Statistica Sinica Preprint No: SS-2020-0418	
Title	Partially Linear Additive Functional Regression
Manuscript ID	SS-2020-0418
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0418

Complete List of Authors	Xiaohui Liu, Wenqi Lu, Heng Lian, Yuzi Liu and Zhongyi Zhu
Corresponding Author	Heng Lian
E-mail	henglian@cityu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Partially Linear Additive Functional Regression

Xiaohui Liu[1], Wenqi Lu[2,3], Heng Lian[2,4], Yuzi Liu[1] and Zhongyi Zhu[3]

Jiangxi University of Finance and Economics[1]*, City University of Hong Kong*[2]*,*

Fudan University[3]*, CityU Shenzhen Research Institute*[4]

Abstract: We consider a novel partially linear additive functional regression model in which both a functional predictor and some scalar predictors appear. The functional part has a semiparametric continuously additive form, while the scalar predictors appear in the linear part. The functional part has the optimal convergence rate, and the asymptotic normality of the nonfunctional part is also shown. Simulations and an empirical analysis of a Covid-19 data set demonstrate the performance of the proposed estimator.

Key words and phrases: Convergence rate; Functional data; Penalization; RKHS.

1. Introduction

Regression problems in statistics can be classified as parametric analysis, nonparametric, and semiparametric regressions. Statistical analyses of functional data, or data involving curves defined on a continuous domain, have been studied for decades, originating with the pioneering works on parametric models of, among others, Ramsay (1982) and Ramsay and Dalzell (1991). Nonparamet-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

ric kernel approaches to functional regression, well documented in the monograph of Ferraty and Vieu (2006), have also had a profound impact in this area. Other nonparametric functional regression approaches include those of Preda (2007) and Lian (2007, 2011). Fewer studies have examined semiparametric approaches; here, works include those of Ait-Saidi et al. (2008), Müller and Yao (2008), Chen et al. (2011), McLean et al. (2014), Zhu et al. (2014) and Radchenko et al. (2015), among others.

Several studies have considered the case in which we have both a functional predictor and a more classical set of scalar predictors. Shin (2009) coined the term partial functional linear regression for the proposed model in which both types of predictors are related to the response in a linear fashion, as given by

$$Y = \int_{\mathcal{T}} X(t)\beta(t)dt + \mathbf{Z}^\top \boldsymbol{\theta} + \epsilon, \tag{1.1}$$

where $\beta \in L^2(\mathcal{T})$ is the unknown slope function associated with the functional predictor X, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is the multivariate slope associated with the multivariate predictor $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$, with p assumed to be fixed. Kong et al. (2016) extended the model to incorporate multiple functions and high-dimensional multivariate predictors. On the other hand, Aneiros-Perez and Vieu (2006) considered a nonparametric alternative, where $\int_{\mathcal{T}} X(t)\beta(t)dt$ in (1.1) is

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

replaced by a general $m(X)$ using a nonparametric function $m : L^2(\mathcal{T}) \to \mathbb{R}$.

In this study, we use a semiparametric approach for the functional part proposed in McLean et al. (2014) and Müller et al. (2013). More specifically, we assume the model

$$Y = \int_{\mathcal{T}} F_0(t, X(t))dt + \mathbf{Z}^\top \boldsymbol{\theta}_0 + \epsilon. \tag{1.2}$$

The functional part above is a continuous version of the more familiar additive form $\sum_{i=1}^{J} F_j(t_j, X(t_j))$ when $J \to \infty$. For nonfunctional data analysis, additive models are frequently used to address the curse of dimensionality in multivariate nonparametric regression (Stone, 1986, Liang and Li, 2009, Xue and Liang, 2010, Wang et al., 2014). In general, additive models offer increased flexibility and potentially lower estimation bias than linear models. Furthermore, they have less variance in estimation and are less susceptible to the curse of dimensionality than are models that make no additivity assumptions. The same can be said about our functional extension (1.2).

We focus on a penalized estimation of the partially linear additive functional model in a reproducing kernel Hilbert space (RKHS) framework, as in Cai and Yuan (2012) and Wang and Ruppert (2015). Methodologically, this extension is related to the large body of literature on partially linear models and partially lin-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

ear additive models for nonfunctional data (Huang, 1999, Liang and Li, 2009). However, for functional data, the setups differ from those used in a nonparametric or semiparametric regression. Although Hall and Horowitz (2007) stated their rate is "generic to a large class of noisy inverse problems," this does not mean our theory can be obtained directly from theirs, because our model is in an RKHS framework. However, there are qualitative similarities, such as the rates being faster if the estimation target is smoother. Compared with Wang and Ruppert (2015), a main contribution of our work is to extend the model to include the case with a linear part, and to establish its asymptotic normality. We also incorporate an additional smoothness parameter (denoted as r), whereas Wang and Ruppert (2015) only considered $r = 0$. This requires a more careful analysis of the bias. For the parametric part, the need to use a profiling strategy also makes the proof more challenging. In addition, a minor point is that we try to clarify the issue of the unidentifiability of F_0, as mentioned in Wang and Ruppert (2015); however, its implications for the theoretical results are not clear.

We establish that the functional part of the estimator has the optimal estimation error, with the same rate as that in Wang and Ruppert (2015), whereas the linear part has the parametric rate. The methodology and theoretical results are presented in Section 2, with the proofs provided in the Supplementary Material. Section 3 reports our simulation results, and presents an empirical analysis of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

a real data set to illustrate our proposed approach. We conclude the paper in Section 4.

Finally, we list some notation and properties for the different norms used. For any operator $\mathcal{F}$, we use $\mathcal{F}^\top$ to denote its adjoint operator. If $\mathcal{F}$ is self-adjoint and nonnegative definite, $\mathcal{F}^{1/2}$ is its square root, satisfying $\mathcal{F}^{1/2}\mathcal{F}^{1/2} = \mathcal{F}$. For $f \in L^2(\mathcal{T})$ or $L^2(\mathcal{T}^2)$, $\|f\|$ denotes its L^2 norm. For any operator $\mathcal{F}$, $\|\mathcal{F}\|_{op}$ is the operator norm $\|\mathcal{F}\|_{op} := \sup_{\|f\|\leq 1} \|\mathcal{F}f\|$. The trace norm of an operator $\mathcal{F}$ is $tr(\mathcal{F}) = \sum_k \langle (\mathcal{F}^\top\mathcal{F})^{1/2} e_k, e_k\rangle$, for any orthonormal basis $\{e_k\}$. $\mathcal{F}$ is a trace class operator if its trace norm is finite. The Hilbert–Schmidt norm of an operator is $\|\mathcal{F}\|_{hs} = (\sum_{j,k}\langle \mathcal{F}e_j, e_k\rangle^2)^{1/2} = (\sum_j \|\mathcal{F}e_j\|^2)^{1/2}$. An operator is Hilbert–Schmidt if its Hilbert–Schmidt norm is finite. From the definition, it is easy to see that $tr(\mathcal{F}^\top\mathcal{F}) = tr(\mathcal{F}\mathcal{F}^\top) = \|\mathcal{F}\|_{hs}^2$. Furthermore, if $\mathcal{F}$ is a Hilbert–Schmidt operator and $\mathcal{G}$ is a bounded operator, then $\mathcal{F}\mathcal{G}$ is also a Hilbert–Schmidt operator, with $\|\mathcal{F}\mathcal{G}\|_{hs} \leq \|\mathcal{F}\|_{hs}\|\mathcal{G}\|_{op}$.

2. Profiled partially linear additive estimator

We assume $\mathcal{T} = [0, 1]$, without loss of generality. Following Wahba (1990), a RKHS $\mathcal{H} \subseteq L^2(\mathcal{W})$, where $\mathcal{W} = [0, 1] \times \mathcal{X}$ is a Hilbert space of real-valued functions with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ ($\langle\cdot,\cdot\rangle$ denotes the standard L^2 inner product), in which the point evaluation operator $L_w : \mathcal{H} \to \mathbb{R}, L_w(f) = f(w)$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

continuous. The corresponding norm induced by the inner product is denoted by $\|\cdot\|_{\mathcal{H}}$. Here, $\mathcal{X}$ can simply be $[0,1]$ when some strictly increasing distribution function is used to transform the range of $X(t)$ to $[0,1]$. This can be assumed without much loss of generality, because F_0 is a nonparametric function to be estimated. One can also use $\mathcal{X} = \mathbb{R}$, without resorting to such a transformation. By the Riesz representation theorem, this definition implies the existence of a nonnegative-definite square-integrable bivariate function $K(w, w')$, such that $K(w,\cdot) \in \mathcal{H}$ and $\langle K(w,\cdot), f\rangle_{\mathcal{H}} = f(w)$, for every $f \in \mathcal{H}$ and $w \in \mathcal{W}$. With a slight abuse of notation, K also denotes the linear operator $f \in L^2(\mathcal{W}) \to Kf = \iint K(\cdot,(t,x))f(t,x)dtdx$. For later use, we note that $\mathcal{H}$ is identical to the range of $K^{1/2}$.

In this section, we consider error bounds for the partially linear additive functional model (PLAFM), assuming the function F_0 is in some given RKHS $\mathcal{H}$. Given independent and identically distributed (i.i.d.) data $(X_i, \mathbf{Z}_i, Y_i)$, for $i = 1,\ldots,n$, the estimators of F_0 and $\boldsymbol{\theta}_0$ are obtained from

$$(\widehat{F}, \widehat{\boldsymbol{\theta}}) = \underset{F,\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \int F(t, X_i(t))dt - \mathbf{Z}_i^\top \boldsymbol{\theta} \right)^2 + n\lambda \|F\|_{\mathcal{H}}^2. \quad (2.1)$$

Our goal is to establish the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ and the optimal convergence rate of $\widehat{F}$. The challenge here is the former, which requires a profiling

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

technique often used in semiparametric statistics. Computationally, $\widehat{F}$ and $\widehat{\boldsymbol{\theta}}$ can be obtained simultaneously from (2.1). However, theoretically, we need to profile out F in order to study the asymptotic properties of the estimator for $\boldsymbol{\theta}$. More specifically, given any $\boldsymbol{\theta}$, we denote the minimizer of (2.1) for F (regarded as a function of $\boldsymbol{\theta}$) by $\widehat{F}(\cdot;\boldsymbol{\theta})$, or simply $\widehat{F}(\boldsymbol{\theta})$ for simplicity. Then, the final estimators for F and $\boldsymbol{\theta}$ are given by $\widehat{F}(\cdot;\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\theta}}$, respectively, where $\widehat{\boldsymbol{\theta}}$ is obtained from

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \int \widehat{F}(t, X_i(t); \boldsymbol{\theta}) dt - \mathbf{Z}_i^\top \boldsymbol{\theta} \right)^2. \tag{2.2}$$

Note that by the reproducing property, we have

$$\int F(t, x(t)) dt = \int \langle F, K(\cdot, (t, x(t))) \rangle_{\mathcal{H}} \, dt = \left\langle F, \int K(\cdot, (t, x(t))) dt \right\rangle_{\mathcal{H}} \tag{2.3}$$

Let $H = K^{-1/2} F \in L^2(\mathcal{W})$ and $G(x) = \int K^{1/2}(\cdot, (t, x(t))) dt \in L^2(\mathcal{W})$, where $K^{1/2}$ is the square root of K, defined by

$$K(w, w') = \iint K^{1/2}(w, (t, x)) K^{1/2}(w', (t, x)) dt dx. \tag{2.4}$$

Then, $\langle F, \int K(\cdot, (t, x(t))) dt \rangle_{\mathcal{H}} = \langle H, G(x) \rangle$. Thus, in terms of H, the optimiza-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

tion problem (2.1) becomes

$$(\widehat{H}, \widehat{\boldsymbol{\theta}}) = \underset{H \in L^2(\mathcal{W}), \boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \langle H, G(X_i) \rangle - \mathbf{Z}_i^\top \boldsymbol{\theta})^2 + n\lambda \|H\|^2. \tag{2.5}$$

For simplicity of notation, we denote $U = G(X)$ and $U_i = G(X_i)$, and define $f \otimes g : L^2(\mathcal{W}) \to L^2(\mathcal{W})$ by $(f \otimes g)(h) = \langle g, h \rangle f$, for any $h \in L^2(\mathcal{W})$. Let T be the operator $E[U \otimes U]$. We assume that T has a spectral expansion given by

$$T = \sum_{j=1}^{\infty} s_j e_j \otimes e_j,$$

where $s_1 \geq s_2 \geq \cdots \geq 0$ are the eigenvalues, and $\{e_j\}$ are the orthonormalized eigenfunctions (more specific assumptions on s_j are stated in Theorem 1).

Note that we can also define the bivariate function

$$T((t,x),(t',x')) = \iint E[K^{1/2}((t,x),(u,x(u)))K^{1/2}((t',x'),(v,x(v)))]dudv,$$

where $K^{1/2}(\cdot,\cdot)$ is defined in (2.4). Then, it is easy to see that $T = E[U \otimes U]$ is the operator mapping $f \in L^2(\mathcal{W})$ to $\iint T(\cdot,(t,x))f(t,x)dtdx$.

The estimator of the true function F_0 is $\widehat{F} = K^{1/2}\widehat{H}$. For a given $\boldsymbol{\theta}$, the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

minimizer of H in (2.5) has a closed form, and is given by

$$\widehat{H}(\cdot;\boldsymbol{\theta}) = (T_n + \lambda I)^{-1} \left(\sum_{i=1}^{n} (Y_i - \mathbf{Z}_i^\top \boldsymbol{\theta}) U_i / n \right), \tag{2.6}$$

where $T_n := \sum_{i=1}^n U_i \otimes U_i / n$, and I denotes the identity operator.

Before we state our assumptions. The important issue of the unidentifiability of the model needs to be addressed. More specifically, in (1.2), it is easy to see that if $F_0(t,x)$ is replaced by $F_0(t,x) + g(t)$, with some function g, such that $\int_0^1 g(t)dt = 0$, the regression function does not change. This is also shown in (2.3), which again shows that no $g \in \mathcal{H}$ with $\langle g, \int K(.,(t,X(t)))dt \rangle_{\mathcal{H}} = \int g(t)dt = 0$ can be recovered. Let $h = K^{-1/2}g$. Because $\langle g, \int K(.,(t,X(t)))dt \rangle_{\mathcal{H}} = \langle h, U \rangle$, such an h satisfies $\langle h, U \rangle = 0$ almost surely. Thus, we see that h is in the kernel of the operator $T = E[U \otimes U]$. In other words, unidentifiability here simply means T is not invertible (the eigenvectors e_i do not span $L^2(\mathcal{W})$). Solving this difficulty is then easy. We just need to focus on the case $H \in L_0^2(\mathcal{W}) := \{h : h \perp \mathrm{Ker}(T)\}$. By assuming $H_0 \in L_0^2(\mathcal{W})$ (or projecting H_0 onto $L_0^2(\mathcal{W})$ and only considering an estimation for this projected function), we can then consider the error $\|\widehat{H} - H_0\|$, for example.

The following technical assumptions are imposed.

(A1) $(X_i, Y_i, \mathbf{Z}_i, \epsilon_i)$ are i.i.d., $E\|G(X)\|^4 + E\|\mathbf{Z}\|^4 + E\|\epsilon\|^4 < \infty$, $E[\epsilon|X] = 0$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

and $E[\epsilon|\mathbf{Z}] = 0$.

(A2) For all $F \in \mathcal{H}$, we have

$$E\left(\int F(t, X(t))dt\right)^4 \leq C\left(E\left(\int F(t, X(t))dt\right)^2\right)^2,$$

for some constant $C > 0$.

(A3) We assume $H_0 := K^{-1/2}F_0 \in \{\sum_{j=1}^{\infty} b_j e_j : \sum_j b_j^2/s_j^r < \infty\} \cap L_0^2(\mathcal{W})$, for some $r \geq 0$.

These assumptions are mostly standard. (A2) is the same as that in Wang and Ruppert (2015). Note that because $\mathcal{H}$ is the range of $K^{1/2}$, when $r = 0$, (A3) simply states that $F_0 \in \mathcal{H}$. Because $s_j \to 0$, (A3) with a larger r can be regarded as assuming greater smoothness of F_0.Note too that assumption (A3) is the same as saying $H_0 \in Ran(T^{r/2})$, where $Ran(.)$ denotes the range of an operator.

Let $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0p})^\top \in (L^2(\mathcal{W}))^p$ be the minimizer obtained by

$$\gamma_{0j} = \min_{\gamma_j} E\left[\left(Z_j - \int \gamma_j(t, X(t))dt\right)^2\right], \ j = 1, \ldots, p, \tag{2.7}$$

and define $\eta_{ij} = Z_{ij} - \int \gamma_{0j}(t, X(t))dt$, and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})^\top$, where Z_j and Z_{ij} denote the jth components of $\mathbf{Z}$ and $\mathbf{Z}_i$, respectively.

(A4) $g_{0k} := K^{-1/2}\gamma_{0j} \in \{\sum_{j=1}^{\infty} b_j e_j : \sum_j b_j^2/s_j^r < \infty\}$, for $j = 1, \ldots, p$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

(A5) $E[\boldsymbol{\eta}_i] = 0$ and both $E[\mathbf{Z}\mathbf{Z}^\top]$ and $E[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top]$ have eigenvalues bounded and bounded away from zero.

Assumption (A4) means γ_{0j} is sufficiently smooth. Such a projection is often used in semiparametric models; for example, see equation (13) and assumption 3 of Li (2000). In particular, this ensures γ_{0j} can be estimated at some sufficiently fast rate. Assumption (A5) is related to the identifiability of $\boldsymbol{\theta}$. In particular, this means $\mathbf{Z}$ cannot be represented by the functional part. If $\mathbf{Z}$ and X are independent, then $E[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top]$ is indeed positive definite. Thus, the assumption on $\boldsymbol{\eta}$ can also be understood as requiring that the dependence between the two is not too strong.

A main goal of this study is to show that the functional part has the optimal L^2 error and prediction error. The L^2 error is simply $\|\widehat{F} - F_0\|_{\mathcal{H}} = \|\widehat{H} - H_0\|$. The prediction risk is defined as $E^*[\langle \widehat{H} - H_0, U^* \rangle^2] = E^*\left[\left(\int \widehat{F}(t, X^*(t)) - F_0(t, X^*(t))dt\right)^2\right]$, where $U^* = G(X^*)$, with X^* a copy of X, independent of the observed data, and where E^* denotes the expectation over the distribution of X^*. It is clear that the prediction risk can be equivalently written as $\|T^{1/2}(\widehat{H} - H_0)\|^2$.

Theorem 1. *Assume (A1)–(A5). If $s_j \asymp j^{-\alpha}$, for some constant $\alpha > 1$, and $r \in [0, 1]$ in assumptions (A3) and (A4), by setting $\lambda \asymp n^{-\alpha/((1+r)\alpha+1)}$, we have,*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

as $n \to \infty$*,*

$$E^*\langle \widehat{H} - H_0, U^* \rangle^2 = O_p\left(n^{-(1+r)\alpha/((1+r)\alpha+1)}\right). \tag{2.8}$$

Furthermore, when $r = 0$*, we have*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2}),$$

and when $r > 0$*,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}),$$

where $\Sigma_1 = E[(\mathbf{Z} - \langle U, \mathbf{g}_0\rangle)(\mathbf{Z} - \langle U, \mathbf{g}_0\rangle)^\top]$*, and* $\Sigma_2 = E[\epsilon^2(\mathbf{Z} - \langle U, \mathbf{g}_0\rangle)(\mathbf{Z} - \langle U, \mathbf{g}_0\rangle)^\top]$*, where* $\mathbf{g}_0 = K^{-1/2}\boldsymbol{\gamma}_0$*. The symbol* $\xrightarrow{d}$ *denotes convergence in distribution.*

Remark 1. *The rate in* (2.8) *with* $r = 0$ *is the same as that in Wang and Ruppert (2015) (note that Wang and Ruppert (2015) established the rate* $n^{-\frac{2r}{2r+1}}$*, where the parameter* $2r$ *in their notation is the same as our* α*). Our assumption (A3) with* $r = 0$ *is equivalent to saying that* $F_0 \in \mathcal{H}$*, without any additional smoothness property, whereas a larger* r *implies greater smoothness. We do not state the result for* $r > 1$*, but because assumptions (A3) and (A4) holding for* $r > 1$ *implies that they hold for* $r = 1$*, we actually have the rate*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

$E^*\langle \widehat{H} - H_0, U^* \rangle^2 = O_p\left(n^{-2\alpha/(2\alpha+1)}\right)$ *for* $r > 1$. *In other words, smoothness beyond* $r = 1$ *does not improve the convergence rate. We do not know if this is an artefact of our proof. For* $r = 0$*, Wang and Ruppert (2015) showed that for a mean regression of a functional additive model, the minimax rate for* F *in the unit ball of* $\mathcal{H}$ *is* $n^{-\alpha/(\alpha+1)}$*. For* $r > 0$*, we currently do not know whether the rate obtained is minimax optimal, although it is faster than that for* $r = 0$.

Remark 2. *We assume an RKHS in which* F_0 *is known to belong. If* F_0 *does not belong to* $\mathcal{H}$*, there is probably some additional bias we need to deal with, in theory, but we do not currently have any relevant theoretical results. The problem is not unlike that of the standard nonparametric regression, where we usually assume the function is in a certain space, such as a Hölder space, Sobolev space, or Besov space. However, in the functional regression setting in the RKHS framework, we are not aware of any studies on estimators when the RKHS is misspecified. We leave this as an open problem.*

3. Simulations and an application

In this section, we investigate the finite-sample properties of the proposed estimating procedure. Both simulated and real data are used.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

For the functional part, we focus on the RKHS $\mathcal{H}(K)$ with

$$K((s, X_i(s)), (t, X_j(t))) = \varphi\left(\sqrt{(s-t)^2 + (X_i(s) - X_j(t))^2}\right),$$

where $\varphi(\cdot)$ denotes the density function of the standard normal distribution. A similar derivation to that of Theorem 2 of Wang and Ruppert (2015) leads to the estimator of F_0 having the form $\sum_{j=1}^{n} c_j \int_0^1 K((t,x),(s,X_j(s)))ds$, for some $\boldsymbol{c} = (c_1, c_2, \cdots, c_n)^\top \in R^n$. Then, by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{Z}_i^\top \boldsymbol{\theta} - \boldsymbol{\Sigma}_i \boldsymbol{c})^2 + \lambda \boldsymbol{c}^\top \boldsymbol{\Sigma} \boldsymbol{c}$$

with respect to $(\boldsymbol{\theta}^\top, \boldsymbol{c}^\top)^\top$, we obtain the estimators $\hat{\boldsymbol{\theta}}$ and

$$\hat{F}(t,x) = \sum_{j=1}^{n} \hat{c}_j \int_0^1 K((t,x),(s,X_j(s)))ds.$$

Here, $\boldsymbol{\Sigma} = \left(\int_0^1 \int_0^1 K((t,X_i(t)),(s,X_j(s)))dtds\right)_{1\le i,j\le n}$, and $\boldsymbol{\Sigma}_i$ denotes the ith row of Σ. Note that, for a given λ, simple algebra yields

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{\Sigma} & \bar{\boldsymbol{Z}} \end{pmatrix} \begin{pmatrix} n\lambda \boldsymbol{I} + \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\bar{\boldsymbol{Z}} \\ \bar{\boldsymbol{Z}}^\top \boldsymbol{\Sigma} & \bar{\boldsymbol{Z}}^\top \bar{\boldsymbol{Z}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma} \\ \bar{\boldsymbol{Z}}^\top \end{pmatrix} \boldsymbol{Y} =: \boldsymbol{H}(\lambda)\boldsymbol{Y},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

where $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^\top$, $\bar{\boldsymbol{Z}} = (Z_{ij})_{1\le i\le n, 1\le j\le p}$, and $\boldsymbol{I}$ denotes the identity matrix of dimension n. Then, we can select the tuning parameter λ by minimizing the generalized cross-validation score $\text{GCV}(\lambda) = \|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2/(1 - tr(\boldsymbol{H}(\lambda))/n)^2$, as in Wahba(1990).

3.1 Simulations

In this section, we carry out some simulation studies. We generate the data $\{X_i, Y_i\}_{i=1}^n$ from the following model:

$$Y_i = \int_0^1 F_0(t, X_i(t))dt + Z_{1i} + 0.5Z_{2i} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_0^2)$, and $Z_{1i}, Z_{2i} \sim U([0, 5])$ are independent. For F_0, we consider the following three cases:

- Case 1. $F_0(t, X_i(t)) = \beta_0(t)X_i(t)$;
- Case 2. $F_0(t, X_i(t)) = \cos(t - X_i(t) - 5)$;
- Case 3. $F_0(t, X_i(t)) = t\exp(X_i(t))$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

In Case 1, we take

$$\begin{aligned} X_i(t) &= \zeta_1 W_{1i} + \sum_{k=2}^{50} \sqrt{2}\zeta_k W_{ki} \cos(k\pi t), \quad t \in [0,1], \\ \beta_0(t) &= 0.3 + \sum_{k=2}^{50} 4\sqrt{2}(-1)^{k+1} k^{-2} \cos(k\pi t), \end{aligned}$$

and $\zeta_k = (-1)^{k+1} k^{-\nu/2}$, with $\nu = 1.1$, where W_{ik} are independently uniform on $[-\sqrt{3}, \sqrt{3}]$. In Cases 2 and 3, we take $X_i(t) =$

$$\cos(U_{i1}) \sin(\pi t/5) + \sin(U_{i1}) \cos(\pi t/5) + \cos(U_{i2}) \sin(2\pi t/5) + \sin(U_{i2}) \cos(2\pi t/5),$$

where U_{i1}, U_{i2} are i.i.d. from Uniform$[0, 2\pi]$. We investigate four combinations $(n, \sigma) \in \{(50, 0.5), (50, 1), (100, 0.5), (100, 1)\}$ for each case.

For each setting, we repeat the experiment 1000 times. In each repeated experiment, we compute the value of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$, and the root mean squared prediction error,

$$\text{RMSPE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\int_0^1 F_0(t, X_i(t)) dt - \int_0^1 \hat{F}(t, X_i(t)) dt \right)^2},$$

where m denotes the sample size of the test data. Table 1 reports the mean of these two quantities, as well as their standard deviation values, computed in

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

1000 experiments. We compare two models, namely, a partially linear functional regression model (PLFM) and a partially linear additive functional regression model (PLAFM). It turns out that both quantities become smaller as the sample size increases. As expected, the PLAFM outperforms the PLFM in the nonlinear cases in terms of the RMSPE.

Furthermore, we provide four Q–Q plots of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ in Case 2. Figure 1 shows that the empirical distributions of both $\hat{\theta}_1$ and $\hat{\theta}_2$ match the normal distribution quite well.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Table 1: The root mean squared predicted errors for $\hat{F}$ and $\hat{\boldsymbol{\theta}}$ with independent covariates.

RMSPE			$n = 50$		$n = 100$	
	Case	σ	PLAFM	PLFM	PLAFM	PLFM
	1	0.5	0.7848(0.1245)	0.7420(0.1162)	0.7749(0.0878)	0.7239(0.0789)
		1	0.7877(0.1262)	0.7495(0.1223)	0.7801(0.0887)	0.7341(0.0819)
	2	0.5	0.1004(0.0479)	0.1231(0.0103)	0.0889(0.0430)	0.1241(0.0074)
		1	0.1314(0.0798)	0.1239(0.0103)	0.1192(0.0735)	0.1236(0.0073)
	3	0.5	0.5853(0.0687)	0.8061(0.0146)	0.4957(0.0656)	0.8060(0.0107)
		1	0.5872(0.1233)	0.8057(0.0148)	0.4917(0.1203)	0.8057(0.0105)
$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$			$n = 50$		$n = 100$	
	Case	σ	PLAFM	PLFM	PLAFM	PLFM
	1	0.5	0.0988(0.0597)	0.0946(0.0581)	0.0727(0.0438)	0.0624(0.0373)
		1	0.1349(0.0804)	0.1336(0.0827)	0.0895(0.0508)	0.0926(0.0582)
	2	0.5	0.0596(0.0346)	0.0595(0.0314)	0.0429(0.0226)	0.0472(0.0214)
		1	0.1132(0.0696)	0.1109(0.0698)	0.0806(0.0454)	0.0796(0.0432)
	3	0.5	0.1560(0.0309)	0.2060(0.0264)	0.1265(0.0226)	0.1998(0.0180)
		1	0.1863(0.0633)	0.2312(0.0553)	0.1430(0.0444)	0.2110(0.0363)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Table 2: The root mean squared predicted errors for $\hat{F}$ and $\hat{\boldsymbol{\theta}}$ with dependent covariates.

RMSPE			$n = 50$		$n = 100$	
	Case	σ	PLAFM	PLFM	PLAFM	PLFM
	1	0.5	0.6520(0.2106)	0.3762(0.1559)	0.6303(0.1423)	0.2756(0.0862)
		1	0.6609(0.2137)	0.6038(0.1977)	0.6419(0.1448)	0.5682(0.1280)
	2	0.5	0.0376(0.0224)	0.0471(0.0552)	0.0296(0.0142)	0.0313(0.0253)
		1	0.0389(0.0243)	0.0430(0.7449)	0.0303(0.0149)	0.0286(0.0235)
	4	0.5	1.5843(1.0383)	4.7431(1.7110)	1.2490(0.6622)	4.5914(1.0956)
		1	1.4665(1.0068)	4.7207(1.6790)	1.1671(0.6411)	4.5804(1.0828)
$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$			$n = 50$		$n = 100$	
	Case	σ	PLAFM	PLFM	PLAFM	PLFM
	1	0.5	0.7712(0.1652)	0.7421(0.1502)	0.7781(0.1121)	0.7505(0.0957)
		1	0.7715(0.1657)	0.7460(0.1632)	0.7786(0.1127)	0.7528(0.1064)
	2	0.5	0.7564(0.1256)	0.7444(0.1275)	0.7656(0.0831)	0.7540(0.0836)
		1	0.7564(0.1256)	0.0504(0.1269)	0.7655(0.0831)	0.7544 (0.0837)
	4	0.5	0.7452(0.2130)	0.7448(0.2946)	0.7561(0.1326)	0.7585(0.2091)
		1	0.7451(0.2074)	0.7449(0.2938)	0.7559(0.1298)	0.7583(0.2092)

Note that in the previous cases, all covariates Z_1, Z_2, and X are independent. In the following, we investigate a more complex scenario such that $(Z_1, Z_2, W_{i1})^\top \sim N(0, \Sigma)$ with $\Sigma = (1, 0.3, 0; 0.3, 1, 0; 0, 0.3, 1)$, and

$$X_i(t) = \zeta_1 W_{1i} + \sum_{k=2}^{50} \sqrt{2}\zeta_k W_{ki} \cos(k\pi t), \quad t \in [0, 1],$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

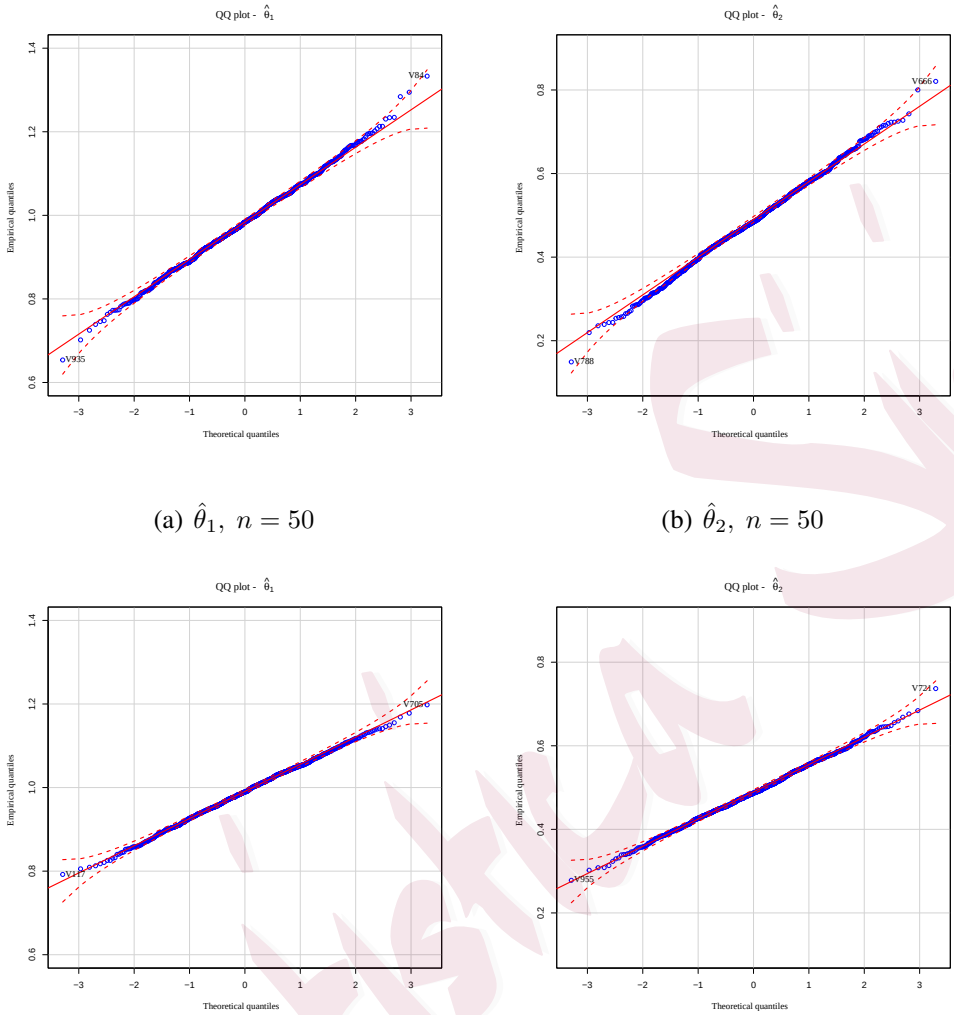
where $\zeta_k = (-1)^{k+1}k^{-\nu/2}$, with $\nu = 1.1$, and W_{ik}, for $k = 2, 3, \cdots, 50$, are independently uniform on $[-\sqrt{3}, \sqrt{3}]$. Obviously, Z_1, Z_2, and X are dependent. Here F_0 takes the same value as Case 1 and 2. We also use a new case,

- Case 4. $F_0(t, X_i(t)) = tX_i^2(t)$.

Using the same estimating procedure, we obtain 1000 estimates of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)^\top$. Table 2 reports the average of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$ and the root mean squared prediction errors, as well as their standard errors. It turns out that we have similar observations to those in Table 1 when Z_1, Z_2, and X are dependent.

Furthermore, we are interested in evaluating the finite-sample performance of the joint asymptotic normality of the parametric part. In descriptive statistics, a graphical method for assessing whether or not observations are generated from a multivariate normal distribution is the DD plot (depth-versus-depth plot), introduced by Liu et al. (1999). We generate the DD plot for $(\widehat{\theta}_1, \widehat{\theta}_2)$ based on the half-space depth, as shown in Figure 2. All points lie close to the diagonal line, which indicates that it is reasonable to conclude that the estimates are roughly jointly normally distributed.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

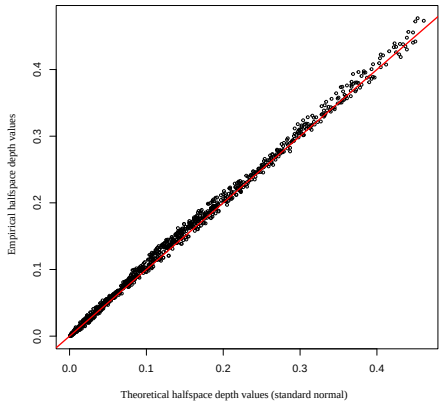


(a) $\hat{\theta}_1,\ n = 50$ (b) $\hat{\theta}_2,\ n = 50$

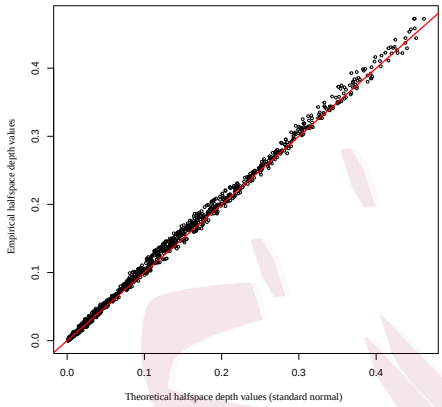
(c) $\hat{\theta}_1,\ n = 100$ (d) $\hat{\theta}_2,\ n = 100$

Figure 1: Q–Q plots for the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of the PLAFM in Case 2 with $\sigma^2 = 1$.

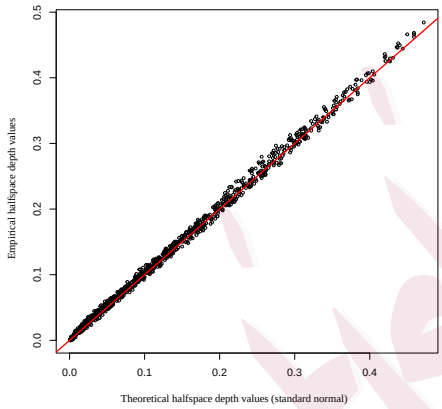
Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418



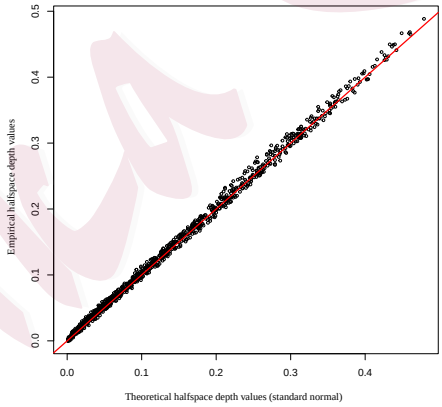
(a) $n = 50,\ \sigma = 0.5$



(b) $n = 50,\ \sigma = 1$



(c) $n = 100,\ \sigma = 0.5$



(d) $n = 100,\ \sigma = 1$

Figure 2: DD plots for the standardized estimates $(\hat{\theta}_1, \hat{\theta}_2)^\top$ of the PLAFM in Case 2 with dependent covariates.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

3.2 An application to a COVID-19 data set

In this section, we illustrate the performance of the proposed model by applying it to a 2019 coronavirus data set, hereafter Covid-19 for convenience. Since the first case reported in Wuhan, China, in December 2019, Covid-19 has emerged as a global public health incident, with a rapid increase in cases and deaths.

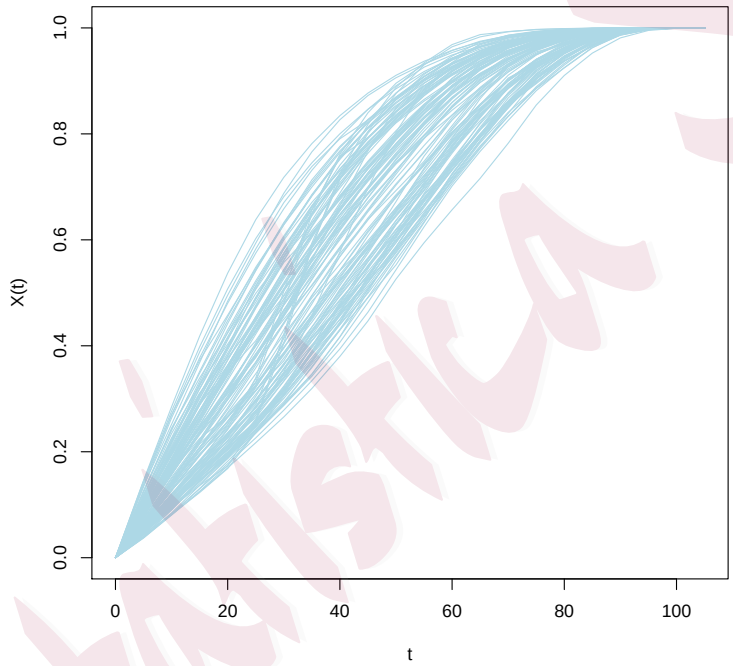


Figure 3: The cumulative probability functions for age in 127 countries or regions outside of Africa.

In the following, we apply the PLAFM to model the relationship between the mortality rates in various countries, and their demographics. Note that the age

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

distribution of a population of a country is related to the level of its medical facilities and economic status. The data show that the proportion of the elderly population in a developed country tends to be higher than those of other countries. Obviously, the demography of a country can be characterized by functional data, with age as the independent variable. Hence, we take the cumulative probability functions for the ages of the populations in countries with reported mortality as the functional covariate. There are 127 countries or regions with reported mortality, which can be downloaded from the World Health Organization (`https://www.who.int/emergencies/diseases/novel-coronavirus-2019`). The values are calculated using the equation mortality = cumulative deaths / cumulative cases. Note that we do not consider data from African countries and some small islands. The Covid-19 data sets update daily, and we collected the cumulative cases until 22 May. Up-to-date data on population by age groups are available at the United Nations Department of Economic and Social Affairs Population Dynamics, World Population Prospects 2019 (`https://population.un.org/wpp/Download/Standard/Population/`). Figure 3 reports the cumulative probability functions of 127 countries or regions outside of Africa. Note that age has been unitized.

Hereafter, we write y_i, for $i = 1, 2, \cdots, 127$, as the mortality of the ith country or region, and $x_i(t)$ is the related cumulative probability at age t. We conduct

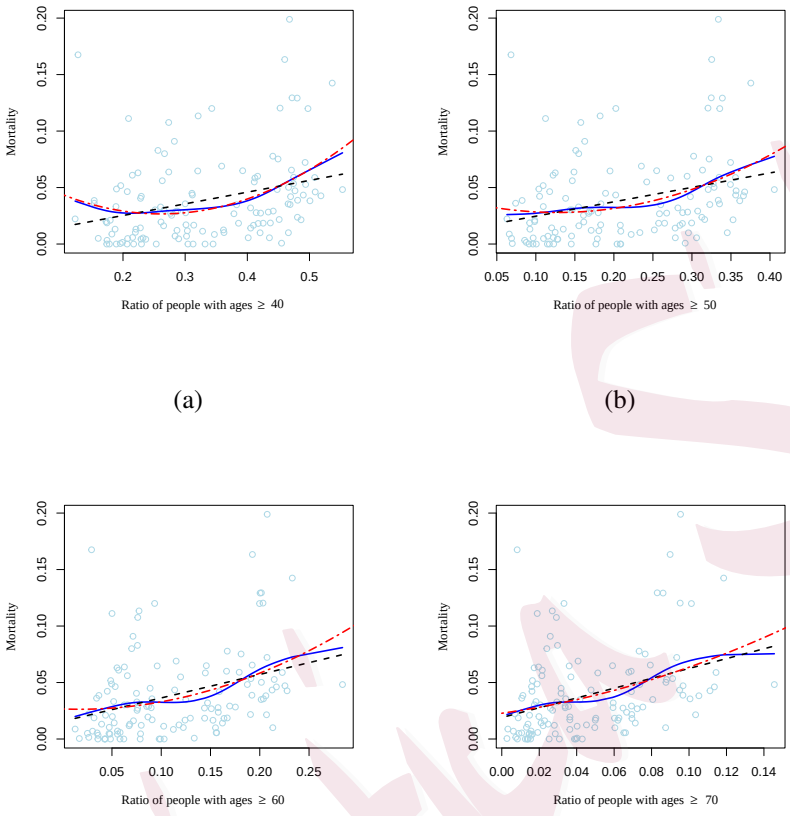
Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

an initial analysis for the marginal relationship between y_i and $1-x_i(t)$ for given $t = 40/105, 50/105, 60/105, 70/105$. It turns out that the relationship between y and $x(t)$ is probably *not* linear at some given point t, and this relationship may vary as t changes. Figure 4 reports scatter plots of mortality versus the ratio of people older than 40, 50, 60, and 70, as well as the lines fitted by the linear, quadratic, and smoothing splines regression models. As shown in Figure 4-(a), it seems better to fit the data using the quadratic model, or some other nonlinear model, than it is to use the linear model. On the other hand, the mortalities are relatively small for some countries from high latitudes, such as Russia and Canada. It is therefore natural to wonder whether mortality is correlated with latitude. The latitude data are taken from `https://www.latlong.net/`. Motivated by this, we consider the following PLAFM for this data set:

$$y_i = \theta_1 z_i + \theta_2 z_i^2 + \int_0^1 F(t, x_i(t))dt + \varepsilon_i, \quad i = 1, 2, \cdots, 127,$$

where z_i denotes the absolute latitude of the capital of the ith country or region. Note that the latitudes of southern hemisphere countries take negative values.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418



(a) (b)

(c) (d)

Figure 4: Scatter plots of mortality versus the ratio of people older than 40, 50, 60, and 70, where the dash line is fitted by the linear model, the dot-dash line is by the quadratic model, and the solid line is by the smoothing splines.

The PLAFM and PLFM are employed to fit this data set. We choose the data with indices from 1 to $m = \lfloor 0.75n \rfloor = 101$ as the training data to fit the models, and use the rest as the test data. It turns out that the test error $\sqrt{\frac{1}{n}\sum_{k=m+1}^{n}(Y_k - \hat{Y}_k)^2}$ for the PLAFM is 0.0556, slightly smaller than the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

0.0572 of the PLFM. The resulting estimate of F is illustrated in Figure 5. Furthermore, the estimate of the coefficients of the linear part in the PLAFM is $\hat{\boldsymbol{\theta}} = (0.0227, -0.0659)^\top$, and that in the PLFM is $(0.0274, -0.0684)^\top$. This indicates that the relationship between mortality and latitude has an inverse U-shape, implying that countries with high or low latitudes have lower Covid-19 mortality rates. Note that the temperature in the low latitude area is relatively high, and the temperature difference between day and night is small. However, in the high latitude area, although the outdoor temperature is low, this temperature difference is also small, owing to the existence of heating systems. In contrast, the temperature difference in the middle latitudes is greater than that of the low and high latitude areas. In medicine, it is known that the temperature difference between day and night affects the mortality rate of respiratory diseases, which is supported by our findings.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

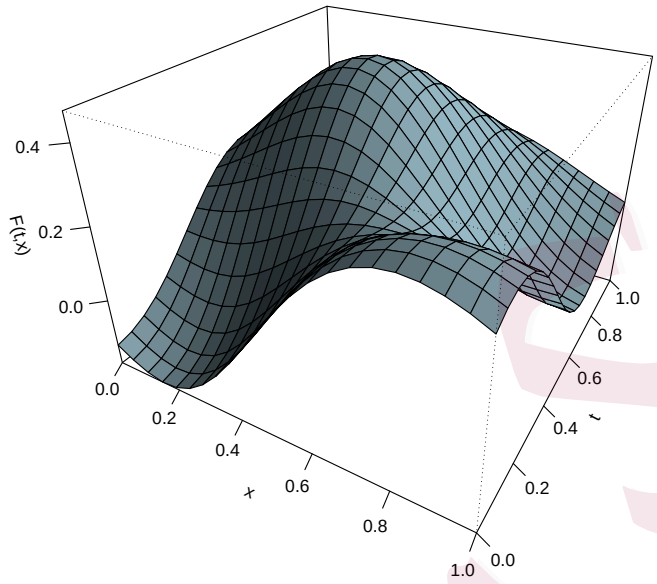


Figure 5: The fitted function $\hat{F}(t, x)$.

4. Conclusion

We have extend the results on optimal prediction for the functional additive regression model based on an RKHS framework to the partially functional additive regression model. When reduced to the purely functional case, the derived convergence rate complements the results in Wang and Ruppert (2015), because we also use stronger smoothness assumptions.

In the existing literature, the functional quantile regression model has also

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

been examined using the functional principle component analysis method. It would be interesting to determine whether the RKHS framework can be adapted for the functional quantile additive regression and the partially functional quantile additive regression.

A notable limitation in our empirical data analysis is that our scalar covariates only include latitude. Other variables, such as mitigation methods, medical facilities, and economic status, may be more relevant. This is left for further work.

Acknowledgments

We sincerely thank the editor Prof. Yuan-chin Chang, Prof. Rong Chen, an associate editor, and two anonymous reviewers for their insightful comments. The research of Heng Lian was supported by Project 11871411 from the NSFC and the Shenzhen Research Institute, City University of Hong Kong, and by Hong Kong RGC general research fund 11301718, 11300519, and 11300721. The research of Xiaohui Liu was supported by the NNSF of China (Grant No.11971208, 11601197), China Postdoctoral Science Foundation funded project (2016M600511, 2017T100475), and NSF of Jiangxi Province (No. 2018ACB21002, 20171ACB21030, 20192BAB201005).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Supplementary Material

The online supplementary material contains the proofs of the theorems.

References

Ait-Saidi, A., Ferraty, F., Kassa, R. and Vieu, P. (2008) Cross-validated estimations in the single-functional index model. *Statistics*, **42**, 475–494.

Aneiros-Perez, G. and Vieu, P. (2006) Semi-functional partial linear regression. *Statistics & Probability Letters*, **76**, 1102–1110.

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

Chen, D., Hall, P. and Mueller, H. G. (2011) Single and multiple index functional regression models with nonparametric link. *Annals of Statistics*, **39**, 1720–1747.

Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer series in statistics. New York, NY: Springer. URLhttp://dx.doi.org/10.1007/0-387-36620-2.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Hall, P. and Horowitz, J. L. (2007) Methodology and convergence rates for functional linear regression. *Annals of Statistics*, **35**, 70–91.

Huang, J. (1999) Efficient estimation of the partly linear additive Cox model. *Annals of Statistics*, **27**, 1536–1563.

Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016) Partially functional linear regression in high dimensions. *Biometrika*, **103**, 147–159.

Li, Q. (2000) Efficient estimation of additive partially linear models. *International Economic Review*, **41**, 1073–1092.

Lian, H. (2007) Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, **35**, 597–606.

Lian, H. (2011) Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, **5**, 31–40.

Liang, H. and Li, R. Z. (2009) Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, **104**, 234–248.

Liu, R. Y., Parelius, J. M., Singh, K. et al. (1999) Multivariate analysis by data

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

depth: descriptive statistics, graphics and inference. *The annals of statistics*, **27**, 783–858.

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.

Müller, H.-G., Wu, Y. and Yao, F. (2013) Continuously additive models for nonlinear functional regression. *Biometrika*, **100**, 607–622.

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

Preda, C. (2007) Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, **137**, 829–840.

Radchenko, P., Qiao, X. and James, G. M. (2015) Index models for sparsely sampled functional data. *Journal of the American Statistical Association*, **110**, 824–836.

Ramsay, J. O. (1982) When the data are functions. *Psychometrika*, **47**, 379–396.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

Journal of the Royal Statistical Society Series B-Methodological, **53**, 539–572.

Shin, H. (2009) Partial functional linear regression. *Journal of Statistical Planning and Inference*, **139**, 3405–3418.

Stone, C. J. (1986) The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**, 590–606.

Wahba, G. (1990) *Spline models for observational data*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

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. *The Annals of Statistics*, **42**, 592–624.

Wang, X. and Ruppert, D. (2015) Model, Optimal prediction in an additive functional model. *Statistica Sinica*, **25**, 567–589.

Xue, L. and Liang, H. (2010) Polynomial spline estimation for a generalized additive coefficient model. *Scandinavian Journal of Statistics*, **37**, 26–46.

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 (Statistical Methodology)*, **76**, 581–603.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0418

School of Statistics, Jiangxi University of Finance and Economics

E-mail: csuliuxh912@gmail.com

Department of Statistics, School of Management, Fudan University, Shanghai 200433, China

E-mail: fdwenqilu@outlook.com

Department of Mathematics, City University of Hong Kong, Hong Kong, China

E-mail: henglian@cityu.edu.hk

School of Statistics, Jiangxi University of Finance and Economics

E-mail: yzliu@hotmail.com

Department of Statistics, School of Management, Fudan University, Shanghai 200433, China

E-mail: zhuz@fudan.edu.cn