2012

Minimax and Adaptive Prediction for Functional Linear Regression

T. Tony Cai
University of Pennsylvania

Ming Yuan

Recommended Citation

Cai, T., & Yuan, M. (2012). Minimax and Adaptive Prediction for Functional Linear Regression. *Journal of the American Statistical Association, 107* (499), 1201-1216. http://dx.doi.org/10.1080/01621459.2012.716337

Minimax and Adaptive Prediction for Functional Linear Regression

Abstract

This article considers minimax and adaptive prediction with functional predictors in the framework of functional linear model and reproducing kernel Hilbert space. Minimax rate of convergence for the excess prediction risk is established. It is shown that the optimal rate is determined jointly by the reproducing kernel and the covariance kernel. In particular, the alignment of these two kernels can significantly affect the difficulty of the prediction problem. In contrast, the existing literature has so far focused only on the setting where the two kernels are nearly perfectly aligned. This motivates us to propose an easily implementable data-driven roughness regularization predictor that is shown to attain the optimal rate of convergence adaptively without the need of knowing the covariance kernel. Simulation studies are carried out to illustrate the merits of the adaptive predictor and to demonstrate the theoretical results.

Keywords

functional linear model, minimax rate of convergence, principal components analysis, reproducing kernel Hilbert space, spectral decomposition

Disciplines

Statistics and Probability

Minimax and Adaptive Prediction for Functional Linear Regression

T. Tony Cai* and Ming Yuan†

University of Pennsylvania and Georgia Institute of Technology

May 18, 2012

Abstract

This paper considers minimax and adaptive prediction with functional predictors in the framework of functional linear model and reproducing kernel Hilbert space. Minimax rate of convergence for the excess prediction risk is established. It is shown that the optimal rate is determined jointly by the reproducing kernel and the covariance kernel. In particular, the alignment of these two kernels can significantly affect the difficulty of the prediction problem. In contrast, the existing literature has so far focused only on the setting where the two kernels are nearly perfectly aligned. This motivates us to propose an easily implementable data-driven roughness regularization predictor that is shown to attain the optimal rate of convergence adaptively without the need of knowing the covariance kernel. Simulation studies are carried out to illustrate the merits of the adaptive predictor and to demonstrate the theoretical results.

Keywords: Adaptive prediction, functional linear model, minimax rate of convergence, principal components analysis, reproducing kernel Hilbert space, spectral decomposition.

*Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.

†H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332. The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234.

1 Introduction

Prediction is an important problem in functional data analysis with a wide range of applications including chemometrics, econometrics, and biomedical studies. See, for example, Ramsay and Silverman (2002; 2005), Ferraty and Vieu (2006), and Ramsay, Hooker and Graves (2009). Consider the functional linear model

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

where Y is a scalar response, $X : \mathcal{T} \to \mathbb{R}$ is a square integrable functional predictor defined over a compact domain $\mathcal{T} \subset \mathbb{R}$, α_0 is the intercept, $\beta_0 : \mathcal{T} \to \mathbb{R}$ is the slope function, and ϵ is random noise with mean 0 and finite variance σ^2. In this paper we focus on the random design where X is a path of a square integrable stochastic process defined over $\mathcal{T}$ and is independent of ϵ . The goal of prediction is to recover the functional η_0:

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

based on a training sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ consisting of n independent copies of (X, Y). Let $\hat{\eta}_n$ be a prediction rule constructed from the training data. Then its accuracy can be naturally measured by the excess risk:

$$\mathcal{E}(\hat{\eta}_n) := \mathbb{E}^* \left[Y^* - \hat{\eta}_n(X^*)\right]^2 - \mathbb{E}^* \left[Y^* - \eta_0(X^*)\right]^2 = \mathbb{E}^* \left[\hat{\eta}_n(X^*) - \eta_0(X^*)\right]^2,$$

where (X^*, Y^*) is a copy of (X, Y) independent of the training data, and $\mathbb{E}^*$ represents expectations taken over X^* and Y^* only. In particular, the rate of convergence of $\mathcal{E}(\hat{\eta}_n)$ as the sample size n increases reflects the difficulty of the prediction problem. A closely related but different problem is that of estimating the intercept α_0 and the slope function β_0.

Many commonly used approaches to the prediction and estimation problems under the functional linear model (1) are based upon the functional principal component analysis (FPCA) (see, e.g., Ramsay and Silverman, 2005; Yao, Müller and Wang, 2005; Cai and Hall, 2006; Hall and Horowitz, 2007). The functional principal components are determined solely by the observed functional predictors X_i. A crucial condition for the success of the FPCA-based methods is that the slope function β_0 is efficiently represented in terms of the leading functional principal components. This condition, however, may not always be true. In many applications it is not realistic to assume that the functional principal components form an efficient basis for the slope function β_0 because the two are typically unrelated. When this condition fails to hold, the low-variance components of the predictor X have non-negligible predictive power and the FPCA-based methods may not

perform well. Similar phenomenon has also been observed in the performance of principle component regression (see, e.g., Jolliffe, 1982) or singular value decomposition methods for linear inverse problems (see, e.g., Donoho, 1995).

For illustration purpose, take the Canadian weather data as an example. It is one of the classical examples in functional linear regression and the goal is to predict the log annual precipitation at 35 different stations based on the average daily temperature. More detailed discussion of the data and analysis can be found in Ramsay and Silverman (2005) and Section 4 of the present paper. The Fourier coefficients of the slope function with respect to the eigenfunctions of the sample covariance are estimated using the FPCA approach. The eigenvalues of the sample covariance and the estimated Fourier coefficients are shown in Figure 1. It is clear that, although the eigenvalues decay nicely, the estimated Fourier coefficients do not decay at all. This is a typical example for the case when the slope function is not well represented by the leading principal components.

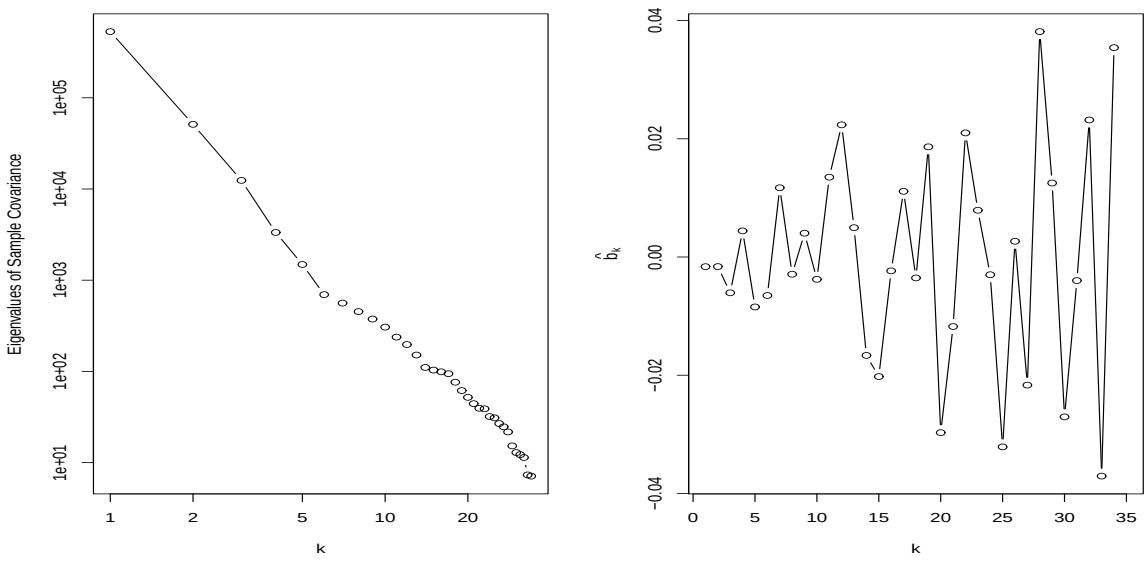

Figure 1: Canadian Weather Data: The eigenvalues of the sample covariance function is given in the left panel. Note that both axes in the left panel are in log scale. The right panel shows the estimated Fourier coefficients of the slope function with respect to the eigenfunctions of the sample covariance function.

In this paper we study the minimax and adaptive prediction problems in the reproducing kernel Hilbert space (RKHS) framework under which the unknown slope function β_0 is assumed to reside in a reproducing kernel Hilbert space $\mathcal{H}(K)$ with a reproducing kernel K. The minimax rate of convergence of the excess prediction risk is established in a general setting with no constraint on

the relationship between the reproducing kernel K and the covariance function C of the random predictor X. It is shown that, under the functional linear model (1), the difficulty of the prediction problem as measured by the minimax rate of convergence depends on both kernels K and C. In particular, the alignment of K and C can significantly affect the optimal rate of convergence. The FPCA-based methods mentioned earlier correspond to the special setting where K and C are assumed to be perfectly aligned, i.e., K and C share a common ordered set of eigenfunctions. The optimal rate of convergence of $\mathcal{E}(\hat{\eta}_n)$ in this case is determined by the rate of decay of the product of the corresponding eigenvalues of K and C (Cai and Hall, 2006). When K and C are not well aligned, as in the Canadian weather data example, the FPCA-based methods may not perform well.

The optimal rate of convergence for the prediction problem is established in two steps. First, a minimax lower bound is derived for the prediction problem. Then a roughness regularization predictor is introduced and is shown to attain the rate of convergence given in the lower bound, when the tuning parameter is appropriately chosen. This estimator is thus rate-optimal. The optimal choice of the tuning parameter depends however on the unknown covariance structure of the predictors X_i. A data-driven procedure for choosing the tuning parameter is then introduced. It is shown that the resulting procedure automatically achieves the optimal rate of convergence for a large collection of covariance functions. The adaptive procedure is easy to implement. Simulation studies are carried out to illustrate the merits of the adaptive predictor and to demonstrate the theoretical results.

The rest of the paper is organized as follows. In Section 2, after basic notation and definitions are reviewed, we establish the optimal rate of convergence for the prediction problem by deriving both minimax lower and upper bounds. A roughness regularization predictor is shown to be rate-optimal when the tuning parameter is appropriately chosen. Section 3 proposes a data-driven method for choosing the tuning parameter and the resulting predictor is shown to adaptively achieve the optimal rate of convergence. Numerical experiments are reported in Section 4 to demonstrate the practical implications of the theoretical developments using both simulated and real data sets. The proofs are given in Section 6. We conclude with some discussions in Section 5.

2 Optimal Rate of Convergence

In this section we establish the minimax rate of convergence of the excess prediction risk. The optimal rate is jointly determined by the reproducing kernel K and the covariance function C and the alignment between K and C plays an important role. We begin by reviewing basic notation

and properties regarding the reproducing kernel K and the covariance function C.

2.1 Notation and definitions

Let $\mathcal{T} \subset \mathbb{R}$ be a compact set. A reproducing kernel $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a real, symmetric, square integrable, and nonnegative definite function. There is a one-to-one correspondence between a reproducing kernel K and a reproducing kernel Hilbert space $\mathcal{H}(K)$ which is a linear functional space endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(K)}$ such that for any $t \in \mathcal{T}$, $K(t, \cdot) \in \mathcal{H}(K)$, and

$$f(t) = \langle K(t, \cdot), f \rangle_{\mathcal{H}(K)}, \qquad \text{for any } f \in \mathcal{H}(K).$$

See, e.g., Wahba (1990). Let $X(\cdot)$ be a square integrable stochastic process defined over $\mathcal{T}$. The covariance function of X is also a real, symmetric, and nonnegative definite function defined as

$$C(s, t) = \mathbb{E}\left([X(s) - \mathbb{E}(X(s))][X(t) - \mathbb{E}(X(t))]\right), \qquad \forall s, t \in \mathcal{T}.$$

The covariance function C is square integrable if $\mathbb{E}\|X\|^2_{\mathcal{L}_2} < \infty$ where

$$\|f\|^2_{\mathcal{L}_2} = \langle f, f \rangle_{\mathcal{L}_2}, \qquad \text{and} \qquad \langle f, g \rangle_{\mathcal{L}_2} = \int_{\mathcal{T}} f(t) g(t) dt.$$

For a real, symmetric, square integrable, and nonnegative definite function $R : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, let $L_R : \mathcal{L}_2 \to \mathcal{L}_2$ denote an integral operator defined by

$$L_R(f)(\cdot) = \langle R(s, \cdot), f \rangle_{\mathcal{L}_2} = \int_{\mathcal{T}} R(s, \cdot) f(s) ds.$$

By the Riesz representation theorem, L_R can also be equivalently defined through

$$\langle f, L_R(g) \rangle_{\mathcal{H}(R)} = \langle f, g \rangle_{\mathcal{L}_2}.$$

The spectral theorem implies that there exist a set of orthonormalized eigenfunctions $\{\psi^R_k : k \geq 1\}$ and a sequence of eigenvalues $\theta^R_1 \geq \theta^R_2 \geq \ldots > 0$ such that

$$R(s, t) = \sum_{k \geq 1} \theta^R_k \psi^R_k(s) \psi^R_k(t), \qquad \forall s, t \in \mathcal{T},$$

and

$$L_R(\psi^R_k) = \theta^R_k \psi^R_k, \qquad k = 1, 2, \ldots.$$

Throughout the paper we shall say that $\{(\theta^R_k, \psi^R_k) : k \geq 1\}$ are the eigenvalue-eigenfunction pairs of the operator R, with the understanding that the pairs are ordered according to the eigenvalues in descending order, $\theta^R_1 \geq \theta^R_2 \geq \ldots$. We say two linear operators (or their corresponding kernels) are

perfectly aligned if they share the same ordered set of eigenfunctions (according to their eigenvalues in descending order).

Let $L_{R^{1/2}}$ be a linear operator defined by

$$L_{R^{1/2}}(\psi_k^R) = \sqrt{\theta_k^R}\psi_k^R,$$

where

$$R^{1/2}(s,t) = \sum_{k\geq 1} \sqrt{\theta_k^R}\psi_k^R(s)\psi_k^R(t), \qquad \forall s,t \in \mathcal{T}.$$

It is clear that $L_{R^{1/2}} = L^{1/2}R$. Moreover, define

$$(R_1R_2)(s,t) = \int_{\mathcal{T}} R_1(s,u)R_2(u,t)dt.$$

Then $L_{R_1R_2} = L_{R_1} \circ L_{R_2}$.

For a given reproducing kernel K and a covariance function C, define the linear operator $L_{K^{1/2}CK^{1/2}}$ by $L_{K^{1/2}CK^{1/2}} = L_{K^{1/2}} \circ L_C \circ L_{K^{1/2}}$, i.e.,

$$L_{K^{1/2}CK^{1/2}}(f) = L_{K^{1/2}}\left(L_C\left(L_{K^{1/2}}(f)\right)\right).$$

If both $L_{K^{1/2}}$ and L_C are bounded linear operators, so is $L_{K^{1/2}CK^{1/2}}$. By the spectral theorem, there exist a sequence of positive eigenvalues $s_1 \geq s_2 \geq \ldots > 0$ and a set of orthonormalized eigenfunctions $\{\varphi_k : k \geq 1\}$ such that

$$K^{1/2}CK^{1/2}(s,t) = \sum_{k\geq 1} s_k\varphi_k(s)\varphi_k(t), \qquad \forall s,t \in \mathcal{T},$$

and

$$L_{K^{1/2}CK^{1/2}}(\varphi_k) = s_k\varphi_k, \qquad k = 1,2,\ldots.$$

It is easy to see that the eigenvalues $\{s_k : k \geq 1\}$ of the linear operator $L_{K^{1/2}CK^{1/2}}$ depend on the eigenvalues of both the reproducing kernel K and the covariance function C as well as the alignment between K and C. We shall show in Sections 2.2 and 2.3 that the difficulty of the prediction problem as measured by the minimax rate of convergence of the excess prediction risk is determined by the decay rate of the eigenvalues $\{s_k : k \geq 1\}$.

For two sequences $\{a_k : k \geq 1\}$ and $\{b_k : k \geq 1\}$ of positive real numbers, we write $a_k \asymp b_k$ if there are positive constants c and C independent of k such that $c \leq a_k/b_k \leq C$ for all $k \geq 1$.

2.2 Minimax lower bound

The optimal rate of convergence of the excess prediction risk will be established in two steps. We first derive a minimax lower bound in this section and then show in Section 2.3 the convergence rate of the lower bound is in fact optimal by constructing a predictor that attains this rate of convergence. The minimax lower bound is given in the following theorem.

Theorem 1 *Suppose the eigenvalues* $\{s_k : k \geq 1\}$ *of the linear operator* $L_{K^{1/2}CK^{1/2}}$ *satisfy* $s_k \asymp k^{-2r}$ *for some constant* $0 < r < \infty$*, then the excess prediction risk satisfies*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\tilde{\eta}}\sup_{\beta_0\in\mathcal{H}(K)}\mathbb{P}\left\{\mathcal{E}(\tilde{\eta}) \geq an^{-\frac{2r}{2r+1}}\right\} = 1, \tag{2}$$

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

It is interesting to compare Theorem 1 with some of the known results in the literature in which the reproducing kernel K and the covariance function C are assumed to be perfectly aligned, i.e., they share the same ordered set of eigenfunctions. The minimax lower bound obtained here generalizes the earlier results for this special case. Let

$$X(t) = \sum_{k=1}^{\infty} Z_k\psi_k(t)$$

be the Karhunen-Loève decomposition of X where $\{Z_k : k \geq 1\}$ are uncorrelated random variables and $\{\psi_k : k \geq 1\}$ is an orthonormal basis of $\mathcal{L}_2(\mathcal{T})$. Set $\theta_k^C = \mathrm{Var}(Z_k)$. We shall assume that θ_k^C are indexed in descending order, $\theta_1^C \geq \theta_2^C \geq \ldots$. Then it is clear that $\{(\theta_k^C, \psi_k) : k \geq 1\}$ also constitutes the eigenvalue-eigenfunction pairs of the covariance function C, i.e.,

$$C(s,t) = \sum_{k=1}^{\infty}\theta_k^C\psi_k(s)\psi_k(t), \qquad \text{for } s,t \in \mathcal{T}.$$

Consider the case where the reproducing kernel K is perfectly aligned with C, i.e.,

$$K(s,t) = \sum_{k=1}^{\infty}\theta_k^K\psi_k(s)\psi_k(t), \qquad \text{for } s,t \in \mathcal{T},$$

with $\theta_1^K \geq \theta_2^K \geq \ldots \geq 0$ being the eigenvalues of K. In this case it is easy to see that

$$L_{K^{1/2}CK^{1/2}}(\psi_k) = \theta_k^K\theta_k^C\psi_k, \qquad k = 1, 2, \ldots,$$

indicating that the eigenvalues $s_k = \theta_k^K\theta_k^C$, $k = 1, 2, \ldots$, and $s_1 \geq s_2 \geq \ldots$. If $\theta_k^K \asymp k^{-2r_1}$ and $\theta_k^C \asymp k^{-2r_2}$, then s_k decays at the rate $k^{-2(r_1+r_2)}$ and by Theorem 1,

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\tilde{\eta}}\sup_{\beta_0\in\mathcal{H}(K)}\mathbb{P}\left\{\mathcal{E}(\tilde{\eta}) \geq an^{-\frac{2(r_1+r_2)}{2(r_1+r_2)+1}}\right\} = 1.$$

This special setting coincides with those considered in Cai and Hall (2006) and Yuan and Cai (2010). Similar results have been established earlier in these papers for this special setting.

In general, however, the eigenvalues of K and C alone cannot determine the decay rate of the eigenvalues of $L_{K^{1/2}CK^{1/2}}$. For example, when $\theta_k^K \asymp k^{-2r_1}$, $\theta_k^C \asymp k^{-2r_2}$ and $\psi_k^C = \psi_{k^2}^K$ for $k = 1, 2, \ldots$, then $s_k \asymp k^{-(4r_1+2r_2)}$.

2.3 Minimax upper bound

Section 2.2 developed a minimax lower bound for the excess prediction risk. We shall now consider the minimax upper bound and show that the lower bound established in Theorem 1 can in fact be achieved. We shall construct a predictor using a roughness regularization method and show that the predictor is asymptotically rate-optimal.

One of the most commonly used methods in nonparametric function estimation is the roughness regularization method (see, e.g., Ramsay and Silverman, 2005) where the intercept α_0 and the slope function β_0 are estimated by

$$\left(\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda}\right) = \underset{a\in\mathbb{R}, b\in\mathcal{H}(K)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - a - \int_{\mathcal{T}} X_i(t) b(t) dt \right)^2 + \lambda \|b\|_{\mathcal{H}(K)}^2 \right\}. \tag{3}$$

Here $\lambda > 0$ is a tuning parameter that balances the tradeoff between the fidelity to the data measured by the sum of squares and the smoothness of the estimate measured by the squared reproducing kernel Hilbert space norm. The estimate $\hat{\beta}_\lambda$ is readily computable even though the minimization in (3) is taken over an infinitely dimensional function space $\mathcal{H}(K)$. More specifically, $\hat{\beta}_{n\lambda}$ can be expressed as

$$\hat{\beta}_{n\lambda}(\cdot) = \sum_{i=1}^{n} c_i \int_{\mathcal{T}} K(t, \cdot) X_i(t) dt \tag{4}$$

for some $c_1, \ldots, c_n \in \mathbb{R}$, and they can be computed together with $\hat{\alpha}_{n\lambda}$ by plugging (4) back to (3). The readers are referred to Yuan and Cai (2010) for more details on the implementation.

Given the estimates $\hat{\alpha}_{n\lambda}$ and $\hat{\beta}_{n\lambda}$, the predictor $\hat{\eta}_{n\lambda}$ is obtained by

$$\hat{\eta}_{n\lambda}(X) = \hat{\alpha}_{n\lambda} + \int_{\mathcal{T}} X(t) \hat{\beta}_{n\lambda}(t) dt.$$

The following theorem states that with an appropriately chosen λ, the predictor $\hat{\eta}_{n\lambda}$ attains the convergence rate of the lower bound given by Theorem 1 and is therefore rate optimal.

Theorem 2 *Assume that there exists a constant $c > 0$ such that for any square integrable function f defined over the domain $\mathcal{T}$,*

$$\mathbb{E}\left(\int_{\mathcal{T}} X(t) f(t) dt \right)^4 \le c \left(\mathbb{E}\left(\int_{\mathcal{T}} X(t) f(t) dt \right)^2 \right)^2. \tag{5}$$

Suppose the eigenvalues $\{s_k : k \geq 1\}$ *satisfy* $s_k \asymp k^{-2r}$ *for some constant* $0 < r < \infty$*, then*

$$\lim_{A\to\infty} \lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{H}(K)} \mathbb{P}\left\{\mathcal{E}(\hat{\eta}_{n\lambda}) \geq An^{-\frac{2r}{2r+1}}\right\} = 0, \tag{6}$$

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

Condition (5) states that linear functionals of X have bounded kurtosis, which is satisfied in particular with $c = 3$ when X follows a Gaussian process.

Theorems 1 and 2 together show that the minimax rate of convergence for the excess prediction risk is

$$n^{-2r/(2r+1)}$$

which is determined by the rate of decay of the eigenvalues of the operator $L_{K^{1/2}CK^{1/2}}$. The optimal rate of convergence depends not only on the eigenvalues of K and C but also on how their eigenfunctions align with each other.

3 Adaptive Prediction

Section 2 established the minimax rate of convergence for the excess prediction risk. As shown in Theorem 2, the optimal rate can be attained by the roughness regularization predictor $\hat{\eta}_{n\lambda}$ when the tuning parameter λ is chosen appropriately. However, the proper choice of λ depends on r, which is unknown since it is determined by the linear operator $L_{K^{1/2}CK^{1/2}}$ and the covariance function C is not known apriori. It is important to develop a data-driven choice of λ that does not require the knowledge of C. In this section, we introduce such an adaptive method for choosing λ.

To motivate our procedure, recall that r represents the decay rate of the eigenvalues of $L_{K^{1/2}CK^{1/2}}$, which can be naturally approximated by $L_{K^{1/2}C_nK^{1/2}}$ where

$$C_n(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t)).$$

Observe that

$$\begin{aligned}
\langle f, L_{K^{1/2}C_nK^{1/2}}g\rangle_{\mathcal{L}_2} &= \langle L_{K^{1/2}}f, L_{C_n}L_{K^{1/2}}g\rangle_{\mathcal{L}_2} \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle L_{K^{1/2}}f, X_i - \bar{X}\rangle_{\mathcal{L}_2}\langle L_{K^{1/2}}g, X_i - \bar{X}\rangle_{\mathcal{L}_2} \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle f, L_{K^{1/2}}X_i - \bar{X}\rangle_{\mathcal{L}_2}\langle g, L_{K^{1/2}}X_i - \bar{X}\rangle_{\mathcal{L}_2} \\
&= \langle f, L_{H_n}g\rangle_{\mathcal{L}_2},
\end{aligned}$$

where

$$H_n(s,t) = \frac{1}{n}\sum_{i=1}^{n}(L_{K^{1/2}}(X_i - \bar{X}))(s)(L_{K^{1/2}}(X_i - \bar{X}))(t).$$

By duality, the eigenvalues of H_n are also the eigenvalues of the Gram matrix $G = (G_{ij})_{1\le i,j\le n}$ where

$$G_{ij} = \frac{1}{n}\langle L_{K^{1/2}}(X_i - \bar{X}), L_{K^{1/2}}(X_i - \bar{X})\rangle_{\mathcal{L}_2} = \frac{1}{n}\int_{\mathcal{T}^2}(X_i - \bar{X})(s)K(s,t)(X_j - \bar{X})(t)dsdt.$$

Information on r can thus be recovered from the eigenvalues of G.

Write

$$\gamma_n(\delta) = \left(\frac{1}{n}\sum_{k\ge 1}\min\{s_k, \delta^2\}\right)^{1/2}$$

and define

$$\rho(C,K) := \inf\left\{\rho \ge n^{-1}\log n : \gamma_n(\delta) \le \rho^{1/2}\delta + \rho,\ \forall\delta \in [0,1]\right\}.$$

It is not hard to see that $\rho \asymp n^{-2r/(2r+1)}$ if $s_k \asymp k^{-2r}$ (see, e.g., Mendelson, 2002). Therefore, we can use an estimate of ρ as our choice of the tuning parameter in defining $\hat{\eta}_\lambda$. To this end, we consider the sample analogue of ρ. Denote by $\hat{s}_1 \ge \hat{s}_2 \ge \ldots \ge \hat{s}_n$ the eigenvalues of G. Write

$$\hat{\rho}(G) := \inf\left\{\rho \ge n^{-1}\log n : \hat{\gamma}_n(\delta) \le \rho^{1/2}\delta + \rho,\ \forall\delta \in [0,1]\right\},$$

where

$$\hat{\gamma}_n(\delta) = \left(\frac{1}{n}\sum_{k=1}^{n}\min\{\hat{s}_k, \delta^2\}\right)^{1/2}.$$

Theorem 3 shows that with high probability $\hat{\rho}$ is of the same order as ρ whenever $\|L_{K^{1/2}}X\|_{\mathcal{L}_2}$ has exponential tails.

Theorem 3 *Assume that there exist some constants c_1, $c_2 > 0$ such that*

$$\mathbb{P}\{\|L_{K^{1/2}}X\|_{\mathcal{L}_2} \ge x\} \le c_1\exp(-c_2x^2), \qquad \text{for } x \ge 0. \tag{7}$$

Then there are constants $0 < c_3 < c_4 < \infty$ such that

$$\mathbb{P}\left\{c_3 \le \frac{\hat{\rho}}{\rho} \le c_4\right\} \to 1$$

as $n \to \infty$.

We note that the tail condition (7) for $\|L_{K^{1/2}}X\|_{\mathcal{L}_2}$ can also be replaced by

$$\mathbb{P}\{\|X\|_{\mathcal{L}_2} \ge x\} \le c_1\exp(-c_2x^2), \qquad \text{for } x \ge 0,$$

because $\|L_{K^{1/2}}X\|_{\mathcal{L}_2} \leq \theta_1^K \|X\|_{\mathcal{L}_2}$. It holds true, in particular, if X is a Gaussian process.

The following result shows that the predictor $\hat{\eta}_{n\lambda}$ with the data-driven choice of the tuning parameter λ indeed adaptively achieves the optimal rate of convergence.

Theorem 4 *Assume that there exist constants $c_1, c_2 > 0$ such that*

$$\mathbb{P}\left\{\|L_{K^{1/2}}X\|_{\mathcal{L}_2} \geq x\right\} \leq c_1 \exp(-c_2 x^2), \qquad \text{for } x \geq 0,$$

and for any $f \in \mathcal{L}_2$

$$\mathbb{E}\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^4 \leq c\left(\mathbb{E}\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^2\right)^2. \tag{8}$$

Suppose the eigenvalues $\{s_k : k \geq 1\}$ satisfy $s_k \asymp k^{-2r}$ for some constant $0 < r < \infty$, then

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\beta_0\in\mathcal{H}(K)} \mathbb{P}\left\{\mathcal{E}(\hat{\eta}_{n\lambda}) \geq An^{-\frac{2r}{2r+1}}\right\} = 0. \tag{9}$$

for any $\lambda \asymp \hat{\rho}$.

In light of Theorem 4, we can choose $\lambda = \hat{\lambda} := c\hat{\rho}$ for some constant $c > 0$ to ensure that the tuning parameter is of the appropriate order. In practice, the constant c can be further fine tuned to ensure enhanced performance. For example, our experience suggests that $\hat{\lambda} = \hat{\rho}/\hat{s}_1^2$ usually leads to satisfactory performance.

4 Numerical Experiments

We turn in this section to the numerical performance of the proposed adaptive predictor and demonstrate the practical implications of the theoretical results developed in the last two sections. We shall begin with simulation results, and then return to the analysis of the Canadian weather data mentioned in the introduction.

4.1 Simulation Study

To fix ideas, we shall focus on the case where $\mathcal{T} = [0, 1]$ and $\mathcal{H}(K)$ consists of functions in the linear span of the cosine basis,

$$f(t) = \sqrt{2}\sum_{k\geq 1} f_k \cos(k\pi t), \qquad t \in [0, 1],$$

such that

$$\sum_{k\geq 1} k^4 f_k^2 < \infty.$$

When endowed with norm

$$\|f\|^2_{\mathcal{H}(K)} = \int (f'')^2 = \int_0^1 \left(\sqrt{2} \sum_{k\geq 1} (k\pi)^2 f_k \cos(k\pi t) \right)^2 = \sum_{k\geq 1} (k\pi)^4 f_k^2,$$

$\mathcal{H}(K)$ forms a reproducing kernel Hilbert space with the reproducing kernel

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

where B_k is the kth Bernoulli polynomial. Here the following fact (Abramowitz and Stegun, 1965) is used:

$$B_{2m}(x) = (-1)^{m-1} 2(2m)! \sum_{k\geq 1} \frac{\cos(2\pi k x)}{(2\pi k)^{2m}}, \qquad \forall x \in [0,1].$$

In this setting, the roughness regularization estimate defined by (3) is given by

$$\left(\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda} \right) = \underset{a\in\mathbb{R}, b\in\mathcal{H}(K)}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left(Y_i - a - \int_{\mathcal{T}} X_i(t) b(t) dt \right)^2 + \lambda \int (b'')^2 \right\}.$$

For brevity, we shall also assume that there is no intercept in the functional linear model (1). The roughness regularization estimate then becomes

$$\hat{\beta}_{n\lambda} = \underset{b\in\mathcal{H}(K)}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left(Y_i - \int_{\mathcal{T}} X_i(t) b(t) dt \right)^2 + \lambda \int (b'')^2 \right\}.$$

To investigate the effect of varying covariance function, the true slope function is taken to be

$$\beta_0(t) = \sum_{k\geq 1} 4\sqrt{2} (-1)^{k-1} k^{-2} \cos(k\pi t).$$

We first consider the effect of varying smoothness of the covariance function. More specifically, set

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

The parameter r_2 controls how fast the eigenvalues of the covariance function C decay and therefore by Theorem 1 determines the optimal rates of convergence. We let the value of r_2 vary between 1 and 3. The functional predictors were simulated from a centered Gaussian process with covariance function C. To comprehend the trend as the sample size increases, the sample size n is chosen to be $n = 32, 64, 128, 256, 512$ and 1024. For each sample size and each value of r_2, we simulate data

from the functional linear model with $\sigma = 0.5$. Both the roughness regularization and functional principal component estimates were computed and the tuning parameters, λ for the former and the number of principal components for latter, were chosen to yield the smallest excess risk so that it reflects the optimal rate achieved by both methods. The experiment was repeated for two hundreds times and the results are summarized in Figure 2.

In this example, both the reproducing kernel and the covariance function share a common ordered set of eigenfunctions. In this case, as shown previously by Cai and Hall (2006) and Yuan and Cai (2010), both methods can attain the optimal rate of convergence. The similarity between the two methods as observed in Figure 2 simply confirms these earlier findings.

One of the key messages from Theorem 1 is that in addition to the decay of the eigenvalues of the reproducing kernel and covariance function, the alignment of the eigenfunctions of the two kernels also plays a crucial role in determining the optimal rate of convergence. To demonstrate the effect of such an alignment, we now consider a slightly different covariance function:

$$C(s,t) = \sum_{k \geq 1} 2\theta_k \cos(k\pi s) \cos(k\pi t)$$

where

$$\theta_k = (|k - k_0| + 1)^{-2}.$$

In this setting, the leading eigenfunctions of C are located around the k_0th eigenfunction of the reproducing kernel and in a certain sense controls the misalignment between the covariance function and the reproducing kernel. We consider $k_0 = 5, 10, 15$ and 20 to appreciate the effect of the alignment. The simulation was carried out in a similar fashion as before and the results are reported in the top panels of Figure 3.

As expected, for larger values of k_0, the functional principal component based approach performs rather poorly when the sample size is small. However, as one can observe from Figure 3, both method appears to converge at similar rates as the sample size increases although the roughness regularization method performs better. The top right panel displays the median relative efficiency of the roughness regularization over the functional principal component based method defined as $\mathcal{E}(\hat{\eta}^{\mathrm{FPCA}})/\mathcal{E}(\hat{\eta}^{\mathrm{REG}})$ where $\hat{\eta}^{\mathrm{REG}}$ and $\hat{\eta}^{\mathrm{FPCA}}$ represent the two estimates respectively. It is evident that the efficiency of the roughness regularization increases with k_0 which reflects how poorly the eigenfunctions of K and C align. In most cases, the roughness regularization estimate outperforms the functional principal component based method by an order of magnitude. It is noteworthy that the performance of FPCA based approach quickly deteriorates as k_0 increases. In particular, when k_0 is greater than the sample size, FPCA approach will fail since the true slope function is

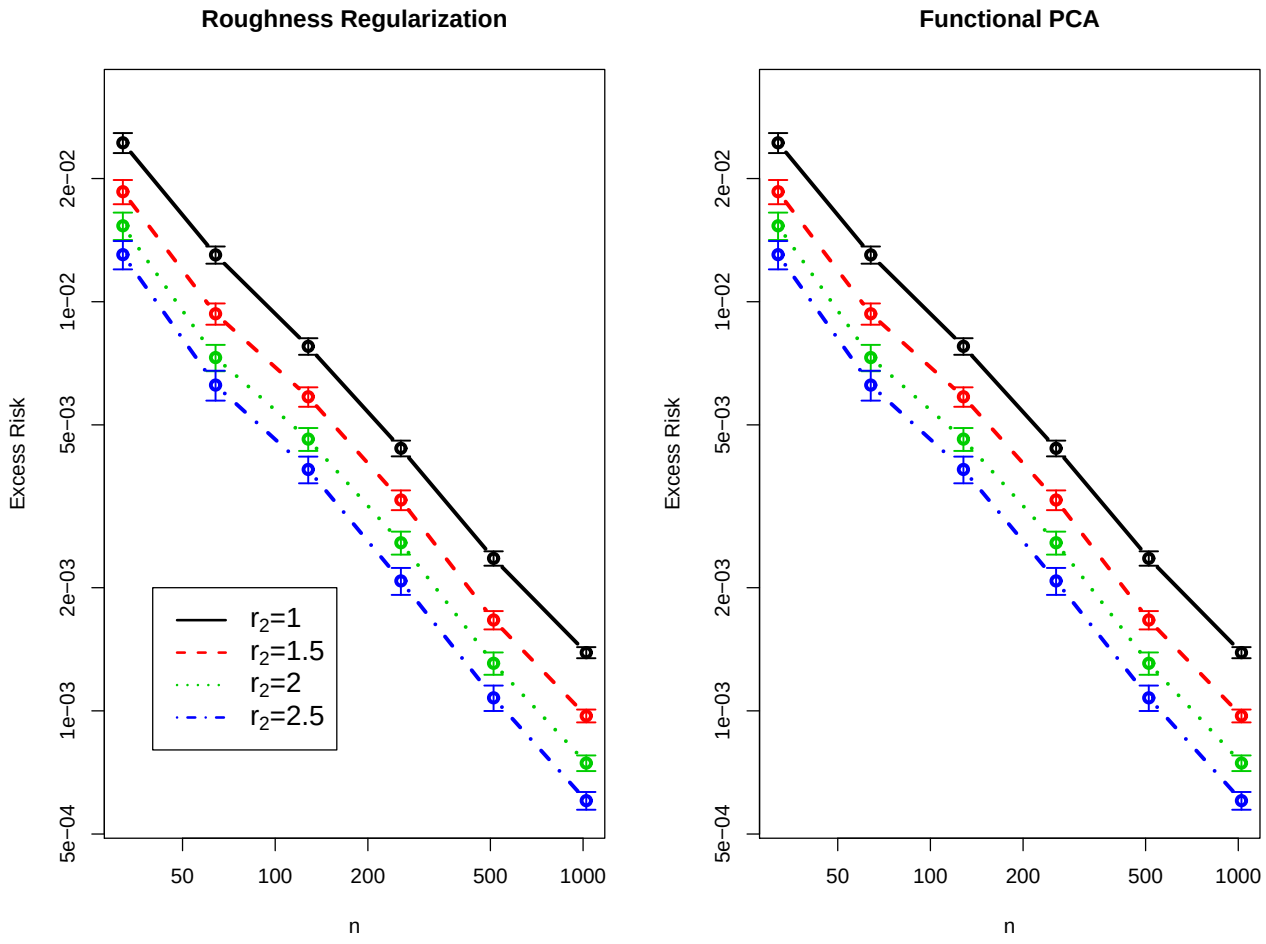

Figure 2: Effect of smoothness of the covariance function: 200 datasets were simulated for each combination of sample size $n = 32, 64, 128, 256, 512$ or 1024 and $r_2 = 1, 1.5, 2, 2.5$ or 3. Both roughness regularization estimate and the functional principal components estimate were computed with tuning parameter chosen to minimize the integrated squared error. The circle corresponds to the excess risk averaged over 200 datasets and error bars correspond to the mean $\pm$ one standard error. Both axes are in log scale.

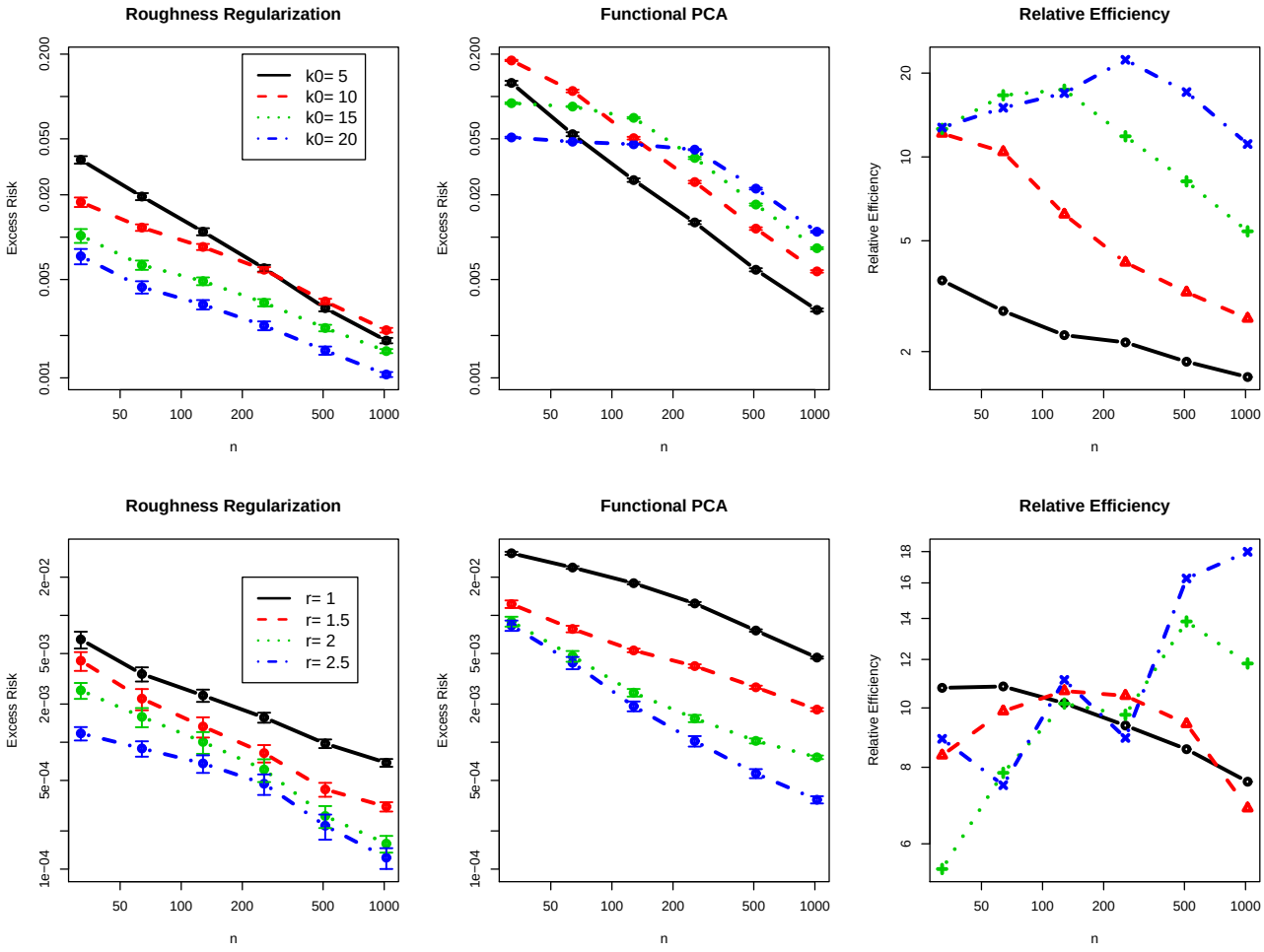

Figure 3: Effect of alignment of eigenfunctions between the reproducing kernel and the covariance function: For the top panels, 200 datasets were simulated for each combination of sample size $n = 32, 64, 128, 256, 512$ or 1024 and the location of the first principal component $k_0 = 5, 10, 15$ or 20. For the bottom panels, 200 datasets were simulated for each combination of sample size $n = 32, 64, 128, 256, 512$ or 1024 and the decay rate of the eigenvalues of C, $r_2 = 1, 1.5, 2, 2.5$. In the left and middle panels, the circle corresponds to the excess risk averaged over 200 datasets and error bars correspond to the mean $\pm$ one standard error. In the rightmost panels, median relative efficiency of the roughness regularization method over functional principal component based method is given for the two simulation settings. Both axes are in log scale.

orthogonal to the leading functional principal components. In this example, we have chosen the largest value of k_0 to be 20 to ensure that k_0 is always smaller than the sample size n, and to give a better contrast of the two methods.

To further demonstrate the effect of alignment between the eigenfunctions of K and C, we also considered a different set of covariance functions. In particular,

$$C(s,t) = \sum_{k\geq 1} 2k^{-2r_2} h_k(s) h_k(t)$$

where h_ks are the Haar functions, i.e.,

$$h_{2^k+l-1}(t) = \begin{cases} 2^{k/2} & t \in [\frac{l-1}{2^k}, \frac{l-1/2}{2^k}) \\ -2^{k/2} & t \in [\frac{l-1/2}{2^k}, \frac{l}{2^k}] \\ 0 & \text{otherwise} \end{cases}$$

for $k = 0, 1, 2, \ldots$ and $l = 1, \ldots$. Different from the previous two simulation settings, the eigenfunctions of C are the Haar basis of $\mathcal{L}_2$ that are different from those of K. We again apply both the roughness regularization and functional principal component based methods to each simulated dataset and summarize the findings in the lower panels of Figure 3. Similar to before, the results are averaged over two hundred simulated datasets for each value of r_2. We observe similar comparison as in the last example with the roughness regularization significantly outperform the functional principal component based method.

We now turn to the choice of the tuning parameter λ_n. The adaptive prediction procedure is applied to the simulated data. For brevity, we focus on the first two simulation settings. For each simulated dataset, the tuning parameter λ is chosen as $\lambda = \hat{\rho}/\hat{s}_1^2$. As mentioned earlier, this particular scaling usually yields fairly reasonable results, although it is plausible that it can be further improved by using other data-driven choice of scaling factors. Figure 4 summarizes the excess risk for all settings. The behavior of the resulting estimate closely resembles those reported earlier, indicating that the adaptive procedure proposed earlier is indeed capable of achieving the optimal rate of convergence.

4.2 Canadian Weather Data

Finally, we revisit the Canadian weather data example. The data set, one of the most popular examples in functional linear regression, contains daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994. The goal is to predict the log annual precipitation based on the average daily temperature. As shown in Figure 1, the application of functional PCA based methods could be problematic since the estimated slope function does not seem to allow for a

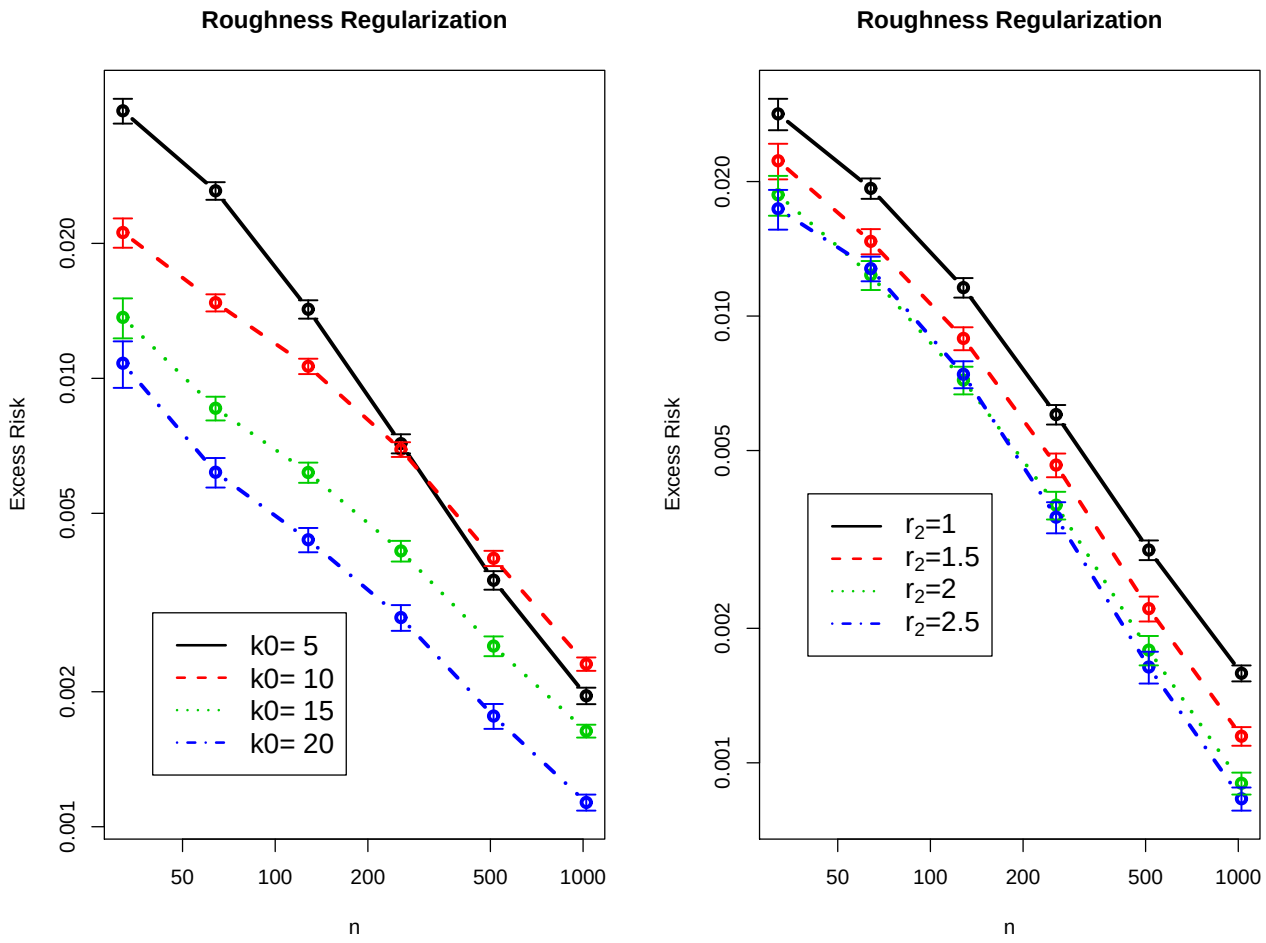

Figure 4: Excess risk of the adaptive prediction: each model used before was considered and the tuning parameter is chosen to be $\hat{\rho}/\hat{s}_1^2$. The left panel shows the result for different combinations of n and k_0 whereas the right panel corresponds to the result for different combinations of n and r_0. The errors bars correspond to the mean $\pm$ one standard error. Both axes are in log scale.

compact representation with respect to the eigenfunctions of the covariance function. Alternatively, we applied the regularization approach discussed earlier to the data. The data were collected on the basis of calendar year. The nature of the data suggests that the predictor function X be periodic on $[0,1]$ (year), and so is the slop function β_0. Let $\mathcal{W}_2^{\rm per}$ be the second order Sobolev space of periodic functions on $[0,1]$, we estimate α_0 and β_0 by

$$(\hat{\alpha},\hat{\beta}) = \underset{a\in\mathbb{R},b\in\mathcal{W}_2^{\rm per}}{\operatorname{argmin}} \left\{ \left[Y_i - a - \int_0^1 X_i(t)b(t)dt \right]^2 + \lambda \int_0^1 \left[b''(t)\right]^2 dt \right\} \tag{10}$$

The space $\mathcal{W}_2^{\rm per}$, endowed with the norm

$$\|b\|^2_{\mathcal{W}_2^{\rm per}} = \left[\int_0^1 b(t)dt\right]^2 + \int_0^1 \left[b''(t)\right]^2 dt,$$

has a reproducing kernel:

$$K(s,t) = 1 - \frac{1}{24}B_4(|s-t|), \qquad \forall s,t\in[0,1],$$

where B_4 is the fourth Bernoulli polynomial (see, e.g., Wahba, 1990). As shown in Yuan and Cai (2010), the solution of (10) can be expressed as

$$\hat{\alpha} = \bar{Y} - \int_0^1 \bar{X}(t)\hat{\beta}(t)dt$$

and

$$\hat{\beta}(\cdot) = c_0 + \sum_{i=1}^n c_i \int_0^1 (X_i - \bar{X})(s)K_1(s,\cdot)ds,$$

where

$$K_1(s,t) = -\frac{1}{24}B_4(|s-t|)$$

is the reproducing kernel of the orthogonal complement of constant functions in $\mathcal{W}_2^{\rm per}$. With this representation, the objective function of (10) becomes quadratic in terms of $c_0, c_1, \ldots, c_n$ and they can solved through

$$(c_0,(c_1,\ldots,c_n)^{\mathsf{T}}) = \underset{a_0\in\mathbb{R},\mathbf{a}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \Sigma\mathbf{a} - a_0\mathbf{u}\|^2 + \lambda\mathbf{a}^{\mathsf{T}}\Sigma\mathbf{a} \right\},$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)^{\mathsf{T}}$, Σ is a $n\times n$ matrix whose (i,j) entry is

$$\Sigma_{ij} = \int_0^1 (X_i - \bar{X})(s)K_1(s,t)(X_j - \bar{X})(t)dsdt$$

and $\mathbf{u}$ is a n dimensional vector with the ith entry given by

$$u_i = \int_0^1 (X_i - \bar{X})(s)ds.$$

The integrals in defining Σ_{ij} and u_i can be approximated by summations for practical purposes. We also used the proposed adaptive procedure to choose the tuning parameter. The estimated slope function is given in left panel of Figure 5. The right panel provides the normal Q-Q plot of the residuals, which suggests a fairly good fit of the data. The plot suggests that the precipitation at Kamloops station, corresponding to the rightmost point in the Q-Q plot, may be overestimated using the functional linear model. This can be attributed to the fact that Kamloops lies deep in the Thompson River Valley. These findings are similar to those reported in Ramsay and Silverman (2005) in which a restricted basis function based, instead of reproducing kernel based, roughness penalization approach was used.

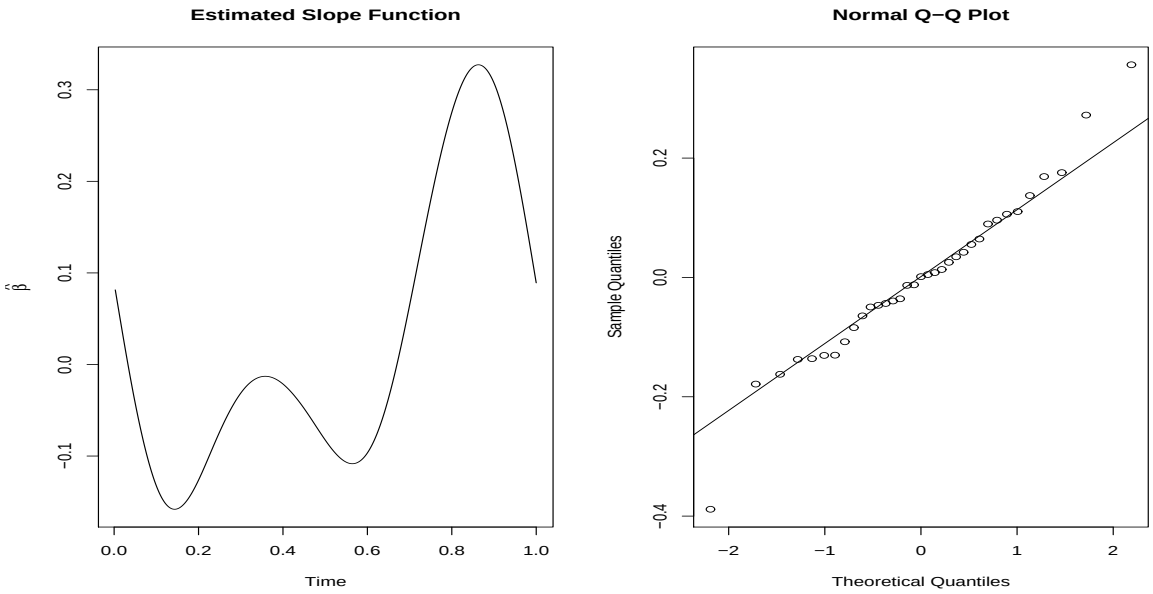

Figure 5: Canadian Weather Data: Smoothness regularization estimate of the slope function with adaptive choice of the tuning parameter. The left panel gives the estimated slope function whereas the right panel shows the Q-Q plot of the residuals.

Note that both the functional PCA based approach and the smoothness regularization approach assume that the slope function comes from a certain class of smooth functions where smoothness is characterized through the decay of the Fourier coefficients under a certain basis. More specifically, the functional PCA takes the Karhunen-Loève basis of X whereas the smoothness regularization approach takes the eigenfunctions of the reproducing kernel K as the basis functions. As seen before, the assumption made by the functional PCA approach might be questionable for the current data example. To further demonstrate the this and validate the appropriateness of the smoothness

regularization approach, we computed the Fourier coefficients of our estimated slope function with respect to both the eigenfunctions of K and C_n. The squared coefficients are given in Figure 6. It again confirms that the assumption made by the smoothness regularization approach regarding the slope function is more reasonable.

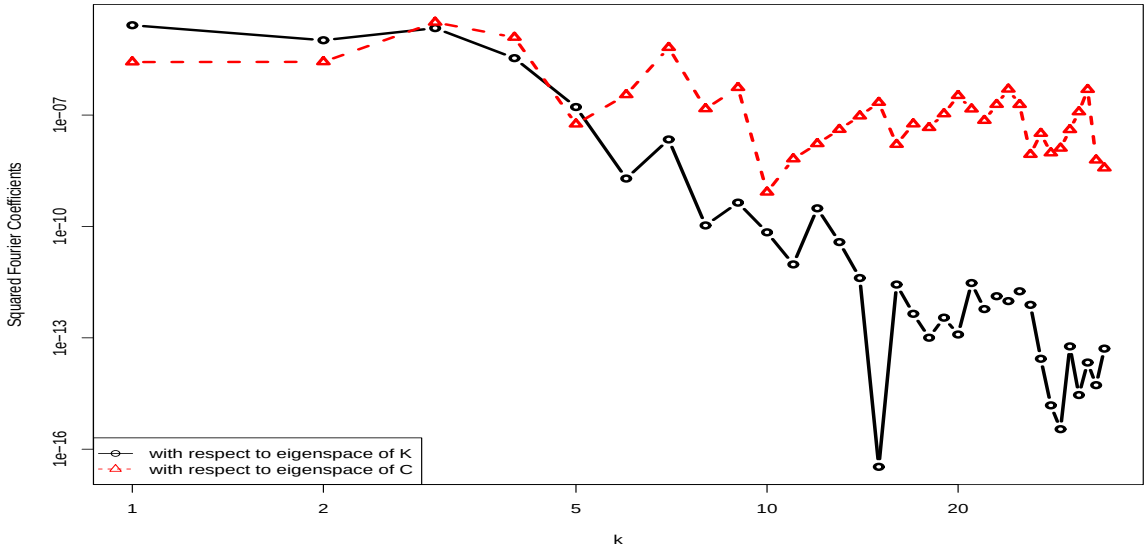

Figure 6: Canadian Weather Data: Squared Fourier coefficients of smoothness regularization estimate of the slope function with respect to the eigenfunctions of the reproducing kernel K or C_n. Both axes are in log scale.

5 Discussions

In the present paper we establish the minimax rate of convergence for prediction with functional predictors in a general setting where the reproducing kernel K and the covariance function C are not necessarily well aligned. The results show that the optimal prediction rate depends not only on the eigenvalues of K and C, but also on the alignment of the eigenfunctions of the two kernels. The prediction problem is more difficult if the eigenfunctions of the two kernels are not appropriately aligned. In contrast, the existing literature typically assumes the two kernels are perfectly or nearly perfectly aligned.

We also show that the method of regularization can achieve the optimal prediction if appropriately tuned. This is to be contrast with existing work in two ways. First, it is worth pointing out that the optimality established here for the method of regularization needs to be contrast with the

functional principal component analysis based approaches. Because the success of these methods hinges upon the assumption that the leading principal components, although entirely determined by the predictor itself, are the most predictive of the response, it is not immediately clear to us if they can also achieve the optimal rate in general when the eigenfunctions of the reproducing kernel and covariance function differ. In addition, earlier work on method of regularization for functional linear regression has assumed that the eigenfunctions of the the reproducing kernel and covariance function are either identical or close to being identical. We show here the optimality of this estimator extends much more generally.

To fix ideas, we have focused in this paper univariate functions, that is, the domain of the slope function and the functional predictors $\mathcal{T} \subset \mathbb{R}$; and used the usual Sobolev space on $\mathcal{T} = [0,1]$ as an working example. Our results, however, apply to the more general reproducing kernel Hilbert spaces when $\mathcal{T}$ is a compact set in an arbitrary Euclidean space. In particular, the optimal rate of convergence and adaptivity hold true for Sobolev spaces on $\mathcal{T} = [0,1]^2$ with the rate of decay r determined by the corresponding reproducing kernel and covariance operator. Such a setting can have useful applications in spatial statistics and image analysis.

6 Proofs

6.1 Proof of Theorem 1

Note that any lower bound for a specific case yields immediately a lower bound for the general case. It therefore suffices to consider the case when $\epsilon \sim N(0, \sigma^2)$. Denote by M the smallest integer greater than $c_0 n^{1/(2r+1)}$ for some constant $c_0 > 0$ to be specified later. For a $\theta = (\theta_{M+1}, \ldots, \theta_{2M}) \in \{0,1\}^M$, let

$$f_\theta = \sum_{k=M+1}^{2M} \theta_k M^{-1/2} L_{K^{1/2}\varphi_k}.$$

First observe that $f_\theta \in \mathcal{H}(K)$ since

$$\begin{aligned} \|f_\theta\|^2_{\mathcal{H}(K)} &= \left\| \sum_{k=M+1}^{2M} \theta_k M^{-1/2} L_{K^{1/2}\varphi_k} \right\|^2_{\mathcal{H}(K)} \\ &= \sum_{k=M+1}^{2M} \theta_k^2 M^{-1} \|L_{K^{1/2}\varphi_k}\|^2_{\mathcal{H}(K)} \\ &\leq \sum_{k=M+1}^{2M} M^{-1} \|L_{K^{1/2}\varphi_k}\|^2_{\mathcal{H}(K)} = 1, \end{aligned}$$

where we used the fact that

$$\begin{aligned}\langle L_{K^{1/2}}\varphi_j, L_{K^{1/2}}\varphi_k\rangle_{\mathcal{H}(K)} &= \langle \varphi_j, L_{K^{1/2}}L_{K^{1/2}}\varphi_k\rangle_{\mathcal{H}(K)} \\ &= \langle \varphi_j, L_K\varphi_k\rangle_{\mathcal{H}(K)} \\ &= \langle \varphi_j, \varphi_k\rangle_{\mathcal{L}_2} = \delta_{jk}.\end{aligned}$$

The Varshamov-Gilbert bound shows that for any $M \geq 8$, there exists a set $\Theta = \{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(N)}\} \subset \{0,1\}^M$ such that

(a) $\theta^{(0)} = (0, \ldots, 0)'$;

(b) $H(\theta, \theta') > M/8$ for any $\theta \neq \theta' \in \Theta$, where $H(\cdot,\cdot)$ is the Hamming distance;

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

We now invoke the results from Tsybakov (2009) to establish the lower bound which is based upon testing multiple hypotheses. To this end, denote by P_θ the joint distribution of $\{(X_i, Y_i) : i \geq 1\}$ with $\beta_0 = f_\theta$. First observe that for any $\theta, \theta' \in \Theta$,

$$\log\left(P_{\theta'}/P_\theta\right) = \frac{1}{\sigma^2}\sum_{i=1}^n \left(Y_i - \int X_i f_\theta\right)\int X_i\left(f_\theta - f_{\theta'}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n\left[\int X_i\left(f_\theta - f_{\theta'}\right)\right]^2.$$

Therefore the Kullback-Leibler distance between P_θ and $P_{\theta'}$ can be given by

$$\mathcal{K}(P_{\theta'}|P_\theta) = \frac{n}{2\sigma^2}\left\|L_{C^{1/2}}\left(f_\theta - f_{\theta'}\right)\right\|_{\mathcal{L}_2}^2 = \frac{n}{2\sigma^2}\left\|\sum_{k=M+1}^{2M}(\theta_k - \theta'_k)M^{-1/2}L_{C^{1/2}}L_{K^{1/2}}\varphi_k\right\|_{\mathcal{L}_2}^2.$$

Note that

$$\langle L_{C^{1/2}}L_{K^{1/2}}\varphi_k, L_{C^{1/2}}L_{K^{1/2}}\varphi_j\rangle_{\mathcal{L}_2} = \langle \varphi_k, L_{K^{1/2}}L_C L_{K^{1/2}}\varphi_j\rangle_{\mathcal{L}_2} = \langle \varphi_k, s_j\varphi_j\rangle_{\mathcal{L}_2} = s_j\delta_{kj}.$$

Hence,

$$\begin{aligned}\mathcal{K}(P_{\theta'}|P_\theta) &= \frac{n}{2\sigma^2}\sum_{k=M+1}^{2M} M^{-1}(\theta_k - \theta'_k)^2 s_k \\ &\leq \frac{ns_M}{2M\sigma^2}\sum_{k=M+1}^{2M}(\theta_k - \theta'_k)^2 \\ &= \frac{2ns_M}{M\sigma^2}H(\theta,\theta') \\ &\leq 2ns_M/\sigma^2 \\ &\leq 2c_2 nM^{-2r}/\sigma^2.\end{aligned}$$

This implies that for any $0 < \alpha < 1/8$,

$$\frac{1}{N}\sum_{j=1}^{N} \mathcal{K}(P_{\theta^{(j)}}|P_{\theta^{(0)}}) \le 2c_2 n M^{-2r}/\sigma^2 \le \alpha \log 2^{M/8} \le \alpha \log N \tag{11}$$

by taking $c_0 = c\alpha^{-1/(2r+1)}$ with a large enough numerical constant $c > 0$.

On the other hand, for any $\theta, \theta' \in \Theta$

$$\begin{aligned}
\|L_{C^{1/2}}(f_{\theta'} - f_\theta)\|_{\mathcal{L}_2}^2 &= \left\| \sum_{k=M+1}^{2M} (\theta_k - \theta_k') M^{-1/2} L_{C^{1/2}} L_{K^{1/2}} \varphi_k \right\|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} M^{-1}(\theta_k - \theta_k')^2 \|L_{C^{1/2}} L_{K^{1/2}} \varphi_k\|_{\mathcal{L}_2}^2 \\
&= \sum_{k=M+1}^{2M} M^{-1}(\theta_k - \theta_k')^2 s_k \\
&\ge M^{-1} s_{2M} \sum_{k=M+1}^{2M} (\theta_k - \theta_k')^2 \\
&= 4M^{-1} s_{2M} H(\theta, \theta') \\
&\ge s_{2M}/2 \\
&\ge c_1 2^{-(2r+1)} M^{-2r} \\
&\ge 2c\alpha^{2r/(2r+1)} n^{-\frac{2r}{2r+1}},
\end{aligned}$$

for some numerical constant $c > 0$. As shown by Tsybakov (2009), this, together with (11), implies that

$$\begin{aligned}
\inf_{\tilde{f}} \sup_{\theta \in \Theta} P_\theta \left\{ \left\| L_{C^{1/2}}(\tilde{f} - f_\theta) \right\|_{\mathcal{L}_2}^2 \ge c\alpha^{2r/(2r+1)} n^{-\frac{2r}{2r+1}} \right\} \\
\ge \frac{\sqrt{N}}{1+\sqrt{N}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log N}} \right).
\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \inf_{\tilde{f}} \sup_{\theta \in \Theta} P_\theta \left\{ \left\| L_{C^{1/2}}(\tilde{f} - f_\theta) \right\|_{\mathcal{L}_2}^2 \ge c\alpha^{2r/(2r+1)} n^{-\frac{2r}{2r+1}} \right\} \ge 1 - 2\alpha,$$

which yields that

$$\lim_{a\to 0} \lim_{n\to\infty} \inf_{\tilde{f}} \sup_{\theta \in \Theta} P_\theta \left\{ \left\| L_{C^{1/2}}(\tilde{f} - f_\theta) \right\|_{\mathcal{L}_2}^2 \ge a n^{-\frac{2r}{2r+1}} \right\} = 1.$$

Now the desired claim follows from the facts that $\{f_\theta : \theta \in \Theta\} \subset \mathcal{H}(K)$, and under P_θ,

$$\mathcal{E}(\tilde{\eta}) = \left\| L_{C^{1/2}}(\tilde{f} - f_\theta) \right\|_{\mathcal{L}_2}^2,$$

where

$$\tilde{\eta} = \int_{\mathcal{T}} X(t)\tilde{f}(t)dt.$$

6.2 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 shall assume that $\mathcal{H}(K)$ is dense in $\mathcal{L}_2$, which ensures that f_0 and $\hat{f}_\lambda$ are uniquely defined, in what follows. The proof in the general case proceeds in exactly the same fashion by restricting ourselves to $\mathcal{L}_2/\ker(L_{K^{1/2}})$.

For brevity, we shall write $T = L_{K^{1/2}CK^{1/2}}$ in what follows. We shall denote by T^ν a linear operator from $\mathcal{L}_2$ to $\mathcal{L}_2$ such that $T^\nu \varphi_k = s_k^\nu \varphi_k$. It is not hard to see that

$$\mathcal{E}(\hat{\eta}) = \|T^{1/2}(\hat{f}_\lambda - f_0)\|_{\mathcal{L}_2}^2.$$

It is also clear that

$$\hat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{L}_2} \left[\frac{1}{n} \sum_{i=1}^n (Y_i - \langle X_i, L_{K^{1/2}} f \rangle_{\mathcal{L}_2})^2 + \lambda \|f\|_{\mathcal{L}_2}^2 \right].$$

Recall that

$$Y_i = \langle X_i, L_{K^{1/2}} f_0 \rangle_{\mathcal{L}_2} + \epsilon_i.$$

Write

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

and $T_n = L_{K^{1/2}} L_{C_n} L_{K^{1/2}}$ where L_{C_n} is an integral operator such that for any $h \in \mathcal{L}_2$,

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

Therefore,

$$\hat{f}_\lambda = (T_n + \lambda \mathbf{1})^{-1} (T_n f_0 + g_n)$$

where $\mathbf{1}$ is the identity operator and

$$g_n = \frac{1}{n} \sum_{i=1}^n \epsilon_i L_{K^{1/2}} X_i.$$

Define

$$f_\lambda = (T + \lambda \mathbf{1})^{-1} T f_0.$$

By triangular inequality,

$$\left\| T^{1/2}(\hat{f}_\lambda - f_0) \right\|_{\mathcal{L}_2} = \left\| T^{1/2}(f_\lambda - f_0) \right\|_{\mathcal{L}_2} + \left\| T^{1/2}(\hat{f}_\lambda - f_\lambda) \right\|_{\mathcal{L}_2}. \tag{12}$$

The first term on the right hand side can be easily bounded. We appeal to the following lemma.

Lemma 1 *For any* $0 < \nu < 1$,

$$\|T^{\nu}(f_\lambda - f_0)\|_{\mathcal{L}_2} \leq (1-\nu)^{1-\nu}\nu^{\nu}\lambda^{\nu}\|f_0\|_{\mathcal{L}_2}. \tag{13}$$

Taking $\nu = 1/2$ in Lemma 1 gives

$$\left\|T^{1/2}(f_\lambda - f_0)\right\|_{\mathcal{L}_2}^2 \leq \frac{1}{4}\lambda\|f_0\|_{\mathcal{L}_2}^2.$$

We now turn to the second term on the right hand side of (12). Observe that

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

Recall that

$$(T_n+\lambda\mathbf{1})\hat{f}_\lambda = T_n f_0 - g_n.$$

Therefore,

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

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

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

Next we make use of another auxiliary lemma.

Lemma 2 *Assume that there exists a constant* $c_3 > 0$ *such that for any* $f \in \mathcal{L}_2$

$$\mathbb{E}\langle X, f\rangle_{\mathcal{L}_2}^4 \leq c_3\left(\mathbb{E}\langle X, f\rangle_{\mathcal{L}_2}^2\right)^2. \tag{14}$$

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

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

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

An application of Lemma 2 yields that

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

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

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

Therefore,

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

By Lemma 1,

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

Together with Lemma 3 stated below, we conclude that

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

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

Lemma 3 *For any* $0 \leq \nu \leq 1/2$,

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

We are now in position to bound $\|T^{1/2}(f_\lambda-\hat{f}_\lambda)\|$. Recall that

$$\begin{aligned}\|T^{1/2}(f_\lambda-\hat{f}_\lambda)\|_{\mathcal{L}_2} \leq\ &\|T^{1/2}(T+\lambda\mathbf{1})^{-1}T(f_\lambda-f_0)\|_{\mathcal{L}_2} + \|T^{1/2}(T+\lambda\mathbf{1})^{-1}(T_n-T)(f_\lambda-f_0)\|_{\mathcal{L}_2}\\ &+\lambda\|T^{1+1/2}f_0\| + \|T^{1/2}(T+\lambda\mathbf{1})^{-1}g_n\|_{\mathcal{L}_2}\\ &+\|T^{1/2}(T+\lambda\mathbf{1})^{-1}(T-T_n)(f_\lambda-\hat{f}_\lambda)\|_{\mathcal{L}_2}.\end{aligned}$$

We now bound the five terms on the right hand side separately. By Lemma 1,

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

We appeal to the following result.

Lemma 4 *Under the conditions of Lemma 2,*

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

By Lemmas 1 and 4,

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

Similarly,

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

By Lemma 3,

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

Together with the fact that $\lambda\|T^{1+1/2}f_0\| = O(\lambda)$, we conclude that

$$\|T^{1/2}(f_\lambda - \hat{f}_\lambda)\|_{\mathcal{L}_2} = O_p\left(n^{-\frac{2r}{2r+1}}\right). \blacksquare$$

6.3 Proof of Theorem 3

Consider the following Rademacher type of process:

$$R_n(b) = \frac{1}{n}\sum_{i=1}^{n} w_i\langle X_i, b\rangle_{\mathcal{L}_2},$$

where w_i's are iid Rademacher random variables, i.e., $\mathbb{P}(w_i = 1) = \mathbb{P}(w_i = -1) = 1/2$. Define

$$\|R_n\|_{\mathcal{B}(\delta)} = \sup_{b\in\mathcal{B}(\delta)} |R_n(b)|,$$

where

$$\mathcal{B}(\delta) = \left\{\beta \in \mathcal{H}(K) : \|b\|_{\mathcal{H}(K)} \leq 1 \text{ and } \|L_{C^{1/2}}b\|_{\mathcal{L}_2} \leq \delta\right\}.$$

Define

$$\rho_1 = \inf\left\{\rho \geq n^{-1}\log n : \mathbb{E}\|R_n\|_{\mathcal{B}(\delta)} \leq \rho^{1/2}\delta + \rho,\ \forall\delta \in [0,1]\right\}.$$

Similarly, write

$$\hat{\rho}_1 = \inf\left\{\rho \geq n^{-1}\log n : \mathbb{E}_w\|R_n\|_{\mathcal{B}(\delta)} \leq \rho^{1/2}\delta + \rho,\ \forall \delta \in [0,1]\right\},$$

where $\mathbb{E}_w$ stands for expectation taken over Rademacher random variables w_i's only. We first note the following result.

Lemma 5 *Under the condition of Theorem 3, there exist constants $c_1, c_2, c_3 > 0$ such that*

$$c_1\gamma_n(\delta) - c_2 n^{-1}(\log n) \leq \mathbb{E}\|R_n\|_{\mathcal{B}} \leq c_3\gamma_n(\delta).$$

It is clear from Lemma 5 that $0 < \inf \rho_1/\rho \leq \sup \rho_1/\rho < \infty$. Following the same argument, it can be shown that $0 < \inf \hat{\rho}/\hat{\rho}_1 \leq \sup \hat{\rho}/\hat{\rho}_1 < \infty$. It now suffices to show that $\hat{\rho}_1/\rho_1$ is also bounded away from 0 and $+\infty$.

For $b \in \mathcal{B}(\delta)$, let $f = L_{K^{-1/2}}b$. Then

$$\begin{aligned} |R_n(b)| &= \left|\frac{1}{n}\sum_{i=1}^n w_i\langle X_i, b\rangle_{\mathcal{L}_2}\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n w_i\langle L_{K^{1/2}}X_i, f\rangle_{\mathcal{L}_2}\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \left|\langle L_{K^{1/2}}X_i, f\rangle_{\mathcal{L}_2}\right|. \end{aligned}$$

By Cauchy-Schwartz inequality, this can be further bounded.

$$|R_n(b)| \leq \frac{1}{n}\sum_{i=1}^n \|L_{K^{1/2}}X_i\|_{\mathcal{L}_2}\|f\|_{\mathcal{L}_2} \leq \frac{1}{n}\sum_{i=1}^n \|L_{K^{1/2}}X_i\|_{\mathcal{L}_2},$$

where the second inequality follows from the fact that $\|f\|_{\mathcal{L}_2} = \|b\|_{\mathcal{H}(K)} \leq 1$. Note that the rightmost hand side converges almost surely to $\mathbb{E}\|L_{K^{1/2}}X\|_{\mathcal{L}_2} < \infty$ by strong law of large numbers. On the other hand,

$$\mathbb{E}\|R_n(b)\|^2 = \frac{1}{n}\mathbb{E}\left(w_i\langle X_i, b\rangle_{\mathcal{L}_2}\right)^2 = \|L_{C^{1/2}}b\|_{\mathcal{L}_2}^2 \leq \delta^2.$$

The rest of the proof follows in the same fashion as that of Koltchinskii and Yuan (2010; Section 3.2) and is therefore omitted. ∎

References

[1] Abramowitz, M. and Stegun, I. (1965), *Handbook of Mathematical Functions*, U.S. Government Printing Office, Washington, D.C.

[2] Cai, T. and Hall, P. (2006), Prediction in functional linear regression, *Annals of Statistics*, **34**, 2159-2179.

[3] Donoho, D.L. (1995), Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition, *Appl. Comput. Harmon. Anal.* **2**, 101-126.

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

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

[6] Jolliffe, I.T. (1982), A note on the use of principal components in regression, *Applied Statistics*, **31**, 300-303.

[7] Koltchinskii, V. and Yuan, M. (2010), Sparsity in multiple kernel learning, *Annals of Statistics*, **36**, 3660-3695.

[8] Mendelson, S. (2002), Geometric parameters of kernel machines. In *COLT 2002. Lecture Notes in Artificial Intelligence* **2375** 29-43, Springer, Berlin.

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

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

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

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

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

[14] Yao, F., Müller, H. and Wang, J. (2005), Functional linear regression analysis for longitudinal data, *Annals of Statistics*, **33**, 2873-2903.

[15] Yuan, M. and Cai, T. T. (2010), A reproducing kernel Hilbert space approach to functional linear regression, *Annals of Statistics*, **38**, 3412-3444.

Appendix – Auxiliary Results

Proof of Lemma 1

Write

$$f_0 = \sum_{k\geq 1} a_k \varphi_k.$$

Then

$$f_\lambda = \sum_{k\geq 1} \frac{s_k a_k}{\lambda + s_k} \varphi_k.$$

Therefore,

$$\|T^\nu(f_\lambda - f_0)\|_{\mathcal{L}_2}^2 = \sum_{k\geq 1} s_k^{2\nu} \left(\frac{\lambda a_k}{\lambda + s_k}\right)^2 \leq \max_{k\geq 1} \frac{\lambda^2 s_k^{2\nu}}{(\lambda + s_k)^2} \sum_{k\geq 1} a_k^2.$$

By Young's inequality, $\lambda + s_k \geq (1-\nu)^{-(1-\nu)} \nu^{-\nu} \lambda^{1-\nu} s_k^\nu$. Hence,

$$\|T^\nu(f_\lambda - f_0)\|_{\mathcal{L}_2}^2 \leq (1-\nu)^{2(1-\nu)} \nu^{2\nu} \lambda^{2\nu} \|f_0\|_{\mathcal{L}_2}^2. \blacksquare$$

Proof of Lemma 2

Recall that

$$\left\|T^\nu (T+\lambda \mathbf{1})^{-1}(T_n - T)T^{-\nu}\right\|_{\text{op}} = \sup_{h:\|h\|_{\mathcal{L}_2}=1} \left|\langle h, T^\nu (T+\lambda \mathbf{1})^{-1}(T_n - T)T^{-\nu} h\rangle_{\mathcal{L}_2}\right|$$

Write

$$h = \sum_{k\geq 1} h_k \varphi_k.$$

Then

$$\begin{aligned}
\langle h, T^\nu (T+\lambda \mathbf{1})^{-1}(T_n - T)T^{-\nu} h\rangle_{\mathcal{L}_2} &= \langle T^\nu (T+\lambda \mathbf{1})^{-1} h, (T_n - T)T^{-\nu} h\rangle_{\mathcal{L}_2} \\
&= \left\langle \sum_{j\geq 1} \frac{s_j^\nu h_j}{s_j + \lambda} \varphi_j, \sum_{k\geq 1} s_k^{-\nu} h_k (T_n - T)\varphi_k \right\rangle_{\mathcal{L}_2} \\
&= \sum_{j,k\geq 1} \frac{s_j^\nu s_k^{-\nu} h_j h_k}{s_j + \lambda} \langle \varphi_j, (T_n - T)\varphi_k\rangle_{\mathcal{L}_2}.
\end{aligned}$$

An application of Cauchy-Schwartz inequality yields that

$$\begin{aligned}
&\left|\sum_{j,k\geq 1} \frac{s_j^\nu s_k^{-\nu} h_j h_k}{s_j + \lambda} \langle \varphi_j, (T_n - T)\varphi_k\rangle_{\mathcal{L}_2}\right| \\
\leq\ &\left(\sum_{j,k\geq 1} h_j^2 h_k^2\right)^{1/2} \left(\sum_{j,k\geq 1} \frac{s_j^{2\nu} s_k^{-2\nu}}{(s_j + \lambda)^2} \langle \varphi_j, (T_n - T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2}
\end{aligned}$$

Hence,

$$\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}(T_n-T)T^{-\nu}\right\|_{\text{op}} \leq \left(\sum_{j,k\geq 1}\frac{s_j^{2\nu}s_k^{-2\nu}}{(s_j+\lambda)^2}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2}. \tag{15}$$

Now consider the expectation of the right hand side. By Jensen's inequality,

$$\mathbb{E}\left(\sum_{j,k\geq 1}\frac{s_j^{2\nu}s_k^{-2\nu}}{(s_j+\lambda)^2}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2} \leq \left(\sum_{j,k\geq 1}\frac{s_j^{2\nu}s_k^{-2\nu}}{(s_j+\lambda)^2}\mathbb{E}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2}.$$

Note that

$$\begin{aligned}
\mathbb{E}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2 &= \mathbb{E}\langle L_{K^{1/2}}\varphi_j,(L_{C_n}-L_C)L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}^2 \\
&\quad \mathbb{E}\langle L_{K^{1/2}}\varphi_j,(L_{C_n}-L_C)L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}^2 \\
&= \mathbb{E}\left(\int_{\mathcal{T}^2}(L_{K^{1/2}}\varphi_j)(s)(C_n(s,t)-C(s,t))(L_{K^{1/2}}\varphi_k)(t)\right)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{T}^2}(L_{K^{1/2}}\varphi_j)(s)(X_i(s)X_i(t)-\mathbb{E}X_i(s)X_i(t))(L_{K^{1/2}}\varphi_k)(t)\right)^2 \\
&= \frac{1}{n}\mathbb{E}\left(\int_{\mathcal{T}^2}(L_{K^{1/2}}\varphi_j)(s)(X(s)X(t)-\mathbb{E}X(s)X(t))(L_{K^{1/2}}\varphi_k)(t)\right)^2 \\
&\leq \frac{1}{n}\mathbb{E}\left(\int_{\mathcal{T}^2}(L_{K^{1/2}}\varphi_j)(s)X(s)X(t)(L_{K^{1/2}}\varphi_k)(t)\right)^2.
\end{aligned}$$

An application of Cauchy-Schwartz inequality yields

$$\begin{aligned}
&\mathbb{E}\langle L_{K^{1/2}}\varphi_j,(L_{C_n}-L_C)L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}^2 \\
\leq\ & \frac{1}{n}\mathbb{E}^{1/2}\left(\int_{\mathcal{T}}(L_{K^{1/2}}\varphi_j)(t)X(t)dt\right)^4\mathbb{E}^{1/2}\left(\int_{\mathcal{T}}(L_{K^{1/2}}\varphi_k)(t)X(t)dt\right)^4 \\
\leq\ & c_3n^{-1}\mathbb{E}\left(\int_{\mathcal{T}}(L_{K^{1/2}}\varphi_j)(t)X(t)dt\right)^2\mathbb{E}\left(\int_{\mathcal{T}}(L_{K^{1/2}}\varphi_k)(t)X(t)dt\right)^2 \\
=\ & c_3n^{-1}\|T^{1/2}\varphi_j\|_{\mathcal{L}_2}^2\|T^{1/2}\varphi_k\|_{\mathcal{L}_2}^2 = c_3n^{-1}s_js_k.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left(\sum_{j,k\geq 1}\frac{s_j^{2\nu}s_k^{-2\nu}}{(s_j+\lambda)^2}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{j,k\geq 1}\frac{s_j^{1+2\nu}s_k^{1-2\nu}}{(s_j+\lambda)^2}\right)^{1/2} \tag{16}$$

Note that $s_k^{1-2\nu}$ is summable because $(1-2\nu)(2r)>1$. We now appeal to the following lemma.

Lemma 6 *If there exist constants $0 < c_1 < c_2 < \infty$ such that $c_1 k^{-2r} < s_k < c_2 k^{-2r}$, then there exist constants $c_3, c_4 > 0$ depending only on c_1, c_2 such that*

$$c_4 \lambda^{-1/(2r)} \leq \sum_{j \geq 1} \frac{s_j^{1+2\nu}}{(\lambda + s_j)^{1+2\nu}} \leq c_3 (1 + \lambda^{-1/(2r)}).$$

Thus, by Lemma 6,

$$\begin{aligned} \left\| T^\nu (T + \lambda \mathbf{1})^{-1} (T_n - T) T^{-\nu} \right\|_{\text{op}} &\leq c \left(n \lambda^{1-2\nu} \right)^{-1/2} \left(\sum_{j \geq 1} \frac{s_j^{1+2\nu}}{(\lambda + s_j)^{1+2\nu}} \right)^{1/2} \\ &\leq c \left(n \lambda^{1-2\nu+1/(2r)} \right)^{-1/2}. \end{aligned}$$

The proof is now completed by Markov inequality. ∎

Proof of Lemma 3

Recall that

$$g_n = \frac{1}{n} \sum_{i=1}^n \epsilon_i L_{K^{1/2}} X_i.$$

Therefore,

$$\begin{aligned} \left\| T^\nu (T + \lambda \mathbf{1})^{-1} g_n \right\|_{\mathcal{L}_2}^2 &= \sum_{k \geq 1} \langle T^\nu (T + \lambda \mathbf{1})^{-1} g_n, \varphi_k \rangle_{\mathcal{L}_2}^2 \\ &= \sum_{k \geq 1} \langle g_n, (T + \lambda \mathbf{1})^{-1} T^\nu \varphi_k \rangle_{\mathcal{L}_2}^2 \\ &= \sum_{k \geq 1} \left\langle \frac{1}{n} \sum_{i=1}^n \epsilon_i L_{K^{1/2}} X_i, \frac{s_k^\nu}{\lambda + s_k} \varphi_k \right\rangle_{\mathcal{L}_2}^2 \\ &= \sum_{k \geq 1} \frac{s_k^{2\nu}}{(\lambda + s_k)^2} \left(\frac{1}{n} \sum_{i=1}^n \epsilon_i \langle X_i, L_{K^{1/2}} \varphi_k \rangle_{\mathcal{L}_2} \right)^2. \end{aligned}$$

Note that

$$\mathbb{E} \left(\epsilon_i \langle X_i, L_{K^{1/2}} \varphi_k \rangle_{\mathcal{L}_2} \right) = \mathbb{E} \left(\epsilon_i \langle X_i, L_{K^{1/2}} \varphi_k \rangle_{\mathcal{L}_2} | X_i \right) = 0.$$

Hence,

$$\begin{aligned}
\mathbb{E}\left\|T^{\nu}(T+\lambda\mathbf{1})^{-1}g_n\right\|_{\mathcal{L}_2}^2 &= \sum_{k\geq 1}\frac{s_k^{2\nu}}{(\lambda+s_k)^2}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle X_i, L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}\right)^2 \\
&= \frac{1}{n}\sum_{k\geq 1}\frac{s_k^{2\nu}}{(\lambda+s_k)^2}\mathbb{E}\left(\epsilon_i\langle X_i, L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}\right)^2 \\
&= \frac{\sigma^2}{n}\sum_{k\geq 1}\frac{s_k^{2\nu}}{(\lambda+s_k)^2}\mathbb{E}\langle X_i, L_{K^{1/2}}\varphi_k\rangle_{\mathcal{L}_2}^2 \\
&= \frac{\sigma^2}{n}\sum_{k\geq 1}\frac{s_k^{1+2\nu}}{(\lambda+s_k)^2} \\
&\leq \frac{\sigma^2}{n\lambda^{1-2\nu}}\sum_{k\geq 1}\frac{s_k^{1+2\nu}}{(\lambda+s_k)^{1+2\nu}} \\
&\leq \frac{c\sigma^2}{n\lambda^{1-2\nu+1/(2r)}}.
\end{aligned}$$

The proof can now be completed by Markov inequality. ∎

Proof of Lemma 4

Similar to (15), by Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left\|T^{1/2}(T+\lambda\mathbf{1})^{-1}(T_n-T)T^{-\nu}\right\|_{\text{op}} \\
= &\sup_{h:\|h\|_{\mathcal{L}_2}=1}\left|\sum_{j,k\geq 1}\frac{s_j s_k^{-\nu}h_j h_k}{s_j+\lambda}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}\right| \\
\leq &\left(\sum_{j,k\geq 1}\frac{s_j s_k^{-2\nu}}{(s_j+\lambda)^2}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2}
\end{aligned}$$

Following a similar argument as that of (16),

$$\mathbb{E}\left(\sum_{j,k\geq 1}\frac{s_j s_k^{-2\nu}}{(s_j+\lambda)^2}\langle\varphi_j,(T_n-T)\varphi_k\rangle_{\mathcal{L}_2}^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{j,k\geq 1}\frac{s_j^2 s_k^{1-2\nu}}{(s_j+\lambda)^2}\right)^{1/2} \leq c(n\lambda^{1/(2r)})^{-1/2}.$$

The proof is now completed by Markov inequality. ∎

Proof of Lemma 5

It is clear that $\mathcal{B}(\delta) = L_{K^{1/2}}(\mathcal{F}(\delta))$ where

$$\mathcal{F}(\delta) = \left\{f\in\mathcal{L}_2 : \|f\|_{\mathcal{L}_2}\leq 1 \text{ and } \|T^{1/2}f\|_{\mathcal{L}_2}^2\leq\delta^2\right\}.$$

Denote

$$\mathcal{G} = \left\{ \sum_{k\geq 1} f_k \varphi_k : \sum_{k\geq 1} \left(\frac{f_k}{\min\{1, \delta/\sqrt{s_k}\}} \right)^2 \leq 1 \right\}$$

It can be easily checked that $\mathcal{G} \subset \mathcal{F} \subset \sqrt{2}\mathcal{G}$. Therefore,

$$\sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)| \leq \|R_n\|_{\mathcal{B}} \leq \sqrt{2} \sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)| .$$

By Jensen's inequality,

$$\mathbb{E} \sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)| \leq \left(\mathbb{E} \sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)|^2 \right)^{1/2} .$$

By Cauchy-Schwartz inequality,

$$\begin{aligned} |R_n (L_{K^{1/2}} f)|^2 &= \left| \sum_{k\geq 1} f_k R_n (L_{K^{1/2}} \varphi_k) \right|^2 \\ &\leq \left(\sum_{k\geq 1} \frac{f_k^2}{\min\{1, \delta^2/s_k\}} \right) \left(\sum_{k\geq 1} \min\{1, \delta^2/s_k\} R_n^2 (L_{K^{1/2}} \varphi_k) \right) \end{aligned}$$

Therefore,

$$\sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)|^2 \leq \left(\sum_{k\geq 1} \min\{1, \delta^2/s_k\} R_n^2 (L_{K^{1/2}} \varphi_k) \right) .$$

Observe that

$$\begin{aligned} \mathbb{E} R_n^2 (L_{K^{1/2}} \varphi_k) &= \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n w_i \langle X_i, L_{K^{1/2}} \varphi_k \rangle_{\mathcal{L}_2} \right)^2 \\ &= \frac{1}{n} \mathbb{E} \langle X_i, L_{K^{1/2}} \varphi_k \rangle_{\mathcal{L}_2}^2 = n^{-1} s_k. \end{aligned}$$

Thus,

$$\mathbb{E} \left(\sup_{f\in\mathcal{G}} |R_n (L_{K^{1/2}} f)|^2 \right) \leq \left(\sum_{k\geq 1} \min\{1, \delta^2/s_k\} \mathbb{E} R_n^2 (L_{K^{1/2}} \varphi_k) \right) = \gamma_n^2(\delta),$$

which implies that

$$\mathbb{E} \|R_n\|_{\mathcal{B}} \leq \sqrt{2} \gamma_n(\delta).$$

To prove the lower bound, we appeal to Hoffman-Jørgensen inequality which suggests that

$$\mathbb{E}^{1/2} \|R_n\|_{\mathcal{B}}^2 \leq c \left(\mathbb{E} \|R_n\|_{\mathcal{B}} + \frac{1}{n} \mathbb{E}^{1/2} \max_{1\leq i\leq n} \sup_{\beta\in\mathcal{F}} \langle X_i, \beta \rangle_{\mathcal{L}_2}^2 \right) .$$

Observe that

$$\begin{aligned}\mathbb{E}^{1/2}\max_{1\le i\le n}\sup_{\beta\in\mathcal{F}}\langle X_i,\beta\rangle^2_{\mathcal{L}_2} &= \mathbb{E}^{1/2}\max_{1\le i\le n}\sup_{f\in\mathcal{G}}\langle L_{K^{1/2}}X_i,f\rangle^2_{\mathcal{L}_2}\\ &\le \mathbb{E}^{1/2}\max_{1\le i\le n}\|L_{K^{1/2}}X_i\|^2_{\mathcal{L}_2}.\end{aligned}$$

Because $\|L_{K^{1/2}}X\|_{\mathcal{L}_2}$ has exponential tails, it can be further bounded by $c\log n$ for some constant $c>0$.

Hence,

$$\begin{aligned}\mathbb{E}\|R_n\|_{\mathcal{B}} &\ge c_1\mathbb{E}^{1/2}\|R_n\|^2_{\mathcal{B}} - c_2n^{-1}(\log n)\\ &\ge c_1\mathbb{E}^{1/2}\left(\sup_{f\in\mathcal{G}}|R_n\left(L_{K^{1/2}}f\right)|^2\right) - c_2n^{-1}(\log n)\\ &\ge c_1\gamma_n(\delta) - c_2n^{-1}(\log n).\ \blacksquare\end{aligned}$$

Proof of Lemma 6

Note that

$$\begin{aligned}\sum_{k\ge1}\frac{s_k^{1+2\nu}}{(\lambda+s_k)^{1+2\nu}} &\le \sum_{k\ge1}\frac{(c_1k^{-2r})^{1+2\nu}}{(\lambda+c_2k^{-2r})^{1+2\nu}}\\ &= c_1^{-2r(1+2\nu)}\sum_{k\ge1}\frac{1}{(c_2+\lambda k^{2r})^{1+2\nu}}\\ &\le c_1^{-2r(1+2\nu)}\left(\frac{1}{c_2}+\int_1^\infty\frac{dx}{(c_2+\lambda x^{2r})^{1+2\nu}}\right)\\ &= c_1^{-2r(1+2\nu)}\left(\frac{1}{c_2}+\lambda^{-\frac{1}{2r}}\int_{\lambda^{1/(2r)}}^\infty\frac{dy}{(c_2+y^{2r})^{1+2\nu}}\right)\\ &\le c_3(1+\lambda^{-1/(2r)}).\end{aligned}$$

Similarly,

$$\begin{aligned}\sum_{k\ge1}\frac{s_k^{1+2\nu}}{(\lambda+s_k)^{1+2\nu}} &\ge c_2^{-2r(1+2\nu)}\sum_{k\ge1}\frac{1}{(c_1+\lambda k^{2r})^{1+2\nu}}\\ &\ge c_2^{-2r(1+2\nu)}\int_1^\infty\frac{dx}{(c_1+\lambda x^{2r})^{1+2\nu}}\\ &= c_2^{-2r(1+2\nu)}\lambda^{-\frac{1}{2r}}\int_{\lambda^{1/(2r)}}^\infty\frac{dy}{(c_1+y^{2r})^{1+2\nu}}\\ &\ge c_4\lambda^{-1/(2r)}.\ \blacksquare\end{aligned}$$