Bernoulli **30**(3), 2024, 1800–1824
https://doi.org/10.3150/23-BEJ1653

Functional linear quantile regression on a two-dimensional domain

NAN ZHANG[1,a], PENG LIU[2,b], LINGLONG KONG[3,c], BEI JIANG[3,d] and JIANHUA Z. HUANG[4,e]

[1]*School of Data Science, Fudan University, Shanghai, China,* [a]*zhangnan@fudan.edu.cn*
[2]*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK,* [b]*p.liu@kent.ac.uk*
[3]*Department of Mathematical and Statistical Sciences University of Alberta, Edmonton, Canada,* [c]*lkong@ualberta.ca,* [d]*bei1@ualberta.ca*
[4]*School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China,* [e]*jhuang@cuhk.edu.cn*

This article considers the functional linear quantile regression which models the conditional quantile of a scalar response given a functional predictor over a two-dimensional domain. We propose an estimator for the slope function by minimizing the penalized empirical check loss function. Under the framework of reproducing kernel Hilbert space, the minimax rate of convergence for the regularized estimator is established. Using the theory of interpolation spaces on a two- or multi-dimensional domain, we develop a novel result on simultaneous diagonalization of the reproducing and covariance kernels, revealing the interaction of the two kernels in determining the optimal convergence rate of the estimator. Sufficient conditions are provided to show that our analysis applies to many situations, for example, when the covariance kernel is from the Matérn class, and the slope function belongs to a Sobolev space. We implement the interior point method to compute the regularized estimator and illustrate the proposed method by applying it to the hippocampus surface data in the ADNI study.

Keywords: Functional linear regression; multi-dimensional domain; quantile regression; rate of convergence; reproducing kernel Hilbert space; simultaneous diagonalization

1. Introduction

Functional linear models are widely used in functional data analysis to associate response and covariate variables. They can be viewed as an extension of the traditional multiple linear regression models to situations involving functional objects (Ramsay and Dalzell, 1991, Ramsay and Silverman, 2005, Wang, Chiou and Müller, 2016). Types of functional regression models are classified by whether the response or covariate is functional or scalar, including functional predictor regression (Cai and Hall, 2006, Cardot, Ferraty and Sarda, 2003, Crambes, Kneip and Sarda, 2009, Hall and Horowitz, 2007, Yuan and Cai, 2010), functional response regression (Liu, Li and Morris, 2020, Zhang et al., 2022), and function-on-function regression (Sun et al., 2018, Yao, Müller and Wang, 2005). See also Morris (2015) for a comprehensive review. Our focus in this article is the functional linear quantile regression model where a scalar response is paired with a functional predictor on a two-dimensional domain.

Since proposed in the seminal work by Koenker and Bassett (1978), quantile regression has drawn great attention in statistical research and applications because of its appealing features. While ordinary least squares regression models the conditional mean of the response given the covariates, quantile regression offers a more comprehensive analysis of the conditional distribution by estimating conditional quantiles at various levels (Koenker, 2005). Moreover, quantile regression imposes no assumption on the error distribution, like Gaussian, and thus provides a more flexible and robust methodology. Compared to the large literature on functional linear mean regression, the interest in developing quantile regression methods for functional data is still growing. Recent works include functional linear quantile regression (Kato, 2012, Li et al., 2021, 2022), function-on-scalar quantile regression (Liu, Li and Morris, 2020), and partially functional quantile regression (Ma et al., 2019, Yao, Sue-Chee and Wang, 2017).

Most methods in functional data analysis focus on stochastic processes defined over a bounded interval. Recently, in many applications including neuroscience, climate science, and chemometrics, it becomes common to collect functional data over higher dimensional domains, such as images and time-space surfaces (Morris, 2015). There is emerging interest in developing an advanced methodology for modeling the multi-dimensional functional data generated from random fields over multi-dimensional domains, see, for example, principal component analysis (Shi et al., 2022, Zhou and Pan, 2014), covariance function estimation (Wang, Wong and Zhang, 2022), and functional regression (Arnone et al., 2019, Morris et al., 2011).

In this article, we study the functional linear quantile regression which models the conditional quantile of a scalar response given a functional predictor over a two-dimensional domain. The estimator of the slope function is obtained by minimizing the penalized check loss function (Koenker, 2005). The non-differentiability of the check loss function makes the theoretical analysis more challenging than in the functional linear regression. We establish the minimax optimal rate of convergence of the regularized estimator in two steps. First, a minimax lower bound is derived from a prediction perspective by evaluating the expected squared prediction error similar to the excess prediction risk in Cai and Yuan (2012). A key observation is that the check loss function is related to the likelihood of the asymmetric Laplace distribution (Koenker and Machado, 1999). It helps characterize the packing property among joint distributions indexed by different slope functions, which in turn leads to an upper bound for the metric entropy in terms of the Kullback-Leibler divergence. The Yang-Barron version of Fano's method then yields the minimax lower bound (Wainwright, 2019, Yang and Barron, 1999). Second, we show that despite the check loss function is not differentiable at the origin, its expected counterpart behaves like a quadratic functional in a neighborhood of the true conditional quantile. Using some empirical processes technique, we prove that the proposed estimator can indeed achieve the rate of convergence in the lower bound, and is thus rate optimal. Moreover, the representer theorem allows us to find the optimal solution to the penalized objective function within a finite-dimensional although the estimation of the slope function is intrinsically an infinite-dimensional problem. We implement the state-of-art interior point algorithm (Koenker, 2005, Koenker et al., 2017) to efficiently compute the regularized estimator, which is further illustrated by analyzing the hippocampus surface data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

One major contribution of this paper is to rigorously investigate the so-called simultaneous diagonalization of the reproducing kernel and the covariance kernel of the functional predictor defined over a two-dimensional or multi-dimensional domain. In one-dimensional case, Yuan and Cai (2010) requires the Sacks-Ylvisaker conditions (Ritter, Wasilkowski and Woźniakowski, 1995) to ensure that the simultaneous diagonalization holds. However, it is largely unknown if an extension of the Sacks-Ylvisaker conditions to a higher dimensional domain is possible. Using the concept of interpolation spaces (Adams and Fournier, 2003, Steinwart and Scovel, 2012), we show that under mild assumptions the reproducing kernel Hilbert space (RKHS) in which the slope function lies is isomorphic to a power space of the RKHS associated with the covariance kernel. Such an isomorphism leads to some useful results, such as common basis expansion and norm equivalence, serving as building blocks for our theoretical analysis. We also present a sufficient condition for the covariance kernel and show that some concrete examples satisfy the regularity assumptions. For example, our analysis holds when the covariance kernel is from the Matérn class (Rasmussen and Williams, 2006), and the slope function belongs to a Sobolev space. To our knowledge, this simultaneous diagonalization result is novel for multi-dimensional functional data analysis, providing a powerful machinery to study the minimax rates of convergence of the regularized estimator in functional data analysis.

There are some earlier works on functional linear quantile regression (Kato, 2012, Li et al., 2021, 2022) and partial functional linear quantile regression (Ma et al., 2019, Yao, Sue-Chee and Wang, 2017). We would like to highlight our contributions in the context of this literature. Kato (2012) studied the estimation of the slope function in functional linear quantile regression using the functional principal components (FPC). A series of papers (Cai and Yuan, 2012, Yuan and Cai, 2010) on functional linear models have pointed out that the success of the FPC approach hinges on an unrealistic assumption that the leading components explaining the most variability of the functional predictor are also effective in representing the slope function. To deal with this limitation of using FPC in the context of functional linear quantile regression, we assume that the slope function resides in an RKHS and unify both the estimation and prediction problems in the RKHS framework. We show that the statistical properties of our regularized estimator are jointly determined by the reproducing kernel and the covariance kernel of the functional predictor and establish its optimal rate of convergence under weaker conditions than Kato (2012). For example, the adequate spacing between adjacent eigenvalues of the covariance function (Assumption (A3), Kato, 2012) is no longer required since our method does not involve estimating the FPCs. Recently, Li et al. (2021) extended the RKHS-based approach to the functional linear quantile regression over a bounded interval. Their analysis requires that the eigenvalues of a composite kernel related to both the reproducing and covariance kernels decay polynomially at a rate. However, this assumption is difficult to verify even in the one-dimensional case, let alone on a multi-dimensional domain. In contrast, we develop the simultaneous diagonalization result revealing deeper connections between the reproducing and covariance kernel. Our analysis can be extended to a multi-dimensional domain with minor modifications, although our current presentation focuses on a two-dimensional domain. For partial functional linear quantile regression, the FPC-based approach is still prevalent in recent works (Ma et al., 2019, Yao, Sue-Chee and Wang, 2017).

The rest of this article is organized as follows. Section 2 introduces the functional linear quantile regression model and the regularized estimator. Section 3 presents the main theoretical results along with the proofs, including the simultaneous diagonalization of the reproducing and covariance kernel functions, the properties of the check loss function, and consequently the rates of convergence for our proposed estimator. We discuss implementation issues in Section 4, including the interior point algorithm and selection of the penalty parameter. In Section 5, we apply the proposed method to the hippocampus surface data in the ADNI study.

2. Functional linear quantile regression in RKHS

Let $Y \in \mathbb{R}$ be the scalar response. The functional predictor $X = (X(t), t \in \mathcal{T})$, is a square integrable random field over a compact domain $\mathcal{T} \subseteq \mathbb{R}^2$. Let $Q_{Y|X}(\cdot|X)$ denote the conditional quantile function of Y given X. Fix $\tau \in (0,1)$ as the quantile level of interest and denote $\eta_0(X) = Q_{Y|X}(\tau|X)$, the conditional quantile of Y given X at level τ. We assume that this conditional quantile is a linear functional of X, that is,

$$\eta_0(X) = \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\, dt, \tag{1}$$

where $\alpha_0 \in \mathbb{R}$ is the intercept, and $\beta_0(\cdot)$ is an unknown function which will be referred to as the slope function. We omit the dependence of η_0, α_0 and β_0 on the quantile level τ in our notation, for the sake of brevity. We also assume that $E(X(t)) = 0$ for simplicity. This is without loss of generality, since in practice, we can always center the data before applying the method presented in this paper.

Let $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ be the check loss function. It is known in the quantile regression literature that the conditional quantile $\eta_0(X)$ minimizes the expected loss

$$\ell_\infty(\eta) = E\{\rho_\tau(Y - \eta(X))\},$$

where minimization is over all η that has the form $\eta(X) = \alpha + \int_{\mathcal{T}} X(t)\beta(t)\,dt$ (Koenker and Bassett, 1978). Given an i.i.d. sample from the joint distribution of (X, Y), denoted as $\{(X_i, Y_i)\}_{i=1}^n$, define the empirical risk function as

$$\ell_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \eta(X_i)\right).$$

It is obvious that $\ell_\infty(\eta) = E\{\ell_n(\eta)\}$. Since minimization of the empirical loss function over all η with the form $\eta(X) = \alpha + \int_{\mathcal{T}} X(t)\beta(t)\,dt$ may lead to overfitting, we shall adopt the regularized estimation technique well-known in the smoothing spline literature (Wahba, 1990).

2.1. Regularized estimation

To define our regularized estimator, we first need to define a quadratic penalty functional of the slope function β. Let $\mathbb{N}_0$ be the set of non-negative integers. For a multi-index $m = (m_1, m_2)^\top \in \mathbb{N}_0^2$, define the length of m as $|m| = m_1 + m_2$ and the factorial as $m! = m_1!m_2!$. Denote by $D^m = D_1^{m_1} D_2^{m_2}$ the differential operator of order $|m|$, where D_1 and D_2 are the (weak) partial differential operators, respectively. For a fixed integer $r > 1$, the thin-plate penalty functional (Duchon, 1977, Wahba, 1990) of order r is defined as

$$J(\beta) = \sum_{|m|=r} \frac{r!}{m!} \|D^m\beta\|^2_{\mathcal{L}_2(\mathcal{T})}. \tag{2}$$

It is also known as the squared Beppo-Levi seminorm (Definition 10.37, Wendland, 2005).

Using the thin-plate penalty functional given in (2), the regularized estimators of the intercept α_0 and the slope function β_0 are defined as the solution to the following penalized empirical risk minimization problem,

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha\in\mathbb{R}, \beta\in\mathcal{L}_2(\mathcal{T})}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \alpha - \int_{\mathcal{T}} X_i(t)\beta(t)\,dt\right) + \lambda J(\beta), \tag{3}$$

where λ is a penalty parameter that controls the trade-off between the fidelity to the data and the roughness of function estimation. We omit the dependence of $\hat{\alpha}$ and $\hat{\beta}$ on n and λ for notation brevity.

Set $\mathcal{H} = \{\beta \in \mathcal{L}_2(\mathcal{T}) : J(\beta) < \infty\}$. The space $\mathcal{H}$ is identical as a set to the Sobolev space of order r on $\mathcal{T}$, which is defined as

$$\mathcal{W}_2^r(\mathcal{T}) = \{\beta \in \mathcal{L}_2(\mathcal{T}) : D^m(\beta) \in \mathcal{L}_2(\mathcal{T}) \text{ for } 0 \le |m| \le r\}, \tag{4}$$

and is equipped with the squared norm

$$\|\beta\|^2_{\mathcal{W}_2^r(\mathcal{T})} = \sum_{|m|=0}^{r} \|D^m\beta\|^2_{\mathcal{L}_2(\mathcal{T})}. \tag{5}$$

For later use, we remark that Sobolev spaces with a non-integer order can be defined by using Besov spaces and the real interpolation method (Definition 7.32 and Remark 7.33, Adams and Fournier, 2003). More precisely, for $q > 0$ and a sufficiently regular domain $\mathcal{T}$, the Sobolev space $\mathcal{W}_2^r(\mathcal{T})$ is defined as the interpolation space $\mathcal{W}_2^q(\mathcal{T}) = [\mathcal{L}_2(\mathcal{T}), \mathcal{W}_2^r(\mathcal{T})]_{q/r,2}$, where r is an integer larger than q. The Sobolev embedding theorem (Remark 7.33 and Theorem 7.34(c), Adams and Fournier, 2003) guarantees that when $q > 1$, the Sobolev space $\mathcal{W}_2^q(\mathcal{T})$ is compactly embedded in $C(\mathcal{T})$, the space of continuous functions on $\mathcal{T}$. In particular, $\mathcal{W}_2^q(\mathcal{T}) \cap C(\mathcal{T})$ is the set of continuous representatives of functions in $\mathcal{W}_2^q(\mathcal{T})$.

2.2. The representer theorem

Let $\mathcal{H}_0 = \{\beta : J(\beta) = 0\}$ be the null space of the quadratic functional J and assume that $\mathcal{H}_0$ is a finite-dimensional linear subspace of $\mathcal{H}$ with basis $\{\xi_k : 1 \le k \le m\}$. Consider the direct sum decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. The space $\mathcal{H}_1$ is a reproducing kernel Hilbert space with $J(\beta)$ a well-defined squared norm restricted in $\mathcal{H}_1$ (Theorem 10.42, Wendland, 2005). Denote by $K_1(\cdot,\cdot)$ the reproducing kernel of $\mathcal{H}_1$.

The regularized estimator $\hat{\beta}(t)$ in (3) is formulated as the minimizer of a penalized criterion in an infinite-dimensional space. The following representer theorem (see, e.g., Kimeldorf and Wahba (1971), Li, Liu and Zhu (2007), Schölkopf, Herbrich and Smola (2001)) suggests that it suffices to search for the minimizer within a finite-dimensional space.

Theorem 2.1. *There exist vectors* $\mathbf{e} = (e_1, \dots, e_m)^\top$ *and* $\mathbf{c} = (c_1, \dots, c_n)^\top$ *such that:*

$$\hat{\beta}(t) = \sum_{k=1}^{m} e_k \xi_k(t) + \sum_{i=1}^{n} c_i \int_{\mathcal{T}} K_1(s,t) X_i(s)\, ds. \tag{6}$$

According to Theorem 2.1, we need only to search over β's with the expression given on the right hand side of (6) when solving (3). For such β's,

$$\int_{\mathcal{T}} X(t)\beta(t)dt = \sum_{k=1}^{m} e_k \int_{\mathcal{T}} X(t)\xi_k(t)dt + \sum_{i=1}^{n} c_i \int_{\mathcal{T}}\int_{\mathcal{T}} X(s)K_1(s,t)X_i(t)\, ds dt.$$

Let $\mathbf{T}$ be the $n \times m$ matrix with

$$T_{ik} = \int_{\mathcal{T}} X_i(t)\xi_k(t)\, dt,$$

where $i = 1, \dots, n$ and $k = 1, \dots, m$. Denote by $\boldsymbol{\Sigma}$ the $n \times n$ matrix with its (i, j)th entry being

$$\Sigma_{ij} = \int_{\mathcal{T}}\int_{\mathcal{T}} X_i(s)K_1(s,t)X_j(t)\, ds dt.$$

The reproducing property of $K_1(\cdot,\cdot)$ leads to

$$J(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \int_{\mathcal{T}}\int_{\mathcal{T}} X_i(s)K_1(s,t)X_j(t)\, ds dt = \mathbf{c}^\top \boldsymbol{\Sigma}\mathbf{c}.$$

Therefore, problem (3) reduces to minimizing the following objective function with respect to $\alpha \in \mathbb{R}, \mathbf{e} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^n$,

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau (Y_i - \alpha - \mathbf{T}_i\mathbf{e} - \boldsymbol{\Sigma}_i\mathbf{c}) + \lambda \mathbf{c}^\top \boldsymbol{\Sigma}\mathbf{c}, \tag{7}$$

where $\mathbf{T}_i$ and $\boldsymbol{\Sigma}_i$ are the ith rows of matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$, respectively.

Note that the objective function (7) is convex in $(\alpha, \mathbf{e}, \mathbf{c})$, so its every local minimum is a global one. However, since (7) is not necessarily strictly convex, the minimum may not be unique, and according to the subgradient optimality condition (Rockafellar, 1997), the minimum is the set of all points where zero is a subgradient of the objective function. When there is no penalization, i.e., $\lambda = 0$, this phenomenon is well-known in the quantile regression literature (e.g., Koenker, Ng and Portnoy, 1994). Since $\mathbf{c}^\top \mathbf{\Sigma} \mathbf{c}$ is strictly convex on the set of $\mathbf{c}$ such that $\mathbf{\Sigma c} \neq 0$, the inclusion of the penalty term substantially reduces the possibility of non-uniqueness of the minimizer. That penalization can help alleviate the non-uniqueness problem in quantile regression which has been confirmed in our numerical studies and also observed in the literature, e.g., Li, Liu and Zhu (2007).

3. Theoretical properties

We now present our main results on the asymptotic properties of the regularized estimator. Subsections 3.1 and 3.3 respectively present and discuss assumptions on the covariance function of the random field predictor X and the smoothness assumptions on the unknown slope function. Subsection 3.2 gives some properties of the check loss function. Subsection 3.4 presents the results on rates of convergence.

We refer the reader to Chapter 4 of Steinwart and Christmann (2008) for the necessary background of reproducing kernel Hilbert spaces. Suppose $\mathcal{T}$, the domain on which the functional predictor X is observed, is a non-empty, bounded open set with Lipschitz boundary.

3.1. Covariance function

Denote the covariance function of the random field $X(t)$ by

$$C(t,t') = E([X(t) - E\{X(t)\}][X(t') - E\{X(t')\}]).$$

We assume that C is continuous on $\mathcal{T} \times \mathcal{T}$ and satisfies $\int_{\mathcal{T}} C(t,t)\, dt < \infty$. Since the covariance function C is positive definite, there is a unique reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}_C$, with C as its reproducing kernel. The inclusion map $I_C : \mathcal{H}_C \to \mathcal{L}_2(\mathcal{T})$ is continuous and, its adjoint operator is the integral operator $L_C : \mathcal{L}_2(\mathcal{T}) \to \mathcal{H}_C$ defined by

$$L_C(f)(\cdot) = \int_{\mathcal{T}} C(t,\cdot) f(t)\, dt, \quad f \in \mathcal{L}_2(\mathcal{T}),$$

which is Hilbert-Schmidt. Since C is the kernel of the integral operator L_C, C is also referred to as a covariance kernel. For later use, define the inner product associated with C as

$$\langle f, g\rangle_C = \langle L_C(f), g\rangle_{\mathcal{L}_2} = \int_{\mathcal{T}} f(t) C(t,t') g(t')\, dt dt', \quad f,g \in \mathcal{L}_2(\mathcal{T}) \tag{8}$$

and the induced norm is $\|f\|_C = \langle f, f\rangle_C^{1/2}$.

The operator $L_C^* L_C$ is compact, positive, and self-adjoint, and hence the spectral theorem for self-adjoint compact operators shows that, there exists a sequence of orthonormal eigenfunctions $\{\phi_k\}_{k=1}^\infty$ and a sequence of non-increasing non-negative eigenvalues $\{\mu_k\}_{k=1}^\infty$ such that $L_C^* L_C(\phi_k) = \mu_k \phi_k$. Moreover, Mercer's theorem yields

$$C(t,t') = \sum_{k=1}^{\infty} \mu_k \phi_k(t) \phi_k(t'), \quad t,t' \in \mathcal{T}, \tag{9}$$

where the convergence is absolute and uniform. Therefore, $\{\mu_k, \phi_k\}$ is also an eigen-system of C. The random field X admits the Karhunen–Loève expansion $X(t) = \sum_{k=1}^{\infty} Z_k \phi_k(t)$, where Z_k's are uncorrelated random variables satisfying $E(Z_k) = 0$ and $E(Z_k^2) = \mu_k$ for $k \geq 1$.

The sample path properties of the random field $X(t)$ is determined by its covariance function C. We next impose a regularity assumption on C through its associated RKHS. We use $\mathcal{W}_2^s(\mathcal{T}) \cap C(\mathcal{T})$ to denote the set of the continuous representatives of functions in the Sobolev space $\mathcal{W}_2^s(\mathcal{T})$, equipped with the Sobolev norm. Two metric spaces are said to be isomorphic if the two sets coincide and the norms of the two spaces are equivalent.

Assumption 1. For some constant $s > 1$, $\mathcal{H}_C$ is isomorphic to $\mathcal{W}_2^s(\mathcal{T}) \cap C(\mathcal{T})$.

We next present a sufficient condition for Assumption 1 and show that some commonly used covariance functions satisfy the condition. When the covariance function is stationary, we can write $C(t, t') = \Phi(t - t')$. The Fourier transform of Φ, defined as $\hat{\Phi}(\omega) = (2\pi)^{-1} \int_{\mathbb{R}^2} \Phi(r) e^{-ir^\top \omega} \, dr$ for $\omega \in \mathbb{R}^2$, is known as the spectral density corresponding to the covariance function. It follows from Corollaries 10.13 and 10.48 of Wendland (2005) that the following is a sufficient condition for Assumption 1.

Assumption 1′. There exist two positive constants $c_1 \leq c_2$ such that

$$c_1(1 + \|\omega\|_2^2)^{-s} \leq \hat{\Phi}(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-s}, \quad \omega \in \mathbb{R}^2.$$

Assumption 1′ essentially requires the tail of the spectral density of C decays as $(1 + \|\omega\|_2^2)^{-s}$ when $\|\omega\| \to \infty$. This requirement is satisfied by the Matérn class of covariance functions (Stein, 1999) which is given by

$$C(t, t') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|t - t'\|}{l} \right)^{\nu} B_\nu \left(\frac{\sqrt{2\nu}\|t - t'\|}{l} \right), \tag{10}$$

with positive smoothness and scale parameters ν and l, where B_ν is a modified Bessel function of the second kind. The Matérn class is widely used in spatial statistics, machine learning, and image analysis. It is a broad class that includes the exponential kernel as a special case and the Gaussian kernel as a limiting case. Stein (1999) recommended it as a canonical class of covariance kernels for modeling spatial random field that works reasonably well in a wide range of circumstances. The corresponding spectral density function is $C_{\nu,l}(2\nu l^{-2} + 4\pi^2\omega^2)^{-\nu-1}$, where $C_{\nu,l}$ is a constant depending only on ν and l (Section 4.2, Rasmussen and Williams, 2006). It follows that Assumption 1′ and therefore Assumption 1 holds with $s = \nu + 1$.

Lemma 3.1. *Under Assumption 1, the eigenvalues of C satisfy $\mu_k \asymp k^{-s}$ for $k \geq 1$.*

Proof. Recall the definitions of the inclusion operator $I_C : \mathcal{H}_C \to \mathcal{L}_2(\mathcal{T})$ and its adjoint operator $L_C = I_C^*$ given in Section 3.1. It is a standard result from functional analysis that the self-adjoint operators $L_C^* L_C$ and $L_C L_C^* = I_C^* I_C$ have the same set of non-zero eigenvalues $\{\mu_k : k \geq 1\}$. Equation 4.4.12 of Carl and Stephani (1990) shows that the kth eigenvalue of $L_C L_C^* = I_C^* I_C$ is equal to the squared approximation number $a_k^2(I_C)$ for $k \geq 1$. Therefore, $\mu_k = a_k^2(I_C)$ for $k \geq 1$.

On the other hand, let $a_k(\mathrm{id})$ denote the kth approximation number of the natural embedding $\mathrm{id} : \mathcal{W}_s^2(\mathcal{T}) \to \mathcal{L}_2(\mathcal{T})$. It follows from Theorem 3.3.4 of Edmunds and Triebel (1996) that $a_k(\mathrm{id}) \asymp k^{-s/2}$ (assuming that $\mathcal{T}$ has a C_∞ boundary).

By the isomorphism under Assumption 1, $a_k(I_C) \asymp a_k(\mathrm{id}) \asymp k^{-s/2}$. We conclude that $\mu_k \asymp k^{-s}$ for $k \geq 1$. □

3.2. The check loss function and the asymmetric Laplace distribution

The check loss function $\rho_\tau(u)$ is not everywhere differentiable in u, making it more challenging to derive the theoretical properties of the regularized estimator defined in (3). Nonetheless, the following Knight's identity (Knight, 1998) presents a useful property of the check loss function. Let $\psi_\tau(u) = \tau - I(u < 0)$.

Lemma 3.2 (Knight's identity). *It holds for all $u, v \in \mathbb{R}$ that*

$$\rho_\tau(u-v) - \rho_\tau(u) = \int_0^v \{I(u \le w) - I(u \le 0)\}\, dw - v\psi_\tau(u).$$

We use Knight's identity to show that the expected loss $\ell_\infty(\eta)$ behaves like a quadratic functional locally around η_0. Denote $W = Y - \eta_0(X)$ and $\Delta = \eta(X) - \eta_0(X)$. Applying Knight's identity with $u = W$ and $v = \Delta$ and taking expectation, we obtain that

$$\ell_\infty(\eta) - \ell_\infty(\eta_0) = E_{\eta_0}\left[\int_0^\Delta \{I(W \le w) - I(W \le 0)\}\, dw - \Delta\psi_\tau(W)\right]. \tag{11}$$

Here, we use the subscript in E_{η_0} to indicate the dependence of the joint distribution of (X,Y) on η_0. Let $F_{W|X}(w|x)$ denote the conditional distribution of W given X. Then by the definition of conditional quantile, $E_{\eta_0}\{\psi_\tau(W)|X\} = 0$ and $F_{W|X}(0|X) = \tau$. Using these results when conditioning on X inside the expectation on the right hand side of (11), we obtain

$$\ell_\infty(\eta) - \ell_\infty(\eta_0) = E_{\eta_0}\left[\int_0^\Delta \{F_{W|X}(w|x) - \tau\}\, dw\right]. \tag{12}$$

Assumption 2. There exist constants C_1 and C_2 such that

$$|F_{W|X}(w|x) - F_{W|X}(0|x)| \ge C_1|w|, \qquad \text{for } |w| \le C_2, \text{ all } x.$$

Assumption 2 guarantees that the conditional distribution cannot be too "flat" at the quantile of interest. A sufficient condition for Assumption 2 is to require that the conditional density is uniformly bounded away from zero in a neighborhood of the conditional quantile of interest, i.e., there exist constants $C_1 > 0$ and $C_2 > 0$ such that uniformly in x, the conditional density $f_{W|X}(w|x) \ge C_1$ holds whenever $|w| \le C_2$. This sufficient (thus stronger) condition is a standard assumption in the literature of quantile regression (e.g., Kato, 2012, Li et al., 2021, Lv et al., 2018, Sherwood and Wang, 2016, Volgushev, Chao and Cheng, 2019).

Assumption 3. For any square integrable function f defined on $\mathcal{T}$,

$$E\left[\int_{\mathcal{T}} X(t)f(t)\, dt\right]^4 \le C_3\left(E\left[\int_{\mathcal{T}} X(t)f(t)\, dt\right]^2\right)^2.$$

Assumption 3 essentially requires that the kurtosis of the random variable $\int_{\mathcal{T}} X(t)f(t)\, dt$ is uniformly bounded for all square integrable f. It is valid with $C_3 = 3$ when X is a Gaussian process. This assumption has been used in the literature on functional linear models (e.g., Cai and Yuan, 2012, Yuan and Cai, 2010). A stronger assumption that X is sub-Gaussian was used in Li et al. (2021), where it is required that for all $f \in \mathcal{L}_2(\mathcal{T})$ and $\zeta \in \mathbb{R}$, $E\exp\{\zeta \int_{\mathcal{T}} X(t)f(t)\, dt\} \le \exp[\zeta^2 E\{\int_{\mathcal{T}} X(t)f(t)\, dt\}^2/2]$.

Lemma 3.3. *Under Assumptions 2 and 3,*

$$\ell_\infty(\eta) - \ell_\infty(\eta_0) \geq \frac{C_1}{4} E\{[\eta(X) - \eta_0(X)]^2\} \tag{13}$$

holds for $\eta(X) = \alpha + \int_{\mathcal{T}} X(t)\beta(t)\,dt$ *and* $\eta_0(X) = \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\,dt$ *satisfying* $E\{[\eta(X) - \eta_0(X)]^2\} \leq C_2^2/(16C_3)$.

This result says that the expected loss $\ell_\infty(\eta)$ is lower bounded by a quadratic functional in a neighborhood of the true conditional quantile η_0. The proof is given at the end of this section.

The check loss function can be viewed as the negative log-likelihood for an asymmetric Laplace distribution (Koenker and Machado, 1999). Let P_{η_0} denote the joint distribution of (X,Y), where the conditional density of $W = Y - \eta_0(X)$ given the functional predictor X is independent of X and given by the following asymmetric Laplace distribution

$$f_{W|X}(w) = \tau(1-\tau)\exp\{-\rho_\tau(w)\}, \quad w \in \mathbb{R}. \tag{14}$$

Let P_η be defined in the same way but with η_0 replaced by η. The Kullback-Leibler divergence between P_{η_0} and P_η is

$$\mathcal{K}(P_{\eta_0}, P_\eta) = E_{\eta_0}\{\log(P_{\eta_0}/P_\eta)\} = \ell_\infty(\eta) - \ell_\infty(\eta_0), \tag{15}$$

where the expectation is taken with respect to P_{η_0}.

Lemma 3.4. $\mathcal{K}(P_{\eta_0}, P_\eta) \leq \tau(1-\tau)E\{[\eta(X) - \eta_0(X)]^2\}$.

The proof is given at the end of this section. This result bounds the Kullback-Leibler divergence of two distributions by the $\mathcal{L}_2$ norm of the corresponding conditional quantiles. It is used as a tool to derive the asymptotic minimax lower bound later in Theorem 3.6. We remark that the asymmetric Laplace distribution assumption used in Lemma 3.4 is only imposed to establish the minimax lower bound, because any lower bound for a specific case immediately leads to a lower bound for the general case (Wainwright, 2019). This restricted distribution assumption is not needed when applying our proposed regularized estimators and studying their asymptotic rates of convergence (upper bound).

Proof of Lemma 3.3. Assumption 2 implies that

$$\int_0^\Delta \{F_{W|X}(w|x) - F_{W|X}(0|x)\}\,dw \geq \frac{C_1}{2}|\Delta|^2 I(|\Delta| < C_2). \tag{16}$$

Recall that

$$\Delta = \eta(X) - \eta_0(X) = (\alpha - \alpha_0) + \int_{\mathcal{T}} X(t)(\beta(t) - \beta_0(t))\,dt.$$

Using the inequality $(a+b)^4 \leq 8(a^4 + b^4)$ and Assumption 3,

$$\begin{aligned} E(\Delta^4) &\leq 8\left[(\alpha - \alpha_0)^4 + E\left\{\int_{\mathcal{T}} X(t)(\beta(t) - \beta_0(t))\,dt\right\}^4\right] \\ &\leq 8C_3\left[(\alpha - \alpha_0)^4 + \left(E\left\{\int_{\mathcal{T}} X(t)(\beta(t) - \beta_0(t))\,dt\right\}^2\right)^2\right]. \end{aligned}$$

Noticing $E(X(t)) = 0$ and $a^2 + b^2 \le (a+b)^2$ for $a, b > 0$, we have that

$$E(\Delta^4) \le 8C_3\left[\left((\alpha-\alpha_0)^2 + E\left\{\int_{\mathcal{T}} X(t)(\beta(t)-\beta_0(t))\,dt\right\}^2\right)^2\right]$$
$$\le 8C_3\left[\left(E\left\{\alpha-\alpha_0 + \int_{\mathcal{T}} X(t)(\beta(t)-\beta_0(t))\,dt\right\}^2\right)^2\right] = 8C_3\{E(\Delta^2)\}^2.$$

Therefore,

$$E\{|\Delta|^2 I(|\Delta| \ge C_2)\} \le C_2^{-2}E(\Delta^4) \le 8C_3C_2^{-2}\{E(\Delta^2)\}^2,$$

and thus

$$E\{|\Delta|^2 I(|\Delta| < C_2)\} \ge E(\Delta^2) - 8C_3C_2^{-2}\{E(\Delta^2)\}^2.$$

Therefore, if $E(\Delta^2) \le C_2^2/(16C_3)$, then $E\{|\Delta|^2 I(|\Delta| < C_2)\} \ge E(\Delta^2)/2$. Combine this with (12) and (16) to obtain the desired result. □

Proof of Lemma 3.4. It follows from (12) and (15) that,

$$\mathcal{K}(P_{\eta_0}, P_\eta) = E_{\eta_0}\left[\int_0^\Delta \{F_{W|X}(w|x) - \tau\}\,dw\right]. \tag{17}$$

When the conditional density is the asymmetric Laplace given in (14), the cumulative distribution function of W conditional on X is

$$F_{W|X}(w|x) = \begin{cases} \tau + (\tau-1)\{\exp(-\tau w) - 1\}, & w > 0, \\ \tau\exp\{(1-\tau)w\}, & w \le 0. \end{cases}$$

By calculation,

$$\int_0^\Delta \{F_{W|X}(w|x) - \tau\}\,dw = \begin{cases} \tau^{-1}(1-\tau)\{\exp(-\tau\Delta) + \tau\Delta - 1\}, & \Delta > 0, \\ \tau(1-\tau)^{-1}[\exp\{(1-\tau)\Delta\} + (\tau-1)\Delta - 1], & \Delta \le 0. \end{cases}$$

Using the fact that $\exp(-z) + z - 1 \le z^2$ for any $z \ge 0$, we obtain that

$$\int_0^\Delta \{F_{W|X}(w|x) - \tau\}\,dw \le \tau(1-\tau)\Delta^2,$$

which together with (17) leads to the desired result. □

3.3. Smoothness assumption on the slope function

In the literature of nonparametric function estimation, it is customary to assume that the unknown function belongs to a Sobolev space of functions. While function in a Sobolev space is only defined almost everywhere, we restrict attention to the continuous version of a function. We make the following assumption on the unknown slope function β_0.

Assumption 4. For some $r > 1$, $\beta_0 \in \mathcal{W}_2^r(\mathcal{T}) \cap C(\mathcal{T})$.

Note that $\mathcal{H} = \mathcal{W}_2^r(\mathcal{T}) \cap C(\mathcal{T})$, equipped with the norm $\|f\|_{\mathcal{H}} = \|f\|_{\mathcal{W}_2^r}$, is an RKHS. To see this, we need only to verify continuity of the evaluation functional $[t]f = f(t)$ for $t \in \mathcal{T}$ (Section 4.2, Steinwart and Christmann, 2008). The Sobolev embedding theorem implies that $|[t]f| \le \sup_{t\in\mathcal{T}} |f(t)| = \|f\|_{C_0} \le C\|f\|_{\mathcal{W}_2^r} = C\|f\|_{\mathcal{H}}$, and thus the continuity of evaluation functionals follows. We call $\mathcal{H}$ the Sobolev RKHS of order r.

Next, using the concept of interpolation spaces introduced by Steinwart and Scovel (2012), the next theorem shows that the Sobolev RKHS $\mathcal{H}$ is isomorphic to a power space of the RKHS $\mathcal{H}_C$. As a result, any $\beta \in \mathcal{H}$ admits a basis expansion using the eigen-functions of the covariance kernel C appearing in (9). Moreover, both of the norms $\|\beta\|_{\mathcal{H}}$ and $\|\beta\|_C$ can be characterized as weighted sums of coefficients in the basis expansion.

Theorem 3.5. *Under Assumptions 1 and 4, $\mathcal{H}$ is isomorphic to $\mathcal{H}_C^{r/s}$, the (r/s)-power space of $\mathcal{H}_C$, which is defined as*

$$\mathcal{H}_C^{r/s} = \left\{ f = \sum_{k=1}^{\infty} a_k \mu_k^{r/(2s)} \phi_k : \|f\|^2_{\mathcal{H}_C^{r/s}} = \sum_{k=1}^{\infty} a_k^2 < \infty \right\},$$

equipped with the inner product

$$\left\langle \sum_{k=1}^{\infty} a_k \mu_k^{r/(2s)} \phi_k, \sum_{k=1}^{\infty} b_k \mu_k^{r/(2s)} \phi_k \right\rangle_{\mathcal{H}_C^{r/s}} = \sum_{k=1}^{\infty} a_k b_k.$$

Any $\beta \in \mathcal{H}$ can be expanded by the eigenfunctions of C such that

$$\beta = \sum_{k=1}^{\infty} b_k \mu_k^{r/(2s)} \phi_k, \qquad b_k = \mu_k^{-r/(2s)} \langle \beta, \phi_k \rangle_{\mathcal{L}_2},$$

where the convergence is in the absolute sense. Furthermore,

$$\|\beta\|_C^2 = \sum_{k=1}^{\infty} \mu_k^{(r+s)/s} b_k^2,$$

and there exist universal constants $c_1, c_2 > 0$ such that

$$c_1 \sum_{k=1}^{\infty} b_k^2 \le \|\beta\|_{\mathcal{H}}^2 \le c_2 \sum_{k=1}^{\infty} b_k^2. \tag{18}$$

Proof. First, we show that $\mathcal{H} \cong \mathcal{H}_C^{r/s}$ where the symbol $\cong$ stands for isomorphism between metric spaces. Choose $q > 1$ be an arbitrary real number satisfying $q > \max\{s, r\}$ and let $\mathcal{H}_Q = \mathcal{W}_2^q(\mathcal{T}) \cap C(\mathcal{T})$ be the Sobolev RKHS space with the measurable and bounded reproducing kernel Q. Suppose Q admits the spectral representation

$$Q(s,t) = \sum_{k=1}^{\infty} \varrho_k \zeta_k(s) \zeta_k(t),$$

where $\varrho_1 \geq \varrho_2 \geq \cdots \geq 0$ are the eigenvalues, and $\{\zeta_k\}_{k=1}^{\infty}$ are the eigenfunctions. According to Steinwart and Christmann (2008, Theorem 4.51), the Mercer representation of the RKHS $\mathcal{H}_Q$ is

$$\mathcal{H}_Q = \left\{ f = \sum_{k=1}^{\infty} a_k \varrho_k^{1/2} \zeta_k : \|f\|_{\mathcal{H}_Q}^2 = \sum_{k=1}^{\infty} a_k^2 < \infty \right\}.$$

For $\alpha \geq 0$, the α-power RKHS is defined by

$$\mathcal{H}_Q^{\alpha} = \left\{ f = \sum_{k=1}^{\infty} a_k \varrho_k^{\alpha/2} \zeta_k : \|f\|_{\mathcal{H}_Q^{\alpha}}^2 = \sum_{k=1}^{\infty} a_k^2 < \infty \right\}.$$

For $0 < \alpha < 1$, the α-power space $\mathcal{H}_Q^{\alpha}$ is characterized in terms of the interpolation spaces of the real method, specifically, $\mathcal{H}_Q^{\alpha} \cong [\mathcal{L}_2(\mathcal{T}), \mathcal{H}_Q]_{\alpha,2}$ (Theorem 4.6, Steinwart and Scovel, 2012).

By considering Sobolev spaces as special cases of Besov spaces, we can apply the interpolation property (Theorem 7.31, Adams and Fournier, 2003) to obtain that $\mathcal{W}_2^s(\mathcal{T}) \cong [\mathcal{L}_2(\mathcal{T}), \mathcal{W}_2^q(\mathcal{T})]_{s/q,2}$ holds for real numbers $q > s > 0$. According to Assumption 1, we may identify $W_2^q(\mathcal{T})$ as $\mathcal{H}_Q$ by using the continuous function as the representer in each equivalence class in $W_2^q(\mathcal{T})$. Therefore, we obtain that

$$\mathcal{H}_C \cong \mathcal{W}_2^s(\mathcal{T}) \cong [\mathcal{L}_2(\mathcal{T}), \mathcal{W}_2^q(\mathcal{T})]_{s/q,2} \cong [\mathcal{L}_2(\mathcal{T}), \mathcal{H}_Q]_{s/q,2} \cong \mathcal{H}_Q^{s/q}.$$

A similar argument yields $\mathcal{H} \cong \mathcal{W}_2^r(\mathcal{T}) \cap C(\mathcal{T}) \cong \mathcal{H}_Q^{r/q}$. On the other hand, it follows from the definition of power RKHSs that

$$\mathcal{H}_Q^{r/q} = \left\{ f = \sum_{k=1}^{\infty} a_k \varrho_k^{r/(2q)} \zeta_k : \|f\|_{\mathcal{H}_Q^{s/q}}^2 = \sum_{k=1}^{\infty} a_k^2 < \infty \right\} = \left[\mathcal{H}_Q^{s/q}\right]^{r/s}.$$

Consequently, we have shown that $\mathcal{H} \cong \mathcal{H}_C^{r/s}$, which is an important relation of representing $\mathcal{H}$ as a power of the RHKS $\mathcal{H}_C$.

Using the eigenfunctions of C, the isomorphism between $\mathcal{H}$ and $\mathcal{H}_C^{r/s}$ leads to the basis expansion of $\beta \in \mathcal{H}$. Moreover, the Mercer representation of $\mathcal{H}_C^{r/s}$ implies that $\|\beta\|_{\mathcal{H}_C^{r/s}}^2 = \sum_{k=1}^{\infty} b_k^2$. The lower and upper bounds of $\|\beta\|_{\mathcal{H}}^2$ follows from the norm equivalence of $\mathcal{H}$ and $\mathcal{H}_C^{r/s}$. On the other hand, the eigenfunctions of C form an orthonormal basis in $\mathcal{L}_2(\mathcal{T})$, and thus

$$\|\beta\|_C^2 = \langle L_C\beta, \beta\rangle_{\mathcal{L}_2} = \left\langle \sum_{k=1}^{\infty} b_k \mu_k^{r/(2s)} L_C(\phi_k), \sum_{k=1}^{\infty} b_k \mu_k^{r/(2s)} \phi_k \right\rangle_{\mathcal{L}_2} = \sum_{k=1}^{\infty} \mu_k^{(r+s)/s} b_k^2.$$

The proof is complete. □

3.4. Rates of convergence

We take a prediction perspective to evaluate the performance of the regularized estimator. Let $X^* = (X^*(t), t \in \mathcal{T})$ be an independent copy of X, which can be interpreted as a future observation of the random field predictor X. Then $\hat{\eta}(X^*) = \hat{\alpha} + \int_{\mathcal{T}} X^*(t)\hat{\beta}(t)\,dt$ is a prediction of the true conditional quantile $\eta_0(X^*) = \alpha_0 + \int_{\mathcal{T}} X^*(t)\beta(t)\,dt$ when the functional predictor is X^*. We define the expected squared prediction error (PE) as

$$\mathrm{PE}(\hat{\eta}) = E_{X^*}\left\{\hat{\eta}(X^*) - \eta_0(X^*)\right\}^2, \tag{19}$$

where the expectation is taken over X^*. It is also referred to as excess risk in the regression setting (Cai and Yuan, 2012). Using the assumption that $E(X(t)) = 0$, we obtain that

$$\mathrm{PE}(\hat{\eta}) = (\hat{\alpha} - \alpha_0)^2 + E_{X^*}\left\{\int_{\mathcal{T}} X^*(t)\hat{\beta}(t)\,dt - \int_{\mathcal{T}} X^*(t)\beta_0(t)\,dt\right\}^2. \tag{20}$$

The PE defined in (20) can be expressed using the norm induced by the inner product defined in (8), which uses the covariance function of X. It is easy to see that

$$\mathrm{PE}(\hat{\eta}) = (\hat{\alpha} - \alpha_0)^2 + \|\hat{\beta} - \beta_0\|_C^2.$$

To simplify the presentation of our theoretical results and the proofs, throughout the rest of this section, we assume that $\alpha_0 = 0$ and remove the α term in the definition of the estimator in (3). Under this simplification, $\mathrm{PE}(\hat{\eta})$ reduces to $\|\hat{\beta} - \beta_0\|_C^2$. This simplification is without loss of generality. When $\alpha_0 \neq 0$, the results hold true by replacing $\|\beta - \beta_0\|_C^2$ with $(\alpha - \alpha_0)^2 + \|\beta - \beta_0\|_C^2$. The same argument goes through with some complications of notation.

Theorem 3.6. *Fix $B > 0$. If Assumptions 1, 2 and 4 hold, we have that*

$$\lim_{a\to 0}\liminf_{n\to\infty}\inf_{\tilde{\beta}}\sup_{\beta_0\in\mathcal{H}:\|\beta_0\|_{\mathcal{H}}\le B} P(\|\tilde{\beta} - \beta_0\|_C^2 \ge a n^{-(r+s)/(r+s+1)}) = 1,$$

where the infimum is taken over all possible estimators $\tilde{\beta}$ based on the observed data $\{(X_i, Y_i) : 1 \le i \le n\}$.

Theorem 3.6 gives the minimax lower bound of convergence. It says that, with high probability, for any estimator $\tilde{\beta}$, one can always find a $\beta_0 \in \mathcal{B}$ with $\|\beta_0\|_{\mathcal{H}} \le B$ such that $\|\tilde{\beta} - \beta_0\|_C^2$ converges to zero at a rate that cannot be faster than $n^{-(r+s)/(r+s+1)}$.

Theorem 3.7. *Fix $B > 0$. Suppose Assumptions 1, 2, 3 and 4 hold true, and $\lambda = O(n^{-(r+s)/(r+s+1)})$. We have that*

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\beta_0\in\mathcal{H}:\|\beta_0\|_{\mathcal{H}}\le B} P(\|\hat{\beta} - \beta_0\|_C^2 \ge A n^{-(r+s)/(r+s+1)}) = 0.$$

Theorem 3.7 gives the minimax upper bound of convergence rate of the regularized estimator $\hat{\beta}$ of the slope function, as defined in (3). It says that $\|\hat{\beta} - \beta_0\|_C^2$ goes to zero in probability at a rate not slower than $n^{-(r+s)/(r+s+1)}$. Since the minimax upper bound matches the lower bound, we say that our regularized estimator achieves the optimal rate of convergence.

3.5. Proofs of results on rates of convergence

Proof of Theorem 3.6. We apply the Yang-Barron version of Fano's method to establish the minimax lower bound (Yang and Barron, 1999). Here, we follow the description given in Section 15.3.5 of Wainwright (2019).

Without loss of generality, consider $B = 1$, so we can focus on the unit ball in $\mathcal{H}$, denoted as $\mathcal{B}(\mathcal{H}) = \{\beta \in \mathcal{H} : \|\beta\|_{\mathcal{H}} \le 1\}$. Note that any lower bound for a specific case immediately leads to a lower bound for the general case. It therefore suffices to consider the case that $W = Y - \int_{\mathcal{T}} X(t)\beta_0(t)\,dt$ follows the asymmetric Laplace distribution with the density given in (14). Let P_β be the joint distribution of $\{(X_i, Y_i) : i = 1, \dots, n\}$ when the true slope function $\beta_0 = \beta$. Denote by $\mathcal{P}$ the collection of probability distributions P_β when the slope function $\beta \in \mathcal{B}(\mathcal{H})$. Let $N_{\mathrm{KL}}(\epsilon, \mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ under the square-root Kullback-Leibler divergence. Let $M(\delta, \mathcal{B}(\mathcal{H}), \|\cdot\|_C)$ be the δ-packing number of $\mathcal{B}(\mathcal{H})$ under the $\|\cdot\|_C$-norm. Following the two-step procedure given on page 513 of Wainwright (2019), we obtain the minimax lower bound by finding a pair (ϵ_n, δ_n) as follows. In the first step, we choose $\epsilon_n > 0$ such that

$$\epsilon_n^2 \ge \log N_{\mathrm{KL}}(\epsilon_n, \mathcal{P}) \tag{21}$$

is satisfied. In the second step, given this choice of ϵ_n, the minimax lower bound (in the $\|\cdot\|_C$-norm) is given by the largest $\delta_n > 0$ such that

$$\log M(\delta_n, \mathcal{B}(\mathcal{H}), \|\cdot\|_C) \ge 4\epsilon_n^2 + 2\log 2. \tag{22}$$

We start with calculating the covering number of $\mathcal{B}(\mathcal{H})$ using the $\|\cdot\|_C$-norm. Based on the norm equivalence of RKHSs $\mathcal{H}$ and $\mathcal{H}_C^{r/s}$ shown in Theorem 3.5, we can express $\mathcal{B}(\mathcal{H})$ with the eigen-system of C as

$$\mathcal{B}(\mathcal{H}) = \left\{ \sum_{k=1}^{\infty} b_k \mu_k^{r/(2s)} \phi_k : \sum_{k=1}^{\infty} b_k^2 \le 1 \right\},$$

and $\|\beta\|_C^2 = \sum_{k=1}^{\infty} \mu_k^{(r+s)/s} b_k^2$, where we omit the universal constants appearing in the norm-equivalence (18) between the RKHSs $\mathcal{H}$ and $\mathcal{H}_C^{r/s}$. By reparameterizing with $a_k = b_k \mu_k^{(r+s)/(2s)}$, we can show that $\mathcal{B}(\mathcal{H})$ equipped with $\|\cdot\|_C$ is isometrically isomorphic to the ellipsoid in $\ell_2(\mathbb{N})$

$$\left\{ (a_k)_{k=1}^{\infty} : \sum_{k=1}^{\infty} \mu_k^{-(r+s)/s} a_k^2 \le 1 \right\},$$

equipped with the usual $\|\cdot\|_{\ell_2(\mathbb{N})}$ norm. On the other hand, Assumption 1 and Lemma 3.1 show that $\mu_k \asymp k^{-s}$. It then follows from the result of Example 5.12 in Wainwright (2019) that $\mathcal{B}(\mathcal{H})$ scales as

$$\log N(\epsilon, \mathcal{B}(\mathcal{H}), \|\cdot\|_C) \asymp \epsilon^{-2/(r+s)}. \tag{23}$$

Next, since the packing number shares the same scaling with ϵ as the covering number, we can find an ϵ-packing set $\{\beta^1, \dots, \beta^M\}$ for $\mathcal{B}(\mathcal{H})$ such that $\|\beta^j - \beta^k\|_C \ge \epsilon$ for any $1 \le j \ne k \le M$, and the packing number $M = M(\epsilon, \mathcal{B}(\mathcal{H}), \|\cdot\|_C)$ satisfies

$$\log M(\epsilon, \mathcal{B}(\mathcal{H}), \|\cdot\|_C) \asymp \epsilon^{-2/(r+s)}. \tag{24}$$

Moreover, by Lemma 3.4, the Kullback-Leibler divergence between P_β and $P_{\beta'}$ can be upper bounded as $\mathcal{K}(P_\beta, P_{\beta'}) \le n\tau(1-\tau)\|\beta - \beta'\|_C^2$; here, the presence of factor n is because the distributions in consideration are for an i.i.d sample of size n. This together with (23) yields

$$\log N_{\mathrm{KL}}(\epsilon, \mathcal{P}) \le \log N(\epsilon/\sqrt{n\tau(1-\tau)}, \mathcal{B}(\mathcal{H}), \|\cdot\|_C) \asymp (\epsilon/\sqrt{n})^{-2/(r+s)}. \tag{25}$$

Finally, we put all these together. It follows from (25) that the choice of $\epsilon_n^2 \asymp n^{1/(r+s+1)}$ satisfies $\epsilon_n^2 \geq \log N_{\mathcal{K}}(\epsilon_n, \mathcal{P})$. Second, for this choice of ϵ_n, it follows from (24) that

$$\log M(\delta_n, \mathcal{B}(\mathcal{H}), \|\cdot\|_C) \asymp \delta_n^{-2/(r+s)} \geq c\{n^{1/(r+s+1)} + 2\log 2\}$$

holds, provided that $\delta_n^2 \asymp n^{-(r+s)/(r+s+1)}$. Comparing these with (21) and (22), we conclude that this δ_n is the desired minimax lower bound under the $\|\cdot\|_C$-norm. □

Proof of Theorem 3.7. Note that, viewed as a linear functional of X, $\eta(X) = \int_{\mathcal{T}} X(t)\beta(t)\,dt$ is determined by β. With a slight abuse of notation, in this proof, we denote the empirical and expected loss functional as $\ell_n(\beta)$ and $\ell_\infty(\beta) = E\ell_n(\beta)$, and the penalty functional as $J(\beta)$.

Because of the convexity of the penalized empirical loss functional $\ell_n(\beta) + \lambda J(\beta)$ to be minimized, we need only to establish the rate of convergence of any local minimizer in the region $\|\beta - \beta_0\|_{\mathcal{H}} \leq 1$, still denoted as $\hat{\beta}$ below. If this local minimizer is not on the boundary of the region, then it is also a global minimizer.

We apply Theorem 3.4.1 of van der Vaart and Wellner (1996). (The consistency of $\hat{\beta}$ is not needed when the two displayed conditions below are valid for all δ.) Denote by $\mathcal{H}_\delta = \{\beta \in \mathcal{H} : \delta/2 < \|\beta - \beta_0\|_C \leq \delta, \|\beta - \beta_0\|_{\mathcal{H}} \leq 1\}$ for any constant $\delta > 0$. According to the cited theorem, if we can show that

$$\inf_{\beta\in\mathcal{H}_\delta} \{\ell_\infty(\beta) - \ell_\infty(\beta_0)\} \gtrsim \delta^2, \tag{26}$$

and

$$E\left\{\sup_{\beta\in\mathcal{H}_\delta} \sqrt{n}\,|(\ell_n - \ell_\infty)(\beta - \beta_0)|\right\} \lesssim \vartheta_n(\delta), \tag{27}$$

for a function $\vartheta_n(\delta)$ satisfying that $\delta^{-\gamma}\vartheta_n(\delta)$ is decreasing for some $\gamma < 2$, then it follows that $\|\hat{\beta} - \beta_0\|_C = O_p(r_n^{-1})$, where r_n satisfies that $r_n^2\vartheta_n(r_n^{-1}) \leq \sqrt{n}$, and $\hat{\beta}$ satisfies that $\ell_n(\hat{\beta}) \leq \ell_n(\beta_0) + O_p(r_n^{-2})$.

Lemma 3.3 immediately implies that the condition (26) holds for $\delta \leq (C_2/\sqrt{16C_3})$. It remains to verify the condition (27), which can be done by investigating the modulus of continuity for the process $\sqrt{n}(\ell_n - \ell_\infty)$. The derivation of the modulus involves the empirical process indexed by the class of functions $\mathcal{G}_\delta = \{g_\beta : \beta \in \mathcal{H}_\delta\}$ where $g_\beta(X,Y) = \rho_\tau\left(Y - \langle X,\beta\rangle_{\mathcal{L}_2}\right) - \rho_\tau\left(Y - \langle X,\beta_0\rangle_{\mathcal{L}_2}\right)$. Using the notations as in van der Vaart and Wellner (1996), we denote by P the joint probability distributions of (X,Y) and by P_n the empirical distribution of $\{X_i,Y_i\}_{i=1}^n$, respectively. Let $\mathbb{G}_n = \sqrt{n}(P_n - P)$ and $\|\mathbb{G}_n\|_{\mathcal{G}_\delta} = \sup_{g\in\mathcal{G}_\delta} |\mathbb{G}_n g|$. Then, the term in the curly brace of the left hand side of (27) can be rewritten as

$$\sup_{\beta\in\mathcal{H}_\delta} \sqrt{n}\,|(\ell_n - \ell_\infty)(\beta - \beta_0)| = \|\mathbb{G}_n\|_{\mathcal{G}_\delta}.$$

Let $\{U_i\}_{i=1}^n$ be an independent and identically distributed sequence of Rademacher variables. The classical symmetrization technique (Lemma 2.3.1, van der Vaart and Wellner, 1996) leads to

$$\begin{aligned} E(\|\mathbb{G}_n\|_{\mathcal{G}_\delta}) &\leq 2\sqrt{n}E\left(\sup_{\beta\in\mathcal{H}_\delta}\left|\frac{1}{n}\sum_{i=1}^n U_i g_\beta(X_i,Y_i)\right|\right) \\ &\leq 8\sqrt{n}E\left(\sup_{\beta\in\mathcal{H}_\delta}\left|\frac{1}{n}\sum_{i=1}^n U_i\langle X_i,\beta-\beta_0\rangle_{\mathcal{L}_2}\right|\right), \end{aligned} \tag{28}$$

where the last inequality follows from the contraction inequality for the Rademacher complexity (Proposition 5.28, Wainwright, 2019) together with the fact that

$$\left|\rho_\tau\left(Y-\langle X,\beta\rangle_{\mathcal{L}_2}\right)-\rho_\tau\left(Y-\langle X,\beta_0\rangle_{\mathcal{L}_2}\right)\right|\le 2\left|\langle X,\beta-\beta_0\rangle_{\mathcal{L}_2}\right|,$$

by Knight's identity in Lemma 3.2.

To handle the inner product $\langle X,\beta-\beta_0\rangle_{\mathcal{L}_2}$ on the right hand side of (28), we appeal to the eigensystem of C because X admits the Karhunen–Loève expansion and $\beta\in\mathcal{H}\cong\mathcal{H}_C^{r/s}$. Recall that C admits the spectral representation (9). On the one hand, for the square integrable stochastic process X with zero mean and the covariance function C, its Karhunen–Loève expansion is $X=\sum_{k=1}^\infty Z_k\phi_k$, where random variables Z_k's are uncorrelated with each other satisfying $E(Z_k)=0$ and $E(Z_k)^2=\mu_k$ for $k\ge 1$. On the other hand, by the norm equivalence between the RKHSs $\mathcal{H}$ and $\mathcal{H}_C^{r/s}$, we can express $\mathcal{H}_\delta$ as

$$\mathcal{H}_\delta=\left\{\sum_{k=1}^\infty b_k\mu_k^{r/(2s)}\phi_k : \delta^2/4<\sum_{k=1}^\infty \mu_k^{(r+s)/s}(b_k-b_{0k})^2\le\delta^2,\sum_{k=1}^\infty(b_k-b_{0k})^2\le 1\right\},$$

where $\beta=\sum_{k=1}^\infty b_k\mu_k^{r/(2s)}\phi_k$, $\beta_0=\sum_{k=1}^\infty b_{0k}\mu_k^{r/(2s)}\phi_k$, and we omit the universal constants appearing in the norm-equivalence (18). Write $\omega_k=\max\{\delta^{-2}\mu_k^{(r+s)/s},1\}$, and it is clear that $\mathcal{H}_\delta$ is contained in the rescaled ball

$$\left\{\sum_{k=1}^\infty b_k\mu_k^{r/(2s)}\phi_k : \sum_{k=1}^\infty \omega_k(b_k-b_{0k})^2\le 2\right\}.$$

Combining the above facts together, we have for $\beta\in\mathcal{H}_\delta$,

$$\langle X,\beta-\beta_0\rangle_{\mathcal{L}_2}=\sum_{k=1}^\infty(b_k-b_{0k})\mu_k^{r/(2s)}Z_k,$$

and the Cauchy-Schwarz inequality further implies that $\langle X,\beta-\beta_0\rangle_{\mathcal{L}_2}^2\le 2\sum_{k=1}^\infty\omega_k^{-1}\mu_k^{r/s}Z_k^2$. Then, it follows that

$$E\left(\sup_{\beta\in\mathcal{H}_\delta}\left|\frac{1}{n}\sum_{i=1}^n U_i\langle X_i,\beta-\beta_0\rangle_{\mathcal{L}_2}\right|\right)^2\le\frac{2}{n}E\left(\sum_{k=1}^\infty\omega_k^{-1}\mu_k^{r/s}Z_k^2\right)=\frac{2}{n}\sum_{k=1}^\infty\min\{\delta^2,\mu_k^{(r+s)/s}\},$$

where the last equality uses the definition of ω_k and the fact that $E(Z_k)^2=\mu_k$ for $k\ge 1$. By Lemma 3.1, $\mu_k\asymp k^{-s}$. We can derive that

$$E\left(\sup_{\beta\in\mathcal{H}_\delta}\left|\frac{1}{n}\sum_{i=1}^n U_i\langle X_i,\beta-\beta_0\rangle_{\mathcal{L}_2}\right|\right)\le\left(\frac{2}{n}\sum_{k=1}^\infty\min\{\delta^2,\mu_k^{(r+s)/s}\}\right)^{1/2}\lesssim n^{-1/2}\delta^{1-1/(r+s)}.$$

Therefore, we conclude from (28) that $E(\|\mathbb{G}_n\|_{\mathcal{G}_\delta})\lesssim\delta^{1-1/(r+s)}$. In (27), we can choose $\vartheta_n(\delta)=\delta^{1-1/(r+s)}$, which satisfies that $\delta^{-\gamma}\vartheta_n(\delta)$ is a decreasing function for $\gamma=1$. Condition (27) has now been verified.

Finally, choose $r_n^2=n^{(r+s)/(r+s+1)}$ such that $r_n^2\vartheta_n(r_n^{-1})\le\sqrt{n}$ holds for every n. By definition, $\hat\beta$ is the minimizer of the penalized empirical loss function $(\ell_n+\lambda J)(\beta)$ such that

$$\ell_n(\hat\beta)+\lambda J(\hat\beta)\le\ell_n(\beta_0)+\lambda J(\beta_0).$$

Recall the definition of the penalty functional $J(\cdot)$, and the fact that $\|\beta_0\|_{\mathcal{H}} = \|\beta_0\|_{\mathcal{W}_2^r}$, $\|\beta_0\|_{\mathcal{H}} \leq B$ implies that $J(\beta_0) \leq cB$ for some constant c. Since $\lambda = O(n^{-(r+s)/(r+s+1)})$, we rearrange the above display to obtain that

$$\ell_n(\hat{\beta}) \leq \ell_n(\beta_0) + \lambda J(\beta_0) - \lambda J(\hat{\beta}) \leq \ell_n(\beta_0) + O_p(n^{-(r+s)/(r+s+1)}).$$

Therefore, we conclude that

$$\|\hat{\beta} - \beta_0\|_C^2 = O_p(n^{-(r+s)/(r+s+1)})$$

by applying Theorem 3.4.1 of van der Vaart and Wellner (1996). □

4. Implementation

4.1. Computation: Interior point algorithm

Using Theorem 2.1, the penalized empirical risk minimization problem (3) defined in an infinite-dimensional function space can be reduced to the problem (7) defined in a finite-dimensional vector space. Numerical computation of the solution of the problem (7) faces several challenges. First, since the check loss function is not differentiable, gradient-based optimization methods cannot be applied. Second, the number of parameters that need to be optimized is usually large in practice, e.g., in our medical imaging application. To be specific, the number of parameters is $n + M + 1$, while the sample size is n. Third, the matrix $\boldsymbol{\Sigma}$ involves calculation of 4-dimensional numerical integrals for n^2 times.

For the spatial function-on-scalar quantile regression, Zhang et al. (2022) developed an iterative ADMM algorithm (Boyd et al., 2011) to obtain the parameters and showed that each individual update has an explicit expression. We can similarly derive an ADMM algorithm for solving the problem (7). However, due to the additional constraint $\mathbf{T}_i^\top \mathbf{c} = 0$ (induced by the thin-plate spline), there is no explicit expression for each individual update in the ADMM iteration. Therefore, the ADMM algorithm is not a feasible approach in our context, and it is necessary to consider an alternative.

We rewrite the problem (7) as follows:

$$\begin{gathered}\underset{\alpha,\mathbf{e},\mathbf{c}}{\arg\min} \sum_{i=1}^{n} \rho_\tau (Y_i - \alpha - \mathbf{T}_i \mathbf{e} - \boldsymbol{\Sigma}_i \mathbf{c}) + \lambda' \mathbf{c}^\top \boldsymbol{\Sigma} \mathbf{c}, \\ \text{subject to } [\mathbf{T}_1^\top, \mathbf{T}_2^\top, \dots, \mathbf{T}_n^\top]\mathbf{c} = \mathbf{0}_{M\times 1},\ i = 1,\dots,n,\end{gathered} \tag{29}$$

where $\lambda' = n\lambda$. Denote $Y_i - \alpha - \mathbf{T}_i e - \boldsymbol{\Sigma}_i c = \varepsilon_i$. We can write

$$\begin{aligned}\varepsilon_i &= u_i - v_i, \\ u_i &= \max(0, \varepsilon_i) = |\varepsilon_i| I(\varepsilon_i > 0), \\ v_i &= \max(0, -\varepsilon_i) = |\varepsilon_i| I(\varepsilon_i < 0).\end{aligned}$$

As a consequence, we can rewrite (29) as the following quadratic programming problem:

$$\begin{gathered}\underset{\mathbf{u},\mathbf{v},\mathbf{c}}{\arg\min}\, \tau \mathbf{u}^\top \mathbf{1} + (1-\tau)\mathbf{v}^\top \mathbf{1} + \lambda' \mathbf{c}^\top \boldsymbol{\Sigma} \mathbf{c}, \\ \text{subject to } Y_i - \alpha - \mathbf{T}_i \mathbf{e} - \boldsymbol{\Sigma}_i \mathbf{c} = u_i - v_i, \\ u_i \geq 0, v_i \geq 0,\ i = 1,\dots,n, \\ [\mathbf{T}_1^\top, \mathbf{T}_2^\top, \dots, \mathbf{T}_n^\top]\mathbf{c} = \mathbf{0}_{M\times 1},\end{gathered} \tag{30}$$

where $\mathbf{u} = [u_1, \ldots, u_n]^\top$, $\mathbf{v} = [v_1, \ldots, v_n]^\top$, and $\mathbf{1} = [1, \ldots, 1]^\top$ are $n \times 1$ column vectors.

In order to transform (30) into a standard form of quadratic programming, we further introduce 6 slack variables, defined as $\alpha^+ = \max(0, \alpha)$, $\alpha^- = \max(0, -\alpha)$, $e^+ = \max(0, e)$, $e^- = \max(0, -e)$, $c^+ = \max(0, c)$, $c^- = \max(0, -c)$. Let

$$\mathbf{x} = [\alpha^+, \alpha^-, e^+, e^-, c^+, c^-, u, v]^\top,$$
$$\mathbf{b} = [0_{1\times1}, 0_{1\times1}, \mathbf{0}_{1\times M}, \mathbf{0}_{1\times M}, \mathbf{0}_{1\times n}, \mathbf{0}_{1\times n}, \tau\mathbf{1}_{1\times n}, (1-\tau)\mathbf{1}_{1\times n}]^\top,$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathbf{0}_{4\times4} & \mathbf{0}_{4\times1} & \mathbf{0}_{4\times1} & \mathbf{0}_{4\times2} \\ \mathbf{0}_{1\times4} & 2\lambda'\boldsymbol{\Sigma} & 2\lambda'\boldsymbol{\Sigma} & \mathbf{0}_{1\times2} \\ \mathbf{0}_{1\times4} & 2\lambda'\boldsymbol{\Sigma} & 2\lambda'\boldsymbol{\Sigma} & \mathbf{0}_{1\times2} \\ \mathbf{0}_{2\times4} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times2} \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{1}_{n\times1} & -\mathbf{1}_{n\times1} & [\mathbf{T}_1, \ldots, \mathbf{T}_M] & -[\mathbf{T}_1, \ldots, \mathbf{T}_M] & [\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n] & -[\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n] & \mathbf{I}_{n\times n} & \mathbf{I}_{n\times n} \\ \mathbf{0}_{M\times1} & \mathbf{0}_{M\times1} & \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} & [\mathbf{T}_1, \ldots, \mathbf{T}_M]^\top & -[\mathbf{T}_1, \ldots, \mathbf{T}_M]^\top & \mathbf{0}_{M\times n} & \mathbf{0}_{M\times n} \end{bmatrix},$$

$$\tilde{\mathbf{Y}} = [Y_1, \ldots, Y_n, \mathbf{0}_{1\times M}]^\top.$$

Then the problem (30) becomes:

$$\begin{aligned} &\arg\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top\tilde{\boldsymbol{\Sigma}}\mathbf{x} + \mathbf{b}^\top\mathbf{x}, \\ &\text{subject to } \mathbf{A}\mathbf{x} = \tilde{\mathbf{Y}}, \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{31}$$

which is a quadratic programming problem in the standard form.

To solve the problem (31), we apply the interior point algorithm (Bertsimas and Tsitsiklis, 1997), which is an attractive approach for solving linear programming, nonlinear programming, and quadratic programming, due to its outstanding efficiency and broad applicability. Roughly speaking, the interior point algorithm iteratively approaches to the optimal value of the objective function from the interior of the feasible regression. One advantage of the interior point algorithm is that it has a polynomial time complexity bound for linear programming (Karmarkar, 1984), which is remarkably faster than the simplex method (Lustig, Marsten and Shanno, 1994). This algorithm was introduced for quantile regression by Portnoy and Koenker (1997), and they showed that the interior point method for solving the linear programming problem for quantile regression can achieve 10- to 100-fold improvement in computational speed, compared with simplex-based methods. For quadratic programming problems, interior point algorithm is also able to exploit problem efficiently and can lead to improved computational complexity bounds (Pearson and Gondzio, 2017, Potra and Wright, 2000).

4.2. Penalty parameter selection

A properly chosen penalty parameter λ is critical to balance the data fidelity and smoothness of the regularized estimator. One natural approach is the widely used leave-one-out cross-validation, but it is not feasible for our problem due to the computation burden of our problem. Another approach is the Generalized Approximate Cross Validation (GACV) criterion developed by Yuan (2006) for choosing the penalty parameter of smoothing splines for nonparametric quantile regression. However, Reiss and Huang (2012) found that GACV often severely overfits for extreme quantiles. Here we use the multifold cross-validation proposed by Zhang (1993), which can avoid the heavy computation of leave-one-out cross validation and do not rely on the approximations of GACV.

The detailed procedure for cross validation is as follows. We first divide the n observation cases into k validation sets $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$ of (roughly) the same size. Then define the k-fold cross-validation criterion as

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in \mathcal{Z}_j} \rho_\tau \left(Y_i - \int_{\mathcal{T}} X_i(t) \hat{\beta}_\lambda^{[-\mathcal{Z}_j]}(t)\, dt \right), \tag{32}$$

where $\hat{\beta}_\lambda^{[-\mathcal{Z}_j]}(t)$ is the estimate based on the observations that exclude $\mathcal{Z}_j$. Generally, a smaller k will produce downward-biased prediction error, while a larger k will increase computational burden as well as produce results that are more variable. The common choice is $k = 5$ or 10, and we use the 5-fold cross validation in our implementation.

5. Application to hippocampus surface data in the ADNI study

We applied the proposed method to analyze the hippocampus surface data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The ADNI study was started in 2004 and was sponsored by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies as well as some nonprofit organizations. The primary goal of the ADNI study is to test whether MRI, PET, other biological biomarkers, clinical and neuropsychological assessments can be combined to measure the process of normal ageing to early mild cognitive impairment (MCI), to late mild cognitive impairment, to dementia or Alzheimer's disease (AD) (Beckett et al., 2015). The data used in this paper were obtained from the Laboratory of Neuro Imaging's Data Archive (https://ida.loni.usc.edu).

Traditional ways to test cognitive impairment is the Mini-Mental State Examination (MMSE), which has been widely used to screen for dementia. The range of MMSE score is 0–30, a score of 20–24 indicates mild cognitive impairment, and a score lower than 20 suggests moderate and severe dementia (Galea and Woodward, 2005). In our study, we analyzed the hippocampus substructure extracted from the baseline T1-weighted MRI scans from the ADNI data set. It is known that the MRI-based measures of atrophy for the hippocampus are strongly correlated with declining cognitive performance (measured by MMSE score), indicating that the hippocampus can serve as the biomarker of AD (Thompson et al., 2004). While other structures can also serve as the biomarkers, such as the whole brain, entorhinal cortex, and ventricular enlargement, hippocampal atrophy is detectable 3 to 5 years before diagnosis (Barnes and Fox, 2014), making it an appealing biomarker. Early diagnosis and intervention of Alzheimer's disease is important since it allows the patient to minimize the disease-related complication as well as to improve quality of life (Santacruz and Swagerty, 2001).

There are many ways to measure the outcome of hippocampal atrophy. Recent studies have shown that surface-based analysis may offer some advantages over other measures such as volume (Qiu et al., 2010). We employed the surface fluid registration-based hippocampal subregional analysis package (Wang et al., 2007, 2009), which left two holes at the front and the back of the hippocampus, and represented the hippocampus as a cylinder such that it can be conformally mapped to a rectangle. As a result, the original 3D surface registration problem degenerated into a 2D surface registration problem. We used radical distance as the outcome measure since it is linked powerfully with the MMSE score at both baseline and follow-up (Riekkinen et al., 1995, Thompson et al., 2004).

The MMSE score has a left-skewed distribution, suggesting that a model focusing on the conditional mean may be not ideal. Applying the popular log transformation on the MMSE score before fitting the conditional mean is not appropriate, since if the original data is left-skewed, then the log transformation will make the original data more left-skewed (Changyong et al., 2014). Therefore, we standardized individual MMSE scores so that they have zero sample mean and unit standard deviation. The standardization has changed the range of the MMSE score from $(6, 20)$ to $(-3.4852, 1.4006)$.

We applied the proposed method based on (1) and (3), to the hippocampus surface data. In this functional linear quantile regression model (referred to as FLQR), the response variable is the (standardized) MMSE score, and the functional covariate is the two-dimensional hippocampus image. The data set contains 798 subjects which belong to three groups, namely, individuals who have AD or Mild Cognitive Impairment (MCI), and healthy controls. Here we only report results of analyzing the healthy controls (denoted as the normal group). A similar analysis has been done for other groups but results are not shown to save space. The normal group has a sample size of $n = 224$. Our focus was on the right hippocampal surface.

We compared our method with the one based on the following least squares problem:

$$(\hat{\alpha}_{LS}, \hat{\beta}_{LS}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathcal{L}_2(\mathcal{T})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \int_{\mathcal{T}} X_i(t) \beta(t)\, dt \right)^2 + \lambda J(\beta), \tag{33}$$

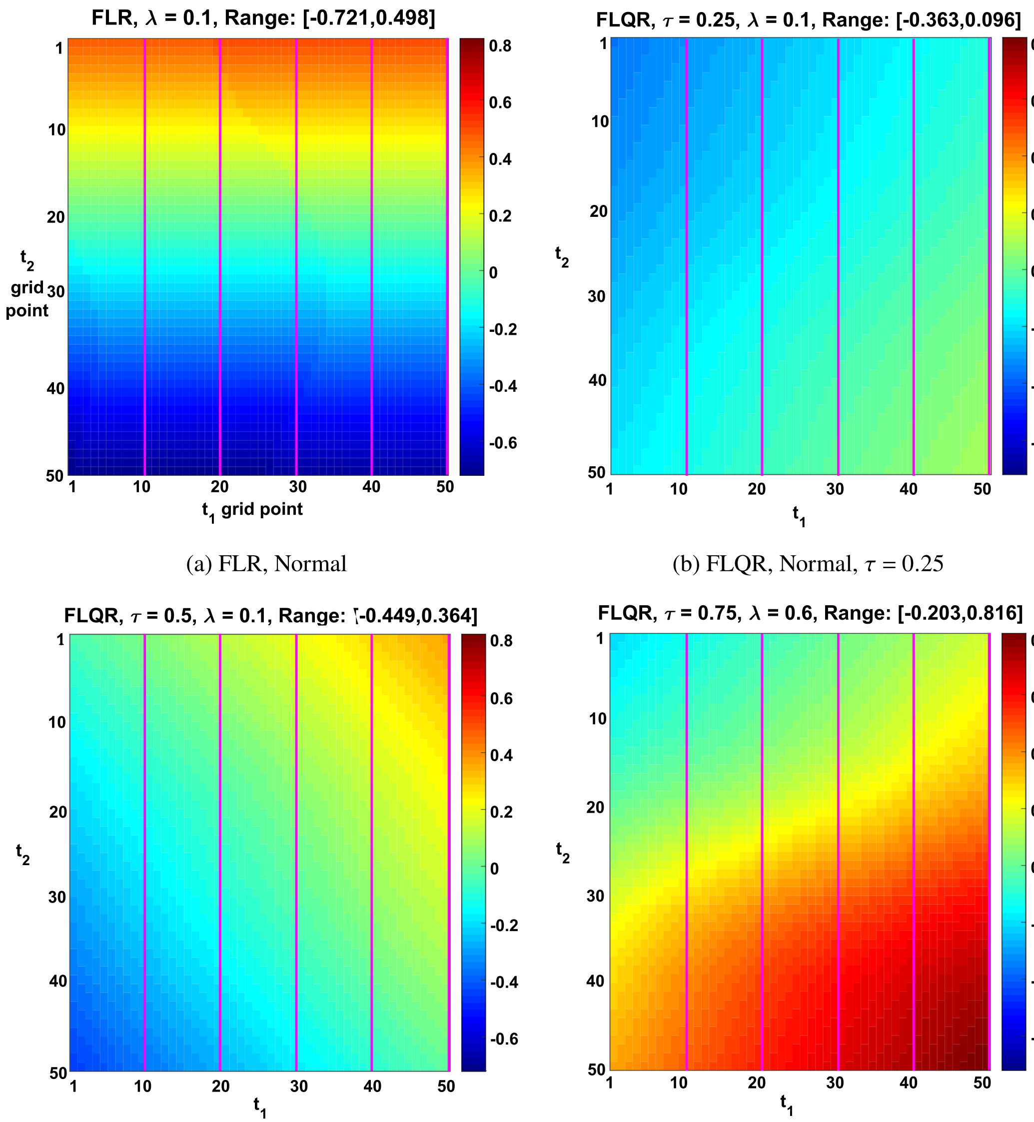
where λ is a penalty parameter that controls the trade-off between the fidelity to the data and roughness of function estimation. We name $\hat{\beta}_{LS}$ as the FLR (functional linear regression) estimate.

The penalty parameter was selected using the 5-fold cross-validation. We found that both FLQR and FLR are not very sensitive to the choice of λ within certain range. Figure 1 shows the estimated slope surface $\hat{\beta}(t_1, t_2)$ for both FLR and FLQR. In this figure, we presented the heatmap using a 50×50 grid. We observed that the fitted slope surfaces for FLR and for FLQR at three different quantile levels ($\tau = 0.25, 0.5, 0.75$) are very different. In particular, the range of the fitted slope surface in the domain of interest is $[-0.721, 0.498]$ for FLR, while for FLQR, it is $[-0.363, 0.096]$, $[-0.449, 0.363]$, $[-0.203, 0.816]$ for $\tau = 0.25, 0.5, 0.75$, respectively. The range of the fitted slope surface by FLR is significantly wider than that by FLQR at all three quantile levels. To further facilitate the comparison, Figure 2 plots the fitted slope surface $\hat{\beta}(t_1, t_2)$ at several selected fixed values of t_1, $t_1 = 0.2, 0.4, 0.6, 0.8, 1$, showing as functions of t_2. The presented functions clear show the differences between the results from FLR and FLQR at different quantile levels.

These comparison results indicate that the conditional mean is not enough in capturing hippocampus substructure for the Alzheimer's disease, i.e., the changing pattern varies among different patient cohort, even for the healthy control group. As a comparison, the quantile regression has the potential to provide a whole picture for diagnosis, with respect to the whole patient group.

6. Discussion

In this article, we study the scalar-on-function quantile regression model where the functional predictor is defined over a two-dimensional domain. We prove that the proposed regularized estimator achieves the minimax optimal convergence rate in terms of prediction risk. Some research directions worth further investigation. First, our result on the simultaneous diagonalization of the reproducing and covariance kernels is of independent interest for multi-dimensional functional data analysis. Existing RKHS-based approach can be adapted to various functional regression models, such as function-on-scalar and function-on-function regression. Second, there are recent advances in penalty parameter selection for scalar-on-scalar quantile regression (Fasiolo et al., 2021, Geraci, 2019, Muggeo et al.,



(a) FLR, Normal

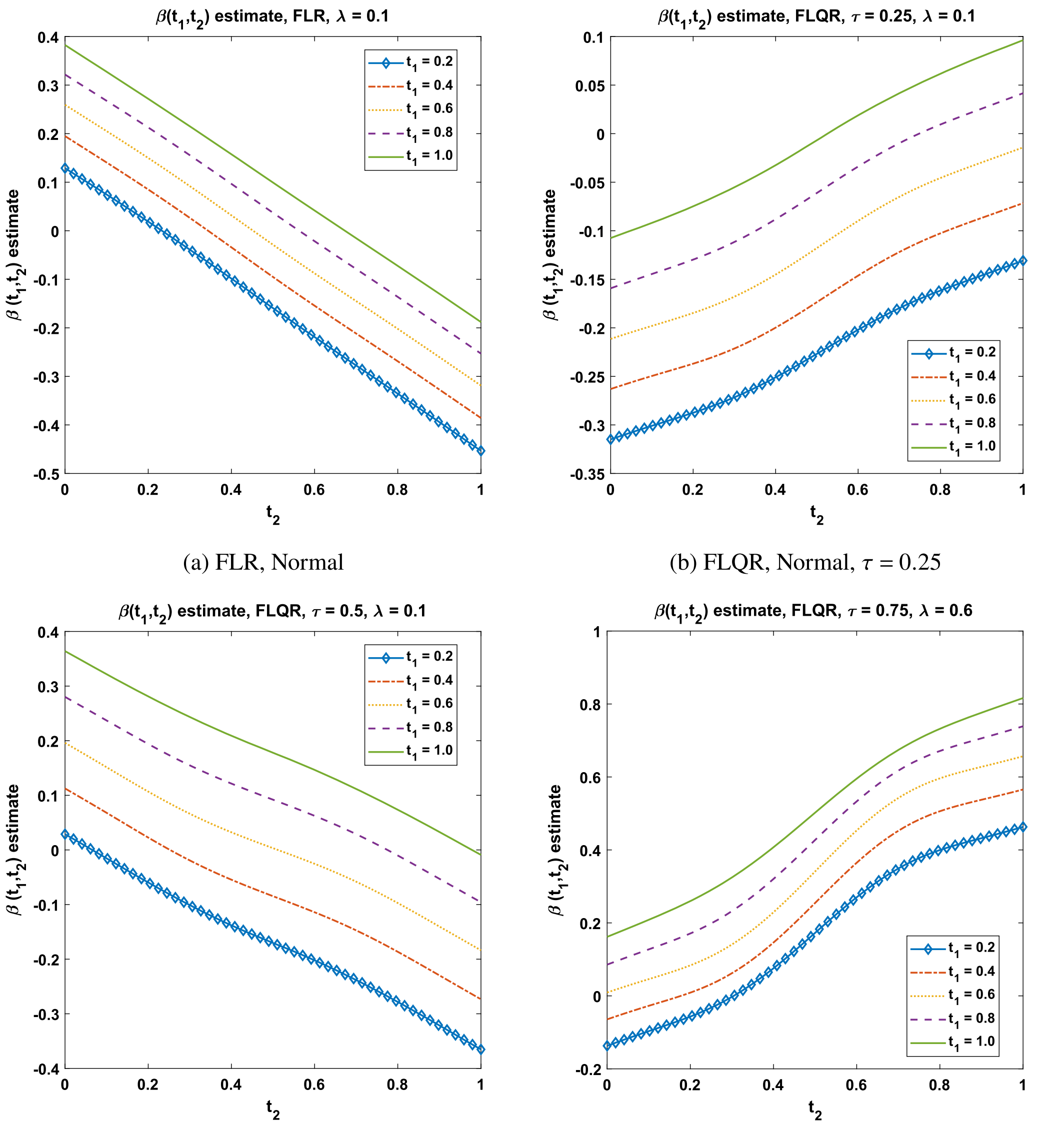
(b) FLQR, Normal, $\tau = 0.25$

(c) FLQR, Normal, $\tau = 0.50$

(d) FLQR, Normal, $\tau = 0.75$

Figure 1. Heatmaps of estimated slope surfaces $\hat{\beta}(t_1,t_2)$ by applying the FLR (Functional Linear Regression) and FLQR (Functional Linear Quantile Regression) on the normal group in the hippocampus surface data. t_1 and t_2 stand for the grid points in plotting $\hat{\beta}(t_1,t_2)$. τ represents the quantile level, λ is the penalty parameter selected by the 5-fold cross-validation, 'Range' means the range of the $\hat{\beta}(t_1,t_2)$ evaluated at the 50×50 grid points.

2021). It is shown that replacing the check loss function with a rounded surrogate loss can be beneficial in particular for more extreme quantile levels. An extension of that approach to functional data analysis can be a promising alternative to the cross validation criterion adopted in this paper.



(a) FLR, Normal

(b) FLQR, Normal, $\tau = 0.25$

(c) FLQR, Normal, $\tau = 0.50$

(d) FLQR, Normal, $\tau = 0.75$

Figure 2. Plots of estimated slope surfaces $\hat{\beta}(t_1, t_2)$ as functions of t_2 for selected t_1 by applying the FLR (Functional Linear Regression) and FLQR (Functional Linear Quantile Regression) on the normal group in the hippocampus surface data. The selected $t_1 = 0.2, 0.4, 0.6, 0.8$ and 1, correspond to grid points 10, 20, 30, 40, and 50 in Figure 1 (shown as vertical lines in magenta color).

References

Adams, R.A. and Fournier, J.J.F. (2003). *Sobolev Spaces*, 2nd ed. *Pure and Applied Mathematics (Amsterdam)* **140**. Amsterdam: Elsevier/Academic Press. MR2424078

Arnone, E., Azzimonti, L., Nobile, F. and Sangalli, L.M. (2019). Modeling spatially dependent functional data via regression with differential regularization. *J. Multivariate Anal.* **170** 275–295. MR3913041 https://doi.org/10.1016/j.jmva.2018.09.006

Barnes, J. and Fox, N.C. (2014). The search for early markers of AD: Hippocampal atrophy and memory deficits. *Int. Psychogeriatr.* **26** 1065–1066. https://doi.org/10.1017/S1041610214000623

Beckett, L.A., Donohue, M.C., Wang, C., Aisen, P., Harvey, D.J., Saito, N. and Initiative, A.D.N. (2015). The Alzheimer's Disease Neuroimaging Initiative phase 2: Increasing the length, breadth, and depth of our understanding. *Alzheimer's Dement.* **11** 823–831.

Bertsimas, D. and Tsitsiklis, J.N. (1997). *Introduction to Linear Optimization*. Athena Scientific.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

Cai, T.T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496 https://doi.org/10.1214/009053606000000830

Cai, T.T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216. MR3010906 https://doi.org/10.1080/01621459.2012.716337

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

Carl, B. and Stephani, I. (1990). *Entropy, Compactness and the Approximation of Operators*. *Cambridge Tracts in Mathematics* **98**. Cambridge: Cambridge Univ. Press. MR1098497 https://doi.org/10.1017/CBO9780511897467

Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H., Ying, L. and TU, X.M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry* **26** 105–109.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344 https://doi.org/10.1214/07-AOS563

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables (Proc. Conf., Math. Res. Inst., Oberwolfach, 1976)*. (W. Schemp and K. Zeller, eds.). *Lecture Notes in Math., Vol. 571* 85–100. Berlin: Springer. MR0493110

Edmunds, D.E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers, Differential Operators*. *Cambridge Tracts in Mathematics* **120**. Cambridge: Cambridge Univ. Press. MR1410258 https://doi.org/10.1017/CBO9780511662201

Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R. and Goude, Y. (2021). Fast calibrated additive quantile regression. *J. Amer. Statist. Assoc.* **116** 1402–1412. MR4309281 https://doi.org/10.1080/01621459.2020.1725521

Galea, M. and Woodward, M. (2005). Mini-mental state examination (MMSE). *Aust. J. Physiother.* **51** 198.

Geraci, M. (2019). Additive quantile regression for clustered data with an application to children's physical activity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1071–1089. MR4002384

Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 https://doi.org/10.1214/009053606000000957

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing* 302–311.

Kato, K. (2012). Estimation in functional linear quantile regression. *Ann. Statist.* **40** 3108–3136. MR3097971 https://doi.org/10.1214/12-AOS1066

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95. MR0290013 https://doi.org/10.1016/0022-247X(71)90184-3

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770. MR1626024 https://doi.org/10.1214/aos/1028144858

Koenker, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge: Cambridge Univ. Press. MR2268657 https://doi.org/10.1017/CBO9780511754098

Koenker, R. and Bassett, G. Jr. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

Koenker, R. and Machado, J.A.F. (1999). Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.* **94** 1296–1310. MR1731491 https://doi.org/10.2307/2669943

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417 https://doi.org/10.1093/biomet/81.4.673

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2017). *Handbook of Quantile Regression*. Boca Raton: CRC Press.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *J. Amer. Statist. Assoc.* **102** 255–268. MR2293307 https://doi.org/10.1198/016214506000000979

Li, R., Lu, W., Zhu, Z. and Lian, H. (2021). Optimal prediction of quantile functional linear regression in reproducing kernel Hilbert spaces. *J. Statist. Plann. Inference* **211** 162–170. MR4122078 https://doi.org/10.1016/j.jspi.2020.06.010

Li, M., Wang, K., Maity, A. and Staicu, A.-M. (2022). Inference in functional linear quantile regression. *J. Multivariate Anal.* **190** Paper No. 104985. MR4402003 https://doi.org/10.1016/j.jmva.2022.104985

Liu, Y., Li, M. and Morris, J.S. (2020). Function-on-scalar quantile regression with application to mass spectrometry proteomics data. *Ann. Appl. Stat.* **14** 521–541. MR4117818 https://doi.org/10.1214/19-AOAS1319

Lustig, I.J., Marsten, R.E. and Shanno, D.F. (1994). Interior point methods for linear programming: Computational state of the art. *ORSA J. Comput.* **6** 1–14. MR1261376 https://doi.org/10.1287/ijoc.6.1.1

Lv, S., Lin, H., Lian, H. and Huang, J. (2018). Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *Ann. Statist.* **46** 781–813. MR3782384 https://doi.org/10.1214/17-AOS1567

Ma, H., Li, T., Zhu, H. and Zhu, Z. (2019). Quantile regression for functional partially linear model in ultra-high dimensions. *Comput. Statist. Data Anal.* **129** 135–147. MR3855203 https://doi.org/10.1016/j.csda.2018.06.005

Morris, J.S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.

Morris, J.S., Baladandayuthapani, V., Herrick, R.C., Sanna, P. and Gutstein, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.* **5** 894–923. MR2840180 https://doi.org/10.1214/10-AOAS407

Muggeo, V.M.R., Torretta, F., Eilers, P.H.C., Sciandra, M. and Attanasio, M. (2021). Multiple smoothing parameters selection in additive regression quantiles. *Stat. Model.* **21** 428–448. MR4307778 https://doi.org/10.1177/1471082X20929802

Riekkinen, P. Jr., Soininen, H., Helkala, E.L., Partanen, K., Laakso, M., Vanhanen, M. and Riekkinen, P. (1995). Hippocampal atrophy, acute THA treatment and memory in Alzheimer's disease. *NeuroReport* **6** 1297–1300. https://doi.org/10.1097/00001756-199506090-00017

Pearson, J.W. and Gondzio, J. (2017). Fast interior point solution of quadratic programming problems arising from PDE-constrained optimization. *Numer. Math.* **137** 959–999. MR3719049 https://doi.org/10.1007/s00211-017-0892-8

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300. MR1619189 https://doi.org/10.1214/ss/1030037960

Potra, F.A. and Wright, S.J. (2000). Interior-point methods. *J. Comput. Appl. Math.* **124** 281–302.

Qiu, A., Brown, T., Fischl, B., Ma, J. and Miller, M.I. (2010). Atlas generation for subcortical and ventricular structures with its applications in shape analysis. *IEEE Trans. Image Process.* **19** 1539–1547. MR2814625 https://doi.org/10.1109/TIP.2010.2042099

Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B* **53** 539–572. MR1125714

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. MR2514435

Reiss, P.T. and Huang, L. (2012). Smoothness selection for penalized quantile regression splines. *Int. J. Biostat.* **8** Art. 10. MR2997664 https://doi.org/10.1515/1557-4679.1381

Ritter, K., Wasilkowski, G.W. and Woźniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *Ann. Appl. Probab.* **5** 518–540. MR1336881

Rockafellar, R.T. (1997). *Convex Analysis*. *Princeton Landmarks in Mathematics*. Princeton, NJ: Princeton Univ. Press. MR1451876

Santacruz, K.S. and Swagerty, D. (2001). Early diagnosis of dementia. *Amer. Fam. Phys.* **63** 703–718.

Schölkopf, B., Herbrich, R. and Smola, A.J. (2001). A generalized representer theorem. In *Computational Learning Theory (Amsterdam, 2001)*. (D. Helmbold and B. Williamson, eds.). *Lecture Notes in Computer Science* **2111** 416–426. Berlin: Springer. MR2042050 https://doi.org/10.1007/3-540-44581-1_27

Sherwood, B. and Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *Ann. Statist.* **44** 288–317. MR3449769 https://doi.org/10.1214/15-AOS1367

Shi, H., Yang, Y., Wang, L., Ma, D., Beg, M.F., Pei, J. and Cao, J. (2022). Two-dimensional functional principal component analysis for image feature extraction. *J. Comput. Graph. Statist.* **31** 1127–1140. MR4513375 https://doi.org/10.1080/10618600.2022.2035738

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. *Springer Series in Statistics*. New York: Springer. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.* **35** 363–417. MR2914365 https://doi.org/10.1007/s00365-012-9153-3

Sun, X., Du, P., Wang, X. and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *J. Amer. Statist. Assoc.* **113** 1601–1611. MR3902232 https://doi.org/10.1080/01621459.2017.1356320

Thompson, P.M., Hayashi, K.M., Zubicaray, G.I.D., Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D.H., Gravano, D., Doddrell, D.M. and Toga, A.W. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *NeuroImage* **22** 1754–1766. https://doi.org/10.1016/j.neuroimage.2004.03.040

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

Volgushev, S., Chao, S.-K. and Cheng, G. (2019). Distributed inference for quantile regression processes. *Ann. Statist.* **47** 1634–1662. MR3911125 https://doi.org/10.1214/18-AOS1730

Wahba, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Philadelphia, PA: SIAM. MR1045442 https://doi.org/10.1137/1.9781611970128

Wainwright, M.J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge: Cambridge Univ. Press. MR3967104 https://doi.org/10.1017/9781108627771

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

Wang, J., Wong, R.K.W. and Zhang, X. (2022). Low-rank covariance function estimation for multidimensional functional data. *J. Amer. Statist. Assoc.* **117** 809–822. MR4436314 https://doi.org/10.1080/01621459.2020.1820344

Wang, Y., Lui, L.M., Gu, X., Hayashi, K.M., Chan, T.F., Toga, A.W., Thompson, P.M. and Yau, S.-T. (2007). Brain surface conformal parameterization using Riemann surface structure. *IEEE Trans. Med. Imag.* **26** 853–865.

Wang, Y., Chan, T.F., Toga, A.W. and Thompson, P.M. (2009). Multivariate tensor-based brain anatomical surface morphometry via holomorphic one-forms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 337–344. Springer.

Wendland, H. (2005). *Scattered Data Approximation*. *Cambridge Monographs on Applied and Computational Mathematics* **17**. Cambridge: Cambridge Univ. Press. MR2131724

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500 https://doi.org/10.1214/aos/1017939142

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660

Yao, F., Sue-Chee, S. and Wang, F. (2017). Regularized partially functional quantile regression. *J. Multivariate Anal.* **156** 39–56. MR3624684 https://doi.org/10.1016/j.jmva.2017.02.001

Yuan, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **50** 813–829. MR2207010 https://doi.org/10.1016/j.csda.2004.10.008

Yuan, M. and Cai, T.T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857 https://doi.org/10.1214/09-AOS772

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21** 299–313. MR1212178 https://doi.org/10.1214/aos/1176349027

Zhang, Z., Wang, X., Kong, L. and Zhu, H. (2022). High-dimensional spatial quantile function-on-scalar regression. *J. Amer. Statist. Assoc.* **117** 1563–1578. MR4480732 https://doi.org/10.1080/01621459.2020.1870984

Zhou, L. and Pan, H. (2014). Principal component analysis of two-dimensional functional data. *J. Comput. Graph. Statist.* **23** 779–801. MR3224656 https://doi.org/10.1080/10618600.2013.827986

Received October 2022 and revised April 2023