Partial Least Squares Regression with Multiple Functional and Scalar Predictors

Jeong Hoon Jang

July 16, 2025

1 Data Objects and Model Formulation

Let $\{X^{(k)}\}_{k=1,\ldots,K}$ be a collection of random functions defined on respective compact domains $\mathcal{T}_k \in \mathbb{R}^{d_k}$ ($d_k \in \mathbb{N}$); i.e., $X^{(k)} : \mathcal{T}_k \to \mathbb{R}$. Assume that each $X^{(k)}$ is in $L^2(\mathcal{T}_k)$, a Hilbert space of square integrable functions with respect to Lebesgue measure dt_k on $\mathcal{T}_k$. Without loss of generality, we further assume that each $\mathcal{T}_k$ is a d_k-dimensional unit interval; for instance $\mathcal{T}_k = [0,1]$ (unit interval) if $d_k = 1$, $\mathcal{T}_k = [0,1] \times [0,1]$ (unit square) if $d_k = 2$, and so forth. Write $X = (X^{(1)}, \ldots, X^{(K)})$ as a multivariate functional object that belongs to $\mathcal{F} = L^2(\mathcal{T}_1) \times \cdots \times L^2(\mathcal{T}_K)$—a cartesian product of individual $L^2(\mathcal{T}_k)$ spaces. Note that if $K = 1$, X reduces to a univariate functional object. We can also express the functional object X evaluated on the multi-dimensional argument $\mathbf{t} = [t_1, \ldots t_k]^T \in \mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_K$ as a K-dimensional vector $X(\mathbf{t}) = [X^{(1)}(t_1), \ldots, X^{(K)}(t_K)]^T$. The inner product of $f_1 = (f_1^{(1)}, \ldots, f_1^{(K)})$ and $f_2 = (f_2^{(1)}, \ldots, f_2^{(K)})$ in $\mathcal{F}$ is defined as $\langle f_1, f_2 \rangle_{\mathcal{F}} = \sum_{k=1}^K \langle f_1^{(k)}, f_2^{(k)} \rangle_{L^2} = \sum_{k=1}^K \int_{\mathcal{T}_k} f_1^{(k)}(t_k) f_2^{(k)}(t_k) dt_k$ with norm $\|f_1\|_{\mathcal{F}} = \langle f_1, f_1 \rangle_{\mathcal{F}}^{1/2} = \{\sum_{k=1}^K \int_{\mathcal{T}_k} f_1^{(k)}(t_k)^2 dt_k\}^{1/2}$.

Let $\mathbf{Z} = [Z_1, \ldots, Z_p]^T$ denote a p-dimensional multivariate scalar data in $\mathbb{R}^p$. We assume

that $\mathbf{Z}$ is a random vector with finite first two moments and equipped with the Euclidean inner product and norm; i.e., for $\mathbf{v_1} = [v_{11}, \ldots, v_{1p}]^T$ and $\mathbf{v_2} = [v_{21}, \ldots, v_{2p}]^T$ in $\mathbb{R}^p$, $\langle \mathbf{v_1}, \mathbf{v_2} \rangle = \mathbf{v_1}^T \mathbf{v_2} = \sum_{r=1}^{p} v_{1r} v_{2r}$, and $\|\mathbf{v}_1\| = \langle \mathbf{v}_1, \mathbf{v}_1 \rangle^{1/2} = (\sum_{r=1}^{p} v_{1r}^2)^{1/2}$.

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

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

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

Our strategy is to formulate a hybrid random object, $\mathbf{W} = (X, \mathbf{Z})$, which combines X and $\mathbf{Z}$ into an ordered pair belonging to $\mathcal{H} = \mathcal{F} \times \mathbb{R}^p$. An alternative notation for the hybrid object can be obtained by evaluating its functional part on $\mathbf{t}$ and expressing it as a $(K+p)$-dimensional vector: $\mathbf{W}[\mathbf{t}] = [X(\mathbf{t}), \mathbf{Z}]^T$, with $X(\mathbf{t}) = [X^{(1)}(t_1), \ldots, X^{(K)}(t_K)]^T \in \mathbb{R}^K$ and $\mathbf{Z} = [Z_1, \ldots, Z_p]^T \in \mathbb{R}^p$. We define the inner product between any two hybrid objects, $\mathbf{h}_1 = (f_1, \mathbf{v}_1)$ and $\mathbf{h}_2 = (f_2, \mathbf{v}_2)$, as

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

with norm $\|\cdot\|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$. 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. Without loss of generality and for the clarity of illustration, all the following theoretical results will be derived for $\omega = 1$. The results remain valid for any positive weights. Then the functional regression model 1 can be

re-expressed in terms of the hybrid object as follows

$$Y = \langle \boldsymbol{\eta}, \mathbf{W} \rangle_{\mathcal{H}} + \epsilon, \tag{2}$$

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

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

$$\mathbf{W}[\mathbf{t}] = \sum_{l=1}^{\infty} \rho_l \boldsymbol{\delta}_l[\mathbf{t}] \quad \text{and} \quad Y = \sum_{l=1}^{\infty} \rho_l \nu_l + \epsilon, \tag{3}$$

where $(\boldsymbol{\delta}_l)_{l \in \mathbb{N}} \in \mathcal{H}$ and $(\nu_l)_{l \in \mathbb{N}} \in \mathbb{R}$ are appropriate bases. The PLS scores are obtained as $\rho_l = \langle \mathbf{W}^{[l]}, \boldsymbol{\xi}_l \rangle_{\mathcal{H}}$, where $\mathbf{W}^{[l]}$ denotes the residualized predictor sequentially derived as the residual of the regression of $\mathbf{W}^{[l-1]}$ on ρ_{l-1} with $\mathbf{W}^{[1]} = \mathbf{W}$, and $\boldsymbol{\xi}_l = (\psi_l, \boldsymbol{\theta}_l) \in \mathcal{H}$ is the PLS component sequentially chosen to maximize the squared covariance between ρ_l and the residualized outcome $Y^{[l]}$—i.e., $\text{Cov}^2(\rho_l, Y^{[l]})$. Here, $Y^{[l]}$ is also sequentially obtained as the residual of the regression of $Y^{[l-1]}$ on ρ_{l-1} with $Y^{[1]} = Y$.

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 (**????**); or (2) a

univariate functional predictor with other scalar predictors (**?**).

2 Naive NIPALS Algorithm

Denote $\{(Y_1, \mathbf{W}_1), \ldots, (Y_n, \mathbf{W}_n)\}$ as n independent data pairs (observed sample) distributed as $(Y, \mathbf{W})$. In this section, we present the extended version of the the conventional nonlinear interative partial least square (NIPALS) algorithm which decomposes the hybrid 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. We first introduce the centered data $Y_i^{[1]} = Y_i - \bar{Y}$ and $\mathbf{W}_i^{[1]} = (X_i^{[1]}, \mathbf{Z}_i^{[1]}) = (X_i - \bar{X}, \mathbf{Z}_i - \bar{\mathbf{Z}}) = \mathbf{W}_i - \bar{\mathbf{W}}$. Here and below, a superscript in square brackets denotes the iteration index of the algorithm. The algorithm iterates through the following steps. For $l = 1, 2, \ldots, L$:

Step 1. Let $\hat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \hat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} = \langle X_i^{[l]}, \hat{\psi}_l \rangle_{\mathcal{F}} + \langle \mathbf{Z}_i^{[l]}, \hat{\boldsymbol{\theta}}_l \rangle$, where the l-th PLS component $\hat{\boldsymbol{\xi}}_l = (\hat{\psi}_l, \hat{\boldsymbol{\theta}}_l) \in \mathcal{H}$ is chosen to maximize $\mathrm{Cov}^2(\rho_{il}, Y_i^{[l]})$. Specifically, we set:

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

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

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

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

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

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

The regression coefficient $\boldsymbol{\eta}$ in model (2) can be written in terms of the PLS scores. First, note that we can show $\hat{\rho}_{il} = \langle \mathbf{W}_i^{[l]}, \hat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} = \langle \mathbf{W}_i, \hat{\boldsymbol{\zeta}}_l \rangle_{\mathcal{H}}$, where $\hat{\boldsymbol{\zeta}}_l = \hat{\boldsymbol{\xi}}_l - \sum_{u=1}^{l-1} \langle \hat{\boldsymbol{\delta}}_u, \hat{\boldsymbol{\xi}}_l \rangle_{\mathcal{H}} \hat{\boldsymbol{\zeta}}_u$ for $l \geq 2$, with $\hat{\boldsymbol{\zeta}}_1 = \hat{\boldsymbol{\xi}}_1$. Then, the decomposition of Y_i in (3) leads to

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

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

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

Here, L can be chosen by cross-validation.

This NIPALS algorithm, however, is performed in a pointwise manner for each observed argument $\mathbf{t}$, so it 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. In the next section, we propose novel strategies for implementing the steps of the NIPALS algorithm. The proposed approach provides an efficient and robust means of producing PLS scores and components 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 gain efficiency, interpretability and predictive accuracy of the PLS.

3 Proposed NIPALS Algorithm

3.1 Step 1: Computing PLS components via Eigenanalysis of the cross-covariance based operator

To simplify notation, we omit the iteration index l, with the understanding that the strategy presented in this section applies to any l-th iteration of the algorithm. We first define the cross-covariance terms between the response and predictors (residualized versions if $l \geq 2$):

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

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

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

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

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

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

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

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

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

Proposition 1 *The operator* $\mathcal{U} = \mathcal{C}_{WY} \circ \mathcal{C}_{YW} = \Sigma_{YW} \otimes \Sigma_{YW} : \mathcal{H} \to \mathcal{H}$ *is positive and self-adjoint. Furthermore, if there exist finite constants* Q_1 *and* Q_2 *such that*

$$\max_{k=1,\ldots,K} \sup_{t_k \in \mathcal{T}_k} \sigma_{YX}^{(k)2}(t_k) < Q_1 \quad \text{and} \quad \max_{r=1,\ldots,p} \sigma_{YZ,r}^2 < Q_2, \tag{4}$$

and if each $\sigma_{YX}^{(k)}$ $(k = 1, \ldots, K)$ *is uniformly continuous in a sense that for any* $\epsilon > 0$*, there exists* $\delta_k > 0$ *such that for every* $t_k, t_k^* \in \mathcal{T}_k$ *with* $|t_k - t_k^*| < \delta_k$*, we have that*

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

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

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

The following proposition suggests a potentially alternative way of obtaining the PLS components through eigen-analysis of $\mathcal{U}$.

Proposition 2

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

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

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

We now propose a practical scheme for implementing the aforementioned strategy, which can ultimately replace Step 1 of the NIPALS Algorithm in Section 2. Our goal here is to

obtain the first eigenfunction $\hat{\boldsymbol{\xi}} = (\hat{\psi}, \hat{\boldsymbol{\theta}})$ of the operator $\hat{\mathcal{U}} = \hat{\Sigma}_{YW} \otimes \hat{\Sigma}_{YW}$—an estimated version of $\mathcal{U}$—by solving the following eigen-equation:

$$\hat{\mathcal{U}}\boldsymbol{\xi} = \hat{\lambda}\boldsymbol{\xi} \quad \Longleftrightarrow \quad (\hat{\mathcal{U}}\boldsymbol{\xi})[\mathbf{t}] = \hat{\lambda}\boldsymbol{\xi}[\mathbf{t}], \;\; \mathbf{t} \in \mathcal{T}, \tag{5}$$

where $\boldsymbol{\xi}[\mathbf{t}] = [\psi(\mathbf{t})^T, \boldsymbol{\theta}^T]^T \in \mathbb{R}^{K+p}$, with $\psi(\mathbf{t}) = [\psi^{(1)}(t_1), \ldots, \psi^{(K)}(t_K)]^T \in \mathbb{R}^K$ and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T \in \mathbb{R}^p$. To achieve this goal, we propose to expand each observed function in terms of M basis functions:

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

where $\mathbf{c}_i^{(k)} = [c_{i1}^{(k)}, \ldots, c_{iM}^{(k)}]^T \in \mathbb{R}^M$ are basis coefficients corresponding to basis functions $\mathbf{b}^{(k)}(t) = [b_1^{(k)}(t_k), \ldots, b_M^{(k)}(t_k)]^T \in \mathbb{R}^M$. This basis expansion can be expressed collectively across n observations as $\mathbf{X}^{(k)}(t_k) = C^{(k)}\mathbf{b}^{(k)}(t_k)$, where $\mathbf{X}^{(k)}(t_k) = [X_1^{(k)}(t_k), \ldots, X_n^{(k)}(t_k)]^T \in \mathbb{R}^n$, and $C^{(k)} = [\mathbf{c}_1^{(k)} \cdots \mathbf{c}_n^{(k)}]^T \in \mathbb{R}^{n \times M}$. This notation can be further extended to accommodate K functional predictors as $X(\mathbf{t}) = CB(\mathbf{t})$, where $X(\mathbf{t}) = [\mathbf{X}^{(1)}(t_1) \cdots \mathbf{X}^{(K)}(t_K)] \in \mathbb{R}^{n \times K}$, $C = [C^{(1)} \cdots C^{(K)}] \in \mathbb{R}^{n \times MK}$, and $B(\mathbf{t}) = blkdiag[\mathbf{b}^{(1)}(t_1), \ldots, \mathbf{b}^{(K)}(t_K)] \in \mathbb{R}^{MK \times K}$. We can also expand the functional part of the PLS component in terms of the same basis functions: $\psi^{(k)}(t_k) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t_k) = \mathbf{a}^{(k)T}\mathbf{b}^{(k)}(t_k)$ and $\psi(\mathbf{t}) = B(\mathbf{t})^T\mathbf{a}$, where $\mathbf{a}^{(k)} = [a_1^{(k)}, \ldots, a_M^{(k)}]^T \in \mathbb{R}^M$, and $\mathbf{a} = [\mathbf{a}^{(1)T}, \ldots, \mathbf{a}^{(K)T}]^T \in \mathbb{R}^{MK}$.

Now let $Z = [\mathbf{Z}_1 \cdots \mathbf{Z}_n] \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = [Y_1, \ldots, Y_n] \in \mathbb{R}^n$ respectively denote the scalar predictors and outcome data. Also $J^{(k)}$ denote a $M \times M$ matrix with (m, n)-th entry equal to $\int_{\mathcal{T}_k} b_m^{(k)}(t) b_n^{(k)}(t) dt$, and set $J = blkdiag[J^{(1)}, \ldots, J^{(K)}] \in \mathbb{R}^{MK \times MK}$. The following proposition presents a practical and efficient approach to obtaining the first eigenfunction $(\hat{\boldsymbol{\xi}})$ of $\hat{\mathcal{U}}$.

Proposition 3 *Given the basis function expansions* $X_i^{(k)}(t) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t)$ *and* $\psi^{(k)}(t) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t)$, $k = 1, \ldots, K$, *solving for the first eigenfunction* $\hat{\boldsymbol{\xi}} = (\hat{\psi}, \hat{\boldsymbol{\theta}}) \in \mathcal{H}$ *in the eigen-equation* (5) *amounts to solving for the first eigenvector* $\tilde{\boldsymbol{\xi}} = [\tilde{\mathbf{u}}^T, \tilde{\boldsymbol{\theta}}^T]^T \in \mathbb{R}^{MK+p}$ *in the following eigen-equation*

$$V\tilde{\boldsymbol{\xi}} = \lambda\tilde{\boldsymbol{\xi}}$$

$$\iff \begin{bmatrix} n^{-2} J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} & n^{-2} J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T Z \\ n^{-2} Z^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} & n^{-2} Z^T \mathbf{y}\mathbf{y}^T Z \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}} \\ \tilde{\boldsymbol{\theta}} \end{bmatrix} = \lambda \begin{bmatrix} \tilde{\mathbf{u}} \\ \tilde{\boldsymbol{\theta}} \end{bmatrix},$$

and then setting $\hat{\boldsymbol{\xi}}[\mathbf{t}] = [\hat{\psi}(\mathbf{t})^T, \hat{\boldsymbol{\theta}}^T]^T \in \mathbb{R}^{K+p}$ *with* $\hat{\psi}(\mathbf{t}) = B(\mathbf{t}) J^{-\frac{1}{2}} \tilde{\mathbf{u}}$ *and* $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$.

Applying this proposition to the first step of the NIPALS algorithm, at each l-th iteration, we can obtain the PLS component $\hat{\boldsymbol{\xi}}_l$ by computing and transforming the first eigenvector of V, whose elements consist of or are derived from the residualized response and predictors: $\mathbf{y}^{[l]}$, $X^{[l]}(\mathbf{t})$ and $Z^{[l]}$.

3.2 Extended Step 1: Regularization of PLS components

Although Proposition 3 provides an efficient and principled way to compute the PLS components $\boldsymbol{\xi}$, their functional part $\psi = (\psi^{(1)}, \ldots, \psi^{(K)})$ may not be smooth, which can lead to overfitting and unstable predictions. Thus, in this section, we introduce an extension of the idea presented in Section 3.1 that yields regularized PLS components. A widely-used approach to regularizing or smoothing a function $\psi^{(k)}$ is to penalize its roughness quantified by the integrated squared second derivative:

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

where the last term is obtained using the basis expansion $\psi^{(k)}(t) = \mathbf{a}^{(k)T}\mathbf{b}^{(k)}(t)$, and $\ddot{J}^{(k)}$ being a $M \times M$ matrix whose (m,n)-th entry is equal to $\int_{\mathcal{T}_k} \ddot{b}_m^{(k)}(t)\ddot{b}_n^{(k)}(t)dt$. The penalty term that simultaneously penalizes the roughness of the K functions in ψ can be obtained by summing up the individual penalty terms as

$$\text{PEN}(\psi) = \sum_{k=1}^{K} \text{PEN}\left(\psi^{(k)}\right) = \sum_{k=1}^{K} \int_{\mathcal{T}_k} \ddot{\psi}^{(k)}(t_k)^2 dt_k = \sum_{k=1}^{K} \mathbf{a}^{(k)T}\ddot{J}^{(k)}\mathbf{a}^{(k)} = \mathbf{a}^T\ddot{J}\mathbf{a},$$

where $\ddot{J} = blkdiag[\ddot{J}^{(1)}, \ldots, \ddot{J}^{(K)}] \in \mathbb{R}^{MK \times MK}$.

Now, maximizing $\text{Cov}^2\left(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y\right)$ is not the only aim; we also want to prevent the roughness $\text{PEN}(\psi)$ of the functional part of the estimated PLS component from being too large. The following proposition provides a generalized Rayleigh quotient which makes explicit this possible conflict and whose maximizer corresponds to a smoothed PLS component that maximizes the squared sample covariance $\widehat{\text{Cov}}^2\left(\langle \mathbf{W}, \boldsymbol{\xi} \rangle_{\mathcal{H}}, Y\right)$ over the class of all functions satisfying sufficient smoothness conditions.

Proposition 4 *Consider the following basis function expansions:* $X_i^{(k)}(t_k) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t_k)$ *and* $\psi^{(k)}(t_k) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t_k)$, $k = 1, \ldots, K$, *Let* $J^* = blkdiag[J, I_p] \in \mathbb{R}^{(MK+p)\times(MK+p)}$, $\ddot{J}^* = blkdiag[\ddot{J}, \mathbf{0}_{p\times p}] \in \mathbb{R}^{(MK+p)\times(MK+p)}$, *and*

$$V^* = \begin{bmatrix} n^{-2}JC^T\mathbf{y}\mathbf{y}^TCJ & n^{-2}JC^T\mathbf{y}\mathbf{y}^TZ \\ n^{-2}Z^T\mathbf{y}\mathbf{y}^TCJ & n^{-2}Z^T\mathbf{y}\mathbf{y}^TZ \end{bmatrix} \in \mathbb{R}^{(MK+p)\times(MK+p)}.$$

Then at each l-th iteration of the first step of the NIPALS algorithm, involving the residualized response and predictors $Y \equiv Y^{[l]}$ and $\mathbf{W} \equiv \mathbf{W}^{[l]}$, the smoothed PLS component can be obtained by solving

$$\max_{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}} \frac{\boldsymbol{\xi}^{*T}V^*\boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*T}J^*\boldsymbol{\xi}^* + \kappa\text{PEN}(\psi)} = \max_{\boldsymbol{\xi}^* \in \mathbb{R}^{MK+p}} \frac{\boldsymbol{\xi}^{*T}V^*\boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*T}(J^* + \kappa\ddot{J}^*)\boldsymbol{\xi}^*}, \tag{6}$$

with respect to $\boldsymbol{\xi}^* = [\mathbf{a}^{*T}, \boldsymbol{\theta}^{*T}]^T \in \mathbb{R}^{MK+p}$*, and then setting* $\hat{\boldsymbol{\xi}}[\mathbf{t}] = [\hat{\psi}(\mathbf{t})^T, \hat{\boldsymbol{\theta}}^T]^T = [\mathbf{a}^{*T}B(\mathbf{t}), \boldsymbol{\theta}^{*T}]^T$*, where* κ *is a positive smoothing parameter that controls the trade-off between the goodness of fit and the amount of smoothness in the functional part* ψ*.*

Smaller κ produces PLS components that fit more closely to the observed data, but selecting too small value may cause overfitting. The estimated PLS component reverts back to the unregularized version of Proposition 3 as $\kappa \to 0$. On the other hand, larger κ places more emphasis on smoothing ψ. In the limit $\kappa \to \infty$, each of the functions is forced to be of the form $\psi^{(k)}(t) = a + bt$ for some constants a and b. In practice, κ can be chosen by cross-validation.

To solve (6) in practice, we first perform a Choleski factorization: $LL^T = J^* + \kappa \ddot{J}^*$, where L is a lower triangular matrix, and then reduce (6) to maximizing the classical Rayleigh quotient

$$\max_{\mathbf{e} \in \mathbb{R}^{MK+p}} \frac{\mathbf{e}^T E \mathbf{e}}{\mathbf{e}^T \mathbf{e}} \iff \max_{\substack{\mathbf{e} \in \mathbb{R}^{MK+p} \\ \|\mathbf{e}\|^2 = 1}} \mathbf{e}^T E \mathbf{e},$$

through the transformations $\mathbf{e} = L^T \boldsymbol{\xi}^*$ and $E = L^{-1} V^* L^{T-1}$. Once $\mathbf{e}$ is obtained as the first eigenvector of E, we set $\boldsymbol{\xi}^* = L^{T-1}\mathbf{e}$ and $\hat{\boldsymbol{\xi}}[\mathbf{t}] = [\mathbf{a}^{*T}B(\mathbf{t}), \boldsymbol{\theta}^{*T}]^T$, which we subsequently renormalize to ensure that $\|\hat{\boldsymbol{\xi}}\|_{\mathcal{H}} = 1$.

3.3 Step 2: Fitting regression model with hybrid response and scalar predictor to compute residualized outcomes

The second step of the NIPALS algorithm is to iteratively obtain the $(l+1)$-th residualized outcome as $\mathbf{W}_i^{[l+1]}[\mathbf{t}] = \mathbf{W}_i^{[l]}[\mathbf{t}] - \hat{\boldsymbol{\delta}}_l[\mathbf{t}]\hat{\rho}_{il}$, where $\hat{\boldsymbol{\delta}}_l[\mathbf{t}]$ is the least squares estimate from regressing the previous (l-th) residualized outcome $\mathbf{W}_i^{[l]}[\mathbf{t}]$ on the corresponding PLS score $\hat{\rho}_{il}$. The pointwise least squares estimate, $\hat{\boldsymbol{\delta}}_l[\mathbf{t}] = (\sum_{i=1}^n \mathbf{W}_i^{[l]}[\mathbf{t}]\hat{\rho}_{il})/(\sum_{i=1}^n \hat{\rho}_{il}^2)$, presented in Section 2, however, is highly inefficient and may not be even feasible for multiple dense and/or

irregular functional data. Moreover, the pointwise least squares estimator can show substantial variability across the domain, which becomes problematic for the entire algorithm as its instability passes on to subsequent residualized outcomes and PLS components. Therefore, in this section, we introduce a novel framework for fitting a regression with a hybrid response and scalar predictor. The framework is readily applicable to dense and irregular functional data, and supports regularization techniques to prevent overfitting and reduce variance.

Again, we omit l in the notations, and let $W[\mathbf{t}]$ be an $n \times (K+p)$ matrix of our hybrid observations; that is,

$$W[\mathbf{t}] = \begin{bmatrix} W_1[\mathbf{t}]^T \\ \vdots \\ W_n[\mathbf{t}]^T \end{bmatrix} = \begin{bmatrix} X_1(\mathbf{t})^T & \mathbf{Z}_1^T \\ \vdots & \vdots \\ X_n(\mathbf{t})^T & \mathbf{Z}_n^T \end{bmatrix} = \begin{bmatrix} X_1^{(1)}(t_1) & \cdots & X_1^{(K)}(t_K) & Z_{11} & \cdots & Z_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ X_n^{(1)}(t_1) & \cdots & X_n^{(K)}(t_K) & Z_{n1} & \cdots & Z_{np} \end{bmatrix}$$

Let $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]^T$ be an n-dimensional vector of PLS scores, and let the $(K+P)$-vector $\boldsymbol{\delta}[\mathbf{t}] = [\pi(\mathbf{t})^T, \boldsymbol{\gamma}^T]^T$ denote the regression coefficient consisting of the functional part $\pi(\mathbf{t}) = [\pi^{(1)}(t_1), \ldots, \pi^{(K)}(t_K)]^T \in \mathbb{R}^K$ and scalar part $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^T \in \mathbb{R}^p$. The hybrid-on-scalar regression model we are interested in fitting is then:

$$W[\mathbf{t}] = \boldsymbol{\rho}\, \boldsymbol{\delta}[\mathbf{t}]^T + \mathbf{e}[\mathbf{t}], \tag{7}$$

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

As in Section 3.1, each of the functions in the hybrid response can be expanded using basis functions as $X(\mathbf{t}) = CB(\mathbf{t})$. Then, augmenting the basis function matrix $B(\mathbf{t})$ with the $p-$dimensional vector of ones $\mathbf{1}_p$ as $B^*[\mathbf{t}] = blkdiag[B(\mathbf{t}), I_p] \in \mathbb{R}^{(MK+p)\times(K+p)}$, and combining the corresponding basis coefficients with the scalar predictors as $C^* = [C^{(1)} \cdots C^{(K)}\ \mathbf{Z}_1 \cdots \mathbf{Z}_p] \in \mathbb{R}^{n\times(MK+p)}$, we can write the hybrid response as: $W[\mathbf{t}] = C^* B^*[\mathbf{t}]$.

Similarly, we expand each functional regression coefficient using the basis functions as $\pi^{(k)}(t_k) = \sum_{m=1}^{M} d_m^{(k)} b_m^{(k)}(t_k) = \mathbf{d}^{(k)T}\mathbf{b}^{(k)}(t_k)$, where $\mathbf{d}^{(k)} = [d_1^{(k)}, \ldots, d_M^{(k)}]^T \in \mathbb{R}^M$ are basis coefficients. Here, a different basis system can be flexibly chosen to expand the regression coefficient, but without loss of generality we use the same system. With the $(MK + p)$-dimensional row vector $\mathbf{d}^* = [\mathbf{d}^{(1)T}, \ldots, \mathbf{d}^{(K)T}, \gamma_1, \ldots, \gamma_p]$ that stacks the basis coefficients and scalar regression coefficients, we can express the linear predictor of the regression as: $\boldsymbol{\rho}\,\boldsymbol{\delta}[\mathbf{t}]^T = \boldsymbol{\rho}\,\mathbf{d}^* B^*[\mathbf{t}]$. Finally, these basis expansions of the hybrid response and the regression coefficient enable us to re-write the hybrid-on-scalar regression model (7) in a finite-dimensional form:

$$C^* B^*[\mathbf{t}] = \boldsymbol{\rho}\,\mathbf{d}^* B^*[\mathbf{t}] + \mathbf{e}[\mathbf{t}]. \tag{8}$$

As done in Section 3.1, we introduce a roughness penalty based on the integrated squared second derivative to regularize the functional part of the regression coefficient. Specifically, we define the following penalty term that simultaneously penalizes the roughness of the K functional regression coefficients:

$$\mathrm{PEN}(\pi) = \sum_{k=1}^{K} \mathrm{PEN}\left(\pi^{(k)}\right) = \sum_{k=1}^{K} \int_{\mathcal{T}_k} \ddot{\pi}^{(k)}(t_k)^2 dt_k = \sum_{k=1}^{K} \mathbf{d}^{(k)T} \ddot{J}^{(k)} \mathbf{d}^{(k)} = \mathbf{d}^* \ddot{J}^* \mathbf{d}^{*T}.$$

In our framework, we do not attempt to regularize the functional part of the hybrid response and allow it to be of high resolution, so as to prevent smoothing away important information in the response data that may impact the regression coefficient estimate.

Based on the model formulation (8) and the penalty term $\mathrm{PEN}(\pi)$, we can now formulate the following penalized least squares criterion whose minimizing solution gives the regularized

estimate of the regression coefficient $\boldsymbol{\delta}[\mathbf{t}]$ in the hybrid-on-scalar regression model (2):

$$\begin{aligned}\text{PENSSE}(\delta) &= \int_{\mathcal{T}} \| C^* B^*[\mathbf{t}] - \boldsymbol{\rho}\, \mathbf{d}^* B^*[\mathbf{t}] \|_F^2 \, d\mathbf{t} + \tau \mathbf{d}^* \ddot{J}^* \mathbf{d}^{*T} \\ &= tr\left(C^{*T} C^* J^*\right) - 2\, tr\left(\mathbf{d}^* J^* C^{*T} \boldsymbol{\rho}\right) + tr\left(\boldsymbol{\rho}^T \boldsymbol{\rho}\, \mathbf{d}^* J^* \mathbf{d}^{*T}\right) + \tau\, tr\left(\mathbf{d}^* \ddot{J}^* \mathbf{d}^{*T}\right), \end{aligned} \tag{9}$$

where tr denotes the trace of a matrix, $\| \cdot \|_F$ is a Frobenius norm, and τ is a positive smoothing parameter that controls the trade-off between the goodness of fit and the amount of smoothness in the functional regression coefficient π. Now taking the derivative of (9) with respect to $\mathbf{d}^*$ and setting the result to zero, we find that $\mathbf{d}^*$ satisfies the system of linear equations

$$\boldsymbol{\rho}^T \boldsymbol{\rho}\, \mathbf{d}^* J^* + \tau \mathbf{d}^* \ddot{J}^* = \boldsymbol{\rho}^T C^* J^*,$$

which leads to the following solutions for $\mathbf{d}^*$ and $\boldsymbol{\delta}[\mathbf{t}]$:

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

The smoothing parameter τ may be chosen by cross-validation.

Appendix: Proofs of Propositions

· Abbreviations

- "CSI": Cauchy–Schwarz inequality

· **Proposition 1**

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

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

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

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

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

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

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

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

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

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

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

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

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

fore, $\mathcal{U}$ is a compact operator.

· **Proposition 2**

Before we prove Proposition 2, we state and prove the following lemma:

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

Proof.

First, we have:

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

Also, we have:

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

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

Now we prove Proposition 2. The singular value decomposition of $\mathcal{C}_{YW}$ is given by

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

Let $\|\cdot\|_{op}$ denote an operator norm. Then, applying Theorem 4.3.4 in **?**, we can show that

$$\|\mathcal{C}_{YW}\|_{op} = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} |\mathcal{C}_{YW}\mathbf{h}|^2 = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} |E(\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}} Y)|^2 = \sup_{\substack{\mathbf{h}\in\mathcal{H} \\ \|\mathbf{h}\|_{\mathcal{H}}=1}} \mathrm{Cov}^2(\langle \mathbf{W}, \mathbf{h}\rangle_{\mathcal{H}}, Y) = \iota_1^2,$$

with maximum attained at $\mathbf{h} = f_{11}$, which is an eigenfunction of $\mathcal{C}_{YW}^*\mathcal{C}_{YW} = \mathcal{C}_{WY} \circ \mathcal{C}_{YW} = \mathcal{U}$ corresponding to the largest eigenvalue ι_1^2.

· **Proposition 3**

Based on outcome data $\mathbf{y}$ and the basis expansions of the functional predictors X, the consistent estimator for σ_{YX} can be written as

$$\hat{\sigma}_{YX}^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i^{(k)}(t)Y_i = \frac{1}{n}\mathbf{y}^T C^{(k)}\mathbf{b}^{(k)}(t),\ t \in \mathcal{T}_k \quad \text{or equivalently,}$$
$$\hat{\sigma}_{YX}(\mathbf{t}) = \frac{1}{n}X(\mathbf{t})^T\mathbf{y} = \frac{1}{n}B(\mathbf{t})^T C^T\mathbf{y},\ \mathbf{t} \in \mathcal{T}.$$

We can also obtain the consistent estimator of σ_{YZ} using the scalar predictors Z and outcome data $\mathbf{y}$ as

$$\hat{\sigma}_{YZ} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z_i}Y_i = \frac{1}{n}Z^T\mathbf{y}.$$

Now consider the left hand side of the eigen-equation (5). Based on the basis expansions of ψ (functional part of $\boldsymbol{\xi}$) and the consistent estimators of σ_{YX} and σ_{YZ}, we have:

$$\begin{aligned}
(\hat{\mathcal{U}}\boldsymbol{\xi})[\mathbf{t}] &= \langle \hat{\Sigma}_{YW}, \boldsymbol{\xi}\rangle_{\mathcal{H}}\, \hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \{\langle \hat{\sigma}_{YX}, \psi\rangle_{\mathcal{F}} + \langle \hat{\sigma}_{YZ}, \boldsymbol{\theta}\rangle\}\hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \left\{\sum_{k=1}^{K}\int_{\mathcal{T}_k} \hat{\sigma}_{YX}^{(k)}(t_k)\psi^{(k)}(t_k)dt_k + \hat{\sigma}_{YZ}^T\boldsymbol{\theta}\right\}\hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \left[\sum_{k=1}^{K}\int_{\mathcal{T}_k}\left\{\frac{1}{n}\mathbf{y}^T C^{(k)}\mathbf{b}^{(k)}(t_k)\right\}\left\{\mathbf{b}^{(k)}(t_k)^T\mathbf{a}^{(k)}\right\}dt_k + \frac{1}{n}\mathbf{y}^T Z\boldsymbol{\theta}\right]\hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \frac{1}{n}\left[\sum_{k=1}^{K}\mathbf{y}^T C^{(k)}\left\{\int_{\mathcal{T}_k}\mathbf{b}^{(k)}(t_k)\mathbf{b}^{(k)}(t_k)^T dt_k\right\}\mathbf{a}^{(k)} + \mathbf{y}^T Z\boldsymbol{\theta}\right]\hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \frac{1}{n}\mathbf{y}^T\left\{\sum_{k=1}^{K} C^{(k)}J^{(k)}\mathbf{a}^{(k)} + Z\boldsymbol{\theta}\right\}\hat{\Sigma}_{YW}[\mathbf{t}] \\
&= \frac{1}{n}\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})\begin{bmatrix}\hat{\sigma}_{YX}(\mathbf{t}) \\ \hat{\sigma}_{YZ}\end{bmatrix} \\
&= \frac{1}{n}\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})\begin{bmatrix}n^{-1}B(\mathbf{t})^T C^T\mathbf{y} \\ n^{-1}Z^T\mathbf{y}\end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})B(\mathbf{t})^T C^T\mathbf{y} \\ n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})Z^T\mathbf{y} \end{bmatrix}$$

Also the write hand side of the eigen-equation (5) can be written as

$$\lambda\boldsymbol{\xi}[\mathbf{t}] = \lambda \begin{bmatrix} \psi(\mathbf{t}) \\ \boldsymbol{\theta} \end{bmatrix} = \lambda \begin{bmatrix} B(\mathbf{t})^T\mathbf{a} \\ \boldsymbol{\theta} \end{bmatrix}.$$

Thus, solving the eigen-equation (5) amounts to solving the following system of equations:

$$\begin{cases} n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})B(\mathbf{t})^T C^T\mathbf{y} = \lambda B(\mathbf{t})^T\mathbf{a} \\ n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})Z^T\mathbf{y} = \lambda\boldsymbol{\theta}. \end{cases}$$

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

$$\begin{cases} n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})C^T\mathbf{y} = \lambda\mathbf{a} \\ n^{-2}\,\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta})Z^T\mathbf{y} = \lambda\boldsymbol{\theta}. \end{cases}$$

This leads to:

$$\begin{cases} n^{-2}\,C^T\mathbf{y}\mathbf{y}^T CJ\mathbf{a} + n^{-2}\,C^T\mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda\mathbf{a} \\ n^{-2}\,Z^T\mathbf{y}\mathbf{y}^T CJ\mathbf{a} + n^{-2}\,Z^T\mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda\boldsymbol{\theta}. \end{cases}$$

Now define $\mathbf{u} = J^{\frac{1}{2}}\mathbf{a}$. Then we have:

$$\begin{cases} n^{-2}\,C^T\mathbf{y}\mathbf{y}^T CJJ^{-\frac{1}{2}}\mathbf{u} + n^{-2}\,C^T\mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda J^{-\frac{1}{2}}\mathbf{u} \\ n^{-2}\,Z^T\mathbf{y}\mathbf{y}^T CJJ^{-\frac{1}{2}}\mathbf{u} + n^{-2}\,Z^T\mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda\boldsymbol{\theta}, \end{cases}$$

which implies

$$
\begin{aligned}
&\begin{cases} n^{-2}\, J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} \mathbf{u} + n^{-2}\, J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda \mathbf{u} \\ n^{-2}\, Z^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} \mathbf{u} + n^{-2}\, Z^T \mathbf{y}\mathbf{y}^T Z\boldsymbol{\theta} = \lambda \boldsymbol{\theta}, \end{cases} \\
&\Longrightarrow \begin{bmatrix} n^{-2}\, J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} & n^{-2}\, J^{\frac{1}{2}} C^T \mathbf{y}\mathbf{y}^T Z \\ n^{-2}\, Z^T \mathbf{y}\mathbf{y}^T C J^{\frac{1}{2}} & n^{-2}\, Z^T \mathbf{y}\mathbf{y}^T Z \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\theta} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\theta} \end{bmatrix} \\
&\Longrightarrow V\tilde{\boldsymbol{\xi}} = \lambda \tilde{\boldsymbol{\xi}}.
\end{aligned}
$$

Therefore, to obtain the first eigenfunction $\hat{\boldsymbol{\xi}} = (\hat{\psi}, \boldsymbol{\theta})$ of $\hat{\mathcal{U}}$, we find the first eigenvector $\tilde{\boldsymbol{\xi}} = [\mathbf{u}^T, \boldsymbol{\theta}^T]^T$ of the $(MK + p) \times (MK + p)$ matrix V, and then set $\psi(\mathbf{t}) = B(\mathbf{t}) J^{-\frac{1}{2}} \mathbf{u}$.

· **Proposition 4**

Firstly, we convert the maximization problem of Proposition 2 to a problem of maximizing a generalized Rayleigh quotient. Specifically, based on the basis function expansions $X_i^{(k)}(t) = \sum_{m=1}^{M} c_{im}^{(k)} b_m^{(k)}(t)$ and $\psi^{(k)}(t) = \sum_{m=1}^{M} a_m^{(k)} b_m^{(k)}(t)$, $k = 1, \ldots, K$, we have:

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

$$= n^{-1}\mathbf{y}^T(CJ\mathbf{a} + Z\boldsymbol{\theta}).$$

This implies that

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

Thus, Proposition 2 translates into solving

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

or equivalently solving

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

where $J^* = blkdiag[J, I_p] \in \mathbb{R}^{MK+p}$. In other words, the PLS component is chosen to maximize the squared sample covariance $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y) = \boldsymbol{\xi}^{*T} V^* \boldsymbol{\xi}^*$ subject to the constraint $\boldsymbol{\xi}^{*T} J^* \boldsymbol{\xi}^* = 1$. Similar in spirit to the regularized principal component analysis (see equation [9.1] of **?**), one way to penalize the squared sample covariance $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y)$ is to divide it by $\{1+\kappa \mathrm{PEN}(\psi)\}$, where $\mathrm{PEN}(\psi)$ is a roughness penalty term. This leads to the penalized squared sample covariance

$$\frac{\boldsymbol{\xi}^{*T} V^* \boldsymbol{\xi}^*}{\boldsymbol{\xi}^{*T} J^* \boldsymbol{\xi}^* + \kappa \mathrm{PEN}(\psi)},$$

whose maximizer corresponds to a smoothed PLS component that maximizes $\widehat{\mathrm{Cov}}^2(\langle \mathbf{W}, \boldsymbol{\xi}\rangle_{\mathcal{H}}, Y)$

over the class of all functions satisfying sufficient smoothness conditions.