2018

Partially Linear Functional Additive Models for Multivariate Functional Data

Raymond K.W. Wong  
*Texas A&M University*

Yehua Li  
*Iowa State University, yehuali@iastate.edu*

Zhengyuan Zhu  
*Iowa State University, zhuz@iastate.edu*

Follow this and additional works at: [https://lib.dr.iastate.edu/stat_las_pubs](https://lib.dr.iastate.edu/stat_las_pubs)  
Part of the [Multivariate Analysis Commons](https://lib.dr.iastate.edu/stat_las_pubs), and the [Statistical Models Commons](https://lib.dr.iastate.edu/stat_las_pubs)

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/stat_las_pubs/126](https://lib.dr.iastate.edu/stat_las_pubs/126). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Partially Linear Functional Additive Models for Multivariate Functional Data

Abstract
We investigate a class of partially linear functional additive models (PLFAM) that predicts a scalar response by both parametric effects of a multivariate predictor and nonparametric effects of a multivariate functional predictor. We jointly model multiple functional predictors that are cross-correlated using multivariate functional principal component analysis (mFPCA), and model the nonparametric effects of the principal component scores as additive components in the PLFAM. To address the high dimensional nature of functional data, we let the number of mFPCA components diverge to infinity with the sample size, and adopt the COmponent Selection and Smoothing Operator (COSSO) penalty to select relevant components and regularize the fitting. A fundamental difference between our framework and the existing high dimensional additive models is that the mFPCA scores are estimated with error, and the magnitude of measurement error increases with the order of mFPCA. We establish the asymptotic convergence rate for our estimator, while allowing the number of components diverge. When the number of additive components is fixed, we also establish the asymptotic distribution for the partially linear coefficients. The practical performance of the proposed methods is illustrated via simulation studies and a crop yield prediction application.

Keywords
Additive model, Functional data, Measurement error, Reproducing kernel Hilbert space, Principal component analysis, Spline

Disciplines
Multivariate Analysis | Statistical Models | Statistics and Probability

Comments

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/stat_las_pubs/126
Partially Linear Functional Additive Models for Multivariate Functional Data

Raymond K. W. Wong, Yehua Li & Zhengyuan Zhu


To link to this article: https://doi.org/10.1080/01621459.2017.1411268

View supplementary material

Accepted author version posted online: 15 Jan 2018.

Submit your article to this journal

Article views: 148

View Crossmark data
Partially Linear Functional Additive Models for Multivariate Functional Data

Raymond K. W. Wong\textsuperscript{1}, Yehua Li\textsuperscript{2} and Zhengyuan Zhu\textsuperscript{2}

\textsuperscript{1}Department of Statistics, Texas A&M University, College Station, TX 77843
\textsuperscript{2}Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011

Abstract

We investigate a class of partially linear functional additive models (PLFAM) that predicts a scalar response by both parametric effects of a multivariate predictor and nonparametric effects of a multivariate functional predictor. We jointly model multiple functional predictors that are cross-correlated using multivariate functional principal component analysis (mFPCA), and model the nonparametric effects of the principal component scores as additive components in the PLFAM. To address the high dimensional nature of functional data, we let the number of mFPCA components diverge to infinity with the sample size, and adopt the COmponent Selection and Smoothing Operator (COSSO) penalty to select relevant components and regularize the fitting. A fundamental difference between our framework and the existing high dimensional additive models is that the mFPCA scores are estimated with error, and the magnitude of measurement error increases with the order of mFPCA. We establish the asymptotic convergence rate for our estimator, while allowing the number of components diverge. When the number of additive components is fixed, we also establish the asymptotic distribution for the partially linear coefficients. The practical performance of the proposed methods is illustrated via simulation studies and a crop yield prediction application.

Key Words: Additive model; Functional data; Measurement error; Reproducing kernel Hilbert space; Principal component analysis; Spline.

Short title: Partially Linear Functional Additive Models.
1 Introduction

As new technology being increasingly used in data collection and storage, many variables are continuously monitored over time and become multivariate functional data (Ramsay and Silverman, 2005; Zhou et al., 2008; Kowal et al., 2017). Extracting useful information from such data for further regression analysis has become a challenging statistical problem. There has been significant amount of recent work devoted to regression models with functional predictors and the most popular model is the functional linear model (James, 2002; Cardot et al., 2003; Müller and Stadtmüller, 2005; Cai and Hall, 2006; Crainiceanu et al., 2009; Li et al., 2010; Cai and Yuan, 2012), where the scalar response variable is assumed to depend on an $L^2$ inner product of the functional predictor with an unknown coefficient function.

Functional data are infinite dimensional vectors in a functional space (Hsing and Eubank, 2015). Due to the richness of information in such data, a simple linear model is often found inadequate and many researchers have investigated nonlinear functional regression models. The most widely used approach is to project functional data into a low-rank functional subspace and use the projections as predictors in a nonlinear model (James and Silverman, 2005; Li and Hsing, 2010a; Yao et al., 2016). The most popular and best understood dimension reduction tool for functional data is the functional principal component analysis (FPCA) (Yao et al., 2005; Hall et al., 2006; Li and Hsing, 2010b). A recent development in nonlinear functional regression model is the functional additive model (Müller and Yao, 2008; Zhu et al., 2014), where FPCA scores are used as predictors in an additive model.

Our research is motivated by a crop yield prediction application in agriculture. Agriculture is a major industry in the U.S., the source of livelihood for millions of farmers and a vital contributor to global food security. Getting timely and reliable predictions on crop production is crucial for planners and policy makers to create appropriate strategies for the
storage, distribution, and trade of agricultural products. The US National Agricultural Statistical Service is the federal agency responsible for providing such statistics to the public, and their in-season crop yield forecast is primarily based on survey data. It is well known that weather has a significant impact on crop yield, and statistical models can be used to relate weather forecast to crop yield prediction (Cadson et al., 1996; Hansen, 2002; Prasad et al., 2006; Lobell and Burke, 2010). Since measurements of meteorological variables, such as maximum and minimum temperatures, are typically available on a daily basis and their effects on yield vary at different growing stage of the crop, it is natural to treat them as functional predictors. Besides the functional predictors, scalar predictors, such as crop management methods, also have a great impact on yield and need to be included in the prediction model.

We propose a partially linear functional additive model (PLFAM) to predict a scalar response variable using both scalar and functional predictors. We use such a model to predict crop yield using the temperature trajectories. Such a model is of fundamental importance in plant science and agricultural economics: it advances our understanding of the relationship between weather conditions and crop yield, help to evaluate the impact of climate change on crop production and assist farmers and stake holders to better predict the future prices of agricultural commodity products and plan their actions accordingly. In many applications including our motivating data example, the functional predictors are strongly correlated to each other. To extract information more efficiently, we jointly model these predictors as a multivariate functional predictor, and perform dimension reduction using multivariate functional principal component analysis (mFPCA) (Ramsay and Silverman, 2005; Chiou et al., 2014). The proposed PLFAM includes the parametric effects of the scalar predictors and additive nonparametric effects of the mFPCA scores. To automatically select significant additive components, we impose COSSO penalties (Lin and Zhang, 2006) to the component
functions and estimate the model in a reproducing kernel Hilbert space (RKHS) framework.

Our approach is different from that of Zhu et al. (2014) in a few important perspectives. On the methodology side, we consider multiple functional predictors, extract informative signals from the functional predictors using mFPCA, and we adopt a semiparametric partially linear structure in our model to take into account the effects of scalar predictors. On the theory side, we allow the number of additive components in the model to diverge to infinity with the sample size, to acknowledge the fact that functional data have infinite number of principal components. Our theory is fundamentally different from those in the high dimensional additive model literature, since our predictors in the additive model are estimated mFPCA scores that are contaminated with measurement errors (Carroll et al., 2006). As we show, the magnitude of measurement error gets higher for higher order principal components. In contrast, Zhu et al. (2014) only allow finite number of principal components in their model. To bound the effect of measurement errors, they also impose some very restrictive conditions which, in effect, limit their estimator in a finite dimensional subspace of the Sobolev space. Our results, on the other hand, does not rely on such artificial assumptions.

The rest of the paper is organized as follows. We describe the model and assumptions in Section 2 and the estimation procedure in Section 3. In Section 4 we investigate the asymptotic properties of the proposed estimator. We illustrate the proposed method with simulation studies in Section 5 and apply it to the motivating data example in Section 6. Some final remarks are collected in Section 7. Technical proofs and additional numerical results are relegated to the supplementary material.
2 Model and Assumptions

Let $Y$ be a scalar random variable associated with a predictor $Z \in \mathbb{R}^p$ and a multivariate functional predictor $X = (X_1, \ldots, X_d)^T$, where $p$ and $d$ are positive integers, and $X_j(t)$ is a stochastic process defined on the time domain $T_j$ for $j = 1, \ldots, d$. For simplicity, we focus on the case $T_j \equiv T$, but having different domains does not affect our methodological nor theoretical developments. Let $\{z_i, x_i\}_{i=1}^n$ be i.i.d. copies of $\{Z, X\}$. Their relationship with the response $\{y_i\}_{i=1}^n$ are modeled as

$$y_i = m(z_i, x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $m$ is the regression function and $\varepsilon_i$ are zero mean errors independent with $\{x_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$. Further we assume $\text{var}(\varepsilon_i) = \sigma^2_{\varepsilon}/\pi_i$, where $\sigma^2_{\varepsilon}$ is an unknown variance parameter and $\pi_i$'s are known positive weights. In our application, the response $y_i$ is the averaged crop yield per acre obtained from a survey, and $\pi_i$ is proportional to the size of the harvest land.

2.1 Multivariate functional principal component analysis

We assume that, with probability 1, the trajectory of $X_j$ is contained in a Hilbert space $X_j$, with inner product $\langle \cdot, \cdot \rangle_{X_j}$ and norm $\| \cdot \|_{X_j}$. We will focus on the case that $X_j$'s are $L^2$ functional spaces and the inner products are $\langle f, g \rangle_{X_j} = \int_T f(t)g(t)dt$ for any $f, g \in X_j$. Let $X = \bigoplus_{j=1}^d X_j$ be the direct sum of the functional spaces, which is a bigger Hilbert space equipped with the induced inner product and norm, i.e. $\langle x_1, x_2 \rangle_X = \sum_{j=1}^d \langle x_{1j}, x_{2j} \rangle_{X_j}$ and $\|x_i\|_X = \langle x_i, x_i \rangle_X^{1/2}$ for any $x_i = (x_{i1}, \ldots, x_{id})^T \in X$, $i = 1, 2$.

Define the mean function of the multivariate functional predictor as $\mu(t) = \mathbb{E}\{X(t)\} = \{\mu_1(t), \ldots, \mu_d(t)\}^T$ where $\mu_j(t) = \mathbb{E}\{X_j(t)\}$. The cross-covariance function between $X_j$ and $X_{j'}$ is $C_{jj'}(s,t) = \mathbb{E}\{(X_j(s) - \mu_j(s))(X_{j'}(t) - \mu_{j'}(t))\}$, and the covariance of $X$ is a $d \times d$
matrix of cross-covariance functions

\[ C(s, t) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))^\top] = \{C_{jj'}(s, t)\}_{j, j' = 1}^d. \]

We assume that \( C \) defines a bounded, self-adjoint, positive semi-definite integral operator (Hsing and Eubank, 2015). Standard operator theory warrants a spectral decomposition

\[ C(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k^\top(t), \]

where \( \lambda_1 \geq \lambda_2 \geq \ldots > 0 \) are the eigenvalues and \( \psi_k = (\psi_{k1}, \ldots, \psi_{kd})^\top \in \mathbb{X} \) are the corresponding eigenfunctions such that \( \langle \psi_k, \psi_{k'} \rangle_{\mathbb{X}} = \int_T \psi_k(t)^\top\psi_{k'}(t)dt = I(k = k') \). By a standard Karhunen-Loève expansion

\[ X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \psi_k(t), \]

where \( \xi_k = \langle X - \mu, \psi_k \rangle_{\mathbb{X}} \) are zero-mean random variables with \( \mathbb{E}(\xi_k \xi_{k'}) = \lambda_k I(k = k') \). The variables \( \xi_k \) are the mFPCA scores of \( X \).

### 2.2 Partially linear functional additive model

Direct estimation of Model (1) suffers from the “curse-of-dimensionality” and is unpractical. Many popular alternative approaches are based on dimension reduction through FPCA and the effects of the functional predictors are modeled through their principal component scores, including the functional linear models (FLM) and the functional additive models (FAM). Our PLFAM model follows a similar strategy and can be considered as a special case of Model (1) with additional structural assumptions.

We denote the sequence of mFPCA scores of \( x_i \) by \( \xi_{i,\infty} = (\xi_{i1}, \xi_{i2}, \ldots)^\top \). Even though
in theory there are infinite number of principal components, the number of eigenfunctions estimated from the sample is at most \( n - 1 \), and as shown in our theory in Section 4.1 even fewer of eigenfunctions are estimated consistently. For these practical reasons, it is a common practice to only use the low-order FPCA scores as predictors in a regression. Denote the truncated mFPCA scores as \( \xi_i = (\xi_{i1}, \ldots, \xi_{is})^\top \), with a positive integer \( s \). To avoid possible scale issues, we instead use the standardized version \( \zeta_{ik} = \Phi(\lambda_k^{-1/2}\xi_{ik}) \), where \( \Phi(\cdot) \) is a continuously differentiable map from \( \mathbb{R} \) to \([0, 1]\). We let \( \Phi(\cdot) \) be the standard Gaussian cumulative distribution function (CDF) in all of our numerical studies. When the distribution of \( \xi \) is close to Gaussian, \( \zeta \) is approximately uniform in \([0, 1]\), which is convenient for nonparametric modeling on the effect of \( \zeta \). Other continuous CDFs can also be used as \( \Phi(\cdot) \), such as the logistic function. Write \( \zeta_{i,\infty} = (\zeta_{i1}, \zeta_{i2}, \ldots) \) and \( \zeta_i = (\zeta_{i1}, \ldots, \zeta_{is}) \).

Assuming that all useful information in the multivariate functional predictor is contained in the first \( s \) principal components, which are related to the response in an additive form, and the covariate effect is linear, then model (1) becomes the following Partially Linear Functional Additive Model (PLFAM)

\[
y_i = m_0(u_i, \zeta_i) + \varepsilon_i = u_i^\top \theta_0 + f_0(\zeta_i) + \varepsilon_i = u_i^\top \theta_0 + \sum_{k=1}^s f_{0k}(\zeta_{ik}) + \varepsilon_i, \tag{2}
\]

where \( \theta_0 \in \mathbb{R}^{p+1} \) and \( u_i = (1, z_i^\top)^\top \). Model (2) bears the functional additive model (FAM) of Müller and Yao (2008) and Zhu et al. (2014) as a special case when the functional predictor \( X(t) \) is univariate \((d = 1)\) and there are no scalar covariates. The partially linear structure is widely used in many popular semiparametric models because it combines the flexibility of nonparametric modeling with easy interpretation of the covariate effects (Carroll et al., 1997; Liu et al., 2011; Wang et al., 2014). In practice, \( u \) can include interactions, quadratic terms and any other low order nonlinear terms as long as their effects are interpretable and
parametric. We show in Section 4 the estimated partially linear coefficient \( \hat{\theta} \) (also referred to as the parametric component of the model) is \( \sqrt{n} \)-consistent and has an asymptotically normal distribution, despite the existence of nonparametric components which converge in a slower rate. This is particularly useful if inference on the parametric effects is of primary interest in the study.

Following Lin and Zhang (2006) and Zhu et al. (2014), we assume that each \( f_{0k} \) belongs to a reproducing kernel Hilbert space (RKHS). We refer interested readers to Wahba (1990) for an introduction of RKHS for penalized regression. The most widely used RKHS is the Sobolev Hilbert space. In such context, an \( l \)-th order Sobolev Hilbert space \( \mathbb{F}^{(l)}[0,1] \) is the collection of functions on \([0,1]\) whose first \((l-1)\)-th derivatives are absolutely continuous and the \(l\)-th derivative belongs to \( L^2[0,1] \), and the corresponding norm is chosen as

\[
\|g\|^2 = \sum_{v=0}^{l-1} \left( \int_0^1 g^{(v)}(t) dt \right)^2 + \int_0^1 g^{(l)}(t)^2 dt \quad \text{for any } g \in \mathbb{F}^{(l)}[0,1].
\]

Let \( \mathbb{F}_k, k = 1, \ldots, s, \) be a sequence of \( l \)-th order Sobolev spaces on \([0,1]\) with reproducing kernels \( R_k \), and we assume \( f_{0k} \in \mathbb{F}_k \). However, the fact that constant functions belong to each \( \mathbb{F}_k \) leads to an identifiability issue. To provide an identifiable parametrization, we note that each \( \mathbb{F}_k \) has an orthogonal decomposition \( \mathbb{F}_k = \{1\} \oplus \overline{\mathbb{F}}_k \) where \( \{1\} \) is the space of all constant functions. From now on, we assume \( m_0 \in \mathbb{M} \equiv \mathbb{I} \oplus \sum_{k=1}^s \overline{\mathbb{F}}_k \), where \( f_{0k} \in \overline{\mathbb{F}}_k \) for \( k = 1, \ldots, s \), and \( \mathbb{I} = \{u^\top \theta : \theta \in \mathbb{R}^{p+1}\} \). For the rest of the paper, we focus on the second order Sobolev space with \( l = 2 \).
3 Estimation and Computation

3.1 Estimation in mFPCA

To start with, we assume that the trajectories of \( x_i(t) \)'s are fully observed. Then the mean and covariance of \( X \) can be estimated by

\[
\hat{\mu}(t) = n^{-1} \sum_{i=1}^{n} x_i(t), \quad \hat{C}(s,t) = n^{-1} \sum_{i=1}^{n} \{ x_i(s) - \hat{\mu}(s) \} \{ x_i(t) - \hat{\mu}(t) \}^\top.
\]  

(3)

Since \( \hat{C} \) has rank \( n - 1 \), it has a spectral decomposition \( \hat{C}(s,t) = \sum_{k=1}^{n-1} \hat{\lambda}_k \hat{\psi}_k(s) \hat{\psi}_k^\top(t) \), where \( \hat{\lambda}_k \) and \( \hat{\psi}_k(t) \) are the sample eigenvalues and eigenfunctions. The estimated mFPCA scores are

\[
\hat{\xi}_{ik} = \langle x_i, \hat{\psi}_k \rangle_X = \sum_{j=1}^{d} \int_{T} x_{ij}(t) \hat{\psi}_{kj}(t) dt, \quad \hat{\zeta}_{ik} = \Phi(\hat{\lambda}_k^{-1/2} \hat{\xi}_{ik}), \quad k = 1, \ldots, d.
\]  

(4)

In practice, we only have discrete noisy observations on \( x_i \)

\[
w_{ijk} = x_{ij}(t_{ijk}) + e_{ijk}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, d, \quad k = 1, \ldots, N_{ij},
\]

where \( e_{ijk} \)'s are independent measurement errors with mean 0 and variance \( \sigma_{e,j}^2, \ j = 1, \ldots, d \).

We will focus on the case where dense measurements are made on each curve such that each functional predictor can be effectively recovered by passing a linear smoother through the discrete observations. Let the recovered functions be \( \tilde{x}_{ij}(t) = \mathcal{S}(t; t_{ij}) w_{ij}, \) where \( w_{ij} = (w_{ij1}, \ldots, w_{ijn_{ij}})^\top \) and \( \mathcal{S}(t; t_{ij}) \) is a linear smoother depending on the design points \( t_{ij} = (t_{ij1}, \ldots, t_{ijn_{ij}})^\top \), e.g. local polynomial or regression splines. The eigenvalues, eigenfunctions and mFPCA scores are estimated by replacing \( x_{ij}(t) \) with \( \tilde{x}_{ij}(t) \) in (3) and (4).

For univariate functional data, this pre-smoothing approach is theoretically justified by \cite{Hall2006}, who show that, when \( \mathcal{S} \) is a local linear smoother and \( N_{\text{min}} = \min_{i,j} N_{ij} > 8 \).
the error incurred by approximating \( x_{ij}(t) \) with \( \tilde{x}_{ij}(t) \) is negligible in \( \hat{\lambda}_k \) and \( \hat{\psi}_k \); \cite{Li2010} further show that this approximation error is negligible to \( \hat{\xi}_{ik} \) if \( N_{\min} > Cn^{5/4} \).

As commented in \cite{Li2010}, there are two sources of error in \( \hat{\xi}_{ik} \): the error caused by approximating \( x_{ij} \) with \( \tilde{x}_{ij} \) and the error in \( \hat{\psi}_k \). If the first type of error prevails, regression analysis using \( \hat{\xi}_{ik} \) will be inconsistent even for linear models. The second type of error, on the other hand, is diminishing to zero as \( n \to \infty \). There are mFPCA methodologies for sparse multivariate functional data (see e.g. \cite{Chiou2014}), but how to consistently estimate FAM or PLFAM when the estimated scores are contaminated with non-diminishing errors is not clear and calls for further research.

In all of our numeric studies, we smooth and register each \( x_{ij} \) on B-splines, pool spline coefficients for each component in \( x_i \) into a longer vector, then the operator \( \hat{C} \) is represented as a high dimensional matrix, and the mFPCA problem reduces to a multivariate PCA problem. For detailed algorithm, we refer the readers to Section 8.5 in \cite{RamsaySilverman2005}.

### 3.2 Estimation of PLFAM with COSSO penalty

Let \( \hat{\zeta}_i = (\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{is})^T \) be a vector of standardized mFPCA scores for \( x_i \) estimated using the procedure in Section 3.1. Since there are potentially infinite number of principal components for \( X \), we choose the truncation point \( s \) to be a large positive number and use a penalized regression method to select the relevant components.

The proposed estimator \( \hat{m} \) is the minimizer of the following penalized loss \( \ell_w(m) \) with respect to \( m \in M \). The loss function is defined as

\[
\ell_w(m) = \frac{1}{n} \sum_{i=1}^{n} \pi_i \{ y_i - m(u_i, \hat{\zeta}_i) \}^2 + \tau_n^2 J(m),
\]

where \( \pi_i \) are the survey weights defined in [1]. Here \( \tau_n^2 \) is a tuning parameter and \( J(m) = \).
$\sum_{i=1}^{s} \|P_k m\|$ with $P_k$ being the projection operator to $\tilde{F}_k$. The penalty $J(m)$ is first proposed in the COSSO framework [Lin and Zhang (2006)] for simultaneous estimation and selection of the nonparametric functions $f_{0k}$'s.

Following [Lin and Zhang (2006)], we minimize (5) by iteratively minimizing its equivalent form

$$\frac{1}{n} \sum_{i=1}^{n} \pi_i \{y_i - m(u_i, \tilde{\zeta}_i)\}^2 + \kappa_0 \sum_{k=1}^{s} \phi_k^{-1} \|P_k m\|^2 + \kappa \sum_{k=1}^{s} \phi_k \tag{6}$$

over $\phi = (\phi_1, \ldots, \phi_s)^T \in [0, \infty)^s$ and $m \in M$, where $\kappa_0 > 0$ is a pre-determined constant and $\kappa$ is a tuning parameter.

The relationship between (5) and (6) is stated in the following lemma, which is an extension of Lemma 2 in [Lin and Zhang (2006)] to partially linear additive model under a weighted least square loss. Its proof is omitted for brevity.

**Lemma 1 (Lemma 2 of [Lin and Zhang (2006)])** Set $\kappa = \tau_4 n / (4 \kappa_0)$. (i) If $\hat{m}$ minimizes (5), set $\hat{\phi}_k = \kappa_0^{-1/2} \kappa^{-1/2} \|P_k \hat{m}\|$; then the pair $(\hat{\phi}, \hat{m})$ minimizes (6). (ii) If $(\hat{\phi}, \hat{m})$ minimizes (6), then $\hat{m}$ minimizes (5).

By representer theorem, the minimizer $\hat{m}(u, \zeta)$ takes the form $u^T \theta + \sum_{k=1}^{s} \phi_k \sum_{i=1}^{n} a_i R_k(\tilde{\zeta}_ik, \zeta_k)$, for $u = (1, z_1, \ldots, z_p)^T \in \mathbb{R}^{p+1}$, $(\zeta_1, \ldots, \zeta_s)^T \in \mathbb{R}^s$, where $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ is a vector of unknown parameters. Then, minimization of (6) is equivalent to minimizing

$$\frac{1}{n} \|\Pi^{1/2} (y - U \theta - \sum_{k=1}^{s} \phi_k R_k a)\|_E^2 + \kappa_0 \sum_{k=1}^{s} \phi_k a^T R_k a + \kappa \sum_{k=1}^{s} \phi_k, \tag{7}$$

where $\| \cdot \|_E$ represents the Euclidean norm, $\Pi = \text{diag}\{\pi_1, \ldots, \pi_n\}$, $y = (y_1, \ldots, y_n)^T$, $U = [u_{ij}]_{i=1,\ldots,n,j=1,\ldots,p+1}$ is a $n \times (p+1)$ design matrix and $R_k = [R_k(\tilde{\zeta}_ik, \zeta_k)]_{i,j=1,\ldots,n}$ is a $n \times n$ matrix for $k = 1, \ldots, s$. For a fixed $\phi$, minimizing (7) with respect to $(\theta, a)$ is similar to solving a weighted ridge regression. For fixed $\theta$ and $a$, let $D$ be the $n \times s$ matrix with the
$k$-th column being $R_k a$, then minimization of (7) with respect to $\phi \in [0, \infty)^s$ becomes

$$\min \frac{1}{n} \left[ \phi^\top D^\top \Pi D \phi - 2 \left\{ D^\top \Pi (y - U \theta) - \frac{1}{2} n \kappa_0 D^\top a \right\}^\top \phi \right]$$

subject to $\sum_{k=1}^s \phi_k < G$ and $\phi \in [0, \infty)^s$,

for some $G > 0$, which is a typical quadratic programming. The practical minimization of (5) is done by iterating over these two minimizations by fixing $(\theta, a)$ and $\phi$ in turn. The algorithm starts with solving $(\theta, a)$ while fixing $\phi = 1$. Empirically, the objective function decreases quickly in the first iteration, which was also observed in Lin and Zhang (2006) and Storlie et al. (2011). To reduce the computational cost, we limit the number of iterations, and follow a one-step update procedure similar to Lin and Zhang (2006).

As discussed in Lin and Zhang (2006) and Storlie et al. (2011), $\kappa_0$ can be fixed at any positive value. We select $\kappa_0$ that minimizes the GCV of the partial spline problem when $\phi = 1$. Let $\hat{\phi}^{(\tau_n)} = (\hat{\phi}_1^{(\tau_n)}, \ldots, \hat{\phi}_n^{(\tau_n)})^\top$, $\hat{a}^{(\tau_n)}$ and $\hat{\theta}^{(\tau_n)}$ be the minimizer of (7) for a fixed $\tau_n$. To select the smoothing parameter $\tau_n$ (or equivalently $G$), we minimize the Bayesian information criterion $n \log(RSS_w(\tau_n)/n) + df(\tau_n) \log(n)$, where the effective degree of freedom $df(\tau_n)$ is the trace of the smoothing matrix in the partial spline problem (7) when $\phi$ is set to $\hat{\phi}^{(\tau_n)}$, and the weighted residual sum of squares is $RSS_w(\tau_n) = \frac{n}{\sum_{i=1}^n \pi_i} \| \Pi^{1/2} (y - U \hat{\theta}^{(\tau_n)} - \sum_{k=1}^s \hat{\phi}_k^{(\tau_n)} R_k a^{(\tau_n)}) \|_E^2$.

4 Theoretical Results

4.1 Basic results for mFPCA

By the theory of Dauxois et al. (1982), $\| \hat{C} - C \|_{op} = O_p(n^{-1/2})$ where the operator norm is defined as $\| A \|_{op} = \sup_{x \in X} \frac{\| A x \|_X}{\| x \|_X}$ for any bounded linear operator $A$ on $X$. To derive the
asymptotic expansion for \( \hat{\xi}_{ik} \), we use the asymptotic expansion of \( \hat{\lambda}_k \) and \( \hat{\psi}_k \) provided by Hsing and Eubank (2015), which is a generalization of those by Hall and Hosseini-Nasab (2006) for univariate functional data to more general Hilbert space random variables. We adopt the following assumptions:

**Assumption 1** (Cai and Hall (2006))

\[ C^{-1} \lambda_k^{-\alpha} \leq \lambda_k \leq C \lambda_k^{-\alpha}, \quad \lambda_k - \lambda_{k+1} \geq C^{-1} \lambda_k^{-1-\alpha}, \quad k = 1, 2, \ldots \]

(8)

To ensure that \( \sum_{k=1}^{\infty} \lambda_k < \infty \), we assume that \( \alpha > 1 \).

**Assumption 2**

\[ \mathbb{E}(\|X\|_4^4) < \infty \quad \text{and there exists a constant } C_\xi > 0 \text{ such that } \mathbb{E}(\xi_k^2 \xi_{k'}^2) \leq C_\xi \lambda_k \lambda_{k'} \quad \text{and } \mathbb{E}(\xi_k^2 - \lambda_k)^2 < C_\xi \lambda_k^2 \text{ for all } k \text{ and } k' \neq k. \]

The polynomial decay rate described in Assumption 1 is a slow decay rate assumption on the eigenvalues and allows \( X(t) \) to be flexibly modeled as a multivariate \( L^2 \) process without strong constraints on the roughness of its sample path. Assumption 2 is a weak moment condition on the functional predictors and is satisfied if \( X(t) \) is a multivariate Gaussian process. Both assumptions are widely used in the functional linear model literature (Cai and Hall, 2006; Cai and Yuan, 2012; Hsing and Eubank, 2015). Define \( \delta_k = \frac{1}{2} \min_{k' \neq k} |\lambda_{k'} - \lambda_k| \), which is no less than \( \frac{1}{2} C^{-1} \lambda^{-1-\alpha} \) under condition (8) and denote \( \Delta = n^{1/2}(\hat{C} - C) \). By Dauxois et al. (1982), \( \Delta \) converges weakly to a Gaussian variable in the space of linear operators and hence \( \|\Delta\|_{op} = O_p(1) \).

**Proposition 1** (Transformed FPC scores) Suppose the transformation function \( \Phi(\cdot) \) has bounded derivative. Under Assumptions 1 and 2, there is a constant \( C > 0 \) such that

\[ \mathbb{E}(\hat{\xi}_{ik} - \xi_{ik})^2 \leq C k^2 / n \quad \text{uniformly for } k \leq J_n, \quad \text{where } J_n = \lfloor (2C_\lambda \|\Delta\|_{op})^{-1/(1+\alpha)} n^{1/(2+2\alpha)} \rfloor. \]
The proof of Proposition 1 is given in Appendix A in the supplementary material. It implies that the estimation error of the principal component score increases as the order of the principal component gets higher. Interestingly, the estimation error is of order \(O_p(n^{-1/2})\), which does not depend on the decay rate \(\alpha\) of the eigenvalues.

4.2 Asymptotic theory for PLFAM

For simplicity, we assume that \(\pi_i = 1\) for \(i = 1, \ldots, n\). We begin by introducing several notations. We write \(P_n\) as the empirical distribution of \((Z, \zeta)\). That is, \(P_n = \sum_{i=1}^{n} \frac{\delta_{z_i, \zeta_i}}{n}\), where \(\delta_{z, \zeta}\) is the delta function at \((z, \zeta)\). Moreover, we denote the distribution of \((Z, \zeta)\) by \(P\). We define the corresponding (squared) empirical norm and inner product as

\[
\|m_1\|^2 = \int m_1^2 dP_n \quad \text{and} \quad (m_1, m_2)_n = \int m_1 m_2 dP_n, \quad \text{for any } m_1, m_2 \in \mathbb{M}.
\]

These notations are extended to measurement errors \(\{\varepsilon_i\}\). For instance, \((\varepsilon, m_1)_n = \sum_{i=1}^{n} \varepsilon_i m_1(u_i, \zeta_i)/n\). Moreover, we write the Euclidean norm for vector as \(\|\cdot\|_E\). To derive the asymptotic properties, we assume that the parametric component is identifiable. More specifically, \(\Sigma = \int uu^t dP\) is non-singular.

**Theorem 1** Suppose, for some \(\beta > 0\), \(\mathbb{E}(\hat{\zeta}_{ik} - \zeta_{ik})^2 \leq C n^{-1} k^{2\beta}\) uniformly for all \(k \leq s\). Assume \(0 < J(m_0) < \infty\), \(\Sigma\) is non-singular and \(\tau_n^{-1} = O_p(\min\{n^{2/5} s^{-6/5}, n^{1/2} s^{-(1/2+\beta)}\})\), we have \(\|\hat{m} - m_0\|_n = O_p(\tau_n)\) and \(J(\hat{m}) = O_p(1)\). If \(J(m_0) = 0\) and \(\tau_n \asymp n^{-1/4} s^3\), \(\|\hat{m} - m_0\|_n = O_p(n^{-1/2})\) and \(J(\hat{m}) = O_p(n^{-1/2} s^{-6})\).

**Remarks:**

1. Under the framework laid out in Assumptions 1 and 2 with \(s = O_p(n^{1/(2(1+\alpha))})\), we have \(\mathbb{E}(\hat{\zeta}_{ik} - \zeta_{ik})^2 \leq C n^{-1} k^2\) uniformly for all \(k \leq s\) followed from Proposition 1. The results in Theorem 1 can be further simplified by identifying \(\beta = 1\). In this case, if \(0 < J(m_0) < \infty\)
and $\tau^{-1} = \mathcal{O}_p(n^{2/5} s^{-6/5})$, we have $\|\hat{m} - m_0\|_n = \mathcal{O}_p(n^{-2/5} s^{6/5})$. If $s$ is fixed, $\|\hat{m} - m_0\|_n = \mathcal{O}_p(n^{-2/5})$ is the optimal nonparametric convergence rate assuming each $f_{0k}$ belongs to a second order Sobolev space.

2. Our result can be considered as an extension of Theorem 1 in [Zhu et al., 2014], where we allow $s \to \infty$ in a rate no faster than $\mathcal{O}_p(n^{1/(2(1+\alpha))})$. The reason for setting such a restriction on the rate of $s$ is that, in order to estimate the principal components consistently, we need the distance between two adjacent eigenvalues to be no smaller than $\|\hat{C} - C\|_{op}$. This is a fundamental difference with classic high dimensional additive models (Meier et al., 2009; Ravikumar et al., 2009; Liu et al., 2011; Wang et al., 2014).

3. The key issue in achieving consistent estimation of PLFAM is to bound the estimation error in $\hat{\zeta}_{ik}$. To achieve this goal, Zhu et al. (2014) assumed (see their Assumption 1)

$$\left| \frac{\partial f(\zeta)}{\partial \zeta_{ik}} \right| = |f'_k(\zeta_{ik})| \leq B_i \|f\|_2 \quad \text{with probability 1}$$

for some independent variables $\{B_i\}_{i=1}^n$ with $\mathbb{E}(B_i^2) < \infty$, where $\|\cdot\|_2$ is the $L_2(P)$-norm. This is a strong assumption that eliminates the possibility $f_k$ belonging to the space spanned by high order Fourier or Demmler-Reinsch basis functions. As an effect, their estimation is restricted in a low dimensional functional space. We, on the other hand, show in Lemma 2 that $\sup_{\zeta \in [0,1]} |f'_k(\zeta)|$ is bounded by the RKHS norm of $f_k$ for all $k \leq s$, and such a result help to control the error caused by the error-contaminated predictor $\hat{\zeta}_i$.

When $s$ is fixed, better asymptotic results can be derived for the regression coefficients.
\[ \gamma = (\gamma_1, \ldots, \gamma_p)^\top = (\theta_2, \ldots, \theta_{p+1})^\top. \]

Define
\[ w(\zeta) = (w_1(\zeta), \ldots, w_p(\zeta))^\top = \text{argmin}_{w_j \in \{1\} \oplus \sum_{k=1}^s \bar{F}_k} \mathbb{E}\| Z - w(\zeta) \|^2, \]
\[ \tilde{w}(z, \zeta) = (\tilde{w}_1(z, \zeta), \ldots, \tilde{w}_p(z, \zeta))^\top = z - w(\zeta), \]
\[ M = (M_{ij})_{i,j=1}^p, \quad \text{where} \quad M_{ij} = \int \tilde{w}_i \tilde{w}_j dP. \quad (9) \]

It is easy to see that \( w(\zeta) \) defines a additive regression of \( Z \) on \( \zeta \), and it can be considered as the projection of \( \mathbb{E}(Z|\zeta) \) on the additive regression space, and therefore
\[ \mathbb{E}\{ \tilde{w}^\top(z, \zeta) g(\zeta) \} = 0 \quad (10) \]

for any \( g(\zeta) = (g_1, \ldots, g_p)^\top(\zeta) \) such that \( g_j(\zeta) \in \{1\} \oplus \sum_{k=1}^s \bar{F}_k \) for \( j = 1, \ldots, p \).

**Theorem 2** Assume the conditions of Theorem 1 hold with \( s < \infty \) being fixed, \( \zeta \) has a non-degenerate joint density on \( [0, 1]^s \) which is bounded above and below, \( \tau_n = o_p(n^{-1/4}) \), and that \( M \) defined in (9) is non-singular. Then \( n^{1/2}(\hat{\gamma} - \gamma_0) \to \text{Normal}(0, M^{-1}(V_1 + V_2)M^{-1}) \) in distribution, where \( V_1 \) and \( V_2 \) are defined in (S.14) of the supplementary material.

**Remark:** As shown in our proof, \( V_1 = \text{cov}\{n^{1/2}(\varepsilon, \tilde{w})_n\} \), and \( M^{-1}V_1M^{-1} \) is the typical asymptotic covariance matrix of \( \hat{\gamma} \) in classic literature of partially linear additive model \( \text{(Wang et al. 2014)} \), where \( \zeta \) is directly observed. The covariance \( V_2 \) is the extra variation, caused by the estimation error in the FPCA score \( \hat{\zeta} \). The two sources of variation are asymptotically independent to each other because the model error \( \varepsilon \) is independent with the error in \( \hat{\zeta} \). A similar effect of FPCA estimation error was discovered by \( \text{(Li et al. 2010)} \), who investigated a simpler functional linear regression model and found that the FPCA error tends to inflate the asymptotic variance of the parametric component even if the functional predictors are fully observed. Our result in Theorem 2 shows the same phenomenon also

15
exists for nonlinear functional regression models such as the PLFAM.

5 Simulation study

We extend the simulation setting of Zhu et al. (2014) to a multivariate functional data setting with an additional vector predictor $Z$. The multivariate functional predictor is

$$x_i(t) = \{x_{i1}(t), x_{i2}(t)\}^\top$$

with

$$x_{i1}(t) = t + \sin(t) + \sum_{k=1}^{10} \xi_{ik}^{(1)} \psi_k^{(1)}(t), \quad x_{i2}(t) = t + \cos(t) + \sum_{k=1}^{10} \xi_{ik}^{(2)} \psi_k^{(2)}(t),$$

where $\xi_{ik}^{(1)} \sim N(0, \varsigma_{2k-1})$, $\xi_{ik}^{(2)} \sim N(0, \varsigma_k)$, $\varsigma_k = 45.25k^{-2}$, $\text{corr}(\xi_{ik}^{(j)}, \xi_{ik'}^{(j)}) = 0$ for $k' \neq k$, and $\psi_k^{(j)}(t) = (1/\sqrt{5}) \sin(\pi kt/10)$ for $t \in T = [0, 10]$, $j = 1, 2$. The equations above define the univariate Karhunen-Loève expansions for the two functional predictors respectively, scores within the same functional predictor are independent, however we allow the scores from different functional predictors to be cross-correlated. We let $\text{corr}(\xi_{ik}^{(1)}, \xi_{ik}^{(2)}) = \varrho$ for $k' = k$ and 0 otherwise, where $\varrho$ is a cross-correlation parameter between 0 and 1.

The mFPCA eigenfunctions are defined through an orthogonalization of the univariate eigenfunctions, as described in Proposition 5 in Happ and Greven (2017). More specifically, suppose the covariance matrix pooling all univariate FPCA scores has the eigenvalue decomposition $\text{var}\left(\{(\xi_i^{(1)})^\top, (\xi_i^{(2)})^\top\}\right) = P Q P^\top$, where $\xi_i^{(j)} = (\xi_{i1}^{(j)}, \ldots, \xi_{i10}^{(j)})^\top$, $j = 1, 2$, $Q = \text{diag}(\lambda_1, \ldots, \lambda_{20})$ and $P^\top P = I$. The $k$-th mFPCA score $\xi_{ik} \sim N(0, \lambda_k)$ is a linear function of the univariate scores $\{(\xi_i^{(1)})^\top, (\xi_i^{(2)})^\top\} p_k$, where $p_k = \{(p_k^{(1)})^\top, (p_k^{(2)})^\top\}^\top$ is the $k$-th column of $P$, and the corresponding mFPCA eigenfunction is $\psi_k(t) = (\psi_{k1}, \psi_{k2})^\top(t)$, where $\psi_{kj}(t) = \{(\psi^{(j)}(t))^\top p_k^{(j)}$, $\psi^{(j)}(t) = (\psi_{1}^{(j)}, \ldots, \psi_{10}^{(j)})^\top(t)$, $j = 1, 2$.

From model (2), we simulate 1000 i.i.d. copies of $\{Y, Z, X(\cdot)\}$, denoted as $\{y_i, z_i, x_i(\cdot)\}_{i=1}^{1000}$, with the first 200 used as training data and the rest as testing data. Observations on $x_i$ are
obtained on a regular grid of 100 points in \( T = [0,10] \) with independent measurement errors following \( N(0,0.2^2) \). For the regression function, we set \( f_0(\zeta) = f_{01}(\zeta_1) + f_{02}(\zeta_2) + f_{04}(\zeta_4) \), where \( f_{01}(\zeta_1) = 3\zeta_1 - 3/2, \ f_{02}(\zeta_2) = \sin\{2\pi(\zeta_2 - 1/2)\} \) and \( f_{04}(\zeta_4) = 8(\zeta_4 - 1/3)^2 - 8/9 \). There are only three non-zero additive component functions in our simulation: \( f_{0k}(\zeta_k) = 0 \) for \( k \notin \{1,2,4\} \). Moreover, we generate the vector predictor \( z_i \) independently from the bivariate uniform distribution over \([0,1]^2\). We consider two settings for the partially linear coefficient \( \theta_0 \): (I) \((1.4,0,0)^\top\) and (II) \((1.4,3,-4)^\top\) and two settings of the correlation parameter \( \varrho \): (i) 0.3 (low correlation) and (ii) 0.9 (high correlation). Combining different setups for \( \varrho \) and \( \theta_0 \), we have four settings: \{\{(i), (I)\}, \{(i), (II)\}, \{(ii), (I)\} and \{(ii), (II)\}\}. The errors \( \varepsilon_i \)'s in the regression model (2) are distributed independently as \( N(0,\sigma^2) \) with \( \sigma^2 \) being 1 for setting (I) and 1.9470 for (II) to achieve the signal-to-noise ratio (SNR) of approximately 2.2. The SNR is defined as \( \text{var}(m_0(\zeta))/\text{var}(\varepsilon) \). For simplicity, all sampling weights \( \pi_i \) are set to be 1. The simulation is repeated 200 times and we fit the following two models to each simulated data set: FAM of [Zhu et al. (2014)], which is also based on COSSO but ignores the effect of \( Z \), and the proposed PLFAM. Throughout this simulation study, \( s \) is chosen to recover at least 99.9% of the total variation in \( \{x_i\} \) and the COSSO tuning parameters are selected by the Bayesian information criterion.

Tables 1 and 2 summarize the results related to component function selection in FAM and PLFAM under the four settings. Due to space constraint, only percentages of model sizes up to 8 and selection percentages of the first 8 component functions are shown. In Table 2 Column “% correct set” corresponds to the percentages of fittings achieving exact selection of \( \hat{f}_1, \hat{f}_2 \) and \( \hat{f}_4 \), while Column “% super set” gives the percentages of fittings that include nonzero \( \hat{f}_1, \hat{f}_2 \) and \( \hat{f}_4 \). Despite a small tendency of over-selection, the COSSO component selection mechanism tends to select parsimonious models and, for each correct component function, the selection percentage is high.
To assess the estimation quality of \( f_{ok}'s \), Table 3 shows the averaged integrated squared errors (AISEs) of the first eight component functions and the overall function \( \hat{f} = \sum_{k=1}^{s} \hat{f}_k \) (without constant term). The integrated squared errors are defined as

\[
\text{ISE}(\hat{f}_k) = \int_0^1 (\hat{f}_k(t) - f_{ok}(t))^2 dt \quad \text{and} \quad \text{ISE}(\hat{f}) = \sum_{k=1}^{s} \int_0^1 (\hat{f}_k(t) - f_{ok}(t))^2 dt.
\]

Notice that, under setting (I) where \( Z \) has zero effect and FAM is the correct model, the PLFAM estimators perform comparably to those of FAM. However, under setting (II), where \( Z \) has non-zero effects, FAM performs significantly worse than PLFAM. This demonstrates the possible risk of ignoring important vector predictors.

We also summarize the prediction errors, and the mean squared errors (MSE) for the estimated partially linear coefficients in Table 4. To show the advantage of mFPCA, we further compare two methods to obtain FPCA scores: the “joint” approach is the mFPCA approach that we advocate; and the “separate” approach is to perform univariate FPCA to each component of \( X \), standardize these scores separately, and then pool all standardized scores together as covariates in the additive model. Both FPCA approaches can be used in conjunction with FAM and PLFAM. The prediction error is computed by \( n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) on the testing data set. To compute the prediction \( \hat{y}_i \) in the test data, we first compute the transformed FPCA scores of \( x_i \) in the test set using the estimates of mean function, eigenvalues and eigenfunctions from the training data, and then plug these scores into the estimated regression \( \hat{m} \). The results in Table 4 suggest that jointly modeling multiple functional predictors leads to smaller MSE’s for \( \hat{\theta} \), and lower prediction errors, as opposed to modeling each functional predictor separately using univariate FPCA. In addition, PLFAM has significant lower prediction errors than FAM under setting (II) when there is a non-zero effect from \( Z \).
In Section C of the supplementary material, we also report the simulation results when $s$ chosen to recover 90% of the total variation. Under this setting, one important component related to $Y$ is close to the 90% cut-off line and often not included as a candidate for COSSO. As a results, a non-zero component function is often failed to be selected, and the resulted models yield higher prediction errors in the test data sets. Based on these results, we recommend to include a large number of components and let the built-in model selection mechanism of COSSO determine the size of the model.

6 Real Data Application

The practical utility of the proposed method is illustrated through an analysis of a crop yield data set from the National Agricultural Statistics Agency (https://quickstats.nass.usda.gov/), which consists of several yield-related variables at the county level (such as annual crop yield in bushels per acre, size of harvested land and the proportion irrigated land to the total harvested land) from 105 counties in Kansas from 1999 to 2011. We have yield-related variables for the two major crops in Kansas, corn and soybean, which are analyzed separately. Variables such as total harvest land and proportion of irrigated land are crop-specific. The weather data (annual averaged precipitation, daily maximum temperature and daily minimum temperature) are gathered from 1123 weather stations in Kansas provided by the National Climatic Data Center (https://www.ncdc.noaa.gov/data-access) and aggregated at the county level.

To apply our model, let $Y$ be the average crop yield per acre (corn or soybean) for a specific year and county; $X_1(t)$ and $X_2(t)$ are the daily maximum and daily minimum temperature trajectories for the same year and county with the time domain $T = [0, 365]$; $Z$ includes proportion of irrigated land in that county and for that particular type of crop,
averaged annul precipitation, and the interaction between the two. In the past several decades, due to sustained improvements in genetics and production technology, there is a consistent increasing trend in the yields of both corn and soybean. To take this effect into consideration, we also include year indicators into $Z$.

Since the response is an average obtained from an agricultural survey, the errors are heteroscedastic with weights $\pi$ equal to the sizes of harvested land. Some earlier work (Smith, 1938; Beran et al., 2013) suggests crop yield may exhibit long range dependency on a scale measured in feet. Our study on the other hand is based on county level aggregated data. The crop yields are usually averaged over tens of thousands of acres within a county and not from a continuous piece of land. At this scale, the spatial correlation is already quite weak and therefore it is reasonable to assume the variance of the average crop yield is proportional to the inverse of the total harvest land. Furthermore, land use rotates between the major crops across years: land used to grow corn this year is usually used to grow soybean the next year. Variables such as the proportion of irrigated land and size of harvest land are different in different years even for the same crop and same county. Even though our theory and methods are developed under the independence assumption, they can still be applied as long as the crop yields are conditionally independent across counties and years, given the local meteorology information, which seems reasonable because of the rotation in land use and because crops of different genotypes are planted in different years.

To illustrate the functional predictors, we show in Figure S.1 of the supplementary material 50 randomly selected trajectories for $X_1(t)$ and $X_2(t)$, with the mean functions $\mu_1(t)$ and $\mu_2(t)$ marked as solid curves in the two panels. As one can see, there are a lot of local fluctuations in the temperature trajectories, which is normal since heat and chill alternate throughout the year. In Figure 1, we show the heat plots for the (cross-) covariance functions. The kernel function for $C_{12}$ shows great resemblance to $C_{11}$ and $C_{22}$, which implies
that the two functional predictors are strongly correlated. This also suggests mFPCA would achieve more efficient dimension reduction than univariate FPCA done separately to the two processes, and the latter would include too much redundant information into the regression model.

6.1 Crop yield prediction experiment

Since our goal of this study is to find the best model for yield prediction, we divide the data into smaller training and validation data sets and compare the prediction of the following 10 competing models.

1. PLFAM(joint): the proposed PLFAM based on mFPCA scores;

2. PLFAM(separate): PLFAM based on univariate FPCA scores from $X_1$ and $X_2$ separately;

3. FAM(joint): FAM based on mFPC scores (without $Z$);

4. FAM(separate): FAM based on univariate FPC scores (without $Z$);

5. FLM-Cov(joint): functional linear model (FLM) based on mFPCA scores, with covariates;

6. FLM-Cov(separate): FLM based on separate univariate FPCA scores, with covariates;

7. FLM(joint): FLM based on joint mFPCA scores (without $Z$);

8. FLM(separate): FLM based on separate FPCA scores (without $Z$);

9. LM: linear model on $Z$ only;

10. LM-GDD: linear model on $Z$ and Growing Degree Days (GDD), to be explained below.
The models we consider can be divided into three categories: (a) functional additive models (Models 1-4), (b) functional linear models (Models 5-8) and (c) non-functional model (Models 9 - 10). For all functional regression models, including those in categories (a) and (b), FPCA scores that account up to 99.9% of the total variation are admitted into the model. For the methods based on separate FPCA on $X_1$ and $X_2$, we include FPCA scores that explain 99.9% of total variation in each functional predictor and thus use twice as many FPCA scores in the regression analysis as the joint modeling methods. For all models in category (a), we rely on the model selection mechanism of COSSO to prevent overfitting and select the tuning parameters by 5-fold cross-validation; for the functional linear models in category (b), we avoid overfitting by introducing ridge penalties, the tuning parameters of which are chosen by generalized cross-validation. It is worth noting that Model 10 serves as the benchmark model for yield prediction with temperature information enters into the model as the GDD variable. GDD is a measure of heat accumulation commonly used to predict plant development (Gilmore and Rogers, 1958; Yang et al., 1995; McMaster and Wilhelm, 1997). Here we adopt the definition used in the EPIC (Erosion Productivity Impact Calculator) plant growth model (Williams et al., 1989), in which GDD is defined as the sum of $\{X_1(t) + X_2(t)\}/2 - T_{\text{base}}$ over growing season, where $T_{\text{base}}$ is the crop-specific base temperature in °C. For corn $T_{\text{base}} = 8$, and for soybean $T_{\text{base}} = 10$. To account for heteroscedasticity, the sizes of harvested land are used as weights in fitting all models.

For each five-year window (i.e., 1999-2003, 2000-2004, . . . , 2007-2011), we pull the data from those five years into a smaller data set. For each five-year data set, we randomly divide it into five subsets, hold out one subset at a time as a validation set, fit the ten models described above to the remaining four subsets, and then use the trained models to predict the responses in the validation data. The mean squared prediction errors are weighted by the sizes of harvested land, averaged over the five validation sets and over all five-year
periods. The averaged overall prediction errors are reported in Table 5. From the table, models without the covariate effects, including FAM(separate), FAM(joint), FLM(separate) and FLM(joint), perform significantly worse than the rest. These results agree with the general belief that irrigation and precipitation are informative in yield prediction, which also stress the importance of extending the FAM of Müller and Yao (2008) to our PLFAM. We can also see that including the functional predictors can reduce the prediction error, and functional regression model such as PLFAM(separate), PLFAM(joint), FLM-Cov(separate) and FLM-Cov(joint) perform better than the non-functional models (LM and LM-GDD). Joint modeling the two functional predictors using mFPCA also leads to lower prediction error for both PLFAM and FLM-Cov. Overall, PLFAM(joint) performs the best in corn yield prediction and achieves comparable result to FLM-Cov(joint) for soybean.

Part of the reason that PLFAM performs slightly worse than FLM in soybean yield prediction is that the nonlinear effect is less significant for soybean and PLFAM requires a larger sample size. In another experiment where we include more years of data in the training set, PLFAM predicts soybean yield better than FLM.

In addition to the 10 models described above, we also consider another 12 models that use $X_1, X_2$ or $(X_1 + X_2)/2$ alone. These models yield higher prediction errors than the proposed PLFAM(joint) model, which utilizes both functional predictors. Even though the two functional covariates in the real data are strongly correlated as suggested by Figure 1, these results show that each covariate does provide additional information that complements the other and it is beneficial to jointly model them. Due to space limitation, these results are presented in Section D of the supplementary material.
6.2 Regression analysis of the whole data

We now apply PLFAM(joint) to the whole data set pooling all available years. For corn yield prediction, we include 52 principal components in the regression model which account for $\sim 99\%$ of variation in the temperature trajectories, and 10 principal components are selected by COSSO. In Figure 2, we show the top 6 most significant principal components; and in Figure 3, we show the corresponding additive component functions $\hat{f}_k(\zeta)$. These components are ranked by the importance of their contribution to $Y$. More specifically, we sort the principal components by the RKHS norm of the component function $\hat{f}_k$. The dashed curves in Figure 3 are the pointwise confidence bands $\hat{f}_k(\zeta) \pm 2 \times se\{\hat{f}_k(\zeta)\}$, and the dotted curves are the 3 times standard error bands. The standard errors are estimated using a bootstrap procedure detailed in the supplementary material.

Since each principal component in mFPCA is a vector of functions $\psi_k(t) = \{\psi_{k1}(t), \psi_{k2}(t)\}^T$, we show $\psi_{k1}(t)$ as the solid curve and $\psi_{k2}(t)$ as the dashed curve in each panel of Figure 2. It is not surprising that $\psi_{k2}(t)$ largely coincides with $\psi_{k1}(t)$, given the observation from the covariance functions that the two processes are strongly correlated. However, the plots do reveal subtle differences between the two temperature trajectories. The component most related to corn yield $\psi_5$ features a temperature pattern with near average daily minimum temperature and lower than average daily maximum temperature during the summer months from May to September. A higher loading on $\psi_5$ means a milder summer, less heat stress and less chance of draught, and corn yield is an increasing function of $\zeta_5$ in Figure 3. In contrast, $\psi_1$ and $\psi_8$ represent hot summers, and crop yield is a decreasing function of their loadings $\zeta_1$ and $\zeta_8$. These are consistent with the findings in Westcott et al. (2013), which conclude that hot July - August weather lowered the corn yield. A prominent feature in $\psi_3$ is warm spring months from January to March, which may lead to less snow coverage on the ground, early insect activities and hence the negative association with corn yield as described.
by $f_3(\zeta)$ in Figure 3. For soybean yield prediction, graphs of the selected eigenfunctions and the corresponding additive component functions are similar to those in Figures 2 and 3, and are hence omitted.

The estimated partially linear coefficients and their bootstrap standard errors for both corn and soybean yield models are summarized in Table 6. As we can see, both the proportion of irrigated land (Irrigate) and precipitation (Prec) have significant positive effects on crop yield. The significant negative interaction means the effect of Prec is mitigated when a big portion of the lands in the county are equipped with irrigation systems. For corn yield prediction, the first and third quartiles for Irrigate are 0.027 and 0.485 respectively. Changing Irrigate from its first quartile to the third, the partial slope on Prec reduces from 167.47 to 151.95.

The bootstrap procedure, provided in the supplementary material, is based on the assumption that the errors in model (2) are independent. To validate this assumption, we also estimate the spatial variogram for each year and temporal autocorrelation for each county based on the residuals of the fitted model, see Figures S.2 and S.3 in the supplementary material. The variograms and ACF’s are contained in their confidence bands based on the assumption of no dependency, which means there is no significant evidence for spatial or temporal correlation.

7 Concluding Remarks

7.1 Our contributions

We have extended the FAM of Müller and Yao (2008) to a class of PLFAM which takes into account of the effects of a multivariate covariate $Z$. As demonstrated in our crop yield application, including the covariate effects significantly improves the prediction accuracy.
The effect of functional predictors are modeled through an additive model on the principal component scores. Since the FPC scores are estimated with error, our theory and methods also shine a new light on the area of additive models with covariate measurement errors.

We have also made a number of important theoretical contributions. First, we develop a more general model framework which includes multivariate functional predictors and multivariate covariates. Second, we allow the number of principal components admitted in the additive model to diverge to infinity, which is fundamentally different from Zhu et al. (2014). Third, we are able to quantify and bound the nuisance from the estimation errors in mF-PCA scores without the artificial assumption in Zhu et al. (2014). Finally, when the number of principal components does not diverge to infinity, we establish root-$n$ consistency and asymptotic normal distribution for the partially linear regression coefficients.

### 7.2 Interpretability of the model

Functional regression models based on principal components are in general hard to interpret, because FPC’s are the maximum modes of variation in the functional predictors which are not necessarily the features most related to the response variable. This is part of the reason that many authors focused on prediction using functional linear model (Cai and Hall 2006; Cai and Yuan 2012). Our proposed PLFAM adopts the philosophy of semiparametric statistics: we model the effects of functional covariates nonparametrically to increase the model flexibility and prediction performance, and model the effects of the multivariate covariates parametrically for better interpretations and statistical inference. Our Theorem 2 provides a basis for statistical inference on the parametric component $\gamma$. There is also another class of functional additive regression models proposed by Müller et al. (2013); McLean et al. (2014); Kim et al. (2017), which offer an alternative view on modeling nonlinear effects of functional covariates.
7.3 mFPCA versus separate FPCA

For multivariate functional data, mFPCA usually provides more efficient dimension reduction than separate FPCA to each functional covariate. However, mFPCA estimates are subject to higher variability due to the need of estimating all cross-covariance functions and performing eigenvalue decomposition on a much larger covariance matrix. When the sample size is small, the extra variation in mFPCA can offset its benefit. There are also other situations where separate FPCA is more preferable, such as when different functional covariates are of different scales or even defined on different domains (Happ and Greven, 2017). Under these situations, our theory and methods can also be easily extended to the model based on separate FPCA scores. A separate FPCA version of model (2) is

\[ y_i = u_i^\top \theta_0 + \sum_{k=1}^{s} \sum_{j=1}^{d} f_{0jk}(\zeta_{ijk}) + \varepsilon_i, \]  

(11)

where \(\zeta_{ijk}\) is the \(k\)th standardized principal component score for \(x_{ij}\). The model can be fitted using the same COSSO algorithm described in Section 3.2 except that the mFPCA scores are replaced by the separate FPCA scores. As long as the separate FPC scores can be estimated with a similar accuracy as assumed in Theorem 1, i.e. \(E(\hat{\zeta}_{ijk} - \zeta_{ijk})^2 \leq Cn^{-1}k^{2/3}\) uniformly for all \(j = 1, \ldots, d\) and \(k \leq s\), the same asymptotic results in Theorems 1 and 2 hold for the model in (11).

Acknowledgement

Wong’s research was partially supported by the National Science Foundation under Grants DMS-1612985 and DMS-1711952 (subcontract). Li’s research was partially supported by the National Science Foundation grant DMS-1317118. Zhu’s research was funded in part by cooperative agreement 68-7482-17-009 between the USDA Natural Resources Conservation
Service and Iowa State University.

References


Table 1: Percentages of fitted model sizes.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Model</th>
<th>% for the following model sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>{ (i), (I) }</td>
<td>FAM</td>
<td>0 0 28 49.5 20.5 1.5 0.5 0</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0 0 24 57.5 17 1 0.5 0</td>
</tr>
<tr>
<td>{ (ii), (I) }</td>
<td>FAM</td>
<td>0 0 20.5 58 16.5 4 1 0</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0 0 19 58.5 18 3.5 0 1</td>
</tr>
<tr>
<td>{ (i), (II) }</td>
<td>FAM</td>
<td>0 6.5 41 39.5 12 0.5 0.5 0</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0 0 22.5 56 18 3 0.5 0</td>
</tr>
<tr>
<td>{ (ii), (II) }</td>
<td>FAM</td>
<td>0 3 44.5 38 12.5 2 0 0</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0 0 22.5 61 12.5 3 1 0</td>
</tr>
</tbody>
</table>

Table 2: Percentages of selected components and, correct and super selection.

| Setting      | Model | % for the following component functions | % correct | % super |
|--------------|-------|----------------------------------------|-----------|
|              |       | \(\hat{f}_1\) \(\hat{f}_2\) \(\hat{f}_3\) \(\hat{f}_4\) \(\hat{f}_5\) \(\hat{f}_6\) \(\hat{f}_7\) \(\hat{f}_8\) | set       | set     |
| { (i), (I) } | FAM   | 100 100 14 93 51.5 2 6 1.5 | 27 93     |
|              | PLFAM | 100 100 14 93 51.5 3 5 1.5 | 23 93     |
| { (ii), (I) }| FAM   | 100 100 20.5 97 51 5.5 1.5 2 | 18.5 97   |
|              | PLFAM | 100 100 20 97.5 53 5.5 1.5 2.5 | 17.5 97.5 |
| { (i), (II) }| FAM   | 100 90 8.5 93.5 34.5 2.5 4.5 4.5 | 35 83.5   |
|              | PLFAM | 100 100 16 97.5 54.5 3.5 3.5 1.5 | 21.5 97.5 |
| { (ii), (II) }| FAM   | 100 93.5 12 94 31.5 2.5 3 1 | 37.5 88    |
|              | PLFAM | 100 99.5 22.5 98 47.5 2.5 2.5 1.5 | 22.5 97.5 |
Table 3: Averaged integrated squared errors.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Model</th>
<th>$\hat{f}_1$</th>
<th>$\hat{f}_2$</th>
<th>$\hat{f}_3$</th>
<th>$\hat{f}_4$</th>
<th>$\hat{f}_5$</th>
<th>$\hat{f}_6$</th>
<th>$\hat{f}_7$</th>
<th>$\hat{f}_8$</th>
<th>$\hat{f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{(i), (I)}</td>
<td>FAM</td>
<td>0.0172</td>
<td>0.1073</td>
<td>0.0057</td>
<td>0.1689</td>
<td>0.1204</td>
<td>0.0001</td>
<td>0.0015</td>
<td>0.0001</td>
<td>0.4292</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0.0175</td>
<td>0.1070</td>
<td>0.0056</td>
<td>0.1689</td>
<td>0.1205</td>
<td>0.0004</td>
<td>0.0014</td>
<td>0.0002</td>
<td>0.4289</td>
</tr>
<tr>
<td>{(ii), (I)}</td>
<td>FAM</td>
<td>0.0198</td>
<td>0.1038</td>
<td>0.0109</td>
<td>0.1290</td>
<td>0.0890</td>
<td>0.0018</td>
<td>0.0004</td>
<td>0.0007</td>
<td>0.3633</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0.0198</td>
<td>0.1046</td>
<td>0.0111</td>
<td>0.1279</td>
<td>0.0896</td>
<td>0.0016</td>
<td>0.0005</td>
<td>0.0011</td>
<td>0.3638</td>
</tr>
<tr>
<td>{(i), (II)}</td>
<td>FAM</td>
<td>0.0330</td>
<td>0.2208</td>
<td>0.0064</td>
<td>0.2197</td>
<td>0.0782</td>
<td>0.0008</td>
<td>0.0026</td>
<td>0.0021</td>
<td>0.5780</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0.0177</td>
<td>0.1072</td>
<td>0.0064</td>
<td>0.1320</td>
<td>0.1130</td>
<td>0.0011</td>
<td>0.0009</td>
<td>0.0005</td>
<td>0.3858</td>
</tr>
<tr>
<td>{(ii), (II)}</td>
<td>FAM</td>
<td>0.0290</td>
<td>0.2035</td>
<td>0.0087</td>
<td>0.2170</td>
<td>0.0841</td>
<td>0.0017</td>
<td>0.0024</td>
<td>0.0005</td>
<td>0.5642</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>0.0179</td>
<td>0.1084</td>
<td>0.0103</td>
<td>0.1398</td>
<td>0.0978</td>
<td>0.0007</td>
<td>0.0008</td>
<td>0.0005</td>
<td>0.3821</td>
</tr>
</tbody>
</table>

Table 4: Prediction errors and mean squared errors for FAM and PLFAM, using separate univariate FPCA scores (columns labelled “separate”) or mFPCA scores (columns labelled “joint”). For prediction errors, means are presented with corresponding standard deviations in parentheses.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Model</th>
<th>Prediction error</th>
<th>Mean squared errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>separate    joint</td>
<td>separate   joint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\theta}_1$</td>
<td>$\hat{\theta}_2$</td>
</tr>
<tr>
<td>{(i), (I)}</td>
<td>FAM</td>
<td>1.55 (0.10) 1.32 (0.13)</td>
<td>-          -      -</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>1.57 (0.11) 1.33 (0.13)</td>
<td>-          -      -</td>
</tr>
<tr>
<td>{(ii), (I)}</td>
<td>FAM</td>
<td>1.65 (0.09) 1.33 (0.12)</td>
<td>-          -      -</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>1.66 (0.09) 1.35 (0.13)</td>
<td>0.0746     0.0911 0.1076</td>
</tr>
<tr>
<td>{(i), (II)}</td>
<td>FAM</td>
<td>3.84 (0.22) 3.63 (0.21)</td>
<td>-          -      -</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>1.59 (0.10) 1.34 (0.13)</td>
<td>0.0678     0.1095 0.0827</td>
</tr>
<tr>
<td>{(ii), (II)}</td>
<td>FAM</td>
<td>3.89 (0.24) 3.69 (0.24)</td>
<td>-          -      -</td>
</tr>
<tr>
<td></td>
<td>PLFAM</td>
<td>1.68 (0.12) 1.35 (0.14)</td>
<td>0.0642     0.0879 0.1092</td>
</tr>
</tbody>
</table>
Table 5: Average of 5-year overall prediction errors.

<table>
<thead>
<tr>
<th></th>
<th>corn</th>
<th>soybean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) functional additive models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLFAM(joint)</td>
<td>298.43</td>
<td>35.64</td>
</tr>
<tr>
<td>PLFAM(separate)</td>
<td>306.50</td>
<td>38.85</td>
</tr>
<tr>
<td>FAM(joint)</td>
<td>830.17</td>
<td>48.54</td>
</tr>
<tr>
<td>FAM(separate)</td>
<td>839.00</td>
<td>51.06</td>
</tr>
<tr>
<td>(b) functional linear models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLM-Cov(joint)</td>
<td>303.81</td>
<td>35.29</td>
</tr>
<tr>
<td>FLM-Cov(separate)</td>
<td>308.57</td>
<td>35.69</td>
</tr>
<tr>
<td>FLM(joint)</td>
<td>704.19</td>
<td>47.31</td>
</tr>
<tr>
<td>FLM(separate)</td>
<td>767.42</td>
<td>50.42</td>
</tr>
<tr>
<td>(c) non-functional model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>391.18</td>
<td>61.74</td>
</tr>
<tr>
<td>LM-GDD</td>
<td>389.76</td>
<td>49.58</td>
</tr>
</tbody>
</table>

Table 6: Estimated regression coefficients (bootstrap standard error) in the PLFAM for crop yield prediction.

<table>
<thead>
<tr>
<th></th>
<th>Irrigate</th>
<th>Prec</th>
<th>Irrigat*Prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>corn</td>
<td>168.38 (6.42)</td>
<td>20.98 (2.72)</td>
<td>-33.87 (3.32)</td>
</tr>
<tr>
<td>soybean</td>
<td>33.30 (3.35)</td>
<td>3.91 (0.70)</td>
<td>-4.88 (1.65)</td>
</tr>
</tbody>
</table>

Note: **Irrigate**: proportion of irrigated land in a county for the specific crop and growing year; **Prec**: averaged precipitation for county and year; **Irrigat*Prec**: the interaction.
Figure 1: Heat plot for the covariance and cross-covariance functions. From bottom to top and from left to right are the kernel functions of the (cross-) covariance operators $C_{jj'}$. 
Figure 2: Corn yield prediction: top 6 principal components selected by COSSO for corn yield prediction, sorted by the decreasing order of the RKHS norm of \( \hat{f}_k(z) \) \( (k = 5, 3, 1, 7, 8, 10) \). Each principal component is a vector \( \psi_k(t) = \{\psi_{k1}(t), \psi_{k2}(t)\}^T \). The solid curve in each panel is \( \psi_{k1}(t) \) and the dashed curve is \( \psi_{k2}(t) \).
Figure 3: Corn yield prediction: top 6 additive component functions $\hat{f}_k(\zeta)$, sorted by the decreasing order of the RKHS norm of $\hat{f}_k(\zeta)$ ($k = 5, 3, 1, 7, 8, 10$).