Functional Partial Least-Squares: Adaptive Estimation and Inference*

Andrii Babii
Department of Economics, UNC-Chapel Hill
and
Marine Carrasco
Department of Economics, University of Montreal
and
Idriss Tsafack
Department of Economics, University of Montreal

August 11, 2025

Abstract

We study the linear regression model with a scalar response and a functional predictor, a canonical example of an ill-posed inverse problem. We show that the functional partial least-squares (PLS) estimator achieves convergence rates that are nearly minimax-optimal over a class of ellipsoids and propose an adaptive early-stopping procedure for selecting the number of PLS components. In addition, we develop a new test that detects parametric local alternatives. The test can be inverted to construct confidence sets for the functional slope parameter. Simulation results show that the estimator performs favorably relative to several existing methods, and that the proposed test has good power. We apply our methodology to evaluate the nonlinear effects of temperature on corn and soybean yields.

Keywords: Functional partial least-squares, inference, rate-optimal and adaptive estimation, functional linear regression, climate science.

*The authors gratefully acknowledge the editor, associate editor, and referees for their helpful and constructive comments. An earlier version of this paper circulated under the title "Theoretical Comparison of Functional Principal Component Analysis and Functional Partial Least Squares." The authors thank participants at the SCSE Conference in Quebec (2019), the EC^2 Conference in Paris (2020), the NBER-NSF Time Series Conference at Rice University (2021), the ESIF Economics and AI+ML Meeting in Ithaca (2024), the Triangle Econometrics Conference (2024), and Econometrics in Rio (2024) for their valuable feedback. We are especially grateful to Ryan Borhani for outstanding research assistance. Marine Carrasco acknowledges partial financial support from NSERC and FQRSC. The authors report there are no competing interests to declare.

1 Introduction

With the increasing availability of data, functional data analysis is now widely used across fields such as chemometrics, climate science, and economics. In this paper, we study a linear functional regression model with a scalar response Y and functional predictor X:

$$Y = \int_0^1 \beta(s)X(s)\mathrm{d}s + \varepsilon, \qquad \mathbf{E}[\varepsilon X] = 0. \tag{1}$$

The primary objective is to estimate the functional slope β to predict the response Y. Both X and β lie in an infinite-dimensional Hilbert space, and the high dimensionality of β leads to an ill-posed inverse problem. Consistent estimation of the slope coefficient thus requires either a dimension-reduction technique or some form of regularization.

Two main approaches have been popular in the functional data analysis literature: penalization methods, e.g., Cardot et al. (2003), Li and Hsing (2007), Crambes et al. (2009), Yuan and Cai (2010), Cai and Yuan (2012), Florens and Van Bellegem (2015), and functional principal component analysis (PCA), e.g., Cardot et al. (1999), Ramsay and Silverman (2002), Cai and Hall (2006), Hall and Horowitz (2007). The PCA-based approach approximates β using a finite expansion over the leading principal components, which correspond to the eigenfunctions of the covariance operator. As noted by Jolliffe (1982), this method performs well when the response Y is primarily correlated with the leading principal components. Moreover, accurate recovery of the principal components requires sufficient separation between the eigenvalues of the covariance operator.

In this paper, we explore an alternative method, functional partial least-squares (PLS). PLS is widely used in statistical learning, see Friedman et al. (2009), Frank and Friedman (1993); economics and finance, see Carrasco and Rossi (2016) and Kelly and Pruitt (2015); chemometrics, see Helland (1988), Wold et al. (1984). However, it has been used less commonly in empirical functional data analysis. The method constructs components as linear

transformations of the predictors X designed to maximize covariance with the response Y. As a result, fewer components are typically needed to achieve good predictive performance compared to PCA. Functional PLS was introduced by Preda and Saporta (2005) to address high-dimensional multicollinearity issues in the scalar-on-function linear model. Several interesting papers compare PCA and various functional PLS estimators using data-driven choices of the number of latent components in simulations; see Reiss and Ogden (2007), Krämer et al. (2008), Baíllo (2009), Aguilera et al. (2010), Febrero-Bande et al. (2017), and Saricam et al. (2022). This literature concludes that the predictive ability of functional PCA and PLS approaches is similar, but functional PLS requires fewer components and provides more accurate estimates of the slope parameter than PCA.

However, the existing literature lacks supporting theory, which is challenging to obtain because the PLS estimator depends nonlinearly on the response variable and is computed iteratively. An important step toward the theoretical analysis of functional PLS was made by Delaigle and Hall (2012), who introduced an alternative PLS (APLS). Nevertheless, it remains unknown whether (1) functional PLS achieves nearly minimax-optimal convergence rates under weak conditions; (2) a rate-adaptive method to select the number of PLS components exists; and (3) inference for functional PLS can be conducted. Deriving an approximation to the distribution of the PLS estimator is particularly challenging because PLS is a nonlinear estimator converging at a nonparametric rate. Our paper fills these gaps in the literature and provides a comprehensive, rigorous theoretical analysis of functional PLS.

The paper makes several original contributions. First, we derive convergence rates for estimation and prediction errors under a source condition. This condition measures the complexity of the problem through the so-called degree of ill-posedness and relates

β to the spectral decomposition of the covariance operator. Our results do not require assuming a separation between adjacent eigenvalues, as is common for PCA (see, e.g., Hall and Horowitz (2007)), and hold even in the presence of repeated eigenvalues. Second, we establish a lower bound on the minimax convergence rate and show that our estimator is nearly minimax-optimal. Because the optimal number of PLS components depends on the unknown degree of ill-posedness, we propose an adaptive early-stopping rule for selecting the number of components. We show that this single early-stopping rule yields a rate-optimal PLS estimator for *both* estimation and prediction errors with high probability. We also characterize how the selected number of components evolves with sample size under various scenarios.

Lastly, we develop a new test and accompanying confidence sets for PLS. We show that the test can detect alternatives converging at a parametric rate. Interestingly, while early stopping is crucial for estimation and prediction to prevent overfitting, it should be avoided for inference. An efficient iterative algorithm to compute the estimator is provided. Our simulation results reveal that the estimator outperforms several alternative methods combined with cross-validation in terms of estimation error and remains competitive in prediction. We also find that our test has excellent power with samples of as few as 100 observations.

To establish our theoretical results, we rely on the inverse problem literature and exploit the close connection between PLS and the conjugate-gradient method, as presented in Hanke (1995), Engl et al. (1996), and Blanchard and Krämer (2016). To the best of our knowledge, this connection has not previously been established in the functional data analysis literature. Recently and independently of our work, Gupta et al. (2023) proposed an estimator in a reproducing kernel Hilbert space (RKHS) generated by a specified kernel

and used the conjugate-gradient method to regularize the solution. Although their method is similar, the results are not directly comparable to ours. Their main result concerns the convergence rate of an estimator in an RKHS, established under different assumptions. In particular, they assume a polynomial decay rate of certain operator eigenvalues, while we do not require any decay assumptions; see also Gupta et al. (2025), who discuss optimal convergence rates with linear spectral regularization schemes under similar assumptions. Similarly, Lin and Cevher (2021) studied conjugate-gradient methods in the context of linear approximations to nonparametric regression with randomized sketches and Nyström sampling. While their results apply to general Hilbert spaces under stronger assumptions, they do not explicitly connect to the functional data analysis literature; for example, their simulations focus on linear approximations of $\mathbf{E}[Y|X] = f(X)$ with real-valued data $Y, X \in \mathbb{R}$ and Sobolev RKHS kernels. Neither paper establishes rate adaptivity or provides a practical early-stopping rule for selecting the number of PLS components.

To illustrate the practical relevance of our results, we apply our method to climate science. Using a fine-grained county-level dataset of U.S. crop yields and temperatures recorded over 70 years, we estimate the impact of temperature on crop yields. We find that the critical temperature at which annual crop yields begin to decay is around 30°C, consistent with Schlenker and Roberts (2009), who relied on highly parameterized least-squares estimators. Our method provides additional insights by examining how the temperature-effect curves have evolved over time. Interestingly, we find that the detrimental effects of high temperatures on corn and soybean yields have diminished over time, likely reflecting farmers' adaptive actions, including the use of more resilient crops and improved irrigation systems. However, this finding is not conclusive once statistical uncertainty is accounted for, given the small number of observed extreme temperatures.

The rest of the paper is organized as follows. Section 2 introduces the functional regression model and the functional PLS estimator. Section 3 establishes the theoretical properties, including the convergence rates for estimation and prediction errors, the minimax lower bound, and the adaptivity of the early-stopping rule. Section 4 develops inference for PLS. Section 5 presents a Monte Carlo study. Section 6 discusses an empirical application to nonlinear temperature effects in agriculture. Section 7 concludes the paper. The Supplementary Material provides all proofs, comparisons to alternative estimators, and additional simulation results. A replication package is also provided.

2 Functional Regression and PLS Estimator

2.1 Functional Linear Regression

Throughout the paper, we consider a generalized version of the functional linear regression model

$$Y = \langle \beta, X \rangle + \varepsilon, \qquad \mathbf{E}[\varepsilon X] = 0,$$

where $(Y, X) \in \mathbb{R} \times \mathbb{H}$, $\beta \in \mathbb{H}$ is the unknown functional slope coefficient, and $(\mathbb{H}, \langle ., . \rangle)$ is a separable Hilbert space with the induced norm $\|.\| = \sqrt{\langle ., . \rangle}$. The model in Equation (1) corresponds to the Hilbert space of square-integrable functions, $\mathbb{H} = L_2[0, 1]$, with the norm induced by the inner product $\langle f, g \rangle = \int_0^1 f(s)g(s)\mathrm{d}s$.

If $\mathbf{E}[X] = 0$, the covariance restriction $\mathbf{E}[\varepsilon X] = 0$ implies that the slope coefficient $\beta \in \mathbb{H}$ solves the moment condition

$$r := \mathbf{E}[YX] = \mathbf{E}[(X \otimes X)\beta] =: K\beta, \tag{2}$$

where $r \in \mathbb{H}$ and $K : \mathbb{H} \to \mathbb{H}$ is a compact covariance operator with summable eigenvalues whenever $\mathbf{E}\|X\|^2 < \infty$. It is well known that the inverse operator K^{-1} is discontinuous

and solving the equation $K\beta = r$ for β is an ill-posed inverse problem; see Carrasco et al. (2007), Engl et al. (1996), and Hanke (1995).

Roughly speaking, there are two popular strategies to regularize such problems:

(a) replace K^{-1} with a continuous operator $R_\alpha(K)$ for some function $R_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\lim_{\alpha \to 0^+} R_\alpha(\lambda) = \lambda^{-1}$.

(b) solve the problem in a finite-dimensional subspace $\mathbb{H}_m \subset \mathbb{H}$, spanned by some fixed basis vectors $h_1, h_2, \ldots, h_m \in \mathbb{H}$.

Examples of (a) include Tikhonov regularization when $R_\alpha(\lambda) = (\alpha + \lambda)^{-1}$, the spectral cutoff when $R_\alpha(\lambda) = \lambda^{-1}\mathbf{1}_{\lambda \geq \alpha}$ and Landweber iterations. On the other hand, the estimators in group (b) often solve the empirical least-squares problem

$$\min_{b \in \mathbb{H}_m} \|\mathbf{y} - T_n b\|_n^2, \tag{3}$$

where $\|v\|_n^2 = v^\top v/n, v \in \mathbb{R}^n$ and we denote $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ and

$$T_n : \mathbb{H} \to \mathbb{R}^n$$

$$b \mapsto (\langle X_1, b\rangle, \ldots, \langle X_n, b\rangle)^\top$$

for an i.i.d. sample $(Y_i, X_i)_{i=1}^n$. The basis $(h_j)_{j=1}^m$ spanning $\mathbb{H}_m$ can be either fixed (e.g. Fourier, polynomials, splines, wavelets) or adaptively constructed from the data.

The data-driven bases are especially attractive since they can adapt to the features of the population represented by the data and can approximate the slope parameter $\beta \in \mathbb{H}$ more efficiently; see Delaigle and Hall (2012). The principal component analysis (PCA)[1] and the partial least-squares (PLS) are two widely used methods to construct adaptive bases in practice. The PCA basis is constructed by identifying the directions in $\mathbb{H}$ where X varies

[1]Using the PCA basis is also related to the spectral cutoff method described in (a).

the most while the PLS basis is constructed in a supervised way taking into account the response variable as well. While the first m elements of the PCA basis $h_1, \ldots, h_m$ usually capture most of the variation of X, these are not necessarily the most important vectors for approximating β or predicting the response variable Y. It is easy to find empirical examples where some of the last few low-variance components *are* important; see Jolliffe (1982) who documented this issue using datasets from economics, climate science, chemical engineering, and meteorology.

2.2 PLS Estimator

The PLS estimator constructs a data-driven basis iteratively maximizing the covariance with the response variable Y; see Preda and Saporta (2005) who introduced it in the functional data analysis setting. The iterative nature of the estimator makes it difficult to analyze its statistical properties. This prompted Delaigle and Hall (2012) to formulate an alternative functional PLS that solves the problem in Equation (3) over the so-called Krylov subspace

$$\mathbb{H}_m = \text{span}\left\{\hat{r}, \hat{K}\hat{r}, \hat{K}^2\hat{r}, \ldots, \hat{K}^{m-1}\hat{r}\right\},$$

where

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \qquad \text{and} \qquad \hat{K} = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$$

are the estimators of r and K; see also Wold et al. (1984), Helland (1988), and Phatak and de Hoog (2002) for the link between PLS and Krylov subspaces.

In this paper, we study a version of the PLS estimator with $m \geq 1$ components, denoted $\hat{\beta}_m$, characterized as a solution to the least-squares problem

$$\min_{b \in \mathbb{H}_m} \|T_n^*(\mathbf{y} - T_n b)\|^2$$

over the Krylov subspace $\mathbb{H}_m$. The least-squares objective function is weighted by the adjoint operator of T_n

$$T_n^* : \mathbb{R}^n \to \mathbb{H}$$

$$\phi = (\phi_1, \ldots, \phi_n)^\top \mapsto \frac{1}{n} \sum_{i=1}^{n} X_i \phi_i$$

and corresponds to minimizing the normal equations (the first-order conditions to the problem in Equation (3)). Equivalently, $\hat{\beta}_m$ fits the empirical counterpart to the equation (2)

$$\min_{b \in \mathbb{H}_m} \left\| \hat{r} - \hat{K} b \right\|^2 \tag{4}$$

since $\hat{r} = T_n^* \mathbf{y}$ and $\hat{K} = T_n^* T_n$. Importantly, the PLS estimator formalized in Equation (4) corresponds to the conjugate-gradient method applied to the self-adjoint operator $\hat{K}$, cf. Hestenes and Stiefel (1952), known for its excellent regularization properties; see also Hanke (1995) and Nemirovski (1986).[2] We provide a more detailed comparison between the two PLS estimators in the Supplementary Material, Section S.1. A related formulation of the PLS in reproducing kernel Hilbert spaces (RKHS) was recently studied in an independent work of Gupta et al. (2023) who focus on the estimation error only and impose assumptions different from ours. Our work can be seen as using a kernel naturally adapted to the data which is unknown in practice.

The estimator is uniquely defined for every $m \leq n_*$, where n_* is the number of distinct nonzero eigenvalues of $\hat{K}$; see Proposition S.1 in the Supplementary Material. It is also easy to see that for every $m \geq 1$, we have $\hat{\beta}_m = \hat{P}_m(\hat{K})\hat{r}$ for a polynomial $\hat{P}_m(\hat{K}) = \sum_{j=1}^{m} a_j \hat{K}^{j-1}$ with coefficients $\mathbf{a} := (a_1, \ldots, a_m)^\top$, where $\hat{P}_0 = 0$ and $\hat{\beta}_0 = 0$. The coefficient vector solves the system of m linear equations

$$\mathbf{K}\mathbf{a} = \mathbf{r},$$

[2]The conjugate-gradient method is one of the most efficient algorithms for solving high-dimensional systems of linear equations; see also (Nocedal and Wright, 1999, Chapter 5) and references therein.

where $\mathbf{K} := \langle \hat{K}^j \hat{r}, \hat{K}^k \hat{r} \rangle_{1 \le j,k \le m}$ and $\mathbf{r} := \langle \hat{K}^j \hat{r}, \hat{r} \rangle_{1 \le j \le m}$. From a practical point of view, it is more efficient to use an iterative conjugate-gradient algorithm that bypasses the (potentially unstable) matrix inversion with an iterative multiplication by the operator $\hat{K}$; see Algorithm 1 in Section 5.

3 Adaptive Estimation

In this section, we will show that the functional PLS estimator achieves the nearly optimal convergence rate on a class of ellipsoids. We consider an early-stopping rule to select the number of PLS components and show that it adapts to the complexity of the ellipsoid. Lastly, we study how rapidly the number of selected components increases with the sample size and make some comparisons to the PCA estimator.

3.1 Optimal Convergence Rates

Since the operators $K : \mathbb{H} \to \mathbb{H}$ and $\hat{K} : \mathbb{H} \to \mathbb{H}$ are self-adjoint and compact, by the spectral theorem

$$K = \sum_{j=1}^{\infty} \lambda_j v_j \otimes v_j \qquad \text{and} \qquad \hat{K} = \sum_{j=1}^{n} \hat{\lambda}_j \hat{v}_j \otimes \hat{v}_j,$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ and $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_n \ge 0$ are the eigenvalues of K and $\hat{K}$ and $(v_j)_{j=1}^{\infty}$ and $(\hat{v}_j)_{j=1}^{n}$ are the corresponding eigenvectors; see Kress (1999), Theorem 15.16. Note that the sample covariance operator $\hat{K}$ is a finite-rank operator with at most $n_* \le n$ distinct nonzero eigenvalues.

For any bounded and measurable function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, we define functions of operators through their spectral decompositions:

$$\phi(K) := \sum_{j=1}^{\infty} \phi(\lambda_j) v_j \otimes v_j \qquad \text{and} \qquad \phi(\hat{K}) := \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j) \hat{v}_j \otimes \hat{v}_j.$$

These definitions are commonly used in the inverse problems literature; see Engl et al. (1996).

The following inequalities for the operator norm will often be used:

$$\|\phi(K)\|_{\text{op}} \leq \sup_{\lambda \in [0,\lambda_1]} |\phi(\lambda)| \qquad \text{and} \qquad \|\phi(\hat{K})\|_{\text{op}} \leq \sup_{\lambda \in [0,\hat{\lambda}_1]} |\phi(\lambda)|, \tag{5}$$

where $\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$.

We next introduce several weak assumptions on the data-generating process.

Assumption 1. *$(X_i, Y_i)_{i=1}^n$ are i.i.d. copies of (X, Y) with $\mathbf{E}[X] = 0$, $\mathbf{E}\|X\|^4 < \infty$, and $\mathbf{E}[\varepsilon^2|X] \leq \sigma^2 < \infty$.*

Assumption 1 imposes mild restrictions on the data-generating process. Note that $\mathbf{E}\|X\|^4 < \infty$ is satisfied when X is a Gaussian process in $\mathbb{H}$. It implies that K is a nuclear operator and, hence, compact.

Assumption 2. *The operator $K : \mathbb{H} \to \mathbb{H}$ does not have zero eigenvalues.*

Assumption 2 ensures that the slope parameter β is identified. If this assumption is violated, our results apply to the identified component of β within the orthogonal complement of the null space of K; see Babii and Florens (2025) and Engl et al. (1996).

Assumption 3. *For some $\mu, C > 0$, the slope parameter β and the operator K belong to the class*

$$\mathcal{S}(\mu, C) = \left\{ \beta \in \mathbb{H},\ K : \mathbb{H} \to \mathbb{H} : \quad \sum_{j=1}^{\infty} \frac{\langle \beta, v_j \rangle^2}{\lambda_j^{2\mu}} \leq C^2, \quad \sum_{j=1}^{\infty} \lambda_j \leq C \right\}.$$

Assumption 3 describes the degree of ill-posedness in terms of the smoothness of β and the smoothing properties of the operator K. The parameter μ is known as the degree of ill-posedness. It restricts the rate of decay of the generalized Fourier coefficients $\langle \beta, v_j \rangle_{j \geq 1}$

relative to the eigenvalues of K. A larger value of μ means that it is easier to estimate the slope coefficient β; see also Carrasco et al. (2007). Recall also that the summability of eigenvalues holds whenever $\mathbf{E}\|X\|^2 < \infty$. Note that Assumption 3 is weaker than what is typically used to analyze the PCA estimators and does not require restricting the spacing between eigenvalues, cf. Hall and Horowitz (2007).

Consider now the so-called residual polynomial $\hat{Q}_m(\lambda) = 1 - \lambda\hat{P}_m(\lambda)$, deriving its name from the identity $\hat{r} - \hat{K}\hat{\beta}_m = \hat{Q}_m(\hat{K})\hat{r}$. It is known that the polynomial, $\hat{Q}_m$, has m distinct real roots, denoted $\hat{\theta}_1 > \hat{\theta}_2 > \cdots > \hat{\theta}_m > 0$. The sum of the inverses of these roots,

$$|\hat{Q}'_m(0)| = \sum_{j=1}^{m} \frac{1}{\hat{\theta}_j},$$

plays an important role in the analysis of the conjugate-gradient regularization; see Lemma S.4 in the Supplementary Material.

Our first result characterizes the convergence rate of the estimation and prediction errors of the PLS estimator.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 are satisfied. Then for every $s \in [0, 1]$, we have*

$$\left\| K^s(\hat{\beta}_m - \beta) \right\|^2 = O_P\left(|\hat{Q}'_m(0)|^{2(1-s)} n^{-1} + |\hat{Q}'_m(0)|^{-2(\mu+s)} + |\hat{Q}'_m(0)|^{-2s} n^{-\mu\wedge 1} \right),$$

provided that $|\hat{Q}'_m(0)| = O_P(n^{1/2})$.

Note that the last condition in Theorem 1 imposes that the number of components m does not increase too fast with the sample size and is not binding. In fact, it is optimal to have $|\hat{Q}'_m(0)| \sim n^{\frac{1}{2(\mu+1)}}$, in which case we obtain the following convergence rate

$$\left\| K^s(\hat{\beta}_m - \beta) \right\|^2 = O_P\left(n^{-\frac{\mu+s}{\mu+1}} \right).$$

When $s = 0$, this shows that the convergence rate of PLS in the Hilbert space norm is of order $n^{-\frac{\mu}{\mu+1}}$. On the other hand, when $s = 1/2$, we obtain the convergence rate of the

out-of-sample prediction error, since

$$\mathbf{E}_X\langle X, \hat{\beta}_m - \beta\rangle^2 = \left\| K^{1/2}(\hat{\beta}_m - \beta) \right\|^2,$$

where $\mathbf{E}_X$ is taken with respect to X, independent of $(Y_i, X_i)_{i=1}^n$.

Remark 1. *Note that the consistency of functional PLS has been established previously in Delaigle and Hall (2012) assuming that the eigenvalues are summable only which can be stated as $\mu = 0$. Characterizing the speed of convergence, however, requires more regularity with $\mu > 0$.*

The following result shows that no estimator can achieve faster than $n^{-\frac{\mu+s}{\mu+1}} \log^{-b} n$ rate on the class $\mathcal{S}(\mu, C)$.

Theorem 2. *For every $s \in [0, 1/2]$, there exists $A < \infty$ such that*

$$\liminf_{n\to\infty} \inf_{\hat{\beta}} \sup_{(\beta,K)\in\mathcal{S}(\mu,C)} \Pr\left(\left\| K^s(\hat{\beta} - \beta) \right\|^2 \geq A n^{-\frac{\mu+s}{\mu+1}} \log^{-b} n \right) > 0,$$

where $b > 2(\mu + s)$ and the infimum is over all estimators.

Therefore, we conclude that the PLS estimator $\hat{\beta}_m$ achieves the nearly optimal convergence rate on $\mathcal{S}(\mu, C)$, simultaneously for the estimation ($s = 0$) and prediction ($s = 1/2$) errors.

Remark 2. *Note that if β is in $\mathcal{S}(\mu, C)$ with $\mu = 1/2$, then it also belongs to the RKHS generated by the covariance operator K. In this case, we show that the minimax-optimal rate of prediction ($s = 1/2$) is $O_P(n^{-2/3})$. The result does not seem to be comparable to Cai and Yuan (2012), since their minimax-optimal $O_P(n^{-2r/(2r+1)})$ rate is obtained under the additional assumption that the eigenvalues decay polynomially fast, i.e. $\lambda_j \sim j^{-r}$.*[3]

[3]Similar assumptions are also made in Gupta et al. (2023), Gupta et al. (2025), and Lin and Cevher (2021).

The value $r = 1$ is not feasible since the existence of the covariance operator holds under $\mathbf{E}\|X\|^2 < \infty$ which is equivalent to $\sum_{j=1}^{\infty} \lambda_j < \infty$. In contrast, our class $\mathcal{S}(\mu, C)$ with $\mu = 1/2$ does not require specifying the decay rate of eigenvalues.

3.2 Adaptive PLS Estimator

Next, we look at the adaptive PLS estimator, where the number of components is selected using the early-stopping rule described in the following assumption.

Assumption 4. *We select $\hat{m}$ as the smallest $m \geq 0$ such that*

$$\left\| \hat{r} - \hat{K}\hat{\beta}_m \right\| \leq \tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}}$$

for $\tau > 1$ and some $\delta \in (0, 1)$.

Assumption 4 states that the selected number of PLS components $\hat{m}$ equals the first non-negative integer m such that the norm of the fitted 'moment' is smaller than a certain threshold; see Supplementary Material, Section S.5 for a practical implementation of this early-stopping rule. Note that the number of selected components is finite since $\hat{m} \leq n_*$, where n_* is the number of distinct nonzero eigenvalues of $\hat{K}$; see Proposition S.1 in the Supplementary Material. In fact, the norm of the 'residual' is zero for $m \geq n_*$ which corresponds to perfect fit.

The following result shows that the early-stopping rule in Assumption 4 is adaptive to the unknown degree of ill-posedness $\mu > 0$.

Theorem 3. *Suppose that Assumptions 1, 2, 3, and 4 hold with $\delta \geq 1/n$. Then*

$$\left\| K^s(\hat{\beta}_{\hat{m}} - \beta) \right\|^2 = O\left((\delta n)^{-\frac{\mu+s}{\mu+1}} \right)$$

with probability at least $1 - \delta$ for every $s \in [0, 1]$.

Taking $\delta_n = 1/\log n$ in Assumption 4, we obtain from Theorem 3 the convergence rate of the estimation and prediction errors of PLS with the number of components selected iteratively with the early-stopping rule:

$$\left\| K^s(\hat{\beta}_{\hat{m}} - \beta) \right\|^2 = O_P\left(\left(\frac{\log n}{n} \right)^{\frac{\mu+s}{\mu+1}} \right).$$

Therefore, the adaptive PLS achieves the nearly optimal convergence rate simultaneously for the estimation and prediction errors without knowing the degree of ill-posedness $\mu > 0$.

3.3 Number of Selected Components

In this section, we look at how rapidly the number of selected components in Assumption 4 increases with the sample size. First, we consider a somewhat conservative bound that does not impose any assumptions on the spectrum of the operator K.

Theorem 4. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq 1/n$ and $\mu \geq 1$. Then with probability at least $1 - \delta$*

$$\hat{m} = O\left((n\delta)^{\frac{1}{4(\mu+1)}} \right).$$

Taking $\delta = 1/\log n$, we obtain from Theorem 4 that $\hat{m} = O_P\left((n/\log n)^{\frac{1}{4(\mu+1)}} \right)$. Next, we consider sharper estimates under additional assumptions imposed on the spectrum of the operator K.

Theorem 5. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq e/n$ and $\mu \geq 1$. Then with probability at least $1 - \delta$*

(i) If $\lambda_j = O(j^{-2\kappa})$ for some $\kappa > 0$, then

$$\hat{m} = O\left((n\delta)^{\frac{1}{4(\kappa+1)(\mu+1)}} \right).$$

(ii) If $\lambda_j = O(q^j)$ for some $q \in (0,1)$, then

$$\hat{m} = O\left(\log(n\delta)\right).$$

Theorem 5 shows that if the eigenvalues decay polynomially fast, then the selected number of components is $\hat{m} = O_P\left(n/\log n\right)^{\frac{1}{4(\kappa+1)(\mu+1)}}$ while in the case of the geometric decay, the number of selected components increases only logarithmically with the sample size. Therefore, the adaptive early-stopping rule will choose a smaller number of components if the eigenvalues of the operator K decay faster and vice versa.

4 Inference

4.1 A Hypothesis Test

We aim to test the null hypothesis against the alternative hypothesis:

$$H_0 : \ \beta = b \qquad \text{vs.} \qquad H_1 : \ \beta \neq b,$$

for some known $b \in \mathbb{H}$. We are also interested in testing against a sequence of local alternative hypotheses

$$H_{1,n} : \ \beta = b + n^{-1/2}\Delta$$

for some $\Delta \in \mathbb{H}$.

Recall that our main result in Theorem 1 shows that

$$\left\| K^s(\hat{\beta}_m - \beta) \right\|^2 = O_P\left(n^{-\frac{\mu+s}{\mu+1}}\right), \qquad \forall s \in [0,1],$$

provided that the number of components m is properly selected. The convergence rate is $O_P(n^{-1})$ for $s = 1$ and it is slower than $O_P(n^{-1})$ for $s < 1$. This suggests that the following statistic

$$T_n = n \left\| \hat{K}(\hat{\beta}_m - b) \right\|^2$$

may have a well-defined asymptotic distribution under H_0.[4]

Let $V = \mathbf{E}[\varepsilon^2 X \otimes X]$ be the variance operator with spectral decomposition

$$V = \sum_{j=1}^{\infty} \omega_j \varphi_j \otimes \varphi_j$$

and let $Z_j \overset{\text{i.i.d.}}{\sim} N(0,1)$. The following theorem establishes the limiting distribution of T_n under the null and alternative hypotheses.

Theorem 6. *Suppose that Assumptions 1 and 2 are satisfied and the PLS estimator is computed using m components such that $\|\hat{r} - \hat{K}\hat{\beta}_m\| = o_P(n^{-1/2})$. Then, under H_0,*

$$T_n \xrightarrow{d} \sum_{j=1}^{\infty} \omega_j Z_j^2,$$

while under H_1, $T_n \xrightarrow{\text{a.s.}} \infty$. Moreover, under $H_{1,n}$,

$$T_n \xrightarrow{d} \sum_{j=1}^{\infty} \omega_j Z_j^2 + 2\sum_{j=1}^{\infty} \omega_j^{1/2} Z_j \langle \varphi_j, K\Delta \rangle + \|K\Delta\|^2.$$

Remark 3. *The requirement $\|\hat{r} - \hat{K}\hat{\beta}_m\| = o_P(n^{-1/2})$ in Theorem 6 can be easily satisfied when the number of components m is sufficiently large. Indeed, we know that $\|\hat{r} - \hat{K}\hat{\beta}_m\| = 0$ for $m = n_*$, where $n_* \leq n$ is the number of nonzero eigenvalues of the sample covariance operator $\hat{K}$, cf. Supplementary Material, Proposition S.1.*

Remark 4. *Theorem 6 illustrates an interesting phenomenon related to the optimal choice of the number of PLS components which amounts to early stopping of the conjugate-gradient method. While it is optimal to select a small number of components to prevent overfitting for estimation, this should be avoided for inference. In fact, it is evident from the proof of Theorem 6 that the overfitted PLS with zero fitting error leads to more accurate asymptotic approximation; cf. Bartlett et al. (2020) for benign overfitting in the linear regression.*

[4]The statistic is similar to Wald statistic which under homoskedasticity roughly corresponds to $s = -1/2$, hence, not converging at the parametric rate.

Remark 5. *The test has more power against the alternatives with larger values of*

$$\|K\Delta\|^2 = \sum_{j=1}^{\infty} \lambda_j^2 \langle \Delta, v_j \rangle^2,$$

where $(\lambda_j, v_j)_{j=1}^{\infty}$ *are the eigenvalues and eigenvectors of the operator* K*. Such alternatives are aligned with directions defined by the leading eigenvectors of* K*, i.e. directions in which it is easier to identify the slope function* β*.*

Remark 6. *Under homoskedasticity,* $\mathbf{E}[\varepsilon^2|X] = \sigma^2$*, we have* $V = \sigma^2 K$*. In this case,* V *has eigenvectors* $\varphi_j = v_j$ *with corresponding eigenvalues* $\omega_j = \sigma^2 \lambda_j$*.*

Theorem 6 implies that the test based on T_n has correct size under the null hypothesis and is consistent under the fixed alternative hypotheses which we state as a trivial corollary below. To that end, let $z_{1-\alpha}$ be the $1-\alpha$ quantile of

$$T = \sum_{j=1}^{\infty} \omega_j Z_j^2.$$

The test rejects H_0 if $T_n > z_{1-\alpha}$. Then, the following result holds:

Corollary 1. *Suppose that the conditions of Theorem 6 are satisfied. Then under* H_0*, we have*

$$\lim_{n\to\infty} \Pr(T_n > z_{1-\alpha}) = \alpha$$

while under H_1*, we have*

$$\lim_{n\to\infty} \Pr(T_n > z_{1-\alpha}) = 1.$$

Moreover, under $H_{1,n}$

$$\lim_{n\to\infty} \Pr(T_n > z_{1-\alpha}) \uparrow 1 \qquad \text{as} \qquad \|K\Delta\| \uparrow \infty.$$

It is also easy to show that the critical value $z_{1-\alpha}$ can be consistently estimated with simple nonparametric bootstrap based on the i.i.d. draws from the sample $(Y_i, X_i)_{i=1}^n$. Alternatively, one can simulate the asymptotic critical values using the eigenvalues $\hat{\omega}_j$ of the sample variance operator $\hat{V}$.

4.2 A Confidence Set

Inverting the test, we obtain a confidence set

$$C_{1-\alpha} = \{b \in \mathbb{H} : \ T_n(b) \leq z_{1-\alpha}\},$$

where $z_{1-\alpha}$ is the quantile of order $1-\alpha$ of $T = \sum_{j=1}^{\infty} \omega_j Z_j^2$. Indeed, since $T_n \xrightarrow{d} T$, the confidence set has the coverage probability $1-\alpha$:

$$\lim_{n\to\infty} \Pr(\beta \in C_{1-\alpha}) = \lim_{n\to\infty} \Pr(T_n(\beta) \leq z_{1-\alpha}) = \Pr(T \leq z_{1-\alpha}) = 1-\alpha.$$

Since β is an infinite-dimensional object, computing this confidence set numerically requires finite-dimensional approximations which can be done as follows. Let $(h_j)_{j=1}^{\infty}$ be a basis of $\mathbb{H}$. Then we approximate $C_{1-\alpha}$ numerically by computing

$$C_{1-\alpha}^J = \left\{ (b_1, \ldots, b_J) \in \mathbb{R}^J : \ T_n\left(\sum_{j=1}^{J} b_j h_j\right) \leq z_{1-\alpha} \right\}.$$

For instance, in functional data analysis, we often have $\mathbb{H} = L_2[0,1]$ in which case, we can use the Fourier basis or Legendre polynomials. If $\mathbb{H} = L_2(\mathbb{R})$, the Hermite polynomials are a natural choice; see Supplementary Material, Section S.5 for numerical implementation and simulation results.

5 Monte Carlo Experiments

In this section, we conduct several Monte Carlo experiments to evaluate the finite-sample performance of the PLS estimator. We simulate i.i.d. samples $(Y_i, X_i)_{i=1}^n$ from the functional linear model

$$Y_i = \int_0^1 X_i(s)\beta(s)\mathrm{d}s + \varepsilon_i, \qquad \varepsilon_i \sim_{\text{i.i.d.}} N(0,1),$$

where the predictors X_i belong to the Hilbert space of square-integrable functions with respect to the Lebesgue measure, denoted $\mathbb{H} = L_2[0,1]$. The functional predictor is generated as

$$X_i(s) = \sum_{j=1}^{100} \sqrt{\lambda_j} u_j v_j(s), \qquad u_j \sim_{\text{i.i.d.}} N(0,1).$$

We specify the slope parameter $\beta \in L_2[0,1]$ and the spectrum $(\lambda_j, v_j)_{j\geq 1}$ correspond to one of the following three models:

- **Model 1:** $\beta(s) = \sum_{j=1}^{100} \beta_j v_j(s)$ with $\beta_j = 4j^{-2.7}$ and $\lambda_j = 2j^{-1.1}$ for all $j \geq 1$, $v_j(s) = \sqrt{2}\cos(j\pi s)$, $j \geq 2$, and $v_1(s) = 1$.
- **Model 2:** same as Model 1, but with $\beta_j = 4$ for all $j = 1, \ldots, 5$.
- **Model 3:** same as Model 1, but with $\lambda_j = 2$ for all $j = 1, \ldots, 5$.

All three models satisfy Assumption 3. Model 2 emphasizes the importance of the first five coefficients β_j, making the estimation and prediction tasks more challenging. Model 3 introduces repeated eigenvalues, causing the first five eigenvectors to be non-identifiable, although the slope parameter remains identifiable.

We compute the PLS estimator using Algorithm 1, which is numerically equivalent to the estimator given by Equation (4); see Hanke (1995), Algorithm 2.1 and Proposition 2.1. This approach avoids direct inversion of the empirical operator $\hat{K}$ by utilizing iterative multiplication, making it suitable for solving high-dimensional linear systems of the form $\hat{K}\hat{\beta} = \hat{r}$ with a symmetric matrix $\hat{K}$.

Algorithm 1: PLS algorithm for solving $\hat{K}\hat{\beta} = \hat{r}$.

Result: $\hat{\beta}_m$

Initialisation: $\hat{\beta}_0 = 0$, $d_0 = e_0 = \hat{r} - \hat{K}\hat{\beta}_0$;

for $j = 0, 1, \ldots, m-1$ **do**

1. Compute step size: $\alpha_j = \frac{\langle e_j, \hat{K}e_j \rangle}{\|\hat{K}d_j\|^2}$;
2. Update slope coefficient: $\hat{\beta}_{j+1} = \hat{\beta}_j + \alpha_j d_j$;
3. Update fitted moment: $e_{j+1} = e_j - \alpha_j \hat{K} d_j$;
4. Compute step size for the conjugate direction update: $\gamma_{j+1} = \frac{\langle e_{j+1}, \hat{K}e_{j+1} \rangle}{\langle e_j, \hat{K}e_j \rangle}$;
5. Update conjugate direction vector: $d_{j+1} = e_{j+1} + \gamma_{j+1} d_j$;

end

The integrals in inner products and the operator K are discretized using a simple Riemann sum approximation over a grid of $T = 200$ equidistant points in $[0, 1]$.[5] The experiments feature $5,000$ replications, with each replication generating samples of size $n = 100$. For each simulation experiment and an estimator $\hat{\beta}$, we compute:

- The integrated squared error (ISE):

$$\mathrm{ISE}(\hat{\beta}) = \int_0^1 |\hat{\beta}(s) - \beta(s)|^2 \mathrm{d}s;$$

- the mean-squared prediction error (MSPE):

$$\mathrm{MSPE}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, \hat{\beta} \rangle)^2,$$

  where the estimator $\hat{\beta}$ is computed from an auxiliary independent sample of size n.

We compare the performance of our functional PLS estimator with the number of components chosen using our early-stopping rule, against several alternative methods:[6]

[5] This gives a satisfactory approximation under weak assumptions. Alternatively, one could employ quadrature rules with fewer points to reduce computational load.

[6] See Supplementary Material, Section S.5, for more details on the practical implementation.

1. A spline estimator, jointly selecting the spline degree and smoothing parameter via generalized cross-validation (GCV), following Crambes et al. (2009).

2. A principal component regression (PCA) estimator, where component selection is also done by GCV.

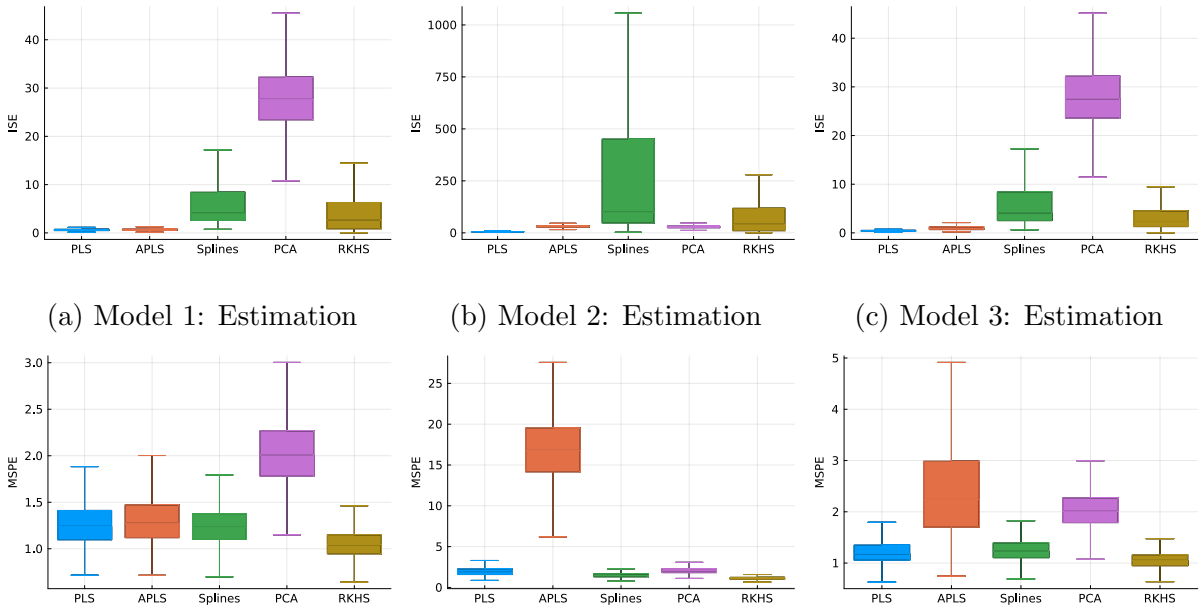
3. A reproducing kernel Hilbert space (RKHS) estimator using the same kernel as in Yuan and Cai (2010), equation (16).

4. The alternative PLS method from Delaigle and Hall (2012), selecting components through 5-fold cross-validation.

Note that Crambes et al. (2009), Proposition 2, establishes that the optimal rate of prediction for the smoothing spline estimator can be attained adaptively using generalized cross-validation. On the other hand, the kernel function employed in Yuan and Cai (2010) is not adaptively selected from the data. We are not aware of any adaptivity results for other competing estimators, including PCA and the alternative PLS estimator with cross-validation. In contrast, our method for selecting the number of PLS components is adaptive for both estimation and prediction according to Theorem 3.

Figure 1 summarizes the distributions of estimation errors (ISE) and prediction errors (MSPE) from $5,000$ simulations. The early-stopped PLS estimator achieves the lowest median estimation error, with consistently lower variability across all models. The penalized spline estimator performs effectively when coefficients decay rapidly but exhibits notably larger errors in Model 2, which involves high-frequency components. Nevertheless, it shows excellent predictive performance consistent with Crambes et al. (2009). The PLS estimator also demonstrates competitive predictive accuracy, although it does not yield the lowest median prediction error across all simulation designs. The RKHS estimator achieves

good prediction performance but inferior estimation across all models. The alternative PLS estimator generally performs comparably to our proposed PLS approach but exhibits notably larger prediction errors for Models 2 and 3. These higher errors may be attributed to numerical instability associated with finite-precision arithmetic, as discussed in Delaigle and Hall (2012). Finally, PCA demonstrates relatively modest performance across all the considered models. Moreover, we find in simulations that the FPLS estimator outperforms FPCA using a smaller number of components offering a more parsimonious representation of the data; see also Frank and Friedman (1993) for related discussion.



(a) Model 1: Estimation (b) Model 2: Estimation (c) Model 3: Estimation

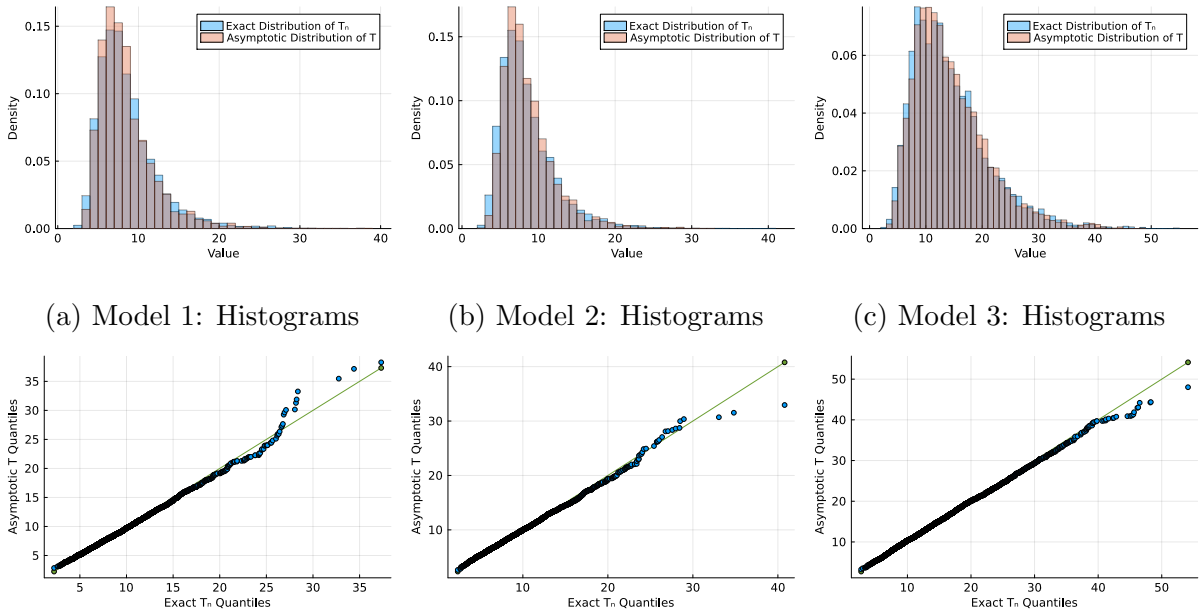
(d) Model 1: Prediction (e) Model 2: Prediction (f) Model 3: Prediction

Figure 1: Estimation and prediction accuracy. Top row: integrated squared error (ISE) for the slope coefficient. Bottom row: mean squared prediction error (MSPE). Each boxplot summarizes 5,000 Monte Carlo replications.

Next, we evaluate the performance of the proposed test. We first assess whether the asymptotic distribution provides a good approximation to the finite-sample distribution under the null hypothesis H_0. To ensure that the late stopping conditions of Theorem 6

are satisfied, we employ PLS with $m = 70$ components.[7] The results are presented in Figure 2.

For each of the three models, the top row in Figure 2 displays the finite-sample distribution of T_n (in blue) under H_0, overlaid with the simulated $\sum_{j=1}^{100} \omega_j Z_j^2$ (in orange) obtained from Theorem 6. The bottom row presents the corresponding Q–Q plots, comparing the empirical quantiles of T_n to the asymptotic quantiles of T. Overall, the simulated asymptotic distribution closely matches the finite-sample distribution, with only minor discrepancies observed in the extreme right tail.



(a) Model 1: Histograms (b) Model 2: Histograms (c) Model 3: Histograms

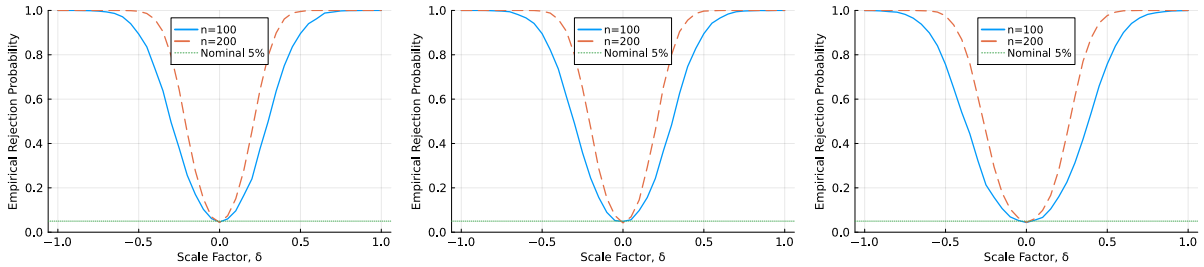
(d) Model 1: Q–Q plot (e) Model 2: Q–Q plot (f) Model 3: Q–Q plot

Figure 2: Accuracy of asymptotic approximations. Top row: finite-sample distribution of T_n under H_0 (blue histogram) overlaid with its asymptotic limit (orange histogram). Bottom row: Q–Q plots comparing empirical quantiles of T_n with the corresponding asymptotic quantiles.

To study the finite-sample power of the test, we simulate the power curves corresponding

[7] We find that for these models the fitting error does not substantially decrease after $m = 10$ components.

to the local alternatives $\beta(s) + \delta s$ with deviations measured by the scale factor $\delta \in [-1, 1]$. $\delta = 0$ corresponds to H_0 while $|\delta| > 0$ to H_1. We also consider doubling the sample size from $n = 100$ to $n = 200$. The results are displayed in Figure 3, confirming that the test has more power once the null and the alternative hypotheses become sufficiently separated. The power also increases with the sample size as expected. Lastly, we provide additional simulation results in the Supplementary Material.



(a) Model 1 (b) Model 2 (c) Model 3

Figure 3: Simulated power curves. Power curves for the three models corresponding to sample size $n = 100$ (blue solid curve) and $n = 200$ (red dashed curve). The green dotted line indicates the 5% nominal significance level.

6 Nonlinear Temperature Effects in U.S. Agriculture

The global surface temperature has increased by 1.1°C above pre-industrial levels and could increase up to 3.6–4.5°C by the end of the century if current CO_2 emissions rise steadily according to the latest studies; see IPCC (2021). Global warming will likely lead to more frequent and severe heat waves, altered precipitation patterns, and intensified droughts. Of all major sectors, agriculture is arguably the most sensitive to climate change. While constituting a modest share of developed economies, it is vital for food security. Indeed, the intensified droughts could cause food shortages which in turn may exacerbate mass

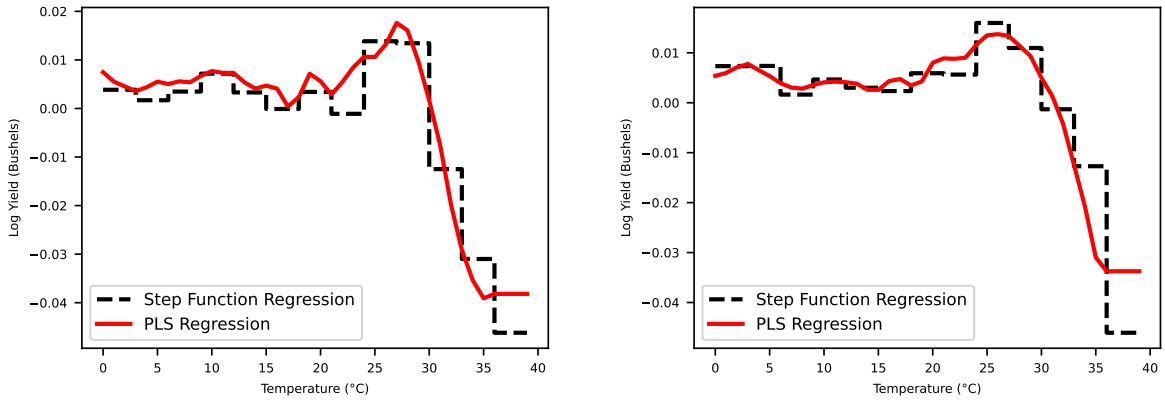
migration and violent conflicts. Some have argued that many regions are already warmer than optimal for agriculture in many parts of Asia, Africa, and Latin America; see Nordhaus (2013).

Determining the precise functional form of the relationship between crop yields and temperature has recently attracted lots of attention; see Schlenker and Roberts (2006, 2009) and Cui et al. (2024). We argue that the methodology used to estimate such nonlinear temperature effects can be understood as a functional linear regression, where the outcome Y_{it} is the log-yield of a crop of a county i in a year t, measured in bushels per acre, and the functional regressor $(X_{it}(s))_{s\in[0,40]}$ is a temperature curve, representing the crop exposure to temperatures between 0°C and 40°C during the growing season. Following Schlenker and Roberts (2009), the temperature curves are computed at discrete points $s_j \in \{0, 1, \ldots, 40\}$ as $X_{it}(s_j) = \Phi_{it}(s_j + 1) - \Phi_{it}(s_j)$, where $\Phi_{it}(s_j)$ is the length of time (measured in days) the crop was continuously exposed to temperature larger than s_j for county i in year t.

We focus on corn and soybeans, which are the two major crops grown in the U.S. The dataset consists of fine-scale county-level crop yields and weather outcomes, spanning U.S. counties east of the 100th meridian from 1950 to 2020. We use the same set of controls as in Schlenker and Roberts (2009), namely: a constant, precipitation measured in mm from March through August, precipitation^2, county fixed effects, and a state-specific quadratic time trend to capture technological change. The crop yields Y and the temperature curve X are regressed on these controls to obtain the residuals, which are subsequently used for the functional data analysis.

The slope coefficient is then estimated using: 1) our functional PLS estimator; and 2) a highly parameterized least-squares estimator with a step function approximation as in Schlenker and Roberts (2009). The latter fits a separate temperature effect for each 3°C bin

from 0°C to 40°C, hence, it involves 13 parameters. On the other hand, our early-stopping rule finds $\hat{m} = 4$ functional PLS components both for corn and soybeans; see Appendix Section S.5 for more details on the implementation.



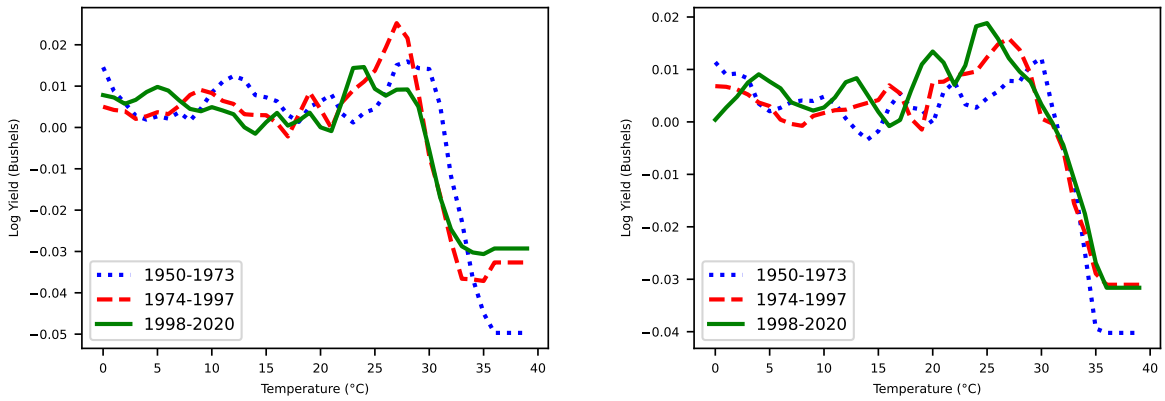
(a) Corn Yield

(b) Soybean Yield

Figure 4: Nonlinear temperature effects. The figure shows the estimated nonlinear relationship between temperature and crop yields fitted using functional PLS (red solid curve) and step function approximation (black dashed curve).

Figure 4 displays the estimated functional slope coefficient β corresponding to our functional PLS (red solid curve) and step function approximation (black dashed curve) for corn and soybeans. We find that the critical temperature after which the crop yields start declining is around 29–30°C which is similar to findings reported in Schlenker and Roberts (2009).

Lastly, we look at how the nonlinear temperature effects have changed over time. Figure 5 reports the estimated functional slope coefficient splitting the data into three subsamples: 1950–1973 (blue dotted curve), 1974–1997 (red dashed curve), and 1998–2020 (green solid curve). The results indicate that the negative temperature effects were larger during 1950–1973 compared to the most recent 22 years, especially for the extreme temperatures.



(a) Corn Yield

(b) Soybean Yield

Figure 5: Adaptation. The figure plots the estimated nonlinear relationship between temperature and crop yields for three subsamples: 1950–1973, 1974–1997, and 1998–2020.

The mitigation of extreme temperature effects may come from two sources: adaptation and CO_2 fertilization. CO_2 fertilization effects observed in our sample are likely to be small; see Nordhaus (2013) who argues that doubling the atmospheric concentration of CO_2 would increase crop yields by $10 - -15\%$ only. In contrast, the adaptation effect is likely to dominate over time. It can be attributed to the actions taken by farmers, such as adjusting the sowing and harvesting dates to maximize yields, using more resilient crops, or building efficient irrigation systems. Our results, therefore, suggest some evidence of adaptation in U.S. agriculture which has been reported in Burke and Emerick (2016) without properly accounting for nonlinearities. However, our confidence sets are too wide to draw conclusive inferences once statistical uncertainty is accounted for because there are only a few observations of extreme temperatures in the data.

7 Conclusions

This paper reframes functional partial least squares (PLS) as a conjugate-gradient regularization scheme for ill-posed functional linear regression. Under mild conditions we prove that the resulting estimator achieves the minimax convergence rate (up to a logarithmic factor) and introduce a fully data-driven early-stopping rule that adapts automatically to the unknown degree of ill-posedness. We also propose a novel test and a confidence set for the functional coefficient β.

Monte Carlo evidence indicates that the estimator matches—or exceeds—the estimation and prediction accuracy of leading alternatives, particularly functional PCA, while requiring substantially fewer components; the stopping rule consistently adjusts to the prevailing signal-to-noise structure.

Applied to seven decades of U.S. county-level crop data, the method suggests that the adverse impact of high temperatures on corn and soybean yields has weakened since the 1950s, hinting at adaptive responses in U.S. agriculture. Because observations of extreme heat are sparse, however, firmer conclusions will require richer data or refinements to inference procedures.

References

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

Andrii Babii and Jean-Pierre Florens. Is completeness necessary? Estimation in nonidentified linear models. *Econometric Theory (forthcoming)*, 2025.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Amparo Baíllo. A note on functional linear regression. *Journal of Statistical Computation and Simulation*, 79(5):657–669, 2009.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

Marshall Burke and Kyle Emerick. Adaptation to climate change: Evidence from US agriculture. *American Economic Journal: Economic Policy*, 8(3):106–140, 2016.

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

T Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

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

Marine Carrasco and Barbara Rossi. In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, 34(3):313–338, 2016.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Christophe Crambes, Alois Kneip, and Pascal Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72, 2009.

Xiaomeng Cui, Bulat Gafarov, Dalia Ghanem, and Todd Kuffner. On model selection criteria for climate change impact studies. *Journal of Econometrics*, 239(1):105511, 2024.

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

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, 1996.

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

Jean-Pierre Florens and Sébastien Van Bellegem. Instrumental variable estimation in functional linear models. *Journal of Econometrics*, 186(2):465–476, 2015.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

Naveen Gupta, S Sivananthan, and Bharath K Sriperumbudur. Convergence analysis of ker-

nel conjugate gradient for functional linear regression. *arXiv preprint arXiv:2310.02607*, 2023.

Naveen Gupta, S Sivananthan, and Bharath K Sriperumbudur. Optimal rates for functional linear regression with general regularization. *Applied and Computational Harmonic Analysis*, 76:101745, 2025.

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

Martin Hanke. *Conjugate gradient type methods for ill-posed problems*, volume 327 of *Pitman Research Notes in Mathematics Series*. CRC Press, 1995.

Inge S Helland. On the structure of partial least squares regression. *Communications in statistics - Simulation and Computation*, 17(2):581–607, 1988.

Magnus R Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

IPCC. Summary for policymakers. In: Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the Intergovernmental Panel on Climate Change (IPCC). pages 3–32, 2021.

Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, (3):300–303, 1982.

Bryan Kelly and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316, 2015.

Nicole Krämer, Anne-Laure Boulesteix, and Gerhard Tutz. Penalized partial least squares

with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):60–69, 2008.

Rainer Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences*. Springer, 1999.

Yehua Li and Tailen Hsing. On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98(9):1782–1804, 2007.

Junhong Lin and Volkan Cevher. Kernel conjugate gradient methods with random projections. *Applied and Computational Harmonic Analysis*, 55:223–269, 2021.

Arkadi S Nemirovski. On regularizing properties of the conjugate gradient method for ill-posed problems (in russian). *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 26(3):332–347, 1986.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

William Nordhaus. *The climate casino: Risk, uncertainty, and economics for a warming world*. Yale University Press, 2013.

Aloke Phatak and Frank de Hoog. Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16(7):361–367, 2002.

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

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002.

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

Semanur Saricam, Ufuk Beyaztas, Baris Asikgil, and Han Lin Shang. On partial least-squares estimation in scalar-on-function regression models. *Journal of Chemometrics*, 36 (12):e3452, 2022.

Wolfram Schlenker and Michael J Roberts. Nonlinear effects of weather on corn yields. *Review of Agricultural Economics*, 28(3):391–398, 2006.

Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009.

Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

Ming Yuan and T Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

SUPPLEMENTARY MATERIAL

Functional Partial Least-Squares: Adaptive Estimation and Inference

Andrii Babii
Department of Economics, UNC-Chapel Hill
and
Marine Carrasco
Department of Economics, University of Montreal
and
Idriss Tsafack
Department of Economics, University of Montreal

August 11, 2025

S.1 Notation and Preliminary Results

In this section, we describe the notation and collect several propositions and lemmas.

Notation: For two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we will use $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \leq cb_n$ for all $n \geq 1$. We will also use $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two real numbers, we use $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $k \in \mathbb{Z}$, consider the following measure

$$\hat{\mu}_k = \sum_{j=1}^{n_*} \hat{\lambda}_j^k \langle \hat{r}, \hat{v}_j \rangle^2 \delta_{\hat{\lambda}_j},$$

where δ_x is the Dirac measure at $x \in \mathbb{R}$ and $(\hat{\lambda}_j, \hat{v}_j)_{j=1}^{n_*}$ is the eigendecomposition of $\hat{K}$. For $\phi, \psi : [0, \hat{\lambda}_1] \to \mathbb{R}$, define

$$\begin{aligned} [\phi, \psi]_k &= \int_0^{\hat{\lambda}_1} \phi(\lambda)\psi(\lambda) \mathrm{d}\hat{\mu}_k(\lambda) \\ &= \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j)\psi(\hat{\lambda}_j)\hat{\lambda}_j^k \langle \hat{r}, \hat{v}_j \rangle^2. \end{aligned} \tag{S.1}$$

Lastly, let

$$\Pi_a = \sum_{j:\hat{\lambda}_j \leq a} \hat{v}_j \otimes \hat{v}_j$$

be the orthogonal projection operator on the eigenspaces of $\hat{K}$ corresponding to eigenvalues smaller or equal to a.

The following proposition states that the PLS estimator $\hat{\beta}_m$ is unique for every $m \leq n_*$ and that the tuning parameter selected in Assumption 4 does not exceed the number of unique nonzero eigenvalues, n_*; see also Blanchard and Krämer (2016) for a kernel regression setting.

Proposition S.1. *The solution in Equation (4) is unique for every $m \leq n_*$. Moreover, $\hat{m} \leq n_*$.*

Proof of Proposition S.1. Let $\mathcal{P}_m$ be the space of real polynomials of degree at most m and let $\mathcal{P}_m^0$ be the affine set of polynomials whose constant term equals one. The PLS problem in Equation (4) amounts to fitting a polynomial of degree $m-1$ solving

$$\hat{P}_m \in \underset{\phi \in \mathcal{P}_{m-1}}{\arg\min} \left\| \left[I - \hat{K}\phi(\hat{K}) \right] \hat{r} \right\|^2$$

or equivalently a residual polynomial $\hat{Q}_m(\lambda) = 1 - \lambda \hat{P}_m(\lambda)$ solving

$$\hat{Q}_m \in \underset{\phi \in \mathcal{P}_m^0}{\arg\min} \left\| \phi(\hat{K})\hat{r} \right\|^2. \tag{S.2}$$

By Parseval's identity, for every $\phi : [0, \hat{\lambda}_1] \to \mathbb{R}$, the objective function can be written as

$$\left\| \phi(\hat{K})\hat{r} \right\|^2 = \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j)^2 \langle \hat{r}, \hat{v}_j \rangle^2 = [\phi, \phi]_0, \tag{S.3}$$

where $[.,.]_0$ is defined in Equation (S.1). Therefore $\hat{Q}_m$ minimizes $\phi \mapsto [\phi, \phi]_0$ on $\mathcal{P}_m^0$. It is easy to see that $[.,.]_0$ is an inner product for every $m < n_*$. Therefore, $\hat{Q}_m$ is the unique projection of zero on a closed convex set $\mathcal{P}_m^0 \subset \mathcal{P}_m$ with respect to $[.,.]_0$. For $m = n_*$, $[.,.]_0$ is not an inner product because we can take an n_*-degree polynomial $\phi \neq 0$ with roots equal to the distinct n_* eigenvalues of $\hat{K}$, so that $[\phi, \phi]_0 = 0$. However, such a polynomial is unique. Therefore, $\hat{P}_m$ and $\hat{\beta}_m$ are unique for every $m \leq n_*$. This also shows that the PLS objective function attains 0 for $m \geq n_*$, hence, the tuning parameter in Assumption 4 satisfies $\hat{m} \leq n_*$. □

We will need the following tail inequality in Hilbert spaces.

Lemma S.1. *Let $(\xi_i)_{i=1}^n$ be independent random variables in a Hilbert space $(\mathbb{H}, \langle .,. \rangle)$ with the induced norm $\|.\|$. Suppose that $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\|\xi_i\|^2 < \infty$. Then for every $\gamma \in (0,1)$*

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\| \leq \sqrt{\frac{\sum_{i=1}^n \mathbf{E}\|\xi_i\|^2}{\gamma n^2}} \right) \geq 1 - \gamma.$$

Proof of Lemma S.1. By Markov's inequality, $\forall u > 0$

$$\begin{aligned}\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\| > u\right) &\leq u^{-2}\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\|^2 \\ &= \frac{1}{u^2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left\langle\xi_i,\xi_j\right\rangle \\ &= \frac{1}{u^2n^2}\sum_{i=1}^{n}\mathbf{E}\left\|\xi_i\right\|^2,\end{aligned}$$

where the last two lines follow because the summands are independent and mean-zero. Setting $\gamma = \sum_{i=1}^{n}\mathbf{E}\|\xi_i\|^2/(n^2u^2)$ and solving for u gives the result. □

Lemma S.1 allows us to control the tail probabilities for the PLS residual as well as the covariance operator errors on an event with probability at least $1-\gamma$.

Lemma S.2. *Suppose that Assumption 1 is satisfied. Then for every $\gamma \in (0,1)$*

$$\left\|\hat{r} - \hat{K}\beta\right\| \leq \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} \qquad \textit{and} \qquad \left\|\hat{K} - K\right\|_{\mathrm{HS}} \leq \sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}}$$

with probability at least $1-\gamma$, where $\|.\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm.

Proof of Lemma S.2. We will apply Lemma S.1. First, we note that

$$\left\|\hat{r} - \hat{K}\beta\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right\|,$$

where $\mathbf{E}\|\varepsilon_i X_i\|^2 \leq \sigma^2\mathbf{E}\|X_i\|^2 < \infty$ under Assumption 1. Then by Lemma S.1 with probability at least $1-\gamma/2$, we have $\|\hat{r} - \hat{K}\beta\| \leq \sigma\sqrt{2\mathbf{E}\|X\|^2/\gamma n}$. Second, the space of Hilbert-Schmidt operators is a Hilbert space and

$$\left\|\hat{K} - K\right\|_{\mathrm{HS}} = \left\|\frac{1}{n}\sum_{i=1}^{n}X_i\otimes X_i - \mathbf{E}[X_i\otimes X_i]\right\|_{\mathrm{HS}},$$

where $\mathbf{E}\|X_i\otimes X_i - \mathbf{E}[X_i\otimes X_i]\|_{\mathrm{HS}}^2 \leq \mathbf{E}\|X_i\otimes X_i\|_{\mathrm{HS}}^2 = \mathbf{E}\|X_i\|^4$. Then by Lemma S.1 with probability at least $1-\gamma/2$, we have $\|\hat{K} - K\|_{\mathrm{HS}} \leq \sqrt{2\mathbf{E}\|X\|^4/\gamma n}$. The result follows by the union bound. □

We will also use the following two inequalities known in perturbation theory.

Lemma S.3. *Let $A : \mathbb{H} \to \mathbb{H}$ and $B : \mathbb{H} \to \mathbb{H}$ be two nonnegative self-adjoint Hilbert-Schmidt operators. Then*

$$\|A^\mu - B^\mu\|_{\text{op}} \leq c_\mu \|A - B\|_{\text{op}}^\mu, \qquad 0 < \mu < 1$$

and

$$\|A^\mu - B^\mu\|_{\text{HS}} \leq \mu \nu^{\mu-1} \|A - B\|_{\text{HS}}, \qquad \mu \geq 1,$$

where $\nu = \|A\|_{\text{op}} \vee \|B\|_{\text{op}}$.

Proof of Lemma S.3. See Aleksandrov and Peller (2016), Theorem 1.7.2 for the first inequality. The second inequality follows from Aleksandrov and Peller (2016), Theorem 3.5.1. □

As an immediate consequence of Lemmas S.2 and S.3, since $\|.\|_{\text{op}} \leq \|.\|_{\text{HS}}$, for every $\gamma \in (0, 1)$, we have

$$\left\| \hat{K}^\mu - K^\mu \right\|_{\text{op}} \leq (c_\mu \mathbf{1}_{\mu \leq 1} + \mu \nu^{\mu-1} \mathbf{1}_{\mu > 1}) \left(\frac{2\mathbf{E}\|X\|^4}{\gamma n} \right)^{\frac{\mu \wedge 1}{2}} \tag{S.4}$$

on an event that holds with probability at least $1 - \gamma$, where $\nu = \|\hat{K}\|_{\text{op}} \vee \|K\|_{\text{op}}$.

The following Lemma presents some useful results on the residual polynomials $\hat{Q}_m(\lambda) = 1 - \lambda \hat{P}_m(\lambda)$; see Engl et al. (1996) and Hanke (1995). For the completeness of the presentation, we sketch proofs for the key results and refer to the aforementioned monographs for others.

Lemma S.4. *Let $m \leq n_*$ be a positive integer. Then*

(i) $\hat{Q}_m$ has m distinct positive real roots $\hat{\theta}_1 > \hat{\theta}_2 > ... > \hat{\theta}_m > 0$.

(ii) $\hat{Q}_m$ is positive, decreasing, and convex on $[0, \hat{\theta}_m]$.

(iii) $(\hat{Q}_l)_{l=0}^{n_*}$ *are orthogonal with respect to* $[.,.]_1$.

(iv) $|\hat{Q}'_m(0)|^{-1} \leq \hat{\theta}_m$.

(v) $\hat{Q}_m(\lambda) = \hat{Q}_{m-1}(\lambda)(1 - \lambda/\hat{\theta}_m)$.

(vi) $\sup_{\lambda \in [0,\hat{\theta}_m]} \lambda^\delta \hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m - \lambda)} \leq (2\delta)^\delta |\hat{Q}'_m(0)|^{-\delta}$ *for every* $\delta \geq 0$ *with* $0^0 := 1$.

Proof of Lemma S.4. (i) is known in the theory of orthogonal polynomials; see Engl et al. (1996), Appendix A.2. For (iii) and (vi), see Engl et al. (1996), Corollary 7.4, and equation (7.8).

Note that since $\hat{Q}_m(0) = 1$, we can write

$$\hat{Q}_m(\lambda) = \prod_{j=1}^{m} \left(1 - \frac{\lambda}{\hat{\theta}_j}\right).$$

This equation implies (v). Moreover, for all $\lambda \in [0, \hat{\theta}_m]$, we have $\hat{Q}_m(\lambda) \geq 0$ and by (i)

$$\hat{Q}'_m(\lambda) = -\sum_{k=1}^{m} \frac{1}{\hat{\theta}_k} \prod_{j \neq k} \left(1 - \frac{\lambda}{\hat{\theta}_j}\right) \leq 0.$$

We also have $\hat{Q}''_m(\lambda) \geq 0$ for all $\lambda \in [0, \hat{\theta}_m]$ which proves (ii).

(iv) follows from (i) and

$$|\hat{Q}'_m(0)| = \sum_{k=1}^{m} \frac{1}{\hat{\theta}_k} \geq \frac{1}{\hat{\theta}_m}.$$

□

The following lemma allows us to control the residual norm of the PLS estimator.

Lemma S.5. *Suppose that Assumptions 1, 2, and 3 are satisfied. Then for every* $1 \leq m \leq n_*$ *and* $\gamma \in [1/n, 1)$

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \lesssim \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + |\hat{Q}'_m(0)|^{-1}\sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}} + |\hat{Q}'_m(0)|^{-(\mu+1)}$$

on an event with probability at least $1 - \gamma$.

Proof of Lemma S.5. Let $\varphi_m(\lambda) := \hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m - \lambda)}$. We will first show that the following inequality holds

$$\begin{aligned}\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\| \\ &\leq \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{r}\right\|,\end{aligned}$$

where the first line uses $\hat{\beta}_m = \hat{P}_m(\hat{K})\hat{r}$ and $\hat{Q}_m(\lambda) = 1 - \lambda\hat{P}_m(\lambda)$. The inequality can be deduced from the proof of Theorem 7.9 in Engl et al. (1996). For completeness, we provide an argument suitably tailored to our setting below. By Lemma S.4 (iii) and (v), since the polynomials $(\hat{Q}_m)_{m\geq 0}$ are orthogonal with respect to $[.,.]_1$ (see Equation (S.1)) we get for $m \geq 1$

$$\begin{aligned}0 &= \int_0^{\hat{\lambda}_1} \hat{Q}_m(\lambda)\hat{Q}_{m-1}(\lambda)\mathrm{d}\hat{\mu}_1(\lambda) \\ &= \hat{\theta}_m \int_0^{\hat{\lambda}_1} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \hat{\theta}_m \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) + \hat{\theta}_m \int_{\hat{\theta}_m}^{\hat{\lambda}_1} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda).\end{aligned}$$

Since $\hat{\theta}_m > 0$, by Lemma S.4 (i), this shows that

$$\int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) = \int_{\hat{\theta}_m}^{\hat{\lambda}_1} \hat{Q}_m^2(\lambda)\frac{\lambda}{\lambda - \hat{\theta}_m}\mathrm{d}\hat{\mu}_0(\lambda)$$

and so by Equation (S.3)

$$\begin{aligned}\left\|\hat{Q}_m(\hat{K})\hat{r}\right\|^2 &= \int_0^{\infty} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_{\hat{\theta}_m}^{\infty} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) \\ &\leq \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_{\hat{\theta}_m}^{\infty} \hat{Q}_m^2(\lambda)\frac{\lambda}{\lambda - \hat{\theta}_m}\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\hat{\theta}_m}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{r}\right\|^2\end{aligned}$$

where the third line follows since $1 \leq \lambda/(\lambda - \hat{\theta}_m)$ for all $\lambda \geq \hat{\theta}_m$.

Therefore,

$$\begin{aligned}\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &\leq \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{r}\right\| \\ &\leq \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\beta\right\| + \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})(\hat{r} - \hat{K}\beta)\right\|.\end{aligned}$$

Under Assumption 3, $\beta = K^\mu w$ with $\|w\| \leq C$, so that

$$\begin{aligned}\left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\beta\right\| &= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}K^\mu w\right\| \\ &= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\hat{K}^\mu w\right\| + \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\left[K^\mu - \hat{K}^\mu\right]w\right\| \\ &\leq \sup_{\lambda\in[0,\hat{\theta}_m]}\left|\varphi_m(\lambda)\lambda^{1+\mu}\right| C + \sup_{\lambda\in[0,\hat{\theta}_m]}|\varphi_m(\lambda)\lambda|\left\|\hat{K}^\mu - K^\mu\right\|_{\text{op}} C \\ &\leq (2\mu+2)^{\mu+1}|\hat{Q}'_m(0)|^{-(\mu+1)}C \\ &\qquad + 2|\hat{Q}'_m(0)|^{-1}(c_\mu \mathbf{1}_{\mu\leq 1} + \mu\nu^{\mu-1}\mathbf{1}_{\mu>1})\left(\frac{2\mathbf{E}\|X\|^4}{\gamma n}\right)^{\frac{\mu\wedge 1}{2}} C,\end{aligned}$$

where the last line follows by Lemma S.4 (vi) and the inequality in Equation (S.4) on an event with probability at least $1-\gamma$. By Lemma S.2

$$\begin{aligned}\nu = \|\hat{K}\|_{\text{op}} \vee \|K\|_{\text{op}} &\leq \|K\|_{\text{op}} + \left\|\hat{K} - K\right\|_{\text{op}} \\ &\leq \lambda_1 + \sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}} \lesssim 1\end{aligned} \tag{S.5}$$

on an event with probability at least $1-\gamma$.

Lastly,

$$\begin{aligned}\left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})(\hat{r} - \hat{K}\beta)\right\| &\leq \sup_{\lambda\in[0,\hat{\theta}_m]}|\varphi_m(\lambda)|\left\|\hat{r} - \hat{K}\beta\right\| \\ &= \sup_{\lambda\in[0,\hat{\theta}_m]}\left|\hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m - \lambda)}\right|\left\|\hat{r} - \hat{K}\beta\right\| \\ &\leq \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}},\end{aligned}$$

where the last line follows by Lemma S.4 (vi) with $\delta = 0$ and Lemma S.2. □

The next lemma provides an upper bound for the derivative of the residual polynomial of degree selected by the stopping rule in Assumption 4 with some fixed $\delta \in (0,1)$.

Lemma S.6. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq 1/n$. Then*

$$|\hat{Q}'_{\hat{m}}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$$

on an event with probability at least $1-\delta$.

Proof of Lemma S.6. We have

$$|\hat{Q}'_{\hat{m}}(0)| \leq |\hat{Q}'_{\hat{m}-1}(0)| + |\hat{Q}'_{\hat{m}}(0) - \hat{Q}'_{\hat{m}-1}(0)|,$$

where each of the two terms will be bounded separately.

By virtue of Assumption 4

$$\begin{aligned}\tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} &\leq \left\|\hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1}\right\| \\ &\leq c\left\{\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + |\hat{Q}'_{\hat{m}-1}(0)|^{-1}\sqrt{\frac{2\mathbf{E}\|X\|^4}{\delta n}} + |\hat{Q}'_{\hat{m}-1}(0)|^{-(\mu+1)}\right\},\end{aligned}$$

where the second line follows by Lemma S.5 for some $c > 0$.

Therefore,

$$(\tau - c)\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq c\max\left\{|\hat{Q}'_{\hat{m}-1}(0)|^{-1}\sqrt{\frac{1}{\delta n}}, |\hat{Q}'_{\hat{m}-1}(0)|^{-(\mu+1)}\right\}.$$

If the first term inside the maximum is larger, then $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim 1$ while if the second term is larger, then $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$ provided that $\tau > c$. Therefore, we always have $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$.

For the second term, by Hanke (1995), Corollary 2.6, for every $1 \leq m \leq n_*$

$$0 \leq \hat{Q}'_{m-1}(0) - \hat{Q}'_m(0) = \frac{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0 - [\hat{Q}_m, \hat{Q}_m]_0}{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1} \leq \frac{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0}{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1}, \tag{S.6}$$

where $(\hat{Q}^{[2]}_l)_{l\geq 0}$ are the polynomials orthogonal with respect to $[.,.]_2$ and constant equal to 1; see Equation (S.1).

Take $a \in (0, \hat{\theta}_{m-1}]$ and let $\hat{K}^+ = \sum_{j=1}^{n_*} \hat{v}_j \otimes \hat{v}_j / \hat{\lambda}_j$ be the generalized inverse of $\hat{K}$. Then

$$\begin{aligned}
\sqrt{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0} &= \left\| \hat{Q}_{m-1}(\hat{K})\hat{r} \right\| \leq \left\| \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r} \right\| \\
&\leq \left\| \Pi_a \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r} \right\| + \left\| \Pi_a^{\perp} \sqrt{\hat{K}^+} \hat{K}^{1/2} \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r} \right\| \\
&\leq \left\| \Pi_a \hat{Q}^{[2]}_{m-1}(\hat{K}) \right\|_{\text{op}} \|\Pi_a \hat{r}\| + \left\| \Pi_a^{\perp} \sqrt{\hat{K}^+} \right\|_{\text{op}} \left\| \hat{K}^{1/2} \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r} \right\| \\
&\leq \sup_{\lambda \in [0,a]} \left| \hat{Q}^{[2]}_{m-1}(\lambda) \right| \|\Pi_a \hat{r}\| + \sup_{\lambda \geq a} \frac{1}{\sqrt{\lambda}} \left\| \hat{K}^{1/2} \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r} \right\| \\
&\leq \|\Pi_a \hat{r}\| + \sqrt{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1 / a},
\end{aligned}$$

where the second line holds since $\hat{Q}_{m-1}$ solves the problem in Equation (S.2) over $\mathcal{P}^0_{m-1}$, and the last line since $|\hat{Q}^{[2]}_{m-1}(\lambda)| \leq 1, \forall \lambda \in [0, a]$; see the proof of Lemma S.4.

Next, under Assumption 3, $\beta = K^{\mu} w$ with $\|w\| \leq C$, so that

$$\begin{aligned}
\|\Pi_a \hat{r}\| &\leq \left\| \Pi_a (\hat{r} - \hat{K}\beta) \right\| + \left\| \Pi_a \hat{K} \beta \right\| \\
&\leq \left\| \hat{r} - \hat{K}\beta \right\| + \left\| \Pi_a \hat{K} \hat{K}^{\mu} w \right\| + \left\| \Pi_a \hat{K} (\hat{K}^{\mu} - K^{\mu}) w \right\| \\
&\leq \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \sup_{\lambda \in [0,a]} \lambda^{1+\mu} C + a \left\| \hat{K}^{\mu} - K^{\mu} \right\|_{\text{op}} C \\
&\lesssim \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + a^{\mu+1} + a \left(\frac{1}{\delta n} \right)^{\frac{\mu \wedge 1}{2}},
\end{aligned}$$

where we use the inequality in Equation (S.4) with $\gamma = \delta$. Take $a = (c_1 \sigma \sqrt{2\mathbf{E}\|X\|^2 / \delta n})^{1/(\mu+1)}$ with a sufficiently small $c_1 > 0$, so that $a \leq |\hat{Q}'_{\hat{m}-1}(0)|^{-1} \leq \hat{\theta}_{\hat{m}-1}$, cf. Lemma S.4 (iv). Such a constant exists since as we've already shown $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim (n\delta)^{\frac{1}{2(\mu+1)}}$. Then for some $c_3 > 0$

$$\begin{aligned}
\sqrt{[\hat{Q}_{\hat{m}-1}, \hat{Q}_{\hat{m}-1}]_0} &\leq c_3 \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \sqrt{[\hat{Q}^{[2]}_{\hat{m}-1}, \hat{Q}^{[2]}_{\hat{m}-1}]_1 / a} \\
&\leq \frac{c_3}{\tau} \left\| \hat{r} - \hat{K} \hat{\beta}_{\hat{m}-1} \right\| + \sqrt{[\hat{Q}^{[2]}_{\hat{m}-1}, \hat{Q}^{[2]}_{\hat{m}-1}]_1 / a} \\
&= \frac{c_3}{\tau} \sqrt{[\hat{Q}_{\hat{m}-1}, \hat{Q}_{\hat{m}-1}]_0} + \sqrt{[\hat{Q}^{[2]}_{\hat{m}-1}, \hat{Q}^{[2]}_{\hat{m}-1}]_1 / a},
\end{aligned}$$

where we use Assumption 4 and Equation (S.3). If τ is selected so that $\tau > c_3$ in Assumption 4, then

$$[\hat{Q}_{\hat{m}-1}, \hat{Q}_{\hat{m}-1}]_0 \leq \left(\frac{\tau}{\tau - c_3} \right)^2 [\hat{Q}^{[2]}_{\hat{m}-1}, \hat{Q}^{[2]}_{\hat{m}-1}]_1 / a.$$

Plugging this into Equation (S.6) and with our choice of a, we get

$$\left|\hat{Q}'_m(0) - \hat{Q}'_{m-1}(0)\right| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}.$$

□

S.2 Proofs of Main Results

In this section, we provide detailed proofs of theorems.

Proof of Theorem 1. Take any $m \leq n_*$, $\gamma \in (0,1)$, and let $a > 0$ be such that $a \leq |\hat{Q}'_m(0)|^{-1}$. By Lemma S.4 (iv) this ensures that $a \leq \hat{\theta}_m$ which we use repeatedly in the proof.

Decompose

$$\hat{\beta}_m - \beta = \Pi_a \hat{P}_m(\hat{K})(\hat{r} - \hat{K}\beta) + \Pi_a \left[\hat{P}_m(\hat{K})\hat{K} - I\right]\beta + \Pi_a^{\perp}(\hat{\beta}_m - \beta),$$

where $\Pi_a = \sum_{j:\hat{\lambda}_j \leq a} \hat{v}_j \otimes \hat{v}_j$ and $\Pi_a^{\perp} = I - \Pi_a$. Then for $s \in [0,1]$, we have

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| &\leq \left\|\Pi_a \hat{K}^s \hat{P}_m(\hat{K})(\hat{r} - \hat{K}\beta)\right\| + \left\|\Pi_a \hat{K}^s \hat{Q}_m(\hat{K})\beta\right\| + \left\|\Pi_a^{\perp}\hat{K}^s(\hat{\beta}_m - \beta)\right\| \\ &=: I + II + III.\end{aligned}$$

We will derive an upper bound for each of these three terms separately. For the first term, note that for every $s \in [0,1]$,

$$\begin{aligned} I &\leq \left\|\Pi_a \hat{K}^s \hat{P}_m(\hat{K})\right\| \left\|\hat{r} - \hat{K}\beta\right\| \\ &\leq \sup_{\lambda \in [0,a]} |\lambda^s \hat{P}_m(\lambda)| \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} \\ &\leq a^s |\hat{Q}'_m(0)| \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}}, \end{aligned}$$

where the second line follows on an event with probability at least $1 - \gamma$ by Lemma S.2 and Equation (5), and the last one by the convexity of $\hat{Q}_m$ on $[0, a]$:

$$\hat{P}_m(\lambda) = \frac{1 - \hat{Q}_m(\lambda)}{\lambda} \leq -\hat{Q}'_m(0),$$

and $\hat{Q}_m(\lambda) \leq \hat{Q}_m(0) = 1$ for every $\lambda \in [0, a]$; see Lemma S.4 (ii).

For the second term, under Assumption 3, we have $\beta = K^\mu w$ with $\|w\| \leq C$, so that

$$\begin{aligned} II &= \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) K^\mu w \right\| \\ &\leq \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) \hat{K}^\mu w \right\| + \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) \left[K^\mu - \hat{K}^\mu \right] w \right\| \\ &\leq \sup_{\lambda \in [0,a]} |\lambda^{\mu+s} \hat{Q}_m(\lambda)| C + \sup_{\lambda \in [0,a]} |\lambda^s \hat{Q}_m(\lambda)| \left\| K^\mu - \hat{K}^\mu \right\|_{\text{op}} C \\ &\leq a^{\mu+s} C + a^s (c_\mu \mathbf{1}_{\mu \leq 1} + \mu \nu^{\mu-1} \mathbf{1}_{\mu > 1}) \left(\frac{2\mathbf{E}\|X\|^4}{\gamma n} \right)^{\frac{\mu \wedge 1}{2}} C, \end{aligned}$$

where the last line follows since $|\hat{Q}_m(\lambda)| \leq 1$ by Lemma S.4 (ii) and Equation (S.4) on an event with probability at least $1 - \gamma$.

Lastly, let $\hat{K}^+ = \sum_{j=1}^{n_*} \hat{v}_j \otimes \hat{v}_j / \hat{\lambda}_j$ be the generalized inverse of $\hat{K}$. Then we bound the third term as follows

$$\begin{aligned} III &\leq \left\| \Pi_a^\perp \hat{K}^s \hat{K}^+ \right\| \left\| \hat{K}(\hat{\beta}_m - \beta) \right\| \\ &\leq \sup_{\lambda \geq a} \lambda^{s-1} \left\| (\hat{K}\hat{\beta}_m - \hat{r}) + (\hat{r} - \hat{K}\beta) \right\| \\ &\leq a^{s-1} \left\{ \left\| \hat{K}\hat{\beta}_m - \hat{r} \right\| + \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} \right\}, \end{aligned}$$

where we use Lemma S.2 on an event with probability at least $1 - \gamma$. Combining the three bounds, we obtain for $m \leq n_*$

$$\left\| \hat{K}^s(\hat{\beta}_m - \beta) \right\| \lesssim a^{s-1} \left\{ \left\| \hat{K}\hat{\beta}_m - \hat{r} \right\| + (\gamma n)^{-1/2} \right\} + a^{\mu+s} + a^s (\gamma n)^{-\frac{\mu \wedge 1}{2}} + a^s |\hat{Q}'_m(0)| (\gamma n)^{-1/2}, \tag{S.7}$$

where we use $\nu \lesssim 1$; see Equation (S.5). Taking $a = |\hat{Q}'_m(0)|^{-1}$, this gives

$$\left\| \hat{K}^s(\hat{\beta}_m - \beta) \right\| = O_P \left(|\hat{Q}'_m(0)|^{-s+1} \left\{ \left\| \hat{K}\hat{\beta}_m - \hat{r} \right\| + n^{-1/2} \right\} + |\hat{Q}'_m(0)|^{-(\mu+s)} + |\hat{Q}'_m(0)|^{-s} n^{-\frac{\mu \wedge 1}{2}} \right).$$

By Lemma S.5

$$\left\| \hat{K}\hat{\beta}_m - \hat{r} \right\| = O_P \left(n^{-1/2} + |\hat{Q}'_m(0)|^{-1} n^{-1/2} + |\hat{Q}'_m(0)|^{-(\mu+1)} \right).$$

Therefore,

$$\left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| = O_P\left(|\hat{Q}'_m(0)|^{-s+1}n^{-1/2} + |\hat{Q}'_m(0)|^{-(\mu+s)} + |\hat{Q}'_m(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right), \qquad \forall s \in [0,1].$$

This proves the result if $s = 0$. If $s \in (0,1]$, then

$$\begin{aligned}
\left\|K^s(\hat{\beta}_m - \beta)\right\| &= \left\|\hat{K}^s(\hat{\beta}_m - \beta) + (K^s - \hat{K}^s)(\hat{\beta}_m - \beta)\right\| \\
&\leq \left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| + \left\|\hat{K}^s - K^s\right\|_{\text{op}}\left\|\hat{\beta}_m - \beta\right\| \\
&= O_P\left(|\hat{Q}'_m(0)|^{-s+1}n^{-1/2} + |\hat{Q}'_m(0)|^{-(\mu+s)} + |\hat{Q}'_m(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right) \\
&\quad + O_P\left(|\hat{Q}'_m(0)|n^{-\frac{1+s}{2}} + |\hat{Q}'_m(0)|^{-\mu}n^{-\frac{s}{2}} + n^{-\frac{s+\mu\wedge 1}{2}}\right) \\
&= O_P\left(|\hat{Q}'_m(0)|^{-s+1}n^{-1/2} + |\hat{Q}'_m(0)|^{-(\mu+s)} + |\hat{Q}'_m(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right),
\end{aligned}$$

provided that $|\hat{Q}'_m(0)| = O_P(n^{1/2})$. □

Proof of Theorem 2. We adopt an approach similar to Cai and Yuan (2012), Theorem 1; see also Tsybakov (2009), Chapter 2. Recall that the lower bound for a restricted class of models yields the lower bound for the general case. Therefore, we can assume without loss of generality that $\varepsilon_i|X_i \sim N(0,\sigma^2)$ and $K : \mathbb{H} \to \mathbb{H}$ has a spectral decomposition $(\lambda_j, v_j)_{j\geq 1}$ with $\lambda_1 = 1$ and $\lambda_j = 1/(j\log^a j)$ for $j = 2, 3, \ldots$ for some $a > 1$. We will also consider the family of slope parameters

$$\beta_\theta = Cm^{-1/2}\sum_{l=m+1}^{2m}\theta_l\lambda_l^\mu v_l, \qquad \theta = (\theta_{m+1}, \ldots, \theta_{2m}) \in \{0,1\}^m$$

for some $C > 0$ and m specified below. It is easy to see that by the orthonormality of $(v_l)_{l\geq 1}$

$$\sum_{j=1}^{\infty}\frac{\langle\beta_\theta, v_j\rangle^2}{\lambda_j^{2\mu}} = \frac{C^2}{m}\sum_{j=m+1}^{2m}\theta_j^2 \leq C^2$$

and that

$$\sum_{j=1}^{\infty}\lambda_j = 1 + \sum_{j=2}^{\infty}\frac{1}{j\log^a j} \leq C$$

for some $C > 0$. Therefore, $(\beta_\theta, K) \in \mathcal{S}(\mu, C), \forall \theta \in \{0,1\}^m$, cf. Assumption 3.

Let $H(\theta, \theta') = \sum_{j=1}^m \mathbf{1}\{\theta_j \neq \theta'_j\}$ be the Hamming distance between the binary sequences $\theta, \theta' \in \{0,1\}^m$. By the Varshamov-Gilbert bound, see Tsybakov (2009), Lemma 2.9, if $m \geq 8$, there exists $\{\theta^{(0)}, \ldots, \theta^{(M)}\} \subset \{0,1\}^m$ such that

(a) $\theta^{(0)} = (0, \ldots, 0)$;

(b) $H(\theta^{(j)}, \theta^{(k)}) \geq \frac{m}{8}, \forall\, 0 \leq j < k \leq M$;

(c) $M \geq 2^{m/8}$.

For every $A > 0$,

$$\begin{aligned} &\sup_{(\beta,K)\in\mathcal{S}(\mu,C)} \Pr\left(\left\|K^s(\hat{\beta} - \beta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}\right) \\ &\geq \max_{\theta\in\{\theta^{(0)},\ldots,\theta^{(M)}\}} \Pr\left(\left\|K^s(\hat{\beta} - \beta_\theta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}\right). \end{aligned} \tag{S.8}$$

To obtain the lower bound for the right-hand side of the Equation (S.8), we will use Tsybakov (2009), Theorem 2.5 and a specific choice of $A < \infty$. To that end, we need to check the following conditions:

(i) $\|K^s(\beta_{\theta^{(j)}} - \beta_{\theta^{(k)}})\| \geq 2An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}$ for all $0 \leq j < k \leq M$;

(ii) $P_j << P_0, \forall j = 1, \ldots, M$, where P_j denotes the distribution of $(Y_i, X_i)_{i\geq 1}$ for the slope parameter $\beta_{\theta^{(j)}}$;

(iii) For $\alpha \in (0, 1/8)$,

$$\frac{1}{M}\sum_{j=1}^M KL(P_j, P_0) \leq \alpha \log M,$$

where KL is the Kullback-Leibler divergence between P_j and P_0.

For the first condition, note that since $H(\theta^{(j)}, \theta^{(k)}) = \sum_{l=m+1}^{2m} (\theta_l^{(j)} - \theta_l^{(k)})^2$, we have

$$\begin{aligned}
\|K^s(\beta_{\theta^{(j)}} - \beta_{\theta^{(k)}})\|^2 &= \left\| Cm^{-1/2} \sum_{l=m+1}^{2m} (\theta_l^{(j)} - \theta_l^{(k)}) \lambda_l^{\mu+s} v_l \right\|^2 \\
&= \frac{C^2}{m} \sum_{l=m+1}^{2m} (\theta_l^{(j)} - \theta_l^{(k)})^2 \lambda_l^{2(\mu+s)} \\
&\geq \frac{C^2}{m} \lambda_{2m}^{2(\mu+s)} H(\theta^{(j)}, \theta^{(k)}) \\
&\geq \frac{C^2}{8} \lambda_{2m}^{2(\mu+s)} = \frac{C^2}{8} (2m)^{-2(\mu+s)} \log^{-2a(\mu+s)}(2m) \\
&\geq 4A^2 n^{-\frac{\mu+s}{\mu+1}} \log^{-2a(\mu+s)} n
\end{aligned}$$

where the last two inequalities follow from (b) provided that $m \leq n^{\frac{1}{2(\mu+1)}}$ for some $A > 0$. This verifies (i).

Next, since $Y_i | X_i \sim N(\langle X_i, \beta_{\theta^{(j)}} \rangle, \sigma^2)$ under P_j, we have $P_j << P_0, \forall j = 1, \ldots, M$ with the log-likelihood ratio

$$\log \frac{\mathrm{d}P_j}{\mathrm{d}P_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \langle X_i, \beta_{\theta^{(j)}} \rangle \right) \langle X_i, \beta_{\theta^{(j)}} - \beta_{\theta^{(0)}} \rangle + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \langle X_i, \beta_{\theta^{(j)}} - \beta_{\theta^{(0)}} \rangle^2 .$$

To verify (iii), we compute the Kullback–Leibler divergence:

$$\begin{aligned}
KL(P_j, P_0) &= \int \log \frac{\mathrm{d}P_j}{\mathrm{d}P_0} \mathrm{d}P_j \\
&= \frac{n}{2\sigma^2} \mathbf{E} \langle X_i, \beta_{\theta^{(0)}} - \beta_{\theta^{(j)}} \rangle^2 \\
&= \frac{n}{2\sigma^2} \left\| K^{1/2} (\beta_{\theta^{(0)}} - \beta_{\theta^{(j)}}) \right\|^2 \\
&= \frac{n}{2\sigma^2} \left\| Cm^{-1/2} \sum_{l=m+1}^{2m} \theta_l^{(j)} \lambda_l^{\mu+1/2} v_l \right\|^2 \\
&= \frac{nC^2}{2\sigma^2 m} \sum_{l=m+1}^{2m} \left(\theta_l^{(j)} \right)^2 \lambda_l^{2\mu+1} \\
&\leq \frac{nC^2}{2\sigma^2} \lambda_m^{2\mu+1} = \frac{nC^2}{2\sigma^2} m^{-(2\mu+1)} (\log m)^{-a(2\mu+1)} \\
&\leq \alpha \frac{m}{8} \log 2 = \alpha \log 2^{m/8},
\end{aligned}$$

provided that $m \geq (c_0 n)^{\frac{1}{2(\mu+1)}} (\log m)^{-a\frac{2\mu+1}{2(\mu+1)}}$ with $c_0 = 4C^2/(\sigma^2 \alpha \log 2)$ some $\alpha \in (0, 1/8)$.[1]

[1] To ensure that this constraint holds and that $m \leq n^{\frac{1}{2(\mu+1)}}$, we can take m as a fraction of $n^{\frac{1}{2(\mu+1)}}$.

This verifies (iii) in light of (c).

Therefore, by Tsybakov (2009), Theorem 2.5

$$\liminf_{n\to\infty}\inf_{\hat{\beta}}\max_{\theta\in\{\theta^{(0)},\ldots,\theta^{(M)}\}}\Pr\left(\left\|K^s(\hat{\beta}-\beta_\theta)\right\|\geq An^{-\frac{\mu+s}{2(\mu+1)}}\log^{-a(\mu+s)}n\right)\geq 1-2\alpha>0$$

which implies the result in light of the inequality (S.8). □

Proof of Theorem 3. Setting $m=\hat{m}$ and $\gamma=\delta$ in Equation (S.7), under Assumption 4, we obtain

$$\left\|\hat{K}^s(\hat{\beta}_{\hat{m}}-\beta)\right\|\lesssim a^{s-1}(\delta n)^{-1/2}+a^{\mu+s}+a^s(\delta n)^{-\frac{\mu\wedge 1}{2}}+a^s|\hat{Q}'_{\hat{m}}(0)|(\delta n)^{-1/2} \tag{S.9}$$

for every $a\leq|\hat{Q}'_{\hat{m}}(0)|^{-1}$. Now we will choose the truncation level a. Suppose that $s\in[0,1)$. Then the function $a\mapsto a^{s-1}(\delta n)^{-1/2}+a^{\mu+s}$ is minimized for $a^*=\left\{(\delta n)^{1/2}(\mu+s)/(1-s)\right\}^{-\frac{1}{\mu+1}}$. If $a^*\leq|\hat{Q}'_{\hat{m}}(0)|^{-1}$, we shall choose $a=a^*$, in which case since $\delta\geq 1/n$, we obtain

$$\left\|\hat{K}^s(\hat{\beta}_{\hat{m}}-\beta)\right\|\lesssim(\delta n)^{-\frac{\mu+s}{2(\mu+1)}}+(\delta n)^{-\frac{s+\mu+1}{2(\mu+1)}}|\hat{Q}'_{\hat{m}}(0)|\lesssim(\delta n)^{-\frac{\mu+s}{2(\mu+1)}}.$$

On the other hand, if $a^*>|\hat{Q}'_{\hat{m}}(0)|^{-1}$, we shall choose $a=|\hat{Q}'_{\hat{m}}(0)|^{-1}$. Then

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_{\hat{m}}-\beta)\right\|&\lesssim|\hat{Q}'_{\hat{m}}(0)|^{1-s}(\delta n)^{-1/2}+|\hat{Q}'_{\hat{m}}(0)|^{-(\mu+s)}+|\hat{Q}'_{\hat{m}}(0)|^{-s}(\delta n)^{-\frac{\mu\wedge 1}{2}}\\&\lesssim(\delta n)^{-\frac{\mu+s}{2(\mu+1)}},\end{aligned}$$

where the last line follows from $|\hat{Q}'_{\hat{m}}(0)|\lesssim(\delta n)^{\frac{1}{2(\mu+1)}}$ by Lemma S.6, and from $|\hat{Q}'_{\hat{m}}(0)|^{-1}<a^*\lesssim(\delta n)^{-\frac{1}{2(\mu+1)}}$ and $(\delta n)^{-1}\leq 1$.

If $s=1$, then setting $a=c(\delta n)^{-\frac{1}{2(\mu+1)}}$ in Equation (S.9), for some $c>0$ such that $a\leq|\hat{Q}'_{\hat{m}}(0)|^{-1}$, cf. Lemma S.6, we get

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_{\hat{m}}-\beta)\right\|&\lesssim(\delta n)^{-1/2}+(\delta n)^{-\frac{1+(\mu\wedge 1)(\mu+1)}{2(\mu+1)}}+(\delta n)^{-\frac{\mu+2}{2(\mu+1)}}|\hat{Q}'_{\hat{m}}(0)|\\&\lesssim(\delta n)^{-1/2},\end{aligned}$$

where the last line follows from Lemma S.6 and $(\delta n)^{-1} \leq 1$. Therefore, we've just established for every $s \in [0,1]$

$$\left\| \hat{K}^s (\hat{\beta}_{\hat{m}} - \beta) \right\| \lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}}. \tag{S.10}$$

This proves the statement of the theorem in the special case when $s = 0$. On the other hand, if $s \in (0,1]$, we have

$$\begin{aligned}
\left\| K^s (\hat{\beta}_{\hat{m}} - \beta) \right\| &\leq \left\| \hat{K}^s (\hat{\beta}_{\hat{m}} - \beta) \right\| + \left\| \hat{K}^s - K^s \right\|_{\text{op}} \left\| \hat{\beta}_{\hat{m}} - \beta \right\| \\
&\lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}} + (\delta n)^{-\frac{s \wedge 1}{2}} (\delta n)^{-\frac{\mu}{2(\mu+1)}} \\
&\lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}},
\end{aligned}$$

where we use Equation (S.10) and (S.4), and $(\delta n)^{-1} \leq 1$. □

Proof of Theorem 4. For $1 \leq m \leq n_*$ and $\nu \geq 0$, let $\tilde{P}_m^{(\nu)}$ be a Jacobi polynomial of degree m on $[-1,1]$, i.e. a polynomial that is orthogonal with respect to the weight function $\lambda \mapsto (1-\lambda)^\alpha (1+\lambda)^\beta$, where we set $\alpha = -1/2$ and $\beta = 2\nu - 1/2$ with $\nu > 0$. Let $P_m^{(\nu)}(\lambda) = \tilde{P}_m^{(\nu)}(2\lambda/\hat{\lambda}_1 - 1)/\tilde{P}_m^{(\nu)}(-1), \forall \lambda \in [0, \hat{\lambda}_1]$ be a shifted Jacobi polynomial, normalized so that $P_m^{(\nu)}(0) = 1$. By Engl et al. (1996), Appendix A.2, p.294, there exists $c_\nu > 0$ such that

$$|P_m^{(\nu)}(\lambda)| \leq c_\nu (1 + m^2 \lambda)^{-\nu}, \qquad \forall \lambda \in [0, \hat{\lambda}_1],\ m \geq 0. \tag{S.11}$$

Recall also that under Assumption 3, we have $\beta = K^\mu w$ with $\|w\| \leq C$. Then

$$\begin{aligned}
\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})\hat{r}\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})(\hat{r} - \hat{K}\beta)\right\| + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}\beta\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})\right\|_{\mathrm{op}} \left\|\hat{r} - \hat{K}\beta\right\| + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}\hat{K}^\mu w\right\| \\
&\qquad + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}(\hat{K}^\mu - K^\mu)w\right\| \\
&\lesssim \sup_{\lambda\in[0,\hat{\lambda}_1]} |P_m^{(\mu+1)}(\lambda)| \left\|\hat{r} - \hat{K}\beta\right\| + \sup_{\lambda\in[0,\hat{\lambda}_1]} \left|\lambda^{\mu+1} P_m^{(\mu+1)}(\lambda)\right| \\
&\qquad + \sup_{\lambda\in[0,\hat{\lambda}_1]} |\lambda P_m^{(\mu+1)}(\lambda)| \left\|\hat{K}^\mu - K^\mu\right\|_{\mathrm{op}} \\
&\lesssim \left\|\hat{r} - \hat{K}\beta\right\| + m^{-2(\mu+1)} + \hat{\lambda}_1 \left\|\hat{K}^\mu - K^\mu\right\|_{\mathrm{op}} \\
&\lesssim \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + m^{-2(\mu+1)} + \|\hat{K}\|_{\mathrm{op}} \left(\frac{2\mathbf{E}\|X\|^4}{\gamma n}\right)^{\frac{\mu\wedge 1}{2}},
\end{aligned}$$

where the second line follows since $\hat{Q}_m$ minimizes the problem in Equation (S.2); the sixth from the inequality (S.11); and the last by Lemma S.2 and the inequality (S.4) with probability at least $1 - \gamma$ for every $\gamma \in (0, 1)$.

Therefore, if $\mu \geq 1$, under Assumption 1 by Lemma S.2 and the inequality (S.4), we obtain

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq c\left\{\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + m^{-2(\mu+1)}\right\}$$

for some $c > 0$ and $\delta \geq 1/n$ on the event with probability at least $1 - \delta$. According to the stopping rule in the Assumption 4, we also know that

$$\tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq \left\|\hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1}\right\|$$

Therefore,

$$(\tau - c)\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \lesssim (\hat{m} - 1)^{-2(\mu+1)},$$

and whence $\hat{m} \lesssim (\delta n)^{\frac{1}{4(\mu+1)}}$, provided that $\tau > c$. □

Proof of Theorem 5. For some integers $m \geq k \geq 0$, put

$$G_m(\lambda) := \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\hat{\lambda}_j}\right) P_{m-k}^{(\mu+1)}\left(\frac{\lambda}{\hat{\lambda}_{k+1}}\right)$$

where $P_m^{(\nu)}(\lambda) = \tilde{P}_m^{(\nu)}(2\lambda/\hat{\lambda}_{k+1} - 1)/\tilde{P}_m^{(\nu)}(-1)$ is a shifted and normalized Jacobi polynomial on $[0, \hat{\lambda}_{k+1}]$, defined to be zero outside of this interval; see the proof of Theorem 4. Then G_m is an m-th degree polynomial with $G_m(0) = 1$ and

$$\sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} \left|\lambda^{\mu+1} G_m(\lambda)\right| \leq \sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} \left|\lambda^{\mu+1} P_{m-k}^{(\mu+1)}\left(\frac{\lambda}{\hat{\lambda}_{k+1}}\right)\right| \leq \hat{\lambda}_{k+1}^{\mu+1} c_\mu (m-k)^{-2(\mu+1)},$$

where we use the inequality (S.11). Then since under Assumption 3, $\beta = K^\mu w$ with $\|w\| \leq C$, we have with probability at least $1 - \gamma$ for every $\gamma \in (0, 1)$

$$\begin{aligned}
\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\| \leq \left\|G_m(\hat{K})\hat{r}\right\| \\
&\leq \left\|G_m(\hat{K})(\hat{r} - \hat{K}\beta)\right\| + \left\|G_m(\hat{K})\hat{K}\hat{K}^\mu w\right\| + \left\|G_m(\hat{K})\hat{K}(\hat{K}^\mu - K^\mu)w\right\| \\
&\leq \sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} |G_m(\lambda)| \left\|\hat{r} - \hat{K}\beta\right\| + C \sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} |\lambda^{\mu+1} G_m(\lambda)| \\
&\qquad + C \sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} |\lambda G_m(\lambda)| \left\|\hat{K}^\mu - K^\mu\right\|_{\mathrm{op}} \\
&\lesssim \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + \hat{\lambda}_{k+1}^{\mu+1}(m-k)^{-2(\mu+1)} + \|\hat{K}\|_{\mathrm{op}} \left(\frac{2\mathbf{E}\|X\|^2}{\gamma n}\right)^{\frac{\mu \wedge 1}{2}}
\end{aligned}$$

where the first inequality follows since $\hat{Q}_m$ solves the problem in Equation (S.2) and for the last inequality, we use $|G_m(\lambda)| \leq |P_{m-k}^{(\mu+1)}(\lambda/\hat{\lambda}_{k+1})| \lesssim 1$; see Equation (S.11). Recall that $\|\hat{K}\|_{\mathrm{op}} \lesssim 1$ on an event with probability at least $1 - \gamma$ for $\gamma \geq 1/n$; see Lemma S.2. Therefore, since $\mu \geq 1$, we obtain

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq c \left\{\sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \hat{\lambda}_{k+1}^{\mu+1}(\hat{m} - k)^{-2(\mu+1)}\right\}$$

for some $c > 0$ and $\delta \geq 1/n$ on an event with probability at least $1 - \delta$.

According to the stopping rule in the Assumption 4, we also know that for some $\delta \in (0, 1)$

$$\tau\sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq \left\|\hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1}\right\|.$$

Therefore, if $\tau > c$, we obtain

$$(\tau - c)\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq c\hat{\lambda}_{k+1}^{\mu+1}(\hat{m} - k - 1)^{-2(\mu+1)}$$

which implies that

$$\begin{aligned}
\hat{m} - k - 1 &\lesssim \hat{\lambda}_{k+1}^{1/2}(\delta n)^{\frac{1}{4(\mu+1)}} \\
&\leq (\delta n)^{\frac{1}{4(\mu+1)}} \left\{ \lambda_{k+1}^{1/2} + \left\| \hat{K}^{1/2} - K^{1/2} \right\|_{\mathrm{op}} \right\} \qquad \text{(S.12)} \\
&\lesssim (\delta n)^{\frac{1}{4(\mu+1)}} \left\{ \lambda_{k+1}^{1/2} + (\delta n)^{-1/4} \right\},
\end{aligned}$$

where we use Weyl's inequality and Equation (S.4).

Case (i): if $\lambda_k = O(k^{-2\kappa})$ with $\kappa > 0$, we can take $k \sim \hat{m}/2$. In this case Equation (S.12) implies

$$\hat{m} \lesssim (\delta n)^{\frac{1}{4(\mu+1)}} \hat{m}^{-\kappa} + (\delta n)^{-\frac{\mu}{4(\mu+1)}}.$$

If the second term in this upper bound dominates the first one, then $\hat{m} \lesssim (\delta n)^{-\frac{\mu}{4(\mu+1)}} = o(1)$, which is a contradiction. Therefore, $\hat{m} \lesssim (\delta n)^{\frac{1}{4(\kappa+1)(\mu+1)}}$.

Case (ii): if $\lambda_j = O(q^j)$ with $q \in (0, 1)$, we can take $k = \hat{m} - 2$. In this case Equation (S.12) implies

$$1 \lesssim (\delta n)^{\frac{1}{4(\mu+1)}} q^{(\hat{m}-1)/2} + (\delta n)^{-\frac{\mu}{4(\mu+1)}}.$$

If the second term in this upper bound dominates the first one, then $1 \lesssim (\delta n)^{-\frac{\mu}{4(\mu+1)}} = o(1)$, which is a contradiction. Therefore, $\hat{m} \lesssim 1 + \log(\delta n)$ since $\log q < 0$. □

Proof of Theorem 6. Observe that

$$\hat{K}(\hat{\beta}_m - b) = (\hat{r} - \hat{K}b) + (\hat{K}\hat{\beta}_m - \hat{r}).$$

Then

$$\begin{aligned} T_n &= n\left\|\hat{K}(\hat{\beta}_m - b)\right\|^2 \\ &= \left\langle \sqrt{n}\hat{K}(\hat{\beta}_m - b), \sqrt{n}\hat{K}(\hat{\beta}_m - b) \right\rangle \\ &= \|\sqrt{n}(\hat{r} - \hat{K}b)\|^2 + n\|\hat{K}\hat{\beta}_m - \hat{r}\|^2 + 2\left\langle \sqrt{n}(\hat{r} - \hat{K}b), \sqrt{n}(\hat{r} - \hat{K}\hat{\beta}_m) \right\rangle \\ &=: I_n + II_n + III_n. \end{aligned}$$

By assumption, $II_n = o_P(1)$. Under H_0, by the Hilbert space central limit theorem, see Bosq (2000), Theorem 2.7,

$$\sqrt{n}(\hat{r} - \hat{K}b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i X_i \xrightarrow{d} G,$$

where G is a zero-mean Gaussian element in $\mathbb{H}$ with covariance operator V. By the continuous mapping theorem, see Van Der Vaart and Wellner (1996), Theorem 1.3.6

$$I_n \xrightarrow{d} \|G\|^2,$$

By the Karhunen–Loève expansion, $G = \sum_{j=1}^{\infty} \omega_j^{1/2} Z_j \varphi_j$, so we obtain

$$I_n \xrightarrow{d} \sum_{j=1}^{\infty} \omega_j Z_j^2$$

under the null hypothesis. Note that this also shows that $\|\sqrt{n}(\hat{r} - \hat{K}\beta)\| = O_P(1)$. Then, under H_0

$$|III_n| \leq 2\|\sqrt{n}(\hat{r} - \hat{K}b)\| \cdot \sqrt{n}\|\hat{r} - \hat{K}\hat{\beta}_m\| = o_P(1),$$

since $\|\hat{r} - \hat{K}\hat{\beta}_m\| = o_P(n^{-1/2})$. This shows that under the null hypothesis, we have

$$T_n \xrightarrow{d} \sum_{j=1}^{\infty} \omega_j Z_j^2.$$

Next, under the fixed alternative hypothesis $\beta \neq b$, we have

$$\hat{K}(\hat{\beta}_m - b) = (\hat{r} - \hat{K}\beta) + \hat{K}(\beta - b) + (\hat{K}\hat{\beta}_m - \hat{r}).$$

Note that the last term is $o_P(1)$ under the maintained assumptions. On the other hand, by the Hilbert space law of large numbers, see Bosq (2000), Theorem 2.4,

$$\hat{r} - \hat{K}\beta = o_P(1) \qquad \text{and} \qquad \hat{K}(\beta - b) \xrightarrow{p} K(\beta - b).$$

This shows that

$$n^{-1}T_n \xrightarrow{p} \|K(\beta - b)\|^2 > 0$$

as long as $K : \mathbb{H} \to \mathbb{H}$ does not have zero eigenvalues. Therefore, by Slutsky's theorem $T_n \xrightarrow{\text{a.s.}} \infty$.

Lastly, under the local alternative hypothesis, we have

$$\begin{aligned}\sqrt{n}\hat{K}(\hat{\beta}_m - b) &= \sqrt{n}(\hat{r} - \hat{K}\beta) + \hat{K}\Delta + \sqrt{n}(\hat{K}\hat{\beta}_m - \hat{r}) \\ &\xrightarrow{d} G + K\Delta.\end{aligned}$$

This shows that

$$\begin{aligned}T_n &= n\big\|\hat{K}(\hat{\beta}_m - b)\big\|^2 \\ &\xrightarrow{d} \|G + K\Delta\|^2 \\ &= \|G\|^2 + 2\langle G, K\Delta\rangle + \|K\Delta\|^2 \\ &= \sum_{j=1}^{\infty} \omega_j Z_j^2 + 2\sum_{j=1}^{\infty} \omega_j^{1/2} Z_j \langle \varphi_j, K\Delta\rangle + \|K\Delta\|^2,\end{aligned}$$

where the last line follows by the Karhunen–Loève expansion, $G = \sum_{j=1}^{\infty} \omega_j^{1/2} Z_j \varphi_j$. □

Proof of Corollary 1. The first two statements follow trivially from Theorem 6. The last statement follows since by Markov's inequality

$$\lim_{n\to\infty} \Pr(T_n > z_{1-\alpha}) \leq \frac{\sum_{j=1}^{\infty} \omega_j + \|K\Delta\|^2}{z_{1-\alpha}},$$

where we use $Z_j \sim N(0,1)$. □

S.3 Comparison to PCA

In this section, we shed some light on the behavior of functional PLS relative to PCA. We will show that for the same fixed number of components m, PLS fits the empirical moment better than PCA, hence, it may require a smaller number of components to obtain a comparable fit. We also show that the regularization bias part of the estimation and prediction risk of PLS is smaller than the one of the PCA. Therefore, the adaptive PLS basis is better suited for approximating the slope function.

In what follows, we will use

$$\hat{\beta}_m^{\mathrm{PLS}} = \sum_{j=1}^{n_*} \hat{P}_m(\hat{\lambda}_j)\langle \hat{r}, \hat{v}_j\rangle \hat{v}_j \qquad \text{and} \qquad \hat{\beta}_m^{\mathrm{PCA}} = \sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j}\langle \hat{r}, \hat{v}_j\rangle \hat{v}_j$$

to denote the functional PLS and PCA estimators. Note that the PLS estimator uses supervised regularization $\hat{P}_m$ while for the PCA estimator the regularization is fixed to select the terms related to the inverse of the largest m eigenvalues of $\hat{K}$. We will also use

$$\beta_m^{\mathrm{PLS}} = \sum_{j=1}^{\infty} P_m(\lambda_j)\langle r, v_j\rangle v_j \qquad \text{and} \qquad \beta_m^{\mathrm{PCA}} = \sum_{j=1}^{m} \lambda_j^{-1}\langle r, v_j\rangle v_j$$

to denote the population counterparts.

Theorem S.1. *If $n_* = n$, then for every $m \leq n_*$,*

$$\left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PLS}} \right\| \leq \left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PCA}} \right\|.$$

and

$$\left\| K^s(\beta_m^{\mathrm{PLS}} - \beta) \right\| \leq \left\| K^s(\beta_m^{\mathrm{PCA}} - \beta) \right\|, \qquad \forall s \in [0,1].$$

Proof of Theorem S.1. If $m = n_*$, then the PLS objective is zero, so the result is trivial. Suppose that $m < n_*$. Recall that $\hat{Q}_m(\lambda) = 1 - \lambda\hat{P}_m(\lambda)$. Let $(\hat{v}_j)_{j=1}^{\infty}$ form an orthonormal basis of $\mathbb{H}$, where the first n_* terms correspond to the eigenbasis of $\hat{K}$. Then

$\hat{r} = \sum_{j=1}^{\infty} \langle \hat{r}, \hat{v}_j \rangle \hat{v}_j$ and for every $m < n_*$ by Blazère et al. (2014), Proposition 6.2

$$\begin{aligned}
\left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PLS}} \right\|^2 &= \left\| \hat{Q}_m(\hat{K})\hat{r} \right\|^2 \\
&\leq \sum_{j=m+1}^{n_*} \prod_{k=1}^{m} \left(1 - \frac{\hat{\lambda}_j}{\hat{\lambda}_k} \right)^2 \langle \hat{r}, \hat{v}_j \rangle^2 \\
&\leq \sum_{j=m+1}^{n_*} \langle \hat{r}, \hat{v}_j \rangle^2 \\
&= \left\| \sum_{j=1}^{\infty} \langle \hat{r}, \hat{v}_j \rangle \hat{v}_j - \sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j} \langle \hat{r}, \hat{v}_j \rangle \hat{K}\hat{v}_j \right\|^2 \\
&= \left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PCA}} \right\|^2,
\end{aligned}$$

where the third line follows since $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{n_*} > 0$; and the fourth by Parseval's identity.

For the second part, define $Q_m(\lambda) = 1 - \lambda P_m(\lambda)$, where $\beta_m^{\mathrm{PLS}} = P_m(K)r$ solves the population counterpart to the problem in Equation (4). Similarly, since $\beta = \sum_{j=1}^{\infty} \langle \beta, v_j \rangle v_j$ and $K\beta = r$, we have

$$\begin{aligned}
\left\| K^s(\beta_m^{\mathrm{PLS}} - \beta) \right\|^2 &= \| K^s Q_m(K)\beta \|^2 \\
&\leq \sum_{j=1}^{\infty} \lambda_j^{2s} \prod_{k=1}^{m} \left(1 - \frac{\lambda_j}{\lambda_{j_k}} \right)^2 \langle \beta, v_j \rangle^2 \\
&\leq \sum_{j=m+1}^{\infty} \lambda_j^{2s} \langle \beta, v_j \rangle^2 \\
&= \left\| K^s \left(\sum_{j=1}^{\infty} \langle \beta, v_j \rangle v_j - \sum_{j=1}^{m} \frac{1}{\lambda_j} \langle K\beta, v_j \rangle v_j \right) \right\|^2 \\
&= \left\| K^s(\beta - \beta_m^{\mathrm{PCA}}) \right\|^2.
\end{aligned}$$

□

The first part of Theorem S.1 shows that the PLS estimator fits the data better than PCA for the same number of components $1 \leq m \leq n_*$. This is the functional version of a result of Jong (1993); see also Phatak and de Hoog (2002) and Blazère et al. (2014). For

the second part of Theorem S.1, it is worth recalling that the estimation and prediction errors in Theorem 1 can be decomposed as

$$K^s\left(\hat{\beta}_m^{\mathrm{PLS}}-\beta\right)=K^s\left(\hat{\beta}_m^{\mathrm{PLS}}-\beta_m^{\mathrm{PLS}}\right)+K^s\left(\beta_m^{\mathrm{PLS}}-\beta\right),\qquad s\in\{0,1/2\},$$

where the second term is the so-called regularization bias. This shows that the PLS basis is more adapted for approximating the slope β than the PCA basis.

S.4 Comparison to Standard PLS

We provide some comparison between our PLS which is a functional conjugate-gradient method applied to the moment equation with the standard PLS. Note that the standard PLS solves

$$\min_{b\in\mathbb{H}_m}\|\mathbf{y}-T_nb\|_n^2, \tag{S.13}$$

where $\mathbb{H}_m=\mathrm{span}\left\{\hat{r},\hat{K}\hat{r},...,\hat{K}^{m-1}\hat{r}\right\}$. Equivalently, we solve

$$\min_{\alpha_1,...,\alpha_m\in\mathbb{R}}\left\|\mathbf{y}-\sum_{j=1}^m\alpha_jT_n\hat{K}^{j-1}\hat{r}\right\|^2,$$

where with some abuse of notation $\|.\|$ denotes the Euclidean norm. The first order conditions give

$$-\left\langle T_n\hat{K}^{j-1}\hat{r},\mathbf{y}\right\rangle+\left\langle T_n\hat{K}^{j-1}\hat{r},\sum_{l=1}^m\alpha_lT_n\hat{K}^{l-1}\hat{r}\right\rangle=0,\qquad j=1,2,...,m. \tag{S.14}$$

Remark that

$$\begin{aligned}
T_n\hat{K}^{j-1}\hat{r} &= T_n\hat{K}^{j-1}T_n^*\mathbf{y} = T_n\left(T_n^*T_n\right)^{j-1}T_n^*\mathbf{y} = \left(T_nT_n^*\right)^j\mathbf{y}, \\
\left\langle T_n\hat{K}^{j-1}\hat{r}, T_n\hat{K}^{l-1}\hat{r}\right\rangle &= \left\langle \hat{K}^{j-1}\hat{r}, T_n^*T_n\hat{K}^{l-1}\hat{r}\right\rangle \\
&= \left\langle \hat{K}^{j-1}\hat{r}, \hat{K}^{l}\hat{r}\right\rangle \\
&= \left\langle \hat{r}, \hat{K}^{j+l-1}\hat{r}\right\rangle \\
&= \left\langle T_n^*\mathbf{y}, \hat{K}^{j+l-1}T_n^*\mathbf{y}\right\rangle \\
&= \left\langle \mathbf{y}, \left(T_nT_n^*\right)^{j+l}\mathbf{y}\right\rangle \\
&= \mathbf{y}^\top\left(T_nT_n^*\right)^{j+l}\mathbf{y}.
\end{aligned}$$

where $T_nT_n^*$ is the $n \times n$ matrix with (i,j) element $\langle X_i, X_j\rangle$. Let v_1 be the $m \times 1$ vector and M_1 the $m \times m$ matrix such that

$$v_1 = \begin{bmatrix} \langle T_n\hat{r}, \mathbf{y}\rangle \\ \vdots \\ \left\langle T_n\hat{K}^{m-1}\hat{r}, \mathbf{y}\right\rangle \end{bmatrix} = \begin{bmatrix} \mathbf{y}^\top\left(T_nT_n^*\right)\mathbf{y} \\ \vdots \\ \mathbf{y}^\top\left(T_nT_n^*\right)^m\mathbf{y} \end{bmatrix},$$

$$M_1 = \left[\left\langle T_n\hat{K}^{j-1}\hat{r}, T_n\hat{K}^{l-1}\hat{r}\right\rangle\right]_{1\le j,l\le m} = \left[\mathbf{y}^\top\left(T_nT_n^*\right)^{j+l}\mathbf{y}\right]_{1\le j,l\le m}.$$

Equation (S.14) can be rewritten as $M_1\alpha = v_1$ where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_m]^\top$. So that $\hat{\alpha}_{PLS} = M_1^{-1}v_1$, provided that the matrix M_1 is invertible.

Our version of PLS solves

$$\min_{b\in\mathbb{H}_m}\left\|T_n^*\mathbf{y} - T_n^*T_nb\right\|^2$$

where $\mathbb{H}_m = \text{span}\left\{\hat{r}, \hat{K}\hat{r}, ..., \hat{K}^{m-1}\hat{r}\right\}$. This can be understood as the PLS with respect to the weighted norm. Equivalently, we want to solve

$$\min_{\alpha_1,...,\alpha_m\in\mathbb{R}}\left\|T_n^*\mathbf{y} - \sum_{j=1}^m\alpha_j\hat{K}^j\hat{r}\right\|^2.$$

Using the first order condition, we obtain the equation $v_2 = M_2\alpha$ with

$$v_2 = \begin{bmatrix} \left\langle \hat{K}\hat{r}, T_n^*\mathbf{y} \right\rangle \\ \vdots \\ \left\langle \hat{K}^m\hat{r}, T_n^*\mathbf{y} \right\rangle \end{bmatrix} = \begin{bmatrix} \mathbf{y}^\top (T_nT_n^*)^2\,\mathbf{y} \\ \vdots \\ \mathbf{y}^\top (T_nT_n^*)^{m+1}\,\mathbf{y} \end{bmatrix},$$

$$M_2 = \left[\left\langle \hat{K}^j\hat{r}, \hat{K}^l\hat{r} \right\rangle\right]_{j,l} = \left[\mathbf{y}^\top (T_nT_n^*)^{j+l+1}\,\mathbf{y}\right]_{j,l}.$$

Comparing v_1 and v_2, M_1 and M_2, shows that the only difference is in the power of $(T_nT_n^*)$. Note also that M_1 and M_2 are Hankel matrices.

S.5 Additional Simulations: Early Stopping and Confidence Sets

In this section we report additional simulation results for our early-stopping rule and confidence sets.

S.5.1 Early Stopping

Recall that our PLS amounts to fitting the norm of the "sample moment" (or score)

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\|.$$

The norm decreases monotonically to zero in m and it becomes zero when the number of conjugate-gradient steps reaches the number of nonzero eigenvalues of $\hat{K}$, i.e. $m = n_*$. To prevent overfitting, we select the number of PLS components as the first value of m for which the norm of residual drops below a certain threshold:

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq \tau\sigma\sqrt{\frac{2\mathbf{E}\left\|X\right\|^2}{\delta n}}, \tag{S.15}$$

where $\tau > 1$ is a constant and $1 - \delta$ is the confidence level of the rule; see Assumption 4. To describe the practical implementation, we focus on the functional linear model

$$Y_i = \int_0^1 \beta(s) X_i(s) \mathrm{d}s + \varepsilon_i, \qquad i = 1, \ldots, n.$$

We compute the fitted moment as:

$$\hat{r}(u) - (\hat{K}\hat{\beta}_m)(u) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \int_0^1 \hat{\beta}_m(s) X_i(s) \mathrm{d}s \right) X_i(u),$$

To compute the threshold in Equation (S.15), we estimate $\mathbf{E} \|X\|^2$ by $\frac{1}{n} \sum_{i=1}^{n} \int_0^1 X_i^2(s)\, ds$ and σ^2 by $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, where

$$\hat{\varepsilon}_i = Y_i - \int_0^1 \hat{\beta}(s) X_i(s)\, ds$$

and $\hat{\beta}$ is a preliminary estimator of β, described below. All integrals are discretized on a uniform grid of 200 points.

We consider an iterative approach for estimating σ^2 inspired by Chernozhukov et al. (2024), Section 3.A. To that end, we first obtain a pilot estimator of β, denoted $\hat{\beta}_0$. Next, we compute the estimator of σ^2 :

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle X_i, \hat{\beta}_0 \rangle \right)^2.$$

If $n \gg T$, we can use the OLS estimator. Otherwise, any other regularized estimator can be used, e.g. PCA with generalized cross-validation. We set $k = 0$ and specify a small constant $\xi \geq 0$ as a tolerance level and the maximum number of iterations $k_{\max}$. The iterative procedure is described in Algorithm 1.

Algorithm 1: Iterative estimation of σ^2.

Result: $\hat{\sigma}_k^2$

Initialization: ξ, $k_{\max}$;

for $k = 0, 1, \ldots, k_{\max}$ **do**

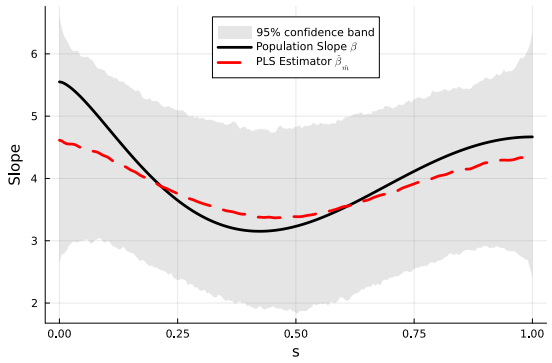
1. Compute $\hat{\beta}$ using the early-stopping rule with σ^2 replaced by $\hat{\sigma}_k^2$;
2. Set $\hat{\sigma}_{k+1}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \hat{\beta}, X_i \rangle)^2$;
3. If $\left|\hat{\sigma}_{k+1}^2 - \hat{\sigma}_k^2\right| \leq \xi$, then stop.

end

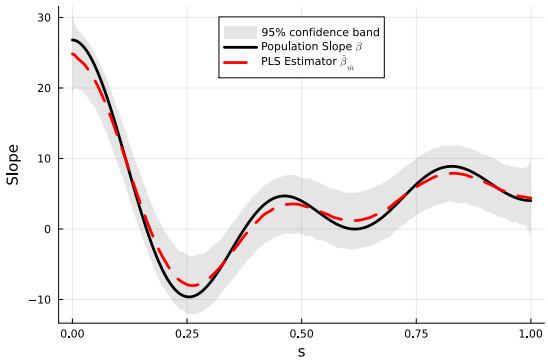
We use the following values in all simulation designs: $\delta = 0.1$, $k_{\max} = 10$, and $\xi = 0.01$. These values correspond to a 90% confidence level, a sufficiently high number of iterations, and a sufficiently low tolerance level respectively. Our theory also tells us that we need $\tau > 1$, so we set $\tau = 1.01$. The same values are also used in the empirical application. We noticed in simulations that the algorithm tends to overestimate σ^2 which leads to underestimated values of m. Figure S.1 reports the population slope coefficient (black solid curve) and the average values of estimated slope coefficients $\hat{\beta}_{\hat{m}}$ (red dashed curve) when the selected number of PLS components $\hat{m}$ for the three simulation designs considered in the paper, Section 5. We also report the median values of $\hat{m}$ selected by our data-driven adaptive rule in each case.

Overall, we can see that the PLS estimator with the number of components selected using early stopping can successfully recover the global shape of the slope parameter across all simulation designs. It is worth noting that the early-stopping rule selects the number of PLS components to achieve the minimax-optimal convergence rate for both the integrated squared error (ISE) and the mean squared prediction error (MSPE). In doing so, it simultaneously balances and minimizes the sum of the squared bias and the variance. This behavior is consistent with the results shown in Figure S.1, where a visible bias is observed

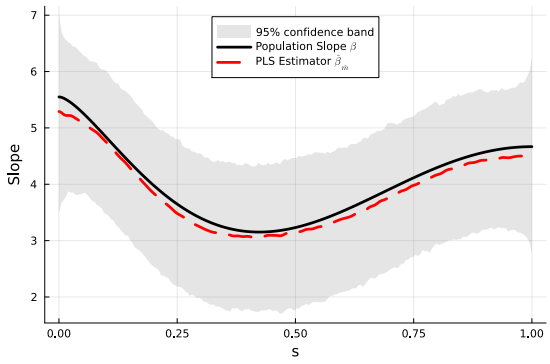
Figure S.1: Early-stopped PLS. True slope function β (solid black curve) and Monte Carlo mean of the early-stopped PLS estimator $\hat{\beta}_{\hat{m}}$ (dashed red curve), with a 95% pointwise confidence band (area shaded in gray), based on 5,000 replications with sample size $n =$ 1,000. The median number of selected functional components ($\hat{m}$) is reported below.



(a) Model 1: median($\hat{m}$) = 1



(b) Model 2: median($\hat{m}$) = 3



(c) Model 3: median($\hat{m}$) = 2

as part of the optimal bias-variance trade-off.

Figure S.2 visualizes our early-stopping rule. The left vertical axis corresponds to the values of MSPE (black solid curve) while the right vertical axis to the early-stopping rule (blue). The number of PLS components is selected as the first m (blue dot) when the norm of the fitted moment (blue dashed curve) drops below the value of the threshold (blue dotted line). We also plot the minimum (black dot) of the simulated MSPE (black solid curve) together with 95% confidence band (area shaded in gray). Given the uncertainty behind the lowest value of the MSPE as well as the norm and the threshold, the early-stopping rule produces reasonable values of tuning parameters that are compatible with the global minimum of MSPE once the statistical uncertainty is accounted for. The simulation results also confirm that there is no reason to select higher values of τ since they would increase the threshold and lead to more conservative choices of $\hat{m}$. We conclude that our early-stopping rule does not lead to overfitting.

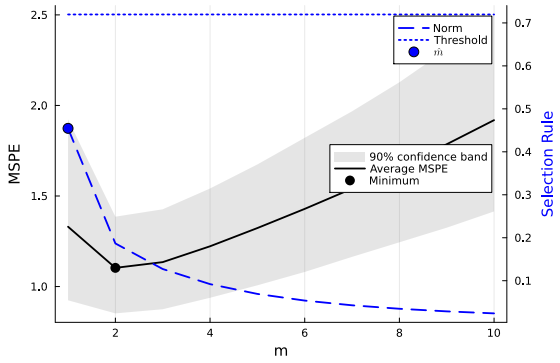
S.5.2 Confidence Sets

To compute the confidence sets using test inversion, we note that if $(h_j)_{j=1}^{\infty}$ is a basis of $L_2[0,1]$, then

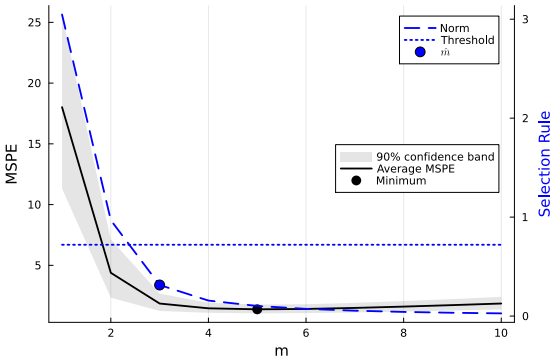
$$\beta(s) = \sum_{j=1}^{\infty} b_j h_j(s)$$

with coefficients $b_j = \langle \beta, h_j \rangle$. Since $b_j \to 0$, to reduce the computational cost, we search over functions β, where the sum is truncated to include only the first five basis functions. We use the cosine basis and create a uniform grid of 20 points on $[0, 4.5]^5$ corresponding to the first five coefficients $b_1, \ldots, b_5$. We simulate the critical value using the approximation to the asymptotic distribution $\sum_{j=1}^{100} \lambda_j Z_j^2$, where $(\lambda_j)_{j=1}^{\infty}$ are the eigenvalues of K, and we use $50,000$ replications to compute the critical value for the 5%-level test.

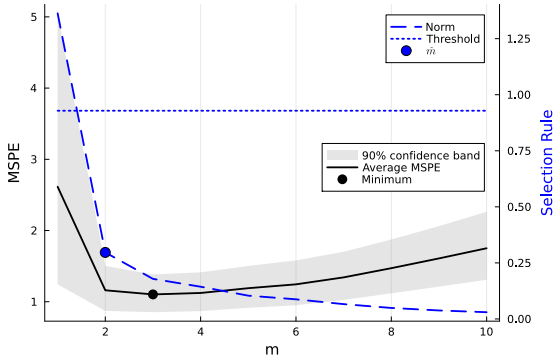
Figure S.2: Adaptive early-stopping rule. Mean MSPE (solid black line) with a 90% Monte Carlo confidence band (area shaded in gray), computed from 5,000 simulations with sample size $n = 1{,}000$. The black dot marks the minimum. The rule selects the smallest m (blue dot) at which the average norm (blue dashed curve) falls below the threshold (blue dotted line); this blue dot therefore indicates the selected number of components, $\hat{m}$. The median number of selected functional components ($\hat{m}$) is reported below.



(a) Model 1: median($\hat{m}$) = 1

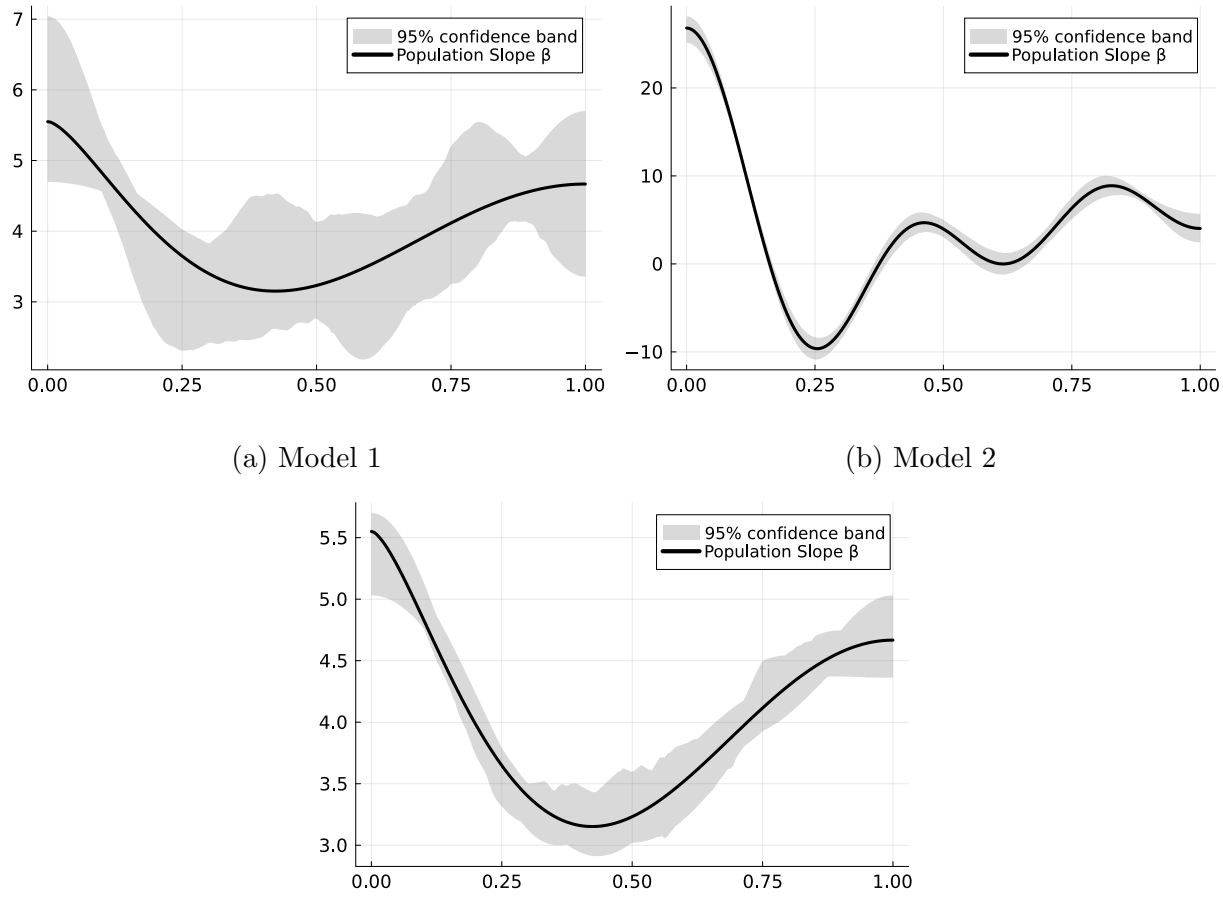


(b) Model 2: median($\hat{m}$) = 3



(c) Model 3: median($\hat{m}$) = 2

Figure S.3 displays the asymptotic 95% confidence bands based on test inversion. These confidence bands are quite informative about the global shape of the estimated slope parameter.



(a) Model 1

(b) Model 2

(c) Model 3

Figure S.3: Asymptotic confidence bands. Pointwise medians of the upper and lower bounds of the 95% confidence band (area shaded in gray), computed across 5,000 Monte Carlo replications.

References

Alexei B Aleksandrov and Vladimir V Peller. Operator lipschitz functions. *Russian Mathematical Surveys*, 71(4):605–702, 2016.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

Mélanie Blazère, Fabrice Gamboa, and Jean-Michel Loubes. A unified framework to study the properties of the PLS vector of regression coefficients. In *The Multiple Facets of Partial Least Squares and Related Methods*, volume 173 of *Springer Proceedings in Mathematics & Statistics*, pages 227–237. Springer, 2014.

Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149 of *Lecture Notes in Statistics*. Springer, 2000.

T Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ML and AI. *arXiv preprint arXiv:2403.02467*, 2024.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, 1996.

Martin Hanke. *Conjugate gradient type methods for ill-posed problems*, volume 327 of *Pitman Research Notes in Mathematics Series*. CRC Press, 1995.

Sijmen De Jong. PLS fits closer than PCR. *Journal of Chemometrics*, 7(6):551–557, 1993.

Aloke Phatak and Frank de Hoog. Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16(7):361–367, 2002.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, 2009.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer, 1996.