathbb{N}}$ of $\mathcal{U}$ in $\mathcal{H}$ such that $\mathcal{U}\boldsymbol{\xi}_{(u)} = \kappa_{(u)}\boldsymbol{\xi}_{(u)}$, where $\{\kappa_{(u)}\}_{u\in\mathbb{N}}$ are the corresponding sequence of eigenvalues that goes to zero as $u \to \infty$, that is, $\kappa_{(1)} \geq \kappa_{(2)} \geq \cdots \geq 0$.

The following proposition introduces the Tucker's Criterion adapted to our scalar-on-hybrid regression model setting based on the aformentioned operators defined in the hybrid space.

Proposition 4

$$\max_{\substack{\boldsymbol{\xi}\in\mathcal{H} \\ \|\boldsymbol{\xi}\|_{\mathcal{H}}=1}} \operatorname{Cov}^2\left(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y\right)$$

is achieved when $\boldsymbol{\xi}$ *is an eigenfunction associated with the largest eigenvalue of* $\mathcal{U}$*.*

The proof of Proposition 4 is provided in Appendix F. In other words, l-th PLS component $\boldsymbol{\xi}_l$ can be obtained as the *first* eigenfunction of $\mathcal{U}^{[l]} = \Sigma_{YW}^{[l]} \otimes \Sigma_{YW}^{[l]}$ whose components are formulated using the l-th residualized response and predictors: $\Sigma_{YW}^{[l]} = E(Y^{[l]}\mathbf{W}^{[l]})$.

5.2 Geometric Properties

In this section, we derive several geometric properties of the PLS components and and scores obtained from the proposed algorithm.

The following proposition states that the PLS components estimated from Proposition 2 are orthonormal with respect to the modified inner product $\langle\langle\cdot\rangle\rangle$ that incorporates the roughness of the functions.

Proposition 5 *The estimated PLS components, $\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \ldots, \widehat{\boldsymbol{\xi}}_L$, from the proposed PLS algorithm are mutually orthonormal with respect to the modified inner product $\langle\langle h_l, h_j\rangle\rangle = \langle h_l, h_j\rangle_{\mathcal{H}} + \sum_{k=1}^{K} \lambda_k \langle \ddot{f}_l^{(k)}, \ddot{f}_j^{(k)}\rangle_{L^2}$ in $\mathcal{H}$ that incorporates the roughness of the functions; i.e.,*

$$\langle\langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j\rangle\rangle = \langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j\rangle_{\mathcal{H}} + \sum_{k=1}^{K} \lambda_k \left\langle \ddot{\widehat{\psi}}_l^{(k)}, \ddot{\widehat{\psi}}_j^{(k)} \right\rangle_{L^2} = \mathbb{1}(l = j).$$

The proof of Proposition 5 is provided in Appendix G. Now let $\widehat{\boldsymbol{\rho}}_l = (\widehat{\rho}_{1l}, \ldots, \widehat{\rho}_{nl})^\top$ denote the n-dimensional vector whose elements consist of estimated l-th PLS scores ($l = 1, \ldots, L$) of n observations. The next proposition states that the vectors of estimated PLS scores $\widehat{\boldsymbol{\rho}}_1, \widehat{\boldsymbol{\rho}}_2, \ldots, \widehat{\boldsymbol{\rho}}_L$ obtained from the proposed algorithm are mutually orthogonal.

Proposition 6 *The vectors of estimated PLS scores, $\widehat{\boldsymbol{\rho}}_1, \widehat{\boldsymbol{\rho}}_2, \ldots, \widehat{\boldsymbol{\rho}}_L$, obtained from the proposed PLS algorithm are mutually orthogonal in the sense that*

$$\widehat{\boldsymbol{\rho}}_l^\top \widehat{\boldsymbol{\rho}}_j = 0 \quad \text{for} \quad l, j \in \{1, \ldots, L\},\ l \neq j.$$

The proof of Proposition 6 is provided in Appendix H.

6 Simulations

To evaluate the superiority of our method under complex dependency structures — specifically, dependencies among functional predictors, among scalar predictors, and between scalar and functional predictors —

Matrix-normal setting. We begin by constructing predictors from a matrix-normal distribution, a framework that offers convenient and flexible control over the dependence structure across both rows and columns. Matrix-normal (MN) models, also known as Kronecker-separable covariance models, provide a principled approach to modeling multivariate data with structured covariance. Specifically, the matrix-normal distribution is defined as

$$\mathbf{X} \sim \mathcal{MN}_{m\times n}(\mathbf{M}; \mathbf{R}, \mathbf{C}),$$

and its log-density is given by

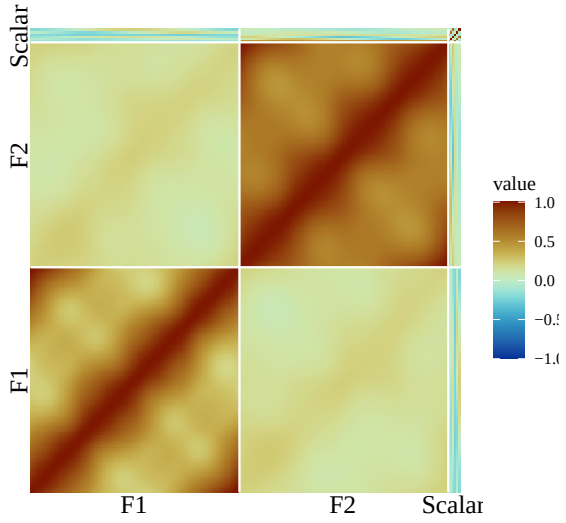
$$\log p(\mathbf{X} \mid \mathbf{M}, \mathbf{R}, \mathbf{C}) = -\frac{mn}{2}\log(2\pi) - \frac{m}{2}\log|\mathbf{C}| - \frac{n}{2}\log|\mathbf{R}| - \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}(\mathbf{X}-\mathbf{M})^\top \mathbf{R}^{-1}(\mathbf{X}-\mathbf{M})\right].$$

The key insight behind Kronecker separability is that if $\mathbf{Y} \sim \mathcal{MN}(\mathbf{M}, \mathbf{R}, \mathbf{C})$, then its vectorized form follows a multivariate normal distribution: $\mathrm{vec}(\mathbf{Y}) \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{C}\otimes\mathbf{R})$, where $\otimes$ denotes the Kronecker product and vec is the vectorization operator.

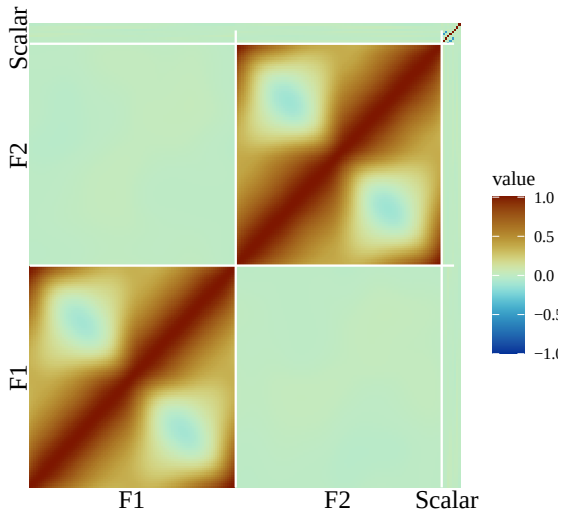
Building on the functional graphical model simulation of Zhu et al. (2016), we generate a mixed graphical model with five nodes, described as follows:

- Nodes F1 and F2: Two functional predictors modeled as Gaussian processes using a truncated Karhunen-Loève expansion, where the eigenbasis consists of Fourier basis functions with a fixed number of basis functions, $M = 9$.
- Nodes S1, S2, and S3: Three scalar predictors, each following an s-dimensional multivariate normal distribution. Unlike the functional predictors, these scalar predictors are modeled directly without basis expansion.

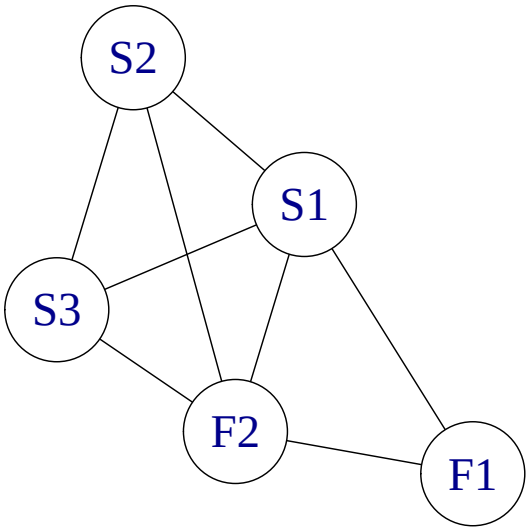
To capture dependencies among predictors, we introduce a graph structure that governs their conditional correlations. We consider two types of graph structures: a weakly con-



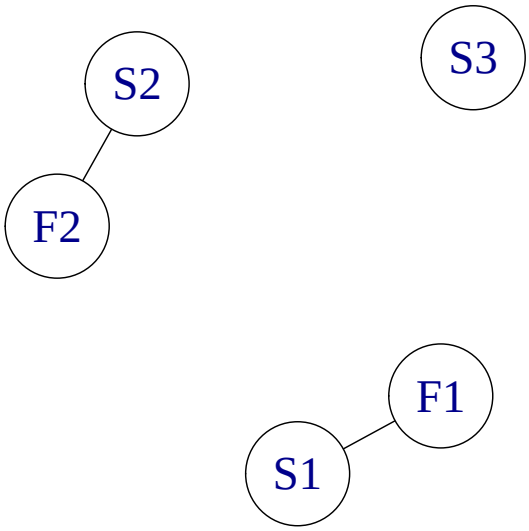
(a) Correlation Matrix - Strong Dependency



(b) Correlation Matrix - Weak Dependency



(c) Graph Structure - Strong Dependency



(d) Graph Structure - Weak Dependency

Figure 1: Comparison of Correlation Matrices and Graph Structures under Strong and Weak Dependencies.

nected graph and a strongly connected graph. In the Gaussian process framework, the precision matrix $\mathbf{R}_0^{-1}$ encodes conditional independence relationships, while its inverse, $\mathbf{R}_0$, represents marginal covariances. This structure extends to a blockwise correlation matrix $\mathbf{R} \in \mathbb{R}^{(2M+3s)\times(2M+3s)}$, where off-diagonal blocks represent correlations between FPC scores and scalar predictor values. Each block (i, j) of $\mathbf{R}$ is given by $(R_0)_{ij}\mathbf{I}_{M_i,M_j}$, where $\mathbf{I}_{M_i,M_j}$ is

a rectangular identity matrix. Here, $M_i = 9$ if node i corresponds to a functional predictor and $M_j = s$ if node j corresponds to a scalar predictor.

For each functional predictor, we assgin M reference eigenvalues (or FPC score) drawn independently from gamma distributions with decreasing means. These reference eigenvalues will later be multiplies by randomly multiplier drawn from correlated multivaraite Normal distribution. These reference eigenvalues are fixed over all samples and all independent repetition of the experiment. We draw it one randomly just to ensure the differentiation between the two functional predictors. we independently draw for each functional predictor from . We then sample zero-mean multivariate normal data from the covariance matrix $\mathbf{R}$. The last $3s$ components are assigned as scalar predictors, while the first $2M$ components, scaled by their corresponding eigenvalues, serve as FPC scores. These scores are then expanded into functional data, evaluated over 100 equally spaced points on $[0, 1]$ using a Fourier basis.

Finally, we introduce structured variability by adding a common mean function, defined as a scaled and shifted sine function. To visualize the generated data, Figure 1 plots functional predictor realizations and heatmaps of sample correlation matrices for both graph structures, illustrating the distinct dependency patterns induced by the precision matrices.

6.1 Simulation shown in february 2025 meeting

To effectively manage the dependencies among functional predictors, scalar covariates, and between functional and scalar predictors, we utilize a Gaussian Markov random field (GMRF) within the framework of Gaussian undirected graphical models. In a GMRF, the off-diagonal elements of the precision matrix capture the conditional correlations between the corresponding components. Given that our setting involves both functional and scalar covariates, we adopt the simulation setup proposed by Kolar et al. (2014), which focuses on mixed attribute Gaussian graphical models.

We construct a graph consisting of two functional predictor nodes and three vector predictor nodes. Below, we sequentially describe the graph structure and the generation process for each node.

Functional Predictors. We consider two functional predictors that share the same precision matrix. Denoted as $\mathbf{\Theta} := (\theta_{t\tilde{t}}) \in \mathbb{R}^{p\times p}$, this precision matrix follows an AR(1) structure with a white noise variance of 0.1^2 and an autoregressive coefficient of $= 0.95$, ensuring a smooth functional trajectory.

More formally, each functional predictor is evaluated at p equally spaced points on $[0, 1]$, with values defined recursively as:

$$X_i^{(k)}(0) = 0, \quad X_i^{(k)}(t) = \rho X_i^{(k)}(t-1) + \varepsilon_{itk}, \quad \varepsilon_{itk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad t = 1, \ldots, p,\ i = 1, \ldots, n,\ k = 1, 2.$$

In this setting, the precision matrix $\mathbf{\Theta}$ exhibits a tridiagonal Toeplitz structure, where the diagonal entries are given by:

$$\theta_{tt} = \begin{cases} \dfrac{1}{\sigma^2(1 - 0.95^2)}, & t = 1, p, \\ \dfrac{1 + 0.95^2}{\sigma^2(1 - 0.95^2)}, & 2 \le t \le p - 1. \end{cases}$$

The off-diagonal entries are:

$$\theta_{t,t+1} = \theta_{t+1,t} = -\frac{\rho}{\sigma^2(1 - \rho^2)}, \quad 1 \le t \le p - 1.$$

Finally, the functional predictors are smoothed using a B-spline basis with 15 basis functions, as described in Section 4.1.

Vector Predictors. We consider s vector predictors, each following a d-dimensional zero-mean Gaussian distribution with a shared precision matrix. This precision matrix, denoted

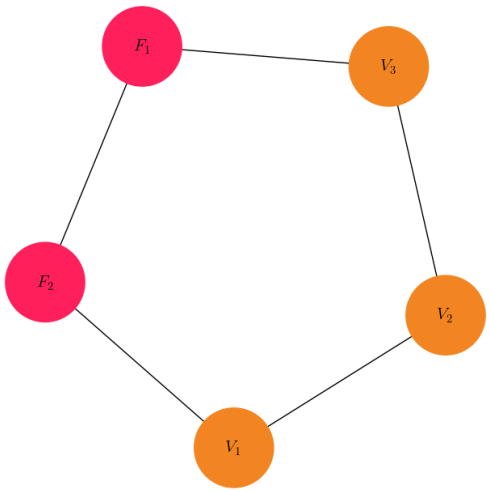


Figure 2

as $\Gamma := (\gamma_{ij}) \in \mathbb{R}^{d\times d}$, follows a Toeplitz structure with exponentially decaying entries, given by:

$$\gamma_{ij} := 0.5^{|i-j|}.$$

Graph Structure. We arrange five nodes in a chain structure, where each node follows a sequential order, and the last node connects back to the first, as illustrated in Figure 2. An edge in the graph indicates that the connected nodes remain correlated when the values of all other nodes are fixed. The resulting marginal dependence structure is significantly more complex than the chain structure itself. We designate nodes F_1 and F_2 as functional predictors and nodes V_1, V_2, and V_3 as vector predictors. To introduce conditional dependencies between the components, we set the off-diagonal blocks of the precision matrix to be $0.5\mathbf{1}$ if the corresponding components are connected in the graph, and zero otherwise. Here, $\mathbf{1}$ denotes a matrix of appropriate dimensions where all elements are equal to 1. Overall, we generate a $2p + 3d$-dimensional multivariate Gaussian distribution with mean zero and a precision matrix Ω, structured as follows:

$$\Omega = \begin{pmatrix} \Omega_F & 0.5\mathbf{1} & 0 & 0 & 0.5\mathbf{1} \\ 0.5\mathbf{1} & \Omega_F & 0.5\mathbf{1} & 0 & 0 \\ 0 & 0.5\mathbf{1} & \Omega_V & 0.5\mathbf{1} & 0 \\ 0 & 0 & 0.5\mathbf{1} & \Omega_V & 0.5\mathbf{1} \\ 0.5\mathbf{1} & 0 & 0 & 0.5\mathbf{1} & \Omega_V \end{pmatrix}$$

For each observation drawn from this multivariate Gaussian distribution, we process

The regression coefficients for the first functional predictor, β_{F_1}, are drawn from a multivariate normal distribution $N(0, 5\mathbf{I}_p)$, with a fixed random seed. The coefficients for the second functional predictor, β_{F_2}, are drawn independently from the same distribution and are also smoothed using a B-spline basis with 10 basis functions. For the vector covariates, the regression coefficients are sampled from $N(0, I_d)$. After calculating the inner product of the covariates and their corresponding coefficients, independent Gaussian noise from $N(0, 0.1^2)$ is added to the generated responses to simulate measurement noise.

The baseline methods are:

- Penalized functional regression (pfr) Goldsmith et al. (2011)
- Principal component regression (fpcr): run both of PCA for multple functional predictors Happ and Greven (2018) and scalar PCA and run OLS on the PC scores.

We compare our method with these baseline methods using $p = 100$. We consider scenarios with $d = 1, 2, 3, 4, 5$ and $n = 100, 200, 300, 400$. For each scenario, we use 70% of the data for training and evaluate prediction performance on the remaining 30% test set, using the root prediction mean squared error as the evaluation metric. For our method and PCR, the maximum number of components are set as 20. The number of components is chosen by 5-fold cross validation. The results, summarized in Table **??**, demonstrate that our method consistently outperforms the baseline methods across all scenarios.

Table 1: Gaussian Markov random field simulation results

Sample Size	Method	$p_{\text{scalar}} = 3$	$p_{\text{scalar}} = 6$	$p_{\text{scalar}} = 9$	$p_{\text{scalar}} = 12$	$p_{\text{scalar}} = 15$
100	fpcr	0.195 (0.028)	0.195 (0.030)	0.207 (0.034)	0.215 (0.036)	0.232 (0.041)
	hybridpls	0.116 (0.016)	0.130 (0.021)	0.162 (0.030)	0.185 (0.032)	0.215 (0.038)
	pfr	0.341 (0.104)	0.323 (0.100)	0.354 (0.102)	0.379 (0.105)	0.417 (0.110)
200	fpcr	0.175 (0.016)	0.171 (0.016)	0.175 (0.017)	0.178 (0.017)	0.182 (0.018)
	hybridpls	0.107 (0.010)	0.110 (0.010)	0.125 (0.014)	0.149 (0.017)	0.162 (0.017)
	pfr	0.201 (0.059)	0.186 (0.031)	0.193 (0.036)	0.204 (0.052)	0.215 (0.064)
300	fpcr	0.169 (0.013)	0.167 (0.012)	0.168 (0.013)	0.171 (0.012)	0.174 (0.013)
	hybridpls	0.104 (0.007)	0.106 (0.007)	0.113 (0.009)	0.137 (0.013)	0.150 (0.013)
	pfr	0.175 (0.018)	0.172 (0.013)	0.173 (0.015)	0.176 (0.013)	0.180 (0.018)
400	fpcr	0.167 (0.011)	0.164 (0.011)	0.167 (0.011)	0.167 (0.011)	0.170 (0.011)
	hybridpls	0.103 (0.006)	0.104 (0.006)	0.110 (0.007)	0.129 (0.011)	0.144 (0.012)
	pfr	0.172 (0.011)	0.169 (0.011)	0.172 (0.011)	0.173 (0.011)	0.175 (0.012)

Figure 3: Enter Caption

7 Data Application

Renal study data. We applied our proposed hybrid functional PLS regression, along with other regression methods, to the Emory renal study data. The study collected data on 226 kidneys (left and right) from 113 subjects, including: (i) baseline renogram curves; (ii) post-furosemide renogram curves; (iii) ordinal ratings of kidney obstruction status (non-obstructed, equivocal, or obstructed) independently assessed by three nuclear medicine experts; (iv) eight kidney-level pharmacokinetic variables derived from radionuclide imaging; and (v) two subject-level variables (age and gender). The subjects had a mean age of 57.8 years (SD = 15.5; range = 18–83), with 54 males (48%) and 59 females (52%). The three experts unanimously classified 153 kidneys as non-obstructed, 5 as equivocal, and 40 as obstructed, while 28 kidneys had discrepant ratings.

The two renogram curves, (i) and (ii), were treated as functional predictors and smoothed using a B-spline basis of order 15. The remaining variables, excluding the diagnosis, were treated as scalar predictors. Given the nature of these variables, we assume they are correlated with the renogram curves but not entirely redundant, as they may contain additional useful information. Finally, the diagnoses provided by the three experts were averaged and

transformed using a min-max logit transformation. We splitted the data into 70% of training data and 30% of testing data, and evalauted the prediction perforamnec by root mean squared error on the test data, normalized by the range of the test data response.

7.1

References

Beyaztas, U. and Shang, H. L. (2020). On function-on-function regression: Partial least squares approach. *Environmental and Ecological Statistics*, 27:95–114.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, 85(1):61–83.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester, West Sussex, UK.

Kolar, M., Liu, H., and Xing, E. P. (2014). Graph estimation from multi-attribute data. *Journal of Machine Learning Research*, 15(51):1713–1750.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B*, 53:233–243.

Silverman, B. W. (1996). Smoothed functional principal components by choice of norm. *The Annals of Statistics*, 24(1):1–24.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842.

Tucker, R. S. (1938). The reasons for price rigidity. *The American Economic Review*, 28(1):41–54.

Wang, Y. (2018). *Partial least squares methods for functional regression models*. PhD thesis, University of North Carolina at Chapel Hill.

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(204):1–27.

A Overview of Appendix

B Notations

- aaa

C Technical lemmas

C.1 old notations

Then we can approximate the functional predictor observations as

$$X_i^{(k)}(t_k) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t_k) = \mathbf{c}_i^{(k)T} \mathbf{b}^{(k)}(t_k), \quad t_k \in \mathcal{T}_k, \quad k = 1, \ldots, K, \quad i = 1, \ldots, n,$$

This basis approximation can be expressed collectively across n observations as: $\widetilde{\mathbf{X}^{(k)}}(t_k) = \widetilde{\mathbf{C}}^{(k)} \mathbf{b}^{(k)}(t_k)$, where $\widetilde{\mathbf{X}}^{(k)}(t_k) = (X_1^{(k)}(t_k), \ldots, X_n^{(k)}(t_k))^\top \in \mathbb{R}^n$, This notation can be further extended to simultaneously express the basis expansions of K functional predictors as

$$\widetilde{X}(\mathbf{t}) = \widetilde{C} B(\mathbf{t}), \quad t \in \mathcal{T}, \tag{21}$$

where $\widetilde{X}(\mathbf{t}) = [\widetilde{\mathbf{X}}^{(1)}(t_1), \cdots, \widetilde{\mathbf{X}}^{(K)}(t_K)] \in \mathbb{R}^{n \times K}$, $\widetilde{C} = [\widetilde{C}^{(1)}, \cdots, \widetilde{\mathbf{C}}^{(k)}] \in \mathbb{R}^{n \times MK}$, and $B(\mathbf{t}) = \text{blkdiag}[\mathbf{b}^{(1)}(t_1), \cdots, \mathbf{b}^{(K)}(t_K)] \in \mathbb{R}^{MK \times K}$.

Furthermore, let $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_n]^\top \in \mathbb{R}^{n \times p}$ denote $n \times p$ matrix of the scalar predictors (observed version when $l = 1$, or residualized version when $l \geq 2$), and let $\widetilde{W}[\mathbf{t}] = [\widetilde{X}(\mathbf{t}), \mathbf{Z}]$ denote $n \times (K + p)$ hybrid predictor matrix that stacks $\widetilde{X}(\mathbf{t})$ and $\mathbf{Z}$ as columns. Then we can also write the hybrid predictor, whose functional part is approximated using the basis functions, as a form of linear transformation similar to (21):

$$\widetilde{W}[\mathbf{t}] = \widetilde{\mathbf{C}}^* B^*[\mathbf{t}], \quad t \in \mathcal{T}. \tag{22}$$

where $B^*[\mathbf{t}] = \text{blkdiag}(B(\mathbf{t}), \mathbf{I}_p) \in \mathbb{R}^{(MK+p) \times (K+p)}$ and $\widetilde{\mathbf{C}}^* = [\widetilde{C}, \mathbf{Z}] \in \mathbb{R}^{n \times (MK+p)}$.

Basis approximation of the PLS component. We can also approximate the functional part of the PLS component in terms of the same basis functions: $\psi^{(k)}(t_k) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t_k) = \mathbf{a}^{(k)\top}\mathbf{b}^{(k)}(t_k)$ and $\psi(\mathbf{t}) = B(\mathbf{t})^\top \mathbf{a} \in \mathbb{R}^K$, where $\mathbf{a}^{(k)} = (a_1^{(k)}, \ldots, a_M^{(k)})^\top \in \mathbb{R}^M$, and $\mathbf{a} = (\mathbf{a}^{(1)\top}, \ldots, \mathbf{a}^{(K)\top})^\top \in \mathbb{R}^{MK}$, so that $\boldsymbol{\xi}[\mathbf{t}] = (\psi(\mathbf{t})^\top, \boldsymbol{\theta}^\top)^\top = (\mathbf{a}^\top B(\mathbf{t}), \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{K+p}$.

Using this definition, we construct the following block-diagonal matrices:

$$\mathbf{J} = \text{blkdiag}(\mathbf{J}^{(1)}, \cdots, \boldsymbol{\Phi}^{(k)}) \in \mathbb{R}^{MK \times MK} \text{ and } \mathbf{J}^* = \text{blkdiag}(\mathbf{J}, \mathbf{I}_p) \in \mathbb{R}^{(MK+p)\times(MK+p)}. \quad (23)$$

Regularization based on second-order derivatives. Our proposed strategy for computing PLS component, detailed in Section 4.1.3, incorporates regularization based on second-order derivatives. To this end, let $\ddot{J}^{(k)}$ denote the $M \times M$ Gram matrix formed by the second derivatives of the basis functions, defined as

$$\ddot{J}^{(k)} = \left[\int_{\mathcal{T}_k} \ddot{b}_m^{(k)}(t) \ddot{b}_n^{(k)}(t)\, dt \right]_{m,n=1}^{M}. \quad (24)$$

and define a block-diagonal matrix $\ddot{J}^* \in \mathbb{R}^{(MK+p)\times(MK+p)}$ as

$$\ddot{J}^* = \text{blkdiag}(\ddot{J}^{(1)}, \cdots, \ddot{J}^{(K)}, 0_{p\times p}). \quad (25)$$

The finite-basis approximation allows us to represent the hybrid predictor observations using the coefficient matrix $\widetilde{\mathbf{C}}^*$, enabling all associated computations to be carried out via matrix operations involving $\widetilde{\mathbf{C}}^*$, J^*, and $\ddot{J}^*$.

D Proofs of Propositions

· Abbreviations

– "CSI": Cauchy–Schwarz inequality

D.1 Proof of new Proposition

$$
\begin{aligned}
\widehat{\mathrm{Cov}}(\langle\widetilde{\mathbf{W}},\boldsymbol{\xi}\rangle_{\mathcal{H}},Y) &= \frac{1}{n}\sum_{i=1}^{n} y_i\langle\widetilde{\mathbf{W}}_i,\boldsymbol{\xi}\rangle_{\mathcal{H}} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i\langle\widetilde{X}_i,\psi\rangle_{\mathcal{F}} + \frac{1}{n}\sum_{i=1}^{n} y_i\langle\mathbf{Z}_i,\boldsymbol{\theta}\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\int_{\mathcal{T}_k} X_i(t_k)\psi^{(k)}(t_k)dt_k\right\} Y_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i^\top\boldsymbol{\theta}Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\int_{\mathcal{T}_k}\left\{\sum_{m=1}^{M} c_{im}^{(k)}b_m^{(k)}(t)\right\}\left\{\sum_{m=1}^{M} a_m^{(k)}b_m^{(k)}(t)\right\}dt_k\right] Y_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i^\top\boldsymbol{\theta}Y_i \\
&= n^{-1}\widetilde{\mathbf{y}}^\top\left[\sum_{k=1}^{K}\widetilde{\mathbf{C}}^{(k)}\left\{\int_{\mathcal{T}_k}\mathbf{b}^{(k)}(t_k)\mathbf{b}^{(k)}(t_k)^\top dt_k\right\}\mathbf{a}^{(k)}\right] + n^{-1}\widetilde{\mathbf{y}}^\top\mathbf{Z}\boldsymbol{\theta} \\
&= n^{-1}\widetilde{\mathbf{y}}^\top(\widetilde{C}J\mathbf{a}) + n^{-1}\widetilde{\mathbf{y}}^\top\mathbf{Z}\boldsymbol{\theta} \\
&= n^{-1}\widetilde{\mathbf{y}}^\top(\widetilde{C}J\mathbf{a} + \mathbf{Z}\boldsymbol{\theta}).
\end{aligned}
$$

D.2 Proof of Propositions 1 and 2

$$
\begin{aligned}
\widehat{\mathrm{Cov}}(\langle\widetilde{\mathbf{W}},\boldsymbol{\xi}\rangle_{\mathcal{H}},Y) &= \frac{1}{n}\sum_{i=1}^{n} y_i\langle\widetilde{\mathbf{W}}_i,\boldsymbol{\xi}\rangle_{\mathcal{H}} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i\langle\widetilde{X}_i,\psi\rangle_{\mathcal{F}} + \frac{1}{n}\sum_{i=1}^{n} y_i\langle\mathbf{Z}_i,\boldsymbol{\theta}\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\int_{\mathcal{T}_k} X_i(t_k)\psi^{(k)}(t_k)dt_k\right\} Y_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i^\top\boldsymbol{\theta}Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\int_{\mathcal{T}_k}\left\{\sum_{m=1}^{M} c_{im}^{(k)}b_m^{(k)}(t)\right\}\left\{\sum_{m=1}^{M} a_m^{(k)}b_m^{(k)}(t)\right\}dt_k\right] Y_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i^\top\boldsymbol{\theta}Y_i \\
&= n^{-1}\widetilde{\mathbf{y}}^\top\left[\sum_{k=1}^{K}\widetilde{\mathbf{C}}^{(k)}\left\{\int_{\mathcal{T}_k}\mathbf{b}^{(k)}(t_k)\mathbf{b}^{(k)}(t_k)^\top dt_k\right\}\mathbf{a}^{(k)}\right] + n^{-1}\widetilde{\mathbf{y}}^\top\mathbf{Z}\boldsymbol{\theta}
\end{aligned}
$$

$$= n^{-1}\widetilde{\mathbf{y}}^\top(\widetilde{C}J\mathbf{a}) + n^{-1}\widetilde{\mathbf{y}}^\top\mathbf{Z}\boldsymbol{\theta}$$

$$= n^{-1}\widetilde{\mathbf{y}}^\top(\widetilde{C}J\mathbf{a} + \mathbf{Z}\boldsymbol{\theta}).$$

we first construct the following matrix:

$$\mathbf{V}^* := n^{-2}\begin{bmatrix} \mathbf{J}\widetilde{\mathbf{C}}^\top\mathbf{y}\mathbf{y}^\top\widetilde{\mathbf{C}}\mathbf{J} & \mathbf{J}\widetilde{\mathbf{C}}^\top\mathbf{y}\mathbf{y}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{y}\mathbf{y}^\top\widetilde{\mathbf{C}}\mathbf{J} & \mathbf{Z}^\top\mathbf{y}\mathbf{y}^\top\mathbf{Z} \end{bmatrix} = n^{-2}\mathbf{J}^*\widetilde{\mathbf{C}}^{*\top}\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top\widetilde{\mathbf{C}}^*\mathbf{J}^* \in \mathbb{R}^{(MK+p)\times(MK+p)}. \quad (26)$$

Here, $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y)$ is expressed as $\boldsymbol{\xi}^{*\top}\mathbf{V}^*\boldsymbol{\xi}^*$ based on the aforementioned basis approximations. Thus, finding $\boldsymbol{\xi}^*$ that maximizes $\boldsymbol{\xi}^{*\top}\mathbf{V}^*\boldsymbol{\xi}^*$ amounts to finding $\widehat{\boldsymbol{\xi}}$ that maximizes the sample squared covariance $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y)$, which is the first step of the PLS algorithm. The constraint $\boldsymbol{\xi}^{*\top}J^*\boldsymbol{\xi}^* = 1$ scales the obtained PLS component to have a unit norm, that is, $\langle\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\xi}}\rangle_{\mathcal{H}} = 1$.

and $\boldsymbol{\xi}^* = (\mathbf{a}^\top, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{MK+p}$. Then, the l-th PLS component, $\widehat{\boldsymbol{\xi}} \equiv \widehat{\boldsymbol{\xi}}_l$ can be obtained by solving

$$\underset{\boldsymbol{\xi}^*\in\mathbb{R}^{MK+p}}{\arg\max}\ \frac{\boldsymbol{\xi}^{*\top}\mathbf{V}^*\boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*\top}J^*\boldsymbol{\xi}^*}, \quad \textit{or equivalently,} \quad \underset{\boldsymbol{\xi}^*\in\mathbb{R}^{MK+p}}{\arg\max}\ \boldsymbol{\xi}^{*\top}\mathbf{V}^*\boldsymbol{\xi}^* \quad \text{subject to} \quad \boldsymbol{\xi}^{*\top}J^*\boldsymbol{\xi}^* = 1, \quad (27)$$

with respect to $\widehat{\boldsymbol{\xi}}^* = (\widehat{\mathbf{a}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$

We show that the solutions to the generalized eigenvalue problems (27) and (15) maximize the covariance with the response. Based on the basis function expansions $X_i^{(k)}(t) = \sum_{m=1}^M c_{im}^{(k)} b_m^{(k)}(t)$ and $\psi^{(k)}(t) = \sum_{m=1}^M a_m^{(k)} b_m^{(k)}(t)$, $k = 1, \ldots, K$, we can express the sample covariance between the PLS score and the outcome, which needs to be maximized to estimate

the corresponding PLS component, as:

$$
\begin{aligned}
\widehat{\mathrm{Cov}}(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y) &= \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{W}_i, \boldsymbol{\xi} \rangle_{\mathcal{H}} Y_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \psi \rangle_{\mathcal{F}} Y_i + \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{Z}_i, \boldsymbol{\theta} \rangle Y_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} \int_{\mathcal{T}_k} X_i(t_k) \psi^{(k)}(t_k) dt_k \right\} Y_i + \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i^\top \boldsymbol{\theta} Y_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \int_{\mathcal{T}_k} \left\{ \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t) \right\} \left\{ \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t) \right\} dt_k \right] Y_i + \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i^\top \boldsymbol{\theta} Y_i \\
&= n^{-1} \widetilde{\mathbf{y}}^\top \left[\sum_{k=1}^{K} \widetilde{\mathbf{C}}^{(k)} \left\{ \int_{\mathcal{T}_k} \mathbf{b}^{(k)}(t_k) \mathbf{b}^{(k)}(t_k)^\top dt_k \right\} \mathbf{a}^{(k)} \right] + n^{-1} \widetilde{\mathbf{y}}^\top \mathbf{Z} \boldsymbol{\theta} \\
&= n^{-1} \widetilde{\mathbf{y}}^\top (\widetilde{C} J \mathbf{a}) + n^{-1} \widetilde{\mathbf{y}}^\top \mathbf{Z} \boldsymbol{\theta} \\
&= n^{-1} \widetilde{\mathbf{y}}^\top (\widetilde{C} J \mathbf{a} + \mathbf{Z} \boldsymbol{\theta}).
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y) &= n^{-2} \widetilde{\mathbf{y}}^\top (\widetilde{C} J \mathbf{a} + \mathbf{Z} \boldsymbol{\theta})(\widetilde{C} J \mathbf{a} + \mathbf{Z} \boldsymbol{\theta})^\top \widetilde{\mathbf{y}} \\
&= n^{-2} \{ \mathbf{a}^\top J \widetilde{C}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \widetilde{C} J \mathbf{a} + \mathbf{a}^\top J \widetilde{C}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \mathbf{Z} \boldsymbol{\theta} + \boldsymbol{\theta} \mathbf{Z}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \widetilde{C} J \mathbf{a} + \boldsymbol{\theta}^\top \mathbf{Z}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \mathbf{Z} \boldsymbol{\theta} \} \\
&= \begin{bmatrix} \mathbf{a}^\top & \boldsymbol{\theta}^\top \end{bmatrix} \begin{bmatrix} n^{-2} J \widetilde{C}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \widetilde{C} J & n^{-2} J \widetilde{C}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \mathbf{Z} \\ n^{-2} \mathbf{Z}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \widetilde{C} J & n^{-2} \mathbf{Z}^\top \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^\top \mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \boldsymbol{\theta} \end{bmatrix} \\
&\equiv \boldsymbol{\xi}^{*T} \mathbf{V}^* \boldsymbol{\xi}^*.
\end{aligned}
$$

Also, the unit norm constraint of the PLS components in $\mathcal{H}$ can be translated into:

$$
\begin{aligned}
\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_{\mathcal{H}} &= \sum_{k=1}^{K} \int_{\mathcal{T}} \psi^{(k)}(t_k) \psi^{(k)}(t_k) dt_k + \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \sum_{k=1}^{K} \int_{\mathcal{T}} \mathbf{a}^{(k)\top} \mathbf{b}^{(k)}(t_k) \mathbf{b}^{(k)}(t_k)^\top \mathbf{a}^{(k)} dt_k + \boldsymbol{\theta}^\top \boldsymbol{\theta}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \mathbf{a}^{(k)\top} J^{(k)} \mathbf{a}^{(k)} + \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \mathbf{a}^\top J \mathbf{a} + \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \boldsymbol{\xi}^{*\top} J^* \boldsymbol{\xi}^* \\
&= 1
\end{aligned}
$$

Thus, the first step of the PLS algorithm translates into solving

$$
\max_{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}} \ \boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^* \quad \text{subject to} \quad \boldsymbol{\xi}^{*\top} J^* \boldsymbol{\xi}^* = 1,
$$

or equivalently solving

$$
\max_{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}} \ \frac{\boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*\top} J^* \boldsymbol{\xi}^*},
$$

where $J^* = \text{blkdiag}[J, \mathbf{I}_p] \in \mathbb{R}^{MK+p}$. In other words, the PLS component is chosen to maximize the squared sample covariance $\widehat{\text{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y) = \boldsymbol{\xi}^{*\top} \mathbf{V}^* \boldsymbol{\xi}^*$ subject to the constraint $\boldsymbol{\xi}^{*\top} J^* \boldsymbol{\xi}^* = 1$.

JM: Another way of computing (27), rather than solving an eigenproblem, is as follows. As shown in the proof in Appendix D.2, the empirical covariance is simplified as:

$$
\begin{aligned}
\widehat{\text{Cov}}(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y) &= \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{W}_i, \boldsymbol{\xi} \rangle_{\mathcal{H}} Y_i \\
&= n^{-1} \widetilde{\mathbf{y}}^\top (\widetilde{C} J \mathbf{a} + \mathbf{Z} \boldsymbol{\theta}). \\
&= \frac{1}{n} \tilde{\mathbf{y}}^\top \begin{bmatrix} \widetilde{C} J & \mathbf{Z} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{a} \\ \theta \end{bmatrix}}_{:= \boldsymbol{\beta}}.
\end{aligned}
$$

Let us denote $\mathbf{b} = \widetilde{\mathbf{y}}^\top \left[\widetilde{C}J\ \mathbf{Z}\right]$. This is the weighted row sum of $\left[\widetilde{C}J\ \mathbf{Z}\right]$, weighted by $\widetilde{\mathbf{y}}$. Then using the symmetricity of J^*, we solve a linear system $J^*\mathbf{d} = \mathbf{b}^\top$ to obtain $\mathbf{d}$. Then we have

$$\begin{aligned}\max_{\sqrt{\boldsymbol{\beta}J^*\boldsymbol{\beta}}=1} \widehat{\mathrm{Cov}}(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y) &= \max_{\sqrt{\boldsymbol{\beta}J^*\boldsymbol{\beta}}=1} \frac{1}{n}\mathbf{d}^\top J^*\boldsymbol{\beta} \\ &\overset{(i)}{\leq} \max_{\sqrt{\boldsymbol{\beta}J^*\boldsymbol{\beta}}=1} \frac{1}{n}\sqrt{\mathbf{d}^\top J^*\mathbf{d}}\sqrt{\boldsymbol{\beta}J^*\boldsymbol{\beta}} \\ &\overset{(ii)}{=} \frac{1}{n}\sqrt{\mathbf{d}^\top J^*\mathbf{d}},\end{aligned}$$

where step (i) uses the Cauchy-Schwarz inequality with the inner product $\langle x, y\rangle_{J^*} = x^\top J^* y$, and step (ii) uses the norm constraint. By the Cauchy-Schwarz inequality, the maximum displayed in the last term is obtained by a unit vector (with respect to $\langle\cdot,\cdot\rangle_{J^*}$) parallel to $\mathbf{d}$. Therefore we have

$$\begin{bmatrix}\mathbf{a}\\ \theta\end{bmatrix} = \frac{1}{\sqrt{\mathbf{d}^\top J^*\mathbf{d}}}\mathbf{d},$$

which only involves solving one linear equation.

E Proof of Proposition 3

$\mathcal{U}$ is *positive* because for every $\mathbf{h} \in \mathcal{H}$,

$$\langle \mathcal{U}\mathbf{h}, \mathbf{h}\rangle_{\mathcal{H}} = \langle \langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}} \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}} = \langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}}^2 \geq 0.$$

$\mathcal{U}$ is *self-adjoint* because for any $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$,

$$\langle \mathcal{U}\mathbf{h}_1, \mathbf{h}_2\rangle_{\mathcal{H}} = \langle \langle \mathbf{h}_1, \Sigma_{YW}\rangle_{\mathcal{H}} \Sigma_{YW}, \mathbf{h}_2\rangle_{\mathcal{H}}$$

$$= \langle \mathbf{h}_1, \Sigma_{YW} \rangle_{\mathcal{H}} \langle \Sigma_{YW}, \mathbf{h}_2 \rangle_{\mathcal{H}}$$

$$= \langle \mathbf{h}_1, \langle \Sigma_{YW}, \mathbf{h}_2 \rangle_{\mathcal{H}} \Sigma_{YW} \rangle_{\mathcal{H}}$$

$$= \langle \mathbf{h}_1, \mathcal{U}\mathbf{h}_2 \rangle_{\mathcal{H}}.$$

Next, We will show that an image of a bounded family in $\mathcal{H}$ under $\mathcal{U}$ is uniformly bounded and equicontinuous, and then apply the Arzelá-Ascoli theorem to show $\mathcal{U}$ is *compact*. Let $\mathcal{B} = \{\mathbf{h} \in \mathcal{H} : \|\mathbf{h}\|_{\mathcal{H}}^2 \leq B\}$ denote a bounded family in $\mathcal{H}$ for some constant $0 < B < \infty$. Clearly, $\mathbf{h} = (f, \mathbf{v}) \in \mathcal{B}$ also implies $\|f\|_{\mathcal{F}}^2 \leq B$ and $\|\mathbf{v}\|^2 \leq B$. Define $\mathcal{I} = \{\mathcal{U}\mathbf{h} : \mathbf{h} \in \mathcal{B}\}$ as the image of $\mathcal{B}$ under $\mathcal{U}$. We will first show that $\mathcal{I}$ is a family of uniformly bounded functions with respect to arguments on $\mathcal{T}$. Let $\mu(\mathcal{T}_k)$ denote a Lebesgue measure of $\mathcal{T}_k$, and let $T = \max_{k=1,\ldots,K} \mu(\mathcal{T}_k)$. Then for any $g \in \mathcal{I}$ and $\mathbf{t} \in \mathcal{T}$,

$$\begin{aligned}
\|g[\mathbf{t}]\|^2 &= \|(\mathcal{U}\mathbf{h})[\mathbf{t}]\|^2 \\
&= \|\langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} \, \Sigma_{YW}[\mathbf{t}]\|^2 \\
&= \left(\langle \sigma_{YX}, f \rangle_{\mathcal{F}} + \langle \sigma_{YZ}, \mathbf{v} \rangle\right)^2 \|[\sigma_{YX}^\top(\mathbf{t}), \sigma_{YZ}^\top]^\top\|^2 \\
&\leq \left(\langle \sigma_{YX}, \sigma_{YX} \rangle_{\mathcal{F}}^{1/2} \langle f, f \rangle_{\mathcal{F}}^{1/2} + \langle \sigma_{YZ}, \sigma_{YZ} \rangle^{1/2} \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}\right)^2 \|[\sigma_{YX}^\top(\mathbf{t}), \sigma_{YZ}^\top]^\top\|^2 \quad (\because \text{CSI}) \\
&= \left[\left\{\sum_{k=1}^{K} \int_{\mathcal{T}_k} \sigma_{YX}^{(k)2}(t_k) dt_k\right\}^{\frac{1}{2}} \|f\|_{\mathcal{F}} + \left(\sum_{r=1}^{p} \sigma_{YZ,r}^2\right)^{\frac{1}{2}} \|\mathbf{v}\|\right]^2 \left(\sum_{k=1}^{K} \sigma_{YX}^{(k)2}(t_k) + \sum_{r=1}^{p} \sigma_{YZ,r}^2\right) \\
&\leq \left\{\left(\sum_{k=1}^{K} \int_{\mathcal{T}_k} Q_1 dt_k\right)^{\frac{1}{2}} B^{\frac{1}{2}} + \left(\sum_{r=1}^{p} Q_2\right)^{\frac{1}{2}} B^{\frac{1}{2}}\right\}^2 \left(\sum_{k=1}^{K} Q_1 + \sum_{r=1}^{p} Q_2\right) \\
&= B\left\{(KTQ_1)^{\frac{1}{2}} + (PQ_2)^{\frac{1}{2}}\right\}^2 (KQ_1 + pQ_2) < \infty.
\end{aligned}$$

We now show that $\mathcal{I}$ is equicontinuous. For $\epsilon > 0$, define

$$\tilde{\epsilon} = \frac{\epsilon}{(KB)^{1/2}\{(KTQ_1)^{1/2} + (PQ_2)^{1/2}\}}$$

By the continuity assumption of $\sigma_{YX}^{(k)}$, there exists $\delta_k > 0$ such that

$$|t_k - t_k^*| < \delta_k \implies |\sigma^{(k)}(t_k) - \sigma^{(k)}(t_k^*)| < \tilde{\epsilon},$$

for all $k = 1, \ldots, K$. Set $\delta = \min_{k=1,\ldots,K} \delta_k$, and let $\|\mathbf{t} - \mathbf{t}^*\| < \delta$, with $\mathbf{t} = (t_1, \ldots, t_K)$ and $\mathbf{t}^* = (t_1^*, \ldots, t_K^*)$. Clearly, $|t_k - t_k^*| < \delta$ for all $k = 1, \ldots, K$, so for $g \in \mathcal{I}$, we can establish that

$$\begin{aligned}
\|g[\mathbf{t}] - g[\mathbf{t}^*]\|^2 &= \|(\mathbf{Uh})[\mathbf{t}] - (\mathbf{Uh})[\mathbf{t}^*]\|^2 \\
&= \|\langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} \Sigma_{YW}[\mathbf{t}] - \langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} \Sigma_{YW}[\mathbf{t}^*]\|^2 \\
&= \langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}}^2 \|\Sigma_{YW}[\mathbf{t}] - \Sigma_{YW}[\mathbf{t}^*]\|^2 \\
&= \left(\langle \sigma_{YX}, f \rangle_{\mathcal{F}} + \langle \sigma_{YZ}, \mathbf{v} \rangle\right)^2 \left\| \begin{bmatrix} \sigma_{YX}(\mathbf{t}) \\ \sigma_{YZ} \end{bmatrix} - \begin{bmatrix} \sigma_{YX}(\mathbf{t}^*) \\ \sigma_{YZ} \end{bmatrix} \right\|^2 \\
&\overset{CSI}{\leq} \left(\langle \sigma_{YX}, \sigma_{YX} \rangle_{\mathcal{F}}^{1/2} \langle f, f \rangle_{\mathcal{F}}^{1/2} + \langle \sigma_{YZ}, \sigma_{YZ} \rangle^{1/2} \langle \mathbf{v}, \mathbf{v} \rangle^{1/2} \right)^2 \left\| \begin{bmatrix} \sigma_{YX}(\mathbf{t}) - \sigma_{YX}(\mathbf{t}^*) \\ \mathbf{0} \end{bmatrix} \right\|^2 \\
&= \left[\left\{ \sum_{k=1}^{K} \int_{\mathcal{T}_k} \sigma_{YX}^{(k)2}(t_k) dt_k \right\}^{\frac{1}{2}} \|f\|_{\mathcal{F}} + \left(\sum_{r=1}^{p} \sigma_{YZ,r}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\| \right]^2 \sum_{k=1}^{K} \left\{ \sigma_{YX}^{(k)}(t_k) - \sigma_{YX}^{(k)}(t_k^*) \right\}^2 \\
&\leq B \left\{ (KTQ_1)^{\frac{1}{2}} + (PQ_2)^{\frac{1}{2}} \right\} K \tilde{\epsilon}^2 = \epsilon^2.
\end{aligned}$$

Now we can apply the Arzelá-Ascoli theorem to conflude that for any bounded sequence $\{\mathbf{h}_n\}_{n \in \mathbb{N}} \in \mathcal{B}$, the sequence $\{\mathbf{g}_n = \mathcal{U}\mathbf{h}_n\}_{n \in \mathbb{N}} \in \mathcal{I}$ contains a convergent subsequence. Therefore, $\mathcal{U}$ is a compact operator.

F Proof of Proposition 4

Before we prove Proposition 4, we state and prove the following lemma:

Lemma 1 *$\mathcal{C}_{WY}$ is an adjoint operator of $\mathcal{C}_{YW}$. That is, $\mathcal{C}_{WY} = \mathcal{C}_{YW}^*$.*

Proof.

First, we have:

$$\langle \mathcal{C}_{YW}\mathbf{h}, d\rangle = E\{\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}} Y\}d$$

Also, we have:

$$\langle \mathbf{h}, \mathcal{C}_{WY}d\rangle_{\mathcal{H}} = \langle \mathbf{h}, E(Y\mathbf{W})d\rangle_{\mathcal{H}} = E\langle \mathbf{h}, Y\mathbf{W}d\rangle_{\mathcal{H}} = E\{\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}} Y\}d.$$

Thus, $\langle \mathcal{C}_{YW}\mathbf{h}, d\rangle = \langle \mathbf{h}, \mathcal{C}_{WY}d\rangle_{\mathcal{H}}$, and this implies $\mathcal{C}_{WY} = \mathcal{C}_{YW}^*$.

Now we prove Proposition 4. The singular value decomposition of $\mathcal{C}_{YW}$ is given by

$$\mathcal{C}_{YW} = \sum_{j=1}^{\infty} \iota_j (f_{1j} \otimes f_{2j}).$$

Let $\|\cdot\|_{op}$ denote an operator norm. Then, applying Theorem 4.3.4 in Hsing and Eubank (2015), we can show that

$$\|\mathcal{C}_{YW}\|_{op} = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} |\mathcal{C}_{YW}\mathbf{h}|^2 = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} |E(\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}} Y)|^2 = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} \mathrm{Cov}^2(\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}}, Y) = \kappa_1^2,$$

with maximum attained at $\mathbf{h} = f_{11}$, which is an eigenfunction of $\mathcal{C}_{YW}^* \circ \mathcal{C}_{YW} = \mathcal{C}_{WY} \circ \mathcal{C}_{YW} = \mathcal{U}$ corresponding to the largest eigenvalue κ_1^2.

G Proof of Proposition 5

Firstly, it is obvious by the unit-norm constraint enforced in Proposition 4 that:

$$\begin{aligned}
\langle\langle \hat{\boldsymbol{\xi}}_l, \hat{\boldsymbol{\xi}}_l \rangle\rangle &= \langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j \rangle_{\mathcal{H}} + \sum_{k=1}^{K} \lambda_k \left\langle \overset{\cdot\cdot}{\widehat{\psi}}{}_l^{(k)}, \overset{\cdot\cdot}{\widehat{\psi}}{}_j^{(k)} \right\rangle_{L^2} \\
&= \sum_{k=1}^{K} \int_{\mathcal{T}_k} \widehat{\psi}_l^{(k)}(t_k) \widehat{\psi}_j^{(k)}(t_k) dt_k + \sum_{r=1}^{p} \widehat{\theta}_{lr} \widehat{\theta}_{jr} + \sum_{k=1}^{K} \lambda_k \int_{\mathcal{T}_k} \overset{\cdot\cdot}{\widehat{\psi}}{}_l^{(k)}(t_k) \overset{\cdot\cdot}{\widehat{\psi}}{}_j^{(k)}(t_k) dt_k \\
&= \sum_{k=1}^{K} \int_{\mathcal{T}_k} \hat{\mathbf{a}}_l^{(k)\top} \mathbf{b}^{(k)}(t_k) \mathbf{b}^{(k)\top}(t_k) \hat{\mathbf{a}}_l^{(k)} dt_k + \boldsymbol{\theta}_l^\top \boldsymbol{\theta}_l + \sum_{k=1}^{K} \lambda_k \int_{\mathcal{T}_k} \hat{\mathbf{a}}_l^{(k)\top} \ddot{\mathbf{b}}^{(k)}(t_k) \ddot{\mathbf{b}}^{(k)\top}(t_k) \hat{\mathbf{a}}_l^{(k)} dt_k \\
&= \sum_{k=1}^{K} \hat{\mathbf{a}}_l^{(k)\top} J^{(k)} \hat{\mathbf{a}}_l^{(k)} + \boldsymbol{\theta}_l^\top \boldsymbol{\theta}_l + \sum_{k=1}^{K} \lambda_k \hat{\mathbf{a}}_l^{(k)\top} \ddot{J}^{(k)} \hat{\mathbf{a}}_l^{(k)} \\
&= \hat{\boldsymbol{\xi}}_l^{*\top} J^* \hat{\boldsymbol{\xi}}_l^* + \hat{\boldsymbol{\xi}}_l^{*\top} \Lambda \ddot{J}^* \hat{\boldsymbol{\xi}}_l^* = \hat{\boldsymbol{\xi}}_l^{*\top} (J^* + \Lambda \ddot{J}^*) \hat{\boldsymbol{\xi}}_l^* = 1
\end{aligned}$$

where $\ddot{\mathbf{b}}^{(k)}(t_k) = (\ddot{b}_1^{(k)}(t_k), \ldots, \ddot{b}_M^{(k)}(t_k))^\top$

Now to prove the orthogonality, we first express the l-th residualized hybrid predictor in terms of its previous $(l-1)$-th version:

$$\widetilde{W}^{[l]}[\mathbf{t}] = \widetilde{W}^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\delta}}_l[\mathbf{t}]^\top$$

which, after applying the finite-dimensional form (9), implies that

$$\begin{aligned}
\widetilde{C}^{*[l]} B^*[\mathbf{t}] &= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\mathbf{d}}_{l-1}^* B^*[\mathbf{t}] \\
&= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}^\top \widetilde{C}^{*[l-1]} J^* \left(\widehat{\boldsymbol{\rho}}_{i,l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1} J^* \right)^{-1} B^*[\mathbf{t}] \\
&= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \left(\frac{\widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1}} \right) \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] \\
&= \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1}} \right) \widetilde{C}^{*[l-1]} B^*[\mathbf{t}].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\widetilde{W}^{[l]}[\mathbf{t}] &= \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}} \right) \widetilde{W}^{[l-1]}[\mathbf{t}] \\
&= \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}} \right) \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-2}\widehat{\boldsymbol{\rho}}_{l-2}^\top}{\widehat{\boldsymbol{\rho}}_{l-2}^\top\widehat{\boldsymbol{\rho}}_{l-2}} \right) \widetilde{W}^{[l-2]}[\mathbf{t}] \\
&= P \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{j}\widehat{\boldsymbol{\rho}}_{j}^\top}{\widehat{\boldsymbol{\rho}}_{j}^\top\widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{W}^{[j]}[\mathbf{t}]
\end{aligned}$$

where P is an $n \times n$ matrix defined as

$$P = \prod_{h=j+1}^{l-1} \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{h}\widehat{\boldsymbol{\rho}}_{h}^\top}{\widehat{\boldsymbol{\rho}}_{h}^\top\widehat{\boldsymbol{\rho}}_{h}} \right)$$

Then for $j < l$,

$$\begin{aligned}
\int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}]\widehat{\boldsymbol{\xi}}_j[\mathbf{t}]d\mathbf{t} &= \int_{\mathcal{T}} P \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{j}\widehat{\boldsymbol{\rho}}_{j}^\top}{\widehat{\boldsymbol{\rho}}_{j}^\top\widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\boldsymbol{\xi}}_k[\mathbf{t}]d\mathbf{t} \\
&= P \left\{ \int_{\mathcal{T}} \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\boldsymbol{\xi}}_j[\mathbf{t}]d\mathbf{t} - \frac{\widehat{\boldsymbol{\rho}}_{j}\widehat{\boldsymbol{\rho}}_{j}^\top}{\widehat{\boldsymbol{\rho}}_{j}^\top\widehat{\boldsymbol{\rho}}_{j}} \int_{\mathcal{T}} \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\boldsymbol{\xi}}_j[\mathbf{t}]d\mathbf{t} \right\} \\
&= P \left(\widehat{\boldsymbol{\rho}}_j - \frac{\widehat{\boldsymbol{\rho}}_{j}\widehat{\boldsymbol{\rho}}_{j}^\top}{\widehat{\boldsymbol{\rho}}_{j}^\top\widehat{\boldsymbol{\rho}}_{j}} \widehat{\boldsymbol{\rho}}_j \right) \\
&= P(\widehat{\boldsymbol{\rho}}_j - \widehat{\boldsymbol{\rho}}_j) \\
&= \mathbf{0} \in \mathbb{R}^n,
\end{aligned}$$

noting that $\int_{\mathcal{T}} \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\boldsymbol{\xi}}_j[\mathbf{t}]d\mathbf{t} = \widehat{\boldsymbol{\rho}}_j$. Now by the generalized eigenvalue problem (i.e., generalized Rayleigh quotient problem) presented in Proposition 2, we have that for $j < l$,

$$\begin{aligned}
\langle\langle \widehat{\boldsymbol{\xi}}_l, \widehat{\boldsymbol{\xi}}_j \rangle\rangle &= \widehat{\boldsymbol{\xi}}_l^{*\top} (J^* + \Lambda \ddot{J}^*) \widehat{\boldsymbol{\xi}}_j^* \\
&= \frac{1}{\kappa_l} (V_l^* \widehat{\boldsymbol{\xi}}_l^*)^\top \widehat{\boldsymbol{\xi}}_j^*
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\kappa_l} \widehat{\boldsymbol{\xi}}_l^{*\top} V_l^* \widehat{\boldsymbol{\xi}}_j^* \\
&= \frac{1}{\kappa_l} \widehat{\boldsymbol{\xi}}_l^{*\top} \left(n^{-2} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \widetilde{C}^{*[l]} J^* \right) \widehat{\boldsymbol{\xi}}_j^* \\
&= \frac{1}{n^2 \kappa_l} \widehat{\boldsymbol{\xi}}_l^{*\top} \; J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \widetilde{C}^{*[l]} \left(\int_{\mathcal{T}} B^*[\mathbf{t}] B^*[\mathbf{t}]^\top d\mathbf{t} \right) \widehat{\boldsymbol{\xi}}_j^* \\
&= \frac{1}{n^2 \kappa_l} \widehat{\boldsymbol{\xi}}_l^{*\top} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \left(\int_{\mathcal{T}} \widetilde{C}^{*[l]} B^*[\mathbf{t}] B^*[\mathbf{t}]^\top \widehat{\boldsymbol{\xi}}_j^* d\mathbf{t} \right) \\
&= \frac{1}{n^2 \kappa_l} \widehat{\boldsymbol{\xi}}_l^{*\top} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \left(\int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}] \widehat{\boldsymbol{\xi}}_j[\mathbf{t}] d\mathbf{t} \right) \\
&= 0,
\end{aligned}
$$

where the last equality follows from the above result that $\int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}] \widehat{\boldsymbol{\xi}}_j[\mathbf{t}] d\mathbf{t} = \mathbf{0}$.

H Proof of Proposition 6

First note that $\widehat{\boldsymbol{\rho}}_l = \int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}] \widehat{\boldsymbol{\xi}}_l[\mathbf{t}] d\mathbf{t} = \widetilde{C}^{*[l]} J^* \widehat{\boldsymbol{\xi}}_l^*$ for $l = 1, \ldots, L$. Then, as above, we first express the l-th residualized hybrid predictor in terms of its previous $(l-1)$-th version:

$$
\widetilde{W}^{[l]}[\mathbf{t}] = \widetilde{W}^{[l]}[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\delta}}_l[\mathbf{t}]^\top
$$

which, after applying the finite-dimensional form (9), implies that

$$
\begin{aligned}
\widetilde{C}^{*[l]} B^*[\mathbf{t}] &= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\mathbf{d}}_{l-1}^* B^*[\mathbf{t}] \\
&= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}^\top \widetilde{C}^{*[l-1]} J^* \left(\widehat{\boldsymbol{\rho}}_{i,l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1} J^* \right)^{-1} B^*[\mathbf{t}] \\
&= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \frac{\widehat{\boldsymbol{\rho}}_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}^\top \widetilde{C}^{*[l-1]} B^*[\mathbf{t}]}{\widehat{\boldsymbol{\rho}}_{l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1}} \\
&= \widetilde{C}^{*[l-1]} B^*[\mathbf{t}] - \frac{\widetilde{C}^{*[l-1]} J^* \widehat{\boldsymbol{\xi}}_{l-1}^* \widehat{\boldsymbol{\rho}}_{l-1}^\top \widetilde{C}^{*[l-1]} B^*[\mathbf{t}]}{\widehat{\boldsymbol{\rho}}_{l-1}^\top \widehat{\boldsymbol{\rho}}_{l-1}},
\end{aligned}
$$

Since the equation should hold for all argument values $\mathbf{t} \in \mathcal{T}$, we can write:

$$\begin{aligned}
\widetilde{C}^{*[l]} &= \widetilde{C}^{*[l-1]} - \frac{\widetilde{C}^{*[l-1]} J^* \widehat{\boldsymbol{\xi}}^*_{l-1} \widehat{\boldsymbol{\rho}}^\top_{l-1} \widetilde{C}^{*[l-1]}}{\widehat{\boldsymbol{\rho}}^\top_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}} \\
&= \widetilde{C}^{*[l-1]} \left(I_{MK+p} - \frac{J^* \widehat{\boldsymbol{\xi}}^*_{l-1} \widehat{\boldsymbol{\rho}}^\top_{l-1} \widetilde{C}^{*[l-1]}}{\widehat{\boldsymbol{\rho}}^\top_{l-1} \widehat{\boldsymbol{\rho}}_{l-1}} \right) \\
&= \widetilde{C}^{*[j]} \left(I_{MK+p} - \frac{J^* \widehat{\boldsymbol{\xi}}^*_{j} \widehat{\boldsymbol{\rho}}^\top_{j} \widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}^\top_{j} \widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{P},
\end{aligned}$$

where $\widetilde{P}$ is a $(MK+p) \times (MK+p)$ matrix defined as

$$\widetilde{P} = \prod_{h=j+1}^{l-1} \left(I_{MK+p} - \frac{J^* \hat{\boldsymbol{\xi}}^*_{h} \widehat{\boldsymbol{\rho}}^\top_{h} \widetilde{C}^{*[h]}}{\widehat{\boldsymbol{\rho}}^\top_{h} \widehat{\boldsymbol{\rho}}_{h}} \right)$$

Then, for $j < l$,

$$\begin{aligned}
\widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[l]} &= \widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} \left(I_{MK+p} - \frac{J^* \widehat{\boldsymbol{\xi}}^*_{j} \widehat{\boldsymbol{\rho}}^\top_{j} \widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}^\top_{j} \widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{P} \\
&= \left(\widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} - \frac{\widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} J^* \widehat{\boldsymbol{\xi}}^*_{j} \widehat{\boldsymbol{\rho}}^\top_{j} \widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}^\top_{j} \widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{P} \\
&= \left(\widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} - \frac{\widehat{\boldsymbol{\rho}}^\top_j \widehat{\boldsymbol{\rho}}_j \widehat{\boldsymbol{\rho}}^\top_{j} \widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}^\top_{j} \widehat{\boldsymbol{\rho}}_{j}} \right) \widetilde{P} \\
&= \left(\widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} - \widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[j]} \right) \widetilde{P} \\
&= \mathbf{0} \in \mathbb{R}^{1 \times (MK+p)},
\end{aligned}$$

Therefore, for $j < l$,

$$\widehat{\boldsymbol{\rho}}^\top_j \widehat{\boldsymbol{\rho}}_l = \widehat{\boldsymbol{\rho}}^\top_j \widetilde{C}^{*[l]} J^* \widehat{\boldsymbol{\rho}}_l = \mathbf{0} J^* \widehat{\boldsymbol{\rho}}_l = 0.$$

Thus, the proposition is proved.

I Simulations

References

Beyaztas, U. and Shang, H. L. (2020). On function-on-function regression: Partial least squares approach. *Environmental and Ecological Statistics*, 27:95–114.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, 85(1):61–83.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester, West Sussex, UK.

Kolar, M., Liu, H., and Xing, E. P. (2014). Graph estimation from multi-attribute data. *Journal of Machine Learning Research*, 15(51):1713–1750.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B*, 53:233–243.

Silverman, B. W. (1996). Smoothed functional principal components by choice of norm. *The Annals of Statistics*, 24(1):1–24.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842.

Tucker, R. S. (1938). The reasons for price rigidity. *The American Economic Review*, 28(1):41–54.

Wang, Y. (2018). *Partial least squares methods for functional regression models*. PhD thesis, University of North Carolina at Chapel Hill.

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(204):1–27.