HYBRID PARTIAL LEAST SQUARES REGRESSION WITH MULTIPLE FUNCTIONAL AND SCALAR PREDICTORS

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

[1]*Data Sciences and Operations Department, Marshall School of Business, University of Southern California,* [a]*jongmin.mun@marshall.usc.edu*

[2]*Department of Biostatistics and Data Science, 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 (scalar data) for each kidney. To effectively analyze such distinct yet related physiological signals, we construct a joint linear regression model incorporating both functional and scalar-valued covariates.

$$Y_i = \boldsymbol{\beta}^\top \mathbf{Z}_i + \sum_{k=1}^{K} \int \beta_k(t) X_{ik}(t) dt + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where Y_i is a scalar response, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ is a Euclidean vector covariate, $X_{i1}(t), \ldots, X_{iK}(t)$ are functional predictors that belong to $L^2[0,1]$, and ϵ_i is observational noise. In other words, this is a scalar-on-hybrid regression where the hybrid covariates belong to $L^2[0,1]^{\otimes K} \times \mathbb{R}^p$. For notational convenience, we assume throughout this paper that the responses and predictors have been centered, allowing the intercept term to be ignored.

This hybrid model poses two-fold challenges. First, estimating the infinite-dimensional slope function $\beta_k(t)$ is an ill-posed problem requiring dimension reduction or regularization based on structural assumption, such as derivatives and covariance patterns. Second, 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, Ferraty and Sarda, 2003; Cai and Hall, 2006) - Wavelet(Zhao, Ogden and Reiss, 2012)

Keywords and phrases: dimension reduction, functional data analysis, multiple data modalities, multivariate data analysis, multivariate functional data, partial least square.

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, Galeano and González-Manteiga, 2017a), but it overlooks potentially strong correlations between functional and scalar components, leading to multicollinearity and sub-optimal 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, Aguilera-Morillo and Preda, 2016; Febrero-Bande, Galeano and González-Manteiga, 2017a; 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. Background on partial least squares and its extension to hybrid predictors. JM: still revising For intuition, let us return to the high-dimensional Euclidean predictor setting $Y_i = \boldsymbol{\beta}^\top \mathbf{Z}_i + \epsilon_i$. A common way to address ill-posedness and correlation is to approximate the high-dimensional vector $\mathbf{Z}_i$ using a low-dimensional vector $\left(\widehat{\rho}_1^{[1]}, \ldots, \widehat{\rho}_1^{[L]}\right)^\top$. To retain the regression relationship, the l-th PLS direction $\hat{\boldsymbol{\xi}}_l$ solves:

$$\max_{\boldsymbol{\alpha}} \widehat{\mathrm{Cov}}^2\left(\{\langle \boldsymbol{\alpha}, \mathbf{Z}_i\rangle, Y_i\}_{i=1}^n\right) \text{ s.t. } \|\boldsymbol{\alpha}\|_2 = 1,\ \boldsymbol{\alpha}^\top \widehat{\mathrm{Cov}}^2\left(\{\mathbf{Z}_i\}_{i=1}^n\right)\hat{\boldsymbol{\xi}}_j = 0, \quad j = 1, \ldots, l-1,$$

where the two $\widehat{\mathrm{Cov}}^2$ denote sample cross-covariance and sample covariance, respectively. A standard algorithm for solving this problem, called nonlinear iterative partial least squares (NIPALS) is presented in Algorithm 1. (4)

We can express the predictor W_i and the regression coefficient β using an orthonormal basis $\psi_1, \psi_2, \ldots$ of $\mathcal{H}$, so that $W_i = \sum_j \langle W_i, \psi_j\rangle_{\mathcal{H}}\, \psi_j$ and $\beta = \sum_j \langle \beta, \psi_j\rangle_{\mathcal{H}}\, \psi_j$. To make the problem tractable, we truncate these expansions to the first L terms. However, if the basis is chosen without considering the data, the most important regression information may not be captured within these first L terms. The core idea of partial least squares (PLS) is simultaneously learn a data-driven, regression-aware orthonormal basis while also performing the regression. This results in a simultaneous decomposition of the predictor and outcome in terms of uncorrelated PLS scores

$$W_i = \sum_{l=1}^{\infty} \rho_l \delta_l \quad \text{and} \quad Y_i = \sum_{l=1}^{\infty} \rho_l \nu_l + \epsilon_i, \tag{2}$$

where $(\boldsymbol{\delta}_l)_{l\in\mathbb{N}} \in \mathcal{H}$ and $(\nu_l)_{l\in\mathbb{N}} \in \mathbb{R}$ are appropriate bases.

Algorithm 1 Scalar partial least squares

Standardize each $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ so that each feature have mean zero and variance one. Standardize $Y_1, \ldots, Y_n$.
for $l = 1, 2, \ldots, L$ **do**
PLS direction and score estimation:
  $\widehat{\boldsymbol{\xi}}_l \leftarrow \arg\max_{\boldsymbol{\alpha}} \widehat{\mathrm{Cov}}^2\big(\{\langle \boldsymbol{\alpha}, \mathbf{Z}_i^{[l]} \rangle, Y_i^{[l]}\}_{i=1}^n\big)$ s.t. $\|\boldsymbol{\alpha}\|_2 = 1$ ▷ PLS direction
  $\widehat{\rho}_i^{[l]} \leftarrow \langle \hat{\boldsymbol{\xi}}_l, \mathbf{Z}_i \rangle,\ i = 1, \ldots, n$ ▷ PLS score
Residualization:
  $\nu^{[l]} \leftarrow \frac{\sum_{i=1}^n Y_i^{[l]} \widehat{\rho}_i^{[l]}}{\sum_{i=1}^n \widehat{\rho}_i^{[l]2}}$
  $Y_i^{[l+1]} \leftarrow Y_i^{[l]} - \nu^{[l]} \widehat{\rho}_i^{[l]}\ i = 1, \ldots, n$
  $\widehat{\boldsymbol{\delta}}^{[l]} \leftarrow \frac{1}{\sum_{i=1}^n \widehat{\rho}_i^{[l]2}} \sum_{i=1}^n \widehat{\rho}_i^{[l]} \mathbf{Z}_i^{[l]}$.
  $\mathbf{Z}_i^{[l+1]} \leftarrow \mathbf{Z}_i^{[l]} - \widehat{\rho}_i^{[l]} \widehat{\boldsymbol{\delta}}^{[l]}$
Regression coefficient recovery:
$\widehat{\boldsymbol{\zeta}}^{[1]} \leftarrow \widehat{\boldsymbol{\xi}}^{[1]}$
for $l = 2, \ldots, L$ **do**
  $\widehat{\boldsymbol{\zeta}}^{[l]} \leftarrow \widehat{\boldsymbol{\xi}}^{[l]} - \sum_{u=1}^{l-1} \langle \widehat{\boldsymbol{\delta}}^{[u]}, \widehat{\boldsymbol{\xi}}^{[l]} \rangle \widehat{\boldsymbol{\zeta}}^{[u]}$
$\widehat{\boldsymbol{\beta}} \leftarrow \sum_{l=1}^{L} \widehat{\nu}^{[l]} \widehat{\boldsymbol{\zeta}}^{[l]}$
Output: the regression coefficient estimate

3. Population PLS and Tucker's Criterion in hybrid regression setting. JM: Previously, was in theoretical property section, now procedes the method section, to motivate the estimation procedure

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 $\langle \xi_l, \xi_j \rangle_{\mathcal{H}} = \mathbb{1}(l = j)$.

Challenges

1. Independent variables consist of multiple highly structured images and scalar predictors.
2. Our sample size is small compared to the dimension and number of functional and scalar predictors.
3. Existing partial least squares (PLS) methods can only accommodate (i) univariate or multivariate functional predictors without any scalar predictors (Preda and Saporta, 2005; Delaigle and Hall, 2012; Febrero-Bande, Galeano and González-Manteiga, 2017b; Beyaztas and Shang, 2020); or (2) a univariate functional predictor with other scalar predictors (Wang, 2018).

The fundamental property of classical scalar PLS is that the PLS component direction corresponds to the first eigenvector of the squared covariance matrix, satisfying the Tucker's Criterion (Tucker, 1938). Under a mild assumption, We derive an analogous result in our scalar-on-hybrid regression model setting (11). This justifies our estimation procedure introduced in Section 5. We omit l in the notations and first define the cross-covariance term between the response and the functional predictors (observed versions if $l = 1$ and residualized versions if $l \geq 2$):

$$\sigma_{YX} = \big(\sigma_{YX,1}(t), \ldots, \sigma_{YX,K}(t)\big) := \big(\mathbb{E}[Y_1 X_{11}], \ldots, \mathbb{E}[Y_1 X_{1K}]\big) \in \mathcal{F}^{\otimes K}, \tag{3}$$

and between the response and the scalar predictor:

$$\sigma_{YZ} = (\sigma_{YZ,1}, \ldots, \sigma_{YZ,p})^\top := \big(\mathbb{E}[Y_1 Z_{11}], \ldots, \mathbb{E}[Y_1 Z_{1p}]\big)^\top \in \mathbb{R}^p.$$

Based on these definitions, we define the cross-covariance term between the response and the hybrid predictor as

$$\Sigma_{YW} := \mathbb{E}[Y_1 W_1] = \big(\sigma_{YX,1}, \ldots, \sigma_{YX,K}, \sigma_{YZ}\big) \in \mathcal{H}.$$

Based on these definitions, we introduce two cross-covariance operators and their properties.

LEMMA 3.1. *Define the operator $\mathcal{C}_{YW} = \mathbb{E}(W_1 \otimes_{\mathcal{H}} Y_1) : \mathcal{H} \to \mathbb{R}$ such that for any $h = (f_1, \ldots, f_K, \mathbf{v}) \in \mathcal{H}$, it maps h to a real number as:*

$$\mathcal{C}_{YW} h := \mathbb{E}\left[\langle W_1, h\rangle_{\mathcal{H}} Y_1\right] = \langle \Sigma_{YW}, h\rangle_{\mathcal{H}} = \sum_{k=1}^{K} \int_0^1 \sigma_{YX,k}(t)\, f_k(t)\, dt + \sigma_{YW}^\top \mathbf{v}.$$

JM: instead of showing $\mathcal{U}$ is compact, let's show $\mathcal{C}_{YW}$ is compact. This way, we don't need uniform continuiity of σ_{YX}. If there exist finite constants Q_1 and Q_2 such that

$$\max_{k=1,\ldots,K} \sup_{t\in[0,1]} \sigma^2_{YX,k}(t) < Q_1 \quad \text{and} \quad \max_{r=1,\ldots,p} \sigma^2_{YZ,r} < Q_2, \tag{4}$$

the operator $\mathcal{C}_{YW}$ is a compact operator.

Proof of Lemma 3.1 is provided in Appendix D.1.

LEMMA 3.2. *Define $\mathcal{C}_{WY} = \mathbb{E}[Y \otimes W] : \mathbb{R} \to \mathcal{H}$, which maps $d \in \mathbb{R}$ to a hybrid object in $\mathcal{H}$ as follows:*

$$\mathcal{C}_{WY} d := \mathbb{E}\big[\langle Y_1, d\rangle W_1\big] = \mathbb{E}[Y_1 W_1 d] = d\,\Sigma_{YW}.$$

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

The proof of Lemma 3.2 is provided in Appendix D.2.
Based on these two operators, we define a positive operator as follows:

LEMMA 3.3 (Composite cross-covariance operator). *Define $\mathcal{U} := \mathcal{C}_{WY} \circ \mathcal{C}_{YW} : \mathcal{H} \to \mathcal{H}$ as an operator which performs the following mapping:*

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

In other words, $\mathcal{U} = \Sigma_{YW} \otimes \Sigma_{YW}$. Then $\mathcal{U}$ is a self-adjoint and positive-semidefinite operator. Under the conditions of Lemma 3.1, $\mathcal{U}$ is a compact operator.

The proof of Lemma 3.3 is provided in Appendix D.3. By the Hilbert-Schmidt theorem (e.g., Theorem 4.2.4 in Hsing and Eubank, 2015), Lemma 3.3 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 theorem introduces the population-level PLS and Tucker's Criterion, adapted to our scalar-on-hybrid regression model setting, based on the aforementioned operators defined in the hybrid space.

THEOREM 3.4 (Population PLS and Tucker's criterion). *Under the conditions of Lemma 3.1, the constrained maximum*

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

is attained by the eigenfunction associated with the largest eigenvalue of the operator $\mathcal{U}$.

The proof of Theorem 3.4 is provided in Appendix D.4. 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: $\Sigma_{YW}^{[l]} = E(Y^{[l]}W^{[l]})$.

4. 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

$$Y_i^{[1]} = Y_i - \bar{Y} \text{ and } \mathbf{W}_i^{[1]} = (X_i^{[1]}, \mathbf{Z}_i^{[1]}) = (X_i - \bar{X}, \mathbf{Z}_i - \bar{\mathbf{Z}}) = \mathbf{W}_i - \bar{\mathbf{W}}. \tag{5}$$

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 **??**.

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

$$\widehat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \widehat{\xi}_l\rangle_{\mathcal{H}} = \langle X_i^{[l]}, \widehat{\psi}_l\rangle_{\mathcal{F}} + \langle \mathbf{Z}_i^{[l]}, \widehat{\boldsymbol{\theta}}_l\rangle, \tag{6}$$

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

$$\widehat{\xi}_l(\mathbf{t}) = \frac{\sum_{i=1}^n Y_i^{[l]}\mathbf{W}_i^{[l]}(\mathbf{t})}{\|\sum_{i=1}^n Y_i^{[l]}\mathbf{W}_i^{[l]}\|_{\mathcal{H}}} = \left[\frac{\sum_{i=1}^n Y_i^{[l]}X_i^{[l]}(\mathbf{t})}{\|\sum_{i=1}^n Y_i^{[l]}\mathbf{W}_i^{[l]}\|_{\mathcal{H}}}, \frac{\sum_{i=1}^n Y_i^{[l]}\mathbf{Z}_i^{[l]}}{\|\sum_{i=1}^n Y_i^{[l]}\mathbf{W}_i^{[l]}\|_{\mathcal{H}}}\right]^\top, \tag{7}$$

where the normalizing factor is computed as

$$\|\sum_{i=1}^n Y_i^{[l]}\mathbf{W}_i^{[l]}\|_{\mathcal{H}} = [\sum_{i=1}^n\sum_{j=1}^n Y_i^{[l]}Y_j^{[l]}\{\langle X_i^{[l]}, X_j^{[l]}\rangle_{\mathcal{F}} + \langle \mathbf{Z}_i^{[l]}, \mathbf{Z}_j^{[l]}\rangle\}]^{1/2}. \tag{8}$$

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.

5. Proposed PLS Algorithm. The hybrid random predictor and corresponding inner product is formally defined as follows.

DEFINITION 5.1 (Hybrid predictor and inner product). We define a hybrid predictor W_i as an ordered pair combining the functional predictors $X_{i1}(t), \ldots, X_{iK}(t)$ and a scalar predictor $\mathbf{Z}_i$:

$$W_i := (X_{i1}, \ldots, X_{iK}, \mathbf{Z}_i) \in \mathcal{H} := \mathcal{F}^{\otimes K} \times \mathbb{R}^p \tag{9}$$

The inner product between two hybrid predictors, $W_1 = (X_{11}, \ldots, X_{1K}, \mathbf{Z}_1)$ and $W_2 = (X_{21}, \ldots, X_{2K}, \mathbf{Z}_2)$, is then defined as:

$$\langle W_1, W_2 \rangle_{\mathcal{H}} := \sum_{k=1}^{K} + \omega \mathbf{Z}_1^\top \mathbf{Z}_2, \text{ where } \int_0^1 X_{1k}(t) X_{2k}(t)\, dt \tag{10}$$

for $\omega > 0$. The corresponding norm $\|\cdot\|_{\mathcal{H}}$ is derived from this inner product as $\langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$.

In (10), 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 5.2). 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.

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.

Then the joint regression model (1) can be concisely written as

$$Y_i = \langle \beta, W_i \rangle_{\mathcal{H}} + \epsilon_i, \text{ where } \beta := \big(\beta_1(t), \ldots, \beta_K(t), \boldsymbol{\beta}\big) \in \mathcal{H}. \tag{11}$$

To address the limitations outlined in Section 4, 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 5.3.1 and 5.3.2, respectively. After a suitable number of iterations, the hybrid regression coefficient is estimated, as described in Section 5.4. 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.

5.1. *Preliminary step 1: 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 $\mathcal{B} := \operatorname{span}\big(b_1(t), \ldots, b_M(t)\big)$. 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 5.3.1). 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:

$$\xi_j(t) = \sum_{m=1}^{M} \gamma_{jm}\, b_m(t), \quad \delta_j(t) = \sum_{m=1}^{M} \pi_{jm} b_m(t) \tag{12}$$

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

$$\widetilde{\mathcal{H}} := \mathcal{B}^K \times \mathbb{R}^p \subset \mathcal{H}. \tag{13}$$

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

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

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

$$\boldsymbol{\theta}_{ij} := (\theta_{ij1}, \ldots, \theta_{ijM})^\top,\ \boldsymbol{\eta}_j := (\eta_{j1}, \ldots, \eta_{jM})^\top,\ \boldsymbol{\gamma}_j := (\gamma_{j1}, \ldots, \gamma_{jM})^\top,\ \boldsymbol{\pi}_j := (\pi_{j1}, \ldots, \pi_{jM})^\top \tag{14}$$

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

$$\Theta_j := (\boldsymbol{\theta}_{1j}, \ldots, \boldsymbol{\theta}_{nj})^\top \in \mathbb{R}^{n \times M}, \tag{15}$$

and construct the full design matrix

$$\Theta := (\Theta_1, \ldots, \Theta_K, \mathbf{Z}) \in \mathbb{R}^{n \times (MK+p)}. \tag{16}$$

Let us denote the response vector as $\mathbf{y} := (y_1, \ldots, y_n)^\top$. Let $B, B'' \in \mathbb{R}^{M \times M}$ denote the Gram matrices of the basis functions and their second derivatives, with entries

$$B_{m,m'} := \int_0^1 b_m(t)\, b_{m'}(t)\, dt, \quad B''_{m,m'} := \int_0^1 b''_m(t)\, b''_{m'}(t)\, dt, \tag{17}$$

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

$$\mathbb{B} := \operatorname{blkdiag}(B, \ldots, B, I_p), \quad \mathbb{B}'' := \operatorname{blkdiag}(B'', \ldots, B'', I_p), \tag{18}$$

Then the full data for the hybrid PLS problem at the l-th iteration can be represented by the tuple

$$(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y}) \in \mathbb{R}^{(MK+p)\times(MK+p)} \times \mathbb{R}^{(MK+p)\times(MK+p)} \times \mathbb{R}^{n\times(MK+p)} \times \mathbb{R}^n, \tag{19}$$

with the index l omitted for brevity.

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.

5.2. *Preliminary step 2 : data preprocessing.* JM: notation not changed yet 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 (10) 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 $\omega^{1/2}$.

5.3. *Iterative steps.* The iterative process presented here yields an orthonormal hybrid basis that effectively captures the predictor-response relationships. It proceeds through two intermediate steps: the estimation of the PLS component direction (Section 5.3.1) and residualization (Section 5.3.2). The properties of the resulting estimates are introduced in Section 6.

5.3.1. *Iterative step 1: regularized estimation of PLS component direction.* We begin by formally introducing the core optimization problem pertinent to the PLS direction estimation, which is formulated as a generalized Rayleigh quotient (Proposition 5.2). Building upon this foundational concept, we present our regularized PLS component direction estimation step that promotes smoothness (Proposition 5.3). Furthermore, we detail an efficient computational scheme (Proposition 5.4).

Core optimization problem. We present the core optimization problem that directly estimates the hybrid PLS component direction. It fully leverages the continuous nature of the functional components and the function-scalar hybrid structure. 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

$$\widehat{\mathrm{Cov}}(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y) := \frac{1}{n} \sum_{i=1}^n y_i \langle \widetilde{W}_i, \xi \rangle_{\mathcal{H}}. \tag{20}$$

We denote the estimated PLS component direction as

$$\hat{\xi} := \arg\max_{\xi \in \widetilde{\mathcal{H}}} \widehat{\mathrm{Cov}}^2\big(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y\big) \; s.t. \; \|\xi\|_{\mathcal{H}} = 1. \tag{21}$$

Here, $\hat{\xi} \in \widetilde{\mathcal{H}}$ is an ordered pair expanded as:

$$\hat{\xi} = \big(\hat{\xi}_1(t), \ldots, \hat{\xi}_K(t), \hat{\boldsymbol{\zeta}}\big) = \left(\sum_{m=1}^{M} \hat{\gamma}_{1m}\, b_m(t), \ldots, \sum_{m=1}^{M} \hat{\gamma}_{Km}\, b_m(t), \hat{\boldsymbol{\zeta}}\right),$$

where $\hat{\boldsymbol{\zeta}}$ is the scalar part. Obtaining these coefficients is equivalent to solving the maximization problem (21). The following proposition formulates this coefficients obtaining procedure as a generalized Rayleigh quotient:

PROPOSITION 5.2. *Let $(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y})$ denote the observed data defined in (19). At the l-th iteration of the PLS algorithm, the coefficients of the squared covariance maximizer defined in (21), is obtained as*

$$\left(\hat{\gamma}_{11}, \ldots, \hat{\gamma}_{1M}, \ldots, \hat{\gamma}_{K1}, \ldots, \hat{\gamma}_{KM}, \hat{\boldsymbol{\zeta}}^\top\right)^\top = \arg\max_{\boldsymbol{\xi} \in \mathbb{R}^{MK+p}} \boldsymbol{\xi}^\top V \boldsymbol{\xi} \quad \textit{subject to} \quad \boldsymbol{\xi}^\top \mathbb{B} \boldsymbol{\xi} = 1. \tag{22}$$

where

$$V := \frac{1}{n^2} (\mathbb{B}\Theta^\top \mathbf{y})(\mathbb{B}\Theta^\top \mathbf{y})^\top \in \mathbb{R}^{(MK+p)\times(MK+p)}. \tag{23}$$

The proof of Proposition 5.2 is provided in Appendix E.

Proposed regularized estimation procedure. Although Proposition 5.2 efficiently estimates the PLS component direction $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_K, \hat{\boldsymbol{\zeta}})$, 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

$$\mathrm{PEN}(\hat{\xi}_j) := \int_0^1 \big\{\hat{\xi}_j''(t)\big\}^2 dt. \tag{24}$$

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 (21) 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.

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:

$$\hat{\xi} := \arg\max_{\xi \in \widetilde{\mathcal{H}}} \widehat{\mathrm{Cov}}^2\left(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y\right) \; s.t. \; \|\xi\|_{\mathcal{H}} + \sum_{j=1}^{K} \lambda_j \, \mathrm{PEN}(\xi_K) = 1. \tag{25}$$

Here, $\hat{\xi} \in \widetilde{\mathcal{H}}$ is an ordered pair expanded as:

$$\hat{\xi} = \left(\hat{\xi}_1(t), \ldots, \hat{\xi}_K(t), \hat{\boldsymbol{\zeta}}\right) = \left(\sum_{m=1}^{M} \hat{\gamma}_{1m} \, b_m(t), \ldots, \sum_{m=1}^{M} \hat{\gamma}_{Km} \, b_m(t), \hat{\boldsymbol{\zeta}}\right),$$

where $\hat{\boldsymbol{\zeta}}$ is the scalar part. This formulation maximizes the squared covariance over a class of smooth functions. Obtaining these coefficients is equivalent to solving the maximization problem (25). The following proposition formulates this coefficients obtaining procedure as a generalized Rayleigh quotient:

PROPOSITION 5.3 (Regularized estimation of PLS component direction). *Let* $(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y})$ *denote the data given at the* l*-th iteration, as defined in* (19)*. Recall from* (23) *that* $V = n^{-2}(\mathbb{B}\Theta^\top \mathbf{y})(\mathbb{B}\Theta^\top \mathbf{y})^\top$ *. Let* $\Lambda \in \mathbb{R}^{(MK+p)\times(MK+p)}$ *be defined as:*

$$\Lambda := \mathrm{blkdiag}(\lambda_1 I_M, \ldots, \lambda_K I_M, 0_{p\times p}), \; \textit{where } \lambda_1, \ldots, \lambda_K \geq 0. \tag{26}$$

Here, $0_{p\times p}$ *denotes the* $p \times p$ *zero matrix. The coefficients of the squared covariance maximizer defined in* (25)*, are obtained as*

$$\left(\hat{\gamma}_{11}, \ldots, \hat{\gamma}_{1M}, \ldots, \hat{\gamma}_{K1}, \ldots, \hat{\gamma}_{KM}, \hat{\boldsymbol{\zeta}}^\top\right)^\top = \arg\max_{\boldsymbol{\xi} \in \mathbb{R}^{MK+p}} \boldsymbol{\xi}^\top V \boldsymbol{\xi} \;\; s.t. \;\; \boldsymbol{\xi}^\top (\mathbb{B} + \Lambda \mathbb{B}'') \boldsymbol{\xi} = 1. \tag{27}$$

The proof of Proposition 5.3 is provided in Appendix F. The constraint $\boldsymbol{\xi}^\top(\mathbb{B} + \Lambda\mathbb{B}'')\boldsymbol{\xi} = 1$ enforces the orthonormality of the estimated PLS component directions with respect to a modified inner product (see Section **??** for details). The smoothing parameter λ_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 5.2. 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 5.3 may be computationally unstable in practice. However, by leveraging the rank-one structure of the matrix V Proposition 5.4 derives a closed-form solution that requires only the solution of linear systems.

PROPOSITION 5.4 (Closed-form solution). *Consider the optimization problem described in Proposition 5.3. Define the following quantities, which depend on the observed data but are not decision variables:*

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

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

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

The proof of Proposition 5.4 is provided in Appendix G. 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.

5.3.2. *Iterative step 2: residualization 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:

$$\hat{\rho}_i^{[l]} := \langle \widetilde{W}_i^{[l]}, \hat{\xi}^{[l]} \rangle_{\mathcal{H}}, \tag{28}$$

using the estimated PLS component direction $\widehat{\xi}^{[l]}$ obtained from Propositions 5.3 and 5.4. 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. To obtain the $(l+1)$-th iteration's responses and hybrid predictors, we regress the (l)-th iteration's responses and hybrid predictors on these PLS scores by least squares and then residualize. Specifically, the $(l+1)$-th predictor is computed as a hybrid-on-scalar linear regression residual. In the same spirit as the PLS component direction estimation step, rather than treating the hybrid object as a long vector of concatenated function evaluations at time points and scalar vectors, we employ a basis expansion approach to fit the entire hybrid object in one step. Therefore, our method is computationally efficient, applicable for dense or irregular functional data, and keeps the residual smooth, as it enforces the regression coefficient, $\delta^{[l]}$, as an element of $\tilde{\mathcal{H}}$ as defined in Equation (13). Consequently, $\delta^{[l]}$ is obtained by minimizing a least squares criterion:

$$\hat{\delta}^{[l]} := \arg\min_{\delta \in \tilde{\mathcal{H}}} \sum_{i=1}^{n} \|\widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]}\, \delta\|_{\mathcal{H}}^2. \tag{29}$$

On the other hand, the $(l+1)$-th response is computed as a scalar-on-scalar linear regression residual. The following proposition demonstrates that this residualization step can be performed simply, analogous to scalar PLS.

LEMMA 5.5 (Closed-form solution). *Let us denote* $\hat{\boldsymbol{\rho}}^{[l]} := (\hat{\rho}_1^{[l]}, \ldots, \hat{\rho}_n^{[l]})^\top$. *The* $(l+1)$*-th iteration's predictors and responses are computed as:*

$$\widetilde{W}_i^{[l+1]} := \widetilde{W}_i^{[l]} - \frac{\widehat{\rho}_{il}}{\|\hat{\boldsymbol{\rho}}^{[l]}\|_2^2} \sum_{i=1}^{n} \widehat{\rho}_{il} \widetilde{W}_i^{[l]}, \quad Y_i^{[l+1]} := Y_i^{[l]} - \frac{\mathbf{y}^{[l]\top} \hat{\boldsymbol{\rho}}^{[l]}}{\|\hat{\boldsymbol{\rho}}^{[l]}\|_2^2} \widehat{\rho}_{il}. \tag{30}$$

Proof of Lemma 5.5 is provided in Appendix H. JM: proof needs to be revised Since $\widetilde{W}_i^{[l]}$ and $Y_i^{[l]}$ are assumed to have a sample mean of zero, their respective residuals, $\widetilde{W}_i^{[l+1]}$ and $Y_i^{[l+1]}$, also maintain a zero sample mean.

5.4. *Final step: estimating the hybrid regression coefficient.* JM: notation not yet changed The regression coefficient $\boldsymbol{\eta}$ in model (11) can be written in terms of the estimated PLS components and scores. First, note that we can show $\widehat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \widehat{\xi}_l \rangle_{\mathcal{H}} = \langle \mathbf{W}_i, \widehat{\boldsymbol{\zeta}}_l \rangle_{\mathcal{H}}$, where $\widehat{\boldsymbol{\zeta}}_l = \widehat{\xi}_l - \sum_{u=1}^{l-1} \langle \widehat{\boldsymbol{\delta}}_u, \widehat{\xi}_l \rangle_{\mathcal{H}} \widehat{\boldsymbol{\zeta}}_u$ for $l \geq 2$, with $\widehat{\boldsymbol{\zeta}}_1 = \widehat{\xi}_1$. Then, the decomposition of Y_i in (2) leads to

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

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

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

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

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

A fundamental property of partial least squares is that its derived bases are orthonormal. Our regularized estimates preserve this property, with respect to a modified inner product that incorporates the roughness penalty, defines as follows:

DEFINITION 6.1 (Roughness-sensitive inner product). Given two hybrid predictors $W_1 = (X_{11}, \ldots, X_{1K}, \mathbf{Z}_1)$ and $W_2 = (X_{21}, \ldots, X_{2K}, \mathbf{Z}_2)$, both elements of $\mathcal{H}$ as defined in Definition 9, and a roughness penalty matrix $\Lambda = \text{blkdiag}(\lambda_1 I_M, \ldots, \lambda_K I_M, 0_{p\times p})$, the roughness-sensitive inner product between W_1 and W_2 is defined as:

$$\langle W_1, W_2 \rangle_{\mathcal{H},\Lambda} := \sum_{k=1}^{K} \int_0^1 X_{1k}(t) X_{2k}(t)\, dt + \sum_{k=1}^{K} \lambda_k \int_0^1 X''_{1k}(t) X''_{2k}(t)\, dt + \mathbf{Z}_1^\top \mathbf{Z}_2. \tag{31}$$

Based on this inner product, the following proposition states that the PLS component directions estimated from Proposition 5.3 are orthonormal.

PROPOSITION 6.2 (Orthonormality of estimated PLS component directions). *The PLS component directions $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_L$, estimated via Proposition 5.3 with a roughness penalty*

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 $\Lambda \in \mathbb{R}^{(MK+p)\times(MK+p)}$ (26)

1: $\widetilde{C}^{*[1]} \leftarrow \widetilde{C}^*$
2: **for** $l = 1, 2, \ldots T$ **do**
3: $\quad \mathbf{V}^* \leftarrow n^{-2} J^* \widetilde{C}^{*\top} \widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^\top \widetilde{C}^* J^*$ ▷ (??)
4: $\quad \widehat{\xi}_l^* \leftarrow \arg\max_{\xi^* \in \mathbb{R}^{MK+p}} \xi^{*\top} \mathbf{V}^* \xi^* \quad$ subject to $\quad \xi^{*\top}(J^* + \Lambda \ddot{J}^*)\xi^* = 1$ ▷ (27)
5: $\quad \widehat{\boldsymbol{\rho}}_l \leftarrow \widetilde{C}^* J^* \widehat{\xi}_l^*$ ▷ $\in \mathbb{R}^n$; (??)
6: Output a

matrix $\Lambda = \mathrm{blkdiag}(\lambda_1 I_M, \ldots, \lambda_K I_M, 0_{p\times p})$*, are mutually orthonormal with respect to the inner product* $\langle\cdot,\cdot\rangle_{\mathcal{H},\Lambda}$*. That is,*

$$\langle \widehat{\xi}_{l_1}, \widehat{\xi}_{l_2}\rangle_{\mathcal{H},\Lambda} = \mathbb{1}(l_1 = l_2), \quad l_1, l_2 = 1, \ldots, L.$$

The proof of Proposition 6.2 is provided in Appendix I.

The next proposition states that the vectors of estimated PLS scores for different iteration numbers are mutually orthogonal.

PROPOSITION 6.3. *Recall from Lemma 5.5 that* $\widehat{\boldsymbol{\rho}}^{[l]}$ *denote the* n*-dimensional vector whose elements consist of the* l*-th estimated PLS scores (*$l = 1, \ldots, L$*) of* n *observations. The vectors* $\widehat{\boldsymbol{\rho}}^{[1]}, \widehat{\boldsymbol{\rho}}^{[2]}, \ldots, \widehat{\boldsymbol{\rho}}^{[L]}$ *are mutually orthogonal in the sense that*

$$\widehat{\boldsymbol{\rho}}^{[l_1]\top}\widehat{\boldsymbol{\rho}}^{[l_2]} = 0 \quad \text{for} \quad l_1, l_2 \in \{1, \ldots, L\},\ l_1 \neq l_2.$$

The proof of Proposition 6.3 is provided in Appendix J.

7. 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

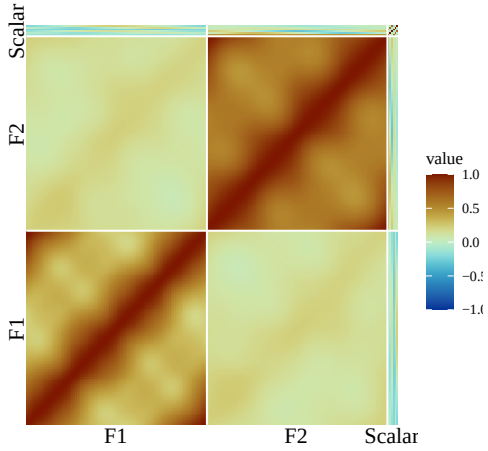
$$\log p(\mathbf{X} \mid \mathbf{M}, \mathbf{R}, \mathbf{C}) = -\frac{mn}{2}\log(2\pi) - \frac{m}{2}\log|\mathbf{C}| - \frac{n}{2}\log|\mathbf{R}| - \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}(\mathbf{X}-\mathbf{M})^\top \mathbf{R}^{-1}(\mathbf{X}-\mathbf{M})\right].$$

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.

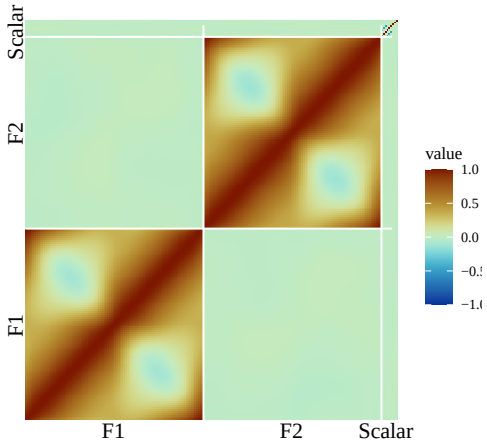
Building on the functional graphical model simulation of Zhu, Strawn and Dunson (2016), we generate a mixed graphical model with five nodes, described as follows:

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

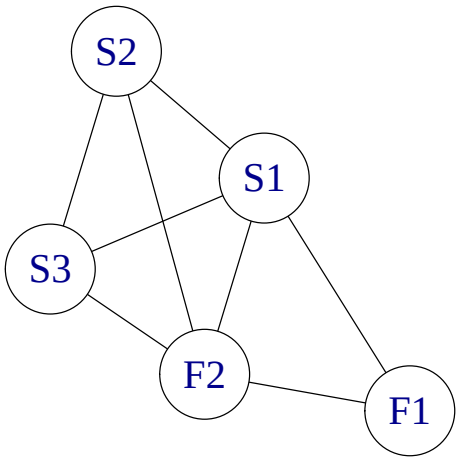
To capture dependencies among predictors, we introduce a graph structure that governs their conditional correlations. We consider two types of graph structures: a weakly connected graph and a strongly connected graph. In the Gaussian process framework, the precision matrix $\mathbf{R}_0^{-1}$ encodes 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



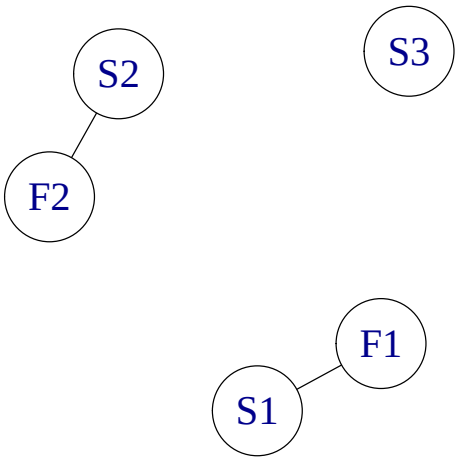
(a) Correlation Matrix - Strong Dependency



(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.

$\mathbf{I}_{M_i,M_j}$ 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.

7.1. *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, Liu and Xing (2014), which focuses on mixed attribute Gaussian graphical models.

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

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

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

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

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

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

The off-diagonal entries are:

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

Finally, the functional predictors are smoothed using a B-spline basis with 15 basis functions, as described in Section 5.3.1.

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

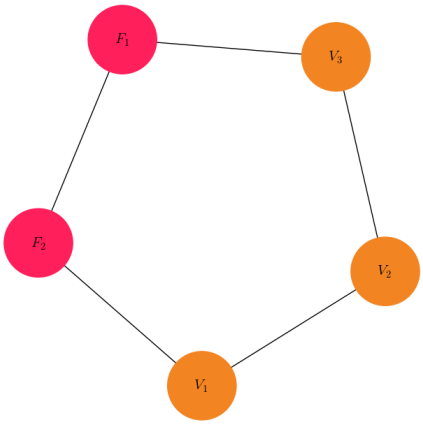


Fig 2

corresponding components are connected in the graph, and zero otherwise. Here, $\mathbf{1}$ denotes a matrix of appropriate dimensions where all elements are equal to 1. Overall, we generate a $2p + 3d$-dimensional multivariate Gaussian distribution with mean zero and a precision matrix Ω, structured as follows:

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

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

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

The baseline methods are:

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

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

8. 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

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

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.

8.1.

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.

Funding. The first author was supported by NSF Grant DMS-??-??????.
The second author was supported in part by NIH Grant ??????????.

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** 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** 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** 2159–2179.

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica* **13** 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** 1645–1663.

DELAIGLE, A. and HALL, P. (2012). Methodology and Theory for Partial Least Squares Applied to Functional Data. *The Annals of Statistics* **40** 322–352.

FEBRERO-BANDE, M., GALEANO, P. and GONZÁLEZ-MANTEIGA, W. (2017a). Functional Principal Component Regression and Functional Partial Least-Squares Regression: An Overview and a Comparative Study. *International Statistical Review* **85** 61–83.

FEBRERO-BANDE, M., GALEANO, P. and GONZÁLEZ-MANTEIGA, W. (2017b). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review* **85** 61-83.

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

GOLDSMITH, J., BOBB, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011). Penalized Functional Regression. *Journal of Computational and Graphical Statistics* **20** 830–851. https://doi.org/10.1198/jcgs.2010.10007

HALL, P. and HOROWITZ, J. L. (2007). Methodology and Convergence Rates for Functional Linear Regression. *The Annals of Statistics* **35** 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** 649–659.

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, 1st edition ed. Wiley.

JANG, J. H. (2021). Principal Component Analysis of Hybrid Functional and Vector Data. *Statistics in Medicine* **40** 5152–5173. https://doi.org/10.1002/sim.9117

KOLAR, M., LIU, H. and XING, E. P. (2014). Graph Estimation From Multi-Attribute Data. *Journal of Machine Learning Research* **15** 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** 814–840.

PREDA, C. and SAPORTA, G. (2005). PLS Regression on a Stochastic Process. *Computational Statistics & Data Analysis* **48** 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** 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** e3452.

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

TUCKER, R. S. (1938). The reasons for price rigidity. *The American Economic Review* **28** 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** 600–617.

ZHU, H., STRAWN, N. and DUNSON, D. B. (2016). Bayesian Graphical Models for Multivariate Functional Data. *Journal of Machine Learning Research* **17** 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

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

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

Furthermore, let $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_n]^\top \in \mathbb{R}^{n \times p}$ denote $n \times p$ matrix of the scalar predictors (observed version when $l = 1$, or residualized version when $l \geq 2$), and let $\widetilde{W}(\mathbf{t}) = [\widetilde{X}(\mathbf{t}), \mathbf{Z}]$ denote $n \times (K + p)$ hybrid predictor matrix that stacks $\widetilde{X}(\mathbf{t})$ and $\mathbf{Z}$ as columns. Then we can also write the hybrid predictor, whose functional part is approximated using the basis functions, as a form of linear transformation similar to (32):

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

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

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:

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

C.1.0.2. Regularization based on second-order derivatives.. Our proposed strategy for computing PLS component, detailed in Section 5.3.1, incorporates regularization based on second-order derivatives. To this end, let $\ddot{J}^{(k)}$ denote the $M \times M$ Gram matrix formed by the second derivatives of the basis functions, defined as

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

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

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

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 SECTION 3

D.1. Proof of Lemma 3.1 . JM: An alternative proof to the compactness of $\mathcal{U}$; rather than directly showing $\mathcal{U}$ is compact, we show $\mathcal{C}_{YW}$ is compact

PROOF. Any bounded linear functional from a Hilbert space to $\mathbb{R}$ is a compact operator. This is because the image of any bounded set under such a functional is a bounded set in $\mathbb{R}$, and by Bolzano-Weierstrass Theorem, every bounded sequence of real numbers has a convergent subsequence. Therefore, to show that $\mathcal{C}_{YW}$ is a compact operator, it suffices to show that $\mathcal{C}_{YW}$ is a bounded linear functional.

Linearity. The operator $\mathcal{C}_{YW} : \mathcal{H} \to \mathbb{R}$ is defined as $\mathcal{C}_{YW}h = \langle \Sigma_{YW}, h\rangle_{\mathcal{H}}$. For any $h_1, h_2 \in \mathcal{H}$ and scalar $c \in \mathbb{R}$, the linearity of the inner product implies:

$$\mathcal{C}_{YW}(h_1 + h_2) = \langle \Sigma_{YW}, h_1 + h_2\rangle_{\mathcal{H}} = \langle \Sigma_{YW}, h_1\rangle_{\mathcal{H}} + \langle \Sigma_{YW}, h_2\rangle_{\mathcal{H}} = \mathcal{C}_{YW}h_1 + \mathcal{C}_{YW}h_2,$$
$$\mathcal{C}_{YW}(ch) = \langle \Sigma_{YW}, ch\rangle_{\mathcal{H}} = c\langle \Sigma_{YW}, h\rangle_{\mathcal{H}} = c\mathcal{C}_{YW}h.$$

Thus, $\mathcal{C}_{YW}$ is a linear functional.

Boundedness. To show that $\mathcal{C}_{YW}$ is bounded, we need to find a finite constant M such that $|\mathcal{C}_{YW}h| \le M\|h\|_{\mathcal{H}}$ for all $h \in \mathcal{H}$. By the Cauchy-Schwarz inequality, we have:

$$|\mathcal{C}_{YW}h| = |\langle \Sigma_{YW}, h\rangle_{\mathcal{H}}| \le \|\Sigma_{YW}\|_{\mathcal{H}}\|h\|_{\mathcal{H}}.$$

Now, we must verify that $\|\Sigma_{YW}\|_{\mathcal{H}}$ is finite. Recall that

$$\Sigma_{YW} = (\sigma_{YX,1}(t), \ldots, \sigma_{YX,K}(t), \sigma_{YZ,1}, \ldots, \sigma_{YZ,p}).$$

Let $\mu([0,1])$ denote a Lebesgue measure of $[0,1]$, and let $T = \max_{k=1,\ldots,K} \mu([0,1])$. The norm of Σ_{YW} in $\mathcal{H}$ is given by:

$$\|\Sigma_{YW}\|_{\mathcal{H}}^2 = \sum_{k=1}^{K} \int_0^1 \sigma_{YX,k}(t)^2\, dt + \sum_{r=1}^{p} \sigma_{YZ,r}^2 < KTQ_1 + pQ_2 < \infty,$$

where the last inequality uses the condition (4). Let $M = \sqrt{KTQ_1 + pQ_2}$. Thus, we have shown that $|\mathcal{C}_{YW}h| \le M\|h\|_{\mathcal{H}}$ for a finite constant M. This completes the proof of Lemma 3.1.

□

D.2. Proof of Lemma 3.2.

PROOF. For $h \in \mathcal{H}$ and $d \in \mathbb{R}$, we have:

$$\langle \mathcal{C}_{YW}h, d\rangle = \mathbb{E}[\langle W_1, h\rangle_{\mathcal{H}} Y_1]\, d = \mathbb{E}\langle Y_1 W_1 d,\, h\rangle_{\mathcal{H}} = \langle h,\, \mathbb{E}[Y_1 W_1 d]\rangle_{\mathcal{H}} = \langle h, \mathcal{C}_{WY}d\rangle_{\mathcal{H}}$$

Thus, we have $\langle \mathcal{C}_{YW}h, d\rangle = \langle h, \mathcal{C}_{WY}d\rangle_{\mathcal{H}}$, which implies $\mathcal{C}_{WY} = \mathcal{C}_{YW}^*$. This completes the proof of Lemma 3.2. □

D.3. Proof of Lemma 3.3.

PROOF. We show that $\mathcal{U}$ is self-adjoint, positive-semidefinite, and compact, in turn.

Self-adjoint. For any $h_1, h_2 \in \mathcal{H}$, we have

$$\begin{aligned}\langle \mathcal{U}h_1,\, h_2\rangle_{\mathcal{H}} = \big\langle \langle h_1, \Sigma_{YW}\rangle_{\mathcal{H}}\Sigma_{YW},\, h_2\big\rangle_{\mathcal{H}} &= \langle h_1, \Sigma_{YW}\rangle_{\mathcal{H}}\langle \Sigma_{YW}, h_2\rangle_{\mathcal{H}} \\ &= \big\langle h_1,\, \langle \Sigma_{YW}, h_2\rangle_{\mathcal{H}}\Sigma_{YW}\big\rangle_{\mathcal{H}} \\ &= \langle h_1, \mathcal{U}h_2\rangle_{\mathcal{H}}.\end{aligned}$$

Positive-semidefinite. For every $h \in \mathcal{H}$, we have

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

Compact. By Lemma 3.1, $\mathcal{C}_{YW}$ is a compact operator. By Theorem 4.1.3 of Hsing and Eubank (2015), the composition of two operators is compact if either operator is compact. Therefore $\mathcal{U} := \mathcal{C}_{WY} \circ \mathcal{C}_{YW}$ is a compact operator. This completes the proof of Lemma 3.3. □

D.4. Proof of Theorem 3.4.

PROOF. by Theorem 4.3.1 of Hsing and Eubank (2015), since $\mathcal{C}_{YW}$ is compact by Lemma 3.1, it has the singular value decomposition of $\mathcal{C}_{YW}$, given by:

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

Let $\|\cdot\|_{\text{op}}$ denote an operator norm. By 4.3.4 in Hsing and Eubank (2015), we have

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

with maximum attained at $h = f_{11}$, 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. This completes the proof of Theorem 3.4. □

APPENDIX E: PROOF OF PROPOSITIONS 5.2 AND 5.3

PROOF. Let $\xi = (\xi_1, \ldots, \xi_K, \boldsymbol{\zeta}) \in \mathcal{H}$. 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:

$$\begin{aligned}\widehat{\text{Cov}}(\langle \widetilde{W}, \xi\rangle_{\mathcal{H}}, Y) &= \frac{1}{n}\sum_{i=1}^{n} y_i \langle \widetilde{W}_i, \xi\rangle_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i \left(\sum_{j=1}^{K} \langle \widetilde{X}_{ij}, \xi_j\rangle + \mathbf{Z}_i^\top \boldsymbol{\zeta}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i \left(\sum_{k=1}^{K} \int_0^1 \widetilde{X}_{ij}(t)\, \xi_j(t)\, dt\right) + \frac{1}{n}\sum_{i=1}^{n} y_i (\mathbf{Z}_i^\top \boldsymbol{\zeta})\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} y_i \left\{ \sum_{k=1}^{K} \int_0^1 \left(\sum_{m=1}^{M} \theta_{ijm} b_m(t) \right) \left(\sum_{m'=1}^{M} d_{jm'} b_{m'}(t) \right) dt \right\} + \frac{1}{n}\mathbf{y}^\top Z\boldsymbol{\zeta} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} \theta_{ijm} d_{jm'} \int_0^1 b_m(t) b_{m'}(t) dt \right\} + \frac{1}{n}\mathbf{y}^\top Z\boldsymbol{\zeta} \\
&= \sum_{k=1}^{K} \left\{ \frac{1}{n}\sum_{i=1}^{n} y_i (\boldsymbol{\theta}_{ij}^\top B\boldsymbol{\gamma}_j) \right\} + \frac{1}{n}\mathbf{y}^\top Z\boldsymbol{\zeta} \\
&= \frac{1}{n}\left(\sum_{k=1}^{K} \mathbf{y}^\top \Theta_j B\boldsymbol{\gamma}_j + \mathbf{y}^\top Z\boldsymbol{\zeta} \right)
\end{aligned}
$$

Building on this computation, the squared empirical covariance is expressed as the quadratic form involving the matrix V defined in (23):

$$
\begin{aligned}
\widehat{\mathrm{Cov}}^2(\langle \widetilde{W}, \xi \rangle_{\mathcal{H}}, Y) &= \frac{1}{n^2}\left(\sum_{k=1}^{K} \mathbf{y}^\top \Theta_j B\boldsymbol{\gamma}_j + \mathbf{y}^\top Z\boldsymbol{\zeta} \right)^2 \\
&= \frac{1}{n^2}(\mathbf{y}^\top \Theta \mathbb{B}\boldsymbol{\xi})^2 \\
&= \frac{1}{n^2}\boldsymbol{\xi}^\top (\mathbb{B}\Theta^\top \mathbf{y})(\mathbb{B}\Theta^\top \mathbf{y})^\top \boldsymbol{\xi} \\
&= \boldsymbol{\xi}^\top V \boldsymbol{\xi}. \qquad (37)
\end{aligned}
$$

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

$$
\begin{aligned}
\langle \xi, \xi \rangle_{\mathcal{H}} &= \sum_{j=1}^{K} \int_0^1 \xi_j(t)\, \xi_j(t)\, dt + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} \\
&= \sum_{j=1}^{K} \int_0^1 \left(\sum_{m=1}^{M} d_{jm} b_m(t) \right) \left(\sum_{m'=1}^{M} d_{jm'} b_{m'}(t) \right) dt + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} \\
&= \sum_{j=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} d_{jm} d_{jm'} \int_0^1 b_m(t) b_{m'}(t) dt + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} \\
&= \sum_{j=1}^{K} \boldsymbol{\gamma}_j^\top B\boldsymbol{\gamma}_j + \boldsymbol{\zeta}^\top \boldsymbol{\zeta}, \\
&= \boldsymbol{\xi}^\top \mathbb{B}\boldsymbol{\xi} \qquad (38)
\end{aligned}
$$

Therefore, the covariance maximization problem (21) is equivalent to the generalized Raleigh quotient (22). This completes the proof of Proposition 5.2.

APPENDIX F: PROOF OF PROPOSITION 5.3

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

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

$$
\begin{aligned}
&= \sum_{j=1}^{K} \lambda_j \int_0^1 \{\sum_{l=1}^{M} \theta_{ijl}\, b_l''(t)\}^2 dt \\
&= \sum_{j=1}^{K} \lambda_j \sum_{l=1}^{M}\sum_{m=1}^{M} \theta_{ijl}\theta_{ijm} \int_0^1 b_l''(t) b_m''(t) dt \\
&= \sum_{j=1}^{K} \lambda_j \boldsymbol{\gamma}_j^\top B'' \boldsymbol{\gamma}_j \\
&= \boldsymbol{\xi}^\top \Lambda \mathbb{B}'' \boldsymbol{\xi}, \qquad (39)
\end{aligned}
$$

where $\boldsymbol{\xi}$ is the concatenated coefficient vector form of ξ as defined in (**??**), $\mathbb{B}''$ is defined in (18), and Λ is defined in (26).

Combining this with (38), we can compute

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

This computation along with (37) implies that the covariance maximization problem (25) is equivalent to the generalized Raleigh quotient (27). This completes the proof of Proposition 5.3.

□

APPENDIX G: PROOF OF PROPOSITION 5.4

PROOF. The optimization problem (27) can be written as

$$
\begin{aligned}
&\max_{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_K \in \mathbb{R}^M,\, \boldsymbol{\zeta} \in \mathbb{R}^p} \frac{1}{n^2} \left(\sum_{j=1}^{K} \mathbf{y}^\top \Theta_j B \boldsymbol{\gamma}_j + \mathbf{y}^\top Z \boldsymbol{\zeta} \right)^2, \\
&\text{subject to} \quad \sum_{j=1}^{K} \boldsymbol{\gamma}_j^\top (B + \lambda_j B'') \boldsymbol{\gamma}_j + \boldsymbol{\zeta}^\top \boldsymbol{\zeta} = 1.
\end{aligned}
$$

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

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

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

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 5.1 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 G.

□

APPENDIX H: PROOF OF LEMMA 5.5

For notational simplicity, we omit the iteration superscript $[l]$. Recall from (28) that $\hat{\rho}_i = \langle \widetilde{W}_i, \hat{\xi} \rangle_{\mathcal{H}}$. Using the basis expansion coefficient notation from (12), (14), (15), and (17), the full vector of PLS scores $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_n)^\top$ is computed through the following matrix multiplication:

$$\hat{\boldsymbol{\rho}}^\top = \sum_{k=1}^{K} (\hat{\boldsymbol{\gamma}}_j)^\top B\, \Theta_j^\top + \boldsymbol{\zeta}^\top \mathbf{Z}^\top. \tag{42}$$

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

$$\mathrm{PENSSE}(\delta) := \sum_{i=1}^{n} \langle \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]}\,\delta,\, \widetilde{W}_i^{[l]} - \hat{\rho}_i^{[l]}\,\delta\rangle_{\mathcal{H}} + \tau \sum_{j=1}^{K} \mathrm{PEN}(\delta_j). \tag{43}$$

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

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 (**??**). 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:

$$\begin{aligned} B_1(\boldsymbol{\delta}) &= -2 \sum_{i=1}^{n} \hat{\rho}_i \langle \delta,\, \widetilde{W}_i\rangle_{\mathcal{H}} = -2\hat{\boldsymbol{\rho}}^\top \left(\sum_{j=1}^{K} \Theta_j B \boldsymbol{\pi}_j + \mathbf{Z}\boldsymbol{\chi} \right), \\ B_2(\boldsymbol{\delta}) &= (\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) \langle \delta,\, \delta\rangle_{\mathcal{H}} = (\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) \left(\sum_{j=1}^{K} \pi_j^\top B \pi_j + \boldsymbol{\chi}^\top \boldsymbol{\chi} \right), \\ B_3(\boldsymbol{\delta}) &= \tau \sum_{j=1}^{K} \mathrm{PEN}(\delta_j) = \tau \boldsymbol{\delta} B'' \boldsymbol{\delta}, \end{aligned}$$

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

$$\begin{aligned} &\nabla_{\boldsymbol{\pi}_j} B_1(\boldsymbol{\delta}) = -2B\Theta_j^\top \hat{\boldsymbol{\rho}},\ j = 1, \ldots, K, \quad \nabla_{\boldsymbol{\chi}} B_1(\boldsymbol{\delta}) = -2\mathbf{Z}^\top \hat{\boldsymbol{\rho}}, \\ &\nabla_{\boldsymbol{\pi}_j} B_2(\boldsymbol{\delta}) = 2(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) B \boldsymbol{\pi}_j,\ j = 1, \ldots, K, \quad \nabla_{\boldsymbol{\chi}} B_2(\boldsymbol{\delta}) = 2(\hat{\boldsymbol{\rho}}^\top \hat{\boldsymbol{\rho}}) \boldsymbol{\chi}, \\ &\nabla_{\boldsymbol{\pi}_j} B_3(\boldsymbol{\delta}) = 2\tau B'' \boldsymbol{\pi}_j,\ j = 1, \ldots, K, \quad \nabla_{\boldsymbol{\chi}} B_3(\boldsymbol{\delta}) = 0. \end{aligned}$$

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 5.1 that both B' and B'' are positive definite, $\mathrm{PENSSE}(\delta)$ is convex. Consequently, the gradient vanishes at its unique minimizer:

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

providing the following closed-form solution:

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

The solution to the optimization problem (29), denoted $\hat{\delta}^{[l]} \in \tilde{\mathcal{H}}$, has basis coefficients

$$\hat{\boldsymbol{\pi}}_j^{[l]} = \frac{1}{\|\hat{\boldsymbol{\rho}}^{[l]}\|_2^2} \Theta_j^{[l]\top} \hat{\boldsymbol{\rho}}^{[l]},\ j = 1, \ldots, K,$$

and scalar parts given by:

$$\hat{\boldsymbol{\chi}}^{[l]} = \frac{1}{\|\hat{\boldsymbol{\rho}}^{[l]}\|_2^2}\mathbf{Z}^{[l]\top}\hat{\boldsymbol{\rho}}^{[l]}.$$

This completes the proof of Lemma 5.5.

APPENDIX I: PROOF OF PROPOSITION 6.2

PROOF. The unit norm condition is trivially met by the constraint $1 = \hat{\boldsymbol{\xi}}_l^\top(\mathbb{B} + \Lambda\mathbb{B}'')\hat{\boldsymbol{\xi}}_l$ enforced in (27), because we have:

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_l^\top(\mathbb{B} + \Lambda\mathbb{B}'')\hat{\boldsymbol{\xi}}_l &= \hat{\boldsymbol{\xi}}_l^\top\mathbb{B}\hat{\boldsymbol{\xi}}_l + \hat{\boldsymbol{\xi}}_l^\top\Lambda\mathbb{B}''\hat{\boldsymbol{\xi}}_l \\
&\overset{(i)}{=} \sum_{j=1}^{K}\int_0^1 \hat{\xi}_l(t)\,\hat{\xi}_l(t)\,dt + \hat{\boldsymbol{\zeta}}^\top\hat{\boldsymbol{\zeta}} + \hat{\boldsymbol{\xi}}^\top\Lambda\mathbb{B}''\hat{\boldsymbol{\xi}} \\
&\overset{(ii)}{=} \sum_{j=1}^{K}\int_0^1 \hat{\xi}_l(t)\,\hat{\xi}_l(t)\,dt + \hat{\boldsymbol{\zeta}}^\top\hat{\boldsymbol{\zeta}} + \sum_{j=1}^{K}\lambda_l\int_0^1\{\hat{\xi}_l''(t)\}^2 dt \\
&= \langle\widehat{\xi}_l, \widehat{\xi}_l\rangle_{\mathcal{H},\Lambda}, \qquad (44)
\end{aligned}$$

where step (i) uses (38) and step (ii) uses (39).

Now to switch our gears toward the the orthogonality. For $l_1 > l_2$, we have

$$\begin{aligned}
\langle\widehat{\xi}_{l_1}, \widehat{\xi}_{l_2}\rangle_{\mathcal{H},\Lambda} &\overset{(i)}{=} \hat{\boldsymbol{\xi}}_{l_1}^\top(\mathbb{B} + \Lambda\mathbb{B}'')\hat{\boldsymbol{\xi}}_{l_2} \\
&\overset{(ii)}{=} \frac{1}{\kappa_{l_1}}(V^{[l_1]}\widehat{\boldsymbol{\xi}}_{l_1})^\top\hat{\boldsymbol{\xi}}_{l_2} \\
&= \frac{1}{\kappa_{l_1}}\widehat{\boldsymbol{\xi}}_{l_1}^\top V^{[l_1]}\hat{\boldsymbol{\xi}}_{l_2} \\
&\overset{(iii)}{=} \frac{1}{n^2\kappa_{l_1}}\widehat{\boldsymbol{\xi}}_{l_1}^\top(\mathbb{B}\Theta^{[l_1]\top}\mathbf{y})(\mathbb{B}\Theta^{[l_1]\top}\mathbf{y})^\top\hat{\boldsymbol{\xi}}_{l_2} \\
&= \frac{1}{n^2\kappa_{l_1}}\widehat{\boldsymbol{\xi}}_{l_1}^\top(\mathbb{B}\Theta^{[l_1]\top}\mathbf{y})\mathbf{y}^\top\Theta^{[l_1]}\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2} \\
&= c\mathbf{y}^\top(\Theta^{[l_1]}\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2}),
\end{aligned}$$

where step (i) uses (44), step (ii) uses the fact that $\hat{\boldsymbol{\xi}}_{l_1}$ is a generalized eigenvector (κ_{l_1} denotes the corresponding generalized eigenvalue), as presented in Proposition 5.3, and step (iii) uses the definition of V matrix given in (23). Here, c represents a scalar that condenses all multiplicative terms preceding $\mathbf{y}^\top$.

Orthogonality can be demonstrated by showing that $\Theta^{[l_1]}\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2} = \mathbf{0} \in \mathbb{R}^n$. From the construction in (16), (15) and (18), the ith entry of $\Theta^{[l_1]}\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2}$ is

$$\begin{aligned}
(\Theta^{[l_1]}\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2})_i &= (\theta_{i11}^{[l_1]}, \ldots, \theta_{i1M}^{[l_1]}, \ldots, \theta_{iK1}^{[l_1]}, \ldots, \theta_{iKM}^{[l_1]}, Z_{i1}^{[l_1]}, \ldots, Z_{ip}^{[l_1]})\mathbb{B}\hat{\boldsymbol{\xi}}_{l_2} \\
&= \sum_{k=1}^{K}\boldsymbol{\theta}_{ik}^{[l_1]}B\hat{\boldsymbol{\gamma}}_k^{[l_2]} + \mathbf{Z}_i^{[l_1]\top}\hat{\boldsymbol{\zeta}}^{[l_2]} \\
&= \langle\widetilde{W}_i^{[l_1]}, \hat{\xi}^{[l_2]}\rangle_{\mathcal{H}}, \qquad (45)
\end{aligned}$$

where the last equality uses the computation of the ith entry in (42). To expand the last quantity, we use a recursive relationship derived as follows. Recall from Lemma 5.5 that we have

$$\widetilde{W}_i^{[l_1]} := \widetilde{W}_i^{[l_1-1]} - \frac{\widehat{\rho}_{i\,l_1-1}}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2} \sum_{i=1}^{n} \widehat{\rho}_{i\,l_1-1} \widetilde{W}_i^{[l_1-1]}. \tag{46}$$

Let us denote, with some abuse of notation:

$$\widetilde{\boldsymbol{W}}^{[l_1]} := \big(\widetilde{W}_1^{[l_1-1]}, \ldots, \widetilde{W}_n^{[l_1-1]}\big)^\top \in \tilde{\mathcal{H}}^{\otimes n}, \tag{47}$$

and

$$\langle \widetilde{\boldsymbol{W}}^{[l_1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} := \left(\langle \widetilde{W}_1^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}}, \ldots, \langle \widetilde{W}_n^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} \right)^\top \in \mathbb{R}^n. \tag{48}$$

Recalling (45), to achieve our goal of showing $\Theta^{[l_1]} \mathbb{B} \hat{\boldsymbol{\xi}}_{l_2} = \mathbf{0} \in \mathbb{R}^n$, it suffices to show $\langle \widetilde{\boldsymbol{W}}^{[l_1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} = \mathbf{0}$. Using these notations and (46), we have, with some abuse of notation,

$$\widetilde{\boldsymbol{W}}^{[l_1]} = \widetilde{\boldsymbol{W}}^{[l_1-1]} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2} \hat{\boldsymbol{\rho}}^{[l_1-1]} \hat{\boldsymbol{\rho}}^{[l_1-1]\top} \right) \widetilde{\boldsymbol{W}}^{[l_1-1]}, \tag{49}$$

and thus

$$\langle \widetilde{\boldsymbol{W}}^{[l_1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} = \langle \widetilde{\boldsymbol{W}}^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2} \hat{\boldsymbol{\rho}}^{[l_1-1]} \hat{\boldsymbol{\rho}}^{[l_1-1]\top} \right) \langle \widetilde{\boldsymbol{W}}^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}}.$$

Clearly, the right-hand side represents a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^n$, which we denote as $P^{[l_1-1]}$. Thus, we have $\langle \widetilde{\boldsymbol{W}}^{[l_1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} = P^{[l_1-1]} \langle \widetilde{\boldsymbol{W}}^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}}$. Repeatedly using this relationship, we have

$$\begin{aligned}
\langle \widetilde{\boldsymbol{W}}^{[l_1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} &= P^{[l_1-1]} \langle \widetilde{\boldsymbol{W}}^{[l_1-1]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} \\
&= P^{[l_1-1]} P^{[l_1-2]} \langle \widetilde{\boldsymbol{W}}^{[l_1-2]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} \\
&= \underbrace{P^{[l_1-1]} P^{[l_1-2]} \cdots P^{[l_2+1]}}_{:=P} P^{[l_2]} \langle \widetilde{\boldsymbol{W}}^{[l_2]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} \\
&= P \left\{ \langle \widetilde{\boldsymbol{W}}^{[l_2]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_2]}\|_2^2} \hat{\boldsymbol{\rho}}^{[l_2]} \hat{\boldsymbol{\rho}}^{[l_2]\top} \right) \langle \widetilde{\boldsymbol{W}}^{[l_2]}, \hat{\xi}^{[l_2]} \rangle_{\mathcal{H}} \right\} \\
&= P \left\{ \hat{\boldsymbol{\rho}}^{[l_2]} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2} \hat{\boldsymbol{\rho}}^{[l_2]} \hat{\boldsymbol{\rho}}^{[l_2]\top} \right) \hat{\boldsymbol{\rho}}^{[l_2]} \right\} \\
&= P \left\{ \hat{\boldsymbol{\rho}}^{[l_2]} - \left(\frac{\hat{\boldsymbol{\rho}}^{[l_2]\top} \hat{\boldsymbol{\rho}}^{[l_2]}}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2} \right) \hat{\boldsymbol{\rho}}^{[l_2]} \right\} \\
&= P \left\{ \hat{\boldsymbol{\rho}}^{[l_2]} - \hat{\boldsymbol{\rho}}^{[l_2]} \right\} = P\mathbf{0} = \mathbf{0}.
\end{aligned}$$

This completes the proof of Proposition 6.2. □

APPENDIX J: PROOF OF PROPOSITION 6.3

PROOF. For $l_1 < l_2$, we have

$$\begin{aligned}\widehat{\boldsymbol{\rho}}^{[l_1]\top}\widehat{\boldsymbol{\rho}}^{[l_2]} &= \sum_{i=1}^{n}\widehat{\rho}_i^{[l_1]}\,\widehat{\rho}_i^{[l_2]}\\ &= \sum_{i=1}^{n}\widehat{\rho}_i^{[l_1]}\,\langle\widetilde{W}_i^{[l_2]},\hat{\xi}^{[l_2]}\rangle_{\mathcal{H}}\\ &= \langle\sum_{i=1}^{n}\widehat{\rho}_i^{[l_1]}\,\widetilde{W}_i^{[l_2]},\hat{\xi}^{[l_2]}\rangle_{\mathcal{H}}\end{aligned}$$

Therefore, to show that $\widehat{\boldsymbol{\rho}}^{[l_1]\top}\widehat{\boldsymbol{\rho}}^{[l_2]}=0$, it suffices to show that $\sum_{i=1}^{n}\widehat{\rho}_i^{[l_1]}\,\widetilde{W}_i^{[l_2]}=0\in\mathcal{H}$, where this zero element represents an ordered pair of K zero functions and a zero matrix. Using Lemma 5.5, notations from (47) and (48), and equation (49), we have:

$$\begin{aligned}\widetilde{\boldsymbol{W}}^{[l_1]} &= \widetilde{\boldsymbol{W}}^{[l_1-1]} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2}\hat{\boldsymbol{\rho}}^{[l_1-1]}\hat{\boldsymbol{\rho}}^{[l_1-1]\top}\right)\widetilde{\boldsymbol{W}}^{[l_1-1]}\\ &= \widetilde{\boldsymbol{W}}^{[l_1-1]} - \left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2}\langle\widetilde{\boldsymbol{W}}^{[l_1-1]},\hat{\xi}^{[l_1-1]}\rangle_{\mathcal{H}}\hat{\boldsymbol{\rho}}^{[l_1-1]\top}\right)\widetilde{\boldsymbol{W}}^{[l_1-1]}.\end{aligned}$$

Using this relationship and with some abuse of notation, we have:

$$\begin{aligned}\sum_{i=1}^{n}\widehat{\rho}_i^{[l_1]}\,\widetilde{W}_i^{[l_2]} &= \hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1]}\\ &= \hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]} - \hat{\boldsymbol{\rho}}^{[l_2]\top}\left(\frac{1}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2}\langle\widetilde{\boldsymbol{W}}^{[l_1-1]},\hat{\xi}^{[l_1-1]}\rangle_{\mathcal{H}}\hat{\boldsymbol{\rho}}^{[l_1-1]\top}\right)\widetilde{\boldsymbol{W}}^{[l_1-1]}\\ &= \hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]} - \frac{\langle\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]},\hat{\xi}^{[l_1-1]}\rangle_{\mathcal{H}}}{\|\hat{\boldsymbol{\rho}}^{[l_1-1]}\|_2^2}\left(\hat{\boldsymbol{\rho}}^{[l_1-1]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]}\right)\\ &= h^{[l_1-1]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]}\right)\text{ (say)}.\end{aligned}$$

Here, the function $h^{[l_1-1]}:\tilde{\mathcal{H}}\mapsto\tilde{\mathcal{H}}$ maps the zero element of $\tilde{\mathcal{H}}$ to itself, where the zero element represents an ordered pair of K zero functions and a zero matrix. Using this relationship, we have

$$\begin{aligned}\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1]} &= h^{[l_1-1]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-1]}\right)\\ &= h^{[l_1-1]}\circ h^{[l_1-2]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_1-2]}\right)\\ &= h^{[l_1-1]}\circ h^{[l_1-2]}\circ\ldots\circ h^{[l_2]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]}\right).\end{aligned}$$

Therefore, to show that $\widehat{\boldsymbol{\rho}}^{[l_1]\top}\widehat{\boldsymbol{\rho}}^{[l_2]}=0$, it suffices to show $h^{[l_2]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]}\right)=0$:

$$\begin{aligned}h^{[l_2]}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]}\right) &= \hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]} - \frac{\langle\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]},\hat{\xi}^{[l_2]}\rangle_{\mathcal{H}}}{\|\hat{\boldsymbol{\rho}}^{[l_2]}\|_2^2}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]}\right)\\ &= \hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]} - \frac{\hat{\boldsymbol{\rho}}^{[l_2]\top}\langle\widetilde{\boldsymbol{W}}^{[l_2]},\hat{\xi}^{[l_2]}\rangle_{\mathcal{H}}}{\|\hat{\boldsymbol{\rho}}^{[l_2]}\|_2^2}\left(\hat{\boldsymbol{\rho}}^{[l_2]\top}\widetilde{\boldsymbol{W}}^{[l_2]}\right)\end{aligned}$$

$$
\begin{aligned}
&= \hat{\boldsymbol{\rho}}^{[l_2]\top} \widetilde{\boldsymbol{W}}^{[l_2]} - \frac{\hat{\boldsymbol{\rho}}^{[l_2]\top} \hat{\boldsymbol{\rho}}^{[l_2]}}{\|\hat{\boldsymbol{\rho}}^{[l_2]}\|_2^2} (\hat{\boldsymbol{\rho}}^{[l_2]\top} \widetilde{\boldsymbol{W}}^{[l_2]}) \\
&= \hat{\boldsymbol{\rho}}^{[l_2]\top} \widetilde{\boldsymbol{W}}^{[l_2]} - \hat{\boldsymbol{\rho}}^{[l_2]\top} \widetilde{\boldsymbol{W}}^{[l_2]} \\
&= 0.
\end{aligned}
$$

This completes the proof of Proposition 6.3. □