Statistica Sinica **25** (2015), 567-589
doi:http://dx.doi.org/10.5705/ss.2013.074

OPTIMAL PREDICTION IN AN ADDITIVE FUNCTIONAL MODEL

Xiao Wang and David Ruppert

Purdue University and Cornell University

Abstract: The functional generalized additive model (FGAM), also known as the continuous additive model (CAM), provides a more flexible functional regression model than the well-studied functional linear regression model. This paper restricts attention to the FGAM with identity link and additive errors, which we will call the additive functional model and is a generalization of the functional linear model. We study the minimax rate of convergence of predictions from the additive functional model in the framework of reproducing kernel Hilbert spaces. It is shown that the optimal rate is determined by the decay rate of the eigenvalues of a certain kernel function, which in turn is determined by the reproducing kernel and the joint distribution of any two points in the random predictor function. In the special case of the functional linear model, this kernel function is jointly determined by the covariance function of the predictor function and the reproducing kernel. The easily implementable roughness-regularized predictor is shown to achieve the optimal rate of convergence. Numerical studies are carried out to illustrate the merits of the predictor. Our simulations and real data examples demonstrate a competitive performance against the existing approach.

Key words and phrases: Functional regression, minimax rate of convergence, principal component analysis, reproducing kernel Hilbert space.

1. Introduction

Functional regression, in particular functional linear regression (FLR), has been studied extensively. Recent synopses include Ramsay and Silverman (2002, 2005), Ferraty and Vieu (2006), and Ramsay, Hooker, and Graves (2009). Let $X(\cdot)$ be a random process defined on $[0,1]$ and Y be the univariate response variable. Typically, t is restricted to a compact interval, so the assumption that $t \in [0,1]$ causes no loss of generality. Suppose we observe n i.i.d. copies of $\big(Y, X\big)$, $\big(Y_i, X_i\big)$, $i = 1, \ldots, n$. The functional linear regression model assumes that

$$Y_i = \alpha_0 + \int_0^1 \beta_0(t) X_i(t) dt + \epsilon_i, \tag{1.1}$$

where $\alpha_0 \in \mathbb{R}$ is the coefficient constant, $\beta_0 : [0,1] \to \mathbb{R}$ is the slope function, and the ϵ_i are i.i.d. random errors with $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2,\ 0 < \sigma^2 < \infty$. One of the popular methods for estimating functional linear models (FLMs) is based on functional principal component analysis (FPCA) (see, e.g., James (2002), Yao, Müller, and Wang (2005), Cai and Hall (2006), Li and Hsing (2007), Hall and Horowitz (2007)). In addition, methods of regularization have been applied to the FLM (see, e.g., Crambes, Kneip, and Sarda (2009), Yuan and Cai (2010), Cai and Yuan (2012)).

Due to the limitation inherent in the assumed linearity of (1.1), Ferraty and Vieu (2006) extended this model to nonparametric functional models and Müller and Yao (2008) discussed functional models that are additive in the functional principal component scores of the predictor functions. Recently, McLean et al. (2014) proposed a new model called a functional generalized additive model (FGAM). The same model was studied by Müller, Wu, and Yao (2012), there termed the continuously additive model (CAM). We study the special case of the FGAM with the identify link and continuous errors so that

$$Y_i = \int_0^1 F_0\Big(t, X_i(t)\Big)dt + \epsilon_i, \tag{1.2}$$

where $F_0(\cdot,\ \cdot) : [0,1]^2 \to \mathbb{R}$ is a bivariate function. Because F_0 is nonlinear, $X(t)$ can be replaced by $G\{X(t)\}$ for a smooth transformation G. Since G can be strictly increasing function from the entire real line to $[0,1]$, assuming that $X(t) \in [0,1]$ causes no loss of generality. (In McLean et al. (2014), $G = G_t$ is allowed to depend on t and is an estimate of the CDF of $X(t)$, but we do not pursue this refinement here.) Model (1.2) is called the additive functional model and contains (1.1) as a special case with $F_0(t,x) = \alpha_0 + x\beta_0(t)$. The additive functional model offers increased flexibility compared to (1.1), while still facilitating interpretation and estimation. In McLean et al. (2014), computational issues of this model were studied and F_0 was estimated using tensor-product B-splines with roughness penalties Eilers and Marx (1996). In Müller, Wu, and Yao (2012), a piecewise constant function was fit to F_0 and the asymptotic properties, e.g., consistency and asymptotic normality, of predictions based on $\widehat{F}_0$ were studied.

We study minimax prediction. The unknown bivariate function F_0 is assumed to reside in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}(K)$ with a reproducing kernel $K : [0,1]^2 \times [0,1]^2 \to \mathbb{R}$. The goal of prediction is to recover the functional η_0:

$$\eta_0(X) = \int_0^1 F_0\Big(t, X(t)\Big)dt,$$

based on the training sample (Y_i, X_i), $i = 1, \ldots, n$. Let $\widehat{F}_n$ be an estimate of F_0 from the training data. Then its accuracy can be measured by the excess risk:

$$\mathfrak{R}_n := \mathbb{E}^*\Big[Y_{n+1} - \int_0^1 \widehat{F}_n\Big(t, X_{n+1}(t)\Big)dt\Big]^2 - \mathbb{E}^*\Big[Y_{n+1} - \int_0^1 F_0\Big(t, X_{n+1}(t)\Big)dt\Big]^2$$

$$=\mathbb{E}^*\Big\{\int_0^1\Big[\widehat{F}_n(t,X_{n+1}(t))-F_0(t,X_{n+1}(t))\Big]dt\Big\}^2,$$

where (Y_{n+1},X_{n+1}) possesses the same distribution as (Y_i,X_i) and is independent of (Y_i,X_i), $i=1,\ldots,n$, and $\mathbb{E}^*$ represents taking expectation over (Y_{n+1},X_{n+1}) only. We study the rate of convergence of $\mathfrak{R}_n$ as n increases, which reflects the difficulty of the prediction problem. A closely related problem is to estimate the bivariate function F_0. F_0 is not identifiable without constraints since, for example, for any function $b(t)$, one can add $b(t)-\int_0^1 b(t)dt$ to F_0 without changing $\int F\Big(t,X(t)\Big)dt$.

The optimal rate of convergence for the prediction problem is to be established. The spectral theorem admits that there exist a set of orthonormalized eigenfunctions $\{\psi_k : k\geq 1\}$ and a sequence of eigenvalues $\kappa_1\geq\kappa_2\geq\cdots>0$ such that

$$K\Big((t,x);(s,y)\Big)=\sum_{k=1}^{\infty}\kappa_k\psi_k(t,x)\psi_k(s,y),$$

$$K(\psi_k):=\int\!\!\int K\Big(\cdot;(s,y)\Big)\psi_k(s,y)dsdy=\kappa_k\psi_k.$$

It is shown that under model (1.2), the difficulty of the prediction problem as measured by the minimax rate of convergence depends on the decay rate of the eigenvalues of the kernel $C:[0,1]^2\times[0,1]^2\to\mathbb{R}$ defined by

$$C\Big((t,x);(s,y)\Big)=\int\!\!\int\mathbb{E}\Big\{K^{1/2}\Big((t,x);(u,X(u))\Big)\ \ K^{1/2}\Big((s,y);(v,X(v))\Big)\Big\}dudv, \tag{1.3}$$

where $K^{1/2}\Big((t,x);(s,y)\Big)=\sum_{k=1}^{\infty}\kappa_k^{1/2}\psi_k(t,x)\psi_k(s,y)$. A minimax lower bound is first derived for the prediction problem. Then a roughness-regularized predictor is introduced and is shown to attain the rate of convergence given in the lower bound.

Example 1. We restrict the bivariate function F to the specific form $F(t,x)=\beta(t)x$, where β belongs to a reproducing kernel Hilbert space $\mathcal{H}(\widetilde{K})$ with the reproducing kernel $\widetilde{K}:[0,1]\times[0,1]\to\mathbb{R}$. This essentially provides us with a functional linear regression model. Let

$$\widetilde{K}(t,s)=\sum_{k=1}^{\infty}\varsigma_k\varphi_k(t)\varphi_k(s),$$

where the (ς_k,φ_k) are the eigenvalue and eigenfunction pairs. It is not hard to see that

$$K\Big((t,x);(s,y)\Big)=3\widetilde{K}(t,s)xy=\sum_{k=1}^{\infty}\kappa_k\psi_k(t,x)\psi_k(s,y),$$

where $\kappa_k = \varsigma_k$ and $\psi_k(t,x) = \sqrt{3}x\varphi_k(t)$. Therefore,

$$C\Big((t,x);(s,y)\Big) = 3xy \iint \widetilde{K}^{1/2}(t,u)G(u,v)\widetilde{K}^{1/2}(v,s)dudv, \tag{1.4}$$

where $G(u,v) = \operatorname{cov}(X(u), X(v))$ is the covariance function of X. So, the eigenvalues of C have the same decay rate as the eigenvalues of the kernel $\widetilde{K}^{1/2}G\widetilde{K}^{1/2}$ defined as the integral on the right-hand side of (1.4). As in Yuan and Cai (2010) and Cai and Yuan (2012), the decay rate of the eigenvalues of $\widetilde{K}^{1/2}G\widetilde{K}^{1/2}$ is assumed to be of k^{-2r}.

Example 2. Let $\phi_1(t) = 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t)$, $k \geq 1$. Take $\psi_k(t,x) = \phi_k(t)\phi_k(x)$, $k \geq 1$. Here $\{\psi_k : k \geq 1\}$ is a set of orthonormal basis functions and we construct the reproducing kernel

$$K\Big((t,x);(s,y)\Big) = \sum_{k=1}^{\infty} \kappa_k \psi_k(t,x)\psi_k(s,y),$$

where κ_k is the kth eigenvalue. Therefore,

$$C\Big((t,x);(s,y)\Big) = \sum_{k,\ell=1}^{\infty} \kappa_k^{1/2} a_{k\ell} \kappa_\ell^{1/2} \phi_k(t)\phi_k(s)\phi_\ell(s)\phi_\ell(y),$$

where

$$a_{k\ell} = \iint \mathbb{E}\Big(\phi_k(X(u))\phi_\ell(X(v))\Big)dudv.$$

It is possible to obtain the eigenvalue decay rate for C numerically. Numerical computation suggests that the eigenvalues of C have a polynomial decay rate if both the eigenvalues of X and the κ_k have the polynomial decay rates.

Example 3. Consider the thin-plate splines space $\mathcal{H} = \{F : L_m^2(F) < \infty\}$, where

$$L_m^2(F) = \sum_{\alpha_1+\alpha_2=m} \frac{m!}{\alpha_1!\alpha_2!} \int_0^1\int_0^1 \Big(\frac{\partial^m F}{\partial t^{\alpha_1}\partial x^{\alpha_2}}\Big)^2 dtdx. \tag{1.5}$$

The decay rate of the eigenvalues of the thin-plate splines reproducing kernel is of order k^{-m} Utreras (1988). The derivation of reproducing kernels for thin-plate splines requires some advanced knowledge of differential equations; details can be found in Duchon (1977) and Meinguet (1979) and references cited therein. However, the function

$$J\Big((t-x)^2 + (x-y)^2\Big),$$

where $J(x) = x^{2m-2}\log x$ acts like a reproducing kernel in our approach to the computation of thin-plate splines, and is called a semi-kernel. It follows from (1.3) that

$$C\Big((t,x);(s,y)\Big)$$
$$=\iiiint K^{1/2}\Big((t,x);(u,z_1)\Big)\ K^{1/2}\Big((s,y);(v,z_2)\Big)g\Big((u,z_1);(v,z_2)\Big)dudvdz_1dz_2,$$

where $g\Big((u,z_1);(v,z_2)\Big)$ is the joint density function of $(X(u),X(v))$ evaluated at (z_1,z_2). The kernel function C depends on both K and distribution of $(X(u),X(v))$ in a complicated way, so knowing the decay rates of the eigenvalues of K or of the covariance kernel of X does not determine the eigenvalue decay rate of C. One can investigate the behavior of the eigenvalues of C empirically.

The paper is organized as follows. Section 2 establishes the minimax lower bound for the rate of convergence of the excess risk. Section 3 develops a predictor using a roughness regularization method and shows that this predictor is rate-optimal. Section 4 conducts a Monte Carlo study to validate our method and we also illustrate it using two data examples. Section 5 discusses the software used in the computations. Some discussion is provided in Section 6. Proofs are in the Appendix.

2. Minimax Lower Bound

In this section, we establish the minimax lower bound for the rate of convergence of the excess risk.

We assume that the unknown F_0 resides in a reproducing kernel Hilbert space $\mathcal{H}(K)$ with a reproducing kernel K. Here $\mathcal{H}(K)$ is a linear functional space endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}(K)}$ such that

$$F(t,x)=\Big\langle K\Big((t,x);\cdot\Big),F\Big\rangle_{\mathcal{H}(K)},\quad \text{for any } F\in\mathcal{H}(K).$$

There is a one-to-one relationship between K and $\mathcal{H}(K)$. Assume that the kernel function C admits the spectral decomposition,

$$C\big((t,x);(s,y)\big)=\sum_{j=1}^{\infty}\rho_j\phi_j(t,x)\phi_j(s,y),$$

where the ρ_j are the positive eigenvalues with a decreasing order and the ϕ_j are the corresponding orthonormal eigenfunctions. We assume $\rho_k\asymp k^{-2r}$ for some constant $0<r<\infty$, where for two sequences $a_k,b_k>0$, $a_k\asymp b_k$ means that a_k/b_k is bounded away from zero and infinity as $k\to\infty$.

Theorem 1. *If the eigenvalues $\{\rho_k:k\geq 1\}$ of the kernel C in (1.3) satisfy $\rho_k\asymp k^{-2r}$ for some constant $0<r<\infty$. Then the excess prediction risk satisfies*

$$\lim_{c\to 0}\lim_{n\to\infty}\inf_{\tilde\eta}\sup_{F_0\in\mathcal{H}(K)}\mathbb{P}\Big(\mathfrak{R}_n\geq cn^{-2r/(2r+1)}\Big)=1, \tag{2.1}$$

where the infimum is taken over all possible predictors $\tilde{\eta}$ *based on* $\{(Y_i, X_i) : i = 1, \ldots, n\}$.

One can compare Theorem 1 with some known results in the literature when the functional linear regression model is the true model. In Example 1, we restrict the bivariate function F to the specific form $F(t,x) = \beta(t)x$, where β belongs to a reproducing kernel Hilbert space $\mathcal{H}(\tilde{K})$ with the reproducing kernel $\widetilde{K} : [0,1] \times [0,1] \to \mathbb{R}$. This essentially provides us a functional linear regression model. Here the eigenvalues of C have the same decay rate as the eigenvalues of $\widetilde{K}^{1/2} G \widetilde{K}^{1/2}$, where $G(u,v) = \mathrm{cov}(X(u), X(v))$ is the covariance function of X. This special setting coincides with those considered in Yuan and Cai (2010) and Cai and Yuan (2012). Similar results have been established earlier there.

3. A Roughness Regularized Estimate

In this section, we develop a predictor using a roughness regularization method and establish that it achieves the optimal rate established in Theorem 1.

3.1. Computation

Take the estimate $\widehat{F}_{n\lambda}$ of F_0 as the minimizer of the functional

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \int_0^1 F(t, X_i(t))dt\Big)^2 + \lambda J(F), \tag{3.1}$$

where λ is the tuning parameter and $J(\cdot)$ is a squared semi-norm on $\mathcal{H}(K)$. The estimate $\widehat{F}_{n\lambda}$ can be computed explicitly over the infinite dimensional function space $\mathcal{H}(K)$. This observation is important for both numerical implementation of our procedure and our asymptotic analysis.

Let $\mathcal{H}_0 = \{F \in \mathcal{H} : J(F) = 0\}$. Assume that $\{\xi_1, \ldots, \xi_N\}$ is the orthonormal basis of $\mathcal{H}_0$ with $N = \dim(\mathcal{H}_0) < \infty$. Let $\mathcal{H}_1$ be its orthogonal complement in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$.

Theorem 2. *The minimizer of* (3.1) *over* $\mathcal{H}(K)$ *can be written as*

$$\widehat{F}_{n\lambda}(t,x) = \sum_{j=1}^{N} d_j \xi_j(t,x) + \sum_{i=1}^{n} c_i \int_0^1 K\Big((t,x);(s, X_i(s))\Big)ds, \tag{3.2}$$

for some $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ *and* $d = (d_1, \ldots, d_N)^T \in \mathbb{R}^N$.

Denote by Σ the $n \times n$ matrix with $(\Sigma)_{ij} = \iint K\Big((t, X_j(t));(s, X_i(s))\Big)dtds$, and by Ξ the $n \times N$ matrix with $(\Xi)_{ij} = \int \xi_j(t, X_i(t))dt$. Then, (3.1) may be written in the matrix form

$$\frac{1}{n}\|Y - \Xi d - \Sigma c\|_2^2 + \lambda c^T \Sigma c, \tag{3.3}$$

where $J(F) = c^T \Sigma c$. It is easy to see that the solution of the linear system

$$(\Sigma + n\lambda I)c + \Xi d = Y, \tag{3.4}$$

$$\Xi^T \Sigma c + \Xi^T \Xi d = \Xi^T Y, \tag{3.5}$$

is a minimizer of (3.3). It follows from (3.4) and (3.5) that $\Xi^T c = 0$. Suppose Ξ is of full column rank. Let

$$\Xi = QR^* = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

be the QR-decomposition of Ξ with Q orthogonal and R upper-triangular. From $\Xi^T c = 0$, $Q_1^T c = 0$, so $c \perp \mathrm{col}(Q_1)$, the column space of Q_1. Since Q is orthogonal, $c \in \mathrm{col}(Q_2)$, and $c = Q_2 Q_2^T c$ because $Q_2 Q_2^T$ projects onto $\mathrm{col}(Q_2)$. Simple algebra gives

$$\widehat{c} = Q_2(Q_2^T \Sigma Q_2 + n\lambda I)^{-1} Q_2^T Y,$$
$$\widehat{d} = R^{-1}(Q_1^T Y - Q_1^T \Sigma c).$$

The linear system (3.4) and (3.5) as the basis for computation first appeared in Wahba and Wedelberger (1980) We have adopted a similar approach and applied the QR-decomposition to obtain an explicit solution of the optimization problem (3.3).

3.2. Rate of convergence

In this section, we turn to the asymptotic properties of the estimate $\widehat{F}_{n\lambda}$.

Theorem 3. *Assume that for any $F \in L_2([0,1]^2)$*

$$\mathbb{E}\Big(\int F(t, X(t))dt\Big)^4 \le c\Big(\mathbb{E}\Big(\int F(t, X(t))dt\Big)^2\Big)^2 \tag{3.6}$$

for a positive constant c. Then, when λ is of order $n^{-2r/(2r+1)}$,

$$\lim_{A\to\infty} \lim_{n\to\infty} \sup_{F_0 \in \mathcal{H}(K)} \mathbb{P}\Big\{\mathfrak{R}_n \ge A n^{-2r/(2r+1)}\Big\} = 0. \tag{3.7}$$

We have made the additional assumption (3.6) on X. For the functional linear regression model when $F(t,x) = \beta(t)x$, (3.6) shows that, for any $\beta \in L_2([0,1])$, $\mathbb{E}\big(\int \beta(t)X(t)dt\big)^4 \le c\Big(\mathbb{E}\big(\int \beta(t)X(t)dt\big)^2\Big)^2$, which states that linear functionals of X have bounded kurtosis. In general, (3.6) states that such special nonlinear functionals $F\big(\cdot, X(t)\big)$ of X have bounded kurtosis.

It follows from both Theorem 1 and Theorem 3 that the minimax rate of convergence for the excess prediction $\mathfrak{R}_n$ is of order $n^{-2r/(2r+1)}$, which is determined by the decay rate of the eigenvalues of the kernel C.

3.3. Optimal choice of λ

Let $\widehat{Y} = \Big(\eta_{\widehat{F}_\lambda}(X_1), \ldots, \eta_{\widehat{F}_\lambda}(X_n)\Big)^T$. Since the regularized estimator is a linear estimator in Y, $\widehat{Y} = H(\lambda)Y$, where $H(\lambda)$ is called the hat matrix depending on λ. Some algebra yields

$$H(\lambda) = I - n\lambda Q_2(Q_2^T \Sigma Q_2 + n\lambda I)^{-1} Q_2^T.$$

We select the tuning parameter λ that minimizes the generalized cross-validation score (Wahba (1990)),

$$\mathrm{GCV}(\lambda) = \frac{\|\widehat{Y} - Y\|_2^2/n}{\Big\{1 - \mathrm{tr}(H(\lambda))/n\Big\}^2}. \tag{3.8}$$

Choosing λ by minimizing GCV has worked well in our numerical studies.

4. Numerical Results

In our numerical studies, we compared the numerical performance of the proposed predictor with some well-known existing predictors.

For our estimator, we focus on a RKHS $\mathcal{H}(K)$ with a squared seminorm (1.5). In this setting, the optimal solution of the roughness-regularized estimate can be written as

$$F(t,x) = \sum_{j=1}^N d_j \xi_j(t,x) + \sum_{i=1}^n c_i \int J\Big(\sqrt{(t-s)^2 + (x - X_i(s))^2}\Big) ds, \tag{4.1}$$

where $\xi_j(t,x) = t^{\gamma_1} x^{\gamma_2}$ for some pair of integers γ_1, γ_2 with $0 \le \gamma_1 + \gamma_2 < m$ and N is the number of such pairs. Let $\hat{c}$ and $\hat{d}$ be the estimates from the training data. Then, for any random function X, the predicted response is

$$\eta_{\widehat{F}}(X) = \sum_{j=1}^N \hat{d}_j \int \xi_j(t, X(t)) dt + \sum_{i=1}^n \hat{c}_i \iint J\Big(\sqrt{(t-s)^2 + (X(t) - X_i(s))^2}\Big) dt ds.$$

In particular, when $m = 2$, we have $N = 3$, and $\xi_1(t,x) = 1$, $\xi_2(t,x) = t$, $\xi_3(t,x) = x$, and $J(x) = x^2 \log x$. Here $\int \xi_1(t, X(t)) dt = 1$ and $\int \xi_2(t, X(t)) dt = 1/2$. To avoid an identifiability problem, we estimate d_1 by $\hat{d}_1 = n^{-1} \sum_{i=1}^n Y_i$. In the following, we use thin-plate splines with $m = 2$ to fit the data.

Table 1. The root mean squared prediction errors (RMSPE) of three estimators for a functional linear regression model where $Y = \int_0^1 \beta_0(t)X(t)dt + \epsilon$. "FPCA/FLR" is an estimator for the functional linear model based on functional principal components analysis. "FGAM/P-spline" is the estimator in McLean et al. (2014). "ThinSpline" is our proposed estimator using a thin-plate spline.

ξ_j	σ	ν	FPCA/FLR	FGAM/P-Spline	ThinSpline
Well Spaced	0.5	1.1	0.61	0.82	0.68
		2.0	0.52	0.55	0.56
	1.0	1.1	1.21	1.65	1.20
		2.0	1.04	1.09	1.08
Closed Spaced	0.5	1.1	0.52	0.53	0.52
		2.0	0.54	0.55	0.56
	1.0	1.1	1.03	1.07	1.03
		2.0	1.06	1.05	1.04

4.1. Simulations

Our first simulation study compares our estimator with two competitors. The first of these competitors uses the functional principal component analysis (FPCA) approach to fitting a FLR; we call this estimator FPCA/FLR. The second method uses the P-spline approach in McLean et al. (2014) to fit an FGAM, where one estimates F using tensor-product B-splines with roughness penalties; we call this estimator FGAM/P-spline.

The simulation setting is the same as in Hall and Horowitz (2007) and McLean et al. (2014). The random predictor function X was generated as

$$X(t) = \zeta_1 Z_1 + \sum_{k=2}^{50} \sqrt{2}\, \zeta_k Z_k \cos(k\pi t), \quad t \in [0,1],$$

the Z_k independently uniform on $[-\sqrt{3}, \sqrt{3}]$. The ζ_k^2, eigenvalues of the covariance function of X, were well spaced case, $\zeta_k = (-1)^{k+1}k^{-\nu/2}$ with $\nu = 1.1$ and 2, or closely spaced case, $\zeta_1 = 1$, $\zeta_j = 0.2(-1)^{j+1}(1 - 0.0001j)$ for $j = 2,3,4$, and $\zeta_{5j+k} = 0.2(-1)^{5j+k+1}(5j)^{-\nu/2} - 0.0001k$ for $j \geq 1$ and $0 \leq k \leq 4$. The coefficient function β_0 was

$$\beta_0(t) = 0.3 + \sum_{k=2}^{50} 4\sqrt{2}(-1)^{k+1}k^{-2} \cos(k\pi t), \quad t \in [0,1].$$

The simulation study had the FLR model as the true model. The response variable Y was simulated from the model: $Y = \int_0^1 \beta_0(t)X(t)dt + \epsilon$, where the error $\epsilon \sim N(0, \sigma^2)$, with $\sigma = 0.5$ and 1. The performance of different estimators was measured by the root mean squared prediction error, RMSPE =

Table 2. The root mean squared prediction errors (RMSPE) based on three different estimators for two nonlinear functional regression models. PCF is the piecewise constant fit of Müller, Wu, and Yao (2012).

Model	σ	FPCA/FLR	PCF/CAM	ThinSpline
$Y=\int_0^{10}\cos\{t-X(t)-5\}dt+\epsilon$	2	2.434 (0.018)	2.200 (0.056)	2.108 (0.062)
	1	1.723 (0.013)	1.156 (0.037)	1.127 (0.035)
	0.5	1.494 (0.011)	0.680 (0.035)	0.569 (0.026)
$Y=\int_0^{10} t\exp\{X(t)\}dt+\epsilon$	1	9.828 (0.106)	1.119 (0.029)	1.108 (0.031)

$\sqrt{d^{-1}\sum_{i=1}^{d}\left(\widehat{Y}_i-Y_i\right)^2}$, where d is the sample size of the test data and the $\widehat{Y}_i$ are predicted values. Each training set contained 67 curves and 33 curves were used for the test set. For each setting, the experiment was repeated 1,000 times. The results of simulations are summarized in Table 1. We observe that our ThinSpline estimator performed nearly identically to the FPCA/FLR estimator, even though this is an ideal setting for the latter since the FLR model holds. Our ThinSpline estimator slightly outperformed the FGAM/P-spline estimator.

We performed a simulation study to compare our estimate with the piecewise constant fit (PCF) proposed in Müller, Wu, and Yao (2012) for fitting what they call the continuously additive model (CAM); we denote this estimator by PCF/CAM. The simulation setting was the same as that in Müller, Wu, and Yao (2012). The predictor functions were generated according to

$$X(t)=\cos(U_1)\sin(\frac{1}{5}\pi t)+\sin(U_1)\cos(\frac{1}{5}\pi t)+\cos(U_2)\sin(\frac{2}{5}\pi t)+\sin(U_2)\cos(\frac{2}{5}\pi t)$$

for $t\in[0,10]$, where U_1 and U_2 were i.i.d. from Uniform$[0,2\pi]$. The sample size for the training data was $n=200$ and for the testing data was $d=$1,000. The data were generated from $Y=\int_0^{10}\cos\{t-X(t)-5\}dt+\epsilon$; or $Y=\int_0^{10} t\exp\{X(t)\}dt+\epsilon$, where $\epsilon\sim N(0,\sigma^2)$. For each setting, the experiment was repeated 50 times. The means and the corresponding standard deviation of the root mean squared prediction error are given in Table 2. As expected, the FPCA/FLR performs poorly for these two non-FLR examples as it has large prediction errors. In addition, our ThinSpline estimator outperforms the PCF/CAM estimator proposed in Müller, Wu, and Yao (2012). An additional tuning data set with sample size 200 was used to select the needed regularization parameter in the original simulation of PCF/CAM by Müller, Wu, and Yao (2012). A benefit of our approach is that we do not require this tuning data set in our simulations.

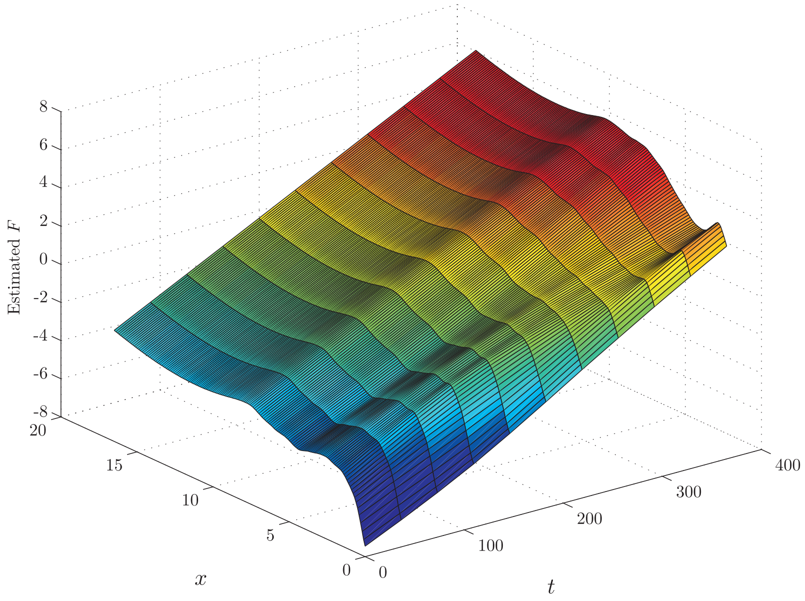

Figure 1. Estimated surface $\widehat{F}_{n\lambda}(t,x)$ from the Canadian weather data.

Table 3. The root mean squared prediction errors based on the estimate (4.2), FGAM/P-spline, and the proposed predictor for Canadian weather data.

	SS/FLR	FGAM/P-spline	ThinSpline
RMSPE	0.3014(0.1244)	0.1127 (0.1002)	0.1110(0.0917)

4.2. Application: Canadian weather data

The Canadian weather data example is revisited here. The dataset contains daily temperature and precipitation at 35 different locations in Canada averaged over years 1960 to 1994. The goal is to predict the log of the average annual precipitation based on the average daily temperature. In Cai and Yuan (2012) it was shown that the functional PCA approach to fitting a FLR could be problematic, since the eigenfunctions corresponding to the leading eigenvalues of the covariance function seem not to represent the estimated coefficient function well. We compare our method with a smoothing spline estimate for fitting the functional linear regression model. Here the estimate, which we call SS/FLR (smoothing spline, functional linear regression), is

$$(\hat{\alpha},\hat{\beta}) = \arg\min\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\alpha-\int_0^1 X_i(t)\beta(t)dt\right)^2+\lambda\int_0^1(\beta''(t))^2dt\right\}. \quad (4.2)$$

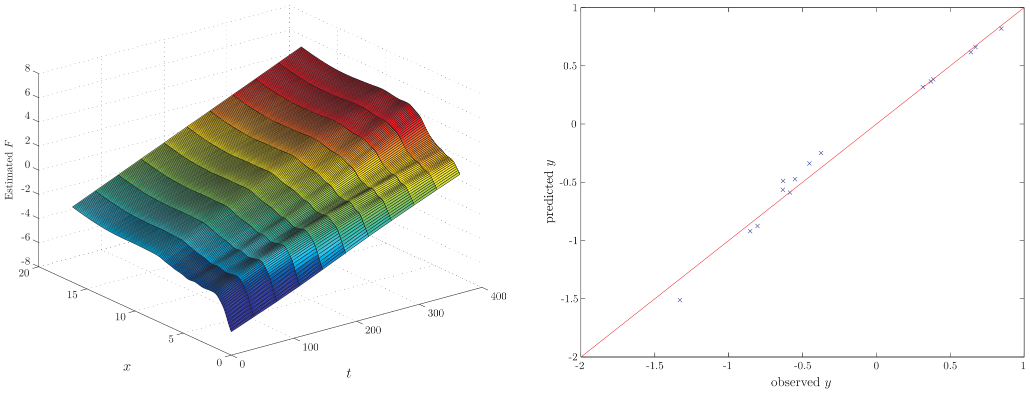

Figure 2. Left: Estimated surface $\widehat{F}_\lambda(t,x)$ from the training data using ThinSpline; Right: the predicted response using ThinSpline versus the observed response for the testing data.

Figure 1 shows the estimated $\widehat{F}_{n\lambda}$ using ThinSpline applied to the complete data. In order to study the performance of these estimators, we randomly split the initial sample into two sub-samples: learning sample, (X_i, Y_i), $i = 1, \ldots, n_\ell$ with $n_\ell = 20$, was used to determine the estimated coefficient function $\hat{\beta}_\lambda$ and the estimator $\widehat{F}_{n\lambda}$ and test sample, (X_i, Y_i), $i = n_\ell + 1, \ldots, n$, with $n - n_\ell = 15$, used to evaluate the quality of the estimation. The left panel of Figure 2 displays the estimated $\widehat{F}_{n\lambda}$ from applying ThinSpline to the training data set and the right panel of Figure 2 shows the predicted response from ThinSpline versus the observed response for the testing data using the estimate from the training data. The points are very close to the diagonal line which indicates a good fit. We have repeated this procedure 200 times. The mean and the corresponding standard deviations of the root mean squared prediction errors based on (4.2) and our proposed predictor are reported in Table 3.

The prediction error using either FGAM/P-spline or ThinSpline is considerably less than for FPCA/FLR. The FGAM/P-spline and ThinSpline estimators have similar prediction errors, as expected. Further study of the goodness-of-fit of different models is an important research topic that we will pursue.

4.3. Application: CA air quality data

Air pollutants cause serious health problems, and modeling different ground level air pollutants has been important research for many years. In May 2011, the California Air Resources Board released the "2011 Air Quality Data", which include 30 years of data (1980–2009). This database, available at `http://www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm`, contains hourly concentrations of pollutants at different locations in California from 1980 to 2009. We focused on the

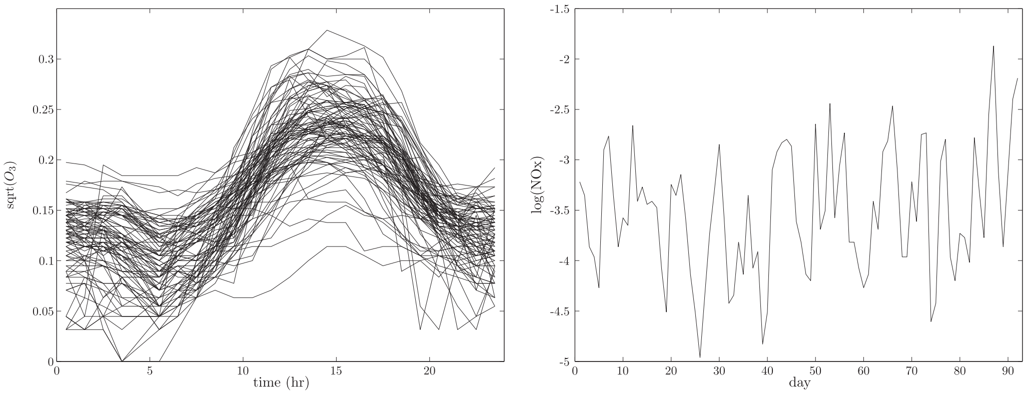

Figure 3. Left: Daily trajectories of ground-level concentrations of ozone in the city of Sacramento in the Summer of 2005; Right: The maximum level of the ground-level concentrations of oxides of nitrogen at each day in the Summer of 2005.

effect of the trajectories of ozone (O3) on the maximum level of oxides of nitrogen (NOx) in the city of Sacramento (site 3011 in the database) between June 1 and August 31 of 2005. The total sample size is $n = 92$. The left panel of Figure 3 displays the daily trajectories of ground-level concentrations of ozone in the city of Sacramento in the summer of 2005. For most days, we have observations at every hour but there are a few days with some missing observations. The right panel of Figure 3 gives the maximum level of the ground-level concentrations of oxides of nitrogen at each day during the summer of 2005 in Sacramento.

Figure 4 shows the estimated $\hat{F}_{n\lambda}$ when using the complete data. It displays a highly nonlinear pattern in x, which suggests that the functional linear model does not fit the data well. To assess the goodness of fit of the additive functional model, the left panel of Figure 5 plots the residuals on the vertical axis and the fitted responses on the horizontal axis. It shows the points are randomly dispersed around the horizontal axis and do not show any typical pattern. The right panel of Figure 5 plots the fitted values versus the observed responses. The points are close to the diagonal line indicating a good fit.

Another interest is to compare the estimation surface. We compared our estimate of F with the estimated F from McLean et al. (2014) that is shown in Figure 6. When we compare the prediction errors of these two estimates, they are almost the same. Estimating F is a different question than the prediction issue discussed here. Thus, for the functional linear regression model, Crambes, Kneip, and Sarda (2009) pointed out that we may not be able to consistently estimate the slope function β without linking the smoothness of β and of the curves X_i. However, in terms of prediction, we can have the consistent result. We believe

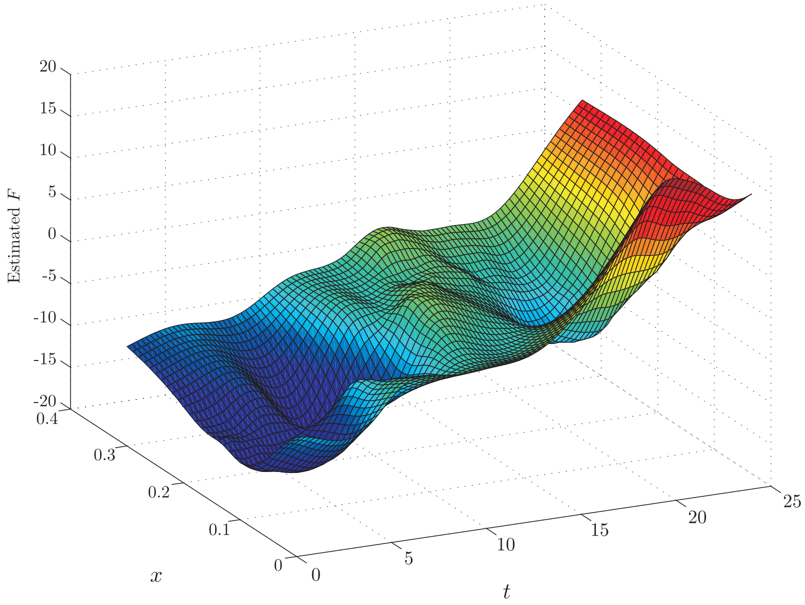

Figure 4. Estimated surface $\widehat{F}_{n\lambda}(t,x)$ from the air quality data.

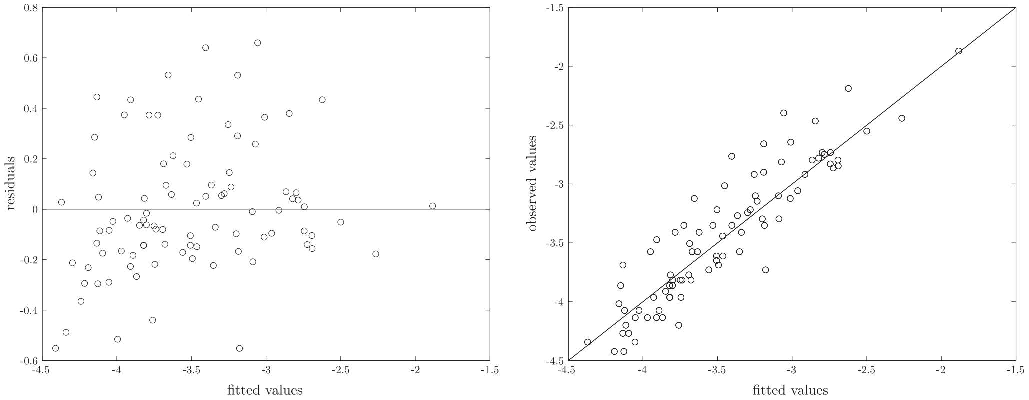

Figure 5. Left: Residual plot; Right: Fitted values versus the observed responses.

that this happens here and we cannot estimate F consistently without additional assumptions on the connection of the smoothness of F and the distribution of $(X(t), X(s))$. Further research is needed to work out the asymptotic distribution of the prediction error for making statistical inferences.

We also compared the performance of the smoothing spline FLR (4.2), FGAM/P-spline, and ThinSpline estimators. The 92 observations were randomly split into training sets of size 60 and test sets of size 32. We repeated this procedure 1,000 times. The mean and the corresponding standard deviations of

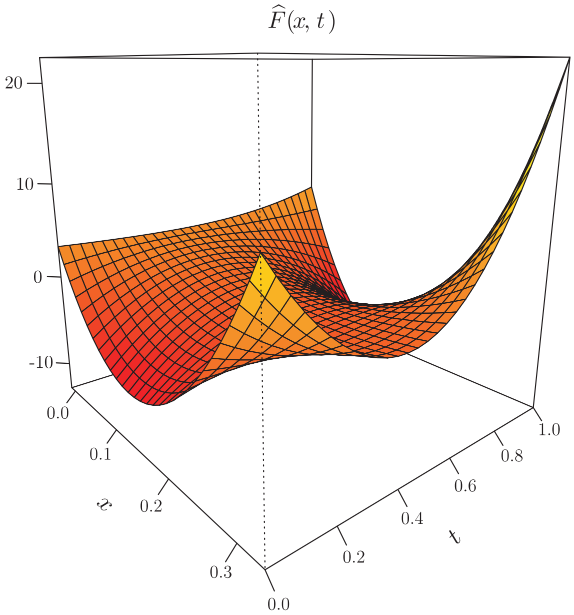

Figure 6. Estimated surface F using P-splines from McLean et al. (2014).

Table 4. The root mean squared prediction errors based on the functional linear regression (FLR) model, the P-splines model, and the additive functional model (ThinSpline) for the air quality data.

	SS/FLR	FGAM/P-spline	ThinSpline
RMSPE	0.9450 (1.6539)	0.6203 (0.0937)	0.6148 (0.0985)

the root mean squared prediction error based on the two models are reported in Table 4. As expected, our ThinSpline estimator outperforms the smoothing spline FLR estimation and displays a similar performance as the FGAM/P-spline estimator.

5. A Note on Computations

The FGAM/P-spline estimator was computed using the `fgam()` function in R's `refund` package. We wrote our own Matlab programs to compute the smoothing spline FLR and ThinSpline estimators. The FPCA/FLR estimators were computed using the software called PACE `http://www.stat.ucdavis.edu/PACE/`.

6. Discussion

We have established the minimax rate of convergence for prediction in the continuous-additive functional regression model. The minimax theory in the existing literature on the functional linear regression model is a special setting of the current study.

We have focused on the continuously-additive functional model with the squared error loss in this paper. The method of regularization can be easily extended to handle other models such as the generalized regression model (Cardot and Sarda (2005); Du and Wang (2013)). We leave these extensions for future papers.

In our simulation study, the only estimator based on the RKHS approach used thin-plate splines. For the case of univariate regression, Wang, Shen, and Ruppert (2011) showed that a smoothing spline and a P-spline are asymptotically equivalent. A similar asymptotic equivalence is expected to hold for bivariate regression. So, it was expected that in simulations, the performance of the ThinSpline estimator would be similar to that of the FGAM/P-spline estimator of McLean et al. (2014). Our results can be applied to more general reproducing kernel Hilbert spaces.

As estimating F_0 itself is a different problem than the prediction problem discussed here. For the functional linear regression model we may not estimate the coefficient function β_0 consistently without additional conditions linking the smoothness of β_0 and of the curves X_i (Crambes, Kneip, and Sarda (2009)). One might assume, for example, that the reproducing kernel K and the covariance kernel G share the same set of eigenfunctions. Under this assumption, we can estimate β_0 consistently (Yuan and Cai (2010)). The question of when we can estimate F_0 consistently under the continuously-additive functional model deserves study. The issue is important, for example, to developing a test of linearity of F_0.

Acknowledgement

We thank two referees, Associate Editor, and the Editor for constructive and insightful comments. Xiao Wang's research is supported by US NSF grants CMMI-1030246 and DMS-1042967.

Appendices

A.1. Proof of Theorem 1

In the proofs, c_i, $i = 1, 2, \ldots$ are generic constants that can change from line to line.

Since any lower bound for a specific case yields a lower bound for the general case, to establish lower bounds, we only study the case when the ϵ_i are i.i.d. $N(0, \sigma^2)$. Fix $\alpha \in (0, 1/8)$. From Theorem 2.5 in Tsybakov (2009), in order to establish the minimax lower bound for $\mathfrak{R}_n$, for each n we need to find functions $\{F_{jn},\, j = 0, \ldots, M\}$ satisfying the following:

(a) $F_{jn} \in \mathcal{H}(K)$, $j = 0, \ldots, M$,

(b) $\mathbb{E}^*\Big\{ \int_0^1 \Big[F_{jn}(t, X_{n+1}(t)) - F_{kn}(t, X_{n+1}(t))\Big] dt\Big\}^2 \geq 2s$, for $0 \leq j < k \leq M$,

(c) $M^{-1} \sum_{j=1}^{M} \mathcal{K}(P_j, P_0) \leq \alpha \log M$, where P_j denotes the joint distribution of $\{(Y_i, X_i) : i = 1, \ldots n\}$ when $F_0 = F_{jn}$, and $\mathcal{K}(\cdot, \cdot)$ is the Kullback-Leibler distance between two probability measures.

We will specify $M \to \infty$ and $s \to 0$ later. If (a), (b), and (c) are satisfied, then the minimax lower bound for the rate of convergence of $\mathfrak{R}_n$ has the same order as s.

First we verify part (a). Let m be the smallest integer greater than $c_0 n^{1/(2r+1)}$ for some positive constant c_0 to be specific later. For a $\omega = (\omega_{m+1}, \ldots, \omega_{2m}) \in \{0, 1\}^m$, let

$$F_\omega = \sum_{j=m+1}^{2m} \omega_j m^{-1/2} K^{1/2}(\phi_j).$$

$F_\omega \in \mathcal{H}(K)$ for all ω if $K^{1/2}(\phi_j) \in \mathcal{H}(K)$ for all j. Thus, we need to show that $\big\langle K^{1/2}(\phi_j), K\big(\cdot, (t, x)\big)\big\rangle_{L_2} = K^{1/2}(\phi_j)(t, x)$. This holds since

$$\begin{aligned}\big\langle K^{1/2}(\phi_j), K\big(\cdot, (t, x)\big)\big\rangle_{L_2} &= \big\langle K(\phi_j), K^{1/2}\big(\cdot, (t, x)\big)\big\rangle_{L_2} \\ &= \big\langle \phi_j, K^{1/2}\big(\cdot, (t, x)\big)\big\rangle_{L_2} = K^{1/2}(\phi_j)(t, x).\end{aligned}$$

We also have

$$\langle K^{1/2}(\phi_j), K^{1/2}(\phi_k)\rangle_{\mathcal{H}(K)} = \langle \phi_j, K(\phi_k)\rangle_{\mathcal{H}(K)} = \langle \phi_j, \phi_k\rangle_{L_2} = \delta_{jk},$$

where $\delta_{jk} = 1$ for $j = k$, and 0 for $j \neq k$.

The Varshamov-Gilbert bound (see Tsybakov (2009, p.104)) shows that, for $m \geq 8$, there exists a subset $\Omega = \{\omega^0, \omega^1, \ldots, \omega^M\} \subseteq \{0, 1\}^m$ such that $\omega^0 = \{0, \ldots, 0\}$,

$$d(\omega^j, \omega^k) \geq \frac{m}{8}, \quad \forall\, 0 \leq j < k \leq M, \tag{A.1}$$

where $d(\omega^j, \omega^k) = \sum_{i=m+1}^{2m} I(\omega_i^j \neq \omega_i^k)$ is the Hamming distance between ω^j and ω^k, and $M \geq 2^{m/8}$.

To verify part (b), for $\omega, \omega' \in \Omega$, direct calculation yields that

$$\mathbb{E}^*\Big\{ \int_0^1 \Big[F_\omega(t, X_{n+1}(t)) - F_{\omega'}(t, X_{n+1}(t))\Big] dt\Big\}^2$$

$$
\begin{aligned}
&= \sum_{j=m+1}^{2m} \sum_{k=m+1}^{2m} m^{-1}(\omega_j - \omega_j')(\omega_k - \omega_k') \\
&\qquad\qquad \iint \mathbb{E}^* \Big[K^{1/2}(\phi_j)(t, X(t)) K^{1/2}(\phi_k)(s, X(s)) \Big] dt ds \\
&= \sum_{k=m+1}^{2m} m^{-1}(\omega_k - \omega_k')^2 \rho_k \geq \ m^{-1} \rho_{2m} d(\omega, \omega') \\
&\geq c_1 m^{-1}(2m)^{-2r} \frac{m}{8} \geq c_2 n^{-2r/(2r+1)}
\end{aligned}
$$

by (A.1), $\rho_k \asymp k^{-2r}$, and the definition of m. Hence, s in part (b) is of order $n^{-2r/(2r+1)}$.

Next, observe that for any $\omega, \omega' \in \Omega$,

$$
\begin{aligned}
\log\Big(\frac{P_{F_{\omega'}}}{P_{F_\omega}}\Big) =& \frac{1}{\sigma^2} \sum_{i=1}^{n} \Big(Y_i - \int F_\omega(t, X(t)) dt \Big) \int \Big\{ F_\omega(t, X(t)) - F_{\omega'}(t, X(t)) \Big\} dt \\
&- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big[\int \Big\{ F_\omega(t, X(t)) - F_{\omega'}(t, X(t)) \Big\} dt \Big]^2 .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathcal{K}(P_{F_{\omega'}}, P_{F_\omega}) &= \frac{n}{2\sigma^2} \mathbb{E}^* \Big[\int \Big\{ F_\omega(t, X(t)) - F_{\omega'}(t, X(t)) \Big\} dt \Big]^2 \\
&= \frac{n}{2\sigma^2} \sum_{k=m+1}^{2m} m^{-1}(\omega_k - \omega_k')^2 \rho_k \\
&\leq \frac{n}{2\sigma^2} \rho_m \sum_{k=m+1}^{2m} m^{-1}(\omega_k - \omega_k')^2 \leq \frac{n}{2\sigma^2} m^{-2r} \leq c_3 n^{1/(2r+1)} .
\end{aligned}
$$

Since m is the smallest integer greater than $c_0 n^{1/(2r+1)}$, this implies that

$$
\frac{1}{M} \sum_{j=1}^{M} \mathcal{K}(P_j, P_0) \leq c_3 n^{1/(2r+1)} \leq \alpha \log M,
$$

if we choose $c_0 \geq 8c_3/(\alpha \log 2)$ and $M = 2^{m/8}$. This completes the proof of Theorem 1.

A.2. Proofs of Theorem 2 and Theorem 3

Proof of Theorem 2. Define the subspace of $\mathcal{H}$,

$$
\overline{\mathcal{H}}_1 = \text{span}\Big\{ \int K\Big((t, x); (s, X_i(s))\Big) ds : i = 1, \ldots, n \Big\}.
$$

Note that $\overline{\mathcal{H}}_1$ is a closed linear subspace of $\mathcal{H}_1$. For any $F \in \mathcal{H}$, one may write $F = F_0 + F_1 + \delta$, where $F_0 \in \mathcal{H}_0$, $F_1 \in \overline{\mathcal{H}}_1$ and $\delta \in \mathcal{H}_1 \ominus \overline{\mathcal{H}}_1$. Observe that

$$\eta_F(X_i) = \int F(t, X_i(t))dt = \eta_{F_0+F_1}(X_i),$$

because

$$\eta_\delta(X_i) = \Big\langle \int K\Big((\cdot; (s, X_i(s))\Big)ds, \delta\Big\rangle_{\mathcal{H}(\mathcal{K})} = 0.$$

Further, due to orthogonality, $\|F\|^2_{\mathcal{H}(\mathcal{K})} = \|F_0 + F_1\|^2_{\mathcal{H}(\mathcal{K})} + \|\delta\|^2_{\mathcal{H}(\mathcal{K})}$ and $\|F_0 + F_1\|^2_{\mathcal{H}(\mathcal{K})} \le \|F\|^2_{\mathcal{H}(\mathcal{K})}$. Therefore, the minimum of (3.1) must belong to the linear space $\mathcal{H}_0 \oplus \overline{\mathcal{H}}_1$.

Proof of Theorem 3. Note that $L_2(K^{1/2}) = \mathcal{H}(K)$. So there exist G_0 and $\hat{G}_\lambda$ such that $F_0 = K^{1/2}(G_0)$ and $\widehat{F}_{n\lambda} = K^{1/2}(\hat{G}_\lambda)$. Therefore,

$$\begin{aligned}\eta_{F_0}(X) &= \int F_0(t, X(t))dt = \int \Big\langle K\Big(\cdot; (s, X(s)\Big), F_0\Big\rangle_{\mathcal{H}(\mathcal{K})} ds \\ &= \int \Big\langle K^{1/2}\Big(\cdot; (s, X(s))\Big), G_0\Big\rangle_{L_2} ds,\end{aligned}$$

$$\mathfrak{R}_n = \mathbb{E}^*\Big| \int \Big\langle K^{1/2}\Big(\cdot; (s, X(s))\Big), \hat{G}_\lambda - G_0\Big\rangle_{L_2} ds\Big|^2 = \Big\|\hat{G}_\lambda - G_0\Big\|^2_C,$$

where

$$\Big\|G\Big\|^2_C = \int \cdots \int G\Big((t,x); (u_1, z_1)\Big) C\Big((u_1, z_1); (u_2, z_2)\Big) G\Big((u_2, z_2); (s, y)\Big).$$

Write

$$C_n\big((t,x); (s,y)\big) = \frac{1}{n}\sum_{i=1}^n \iint K^{1/2}\Big((t,x); (u, X_i(u))\Big) K^{1/2}\Big((s,y); (v, X_i(v))\Big) du dv.$$

Recall that $Y_i = \int \Big\langle K^{1/2}\Big(\cdot; (s, X_i(s))\Big), G_0\Big\rangle ds + \epsilon_i$. Write $g_n = (1/n)\sum_{i=1}^n \epsilon_i \int K^{1/2}\Big(\cdot; (s, X(s))\Big)ds$, $\hat{G}_\lambda = \Big(C_n + \lambda I\Big)^{-1}\Big(C_n(G_0) + g_n\Big)$. Define $G_\lambda = \Big(C + \lambda I\Big)^{-1} C(G_0)$. It follows from triangle inequality that

$$\Big\|\hat{G}_\lambda - G_0\Big\|_C \le \Big\|G_\lambda - G_0\Big\|_C + \Big\|\hat{G}_\lambda - G_\lambda\Big\|_C. \tag{A.2}$$

For the first term on the right hand side of (A.2), write $G_0 = \sum_{k=1}^\infty a_k \phi_k$. Then,

$$\begin{aligned}G_\lambda &= \sum_{k=1}^\infty \frac{a_k \rho_k}{\lambda + \rho_k}\phi_k, \\ \Big\|G_\lambda - G_0\Big\|^2_C &= \sum_{k=1}^\infty \frac{\lambda^2 a_k^2 \rho_k}{(\lambda + \rho_k)^2} \le \lambda^2 \max_{k\ge 1} \frac{\rho_k}{(\lambda + \rho_k)^2} \sum_{k=1}^\infty a_k^2 = O(\lambda)\Big\|G_0\Big\|^2_{L_2}.\end{aligned}$$

For the second term on the right hand side of (A.2), with $\Big(C_n + \lambda I\Big)\hat{G}_\lambda = C_n(G_0) + g_n$, We have

$$\begin{aligned} G_\lambda - \hat{G}_\lambda &= (C+\lambda I)^{-1}(C_n+\lambda I)(G_\lambda - \hat{G}_\lambda) + (C+\lambda I)^{-1}(C - C_n)(G_\lambda - \hat{G}_\lambda) \\ &= (C+\lambda I)^{-1}(C_n+\lambda I)G_\lambda - (C+\lambda I)^{-1}C_nG_0 - (C+\lambda I)^{-1}g_n \\ &\qquad + (C+\lambda I)^{-1}(C-C_n)(G_\lambda - \hat{G}_\lambda) \\ &= (C+\lambda I)^{-1}C_n(G_\lambda - G_0) + \lambda(C+\lambda I)^{-2}CG_0 - (C+\lambda I)^{-1}g_n \\ &\qquad + (C+\lambda I)^{-1}(C-C_n)(G_\lambda - \hat{G}_\lambda) \\ &= (C+\lambda I)^{-1}C(G_\lambda - G_0) + \lambda(C+\lambda I)^{-2}CG_0 - (C+\lambda I)^{-1}g_n \\ &\qquad + (C+\lambda I)^{-1}(C_n - C)(G_\lambda - G_0) \\ &\qquad + (C+\lambda I)^{-1}(C-C_n)(G_\lambda - \hat{G}_\lambda) \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}. \end{aligned}$$

We bound the five terms on the right side separately. Direct calculation yields

$$\begin{aligned} \Big\|\mathrm{I}\Big\|_C^2 = \Big\|(C+\lambda I)^{-1}C(G_\lambda - G_0)\Big\|_C^2 &= \lambda^2 \sum_{k=1}^{\infty} \frac{a_k^2\rho_k^3}{(\lambda+\rho_k)^4} \\ &\le \lambda^2 \max_{k\ge 1} \frac{\rho_k^3}{(\lambda+\rho_k)^4} \sum_{k=1}^{\infty} a_k^2 = O(\lambda)\Big\|G_0\Big\|_{L_2}^2. \end{aligned}$$

Similarly,

$$\Big\|\mathrm{II}\Big\|_C^2 = \Big\|\lambda(C+\lambda I)^{-2}CG_0\Big\|_C^2 = \lambda^2 \sum_{k=1}^{\infty} \frac{a_k^2\rho_k^3}{(\lambda+\rho_k)^4} \le O(\lambda)\Big\|G_0\Big\|_{L_2}^2.$$

We make use three of auxiliary results whose proofs are similar to ones in Cai and Yuan (2012), so we omit the details. If there exists a constant $c > 0$ such that

$$\mathbb{E}\Big(\int F(t, X(t))dt\Big)^4 \le c\Big(\mathbb{E}\Big(\int F(t, X(t))dt\Big)^2\Big)^2,$$

for any $\nu > 0$ such that $2r(1-2\nu) > 1$, then

$$\Big\|C^\nu(C+\lambda I)^{-1}(C-C_n)C^{-\nu}\Big\|_{op} = O_p\Big(\Big(n\lambda^{1-2\nu+1/(2r)}\Big)^{-1/2}\Big), \qquad \text{(A.3)}$$

$$\Big\|C^{1/2}(C+\lambda I)^{-1}(C-C_n)C^{-\nu}\Big\|_{op} = O_p\Big(\Big(n\lambda^{1/(2r)}\Big)^{-1/2}\Big), \qquad \text{(A.4)}$$

where $\|\cdot\|_{op}$ stands for the usual operator norm. Further, for any $0 \le \nu \le 1/2$

$$\Big\|C^\nu(C+\lambda I)^{-1}g_n\Big\|_{L_2} = O_p\Big(\Big(n\lambda^{1-2\nu+1/(2r)}\Big)^{-1/2}\Big). \qquad \text{(A.5)}$$

Using (A.3) we have

$$\begin{aligned}
&\Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2}^2\\
&\qquad\le \Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)C^{-\nu}\Big\|_{\mathrm{op}}\ \Big\|C^{\nu}(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2}^2\\
&\qquad\le o_p(1)\Big\|C^{\nu}(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2}^2,
\end{aligned}$$

whenever $\lambda \ge cn^{-2r/(2r+1)}$ for some constant $c>0$. Similarly,

$$\begin{aligned}
&\Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)(G_\lambda-G_0)\Big\|_{L_2}^2\\
&\qquad\le \Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)C^{-\nu}\Big\|_{\mathrm{op}}\ \Big\|C^{\nu}(G_\lambda-G_0)\Big\|_{L_2}^2\\
&\qquad\le o_p(1)\Big\|C^{\nu}(G_\lambda-G_0)\Big\|_{L_2}^2.
\end{aligned}$$

So, for $0<\nu<1/2-1/(4r)$,

$$\begin{aligned}
\Big\|C^{\nu}(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2} \le{}& \Big\|C^{\nu}(C+\lambda I)^{-1}C(G_\lambda-G_0)\Big\|_{L_2}\\
&+\Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)(G_\lambda-G_0)\Big\|_{L_2}\\
&+\lambda\|C^{1+\nu}G_0\|_{L_2}+\|C^{\nu}(C+\lambda I)^{-1}g_n\|_{L_2}\\
&+\Big\|C^{\nu}(C+\lambda I)^{-1}(C-C_n)(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2}\\
={}& O_p\Big(\lambda^{\nu}+\Big(n\lambda^{1-2\nu+1/(2r)}\Big)^{-1/2}\Big)=O_p(\lambda^{\nu}),
\end{aligned}$$

when $c_1n^{-2r/(1+2r)}\le\lambda\le c_2n^{-2r/(1+2r)}$ for $0<c_1<c_2<\infty$. Next,

$$\begin{aligned}
\big\|\mathrm{IV}\big\|_C &= \Big\|(C+\lambda I)^{-1}(C_n-C)(G_\lambda-G_0)\Big\|_C\\
&= \Big\|C^{1/2}(C+\lambda I)^{-1}(C_n-C)(G_\lambda-G_0)\Big\|_{L_2}\\
&\le \Big\|C^{1/2}(C+\lambda I)^{-1}(C_n-C)C^{-\nu}\Big\|\|T^{\nu}(G_\lambda-G_0)\|_{L_2}\\
&\le O_p\Big((n\lambda^{1/(2r)})^{-1/2}\lambda^{\nu}\Big)=o_p\Big((n\lambda^{1/(2r)})^{-1/2}\Big).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\big\|\mathrm{V}\big\|_C &= \Big\|(C+\lambda I)^{-1}(C_n-C)(G_\lambda-\hat{G}_\lambda)\Big\|_C\\
&= \Big\|C^{1/2}(C+\lambda I)^{-1}(C_n-C)(G_\lambda-\hat{G}_\lambda)\Big\|_{L_2}
\end{aligned}$$

$$\leq \Big\|C^{1/2}(C+\lambda I)^{-1}(C_n - C)C^{-\nu}\Big\| \|T^{\nu}(G_\lambda - \hat{G}_\lambda)\|_{L_2} \leq O_p\Big((n\lambda^{1/(2r)})^{-1/2}\lambda^{\nu}\Big)$$
$$= o_p\Big((n\lambda^{1/(2r)})^{-1/2}\Big).$$

It follows from (A.5) that

$$\big\|\mathrm{III}\big\|_C = \Big\|(C+\lambda I)^{-1}g_n\Big\|_C = \Big\|C^{1/2}(C+\lambda I)^{-1}g_n\Big\|_{L_2} = O_p\Big((n\lambda^{1/(2r)})^{-1/2}\Big).$$

Combining, we conclude that, if λ is of order $n^{-2r/(2r+1)}$, then $\|G_\lambda - \hat{G}_\lambda\|_C = O_P(n^{-2r/(2r+1)})$.

References

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2158-2179.

Cai, T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. of Amer. Statist. Assoc.* **107**, 1201-1216.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92**, 24-41

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

Du, P. and Wang, X. (2013). Penalized functional linear regression. *Statist. Sinica* **24**, 1017-1041.

Duchon, J. (1977). Spline minimizing rotation-invariate semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables* (Edited by W. Schemp and K. Zeller), 85-100. Springer-Verlag, Berlins.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statist. Sci.* **11**, 89-121.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations.* Springer, New York.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statis.* **35**, 70-91.

James, G. (2002). Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64**, 411-432.

Li, Y. and Hsing, T. (2007). On the rate of convergence in functional linear regression. *J. Multivariate Anal.* **98**, 1782-1804.

McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F. and Ruppert, D. (2014). Functional generalized additive models. *J. Comput. Graph. Statist.* **23**, 249-269.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *J. Appl. Math. Phys.* (ZIMP) **30**, 292-304.

Müller, H. G., Wu, Y. and Yao, F. (2012). Continuously additive models for nonlinear functional regression. *Biometrika*, to appear.

Müller, H. G. and Yao, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103**, 1534-1544.

Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and Matlab*. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer, New York.

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

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

Utreras, F. (1988). Convergence rates for multivariate smoothing spline functions. *J. Approx. Theory* **52**, 1-27.

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

Wahba, G. and Wedelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Month. Weather Rev.* **108**, 1122-1145.

Wang, X., Shen, J. and Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic J. Statist.* **5**, 1-17.

Yao, F., Müller, H. and Wang, J. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-2903.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412-3444.

Department of Statistics, Purdue University, West Lafayette, IN 47907-2066, USA.

E-mail: wangxiao@purdue.edu

Department of Statistical Science and School of Operations Research and Information Engineering, Cornell University, Comstock Hall, Ithaca, NY 14853, USA.

E-mail: dr24@cornell.edu

(Received April 2013; accepted March 2014)