PARTIAL LEAST SQUARES REGRESSION WITH MULTIPLE FUNCTIONAL AND SCALAR PREDICTORS

BY JONGMIN MUN [1,a] AND JEONG HOON JANG [2,b]

[1]*University of Southern California,* [a]*jongmin.mun@marshall.usc.edu*

[2]*University of Texas Medical Branch,* [b]*jejang@utmb.edu*

Modern biomedical studies increasingly collect both functional and vector-valued predictors for each subject, introducing challenges due to high dimensionality and cross-modality correlations. Existing methods often model these components separately, failing to capture their joint predictive structure. We propose a hybrid partial least squares (PLS) regression framework that integrates functional and scalar predictors within a unified Hilbert space. By iteratively maximizing empirical covariance with the response, our method identifies low-dimensional directions that capture both within- and between-modality variation. The procedure is computationally efficient, requiring only to solve linear systems at each step. We provide theoretical guarantees and demonstrate the effectiveness of our approach through simulations and an application to clinical outcome prediction using renal imaging and scalar covariates from the Emory University renal study.

1. Introduction. Modern biomedical studies frequently collect diverse data types from each subject. As an illustrative example, the Emory University renal study (Chang et al., 2020; Jang, 2021) records both multiple renogram curves (functional data) and multiple renogram variables (vector data) for each kidney. These modalities provide distinct yet related physiological signals. To effectively analyze such heterogeneous data, we need statistical methods capable of jointly modeling functional and vector-valued covariates. This approach accounts for both shared and modality-specific variations linked to clinical outcomes. Our work focuses on developing a method to fit the following hybrid functional linear regression model:

$$(1) \qquad Y_i = \boldsymbol{\beta}^\top \mathbf{Z}_i + \sum_{j=1}^{K} \int \beta_j(t) X_{ij}(t) dt + \epsilon_i, \quad i = 1, \ldots, n,$$

where Y is a scalar response, $\mathbf{Z}$ is the Euclidean vector covariate and $X_{i1}(t), \ldots, X_{iK}(t)$ are functional predictors. For notational convenience, we assume throguth this paper that the responsd sand preedictoras ahve been centered so the intercept can be ignored.

Regression with functional and scalar predictors poses several challenges. Estimating the infinite-dimensional slope function is an ill-posed problem requiring regularization or dimension reduction. Scalar predictors are often high-dimensional and multicollinear, further complicating estimation. Moreover, strong correlations between functional and scalar predictors are common but typically overlooked by separate modeling approaches. These issues call for a unified framework that addresses ill-posedness, high-dimensionality, and cross-modality correlation jointly.

Many approaches have been proposed 1. Roughness Penalty 2. basis approach (power-series, B-splines, wavelets) - Power Series splines (Goldsmith et al., 2011) - B-splines (Cardot et al., 2003; Cai and Hall, 2006) - Wavelet(Zhao et al., 2012)

One possible approach is to apply principal component analysis, FPCA or MFPCA for the functional predictors and standard PCA for the scalar predictors, followed by classical multivariate regression on the resulting scores. This PCA regression strategy has been well-studied for functional predictors alone (Hall and Horowitz, 2007; Reiss and Ogden, 2007; Febrero-Bande et al., 2017), but it overlooks potentially strong correlations between functional and scalar components, leading to multicollinearity and suboptimal predictive performance. Outside the regression context, Jang (2021) proposed a joint PCA method that accounts for correlation between functional and scalar data. However, the resulting components are not informed by the response and may fail to capture directions of maximal correlation with the outcome.

Partial least squares (PLS) regression offers a powerful alternative by explicitly seeking low-dimensional projections of the predictors that maximize their covariance with the outcome. While PLS has been extensively studied in the context of functional data, both for single and multiple functional predictors (Preda and Saporta, 2005; Reiss and Ogden, 2007; Aguilera et al., 2010; Delaigle and Hall, 2012; Aguilera et al., 2016; Febrero-Bande et al., 2017; Beyaztas and Lin Shang, 2022; Saricam et al., 2022; Mutis et al., 2025), a unified framework capable of simultaneously handling mixed modalities of functional and scalar covariates remains underdeveloped.

In this paper, we propose a hybrid partial least squares (PLS) regression framework that integrates functional and vector predictors in a principled and coherent manner. To extract predictive structure from jointly observed and potentially correlated data types, we define a Hilbert space that treats functional and vector components as a single hybrid object, equipped with a suitable inner product. The hybrid PLS direction is then obtained by iteratively maximizing the empirical covariance with the response, subject to a unit-norm constraint in this Hilbert space.

The framework is readily applicable to dense and irregular functionaldata, and supports regularization techniques to prevent overfitting and reduce variance.

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 $\mathcal{F} = L^2(\mathcal{T}_1) \times \cdots \times L^2([0,1])$—a cartesian product of individual $L^2([0,1])$ 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 $\mathbf{t} = (t_1, \ldots, t_k)^\top \in \mathcal{T} = \mathcal{T}_1 \times \cdots \times [0,1]$ 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 $\langle f_1, f_2 \rangle_{\mathcal{F}} = \sum_{k=1}^{K} \langle f_1^{(k)}, f_2^{(k)} \rangle_{L^2} = \sum_{k=1}^{K} \int_0^1 f_1^{(k)}(t_k) f_2^{(k)}(t_k) dt_k$ with norm $\|f_1\|_{\mathcal{F}} = \langle f_1, f_1 \rangle_{\mathcal{F}}^{1/2} = \{\sum_{k=1}^{K} \int_0^1 f_1^{(k)}(t_k)^2 dt_k\}^{1/2}$.

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 $\|\mathbf{v_1}\| = \langle \mathbf{v_1}, \mathbf{v_1} \rangle^{1/2} = (\sum_{r=1}^{p} v_{1r}^2)^{1/2}$.

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 2.1 (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

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle_{\mathcal{H}} := \langle f_1, f_2 \rangle_{\mathcal{F}} + \omega \langle \mathbf{v_1}, \mathbf{v_2} \rangle = \sum_{k=1}^{K} \int_0^1 f_1^{(k)}(t_k) f_2^{(k)}(t_k) dt_k + \omega \sum_{r=1}^{p} v_{1r} v_{2r}, \tag{3}$$

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 4.5). 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 $\rho_l = \langle \mathbf{W}^{[l]}, \xi_l \rangle_{\mathcal{H}}$, 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 $\xi_l = (\psi_l, \boldsymbol{\theta}_l) \in \mathcal{H}$ 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., $\mathrm{Cov}^2(\rho_l, Y^{[l]}) = \mathrm{Cov}^2(\langle \mathbf{W}^{[l]}, \xi_l \rangle_{\mathcal{H}}, Y^{[l]})$. 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 $\{\xi_l\}_{l=1}^{\infty}$, are orthonormal in a sense that $\langle \xi_l, \xi_j \rangle_{\mathcal{H}} = \mathbb{K}(l = j)$.

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).

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 $\mathbf{t}_1, \ldots, \mathbf{t}_m \in [0,1]^K$

1: **for** $i = 1, 2, \ldots n$ **do** $\triangleright$ Centering (6)

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

3: **for** $l = 1, 2, \ldots L$ **do**

  PLS component computation step:

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

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

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

  Orthogonalization step:

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

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

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

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

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

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

---

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 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

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

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

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

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

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

where the normalizing factor is computed as

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

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 4.2 and 4.3, respectively. After a suitable number of iterations, the hybrid regression coefficient is estimated, as described in Section 4.4. 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. *Preliminary step: Finite-basis Approximation*. Let $b_m(t)$ be a twice-differentiable basis of $L^2([0,1])$ whose second derivatives are also linearly independent, for example, cubic B-splines, the Fourier basis, or an orthonormal polynomial basis of degree greater than three. Using this basis, the jth functional predictor, regression coefficient, PLS component direction, and orthogonalization regression coefficient (with iteration indices suppressed) are represented as follows:

$$
X_{ij}(t) = \sum_{m=1}^{\infty} \theta_{ijm} b_m(t), \ \beta_j(t) = \sum_{m=1}^{\infty} \eta_{jm} b_m(t), \ \xi_j(t) = \sum_{m=1}^{\infty} \gamma_{jm} b_m(t), \ \delta_j(t) = \sum_{m=1}^{\infty} \pi_{jm} b_m(t)
$$

In practice, the full set of coefficients can not be obtained with finite sample size, as functional data are measured on a finite grid. Thus, we truncate the expansion at M terms and define the truncated basis space $\mathcal{B} := \mathrm{span}\big(b_1(t), \ldots, b_M(t)\big)$. We choose a moderately large M (e.g., 15 or 20) to capture functional variation without fine-tuning, as smoothness is handled via

penalization (see Section 4.2.2). The truncated expansions of the predictor and coefficient are denoted as

$$\widetilde{X}_{ij}(t) := \sum_{m=1}^{M} \theta_{ijm}\, b_m(t), \quad \tilde{\beta}_j(t) := \sum_{m=1}^{M} \eta_{jm}\, b_m(t),$$

and our suggest method restricts each PLS component direction and orthogonalization regression coefficient to admit the following expansion:

$$(10) \qquad \xi_j(t) = \sum_{m=1}^{M} \gamma_{jm}\, b_m(t), \quad \delta_j(t) = \sum_{m=1}^{M} \pi_{jm} b_m(t)$$

This implies that all computations in this paper are carried out entirely within the subspace

$$(11) \qquad \widetilde{\mathcal{H}} := \mathcal{B}^K \times \mathbb{R}^p \subset \mathcal{H}.$$

The ith hybrid predictor, projected on $\widetilde{\mathcal{H}}$, is represented by the tuple

$$\widetilde{W}_i := (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{iK}, \mathbf{Z}_i).$$

Let $\boldsymbol{\theta}_{ij}$, $\boldsymbol{\eta}_j$, $\boldsymbol{\gamma}_j$, and $\boldsymbol{\pi}_j$ denote the M-dimensional vectors of coefficients:

$$(12)$$
$$\boldsymbol{\theta}_{ij} := (\theta_{ij1}, \ldots, \theta_{ijM})^\top, \; \boldsymbol{\eta}_j := (\eta_{j1}, \ldots, \eta_{jM})^\top, \; \boldsymbol{\gamma}_j := (\gamma_{j1}, \ldots, \gamma_{jM})^\top, \; \boldsymbol{\pi}_j := (\pi_{j1}, \ldots, \pi_{jM})^\top$$

For the predictors, we stack the coefficient vectors across observations into the matrix

$$(13) \qquad \Theta_j := (\boldsymbol{\theta}_{1j}, \ldots, \boldsymbol{\theta}_{nj})^\top \in \mathbb{R}^{n \times M}.$$

The full data for the hybrid PLS problem can be represented by the tuple

$$(14) \qquad (\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y}) \in \mathbb{R}^{(MK+p) \times (MK+p)} \times \mathbb{R}^{(MK+p) \times (MK+p)} \times \mathbb{R}^{n \times (MK+p)} \times \mathbb{R}^n,$$

where $B, B'' \in \mathbb{R}^{M \times M}$ are the Gram matrices of the basis functions and their second derivatives, with entries

$$(15) \qquad B_{m,m'} := \int_0^1 b_m(t)\, b_{m'}(t)\, dt, \quad B''_{m,m'} := \int_0^1 b''_m(t)\, b''_{m'}(t)\, dt,$$

for $m, m' = 1, \ldots, M$. We then define the block-diagonal matrices

$$(16) \qquad \mathbb{B} := \mathrm{blkdiag}(B, \ldots, B, I_p), \quad \mathbb{B}'' := \mathrm{blkdiag}(B'', \ldots, B'', I_p),$$

the full design matrix $\Theta := (\Theta_1, \ldots, \Theta_K, Z) \in \mathbb{R}^{n \times (MK+p)}$, and the response vector $\mathbf{y} := (y_1, \ldots, y_n)^\top$.

REMARK 2. While different bases could be used for each functional predictor, we adopt a common basis for simplicity. The definitions of $\mathbb{B}$ and $\mathbb{B}''$ remain general enough to accommodate distinct bases if needed.

4.2. *Step 1: Computing smoothed PLS component direction.* We begin with the first step of the iterative algorithm: estimating the smooth PLS component direction. Section 4.2.1 introduces the core optimization problem for this step, formulated as a generalized Rayleigh quotient. Building on this, Section 4.2.2 presents a modified version that promotes smoothness, which is the estimation procedure for our proposed method, along with a efficient computation scheme.

4.2.1. *Core optimization problem for PLS component direction estimation.* Building on the basis function approximation in Section 4.1, we now present the core optimization problem that directly estimates the hybrid PLS component direction. This strategy fully leverages the continuous nature of the functional components and the function-scalar hybrid structure, and can be implemented using simple matrix-vector multiplications.

We describe the strategy at the l-th iteration. The PLS component direction is estimated by the unit-norm direction $\xi^{[l]} \in \widetilde{\mathcal{H}}$ that maximizes the squared empirical covariance, which quantifies the linear dependence between the PLS scores $\langle \widetilde{W}_1, \xi \rangle_{\mathcal{H}}, \ldots, \langle \widetilde{W}_n, \xi \rangle_{\mathcal{H}}$ and the responses $y_1, \ldots, y_n$, defined as

$$(17) \qquad \widehat{\mathrm{Cov}}(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y) := \frac{1}{n} \sum_{i=1}^{n} y_i \langle \widetilde{W}_i, \xi \rangle_{\mathcal{H}}.$$

We denote the estimated PLS component direction as

$$(18) \qquad \hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_K, \hat{\zeta}) := \arg\max_{\xi \in \widetilde{\mathcal{H}}} \widehat{\mathrm{Cov}}^2 \big(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y \big) \ s.t. \ \|\xi\|_{\mathcal{H}} = 1,$$

where each functional part $\hat{\xi}_j$ is expanded as $\hat{\xi}_j(t) := \sum_{m=1}^{M} \hat{\gamma}_{jm} b_m(t)$, and $\hat{\zeta}$ is the scalar part. This maximizer can also be represented as a concatenated Euclidean vector

$$(19) \qquad (\hat{\boldsymbol{\gamma}}_1^{\top}, \ldots, \hat{\boldsymbol{\gamma}}_K^{\top}, \hat{\boldsymbol{\zeta}}^{\top})^{\top} \in \mathbb{R}^{MK+p},$$

where the coefficient vectors $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_K$ are defined as in (12). Proposition 4.1 reformulates the problem (18) as a generalized Rayleigh quotient.

PROPOSITION 4.1. *Let* $(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y})$ *denote the observed data defined in* (14). *At the l-th iteration of the PLS algorithm, the unit-norm maximizer of the squared empirical covariance defined in* (18)*, represented as the Euclidean vector form as in* (19)*, is obtained by solving*

$$(20) \qquad \left(\hat{\boldsymbol{\gamma}}_1^{\top}, \ldots, \hat{\boldsymbol{\gamma}}_K^{\top}, \hat{\boldsymbol{\zeta}}^{\top} \right)^{\top} = \arg\max_{\boldsymbol{\xi} \in \mathbb{R}^{MK+p}} \boldsymbol{\xi}^{\top} V \boldsymbol{\xi} \quad \text{subject to} \quad \boldsymbol{\xi}^{\top} \mathbb{B} \boldsymbol{\xi} = 1.$$

where

$$(21) \qquad V := \frac{1}{n^2} (\mathbb{B}\Theta^{\top}\mathbf{y})(\mathbb{B}\Theta^{\top}\mathbf{y})^{\top} \in \mathbb{R}^{(MK+p) \times (MK+p)}.$$

The proof of Proposition 4.1 is provided in Appendix D.

4.2.2. *Proposed Approach: Regularized and Smoothed Computation.* Although Proposition 4.1 efficiently estimates the PLS component direction $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_K, \hat{\zeta})$, the functional components $\hat{\xi}_1, \ldots, \hat{\xi}_K$ may not be smooth, potentially leading to overfitting and unstable predictions. To address this, we propose a regularized extension that balances predictive performance with smoothness. Specifically, we penalize the roughness of each $\hat{\xi}_j$ using its integrated squared second derivative

$$(22) \qquad \mathrm{PEN}(\hat{\xi}_j) := \int_0^1 \big\{ \hat{\xi}_j''(t) \big\}^2 dt.$$

Instead of solely maximizing the squared empirical covariance $\widehat{\mathrm{Cov}}^2 \big(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y \big)$, we incorporate this roughness penalty to simultaneously control the complexity of the estimated

functional components. One possible approach is to extend the smoothed functional PCA framework of Rice and Silverman (1991) by modifying the objective in (18) to

$$\widehat{\mathrm{Cov}}^2\big(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y\big) - \sum_{j=1}^{K} \lambda_j \,\mathrm{PEN}(\xi_j),$$

where the smoothing parameters $\{\lambda_j\}_{j=1}^K$ control the trade-off between maximizing covariance and penalizing roughness. However, this approach Rice and Silverman (1991) assumes that the functional predictors admit an orthogonal expansion in the L^2 sense.

Proposed estimation procedure. To avoid the orthogonal basis assumption of Rice and Silverman (1991), we adopt the strategy of Silverman (1996), which replaces the standard orthonormality constraint with a weaker one based on a modified inner product that incorporates roughness. Accordingly, our estimation procedure at the l-th iteration, iteration index omitted and assuming the observations have been residualized in previous steps, solves the following optimization problem:

$$(23) \quad \hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_K, \hat{\zeta}) := \arg \max_{\xi \in \widetilde{\mathcal{H}}} \widehat{\mathrm{Cov}}^2\big(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y\big) \ s.t. \ \|\xi\|_{\mathcal{H}} + \sum_{j=1}^{K} \lambda_j \,\mathrm{PEN}(\xi_K) = 1.$$

This formulation maximizes the squared covariance over a class of smooth functions. Proposition 4.2 expresses this problem as a generalized Rayleigh quotient.

PROPOSITION 4.2. *Let* $(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y})$ *denote the observed data defined in* (14). *At the l-th iteration of the PLS algorithm, the maximizer of the squared empirical covariance defined in* (23), *represented as the Euclidean vector form as in* (19), *is obtained by solving*

$$(24) \quad \left(\hat{\gamma}_1^\top, \ldots, \hat{\gamma}_K^\top, \hat{\zeta}^\top\right)^\top = \arg \max_{\boldsymbol{\xi} \in \mathbb{R}^{MK+p}} \boldsymbol{\xi}^\top V \boldsymbol{\xi} \quad s.t. \quad \boldsymbol{\xi}^\top (\mathbb{B} + \Lambda \mathbb{B}'') \boldsymbol{\xi} = 1.$$

where we recall that $V = n^{-2}(\mathbb{B}\Theta^\top \mathbf{y})(\mathbb{B}\Theta^\top \mathbf{y})^\top$ *is defined in* (21), *and* $\Lambda \in \mathbb{R}^{(MK+p) \times (MK+p)}$ *is given by*

$$(25) \quad \Lambda := \mathrm{blkdiag}(\lambda_1 I_M, \ldots, \lambda_K I_M, 0_{p \times p}),$$

where $0_{p \times p}$ *denotes the* $p \times p$ *zero matrix.*

The proof of Proposition 4.2 is provided in Appendix E. The smoothing parameter $\lambda_k > 0$ $(k = 1, \ldots, K)$ balances goodness of fit and smoothness in $\hat{\xi}_j$. Smaller λ_k yields components that better fit the data but risks overfitting; setting $\lambda_k = 0$ recovers the unregularized solution in Proposition 4.1. Larger λ_k enforces greater smoothness, and in the limit $\lambda_k \to \infty$, $\hat{\xi}_j(t)$ approaches a linear form $a + bt$. In practice, both $\{\lambda_k\}$ and the number of components L can be selected via cross-validation using a predictive criterion such as mean squared error.

Computation. The generalized eigenproblem presented in Proposition 4.2 may be computationally unstable in practice. However, by leveraging the rank-one structure of the matrix V Proposition 4.3 derives a closed-form solution that requires only the solution of linear systems.

PROPOSITION 4.3 (Closed-form solution). *Consider the optimization problem described in Proposition 4.2. Define the following quantities, which depend on the observed data but are not decision variables:*

$$\mathbf{u}_j := B\Theta_j^\top \mathbf{y} \in \mathbb{R}^M \quad for \ j = 1, \ldots, K, \quad and \quad \mathbf{v} := \mathbf{Z}^\top \mathbf{y} \in \mathbb{R}^p.$$

Let

$$q := \sum_{j=1}^{K} \mathbf{u}_j^\top (B + \lambda_j B'')^{-1} \mathbf{u}_j + \mathbf{v}^\top \mathbf{v}.$$

Then the unique (up to sign) solution to the regularized maximization problem is given in closed form by

$$\hat{\boldsymbol{\gamma}}_j = \frac{1}{\sqrt{q}}(B + \lambda_j B'')^{-1}\mathbf{u}_j \quad \text{for } j = 1, \ldots, K, \quad \text{and} \quad \hat{\boldsymbol{\zeta}} = \frac{1}{\sqrt{q}}\mathbf{v}.$$

The proof of Proposition 4.3 is provided in Appendix F. The expressions above involve solving linear systems for the functional and scalar components separately, followed by normalization by a common factor. Although the unnormalized coefficients are obtained independently, the normalization step couples the functional and scalar parts, allowing them to influence one another. This coupling enables the procedure to capture the correlation between the functional and scalar components of the PLS direction.

4.3. *Step 2: Computing Residualized Predictors via Hybrid-on-Scalar Regression.* The l-th iteration's second step involves residualization of both predictors and responses. We first compute the individual PLS score:

$$(26) \qquad\qquad \hat{\rho}_i^{[l]} := \langle \widetilde{W}_i^{[l]}, \hat{\xi}^{[l]} \rangle_{\mathcal{H}},$$

using the estimated PLS component direction $\widehat{\xi}^{[l]}$ obtained from Propositions 4.2 and 4.3. Since $\widetilde{W}_i^{[l]}$ are assumed to have a sample mean of zero, these PLS scores will also have a sample mean of zero. Using these PLS scores as scalar predictor observations, the $(l+1)$-th iteration's residualized response is computed as:

$$(27) \qquad\qquad Y_i^{[l+1]} := Y_i^{[l]} - \frac{(\mathbf{y}^{[l]})^\top \hat{\boldsymbol{\rho}}^{[l]}}{(\hat{\boldsymbol{\rho}}^{[l]})^\top \hat{\boldsymbol{\rho}}^{[l]}} \widehat{\rho}_{il},$$

which is equivalent to using the least squares estimate from regressing $Y_i^{[l]}$ on $\widehat{\rho}_{il}$, a procedure identical to that found in the naive PLS algorithm (Section 3). Since $Y_i^{[l]}$ are assumed to have a sample mean of zero, these residuals $Y_i^{[l+1]}$ also have a sample mean of zero.

For predictor residualization, we fit a hybrid-on-scalar regression model where the predictor is the PLS score for each observation. Specifically, we fit $\widetilde{W}_i^{[l]} = \hat{\rho}_i^{[l]} \delta^{[l]} + E_i^{[l]}$, and then take the residuals as the predictor for the next iteration. Here, $\delta^{[l]}$ and $E_i^{[l]}$ are the regression coefficient and error terms, respectively. Performing this pointwisely, as in Section 3, is highly inefficient and potentially infeasible for dense or irregular functional data. Moreover, the results may lack smoothness, introducing instability that propagates to subsequent results and compromises the entire algorithm. Therefore, we employ a basis expansion approach to fit the entire hybrid object in one step. We assume both $\delta^{[l]}$ and $E_i^{[l]}$ are elements of $\widetilde{\mathcal{H}}$, as defined in (11). To ensure the stability of the entire PLS estimation procedure, the estimated regression coefficient $\hat{\delta}^{[l]} = (\hat{\delta}_1^{[l]}, \ldots, \hat{\delta}_K^{[l]}, \hat{\chi}^{[l]}) \in \widetilde{\mathcal{H}}$ should have smooth functional parts, consistent with Section 4.2.2. Consequently, $\delta^{[l]}$ is obtained by minimizing a penalized least squares criterion, where the functional part is penalized by the integrated squared second derivative defined in (22):

$$(28) \qquad\qquad \hat{\delta}^{[l]} := \arg\min_{\delta \in \widetilde{\mathcal{H}}} \text{PENSSE}(\delta),$$

where

$$(29) \qquad \mathrm{PENSSE}(\delta) := \sum_{i=1}^{n} \langle \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]} \delta, \ \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]} \delta \rangle_{\mathcal{H}} + \tau \sum_{j=1}^{K} \mathrm{PEN}(\delta_j).$$

where $\tau > 0$ is a smoothing parameter controlling the trade-off between fit accuracy and functional regression coefficient smoothness. The following proposition shows that the solution of the problem 28 can be obtained by solving linear equations.

LEMMA 4.4 (Closed-form solution). *Let us denote $\hat{\boldsymbol{\rho}}^{[l]} := (\hat{\rho}_1^{[l]}, \ldots, \hat{\rho}_n^{[l]})^{\top}$. The solution to the optimization problem* (28)*, denoted $\hat{\delta}^{[l]} \in \tilde{\mathcal{H}}$, has basis coefficients and scalar parts given by:*

$$\hat{\boldsymbol{\pi}}_j^{[l]} = \left(\|\hat{\boldsymbol{\rho}}^{[l]}\|_2^2 B + \tau B'' \right)^{-1} B(\Theta_j^{[l]})^{\top} \hat{\boldsymbol{\rho}}^{[l]}, \ j = 1, \ldots, K, \quad \hat{\boldsymbol{\chi}}^{[l]} = (\mathbf{Z}^{[l]})^{\top} \hat{\boldsymbol{\rho}}^{[l]}.$$

Proof of Lemma 4.4 is provided in Appendix G.

The $(l+1)$-th iteration's residualized predictor is then computed as:

$$(30) \qquad \widetilde{W}_i^{[l+1]} = \widetilde{W}_i^{[l]} - \hat{\rho}_{il} \hat{\delta}_l.$$

Since $\widetilde{W}_i^{[l]}$ are assumed to have a sample mean of zero, these residuals $\widetilde{W}_i^{[l+1]}$ also have a sample mean of zero.

4.4. *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 $\hat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \hat{\xi}_l \rangle_{\mathcal{H}} = \langle \mathbf{W}_i, \hat{\boldsymbol{\zeta}}_l \rangle_{\mathcal{H}}$, where $\hat{\boldsymbol{\zeta}}_l = \hat{\xi}_l - \sum_{u=1}^{l-1} \langle \hat{\boldsymbol{\delta}}_u, \hat{\xi}_l \rangle_{\mathcal{H}} \hat{\boldsymbol{\zeta}}_u$ for $l \geq 2$, with $\hat{\boldsymbol{\zeta}}_1 = \hat{\xi}_1$. Then, the decomposition of Y_i in (5) leads to

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

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

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

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

Algorithm 2 Hybrid PLS (under construction)

Require: Response $\mathbf{y} \in \mathbb{R}^n$, hybrid predictors approximation coefficient $\widetilde{C}^* \in \mathbb{R}^{n \times (MK+p)}$ (33), Gram matrix with respect to basis functions $J^* \in \mathbb{R}^{(MK+p) \times (MK+p)}$ (34), Gram matrix with respect to second derivatives of the basis functions $\ddot{J}^* \in \mathbb{R}^{(MK+p) \times (MK+p)}$ (36), regularization matrix $\Lambda \in \mathbb{R}^{(MK+p) \times (MK+p)}$ (25)

1: $\widetilde{C}^{*[1]} \leftarrow \widetilde{C}^*$
2: **for** $\mathrm{l} = 1, 2, \ldots T$ **do**
3: $\quad \mathbf{V}^* \leftarrow n^{-2} J^* \widetilde{C}^{*\top} \widetilde{\mathbf{y}} \widetilde{\mathbf{y}}^{\top} \widetilde{C}^* J^* \qquad\qquad\qquad\qquad\qquad\qquad \triangleright \ \textbf{(??)}$
4: $\quad \hat{\xi}_l^* \leftarrow \arg\max_{\xi^* \in \mathbb{R}^{MK+p}} \xi^{*\top} \mathbf{V}^* \xi^* \quad \text{subject to} \quad \xi^{*\top}(J^* + \Lambda \ddot{J}^*)\xi^* = 1 \qquad \triangleright \ (24)$
5: $\quad \hat{\boldsymbol{\rho}}_l \leftarrow \widetilde{C}^* J^* \hat{\xi}_l^* \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \triangleright \in \mathbb{R}^n; \ \textbf{(??)}$
6: Output a

4.5. *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 $X_i = (X_i^{(1)}, \ldots, X_i^{(K)})$ 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. *Orthonormality.* The constraint $\xi^{*\top}(J^* + \Lambda \ddot{J}^*)\xi^* = 1$ enforces the following modified orthonormality of the estimated PLS components $(\widehat{\xi}_1, \widehat{\xi}_2, \ldots)$ 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:

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

$$= \sum_{k=1}^K \int_0^1 \widehat{\psi}_l^{(k)}(t_k)\widehat{\psi}_j^{(k)}(t_k)dt_k + \sum_{r=1}^p \widehat{\theta}_{lr}\widehat{\theta}_{jr} + \sum_{k=1}^K \lambda_k \int_0^1 \ddot{\widehat{\psi}}_l^{(k)}(t_k)\ddot{\widehat{\psi}}_j^{(k)}(t_k)dt_k$$

$$= \mathbb{K}(l = j).$$

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 5.3), along with other properties.

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:

$$\mathcal{C}_{YW}\mathbf{h} = E\{\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}} Y\} = \langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}} = \sum_{k=1}^{K} \int_0^1 \sigma_{YX}^{(k)}(t_k) f^{(k)}(t_k) dt_k + E(Y\mathbf{Z}^\top)\mathbf{v}$$

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

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 5.1.   *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*

$$(31) \qquad \max_{k=1,\ldots,K} \sup_{t_k \in [0,1]} \sigma_{YX}^{(k)2}(t_k) < Q_1 \quad \text{and} \quad \max_{r=1,\ldots,p} \sigma_{YZ,r}^2 < Q_2,$$

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 $t_k, t_k^* \in [0,1]$ 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 5.1 is provided in Appendix H. By the Hilbert-Schmidt theorem (e.g., Theorem 4.2.4 in Hsing and Eubank, 2015), Proposition 5.1 guarantees the existence of a complete orthonormal system of eigenfunctions $\{\xi_{(u)}\}_{u \in \mathbb{N}}$ of $\mathcal{U}$ in $\mathcal{H}$ such that $\mathcal{U}\xi_{(u)} = \kappa_{(u)}\xi_{(u)}$, 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 5.2.

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

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

The proof of Proposition 5.2 is provided in Appendix I. In other words, l-th PLS component ξ_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 4.2 are orthonormal with respect to the modified inner product $\langle\langle\cdot\rangle\rangle$ that incorporates the roughness of the functions.

PROPOSITION 5.3.   *The estimated PLS components,* $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_L$, *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.,*

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

The proof of Proposition 5.3 is provided in Appendix J. 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 5.4.   *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 5.4 is provided in Appendix K.

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:
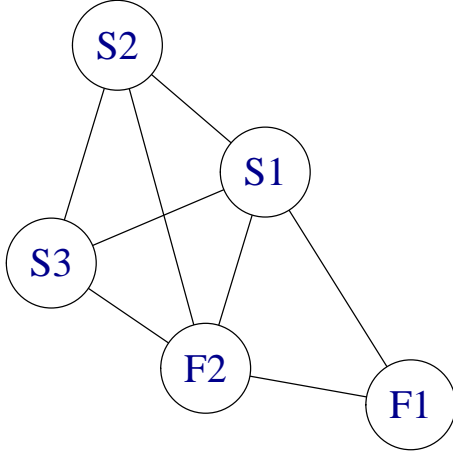
- Nodes F1 and F2: Two functional predictors modeled as Gaussian processes using a truncated Karhunen-Loève expansion, where the eigenbasis consists of Fourier basis functions with a fixed number of basis functions, $M = 9$.
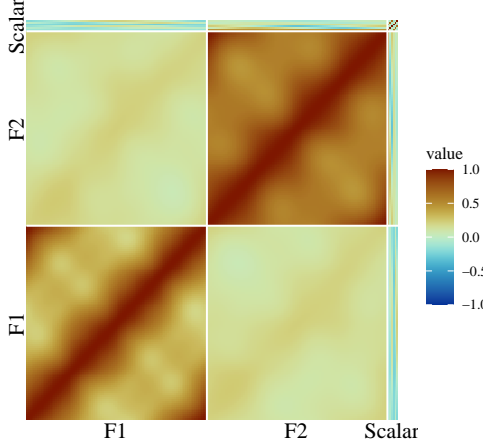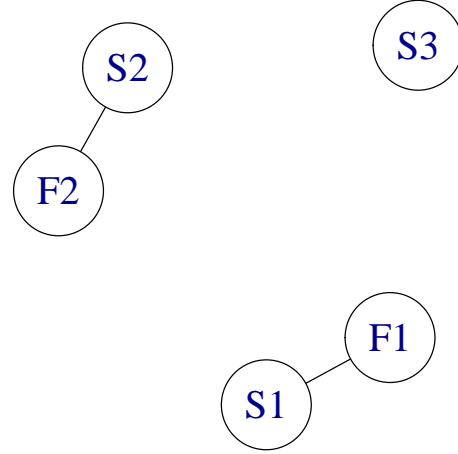
(a) Correlation Matrix - Strong Dependency
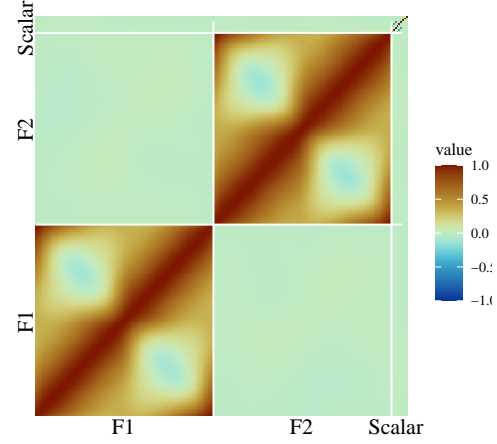
(b) Correlation Matrix - Weak Dependency

(c) Graph Structure - Strong Dependency

(d) Graph Structure - Weak Dependency

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

- 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 connected 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 $\boldsymbol{\Theta} := (\theta_{t\tilde{t}}) \in \mathbb{R}^{p \times p}$, 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 $\boldsymbol{\Theta}$ exhibits a tridiagonal Toeplitz structure, where the diagonal entries are given by:

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

The off-diagonal entries are:

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

Finally, the functional predictors are smoothed using a B-spline basis with 15 basis functions, as described in Section 4.2.

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
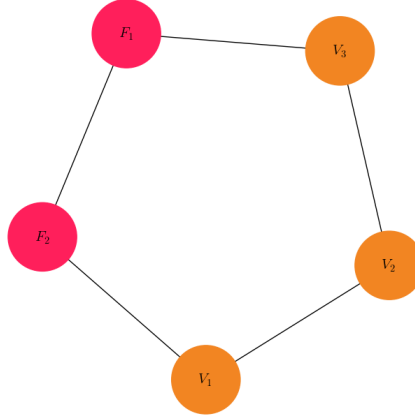
Fig 2

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)

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)

Fig 3: Enter Caption

- 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.

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

Aguilera, A. M., Aguilera-Morillo, M. C., and Preda, C. (2016). Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92.

18

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*. Wiley-Interscience, Hoboken, N.J.

Beyaztas, U. and Lin Shang, H. (2022). A robust functional partial least squares for scalar-on-multiple-function regression. *Journal of Chemometrics*, 36(4):e3394.

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

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591.

Chang, C., Jang, J. H., Manatunga, A., Taylor, A. T., and Long, Q. (2020). A bayesian latent class model to predict kidney obstruction in the absence of gold standard. *Journal of the American Statistical Association*, 115(532):1645–1663.

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.

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.

Gockenbach, M. S. (2010). *Finite-Dimensional Linear Algebra*. CRC Press, Boca Raton, FL, 1st edition edition.

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.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

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.

Jang, J. H. (2021). Principal component analysis of hybrid functional and vector data. *Statistics in Medicine*, 40(24):5152–5173.

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

Mutis, M., Beyaztas, U., Karaman, F., and Lin Shang, H. (2025). On function-on-function linear quantile regression. *Journal of Applied Statistics*, 52(4):814–840.

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

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

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.

Saricam, S., Beyaztas, U., Asikgil, B., and Shang, H. L. (2022). On partial least-squares estimation in scalar-on-function regression models. *Journal of Chemometrics*, 36(12):e3452.

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

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.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

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.

APPENDIX A: OVERVIEW OF APPENDIX

APPENDIX B: NOTATIONS

- aaa

APPENDIX C: TECHNICAL LEMMAS

C.1. old notations.   Then we can approximate the functional predictor observations as

$$X_i^{(k)}(t_k) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t_k) = \mathbf{c}_i^{(k)T} \mathbf{b}^{(k)}(t_k),\ t_k \in [0,1],\ k=1,\ldots,K,\ i=1,\ldots,n,$$

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

$$(32) \qquad\qquad \widetilde{X}(\mathbf{t}) = \widetilde{C}B(\mathbf{t}), \quad t \in \mathcal{T},$$

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 $B(\mathbf{t}) = \mathrm{blkdiag}[\mathbf{b}^{(1)}(t_1),\cdots,\mathbf{b}^{(K)}(t_K)] \in \mathbb{R}^{MK \times K}$.

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 (32):

$$(33) \qquad\qquad \widetilde{W}[\mathbf{t}] = \widetilde{\mathbf{C}}^* B^*[\mathbf{t}], \quad t \in \mathcal{T}.$$

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

C.1.0.1. 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 $\xi[\mathbf{t}] = (\psi(\mathbf{t})^{\top}, \boldsymbol{\theta}^{\top})^{\top} = (\mathbf{a}^{\top} B(\mathbf{t}), \boldsymbol{\theta}^{\top})^{\top} \in \mathbb{R}^{K+p}$.

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

$$(34)$$
$$\mathbf{J} = \mathrm{blkdiag}(\mathbf{J}^{(1)},\cdots,\boldsymbol{\Phi}^{(k)}) \in \mathbb{R}^{MK \times MK} \text{ and } \mathbf{J}^* = \mathrm{blkdiag}(\mathbf{J}, \mathbf{I}_p) \in \mathbb{R}^{(MK+p) \times (MK+p)}.$$

C.1.0.2. Regularization based on second-order derivatives..   Our proposed strategy for computing PLS component, detailed in Section 4.2.2, 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

$$(35) \qquad\qquad \ddot{J}^{(k)} = \left[\int_0^1 \ddot{b}_m^{(k)}(t) \ddot{b}_n^{(k)}(t)\, dt \right]_{m,n=1}^{M}.$$

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

$$(36) \qquad\qquad \ddot{J}^* = \mathrm{blkdiag}(\ddot{J}^{(1)},\cdots,\ddot{J}^{(K)}, 0_{p \times p}).$$

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 $\ddot{J}^*$.

APPENDIX D: PROOF OF PROPOSITIONS 4.1 AND 4.2

PROOF. Let $\xi = (\xi_1, \ldots, \xi_K, \zeta) \in \mathcal{H}$. Assume that each functional component $\xi_j \in L^2([0,1])$ lies in the span of the basis functions $b_1(t), \ldots, b_M(t)$ and can be written as

$$\xi_j(t) := \sum_{m=1}^{M} d_{jm} b_m(t), \quad j = 1, \ldots, K,$$

with coefficient vectors $\boldsymbol{\gamma}_j := (d_{j1}, \ldots, d_{jM})^\top \in \mathbb{R}^M$. Then, the empirical covariance is computed as follows:

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

$$= \frac{1}{n} \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{K} \langle \widetilde{X}_{ij}, \xi_j \rangle + \mathbf{Z}_i^\top \zeta \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} y_i \left(\sum_{k=1}^{K} \int_0^1 \widetilde{X}_{ij}(t) \, \xi_j(t) \, dt \right) + \frac{1}{n} \sum_{i=1}^{n} y_i (\mathbf{Z}_i^\top \zeta)$$

$$= \frac{1}{n} \sum_{i=1}^{n} y_i \left\{ \sum_{k=1}^{K} \int_0^1 \left(\sum_{m=1}^{M} \theta_{ijm} b_m(t) \right) \left(\sum_{m'=1}^{M} d_{jm'} b_{m'}(t) \right) dt \right\} + \frac{1}{n} \mathbf{y}^\top Z \zeta$$

$$= \frac{1}{n} \sum_{i=1}^{n} y_i \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} \theta_{ijm} d_{jm'} \int_0^1 b_m(t) b_{m'}(t) dt \right\} + \frac{1}{n} \mathbf{y}^\top Z \zeta$$

$$= \sum_{k=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} y_i (\boldsymbol{\theta}_{ij}^\top B \boldsymbol{\gamma}_j) \right\} + \frac{1}{n} \mathbf{y}^\top Z \zeta$$

$$= \frac{1}{n} \left(\sum_{k=1}^{K} \mathbf{y}^\top \Theta_j B \boldsymbol{\gamma}_j + \mathbf{y}^\top Z \zeta \right)$$

Building on this computation, the squared empirical covariance is expressed as the quadratic form involving the matrix V defined in (21):

$$\widehat{\mathrm{Cov}}^2(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y) = \frac{1}{n^2} \left(\sum_{k=1}^{K} \mathbf{y}^\top \Theta_j B \boldsymbol{\gamma}_j + \mathbf{y}^\top Z \zeta \right)^2$$

$$= \frac{1}{n^2} (\mathbf{y}^\top \Theta \mathbb{B} \boldsymbol{\xi})^2$$

$$= \frac{1}{n^2} \boldsymbol{\xi}^\top (\mathbb{B} \Theta^\top \mathbf{y})(\mathbb{B} \Theta^\top \mathbf{y})^\top \boldsymbol{\xi}$$

$$(37) \qquad\qquad = \boldsymbol{\xi}^\top V \boldsymbol{\xi}.$$

The squared norm of ξ in the hybrid Hilbert space $\mathcal{H}$ is computed as follows:

$$\langle \xi, \xi \rangle_{\mathcal{H}} = \sum_{j=1}^{K} \langle \xi_j, \xi_j \rangle + \zeta^\top \zeta$$

$$= \sum_{j=1}^{K} \int_0^1 \left(\sum_{m=1}^{M} d_{jm} b_m(t) \right) \left(\sum_{m'=1}^{M} d_{jm'} b_{m'}(t) \right) dt + \zeta^\top \zeta$$

$$= \sum_{j=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} d_{jm} d_{jm'} \int_0^1 b_m(t) b_{m'}(t) dt + \boldsymbol{\zeta}^\top \boldsymbol{\zeta}$$

$$= \sum_{j=1}^{K} \boldsymbol{\gamma}_j^\top B \boldsymbol{\gamma}_j + \boldsymbol{\zeta}^\top \boldsymbol{\zeta},$$

(38) $$= \boldsymbol{\xi}^\top \mathbb{B} \boldsymbol{\xi}$$

Therefore, the covariance maximization problem (18) is equivalent to the generalized Raleigh quotient (20). This completes the proof of Proposition 4.1.

APPENDIX E: PROOF OF PROPOSITION 4.2

The penalization term $\sum_{j=1}^{K} \lambda_j \operatorname{PEN}(\xi_K)$ can be written in matrix form as:

$$\sum_{j=1}^{K} \lambda_j \operatorname{PEN}(\xi_j) = \sum_{j=1}^{K} \lambda_j \int_0^1 \{\hat{\xi}_j''(t)\}^2 dt$$

$$= \sum_{j=1}^{K} \lambda_j \int_0^1 \{\sum_{l=1}^{M} \theta_{ijl} b_l''(t)\}^2 dt$$

$$= \sum_{j=1}^{K} \lambda_j \sum_{l=1}^{M} \sum_{m=1}^{M} \theta_{ijl} \theta_{ijm} \int_0^1 b_l''(t) b_m''(t) dt$$

$$= \sum_{j=1}^{K} \lambda_j \boldsymbol{\gamma}_j^\top B'' \boldsymbol{\gamma}_j$$

(39) $$= \boldsymbol{\xi}^\top \Lambda \mathbb{B}'' \boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is the concatenated coefficient vector form of ξ as defined in (19), $\mathbb{B}''$ is defined in (16), and Λ is defined in (25).

Combining this with (38), we can compute

$$\|\xi\|_{\mathcal{H}} + \sum_{j=1}^{K} \lambda_j \operatorname{PEN}(\xi_K) = \boldsymbol{\xi}^\top \mathbb{B} \boldsymbol{\xi} + \boldsymbol{\xi}^\top \Lambda \mathbb{B}'' \boldsymbol{\xi} = \boldsymbol{\xi}^\top (\mathbb{B} + \Lambda \mathbb{B}'') \boldsymbol{\xi}.$$

This computation along with (37) implies that the covariance maximization problem (23) is equivalent to the generalized Raleigh quotient (24). This completes the proof of Proposition 4.2.

$\square$

APPENDIX F: PROOF OF PROPOSITION 4.3

PROOF. The optimization problem (24) can be written as

$$\max_{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_K \in \mathbb{R}^M, \boldsymbol{\zeta} \in \mathbb{R}^p} \frac{1}{n^2} \left(\sum_{j=1}^{K} \mathbf{y}^\top \Theta_j B \boldsymbol{\gamma}_j + \mathbf{y}^\top Z \boldsymbol{\zeta} \right)^2,$$

$$\text{subject to} \quad \sum_{j=1}^{K} \boldsymbol{\gamma}_j^\top (B + \lambda_j B'') \boldsymbol{\gamma}_j + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} = 1.$$

Let $\mathbf{u}_j := B\Theta_j^\top \mathbf{y} \in \mathbb{R}^M$ and $\mathbf{v} := Z^\top \mathbf{y} \in \mathbb{R}^p$. These are fixed problem data. Then the objective becomes

$$\frac{1}{n^2}\left(\sum_{j=1}^K \mathbf{u}_j^\top \boldsymbol{\gamma}_j + \mathbf{v}^\top \boldsymbol{\zeta}\right)^2.$$

The objective function is continuous. The constraint defines the boundary of an ellipsoid in a finite-dimensional Euclidean space, the feasible set is compact. By the Weierstrass Extreme Value Theorem (see, e.g., Theorem 2.3.1 in Bazaraa et al. 2006), a global maximizer exists. The constraint function is continuously differentiable and its gradient vanishes only at the origin, which is not feasible. Thus the gradient is nonzero at all feasible points. Hence, the Linear Independence Constraint Qualification (LICQ) holds, and the Karush-Kuhn-Tucker (KKT) conditions are necessary for local optimality (see, e.g., Theorem 5.3.1 in Bazaraa et al. 2006).

Define the Lagrangian:

$$\mathcal{L}(\{\boldsymbol{\gamma}_j\}, \boldsymbol{\zeta}, \mu) := \left(\sum_{j=1}^K \mathbf{u}_j^\top \boldsymbol{\gamma}_j + \mathbf{v}^\top \boldsymbol{\zeta}\right)^2 - \mu\left[\sum_{j=1}^K \boldsymbol{\gamma}_j^\top (B + \lambda_j B'')\boldsymbol{\gamma}_j + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} - 1\right].$$

Let $s := \dfrac{2}{n^2}\left(\sum_{j=1}^K \mathbf{u}_j^\top \boldsymbol{\gamma}_j + \mathbf{v}^\top \boldsymbol{\zeta}\right)$. The KKT conditions require that:

(40) $$\nabla_{\boldsymbol{\gamma}_j}\mathcal{L} = s\mathbf{u}_j - 2\mu(B + \lambda_j B'')\boldsymbol{\gamma}_j = 0, \quad \text{for } j = 1, \ldots, K,$$

(41) $$\nabla_{\boldsymbol{\zeta}}\mathcal{L} = s\mathbf{v} - 2\mu\boldsymbol{\zeta} = 0.$$

From (40) and (41), we have

$$(B + \lambda_j B'')\boldsymbol{\gamma}_j = \frac{s}{2\mu}\mathbf{u}_j, \quad \boldsymbol{\zeta} = \frac{s}{2\mu}\mathbf{v}, \quad \text{for } j = 1, \ldots, K.$$

This implies that for any local maximizer, there exists $c \neq 0$ such that

$$\boldsymbol{\gamma}_j = c(B + \lambda_j B'')^{-1}\mathbf{u}_j, \quad \boldsymbol{\zeta} = c\mathbf{v}, \quad \text{for } j = 1, \ldots, K.$$

Since we assumed in Section 4.1 that the functions $\{b_m(t)\}$ and their second derivatives are linearly independent, both B and B'' are positive definite (see, for example, Theorem 273 of Gockenbach, 2010). As positive definiteness is preserved under conic combinations, $B + \lambda_j B''$ is also positive definite.

A local maximizer must also be primal feasible. Substituting the conditions above in to the primal constraint:

$$\sum_{j=1}^K \left(c(B + \lambda_j B'')^{-1}\mathbf{u}_j\right)^\top (B + \lambda_j B'')\left(c(B + \lambda_j B'')^{-1}\mathbf{u}_j\right) + c^2 \mathbf{v}^\top \mathbf{v} = 1$$

$$\Rightarrow c^2\left(\sum_{j=1}^K \mathbf{u}_j^\top (B + \lambda_j B'')^{-1}\mathbf{u}_j + \mathbf{v}^\top \mathbf{v}\right) = 1.$$

Solving for c^2 gives:

$$c^2 = \left(\sum_{j=1}^K \mathbf{u}_j^\top (B + \lambda_j B'')^{-1}\mathbf{u}_j + \mathbf{v}^\top \mathbf{v}\right)^{-1}.$$

Hence, the unique maximizer (up to sign) is:

$$\boldsymbol{\gamma}_j^* = \frac{1}{\sqrt{q}}(B + \lambda_j B'')^{-1}\mathbf{u}_j, \quad \boldsymbol{\zeta}^* = \frac{1}{\sqrt{q}}\mathbf{v},$$

where

$$q := \sum_{j=1}^{K} \mathbf{u}_j^\top (B + \lambda_j B'')^{-1}\mathbf{u}_j + \mathbf{v}^\top \mathbf{v}.$$

This completes the proof of Proposition F.

$\square$

APPENDIX G: PROOF OF LEMMA 4.4

For notational simplicity, we omit the iteration superscript $[l]$. Using the basis expansion coefficient notation from (10), (12), (13), and (15), the full vector of PLS scores $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_n)^\top$ where $\hat{\rho}_1$ is defined in (26), is computed through the following matrix multiplication:

$$\hat{\boldsymbol{\rho}}^\top = \sum_{k=1}^{K} (\hat{\boldsymbol{\gamma}}_j)^\top B \Theta_j^\top + \boldsymbol{\zeta}^\top \mathbf{Z}^\top.$$

The criterion $\text{PENSSE}(\delta)$ can be decomposed as follows:

$$\text{PENSSE}(\delta) = \sum_{i=1}^{n} \langle \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]}\delta, \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]}\delta \rangle_{\mathcal{H}} + \tau \sum_{j=1}^{K} \text{PEN}(\delta_j)$$

$$= \underbrace{\sum_{i=1}^{n} \langle \widetilde{W}_i^{[l]}, \widetilde{W}_i^{[l]} \rangle_{\mathcal{H}}}_{A} \underbrace{-2\sum_{i=1}^{n} \hat{\rho}_i^{[l]}\langle \delta, \widetilde{W}_i^{[l]} \rangle_{\mathcal{H}}}_{B_1(\boldsymbol{\delta})} \underbrace{+(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}})\langle \delta, \delta \rangle_{\mathcal{H}}}_{B_2(\boldsymbol{\delta})} + \underbrace{\tau \sum_{j=1}^{K} \text{PEN}(\delta_j)}_{B_3(\boldsymbol{\delta})}$$

Here, part A does not contain δ, so it does not contribute to the minimization problem.

Let $\boldsymbol{\delta} := (\pi_1^\top, \ldots, \pi_K^\top, \boldsymbol{\chi}^\top)^\top$ be the $MK + p$-dimensional concatenated vector of basis coefficients and the scalar part of $\boldsymbol{\delta}$, as presented in (19). Abusing notation, we treat PENSSE as a function of $\boldsymbol{\delta}$. Next, We will now demonstrate that the combined term $B_1(\boldsymbol{\delta}) + B_2(\boldsymbol{\delta}) + B_3(\boldsymbol{\delta})$ is a quadratic function of $\boldsymbol{\delta}$, by expanding its component functions:

$$B_1(\boldsymbol{\delta}) = -2\sum_{i=1}^{n} \hat{\rho}_i \langle \delta, \widetilde{W}_i \rangle_{\mathcal{H}} = -2\hat{\boldsymbol{\rho}}^\top \left(\sum_{j=1}^{K} \Theta_j B \pi_j + \mathbf{Z}\boldsymbol{\chi} \right),$$

$$B_2(\boldsymbol{\delta}) = (\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}})\langle \delta, \delta \rangle_{\mathcal{H}} = (\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) \left(\sum_{j=1}^{K} \pi_j^\top B \pi_j + \boldsymbol{\chi}^\top \boldsymbol{\chi} \right),$$

$$B_3(\boldsymbol{\delta}) = \tau \sum_{j=1}^{K} \text{PEN}(\delta_j) = \tau \boldsymbol{\delta} B'' \boldsymbol{\delta},$$

where the expansion of $B_3(\boldsymbol{\delta})$ leverages the computation in (39). Now we compute the gradients:

$$\nabla_{\boldsymbol{\pi}_j} B_1(\boldsymbol{\delta}) = -2B\Theta_j^\top \hat{\boldsymbol{\rho}}, \ j = 1, \ldots, K, \quad \nabla_{\boldsymbol{\chi}} B_1(\boldsymbol{\delta}) = -2\mathbf{Z}^\top \hat{\boldsymbol{\rho}},$$

$$\nabla_{\boldsymbol{\pi}_j} B_2(\boldsymbol{\delta}) = 2(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) B \boldsymbol{\pi}_j, \ j=1,\ldots,K, \quad \nabla_{\boldsymbol{\chi}} B_2(\boldsymbol{\delta}) = 2\boldsymbol{\chi},$$

$$\nabla_{\boldsymbol{\pi}_j} B_3(\boldsymbol{\delta}) = 2\tau B'' \boldsymbol{\pi}_j, \ j=1,\ldots,K, \quad \nabla_{\boldsymbol{\chi}} B_3(\boldsymbol{\delta}) = 0.$$

The Hessian of $\mathrm{PENSSE}(\boldsymbol{\delta})$ is then given by $2(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) B + 2\tau B''$. Since we assumed in Section 4.1 that both B' and B'' are positive definite, $\mathrm{PENSSE}(\delta)$ is convex. convex. Consequently, the gradient vanishes at its unique minimizer:

$$\nabla_{\boldsymbol{\pi}_j} \mathrm{PENSSE}(\hat{\boldsymbol{\delta}}) = -2B\Theta_j^\top \hat{\boldsymbol{\rho}} + 2(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}})B\hat{\boldsymbol{\pi}}_j + 2B''\hat{\boldsymbol{\pi}}_j = 0, \ j=1,\ldots,K,$$

$$\nabla_{\boldsymbol{\chi}} \mathrm{PENSSE}(\hat{\boldsymbol{\delta}}) = -2\mathbf{Z}^\top \hat{\boldsymbol{\rho}} + 2\hat{\boldsymbol{\chi}} = 0,$$

providing the following closed-form solution:

$$\hat{\boldsymbol{\pi}}_j = \left\{ (\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}})B + \tau B'' \right\}^{-1} B\Theta_j^\top \hat{\boldsymbol{\rho}} \ j=1,\ldots,K, \quad \hat{\boldsymbol{\chi}} = \mathbf{Z}^\top \hat{\boldsymbol{\rho}}.$$

This completes the proof of Lemma 4.4.

APPENDIX H: PROOF OF PROPOSITION 5.1

$\mathcal{U}$ 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 $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$,

$$\begin{aligned} \langle \mathcal{U}\mathbf{h}_1, \mathbf{h}_2 \rangle_{\mathcal{H}} &= \langle \langle \mathbf{h}_1, \Sigma_{YW} \rangle_{\mathcal{H}} \Sigma_{YW}, \mathbf{h}_2 \rangle_{\mathcal{H}} \\ &= \langle \mathbf{h}_1, \Sigma_{YW} \rangle_{\mathcal{H}} \langle \Sigma_{YW}, \mathbf{h}_2 \rangle_{\mathcal{H}} \\ &= \langle \mathbf{h}_1, \langle \Sigma_{YW}, \mathbf{h}_2 \rangle_{\mathcal{H}} \Sigma_{YW} \rangle_{\mathcal{H}} \\ &= \langle \mathbf{h}_1, \mathcal{U}\mathbf{h}_2 \rangle_{\mathcal{H}}. \end{aligned}$$

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 $\mu([0,1])$ denote a Lebesgue measure of $[0,1]$, and let $T = \max_{k=1,\ldots,K} \mu([0,1])$. Then for any $g \in \mathcal{I}$ and $\mathbf{t} \in \mathcal{T}$,

$$\begin{aligned} \|g[\mathbf{t}]\|^2 &= \|(\mathcal{U}\mathbf{h})[\mathbf{t}]\|^2 \\ &= \|\langle \Sigma_{YW}, \mathbf{h} \rangle_{\mathcal{H}} \Sigma_{YW}[\mathbf{t}]\|^2 \\ &= (\langle \sigma_{YX}, f \rangle_{\mathcal{F}} + \langle \sigma_{YZ}, \mathbf{v} \rangle)^2 \|[\sigma_{YX}^\top(\mathbf{t}), \sigma_{YZ}^\top]^\top\|^2 \\ &\leq \left(\langle \sigma_{YX}, \sigma_{YX} \rangle_{\mathcal{F}}^{1/2} \langle f, f \rangle_{\mathcal{F}}^{1/2} + \langle \sigma_{YZ}, \sigma_{YZ} \rangle^{1/2} \langle \mathbf{v}, \mathbf{v} \rangle^{1/2} \right)^2 \|[\sigma_{YX}^\top(\mathbf{t}), \sigma_{YZ}^\top]^\top\|^2 \quad (\because \mathrm{CSI}) \\ &= \left[\left\{ \sum_{k=1}^K \int_0^1 \sigma_{YX}^{(k)2}(t_k) dt_k \right\}^{\frac{1}{2}} \|f\|_{\mathcal{F}} + \left(\sum_{r=1}^p \sigma_{YZ,r}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\| \right]^2 \left(\sum_{k=1}^K \sigma_{YX}^{(k)2}(t_k) + \sum_{r=1}^p \sigma_{YZ,r}^2 \right) \\ &\leq \left\{ \left(\sum_{k=1}^K \int_0^1 Q_1 dt_k \right)^{\frac{1}{2}} B^{\frac{1}{2}} + \left(\sum_{r=1}^p Q_2 \right)^{\frac{1}{2}} B^{\frac{1}{2}} \right\}^2 \left(\sum_{k=1}^K Q_1 + \sum_{r=1}^p Q_2 \right) \\ &= B \left\{ (KTQ_1)^{\frac{1}{2}} + (PQ_2)^{\frac{1}{2}} \right\}^2 (KQ_1 + pQ_2) < \infty. \end{aligned}$$

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

$$\|g[\mathbf{t}] - g[\mathbf{t}^*]\|^2 = \|(\mathbf{U}\mathbf{h})[\mathbf{t}] - (\mathbf{U}\mathbf{h})[\mathbf{t}^*]\|^2$$

$$= \|\langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}} \Sigma_{YW}[\mathbf{t}] - \langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}} \Sigma_{YW}[\mathbf{t}^*]\|^2$$

$$= \langle \Sigma_{YW}, \mathbf{h}\rangle_{\mathcal{H}}^2 \|\Sigma_{YW}[\mathbf{t}] - \Sigma_{YW}[\mathbf{t}^*]\|^2$$

$$= (\langle \sigma_{YX}, f\rangle_{\mathcal{F}} + \langle \sigma_{YZ}, \mathbf{v}\rangle)^2 \left\| \begin{bmatrix} \sigma_{YX}(\mathbf{t}) \\ \sigma_{YZ} \end{bmatrix} - \begin{bmatrix} \sigma_{YX}(\mathbf{t}^*) \\ \sigma_{YZ} \end{bmatrix} \right\|^2$$

$$\overset{CSI}{\le} \left(\langle \sigma_{YX}, \sigma_{YX}\rangle_{\mathcal{F}}^{1/2} \langle f, f\rangle_{\mathcal{F}}^{1/2} + \langle \sigma_{YZ}, \sigma_{YZ}\rangle^{1/2} \langle \mathbf{v}, \mathbf{v}\rangle^{1/2} \right)^2 \left\| \begin{bmatrix} \sigma_{YX}(\mathbf{t}) - \sigma_{YX}(\mathbf{t}^*) \\ \mathbf{0} \end{bmatrix} \right\|^2$$

$$= \left[\left\{ \sum_{k=1}^{K} \int_0^1 \sigma_{YX}^{(k)2}(t_k)dt_k \right\}^{\frac{1}{2}} \|f\|_{\mathcal{F}} + \left(\sum_{r=1}^{p} \sigma_{YZ,r}^2 \right)^{\frac{1}{2}} \|\mathbf{v}\| \right]^2 \sum_{k=1}^{K} \left\{ \sigma_{YX}^{(k)}(t_k) - \sigma_{YX}^{(k)}(t_k^*) \right\}^2$$

$$\le B \left\{ (KTQ_1)^{\frac{1}{2}} + (PQ_2)^{\frac{1}{2}} \right\} K\tilde{\epsilon}^2 = \epsilon^2.$$

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.

APPENDIX I: PROOF OF PROPOSITION 5.2

Before we prove Proposition 5.2, we state and prove the following lemma:

LEMMA I.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 5.2. 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.

APPENDIX J: PROOF OF PROPOSITION 5.3

Firstly, it is obvious by the unit-norm constraint enforced in Proposition 5.2 that:

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

$$= \sum_{k=1}^{K}\int_0^1 \widehat{\psi}_l^{(k)}(t_k)\widehat{\psi}_j^{(k)}(t_k)dt_k + \sum_{r=1}^{p}\widehat{\theta}_{lr}\widehat{\theta}_{jr} + \sum_{k=1}^{K}\lambda_k\int_0^1 \ddot{\hat{\psi}}_l^{(k)}(t_k)\ddot{\hat{\psi}}_j^{(k)}(t_k)dt_k$$

$$= \sum_{k=1}^{K}\int_0^1 \hat{\mathbf{a}}_l^{(k)\top}\mathbf{b}^{(k)}(t_k)\mathbf{b}^{(k)\top}(t_k)\hat{\mathbf{a}}_l^{(k)}dt_k + \boldsymbol{\theta}_l^\top\boldsymbol{\theta}_l + \sum_{k=1}^{K}\lambda_k\int_0^1 \hat{\mathbf{a}}_l^{(k)\top}\ddot{\mathbf{b}}^{(k)}(t_k)\ddot{\mathbf{b}}^{(k)\top}(t_k)\hat{\mathbf{a}}_l^{(k)}dt_k$$

$$= \sum_{k=1}^{K}\hat{\mathbf{a}}_l^{(k)\top}J^{(k)}\hat{\mathbf{a}}_l^{(k)} + \boldsymbol{\theta}_l^\top\boldsymbol{\theta}_l + \sum_{k=1}^{K}\lambda_k\hat{\mathbf{a}}_l^{(k)\top}\ddot{J}^{(k)}\hat{\mathbf{a}}_l^{(k)}$$

$$= \hat{\xi}_l^{*\top}J^*\hat{\xi}_l^* + \hat{\xi}_l^{*\top}\Lambda\ddot{J}^*\hat{\xi}_l^* = \hat{\xi}_l^{*\top}(J^* + \Lambda\ddot{J}^*)\hat{\xi}_l^* = 1$$

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

$$\widetilde{C}^{*[l]}B^*[\mathbf{t}] = \widetilde{C}^{*[l-1]}B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\gamma}}_{l-1}^* B^*[\mathbf{t}]$$

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

$$= \widetilde{C}^{*[l-1]}B^*[\mathbf{t}] - \left(\frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}}\right)\widetilde{C}^{*[l-1]}B^*[\mathbf{t}]$$

$$= \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}}\right)\widetilde{C}^{*[l-1]}B^*[\mathbf{t}].$$

Thus, we have

$$\widetilde{W}^{[l]}[\mathbf{t}] = \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}}\right)\widetilde{W}^{[l-1]}[\mathbf{t}]$$

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

$$= P \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_j \widehat{\boldsymbol{\rho}}_j^\top}{\widehat{\boldsymbol{\rho}}_j^\top \widehat{\boldsymbol{\rho}}_j} \right) \widetilde{W}^{[j]}[\mathbf{t}]$$

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 $j < l$,

$$\int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}]\widehat{\xi}_j[\mathbf{t}]d\mathbf{t} = \int_{\mathcal{T}} P \left(I_n - \frac{\widehat{\boldsymbol{\rho}}_j \widehat{\boldsymbol{\rho}}_j^\top}{\widehat{\boldsymbol{\rho}}_j^\top \widehat{\boldsymbol{\rho}}_j} \right) \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\xi}_k[\mathbf{t}]d\mathbf{t}$$

$$= P \left\{ \int_{\mathcal{T}} \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\xi}_j[\mathbf{t}]d\mathbf{t} - \frac{\widehat{\boldsymbol{\rho}}_j \widehat{\boldsymbol{\rho}}_j^\top}{\widehat{\boldsymbol{\rho}}_j^\top \widehat{\boldsymbol{\rho}}_j} \int_{\mathcal{T}} \widetilde{W}^{[j]}[\mathbf{t}]\widehat{\xi}_j[\mathbf{t}]d\mathbf{t} \right\}$$

$$= P \left(\widehat{\boldsymbol{\rho}}_j - \frac{\widehat{\boldsymbol{\rho}}_j \widehat{\boldsymbol{\rho}}_j^\top}{\widehat{\boldsymbol{\rho}}_j^\top \widehat{\boldsymbol{\rho}}_j} \widehat{\boldsymbol{\rho}}_j \right)$$

$$= P(\widehat{\boldsymbol{\rho}}_j - \widehat{\boldsymbol{\rho}}_j)$$

$$= \mathbf{0} \in \mathbb{R}^n,$$

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

$$\langle\langle \widehat{\xi}_l, \widehat{\xi}_j \rangle\rangle = \widehat{\xi}_l^{*\top} (J^* + \Lambda \ddot{J}^*) \widehat{\xi}_j^*$$

$$= \frac{1}{\kappa_l} (V_l^* \widehat{\xi}_l^*)^\top \widehat{\xi}_j^*$$

$$= \frac{1}{\kappa_l} \widehat{\xi}_l^{*\top} V_l^* \widehat{\xi}_j^*$$

$$= \frac{1}{\kappa_l} \widehat{\xi}_l^{*\top} \left(n^{-2} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \widetilde{C}^{*[l]} J^* \right) \widehat{\xi}_j^*$$

$$= \frac{1}{n^2 \kappa_l} \widehat{\xi}_l^{*\top} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \widetilde{C}^{*[l]} \left(\int_{\mathcal{T}} B^*[\mathbf{t}] B^*[\mathbf{t}]^\top d\mathbf{t} \right) \widehat{\xi}_j^*$$

$$= \frac{1}{n^2 \kappa_l} \widehat{\xi}_l^{*\top} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \left(\int_{\mathcal{T}} \widetilde{C}^{*[l]} B^*[\mathbf{t}] B^*[\mathbf{t}]^\top \widehat{\xi}_j^* d\mathbf{t} \right)$$

$$= \frac{1}{n^2 \kappa_l} \widehat{\xi}_l^{*\top} J^* \widetilde{C}^{*[l]\top} \widetilde{\mathbf{y}}^{[l]} \widetilde{\mathbf{y}}^{[l]\top} \left(\int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}]\widehat{\xi}_j[\mathbf{t}]d\mathbf{t} \right)$$

$$= 0,$$

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

APPENDIX K: PROOF OF PROPOSITION 5.4

First note that $\widehat{\boldsymbol{\rho}}_l = \int_{\mathcal{T}} \widetilde{W}^{[l]}[\mathbf{t}]\widehat{\xi}_l[\mathbf{t}]d\mathbf{t} = \widetilde{C}^{*[l]} J^* \widehat{\xi}_l^*$ 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:

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

28

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

$$\widetilde{C}^{*[l]}B^*[\mathbf{t}] = \widetilde{C}^{*[l-1]}B^*[\mathbf{t}] - \widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\gamma}}_{l-1}^*B^*[\mathbf{t}]$$

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

$$= \widetilde{C}^{*[l-1]}B^*[\mathbf{t}] - \frac{\widehat{\boldsymbol{\rho}}_{l-1}\widehat{\boldsymbol{\rho}}_{l-1}^\top\widetilde{C}^{*[l-1]}B^*[\mathbf{t}]}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}}$$

$$= \widetilde{C}^{*[l-1]}B^*[\mathbf{t}] - \frac{\widetilde{C}^{*[l-1]}J^*\widehat{\xi}_{l-1}^*\widehat{\boldsymbol{\rho}}_{l-1}^\top\widetilde{C}^{*[l-1]}B^*[\mathbf{t}]}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}},$$

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

$$\widetilde{C}^{*[l]} = \widetilde{C}^{*[l-1]} - \frac{\widetilde{C}^{*[l-1]}J^*\widehat{\xi}_{l-1}^*\widehat{\boldsymbol{\rho}}_{l-1}^\top\widetilde{C}^{*[l-1]}}{\widehat{\boldsymbol{\rho}}_{l-1}^\top\widehat{\boldsymbol{\rho}}_{l-1}}$$

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

$$= \widetilde{C}^{*[j]}\left(I_{MK+p} - \frac{J^*\widehat{\xi}_j^*\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}_j^\top\widehat{\boldsymbol{\rho}}_j}\right)\widetilde{P},$$

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

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

Then, for $j < l$,

$$\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[l]} = \widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}\left(I_{MK+p} - \frac{J^*\widehat{\xi}_j^*\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}_j^\top\widehat{\boldsymbol{\rho}}_j}\right)\widetilde{P}$$

$$= \left(\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]} - \frac{\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}J^*\widehat{\xi}_j^*\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}_j^\top\widehat{\boldsymbol{\rho}}_j}\right)\widetilde{P}$$

$$= \left(\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]} - \frac{\widehat{\boldsymbol{\rho}}_j^\top\widehat{\boldsymbol{\rho}}_j\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}}{\widehat{\boldsymbol{\rho}}_j^\top\widehat{\boldsymbol{\rho}}_j}\right)\widetilde{P}$$

$$= \left(\widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]} - \widehat{\boldsymbol{\rho}}_j^\top\widetilde{C}^{*[j]}\right)\widetilde{P}$$

$$= \mathbf{0} \in \mathbb{R}^{1\times(MK+p)},$$

Therefore, for $j < l$,

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

Thus, the proposition is proved.

APPENDIX L: SIMULATIONS

APPENDIX M: INTRODUCTION

This template helps you to create a properly formatted LATEX 2_ε manuscript. Prepare your paper in the same style as used in this sample .pdf file. Try to avoid excessive use of italics and bold face. Please do not use any LATEX 2_ε or TEX commands that affect the layout or formatting of your document (i.e., commands like `\texttheight`, `\textwidth`, etc.).

APPENDIX N: SECTION HEADINGS

Here are some sub-sections:

N.1. A sub-section. Regular text.

N.1.1. *A sub-sub-section.* Regular text.

APPENDIX O: TEXT

O.1. Lists. The following is an example of an *itemized* list, two levels deep.

- This is the first item of an itemized list. Each item in the list is marked with a "tick." The document style determines what kind of tick mark is used.
- This is the second item of the list. It contains another list nested inside it.
  – This is the first item of an itemized list that is nested within the itemized list.
  – This is the second item of the inner list. LATEX allows you to nest lists deeper than you really should.
  This is the rest of the second item of the outer list.
- This is the third item of the list.

The following is an example of an *enumerated* list of one level.

  (i) This is the first item of an enumerated list.
 (ii) This is the second item of an enumerated list.

The following is an example of an *enumerated* list, two levels deep.

  1. This is the first item of an enumerated list. Each item in the list is marked with a "tick." The document style determines what kind of tick mark is used.
  2. This is the second item of the list. It contains another list nested inside of it.

     (i) This is the first item of an enumerated list that is nested within.
     (ii) This is the second item of the inner list. LATEX allows you to nest lists deeper than you really should.

This is the rest of the second item of the outer list.
  3. This is the third item of the list.

O.2. Punctuation. Dashes come in three sizes: a hyphen, an intra-word dash like "*U*-statistics" or "the time-homogeneous model"; a medium dash (also called an "en-dash") for number ranges or between two equal entities like "1–2" or "Cauchy–Schwarz inequality"; and a punctuation dash (also called an "em-dash") in place of a comma, semicolon, colon or parentheses—like this.

Generating an ellipsis . . . with the right spacing around the periods requires a special command.

O.3. Citation.

APPENDIX P: FONTS

Please use text fonts in text mode, e.g.:

Roman
Italic
Bold
SMALL CAPS
Sans serif
`Typewriter`

Please use mathematical fonts in mathematical mode, e.g.:

$\mathrm{ABCabc123}$
$ABCabc123$
$\mathbf{ABCabc123}$
$\boldsymbol{ABCabc123\alpha\beta\gamma}$
$\mathcal{ABC}$
$\mathbb{ABC}$
$\mathsf{ABCabc123}$
$\mathtt{ABCabc123}$
$\mathfrak{ABCabc123}$

Note that \mathcal, \mathbb belongs to capital letters-only font typefaces.

APPENDIX Q: NOTES

Footnotes[1] pose no problem.[2]

APPENDIX R: QUOTATIONS

Text is displayed by indenting it from the left margin. There are short quotations

> This is a short quotation. It consists of a single paragraph of text. There is no paragraph indentation.

and longer ones.

> This is a longer quotation. It consists of two paragraphs of text. The beginning of each paragraph is indicated by an extra indentation.
>
> This is the second paragraph of the quotation. It is just as dull as the first paragraph.

APPENDIX S: ENVIRONMENTS

S.1. Examples for `plain`-style environments.

AXIOM 1. *This is the body of Axiom 1.*

PROOF. This is the body of the proof of the axiom above. □

CLAIM 2. *This is the body of Claim 2. Claim 2 is numbered after Axiom 1 because we used* [axiom] *in* \newtheorem.

---

[1]This is an example of a footnote.

[2]Note that footnote number is after punctuation.

TABLE 1
The spherical case ($I_1 = 0$, $I_2 = 0$)

Equil. points	x	y	z	C	S
L_1	-2.485252241	0.000000000	0.017100631	8.230711648	U
L_2	0.000000000	0.000000000	3.068883732	0.000000000	S
L_3	0.009869059	0.000000000	4.756386544	-0.000057922	U
L_4	0.210589855	0.000000000	-0.007021459	9.440510897	U
L_5	0.455926604	0.000000000	-0.212446624	7.586126667	U
L_6	0.667031314	0.000000000	0.529879957	3.497660052	U
L_7	2.164386674	0.000000000	-0.169308438	6.866562449	U
L_8	0.560414471	0.421735658	-0.093667445	9.241525367	U
L_9	0.560414471	-0.421735658	-0.093667445	9.241525367	U
L_{10}	1.472523232	1.393484549	-0.083801333	6.733436505	U
L_{11}	1.472523232	-1.393484549	-0.083801333	6.733436505	U

THEOREM S.1. *This is the body of Theorem S.1. Theorem S.1 numbering is dependent on section because we used* [section] *after* \newtheorem.

THEOREM S.2 (Title of the theorem). *This is the body of Theorem S.2. Theorem S.2 has additional title.*

LEMMA S.3. *This is the body of Lemma S.3. Lemma S.3 is numbered after Theorem S.2 because we used* [theorem] *in* \newtheorem.

PROOF OF LEMMA S.3. This is the body of the proof of Lemma S.3.   □

S.2. Examples for `definition`-style environments.

DEFINITION S.4. This is the body of Definition S.4. Definition S.4 is numbered after Lemma S.3 because we used [theorem] in \newtheorem.

EXAMPLE. This is the body of the example. Example is unnumbered because we used \newtheorem* instead of \newtheorem.

FACT. This is the body of the fact. Fact is unnumbered because we used \newtheorem* instead of \newtheorem.

APPENDIX T: TABLES AND FIGURES

Cross-references to labeled tables: As you can see in Table 1 and also in Table 2.
Sample of cross-reference to figure. Figure 4 shows that it is not easy to get something on paper.

APPENDIX U: EQUATIONS AND THE LIKE

Two equations:

$$(42) \qquad C_s = K_M \frac{\mu / \mu_x}{1 - \mu / \mu_x}$$

and

$$(43) \qquad G = \frac{P_{\text{opt}} - P_{\text{ref}}}{P_{\text{ref}}} 100(\%).$$

TABLE 2
Sample posterior estimates for each model

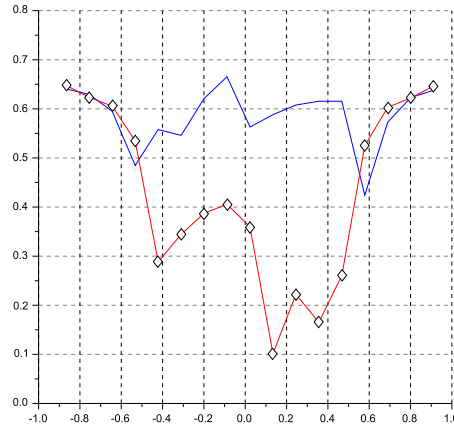| Model | Parameter | Mean | Std. dev. | Quantile | | |
				2.5%	50%	97.5%
Model 0	β_0	-12.29	2.29	-18.04	-11.99	-8.56
	β_1	0.10	0.07	-0.05	0.10	0.26
	β_2	0.01	0.09	-0.22	0.02	0.16
Model 1	β_0	-4.58	3.04	-11.00	-4.44	1.06
	β_1	0.79	0.21	0.38	0.78	1.20
	β_2	-0.28	0.10	-0.48	-0.28	-0.07
Model 2	β_0	-11.85	2.24	-17.34	-11.60	-7.85
	β_1	0.73	0.21	0.32	0.73	1.16
	β_2	-0.60	0.14	-0.88	-0.60	-0.34
	β_3	0.22	0.17	-0.10	0.22	0.55

Fig 4: Pathway of the penicillin G biosynthesis.

Equation arrays:

$$\frac{dS}{dt} = -\sigma X + s_F F, \tag{44}$$

$$\frac{dX}{dt} = \mu X, \tag{45}$$

$$\frac{dP}{dt} = \pi X - k_h P, \tag{46}$$

$$\frac{dV}{dt} = F. \tag{47}$$

One long equation:

$$
\begin{aligned}
\mu_{\text{normal}} &= \mu_x \frac{C_s}{K_x C_x + C_s} \\
&= \mu_{\text{normal}} - Y_{x/s}\big(1 - H(C_s)\big)(m_s + \pi/Y_{p/s}) \\
&= \mu_{\text{normal}}/Y_{x/s} + H(C_s)(m_s + \pi/Y_{p/s}).
\end{aligned}
\tag{48}
$$

APPENDIX: APPENDICES MUST BE MOVED TO A SUPPLEMENT FILE

APPENDIX: ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

APPENDIX: FUNDING

SUPPLEMENTARY MATERIAL

Title of Supplement A
Short description of Supplement A.

Title of Supplement B
Short description of Supplement B.

REFERENCES

Aguilera, A. M., Aguilera-Morillo, M. C., and Preda, C. (2016). Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92.

Aguilera, A. M., Escabias, M., Preda, C., and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*. Wiley-Interscience, Hoboken, N.J.

Beyaztas, U. and Lin Shang, H. (2022). A robust functional partial least squares for scalar-on-multiple-function regression. *Journal of Chemometrics*, 36(4):e3394.

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

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–591.

Chang, C., Jang, J. H., Manatunga, A., Taylor, A. T., and Long, Q. (2020). A bayesian latent class model to predict kidney obstruction in the absence of gold standard. *Journal of the American Statistical Association*, 115(532):1645–1663.

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.

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.

Gockenbach, M. S. (2010). *Finite-Dimensional Linear Algebra*. CRC Press, Boca Raton, FL, 1st edition edition.

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.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

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.

Jang, J. H. (2021). Principal component analysis of hybrid functional and vector data. *Statistics in Medicine*, 40(24):5152–5173.

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

Mutis, M., Beyaztas, U., Karaman, F., and Lin Shang, H. (2025). On function-on-function linear quantile regression. *Journal of Applied Statistics*, 52(4):814–840.

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

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

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.

Saricam, S., Beyaztas, U., Asikgil, B., and Shang, H. L. (2022). On partial least-squares estimation in scalar-on-function regression models. *Journal of Chemometrics*, 36(12):e3452.

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

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.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

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.