The Canadian Journal of Statistics

Vol. xx, No. yy, 2021, Pages 1–52

La revue canadienne de statistique

Reproducing kernel-based functional linear expectile regression

Meichen LIU[1], Matthew PIETROSANU[1], Peng LIU[2], Bei JIANG[1], Xingcai ZHOU[3] and Linglong KONG[1]*

[1]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*

[2]*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, United Kingdom CT2 7FS*

[3]*School of Statistics and Data Science, Nanjing Audit University, Nanjing, Jiangsu, China 211085*

Key words and phrases: Expectile regression; functional data analysis; heteroscedasticity; reproducing kernel Hilbert space.

MSC 2010: Primary 62G08

Abstract: Expectile regression is a useful alternative to conditional mean and quantile regression for characterizing a conditional response distribution, especially when the distribution is asymmetric or when its tails are of interest. In this article, we propose a class of scalar-on-function linear expectile regression models where the functional slope parameter is assumed to reside in a reproducing kernel Hilbert space (RKHS). Our perspective addresses numerous drawbacks to existing estimators based on functional principal components analysis (FPCA), which make implicit assumptions about RKHS eigenstructure. We show that our proposed estimator can achieve

an optimal rate of convergence by establishing asymptotic minimax lower and upper bounds on prediction error. Under this framework, we propose a flexible implementation based on the alternating direction method of multipliers algorithm. Simulation studies and an analysis of real-world neuroimaging data validate our methodology and theoretical findings and, furthermore, suggest its superiority over FPCA-based approaches in numerous settings. *The Canadian Journal of Statistics* xx: 1–52; 2021

1. INTRODUCTION

Functional data have grown ubiquitous in medical data analysis, biology, and image and signal processing, among many other fields (Ramsay and Silverman, 2006; Li et al., 2018; Wang et al., 2019; Yu et al., 2019). While intrinsically functional, this type of data is almost always observed discretely over a grid, where the number of grid points is often larger than the number of observations. Due to the spatial or temporal nature of this grid, observations at nearby grid points are often highly correlated. Specialized techniques are consequently crucial in the proper analysis of functional data.

Traditional analytic approaches typically assume that errors are independent and identically distributed (i.i.d.) with a symmetric and homoscedastic density (Gu and Hui, 2016). These assumptions cannot be guaranteed in practice, particularly in high-dimensional settings (Li and Yao, 2019). As typical examples, consider modelling meteorological outcomes (e.g., from the Canadian weather

* *Author to whom correspondence may be addressed.*

E-mail: lkong@ualberta.ca

dataset, as in Ramsay and Silverman, 2006 or Cai and Yuan, 2012) or clinical outcomes (e.g., Mini Mental State Examination (MMSE) scores, a clinical survey-based measure used to quantify Alzheimer's disease severity, as in Jack et al., 2008). In these and many other settings, there is no guarantee that the conditional response distribution will be symmetric, much less Gaussian. It is more often the case that stochastic error terms are heteroscedastic and that the conditional response distribution is highly skewed or heavy-tailed. As noted in Newey and Powell (1987), heteroscedastic errors lead to inefficient or inconsistent parameter and covariance estimation.

From a practical standpoint, in regression settings where error is heteroscedastic or asymmetric, several estimators may be required for a satisfactory picture of the relationship between the response variable and model predictors. Each of these estimators may speak to a different notion of the location of the conditional response distribution, such as its different quantiles levels. Neuroimaging data analysis is one such setting where responses at multiple extreme levels, representing outlying or abnormal cases, are of more practical interest than, say, a single conditional mean. Pietrosanu et al. (2021) further emphasizes the particular need for functional tools not focused solely on conditional mean estimation in neuroimaging data analysis and more general fields of application.

Motivated by the dependence of traditional coefficient estimators on error homoscedasticity and symmetry assumptions, Newey and Powell (1987) first in-

troduced expectile regression, also called asymmetric least squares regression (Waltrup et al., 2015; Gu and Hui, 2016). Expectiles are analogous to quantiles and can similarly be computed for a random variable Y at any level $\tau \in (0, 1)$, but are determined by the tail expectations rather than the tail probabilities of a distribution. While quantile regression has a strong intuitive appeal, well-studied robustness properties, and broad applications in a variety of research fields (Koenker, 2017), Newey and Powell (1987) motivates expectiles by pointing out three major drawbacks to quantile estimators: their nondifferentiability, their relative inefficiency for near-Gaussian error distributions, and the difficulty inherent in computing their covariance.

It is then a natural development to consider expectile regression with functional predictors (i.e., in a "scalar-on-function" framework). In this article, we are concerned with the model

$$Y = \int_{\mathcal{T}} X(t)\beta_0(t)\ \mathrm{d}t + \varepsilon, \tag{1}$$

where Y is a scalar response, $X : \mathcal{T} \to \mathbb{R}$ is a square-integrable stochastic processs, and $\beta_0 : \mathcal{T} \to \mathbb{R}$ is the slope function. We assume that the domain $\mathcal{T}$ is a compact subset of a Euclidean space.

Most recent approaches to functional linear regression are based on functional principal components analysis (FPCA) (Hall and Horowitz, 2007). FPCA ultimately relies on an efficient representation of β_0 in terms of the leading functional principal components of X (Cai and Yuan, 2012). However, these func-

tional principal components might not form an appropriate basis to express β_0 or might have little predictive power. Consequently, FPCA-based methods might not perform well. In practice, this phenomenon has been observed for functional data in the Canadian weather dataset (Ramsay and Silverman, 2006; Cai and Yuan, 2012), in the Alzheimer's Disease Neuroimaging Initiative (ADNI) data analyzed in this article, and more generally, in principal components regression and singular value decomposition methods for linear inverse problems (Donoho and Johnstone, 1995). Numerous other works have considered the model in Equation (1) using FPCA-based approaches (Cai and Hall, 2006; Hall and Horowitz, 2007; Crambes et al., 2009; James et al., 2009; Schnabel and Eilers, 2009; Guo et al., 2015; Liao et al., 2019).

In this article, we study instead the functional linear expectile regression model from the perspective of a reproducing kernel Hilbert space (RKHS): we assume that the slope function β_0 resides in an RKHS $\mathcal{H}(K)$. In this more general framework, the functional covariance operator C and the reproducing kernel K of the RKHS are not required to be related. This assumption differs from the implicit requirements in FPCA-based frameworks that the ordered eigenfunctions of K and C perfectly coincide. FPCA-based approaches further assume that the slope function β_0 can be efficiently represented in terms of leading functional principal components (Yuan and Cai, 2010; Cai and Yuan, 2012). RKHS-based estimators, such as those proposed in this article, circumvent this restriction.

As illustrated in Yuan and Cai (2010), the eigenstructure of the RKHS plays an important role in estimation and prediction, making RKHS-based methods more difficult to implement. To our knowledge, the literature on RKHS-based approaches to functional data analysis is limited. Cheng and Shang (2015) considered a joint asymptotic framework for studying semi-nonparametric regression models where (finite-dimensional) Euclidean parameters and (infinite-dimensional) functional parameters are both of interest: the authors derived convergence rates for estimators of both. Qu et al. (2016) studied functional Cox models with right-censored data in the presence of both functional and scalar covariates in an RKHS framework. Notably, the authors proved that their functional coefficient estimator achieves the minimax optimal rate of convergence in penalized log partial likelihood settings. Li et al. (2007) derived various asymptotic results regarding kernel quantile regression (KQR) and proposed an efficient algorithm to compute entire KQR solution paths.

In this article, we propose a regularized estimator for the functional linear expectile regression model in an RKHS framework. Specifically, unlike existing FPCA-based approaches to expectile regression, we use the reproducing kernel to approximate functional effects and capture local features. Theoretically (when the eigenfunctions of K and C agree) and empirically (regardless of this agreement) we find that our estimators exhibit stronger convergence rates relative to FPCA-based estimators. We further incorporate shrinkage penalties as

a means to improve estimate interpretability and generalizability for prediction. We derive upper and lower bounds for minimax convergence in prediction error and establish minimax convergence rate optimality for our proposed estimator. We demonstrate that RKHS-based methods simplify functional coefficient estimate regularization (e.g., via smoothness, sparsity, or Tikhonov penalties) and allow model estimation to be formulated as a convex optimization problem. Our RKHS-based estimator can thus be efficiently computed: the alternating direction method of multipliers (ADMM) algorithm we apply makes our procedure simple to implement and allows us to incorporate existing computational techniques for smoothing splines.

The remainder of the article is organized as follows. In Section 2, we discuss expectile regression and RKHSs and establish the minimax optimality of our proposed estimator. In Section 3, we reformulate model estimation as a convex optimization problem and derive an ADMM iterative update scheme using a finite-dimensional representation of the slope function obtained via the representer theorem. Section 4 investigates finite-sample performance through simulation studies and a real-world data analysis, the latter using data from the ADNI (Jack et al., 2008). A subsequent appendix contains technical proofs of this article's main results.

As notation to be used throughout this article, let $\|\cdot\|_2$ denote the Euclidean L_2 norm. For two positive real sequences $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$, we write $a_k \asymp b_k$

to indicate that the sequence of ratios $(a_k/b_k)_{k\in\mathbb{N}}$ is bounded away from zero and infinity.

2. THEORETICAL PROPERTIES

We first introduce functional linear expectile regression, our proposed estimator, and the setting where $\beta_0 \in \mathcal{H}(K)$. Following this, we derive upper and lower bounds for the minimax rate of convergence in prediction error and establish the minimax optimality of our proposed estimator.

2.1. Expectiles and functional linear expectile regression

Let Y be a random variable with a distribution function F and a finite mean. The τth expectile $\mu_\tau = \mu_\tau(F)$ of Y, as defined by Newey and Powell (1987), is

$$\mu_\tau(F) = \arg\min_{\eta\in\mathbb{R}} E_Y\, r_\tau(y-\eta),$$

for $\tau \in (0,1)$, where $r_\tau(y-\eta) = |\tau - \mathbb{1}(y<\eta)|(y-\eta)^2$.

Expectiles share many desirable characteristics of quantiles and various additional computational advantages (Newey and Powell, 1987). Jones (1994) showed that the expectiles of a distribution F are the quantiles of a distribution G defined explicitly as

$$G(y) = \frac{P(y) - yF(y)}{2(P(y) - yF(y)) + (y-\mu)},$$

where $P(y) = \int_{-\infty}^{y} x\ \mathrm{d}F(x)$ and $\mu = \int_{-\infty}^{\infty} x\ \mathrm{d}F(x)$.

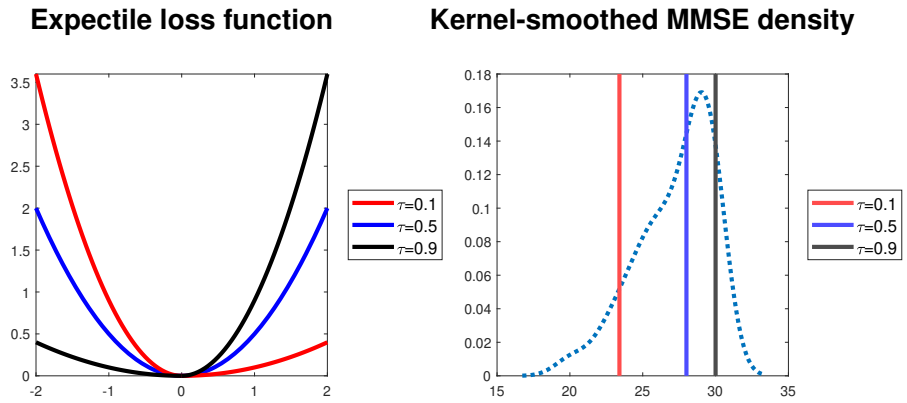

FIGURE 1: (Left) The expectile loss function for $\tau = 0.1, 0.5, 0.9$ in red, blue, and black, respectively. (Right) Kernel-smoothed estimate of the MMSE score density function (dotted blue) from the ADNI dataset. The corresponding expectiles at $\tau = 0.1, 0.5, 0.9$ are indicated in solid red, blue, and black, respectively.

As a generalization of ordinary mean regression, expectile regression is known to be statistically more efficient than quantile regression when standard assumptions such as error homoscedasticity are not severely violated (Liao et al., 2019). Unlike quantile regression, expectile regression uses a smooth loss function which, in terms of general computation, is considerably easier to optimize (Gu and Hui, 2016). Holzmann and Klar (2016) and Krätschmer and Zähle (2017) explored the asymptotic properties of sample expectiles and established their uniform consistency under the assumption of a finite mean. Unlike quantiles, expectiles are also guaranteed to be unique under this assumption. The asymptotic normality of the sample expectile estimator follows directly with the additional assumption of a finite second moment. Similar to quantiles, expectiles characterize and give more insight into a distribution of interest.

Figure 1 illustrates the expectile loss function at $\tau = 0.1, 0.5, 0.9$. When $\tau < 0.5$, the cost of a positive error is lower than that of a negative one, encouraging a smaller expectile μ_τ. Larger expectiles are correspondingly encouraged when $\tau > 0.5$. At $\tau = 0.5$, the loss $r_{0.5}$ is equivalent to the least squares loss and recovers the mean of the distribution. In settings where the distribution of Y is highly skewed rather than symmetric, τ can be chosen to obtain a more desirable location estimate. This is illustrated in Figure 1 using MMSE data from the ADNI.

In this article, we are primarily interested in establishing the convergence properties of our proposed regularized sample estimator for β_0 in the functional linear expectile regression model of Equation 1,

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} r_\tau \left(y_i - \int_{\mathcal{T}} x_i(t)\beta(t)\ \mathrm{d}t \right) + \lambda J(\beta), \tag{2}$$

where $\{(x_i, y_i) : i = 1, \ldots, n\}$ is a set of observed training data, J is a penalty function assessing the "plausibility" of a candidate β, and $\lambda \geq 0$ is a tuning parameter controlling the strength of the penalty J. For convenience, we suppress notation indicating implicit dependence on τ.

2.2. RKHS

We assume that the slope function β_0 resides in an RKHS $\mathcal{H} = \mathcal{H}(K)$, a subspace of square-integrable functions with the domain $\mathcal{T}$, equipped with a reproducing kernel K. The canonical example of $\mathcal{H}(K)$ is a Sobolev space: assuming, without

loss of generality, that $\mathcal{T} = [0, 1]$, the Sobolev space of order r (Golub et al., 1979) can be defined as

$$\mathcal{W}_2^r = \mathcal{W}_2^r([0,1]) = \{\beta : [0,1] \to \mathbb{R} : \beta, \beta^{(1)}, \ldots, \beta^{(r-1)} \text{ are absolutely continuous and } \beta^{(r)} \in \mathcal{L}_2\}.$$

One squared norm that will make $\mathcal{W}_2^r$ an RKHS (Brézis, 2011) is $\sum_{j=0}^{r-1}\left\{\int_{\mathcal{T}} \beta^{(j)}(t)\ \mathrm{d}t\right\}^2 + \int_{\mathcal{T}}\{\beta^{(r)}(t)\}^2\ \mathrm{d}t$. The penalty functional J on the slope function β can be conveniently defined as the squared norm or semi-norm associated with $\mathcal{H}$ (Cai and Yuan, 2011): one possible choice is $J(\beta) = \int_0^1 [\beta^{(r)}(t)]^2\ \mathrm{d}t$. The null space of J, defined as $\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$, forms a finite-dimensional linear subspace of $\mathcal{H}$ with some orthonormal basis $(\xi_1, \xi_2, \ldots, \xi_M)$, where $M = \dim(\mathcal{H}_0)$. The orthogonal complement $\mathcal{H}_1$ of the null space $\mathcal{H}_0$ is such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. It can be shown that $\mathcal{H}_1$ also forms an RKHS with the same inner product as $\mathcal{H}$, but restricted to $\mathcal{H}_1$. More generally, for any $\beta \in \mathcal{H}$, there exists $\beta_1 \in \mathcal{H}_0$ and $\beta_2 \in \mathcal{H}_1$ such that the decomposition $\beta = \beta_1 + \beta_2$ is unique (Nosedal-Sanchez et al., 2012; Gu, 2013). Let K be the reproducing kernel of $\mathcal{H}_1$ such that $J(\beta_2) = \|\beta_2\|_{\mathcal{H}}^2 = \|\beta\|_K^2$, defined as the RKHS norm of β. Consequently, as we will demonstrate, we can find a finite-dimensional representation for the functional slope coefficient β_0.

We next consider the two kernels crucial to the estimation process. First, recalling that $\mathcal{T} \subset \mathbb{R}$ is a compact set, a reproducing kernel $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a real, symmetric, square-integrable, nonnegative-definite function. There is

a one-to-one correspondence between a reproducing kernel K and an RKHS $\mathcal{H}(K)$. Mercer's theorem implies that K admits the spectral decomposition $K(s,t) = \sum_{k=1}^{\infty} \varrho_k \varphi_k(s)\varphi_k(t)$, where the eigenvalues $(\varrho_k)_{k\in\mathbb{N}}$ are in nonincreasing order and $(\varphi_k)_{k\in\mathbb{N}}$ are the corresponding eigenfunctions.

For any real, square-integrable, semidefinite function R, define $L_R : \mathcal{L}_2 \to \mathcal{L}_2$ as the linear integral operator $L_R(f)(\cdot) = \langle R(s,t), f\rangle_{\mathcal{L}^2(\mathcal{T})} = \int_{\mathcal{T}} R(s,\cdot)f(s)\ \mathrm{d}s$. By the spectral theorem, there exists a sequence of orthonormal eigenfunctions $(\psi_k^R)_{k\in\mathbb{N}}$ and a corresponding sequence of nonincreasing eigenvalues $(\theta_k^R)_{k\in\mathbb{N}}$ such that $R(s,t) = \sum_{k\in\mathbb{N}} \theta_k^R \psi_k^R(s)\psi_k^R(t)$ for all $s,t \in \mathcal{T}$, and $L_R(\psi_k^R) = \theta_k^R\psi_k^R$ for $k \in \mathbb{N}$. Additionally, for all $s,t \in \mathcal{T}$, $R^{1/2}(s,t) = \sum_{k\in\mathbb{N}} \sqrt{\theta_k^R}\psi_k^R(s)\psi_k^R(t)$. We say that two linear operators are aligned if they share the same ordered (i.e., with corresponding eigenvalues in nonincreasing order) sequence of eigenfunctions.

Let $L_{R^{1/2}}$ be the linear operator defined by $L_{R^{1/2}}(\psi_k^R) = \sqrt{\theta_k^R}\psi_k^R$. It is clear that $L_{R^{1/2}} = (L_R)^{1/2}$. Further defining $(R_1R_2)(s,t) = \int_{\mathcal{T}} R_1(s,u)R_2(u,t)\ \mathrm{d}u$, it follows that $L_{R_1R_2} = L_{R_1} \circ L_{R_2} = L_{R_2} \circ L_{R_1}$.

With the previous results in mind, consider the covariance kernel $C : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ for X, defined as $C(s,t) = E\left([X(s) - EX(s)]\,[X(t) - EX(t)]\right)$. Of course, we require that the covariance kernel C be continuous and square-integrable over $\mathcal{T} \times \mathcal{T}$. Similar to K, C admits the spectral decomposition $C(s,t) = \sum_{k=1}^{\infty} \mu_k\phi_k(s)\phi_k(t)$. The two eigenfunction sequences $(\varphi_k)_{k\in\mathbb{N}}$ and

$(\phi_k)_{k\in\mathbb{N}}$ are different in general. However, under certain conditions, K and C can be simultaneously diagonalized (Conrad, 2014).

Using the eigenstructures of the reproducing and covariance kernels K and C, we can define the linear operator $L_{K^{1/2}CK^{1/2}}$ in a compositional fashion as $L_{K^{1/2}CK^{1/2}} = L_{K^{1/2}} \circ L_C \circ L_{K^{1/2}}$. By the spectral theorem, $K^{1/2}CK^{1/2}$ has the spectral decomposition $K^{1/2}CK^{1/2}(s,t) = \sum_{k=1}^{\infty} \nu_k \zeta_k(s)\zeta_k(t)$, where the sequence of eigenvalues $(\nu_k)_{k\in\mathbb{N}}$ is arranged in nonincreasing order and $(\zeta_k)_{k\in\mathbb{N}}$ is the corresponding sequence of orthonormal eigenfunctions. Obviously, the eigenvalues $(\nu_k)_{k\in\mathbb{N}}$ are determined by the eigenvalues of both K and C and the alignment of their respective eigenfunctions. We will eventually show that the convergence rate of our proposed estimator is related to the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$.

Before discussing estimation of the functional coefficient β_0 over $\mathcal{H}(K)$, we impose two basic assumptions on the reproducing and covariance kernels, whose eigenstructures determine the optimal convergence rate.

(A1) The eigenvalues of $K^{1/2}CK^{1/2}$ satisfy $\nu_k \asymp k^{-2r}$ for some $r > 0$.

(A2) For any square-integrable function f,

$$E\left[\int_{\mathcal{T}} [X(t) - EX(t)]\, f(t)\, \mathrm{d}t\right]^4 \leq c\left(E\left[\int_{\mathcal{T}} [X(t) - EX(t)]\, f(t)\, \mathrm{d}t\right]^2\right)^2$$

for some constant $c > 0$.

Assumption A1 pertains to the decay rate of ν_k. As already discussed, this rate is determined by the eigenstructures of the kernels K and C, specifically, their individual eigenvalue decay rates and the alignment between their eigenfunctions. The eigenvalues of the covariance kernel C obey $\mu_k \asymp k^{-2r^C}$ if the Sacks–Ylvisaker condition of order $r^C - 1$ is satisfied for some integer $r^C \geq 1$ (Yuan and Cai, 2010; Ritter et al., 1995). As an example, the Ornstein-Uhlenbeck covariance kernel $C(s,t) = \exp(-|s-t|)$ has $r^C = 1$. For Sobolev spaces, various covariance functions are known to satisfy the Sacks–Ylvisaker condition (Ritter et al., 1995). Concerning the eigenvalue decay rate of the kernel K, if $\mathcal{H}$ is the r^Kth order Sobolev space $\mathcal{W}_2^{r^K}$, it is known that $\varrho_k \asymp k^{-2r^K}$ (Micchelli and Wahba, 1979).

When K and C are aligned, i.e., when they share a common ordered eigenfunction set so that $\phi_k = \varphi_k$ for $k \in \mathbb{N}$ (Cai and Yuan, 2012), it follows that $r = r^C + r^K$ in Assumption A1. However, if K and C are not aligned, then the eigenvalues of the two operators alone cannot determine the order r. For example, the eigenvalues for the Sobolev class $\mathcal{W}_2^r$ for $r > 1/2$ follow a polynomial decay rate.

Assumption A2 restricts the fourth moment of the linear functional $\int_{\mathcal{T}} X(t)f(t)\,\mathrm{d}t$, ensuring bounded kurtosis. When X is a Gaussian process, for example, Assumption A2 is satisfied with $c = 3$.

We take $\mathcal{H}(K) = \mathcal{W}_2^2$ and define the penalty function as $J(\beta) = \int_{\mathcal{T}} [\beta''(t)]^2 \, \mathrm{d}t = \|\beta\|_K^2$. Consequently, $\mathcal{H}_0$ is the linear space spanned by $\xi_1(t) = 1$ and $\xi_2(t) = t$.

The accuracy of $\hat{\beta}_n$ can be measured via the squared RKHS norm associated with the covariance kernel C (Yuan and Cai, 2010), as

$$\left\| \hat{\beta}_n - \beta_0 \right\|_C^2 = E_{X^*} \left(\int X^*(t) \hat{\beta}_n(t) \, \mathrm{d}t - \int X^*(t) \beta_0(t) \, \mathrm{d}t \right)^2,$$

where X^* is an independent copy of X and the expectation on the right-hand side is taken over X^*. The above quantity measures the mean squared prediction error for a random, future observation of X.

Theorem 1. (Minimax lower bound) *Under Assumption A1,*

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}_n} \sup_{\beta_0 \in \mathcal{H}(K)} \mathbb{P}_{\beta_0} \left\{ \|\hat{\beta}_n - \beta_0\|_C \geq a n^{-\frac{2r}{2r+1}} \right\} = 1, \tag{3}$$

where the infimum is taken over all possible estimators $\hat{\beta}_n$ computed from the training data.

Theorem 2. (Minimax upper bound) *Under Assumptions A1 and A2,*

$$\lim_{A \to 0} \limsup_{n \to \infty} \sup_{\beta_0 \in \mathcal{H}(K)} \mathbb{P}_{\beta_0} \left\{ \|\hat{\beta}_n - \beta_0\|_C \geq A n^{-\frac{2r}{2r+1}} \right\} = 0. \tag{4}$$

provided that the tuning parameter satisfies $\lambda \asymp n^{-2r/(2r+1)}$.

By Theorems 2 and 3, the regularized estimator $\hat{\beta}_n$ is minimax rate optimal: the minimax rate of convergence for the prediction error is $n^{-2r/(2r+1)}$. As discussed previously, this optimal rate of convergence depends jointly on the eigenvalue decay rate of the operator $L_{K^{1/2}CK^{1/2}}$ (i.e., of the eigenvalues of K and C) through r and, more importantly, on alignment between the eigenfunctions of K and C.

3. COMPUTATION

In this section, we propose an efficient computational approach for model estimation using the ADMM algorithm. We begin with an application of the representer theorem to establish that the proposed estimator lies in a finite-dimensional subspace. We subsequently discuss hyperparameter tuning and propose our estimation algorithm.

3.1. Representer theorem

Theorem 3. (Representer theorem) *Let $(\xi_1, \ldots, \xi_M)$ be a basis of $\mathcal{H}_0$. There exist vectors $e = (e_1, \ldots, e_M)^\top$ and $c = (c_1, \ldots, c_n)^\top$ allowing the solution $\hat{\beta}_n$ to the problem in Equation (2) to be expressed as*

$$\hat{\beta}_n(t) = \sum_{i=1}^{M} e_i \xi_i(t) + \sum_{k=1}^{n} c_k \int_{\mathcal{T}} K(s,t) X_k(t)\, \mathrm{d}s. \tag{5}$$

Theorem 3 is a generalization of the well-known representer lemma for smoothing splines (Wahba, 1990). Although the minimization over $\hat{\beta}_n$ in Equation (2) is taken over an infinite-dimensional space $\mathcal{H}(K)$, the above result im-

plies that the solution lies in a finite-dimensional subspace. Thus, it suffices to estimate the coefficients e and c in Equation (5). By Theorem 3, we can conclude that

$$\int_{\mathcal{T}} X(t)\beta(t)dt = \sum_{i=1}^{M} e_i \int_{\mathcal{T}} X(t)\xi_i(t)\ \mathrm{d}t + \sum_{k=1}^{n} c_k \int_{\mathcal{T}}\int_{\mathcal{T}} X(t)K(s,t)X_k(s)\ \mathrm{d}s\ \mathrm{d}t.$$

Let $Y = (Y_1, Y_2, \ldots, Y_n)^\top$ and let T represent the $n \times M$ matrix with the (i,j)th entry $T_{ij} = \int_{\mathcal{T}} X_i(t)\xi_j(t)\ \mathrm{d}t$ for $i = 1, \ldots, n$ and $j = 1, \ldots, M$. Similarly, let Σ be the $n \times n$ matrix with the (i,j)th entry $\Sigma_{ij} = \int_{\mathcal{T}}\int_{\mathcal{T}} X_i(t)K(s,t)X_j(s)\ \mathrm{d}s\ \mathrm{d}t$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. It follows from the reproducing property that

$$J(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \int_{\mathcal{T}}\int_{\mathcal{T}} X_i(t)K(s,t)X_j(s)\ \mathrm{d}s\ \mathrm{d}t = c^\top \Sigma c.$$

We make use of this representation in the following subsections for model estimation.

3.2. Hyperparameter tuning

As with most smoothing methods, the selection of the tuning parameter λ influences the performance of the regularized estimator $\hat{\beta}_n$. There are various tools available for this task, such as K-fold cross-validation (Kohavi, 1995), the Bayesian information criterion (BIC), generalized maximum likelihood (Wahba, 1990), and generalized cross-validation (GCV) (Golub et al., 1979).

In this article, unless otherwise noted, we employ GCV as a practical criterion for choosing the optimal tuning parameter value. Because the regularized esti-

mator is a linear estimator and can be written as $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) = H(\lambda)y = Te + \Sigma c$, where $H(\lambda)$ is the "hat matrix" for a particular value of λ, we may select the the value of λ that minimizes (Wahba, 1990)

$$GCV(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} r_\tau(\hat{y}_i - y_i)}{(1 - \mathrm{Tr}(H(\lambda))/n)^2}.$$

3.3. ADMM algorithm

We next apply the ADMM algorithm to estimate the functional linear expectile model. Pseudocode for the proposed estimation procedure is provided in Algorithm 1.

Developed in the 1970s and further summarized in Boyd (2010), the ADMM algorithm is a simple and efficient approach for solving convex optimization problems. It has found renewed popularity in large-scale computing through its ability to decentralize large, global problems into small, local ones. The ADMM algorithm has been employed in quantile regression (Gu et al., 2018), two-way functional hazard models (Li et al., 2018), and Gaussian graphical models (Ma et al., 2020), to name only a few applications.

Using the results and notation of Section 3.1, the optimization problem in Equation (1) can be reformulated as a convex optimization problem with respect to e, c, and the auxiliary variable u as

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{n} \textstyle\sum_{i=1}^{n} r_\tau(y_i - u_i) + \lambda c^\top \Sigma c \\ &\text{subject to} \quad u_i = T_i e + \Sigma_i c, \quad i = 1, \ldots, n, \end{aligned} \tag{6}$$

where T_i and Σ_i denote the ith rows of T and Σ, respectively. The scaled ADMM algorithm (Boyd, 2010) uses an objective function defined by the augmented Lagrangian form of the above problem,

$$\begin{aligned} L_\sigma(u, e, c, h) =& \frac{1}{n}\sum_{i=1}^{n} r_\tau(Y_i - u_i) + \lambda c^\top \Sigma c + \\ & \frac{\sigma}{2}\sum_{i=1}^{n}(u_i - T_i e - \Sigma_i c + h_i)^2 - \frac{\sigma}{2}\sum_{i=1}^{n} h_i^2, \end{aligned}$$

which we aim to minimize over $u = (u_1, \ldots, u_n)^\top$, e, c, and $h = (h_1, \ldots, h_n)^\top$ without restriction.

The scaled ADMM update scheme for the $(k+1)$th iteration is straightforward to derive:

$$\begin{aligned} u_i^{k+1} &= \arg\min_{u_i} \left\{ \frac{1}{n}\sum_{i=1}^{n} r_\tau(Y_i - u_i) + \frac{\sigma}{2}(u_i - T_i e^k - \Sigma_i c^k + h_i^k)^2 \right\} \\ &= \begin{cases} \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2\tau y_i}{\sigma + 2\tau}, & y_i \geq u_i \\ \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2(1-\tau) y_i}{\sigma + 2(1-\tau)}, & y_i < u_i \end{cases} \\ (e^{k+1},\ c^{k+1}) &= \arg\min_{e_i, c_i} \left\{ \lambda c^\top \Sigma c + \frac{\sigma}{2}\left(u_i^{k+1} - T_i e - \Sigma_i c + h_i^k\right)^2 \right\} \\ h_i^{k+1} &= h_i^k + u_i^{k+1} - T_i e^{k+1} - \Sigma_i c^{k+1}. \end{aligned}$$

The update step for (e, c) above can be explicitly solved using the sub-iterations

$$e^{k+1} = (T^\top T)^{-1} \left[\sum_{i=1}^{n} T_i^\top (u_i^{k+1} - \Sigma_i c^k + h_i^k) \right],$$

$$c^{k+1} = (2\lambda\Sigma/\sigma + \Sigma^\top \Sigma)^{-1} \left[\sum_{i=1}^{n} \Sigma_i^\top (u_i^{k+1} - T_i e^{k+1} + h_i^k) \right].$$

Stopping conditions for the proposed scheme can be defined in terms of the size of the problem's primal and dual residuals: we terminate the algorithm when $r^k = \|u - Te - \Sigma c\| \leq \epsilon_{\text{dual}}$ and $s^k = \sigma \left(T(e^{k+1} - e^k) + \Sigma(c^{k+1} - c^k) \right) \leq \epsilon_{\text{pri}}$. Here, $\epsilon_{pri} = \sqrt{n}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max(\|u\|_2, \|Te + \Sigma c\|_2) > 0$ and $\epsilon_{\text{dual}} = \sqrt{n}\epsilon_{\text{abs}} + \epsilon_{\text{rel}}\|h\|_2 > 0$ are feasibility tolerances for the primal and dual feasibility conditions, where $\epsilon_{\text{abs}} > 0$ and $\epsilon_{\text{rel}} > 0$ are absolute and relative tolerances, respectively. In all of the numerical studies presented in this paper, we follow the suggestion in Boyd (2010) by fixing $\epsilon_{\text{rel}} = 10^{-4}$, $\epsilon_{\text{abs}} = 10^{-2}$, and $\sigma = 2$.

3.4. Convergence of the ADMM algorithm

We next apply a general result of Boyd (2010) to verify the convergence of our proposed ADMM-based approach for estimating the functional linear expectile regression model. For convenience, we return to a more-general formulation of the ADMM algorithm:

$$\begin{aligned} &\text{minimize } F(x, z) = f(x) + g(z) \\ &\text{subject to } G(x, z) = Ax + Bz - c = 0, \end{aligned} \tag{7}$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, where $A \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times m}$, and $c \in \mathbb{R}^p$ (Boyd, 2010). For our setting, x and z correspond to u and $(e^\top, c^\top)^\top$; f and g to the empirical

Algorithm 1 ADMM algorithm for functional linear expectile regression.

Input: u^0, e^0, c^0, h^0 (initial estimates); σ (step size parameter); λ (tuning parameter)

repeat

  for $i = 1, \ldots, n$ **do**

    if $u_i^k \leq y_i^k$ **then**

      $u_i^{k+1} \leftarrow \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2\tau y_i}{\sigma + 2\tau}$

    else

      $u_i^{k+1} \leftarrow \dfrac{\sigma(T_i e^k + \Sigma_i c^k - h_i^k) + 2(1-\tau) y_i}{\sigma + 2(1-\tau)}$

    end if

  end for

  $e^{k+1} \leftarrow (T^\top T)^{-1} \left[\sum_{i=1}^n T_i^\top (u_i^{k+1} - \Sigma_i c^k + h_i^k)\right]$

  $c^{k+1} \leftarrow (2\lambda\Sigma/\sigma + \Sigma^\top \Sigma)^{-1} \left[\sum_{i=1}^n \Sigma_i^\top (u_i^{k+1} - T_i e^{k+1} + h_i^k)\right]$

  $h_i^{k+1} \leftarrow h_i^k + u_i^{k+1} - T_i e^{k+1} - \Sigma_i c^{k+1}$

until stopping criteria are met

compute estimated slope function $\hat{\beta}$ from the optimal e, c

Output: $L_\sigma(u, e, c, h)$, $\hat{\beta}$

expectile loss $\frac{1}{n}\sum_{i=1}^n r_\tau(Y_i - u_i)$ and $\lambda c^\top \Sigma c$; and A, B, and c to the identity matrix I, $[T_i, \Sigma_i]$, and 0, respectively. To guarantee convergence, we verify two additional conditions, referring to Assumptions 1 and 2 of Boyd (2010).

First, we require that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, proper, and convex. This requirement is naturally satisfied for our for-

mulation in (6).

Second, we require that the nonaugmented Lagrangian $L_0(x, z, y) = f(x) + g(z) + y^\top(Ax + Bz - c)$ has a saddle point, i.e., that there exists a (not necessarily unique) $(x^\star, z^\star, y^\star)$ satisfying $L_0(x^\star, z^\star, y) \leq L_0(x^\star, z^\star, y^\star) \leq L_0(x, z, y^\star)$ for all (x, z, y). The existence of a saddle point follows immediately from the saddle point theorem (Mohri et al., 2018, Theorem B.29) and the fact that we are optimizing over a real space (specifically, with a nonempty interior), that $G(x, z)$ is affine, and that $G(0, 0) = 0$ (since $c = 0$).

Consequently, we can guarantee that the estimate and objective function iterates in our ADMM-based implementation will converge to the solution and optimal value, respectively, of the original problem. The particular benefit of an ADMM-based approach is that each update has a closed form, which speeds up numerical computation relative to traditional interior point methods or other generic algorithms. Indeed, empirical results in existing literature have illustrated the clear superiority that ADMM-based algorithms have in a variety of settings (Chen and Wei, 2005; Pietrosanu et al., 2020).

4. NUMERICAL EXPERIMENTS

We now investigate the finite-sample performance of our proposed estimators. We are specifically interested in comparisons between our proposed estimator and one using an FPCA-based approach that uses the first four leading eigenfunctions (Ramsay and Silverman, 2006; Yuan and Cai, 2010; Kato, 2012). Three

sets of simulations in Section 4.1 examine the effects of eigenvalue decay, kernel alignment, and various error distributions on the convergence of both estimators.

4.1. Simulation studies

In the following sets of simulation studies, we consider $\mathcal{T} = [0, 1]$ and let $\mathcal{H} = \mathcal{H}(K)$ be the set of functions in the linear span of the cosine basis (Cai and Yuan, 2012), i.e., $\mathcal{H}(K) = \{g(t) = \sqrt{2}\sum_{k\in\mathbb{N}} g_k \cos(k\pi t) : g_k \in \mathbb{R}, k \in \mathbb{N}\} \subset \mathcal{W}_2^2$. When endowed with the squared norm

$$\|f\|_{\mathcal{H}(K)}^2 = \int_{\mathcal{T}} (f'')^2 = \int_0^1 \left(\sqrt{2}\sum_{k\in\mathbb{N}} (k\pi)^2 g_k \cos k\pi t\right)^2 = \sum_{k\in\mathbb{N}} (k\pi)^4 g_k^2,$$

$\mathcal{H}$ is an RKHS with the reproducing kernel

$$\begin{aligned} K(s,t) &= \sum_{k\in\mathbb{N}} 2(k\pi)^{-4} \cos(k\pi s)\cos(k\pi t) \\ &= -\frac{1}{3}\left(B_4(|s-t|/2) + B_4((s+t)/2)\right), \end{aligned}$$

where B_k is the kth Bernoulli polynomial

$$B_{2m}(x) = (-1)^{m-1} 2(2m)! \sum_{k\in\mathbb{N}} \frac{\cos(2\pi k x)}{(2\pi k)^{2m}},$$

for $x \in [0, 1]$. Additionally, we choose $(\xi_1(t) = 1, \xi_2(t) = t)$ as the basis for the null space $\mathcal{H}_0$.

To quantify the behaviour of varying coefficient estimates, we calculate prediction error (PE) on a test dataset $\{(x_i^*, y_i^*) : i = 1, \ldots, n^*\}$, given by

$$\text{PE}_\tau = \left(\frac{1}{n^*}\sum_{j=1}^{n^*} \left\| \int_{\mathcal{T}} x_j^*(t)\hat{\beta}_n(t)\, \mathrm{d}t - \int_{\mathcal{T}} x_j^*(t)\beta_0(t)\, \mathrm{d}t \right\|_2^2\right)^{1/2}.$$

As a more direct comparison between the RKHS- and FPCA-based estimators, we also report relative prediction error, defined by $\text{PE}_\tau^{\text{FPCA}}/\text{PE}_\tau^{\text{RKHS}}$, where $\text{PE}_\tau^{\text{FPCA}}$ and $\text{PE}_\tau^{\text{RKHS}}$ represent prediction errors for the two methods. In all simulation studies, results are averaged over 100 simulated training and test datasets.

In the first simulation study, we focus primarily on the effect of eigenvalue decay rate. We define the covariance operator as

$$C(s,t) = \sum_{k=1}^{50} 2k^{-2r_2} \cos(k\pi s)\cos(k\pi t),$$

where $r_2 = 1, 2, 3$ imposes different decay rates on the eigenvalues of C: a larger value of r_2 yields stronger eigenvalue decay. In this setting, the two kernels, K and C, share the same ordered set of eigenfunctions.

We follow the data generation procedure in Hall and Horowitz (2007) and Cai and Yuan (2012). The response is generated as $Y = \int_0^1 X(t)\beta_0(t)\ \mathrm{d}t + \varepsilon$, with $\beta_0(t) = \sum_{k=1}^{50} \beta_k\phi_k(t)$; $\beta_k = 4(-1)^{k+1}k^{-2}$ and $\phi_k(t) = \sqrt{2}\cos(k\pi t)$ for $k = 1, \ldots, 50$; and $\varepsilon \sim N(0, 0.5)$. The functional covariate is generated as $X(t) = \sum_{k=1}^{50} \gamma_k U_k \phi_k(t)$, where $\gamma_k = (-1)^{k+1}k^{-r_2}$ and $U_k \overset{\text{i.i.d.}}{\sim} U[-\sqrt{3}, \sqrt{3}]$. The U_ks have a mean of zero and unit variance and each X is observed at 101 equally spaced grid points on $[0, 1]$. We emphasize that the data generation process is ultimately driven by the choice of the covariance operator.

Results for the first simulation are presented in Figure 2. First, the generally positive performance of the FPCA-based estimator is not surprising, as β_0 is a linear combination of the leading eigenfunctions of the functional covariate X.

Nonetheless, our RKHS-based estimator demonstrates higher relative predictive performance except in certain settings with $r_2 = 1$, where the eigenvalue decay rate is small. In these settings, the standard errors of the PE and relative PE measures across simulations are typically small. Together, these results suggest a systematically lower PE for the proposed method. The PE of both estimators generally decreases as r_2 increases, as expected. Both methods appear to converge at similar rates as the sample size increases, although the RKHS-based estimator again outperforms the FPCA-based one.

In the second simulation study, we are primarily interested in how alignment between the reproducing kernel K and the covariance kernel C influences the performance of the RKHS- and FPCA-based estimators. We define the covariance kernel in this setting as

$$C(s,t) = \sum_{k=1}^{50} 2(|k-k_0|+1)^{-2}\cos(k\pi s)\cos(k\pi t).$$

To control the extent of the alignment between K and C, the leading eigenfunctions of C are located around the k_0th eigenfunction of the reproducing kernel K: we consider $k_0 = 5, 10, 20$, with larger values of k_0 corresponding to worse alignment (Cai and Yuan, 2012). In all other aspects, the data generation process matches that of the first simulation study.

Figure 3 presents results for the second simulation study. As expected, the FPCA-based estimator generally shows worse PE relative to the RKHS-based estimator. We observe that relative PE increases with worsened alignment, most

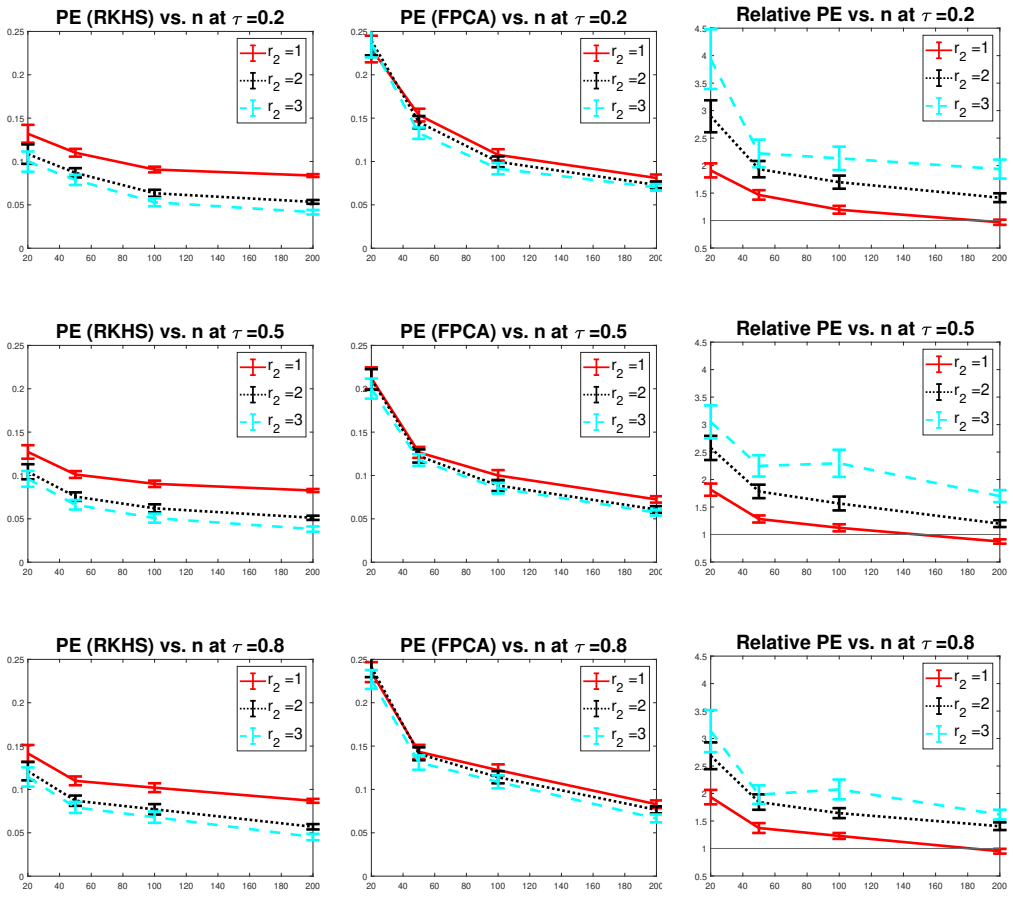

FIGURE 2: Effect of covariance kernel eigenvalue decay rate on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the first simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$.

notably when $k_0 = 20$. Furthermore, with increasing k_0, poor alignment between K and C seems to have a significant impact on the FPCA-based estimator but little effect on the proposed RKHS-based one. The standard error for relative PE is large in some settings, but still leads us to conclude that the proposed method gives systematically better PE. These empirical results are consistent with our

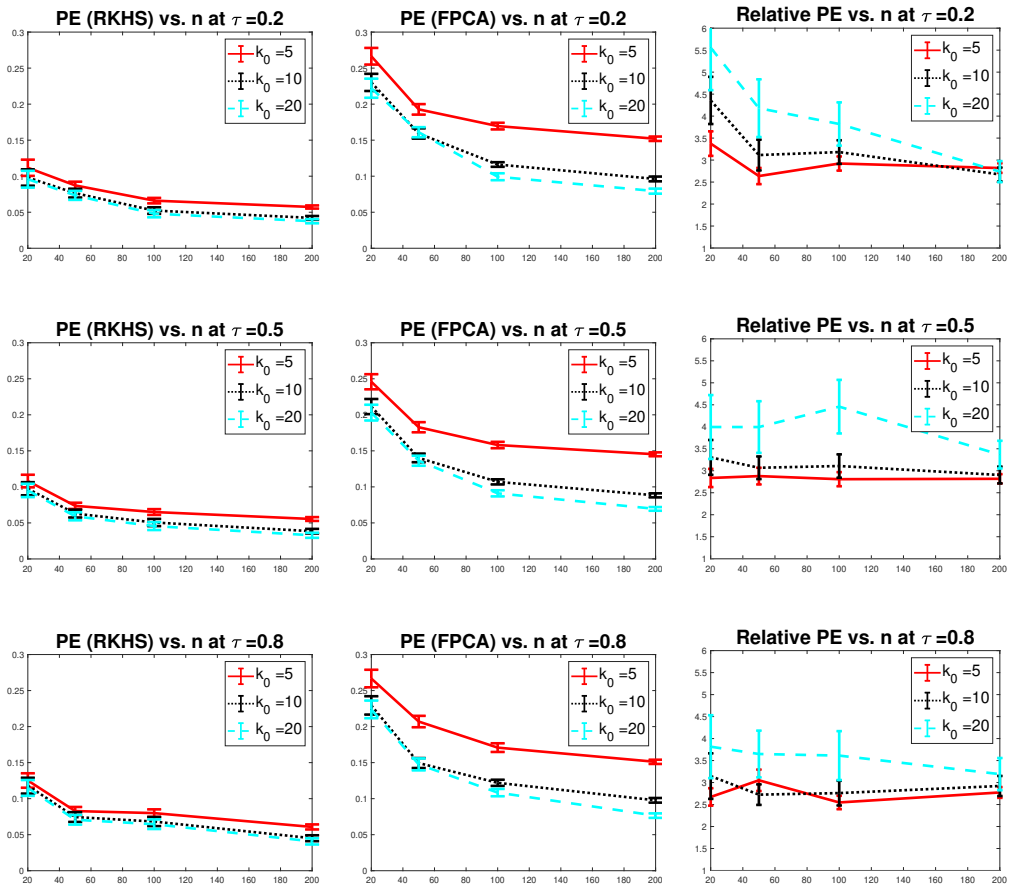

FIGURE 3: Effect of reproducing and covariance kernel alignment on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the second simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$.

theoretical expectations and illustrate the merit of our RKHS-based perspective.

In the third simulation study, we investigate the ability of our proposed approach to cope with different types of error distributions. Specifically, we consider distributions that are either heteroscedastic or asymmetric.

We use the same setup as the first simulation study (excepting the distribution of ε), with $r_2 = 2$. As asymmetric error distributions, we take $\varepsilon \sim \text{Gamma}(2, 0.2)$ and $\varepsilon \sim \text{Beta}(5, 1)$ for left- and right-skewed errors, respectively. Heteroscedastic errors are sampled as a mixture of $N(0, 0.25)$, $N(0, 0.375)$, and $N(0, 0.5)$ distributions, representing a simple case with three heteroscedastic groups.

Results are presented in Figure 4 for the third simulation study. As a general trend, our proposed RKHS-based estimator shows better performance than the FPCA-based estimator, with relative PE typically falling between one and four. Standard error for relative PE is moderate across the different settings but is again suggestive of a systematically lower PE for the proposed RKHS-based estimator. In the setting with right-skewed errors, PE for both estimators is relatively smaller when $\tau = 0.2$ than when $\tau = 0.8$: this result is reversed for left-skewed errors. These results, for both asymmetric and heteroscedastic error distributions, demonstrate the power of expectile regression in dealing with various error distributions, relative to methods that focus on conditional mean estimation. This simulation study highlights the versatility of our expectile model in cases of model error misspecification.

4.2. Application to ADNI data

We next apply the proposed RKHS-based estimator in an analysis of MMSE scores from 199 patients in the ADNI dataset. In the functional linear model,

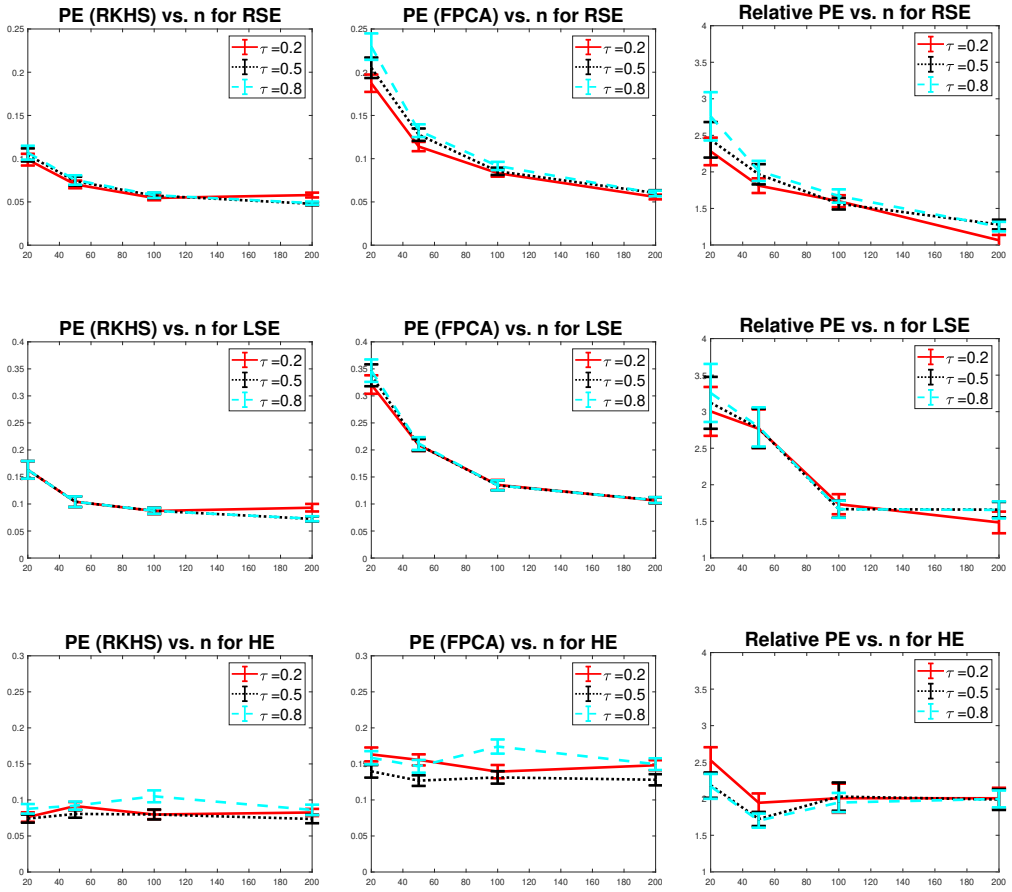

FIGURE 4: Effect of abnormal errors on the RKHS- and FPCA-based estimators at $\tau = 0.2, 0.5, 0.8$ in the third simulation study. From left to right, the three columns show PE for the RKHS- and FPCA-based estimators and relative PE between both (with values above one favouring the proposed estimator). Error bars correspond to average PE $\pm$ SE, evaluated over 100 replications. In each subplot, the horizontal axis represents the size n of the training dataset, considered at $n = 20, 50, 100, 200$. RSE, LSE, and HE indicate left-skewed, right-skewed, and heteroscedastic error distributions, respectively.

the response Y is MMSE score while the functional predictor X is fractional anisotropy (FA) as a function of distance along the midsagittal corpus callosum skeleton (scaled to $\mathcal{T} = [0, 1]$). The corresponding functional linear model is

$$MMSE = \int_0^1 \beta_0(t) FA(t)\ \mathrm{d}t + \varepsilon.$$

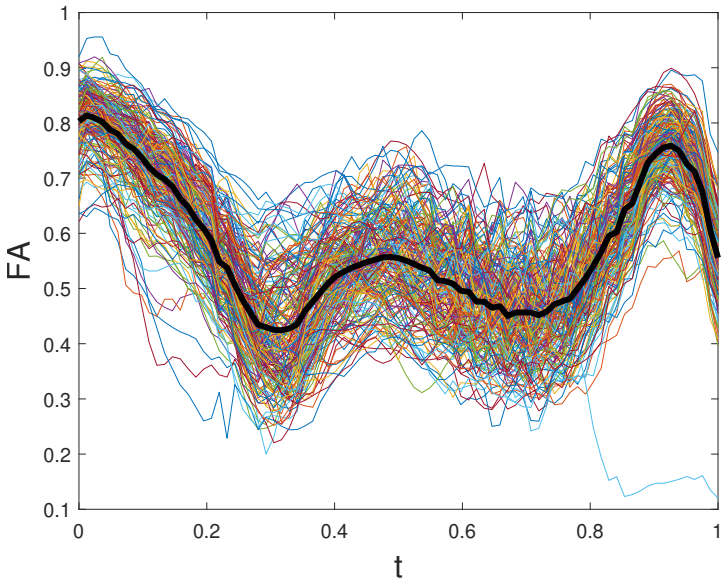

FIGURE 5: FA curves evaluated along the corpus callosum skeleton. The solid black line is the mean FA curve.

Figure 5 plots FA values, observed at 83 grid points, for all 199 patients. For tuning and evaluating both estimators, approximately 80% of the data is used for four-fold cross validation while the remaining 20% is held out as a test set. Context and the visualization of the neuroimaging data in Figure 5 suggest that the functional predictor $X = \text{FA}$ may be periodic on $[0, 1]$.

We let $\mathcal{H}(K) = \mathcal{W}_2^{\text{per}}$ be the second-order Sobolev space of periodic functions on $[0, 1]$, endowed with the norm $\|b\|_{\mathcal{H}}^2 = \left[\int_0^1 b(t)dt\right]^2 + \int_0^1 [b''(t)]^2\, dt$ and the reproducing kernel $K(s, t) = 1 - \frac{1}{24}B_4(|s - t|)$, where B_4 is the fourth Bernoulli polynomial (Wahba, 1990).

Estimates obtained using our proposed method at the expectile levels $\tau = 0.1, \ldots, 0.9$ are shown in Figure 6. As expected, for any fixed t, $\hat{\beta}_\tau(t)$ increases with τ. For the sake of practical interpretation, it is useful that these functional

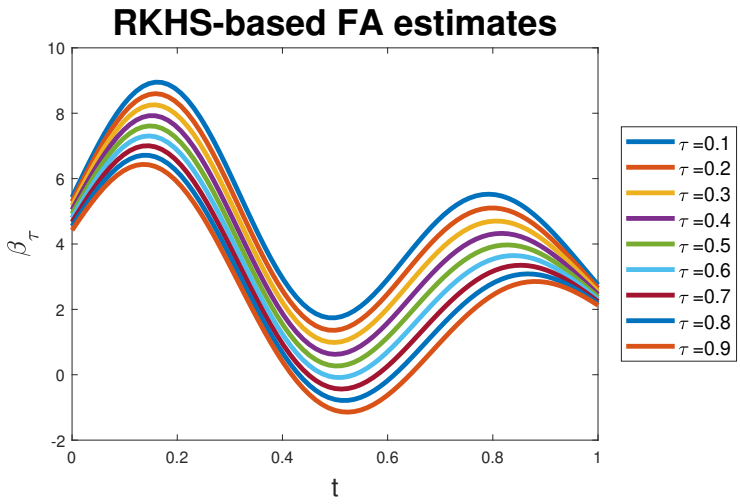

FIGURE 6: RKHS-based estimates $\hat{\beta}_\tau$ at $\tau = 0.1, \ldots, 0.9$ in the ADNI data analysis, describing the functional effect of FA on MMSE score.

estimates do not cross each other.

We also considered FPCA-based estimates obtained using 4, 6, 8, and 10 functional principal components. These estimates, illustrated in Figure 7, are clearly not ideal for at least a couple reasons. First, the FPCA-based estimates cross each other, unlike the RKHS-based estimates in Figure 6. This "crossing problem" is further discussed in He (1997) in the context of quantile regression. Second, the FPCA-based estimates are sensitive to the user-specified number of principal components. The discrete nature of this hyperparameter makes it difficult to tune finely, unlike the continuous hyperparameter λ in our RKHS-based approach.

Table 1 moreover shows that, at each expectile level considered, the proposed RKHS-based estimator outperforms the FPCA-based one in predicting MMSE. These results emphasize the practical importance and advantages of our RKHS-

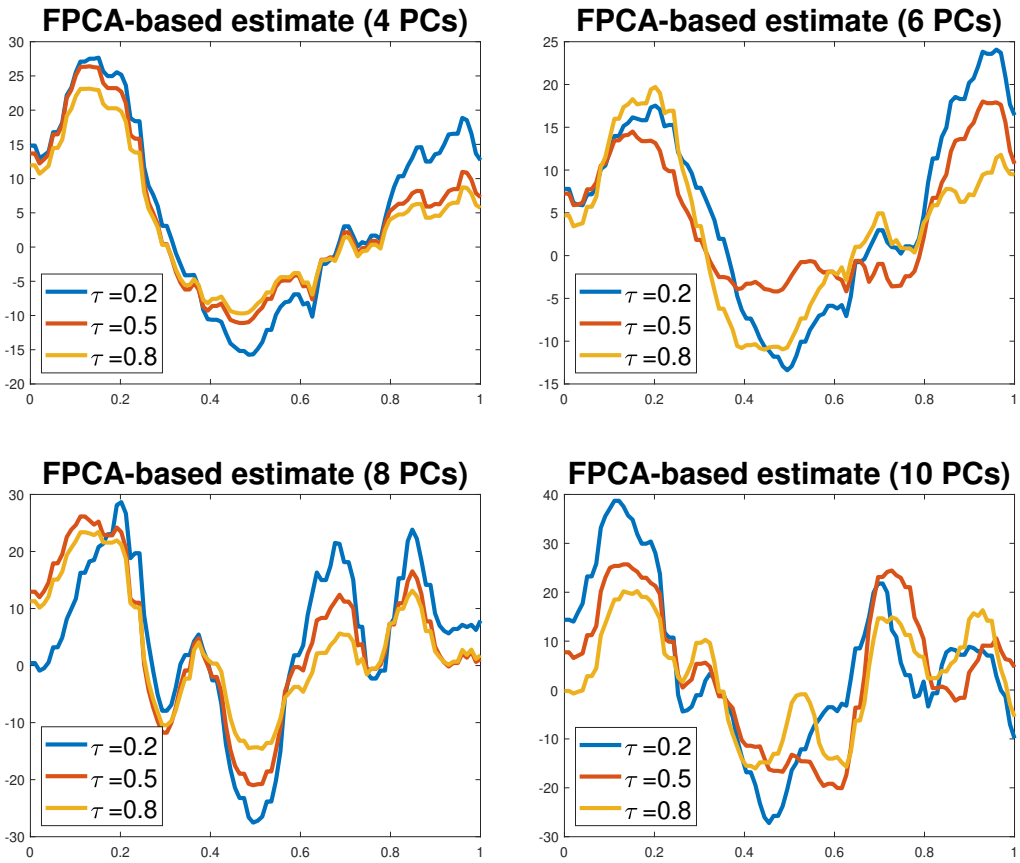

FIGURE 7: FPCA-based estimates $\hat{\beta}_\tau$ at $\tau = 0.2, 0.5, 0.8$ in the ADNI data analysis, describing the functional effect of FA on MMSE score. The number of functional principal components (PCs) used is indicated in each subplot: 4, 6, 8, and 10 PCs explain 79.9%, 86.0%, 89.3%, and 91.5%, respectively, of the observed variance in functional FA.

based approach in functional linear expectile regression.

As an informal aside (due to the computation time involved), we also compared the computational efficiency of different implementations of our proposed RKHS-based estimator. Our first implementation is as presented in Section 3.3 using the ADMM algorithm while the second uses an interior point (IP) algorithm (Mehrotra, 1992). The latter is popularly applied to constrained optimiza-

TABLE 1: Test set prediction error in the ADNI analysis for the RKHS- and FPCA-based predictors at $\tau = 0.1, \ldots, 0.9$.

Expectile level τ	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
RKHS PE (SE)	0.9954	0.9936	0.9922	0.9913	0.9907	0.9905	0.9907	0.9911	0.9918
	(0.0153)	(0.0152)	(0.0151)	(0.015)	(0.0150)	(0.0149)	(0.0149)	(0.0148)	(0.0148)
FPCA PE (SE)	1.0164	1.0161	1.0058	1.0006	0.9999	0.9982	1.0000	1.0004	1.0008
	(0.0174)	(0.0174)	(0.0163)	(0.0164)	(0.0161)	(0.016)	(0.0160)	(0.0161)	(0.0157)

tion problems. We found our ADMM implementation to be far superior to the IP implementation: the latter typically requires at least 100 times more computation time than the former until convergence. These results can be made available on request.

5. DISCUSSION

In this paper, we proposed a regularized estimator for the functional linear expectile regression model under an RKHS framework. We derived upper and lower bounds for the minimax rate of convergence of prediction error and established the minimax optimality of our proposed estimator. While most existing approaches to functional linear expectile regression rely on FPCA, we argue that these approaches are too restrictive in their assumption regarding eigenvalue spacing. Additionally, FPCA-based methods rely on the assumption that leading principal components (which are determined by only the functional predictor X and not the response Y) are predictive of the response: in practice, this assumption is typically not valid.

We demonstrated the general superiority of our proposed RKHS-based approach in three sets of simulation studies and an application to an ADNI neuroimaging dataset. In particular, we illustrated the degradation of FPCA-based estimators when its implicit assumptions regarding the eigenstructures of the reproducing and covariance kernels are violated. Our results showed that both eigenfunction alignment and eigenvalue decay rates between the reproducing and covariance kernels have an important impact on estimator performance.

For the sake of illustration, we focused on a univariate functional predictor X with a domain $\mathcal{T}$ that is a compact subset of $\mathbb{R}$. We took $\mathcal{T} = [0, 1]$ and used the corresponding canonical Sobolev space as a working example. Our theoretical results apply nonetheless to more general RKHSs, provided that $\mathcal{T}$ remains a compact subset of an arbitrary Euclidean space. For example, the derived optimal convergence rate still holds for Sobolev spaces on $\mathcal{T} = [0, 1]^2$, e.g., for imaging data, with the decay rate r determined by the corresponding reproducing and covariance kernels. The developments in this article thus have wide applications in spatial statistics, 2D and 3D image analysis, and longitudinal data analysis.

Settings where the reproducing and covariance kernels are not well aligned (i.e., in the sense of their eigenfunctions) are interesting topics for future work. As suggested by our ADNI analysis, another natural generalization of our approach is the inclusion of scalar predictors, e.g., age, gender, and diagnosis status, for a partial functional expectile regression model. While it is straightforward to

accommodate scalar covariate effect estimation from an algorithmic perspective, the optimality of the corresponding estimators requires more work to establish. Informally (and with results available on request), PE for the RKHS- and FPCA-based estimators are comparable when scalar age, gender, and diagnosis status effects are included in the model. We suspect that this decrease in relative PE can be attributed to the relative complexity of the two models and possibly the overwhelming usefulness of these scalar covariates as predictors. We feel that the full impact of scalar predictors on empirical performance, such as in high-dimensional settings, should be investigated in future work.

Acknowledgements

This research was partially supported by the National Social Science Fund of China (19BTJ034), the National Natural Science Foundation of China (12171242), and the China Postdoctoral Science Foundation (2018T110422 and 2016M590396). We are grateful to the editor, associate editor, and two anonymous referees for their many helpful comments and suggestions.

BIBLIOGRAPHY

Boyd, S. (2010), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122.

Brézis, H. (2011), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, NY.

Cai, T. T. and Hall, P. (2006), 'Prediction in functional linear regression', *The Annals of Statistics* **34**(5), 2159–2179.

Cai, T. T. and Yuan, M. (2011), 'Optimal estimation of the mean function based on discretely sampled functional data: Phase transition', *The Annals of Statistics* **39**(5), 2330–2355.

Cai, T. T. and Yuan, M. (2012), 'Minimax and adaptive prediction for functional linear regression', *Journal of the American Statistical Association* **107**(499), 1201–1216.

Chen, C. and Wei, Y. (2005), 'Computational issues for quantile regression', *Sankhyā: The Indian Journal of Statistics* **67**(2), 399–417.

Cheng, G. and Shang, Z. (2015), 'Joint asymptotics for semi-nonparametric regression models with partially linear structure', *The Annals of Statistics* **43**(3), 1351–1390.

Conrad, K. (2014), The minimal polynomial and some applications, Technical report, Department of Mathematics, University of Connecticut.
URL: https://kconrad.math.uconn.edu/blurbs/linmultialg/minpolyandappns.pdf

Crambes, C., Kneip, A. and Sarda, P. (2009), 'Smoothing splines estimators for functional linear regression', *The Annals of Statistics* **37**(1), 35–72.

Donoho, D. L. and Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* **90**(432), 1200–1224.

Golub, G. H., Heath, M. and Wahba, G. (1979), 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics* **21**(2), 215–223.

Gu, C. (2013), *Smoothing Spline ANOVA Models*, Springer, New York, NY.

Gu, Y., Fan, J., Kong, L., Ma, S. and Zou, H. (2018), 'ADMM for high-dimensional sparse penalized quantile regression', *Technometrics* **60**(3), 319–331.

Gu, Y. and Hui, Z. (2016), 'High-dimensional generalizations of asymmetric least squares regression and their applications', *The Annals of Statistics* **44**(6), 2661–2694.

Guo, M., Zhou, L., Huang, J. Z. and Härdle, W. K. (2015), 'Functional data analysis of generalized regression quantiles', *Statistics and Computing* **25**(2), 189–202.

Hall, P. and Horowitz, J. L. (2007), 'Methodology and convergence rates for functional linear regression', *The Annals of Statistics* **35**(1), 70–91.

He, X. (1997), 'Quantile curves without crossing', *The American Statistician* **51**(2), 186–192.

Holzmann, H. and Klar, B. (2016), 'Expectile asymptotics', *Electronic Journal of Statistics* **10**(2), 2355–2371.

Jack, Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C. et al. (2008), 'The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods', *Journal of Magnetic Resonance Imaging* **27**(4), 685–691.

James, G. M., Wang, J. and Zhu, J. (2009), 'Functional linear regression that's interpretable', *The Annals of Statistics* **37**(5A), 2083–2108.

Jones, M. C. (1994), 'Expectiles and M-quantiles are quantiles', *Statistics & Probability Letters* **20**(2), 149–153.

Kato, K. (2012), 'Estimation in functional linear quantile regression', *The Annals of Statistics* **40**(6), 3108–3136.

Koenker, R. (2017), 'Quantile regression: 40 years on', *Annual Review of Economics* **9**, 155–176.

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2', Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 1137–1143.

Krätschmer, V. and Zähle, H. (2017), 'Statistical inference for expectile-based risk measures', *Scandinavian Journal of Statistics* **44**(2), 425–454.

Li, G., Huang, J. Z. and Shen, H. (2018), 'To wait or not to wait: Two-way functional hazards model for understanding waiting in call centers', *Journal of the American Statistical Association* **113**(524), 1503–1514.

Li, Y., Liu, Y. and Zhu, J. (2007), 'Quantile regression in reproducing kernel Hilbert spaces', *Journal of the American Statistical Association* **102**(477), 255–268.

Li, Z. and Yao, J. (2019), 'Testing for heteroscedasticity in high-dimensional regressions', *Econometrics and Statistics* **9**, 122–139.

Liao, L., Park, C. and Choi, H. (2019), 'Penalized expectile regression: An alternative to penalized quantile regression', *Annals of the Institute of Statistical Mathematics* **71**(2), 409–438.

Ma, C., Lu, J. and Liu, H. (2020), 'Inter-subject analysis: A partial Gaussian graphical model approach', *Journal of the American Statistical Association* **116**, 1–57.

Mehrotra, S. (1992), 'On the implementation of a primal-dual interior point method', *SIAM Journal on Optimization* **2**(4), 575–601.

Micchelli, C. A. and Wahba, G. (1979), Design problems for optimal surface interpolation, Technical Report ADA070012, Department of Statistics, University of Wisconsin–Madison. **URL:** https://apps.dtic.mil/sti/citations/ADA070012

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018), *Foundations of Machine Learning*, MIT press, Cambridge, MA.

Newey, W. and Powell, J. (1987), 'Asymmetric least squares estimation and testing', *Econometrica* **55**(4), 819–847.

Nosedal-Sanchez, A., Storlie, C. B., Lee, T. C. and Christensen, R. (2012), 'Reproducing kernel Hilbert spaces for penalized regression: A tutorial', *The American Statistician* **66**(1), 50–60.

Pietrosanu, M., Gao, J., Kong, L., Jiang, B. and Niu, D. (2020), 'Advanced algorithms for penalized quantile and composite quantile regression', *Computational Statistics* **36**, 333–346.

Pietrosanu, M., Shu, H., Jiang, B., Kong, L., Heo, H., He, Q., Gilmore, J. and Zhu, H. (2021), 'Estimation for the bivariate quantile varying coefficient model with application to diffusion tensor imaging data analysis', *Biostatistics* p. kxab031.

Qu, S., Wang, J.-L. and Wang, X. (2016), 'Optimal estimation for the functional Cox model', *The Annals of Statistics* **44**(4), 1708–1738.

Ramsay, J. O. and Silverman, B. W. (2006), *Functional Data Analysis*, Springer, New York, NY.

Ritter, K., Wasilkowski, G. W. and Wozniakowski, H. (1995), 'Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions', *The Annals of Applied Probability* **5**(2), 518–540.

Schnabel, S. K. and Eilers, P. H. (2009), 'Optimal expectile smoothing', *Computational Statistics & Data Analysis* **53**(12), 4168–4177.

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, Springer Science & Business Media, New York, NY.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Waltrup, L. S., Sobotka, F., Kneib, T. and Kauermann, G. (2015), 'Expectile and quantile regression—David and Goliath?', *Statistical Modelling* **15**(5), 433–456.

Wang, Y., Kong, L., Jiang, B., Zhou, X., Yu, S., Zhang, L. and Heo, G. (2019), 'Wavelet-based lasso in functional linear quantile regression', *Journal of Statistical Computation and Simulation* **89**(6), 1111–1130.

Yu, D., Zhang, L., Mizera, I., Jiang, B. and Kong, L. (2019), 'Sparse wavelet estimation in quantile regression with multiple functional predictors', *Computational Statistics & Data Analysis* **136**, 12–29.

Yuan, M. and Cai, T. T. (2010), 'A reproducing kernel Hilbert space approach to functional linear regression', *The Annals of Statistics* **38**(6), 3412–3444.

APPENDIX

Proof of Theorem 1. Recall the functional model $Y = \int_{\mathcal{T}} X(t)\beta_0(t)\,\mathrm{d}t + \varepsilon$ specified in the main text. Fix an expectile level $\tau \in (0,1)$ and assume that ε follows an asymmetric normal distribution with the density function

$$f(\varepsilon) = \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}} \frac{1}{\sqrt{\pi\sigma^2}} \exp\{-r_\tau(\varepsilon/\sigma)\}, \tag{1}$$

where $r_\tau(u) = |\tau - I(u<0)|u^2$. Further assume that β_0 belongs to an RKHS $\mathcal{H}(K)$.

Consider the functional space

$$\mathcal{H}^* = \left\{ \beta = \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}}\zeta_k : (b_{M+1},\ldots,b_{2M}) \in \{0,1\}^M \right\},$$

where $(\zeta_k)_{k\in\mathbb{N}}$ is a sequence of orthonormal eigenfunctions of $K^{1/2}CK^{1/2}$. The function $\|\cdot\|_K$ is a semi-norm on $\mathcal{H}(K)$ and M is some large number to be discussed later. For any $\beta \in \mathcal{H}^*$, observe that

$$\begin{aligned} J(\beta) = \|\beta\|_K^2 &= \left\| \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}}\zeta_k \right\|_K^2 \\ &= \sum_{k=M+1}^{2M} b_k^2 M^{-1} \left\| L_{K^{1/2}}\zeta_k \right\|_K^2 \\ &\le \sum_{k=M+1}^{2M} M^{-1} \left\| L_{K^{1/2}}\zeta_k \right\|_K^2 \\ &= 1, \end{aligned}$$

which follows from the fact that $\langle L_{K^{1/2}}\zeta_k, L_{K^{1/2}}\zeta_l\rangle_K = \langle L_K\zeta_k, \zeta_l\rangle_K = \langle \zeta_k, \zeta_l\rangle_{\mathcal{L}_2} = \delta_{kl}$. Therefore, $\mathcal{H}^* \subset \mathcal{H}(K) = \{\beta : \|\beta\|_K < \infty\}$.

The Gilbert-Varshamov bound (Tsybakov, 2008, Lemma 2.9) establishes that, for any $M \geq 8$, there exists a set $\{b^{(0)}, b^{(1)}, \ldots, b^{(N)}\} \subset \{0,1\}^M$ such that

(i) $b^{(0)} = (0, \ldots, 0)^\top$;

(ii) $H(b^{(i)}, b^{(j)}) \geq M/8$ for any distinct $b^{(i)}, b^{(j)} \in \mathcal{B}$, where $H(\cdot\,,\cdot\,)$ denotes Hamming distance; and

(iii) $N \geq 2^{M/8}$.

Define the subset

$$\mathcal{B} = \left\{ \beta^{(0)}, \ldots, \beta^{(N)} : \beta^{(i)} = \sum_{k=M+1}^{2M} b^{(i)}_{k-M} M^{-1/2} L_{K^{1/2}} \zeta_k, i = 1, \ldots, N \right\} \subset \mathcal{H}^*$$

and let M be the smallest integer greater than $c_0 n^{1/(2r+1)}$ for some constant $c_0 > 0$. Then for i and j satisfying $0 \leq i \leq j \leq N$,

$$\begin{aligned}
\left\|\beta^{(i)} - \beta^{(j)}\right\|_C^2 &= \left\| L_{C^{1/2}} \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)}) M^{-1/2} L_{K^{1/2}\zeta_k} \right\|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \| L_{C^{1/2}} L_{K^{1/2}\zeta_k} \|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \nu_k. \\
&\geq \nu_{2M} M^{-1} \sum_{k=1}^{M} (b_k^{(i)} - b_k^{(j)})^2 \\
&= 4\nu_{2M} M^{-1} H(b^{(i)}, b^{(j)}) \\
&\geq \nu_{2M}/2 \\
&\geq c_1 2^{-(2r+1)} M^{-2r} \\
&\geq 2c\alpha^{2r/(2r+1)} n^{-2r/(2r+1)},
\end{aligned}$$

where $c > 0$ is some constant.

We apply the results of Tsybakov (2008) to establish a lower bound based on multiple hypothesis testing. Under the assumption that the slope function β_0 belongs to the subset $\mathcal{B}$, we construct a subset $\{\beta^{(0)}, ..., \beta^{(N)}\} \subset \mathcal{H}^*$ with N increasing in n such that, for some positive constant c and for i and j such that $0 \leq i \leq j \leq N$,

$$\|\beta^{(i)} - \beta^{(j)}\|_C^2 \geq c\alpha^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}} \tag{2}$$

and

$$\frac{1}{N}\sum_{j=1}^{N}\mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) \leq \alpha \log N, \tag{3}$$

where P_β denotes the joint conditional distribution of Y given X and KL represents Kullback-Leibler divergence. By Theorem 2.5 of Tsybakov (2008), it follows that

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathcal{H}^*} \mathbb{P}(\|\beta^{(i)} - \beta^{(j)}\|_C^2 \geq c\alpha^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}}) \geq \frac{\sqrt{N}}{\sqrt{N}+1}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log N}}\right). \tag{4}$$

Note that $M, N \to \infty$ as $n \to \infty$. This implies that the right-hand side of (4) can be made arbitrarily close to 1 as $n \to \infty$ and $\alpha \to 0$. We conclude that

$$\lim_{\alpha \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}} \sup_{\beta_0 \in \mathcal{H}^*} \mathbb{P}(\|\beta^{(i)} - \beta^{(j)}\|_C^2 \geq a n^{-\frac{2r}{2r+1}}) = 1. \tag{5}$$

This lower bound for the asymmetric normal distribution yields a lower bound for general error distributions. Let P_j, for $j = 1, \ldots, N$, represent the joint distribution of the observed sample $\{(x_k, y_k) : k = 1, \ldots, n\}$ under the assumption that $\beta_0 = \beta^{(j)}$. It follows that

$$P_j = \prod_{k=1}^{n} \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}} \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-r_\tau\left(\frac{y_k - \int_{\mathcal{T}} x_k(t)^\top \beta^{(j)}(t)}{\sigma}\right)\right\}. \tag{6}$$

The Kullback-Leibler divergence between $P_{\beta^{(i)}}$ and $P_{\beta^{(j)}}$ is

$$\begin{aligned}\mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) &= E_{\beta^{(i)}} \log(P_{\beta^{(i)}}/P_{\beta^{(j)}}) \\ &= nE_{\beta^{(i)}}\left[r_\tau\left(\frac{Y - \int_{\mathcal{T}} X(t)\beta^{(j)}(t)\ \mathrm{d}t}{\sigma}\right) - r_\tau\left(\frac{Y - \int_{\mathcal{T}} X(t)^\top\beta^{(i)}(t)\ \mathrm{d}t}{\sigma}\right)\right] \\ &\leq n\max(\tau, 1-\tau)\left(\int_{\mathcal{T}} X(t)^\top(\beta^{(j)}(t) - \beta^{(i)}(t))\ \mathrm{d}t\right)^2.\end{aligned}$$

The inequality above holds since, defining $\mu^i = \int_{\mathcal{T}} X(t)^\top\beta^{(i)}(t)\ \mathrm{d}t$,

$$\begin{aligned}&E_{\beta^{(i)}}\left[r_\tau\left(\frac{Y-\mu^j}{\sigma}\right) - r_\tau\left(\frac{Y-\mu^i}{\sigma}\right)\right] \\ &\quad= \int_{\mu^i}^{\infty} \tau\left[\left(\frac{y-\mu^j}{\sigma}\right)^2 - \left(\frac{y-\mu^i}{\sigma}\right)^2\right] f(y-\mu^i)\ \mathrm{d}y \\ &\qquad+ \int_{-\infty}^{\mu^i} (1-\tau)\left[\left(\frac{y-\mu^j}{\sigma}\right)^2 - \left(\frac{y-\mu^i}{\sigma}\right)^2\right] f(y-\mu^i)\ \mathrm{d}y \\ &\qquad+ \int_{\mu^j}^{\mu^i} (2\tau-1)\left(\frac{y-\mu^j}{\sigma}\right)^2 f(y-\mu^i)\ \mathrm{d}y,\end{aligned}$$

where

$$\int_{\mu^j}^{\mu^i} (2\tau-1)\left(\frac{y-\mu^j}{\sigma}\right)^2 f(y-\mu^i)\ \mathrm{d}y \leq |1-2\tau|\left(\frac{\mu^i-\mu^j}{\sigma}\right)^2 \int_{\mu^j}^{\mu^i} f(y-\mu^i)\ \mathrm{d}y.$$

Thus,

$$\begin{aligned}
\mathrm{KL}(P_{\beta^{(i)}} \mid P_{\beta^{(j)}}) &\leq n \max(\tau, 1-\tau) \left(\int_{\mathcal{T}} X_k(t)^\top (\beta^{(j)}(t) - \beta^{(i)}(t)) \, \mathrm{d}t \right)^2 \\
&= n \max(\tau, 1-\tau) \| L_{c^{1/2}} (\beta^{(j)}(t) - \beta^{(i)}(t)) \|_{\mathcal{L}_2}^2 \\
&= n \max(\tau, 1-\tau) \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} \nu_k \\
&\leq n \max(\tau, 1-\tau) \nu_M M^{-1} \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 \\
&= 4n \max(\tau, 1-\tau) \nu_M M^{-1} H(b^{(i)}, b^{(j)}) \\
&\leq 4n \max(\tau, 1-\tau) \nu_M \\
&\leq 4c_2 n \max(\tau, 1-\tau) M^{-2r}.
\end{aligned}$$

Consequently, when $0 < \alpha < 1/8$,

$$\frac{1}{N} \sum_{j=1}^{N} KL(P_j \mid P_0) \leq 4c_2 n \max(\tau, 1-\tau) M^{-2r} \leq \alpha \log 2^{M/8} \leq \alpha \log N.$$

By taking M to be the smallest integer greater than $c_2 \alpha^{-1/(2r+1)} n^{1/(2r+1)}$ with $c_2 = (8c_1 \log 2)^{1/(2r+1)}$, the desired result follows. ■

Proof of Theorem 2. Recall that $L_{K^{1/2}}(\mathcal{L}_2) = \mathcal{H}(K)$. Therefore, there exist $f_0, \hat{f} \in \mathcal{L}_2$ such that $\beta_0 = L_{K^{1/2}} f_0$ and $\hat{\beta}_\lambda = L_{K^{1/2}} \hat{f}_\lambda$. For brevity, we assume that $\mathcal{H}(K)$ is dense in $\mathcal{L}_2$, which ensures that f_0 and $\hat{f}_\lambda$ are uniquely defined. The proof in the general case proceeds in exactly the same fashion by restricting consideration to $\mathcal{L}_2 / \ker(L_{K^{1/2}})$.

For brevity, define $T = L_{K^{1/2}CK^{1/2}}$. Let T^ν denote a linear operator from $\mathcal{L}_2$ to $\mathcal{L}_2$ such that $T^\nu \varphi_k = s_k^\nu \varphi_k$. Prediction error can then be written as

$$\|\hat{\beta} - \beta_0\|_C^2 = \left\| T^{1/2} \left(\hat{f}_\lambda - f_0 \right) \right\|_{\mathcal{L}_2}^{2.}$$

and, furthermore,

$$\hat{f}_\lambda = \underset{f \in \mathcal{L}_2}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} r_\tau \left(y_i - \langle x_i, L_{K^{1/2}} f \rangle_{\mathcal{L}_2} \right)^2 + \lambda \|f\|_{\mathcal{L}_2}^2 \right].$$

Recalling that $y_i = \langle x_i, L_{K^{1/2}} f_0 \rangle_{\mathcal{L}_2} + \varepsilon_i$,

$$C_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} e_i x_i(s) x_i(t),$$

where $e_i = \tau$ if $y_i \geq \left\langle x_i, L_{K^{1/2}} \hat{f}_\lambda \right\rangle_{\mathcal{L}_2}$ and $e_i = 1 - \tau$ otherwise. Define $T_n = L_{K^{1/2}} L_{C_n} L_{K^{1/2}}$, where L_{C_n} is an integral operator such that, for any $h \in \mathcal{L}_2$,

$$L_{C_n} h(\cdot) = \int_{\mathcal{T}} C_n(s, \cdot) h(s) \, \mathrm{d}s.$$

Consequently, $\hat{f}_\lambda = (T_n + \lambda \mathbf{1})^{-1} (T_n f_0 + g_n)$, where $\mathbf{1}$ is the identity operator and $g_n = \frac{1}{n} \sum_{i=1}^{n} e_i \epsilon_i L_{K^{1/2}} x_i$.

Next, define $f_\lambda = (T + \lambda \mathbf{1})^{-1} T f_0$. By the triangle inequality,

$$\left\| T^{1/2} \left(\hat{f}_\lambda - f_0 \right) \right\|_{\mathcal{L}_2} = \left\| T^{1/2} (f_\lambda - f_0) \right\|_{\mathcal{L}_2} + \left\| T^{1/2} \left(\hat{f}_\lambda - f_\lambda \right) \right\|_{\mathcal{L}_2}. \quad (7)$$

The first term on the right-hand side can be easily bounded. To proceed, we appeal to the following lemma.

Lemma A1. *For* $0 < \nu < 1$, $\left\| T^v (f_\lambda - f_0) \right\|_{\mathcal{L}_2} \leq (1 - \nu)^{1-\nu} \nu^\nu \lambda^\nu \left\| f_0 \right\|_{\mathcal{L}_2}$.

Taking $\nu = 1/2$ in Lemma A1 establishes that $\left\|T^{1/2}\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2}^2 \leq \frac{1}{4}\lambda \left\|f_0\right\|_{\mathcal{L}_2}^2$.

We now turn to the second term on the right-hand side of Equation (7). Observe that

$$f_\lambda - \hat{f}_\lambda = \left(T + \lambda \mathbf{1}\right)^{-1}\left(T_n + \lambda \mathbf{1}\right)\left(f_\lambda - \hat{f}_\lambda\right) + \left(T + \lambda \mathbf{1}\right)^{-1}\left(T - T_n\right)\left(f_\lambda - \hat{f}_\lambda\right)$$

and that $\left(T_n + \lambda \mathbf{1}\right)\hat{f}_\lambda = T_n f_0 - g_n$. Therefore,

$$\begin{aligned} f_\lambda - \hat{f}_\lambda &= \left(T + \lambda \mathbf{1}\right)^{-1} T_n\left(f_\lambda - f_0\right) + \lambda\left(T + \lambda \mathbf{1}\right)^{-1} f_\lambda + \left(T + \lambda \mathbf{1}\right)^{-1} g_n \\ &\quad + \left(T + \lambda \mathbf{1}\right)^{-1}\left(T - T_n\right)\left(f_\lambda - \hat{f}_\lambda\right) \\ &= \left(T + \lambda \mathbf{1}\right)^{-1} T\left(f_\lambda - f_0\right) + \left(T + \lambda \mathbf{1}\right)^{-1}\left(T_n - T\right)\left(f_\lambda - f_0\right) + \lambda\left(T + \lambda \mathbf{1}\right)^{-1} f_\lambda \\ &\quad + \left(T + \lambda \mathbf{1}\right)^{-1} g_n + \left(T + \lambda \mathbf{1}\right)^{-1}\left(T - T_n\right)\left(f_\lambda - \hat{f}_\lambda\right) \end{aligned}$$

We first consider bounding $\left\|T^\nu\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2}$ for some $\nu \in (0, 1/2 - 1/(4r))$. By the triangle inequality,

$$\begin{aligned} \left\|T^\nu\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2} \leq{}& \left\|T^\nu\left(T + \lambda \mathbf{1}\right)^{-1} T\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\ &+ \left\|T^\nu\left(T + \lambda \mathbf{1}\right)^{-1}\left(T_n - T\right)\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\ &+ \lambda\left\|T^\nu\left(T + \lambda \mathbf{1}\right)^{-1} f_\lambda\right\|_{\mathcal{L}_2} + \left\|T^\nu\left(T + \lambda \mathbf{1}\right)^{-1} g_n\right\|_{\mathcal{L}_2} \\ &+ \left\|T^\nu\left(T + \lambda \mathbf{1}\right)^{-1}\left(T - T_n\right)\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2}. \end{aligned}$$

Lemma A2. *Assume that there exists a constant $c_3 > 0$ such that, for any $f \in \mathcal{L}_2$, $E\langle X, f\rangle_{\mathcal{L}_2}^4 \leq c_3(E\langle X, f\rangle_{\mathcal{L}_2}^2)^2$. Then for any $\nu > 0$ such that $2r(1-2\nu) > 1$,*

$$\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n - T\right)T^{-\nu}\right\|_{op} = O_p\left(\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right),$$

where $\|\cdot\|_{op}$ denotes the usual operator norm, i.e., $\|U\|_{op} = \sup_{\{h:\|h\|_{\mathcal{L}_2}=1\}}\|Uh\|_{\mathcal{L}_2}$ for an operator $U : \mathcal{L}_2 \to \mathcal{L}_2$.

By an application of Lemma A2,

$$\begin{aligned}
&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T - T_n\right)\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2} \\
&\qquad \leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T - T_n\right)T^{-\nu}\right\|_{\text{op}} \left\|T^{\nu}\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2} \\
&\qquad \leq o_p(1)\left\|T^{\nu}\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2}
\end{aligned}$$

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

$$\begin{aligned}
&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n - T\right)\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\
&\qquad \leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}\left(T_n - T\right)T^{-\nu}\right\|_{\text{op}} \left\|T^{\nu}\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\
&\qquad \leq o_p(1)\left\|T^{\nu}\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\|T^{\nu}\left(f_\lambda - \hat{f}_\lambda\right)\right\|_{\mathcal{L}_2} = O_p\Big(&\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\
&+ \lambda\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}f_\lambda\right\|_{\mathcal{L}_2} + \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_n\right\|_{\mathcal{L}_2}\Big).
\end{aligned}$$

By Lemma A1,

$$\begin{aligned}\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2} &\leq \left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}T^{1-\nu}\right\|_{\text{op}}\left\|T^{\nu}\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2}\\ &\leq \left\|T^{\nu}\left(f_{\lambda}-f_{0}\right)\right\|_{\mathcal{L}_2}\\ &\leq (1-\nu)^{1-\nu}\nu^{\nu}\lambda^{\nu}\left\|f_{0}\right\|_{\mathcal{L}_2}.\end{aligned}$$

Lemma A3. *When* $0 \leq \nu \leq 1/2$*,*

$$\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_{n}\right\|_{L_2} = O_{p}\left(\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right).$$

Lemma A3 and the preceding result imply that

$$\left\|T^{\nu}\left(f_{\lambda}-\hat{f}_{\lambda}\right)\right\|_{\mathcal{L}_2} = O_{p}\left(\lambda^{\nu}+\left(n\lambda^{1-2\nu+1/(2r)}\right)^{-1/2}\right) = O_{p}\left(\lambda^{\nu}\right),$$

provided that $c_1 n^{-2r/(2r+1)} \leq \lambda \leq c_2 n^{-2r/(2r+1)}$ for some constants c_1 and c_2 satisfying $0 < c_1 < c_2 < \infty$.

Recall that

$$\begin{aligned}\left\|T^{1/2}\big(f_{\lambda}-\hat{f}_{\lambda}\big)\right\|_{\mathcal{L}_2} =& \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T\big(f_{\lambda}-f_{0}\big)\right\|_{\mathcal{L}_2}\\ &+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_{n}-T\big)\big(f_{\lambda}-f_{0}\big)\right\|_{\mathcal{L}_2}\\ &+\lambda\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}f_{\lambda}\right\|_{\mathcal{L}_2}+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}g_{n}\right\|_{\mathcal{L}_2}\\ &+\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T-T_{n}\big)\big(f_{\lambda}-\hat{f}_{\lambda}\big)\right\|_{\mathcal{L}_2},\end{aligned}$$

so we can bound $\left\|T^{1/2}\left(f_\lambda - \hat{f}_\lambda\right)\right\|$ by bounding the five terms on the right-hand side of the above equation. By Lemma A1,

$$\begin{aligned}\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} &\leq \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T^{1/2}\right\|_{\text{op}}\left\|T^{1/2}\left(f_\lambda - f_0\right)\right\|_{\mathcal{L}_2} \\ &\leq \frac{1}{2}\lambda^{1/2}\left\|f_0\right\|_{\mathcal{L}_2}.\end{aligned}$$

Lemma A4. *Under the conditions of Lemma A2,*

$$\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\left(T_n - T\right)T^{-\nu}\right\|_{op} = O_p\left(\left(n\lambda^{1/(2r)}\right)^{-1/2}\right).$$

By Lemmas A1 and A4,

$$\begin{aligned}&\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n - T\big)\big(f_\lambda - f_0\big)\right\|_{\mathcal{L}_2} \\ &\quad\leq \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n - T\big)T^{-\nu}\right\|_{\text{op}}\left\|T^{\nu}\big(f_\lambda - f_0\big)\right\|_{\mathcal{L}_2} \\ &\quad\leq O_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\lambda^{\nu}\big) \\ &\quad= o_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\big).\end{aligned}$$

Similarly,

$$\begin{aligned}&\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n - T\big)\big(f_\lambda - \hat{f}_\lambda\big)\right\|_{\mathcal{L}_2} \\ &\quad\leq \left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}\big(T_n - T\big)T^{-\nu}\right\|_{\text{op}}\left\|T^{\nu}\big(f_\lambda - \hat{f}_\lambda\big)\right\|_{\mathcal{L}_2} \\ &\quad\leq O_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\lambda^{\nu}\big) \\ &\quad= o_p\big(\big(n\lambda^{1/(2r)}\big)^{-1/2}\big).\end{aligned}$$

By Lemma A3, $\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}g_n\right\|_{\mathcal{L}_2} = O_p\left(\left(n\lambda^{1/(2r)}\right)^{-1/2}\right)$.

Finally, together with the fact that $\lambda \left\| T^{1/2}(T+\lambda \mathbf{1})^{-1} f_\lambda \right\|_{\mathcal{L}_2} = O(\lambda)$, we conclude that $\|T^{1/2}(f_\lambda - \hat{f}_\lambda)\|_{\mathcal{L}_2} = O_p(n^{-\frac{2r}{2r+1}})$, as desired. ■

Received 31 January 2021

Accepted 14 September 2021