Optimal prediction for additive function-on-function regression

Matthew Reimherr* **and Bharath Sriperumbudur**[†]

Department of Statistics
Pennsylvania State University
University Park, PA, USA
e-mail: mreimherr@psu.edu*;* bks18@psu.edu

Bahaeddine Taoufik

Department of Mathematics
University of Lynchburg
Lynchburg, VA, USA
e-mail: Taoufik_B@lynchburg.edu

Abstract: As with classic statistics, functional regression models are invaluable in the analysis of functional data. While there are now extensive tools with accompanying theory available for linear models, there is still a great deal of work to be done concerning nonlinear models for functional data. In this work we consider the Additive Function-on-Function Regression model, a type of nonlinear model that uses an additive relationship between the functional outcome and functional covariate. We present an estimation methodology built upon Reproducing Kernel Hilbert Spaces, and establish optimal rates of convergence for our estimates in terms of prediction error. We also discuss computational challenges that arise with such complex models, developing a representer theorem for our estimate as well as a more practical and computationally efficient approximation. Simulations and an application to Cumulative Intraday Returns around the 2008 financial crisis are also provided.

MSC 2010 subject classifications: Primary 62G08, 62G20.
Keywords and phrases: Nonlinear regression, functional data analysis, reproducing kernel Hilbert space, minimax convergence.

Contents

1. Introduction

Functional data analysis (FDA) concerns the statistical analysis of data where one of the variables of interest is a function. FDA has seen rapidly increasing interest over the last few decades and has successfully been applied to a variety of fields, including economics, finance, the geosciences, and the health sciences. One of the most fundamental tools in statistics is linear regression, as such, it has been a major area of research in FDA. While the literature is too vast to cover here, we refer readers to Ramsay and Silverman [27], Ramsay et al. [28], Horváth and Kokoszka [8], Kokoszka and Reimherr [16], which provide introductions to FDA, as well as Morris [21], which provides a broad overview of methods for functional linear regression.

A major challenge of functional regression is handling functional predictors. At least conceptually, a functional predictor means having a large number (theoretically infinite) of predictors that are all highly correlated. To handle such a setting, certain regularity conditions are imposed to make the problem tractable. Most of these conditions are directly or indirectly related to the smoothness of the parameter being estimated. However, the convergence rates of the resulting estimators then depend heavily on these assumptions, and the rates are not parametric when the predictor is infinite dimensional.

One of the most well studied models in FDA is the functional linear model. Commonly, one distinguishes between function-on-scalar, scalar-on-function, and function-on-function regression when discussing such models, with first term denoting the type of response and the second term denoting the type of covariate. The convergence rates for function-on-scalar regression are usually much faster than for the scalar-on-function or function-on-function. Methodological, theoretical, and computational issues related to functional linear models are now well understood. More recently, there has been a growing interest in developing nonlinear regression models. While it is natural to begin examining nonlinear models after establishing the framework for linear ones, there is also a practical need for such models. Functional data may contain complicated temporal dynamics, which may exhibit nonlinear patterns that are not well modeled assuming linearity; Fan et al. [5] examine this issue deeply.

Nonlinear regression methods for FDA have received a fair amount of attention for the scalar-on-function setting, while function-on-function regression models, where the relationship between the response and covariates is believed to be nonlinear, have received considerably less attention. Concerning nonlinear

scalar-on-function regression, James and Silverman [10] introduced a functional single index model, where the outcome is related to a linear functional of the predictor through a nonlinear transformation. This work would later be extended in Fan et al. [5], allowing for a potentially high-dimensional number of a functional predictors. Preda [26] explored fitting a fully nonlinear model using reproducing kernel Hilbert spaces (RKHS). In contrast, Müller et al. [22] simplified the form of the nonlinear relationship by introducing the functional additive model, which combines ideas from functional linear models and scalar additive models [7]. Optimal convergence rates for the functional additive model were then established by Wang and Ruppert [32], which generalized the work of Cai and Yuan [3] in the linear case. An alternative to the functional additive model was given in Zhu et al. [36] who first expressed the functional predictor using functional principal components analysis, FPCA, and then built an additive model between the outcome and scores. An extension to generalized linear models can be found in [20, 4].

Moving to function-on-function regression, Lian [18] extended the work of Preda [26] to functional outcomes, which was then also considered in Kadri et al. [11]. Most relevant to the present paper is the work of Scheipl et al. [29] who extended the work of Müller et al. [22] by introducing an additive model for function-on-function regression. They used a general trivariate tensor product basis approach for estimation, which allowed them to rely on *GAM* from the *MGCV* package in R to carry out the computation, as is implemented in the *Refund* package. Ma and Zhu [19], examining the same model, considered a binning estimation technique combined with FPCA. In addition, they were able to prove convergence of their estimators, but made no mention of optimality while also needing a great deal of assumptions which are challenging to interpret. Another estimation technique was examined in Kim et al. [13], which was similar to the trivariate tensor product approach of Scheipl et al. [29], but two of the bases are explicitly assumed to be orthogonal B-splines, while the third comes from an FPCA expansion. However, as with Scheipl et al. [29], no theoretical justification is provided. Lastly, in very recent work, [30] considered the case of using an RKHS framework to estimate a function-on-function linear model. Extending the work the [3], they were able to establish the optimality of their procedure. Our work can be viewed as extending this work to nonlinear relationships via a function-on-function additive model.

The goal of this work is to develop a penalized regression framework based on Reproducing Kernel Hilbert Spaces, RKHS, for fitting the additive function-on-function regression model, AFFR [29]. A major contribution of this work is to provide optimal convergence rates of our estimators in terms of prediction error, and that this rate is the same as for the scalar outcome setting [32]. We also discuss computational aspects of our approach, as the RKHS structure allows for a fairly efficient computation as compared to the trivariate tensor product bases that have been used previously. Background and the model are introduced in Section 2. Computation is discussed in Section 3, while theory is presented in Section 4. We conclude with a numeric study consisting of simulations and an application to financial data.

2. Model and background

We assume that we observe i.i.d pairs $\{(X_i(t), Y_i(t)) : i = 1, \ldots, n, \ t \in [0, 1]\}$. The functions could be observed on other intervals, but as long as they are closed and bounded, then they can always be rescaled to be $[0, 1]$, thus it is common in FDA to work on the unit interval. Both the outcome, $Y_i(t)$, and $X_i(t)$ are assumed to be completely observed functions, a practice sometimes referred to as *dense functional data analysis* [16]; practically this means that the curve reconstruction contributes a comparatively small amount of uncertainty to the final parameter estimates. More rigorous definitions can be found in Cai and Yuan [2], Li et al. [17], Zhang et al. [35]. For sparsely observed curves, it is usually better to use more tailored approaches such as PACE [34], FACE [33], or MISFIT [25].

The additive function-on-function regression model is defined as

$$Y_i(t) = \int_0^1 g(t, s, X_i(s))\, ds + \varepsilon_i(t).$$

We assume that the functions X_i, ε_i, and Y_i are elements of $L^2[0, 1]$, which is a real separable Hilbert space. The trivariate function, $g(t, s, x)$ is assumed to be an element of an RKHS, $\mathbb{K}$.

Recall that an RKHS is a Hilbert space that possesses the *reproducing property*, namely, we assume that $\mathbb{K}$ is a Hilbert space of functions from $[0, 1] \times [0, 1] \times \mathbb{R} \to \mathbb{R}$, and that there exists a kernel function $k(t, s, x, t', s', x') = k_{t,s,x}(t', s', x')$ that satisfies

$$f(t, s, x) = \langle k_{t,s,x}, f \rangle_{\mathbb{K}},$$

for any $f \in \mathbb{K}$. There is a one-to-one correspondence between $\mathbb{K}$ and k, thus choosing the kernel function completely determines the resulting RKHS. The functions in $\mathbb{K}$ inherit properties from k, in particular, one can choose k so that the functions in $\mathbb{K}$ possess some number of derivatives, or satisfy some boundary conditions. In addition, many Sobolev spaces, which are commonly used to enforce smoothness conditions, are also RKHS's. We refer an interested reader to Berlinet and Thomas-Agnan [1] for further details.

We propose to estimate g by minimizing the following penalized objective:

$$RSS_\lambda(g) = \sum_{i=1}^n \int_0^1 \left(Y_i(t) - \int_0^1 g(t, s, X_i(s))\, ds \right)^2 dt + \lambda \|g\|_{\mathbb{K}}^2, \qquad (1)$$

i.e.,

$$\hat{g} = \arg \inf_{g \in \mathbb{K}} RSS_\lambda(g),$$

where $\lambda > 0$. As we will see in the next section, an explicit solution to this minimization problem exists due to the reproducing property. However, we will also discuss using FPCA to help reduce the computational burden.

3. Computation

One of the benefits of using RKHS methods is that one can often get an exact solution to the corresponding minimization problem such as the one in (1), due to the representer theorem [14]. This also turns out to be the case here, however, later on we will discuss using a slightly modified version that still works well and is easier to compute. The expression we derive is quite a bit simpler than the analogs derived in [3, 32, 30]; this is partly due to our use of functional principal components, which simplify the expression and also provide an avenue for reducing the computational complexity of the problem, and also due to our use of the RKHS norm penalty when fitting the model (where as others used a more general penalty term).

Using the reproducing property we have

$$\langle k_{t,s,X_i(s)}, g \rangle_{\mathbb{K}} = g(t, s, X_i(s)) \quad \text{for} \quad i = 1, 2, ..., n.$$

We then have that

$$\int_0^1 g(t, s, X_i(s)) ds = \int_0^1 \langle g, k_{t,s,X_i(s)} \rangle_{\mathbb{K}} ds = \left\langle g, \int_0^1 k_{t,s,X_i(s)} ds \right\rangle_{\mathbb{K}}, \quad (2)$$

which is justified by the integrability constraints inherent in Assumption 1(*iii*), discussed in the next section. Let $\hat{v}_1, \hat{v}_2, ..., \hat{v}_n$ denote the empirical functional principal components, EFPC's, of $Y_1, Y_2, ..., Y_n$. Then, assuming the Y_i's are centered, it is a basic fact of PCA that $\text{span}\{\hat{v}_1, ..., \hat{v}_n\} = \text{span}\{Y_1, ..., Y_n\}$. Recall that it is also a basic fact from linear algebra that the $\hat{v}_1, \hat{v}_2, ..., \hat{v}_n$ can be completed to form a full orthonormal basis (all of the additional functions will have an empirical eigenvalue of 0). We then apply Parseval's identity to obtain

$$\sum_{i=1}^n \int_0^1 \left(Y_i(t) - \int_0^1 g(t, s, X_i(s)) ds \right)^2 dt$$

$$= \sum_{i=1}^n \sum_{j=1}^\infty \left(\langle Y_i, \hat{v}_j \rangle - \left\langle g, \int_0^1 \int_0^1 k_{t,s,X(s)} \hat{v}_j(t) dt ds \right\rangle_{\mathbb{K}} \right)^2.$$

Define the subspace (of $\mathbb{K}$)

$$\mathcal{H}_1 = \text{span} \left\{ \int_0^1 \int_0^1 k_{t,s,X_i(s)} \hat{v}_j(t) dt ds, \ i = 1, 2, ..., n, \ j = 1, ..., n \right\},$$

as well as its orthogonal compliment $\mathcal{H}_1^\perp$. The space $\mathbb{K}$ can be decomposed into the direct sum: $\mathbb{K} = \mathcal{H}_1 \oplus \mathcal{H}_1^\perp$, which means that we can write any function $g \in \mathbb{K}$ as $g = g_1 + g_1^\perp$, with $g_1 \in \mathcal{H}_1$ and $g_1^\perp \in \mathcal{H}_1^\perp$. Using this decomposition we have that, for $1 \le j \le n$,

$$\left\langle g, \int_0^1 \int_0^1 k_{t,s,X_i(s)} \hat{v}_j(t) dt ds \right\rangle_{\mathbb{K}} = \left\langle g_1, \int_0^1 \int_0^1 k_{t,s,X_i(s)} \hat{v}_j(t) dt ds \right\rangle_{\mathbb{K}}. \quad (3)$$

Since $\|g\|_{\mathbb{K}}^2 = \|g_1\|_{\mathbb{K}}^2 + \|g_1^\perp\|_{\mathbb{K}}^2$, it follows from (1) and (3) that $\hat{g} \in \mathcal{H}_1$ and so has the form

$$\hat{g}(t, s, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \int_0^1 \int_0^1 k\left((t, s, x); (t', s', X_i(s'))\right) \hat{v}_j(t') dt' ds'.$$

Note that this same expression would hold if we replaced the $\{v_j(t)\}$ with $\{Y_j(t)\}$ (since they span the same space), however, it would not hold for an arbitrary basis. We use the FPCs for computational reasons as we discuss at the end of the section. To compute the estimate, $\hat{g}$, we only need to compute the coefficients $\{\alpha_{ij}\}$. As usual, the coefficients α_{ij} can be computed via a type of ridge regression. Note that

$$\left\langle \hat{g}, \int_0^1 \int_0^1 k_{t,s,X_i(s)} \hat{v}_j(t) dt ds \right\rangle_{\mathbb{K}}$$
$$= \sum_{i'=1}^{n} \sum_{j'=1}^{n} \alpha_{i'j'} \int_0^1 \int_0^1 \int_0^1 \int_0^1 \langle k_{t,s,X_{i'}(s)}, k_{t',s',X_i(s')} \rangle_{\mathbb{K}} \hat{v}_{j'}(t) \hat{v}_j(t') dt ds dt' ds'$$
$$= \sum_{i'=1}^{n} \sum_{j'=1}^{n} \alpha_{i'j'} \int_0^1 \int_0^1 \int_0^1 \int_0^1 k(t, s, X_{i'}(s); t', s', X_i(s')) \hat{v}_{j'}(t) \hat{v}_j(t') dt ds dt' ds'.$$

Define

$$A_{iji'j'} = \int_0^1 \int_0^1 \int_0^1 \int_0^1 k(t, s, X_{i'}(s); t', s', X_i(s')) \hat{v}_{j'}(t) \hat{v}_j(t') dt ds dt' ds'.$$

Turning to the norm in the penalty we can use the same arguments to show that

$$\|\hat{g}\|_{\mathbb{K}}^2 = \langle \hat{g}, \hat{g} \rangle_{\mathbb{K}} = \sum_{iji'j'} \alpha_{ij} A_{iji'j'} \alpha_{i'j'}.$$

Thus the minimization problem can be phrased as

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(Y_{ij} - \sum_{i'j'} A_{iji'j'} \alpha_{i'j'} \right)^2 + \lambda \sum_{iji'j'} \alpha_{ij} A_{iji'j'} \alpha_{i'j'}.$$

We now vectorize the problem by stacking the columns of Y_{ij} and α_{ij}, denoted as $\mathbf{Y}_V$ and α_V. We also turn the array $A_{iji'j'}$ into a matrix $\mathbf{A}_V$, by collapsing the corresponding dimensions. We can then phrase the minimization problem as

$$(\mathbf{Y}_V - \mathbf{A}_V \alpha_V)^\top (\mathbf{Y}_V - \mathbf{A}_V \alpha_V) + \lambda \alpha_V^\top \mathbf{A}_V \alpha_V.$$

Thus, the final estimate can be expressed as

$$\hat{\alpha}_V = (\mathbf{A}_V^\top \mathbf{A}_V + \lambda \mathbf{A}_V)^{-1} \mathbf{A}_V \mathbf{Y}_V.$$

Note that we are estimating n^2 parameters and inverting an $n^2 \times n^2$ matrix. Thus for computational convenience, it is often useful to truncate the EFPCs at

some value $J < n$. However, even without truncating this approach still has the potential to lead to less parameters than the basis methods of Scheipl et al. [29], where the number of parameters to estimate is m^3, with m being the number of basis functions used in their tensor product basis. In contrast, our approach yields n^2 parameters, and combined with an FPCA, this can be reduced to nJ with relatively little loss in practical predictive performance. There is also the possibility of using an eigen-expansion on k to reduce the computational complexity even further [24], though we don't pursue that here.

3.1. Alternative Domains

While our work is focused primarily on the "classic" function-on-function paradigm, we briefly mention in this section an easy way to modify the kernels to allow for more complex domains. In particular, one major concern brought up by a referee is when both $X_i(t)$ and $Y_i(t)$ are observed concurrently. In that case, the classic approach would actually use future values of the covariate to predict present values of the outcome. Interestingly, we need only make a very slight adjustment to the kernels to handle such a setting.

The goal here is to adjust the model such that

$$Y_i(t) = \int_0^t g(t, s, X_i(s)) \, ds + \varepsilon_i(t) \qquad 0 \le t \le 1, \tag{4}$$

or equivalently to require that $g(t, s, X_i(s)) = 0$ if $s > t$. More generally, we can allow the domain of X used to predict Y to change arbitrarily with t. Let $\{A_t \subset [0, 1] : 0 \le t \le 1\}$ be a collection of (measurable) subsets of the unit interval. Fitting (4) is equivalent to taking $A_t = [0, t]$, which is what we use to highlight this approach in Section 6. We aim to fit the more general model

$$Y_i(t) = \int_{A_t} g(t, s, X_i(s)) \, ds + \varepsilon_i(t) \qquad 0 \le t \le 1.$$

Interestingly, this can be done through a simple modification of the kernel. In particular, we can define a new kernel as

$$\tilde{k}(t, s, x, t', s', x') = 1_{s \in A_t} 1_{s' \in A_{t'}} k(t, s, x, t', s', x').$$

A direct verification shows that $\tilde{k}$ is a valid reproducing kernel as long as the original k was. Then our estimate would take the form

$$\hat{g}(t, s, x) = \sum_{i=1}^n \sum_{j=1}^n \int_0^1 \int_0^1 \tilde{k}(t, s, x; t', s', X_i(s')) \hat{v}_j(t') dt' ds'$$

$$= 1_{s \in A_t} \sum_{i=1}^n \sum_{j=1}^n \int_0^1 \int_{s' \in A_{t'}} k(t, s, x; t', s', X_i(s')) \hat{v}_j(t') ds' dt',$$

which means that $\hat{Y}_{n+1}(t)$ can be computed using only $\{X_{n+1}(s) : s \in A_t\}$ and a very slight modification of our current approach. We illustrate this technique in Section 6.

4. Asymptotic theory

In this section, we demonstrate that the *excess risk*, $\Re_n$ (defined below), of our estimator converges to zero at the optimal rate. Optimal convergence of $\Re_n$, for scalar-on-function linear regression was established by Cai and Yuan [3], while optimal convergence for the continuously additive scalar-on-function regression model was established in Wang and Ruppert [32]. In both cases an RKHS estimation framework was used. Because our model involves a functional response, the form of the excess risk $\Re_n$ is different and requires some serious mathematical extensions over previous works. However, we will show that the convergence rate for our model is the same as the one found in Wang and Ruppert [32].

We begin by defining the excess risk, $\Re_n$. Let $X_{n+1}(t)$ be new predictor which is distributed as, but independent of $(X_i(t))_{i=1}^n$. We let E^* denote the expected value, conditioned on the data $\{(Y_i, X_i) : 1 \le i \le n\}$. Then the excess risk is defined as

$$\Re_n = E^* \left[\int_0^1 \int_0^1 (\hat{g}(t, s, X_{n+1}(s)) - g(t, s, X_{n+1}(s)))^2 \ dtds \right].$$

Note that $\Re_n$ is still a random variable as it is a function of the data. Intuitively, this quantity can be thought of as prediction error, namely, for a future observation, how far away is our prediction from the optimal one where the true g is known. For ease of exposition, we present all of assumptions below, even the ones discussed previously.

Assumption 1. We make the following assumptions.

(i) The observations $\{Y_i(t), X_i(t)\}$ are assumed to satisfy

$$Y_i(t) = \int g(t, s, X_i(s)) \ ds + \varepsilon_i(t)$$

where $\{X_i\}$ and $\{\varepsilon_i\}$ are independent of each other and iid across $i = 1, \dots, n$.

(ii) Denote by $\mathcal{L}_k$ the integral operator with k as its kernel:

$$(\mathcal{L}_k f)(t, s, x) := \int k(t, s, x; t', s', x') f(t', s', x') \ dt'ds'dx'.$$

The kernel, k, which also defines the RKHS, $\mathbb{K}$, is assumed to be symmetric, positive definite, and square integrable.

(iii) Assume that there exists a constant $c > 0$ such that for any $f \in \mathbb{K}$ and $t \in [0, 1]$ we have

$$E \left(\int_0^1 f(t, s, X(s)) \ ds \right)^4 \le c \left[E \left(\int_0^1 f(t, s, X(s)) \ ds \right)^2 \right]^2 < \infty.$$

(iv) Let $\mathcal{L}_k^{1/2}$ denote a square–root of $\mathcal{L}$ (which exists due to Assumption 1(ii)) and define $k_{t,s,x}^{1/2} := \mathcal{L}_k^{-1/2} k_{t,s,x}$. Define the operator, C, as

$$C(f) = E \left[\int \int \int k_{t,s,X_i(s)}^{1/2} \langle k_{t,s',X_i(s')}^{1/2}, f \rangle_{L^2} \ dsds'dt \right].$$

Assume that the eigenvalues $\{\rho_k : k \geq 1\}$ of C satisfy $\rho_k \asymp k^{-2r}$ for some constant $r > 1/2$.

(v) There exists a constant $M > 0$ such that, for all $t \in [0, 1]$ and $i = 1, \ldots, M$

$$\mathrm{E}(\epsilon_i^2(t)) \leq M < \infty.$$

(vi) The function g lies in Ω, which we assume is a closed bounded ball in $\mathbb{K}$.

We are now in a position to state our main result.

Theorem 1. *If Assumption 1 holds and the penalty parameter, λ, is chosen such that $\lambda \asymp n^{-\frac{2r}{2r+1}}$ then we have that*

$$\lim_{A \to \infty} \lim_{n \to \infty} \sup_{g \in \Omega} \mathbb{P}\left(\Re_n \geq A n^{-\frac{-2r}{2r+1}}\right) = 0.$$

Before interpreting this result, let us discuss each of the assumptions individually. Assumption $1(i)$ explicitly defines the model we are considering. Assumption $1(ii)$ ensures that the kernel has a spectral decomposition via Mercer's theorem, which will be used extensively. Assumption $1(iii)$ is fairly typical in these sorts of asymptotics, assuming that the fourth moment is bounded by a constant times the square of the second. Assumption $1(iv)$ introduces a central quantity that is used extensively in the proofs. While not immediately obvious, this assumption basically states how "smooth" or "regular" the function g is, as g must lie in $\mathbb{K}$, whose kernel contributes to C. In such results it is common for X to contribute to the asymptotic behavior as the prediction error depends on the complexity of the X. Note that $k_{t,s,x}^{1/2}$ is a well defined quantity and it is easy to show via the reproducing property that it is an element of $L^2([0,1]^2 \times \mathbb{R})$. The operator C does depend on the choice of the square-root $\mathcal{L}_k^{1/2}$ (which is not a unique choice), however its eigenvalues do not. Assumption $1(v)$ simply assumes that the point-wise variance of the errors is bounded, while the last assumption requires that the true function lie in a ball in $\mathbb{K}$, which is used to control the bias of the estimate.

The rate given in Theorem 1 is the same as was found in the scalar outcome case in Wang and Ruppert [32], thus we know that this is the minimax rate of convergence. In our case, as well as in Wang and Ruppert [32] and Cai and Yuan [3], it is the interaction between the covariance of X and the kernel k which determines the optimal rate. The proof is quite extensive and given in the appendix. The idea of the proof is to rephrase the estimate using operator notation instead of the representation theorem. The difference between the estimate and truth is then split into a bias/variance decomposition. Bounding the bias turns out to be relatively straight forward. Bounding the variance is done by decomposing it into five more manageable pieces, and then bounding each of them separately. Our task is complicated by the fact that the errors and response are now functions, where as in both Wang and Ruppert [32] and Cai and Yuan [3] they were scalars. This requires extending many of the lemmas to this new setting, as well as using some completely new arguments to get the necessary bounds in place.

5. Simulation study

Here we investigate the prediction performance of AFFR. We compare it with a linear model estimated in one of two ways. The first way will be denoted as LMR (linear model reduced) and LMF (linear model full), where both use FPCA to reduce the dimension of the predictors, but LMR also reduces the dimension of the outcome, while LMF does not. To implement our approach we relied heavily on the *TensorA* package [31] in R, which allowed us to carryout various tensor products very quickly.

We consider three different settings for $g(t, s, x)$ one linear and two nonlinear forms:

(a) Scenario (a): $g(t, s, x) = tsx$,
(b) Scenario (b): $g(t, s, x) = t + s + x^2$,
(c) Scenario (c): $g(t, s, x) = tsx^2 + x^4$.

In all settings, the predictors $X_i(t)$ and errors $\epsilon_i(t)$ are taken to be iid Gaussian processes with mean 0 and the following covariance function from the Matérn family:

$$C(t, s) = \left(1 + \frac{\sqrt{5}|t-s|}{\rho} + \frac{5|t-s|^2}{3\rho^2}\right) \exp\left(-\frac{\sqrt{5}|t-s|}{\rho}\right),$$

where $\rho = 1/4$. For the RKHS we considered both the Gaussian kernel

$$k\left((x, y, z), (x', y', z')\right) = e^{-\delta\left[(x-x')^2 + (y-y')^2 + (z-z')^2\right]},$$

and exponential kernel

$$k\left((x, y, z), (x', y', z')\right) = e^{-\delta\left[|x-x'| + |y-y'| + |z-z'|\right]},$$

where δ is the range parameter. We will examine the sensitivity of our approach to this parameter in Tables 2 and 3. All of the curves ($X_i(t)$, $Y_i(t)$, and $\varepsilon_i(t)$) were simulated on a $M = 50$ equispaced grid between 0 and 1. The data is approximated using $K = 100$ B-splines. We denote by J_X and J_Y the number of principal components of X and Y respectively. These steps are carried out using the *Data2fd* and *pca.fd* functions in the *R* package *fda*. Our approach uses an FPCA on Y only, but the LMR approach uses the FPCs for both X and Y. The common recommendations for choosing J_Y is either to use some cutoff for explained variability (commonly 85%) or to look for an elbow in the scree plot (J_X can also be chosen the same way or using a model based criteria such as BIC) [16]. Using an 85% cutoff here results in 3 FPCs for our simulations, though we also include 6 and 9 to show that our approach is not very sensitive to this choice as long as a large proportion of variability is explained. However, one should note the trade offs when choosing J_Y. In general, the major gain in choosing a smaller J_Y is faster computation, which is nontrivial for this problem. The major loss is that one "gives up" on some proportion of the variability in Y.

For example, if the FPCs explain 95% of the variability, then one immediately gives up on predicting that remaining 5%. This is a different consideration than when choosing FPCs for predictors. In general, users can tailor this choice to their data; if one expects very accurate predictions then a larger J_Y can be helpful so that one does not lose prediction accuracy, while if it is known a-priori that the prediction accuracy will be low, then J_Y can be safely made smaller.

To evaluate the different approaches, we used 1000 repetitions of every scenario. In each case we generate 150 curves to fit the different models and then generated another 150 curves to evaluate *out-of-sample* prediction error. The metric for determining prediction performance we denote as RPE, for relative prediction error. This metric denotes the improvement of the predictions over just using the mean, and can be thought of as a type of out-of-sample R^2. An RPE of 0 implies that the model shows no improvement over just using the mean, while an RPE of 1 means the predictions are perfect. More precisely, we first compute the Mean Squared Prediction error as:

$$MSPE = \sum_{i=1}^{n} \|Y_i - \widehat{Y}_i\|_{L^2}^2,$$

where $\widehat{Y}_i$ is a predicted value using one of the three discussed models or simply the mean. The RPE is then defined as

$$RPE = \frac{MSPE_{mean} - MSPE}{MSPE_{mean}},$$

where $MSPE_{mean}$ denotes the MSPE using a mean only model. Note that even in the mean only model, all parameters are estimated on the initial 150 curves and prediction is then evaluated on the second 150. Therefore, it is actually possible to have a numerically negative RPE if an approach isn't predicting any better than just using the mean.

The RPEs of LMR and LMF for the three models (a), (b), and (c) are summarized in Table 1. For both models, we took $J_X = 3$, which explained over 85% of the variability of the predictors and for LMR we took $J_Y = 3$ PCs for the outcome as well. The RPEs for our approach with $\delta = \{2^{-3}, 2^{-2}, 2^{-1}, 1, 2\}$ and $J_Y = 3, 6, 9$ are summarized in Tables 2 and 3, which represent the Gaussian and exponential kernels respectively. An initial look at the tables confirms much of what one would expect. When the true model is linear, the two linear approaches work best, resulting in about twice the RPE of AFFR. However, when moving to the two nonlinear models, the AFFR approach does substantially better. This increased performance is seen for any choice of J_Y and δ. Furthermore, the prediction performance seems relatively robust to the choice of J_Y, δ, and even the kernel. In the case of J_Y this is not so surprising as over 90% of the variability of the Y_i is explained by the first three FPCs. In contrast, there is some sensitivity to the choice of δ, but it is relatively weak given how much we are changing δ in each row. In our application section we set δ using a type of median, but one could also refit the model with a few different δ and choose

the one with the best prediction performance. Given how consistent the AFFR predictions are, trying a few δ appears to be satisfactory, and large grid searches can be avoided.

	Scenario (a)	Scenario (b)	Scenario (c)
LMR	0.045	0.030	0.060
LMF	0.045	0.029	0.060

TABLE 1

Relative prediction errors, RPE, for the two linear models. For both, the number of FPCs for the predictor is $J_X = 3$. LMR also reduces the dimension of the outcome with $J_Y = 3$ FPCs.

	Scenario (a)			Scenario (b)			Scenario (c)		
	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$
$\delta = 2^{-3}$	0.025	0.026	0.026	0.379	0.379	0.379	0.840	0.840	0.845
$\delta = 2^{-2}$	0.024	0.025	0.025	0.370	0.370	0.370	0.816	0.804	0.815
$\delta = 2^{-1}$	0.023	0.024	0.023	0.360	0.361	0.361	0.847	0.831	0.830
$\delta = 2^{0}$	0.022	0.023	0.021	0.346	0.347	0.347	0.83	0.83	0.83
$\delta = 2^{1}$	0.020	0.021	0.019	0.328	0.328	0.400	0.808	0.808	0.790
PEV	90.45%	99.12%	99.88%	90.82%	99.10%	99.84%	91.25%	99.22%	99.87%

TABLE 2

Relative prediction error, RPE, for AFFR using a Gaussian kernel and with different kernel parameter values, δ. In every case the penalty parameter, λ, is chosen using cross-validation. PEV indicates the proportion of explained variance of Y for the corresponding number of FPCs, J_Y.

	Model (a)			Model (b)			Model (c)		
	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$	$J_Y = 3$	$J_Y = 6$	$J_Y = 9$
$\delta = 2^{-3}$	0.021	0.023	0.020	0.368	0.379	0.379	0.774	0.789	0.775
$\delta = 2^{-2}$	0.022	0.023	0.023	0.361	0.357	0.359	0.813	0.805	0.815
$\delta = 2^{-1}$	0.022	0.023	0.023	0.350	0.349	0.351	0.829	0.813	0.818
$\delta = 2^{0}$	0.021	0.022	0.022	0.338	0.332	0.334	0.780	0.800	0.792
$\delta = 2^{1}$	0.020	0.019	0.019	0.300	0.304	0.302	0.743	0.752	0.749
PEV	90.45%	99.12%	99.88%	90.82%	99.10%	99.84%	91.25%	99.22%	99.87%

TABLE 3

Relative prediction error, RPE, for AFFR using an exponential kernel and with different kernel parameter values, δ. In every case the penalty parameter, λ, is chosen using cross-validation. PEV indicates the proportion of explained variance of Y for the corresponding number of FPCs, J_Y.

As a final illustration of the efficacy of AFFR, we provide several plots to help visualize the performance. In Figure 1 we plot several realizations of Y_i and their corresponding (out of sample) predictions using the optimal prediction, $E[Y(t)|X]$, AFFR, and the linear model without reducing the dimension of the Y. We consider only the Gaussian kernel and take $\delta = 1/4$. For the nonlinear scenarios (rows 2 and 3), one can clearly see the RPE results reflected in the predictions as AFFR is much closer to the optimal prediction. In Figure 2 we plot several realizations of $\hat{g}(t, s, X_i(s))$, which are again done out of sample along with the true value of $g(t, s, X_i(s))$. Plotting in this way allows us to visualize g using surfaces, where as plotting $g(t, s, x)$ would be challenging since

the domain has three coordinates. As we can see, the estimates are quite close to the true values, capturing the nonlinear structure quite well.
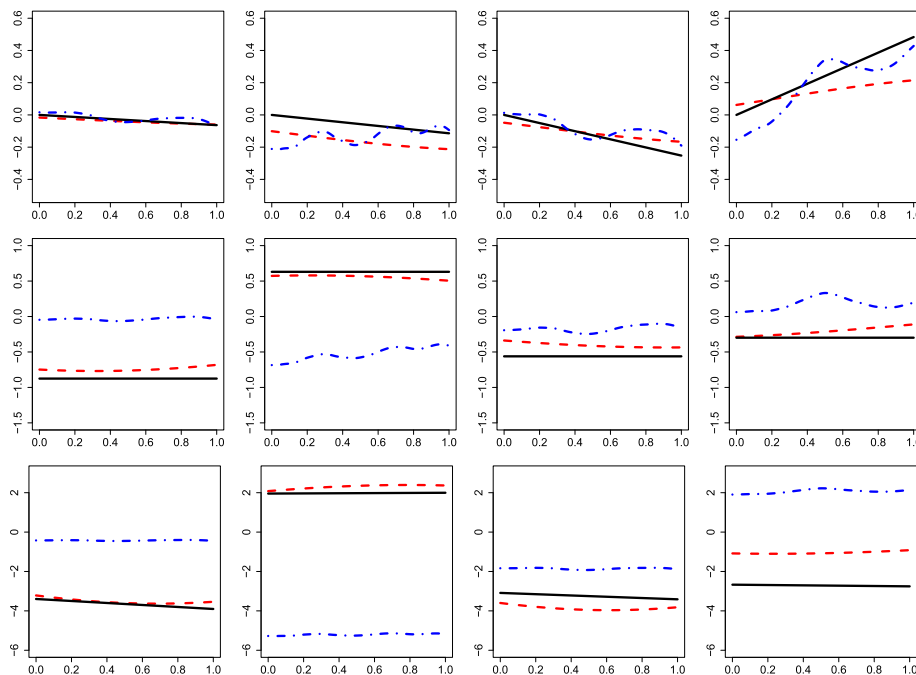

FIG 1. *Plots of the optimal prediction $E[Y(t)|X]$ (black), prediction using AFFR $\hat{Y}(t)$ (red dashed), and prediction using the unreduced linear model $\hat{Y}_{LM}(t)$ (blue dashed). The four plots on the top row correspond to the scenario (a), which is linear. The four plots in the middle row correspond to the scenario (b), which is nonlinear. The four plots in the bottom row correspond to the scenario (c) which is also nonlinear.*

6. Application to cumulative intraday data

We conclude with an illustration of our approach applied to real data. Cumulative Intra-Day Returns (CIDR's) consist of daily stock prices that are normalized to start at zero at the beginning of each trading day. FDA methods have been useful in analyzing such data [6, 15, 9], given the density at which stock prices can be observed. Let $P_i(t_j)$ denote the price of a stock on day i and time of day t_j. The CIDRs are then defined as

$$R_i(t_j) = 100 \left[\ln P_i(t_j) - \ln P_i(t_1) \right], i = 1, ..., n, \ j = 1, ..., M.$$

The CIDRs are observed each minute throughout the trading day. This corresponds to $M = 390$ minutes (9:30 am-4:00 pm EST) of trading time for each trading day of the New York Stock Exchange, NYSE. In this application study,
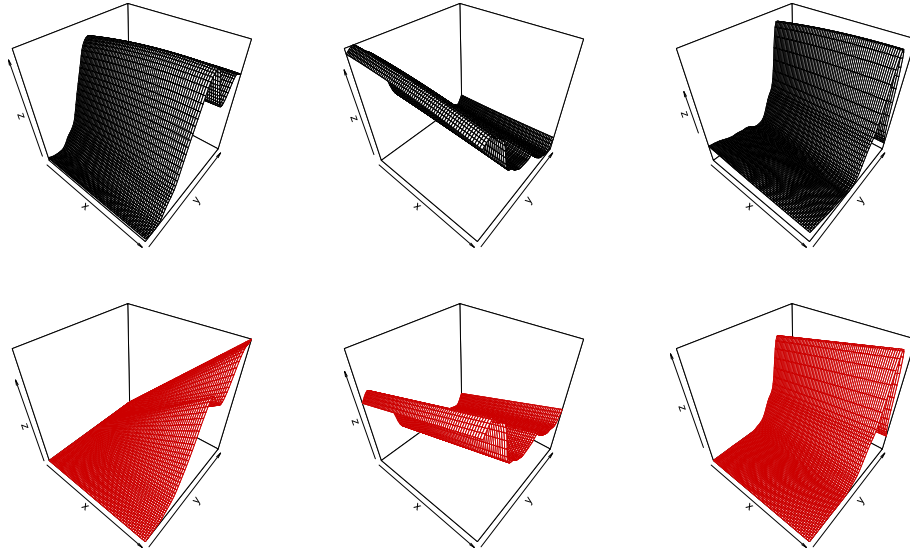
FIG 2. *The top row plots one realization of $g(t, s, X(s))$ for models (a), (b), and (c) respectively. The bottom row plots the corresponding (out of sample) prediction $\hat{g}(t, s, X(s))$.*

we deal with the CIDR's of two of the most important US market indexes: Standard & Poor's 500 Index (S&P 500) and the Dow Jones Industrial Average (DJ). Also, we consider two individual stocks: General Electric Company (GE) and International Business Machines Corporation (IBM). The study period of the data consists of three periods in relation to the 2007–2008 financial crisis, denoted as *Before* (06/13/2006-04/10/2007), *During* (11/01/2007-07/28/2008), and *After* (01/04/2010-10/1/2010). These periods each contain 270 calendar days.

We investigate the performance of the market indexes, S&P 500 and DJ, in predicting GE and IBM for the three periods. Understanding such relationships is imperative for developing financial portfolios as many strategies consist of balancing buying/shorting certain stocks with buying/shorting market indices [23]. We fit four different models; the first two are our discussed models, AFFR, one based on using the full $X_i(t)$ to predict $Y_i(t)$ (AFFR) and one where only the current and past values are used (AFFR Pre) as described in Section 3.1. The other two methods are the linear models. The first linear model uses an FPCA on both the outcome and predictor (5 PCs for both) and then fits a multivariate linear model, while the second linear model only uses FPCA on the predictor (5 PCs) [16]. To evaluate the prediction performance for each period we split each period into 3 equal folds and use a type K-fold cross-validation. The model is fit on two folds, while prediction is then evaluated on the third. We use the Gaussian kernel from Section 5 and the smoothing parameter selected via Generalized Cross-Validation. Prediction performance is then averaged over the 3 folds. To provide a more readily interpretable metric for prediction performance, we use the same RPE metric given in Section 5, which

denotes the relative performance of a model with respect to a mean only model. A value of 1 means perfect prediction, while a value of 0 indicates that the model is doing no better than just using the mean. The results are summarized in Table 4.

As we can see, all models perform better during and after the crisis. This suggests that the behavior of the market had not returned to its pre-crisis characteristics. Looking at Figure 3, we can clearly see that the volatility increases during and after the financial crises. This suggests that the overall "market" effect on the stocks is stronger during periods of high-volatility. When comparing the four different models, the linear models do nearly the same, which is to be expected since 5 PCs explains over 90% of the variability of the stocks. The AFFR model is not too far behind, but does noticeably worse in every setting. This suggests that the relationship between the discussed stocks and the indices is approximately linear; if there are any nonlinear relationships then they are either very minor deviations from linearity or are not well captured by an additive structure. The results of AFFR using only current and past values of $X_i(t)$ to predict $Y_i(t)$ (AFFR Pre) does substantially worse before the crises. Interestingly, during and after its performance is closer to AFFR, though some relationships it still does not capture well. Thus suggests that, unsurprisingly, knowing the future values of $X_i(t)$ is very helpful for predicting current values of $Y_i(t)$, though this is obviously impractical. During the financial crises, many stocks are likely being driven by large market level effects. In this setting, AFFR Pre, does quite well, even beating AFFR slightly in some settings, suggesting that the simpler structure has actually helped with prediction.

Period	Before				During				After			
Model	AFFR	AFFR Pre	LM Red	LM Full	AFFR	AFFR Pre	LM Red	LM Full	AFFR	AFFR Pre	LM Red	LM Full
GE on DJ	0.133	5.124e-06	0.191	0.191	0.459	0.311	0.536	0.548	0.500	0.421	0.501	0.512
GE on SP	1.325e-07	4.216e-14	0.184	0.183	0.273	0.253	0.458	0.472	0.510	0.436	0.487	0.497
IBM on DJ	0.092	1.645e-03	0.182	0.184	0.274	0.350	0.486	0.495	0.364	0.011	0.402	0.412
IBM on SP	0.079	1.251e-11	0.180	0.180	0.213	0.272	0.373	0.384	0.296	0.009	0.343	0.351

TABLE 4

Prediction performance of four models: AFFR (our model), AFFR Pre (modifies domain to avoid using future values), LM Red (linear model with PCA in both the outcome and predictor), and LM Full (linear model with PCA on the predictor only). The top row corresponds to predicting GE based on DJ, the second corresponds to prediction GE from SP, and so on. Each number denotes the relative increase in out-of-sample prediction performance over a mean only model, with 100% denoting perfect prediction and 0% denoting no increase over just using the mean.

7. Conclusions and future work

In this paper we have presented a new RKHS framework for estimating an additive function-on-function regression model, that is better able to account for complex nonlinear dynamics in functional regression models than classic linear models. We showed that the estimator is minimax in the sense that it achieves an optimal rate of convergence in terms of prediction error. In addition, computing the estimate is computationally efficient, especially if dimension reduction is incorporated.
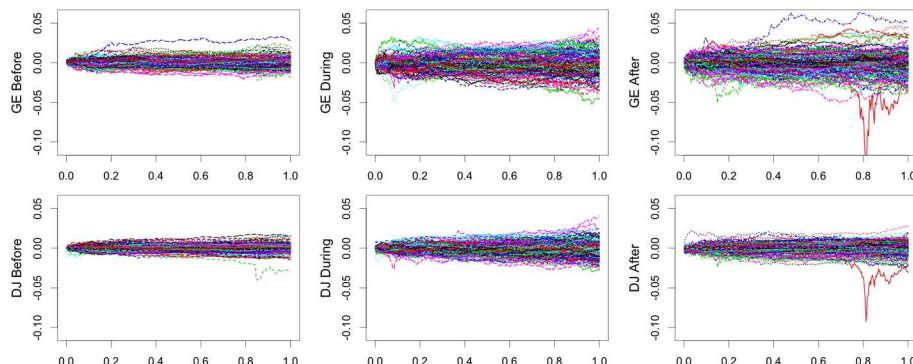
FIG 3. *Plots of the intraday cumulative returns for the Dow Jones Index (top) and General Electric (bottom) before (left), during (middle), and after (right) the 2008 financial crisis.*

Nonlinear models for functional data have recently received a great deal of attention, however, there are still a number of interesting questions that remain open. One that is especially relevant to the work presented here concerns further statistical properties of the estimate, $\hat{g}$. In particular, convergence rates of $\hat{g}$ as well as its asymptotic distribution would be especially interesting for quantifying the estimation uncertainty in practice. Using such tools, one could also construct confidence/prediction bands, which would be of great use in practice.

Another nontrivial extension would be to curves that are observed sparsely. Nonlinear models in FDA often require that the curves be observed or at least consistently estimated. However, for some data this is unrealistic and there is a great deal of uncertainty related to imputing the curves.

Lastly, extensions to more complex settings would also be of interest. For example, the handling of more complex domains, e.g. space or space-time. In these cases, the minimax rates usually depend on the dimension of the domain. Another important extension would be to functional binary or categorical outcomes (as opposed to quantitative) would be of interest as one must incorporate tools from functional glms.

Appendix A: Proof of Theorem 1

A.1. Excess risk

We begin by expressing the excess risk in an alternative form. Recall that $k(t, s, x; t', s', x')$ is the kernel function used to define the RKHS, $\mathbb{K}$. This kernel can be viewed as the kernel of an integral operator, $\mathcal{L}_k$, which maps $L^2([0, 1]^2 \times \mathbb{R}) \to \mathbb{K} \subset L^2([0, 1]^2 \times \mathbb{R})$. In particular

$$(\mathcal{L}_k f)(t, s, x) = \int \int \int k(t, s, x; t', s', x') f(t', s', x') \, dt' ds' dx'.$$

From here on, for simplicity, we will denote $L^2([0,1]^2 \times \mathbb{R})$ as simply L^2. By Assumption 1, $\mathcal{L}_k$ is a positive definite, compact operator, which is also self-adjoint in the sense that $\langle f, \mathcal{L}_k g \rangle = \langle \mathcal{L}_k f, g \rangle$, for any f and g in L^2. We can therefore define a square-root of $\mathcal{L}_k$, denoted as $\mathcal{L}_k^{1/2}$ that satisfies

$$f_1 \in \mathbb{K} \Longleftrightarrow \mathcal{L}_k^{-\frac{1}{2}} f_1 \in L^2 \quad \text{and} \quad \mathcal{L}_k^{\frac{1}{2}} f_2 \in \mathbb{K} \Longleftrightarrow f_2 \in L^2.$$

Note that if $\mathcal{L}_k$ has a nontrivial null space, then $\mathcal{L}_k^{-1/2}$ can still be well defined since assuming $f \in \mathbb{K}$ means that f is orthogonal to the null space of $\mathcal{L}$. Recall that one can also move between the $\mathbb{K}$ and L^2 inner product as follows

$$\langle f, g \rangle_{\mathbb{K}} = \langle \mathcal{L}_k^{-1/2} f, \mathcal{L}_k^{-1/2} g \rangle_{L^2} = \langle f, \mathcal{L}_k^{-1} g \rangle_{L^2}.$$

We refer the interested reader to Kennedy and Sadeghi [12] for more details.

Let $\hat{g}$ denote our estimate of the true function, g. We then define the following

$$k_{t,s,X(s)}^{\frac{1}{2}} = \mathcal{L}_k^{-\frac{1}{2}} k_{t,s,X(s)}, \quad h = \mathcal{L}_k^{-\frac{1}{2}} g \quad \text{and} \quad \hat{h} = \mathcal{L}_k^{-\frac{1}{2}} \hat{g}.$$

Using the reproducing property, we have that

$$g(t,s,X(s)) = \langle k_{t,s,X(s)}^{\frac{1}{2}}, h \rangle_{L^2} \quad \text{and} \quad \hat{g}(t,s,X(s)) = \langle k_{t,s,X(s)}^{\frac{1}{2}}, \hat{h} \rangle_{L^2}.$$

Now define the random operator, $T : L^2 \to L^2$ as

$$T_{t,s,s'} = k_{t,s,X_{n+1}(s)}^{1/2} \otimes k_{t,s',X_{n+1}(s')}^{1/2},$$

where $\otimes$ denotes the tensor product, and the resulting object is interpreted as an operator:

$$T_{t,s,s'}(f) = k_{t,s,X_{n+1}(s)}^{1/2} \langle k_{t,s',X_{n+1}(s')}^{1/2}, f \rangle_{L^2}.$$

We also define a second operator, which integrates out t, s, and s', and takes an expectation over X_{n+1}:

$$C = \mathrm{E} \left[\int_0^1 \int_0^1 \int_0^1 T_{t,s,s'} \, ds \, ds' \, dt \right]. \tag{5}$$

Note that C is a symmetric, positive definite, compact operator, and thus has a spectral decomposition

$$C = \sum_{k=1}^{\infty} \rho_k (\phi_k \otimes \phi_k), \tag{6}$$

where $\rho_k \geq 0$ and $\phi_k \in L^2$ are, respectively, the eigenvalues and eigenfunctions of C. This decomposition will be used later on.

As we said before, denote by E^* the expected value conditioned on the data $(X_1, Y_1), ..., (X_n, Y_n)$. The excess risk can be written as

$$\Re_n = E^* \int_0^1 \left(\int_0^1 [\hat{g}(t, s, X_{n+1}(s)) - g(t, s, X_{n+1}(s)))] \, ds \right)^2 dt$$

$$= E^* \int_0^1 \left(\int_0^1 \left[\langle k_{t,s,X_{n+1}(s)}^{\frac{1}{2}}, \hat{h}_\lambda \rangle_{L^2} - \langle k_{t,s,X_{n+1}(s)}^{\frac{1}{2}}, h \rangle_{L^2} \right] ds \right)^2 dt$$

$$= E^* \int_0^1 \left(\int_0^1 \langle k_{t,s,X_{n+1}(s)}^{\frac{1}{2}}, \hat{h}_\lambda - h \rangle_{L^2} ds \right)^2 dt$$

$$= E^* \int_0^1 \int_0^1 \int_0^1 \langle k_{t,s,X_{n+1}(s)}^{\frac{1}{2}}, \hat{h} - h \rangle_{L^2} \langle k_{t,s',X_{n+1}(s')}^{\frac{1}{2}}, \hat{h} - h \rangle_{L^2} \, ds \, ds' \, dt$$

$$= E^* \int_0^1 \int_0^1 \int_0^1 \langle T_{t,s,s'}(\hat{h} - h), \hat{h} - h \rangle_{L^2} \, ds \, ds' \, dt$$

$$= \langle C(\hat{h} - h), \hat{h} - h \rangle_{L^2} = \|\hat{h} - h\|_C^2.$$

Thus, the excess risk can be expressed as sort of a weighted L^2 norm, where the operator C defines the weights, which is composed of the kernel and the distribution of X_{n+1}.

A.2. Re-expressing the estimator

In this section we define an alternative form for the estimator $\hat{g}$, which was given in Section 3. In particular, instead of using the reproducing property, we will write down the estimator using operators. To do this, we will take derivatives of $RSS_\lambda(g)$ with respect to g. Since these are functions, we mean the Fréchet derivative or strong derivative. Note that $RSS_\lambda(g)$ is a convex differentiable functional over $\mathbb{K}$. However, so that we are working with L^2 instead of $\mathbb{K}$, we use $\widetilde{RSS}_\lambda(h) := RSS_\lambda(\mathcal{L}_k^{1/2} h)$, where $h = \mathcal{L}_k^{-1/2} g$:

$$\widetilde{RSS}_\lambda(h) = \sum_{i=1}^n \int_0^1 \left(Y_i(t) - \int_0^1 \langle h, k_{t,s,X_i(s)}^{1/2} \rangle_{L^2} ds \right)^2 dt + \lambda \|h\|_{L^2}^2.$$

Now $\widetilde{RSS}_\lambda(h)$ is a convex differentiable functional over L^2. Thus, when taking the derivative, we are using the topology of L^2 not $\mathbb{K}$.

To take the derivative of $\widetilde{RSS}_\lambda(h)$ we first focus on the penalty, which is easier. We have that

$$\frac{\partial}{\partial h} \|h\|_{L^2}^2 = 2h.$$

Turning to the first term in $\widetilde{RSS}_\lambda(h)$ we first define the empirical quantities

$$T_{i;t,s,s'} = k_{t,s,X_i(s)}^{1/2} \otimes k_{t,s',X_i(s')}^{1/2}$$

and

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \int_0^1 \int_0^1 T_{i;t,s,s'} \, ds \, ds' \, dt. \tag{7}$$

Now we can apply a chain rule to obtain

$$\frac{\partial}{\partial h} \left[\frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left(Y_i(t) - \int_0^1 \langle h, k_{t,s,X_i(s)}^{1/2} \rangle_{L^2} ds \right)^2 dt \right]$$

$$= -\frac{2}{n} \sum_{i=1}^{n} \int_0^1 \int_0^1 Y_i(t) k_{t,s,X_i(s)}^{\frac{1}{2}} ds \, dt + 2C_n h.$$

For notational simplicity, define

$$\Gamma_{k^{1/2},Y} = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \int_0^1 Y_i(t) k_{t,s,X_i(s)}^{\frac{1}{2}} ds \, dt.$$

So, we finally have that

$$\frac{\partial}{\partial h} \widetilde{RSS}_\lambda(h) = -2\Gamma_{k^{1/2},Y} + 2C_n h + 2\lambda h,$$

which yields the estimate

$$\hat{h} = (C_n + \lambda I)^{-1} \Gamma_{k^{1/2},Y}, \tag{8}$$

where I is the identity operator.

A.3. Proof of Theorem 1 - controlling bias

Using Assumption 1 we can express

$$Y_i(t) = \int \langle k_{t,s,X_i(s)}^{1/2}, h \rangle_{L^2} + \varepsilon_i(t).$$

and we therefore have that

$$\Gamma_{k^{1/2},Y} = C_n(h) + f_n$$

where

$$f_n = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \int_0^1 \epsilon_i(t) k_{t,s,X_i(s)}^{\frac{1}{2}} ds \, dt.$$

This implies that $\hat{h}$ from (8) can be expressed as

$$\hat{h} = (C_n + \lambda I)^{-1} C_n(h) + (C_n + \lambda I)^{-1} f_n.$$

We introduce an intermediate quantity, h_λ, which is given by

$$h_\lambda = (C + \lambda I)^{-1} C(h),$$

where C is defined in (5). The difference between h_λ and h represents the bias of the estimator $\hat{h}$. Balancing this quantity with the variance, discussed in the next section, is called the *bias-variance trade off* a common term in nonparametric smoothing. Inherently, the idea is that to achieve an optimal $\hat{h}$ we have to balance both the bias and variance so that neither one is overly large.

Using the eigenfunctions of C as a basis, we can write

$$h = \sum_{k=1}^{\infty} a_k \phi_k.$$

Since C and $C + I$ have the same eigenfunctions, it follows that we can express

$$C + \lambda I = \sum_{k=1}^{\infty} (\lambda + \rho_k)(\phi_k \otimes \phi_k) \implies (C + \lambda I)^{-1} = \sum_{k=1}^{\infty} (\lambda + \rho_k)^{-1}(\phi_k \otimes \phi_k).$$

So we have that h_λ can be expressed as

$$h_\lambda = (C + \lambda I)^{-1} C(h) = \sum_{k=1}^{\infty} \frac{a_k \rho_k}{\lambda + \rho_k} \phi_k.$$

So the difference, $h_\lambda - h$ can be written as

$$h_\lambda - h = -\sum_{k=1}^{\infty} \frac{\lambda a_k}{\lambda + \rho_k} \phi_k. \tag{9}$$

The bias is therefore given by

$$\|h_\lambda - h\|_C^2 = \sum_{k=1}^{\infty} \frac{\lambda^2 a_k^2 \rho_k}{(\lambda + \rho_k)^2} \le \lambda^2 \max_{k \ge 1} \frac{\rho_k}{(\lambda + \rho_k)^2} \sum_{k=1}^{\infty} a_k^2 = \lambda^2 \|h\|_{L^2}^2 \max_{k \ge 1} \frac{\rho_k}{(\lambda + \rho_k)^2}.$$

It is easy to verify that the maximum of $F(x) = x/(\lambda + x)^2$ is achieved at $x = \lambda$ with the maximum value being $\frac{1}{4\lambda}$. We can therefore bound the bias as

$$\|h_\lambda - h\|_C^2 \le \frac{\lambda \|h\|_{L^2}^2}{4}.$$

In the statement of Theorem 1 we assume that $\lambda \asymp n^{2r/(2r+1)}$, which implies that the bias is of the order $n^{-\frac{2r}{2r+1}} O(1)$. We will show in the next section that the variance of our estimate achieves the same order.

A.4. Proof of Theorem 1 - controlling variability

Controlling the variability of the estimates, $\|\hat{h} - h_\lambda\|_C$ follows similar arguments as controlling the bias. However, there are many more terms which must be analyzed separately. In particular, we decompose $\hat{h} - h_\lambda$ into five separate components:

$$\hat{h} - h_\lambda = T_1 + T_2 + T_3 + T_4 + T_5, \tag{10}$$

where the T_i terms are given by

$$
\begin{aligned}
T_1 &= (C + \lambda I)^{-1} C(h_\lambda - h), \\
T_2 &= \lambda (C + \lambda I)^{-2} C(h), \\
T_3 &= -(C + \lambda I)^{-1} f_n, \\
T_4 &= (C + \lambda I)^{-1} (C_n - C)(h_\lambda - h), \\
T_5 &= (C + \lambda I)^{-1} (C - C_n)(h_\lambda - \hat{h}).
\end{aligned}
$$

While a bit tedious, it only requires linear algebra and repeated calls to the definitions of $\hat{h}$ and h_λ to verify (10), we thus omit the details here. We now develop bounds for each term, $\|T_i\|_C$, separately. For the first four, it turns out to be convenient to bound $\|C^\nu T_i\|_{L^2}$ for $0 < \nu \le 1/2$, as these bounds will be needed for the final term T_5. Notice that when $\nu = 1/2$ we have $\|C^\nu T_i\|_{L^2} = \|T_i\|_C$.

1. Using the eigenfunctions of C to express $h_\lambda - h$ as in (9), we get that

$$T_1 = -\sum_{k=1}^\infty \frac{\lambda a_k \rho_k}{(\lambda + \rho_k)^2} \phi_k.$$

We then have that

$$\|C^\nu T_1\|_{L^2}^2 = \sum_{k=1}^\infty \frac{\lambda^2 a_k^2 \rho_k^{2(1+\nu)}}{(\lambda + \rho_k)^4} \le \lambda^2 \max_{k \ge 1} \frac{\rho_k^{2(1+\nu)}}{(\lambda + \rho_k)^4} \|h\|_{L^2}^2.$$

Again, it is a basic calculus exercise to show that

$$\max_{k \ge 1} \frac{\rho_k^{2(1+\nu)}}{(\lambda + \rho_k)^4} \le \frac{\left(\lambda \frac{1+\nu}{1-\nu}\right)^{2(1+\nu)}}{\left(\lambda + \lambda \frac{1+\nu}{1-\nu}\right)^4} = \frac{(1-\nu)^{2(1-\nu)}(1+\nu)^{2(1+\nu)}}{16} \frac{1}{\lambda^{2-2\nu}}.$$

We thus have the bound

$$\|C^\nu T_1\|_{L^2}^2 \le c\lambda^{2\nu} \|h\|_{L^2}^2, \tag{11}$$

where c is a constant that depends only on ν.

2. Using the same arguments as in the previous step, we have that

$$\|C^\nu T_2\|_{L^2}^2 = \sum_{k=1}^\infty \frac{\lambda^2 a_k^2 \rho_k^{2(1+\nu)}}{(\lambda + \rho_k)^4} \leq \lambda^2 \max_{k \geq 1} \frac{\rho_k^{2(1+\nu)}}{(\lambda + \rho_k)^4} \sum_{k=1}^\infty a_k^2 \leq c\lambda^{2\nu} \|h\|_{L^2}^2.$$

(12)

3. Turning to T_3, we apply Lemma 1 with $0 < \nu \leq 1/2$ to obtain

$$\|C^\nu T_3\|_{L^2}^2 = \|C^\nu (C + \lambda I)^{-1} f_n\|_{L^2} = \frac{1}{n\lambda^{1-2\nu+1/2r}} O_p(1),$$

where r is defined as in Assumption 1. By the statement of Theorem 1 it follows that $n\lambda^{1+\frac{1}{2r}}$ tends to a nonzero constant, meaning that

$$\frac{1}{n\lambda^{1-2\nu+\frac{1}{2r}}} \asymp \lambda^{2\nu} \to 0,$$

(13)

since $\lambda \to 0$. Thus we have that $\|C^\nu T_3\|_{L^2} = O_p(\lambda^\nu)$.

4. To bound T_4 we first fix a second value $\nu > \nu_2 > 0$ that satisfies $2r(1 - 2\nu_2) > 1$, or equivalently $\nu_2 < (2r-1)/4r$, as well as $4r(2\nu_2 + 2\nu) > 1$, which is possible as long as $r > 1/2$ (Assumption 1). We now apply a basic operator inequality

$$\|C^\nu T_4\|_{L^2} = \|C^\nu (C + \lambda I)^{-1}(C_n - C)(h_\lambda - h)\|_{L^2}$$
$$\leq \|C^\nu (C + \lambda I)^{-1}(C_n - C)C^{-\nu_2}\|_{op} \|C^{\nu_2}(h_\lambda - h)\|_{L^2}.$$

and then apply Lemmas 3 and 4 to obtain

$$\|C^\nu T_4\|_{L^2}^2 \leq O_p\left(\left(n\lambda^{1-2\nu+\frac{1}{2r}}\right)^{-1}\right) O_p\left(\lambda^{2\nu_2}\right) = \frac{1}{n\lambda^{1-2\nu+\frac{1}{2r}}} o_p(1), \quad (14)$$

since $\nu_2 > 0$ and $\lambda \to 0$. Using (13) we conclude that $\|C^\nu T_4\|_{L^2} = O_p(\lambda^\nu)$.

5. The last term is the most involved to bound and the reason why the previous four bounds involved $C^\nu T_i$. We begin by expressing

$$\|T_5\|_C = \|C^{\frac{1}{2}}(C + \lambda I)^{-1}(C - C_n)(h_\lambda - \hat{h})\|_{L^2}$$
$$\leq \|C^{\frac{1}{2}}(C + \lambda I)^{-1}(C - C_n)C^{-\nu}\|_{op} \|C^\nu(h_\lambda - \hat{h})\|_{L^2}.$$

Here $\nu > 0$ is chosen to satisfy $2r(1 - 2\nu) > 1$. Applying Lemma 3, we have that

$$\|T_5\|_C^2 \leq \frac{1}{n\lambda^{1/2r}} O_p(1) \|C^\nu(h - \hat{h})\|_{L^2}^2.$$

We have now, in some sense, looped back and are dealing with the term $h - \hat{h}$. Using (10) we have

$$\|C^\nu(h - \hat{h})\|_{L^2} \leq \|C^\nu T_1\|_{L^2} + \|C^\nu T_2\|_{L^2} + \|C^\nu T_3\|_{L^2} + \|C^\nu T_4\|_{L^2}$$
$$+ \|C^\nu T_5\|_{L^2}.$$

(15)

The first four terms we already have bounds for, so we need only focus on the last, which again, has looped back to our original term. We now apply Lemma 2 to obtain

$$\|C^\nu T_5\|_{L^2}^2 = \|C^\nu(C+\lambda I)^{-1}(C-C_n)(h-\hat{h})\|_{L^2}^2 \le \frac{1}{n\lambda^{1-2\nu+\frac{1}{2r}}}\|C^\nu(h-\hat{h})\|_{L^2}^2.$$

Combining the above with (15) we have that

$$\|C^\nu(h-\hat{h})\|_{L^2}\left(1 - \frac{1}{n\lambda^{1-2\nu+\frac{1}{2r}}}\right) \le \|C^\nu T_1\|_{L^2} + \|C^\nu T_2\|_{L^2} + \|C^\nu T_3\|_{L^2}$$
$$+ \|C^\nu T_4\|_{L^2}.$$

Using (13) it thus follows that

$$\|C^\nu(h-\hat{h})\|_{L^2} = O_p(1)(\|C^\nu T_1\|_{L^2} + \|C^\nu T_2\|_{L^2} + \|C^\nu T_3\|_{L^2} + \|C^\nu T_4\|_{L^2}),$$

and applying steps 1-4 we get that

$$\|C^\nu(h-\hat{h})\|_{L^2} = O_p(\lambda^\nu) = o_p(1),$$

and we finally have that

$$\|T_5\|_C = \frac{1}{n\lambda^{\frac{1}{2r}}}o_p(1).$$

We can now combine Steps 1-4, taking $\nu = 1/2$, with step 5 to finally conclude that

$$\Re_n^2 = \|\hat{h} - h\|_C^2 = \lambda^2 O_p(1) = n^{-\frac{2r}{2r+1}}O_p(1).$$

Combined with the results of Section A.3, this concludes the proof.

A.5. Auxiliary lemmas

Here we state four lemmas which are generalizations of ones used in Cai and Yuan [3] and Wang and Ruppert [32].

Lemma 1. *If Assumption 1 holds then for any $0 \le \nu \le \frac{1}{2}$*

$$\|C^\nu(C+\lambda I)^{-1}f_n\|_{L^2} = O_p\left(\left(n\lambda^{1-2\nu+\frac{1}{2r}}\right)^{-\frac{1}{2}}\right).$$

Lemma 2. *Let Assumption 1 hold. Then for any $\nu > 0$ such that $2r(1-2\nu) > 1$, we have that*

$$\|C^\nu(C+\lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op} = O_p\left(\left(n\lambda^{1-2\nu+\frac{1}{2r}}\right)^{-\frac{1}{2}}\right),$$

where $\|.\|_{op}$ represents the usual operator norm i.e., $\|A\|_{op} = \sup_{h:\|h\|_{L^2}=1}\|Ah\|$.

Lemma 3. *Let Assumption 1 hold and fix $0 < \nu < \nu_2$ to be any two values that satisfy $2r(1 - 2\nu) > 1$ and $4r(\nu_2 + \nu) > 1$, then we have that*

$$\|C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op} = O_p\left(\left(n\lambda^{1-2\nu_2+\frac{1}{2r}}\right)^{-\frac{1}{2}}\right).$$

Lemma 4. *(3, Lemma 1) For any $0 < \nu < 1$,*

$$\|C^{\nu}(h_\lambda - h)\|_{L^2} \leq (1 - \nu)^{1-\nu}\nu^{\nu}\lambda^{\nu}\|h\|_{L^2}.$$

Lemma 5. *Fix $\nu > 0$ and $\nu_2 > 0$ such that $4r(\nu_2 + \nu) > 1$. 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*

$$c_4\lambda^{\frac{-1}{2r}-1+2\nu_2} \leq \sum_{j=1}^{\infty} \frac{s_j^{2\nu_2+2\nu}}{(\lambda + s_j)^{1+2\nu}} \leq c_3(1 + \lambda^{\frac{-1}{2r}-1+2\nu_2}).$$

Proof of Lemma 1

Recall that

$$f_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^1\int_0^1 \epsilon_i(t)k_{t,s,X_i(s)}^{\frac{1}{2}}dsdt.$$

Using Parseval's identity we have that

$$\|C^{\nu}(C + \lambda I)^{-1}f_n\|_{L^2}^2 = \sum_{k=1}^{\infty}\frac{\rho_k^{2\nu}}{(\lambda + \rho_k)^2}\langle f_n, \phi_k\rangle^2.$$

Taking expected values yields

$$\mathrm{E}\,\|C^{\nu}(C + \lambda I)^{-1}f_n\|_{L^2}^2$$

$$= \frac{1}{n}\sum_{k=1}^{\infty}\frac{\rho_k^{2\nu}}{(\lambda + \rho_k)^2}\,\mathrm{E}\left(\int_0^1\int_0^1 \epsilon(t)\langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k\rangle_{L^2}dsdt\right)^2.$$

By Jensen's inequality we have

$$\left(\int_0^1\int_0^1 \epsilon(t)\langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k\rangle_{L^2}dsdt\right)^2$$

$$\leq \int_0^1\left(\int_0^1 \epsilon(t)\langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k\rangle_{L^2}ds\right)^2 dt$$

$$= \int_0^1\int_0^1\int_0^1 \epsilon^2(t)\langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k\rangle_{L^2}\langle k_{t,s^*,X(s^*)}^{\frac{1}{2}}, \phi_k\rangle_{L^2}dsds^*dt.$$

Using the assumed independence between ϵ and X, as well as the assumption that $\mathrm{E}(\epsilon^2(t)) \leq M$, where M is a constant, we obtain

$$
\mathrm{E}\left(\int_0^1 \int_0^1 \epsilon(t) \langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k \rangle_{L^2} ds dt \right)^2
$$

$$
\leq M\,\mathrm{E}\left(\int_0^1 \int_0^1 \int_0^1 \langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_k \rangle_{L^2} \langle k_{t,s^*,X(s^*)}^{\frac{1}{2}}, \phi_k \rangle_{L^2} ds ds^* dt \right)
$$

$$
= M \langle C(\phi_k), \phi_k \rangle = M \rho_k.
$$

Since $0 \leq \nu \leq \frac{1}{2}$ and both ρ_k and λ are positive, we can obtain the bound

$$
\mathrm{E}\,\|C^\nu (C + \lambda I)^{-1} f_n\|_{L^2}^2 \leq \frac{M}{n} \sum_{k=1}^\infty \frac{\rho_k^{2\nu+1}}{(\lambda + \rho_k)^2} \leq \frac{M}{n\lambda^{1-2\nu}} \sum_{k=1}^\infty \frac{\rho_k^{2\nu+1}}{(\lambda + \rho_k)^{1+2\nu}}.
$$

Now we apply Lemma 5 with $\nu_2 = 1/2$ to obtain

$$
\mathrm{E}\,\|C^\nu (C + \lambda I)^{-1} f_n\|_{L^2}^2 \leq \frac{c^*}{n\lambda^{1-2\nu+\frac{1}{2r}}},
$$

where c^* is a constant. An application of Markov's inequality completes the proof.

Proof of Lemma 2

By definition

$$
\|C^\nu (C + \lambda I)^{-1} (C_n - C) C^{-\nu}\|_{op}^2 = \sup_{f : \|f\|_{L^2} = 1} \|C^\nu (C + \lambda I)^{-1} (C_n - C) C^{-\nu} f\|_{L^2}^2.
$$

Fix $f \in L^2$ such that $\|f\|_{L^2} = 1$. We can expand f as

$$
f = \sum_{k=1}^\infty f_k \phi_k,
$$

By Parseval's identity we have

$$
\|C^\nu (C + \lambda I)^{-1} (C_n - C) C^{-\nu} f\|^2 = \sum_{j=1}^\infty \left[\frac{\rho_j^\nu}{\rho_j + \lambda} \sum_{k=1}^\infty f_k \rho_k^{-\nu} \langle (C_n - C) \phi_k, \phi_j \rangle_{L^2} \right]^2.
$$

Applying the Cauchy-Schwartz inequality and using the fact that $\|f\|_{L^2} = 1$ we have that

$$
\sum_{k=1}^\infty f_k \rho_k^{-\nu} \langle (C_n - C) \phi_k, \phi_j \rangle_{L^2} \leq \left(\sum_{k=1}^\infty \rho_k^{-2\nu} \langle (C_n - C) \phi_k, \phi_j \rangle_{L^2}^2 \right)^{1/2}.
$$

So we can bound the operator norm as

$$\|C^\nu(C+\lambda I)^{-1}(C_n-C)C^{-\nu}\|_{op}^2 \leq \sum_{k=1}^\infty \sum_{j=1}^\infty \frac{\rho_k^{-2\nu}\rho_j^{2\nu}}{(\lambda+\rho_j)^2}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2.$$

Applying Jensen's equality we get that

$$\mathrm{E}\left(\sum_{k=1}^\infty \sum_{j=1}^\infty \frac{\rho_k^{-2\nu}\rho_j^{2\nu}}{(\lambda+\rho_j)^2}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2\right)^{\frac{1}{2}}$$

$$\leq \left(\sum_{k=1}^\infty \sum_{j=1}^\infty \frac{\rho_k^{-2\nu}\rho_j^{2\nu}}{(\lambda+\rho_j)^2}\mathrm{E}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2\right)^{\frac{1}{2}}.$$

Using the definition of C_n from (7) we have that

$$\mathrm{E}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2 \leq \mathrm{E}\langle\phi_j,C_n\phi_k\rangle_{L^2}^2$$

$$= \frac{1}{n}\mathrm{E}\left(\int_0^1\int_0^1\int_0^1\langle k_{t,s,X(s)}^{\frac{1}{2}},\phi_j\rangle\langle k_{t,s^*,X(s^*)}^{\frac{1}{2}},\phi_k\rangle dsds^*dt\right)^2.$$

Note the first inequality follows from the fact that C is the mean of C_n and thus replacing C above with any other quantity cannot decrease it (since it is minimized when using C). One can show this using basic calculus arguments over Hilbert spaces, thus we omit the details here. By applying Cauchy-Schwartz inequality and Fubini's theorem we have

$$\mathrm{E}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2$$

$$\leq \frac{1}{n}\mathrm{E}\left(\left[\int_0^1\left(\int_0^1\langle k_{t,s,X(s)}^{\frac{1}{2}},\phi_j\rangle ds\right)^2 dt\right]\left[\int_0^1\left(\int_0^1\langle k_{t^*,s^*,X(s^*)}^{\frac{1}{2}},\phi_k\rangle ds^*\right)^2 dt^*\right]\right)$$

$$= \frac{1}{n}\int_0^1\int_0^1\mathrm{E}\left[\left(\int_0^1\langle k_{t,s,X(s)}^{\frac{1}{2}},\phi_j\rangle ds\right)^2\left(\int_0^1\langle k_{t^*,s^*,X(s^*)}^{\frac{1}{2}},\phi_k\rangle ds^*\right)^2\right]dtdt^*.$$

Using Cauchy-Schwartz inequality again

$$\mathrm{E}\langle\phi_j,(C_n-C)\phi_k\rangle_{L^2}^2$$

$$\leq \frac{1}{n}\int_0^1\int_0^1\mathrm{E}^{\frac{1}{2}}\left(\int_0^1\langle k_{t,s,X(s)}^{\frac{1}{2}},\phi_j\rangle ds\right)^4\mathrm{E}^{\frac{1}{2}}\left(\int_0^1\langle k_{t^*,s^*,X(s^*)}^{\frac{1}{2}},\phi_k\rangle ds^*\right)^4 dtdt^*.$$

Note that we can move to the $\mathbb{K}$ inner product to obtain:

$$\langle k_{t,s,X(s)}^{\frac{1}{2}},\phi_k\rangle_{L^2} = \langle k_{t,s,X(s)},\mathcal{L}_k^{1/2}\phi_k\rangle_{\mathbb{K}} = (\mathcal{L}_k^{1/2}\phi_k)(t,s,X(s))$$

and $\mathcal{L}^{1/2}\phi_k$ is a function in $\mathbb{K}$, thus we can apply Assumption 1.4 to obtain

$$
\mathrm{E}\langle\phi_j, (C_n - C)\phi_k\rangle_{L^2}^2 \leq \frac{c}{n}\, \mathrm{E}\left[\int_0^1\left(\int_0^1 \langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_j\rangle ds\right)^2 dt\right]
$$
$$
\times \mathrm{E}\left[\int_0^1\left(\int_0^1 \langle k_{t^*,s^*,X(s^*)}^{\frac{1}{2}}, \phi_k\rangle ds^*\right)^2 dt^*\right].
$$

It is easy to see that

$$
\mathrm{E}\left[\int_0^1\left(\int_0^1 \langle k_{t,s,X(s)}^{\frac{1}{2}}, \phi_j\rangle ds\right)^2 dt\right] = \rho_j.
$$

Now we obtain

$$
\mathrm{E}\langle\phi_j, (C_n - C)\phi_k\rangle_{L^2}^2 \leq cn^{-1}\rho_j\rho_k.
$$

Therefore,

$$
\mathrm{E}\left(\sum_{k=1}^\infty\sum_{j=1}^\infty \frac{\rho_k^{-2\nu}\rho_j^{2\nu}}{(\lambda + \rho_j)^2}\langle\phi_j, (C_n - C)\phi_k\rangle_{L^2}\right)^{\frac{1}{2}} \leq \left(\frac{c}{n}\sum_{k=1}^\infty\sum_{j=1}^\infty \frac{\rho_k^{1-2\nu}\rho_j^{1+2\nu}}{(\lambda + \rho_j)^2}\right)^{\frac{1}{2}}.
$$

Note that

$$
\sum_{k=1}^\infty\sum_{j=1}^\infty \frac{\rho_k^{1-2\nu}\rho_j^{1+2\nu}}{(\lambda + \rho_j)^2} = \sum_{k=1}^\infty \rho_k^{1-2\nu}\sum_{j=1}^\infty \frac{\rho_j^{1+2\nu}}{(\lambda + \rho_j)^2}.
$$

Since $2r(1 - 2\nu) > 1$ and $\rho_k < c_2 k^{-2r}$ we have

$$
\sum_{k=1}^\infty \rho_k^{1-2\nu} \leq c_2 \sum_{k=1}^\infty k^{-2r(1-2\nu)} = c^{**} < \infty.
$$

Finally, by applying Lemma 5 with $\nu_2 = 1/2$ we obtain

$$
\mathrm{E}\,\|C^\nu(C + \lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op} \leq \gamma(n\lambda^{1-2\nu+\frac{1}{2r}})^{-\frac{1}{2}},
$$

An application of Markov's inequality completes the proof.

Proof of Lemma 3

Recall that

$$
\|C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op}^2
$$
$$
= \sup_{h:\|h\|_{L^2}=1}\|C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}h\|_{L^2}^2.
$$

Note that

$$C^{\nu_2}(C + \lambda I)^{-1} = \sum_{j=1}^{\infty} \frac{\rho_j^{\nu_2}}{\rho_j + \lambda}(\phi_j \otimes \phi_j),$$

and recall that from the proof of Lemma 2,

$$(C_n - C)C^{-\nu}h = \sum_{k=1}^{\infty} a_k \rho_k^{-\nu}(C_n - C)\phi_k.$$

Therefore

$$C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}h = \sum_{j=1}^{\infty} \frac{\rho_j^{\nu_2}}{\rho_j + \lambda}\langle \sum_{k=1}^{\infty} a_k \rho_k^{-\nu}(C_n - C)\phi_k, \phi_j \rangle_{L^2}\phi_j.$$

By Parseval's identity we obtain

$$\|C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op}^2 = \sum_{j=1}^{\infty} \left[\frac{\rho_j^{\nu_2}}{\rho_j + \lambda}\langle \sum_{k=1}^{\infty} a_k \rho_k^{-\nu}(C_n - C)\phi_k, \phi_j \rangle_{L^2} \right]^2$$

By Cauchy-Schwartz inequality and using the same steps in the proof of Lemma 2, we have

$$\|C^{\nu_2}(C + \lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op}^2 \leq \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\rho_k^{-2\nu}\rho_j^{2\nu_2}}{(\lambda + \rho_j)^2}\langle \phi_j, (C_n - C)\phi_k \rangle_{L^2}^2.$$

Using the same arguments in the proof of Lemma 2, we obtain

$$\mathrm{E}\left(\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\rho_k^{-2\nu}\rho_j^{2\nu_2}}{(\lambda + \rho_j)^2}\langle \phi_j, (C_n - C)\phi_k \rangle_{L^2} \right)^{\frac{1}{2}} \leq \left(\frac{c}{n}\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\rho_k^{1-2\nu}\rho_j^{1+2\nu_2}}{(\lambda + \rho_j)^2} \right)^{\frac{1}{2}}.$$

Note that

$$\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\rho_k^{1-2\nu}\rho_j^{1+2\nu_2}}{(\lambda + \rho_j)^2} = \sum_{k=1}^{\infty}\rho_k^{1-2\nu}\sum_{j=1}^{\infty} \frac{\rho_j^{1+2\nu_2}}{(\lambda + \rho_j)^2}.$$

Note that the condition $2r(1 - 2\nu) > 1$ implies $\nu < \frac{1}{2} - \frac{1}{2r} < \frac{1}{2}$. We therefore have

$$\left(\frac{\lambda + \rho_j}{\rho_j} \right)^{2\nu - 1} \leq 1.$$

It follows that

$$\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{\rho_k^{1-2\nu}\rho_j^{1+2\nu_2}}{(\lambda + \rho_j)^2} \leq \sum_{k=1}^{\infty}\rho_k^{1-2\nu}\sum_{j=1}^{\infty} \frac{\rho_j^{2\nu+2\nu_2}}{(\lambda + \rho_j)^{2\nu+1}}.$$

Recall that

$$\sum_{k=1}^{\infty} \rho_k^{1-2\nu} \le c_2 \sum_{k=1}^{\infty} k^{-2r(1-2\nu)} = c^{**} < \infty.$$

Therefore

$$\mathrm{E}\, \|C^{\nu_2}(C+\lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op} \le \left(\frac{cc^{**}}{n} \sum_{j=1}^{\infty} \frac{\rho_j^{2\nu+2\nu_2}}{(\lambda+\rho_j)^{2\nu+1}} \right)^{\frac{1}{2}}.$$

By applying Lemma 5 we have that

$$\mathrm{E}\, \|C^{\frac{1}{2}}(C+\lambda I)^{-1}(C_n - C)C^{-\nu}\|_{op} \le \beta(n\lambda^{\frac{1}{2r}+1-2\nu_2})^{-\frac{1}{2}},$$

where β is a constant. An application of the Markov inequality completes the proof.

A.6. Proof of Lemma 5

Using the same arguments as in [3], we get that

$$
\begin{aligned}
\sum_{j=1}^{\infty} \frac{s_j^{2\nu_2+2\nu}}{(\lambda+s_j)^{1+2\nu}} &= \sum_{j=1}^{\infty} \frac{s_j^{1+2\nu}}{(\lambda+s_j)^{1+2\nu}} s_j^{2\nu_2-1} \\
&\le \sum_{j=1}^{\infty} \frac{c_1^{1+2\nu} k^{-2r(1+2\nu)}}{(\lambda+c_2 k^{-2r})^{1+2\nu}} k^{-2r(2\nu_2-1)} \\
&= c_1^{1+2\nu} \sum_{j=1}^{\infty} \frac{k^{-2r(2\nu_2-1)}}{(\lambda k^{2r}+c_2)^{1+2\nu}} \\
&\le c_1^{1+2\nu} \left(\frac{1}{c_2} + \int_1^{\infty} \frac{x^{-2r(2\nu_2-1)}}{(\lambda x^{2r}+c_2)^{1+2\nu}}\, dx \right) \\
&= c_1^{1+2\nu} \left(\frac{1}{c_2} + \lambda^{2\nu_2-1-1/2r} \int_{\lambda^{1/2r}}^{\infty} \frac{y^{-2r(2\nu_2-1)}}{(y^{2r}+c_2)^{1+2\nu}}\, dy \right).
\end{aligned}
$$

For the integral to be finite, it is enough if $2r(2\nu_2+2\nu) \ge 1+\delta$, for some $\delta > 0$, as the integrand will go to zero faster than $y^{-(1+\delta)}$. The argument for the lower bound follows the same arguments.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert spaces in Probability and Statistics*. Springer Science & Business Media, 2011. MR2239907

[2] T. T. Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5):2330–2355, 2011. MR2906870

[3] T. T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012. MR3010906

[4] P. Du and X. Wang. Penalized likelihood functional regression. *Statistica Sinica*, pages 1017–1041, 2014. MR3236451

[5] Y. Fan, G. M. James, and P. Radchenko. Functional additive regression. *The Annals of Statistics*, 43(5):2296–2325, 2015. MR3396986

[6] R. Gabrys, L. Horváth, and P. Kokoszka. Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105(491):1113–1125, 2010. MR2752607

[7] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Wiley Online Library, 1990. MR1082147

[8] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*, volume 200. Springer, 2012. MR2920735

[9] L. Horváth, P. Kokoszka, and G. Rice. Testing stationarity of functional time series. *Journal of Econometrics*, 179(1):66–82, 2014. MR3153649

[10] G. M. James and B. W. Silverman. Functional adaptive model estimation. *Journal of the American Statistical Association*, 100(470):565–576, 2005. MR2160560

[11] H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: A functional RKHS approach. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 374–380, 2010.

[12] R. A. Kennedy and P. Sadeghi. *Hilbert Space Methods in Signal Processing*. Cambridge University Press, 2013. MR3052450

[13] J. S. Kim, A.-M. Staicu, A. Maity, R. J. Carroll, and D. Ruppert. Additive function-on-function regression. *Journal of Computational and Graphical Statistics*, 27:234–244, 2018. MR3788315

[14] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971. MR0290013

[15] P. Kokoszka and M. Reimherr. Predictability of shapes of intraday price curves. *The Econometrics Journal*, 16(3):285–308, 2013. MR3146768

[16] P. Kokoszka and M. Reimherr. *Introduction to Functional Data Analysis*. CRC Press, 2017. MR3793167

[17] Y. Li, T. Hsing, et al. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010. MR2766854

[18] H. Lian. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics*, 35(4):597–606, 2007. MR2381399

[19] H. Ma and Z. Zhu. Continuously dynamic additive models for functional

data. *Journal of Multivariate Analysis*, 150:1–13, 2016. MR3534898

[20] M. W. McLean, G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269, 2014. MR3173770

[21] J. S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

[22] H.-G. Müller, Y. Wu, and F. Yao. Continuously additive models for nonlinear functional regression. *Biometrika*, pages 607–622, 2013. MR3094440

[23] J. G. Nicholas. *Market Neutral Investing*. Bloomberg Press Princeton, NJ, 2000.

[24] A. Parodi and M. Reimherr. FLAME: Simultaneous variable selection and smoothing for function-on-scalar regression. Technical report, Pennsylvania State University, 2017.

[25] J. Petrovich, M. Reimherr, and C. Daymont. Functional regression models with highly irregular designs. Technical report, Pennsylvania State University, 2018. (https://arxiv.org/abs/1805.08518).

[26] C. Preda. Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137(3):829–840, 2007. MR2301719

[27] J. O. Ramsay and B. Silverman. *Functional Data Analysis*. Wiley Online Library, 2006. MR2168993

[28] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB*. Springer Science & Business Media, 2009. MR3645102

[29] F. Scheipl, A.-M. Staicu, and S. Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015. MR3357391

[30] X. Sun, P. Du, X. Wang, and P. Ma. Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *Journal of the American Statistical Association*, page Accepted, 2017.

[31] K. G. van den Boogaart. tensorA: Advanced tensors arithmetic with named indices. *R package version 0.31*, 2007. http://CRAN.R-project.org/package=tensorA.

[32] X. Wang and D. Ruppert. Optimal prediction in an additive functional model. *Statistica Sinica*, 25:567–589, 2015. MR3379089

[33] L. Xiao, C. Li, W. Checkley, and C. Crainiceanu. Fast covariance estimation for sparse functional data. *Statistics and Computing*, pages 1–12, 2017. MR3761337

[34] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005. MR2160561

[35] X. Zhang, J.-L. Wang, et al. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016. MR3546451

[36] H. Zhu, F. Yao, and H. H. Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):581–603, 2014. MR3210729