Biometrics 000, 000–000 DOI: 000
000 0000

Hybrid Partial Least Squares Regression with Multiple Functional and Scalar Predictors

Summary: Motivated by renal imaging data with accompanying side information, we propose a hybrid partial least squares (PLS) regression method to capture cross-modality correlations in settings with both functional and scalar predictors. Our approach extends the nonlinear iterative PLS (NIPALS) algorithm to a hybrid Hilbert space, embedding functional and scalar predictors into a unified vector representation and iteratively maximizing their empirical cross-covariance with the response. The resulting low-dimensional representations effectively capture within- and between-modality variation and predictor-response correlation. Out method preserves the desirable theoretical properties of PLS. Moreover, the procedure is computationally efficient, requiring only the solution of linear systems at each iteration.
Key words: Dimension reduction; Functional data analysis; Multiple data modalities; Multivariate data analysis; Multivariate functional data; Partial least square

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Biomedical studies often collect multi-modal data. Our running example is the Emory University renal study (Chang et al., 2020; Jang, 2021), which records two renogram curves (functional data) along with side information (Euclidean vector) for each kidney. Such distinct yet related physiological signals can be represented as functional- and scalar-valued i.i.d. covariates in a linear regression model:

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

where Y_i is a scalar response, $X_{i1}(t), \ldots, X_{iK}(t) \in \mathbb{L}^2[0,1]$ are functional predictors, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ is a Euclidean vector covariate, and ϵ_i denotes observational noise. For notational convenience, we assume throughout that the responses and predictors are centered, so the intercept term can be omitted. The key challenge here is the strong correlations between functional and scalar predictors, in addition to the ill-posedness of the infinite-dimensional slope functions $\beta_k(t)$ and the high dimensionality of the predictors $(K + p)$. This paper addresses all three issues in a unified manner by introducing a hybrid partial least squares (PLS) regression framework defined on a novel hybrid Hilbert space.

Previous works and limitations. Basis expansion with regularization of roughness addresses the ill-posedness of functional data by exploiting their smoothness (Cardot et al., 2003; Aguilera et al., 2016; Cai and Hall, 2006; Zhao et al., 2012). High dimensionality can be addressed by using derived inputs, with PCA regression as a straightforward example. We can apply PCA separately to functional (e.g., Happ and Greven 2018) and scalar predictors, followed by multivariate regression on the combined scores. While well studied for functional predictors alone (Hall and Horowitz, 2007; Reiss and Ogden, 2007; Febrero-Bande et al., 2017), PCA regression may fail to capture the core regression relationship since the derived inputs are not guided by the response. Partial least squares (PLS) regression is a powerful alternative that constructs orthogonal latent components from predictors to

maximize covariance with the response. For functional predictors, Preda and Saporta (2005) introduced PLS regression for a single predictor, with subsequent extensions incorporating basis approximations and roughness penalties (Reiss and Ogden, 2007; Aguilera et al., 2010, 2016), and multiple functional predictors (Beyaztas and Lin Shang, 2022). Theoretical and computational advances were developed by Delaigle and Hall (2012) and Saricam et al. (2022), respectively; see Febrero-Bande et al. (2017) for a review. These approaches ignore correlations between functional and scalar components, leading to multicollinearity and unreliable prediction. Multimodal correlations have mainly been explored in unsupervised settings, such as graphical models, precision matrix estimation, and joint PCA (Kolar et al., 2014; Geng et al., 2020; Jang, 2021). However, these methods are response-agnostic and may miss outcome-relevant correlations.

Our contributions. To address these gaps, we propose a hybrid PLS regression framework that unifies basis expansion, PLS, and functional-scalar correlations. Treating functional and Euclidean components as a single vector in a Hilbert space with a suitable inner product, we iteratively extract joint directions that maximize covariance with the response under a unit-norm constraint. The framework handles dense or irregular functional data, supports regularization to reduce overfitting, and comes with theoretical guarantees.

2. Hybrid Hilbert Space

Our framework leverages that the nonlinear iterative PLS (NIPALS; Wold, 1975) algorithm, a standard PLS fitting method, operates solely via inner-product space operations. We define a Hilbert space for hybrid predictors with a complete inner product to extend the algorithm.

THEOREM 1 (Hybrid space): *An element* $h \in \mathbb{H} := (\mathbb{L}^2[0,1])^K \times \mathbb{R}^p$ *is an ordered tuple* $h = (f_1, \ldots, f_K, \mathbf{u})$*, where* $f_k \in \mathbb{L}^2[0,1]$ *for* $k = 1, \ldots, K$ *and* $\mathbf{u} \in \mathbb{R}^p$*. We define an inner*

product on $\mathbb{H}$ *for any two elements* $h_1 = (f_1, \ldots, f_K, \mathbf{u})$ *and* $h_2 = (g_1, \ldots, g_K, \mathbf{v})$ *as:*

$$\langle h_1, h_2 \rangle_{\mathbb{H}} := \sum_{k=1}^{K} \int_0^1 f_k(t) g_k(t)\, dt + \mathbf{u}^\top \mathbf{v}. \tag{2}$$

The inner product induces a norm $\|\cdot\|_{\mathbb{H}}$ *on the space, defined as* $\|h\|_{\mathbb{H}} := \langle h, h \rangle_{\mathbb{H}}^{1/2}$*, and a corresponding metric* $d(h_1, h_2) = \|h_1 - h_2\|_{\mathbb{H}}$*. Then the hybrid inner product space* $\mathbb{H}$ *is a separable Hilbert space.*

Definition 1 (Hybrid Predictor): For the Hilbert space $\mathbb{H}$ defined in Definition 1, The Borel σ-field on $\mathbb{H}$, denoted $\mathfrak{B}(\mathbb{H})$, is the smallest σ-field containing the class $\mathfrak{M}$ of all sets of the form $\{q \in \mathbb{H} \mid \langle q, h \rangle \in O\}$, for any $h \in \mathbb{H}$ and any open subset $O \subseteq \mathbb{R}$ (details can be found in Theorem 7.1.1 of Hsing and Eubank 2015). A hybrid predictor $W_i = (X_{i1}(t), \ldots, X_{iK}(t), \mathbf{Z}_i)$ is a measurable mapping from a probability space $(\Omega, \mathfrak{F}, P)$ into $\big(\mathbb{H}, \mathfrak{B}(\mathbb{H})\big)$.

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

$$Y_i = \langle \beta, W_i \rangle_{\mathbb{H}} + \epsilon_i, \text{ where } \beta := \big(\beta_1(t), \ldots, \beta_K(t), \boldsymbol{\beta}\big) \in \mathbb{H}. \tag{3}$$

3. Proposed PLS Algorithm

Our approach efficiently computes PLS components and scores for multiple dense or irregular functional predictors alongside scalar predictors. It incorporates regularization to exploit structural relationships within and between the functions, to prevent overfitting and improve interpretability. Each iteration involves two subroutines: regularized component estimation (Section **??**) and residualization (Section **??**) After iteration terminates the hybrid regression coefficient is estimated (Section **??**).

3.1 *Preliminary step 1: finite-basis approximation*

Let $\{b_m(t)\}$ be a twice-differentiable basis of $\mathbb{L}^2([0,1])$ whose second derivatives are also linearly independent, for example, cubic B-splines, Fourier, or orthonormal cubic polynomi-

als. Fourier coefficients of functional predictors $X_{ij}(t)$, functional regression coefficients $\beta_j(t)$, PLS directions, $\xi_j(t)$, and residualization coefficients $\delta_j(t)$ (with iteration indices suppressed) with respect to $\{b_m(t)\}$ are denoted as $\{\theta_{ijm}\}, \{\eta_{jm}\}, \{\gamma_{jm}\}$, and $\{\pi_{jm}\}$. For practicality, we truncate the expansion at a moderate M (e.g., 15-20) to capture most functional variation, while smoothness is handled via penalization (Section **??**). The truncated expansion for the functional data is denoted as are denoted as $\widetilde{X}_{ij}(t) := \sum_{m=1}^{M} \theta_{ijm} b_m(t)$, but for the other funtions, we denote

$$\beta_j(t) := \sum_{m=1}^{M} \eta_{jm} b_m(t), \xi_j(t) := \sum_{m=1}^{M} \gamma_{jm} b_m(t), \delta_j(t) = \sum_{m=1}^{M} \pi_{jm} b_m(t), \tag{4}$$

since these functions are estimated, using a hat over the tilde notation would be cumbersome. Let Let $\boldsymbol{\theta}_{ij}$, $\boldsymbol{\eta}_j$, $\boldsymbol{\gamma}_j$, and $\boldsymbol{\pi}_j$ denote these coefficients as M-dimensional vectors. The truncations imply that all computations in this paper are carried out entirely within the subspace

$$\widetilde{\mathbb{H}} := \operatorname{span}\big(b_1(t), \ldots, b_M(t)\big)^K \times \mathbb{R}^p \subset \mathbb{H}. \tag{5}$$

The ith hybrid predictor, projected on $\widetilde{\mathbb{H}}$, is represented by the tuple $\widetilde{W}_i := (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{iK}, \boldsymbol{Z}_i)$. For notational convenience, we stack the predictor coefficient vectors as

$$\Theta_j := (\boldsymbol{\theta}_{1j}, \ldots, \boldsymbol{\theta}_{nj})^\top \in \mathbb{R}^{n \times M}, \quad \Theta := (\Theta_1, \ldots, \Theta_K, \mathbf{Z}) \in \mathbb{R}^{n \times (MK+p)}. \tag{6}$$

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

$$B_{m,m'} := \int_0^1 b_m(t)\, b_{m'}(t)\, dt, \quad B''_{m,m'} := \int_0^1 b''_m(t)\, b''_{m'}(t)\, dt, \tag{7}$$

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

$$\mathbb{B} := \operatorname{blkdiag}(B, \ldots, B, I_p), \quad \mathbb{B}'' := \operatorname{blkdiag}(B'', \ldots, B'', I_p), \tag{8}$$

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

$$(\mathbb{B}, \mathbb{B}'', \Theta, \mathbf{y}) \in \mathbb{R}^{(MK+p)\times(MK+p)} \times \mathbb{R}^{(MK+p)\times(MK+p)} \times \mathbb{R}^{n\times(MK+p)} \times \mathbb{R}^n, \tag{9}$$

with the index l omitted for brevity.

REMARK 1: Different bases could be used for each functional predictor to handle multiple dense or irregular functional predictors. We adopt a common basis for simplicity. The definitions of $\mathbb{B}$ and $\mathbb{B}''$ remain general enough to accommodate distinct bases if needed.

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.

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

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. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review* **85,** 61–83.

Geng, S., Kolar, M., and Koyejo, O. (2020). Joint Nonparametric Precision Matrix Estimation with Confounding. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 378–388. PMLR.

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*. Wiley, 1st edition edition.

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

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

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.

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.

Wold, HERMAN. (1975). Path models with latent variables: The NIPALS approach*. In Blalock, H. M., Aganbegian, A., Borodkin, F. M., Boudon, R., and Capecchi, V., editors, *Quantitative Sociology*, International Perspectives on Mathematical and Statistical Modeling, pages 307–357. Academic Press.

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.