Partially Linear Functional Additive Models for Multivariate Functional Data

Raymond K. W. Wong[1], Yehua Li[2] and Zhengyuan Zhu[2]

[1]Department of Statistics, Texas A&M University, College Station, TX 77843

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

$$y_i = m(\boldsymbol{z}_i, \boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where m is the regression function and ε_i are zero mean errors independent with $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{z}_i\}_{i=1}^n$. Further we assume $\mathrm{var}(\varepsilon_i) = \sigma_\varepsilon^2/\pi_i$, where σ_ε^2 is an unknown variance parameter and π_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 π_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 $\mathbb{X}_j$, with inner product $\langle \cdot, \cdot \rangle_{\mathbb{X}_j}$ and norm $\| \cdot \|_{\mathbb{X}_j}$. We will focus on the case that $\mathbb{X}_j$'s are L^2 functional spaces and the inner products are $\langle f, g \rangle_{\mathbb{X}_j} = \int_{\mathcal{T}} f(t)g(t)dt$ for any $f, g \in \mathbb{X}_j$. Let $\mathbb{X} = \bigoplus_{j=1}^d \mathbb{X}_j$ be the direct sum of the functional spaces, which is a bigger Hilbert space equipped with the induced inner product and norm, i.e. $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_{\mathbb{X}} = \sum_{j=1}^d \langle x_{1j}, x_{2j} \rangle_{\mathbb{X}_j}$ and $\|\boldsymbol{x}_1\|_{\mathbb{X}} = \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle_{\mathbb{X}}^{1/2}$ for any $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\intercal \in \mathbb{X}$, $i = 1, 2$.

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

matrix of cross-covariance functions

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

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

$$\mathcal{C}(s,t) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\psi}_k(s) \boldsymbol{\psi}_k^\intercal(t),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ are the eigenvalues and $\boldsymbol{\psi}_k = (\psi_{k1}, \ldots, \psi_{kd})^\intercal \in \mathbb{X}$ are the corresponding eigenfunctions such that $\langle \boldsymbol{\psi}_k, \boldsymbol{\psi}_{k'} \rangle_{\mathbb{X}} = \int_{\mathcal{T}} \boldsymbol{\psi}_k(t)^\intercal \boldsymbol{\psi}_{k'}(t) dt = I(k = k')$. By a standard Karhunen-Loève expansion

$$\boldsymbol{X}(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{\infty} \xi_k \boldsymbol{\psi}_k(t),$$

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

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 $\boldsymbol{x}_i$ by $\boldsymbol{\xi}_{i,\infty} = (\xi_{i1}, \xi_{i2}, \ldots)^\intercal$. 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 $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{is})^\intercal$, 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 ξ is close to Gaussian, ζ is approximately uniform in $[0,1]$, which is convenient for nonparametric modeling on the effect of ζ. Other continuous CDFs can also be used as $\Phi(\cdot)$, such as the logistic function. Write $\boldsymbol{\zeta}_{i,\infty} = (\zeta_{i1}, \zeta_{i2}, \ldots)$ and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{is})$.

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)

$$y_i = m_0(\boldsymbol{u}_i, \boldsymbol{\zeta}_i) + \varepsilon_i = \boldsymbol{u}_i^\intercal \boldsymbol{\theta}_0 + f_0(\boldsymbol{\zeta}_i) + \varepsilon_i = \boldsymbol{u}_i^\intercal \boldsymbol{\theta}_0 + \sum_{k=1}^{s} f_{0k}(\zeta_{ik}) + \varepsilon_i, \tag{2}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^{p+1}$ and $\boldsymbol{u}_i = (1, \boldsymbol{z}_i^\intercal)^\intercal$. 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, $\boldsymbol{u}$ 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 $\widehat{\boldsymbol{\theta}}$ (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 belongs 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 $\mathbb{F}_k = \{1\} \oplus \bar{\mathbb{F}}_k$ where $\{1\}$ is the space of all constant functions. From now on, we assume $m_0 \in \mathbb{M} = \mathbb{I} \oplus \sum_{k=1}^{s} \bar{\mathbb{F}}_k$, where $f_{0k} \in \bar{\mathbb{F}}_k$ for $k = 1, \ldots, s$, and $\mathbb{I} = \{\boldsymbol{u}^\intercal \boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{p+1}\}$. 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 $\boldsymbol{x}_i(t)$'s are fully observed. Then the mean and covariance of $\boldsymbol{X}$ can be estimated by

$$\widehat{\boldsymbol{\mu}}(t) = n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i(t), \quad \widehat{\mathcal{C}}(s,t) = n^{-1}\sum_{i=1}^{n}\{\boldsymbol{x}_i(s)-\widehat{\boldsymbol{\mu}}(s)\}\{\boldsymbol{x}_i(t)-\widehat{\boldsymbol{\mu}}(t)\}^{\intercal}. \tag{3}$$

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

$$\widehat{\xi}_{ik} = \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\psi}}_k\rangle_{\mathbb{X}} = \sum_{j=1}^{d}\int_{\mathcal{T}} x_{ij}(t)\widehat{\psi}_{kj}(t)dt, \quad \widehat{\zeta}_{ik} = \Phi(\widehat{\lambda}_k^{-1/2}\widehat{\xi}_{ik}), \quad k = 1,\ldots,d. \tag{4}$$

In practice, we only have discrete noisy observations on $\boldsymbol{x}_i$

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

where e_{ijk}'s are independent measurement errors with mean 0 and variance $\sigma^2_{e,j}$, $j = 1,\ldots,d$. 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 $\widetilde{x}_{ij}(t) = \mathfrak{S}(t;\boldsymbol{t}_{ij})\mathfrak{w}_{ij}$, where $\mathfrak{w}_{ij} = (w_{ij1},\ldots,w_{ij,N_{ij}})^{\intercal}$ and $\mathfrak{S}(t;\boldsymbol{t}_{ij})$ is a linear smoother depending on the design points $\boldsymbol{t}_{ij} = (t_{ij1},\ldots,t_{ijN_{ij}})^{\intercal}$, e.g. local polynomial or regression splines. The eigenvalues, eigenfunctions and mFPCA scores are estimated by replacing $x_{ij}(t)$ with $\widetilde{x}_{ij}(t)$ in (3) and (4).

For univariate functional data, this pre-smoothing approach is theoretically justified by Hall et al. (2006), who show that, when $\mathfrak{S}$ is a local linear smoother and $N_{\min} = \min_{i,j}N_{ij} >$

$Cn^{1/4}$, the error incurred by approximating $x_{ij}(t)$ with $\widetilde{x}_{ij}(t)$ is negligible in $\widehat{\lambda}_k$ and $\widehat{\psi}_k$; Li et al. (2010) further show that this approximation error is negligible to $\widehat{\xi}_{ik}$ if $N_{\min} > Cn^{5/4}$. As commented in Li et al. (2010), there are two sources of error in $\widehat{\xi}_{ik}$: the error caused by approximating x_{ij} with $\widetilde{x}_{ij}$ and the error in $\widehat{\psi}_k$. If the first type of error prevails, regression analysis using $\widehat{\xi}_{ik}$ 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. Chiou et al. (2014)), 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 $\boldsymbol{x}_i$ into a longer vector, then the operator $\widehat{\mathcal{C}}$ 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 Ramsay and Silverman (2005).

3.2 Estimation of PLFAM with COSSO penalty

Let $\widehat{\boldsymbol{\zeta}}_i = (\widehat{\zeta}_{i1}, \ldots, \widehat{\zeta}_{is})^\intercal$ be a vector of standardized mFPCA scores for $\boldsymbol{x}_i$ estimated using the procedure in Section 3.1. Since there are potentially infinite number of principal components for $\boldsymbol{X}$, 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 $\widehat{m}$ is the minimizer of the following penalized loss $\ell_w(m)$ with respect to $m \in \mathbb{M}$. The loss function is defined as

$$\ell_w(m) = \frac{1}{n}\sum_{i=1}^{n} \pi_i \{y_i - m(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 + \tau_n^2 J(m), \tag{5}$$

where π_i are the survey weights defined in (1). Here τ_n^2 is a tuning parameter and $J(m) =$

$\sum_{i=1}^{s} \|\mathcal{P}_k m\|$ with $\mathcal{P}_k$ being the projection operator to $\bar{\mathbb{F}}_k$. 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

$$\frac{1}{n}\sum_{i=1}^{n} \pi_i\{y_i - m(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 + \kappa_0 \sum_{k=1}^{s} \phi_k^{-1} \|\mathcal{P}_k m\|^2 + \kappa \sum_{k=1}^{s} \phi_k \tag{6}$$

over $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s)^\intercal \in [0, \infty)^s$ and $m \in \mathbb{M}$, where $\kappa_0 > 0$ is a pre-determined constant and κ 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 $\kappa = \tau_n^4/(4\kappa_0)$. (i) If $\widehat{m}$ minimizes (5), set $\widehat{\phi}_k = \kappa_0^{1/2}\kappa^{-1/2}\|\mathcal{P}_k\widehat{m}\|$; then the pair $(\widehat{\boldsymbol{\phi}}, \widehat{m})$ minimizes (6). (ii) If $(\widehat{\boldsymbol{\phi}}, \widehat{m})$ minimizes (6), then $\widehat{m}$ minimizes (5).*

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

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

where $\|\cdot\|_E$ represents the Euclidean norm, $\Pi = \mathrm{diag}\{\pi_1, \ldots, \pi_n\}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$, $\boldsymbol{U} = [u_{ij}]_{i=1,\ldots,n, j=1,\ldots,p+1}$ is a $n \times (p+1)$ design matrix and $\boldsymbol{R}_k = [R_k(\widehat{\zeta}_{ik}, \widehat{\zeta}_{jk})]_{i,j=1,\ldots,n}$ is a $n \times n$ matrix for $k = 1, \ldots, s$. For a fixed $\boldsymbol{\phi}$, minimizing (7) with respect to $(\boldsymbol{\theta}, \boldsymbol{a})$ is similar to solving a weighted ridge regression. For fixed $\boldsymbol{\theta}$ and $\boldsymbol{a}$, let $\boldsymbol{D}$ be the $n \times s$ matrix with the

k-th column being $\boldsymbol{R}_k\boldsymbol{a}$, then minimization of (7) with respect to $\boldsymbol{\phi} \in [0,\infty)^s$ becomes

$$\min \frac{1}{n}\left[\boldsymbol{\phi}^\intercal \boldsymbol{D}^\intercal \Pi \boldsymbol{D}\boldsymbol{\phi} - 2\left\{\boldsymbol{D}^\intercal\Pi(\boldsymbol{y}-\boldsymbol{U}\boldsymbol{\theta}) - \frac{1}{2}n\kappa_0\boldsymbol{D}^\intercal\boldsymbol{a}\right\}^\intercal \boldsymbol{\phi}\right]$$
$$\text{subject to } \sum_{k=1}^{s}\phi_k < G \text{ and } \boldsymbol{\phi}\in[0,\infty)^s,$$

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 $(\boldsymbol{\theta}, \boldsymbol{a})$ and $\boldsymbol{\phi}$ in turn. The algorithm starts with solving $(\boldsymbol{\theta}, \boldsymbol{a})$ while fixing $\boldsymbol{\phi} = \mathbf{1}$. 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), κ_0 can be fixed at any positive value. We select κ_0 that minimizes the GCV of the partial spline problem when $\boldsymbol{\phi} = \mathbf{1}$. Let $\widehat{\boldsymbol{\phi}}^{(\tau_n)} = (\widehat{\phi}_1^{(\tau_n)}, \ldots, \widehat{\phi}_n^{(\tau_n)})^\intercal$, $\widehat{\boldsymbol{a}}^{(\tau_n)}$ and $\widehat{\boldsymbol{\theta}}^{(\tau_n)}$ be the minimizer of (7) for a fixed τ_n. To select the smoothing parameter τ_n (or equivalently G), we minimize the Bayesian information criterion $n\log(\mathrm{RSS}_w(\tau_n)/n) + \mathrm{df}(\tau_n)\log(n)$, where the effective degress of freedom $\mathrm{df}(\tau_n)$ is the trace of the smoothing matrix in the partial spline problem (7) when $\boldsymbol{\phi}$ is set to $\widehat{\boldsymbol{\phi}}^{(\tau_n)}$, and the weighted residual sum of squares is $\mathrm{RSS}_w(\tau_n) = \frac{n}{\sum_{i=1}^n \pi_i}\|\Pi^{1/2}(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\theta}}^{(\tau_n)} - \sum_{k=1}^s \widehat{\phi}_k^{(\tau_n)}\boldsymbol{R}_k\widehat{\boldsymbol{a}}^{(\tau_n)})\|_E^2$.

4 Theoretical Results

4.1 Basic results for mFPCA

By the theory of Dauxois et al. (1982), $\|\widehat{\mathcal{C}} - \mathcal{C}\|_{\mathrm{op}} = O_p(n^{-1/2})$ where the operator norm is defined as $\|\mathcal{A}\|_{\mathrm{op}} = \sup_{\boldsymbol{x}\in\mathbb{X}} \frac{\|\mathcal{A}\boldsymbol{x}\|_{\mathbb{X}}}{\|\boldsymbol{x}\|_{\mathbb{X}}}$ for any bounded linear operator $\mathcal{A}$ on $\mathbb{X}$. To derive the

asymptotic expansion for $\widehat{\xi}_{ik}$s, we use the asymptotic expansion of $\widehat{\lambda}_k$ and $\widehat{\boldsymbol{\psi}}_k$ 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))

$$C_\lambda^{-1}k^{-\alpha} \leq \lambda_k \leq C_\lambda k^{-\alpha}, \quad \lambda_k - \lambda_{k+1} \geq C_\lambda^{-1}k^{-1-\alpha}, \quad k = 1, 2, \ldots. \tag{8}$$

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

Assumption 2 $\mathbb{E}(\|\boldsymbol{X}\|_{\mathbb{X}}^4) < \infty$ *and there exists a constant* $C_\xi > 0$ *such that* $\mathbb{E}(\xi_k^2\xi_{k'}^2) \leq C_\xi\lambda_k\lambda_{k'}$ *and* $\mathbb{E}(\xi_k^2 - \lambda_k)^2 < C_\xi\lambda_k^2$ *for all* k *and* $k' \neq k$*.*

The polynomial decay rate described in Assumption 1 is a slow decay rate assumption on the eigenvalues and allows $\boldsymbol{X}(t)$ 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 $\boldsymbol{X}(t)$ 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 $\frac{1}{2}C_\lambda^{-1}k^{-1-\alpha}$ under condition (8) and denote $\Delta = n^{1/2}(\widehat{\mathcal{C}} - \mathcal{C})$. By Dauxois et al. (1982), Δ converges weakly to a Gaussian variable in the space of linear operators and hence $\|\Delta\|_{\text{op}} = \mathcal{O}_p(1)$.

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* $\mathbb{E}(\widehat{\zeta}_{ik} - \zeta_{ik})^2 \leq Ck^2/n$ *uniformly for* $k \leq J_n$*, where* $J_n = \lfloor(2C_\lambda\|\Delta\|_{\text{op}})^{-1/(1+\alpha)}n^{1/(2+2\alpha)}\rfloor$*.*

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 $\mathcal{O}_p(n^{-1/2}k)$, which does not depend on the decay rate α 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 $(\boldsymbol{Z}, \boldsymbol{\zeta})$. That is, $P_n = \sum_{i=1}^n \delta_{\boldsymbol{z}_i,\boldsymbol{\zeta}_i}/n$, where $\delta_{\boldsymbol{z},\boldsymbol{\zeta}}$ is the delta function at $(\boldsymbol{z}, \boldsymbol{\zeta})$. Moreover, we denote the distribution of $(\boldsymbol{Z}, \boldsymbol{\zeta})$ by P. We define the corresponding (squared) empirical norm and inner product as

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

These notations are extended to measurement errors $\{\varepsilon_i\}$. For instance, $(\varepsilon, m_1)_n = \sum_{i=1}^n \varepsilon_i m_1(\boldsymbol{u}_i, \boldsymbol{\zeta}_i)/n$. 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, $\boldsymbol{\Sigma} = \int \boldsymbol{u}\boldsymbol{u}^\intercal dP$ is non-singular.

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

Remarks:

1. Under the framework laid out in Assumptions 1 and 2, with $s = \mathcal{O}_p(n^{1/\{2(1+\alpha)\}})$, we have $\mathbb{E}(\widehat{\zeta}_{ik} - \zeta_{ik})^2 \leq Cn^{-1}k^2$ 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 $\tau_n^{-1} = \mathcal{O}_p(n^{2/5}s^{-6/5})$, we have $\|\widehat{m} - m_0\|_n = \mathcal{O}_p(n^{-2/5}s^{6/5})$. If s is fixed, $\|\widehat{m} - m_0\|_n = \mathcal{O}_p(n^{-2/5})$ 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 $\|\widehat{\mathcal{C}} - \mathcal{C}\|_{\text{op}}$. 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 $\widehat{\zeta}_{ik}$. To achieve this goal, Zhu et al. (2014) assumed (see their Assumption 1)

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

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 $\widehat{\boldsymbol{\zeta}}_i$.

When s is fixed, better asymptotic results can be derived for the regression coefficients

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\intercal = (\theta_2, \ldots, \theta_{p+1})^\intercal$. Define

$$\begin{aligned}
&\boldsymbol{w}(\boldsymbol{\zeta}) = (w_1(\boldsymbol{\zeta}), \ldots, w_p(\boldsymbol{\zeta}))^\intercal = \operatorname{argmin}_{\substack{w_j \in \{1\} \oplus \sum_{k=1}^s \bar{\mathbb{F}}_k \\ j=1,\ldots,p}} \mathbb{E}\|\boldsymbol{Z} - \boldsymbol{w}(\boldsymbol{\zeta})\|^2, \\
&\widetilde{\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\zeta}) = (\widetilde{w}_1(\boldsymbol{z}, \boldsymbol{\zeta}), \ldots, \widetilde{w}_p(\boldsymbol{z}, \boldsymbol{\zeta}))^\intercal = \boldsymbol{z} - \boldsymbol{w}(\boldsymbol{\zeta}), \\
&\boldsymbol{M} = (M_{ij})_{i,j=1}^p, \qquad \text{where } M_{ij} = \int \widetilde{w}_i \widetilde{w}_j dP.
\end{aligned} \tag{9}$$

It is easy to see that $\boldsymbol{w}(\boldsymbol{\zeta})$ defines a additive regression of $\boldsymbol{Z}$ on $\boldsymbol{\zeta}$, and it can be considered as the projection of $\mathbb{E}(\boldsymbol{Z}|\boldsymbol{\zeta})$ on the additive regression space, and therefore

$$\mathbb{E}\{\widetilde{\boldsymbol{w}}^\intercal(\boldsymbol{z}, \boldsymbol{\zeta})\boldsymbol{g}(\boldsymbol{\zeta})\} = 0 \tag{10}$$

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

Theorem 2 *Assume the conditions of Theorem 1 hold with $s < \infty$ being fixed, $\boldsymbol{\zeta}$ 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 $\boldsymbol{M}$ defined in (9) is non-singular. Then $n^{1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \to \text{Normal}(\boldsymbol{0}, \boldsymbol{M}^{-1}(\boldsymbol{V}_1 + \boldsymbol{V}_2)\boldsymbol{M}^{-1})$ in distribution, where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are defined in (S.14) of the supplementary material.*

Remark: As shown in our proof, $\boldsymbol{V}_1 = \text{cov}\{n^{1/2}(\varepsilon, \widetilde{\boldsymbol{w}})_n\}$, and $\boldsymbol{M}^{-1}\boldsymbol{V}_1\boldsymbol{M}^{-1}$ is the typical asymptotic covariance matrix of $\widehat{\boldsymbol{\gamma}}$ in classic literature of partially linear additive model (Wang et al., 2014), where $\boldsymbol{\zeta}$ is directly observed. The covariance $\boldsymbol{V}_2$ is the extra variation, caused by the estimation error in the FPCA score $\widehat{\boldsymbol{\zeta}}$. The two sources of variation are asymptotically independent to each other because the model error ε is independent with the error in $\widehat{\boldsymbol{\zeta}}$. A similar effect of FPCA estimation error was discovered by 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

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 $\boldsymbol{Z}$. The multivariate functional predictor is $\boldsymbol{x}_i(t) = \{x_{i1}(t), x_{i2}(t)\}^\intercal$ with

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

where $\xi_{ik}^{(1)} \sim N(0, \varsigma_{2k-1})$, $\xi_{ik}^{(2)} \sim N(0, \varsigma_{2k})$, $\varsigma_k = 45.25k^{-2}$, $\text{corr}(\xi_{ik}^{(j)}, \xi_{ik'}^{(j)}) = 0$ for $k' \neq k$, and $\psi_k^{(j)}(t) = (1/\sqrt{5})\sin(\pi k t/10)$ for $t \in \mathcal{T} = [0, 10]$, $j = 1, 2$. 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 ϱ 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 $\text{var}\left(\{(\boldsymbol{\xi}_i^{(1)})^\intercal, (\boldsymbol{\xi}_i^{(2)})^\intercal\}^\intercal\right) = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}^\intercal$, where $\boldsymbol{\xi}_i^{(j)} = (\xi_{i1}^{(j)}, \ldots, \xi_{i10}^{(j)})^\intercal$, $j = 1, 2$, $\boldsymbol{Q} = \text{diag}\{\lambda_1, \ldots, \lambda_{20}\}$ and $\boldsymbol{P}^\intercal\boldsymbol{P} = \boldsymbol{I}$. The k-th mFPCA score $\xi_{ik} \sim N(0, \lambda_k)$ is a linear function of the univariate scores $\{(\boldsymbol{\xi}_i^{(1)})^\intercal, (\boldsymbol{\xi}_i^{(2)})^\intercal\}\boldsymbol{p}_k$, where $\boldsymbol{p}_k = \{(\boldsymbol{p}_k^{(1)})^\intercal, (\boldsymbol{p}_k^{(2)})^\intercal\}^\intercal$ is the k-th column of $\boldsymbol{P}$, and the corresponding mFPCA eigenfunction is $\boldsymbol{\psi}_k(t) = (\psi_{k1}, \psi_{k2})^\intercal(t)$, where $\psi_{kj}(t) = \{\boldsymbol{\psi}^{(j)}(t)\}^\intercal \boldsymbol{p}_k^{(j)}$, $\boldsymbol{\psi}^{(j)}(t) = (\psi_1^{(j)}, \ldots, \psi_{10}^{(j)})^\intercal(t)$, $j = 1, 2$.

From model (2), we simulate 1000 i.i.d. copies of $\{Y, \boldsymbol{Z}, \boldsymbol{X}(\cdot)\}$, denoted as $\{y_i, \boldsymbol{z}_i, \boldsymbol{x}_i(\cdot)\}_{i=1}^{1000}$, with the first 200 used as training data and the rest as testing data. Observations on $\boldsymbol{x}_i$ are

obtained on a regular grid of 100 points in $\mathcal{T} = [0, 10]$ with independent measurement errors following $N(0, 0.2^2)$. For the regression function, we set $f_0(\boldsymbol{\zeta}) = f_{01}(\zeta_{i1}) + f_{02}(\zeta_{i2}) + f_{04}(\zeta_{i4})$, 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 $\boldsymbol{z}_i$ independently from the bivariate uniform distribution over $[0, 1]^2$. We consider two settings for the partially linear coefficient $\boldsymbol{\theta}_0$: (I) $(1.4, 0, 0)^\intercal$ and (II) $(1.4, 3, -4)^\intercal$ and two settings of the correlation parameter ϱ: (i) 0.3 (low correlation) and (ii) 0.9 (high correlation). Combining different setups for ϱ and $\boldsymbol{\theta}_0$, we have four settings: {(i), (I)}, {(i), (II)}, {(ii), (I)} and {(ii), (II)}. The errors ε_i's in the regression model (2) are distributed independently as $N(0, \sigma^2)$ with σ^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 $\mathrm{var}(m_0(\boldsymbol{\zeta}))/\mathrm{var}(\varepsilon)$. For simplicity, all sampling weights π_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 $\boldsymbol{Z}$, and the proposed PLFAM. Throughout this simulation study, s is chosen to recover at least 99.9% of the total variation in $\{\boldsymbol{x}_i\}$ 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 $\widehat{f}_1$ $\widehat{f}_2$ and $\widehat{f}_4$, while Column "% super set" gives the percentages of fittings that include nonzero $\widehat{f}_1$, $\widehat{f}_2$ and $\widehat{f}_4$. 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_{0k}'s, Table 3 shows the averaged integrated squared errors (AISEs) of the first eight component functions and the overall function $\widehat{f} = \sum_{k=1}^{s} \widehat{f}_k$ (without constant term). The integrated squared errors are defined as

$$\mathrm{ISE}(\widehat{f}_k) = \int_0^1 \{\widehat{f}_k(t) - f_{0k}(t)\}^2 dt \quad \text{and} \quad \mathrm{ISE}(\widehat{f}) = \sum_{k=1}^{s} \int_0^1 \{\widehat{f}_k(t) - f_{0k}(t)\}^2 dt.$$

Notice that, under setting (I) where $\boldsymbol{Z}$ has zero effect and FAM is the correct model, the PLFAM estimators perform comparably to those of FAM. However, under setting (II), where $\boldsymbol{Z}$ 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 $\boldsymbol{X}$, 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 $n^{-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ on the testing data set. To compute the prediction $\widehat{y}_i$ in the test data, we first compute the transformed FPCA scores of $\boldsymbol{x}_i$ 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 $\widehat{m}$. The results in Table 4 suggest that jointly modeling multiple functional predictors leads to smaller MSE's for $\widehat{\boldsymbol{\theta}}$, 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 $\boldsymbol{Z}$.

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 $\mathcal{T} = [0, 365]$; $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$.

Since the response is an average obtained from an agricultural survey, the errors are heteroscedastic with weights π 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 1 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 2, we show the heat plots for the (cross-) covariance functions. The kernel function for $\mathcal{C}_{12}$ shows great resemblance to $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$, 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 $\boldsymbol{Z}$);
4. FAM(separate): FAM based on univariate FPC scores (without $\boldsymbol{Z}$);
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 $\boldsymbol{Z}$);
8. FLM(separate): FLM based on separate FPCA scores (without $\boldsymbol{Z}$);
9. LM: linear model on $\boldsymbol{Z}$ only;
10. LM-GDD: linear model on $\boldsymbol{Z}$ 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 $[\{X_1(t) + X_2(t)\}/2 - T_{base}]_+$ over growing season, where T_{base} is the crop-specific base temperature in °C. For corn $T_{base} = 8$, and for soybean $T_{base} = 10$. 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 2, 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 3, we show the top 6 most significant principal components; and in Figure 4, we show the corresponding additive component functions $\widehat{f}_k(\zeta)$. 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 $\widehat{f}_k$. The dashed curves in Figure 4 are the pointwise confidence bands $\widehat{f}_k(\zeta) \pm 2 \times se\{\widehat{f}_k(\zeta)\}$, 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 $\boldsymbol{\psi}_k(t) = \{\psi_{k1}(t), \psi_{k2}(t)\}^\intercal$, we show $\psi_{k1}(t)$ as the solid curve and $\psi_{k2}(t)$ as the dashed curve in each panel of Figure 3. 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 $\boldsymbol{\psi}_5$ 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 $\boldsymbol{\psi}_5$ means a milder summer, less heat stress and less chance of draught, and corn yield is an increasing function of ζ_5 in Figure 4. In contrast, $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_8$ represent hot summers, and crop yield is a decreasing function of their loadings ζ_1 and ζ_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 $\boldsymbol{\psi}_3$ is warm spring months from January to March. Hollinger et al. (1994) showed that warmer temperature during the period from planting to tassel initiation (the first 20 to 30 days after

planting) resulted in lower corn yields. This may be due to less snow coverage on the ground and more early insect activities. Our estimated $f_3(\zeta)$ in Figure 4 confirmed this finding, with corn yield a decreasing funciton of ζ_3 when ζ_3 is greater than 0.6. For soybean yield prediction, graphs of the selected eigenfunctions and the corresponding additive component functions are similar to those in Figures 3 and 4, 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 signifiant 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.1 and S.2 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 $\boldsymbol{Z}$. 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 mFPCA 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 $\boldsymbol{\gamma}$. 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

$$y_i = \boldsymbol{u}_i^\intercal \boldsymbol{\theta}_0 + \sum_{k=1}^{s} \sum_{j=1}^{d} f_{0jk}(\zeta_{ijk}) + \varepsilon_i, \tag{11}$$

where ζ_{ijk} is the kth 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. $\mathbb{E}(\widehat{\zeta}_{ijk} - \zeta_{ijk})^2 \leq Cn^{-1}k^{2\beta}$ 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).

References

Beran, J., Feng, Y., Ghosh, S., and Kulik, R. (2013). *Long-Memory Processes*. Springer, New York.

Brezis, H. (2010). *Functional analysis, Sobolev spaces and partial differential equations*. Springer, New York.

Cadson, R., Todey, D. P., and Taylor, S. E. (1996). Midwestern corn yield and weather in relation to extremes of the southern oscillation. *Journal of Production Agriculture*, 9(3):347–352.

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

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

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL.

Chiou, J.-M., Y.-T., C., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica*, 24:1571–1596.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104:155–1561.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154.

Gilmore, E. and Rogers, J. (1958). Heat units as a method of measuring maturity in corn. *Agronomy Journal*, 50(10):611–615.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B*, 68(1):109–126.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34:1493–1517.

Hansen, J. W. (2002). Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges. *Agricultural Systems*, 74(3):309–330.

Happ, C. and Greven, S. (2017). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of American Statistical Association*, page to appear.

Hollinger, S. E., Changnon, S. A., et al. (1994). Response of corn and soybean yields to precipitation augmentation, and implications for weather modification in illinois. *Bulletin/Illinois State Water Survey; no. 73*.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley.

James, G. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society, Series B*, 64:411–432.

James, G. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, 100:565–576.

Kim, J., Staicu, A.-M., Maity, A., Carroll, R. J., and Ruppert, D. (2017). Additive function-on-function regression. *Journal of Computational and Graphical Statistics*, to appear.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017). A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112:733–744.

Li, Y. and Hsing, T. (2010a). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Annals of Statistics*, 38:3028–3062.

Li, Y. and Hsing, T. (2010b). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 38:3321–3351.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105:621–633.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21:1225–1248.

Lobell, D. B. and Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, 25(3):1014–1035.

McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23:249 – 269.

McMaster, G. S. and Wilhelm, W. (1997). Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology*, 87(4):291–300.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821.

Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33:774–805.

Müller, H.-G., Wu, Y., and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, 103:607–622.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Nirenberg, L. (1959). On elliptic partial differential equations. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 13(2):115–162.

Prasad, A. K., Chai, L., Singh, R. P., and Kafatos, M. (2006). Crop yield estimation model for iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer, New York, 2nd edition.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statististical Society, Series B*, pages 1009–1030.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28:1–23.

Storlie, C. B., Bondell, H. D., Reich, B. J., and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21(2):679.

van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, L., Xue, L., Qu, A., and Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *Annals of Statistics*, 42:592–624.

Westcott, P. C., Jewison, M., et al. (2013). Weather effects on expected corn and soybean yields. *Washington DC: USDA Economic Research Service FDS-13g-01*.

Williams, J., Jones, C., Kiniry, J., and Spanel, D. (1989). The epic crop growth model. *Transactions of the ASAE*, 32(2):497–0511.

Yang, S., Logan, J., and Coffey, D. L. (1995). Mathematical formulae for calculating the base temperature for growing degree days. *Agricultural and Forest Meteorology*, 74(1):61–74.

Yao, F., Lei, E., and Wu, Y. (2016). Effective dimension reduction for sparse functional data. *Biometrika*, page to appear.

Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95:601–619.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *Journal of the Royal Statistical Society: Series B*, 76(3):581–603.

Table 1: Percentages of fitted model sizes.

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

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

Setting	Model	% for the following component functions								% correct	% super
		$\widehat{f}_1$	$\widehat{f}_2$	$\widehat{f}_3$	$\widehat{f}_4$	$\widehat{f}_5$	$\widehat{f}_6$	$\widehat{f}_7$	$\widehat{f}_8$	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.

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

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.

Setting	Model	Prediction error		Mean squared errors					
		separate	joint	separate			joint		
				$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$
{(i), (I)}	FAM	1.55 (0.10)	1.32 (0.13)	-	-	-	-	-	-
	PLFAM	1.57 (0.11)	1.33 (0.13)	0.0746	0.0911	0.1076	0.06	0.0751	0.0831
{(ii), (I)}	FAM	1.65 (0.09)	1.33 (0.12)	-	-	-	-	-	-
	PLFAM	1.66 (0.09)	1.35 (0.13)	0.0678	0.1095	0.0827	0.0585	0.0888	0.0681
{(i), (II)}	FAM	3.84 (0.22)	3.63 (0.21)	-	-	-	-	-	-
	PLFAM	1.59 (0.10)	1.34 (0.13)	0.0639	0.1023	0.0894	0.0545	0.0935	0.0696
{(ii), (II)}	FAM	3.89 (0.24)	3.60 (0.24)	-	-	-	-	-	-
	PLFAM	1.68 (0.12)	1.35 (0.14)	0.0642	0.0879	0.1092	0.0526	0.069	0.0851

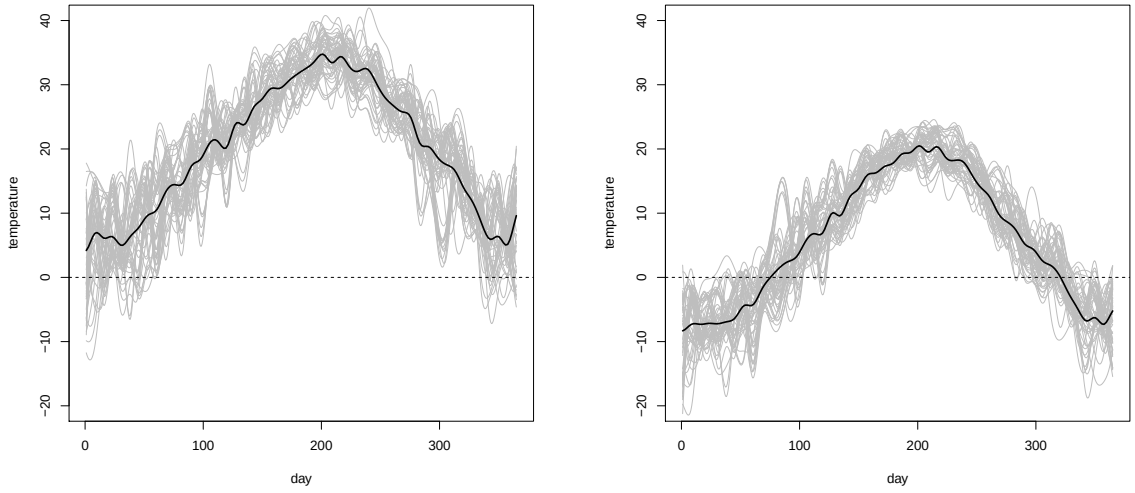
Table 5: Average of 5-year overall prediction errors.

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

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

	`Irrigate`	`Prec`	`Irrigat*Prec`
corn	168.38 (6.42)	20.98 (2.72)	-33.87 (3.32)
soybean	33.30 (3.35)	3.91 (0.70)	-4.88 (1.65)

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.



(a) $X_1(t)$ (b) $X_2(t)$

Figure 1: 50 randomly selected trajectories for daily maximum and daily minimum temperature. The solid dark curve in each panel is the mean function.

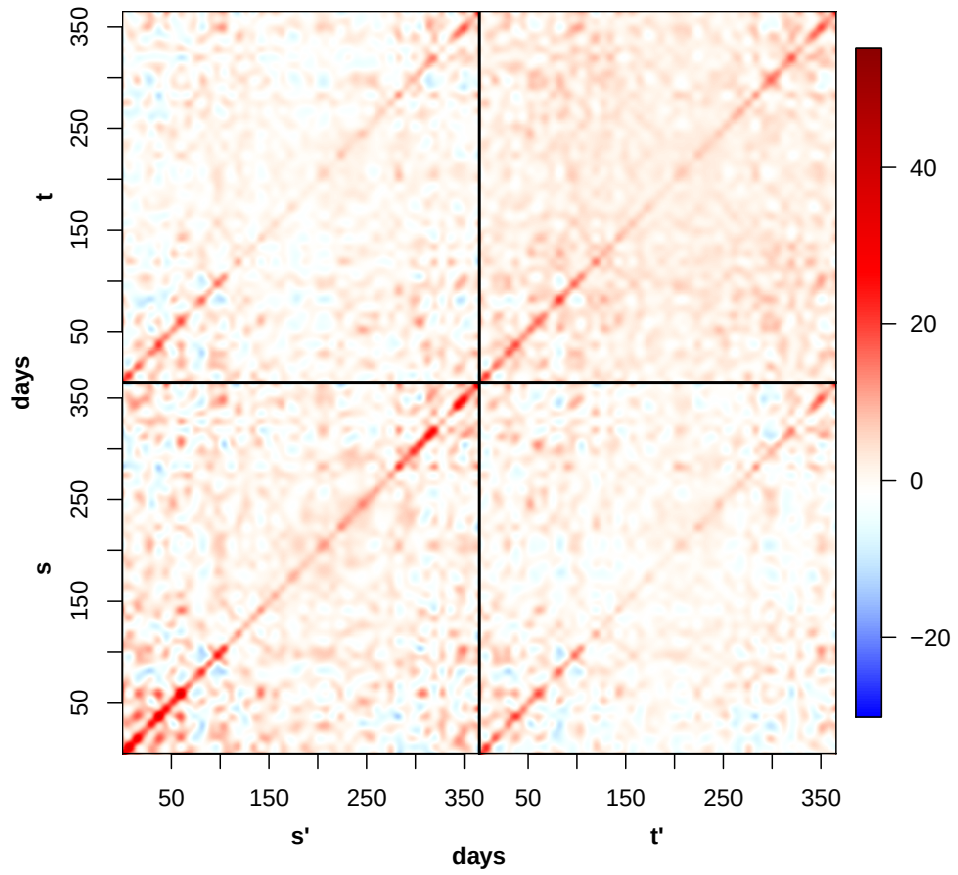


Figure 2: Heat plot for the covariance and cross-covariance functions. From bottom to top and from left to right are the kernel functions of the (cross-) covariance operators $\mathcal{C}_{jj'}$.

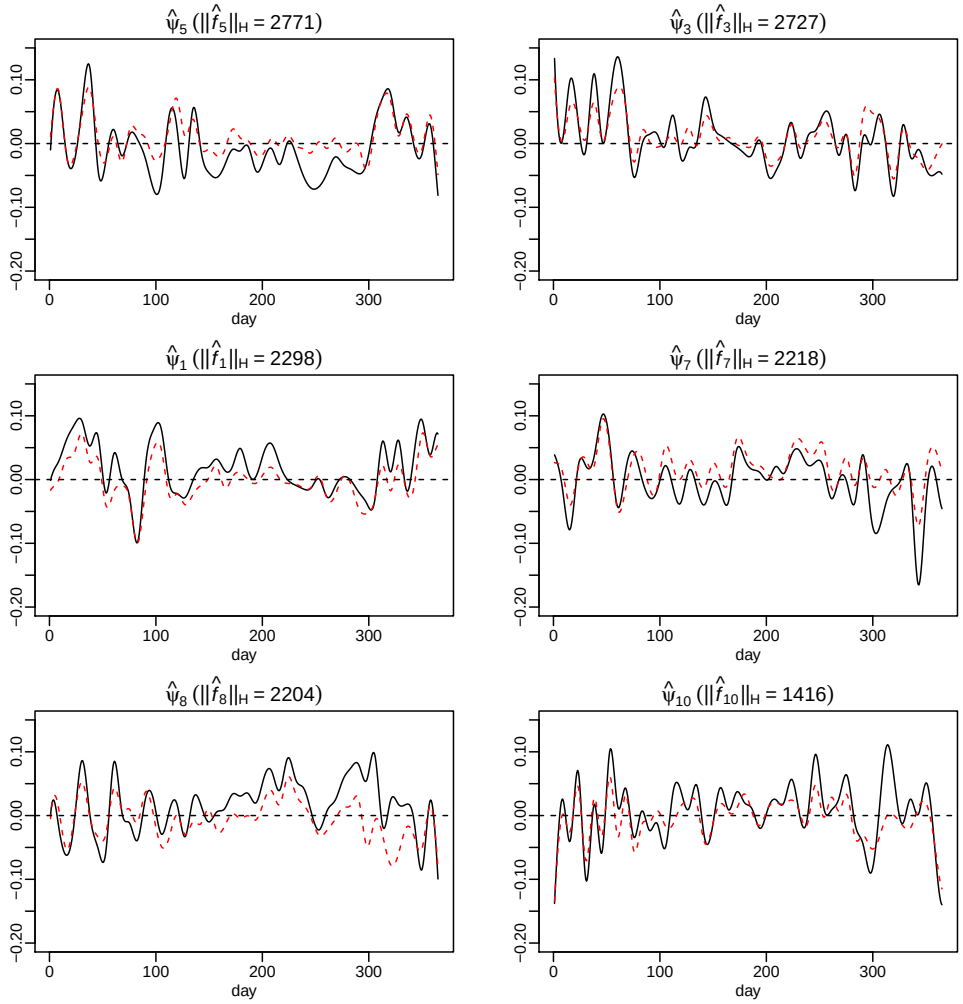


Figure 3: Corn yield prediction: top 6 principal components selected by COSSO for corn yield prediction, sorted by the decreasing order of the RKHS norm of $\widehat{f}_k(\zeta)$ $(k = 5, 3, 1, 7, 8, 10)$. Each principal component is a vector $\boldsymbol{\psi}_k(t) = \{\psi_{k1}(t), \psi_{k2}(t)\}^\intercal$. The solid curve in each panel is $\psi_{k1}(t)$ and the dashed curve is $\psi_{k2}(t)$.

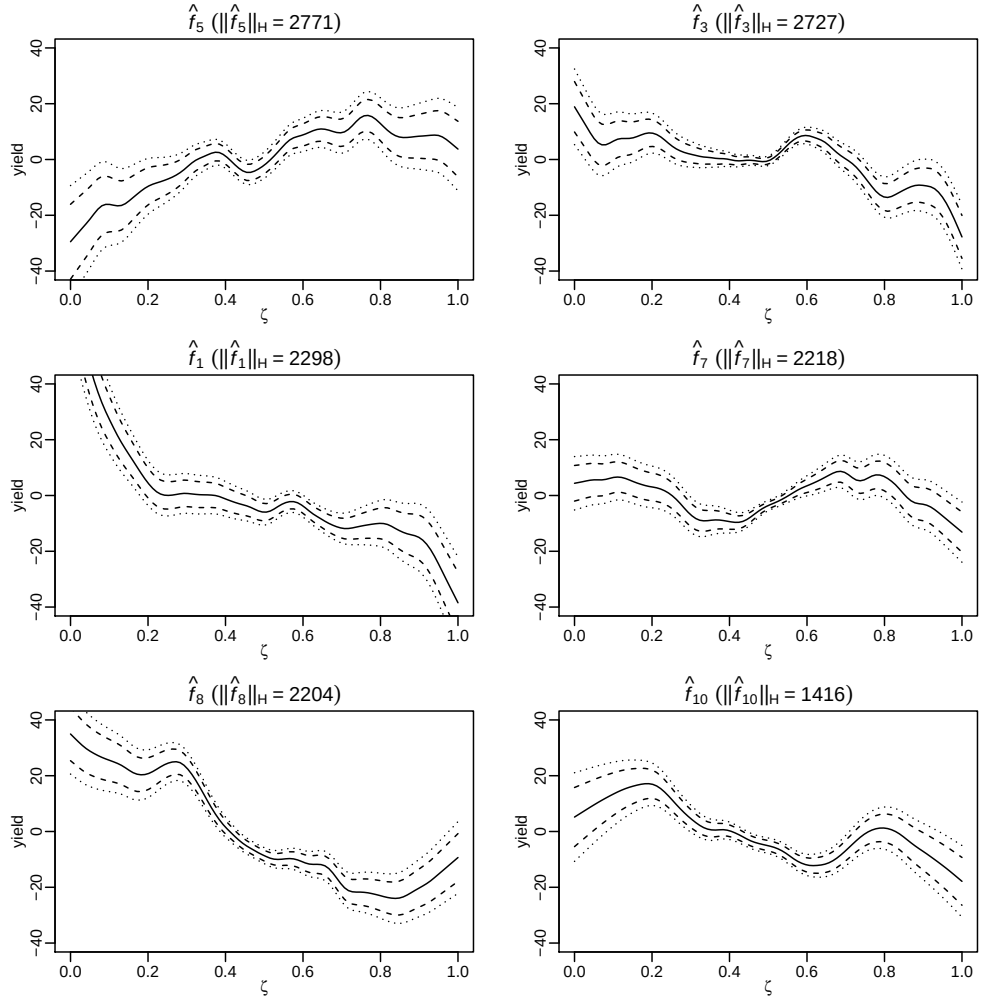


Figure 4: Corn yield prediction: top 6 additive component functions $\widehat{f}_k(\zeta)$, sorted by the decreasing order of the RKHS norm of $\widehat{f}_k(\zeta)$ ($k = 5, 3, 1, 7, 8, 10$).

Supplementary Material for
Partially Linear Functional Additive Models for Multivariate Functional Data

Raymond K. W. Wong[1], Yehua Li[2] and Zhengyuan Zhu[2]

[1]Department of Statistics, Texas A&M University, College Station, TX 77843

[2]Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011

The supplementary material is organized as follows. We provide a proof for Proposition 1 in Section A, theory for PLFAM (including proofs of Theorems 1 and 2) in Section B, additional simulation and data analysis results in Sections C and D, and the bootstrap procedure for standard error estimation in Section E.

A Theory for mFPCA

Proof of Proposition 1 We use C as a generic notation for positive constant. For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n \lesssim b_n$ to denote that a_n is bounded by b_n omitting some negligible terms. Recall that $\Delta = n^{1/2}(\widehat{\mathcal{C}} - \mathcal{C})$, and under Assumption 2 we have $\mathbb{E}\|\Delta\|_{\text{op}}^2 < \infty$.

Asymptotic expansions for the empirical eigenfunctions and eigenvalues similar to (2.8) and (2.9) in Hall and Hosseini-Nasab (2006) also hold for multivariate FPCA. For any k such that $\delta_k > n^{-1/2}\|\Delta\|_{\text{op}}$,

$$\begin{aligned} \widehat{\lambda}_k - \lambda_k &= n^{-1/2}\langle \Delta\boldsymbol{\psi}_k, \boldsymbol{\psi}_k\rangle_{\mathbb{X}} + \Lambda_{nk} \times \{1 + o_p(1)\}, \\ \widehat{\boldsymbol{\psi}}_k(t) - \boldsymbol{\psi}_k(t) &= \left\{ n^{-1/2} \sum_{j\neq k} (\lambda_k - \lambda_j)^{-1} \boldsymbol{\psi}_j \langle \Delta\boldsymbol{\psi}_k, \boldsymbol{\psi}_k\rangle_{\mathbb{X}} \right\} \times \{1 + \mathcal{O}_p(n^{-1/2}\delta_k^{-1})\}, \quad \text{(S.1)} \end{aligned}$$

where $\Lambda_{nk} = n^{-1}\sum_{j\neq k}(\lambda_k - \lambda_j)^{-1}(\langle \Delta\boldsymbol{\psi}_j, \boldsymbol{\psi}_k\rangle_{\mathbb{X}})^2 = n^{-1}\sum_{j\neq k}(\lambda_k - \lambda_j)^{-1}(n^{-1/2}\sum_{i=1}^n \xi_{ij}\xi_{ik})^2$. It is easy to see that $\mathbb{E}|\Lambda_{nk}| \leq (n\delta_k)^{-1}\sum_{j\neq k}\lambda_j\lambda_k \leq C(n\delta_k)^{-1}\lambda_k$ for all k.

Since $\widehat{\xi}_{ik} = \langle \boldsymbol{x}_i, \boldsymbol{\psi}_k\rangle_{\mathbb{X}}$, by the expansion (S.1),

$$\widehat{\xi}_{ik} - \xi_{ik} = A_{ik} \times \{1 + o_p(1)\} \quad \text{for all } k \leq J_n,$$

where $A_{ik} = n^{-1/2}\sum_{j\neq k}(\lambda_k - \lambda_j)^{-1}\xi_{ij}\langle \Delta\boldsymbol{\psi}_k, \boldsymbol{\psi}_j\rangle_{\mathbb{X}} = \sum_{j\neq k}(\lambda_k - \lambda_j)^{-1}\xi_{ij}(\frac{1}{n}\sum_{i_1=1}^n \xi_{i_1k}\xi_{i_1j})$.

Next, we calculate the order of A_{ik}. Denote $[x]$ as the integer part of x. By Assumption 1, $\lambda_j - \lambda_{j+1} \geq C_\lambda^{-1} j^{-\alpha-1}$,

$$\begin{aligned}
\lambda_j - \lambda_k &\geq C_\lambda^{-1}\sum_{l=j}^{k-1} l^{-\alpha-1} \geq C_\lambda^{-1}\int_j^k x^{-\alpha-1}dx \geq \frac{1}{C_\lambda\alpha}(j^{-\alpha} - k^{-\alpha}) \qquad \text{for } j < k; \\
\lambda_k - \lambda_j &\geq C_\lambda^{-1}\sum_{l=k}^{j-1} l^{-\alpha-1} \geq C_\lambda^{-1}\int_k^j x^{-\alpha-1}dx \geq \frac{1}{C_\lambda\alpha}(k^{-\alpha} - j^{-\alpha}) \qquad \text{for } j > k. \quad \text{(S.2)}
\end{aligned}$$

By Assumption 2 $\mathbb{E}(\frac{1}{n}\sum_{i_1=1}^n \xi_{i_1k}\xi_{i_1j})^2 \le C\lambda_k\lambda_j/n$ for all k and j, and by (S.2)

$$\begin{aligned}
\mathbb{E}(A_{ik}^2) &\lesssim \frac{C}{n}\sum_{j\neq k}(\lambda_k-\lambda_j)^{-2}\lambda_k\lambda_j^2 \\
&= \frac{C}{n}(\sum_{j<k}+\sum_{j>k})(\lambda_k-\lambda_j)^{-2}\lambda_k\lambda_j^2 \\
&\le \frac{C\lambda_k}{n}\Bigg\{\sum_{j=1}^{[(1-a)k]}\left(\frac{C_\lambda^2\alpha j^{-\alpha}}{j^{-\alpha}-k^{-\alpha}}\right)^2+\left(\sum_{j=[(1-a)k]+1}^{k-1}+\sum_{j=k+1}^{[(1+b)k]}\right)\frac{C_\lambda^2 j^{-2\alpha}}{C_\lambda^{-2}k^{-2\alpha-2}} \\
&\quad +\sum_{j=[(1+b)k]+1}^{\infty}\left(\frac{C_\lambda^2\alpha j^{-\alpha}}{k^{-\alpha}-j^{-\alpha}}\right)^2\Bigg\} \qquad \text{(for some } a,b\in(0,1)\text{)} \\
&\lesssim \frac{C\lambda_k}{n}\Bigg\{\sum_{j=1}^{[(1-a)k]}\left(\frac{1}{1-(j/k)^\alpha}\right)^2+\sum_{j=[(1+b)k]+1}^{\infty}\left(\frac{1}{(j/k)^\alpha-1}\right)^2+[(a+b)k]k^2\Bigg\} \\
&\lesssim \frac{Ck\lambda_k}{n}\Bigg\{\int_0^{1-a}(1-x^\alpha)^{-2}dx+\int_{(1+b)}^{\infty}(x^\alpha-1)^{-2}dx\Bigg\}+C(a+b)k^{3-\alpha}/n \\
&\lesssim \frac{Ck^{1-\alpha}}{n}\Bigg\{\int_0^{(1-a)}(1-y)^{-2}dy+\int_{(1+b)}^{\infty}(y-1)^{-2}dy\Bigg\}+C(a+b)k^{3-\alpha}/n \\
&\lesssim \frac{Ck^{1-\alpha}}{n}\{a^{-1}-1+b^{-1}+(a+b)k^2\}.
\end{aligned}$$

We select $a\sim k^{-1}$ and $b\sim k^{-1}$, we get $\mathbb{E}A_{ik}^2 \le Ck^{2-\alpha}/n$ for all k. This implies $\widehat{\xi}_{ik}-\xi_{ik} = \mathcal{O}_p(n^{-1/2}k^{1-\alpha/2})$ uniformly for $k\le J_n$.

On the other hand, by (S.1) we can show

$$\begin{aligned}
&\mathbb{E}\left|\widehat{\lambda}_k-\lambda_k-n^{-1/2}\langle\Delta\boldsymbol{\psi}_k,\boldsymbol{\psi}_k\rangle\right| \lesssim \mathbb{E}|\Lambda_{nk}| \le Cn^{-1}\lambda_k\delta_k^{-1}, \\
&\mathbb{E}(n^{-1/2}\langle\Delta\boldsymbol{\psi}_k,\boldsymbol{\psi}_k\rangle)^2 = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n(\xi_{ik}^2-\lambda_k)\right\}^2 \le C\lambda_k^2n^{-1}.
\end{aligned}$$

This also means $\widehat{\lambda}_k-\lambda_k = \mathcal{O}_p(n^{-1/2}\lambda_k)$ uniformly for all $k\le J_n$. Since $\Phi(\cdot)$ is differentiable

transformation function, using the delta method

$$\begin{aligned}\widehat{\zeta}_{ik} &\approx \Phi[(\xi_{ik}+A_{ik})\{\lambda_k^{-1/2}-(1/2)\lambda_k^{-3/2}(\widehat{\lambda}_k-\lambda_k)\}] \\ &\approx \zeta_{ik}+\Phi'(\xi_{ik}\lambda_{ik}^{-1/2})\{\lambda_k^{-1/2}A_{ik}-\frac{1}{2}\xi_{ik}\lambda_k^{-3/2}(\widehat{\lambda}_k-\lambda_k)\} \\ &= \zeta_{ik}+\mathcal{O}_p(n^{-1/2}k).\end{aligned} \tag{S.3}$$

By the assumption that $|\Phi'(x)| < C$ for all x and the mean-value theorem, one can verify that $\mathbb{E}(\widehat{\zeta}_{ik}-\zeta_{ik})^2 \le Cn^{-1}k^2$ uniformly for all $k \le J_n$.

B Theory for PLFAM

Throughout the theoretical development, we utilize the following representation of a generic function $m \in \mathbb{M}$:

$$m(\boldsymbol{u},\boldsymbol{\zeta}) \;\; = \boldsymbol{u}^\intercal\boldsymbol{\nu}+h(\boldsymbol{u},\boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\boldsymbol{\nu}+\textstyle\sum_{k=1}^s h_k(\boldsymbol{u},\zeta_k),$$

where $h_k \in \mathbb{H}_k = \{h_k \in \mathbb{I}\oplus\bar{\mathbb{F}}_k : \sum_{i=1}^n h_k(\boldsymbol{u}_i,\widehat{\zeta}_{ik})u_{ij}=0, j=1\ldots,p+1\}$ for $k=1,\ldots,s$. Note that the set $\mathbb{H}_k$ depends on $\{\boldsymbol{u}_i\}$ and $\{\widehat{\boldsymbol{\zeta}}_i\}$ and thus is a random set with randomness inherited from them. Write $\boldsymbol{U}=[u_{ij}]_{i=1,\ldots,n,j=1,\ldots,p+1}$. Given $m(\boldsymbol{u},\boldsymbol{\zeta})=\boldsymbol{u}^\intercal\boldsymbol{\theta}+\sum_{k=1}^s f_k(\zeta_k)$, where $f_k \in \bar{\mathbb{F}}_k$, one can transform it into the aforementioned representation by setting $\boldsymbol{\nu}=\boldsymbol{\theta}-\sum_{k=1}^s\boldsymbol{\omega}_k$ and $h_k(\boldsymbol{u},\zeta_k)=\boldsymbol{u}^\intercal\boldsymbol{\omega}_k+f_k(\zeta_k)$, where $\boldsymbol{\omega}_k$ fulfills

$$\frac{1}{n}\boldsymbol{U}^\intercal\boldsymbol{U}\boldsymbol{\omega}_k = -\frac{1}{n}\boldsymbol{U}^\intercal(f_k(\widehat{\zeta}_{k1}),\ldots,f_k(\widehat{\zeta}_{kn}))^\intercal.$$

Similarly, $m_0(\boldsymbol{u},\boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\boldsymbol{\nu}_0+h_0(\boldsymbol{u},\boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\boldsymbol{\nu}_0+\sum_{k=1}^s h_{0k}(\boldsymbol{u},\zeta_k)$ and $\widehat{m}(\boldsymbol{u},\boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\widehat{\boldsymbol{\nu}}+\widehat{h}(\boldsymbol{u},\boldsymbol{\zeta})=\boldsymbol{u}^\intercal\widehat{\boldsymbol{\nu}}+\sum_{k=1}^s\widehat{h}_k(\boldsymbol{u},\zeta_k)$. Moreover, write $\mathbb{H}=\sum_{k=1}^s\mathbb{H}_k$.

Similar to P_n, we write $P_{n,*}$ as the empirical distributions of $(\boldsymbol{Z},\widehat{\boldsymbol{\zeta}})$. That is, $P_{n,*}=$

$\sum_{i=1}^n \delta_{\boldsymbol{z}_i,\widehat{\boldsymbol{\zeta}}_i}/n$. Moreover, we define the corresponding version of (squared) empirical norm and inner product as

$$\|m_1\|_{n,*}^2 = \int m_1^2 dP_{n,*} \quad \text{and} \quad (m_1, m_2)_{n,*} = \int m_1 m_2 dP_{n,*}, \qquad \text{for any } m_1, m_2 \in \mathbb{M}.$$

First, we prove the following proposition about the convergence with respect to the empirical norm $\|\cdot\|_{n,*}$ rather than the intended $\|\cdot\|_n$.

Proposition 2 *Suppose* $s = \mathcal{O}_p(n^{1/\{2(1+\alpha)\}})$ *and* $\mathbb{E}(\widehat{\zeta}_{ik} - \zeta_{ik})^2 \leq Cn^{-1}k^{2\beta}$ *uniformly for all* $k \leq s$. *Further, assume* $J(m_0) < \infty$ *and* $\boldsymbol{\Sigma}$ *is non-singular. If* $\tau_n^{-1} = \mathcal{O}_p(\min\{n^{2/5}s^{-6/5}, n^{1/2}s^{-(\frac{1}{2}+\beta)}\})$, *we have* $\|\widehat{m} - m_0\|_{n,*} = \mathcal{O}_p(\tau_n)$ *and* $J(\widehat{m}) = \mathcal{O}_p(1)$. *If* $J(m_0) = 0$ *and* $\tau_n \asymp n^{-1/4}s^3$, $\|\widehat{m} - m_0\|_{n,*} = \mathcal{O}_p(n^{-1/2})$ *and* $J(\widehat{m}) = \mathcal{O}_p(n^{-1/2}s^{-6})$.

The proof of Proposition 2 is given in Section B.1. By Taylor expansion arguments and convergence of $\widehat{\boldsymbol{\zeta}}$, the convergence results based on $\|\cdot\|_n$ (Theorem 1) is implied by those based on $\|\cdot\|_{n,*}$ (Proposition 2). See Section B.2 for the proof of Theorem 1. With convergence of $\widehat{m}$, we study the parametric part in details and obtain the optimal $\sqrt{n}$-consistency for $\widehat{\boldsymbol{\gamma}}$. The details is shown in Section B.3.

For ease of reading, we collect all other lemmas that are used throughout the subsequent proofs here. Their proofs are deferred to Section B.4.

Lemma 2 *For any* $f(\boldsymbol{\zeta}) = \sum_{k=1}^s f_k(\zeta_k) \in \sum_{k=1}^s \mathbb{F}_k$, *there exists* C_2 *(independent of* s*) such that*

$$\max_{1\leq k\leq s} \sup_{\zeta_k\in[0,1]} \left|\frac{\partial f_k(\zeta_k)}{\partial \zeta_k}\right| \Big/ \|f_k\| \leq C_2. \tag{S.4}$$

Lemma 3 (Entropy result) *Assume* $\boldsymbol{\Sigma}$ *is non-singular. Then there exists constants* C_1 *and* C_1' *such that the events*

$$\liminf_n \left\{ \sup_{\delta>0} \delta^{1/2} H_\infty(\delta, \{h_k \in \mathbb{H}_k : J(h_k) \leq 1\}) \leq C_1 \right\},$$

$$\liminf_n \left\{ \sup_{\delta>0} \delta^{1/2} H_\infty(\delta, \{h \in \mathbb{H} : J(h) \le 1\}) \le C_1 s^{3/2} \right\}$$

and

$$\liminf_n \left\{ \sup_{h \in \mathbb{H}: J(h) \le 1} |h|_\infty \le C_1' s \right\}$$

are of probability 1.

Lemma 4 *Assume* $\mathbf{\Sigma}$ *is non-singular. We have*

$$\sup_{h \in \mathbb{H}} \frac{|(\varepsilon, h - h_0)_{n,*}|}{\|h - h_0\|_{n,*}^{3/4} \{J(h) + J(h_0)\}^{1/4}} = \mathcal{O}_p(n^{-1/2} s^{3/2}),$$

where $m_0(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{u}^\intercal \boldsymbol{\nu}_0 + h_0(\boldsymbol{u}, \boldsymbol{\zeta})$ *with* $\boldsymbol{\nu}_0 \in \mathbb{R}^{p+1}$ *and* $h_0 \in \mathbb{H}$.

B.1 Proof of Proposition 2

Proof of Proposition 2

Expanding the objective function, we have

$$\begin{aligned}
\ell(m) &= \frac{1}{n} \sum_{i=1}^n \{y_i - m(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 + \tau_n^2 J(m) \\
&= \frac{1}{n} \sum_{i=1}^n \{\boldsymbol{u}_i^\intercal \boldsymbol{\nu}_0 + h_0(\boldsymbol{u}_i, \boldsymbol{\zeta}_i) + \varepsilon_i - \boldsymbol{u}_i^\intercal \boldsymbol{\nu} - h(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 + \tau_n^2 J(h) \\
&= \frac{1}{n} \sum_{i=1}^n \{\boldsymbol{u}_i^\intercal (\boldsymbol{\nu}_0 - \boldsymbol{\nu})\}^2 + \frac{2}{n} \sum_{i=1}^n \{\boldsymbol{u}_i^\intercal (\boldsymbol{\nu}_0 - \boldsymbol{\nu})\} \{h_0(\boldsymbol{u}_i, \boldsymbol{\zeta}_i) + \varepsilon_i\} \\
&\qquad + \frac{1}{n} \sum_{i=1}^n \{h_0(\boldsymbol{u}_i, \boldsymbol{\zeta}_i) + \varepsilon_i - h(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 + \tau_n^2 J(h).
\end{aligned}$$

Minimizing ℓ is equivalent to the following two minimizations:

$$\begin{aligned}
\widehat{\boldsymbol{\nu}} &= \arg\min_{\boldsymbol{\nu}\in\mathbb{R}^{p+1}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{u}_i^\intercal(\boldsymbol{\nu}_0-\boldsymbol{\nu})\}^2 + \frac{2}{n}\sum_{i=1}^{n}\{\boldsymbol{u}_i^\intercal(\boldsymbol{\nu}_0-\boldsymbol{\nu})\}\{h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)+\varepsilon_i\} \right\}, \\
\widehat{h} &= \arg\min_{h\in\mathbb{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\{h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)+\varepsilon_i - h(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}^2 + \tau_n^2 J(h) \right\}.
\end{aligned}$$

The first one leads to

$$\frac{1}{n}\boldsymbol{U}^\intercal\boldsymbol{U}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_0) = \frac{1}{n}\boldsymbol{U}^\intercal(\boldsymbol{h}_0+\boldsymbol{\varepsilon}),$$

where $\boldsymbol{U}=[u_{ij}]_{i=1,\ldots,n,j=1,\ldots,p+1}$, $\boldsymbol{h}_0=(h_0(\boldsymbol{u}_1,\boldsymbol{\zeta}_1),\ldots,h_0(\boldsymbol{u}_n,\boldsymbol{\zeta}_n))^\intercal$ and $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\intercal$. By Taylor expansion of h_0 with respect to $\boldsymbol{\zeta}$ at $\widehat{\boldsymbol{\zeta}}_i$ and the fact that $D_{\boldsymbol{\zeta}}h_0 = D_{\boldsymbol{\zeta}}f_0$,

$$\frac{1}{n}\sum_{i=1}^{n}u_{ij}h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i) = \frac{1}{n}\sum_{i=1}^{n}u_{ij}D_{\boldsymbol{\zeta}}f_0(\boldsymbol{\zeta}_i^*)(\boldsymbol{\zeta}_i-\widehat{\boldsymbol{\zeta}}_i) = J(f_0)\,\mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta}) \tag{S.5}$$

where $\boldsymbol{\zeta}_i^*$ lies on the line segment joining $\boldsymbol{\zeta}_i$ and $\widehat{\boldsymbol{\zeta}}_i$; and the last equality follows from the assumption $\mathbb{E}(\widehat{\zeta}_{ik}-\zeta_{ik})^2 \le Cn^{-1}k^{2\beta}$, Lemma 2 and the following calculation

$$\begin{aligned}
|D_{\boldsymbol{\zeta}}f_0(\boldsymbol{\zeta}_i^*)(\boldsymbol{\zeta}_i-\widehat{\boldsymbol{\zeta}}_i)| &= \left|\sum_{k=1}^{s}\frac{\partial}{\partial\zeta_k}f_{0k}(\zeta_{ik}^*)(\widehat{\zeta}_{ik}-\zeta_{ik})\right| \\
&\le \left\{\sum_{k=1}^{s}\left|\frac{\partial}{\partial\zeta_k}f_{0k}(\zeta_{ik}^*)\right|^2\right\}^{1/2}\left\{\sum_{k=1}^{s}(\widehat{\zeta}_{ik}-\zeta_{ik})^2\right)^{1/2} \\
&\le \left(\sum_{k=1}^{s}\|f_{0k}\|^2\right\}^{1/2}\times\left\{\mathcal{O}_p\Big(\sum_{k=1}^{s}n^{-1}k^{2\beta}\Big)\right\}^{1/2} \\
&= \|f_0\|\times\mathcal{O}_p(n^{-1/2}s^{\beta+\frac{1}{2}}).
\end{aligned}$$

Moreover,

$$\frac{1}{n}\sum_{i=1}^{n}u_{ij}\varepsilon_i = \mathcal{O}_p(n^{-1/2}).$$

Since $\boldsymbol{U}^\intercal\boldsymbol{U}/n \to \boldsymbol{\Sigma}$ almost surely (element-wisely) and $\boldsymbol{\Sigma}$ is non-singular, we have $\|\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_0\|_E = \mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta})$. Note that if $J(f_0)=0$, we have $\boldsymbol{U}^\intercal\boldsymbol{h}_0=\boldsymbol{0}$ and $\|\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_0\|_E =$

$\mathcal{O}_p(n^{-1/2})$.

In sequel, we focus on the second optimization. Since $\widehat{h}$ is the minimizer,

$$\frac{1}{n}\sum_{i=1}^{n}\{h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)+\varepsilon_i-\widehat{h}(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}^2+\tau_n^2 J(\widehat{h})\le\frac{1}{n}\sum_{i=1}^{n}\{h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)+\varepsilon_i-h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}^2+\tau_n^2 J(h_0),$$

which leads to

$$\begin{aligned}\|h_0-\widehat{h}\|_{n,*}^2+\tfrac{2}{n}\textstyle\sum_{i=1}^{n}\{h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)-h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}\{h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)-\widehat{h}(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}+\tau_n^2 J(\widehat{h})\\ \le(\varepsilon,\widehat{h}-h_0)_{n,*}+\tau_n^2 J(h_0).\end{aligned}\tag{S.6}$$

Now, we utilize the previous Taylor expansions: For $i=1,\ldots,n$,

$$h_0(\boldsymbol{u}_i,\boldsymbol{\zeta}_i)=h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)+D_{\boldsymbol{\zeta}}f_0(\boldsymbol{\zeta}_i^*)(\boldsymbol{\zeta}_i-\widehat{\boldsymbol{\zeta}}_i).$$

Thus (B.1) becomes

$$\|\widehat{h}-h_0\|_{n,*}^2+\frac{2}{n}\sum_{i=1}^{n}\{\widehat{h}(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)-h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}\{D_{\boldsymbol{\zeta}}f_0(\boldsymbol{\zeta}_i^*)(\widehat{\boldsymbol{\zeta}}_i-\boldsymbol{\zeta}_i)\}+\tau_n^2 J(\widehat{h})\le 2(\varepsilon,\widehat{h}-h_0)_{n,*}+\tau_n^2 J(h_0).\tag{S.7}$$

Now we derive asymptotic order of the following two terms:

$$\begin{aligned}&\left|\frac{2}{n}\sum_{i=1}^{n}\{\widehat{h}(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)-h_0(\boldsymbol{u}_i,\widehat{\boldsymbol{\zeta}}_i)\}\{D_{\boldsymbol{\zeta}}f_0(\boldsymbol{\zeta}_i^*)(\widehat{\boldsymbol{\zeta}}_i-\boldsymbol{\zeta}_i)\}\right|\\ \le\;& 2\|\widehat{h}-h_0\|_{n,*}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{s}\frac{\partial f_{0k}(\zeta_{ik}^*)}{\partial\zeta_{ik}}(\widehat{\zeta}_{ik}-\zeta_{ik})\right\}^2\right)^{1/2}\\ \le\;& J(h_0)\|\widehat{h}-h_0\|_{n,*}\,\mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta});\end{aligned}$$

and by Lemma 4,

$$2(\varepsilon, \widehat{h} - h_0)_{n,*} = \mathcal{O}_p(n^{-1/2}s^{3/2})\|\widehat{h} - h_0\|_{n,*}^{3/4}\{J(\widehat{h}) + J(h_0)\}^{1/4}.$$

Collecting the above results, (S.7) leads to

$$\begin{aligned}\|\widehat{h} - h_0\|_{n,*}^2 + \tau_n^2 J(\widehat{h}) \;\; &\leq \mathcal{O}_p(n^{-1/2}s^{3/2})\|\widehat{h} - h_0\|_{n,*}^{3/4}\{J(\widehat{h}) + J(h_0)\}^{1/4} + \tau_n^2 J(h_0) \\ &\quad + J(h_0)\|\widehat{h} - h_0\|_{n,*}\,\mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta}).\end{aligned}$$

Next, we investigate the following three scenarios where one particular term on the right hand side dominates the other two.

(A) The term $\mathcal{O}_p(n^{-1/2}s^{3/2})\|\widehat{h} - h_0\|_{n,*}^{3/4}\{J(\widehat{h}) + J(h_0)\}^{1/4}$ is the largest: Thus

$$\|\widehat{h} - h_0\|_{n,*}^2 + \tau_n^2 J(\widehat{h}) \leq \mathcal{O}_p(n^{-1/2}s^{3/2})\|\widehat{h} - h_0\|_{n,*}^{3/4}\{J(\widehat{h}) + J(h_0)\}^{1/4}.$$

If $J(\widehat{h}) \geq J(h_0)$, one can deduce that $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(n^{-2/3}s^2)\tau_n^{-2/3}$ and $J(\widehat{h}) = \mathcal{O}_p(n^{-4/3}s^4)\tau_n^{-10/3}$. As for $J(\widehat{h}) < J(h_0)$, we have $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(n^{-2/5}s^{6/5})J(h_0)^{1/5}$ and $J(\widehat{h}) = \mathcal{O}_p(1)J(h_0)$.

(B) The term $\tau_n^2 J(h_0)$ is the largest: Thus

$$\|\widehat{h} - h_0\|_{n,*}^2 + \tau_n^2 J(\widehat{h}) \leq \tau_n^2 J(h_0)\,\mathcal{O}_p(1),$$

which leads to $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(\tau_n)J^{1/2}(h_0)$ and $J(\widehat{h}) = \mathcal{O}_p(1)J(h_0)$.

(C) The term $J(h_0)\|\widehat{h} - h_0\|_{n,*}\,\mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta})$ is the largest: Thus

$$\|\widehat{h} - h_0\|_{n,*}^2 + \tau_n^2 J(\widehat{h}) \leq J(h_0)\|\widehat{h} - h_0\|_{n,*}\,\mathcal{O}_p(n^{-1/2}s^{\frac{1}{2}+\beta}),$$

which leads to

$$\begin{cases} \|\widehat{h} - h_0\|_{n,*} \leq J(h_0)\, \mathcal{O}_p(n^{-1/2} s^{\frac{1}{2}+\beta}), \\ \tau_n^2 J(\widehat{h}) \leq \|\widehat{h} - h_0\|_{n,*}\, \mathcal{O}_p(n^{-1/2} s^{\frac{1}{2}+\beta}). \end{cases}$$

Thus $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(n^{-1/2} s^{\frac{1}{2}+\beta}) J(h_0)$ and $J(\widehat{h}) = \mathcal{O}_p(n^{-1} s^{(1+2\beta)}) \tau_n^{-2} J^2(h_0)$.

By carefully comparing the stochastic orders of terms arising from the above three cases, if $\tau_n^{-1} = \mathcal{O}_p(\min\{n^{2/5} s^{-6/5}, n^{1/2} s^{-(\frac{1}{2}+\beta)}\})$, we have $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(\tau_n)$ and $J(\widehat{h}) = \mathcal{O}_p(1)$. If $J(h_0) = 0$ and $\tau_n \asymp n^{-1/4} s^3$, $\|\widehat{h} - h_0\|_{n,*} = \mathcal{O}_p(n^{-1/2})$ and $J(\widehat{h}) = \mathcal{O}_p(n^{-1/2} s^{-6})$.

B.2 Proof of Theorem 1

Proof of Theorem 1. Let $q = \widehat{m} - m_0$. By Taylor expansion,

$$\begin{aligned} \|q\|_n^2 &= \frac{1}{n} \sum_{i=1}^n \{q(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i) + D_{\boldsymbol{\zeta}} q(\boldsymbol{u}_i, \widetilde{\boldsymbol{\zeta}}_i)(\boldsymbol{\zeta}_i - \widehat{\boldsymbol{\zeta}}_i)\}^2 \\ &= \|q\|_{n,*}^2 + \frac{1}{n} \sum_{i=1}^n \{D_{\boldsymbol{\zeta}} q(\boldsymbol{u}_i, \widetilde{\boldsymbol{\zeta}}_i)(\boldsymbol{\zeta}_i - \widehat{\boldsymbol{\zeta}}_i)\}^2 + \frac{1}{n} \sum_{i=1}^n 2q(\widehat{\boldsymbol{\zeta}}_i) \{D_{\boldsymbol{\zeta}} q(\boldsymbol{u}_i, \widetilde{\boldsymbol{\zeta}}_i)(\boldsymbol{\zeta}_i - \widehat{\boldsymbol{\zeta}}_i)\} \end{aligned}$$

where $\widetilde{\boldsymbol{\zeta}}_i$ lies in the line segment joining $\boldsymbol{\zeta}_i$ and $\widehat{\boldsymbol{\zeta}}_i$. By calculation similar to (S.5), we have

$$\begin{aligned} &\tfrac{1}{n} \textstyle\sum_{i=1}^n \{D_{\boldsymbol{\zeta}} q(\boldsymbol{u}_i, \widetilde{\boldsymbol{\zeta}}_i)(\boldsymbol{\zeta}_i - \widehat{\boldsymbol{\zeta}}_i)\}^2 = J(q)\, \mathcal{O}_p(n^{-1} s^{(1+2\beta)}), \\ &\tfrac{1}{n} \textstyle\sum_{i=1}^n 2q(\widehat{\boldsymbol{\zeta}}_i) \{Dq(\boldsymbol{u}_i, \widetilde{\boldsymbol{\zeta}}_i)(\boldsymbol{\zeta}_i - \widehat{\boldsymbol{\zeta}}_i)\} = \|q\|_{n,*} J(q)\, \mathcal{O}_p(n^{-1/2} s^{\frac{1}{2}+\beta}). \end{aligned}$$

By Proposition 2, if $\tau_n^{-1} = \mathcal{O}_p(\min\{n^{2/5} s^{-6/5}, n^{1/2} s^{-(\frac{1}{2}+\beta)}\})$, $J(\widehat{m}) = \mathcal{O}_p(1)$ and

$$\|q\|_n^2 = \|q\|_{n,*}^2 + \mathcal{O}_p(n^{-1} s^{(1+2\beta)}) + \|q\|_{n,*}\, \mathcal{O}_p(n^{-1/2} s^{\frac{1}{2}+\beta}) = \mathcal{O}_p(\tau_n^2).$$

If $J(m_0) = 0$ and $\tau_n \asymp n^{-1/4} s^3$, $J(\widehat{m}) = \mathcal{O}_p(n^{-1/2} s^{-6})$ from Proposition 2. Similarly as the proof of Proposition 2, write $\widehat{m}(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{u}^\intercal \widehat{\boldsymbol{\nu}} + \widehat{h}(\boldsymbol{u}, \boldsymbol{\zeta})$. In its proof, we show that $\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0\|_E = \mathcal{O}_p(n^{-1/2})$ and $\|h_0\|_n = 0$ (due to $\boldsymbol{U}^\intercal \boldsymbol{h}_0 = \boldsymbol{0}$). By Lemma 3, we have $|\widehat{h}|_\infty =$

$J(\widehat{h})\,\mathcal{O}_p(1) = \mathcal{O}_p(n^{-1/2}s^{-6})$ since $J(\widehat{h}) = J(\widehat{m})\,\mathcal{O}_p(n^{-1/2}s^{-6})$. Since $\boldsymbol{u} \in [0,1]^{p+1}$, $\|q\|_n \leq \|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0\|_E + \|\widehat{h}\|_n = \mathcal{O}_p(n^{-1/2})$.

B.3 Proof of Theorem 2

We first introduce a few Lemmas, the proof of which is relegated to Section B.4.

Lemma 5 *Under the conditions of Theorem 2,* $\|\widehat{m} - m_0\|_2 = o_p(n^{-1/4})$*, where* $\|\cdot\|_2$ *represents the* $L_2(P)$*-norm..*

Lemma 6 *For any* $k = 1, \ldots, s$ *and* $g_k \in \bar{\mathbb{F}}_k$*, we have*

$$\sup_{g_k \in \bar{\mathbb{F}}_k} \frac{\left| \|g_k^{(1)}\|_n^2 - \|g_k^{(1)}\|_2^2 \right|}{\|g_k\|^2} = o_p(1).$$

Lemma 7 *Under the conditions of Theorem 2,* $\|\widehat{f}_k' - f_{0k}'\|_n^2 = o_p(1)$ *for all* $k = 1, \ldots, s$*.*

Proof of Theorem 2. Write $\widehat{m}(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{z}^\intercal \widehat{\boldsymbol{\gamma}} + \widehat{g}(\boldsymbol{\zeta})$ and $m_0(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{z}^\intercal \boldsymbol{\gamma}_0 + g_0(\boldsymbol{\zeta})$ where $\widehat{g}, g_0 \in \sum_{k=1}^s \mathbb{F}_k$ and $\boldsymbol{u} = (1, \boldsymbol{z}^\intercal)^\intercal$. We also write $\widehat{g}_k = \mathcal{P}_k \widehat{g} \in \bar{\mathbb{F}}_k$ and $g_{0k} = \mathcal{P}_k g_0 \in \bar{\mathbb{F}}_k$ for $k = 1, \ldots, s$. Note that $\widehat{g}$ and $\sum_{k=1}^s \widehat{g}_k$ may differ by a constant. Similarly for g_0 and $\sum_{k=1}^s g_{0k}$.

By expanding $\|\widehat{m} - m_0\|_2^2 = \|\widetilde{\boldsymbol{w}}^\intercal(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\|_2^2 + \|\boldsymbol{w}^\intercal(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \widehat{g} - g_0\|_2^2$, we show that $\|\widetilde{\boldsymbol{w}}^\intercal(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\|_2^2 = o_p(n^{-1/4})$ using Lemma 5. By the condition that $\boldsymbol{M}$ is non-singular, we have

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_E = o_p(n^{-1/4}) \qquad \text{and} \qquad \|\widehat{g} - g_0\|_2 = o_p(n^{-1/4}). \tag{S.8}$$

Recall that $\widetilde{\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\zeta}) = \boldsymbol{z} - \boldsymbol{w}(\boldsymbol{\zeta})$. We then define

$$\widehat{m}_{\boldsymbol{\rho}}(\boldsymbol{z}, \boldsymbol{\zeta}) = \widehat{m}(\boldsymbol{u}, \boldsymbol{\zeta}) + \boldsymbol{\rho}^\intercal \widetilde{\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\zeta}) = \boldsymbol{z}^\intercal(\widehat{\boldsymbol{\gamma}} + \boldsymbol{\rho}) + \{\widehat{g}(\boldsymbol{\zeta}) - \boldsymbol{\rho}^\intercal \boldsymbol{w}(\boldsymbol{\zeta})\},$$

for $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^\intercal \in \mathbb{R}^p$. Note that we assume that $w_j \in \sum_{k=1}^s \mathbb{F}_k$ and hence $\widetilde{w}_j \in \mathbb{I} + \sum_{k=1}^s \mathbb{F}_k$. Since $\widehat{m}_{\boldsymbol{\rho}} \in \mathbb{I} + \sum_{k=1}^s \mathbb{F}_k$, there exists a subgradient $\boldsymbol{c} = (c_1, \ldots, c_p)^\intercal$ of $J(\widehat{m}_{\boldsymbol{\rho}})$ with respect to $\boldsymbol{\rho}$ at $\boldsymbol{\rho} = \mathbf{0}$ such that

$$\frac{\partial}{\partial \boldsymbol{\rho}} \left[\frac{1}{n} \sum_{i=1}^n \{y_i - \widehat{m}_{\boldsymbol{\rho}}(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2 \right] \Bigg|_{\boldsymbol{\rho}=\mathbf{0}} + \tau_n^2 \boldsymbol{c} = \mathbf{0}. \tag{S.9}$$

We first analyze the order of the subgradient $\boldsymbol{c}$. Note that $J(\widehat{m}_{\boldsymbol{\rho}}) = \sum_{k=1}^s \|\widehat{g}_k - \sum_{j=1}^p \rho_j w_{jk}\|$ where $w_{jk} = \mathcal{P}_k w_j$. Now we study two cases, $\|\widehat{g}_k\| > 0$ and $\|\widehat{g}_k\| = 0$, separately.

Suppose $\|\widehat{g}_k\| > 0$. Then $\|\widehat{g}_k - \sum_{j=1}^p \rho_j w_{jk}\|$ is differentiable at $\boldsymbol{\rho} = \mathbf{0}$ and its partial derivative with respect to ρ_l at $\boldsymbol{\rho} = \mathbf{0}$ is

$$-\frac{\int_0^1 \widehat{g}_k(t)dt \int_0^1 w_{lk}(t)dt + \int_0^1 \widehat{g}_k'(t)dt \int_0^1 w_{lk}'(t)dt + \int_0^1 \widehat{g}_k''(t) w_{lk}''(t)dt}{\|\widehat{g}_k\|},$$

for $l = 1, \ldots, p$. The numerator is less than or equal to $\|\widehat{g}_k\| \|w_{lk}\|$. Hence the absolute value of this partial derivative is smaller than or equal to $\|w_{lk}\| < \infty$ by the assumption that $J(w_l) < \infty$.

Suppose $\|\widehat{g}_k\| = 0$, which implies that $\widehat{g}_k = 0$. Then

$$\begin{aligned} \left\| \widehat{g}_k - \sum_{j=1}^p \rho_j w_{jk} \right\|^2 &= \left(\int_0^1 \sum_{j=1}^p \rho_j w_{jk}(t)dt \right)^2 + \left(\int_0^1 \sum_{j=1}^p \rho_j w_{jk}'(t)dt \right)^2 + \int_0^1 \left(\sum_{j=1}^p \rho_j w_{jk}''(t) \right)^2 dt \\ &= \boldsymbol{\rho}^\intercal \boldsymbol{N}_k \boldsymbol{\rho}, \end{aligned}$$

where $\boldsymbol{N}_k$ is a $p \times p$ matrix with (i,j)-entry being $\int w_{ik} \int w_{jk} + \int w_{ik}' \int w_{jk}' + \int w_{ik}'' w_{jk}''$. Note that $\boldsymbol{N}_k$ is positive semi-definite. Using subgradient chain rule and the subgradient

formulation of Euclidean norm, the subgradient of $\sqrt{\boldsymbol{\rho}^\intercal \boldsymbol{N}_k \boldsymbol{\rho}}$ with respect to $\boldsymbol{\rho}$ is

$$\begin{cases} \frac{\boldsymbol{N}_k \boldsymbol{\rho}}{\|\boldsymbol{N}_k^{1/2}\boldsymbol{\rho}\|_E}, & \text{if } \boldsymbol{N}_k^{1/2}\boldsymbol{\rho} \neq \boldsymbol{0}; \\ \in \{\boldsymbol{N}_k^{1/2}\boldsymbol{a} : \|\boldsymbol{a}\|_E \leq 1\}, & \text{otherwise.} \end{cases}$$

Recall that we are interested in the case of $\boldsymbol{\rho} = \boldsymbol{0}$. For any $\boldsymbol{a} = (a_1, \ldots, a_p)^\intercal$ such that $\|\boldsymbol{a}\|_E \leq 1$, $\|\boldsymbol{N}_k^{1/2}\boldsymbol{a}\|_\infty \leq \|\boldsymbol{N}_k^{1/2}\boldsymbol{a}\|_E = \|\sum_{j=1}^p a_j w_{jk}\| \leq \sum_{j=1}^p |a_j| \|w_{jk}\| \leq \sum_{j=1}^p \|w_{jk}\| < \infty$, where $\|\cdot\|_\infty$ is the max norm of a vector. Combining results from both cases, $\|\widehat{g}_k\| > 0$ and $\|\widehat{g}_k\| = 0$, we conclude that all entries of $\boldsymbol{c}$ are $\mathcal{O}(1)$.

Now, we go back to (S.9) and study the first term on the right hand side. For $l = 1, \ldots, p$,

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial \rho_l}\left[\frac{1}{n}\sum_{i=1}^n \{y_i - \widehat{m}_{\boldsymbol{\rho}}(\boldsymbol{z}_i, \widehat{\boldsymbol{\zeta}}_i)\}^2\right]\Bigg|_{\boldsymbol{\rho}=\boldsymbol{0}} = -\frac{1}{n}\sum_{i=1}^n \{y_i - \widehat{m}(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}\widetilde{w}_l(\boldsymbol{z}_i, \widehat{\boldsymbol{\zeta}}_i) \\
&= -\frac{1}{n}\sum_{i=1}^n \left[\{y_i - m(\boldsymbol{u}_i, \boldsymbol{\zeta}_i)\} + \{m(\boldsymbol{u}_i, \boldsymbol{\zeta}_i) - m(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\} + \{m(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i) - \widehat{m}(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i)\}\right]\widetilde{w}_l(\boldsymbol{z}_i, \widehat{\boldsymbol{\zeta}}_i), \\
&= -(\varepsilon, \widetilde{w}_l)_n + ((\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal \boldsymbol{w}, \widetilde{w}_l)_n + ((\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal \widetilde{\boldsymbol{w}}, \widetilde{w}_l)_n + (\widehat{g} - g_0, \widetilde{w}_l)_{n,*} \\
&\qquad + \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^s \widetilde{w}_l(\boldsymbol{z}_i, \boldsymbol{\zeta}_i) f'_{0k}(\zeta_{ik})(\widehat{\zeta}_{ik} - \zeta_{ik}) + \mathcal{O}_p(n^{-1}) \\
&= -\mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V} + \mathcal{O}_p(n^{-1}).
\end{aligned}$$

By the asymptotic expansions (S.1) and (S.3),

$$\begin{aligned}
\mathrm{V} &= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^s \widetilde{w}_l(\boldsymbol{z}_i, \boldsymbol{\zeta}_i) f'_{0k}(\zeta_{ik}) \Phi'(\zeta_{ik}) \Bigg\{ n^{-1/2} \sum_{j \neq k} \frac{\zeta_{ij} \lambda_j^{1/2}}{(\lambda_k - \lambda_j)\lambda_k^{1/2}} \langle \Delta \boldsymbol{\psi}_k, \boldsymbol{\psi}_j \rangle \\
&\qquad\qquad - \frac{1}{2} n^{-1/2} \zeta_{ik} \lambda_k^{-1} \langle \Delta \boldsymbol{\psi}_k, \boldsymbol{\psi}_k \rangle \Bigg\} \\
&= n^{-1/2} \sum_{k=1}^s \langle \Delta \boldsymbol{\psi}_k, \boldsymbol{\varpi}_{k,l} \rangle \times \{1 + \mathcal{O}_p(n^{-1/2})\},
\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\varpi}_{k,l} &= \sum_{j\neq k}\mathbb{E}\{\widetilde{w}_l(\boldsymbol{z}_1,\boldsymbol{\zeta}_1)f'_{0k}(\zeta_{1k})\Phi'(\zeta_{1k})\zeta_{1j}\}\lambda_j^{1/2}\lambda_k^{-1/2}(\lambda_k-\lambda_j)^{-1}\boldsymbol{\psi}_j \\ &\quad -\frac{1}{2}\mathbb{E}\{\widetilde{w}_l(\boldsymbol{z}_1,\boldsymbol{\zeta}_1)f'_{0k}(\zeta_{1k}\Phi'(\zeta_{1k})\zeta_{1k}\}\lambda_k^{-1}\boldsymbol{\psi}_k. \qquad \text{(S.10)}\end{aligned}$$

Since Δ converge weakly to a Gaussian random field, it is easy to see that $\mathrm{V} = \mathcal{O}_p(n^{-1/2})$ and is asymptotically normal.

By (10), $\mathbb{E}\{w_j(\boldsymbol{\zeta})\widetilde{w}_l(\boldsymbol{Z},\boldsymbol{\zeta})\} = 0$,

$$\mathrm{II} = \sum_{j=1}^{p}(\widehat{\gamma}_j-\gamma_{0j})(w_j,\widetilde{w}_l)_n = \sum_{j=1}^{p}\mathcal{O}_p(n^{-1/2})(\widehat{\gamma}_j-\gamma_{0j}).$$

Similarly, by law of large numbers,

$$\mathrm{III} = \sum_{j=1}^{p}(\widehat{\gamma}_j-\gamma_{0j})(\widetilde{w}_j,\widetilde{w}_l)_n = \sum_{j=1}^{p}(M_{lj}+o_p(1))(\widehat{\gamma}_j-\gamma_{0j}).$$

Similarly as before, we can show that the event $\liminf_n\{|\widehat{g}-g_0|_\infty/(1+J(\widehat{g})+J(g_0)) \le C_1+1\}$ is of probability 1.

It is easy to see that

$$\begin{aligned}\mathrm{IV} &= (\widehat{g}-g_0,\widetilde{w}_l)_n + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\{\widehat{f}'_k(\zeta_{ik}-f'_{k0}(\zeta_{ik})\}\widetilde{w}_l(\zeta_{ik})(\widehat{\zeta}_{ik}-\zeta_{ik})+o_p(n^{-1/2}) \\ &= (\widehat{g}-g_0,\widetilde{w}_l)_n + o_p(n^{-1/2}) \qquad (\text{ by Lemma 7}).\end{aligned}$$

Next, we study the behavior of $\sqrt{n}(g-g_0,\widetilde{w}_l)_n$ as a function of $\|(g-g_0)\widetilde{w}_l\|_2$. We are going to apply Theorem 2.4 of Mammen and van de Geer (1997). To prepare this, we first derive

some entropy results. Let

$$\mathcal{K} = \left\{ (g - g_0)\widetilde{w}_l : J(g - g_0) \leq 1, g \in \sum_{k=1}^{s} \mathbb{F}_k \right\}.$$

Since $\widetilde{w}_l \in \mathbb{M}$, write $K_6 = |\widetilde{w}_l|_\infty < \infty$. Therefore

$$H_\infty(\delta, \mathcal{K}) \leq H_\infty\left(\frac{\delta}{K_6}, \widetilde{\mathcal{K}}\right) \quad \text{with } \widetilde{\mathcal{K}} = \left\{ g - g_0 : J(g - g_0) \leq 1, g \in \sum_{k=1}^{s} \mathbb{F}_k \right\}.$$

For any $m \in \mathcal{M}$, we can write it in two ways:

$$m(\boldsymbol{u}, \boldsymbol{\zeta}) - m_0(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{z}^\intercal(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + g(\boldsymbol{\zeta}) - g_0(\boldsymbol{\zeta}) = \boldsymbol{u}^\intercal(\boldsymbol{\nu} - \boldsymbol{\nu}_0) + h(\boldsymbol{u}, \boldsymbol{\zeta}) - h_0(\boldsymbol{u}, \boldsymbol{\zeta}). \quad \text{(S.11)}$$

Note that $J(m - m_0) = J(g - g_0) = J(h - h_0)$. If $J(m - m_0) \leq 1$, we can represent $g - g_0$ and $h - h_0$ uniquely as follows:

$$\begin{aligned} g(\boldsymbol{\zeta}) - g_0(\boldsymbol{\zeta}) &= \mu + \sum_{k=1}^{s} \widetilde{r}_k(\zeta_k), \quad \text{(S.12)} \\ h(\boldsymbol{u}, \boldsymbol{\zeta}) - h_0(\boldsymbol{u}, \boldsymbol{\zeta}) &= \sum_{k=1}^{s} \widetilde{h}_k(\boldsymbol{u}, \zeta_k) \quad \text{with} \quad \widetilde{h}_k(\boldsymbol{u}, \zeta_k) = \boldsymbol{u}^\intercal \widetilde{\boldsymbol{\omega}}_k + \widetilde{r}_k(\zeta_k) \in \mathbb{H}_k, \end{aligned}$$

where $\widetilde{r}_k \in \bar{\mathbb{F}}_k$ such that $\sum_{i=1}^{n} \widetilde{r}_k(\zeta_{ik}) = 0$ and $J(\widetilde{r}_k) \leq 1$. Plugging them into (S.11), we show that μ is the first element of $\boldsymbol{\nu} - \boldsymbol{\nu}_0 + \sum_{k=1}^{s} \widetilde{\boldsymbol{\omega}}_k$. Write $\widehat{\mu}$ as μ in (S.12) for $\widehat{g} - g_0$. Recall that the event $\liminf\{\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0\|_E \leq K_1\}$ is of probability 1. Moreover, from the proof of Lemma 3, we have the event $\liminf\{\max_{k=1,\ldots,s} \|\boldsymbol{\omega}_k\|_E \leq L\}$ is of probability 1. Thus $\liminf\{|\widehat{\mu}| \leq K_7\}$ for some constant K_7. Thus we focus on the set

$$\bar{\mathcal{K}} = \left\{ g(\boldsymbol{\zeta}) - g_0(\boldsymbol{\zeta}) = \mu + \sum_{k=1}^{s} \widetilde{r}_k(\zeta_k) : |\mu| \leq K_7, J(g - g_0) \leq 1, g \in \sum_{k=1}^{s} \mathbb{F}_k \right\},$$

where, with probability 1, $\widehat{g} - g_0$ will eventually falls into. We use similar trick in (S.16)

to derive the entropy result for $\widetilde{\mathcal{K}}$ by the decomposition (S.12). It suffices to obtain bound for $H_\infty(\cdot, \{g(\boldsymbol{\zeta}) = \mu : |\mu| \le K_7\})$ and $H_\infty(\cdot, \{\sum_{k=1}^s \widetilde{r}_k : \boldsymbol{u}^\intercal \widetilde{\boldsymbol{\omega}}_k + \widetilde{r}_k(\zeta_k) \in \mathbb{H}_k, \sum_{i=1}^n \widetilde{r}_k(\zeta_{ik}) = 0, J(\widetilde{r}_k) \le 1\})$. The bound for the first entropy is from Lemma 2.5 of van de Geer (2000), while that for the second entropy is derived similarly in the proof of Lemma 3. For simplicity, we skip those details. In the end, we get the event $\liminf_n\{\sup_{\delta>0} \delta^{1/2} H_\infty(\delta, \bar{\mathcal{K}}) \le K_8\}$ is of probability 1. Combining with the above results, we obtain an entropy bound for the set

$$\widehat{\mathcal{K}} = \left\{ \frac{(g - g_0)\widetilde{w}_l}{1 + J(g) + J(g_0)} : g - g_0 = \mu + \sum_{k=1}^s \widetilde{r}_k, |\mu| \le K_7, g \in \sum_{k=1}^s \mathbb{F}_k \right\}.$$

That is, the event $\liminf_n\{\sup_{\delta>0} \delta^{1/2} H_\infty(\delta, \widehat{\mathcal{K}}) \le K_9\}$ is of probability 1.

Note that $\mathbb{E}(g - g_0, \widetilde{w}_l)_n = 0$ since $\mathbb{E}(g(\boldsymbol{\zeta})\widetilde{w}_l(\boldsymbol{z}, \boldsymbol{\zeta})) = 0$ for any $g \in \sum_{k=1}^s \mathbb{F}_k$. Applying Theorem 2.4 of Mammen and van de Geer (1997) to $\widehat{\mathcal{K}}$, we have

$$\text{IV} = o_p(n^{-1/2}),$$

since $J(\widehat{g}) = \mathcal{O}_p(1)$ (Theorem 1) and $\|\widehat{g} - g_0\|_2 = o_p(1)$.

Also, it is simple to show that $\sum_{i=1}^n (y_i - \widehat{m}(\boldsymbol{u}_i, \widehat{\boldsymbol{\zeta}}_i))\, \mathcal{O}_p(n^{-1/2})/n = \mathcal{O}_p(n^{-1/2})$ since $\|\widehat{m} - m_0\|_n = \mathcal{O}_p(\tau_n) = \mathcal{O}_p(n^{-1/4})$. Collecting all the above results, we have, for $l = 1, \ldots, p$,

$$-(\varepsilon, \widetilde{w}_l)_n + \sum_{j=1}^p (M_{lj} + o_p(1))(\widehat{\gamma}_j - \gamma_{0j}) + n^{-1/2} \sum_{k=1}^s \langle \Delta \boldsymbol{\psi}_k, \boldsymbol{\varpi}_{k,l} \rangle + 2\tau_n^2 c_l + o_p(n^{-1/2}) = 0,$$

with $c_l = \mathcal{O}(1)$. Since $\boldsymbol{M}$ is non-singular, we have

$$n^{1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = \boldsymbol{M}^{-1}(\boldsymbol{q}_1 + \boldsymbol{q}_2) + o_p(1), \tag{S.13}$$

where $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jp})^\intercal$, $j = 1, 2$, with $q_{1l} = n^{1/2}(\varepsilon, \widetilde{w}_l)_n$, $q_{2l} = -\sum_{k=1}^{s} \langle \Delta\boldsymbol{\psi}_k, \boldsymbol{\varpi}_{k,l} \rangle$. Put

$$\boldsymbol{V}_1 = \mathrm{cov}(\boldsymbol{q}_1) \quad \text{and } \boldsymbol{V}_2 = \mathrm{cov}(\boldsymbol{q}_2), \tag{S.14}$$

by the central limit theorem $\boldsymbol{q}_1 \to Normal(\boldsymbol{0}, \boldsymbol{V}_1)$ in distribution, and since Δ converge weakly to a Gaussian random field (Dauxois et al., 1982), $\boldsymbol{q}_2 \to Normal(\boldsymbol{0}, \boldsymbol{V}_2)$ in distribution. It is easy to see that $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are asymptotically independent because ε and Δ are independent. The results of the theorem follows from (S.13).

B.4 Proofs of Lemmas

Proof of Lemma 2. For $f_k \in \mathbb{F}_k$ which is a RKHS with the reproducing kernel $R_k(\cdot, \cdot)$

$$\left| \frac{\partial f_k(\zeta_k)}{\partial \zeta_k} \right| = \left| \left\langle f_k(\cdot), \frac{\partial R_k(\zeta_k, \cdot)}{\partial \zeta_k} \right\rangle \right| \leq \|f_k\| \left\| \frac{\partial R_k(\zeta_k, \cdot)}{\partial \zeta_k} \right\|.$$

The reproducing kernel of 2nd order Sobolev Hilbert spaces are $R_k(s,t) = h_1(s)h_1(t) + h_2(s)h_2(t) - h_4(|s-t|)$ where $h_1(t) = t - 1/2$, $h_2(t) = \{h_1^2(t) - 1/12\}/2$ and $h_4(t) = \{h_1^4(t) - h_1^2(t)/2 + 7/240\}/24$. Note that

$$\frac{\partial^2 R_k(s,t)}{\partial s \partial t} = \frac{13}{12} + \left(s - \frac{1}{2}\right)\left(t - \frac{1}{2}\right) - \frac{1}{2}|s-t| + \frac{1}{2}(s-t)^2. \tag{S.15}$$

Now, for any $k \leq s$,

$$\sup_{\zeta \in [0,1]} \left\| \frac{\partial R_k(\zeta, \cdot)}{\partial \zeta} \right\|^2 = \sup_{\zeta \in [0,1]} \left\langle \frac{\partial R_k(\zeta, \cdot)}{\partial \zeta}, \frac{\partial R_k(\zeta, \cdot)}{\partial \zeta} \right\rangle = \sup_{\zeta \in [0,1]} \left. \frac{\partial^2 R_k(s,t)}{\partial s \partial t} \right|_{s=t=\zeta} \leq \frac{4}{3}.$$

Proof of Lemma 3. We will study the entropy result for $\widetilde{\mathbb{H}}_k := \{h_k \in \mathbb{H}_k : J(h_k) \leq 1\}$ first. For $h_k \in \widetilde{\mathbb{H}}_k$, we can represent it uniquely as $h_k(\boldsymbol{u}, \zeta) = \boldsymbol{u}^\intercal \boldsymbol{\omega}_k + r_k(\zeta)$, where $\sum_{i=1}^{n} r_k(\zeta_{ik}) = 0$ and $r_k \in \bar{\mathbb{F}}_k$ with $J(r_k) \leq 1$. Note that if $\mathcal{S}_1$ and $\mathcal{S}_2$ are two sets of functions, we can bound

the uniform entropy of $\mathcal{S}_1 + \mathcal{S}_2$:

$$H_\infty(\delta, \mathcal{S}_1 + \mathcal{S}_2) \leq H_\infty(\delta/2, \mathcal{S}_1) + H_\infty(\delta/2, \mathcal{S}_2). \tag{S.16}$$

Take

$$\mathcal{S}_{k,1} = \left\{ r_k : h_k(\boldsymbol{u}, \zeta) = \boldsymbol{u}^\intercal \boldsymbol{\omega}_k + r_k(\zeta), \sum_{i=1}^n r_k(\zeta_{ik}) = 0, h_k \in \widetilde{\mathbb{H}}_k \right\}$$

and

$$\mathcal{S}_{k,2} = \left\{ g(\boldsymbol{u}) = \boldsymbol{u}^\intercal \boldsymbol{\omega} : h_k(\boldsymbol{u}, \zeta) = \boldsymbol{u}^\intercal \boldsymbol{\omega}_k + r_k(\zeta), \sum_{i=1}^n r_k(\zeta_{ik}) = 0, h_k \in \widetilde{\mathbb{H}}_k \right\}.$$

Note that $\widetilde{\mathbb{H}}_k \subseteq \mathcal{S}_{k,1} + \mathcal{S}_{k,2}$ and thus $H_\infty(\delta, \widetilde{\mathbb{H}}_k) \leq H_\infty(\delta/2, \mathcal{S}_{k,1}) + H_\infty(\delta/2, \mathcal{S}_{k,2})$. By the proof of Lemma A.1 in Lin and Zhang (2006), $|r_k|_\infty \leq 1$ and there exists a constant A such that $H_\infty(\delta, \mathcal{S}_{k,1}) \leq A\delta^{-1/2}$ for all $\delta > 0$.

Now, it remains to obtain results about $H_\infty(\delta, S_{k,2})$. The constraints of $\mathbb{H}_k$ can be written as

$$\frac{1}{n}\boldsymbol{U}^\intercal \boldsymbol{U}\boldsymbol{\omega}_k = -\frac{1}{n}\boldsymbol{U}^\intercal (r_k(\widehat{\zeta}_{k1}), \ldots, r_k(\widehat{\zeta}_{kn}))^\intercal$$

where $\boldsymbol{U} = [u_{ij}]_{i=1,\ldots,n, j=1,\ldots,p+1}$. Note that $\boldsymbol{U}^\intercal \boldsymbol{U}/n \to \boldsymbol{\Sigma}$ almost surely (element-wisely) and $\boldsymbol{\Sigma}$ is non-singular. Write the smallest eigenvalue of $\boldsymbol{\Sigma}$ as σ_1. Let $\mathcal{E}_n$ be the event that $\max_{k=1,\ldots,s} \|\boldsymbol{\omega}_k\|_E \leq L = 2\sqrt{p+1}/\sigma_1$. Combining with $|r_k|_\infty \leq 1$ and $|u_{ij}| \leq 1$, we have

$$\left\| \frac{1}{n}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^\intercal (r_k(\zeta_{k1}), \ldots, r_k(\zeta_{kn}))^\intercal \right\|_E \leq \frac{1}{\sigma_1} \left\| \frac{1}{n}\boldsymbol{U}^\intercal (r_k(\zeta_{k1}), \ldots, r_k(\zeta_{kn}))^\intercal \right\|_E \leq \frac{\sqrt{p+2}}{\sigma_1}$$

for all k. Therefore $P(\liminf_{n\to\infty} \mathcal{E}_n) = 1$. We note that this result hinges on the convergence of $\boldsymbol{U}^\intercal \boldsymbol{U}/n$, which does not depend on s, and thus still holds even s grows with n. Next, for any $\boldsymbol{u} \in [0,1]^{p+1}$ and $\boldsymbol{\omega}, \boldsymbol{\omega}^* \in \mathbb{R}^{p+1}$, $|\boldsymbol{u}^\intercal \boldsymbol{\omega} - \boldsymbol{u}^\intercal \boldsymbol{\omega}^*| \leq \sqrt{p+1}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|_E$. Therefore, on $\mathcal{E}_n$, $H_\infty(\delta, \mathcal{S}_{k,2}) \leq H(\delta/\sqrt{p+1}, \{\boldsymbol{\omega} : \|\boldsymbol{\omega}\|_E \leq L\}, \|\cdot\|_E)$. From Lemma 2.5 of van de Geer (2000), there exists a constant B such that $H(\delta/\sqrt{p+1}, \{\boldsymbol{\omega} : \|\boldsymbol{\omega}\|_E \leq L\}, \|\cdot\|_E) \leq (p+1)\log(1 +$

$4L\sqrt{p+1}/\delta) \leq B\delta^{-1/2}$. Thus $H_\infty(\delta, \{h_k \in \mathbb{H}_k : J(h_k) \leq 1\}) \leq (A+B)\sqrt{2}\delta^{-1/2} = C_1\delta^{-1/2}$ where $C_1 = sqrt2(A+B)$. As a result, on $\mathcal{E}_n$, $H_\infty(\delta, \{\delta, \{h \in \mathbb{H} : J(h) \leq 1\}\}) \leq C_1 s^{3/2}\delta^{-1/2}$ since $J(h) \leq 1$ implies $J(h_k) \leq 1$ for all k. Moreover, on $\mathcal{E}_n$, $\sup_{\{h\in\mathbb{H}:J(h)\leq 1\}} |h|_\infty < sC_1'$ due to $|h_k| \leq C_1' := \sqrt{p+1}L + 1$ for all k.

Proof of Lemma 4. Suppose

$$H_\infty(\delta, \{h \in \mathbb{H} : J(h) \leq 1\}) \leq C_1 s^{3/2}\delta^{-1/2}, \tag{S.17}$$

for all $\delta > 0$, $n \geq 1$ and some constant $C_1 > 0$ not depending on n and s. Then,

$$H\left(\delta, \left\{\frac{h - h_0}{J(h) + J(h_0)} : h \in \mathbb{H}\right\}, \|\cdot\|_{n,*}\right)$$

has the same entropy bound (S.17). The rest follows from the proof of Lemma 8.4 in van de Geer (2000) and Lemma 3 that (S.17) holds eventually with probability 1.

Proof of Lemma 5. By Theorem 1, we have $\|\widehat{m} - m_0\|_n = \mathcal{O}_p(n^{-1/4})$. We will show that $\|\widehat{m} - m_0\|_n$ and $\|\widehat{m} - m_0\|_2$ have the same order.

Recall that, in the proof of Proposition 2, we write $\widehat{m}(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\widehat{\boldsymbol{\nu}} + \widehat{h}(\boldsymbol{u}, \boldsymbol{\zeta})$ and $m(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\boldsymbol{\nu} + h(\boldsymbol{u}, \boldsymbol{\zeta})$. In its proof, using strong laws of large number, we show that $\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0\|_E$ converges to zero almost surely and hence the event $\liminf_n\{\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}_0\|_E \leq K_1\}$ is of probability 1 for some constant K_1. Consider the set

$$\mathcal{J} = \{m - m_0 : \|\boldsymbol{\nu} - \boldsymbol{\nu}_0\|_E \leq K_1, J(h - h_0) \leq 1, m(\boldsymbol{u}, \boldsymbol{\zeta}) = \boldsymbol{u}^\intercal\boldsymbol{\nu} + h(\boldsymbol{u}, \boldsymbol{\zeta}) \in \mathbb{M}\}.$$

We can use the similar trick in (S.16) to derive the entropy result for $\mathcal{J}$ by decomposing a function in $\mathcal{J}$: $m - m_0 = \boldsymbol{u}^\intercal(\boldsymbol{\nu} - \boldsymbol{\nu}_0) + h - h_0$. Next, it suffices to derive uniform entropies

$$H_\infty(\cdot, \{\boldsymbol{u}^\intercal(\boldsymbol{\nu} - \boldsymbol{\nu}_0) : \|\boldsymbol{\nu} - \boldsymbol{\nu}_0\|_E \leq K_1, \boldsymbol{\nu} \in \mathbb{R}^p\}) \quad \text{and} \quad H_\infty(\cdot, \{h - h_0 : J(h - h_0) \leq 1, h \in \mathbb{H}\}).$$

The first one can be handled by Lemma 2.5 of van de Geer (2000) similary as in the proof of Lemma 3 while the second one can be handled by Lemma 3. For simplicity, we skip those details. In the end, we have $\liminf_n\{\sup_{\delta>0}\delta^{1/2}H_\infty(\delta,\mathcal{J})\le K_2\}$ is of probability 1 for some constant K_2. And this implies the entropy results for the set

$$\widetilde{\mathcal{J}}=\left\{\frac{m-m_0}{1+J(m)+J(m_0)}:\|\boldsymbol{\nu}-\boldsymbol{\nu}_0\|_E\le K_1, m(\boldsymbol{u},\boldsymbol{\zeta})=\boldsymbol{u}^\intercal\boldsymbol{\nu}+h(\boldsymbol{u},\boldsymbol{\zeta})\in\mathbb{M}\right\}.$$

Namely, $\liminf_n\{\sup_{\delta>0}\delta^{1/2}H_\infty(\delta,\widetilde{\mathcal{J}})\le K_3\}$ is of probability 1 for some constant K_3.

Using Lemma 3, we can show that the event $\liminf_n\{|\widehat{h}-h_0|_\infty/(1+J(\widehat{h})+J(h_0))\le K_4\}$ is of probability 1 for some constant K_4. (Note that s is assumed to be fixed and thus is assimilated into the constant.) Combining with $\mathbb{P}(\liminf_n\{\|\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}_0\|_E\le K_1\})=1$, we can simply focus on the set

$$\bar{\mathcal{J}}=\left\{\frac{m-m_0}{1+J(m)+J(m_0)}:\|\boldsymbol{\nu}-\boldsymbol{\nu}_0\|_E\le K_1,\frac{|\widehat{h}-h_0|_\infty}{1+J(\widehat{h})+J(h_0)}\le K_4, m\in\mathbb{M}\right\}, \quad \text{(S.18)}$$

where, with probability 1, $(\widehat{m}-m_0)/(1+J(\widehat{m})+J(m_0))$ will eventually fall into. Clearly, we also have that $\liminf_n\{\sup_{\delta>0}\delta^{1/2}H_\infty(\delta,\bar{\mathcal{J}})\le K_3\}$ is of probability 1. It is also easy to show that $\bar{\mathcal{J}}$ is uniformly bounded.

From Theorem 1, we have $\|\widehat{m}-m\|_n=\mathcal{O}_p(n^{-1/4})$. Hence, by applying Lemma 5.16 of van de Geer (2000) on $\bar{\mathcal{J}}$, with $\delta_n=K_5n^{-2/5}$ for some constant K_5, we can show that $\|\widehat{m}-m_0\|_n$ and $\|\widehat{m}-m_0\|_2$ have the same order and thus $\|\widehat{m}-m_0\|_2=o_p(1)$.

Proof of Lemma 6.

Consider $\widehat{\mathbb{F}}'_k=\{f^{(1)}/\|f\|:f\in\bar{\mathbb{F}}_k\}$. By Lemma 2, we have the uniform boundedness of $\widehat{\mathbb{F}}'_k$: $\sup_{f\in\widehat{\mathbb{F}}'_k}\sup_{t\in[0,1]}|f(t)|\le C_2$. Using Lemma 2.4 of van de Geer (2000), it is easy to show that there exists a constant C_3 such that $\sup_{\delta>0}\delta H_\infty(\delta,\widehat{\mathbb{F}}'_k)\le C_3$. Owing to the uniform boundedness of $\widehat{\mathbb{F}}'_k$, $\sup_{\delta>0}\delta H_\infty(\delta,\{f^2:f\in\widehat{\mathbb{F}}'_k\})\le 2C_2C_3$. The desired result then follows

from Lemma 3.6 of van de Geer (2000).

Proof of Lemma 7. Put $q_k = \widehat{f}_k - f_{0k}$, then $\widehat{f} - f_0 = \sum_{j=1}^s q_j$. Since $q_j \in \bar{\mathbb{F}}_j$, $\int_0^1 q_j(t)dt = 0$, and therefore $\|\widehat{f} - f_0\|^2_{L^2[0,1]^s} = \sum_{j=1}^s \|q_j\|^2_{L^2[0,1]}$. By (S.8), $\|\widehat{g} - g_0\|_2 = o_p(1)$. By the assumption that $\boldsymbol{\zeta}$ has non-degenerate, bounded joint density on $[0,1]^s$, $\|\cdot\|_2$ and $\|\cdot\|_{L^2[0,1]^s}$ are equivalent norms, and therefore $\|q_j\|_{L^2[0,1]} = o_p(1)$ for $j = 1, \ldots, s$. By Gagliardo-Nirenberg interpolation inequality (Nirenberg (1959) and Brezis (2010, pp. 313-314)), there exists a constant C_4 such that

$$\|q_k^{(1)}\|_{L^2[0,1]} \le C_4 \|q_k\|^{1/2} \|q_k\|^{1/2}_{L^2[0,1]}.$$

By Theorem 1, $J(\widehat{m}) = \mathcal{O}_p(1)$ and therefore $\|q_k\| = \mathcal{O}_p(1)$. Therefore $\|q_k^{(1)}\|_{L^2[0,1]} = o_p(1)$. Again, because $\|\cdot\|_{L^2[0,1]}$ and $\|\cdot\|_2$ are equivalent norms, $\|q_k^{(1)}\|_2 = o_p(1)$. Finally, by Lemma 6 and $\|q_k\| = \mathcal{O}_p(1)$, we have $\|q_k^{(1)}\|_n^2 = \|q_k^{(1)}\|_{2,k}^2 + \|q_k^{(1)}\|_n^2 - \|q_k^{(1)}\|_{2,k}^2 = \|q_k^{(1)}\|_{2,k}^2 + \|q_k\|^2 o_p(1) = o_p(1)$.

C Additional results for Section 5

Following the suggestion of a referee, we also provide results when s is set to recover 90% of the total variation in $\{\boldsymbol{x}_i\}$, instead of 99.9%. The results are presented in Tables S.1-S.4, which should be compared with Tables 1-4 in the main text. When such a smaller percentage is used, the 4th component, which is related to Y, is near the cut-off point and often not included in the model. As a result, f_4 is often falsely excluded from the model (see Table S.2), and there is a much lower chance for COSSO to select the correct model. We also see much bigger prediction errors in Table S.4 than those in Table 4. Our conclusion is it is best to include as many components as possible and let the model selection mechanism of COSSO determine the size of the model.

D Additional Results for Section 6

Since the two functional predictors in our real data are strongly correlated, we also compare the prediction performance for models using only one functional predictor. Recall that $X_1(t)$ and $X_2(t)$ are the daily maximum and daily minimum temperature trajectories respectively. We denote by $\bar{X}(t) = \{X_1(t) + X_2(t)\}/2$ the mean trajectory. In addition to the models presented in Section 6, we also compare the yield prediction performance of the following 12 models, which use only one of $X_1(t)$, $X_2(t)$ and $\bar{X}(t)$ as the functional predictor. In the prediction experiment described in Section 6.1, the prediction errors of these 12 models are presented in Table S.5. As we can see, the models using only one functional predictor or the average yield higher prediction errors than PLFAM(joint) which jointly model both functional predictors.

1. PLFAM(max): PLFAM based on univariate FPCA scores from X_1;
2. FAM(max): FAM based on univariate FPCA scores from X_1;
3. FLM-Cov(max): FLM based on univariate FPCA scores from X_1, with covariate effects;
4. FLM(max): FLM based on univariate FPCA scores from X_1 (without $\boldsymbol{Z}$);
5. PLFAM(min): PLFAM based on univariate FPCA scores from X_2;
6. FAM(min): FAM based on univariate FPCA scores from X_2;
7. FLM-Cov(min): FLM based on univariate FPCA scores from X_2, with covariate effects;
8. FLM(min): FLM based on univariate FPCA scores from X_2 (without $\boldsymbol{Z}$);
9. PLFAM(mean): PLFAM based on univariate FPCA scores from $\bar{X}$;
10. FAM(mean): FAM based on univariate FPCA scores from $\bar{X}$;

11. FLM-Cov(mean): FLM based on univariate FPCA scores from $\bar{X}$, with covariate effects;

12. FLM(mean): FLM based on univariate FPCA scores from $\bar{X}$ (without $\boldsymbol{Z}$).

We also made the assumption that crop yields in different counties and years are conditional independent given the local meteorology information. To check for possible spatial dependency, we calculate the spatial variograms for each year based on the residuals from the fitted yield prediction model; to check for possible temporal dependency, we also calculate the autocorrelation function (ACF) for each county. Because of limited space, we show the spatial variograms for the first 4 years in Figure S.1 and ACF for the first 4 counties in Figure S.2. These plots are based on the residuals of the corn yield prediction model. Plots for other years and counties and those based on the soybean prediction model are similar. All variograms and ACF's are contained in the confidence band based on the assumption of no dependency, which supports the conditional independence assumption that we make.

E Standard Error Estimation by Bootstrap

To quantify the uncertainties in the estimated model, we estimate the standard errors of both $\widehat{\boldsymbol{\theta}}$ and $\widehat{f}(\boldsymbol{\zeta})$ using bootstrap. In addition to the uncertainties in the regression step, our bootstrap procedure also takes into account the variation in mFPCA. The bootstrap samples are obtained by resampling residuals from both the observations on the functional covariates and the response variables. The procedure is as follows.

1. (Resampling the functional covariates) Recall that the discrete noisy observations on $\boldsymbol{x}_i$ are

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

and the recovered functions from the discrete observations are $\tilde{x}_{ij}(t)$. Let $\widehat{e}_{ijk} = w_{ijk} - \tilde{x}_{ij}(t_{ijk})$ and resample with replacement e^*_{ijk} from $\{\widehat{e}_{ijk} : k = 1, \ldots, N_{ij}\}$ to obtain a bootstrap sample $w^*_{ijk} = \tilde{x}_{ij}(t_{ijk}) + e^*_{ijk}$. Repeat for all i, j, k, to obtain the bootstrap sample $\mathcal{W}^* = \{w^*_{ijk} : i = 1, \ldots, n,\ j = 1, \ldots, d,\ k = 1, \ldots, N_{ijk}\}$ for the functional data.

2. (Resampling the response) Denote $\widehat{y}_i$ as the fitted value of y_i from the original data and define the residuals $\widehat{\varepsilon}_i = \pi_i^{1/2}(y_i - \widehat{y}_i)$. Sample with replacement ε^*_i uniformly from $\{\widehat{\varepsilon}_i : i = 1, \ldots, n\}$ to obtain a bootstrap sample $y^*_i = \widehat{y}_i + \pi_i^{-1/2}\varepsilon^*_i$ of y_i. Denote the bootstrap sample as $\mathcal{Y}^* = \{y^*_i : i = 1, \ldots, n\}$.

3. Apply the mFPCA procedure on $\mathcal{W}^*$ to obtained mFPC scores $\boldsymbol{\zeta}^*$, and then fit the propose PLFAM to $\mathcal{Y}^*$ using $\boldsymbol{\zeta}^*$ and the original $\boldsymbol{Z}$. Denote the estimates from the bootstrap sample as $\widehat{\boldsymbol{\theta}}^*$ and $\widehat{f}^*(\boldsymbol{\zeta})$.

4. Repeat Steps 1- 3 a large number of times and use the sample standard deviations of $\widehat{\boldsymbol{\theta}}^*$ and $\widehat{f}^*(\boldsymbol{\zeta})$ as estimates of the standard errors for $\widehat{\boldsymbol{\theta}}$ and $\widehat{f}(\boldsymbol{\zeta})$.

Table S.1: Percentages of fitted model sizes.

Setting	Model	% for the following model sizes							
		1	2	3	4	5	6	7	8
{(i), (I)}	FAM	1	40	58.5	0.5	0	0	0	0
	PLFAM	1	40	58.5	0.5	0	0	0	0
{(ii), (I)}	FAM	2.5	95.5	2	0	0	0	0	0
	PLFAM	2.5	95.5	2	0	0	0	0	0
{(i), (II)}	FAM	5.5	50	42	2.5	0	0	0	0
	PLFAM	0	37.5	62.5	0	0	0	0	0
{(ii), (II)}	FAM	15	84	1	0	0	0	0	0
	PLFAM	2	97	1	0	0	0	0	0

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

Setting	Model	% for the following component functions								% correct	% super
		$\widehat{f}_1$	$\widehat{f}_2$	$\widehat{f}_3$	$\widehat{f}_4$	$\widehat{f}_5$	$\widehat{f}_6$	$\widehat{f}_7$	$\widehat{f}_8$	set	set
{(i), (I)}	FAM	100	99	2.5	57	0	0	0	0	56.5	57
	PLFAM	100	99	2.5	57	0	0	0	0	56.5	57
{(ii), (I)}	FAM	100	97.5	2	0	0	0	0	0	0	0
	PLFAM	100	97.5	2	0	0	0	0	0	0	0
{(i), (II)}	FAM	100	81	3	57.5	0	0	0	0	41.5	44
	PLFAM	100	100	2.5	60	0	0	0	0	60	60
{(ii), (II)}	FAM	100	85	1	0	0	0	0	0	0	0
	PLFAM	100	98	1	0	0	0	0	0	0	0

Table S.3: Averaged integrated squared errors.

Setting	Model	AISEs for the following component functions								
		$\widehat{f}_1$	$\widehat{f}_2$	$\widehat{f}_3$	$\widehat{f}_4$	$\widehat{f}_5$	$\widehat{f}_6$	$\widehat{f}_7$	$\widehat{f}_8$	$\widehat{f}$
{(i), (I)}	FAM	0.0257	0.0903	0.0020	0.4682	0.0000	0.0000	0.0000	0.0000	0.5861
	PLFAM	0.0258	0.0907	0.0018	0.4682	0.0000	0.0000	0.0000	0.0000	0.5865
{(ii), (I)}	FAM	0.0321	0.1364	0.0026	0.9508	0.0000	0.0000	0.0000	0.0000	1.1219
	PLFAM	0.0324	0.1352	0.0027	0.9508	0.0000	0.0000	0.0000	0.0000	1.1210
{(i), (II)}	FAM	0.0439	0.2211	0.0056	0.4902	0.0000	0.0000	0.0000	0.0000	0.7609
	PLFAM	0.0252	0.0855	0.0015	0.4348	0.0000	0.0000	0.0000	0.0000	0.5470
{(ii), (II)}	FAM	0.0423	0.2158	0.0014	0.9508	0.0000	0.0000	0.0000	0.0000	1.2102
	PLFAM	0.0278	0.1341	0.0009	0.9508	0.0000	0.0000	0.0000	0.0000	1.1136

Table S.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.

Setting	Model	Prediction error		Mean squared errors					
		separate	joint	separate			joint		
				$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$
{(i), (I)}	FAM	1.68 (0.11)	1.68 (0.40)	-	-	-	-	-	-
	PLFAM	1.69 (0.11)	1.70 (0.41)	0.0763	0.0975	0.1097	0.0756	0.1047	0.1095
{(ii), (I)}	FAM	1.68 (0.10)	2.13 (0.12)	-	-	-	-	-	-
	PLFAM	1.69 (0.10)	2.15 (0.13)	0.0667	0.1108	0.0858	0.0767	0.1388	0.1100
{(i), (II)}	FAM	3.94 (0.24)	3.94 (0.36)	-	-	-	-	-	-
	PLFAM	1.71 (0.11)	1.69 (0.39)	0.0688	0.1091	0.0973	0.0686	0.1181	0.0818
{(ii), (II)}	FAM	3.91 (0.25)	4.29 (0.27)	-	-	-	-	-	-
	PLFAM	1.71 (0.11)	2.13 (0.13)	0.0675	0.0897	0.1156	0.079	0.1332	0.1284

Table S.5: Average of 5-year overall prediction errors.

		corn	soybean
(a) functional additive models	PLFAM(joint)	298.43	35.64
	PLFAM(separate)	306.50	38.85
	PLFAM(max)	324.27	38.22
	PLFAM(min)	338.51	44.09
	PLFAM(mean)	330.17	40.93
	FAM(joint)	830.17	48.54
	FAM(separate)	839.00	51.06
	FAM(max)	898.12	51.92
	FAM(min)	997.27	65.48
	FAM(mean)	916.80	57.79
(b) functional linear models	FLM-Cov(joint)	303.81	35.29
	FLM-Cov(separate)	308.57	35.69
	FLM-Cov(max)	317.83	37.52
	FLM-Cov(min)	338.88	42.43
	FLM-Cov(mean)	310.02	37.27
	FLM(joint)	704.19	47.31
	FLM(separate)	767.42	50.42
	FLM(max)	779.56	51.49
	FLM(min)	842.12	61.42
	FLM(mean)	790.96	52.38

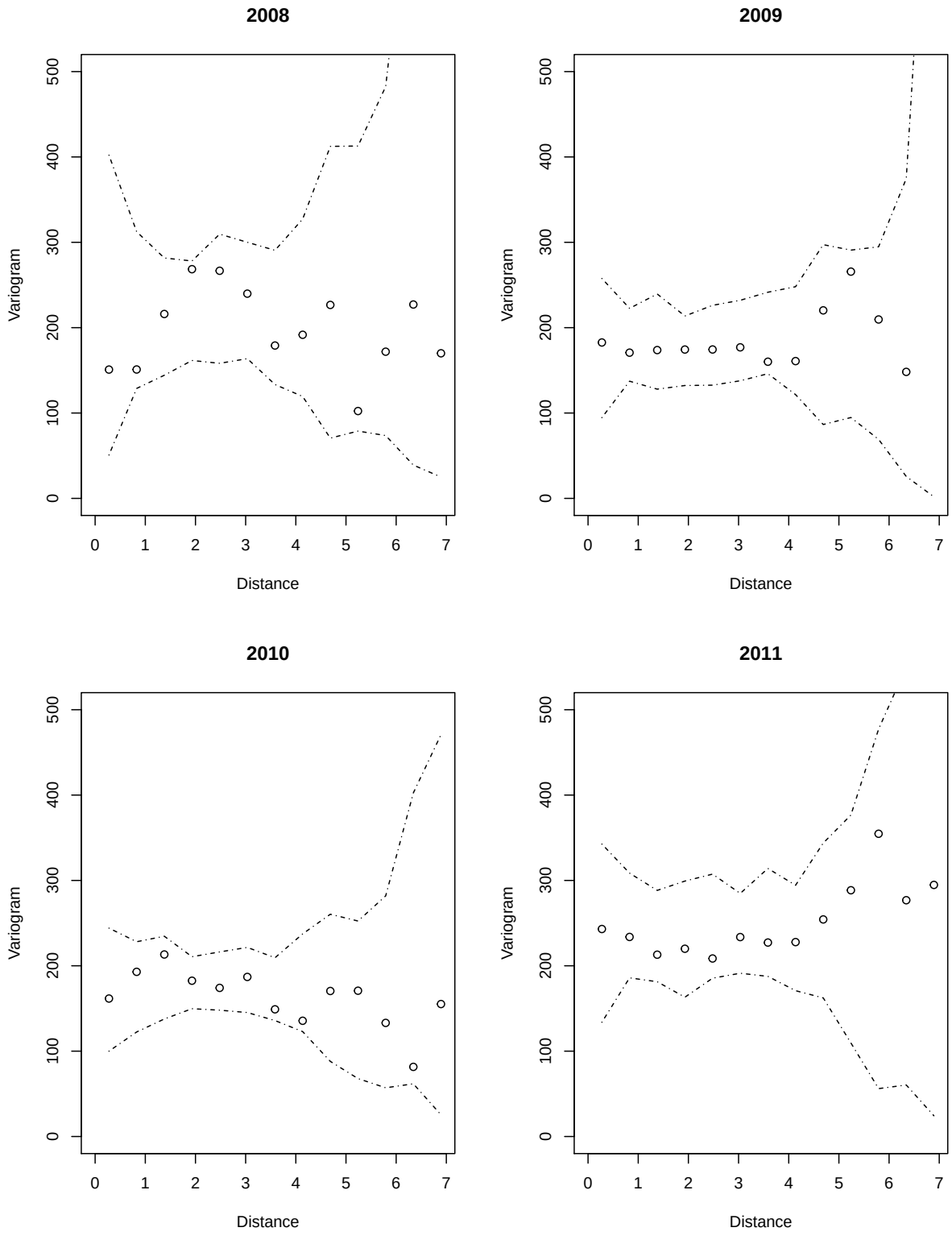


Figure S.1: Spatial variograms for each year from 2008 to 2011, based on the residuals from the corn yield prediction model. The unit in the horizontal axis is degree (in longitude or latitude). The dotted curves are confidence bands based on the assumption of no spatial dependency.

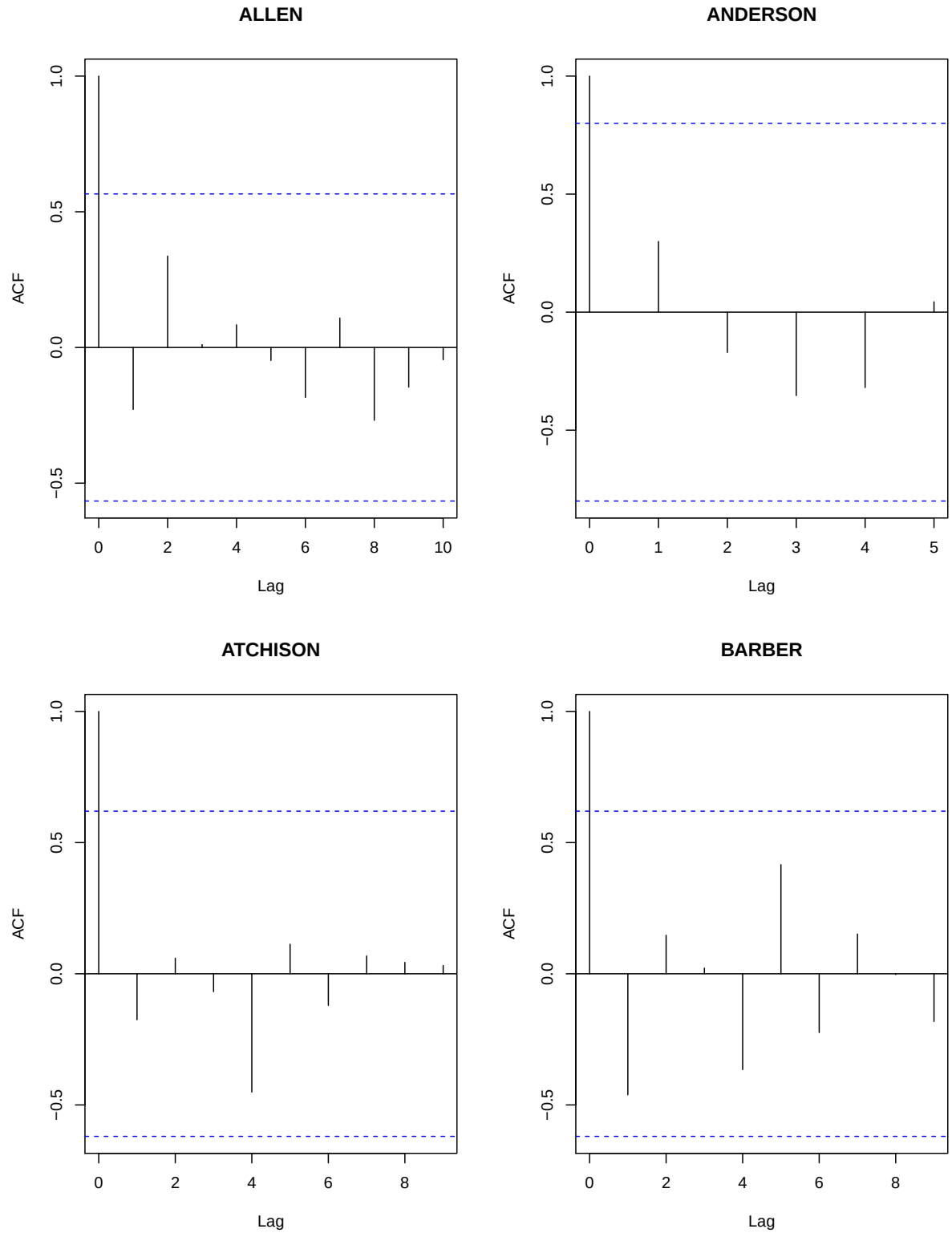


Figure S.2: The ACF plot for the first four counties, based on the residuals from the corn yield prediction model.