Accepted Manuscript

Reduced rank modeling for functional regression with functional responses

Hongmei Lin, Xuejun Jiang, Heng Lian, Weiping Zhang

PII: S0047-259X(18)30460-3
DOI: https://doi.org/10.1016/j.jmva.2018.09.004
Reference: YJMVA 4414

To appear in: *Journal of Multivariate Analysis*

Received date : 6 September 2018

Please cite this article as: H. Lin, et al., Reduced rank modeling for functional regression with functional responses, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.09.004

Reduced rank modeling for functional regression with functional responses

Hongmei Lin[a], Xuejun Jiang[b], Heng Lian[c,], Weiping Zhang[d]

[a]School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China
[b]Department of Mathematics, Southern University of Science and Technology, Shenzhen, China
[c]Department of Mathematics, City University Hong Kong, Hong Kong
[d]Department of Statistics and Finance, University of Science and Technology of China, Hefei, China, 230026

Abstract

This article considers regression problems where both the predictor and the response are functional in nature. Driven by the desire to build a parsimonious model, we consider functional reduced rank regression in the framework of reproducing kernel Hilbert spaces, which can be formulated in the form of linear factor regression with estimated multivariate factors, and achieves dimension reduction in both the predictor and the response spaces. The convergence rate of the estimator is derived. Simulations and real data sets are used to demonstrate the competitive performance of the proposed method.

Keywords: Dimension reduction, Functional data, Functional response, Reproducing kernel Hilbert space.

1. Introduction

It is increasingly common to deal with regression problems in which the predictor, the response or both are functional in nature, with recent contributions include but not limited to [2, 6, 9, 12, 13, 16, 17, 20, 23, 24, 27, 29]. In this article we consider the following functional linear regression model with functional response

$$Y(t) = \mu(t) + \int_0^1 \beta(t, s)X(s)\,ds + \epsilon(t), \tag{1}$$

where Y, X, $\epsilon \in L^2([0, 1])$ and $\mathrm{E}(\epsilon \mid X) = 0$. This problem has been studied in [1, 4, 10, 25, 27]. We assume that the entire functional predictor and response are observed. As the mean function can be trivially estimated in this situation, for simplicity and without loss of generality we assume $\mathrm{E}(Y) = \mathrm{E}(X) = 0$ and thus do not model the intercept $\mu(t)$.

With either scalar or functional responses, there are several approaches to fit the functional linear model given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The traditionally most popular one is to approximate functional variables via basis expansion. Polynomial splines were used in [25] whereas Fourier bases are more appropriate for periodic functional data. Wavelet bases in functional data analysis were popularized because of their suitability for modeling spatially heterogeneous functions [21, 22]. Random basis functions obtained from functional principal component analysis (PCA) are also used, with particular advantages in theoretical analysis [8, 13].

The framework adopted here is based on reproducing kernel Hilbert spaces (RKHS), which were studied in [7, 28], and extended in [18] to functional responses. Assuming that for all $t \in [0, 1]$, $\beta(t, \cdot)$ is in an RKHS $\mathcal{H}_K$ with kernel K, and denoting the L^2 norm and the RKHS norm by $\|\cdot\|$ and $\|\cdot\|_{\mathcal{H}_K}$, respectively, we can estimate β by

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} \left\| Y_i - \int_0^1 \beta(\cdot, s)X_i(s)\,ds \right\|^2 + \lambda \int_0^1 \|\beta(t, \cdot)\|^2_{\mathcal{H}_K}\,dt. \tag{2}$$

Email address: henglian@cityu.edu.hk (Heng Lian)

The single most important special case of RKHS is the second order Sobolev space in which

$$\int_0^1 \|\beta(t,\cdot)\|^2_{\mathcal{H}_K} dt = \int_0^1 \int_0^1 \frac{\partial^2}{\partial s^2}\beta(t,s)dsdt.$$

By Lemma 7.1 in Chapter 8 of [5], any $\beta \in L^2([0,1]^2)$ can be expressed as a converging infinite sum

$$\beta(t,s) = \sum_{j=1}^{\infty} a_j(t)b_j(s)$$

for some non-unique sequences $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ in $L^2([0,1])$. The proposed method here is to truncate this infinite sum at some small integer r which will be called the rank of the model. That is, we assume that for some $r < \infty$,

$$\beta(t,s) = \sum_{j=1}^{r} a_j(t)b_j(s). \tag{3}$$

This is an extension of the traditional reduced rank regression model [3, 15], which postulates that, for all $i \in \{1,\ldots,n\}$, $\mathbf{Y}_i = \mathbf{C}\mathbf{X}_i + \mathbf{E}_i$ with $\text{rank}(\mathbf{C}) \le r$, where $\mathbf{Y}_i \in \mathbb{R}^q$ is the vector of responses, $\mathbf{X}_i \in \mathbb{R}^p$ is the vector of predictors, $\mathbf{E}_i$ is the mean zero noise, and $\mathbf{C}$ is the $q \times p$ coefficient matrix. Since $\text{rank}(\mathbf{C}) \le r$, we can write $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$ with $\mathbf{A}$ and $\mathbf{B}$ being a $q \times r$ and a $p \times r$ matrix, respectively.

If $\mathbf{a}_1, \ldots, \mathbf{a}_r$ and $\mathbf{b}_1, \ldots, \mathbf{b}_r$ are the columns of $\mathbf{A}$ and $\mathbf{B}$, respectively, we can write $\mathbf{C} = \mathbf{a}_1\mathbf{b}_1^\top + \cdots + \mathbf{a}_r\mathbf{b}_r^\top$. Since $\mathbf{A}\mathbf{B}^\top = (\mathbf{A}\mathbf{G})(\mathbf{G}^{-1}\mathbf{B}^\top)$ for any nonsingular $r \times r$ matrix $\mathbf{G}$, it is easy to see that we can assume $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ without loss of generality. Similarly, for functional reduced rank regression, we can assume that the functions $a_1, \ldots, a_r$ are orthogonal to each other with unit norm in $L^2([0,1])$.

To see that this orthonormal property can be achieved for any $\beta(t,s)$, suppose that $\mathbf{A}$ is the $r \times r$ matrix with entries $\int a_j(t)a_{j'}(t)dt$. Let $\mathbf{a} = (a_1, \ldots, a_r)^\top$ and $\mathbf{b} = (b_1, \ldots, b_r)^\top$. Further define $\mathbf{a}' = (a_1', \ldots, a_r')^\top = \mathbf{A}^{-1/2}\mathbf{a}$. Then the matrix with entries $\int a_j'(t)a_{j'}'(t)dt$ is

$$\int_0^1 \mathbf{a}'(t)\{\mathbf{a}'(t)\}^\top dt = \int \mathbf{A}^{-1/2}\mathbf{a}(t)\mathbf{a}^\top(t)\mathbf{A}^{-1/2}dt = \mathbf{I}$$

and hence $a_1', \ldots, a_r'$ are orthonormal. We can then write $\beta(s,t) = a_1'(s)b_1'(t) + \cdots + a_r'(s)b_r'(t)$ with $(b_1', \ldots, b_r')^\top = \mathbf{A}^{1/2}\mathbf{b}$.

Assumption (3) induces a more parsimonious functional regression model and can hopefully increase estimation efficiency if the assumption is true or nearly so. Although theoretically the convergence rate obtained in our analysis does not indicate this, the simulation results demonstrate that this is indeed the case when the assumption is met. Thus our comparison is mainly through numerical studies.

For the estimator (2), Proposition 1 of [18] implies that given the data, the estimator has at most rank $r = n$, which is large and also increases with the sample size. In the current work, due to the imposed low-rank constraint, we estimate β differently from [18], with a_j and b_j estimated separately. The motivation for the low-rank assumption on β is to reduce complexity of the estimation problem (reduce the estimation of a bivariate function to the estimation of multiple univariate functions) and also to gain additional interpretability in modeling.

Reduced rank modeling can be understood from the point of view of dimension reduction. The conditional expectation of Y is

$$\int_0^1 \beta(\cdot,s)X(s)ds = \sum_{j=1}^{r}\left\{\int_0^1 b_j(s)X(s)ds\right\}a_j.$$

Thus the predictor X affects the response through r estimated "factors" $z_j = \int b_j(s)X(s)ds$. Furthermore, only the projection of Y onto the space spanned by $a_1, \ldots, a_r$ can be predictable through X. In this sense, functional reduced rank regression achieves dimension reduction both in the predictor and the response. More explicitly, if the functions $a_1, \ldots, a_r$ are orthonormal in $L^2([0,1])$, the model is equivalent to

$$\forall_{j\in\{1,\ldots,r\}}\ \langle Y, a_j\rangle = \langle X, b_j\rangle + \langle \epsilon, a_j\rangle, \quad \forall_{a\perp\{a_1,\ldots a_r\}}\ \langle Y, a\rangle = \langle \epsilon, a\rangle, \tag{4}$$

where we write $\langle f, g\rangle = \int f(s)g(s)ds$ for the inner product in L^2. Based on this representation, the fitted model is more easily interpretable than the general model (1). The projection of Y along the direction a_j is dependent on the projection of X along the direction b_j. One cannot interpret the model in this way when using the functional PCA approach where a_j, b_j are the eigenfunctions, since a_j (b_j) is obtained on the observed response (predictor) alone and is the direction that the response (predictor) varies the most.

Related to this work is [19], which studied functional reduced-rank regression from a Bayesian point of view. In that work, reduced-rank models with latent variables are considered where the latent variables serve as the unknown common factors relating X and Y. Our work is on reduced-rank regression with manifest variables, which is a direct extension of the classical reduced-rank regression of [3].

The rest of the paper is organized as follows. In Section 2, we discuss the estimation approach of the functional reduced rank regression model and obtain the convergence rate of the estimator. We further discuss the effect of discretization in implementation and theory. Section 3 contains our simulation studies and real data illustrations. We conclude the paper with a discussion in Section 4.

2. Estimation and convergence rate

2.1. Estimator definition and estimation in the population

First, we note that although we choose to work within the RKHS framework, we can similarly estimate the coefficient β with the reduced rank methodology using other approaches such as basis expansion. The main purpose of this work is to propose a new estimator under the reduced rank framework and to compare functional reduced rank regression with general functional linear regression in estimation efficiency and interpretability through numerical studies, instead of comparing RKHS-based estimation with basis-expansion-based estimation. Some comparisons between the RKHS estimator and the functional PCA estimator are made in [7, 18] for general functional linear regression (2).

Given a positive definite kernel $K(s,t)$, the induced RKHS $\mathcal{H}_K$ is the range of the operator $K^{1/2}$, where $K^{1/2}$ is the operator square root of

$$K : f \in L^2([0,1]) \to \int_0^1 K(\cdot, t)f(t)dt \in L^2([0,1]).$$

See, e.g., [26]. As mentioned in Section 3 of Chapter III in [11], we have $\|f\| = \|K^{1/2}f\|_{\mathcal{H}_K}$ for any $f \in L^2([0,1])$. Although we use K to denote both the bivariate kernel function and the associated operator, its meaning should be clear from the context.

Informally, one can think of $\mathcal{H}_K$ as containing functions that possess certain smoothness properties. We assume $b_j \in \mathcal{H}_K$. It is not necessary to impose any smoothness condition on a_j, which does not need to be regularized. Technically, this is because we take the inverse of the covariance operator of X, but not the inverse of the covariance operator of Y. Furthermore, without loss of generality, we assume that $a_1, \ldots, a_r$ form an orthonormal set of functions in $L^2([0,1])$. This will also allow us to derive closed form estimators of a_j, b_j.

In reduced rank regression, (2) reduces to

$$(\{\hat{a}_j\}, \{\hat{b}_j\}) = \underset{a_j \in L^2([0,1]), b_j \in \mathcal{H}_K}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{r} a_j \int_0^1 b_j(s)X_i(s)\,ds \right\|^2 + \lambda \sum_{j=1}^{r} \|b_j\|^2_{\mathcal{H}_K}. \tag{5}$$

In the population, the minimization problem is

$$\mathrm{E}\left\{ \left\| Y - \sum_{j=1}^{r} a_j \int_0^1 b_j(s)X(s)\,ds \right\|^2 \right\},$$

with the constraint $b_j \in \mathcal{H}_K$. Since $b_j \in \mathcal{H}_K$ is equivalent to $b_j = K^{1/2}f_j$ for some $f_j \in L^2([0,1])$, we can write the

above minimization problem in terms of the a_js and f_js as

$$\mathrm{E}\left\{\left\|Y-\sum_{j=1}^{r} a_j \int_0^1 f_j(s)(K^{1/2}X)(s)\,ds\right\|^2\right\}.$$

In the above, given $a_1,\ldots,a_r$, the minimizer for $f_1,\ldots,f_r$ is similar to that in functional linear regression and is given by

$$f_j = T_X^{-1}T_{XY}a_j, \tag{6}$$

where $T_X = \mathrm{E}\{(K^{1/2}X)\otimes(K^{1/2}X)\}$ is the covariance operator of $K^{1/2}X$, which is assumed to be invertible although the inverse is not bounded. Also, $T_{XY} = \mathrm{E}\{(K^{1/2}X)\otimes Y\}$ is the cross-covariance operator, where for $x, y \in L^2([0,1])$, $x\otimes y : L^2([0,1]) \to L^2([0,1])$ is defined by $(x\otimes y)(g) = \langle y, g\rangle x$ for any $g \in L^2([0,1])$. If fact, to see (6), we note that

$$\begin{aligned}\mathrm{E}\left\{\left\|Y-\sum_{j=1}^{r} a_j \int_0^1 f_j(s)(K^{1/2}X)(s)\,ds\right\|^2\right\} &= \mathrm{E}\|Y\|^2 - 2\mathrm{E}\left\langle Y, \sum_{j=1}^{r}\langle f_j, K^{1/2}X\rangle a_j\right\rangle + \left\|\sum_{j=1}^{r}\langle f_j, K^{1/2}X\rangle a_j\right\|^2 \\ &= \mathrm{E}\|Y\|^2 - 2\mathrm{E}\left\langle Y, \sum_{j=1}^{r}\langle f_j, K^{1/2}X\rangle a_j\right\rangle + \sum_{j=1}^{r}\mathrm{E}\langle f_j, K^{1/2}X\rangle^2,\end{aligned}$$

where in the last step above we used that $a_1,\ldots,a_r$ are orthonormal. Thus f_j is the minimizer of

$$\mathrm{E}\langle f_j, K^{1/2}X\rangle^2 - 2\mathrm{E}\langle Y, \langle f_j, K^{1/2}X\rangle a_j\rangle = \langle T_X f_j, f_j\rangle - 2\mathrm{E}\langle Y, a_j\rangle\langle f_j, K^{1/2}X\rangle = \langle T_X f_j, f_j\rangle - 2\langle T_{XY}a_j, f_j\rangle,$$

which implies (6). Using $f_j = T_X^{-1}T_{XY}a_j$, we have

$$\begin{aligned}\left\|Y-\sum_{j=1}^{r} a_j \int_0^1 f_j(s)(K^{1/2}X)(s)\,ds\right\|^2 &= \left\|Y-\sum_{j=1}^{r}\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle a_j\right\|^2 \\ &= \|Y\|^2 - 2\left\langle Y, \sum_{j=1}^{r}\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle a_j\right\rangle + \sum_{j=1}^{r}\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle^2,\end{aligned}$$

where $T_{YX} = \mathrm{E}\{Y\otimes(K^{1/2}X)\}$ is the adjoint operator of T_{XY}. Direct calculations yield

$$\langle Y, \langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle a_j\rangle = \langle Y, a_j\rangle\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle = \langle T_{YX}T_X^{-1}\{(K^{1/2}X)\otimes Y\}a_j, a_j\rangle,$$

and

$$\begin{aligned}\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle^2 = \langle K^{1/2}X, T_X^{-1}T_{XY}a_j\rangle^2 &= \langle (K^{1/2}X\otimes K^{1/2}X)T_X^{-1}T_{XY}a_j, T_X^{-1}T_{XY}a_j\rangle \\ &= \langle T_{YX}T_X^{-1}(K^{1/2}X\otimes K^{1/2}X)T_X^{-1}T_{XY}a_j, a_j\rangle.\end{aligned}$$

Thus minimization of

$$\mathrm{E}\left(\left\|Y-\sum_{j=1}^{r}\langle T_{YX}T_X^{-1}K^{1/2}X, a_j\rangle a_j\right\|^2\right)$$

is equivalent to maximization of $\sum_{j=1}^{r}\langle T_{YX}T_X^{-1}T_{XY}a_j, a_j\rangle$. It is well known that the maximizers $a_1,\ldots,a_r$ are the eigenfunctions of $T_{YX}T_X^{-1}T_{XY}$ associated with the largest r eigenvalues; see, e.g., Theorem 4.2.5 of [14].

For simplicity we assume that the largest r eigenvalues are distinct and that the a_js are uniquely identified. Note that because $\langle K^{1/2}X, T_X^{-1}T_{XY}a_1\rangle = \langle X, b_1\rangle$ is the conditional mean of $\langle Y, a_1\rangle$ — see (4) — and a_1 maximizes $\mathrm{E}\langle K^{1/2}X, T_X^{-1}T_{XY}a\rangle^2 = \langle T_{YX}T_X^{-1}T_{XY}a, a\rangle$, we can interpret a_1 as the direction such that the projection of Y has the largest variability explained by the predictor.

2.2. Empirical estimator and asymptotic properties

Similarly to the previous subsection, for the sample version (5), it can be shown that $\hat{a}_1, \ldots, \hat{a}_r$ can be obtained as the eigenfunctions of $\hat{T}_{YX}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}$ and $\hat{b}_j = K^{1/2}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}\hat{a}_j$, where I is the identity operator,

$$\hat{T}_X = \frac{1}{n}\sum_{i=1}^{n} K^{1/2}X \otimes K^{1/2}X \quad \text{and} \quad \hat{T}_{XY} = \frac{1}{n}\sum_{i=1}^{n} K^{1/2}X \otimes Y$$

are the sample versions of T_X and T_{XY}, respectively. Thus the ridge-type penalty in (5) provides regularization of the inverse of $\hat{T}_X$, given that both $\hat{T}_X$ and T_X do not have a bounded inverse.

We now study the asymptotic property of the estimators. Suppose that the spectral decomposition of T_X is

$$T_X = \sum_{j=1}^{\infty} \gamma_j e_j \otimes e_j,$$

with $\gamma_1 > \gamma_2 > \cdots > 0$. By Theorem 7.2.6 in [14], this is guaranteed if $K^{1/2}X$ is a random element of $L^2([0,1])$ and $\mathrm{E}\|K^{1/2}X\|^2 < \infty$, which is implied by our assumption (A1) below. The following technical assumptions are imposed.

(A1) X is a random element of $L_2([0,1])$, $K \in L^2([0,1]^2)$, $\mathrm{E}\|K^{1/2}X\|^4 < \infty$. The eigenvalues of T_X are all positive and distinct.

(A2) $\beta(t,s) = a_1(t)b_1(s) + \cdots + a_r(t)b_r(t)$ for some fixed integer r, $a_j \in L^2([0,1])$, $b_j \in \mathcal{H}_K$. $\langle a_j, a_{j'}\rangle = \delta_{jj'}$. $T_{YX}T_X^{-1}T_{XY}$ has r distinct nonzero eigenvalues with associated eigenfunctions a_j.

(A3) $\epsilon \in L^2([0,1])$, $\mathrm{E}\langle \epsilon, h\rangle = 0$ and $\mathrm{E}(\langle \epsilon, h\rangle^2) < \infty$ for any $h \in L^2([0,1])$.

Assumptions (A1) implies that T_X is a trace-class operator (i.e., $\sum \gamma_j < \infty$). If T_X has non-trivial kernel space, then only the projection of b_j orthogonal to the kernel space of T_X can be identified. Thus invertibility of T_X is typically assumed in functional linear regression for convenience. In (A2), it can be shown that the expression for $\beta(t,s)$ actually implies that $T_{YX}T_X^{-1}T_{XY}$ has r nonzero eigenvalues, and we further assume that they are distinct so that the a_js are identified. Actually, that the eigenvalues are distinct is only assumed for simplicity so that $\hat{a}_j$ can converge to a_j; otherwise we can only say that the subspace spanned by $\hat{a}_1, \ldots, \hat{a}_r$ will converge to that spanned by $a_1, \ldots, a_r$.

For the convergence of $\hat{a}_j$, we can directly use the L^2 norm. However, for b_j, the risk we consider is $\mathrm{E}^*\langle \hat{b}_j - b_j, X^*\rangle^2 = \|\Sigma_X^{1/2}(\hat{b}_j - b_j)\|^2$, where X^* is a copy of X independent of the training data, E^* is the expectation taken over X^*, and $\Sigma_X = \mathrm{E}(X \otimes X)$ is the covariance operator of X.

Theorem 1. *Under Assumptions (A1)–(A3), we have*

$$\|\hat{a}_j - a_j\|^2 = o_p(1), \quad \|\Sigma_X^{1/2}(\hat{b}_j - b_j)\|^2 = o_p(1), \quad \mathrm{E}^*\left\|\int_0^1 \hat{\beta}(t,s)X^*(s)ds - \int_0^1 \beta(t,s)X^*(s)ds\right\|^2 = o_p(1),$$

where $\hat{\beta}(t,s) = \sum_{j=1}^{r} \hat{a}_j(t)\hat{b}_j(s)$. Obviously, the right-most term is directly related to prediction error.

To get a nontrivial convergence rate, we use the following additional assumptions.

(B1) $\gamma_j \asymp j^{-\alpha}$ for some $\alpha > 1$.

(B2) For all $f \in L^2([0,1])$,

$$\mathrm{E}\left\{\int X(t)f(t)dt\right\}^4 \le c\left[\mathrm{E}\left\{\int X(t)f(t)dt\right\}^2\right]^2.$$

Assumptions (B1) and (B2) also appeared in [7].

Theorem 2. *Under Assumptions (A1)–(A3) and (B1)–(B2), and that* $\lambda \asymp n^{-\alpha/(\alpha+1)}$, *we have* $\|\hat{a}_j - a_j\|^2 = O_p\{n^{-\alpha/(\alpha+1)}\}$, $\|\Sigma_X^{1/2}(\hat{b}_j - b_j)\|^2 = O_p\{n^{-\alpha/(\alpha+1)}\}$, *and*

$$\mathrm{E}^* \left\| \int_0^1 \hat{\beta}(t,s)X^*(s)ds - \int_0^1 \beta(t,s)X^*(s)ds \right\|^2 = O_p\{n^{-\alpha/(\alpha+1)}\}.$$

So far, for notational simplicity, we assumed that K is a positive definite kernel and that the RKHS norm of b_j is used in the penalty in (5). In practice, one can usually decompose $\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ and $\mathcal{H}_1$ are orthogonal subspaces and $\mathcal{H}_0$ is finite-dimensional, and only the projection of b_j on $\mathcal{H}_1$ is used in the penalty. A common example is again the second-order Sobolev space, where $\mathcal{H}_0$ is the space of linear functions. Our theory and methodology can be straightforwardly extended to this case. For our numerical results on the real data, we use the second-order Sobolev space of periodic functions since our functional observations are periodic in nature. In this situation, $\mathcal{H}_0$ contains functions that are constants. We only penalize the projection b_j onto $\mathcal{H}_1$ in our numerical study.

In the Online Supplement, we further discuss how the theory can be adapted to the case when r is not the same as the rank for the true model.

2.3. Computation and theoretical implications of discretization

In our implementation, all operators are approximated by matrices and numerical integration is used. For example, with a grid $(t_1, \ldots, t_p)$ on $[0, 1]$, $K^{1/2}$ is represented by the $p \times p$ matrix $\{K(t_j, t_k)\}$ with $j, k \in \{1, \ldots, p\}$ and then we approximate $K^{1/2}X_i$ by

$$\frac{1}{p}\sum_{k=1}^{p} K^{1/2}(t_j, t_k)X_i(t_k) \tag{7}$$

for all $j \in \{1, \ldots, p\}$. Then $\hat{T}_X$ can be represented by a $p \times p$ matrix with its inverse readily computed after adding λI. Similarly we can represent $\hat{T}_{XY}$ by a matrix and then compute the eigenvector of the matrix approximation of $\hat{T}_{YX}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}$ to get a p-dimensional vector as the discretized version of $\hat{a}_j$.

Our implementation simply treats the discretized functional variable as multi-dimensional vectors and thus multivariate procedures are directly used. Alternatively, in estimating b_j for example, one can also base implementation on a representer theorem as derived in [7], which looks more elegant. However, since modern desktop PCs can easily deal with multivariate variables with large dimensions, there seems to be no convincing advantage to this approach compared to straightforward discretization of the functional variables.

To see the computational complexities, suppose we discretize different functions using a grid of p points. The time complexity for computing the matrix square root of the $p \times p$ matrix $\{K(t_j, t_k)\}$ is $O(p^3)$. The complexity is $O(p^2 n)$ to approximate $K^{1/2}X_i$ by (7), and to construct the discretized version of $\hat{T}_X$. The complexity to compute the inverse of $\hat{T}_X + \lambda I$ is $O(p^3)$. The calculation of eigenvectors takes at most $O(p^3)$. All other steps, including calculating $\hat{b}_j$, takes no more than $O\{p^2(n + p)\}$ and thus the overall complexity is $O\{p^2(n + p)\}$.

We now discuss how discretization impacts our asymptotic theory. Computationally, after discretization, the operators reduce to matrices and eigenfunctions to finite-dimensional eigenvectors.

Let $\check{X}_i(t) = X_i(t_k)$ if $t \in [t_k, t_{k-1})$, and similarly define $\check{Y}_i$. We assume $t_k = (k - 1)/p$ for all $k \in \{1, \ldots, p\}$ for simplicity. Note that $\hat{a}_j$ is the eigenfunction of $\hat{T}_{YX}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}$. The discretization procedure is equivalent to computing the eigenfunction of $\check{T}_{YX}(\check{T}_X + \lambda I)^{-1}\check{T}_{XY}$, denoted by $\check{a}_j$, where $\check{T}_X = \sum_{i=1}^n \check{X}_i \otimes \check{X}_i/n$ and $\check{T}_{XY} = \sum_{i=1}^n \check{X}_i \otimes \check{Y}_i/n$. We have

$$\begin{aligned}\|n(\check{T}_X - \hat{T}_X)\|^2 &= \left\| \sum_{i=1}^n \check{X}_i \otimes \check{X}_i - \sum_{i=1}^n X_i \otimes X_i \right\|^2 \\ &\le 3\left\| \sum_{i=1}^n (X_i - \check{X}_i) \otimes X_i \right\|^2 + 3\left\| \sum_{i=1}^n X_i \otimes (X_i - \check{X}_i) \right\|^2 + 3\left\| \sum_{i=1}^n (X_i - \check{X}_i) \otimes (X_i - \check{X}_i) \right\|^2.\end{aligned}$$

Furthermore, the right-hand term can be bounded above by

$$6\left(\sum_{i=1}^{n}\|X_i-\check{X}_i\|\times\|X_i\|\right)^2+3\left(\sum_{i=1}^{n}\|X_i-\check{X}_i\|^2\right)^2\le 6\left(\sum_{i=1}^{n}\|X_i-\check{X}_i\|^2\right)\left(\sum_{i=1}^{n}\|X_i\|^2\right)+3\left(\sum_{i=1}^{n}\|X_i-\check{X}_i\|^2\right)^2,$$

so that $\|n(\check{T}_X-\hat{T}_X)\|^2=O_p\{n^2(\delta_x+\delta_x^2)\}$, where $\delta_x=\mathrm{E}\|X_i-\check{X}_i\|^2$. Similarly, we can show $\|\check{T}_{XY}-\hat{T}_{XY}\|^2=O_p(\delta_x+\delta_y+\delta_x\delta_y)$, where $\delta_y=\mathrm{E}\|Y_i-\check{Y}_i\|^2$. In the following we assume $\delta_x,\delta_y=o(1)$. Then we have

$$\begin{aligned}\|\check{T}_{YX}(\check{T}_X+\lambda I)^{-1}\check{T}_{XY}-\hat{T}_{YX}(\hat{T}_X+\lambda I)^{-1}\hat{T}_{XY}\|&\\ \le\|\check{T}_{YX}-\hat{T}_{YX}\|\|(\hat{T}_X+\lambda I)^{-1}\hat{T}_{XY}\|+\|\check{T}_{YX}(\check{T}_X+\lambda I)^{-1}(\check{T}_X-\hat{T}_X)(\hat{T}_X+\lambda I)^{-1}\hat{T}_{XY}\|&\\ +\|\check{T}_{YX}(\check{T}_X+\lambda I)^{-1}\|\|\check{T}_{XY}-\hat{T}_{XY}\|=O_p(\sqrt{\delta_x+\delta_y}/\lambda^2).&\end{aligned}$$

Thus $\|\check{a}_j-\hat{a}_j\|^2=O_p\{(\delta_x+\delta_y)/\lambda^4\}$. Let $\check{f}_j=(\check{T}_X+\lambda I)^{-1}\check{T}_{XY}\check{a}_j$. Then

$$\begin{aligned}\|\check{f}_j-\hat{f}_j\|\le\|(\check{T}_X+\lambda I)^{-1}\check{T}_{XY}\|\|\check{a}_j-\hat{a}_j\|+\|(\check{T}_X+\lambda I)^{-1}(\check{T}_{XY}-\hat{T}_{XY})\hat{a}_j\|&\\ +\|(\check{T}_X+\lambda I)^{-1}(\check{T}_X-\hat{T}_X)(\hat{T}_X+\lambda I)^{-1}\|\|\hat{T}_{XY}\hat{a}_j\|=O_p(\sqrt{\delta_x+\delta_y}/\lambda^3).&\end{aligned}$$

If we then define $\check{b}_j=K^{1/2}\check{f}_j$ and $\check{\beta}(t,s)=\sum_{j=1}^{r}\check{a}_j(t)\check{b}_j(s)$, then $\|\Sigma_X^{1/2}(\check{b}_j-\hat{b}_j)\|^2=O_p\{(\delta_x+\delta_y)/\lambda^6\}$ and

$$\mathrm{E}^*\left\|\int\check{\beta}(t,s)X^*(s)ds-\int\hat{\beta}(t,s)X^*(s)dx\right\|^2=O_p\{(\delta_x+\delta_y)/\lambda^6\}.$$

Thus under some smoothness assumptions of the sample paths of X and Y, with large enough p, the error in discrete sampling of functional data can be ignored. For example, one can use assumption (A.2) of [9] to get $\delta_x=p^{-\kappa}$ for some $\kappa>0$.

So far, we have taken into account that the functional data are discretely sampled. Another related problem is that when calculating $\check{b}_j=K^{1/2}\check{f}_j$, we need to compute the operator $K^{1/2}$. This can only be done in practice after discretization to represent K as a matrix. Let $\check{K}$ be the discretized version of K, i.e., $\check{K}(s,t)=K(s_i,t_j)$ for $s\in[s_i,s_{i+1}),t\in[t_i,t_{i+1})$, where $s_i=(i-1)/p$ and $t_i=(i-1)/p$ for all $i\in\{1,\dots,p\}$. Smoothness assumptions will produce bounds for $\|\check{K}-K\|$. For example, if $K(s,t)$ is Lipschitz continuous, we have $\|\check{K}-K\|=O(1/p)$. The difficulty, however, is that we need to bound $\|\check{K}^{1/2}-K^{1/2}\|$, which seems hard, especially because the eigenvalues of K typically converge to zero. We could get around this difficulty by truncating the spectral expansion of $(\check{K})^{1/2}$. In fact, suppose the spectral decomposition of K is $K=\sum s_j\phi_j\otimes\phi_j$ with $s_1>s_2>\cdots>0$, and that of $\check{K}$ is $\check{K}=\sum_j\tilde{s}_j\tilde{\phi}_j\otimes\tilde{\phi}_j$. Then by Theorems 5.1.6 and 5.1.8 of [14], $|s_j-\tilde{s}_j|\le\|\check{K}-K\|$ and $\|\tilde{\phi}_j-\phi_j\|\le C\|\check{K}-K\|/r_j$, where $r_j=\min(s_{j-1}-s_j,s_j-s_{j+1})$. Let the truncated version $\check{K}^{1/2}$ be $\tilde{K}_{(m)}^{1/2}=\sum_{j=1}^{m}\sqrt{\tilde{s}_j}\tilde{\phi}_j\otimes\tilde{\phi}_j$. We have

$$\begin{aligned}\tilde{K}_{(m)}^{1/2}-K^{1/2}&=\sum_{j=1}^{m}\sqrt{\tilde{s}_j}\tilde{\phi}_j\otimes\tilde{\phi}_j-\sum_{j=1}^{m}\sqrt{s_j}\phi_j\otimes\phi_j-\sum_{j=m+1}^{\infty}\sqrt{s_j}\phi_j\otimes\phi_j\\ &=\sum_{j=1}^{m}\sqrt{\tilde{s}_j}(\tilde{\phi}_j\otimes\tilde{\phi}_j-\phi_j\otimes\phi_j)+\sum_{j=1}^{m}(\sqrt{\tilde{s}_j}-\sqrt{s_j})\phi_j\otimes\phi_j-\sum_{j=m+1}^{\infty}\sqrt{s_j}\phi_j\otimes\phi_j.\end{aligned}$$

We have

$$\begin{aligned}\left\|\sum_{j=1}^{m}\sqrt{\tilde{s}_j}(\tilde{\phi}_j\otimes\tilde{\phi}_j-\phi_j\otimes\phi_j)\right\|&\le C\max_{j\in\{1,\dots,m\}}\sqrt{\tilde{s}_j}\|\check{K}-K\|/r_j\le C\max_{j\in\{1,\dots,m\}}(\sqrt{s_j}+\|\check{K}-K\|)\|\check{K}-K\|/r_j,\\ \left\|\sum_{j=1}^{m}(\sqrt{\tilde{s}_j}-\sqrt{s_j})\phi_j\otimes\phi_j\right\|&\le C\max_{j\in\{1,\dots,m\}}|\tilde{s}_j-s_j|/(\sqrt{\tilde{s}_j}+\sqrt{s_j})\le C\max_{j\in\{1,\dots,m\}}\|\check{K}-K\|/\sqrt{s_j}.\end{aligned}$$

Then by letting m be large enough to make $\| \sum_{j=m+1}^{\infty} \sqrt{s_j}\phi_j \otimes \phi_j \|$ arbitrarily small, and then choosing p to be large enough to make $\|\tilde{K} - K\|$ arbitrarily small, we can make $\|\tilde{K}_{(m)}^{1/2} - K^{1/2}\|$ arbitrarily small, so that the rates in Theorems 1–2 are not affected.

Although theoretically this truncation makes it possible to bound $\|\tilde{K}_{(m)}^{1/2} - K^{1/2}\|$, empirically we find the truncation is not necessary and do not use it in our numerical implementation. It remains an open problem whether one can get a bound for $\|\tilde{K}^{1/2} - K^{1/2}\|$.

3. Numerical results

3.1. Simulations

The construction of the RKHS and the covariates in our simulations is the same as that used in [7]. We consider the RKHS with kernel

$$K(s,t) = \sum_{j=1}^{\infty} 2\cos(j\pi s)\cos(j\pi t)/(j\pi)^4,$$

and thus $\mathcal{H}_K$ consists of functions of the form

$$f(t) = \sum_{j=1}^{\infty} f_j \cos(j\pi t),$$

such that $\sum_j j^4 f_j^2 < \infty$. In our implementation, the kernel $K(s,t)$ is approximated by using the first 100 terms in the infinite sum. In this RKHS, we actually have $\|f\|_{\mathcal{H}_K}^2 = \int (f'')^2$. For the covariate kernel $\Sigma_X(s,t) = \mathrm{E}\{X(s)X(t)\}$, we set the covariate kernel to be

$$\Sigma_X(s,t) = \sum_{j=1}^{\infty} 2\theta_j \cos(j\pi s)\cos(j\pi t),$$

where $\theta_j = (|j - j_0| + 1)^{-2}$. When $j_0 = 1$, the two kernels Σ_X and K have the same sequence of eigenfunctions (ordered by eigenvalues). By choosing $j_0 > 1$, we allow the sequences to be different. The error process is generated from

$$\epsilon_i(t) = \sum_{j=2}^{10} \sqrt{2}\eta_i \cos(j\pi t)/j, \quad \eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

As we will see later in the results, if σ is large enough, the error will significantly affect the major mode of variability (as assessed by functional PCA) of the response. Thus the basis obtained from functional PCA on Y does not represent the mode of variability that can be predicted.

The responses are generated from (1) without the intercept term, using the following three examples of coefficient function:

Example 1: $\beta(t,s) = \sin(2\pi t)\cos(2\pi s)$

Example 2: $\beta(t,s) = \{\sin(\pi t) + \sin(2\pi t)\}\{\cos(\pi s) - \cos(2\pi s)\} + \{\sin(\pi t) - \sin(2\pi t)\}\{\cos(\pi s) + \cos(2\pi s)\}$

Example 3: $\beta(t,s) = \sin(\pi t)\cos(\pi s) + \{\sin(\pi t) + \cos(\pi t)\}\cos(2\pi s) + \{\sin(\pi t) + \cos(2\pi t)\}\cos(4\pi s)$

Example 4: $\beta(t,s) = \sum_{j=1}^{50} 4\sqrt{2}(-1)^j \sin(j\pi t)\cos(j\pi s)/j^2$

Thus the first three examples represent models with rank 1, 2 and 3, respectively. Although β in the fourth example also has a finite rank, the purpose is to demonstrate that a small rank suffices to fit the data well if it can provide a good approximation to the truth. In generating the data, the integral is approximated by a Riemannian sum on an equally-spaced grid $(t_1, \ldots, t_{100})$ with 100 points, and all functional variables are also discretized on the same grid of 100 points.

The estimators for a_j are obtained as eigenfunctions of $\hat{T}_{YX}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}$ and estimators for b_j is obtained by $\hat{b}_j = K^{1/2}(\hat{T}_X + \lambda I)^{-1}\hat{T}_{XY}\hat{a}_j$ based on discretization as mentioned in Section 2.3. With our implementation in R, all simulations are completed in about three hours using a HP Z230 workstation running Windows 7 with i7-4790 CPU @ 3.60GHz.

We first examine the results for different values of λ with r set at the true value. We consider $\ln\lambda \in \{-10, \ldots, -1\}$. Although the results could be improved a little by using a finer sequence of λ, we find that such a relatively coarse sequence is usually sufficient.

First we set $n = 100$, $\sigma = 0.25$, 0.5 or 1, $j_0 = 3$ and generate 100 datasets from Example 1. Another 1000 observations are generated from the same model for calculating the prediction error. For $r = 1$ and λ minimizing the prediction error, the top row of Figure 1 compares the estimated b_1 (normalized to have unit norm before visualization) with the estimated first eigenfunction of X and the bottom row compares the estimated a_1 with the estimated first eigenfunction of Y, where the estimates for a_1, b_1 are averaged over the 100 simulations. We see that b_1 is very different from the eigenfunction of X and the former represents the direction of X that is useful for prediction. Looking at the shape of b_1, this can be interpreted as saying that the difference between value of X in the middle and the value of X at the two ends is useful for prediction. When σ (noise) is small, a_1 matches well with the eigenfunction of Y since the variability of Y mainly comes from that of X. As σ increases, the shape of eigenfunctions of Y is increasingly dominated by error variability which does not represent the "explained" variability of the response. Thus a_1 is visually different from the eigenfunction of Y. Whatever the noise, the estimated shape of a_j suggests that X is only useful in predicting the difference between the value of Y on $[0, 1/2]$ and that on $[1/2, 1]$. The eigenfunction of Y only shows the direction of Y that varies the most which is not necessarily the direction that is related to the predictor.

For Example 2, using the same set of parameters, the top row of Figure 2 shows the estimated b_1, b_2 (solid curves) with the first two estimated eigenfunctions of the predictor, and the bottom row of Figure 2 shows the estimated a_1, a_2 (solid curves) with the estimated eigenfunctions of the response. In Figure 2, b_1 and b_2 are scaled to have unit norm. We see that the estimated functions a_1, a_2, b_1, b_2 are indeed different from the eigenfunctions when the noise is large, and the former can be interpreted in a similar fashion as we did for Example 1.

In practice, the parameters r and λ can be tuned via five-fold cross-validation. Now we consider the prediction error with more parameter set-ups and assess the performance of five-fold cross-validation. We set $n = 50$ or 100, $\sigma = 0.25$, 0.5 or 1, $j_0 = 1$ or 3, for a total of 12 scenarios. We search the optimal set of tuning parameters r and λ with r ranging from 1 to 5 and $\ln\lambda \in \{-10, \ldots, -1\}$. Again an independent set of 1000 observations is generated for testing. Besides comparing between reduced rank estimator (5) and the functional linear regression estimator (2), we also compare the prediction error using r and λ selected by cross-validation (the selection of tuning parameters does not use the test data) with the smallest prediction error achieved by different r and λ values (as in the previous investigation without using cross-validation) in Tables 1–4, for the four examples, respectively. Using either r, λ with the smallest prediction error, or obtained by cross-validation, we also compute the mean squared error of $\hat{\beta}$, viz.

$$\text{MSE} = \iint_{[0,1]^2} \{\hat{\beta}(t,s) - \beta(t,s)\}^2 dtds.$$

We see that in all cases under Examples 1–3, the reduced rank regression performs better than functional linear regression without rank reduction of [18], although judging from the standard errors, the differences are not statistically significant. For Example 4, models with rank no more than 5 gives basically the same accuracy as functional linear regression. The estimators with parameters chosen by the cross-validation method perform well with prediction errors and MSE close to that without using cross-validation (the latter directly uses the independently generated testing data for parameter tuning and thus is considered as infeasible estimators). In the column labeled "correct r", we also present the number of times (among 100 repetitions) that the true r (1 or 2) is selected by cross-validation. We see that most of the time the true value of r can be identified.

3.2. Real data

We now illustrate functional reduced rank regression on two common functional datasets.

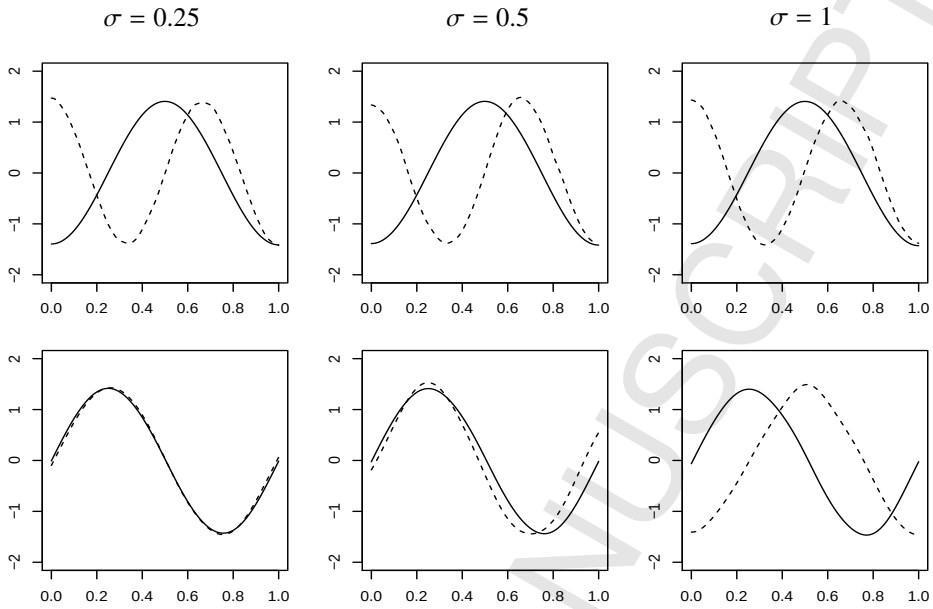

Figure 1: The first row shows the estimated function b_1 (solid line) and the first eigenfunction of X (dashed line), and the second row shows the estimated function a_1 (solid line) and the first eigenfunction of Y (dashed line). The three columns correspond to $\sigma = 0.25, 0.5$ and 1, respectively.

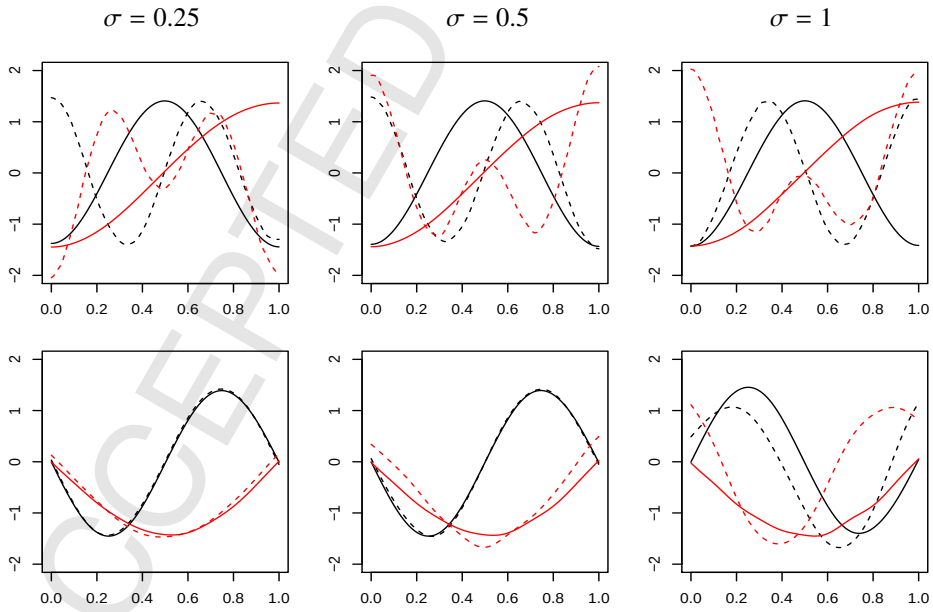

Figure 2: The first row shows the estimated function b_1, b_2 (solid line) and the first two eigenfunction of X (dashed line), and the second row shows the estimated function a_1, a_2 (solid line) and the first two eigenfunction of Y (dashed line). The three columns correspond to $\sigma = 0.25$, 0.5 and 1, respectively. The first directions (a_1, b_1 and first eigenfunction) are shown in black and the second directions in red.

Table 1: Prediction errors and mse for the simulations for the functional linear regression (FLR, Eq. (2), as in [18]) and reduced rank regression (RRR, Eq. (5)), for Example 1. The numbers in brackets are the standard errors computed from simulations.

			Prediction Error				MSE ($\times 10^{-4}$)				
n	j_0	σ	FLR	FLR-CV	RRR	RRR-CV	FLR	FLR-CV	RRR	RRR-CV	correct r
50	1	0.25	0.046	0.048	0.027	0.034	0.82	0.93	0.19	0.29	81
			(0.008)	(0.011)	(0.009)	(0.018)	(0.66)	(0.73)	(0.20)	(0.33)	
50	1	0.5	0.086	0.096	0.055	0.069	2.26	2.94	0.77	1.18	80
			(0.014)	(0.021)	(0.017)	(0.035)	(1.97)	(2.84)	(0.85)	(1.34)	
50	1	1	0.149	0.170	0.108	0.144	4.22	6.01	2.56	4.25	69
			(0.031)	(0.052)	(0.031)	(0.068)	(3.87)	(6.51)	(3.02)	(5.30)	
50	3	0.25	0.052	0.055	0.028	0.033	0.96	1.05	0.21	0.28	86
			(0.009)	(0.012)	(0.009)	(0.016)	(0.79)	(0.90)	(0.22)	(0.31)	
50	3	0.5	0.096	0.105	0.057	0.070	3.45	4.04	0.91	1.22	82
			(0.013)	(0.023)	(0.018)	(0.034)	(2.86)	(3.70)	(1.00)	(1.36)	
50	3	1	0.167	0.194	0.116	0.154	9.84	12.7	4.13	6.07	63
			(0.029)	(0.051)	(0.034)	(0.070)	(8.95)	(12.9)	(4.64)	(7.14)	
100	1	0.25	0.034	0.034	0.019	0.023	0.64	0.63	0.11	0.16	82
			(0.005)	(0.006)	(0.007)	(0.010)	(0.48)	(0.47)	(0.12)	(0.18)	
100	1	0.5	0.060	0.063	0.039	0.046	1.03	1.10	0.44	0.63	82
			(0.011)	(0.013)	(0.014)	(0.020)	(0.86)	(1.06)	(0.46)	(0.69)	
100	1	1	0.108	0.112	0.077	0.092	2.74	2.94	1.38	1.86	80
			(0.023)	(0.029)	(0.026)	(0.039)	(2.41)	(2.79)	(1.58)	(2.34)	
100	3	0.25	0.038	0.038	0.020	0.024	0.57	0.62	0.10	0.15	80
			(0.005)	(0.006)	(0.007)	(0.010)	(0.40)	(0.47)	(0.11)	(0.17)	
100	3	0.5	0.068	0.069	0.040	0.049	1.70	1.83	0.42	0.65	81
			(0.010)	(0.014)	(0.014)	(0.021)	(1.45)	(1.64)	(0.44)	(0.71)	
100	3	1	0.125	0.128	0.081	0.097	5.33	5.63	1.74	2.57	80
			(0.021)	(0.030)	(0.027)	(0.041)	(4.84)	(5.35)	(1.93)	(2.98)	

Table 2: Prediction errors and mse for the simulations for the functional linear regression (FLR, Eq. (2)) and reduced rank regression (RRR, Eq. (5)), for Example 2. The numbers in brackets are the standard errors computed from simulations.

			Prediction Error				MSE ($\times 10^{-4}$)				
n	j_0	σ	FLR	FLR-CV	RRR	RRR-CV	FLR	FLR-CV	RRR	RRR-CV	correct r
50	1	0.25	0.047	0.048	0.039	0.046	0.87	0.91	0.41	0.55	63
			(0.009)	(0.011)	(0.009)	(0.016)	(0.72)	(0.73)	(0.34)	(0.58)	
50	1	0.5	0.092	0.096	0.078	0.091	3.27	3.64	1.41	2.04	62
			(0.017)	(0.023)	(0.018)	(0.032)	(2.77)	(2.97)	(1.20)	(2.13)	
50	1	1	0.173	0.191	0.155	0.186	8.18	11.6	5.00	8.66	59
			(0.032)	(0.043)	(0.036)	(0.064)	(6.84)	(11.4)	(4.55)	(9.26)	
50	3	0.25	0.054	0.055	0.040	0.047	1.01	1.06	0.69	0.84	72
			(0.010)	(0.012)	(0.009)	(0.015)	(0.85)	(0.91)	(0.63)	(0.83)	
50	3	0.5	0.105	0.110	0.080	0.094	3.82	4.16	2.63	3.17	68
			(0.018)	(0.024)	(0.018)	(0.031)	(3.23)	(3.58)	(2.61)	(3.18)	
50	3	1	0.194	0.210	0.162	0.193	13.9	15.7	10.4	13.8	65
			(0.031)	(0.048)	(0.036)	(0.055)	(11.8)	(14.5)	(10.6)	(13.5)	
100	1	0.25	0.034	0.034	0.027	0.030	0.66	0.62	0.22	0.25	75
			(0.005)	(0.005)	(0.006)	(0.007)	(0.50)	(0.46)	(0.18)	(0.23)	
100	1	0.5	0.068	0.069	0.054	0.060	2.43	2.52	0.81	0.96	74
			(0.011)	(0.012)	(0.013)	(0.013)	(1.89)	(1.92)	(0.67)	(0.88)	
100	1	1	0.119	0.127	0.108	0.120	3.68	4.52	2.84	3.83	72
			(0.020)	(0.023)	(0.024)	(0.027)	(3.28)	(4.24)	(2.54)	(3.54)	
100	3	0.25	0.038	0.038	0.028	0.031	0.57	0.61	0.33	0.40	78
			(0.005)	(0.005)	(0.006)	(0.007)	(0.40)	(0.46)	(0.30)	(0.40)	
100	3	0.5	0.076	0.077	0.056	0.062	2.23	2.47	1.28	1.57	78
			(0.011)	(0.012)	(0.013)	(0.015)	(1.59)	(1.91)	(1.18)	(1.57)	
100	3	1	0.135	0.141	0.112	0.126	6.29	7.32	5.13	6.57	75
			(0.019)	(0.025)	(0.024)	(0.031)	(5.32)	(6.46)	(4.75)	(6.79)	

Table 3: Prediction errors and mse for the simulations for the functional linear regression (FLR, Eq. (2)) and reduced rank regression (RRR, Eq. (5)), for Example 3. The numbers in brackets are the standard errors computed from simulations.

			Prediction Error				MSE ($\times 10^{-4}$)				
n	j_0	σ	FLR	FLR-CV	RRR	RRR-CV	FLR	FLR-CV	RRR	RRR-CV	correct r
50	1	0.25	0.062	0.063	0.055	0.061	3.71	4.01	2.22	2.97	90
			(0.010)	(0.011)	(0.012)	(0.017)	(2.38)	(2.80)	(1.87)	(2.82)	
50	1	0.5	0.121	0.126	0.106	0.119	13.8	15.4	8.03	10.8	92
			(0.018)	(0.023)	(0.021)	(0.036)	(8.84)	(10.9)	(6.81)	(9.88)	
50	1	1	0.220	0.232	0.205	0.229	32.0	35.5	26.8	33.3	69
			(0.032)	(0.047)	(0.036)	(0.056)	(26.0)	(29.1)	(22.4)	(39)	
50	3	0.25	0.065	0.067	0.055	0.063	2.50	2.91	1.52	2.18	91
			(0.010)	(0.011)	(0.012)	(0.018)	(1.55)	(1.98)	(1.27)	(2.01)	
50	3	0.5	0.129	0.137	0.110	0.127	9.27	11.3	5.94	8.76	86
			(0.018)	(0.022)	(0.022)	(0.036)	(5.94)	(8.35)	(4.72)	(7.98)	
50	3	1	0.236	0.238	0.216	0.230	19.9	20.5	19.7	20.0	67
			(0.039)	(0.049)	(0.041)	(0.046)	(14.7)	(16.9)	(15.7)	(24.4)	
100	1	0.25	0.045	0.045	0.038	0.040	2.68	2.64	1.43	1.65	90
			(0.005)	(0.005)	(0.006)	(0.007)	(1.59)	(1.73)	(1.27)	(1.53)	
100	1	0.5	0.088	0.091	0.076	0.081	9.52	10.5	5.45	6.46	88
			(0.011)	(0.011)	(0.013)	(0.015)	(5.71)	(7.10)	(4.64)	(6.36)	
100	1	1	0.157	0.165	0.145	0.160	17.6	19.5	15.7	19.2	67
			(0.022)	(0.032)	(0.026)	(0.029)	(13.4)	(15.9)	(13.3)	(20.1)	
100	3	0.25	0.048	0.048	0.039	0.041	1.71	1.83	1.00	1.11	96
			(0.006)	(0.006)	(0.006)	(0.008)	(1.05)	(1.15)	(0.85)	(0.99)	
100	3	0.5	0.090	0.095	0.077	0.084	4.27	5.16	3.27	4.41	92
			(0.011)	(0.012)	(0.013)	(0.015)	(2.98)	(4.13)	(2.63)	(4.05)	
100	3	1	0.163	0.163	0.147	0.161	10.1	11.1	8.94	10.9	69
			(0.023)	(0.023)	(0.025)	(0.030)	(6.90)	(8.06)	(6.93)	(9.51)	

Table 4: Prediction errors and mse for the simulations for the functional linear regression (FLR, Eq. (2)) and reduced rank regression (RRR, Eq. (5)), for Example 4. The numbers in brackets are the standard errors computed from simulations.

			Prediction Error				MSE ($\times 10^{-4}$)			
n	j_0	σ	FLR	FLR-CV	RRR	RRR-CV	FLR	FLR-CV	RRR	RRR-CV
50	1	0.25	0.077	0.076	0.077	0.077	5.50	5.44	5.56	5.53
			(0.003)	(0.003)	(0.003)	(0.003)	(2.51)	(2.42)	(2.49)	(2.43)
50	1	0.5	0.135	0.133	0.135	0.134	10.4	10.3	10.3	10.7
			(0.008)	(0.008)	(0.008)	(0.007)	(6.21)	(6.02)	(5.91)	(5.84)
50	1	1	0.230	0.228	0.230	0.229	19.9	20.0	20.0	20.2
			(0.015)	(0.016)	(0.015)	(0.016)	(11.1)	(11.3)	(11.1)	(11.3)
50	3	0.25	0.085	0.085	0.086	0.085	4.53	4.60	4.54	4.60
			(0.004)	(0.004)	(0.004)	(0.004)	(2.06)	(2.04)	(2.10)	(2.12)
50	3	0.5	0.152	0.149	0.152	0.149	10.3	10.1	10.2	10.1
			(0.008)	(0.008)	(0.008)	(0.008)	(6.27)	(6.20)	(6.13)	(6.06)
50	3	1	0.265	0.262	0.265	0.263	25.6	25.9	25.6	26.1
			(0.014)	(0.016)	(0.014)	(0.016)	(20.2)	(19.8)	(20.1)	(20.0)
100	1	0.25	0.058	0.057	0.058	0.058	3.99	4.05	4.14	4.15
			(0.002)	(0.002)	(0.002)	(0.002)	(1.7)	(1.74)	(1.67)	(1.74)
100	1	0.5	0.101	0.099	0.101	0.099	7.42	7.61	7.40	7.60
			(0.004)	(0.004)	(0.004)	(0.004)	(4.09)	(4.21)	(3.97)	(4.06)
100	1	1	0.174	0.170	0.174	0.170	14.1	14.2	14.1	14.2
			(0.008)	(0.009)	(0.008)	(0.009)	(7.42)	(7.49)	(7.41)	(7.51)
100	3	0.25	0.063	0.063	0.064	0.064	3.22	3.22	3.25	3.30
			(0.002)	(0.002)	(0.002)	(0.002)	(1.28)	(1.25)	(1.52)	(1.35)
100	3	0.5	0.111	0.110	0.111	0.111	6.56	6.53	6.53	6.57
			(0.004)	(0.004)	(0.004)	(0.004)	(3.55)	(3.43)	(3.41)	(3.35)
100	3	1	0.197	0.195	0.197	0.195	15.3	15.0	15.0	15.1
			(0.008)	(0.009)	(0.008)	(0.009)	(10.4)	(9.82)	(10.3)	(9.81)

3.2.1. Canadian weather data

These daily data consist of daily temperature and precipitation measurements recorded in 35 Canadian weather stations. Each observation consists of functional data observed on an equally-spaced grid of 365 points. We treat temperature as the independent variable and the goal is to predict the corresponding precipitation curve given the temperature measurements. As was previously done in [25], we set the dependent variable to be the log-transformed precipitation measurements, and a small positive number is added to the values with 0 precipitation recorded. We pre-smooth data by approximating the functional variables using cubic splines with 6 internal knots and then discretize them using a grid of 100 points. Given the periodic nature of the data, we set $\mathcal{H}_K = \mathcal{W}_2^{\text{per}}$, the second-order Sobolev space of periodic functions on $[0, 1]$. The reproducing kernel is given by $K(s,t) = K_1(s,t) + K_2(s,t)$ with $K_1(s,t) = 1$ and

$$K_2(s,t) = \sum_{j=1}^{\infty} \frac{2}{(2\pi j)^4} \cos\{2\pi j(s-t)\}.$$

We use leave-one-out cross-validation to determine the best tuning parameters r and λ with $\ln\lambda \in \{-10, \ldots, -1\}$ and $r \in \{1, \ldots, 5\}$. The smallest cross-validation error for the reduced rank regression is achieved when $r = 3$. The estimated values of a_1, a_2, a_3 and b_1, b_2 and b_3 are shown in Figure 3. For both the reduced rank regression and the functional linear regression, the cross-validation error is 0.426. Since a_1 (black solid curve in the bottom-left panel of Figure 3) shows larger absolute values in the winter, this means the most significant effect of temperature on precipitation is on the weighted average precipitation level with winter months receiving more weights. For this data set, the estimated principal component functions for Y are very similar to a_j.

3.2.2. Gait data

The Motion Analysis Laboratory at Children's Hospital, San Diego, California, collected these data, which consist of the angles formed by the hip and knee of 39 children over each child's gait cycle. The cycle begins and ends at the point where the heel of the limb under observation strikes the ground. Both sets of functions are periodic and it is of interest to see how the two joints interact. In this application, we use hip angle as the predictor and knee angle as the response. The smallest cross-validation error for the reduced rank regression is achieved when $r = 4$. The estimated values of a_1, a_2, a_3, a_4 and b_1, b_2, b_3, b_4 are shown in Figure 4. For the reduced rank regression the cross-validation error is 4.20, slightly smaller than the error of 4.27 achieved by the functional linear regression. The shapes of a_1, b_1 (black solid curves) indicate that that the contrast between hip angle in the middle of the cycle and that at the beginning and ends is predictive of the contrast between knee angle in the middle of the cycle and that at the beginning and ends. For these data, the estimated principal component functions for Y are visually different from a_j and cannot be interpreted in this way.

4. Conclusion

In this paper, we considered reduced rank regression for functional regression models with functional responses. Instead of estimating a bivariate function, we only need to estimate a finite number of univariate functions. The model is thus more parsimonious and can potentially achieve higher estimation efficiency with finite sample sizes. We can further interpret the model as aiming to achieve dimension reduction for both the predictor and the response. We carried out simulation studies and further illustrated the methodology on two real datasets. Although on the real data there are no significant improvements on prediction accuracy, the reduced rank model allows us to interpret the estimated coefficients more easily than in general functional linear regression models.

Acknowledgments

The authors sincerely thank the Editor-in-Chief, Prof. C. Genest, FRSC, an Associate Editor and a reviewer for their insightful comments that led to a significant improvement of the paper. Jiang's research was partially supported by NSFC (11101432), Natural Science Foundation of Guangdong Province of China (2017A030313012), and Shenzhen Sci-Tech Fund (JCYJ20170307110329106). Lian's research is supported by Hong Kong RGC General Research Fund 11301718 and NSFC No. 11871411. Zhang's research is supported by NSFC 11671374 and 71631006.

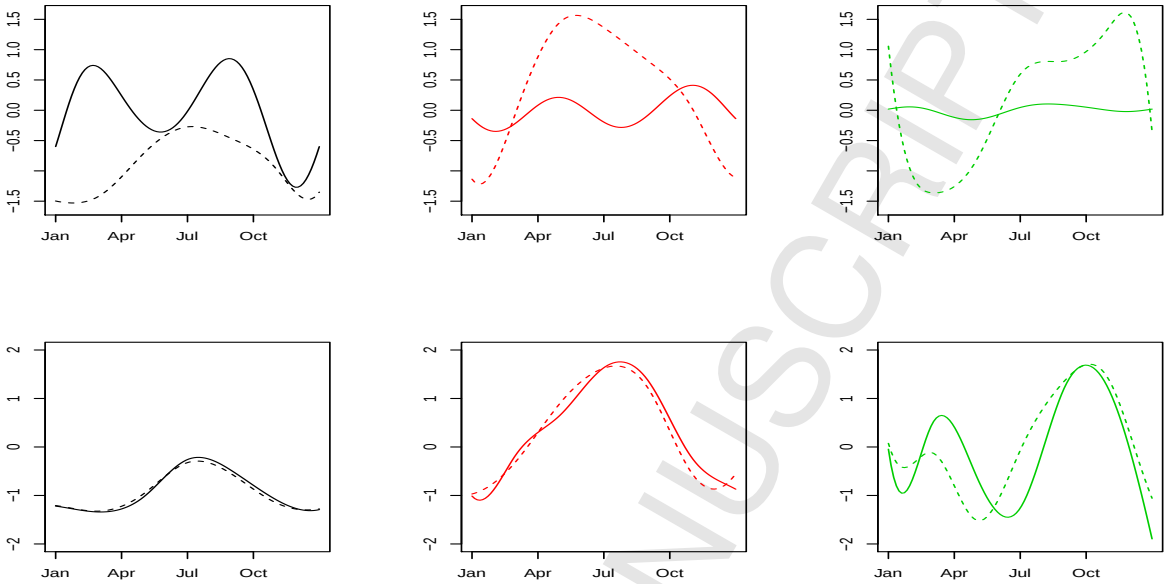

Figure 3: Estimated b_j (first row) and a_j (second row) functions for the gait data shown as solid curves. The dashed curves are the principal component functions for X and Y, respectively.

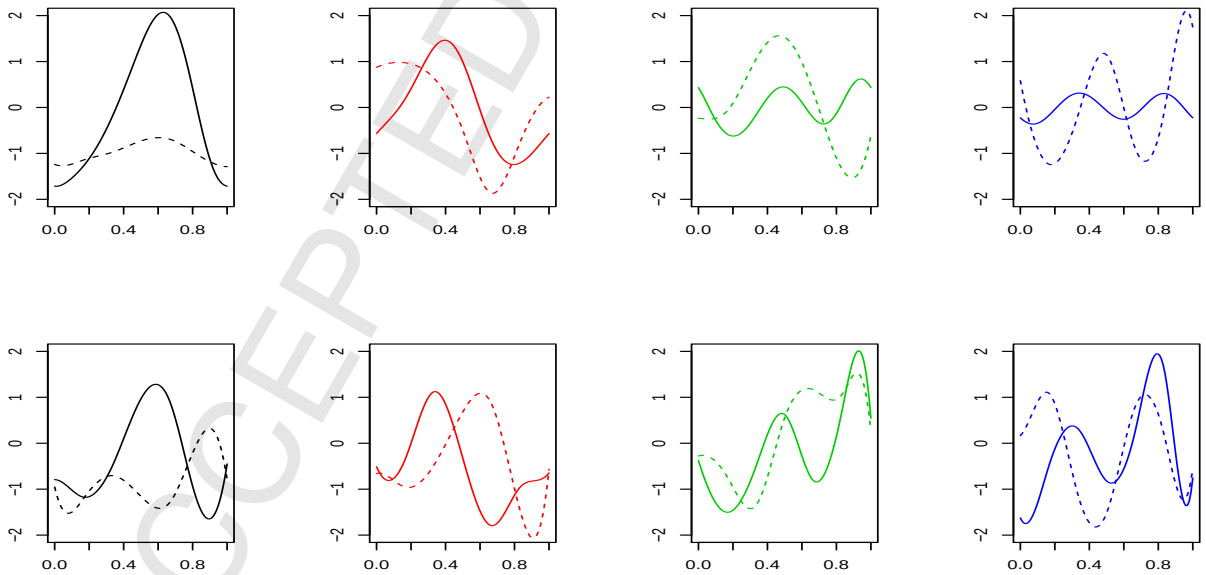

Figure 4: Estimated b_j (first row) and a_j (second row) functions for the gait data shown as solid curves. The dashed curves are the principal component functions for X and Y, respectively.

References

[1] A. Aguilera, F. Ocana, M. Valderrama, Estimation of functional regression models for functional responses by wavelet approximation, International Workshop on Functional and Operatorial Statistics, 2008.

[2] A. Ait-Saidi, F. Ferraty, R. Kassa, P. Vieu, Cross-validated estimations in the single-functional index model, Statistics 42 (2008) 475–494.

[3] T.W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann. Math. Statist. 29 (1951) 327–351.

[4] J. Antoch, L. Prchal, M.R. De Rosa, P. Sarda, Functional linear regression with functional response: Application to prediction of electricity consumption, In: Functional and Operatorial Statistics: Contributions to Statistics, S. Dabo-Niang, F. Ferraty (Eds.), Physica-Verlag HD, 2008.

[5] I. Berezans'kiĭ, Z. Sheftel, G. Us, Functional Analysis, vol. 1, Birkhäuser Verlag, Basel, 1996.

[6] T.T. Cai, P. Hall, Prediction in functional linear regression, Ann. Statist. 34 (2006) 2159–2179.

[7] T.T. Cai, M. Yuan, Minimax and adaptive prediction for functional linear regression, J. Amer. Statist. Assoc. 107 (2012) 1201–1216.

[8] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. 45 (1999) 11–22.

[9] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Ann. Statist. 37 (2009) 35–72.

[10] C. Crambes, A. Mas, Asymptotics of prediction in functional linear regression with functional outputs, Bernoulli 19 (2013) 2627–2651.

[11] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2002) 1–49.

[12] F. Ferraty, W. Gonzalez-Manteiga, A. Martínez-Calvo, P. Vieu, Presmoothing in functional linear regression, Statist. Sinica 22 (2011) 69–94.

[13] P. Hall, J.L. Horowitz, Methodology and convergence rates for functional linear regression, Ann. Statist. 35 (2007) 70–91.

[14] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, Wiley, Chichester, 2015.

[15] A.J. Izenman, Reduced-rank regression for the multivariate linear model, J. Multivariate Anal. 5 (1975) 248–264.

[16] H. Lian, Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces, Canad. J. Statist. 35 (2007) 597–606.

[17] H. Lian, Convergence of functional k-nearest neighbor regression estimate with functional responses, Electron. J. Stat. 5 (2011) 31–40.

[18] H. Lian, Minimax prediction for functional linear regression with functional responses in reproducing kernel Hilbert spaces, J. Multivariate Anal. 140 (2015) 395–402.

[19] A. van der Linde, Reduced rank regression models with latent variables in Bayesian functional data analysis, Bayesian Anal. 6 (2011) 77–126.

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

[21] J.S. Morris, C. Arroyo, B.A. Coull, L.M. Ryan, R. Herrick, S.L. Gortmaker, Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study, J. Amer. Statist. Assoc. 101 (2006) 1352–1364.

[22] J.S. Morris, R.J. Carroll, Wavelet-based functional mixed models, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 68 (2006) 179–199.

[23] H.-G. Müller, Y. Wu, F. Yao, Continuously additive models for nonlinear functional regression, Biometrika 100 (2013) 607–622.

[24] C. Preda, Regression models for functional data by reproducing kernel Hilbert spaces methods, J. Stat. Plann. Inf. 137 (2007) 829–840.

[25] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, 2nd ed., Springer, New York, 2005.

[26] G. Wahba, Spline Models for Observational Data, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[27] F. Yao, H.G. Müller, J.L. Wang, Functional linear regression analysis for longitudinal data, Ann. Statist. 33 (2005) 2873–2903.

[28] M. Yuan, T.T. Cai, A reproducing kernel Hilbert space approach to functional linear regression, Ann. Statist. 38 (2010) 3412–3444.

[29] H. Zhu, F. Yao, H.H. Zhang, Structured functional additive regression in reproducing kernel Hilbert spaces, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 76 (2014) 581–603.