Functional Partial Least-Squares: Optimal Rates and Adaptation*

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

September 20, 2024

Abstract

We consider the functional linear regression model with a scalar response and a Hilbert space-valued predictor, a well-known ill-posed inverse problem. We propose a new formulation of the functional partial least-squares (PLS) estimator related to the conjugate gradient method. We provide the first optimality result for functional PLS showing that the estimator achieves the (nearly) optimal convergence rate on a class of ellipsoids and propose a data-driven early stopping rule that adapts to the unknown degree of ill-posedness. We find in simulations that our estimator performs favorably compared to the principal component regression estimator and requires a smaller number of functional components. Using our estimator, we study the non-linear temperature effect on corn and soybean yields and find some evidence of the adaptation of US agriculture to climate change.

Keywords: Conjugate Gradient Method, Early Stopping, Adaptive Estimation, Functional Linear Regression, Climate Science.

*An older version of this paper has circulated under the name of "Theoretical comparison of the functional principal component analysis and functional partial least-squares". The authors would like to thank the participants of the SCSE conference, Quebec (2019), EC^2 conference, Paris (2020), NBER-NSF Time-Series conference, Rice University (2021), ESIF Economics and AI+ML Meeting, Ithaca (2024), Triangle Econometrics Conference (2024), and Econometrics in Rio (2024) for helpful comments. The authors are also grateful to Ryan Borhani for excellent research assistantship. Marine Carrasco thanks NSERC for partial financial support.

1 Introduction

With the growing availability of data, functional data analysis has been increasingly applied across various fields, including chemometrics, climate science, and economics. In this paper, we consider a linear functional regression with 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 main goal is the estimation of the functional slope β and the prediction of the response Y. Both X and β belong to an infinite dimensional Hilbert space and the high dimensionality of the parameter β leads to an ill-posed inverse problem. To estimate the slope coefficient β consistently, one needs to use either a dimension-reduction device or regularization. The most popular approach has been the functional principal component analysis (PCA); see Cardot et al. (1999), Ramsay and Silverman (2002), Cai and Hall (2006), and Hall and Horowitz (2007). The PCA approach consists of approximating β by the first principal components (PC) corresponding to the eigenfunctions of the covariance operator of X. As noted by Jolliffe (1982), this approach will work best if the response variable Y is correlated with the first PCs.

In this paper, we propose to estimate the slope function by partial least squares (PLS). This method uses components that are linear transformations of the regressors X chosen to maximize the correlation with the response Y. As a result, the number of components required to obtain a good prediction is generally smaller than the number of PCs. It was initially developed to estimate regressions with many regressors, see Wold et al. (1984) and Helland (1988), and has recently been adapted to functional regression by Delaigle and Hall (2012). As the estimator is nonlinear in the response variable and computed iteratively, analyzingf its properties is much more challenging. To establish our theoretical results, we

rely on the inverse problem literature and exploit the close connection between PLS and conjugate gradient as presented in Engl et al. (1996), Hanke (1995), and Blanchard and Krämer (2016).

Our paper makes several original contributions. First, we establish the convergence rate of our estimator and the prediction error under a source condition that relates β and the spectral decomposition of the covariance operator. Our results do not require specifying the distance between the adjacent eigenvalues as commonly done in PCA, see for instance Hall and Horowitz (2007), and hold even in the presence of repeated eigenvalues. Second, we derive the lower bound on the minimax convergence rate and show that our estimator is (nearly) minimax optimal. Given the optimal convergence rate depends on the degree of ill-posedness of the problem which is typically unknown, we propose an adaptive early stopping method for selecting the number of PLS components. We show that this data-driven selection yields an estimator that is rate optimal with high probability. We also characterize how the selected number of components changes with the sample size in various scenarios. A theoretical comparison shows that the bias of the PLS estimator is smaller than or equal to that of the PCA estimator for the same number of components. Moreover, simulations show that the PLS estimator gives a smaller prediction error than PCA in various settings and that the PLS estimator combined with the adaptive stopping rule gives reliable estimates. An efficient algorithm to compute the estimator iteratively is provided.

To illustrate the practical relevance of our results, we apply our method to climate science. Using a fine-grained county-level dataset of US crop yields and temperature, recorded over 70 years, we estimate the impact of temperature on crop yields. We find that the critical temperature after which the annual crop yields start declining is around 30°C

which is similar to Schlenker and Roberts (2009) who rely on highly-parameterized least-squares estimators. We obtain additional insights by comparing how the temperature effect curves change over time. Interestingly, we find that the detrimental temperature effects on corn and soybean yields have decreased over time which is attributed to the adaptive actions taken by farmers, including the use of resilient crops and efficient irrigation systems.

The literature on functional regression is vast. Besides PCA already mentioned, another popular approach consists of using a predetermined basis to approximate the slope function β, including highly parametrized step functions or polynomials. Other papers propose a penalized estimator using Tikhonov regularization, see Florens and Van Bellegem (2015), or using a roughness penalty or similarly a penalty in terms of a norm in a reproducing kernel Hilbert space; see Cardot et al. (2003), Li and Hsing (2007), Yuan and Cai (2010). Reiss and Ogden (2007) combine the projection on the first principal component and a roughness penalty. Finally, in an independent work, Gupta et al. (2023) propose an estimator that belongs to a reproducing kernel Hilbert space generated by a specified kernel and uses a conjugate gradient to regularize the solution. While their method is similar to ours, our results are not directly comparable. Their main result is the convergence rate of an estimator in reproducing kernel Hilbert space, established under assumptions that differ from ours. In particular, they assume a polynomial decay rate of the eigenvalues of certain operators, while we do not need to specify any decay rate for eigenvalues in our work. Their source condition is also different from ours and it is not clear which one is the weakest.

The rest of the paper is organized as follows. Section 2 introduces the functional regression and the PLS estimator. Section 3 establishes the theoretical properties including the convergence rate of estimation and prediction errors, the minimax lower bound, the adaptivity of an early stopping rule, and the bias comparison to PCA. Section 4 presents

a Monte Carlo study. Section 5 describes an empirical application of our estimator to the nonlinear temperature effects in agriculture, and Section 6 concludes. An online appendix contains all the proofs and additional simulation results.

2 Model and 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 $\langle f, g \rangle = \int_0^1 f(s)g(s)$; see also Cai and Hall (2006), Cai and Yuan (2012), Cardot et al. (1999), Cardot et al. (2003), Cardot and Johannes (2010), Comte and Johannes (2012), Crambes et al. (2009), Delaigle and Hall (2012), Hall and Horowitz (2007), among many other important contributions.

For simplicity we assume that $\mathbb{E}[X] = 0$. 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), Hoffmann and Reiss (2008), Klemelä and Mammen (2010), Nemirovski (1986).

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 basis vectors $h_1, h_2, \ldots, h_m \in \mathbb{H}$.

Examples of (a) include the Tikhonov regularization when $R_\alpha(\lambda) = (\alpha+\lambda)^{-1}$, the spectral cut-off when $R_\alpha(\lambda) = \lambda^{-1}\mathbf{1}_{\lambda\geq\alpha}$ and the Landweber iterations; see Carrasco et al. (2007), Cavalier (2011), Engl et al. (1996) for more details. 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 put $\mathbf{y} = (Y_1, \ldots, Y_n)^\top$ 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

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

in practice. The PCA basis is constructed by identifying the directions in $\mathbb{H}$ where X varies 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 the issue on datasets used in 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 Blazère et al. (2014a); Delaigle and Hall (2012); Preda and Saporta (2005); Reiss and Ogden (2007); Wold et al. (1984) for theoretical analysis and Carrasco and Rossi (2016); Kelly and Pruitt (2015) for applications in economics and finance. 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 solving 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 Helland (1988); Phatak and de Hoog (2002) for the link between PLS and Krylov subspaces.

While the functional PLS estimator is popular in practice due to its efficient representation of the data, to the best of our knowledge, it is still unknown whether it is the minimax

optimal estimator. 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 first-order conditions to the problem in equation (3), often called the normal equations. Equivalently, $\hat{\beta}_m$ fits the empirical counterpart to the equation (2)

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

as it is easy to see that $\hat{r} = T_n^* \mathbf{y}$ and $\hat{K} = T_n^* T_n$. Importantly, the PLS estimator formalized in equation (4) corresponds to the conjugate gradient method with a self-adjoint operator $\hat{K}$, cf. Hestenes and Stiefel (1952), known for its excellent regularization properties; see also Blanchard and Krämer (2016); Engl et al. (1996); Hanke (1995); 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 non-zero eigenvalues of $\hat{K}$; see Proposition 1 in the Supplementary Material. It is also

[2]The method of conjugate gradients 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.

easy to see that for every $m \geq 1$,[3] 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 $\mathbf{a} := (a_1, \ldots, a_m)^\top$ solving the system of m linear equations

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

where $\mathbf{K} := \langle \hat{K}^j \hat{r}, \hat{K}^k \hat{r} \rangle_{1 \leq j,k \leq m}$ and $\mathbf{r} := \langle \hat{K}^j \hat{r}, \hat{r} \rangle_{1 \leq j \leq m}$. From the 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 4.

3 Theoretical Properties

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 for the estimator 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n \geq 0$ 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.

[3] We also define $\hat{P}_0 = 0$ and $\hat{\beta}_0 = 0$.

Note that the sample covariance operator $\hat{K}$ is a finite-rank operator with at most $n_* \leq n$ distinct non-zero 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 be often 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 shall introduce several relatively mild assumptions on the distribution of the data next.

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, the focus would shift to the identified component of β within the orthogonal complement of the null space of K; see Babii and Florens (2017) and Engl et al. (1996).

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

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

Assumption 3 describes the complexity of the ill-posed inverse problem 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 decline of the generalized Fourier coefficients $\langle \beta, v_j \rangle_{j \geq 1}$ relatively 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 $\mathbf{E}\|X\|^2 < \infty$.

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 inverse 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 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 $(Y_i, X_i)_{i=1}^n$.

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

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

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

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 $\mathcal{S}(\mu, R, C)$, simultaneously for the estimation ($s = 0$) and prediction ($s = 1/2$) errors.[4]

3.2 Adaptive PLS estimator

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

[4]It is possible to avoid the $1/\log n$ factor by considering the larger class of Hilbert–Schmidt operators.

Assumption 4. *We select $\hat{m}$ such that*

$$\min\left\{m \geq 0 : \ \left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq \tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}}\right\}.$$

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

Assumption 4 states that the PLS iterations stop at the first value $\hat{m}$ for which the norm of the fitted "moment" is smaller than a certain threshold; see Supplementary Material, Section S.4 for a practical implementation of this early stopping rule. Note that the number of iterations is finite since $\hat{m} \leq n_*$, where n_* is the number of distinct non-zero eigenvalues of $\hat{K}$; see Proposition 1 in the Supplementary Material. In fact, the norm of "residual" is zero for $m \geq n_*$ in which case we have perfect overfitting.

The following result shows that the data-driven rule in Assumption 4 is adaptive to the unknown degree of ill-posedness $\mu > 0$.[5]

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 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 $\mu > 0$.

[5]See also Cai and Yuan (2012); Comte and Johannes (2012) for adaptivity results in the functional linear regression model and Blanchard et al. (2018) for the analysis of early stopping rule in abstract inverse problems.

3.3 Number of Selected Components

In this section, we look at how rapidly the number of components selected by the early stopping rule 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 decline polynomially fast, then the selected number of components is $\hat{m} = O_P(n/\log n)^{\frac{1}{4(\kappa+1)(\mu+1)}}$ while in the case of the geometric decline, the number of selected components increases slowly with the sample size. Therefore, the adaptive stopping rule will select a smaller number of components if the eigenvalues of the operator K decline faster and vice versa.

3.4 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 coefficient.

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 6. *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

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

The first part of Theorem 6 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. (2014b).

For the second part of Theorem 6, 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.

4 Monte Carlo Experiments

In this section, we set up several Monte Carlo experiments to evaluate the finite sample performance of the PLS estimator.

We simulate the i.i.d. samples $(Y_i, X_i)_{i=1}^n$ as follows

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

where the predictors 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

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

The slope parameter $\beta\in L^2[0,1]$ and the spectrum of the operator $(\lambda_j,v_j)_{j\geq 1}$ correspond to one of the following four models:

- **Model 1:** $\beta(s)=\sum_{j=1}^{100}\beta_jv_j(s)$ with $\beta_j=4(-1)^{j+1}j^{-1}$, $v_1(s)=1$, $v_j(s)=\sqrt{2}\cos((j-1)\pi s)$, and $\lambda_j=j^{-1.1}$.
- **Model 2:** same as Model 1, but with $\beta_j=4$ for $j=1,\ldots,10$.
- **Model 3:** same as Model 1, but with $\beta_j=4(-1)^{j+1}j^{-1/4}$.

- **Model 4:** same as Model 1, but with $\lambda_j = 2^{-r_j}$ with $r_j = \lceil 0.1j \rceil, \forall j \geq 1$, where $x \mapsto \lceil x \rceil$ is the ceiling function.

Note that the first few high-variance principal components terms are sufficient to capture most of nonlinearities in Model 1. Therefore, this design favors strongly PCA. On the other hand, the low-variance components are important for predicting Y in Model 2, cf. Jolliffe (1982). With Model 3, we would like to see how the methods perform when the Fourier coefficients are not square-summable. Lastly, Model 4 is an example of severely ill-posed problem with repeated eigenvalues. We compute the PLS estimator using the Algorithm 1 which is numerically equivalent to the estimator in equation (4); see Hanke (1995), Algorithm 2.1 and Proposition 2.1. It bypasses the operator inversion with an iterative multiplication by $\hat{K}$ and is designed to solve high-dimensional linear systems, $\hat{K}\hat{\beta} = \hat{r}$, 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 the step size: $\alpha_j = \frac{\langle e_j, \hat{K}e_j \rangle}{\|\hat{K}d_j\|^2}$;
2. Update the slope coefficient: $\hat{\beta}_{j+1} = \hat{\beta}_j + \alpha_j d_j$;
3. Update the fitted moment: $e_{j+1} = e_j - \alpha_j \hat{K} d_j$;
4. Compute the 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 the 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 approximation with Riemann sum on a grid of $T = 100$ equidistant points in $[0, 1]$.[6] The

[6]We also tried to use larger values of T and did not find that finer approximations change substantially

experiments feature $5,000$ replications, where samples of size $n = 1,000$ are generated in each replication. For each experiment, the mean-squared prediction error (MSPE) is computed as

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

where $\hat{\beta}_m$ is obtained from an auxiliary sample of size n, independent of $(Y_i, X_i)_{i=1}^n$.

Figure 1 displays the MSPE using the first $m = 1, 2, \ldots, 15$ components using PLS (orange circles) or PCA (blue crosses). The PLS estimator achieves lower value of MSPE with a smaller number of components, so it offers a more parsimonious representation of the data than PCA across all four models.

We report the exact values of simulated MSPE with highlighted lowest values in Table 1. These results confirm that the PLS estimator achieves the lowest value of MSPE across all designs and it usually does so with a smaller number of components. We find that for prediction, the optimal number of PCA components is $(25, 36, 100, 35)$ while PLS only requires $(5, 7, 13, 4)$ components for Models 1 to 4 respectively. For Model 3, the prediction error of PCA continues to decrease up to $m = 100$ components.[7]

Next, we look at the mean integrated squared error (MISE) computed as

$$\text{MISE}(\hat{\beta}_m) = \mathbf{E}\int_0^1 |\hat{\beta}_m(s) - \beta(s)|^2 \mathrm{d}s,$$

where the expectation is approximated with the simulations. The results in Table 2 show that PLS achieves smaller MISE with a smaller number of functional components. The optimal number of the PCA components is $(34, 34, 100, 35)$ while PLS requires $(4, 6, 13, 3)$ components only for Models 1-4 respectively. The optimal number of components for prediction (MSPE) and estimation (MISE) are similar which is in line with our theoretical

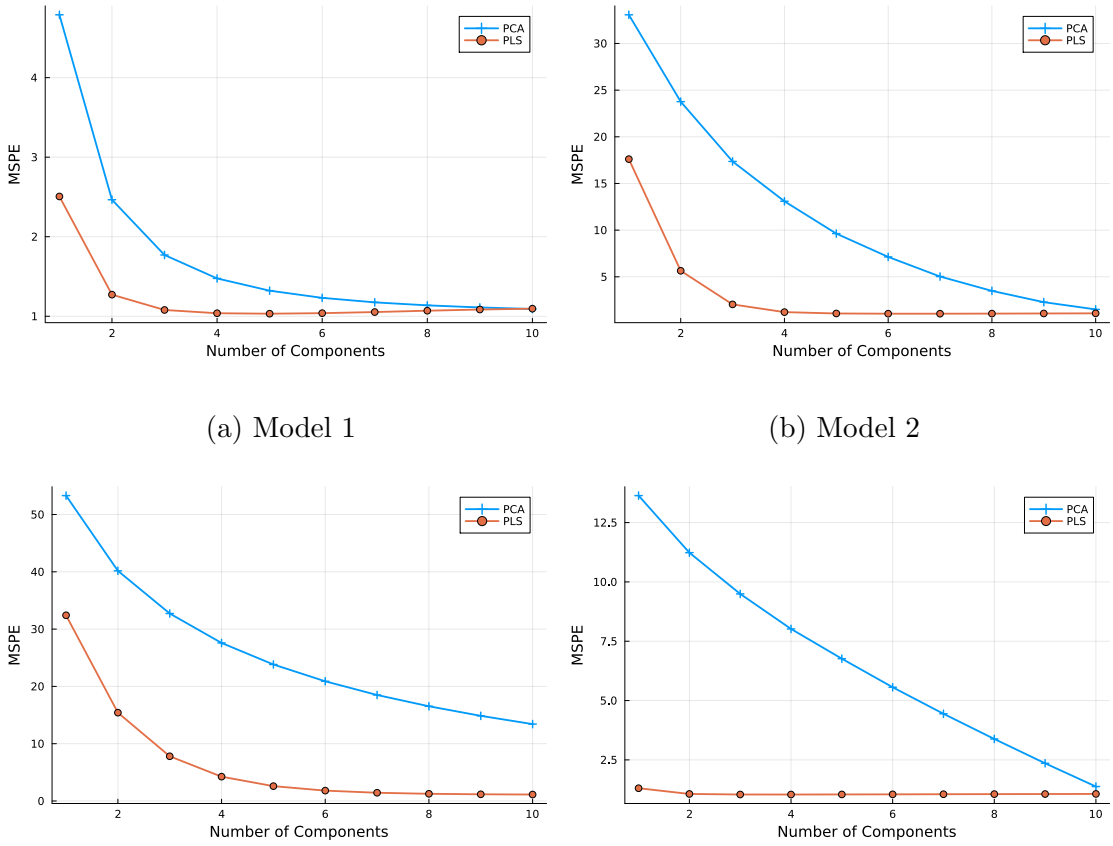
results reported in the paper.

[7]For readability, we report results for the first 10 functional components in the paper.

Table 1: Mean-Squared Prediction Error (MSPE) of PLS and PCA using the first m components, calculated from $1,000$ samples of size $n = 1,000$. The number of components corresponding to the lowest MSPE is highlighted.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	4.7903	2.5062	33.0712	17.608	53.2982	32.4159	13.6365	1.3056
2	2.4651	1.2715	23.7662	5.6453	40.1718	15.4156	11.232	1.0648
3	1.7703	1.079	17.3605	2.0454	32.7232	7.8054	9.4952	1.0413
4	1.4757	1.0379	13.0917	1.2169	27.5816	4.246	8.0192	**1.0402**
5	1.3217	**1.0321**	9.6292	1.0708	23.8259	2.5921	6.7641	1.0440
6	1.2314	1.0390	7.1322	1.0480	20.8922	1.8100	5.5612	1.0473
7	1.1746	1.0528	5.0317	**1.0475**	18.5005	1.4363	4.4428	1.0529
8	1.1370	1.0697	3.4932	1.0576	16.5333	1.2557	3.3814	1.0572
9	1.1107	1.0846	2.2835	1.0717	14.8621	1.1686	2.3566	1.0606
10	1.0913	1.0953	1.5098	1.0859	13.4207	**1.1286**	1.3763	1.0655

Figure 1: Mean-Squared Prediction Error (MSPE) of PLS (orange circles) and PCA (blue crosses) using the first m components, calculated from $5,000$ samples of size $n = 1,000$.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

results. The resulting MISE are substantially lower for PLS across the four models. This suggests that the PLS basis is more suitable for estimating the slope coefficient β.

Tables 3 and 4 display the bias and the variance MISE components. We can see that the remarkable MISE performance of PLS is driven by substantially smaller bias for the same number of functional components. On the other hand, the variance of PCA can sometimes be smaller than that of PLS. This aligns with the result of Theorem 6 confirming that the PLS basis approximates the slope coefficient β better compared to the PCA basis.

Table 2: Mean Integrated Squared Error (MISE) of PLS and PCA using the first m components, calculated from $5,000$ samples of size $n = 1,000$. The number of components corresponding to the lowest MISE is highlighted.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	12.9386	6.8200	158.2132	99.8625	546.089	448.6095	24.9864	1.2872
2	8.0895	2.5672	138.6779	38.3661	518.7172	319.7507	20.7245	0.6244
3	5.8387	1.2160	117.8103	11.1731	494.3239	209.6672	17.5198	**0.4894**
4	4.5358	**0.8009**	98.8206	3.5350	471.3402	130.3394	14.7569	0.4906
5	3.6653	0.8669	79.1779	1.9680	449.7725	79.2572	12.3737	0.5809
6	3.0475	1.3308	61.9748	**1.7267**	429.3939	47.8975	10.075	0.6867
7	2.5884	2.2507	45.0136	1.9509	409.7226	29.2685	7.9137	0.9227
8	2.239	3.5537	30.7397	2.7170	390.9891	18.478	5.8557	1.1535
9	1.9602	4.9199	18.0959	3.8499	372.7052	12.4587	3.8655	1.3964
10	1.7321	6.0644	9.2499	5.1692	355.2951	**9.3763**	1.9564	1.8625

Table 3: Squared Bias part of MISE using the first m components, calculated from $5,000$ samples of size $n = 1,000$. The number of components corresponding to the lowest MISE is highlighted.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	12.8396	6.7512	158.042	99.6102	545.8576	447.9841	21.0837	0.7598
2	7.9718	2.4527	138.2143	37.4805	517.7943	318.1858	15.2567	0.4077
3	5.6967	1.0159	116.6897	9.8767	492.4548	206.5991	11.5159	**0.2703**
4	4.3522	**0.4349**	96.8103	2.122	468.6711	125.7045	8.6575	0.2079
5	3.4581	0.1802	76.2421	0.5579	445.8752	73.4402	6.4482	0.1539
6	2.8112	0.0689	57.8386	**0.2224**	424.042	41.6481	4.5399	0.1287
7	2.3325	0.0259	40.135	0.0889	402.9084	22.8554	3.1748	0.099
8	1.9664	0.0118	25.1972	0.0288	382.3052	12.0892	2.165	0.0833
9	1.6725	0.0059	12.3835	0.0105	363.0781	6.1220	1.5032	0.0715
10	1.4353	0.0031	4.3479	0.0045	343.7048	**2.9256**	1.2050	0.0566

Table 4: Variance part of MISE using the first m components, calculated from $5,000$ samples of size $n = 1,000$. The number of components corresponding to the lowest MISE is highlighted.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	0.1026	0.0827	0.1724	0.2736	0.2483	0.4112	3.837	0.5221
2	0.1486	0.1211	0.5518	0.8766	0.7678	1.5063	5.3783	0.2154
3	0.1766	0.1990	1.1741	1.3028	1.5379	3.1381	5.9298	**0.2180**
4	0.1995	**0.3670**	1.9509	1.4047	2.4680	4.7544	6.0901	0.2815
5	0.2219	0.6883	2.9034	1.4026	3.6991	5.8475	5.9311	0.4254
6	0.2385	1.2648	3.8637	**1.4977**	5.0667	6.3634	5.4788	0.5571
7	0.2572	2.2256	4.8643	1.855	6.5964	6.4811	4.7681	0.8252
8	0.2701	3.5472	5.5178	2.6769	8.1606	6.4353	3.7528	1.0727
9	0.2860	4.9168	5.7629	3.8224	9.8113	6.412	2.4025	1.3285
10	0.2981	6.0604	4.9463	5.143	11.7207	**6.5094**	0.7433	1.8125

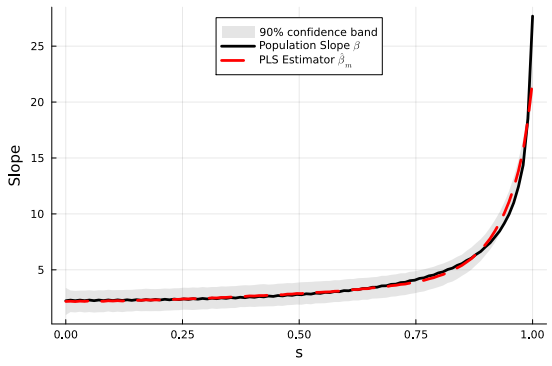
Lastly, Figure 2 reports the mean of the PLS estimator $\hat{\beta}_m$ (dashed red) with the MISE-optimal number of functional components, m, as well as the true slope coefficient (solid black). Remarkably, PLS requires a relatively small number of components to recover accurately the shape of various functions, including those with sharp changes and complex nonlinearities. This is especially important for our empirical application, where the temperature effects on crop yields are believed to exhibit sharp changes for extreme values; see Section 5.

The simulation results also underscore the importance of early stopping for the PLS since there is overfitting for excessively large m. In the Supplementary Material, Section S.4, we report additional simulation results that show that our early stopping rule produces reliable estimates. To conclude, the results of the experiments confirm our theoretical results and illustrate that the supervised PLS bases are well-suited for representing the slope parameter β and for predicting the response variable Y.

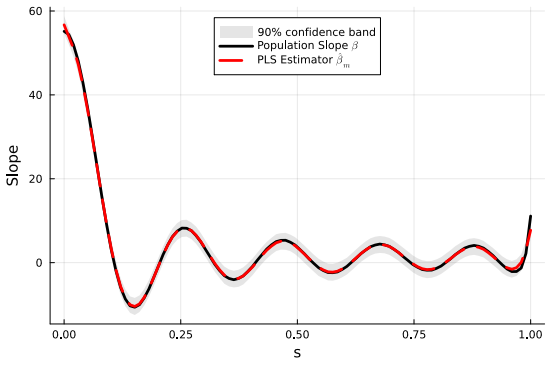
5 Nonlinear Temperature Effects in US Agriculture

The global surface temperature has increased by 1.1°C above pre-industrial levels and could increase up to 3.6°C to 4.5°C by the end of the century if current CO_2 emissions rise steadily according to the latest studies; see IPCC (2021). The global warming will likely lead to more frequent and severe heatwaves, 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 potentially exacerbate mass migration and violent conflicts. Some have argued that the current climates are already warmer than is optimal for agriculture in many parts of Asia, Africa, and Latin

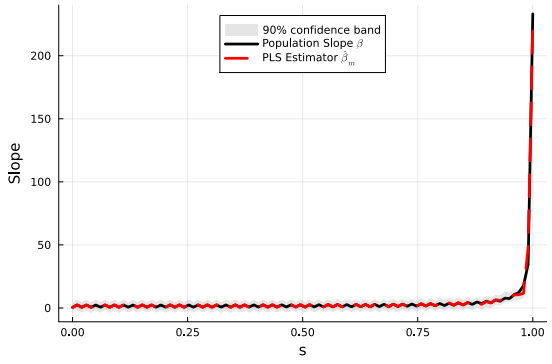
Figure 2: Average PLS estimator with the MISE-optimal number of functional components and the pointwise 90% confidence band, calculated from $5,000$ samples of size $n = 1,000$.



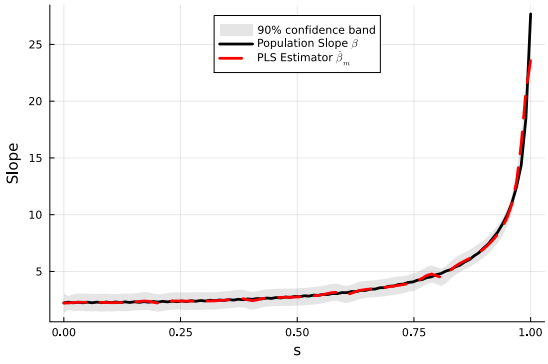
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

America; see Nordhaus (2013).

Determining the precise functional form of the relationship between crop yields and temperature has attracted lots of attention recently; see Schlenker and Roberts (2006, 2009).[8] We argue that the methodology used to estimate such nonlinear temperature effects can be understood as a functional linear regression,[9] where the outcome Y is the log yield of a crop, measured in bushels per acre, and the functional regressor $(X(s))_{s\in[0,40]}$ is a temperature curve, representing the crop exposure to temperatures between 0°C to 40°C during the growing season, measured in degree days, reflecting how much the crop is exposed to particular temperatures.

We focus on corn and soybeans which are the two major crops grown in the US. The dataset is comprised of fine-scale county-level crop yields and weather outcomes, spanning US counties east of the 100 degree meridian from 1950 to 2020.[10] We use the same set of controls as in Schlenker and Roberts (2009), namely (precipitation,precipitation^2,t, t^2), the county dummies, and the interaction between (t, t^2) and state dummies. 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 optimal

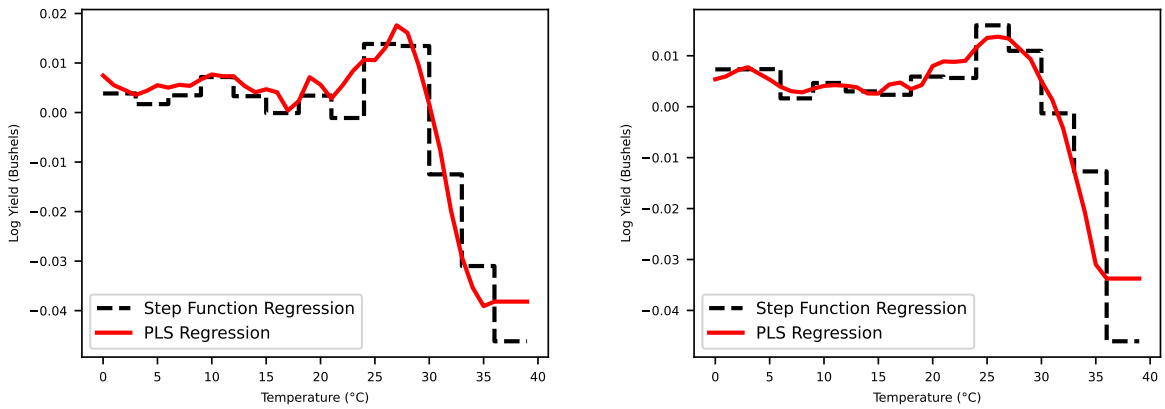
[8]The influential study of Schlenker and Roberts (2009) has more than 4,000 Google Scholar citations at the time of writing.

[9]A similar methodology is also used to quantify the nonlinear temperature effects on mortality in public health studies; see Gasparrini et al. (2015).

[10]The dataset is publicly available at the time of writing at `www.wolfram-schlenker.info/replicationFiles/SchlenkerRoberts2009.zip`.

stopping rule finds $\hat{m} = 4$ functional PLS components both for corn and soybeans; see Appendix Section S.4 for more details on the implementation of the early stopping rule.

Figure 3: Nonlinear relationship between temperature and crop yields fitted using functional PLS (red curve) and step function approximation (black dash).

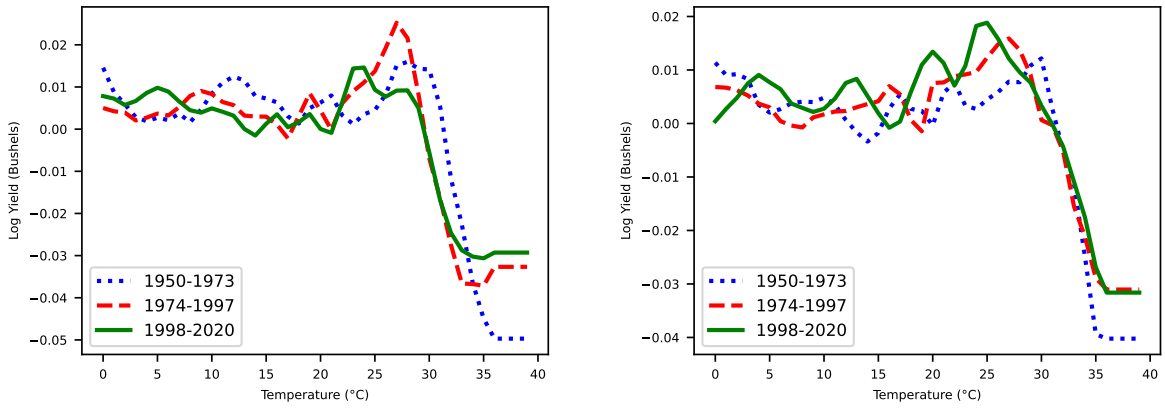


(a) Impact of Temperature on Corn Yield (b) Impact of Temperature on Soybean Yield

Figure 3 displays the estimated functional slope coefficient β corresponding to our functional PLS (red cure) and step function approximation (black dash) 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 4 reports the estimated functional slope coefficient splitting the data into three subsamples: 1950-1973 (blue dot), 1974-1997 (red dash), and 1998-2020 (green 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. The mitigation of extreme temperature effects may come from two sources: the adaptation and the CO_2 fertilization. The CO_2 fertilization effects observed in our sample are likely to be small; see

Figure 4: Adaptation effects in nonlinear relationship between temperature and crop yields.



(a) Impact of Temperature on Corn Yield (b) Impact of Temperature on Soybeans Yield

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 US agriculture which has also been reported in Burke and Emerick (2016) using the aggregate linear regression analysis without accounting for nonlinearities.

6 Conclusions

This paper proposes a new formulation of the functional PLS estimator related to the conjugate gradient method applied to an ill-posed inverse problem with a self-adjoint operator. We provide the first optimality result for functional PLS and consider a rate-adaptive data-driven early stopping rule to select the optimal number of functional components. The

estimator has good estimation and prediction properties for a smaller number of principal components than PCA and the data-driven early stopping rule performs well in simulations. We find in an empirical application that the nonlinear temperature effects on crop yields have slightly decreased since 1950, especially for extreme temperatures. This provides some additional evidence of the adaptation of US agriculture to climate change.

References

Andrii Babii and Jean-Pierre Florens. Is completeness necessary? Estimation in nonidentified linear models. *arXiv preprint arXiv:1709.03473*, 2017.

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

Gilles Blanchard, Marc Hoffmann, and Markus Reiß. Optimal adaptation for early stopping in statistical inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3): 1043–1075, 2018.

Mélanie Blazère, Fabrice Gamboa, and Jean-Michel Loubes. PLS: a new statistical insight through the prism of orthogonal polynomials. *arXiv preprint arXiv:1405.5900*, 2014a.

Mélanie Blazère, Fabrice Gamboa, and Jean-Michel Loubes. A unified framework to study the properties of the PLS vector of regression coefficients. In *International Conference on Partial Least Squares and Related Methods*, pages 227–237. Springer, 2014b.

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 and Jan Johannes. Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, 101(2):395–408, 2010.

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.

Laurent Cavalier. Inverse problems in statistics. In *Inverse Problems and High-Dimensional Estimation: Stats in the Château Summer School, August 31-September 4, 2009*, pages 3–96. Springer, 2011.

Fabienne Comte and Jan Johannes. Adaptive functional linear regression. *The Annals of Statistics*, 40(6):2765–2797, 2012.

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

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. Springer Science & Business Media, 1996.

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

Antonio Gasparrini, Yuming Guo, Masahiro Hashizume, Eric Lavigne, Antonella Zanobetti, Joel Schwartz, Aurelio Tobias, Shilu Tong, Joacim Rocklöv, Bertil Forsberg, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, 386(9991):369–375, 2015.

Naveen Gupta, S Sivananthan, and Bharath K Sriperumbudur. Convergence analysis of kernel conjugate gradient for functional linear regression. *arXiv preprint arXiv:2310.02607*, 2023.

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 linear ill-posed problems*. Pitman Research Notes in Mathematics Series, 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.

Marc Hoffmann and Markus Reiss. Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, (1):310–336, 2008.

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). 2021.

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

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

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.

Jussi Klemelä and Enno Mammen. Empirical risk minimization in inverse problems. *The Annals of Statistics*, 38(1):482–511, 2010.

Rainer Kress. *Linear Integral Equations*, volume 82. Springer Science & Business Media, 1999.

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

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.

Cristian Preda and Gilbert Saporta. Clusterwise PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 49(1):99–108, 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.

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.