JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 0, NO. 0, 1–11
https://doi.org/10.1080/01621459.2017.1356320

Taylor & Francis
Taylor & Francis Group

Optimal Penalized Function-on-Function Regression Under a Reproducing Kernel Hilbert Space Framework

Xiaoxiao Sun[a], Pang Du[b], Xiao Wang[c], and Ping Ma[a]

[a]Department of Statistics, University of Georgia, Athens, GA; [b]Department of Statistics, Virginia Tech, Blacksburg, VA; [c]Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

Many scientific studies collect data where the response and predictor variables are both functions of time, location, or some other covariate. Understanding the relationship between these functional variables is a common goal in these studies. Motivated from two real-life examples, we present in this article a function-on-function regression model that can be used to analyze such kind of functional data. Our estimator of the 2D coefficient function is the optimizer of a form of penalized least squares where the penalty enforces a certain level of smoothness on the estimator. Our first result is the Representer Theorem which states that the exact optimizer of the penalized least squares actually resides in a data-adaptive finite-dimensional subspace although the optimization problem is defined on a function space of infinite dimensions. This theorem then allows us an easy incorporation of the Gaussian quadrature into the optimization of the penalized least squares, which can be carried out through standard numerical procedures. We also show that our estimator achieves the minimax convergence rate in mean prediction under the framework of function-on-function regression. Extensive simulation studies demonstrate the numerical advantages of our method over the existing ones, where a sparse functional data extension is also introduced. The proposed method is then applied to our motivating examples of the benchmark Canadian weather data and a histone regulation study. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2016
Revised June 2017

KEYWORDS
Function-on-Function regression; Minimax convergence rate.; Penalized least squares; Representer theorem; Reproducing kernel Hilbert space

1. Introduction

Functional data have attracted much attention in the past decades (Ramsay and Silverman 2005). Most of the existing literature has only considered the regression models of a scalar response against one or more functional predictors, possibly with some scalar predictors as well. Some of them considered a reproducing kernel Hilbert space framework. For example, Yuan and Cai (2010) provided a thorough theoretical analysis of the penalized functional linear regression model with a scalar response. The article laid the foundation for several theoretical developments including the representer theorem and minimax convergence rates for prediction and estimation for penalized functional linear regression models. In a follow-up, Cai and Yuan (2012) showed that the minimax rate of convergence for the excess prediction risk is determined by both the covariance kernel and the reproducing kernel. Then they designed a data-driven roughness regularization predictor that can achieve the optimal convergence rate adaptively without the knowledge of the covariance kernel. Du and Wang (2014) extended the work of Yuan and Cai (2010) to the setting of a generalized functional linear model, where the scalar response comes from an exponential family distribution.

In contrast to these functional linear regression models with a scalar response, the model with a functional response $Y(t)$ over a functional predictor $X(s)$ has only been scarcely investigated (Ramsay and Silverman 2005; Yao, Müller, and Wang 2005b). Such data with functional responses and predictors are abundant in practice. We shall now present two motivating examples.

Example 1. Canadian Weather Data. Daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994 were collected (Figure 1). The main interest is to use the daily temperature profile to predict the daily precipitation profile for a location in Canada.

Example 2. Histone Regulation Data. Extensive researches have been shown that histone variants, that is, histones with structural changes compared to their primary sequence, play an important role in the regulation of chromatin metabolism and gene activity (Ausió 2006). An ultra-high throughput time course experiment was conducted to study the regulation mechanism during heat stress in *Arabidopsis thaliana*. The genome-wide histone variant distribution was measured by ChIP sequencing (ChIP-seq; Johnson et al. 2007) experiments. We computed histone levels over 350 base pairs (bp) on genomes from the ChIP-seq data, see left panel in Figure 2 . The RNA sequencing (RNA-seq; Wang, Gerstein, and Snyder 2009) experiments measured the expression levels over seven time points within 24 hr, see right panel in Figure 2. Of primary interest is to study the regulation mechanism between gene expression levels over time domain and histone levels over spatial domain.

CONTACT Pang Du pangdu@vt.edu Department of Statistics, Virginia Tech, Blacksburg, VA 24061.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

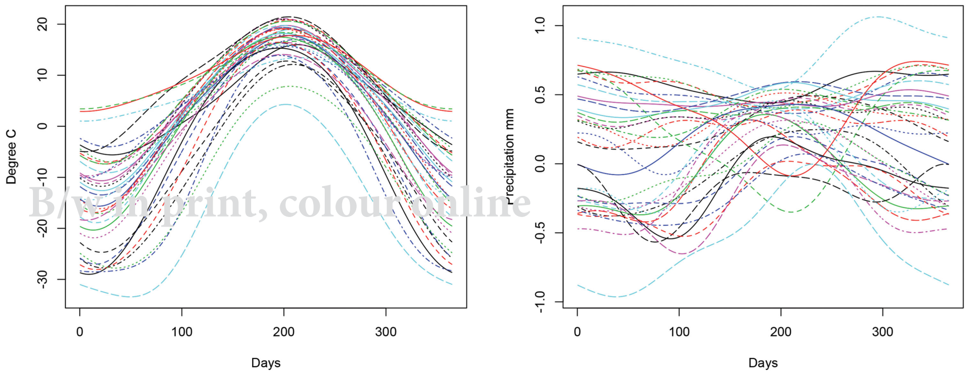


Figure 1. Smoothed trajectories of temperature (celsius) in left panel and the log (base 10) of daily precipitation (millimeter) in right panel. The *x*-axis labels in both panels represent 365 days.

Motivated by the examples, we now present the statistical model. Let $\{(X(s), Y(t)) : s \in I_x, t \in I_y\}$ be two random processes defined, respectively, on $I_x, I_y \subseteq \mathbb{R}$. Suppose n independent copies of (X, Y) are observed: $(X_i(s), Y_i(t))$, $i = 1, \ldots, n$. The functional linear regression model of interest is

$$Y_i(t) = \alpha(t) + \int_{I_x} \beta(t, s) X_i(s) ds + \epsilon_i(t), \quad t \in I_y, \qquad (1)$$

where $\alpha(\cdot) : I_y \to \mathbb{R}$ is the intercept function, $\beta(\cdot, \cdot) : I_y \times I_x \to \mathbb{R}$ is a bivariate coefficient function, and $\epsilon_i(t)$, independent of $X_i(s)$, are iid random error functions with $\mathbb{E}\epsilon_i(t) = 0$ and $\mathbb{E}\|\epsilon_i(t)\|_2^2 < \infty$. Here $\|\cdot\|_2$ denotes the L_2-norm. In Example 1, $Y_i(t)$ and $X_i(t)$ represent the daily precipitation and temperature at station i. In Example 2, the expression levels of gene i over seven time points, $Y_i(t)$, from RNA-seq is used as the functional response. The histone levels of gene i over 350 base pairs (bp), $X_i(s)$, from ChIP-seq is used as the functional predictor.

At a first look, model (1) might give the (wrong) impression of being an easy extension from the model with a scalar response, with the latter obtained from (1) by removing all the t notation. However, the coefficient function in the scalar response case is univariate and thus can be easily estimated by most off-the-shelf smoothing methods. When extended to estimating a bivariate coefficient function $\beta(t, s)$ in (1), many of these smoothing methods may encounter major numerical and/or theoretical difficulties. This partly explains the much less abundance of research in this direction.

Some exceptions though are reviewed below. Cuevas, Febrero, and Fraiman (2002) considered a fixed design case, a different setting from (1) with $Y_i(t)$ and $X_i(s)$ represented and analyzed as sequences. Nonetheless they provided many motivating applications in neuroscience, signal transmission, pharmacology, and chemometrics, where (1) can apply. The historical functional linear model in Malfait and Ramsay (2003) was among the first to study regression of a response functional variable over a predictor functional variable, or more precisely, the history of the predictor function. Ferraty et al. (2011) proposed a simple extension of the classical Nadaraya–Watson estimator to the functional case and derived its convergence rates. They provided no numerical results on the empirical performance of their kernel estimator. Benatia, Carrasco, and Florens (2015) extended ridge regression to the functional setting. However, their estimation relied on an empirical estimate of the covariance process of predictor functions. Theoretically sound as it is, this covariance process estimate is generally not reliable in practice. Consequently, their coefficient surface estimates suffered as shown in their simulation plots. Meyer et al. (2015) proposed a Bayesian function-on-function regression model for multi-level functional data, where the basis expansions of functional parameters were regularized by basis-space prior distributions and a random effect function was introduced to incorporate the with-subject correlation between functional observations.

A popular approach has been the functional principal component analysis (FPCA) as in Yao, Müller, and Wang (2005b) and Crambes and Mas (2013). The approach starts with a

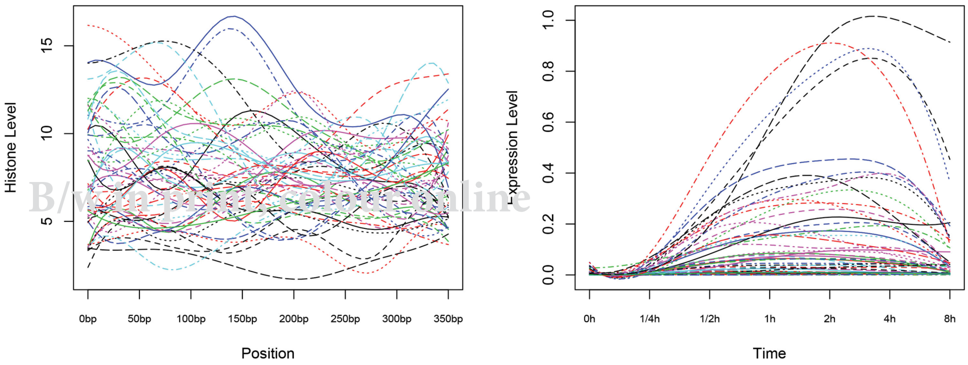


Figure 2. Smoothed trajectories of normalized histone levels in ChIP-seq experiments in left panel and the normalized expression levels in RNA-seq experiments in right panel. The *x*-axis label in the left panel stands for the region of 350 bp. The *x*-axis label in the right panel represents seven time points within 24 hr.

basis representation of $\beta(t, s)$ in terms of the eigenfunctions in the Karhunen–Loève expansions of $Y(t)$ and $X(s)$. Since this representation has infinitely many terms, it is truncated at certain point to obtain an estimable basis expansion of $\beta(t, s)$. Yao, Müller, and Wang (2005b) studied a general data setting where $Y(t)$ and $X(s)$ are only sparsely observed at some random points. They derived the consistency and proposed asymptotic point-wise confidence bands for predicting response trajectories. Crambes and Mas (2013) furthered the theoretical investigation of the FPCA approach by providing a minimax optimal rates in terms of the mean square prediction error. However, the FPCA approach has a couple of critical drawbacks. First, $\beta(t, s)$ is a statistical quantity unrelated to $Y(t)$ or $X(s)$. Hence, the leading eigenfunctions in the *truncated* Karhunen–Loève expansions of $Y(t)$ and $X(s)$ may not be an effective basis for representing $\beta(t, s)$. See, for example, Cai and Yuan (2012) and Du and Wang (2014) for some scalar-response examples where the FPCA approach breaks down when the aforementioned situation happens. Second, the truncation point is integer-valued and thus only has a discrete control on the model complexity. This puts it at disadvantage against the roughness penalty regularization approach, which offers a continuous control via a positive and real-valued smoothing parameter (Ramsay and Silverman 2005, chap. 5).

In this article, we consider a penalized function-on-function regression approach to estimating the bivariate coefficient function $\beta(t, s)$. There have been a few recent developments in the direction of penalized function-on-function regression. Lian (2015) studied the convergence rates of the function-on-function regression model under a reproducing kernel Hilbert space framework. Although his model resembled model (1), he developed everything with the variable t fixed and did not enforce any regularization on the t direction. First, this lack of t-regularization can be problematic since this leaves the noisy errors on the t direction completely uncontrolled and can result in a $\beta(s, t)$ estimate that is very rough on the t direction. Second, this simplification of fixing t essentially reduces the problem to a functional linear model with a scalar response and thus makes all the results in Yuan and Cai (2010) directly transferrable even without calling on any new proofs. In the R package `fda`, Ramsay and his collaborators have implemented a version of penalized B-spline estimation of $\beta(t, s)$ with a fixed smoothing parameter. Ivanescu et al. (2015) considered a penalized function-on-function regression model where the coefficient functions were represented by expansions into some basis system such as tensor cubic B-splines. Quadratic penalties on the expansion coefficients were used to control the smoothness of the estimates. This work provided a nice multiple-predictor-function extension to the function-on-function regression model in the `fda` package. Scheipl and Greven (2016) studied the identifiability issue in these penalized function-on-function regression models. However, this penalized B-spline approach has several well-known drawbacks. First, it is difficult to show any theoretical optimality such as the minimax risk of mean prediction in Cai and Yuan (2012). So its theoretical soundness is hard to justify. Moreover, the B-spline expansion is only an approximate solution to the optimization of the penalized least-square score. Hence, the penalized B-spline estimate is not numerically optimal from the beginning either. These drawbacks can have negative impacts on the numerical performance as we shall see from the simulation results in Section 4.

The penalized function-on-function regression method proposed in this article obtains its estimator of $\beta(t, s)$ through the minimization of penalized least squares on a reproducing kernel Hilbert space that is naturally associated with the roughness penalty. Such a natural formulation through a reproducing kernel Hilbert space offers several advantages comparing with the existing penalized function-on-function regression methods. First, it allows us to establish a Representer Theorem which states that, although the optimization of the penalized least squares is defined on an infinite-dimensional function space, its solution actually resides in a data-adaptive finite-dimensional subspace. This result guarantees an exact solution when the optimization is carried out on this finite-dimensional subspace. This result itself is a nontrivial generalization of the Representer Theorems in the scenarios of nonparametric smooth regression model (Wahba 1990) and the penalized functional regression model with a scalar response (Yuan and Cai 2010). Based on the Representer Theorem, we propose an estimation algorithm which uses penalized least squares and Gaussian quadrature with the Gauss–Legendre rule to estimate the bivariate coefficient function. The smoothing parameter is selected by the generalized cross-validation (GCV) method. Second, the reproducing kernel Hilbert space framework allows us to show that our estimator has the optimal rate of mean prediction since it achieves the minimax convergence rate in terms of the excess risk. This generalizes the results in Cai and Yuan (2012) and Du and Wang (2014) for functional linear regression with a scalar response to the functional response scenario. In the numerical study, we have also considered the problem with sparsely sampled data. Particularly, we introduce an extra presmoothing step before applying the proposed penalized functional regression model. The presmoothing step implements the principal-component-analysis-through-expectation (PACE) method in Yao, Müller, and Wang (2005a). Our extensive simulation studies demonstrate the numerical advantages of our method over the existing ones. In summary, our method has the following distinguishing features: (i) it makes no structural dependence assumptions of $\beta(t, s)$ over the predictor and response processes; (ii) the Representer Theorem guarantees an exact solution instead of an approximation to the optimization of the penalized score; (iii) benefited from the Representer Theorem, we develop a numerically reliable algorithm that has sound performance in simulations; (iv) we show theoretically the estimator achieves the optimal minimax convergence rate in mean prediction.

The rest of the article is organized as follows. In Section 2, we first derive the Representer Theorem showing that the solution of the minimization of penalized least squares can be found in a finite-dimension subspace. In addition, an easily implementable estimation algorithm is considered in Section 2. In Section 3, we prove that our method has the optimal rate of mean prediction. Numerical experiments are reported in Section 4, where we compare our method with the functional linear regression models in Ramsay and Silverman (2005) and Yao, Müller, and Wang (2005b) in terms of prediction accuracy.

Two real data examples, the Canadian weather data, and the histone regulation data are analyzed in Section 5. Discussion in Section 6 concludes the article. Proofs of the theorems are collected in supplementary material.

2. Penalized Functional Linear Regression Method

We first introduce a simplification to model (1). Since model (1) implies that

$$Y_i(t) - \mathbb{E}Y_i(t) = \int_{I_x} \beta(t,s)\{X_i(s) - \mathbb{E}X_i(s)\}ds + \epsilon_i(t), \quad t \in I_y,$$

we may, for simplicity, only consider X and Y to be centered, that is, $\mathbb{E}X = \mathbb{E}Y = 0$. Thus, the functional linear regression model takes the form of

$$Y_i(t) = \int_{I_x} \beta(t,s)X_i(s)ds + \epsilon_i(t), \quad t \in I_y. \tag{2}$$

2.1. The Representer Theorem

Assume that the unknown β resides in a reproducing kernel Hilbert space $\mathcal{H}(K)$ with the reproducing kernel $K : I \times I \to \mathbb{R}$, where $I = I_y \times I_x$. The estimate $\hat{\beta}_n$ can be obtained by minimizing the following penalized least-square functional

$$\frac{1}{n}\sum_{i=1}^{n}\int_{I_y}\left\{Y_i(t) - \int_{I_x}\beta(t,s)X_i(s)ds\right\}^2 dt + \lambda J(\beta) \tag{3}$$

with respect to $\beta \in \mathcal{H}(K)$, where the sum of integrated squared errors represents the goodness of fit, J is a roughness penalty on β, and $\lambda > 0$ is the smoothing parameter balancing the trade-off. When β is a univariate function, a common example for J is $J(\beta) = \int\{\beta''(t)\}^2dt$, the integral of the squared curvature of β. This integral takes a large value when β is rough and has high curvatures. When β is a bivariate function as considered in this article, J is often a combination of multiple integrals, each representing the roughness of a certain part of β; see Example 3. We now establish the Representer Theorem stating that $\hat{\beta}_n$ actually resides in a finite-dimensional subspace of $\mathcal{H}(K)$. This result generalizes Theorem 1 in Yuan and Cai (2010) and facilitates the computation by reducing an infinite-dimensional optimization problem to a finite-dimensional one.

Note that the penalty functional J is a squared seminorm on $\mathcal{H}(K)$. Its null space $\mathcal{H}_0 = \{\beta \in \mathcal{H}(K) : J(\beta) = 0\}$ is a finite-dimensional linear subspace of $\mathcal{H}(K)$. Denote by $\mathcal{H}_1$ its orthogonal complement in $\mathcal{H}(K)$ such that $\mathcal{H}(K) = \mathcal{H}_0 \oplus \mathcal{H}_1$, the *tensor sum* or *direct sum* of $\mathcal{H}_0$ and $\mathcal{H}_1$. That is, for any $\beta \in \mathcal{H}(K)$, there exists a unique decomposition $\beta = \beta_0 + \beta_1$ where $\beta_0 \in \mathcal{H}_0$ and $\beta_1 \in \mathcal{H}_1$. Let $K_0(\cdot,\cdot)$ and $K_1(\cdot,\cdot)$ be the corresponding reproducing kernels of $\mathcal{H}_0$ and $\mathcal{H}_1$. Then K_0 and K_1 are both nonnegative definite operators on L_2, and $K = K_0 + K_1$. In fact, the penalty term $J(\beta) = \|\beta\|^2_{K_1} = \|\beta_1\|^2_{K_1}$. By the theory of reproducing kernel Hilbert spaces, $\mathcal{H}(K)$ has a tensor product decomposition $\mathcal{H}(K) = \mathcal{H}_y(K_y) \otimes \mathcal{H}_x(K_x)$. That is, given the respective bases $\{f_1, f_2, \ldots,\}$ and $\{g_1, g_2, \ldots,\}$ of $\mathcal{H}_y(K_y)$ and $\mathcal{H}_x(K_x)$ any function $\beta(t,s) \in \mathcal{H}(K)$ can be uniquely written as $\beta(t,s) = \sum_j c_j f_j(t)g_j(s)$ for some coefficients c_j. Here $\mathcal{H}_y(K_y)$ is the reproducing kernel Hilbert space with a reproducing kernel $K_y : I_y \times I_y \to \mathbb{R}$, and $\mathcal{H}_x(K_x)$ is the reproducing kernel Hilbert space with a reproducing kernel $K_x : I_x \times I_x \to \mathbb{R}$. For the reproducing kernels, we have $K(t,s) = K_y(t)K_x(s)$. Note that the functions in $\mathcal{H}_y(K_y)$ and $\mathcal{H}_x(K_x)$ are univariate and defined, respectively, on I_y and I_x. Similar to the decomposition of $\mathcal{H}$ and K, we have the tensor sum decompositions of the marginal subspaces $\mathcal{H}_y(K_y) = \mathcal{H}_{0y} \oplus \mathcal{H}_{1y}$ and $\mathcal{H}_x(K_x) = \mathcal{H}_{0x} \oplus \mathcal{H}_{1x}$, and the orthogonal decompositions of the marginal reproducing kernels $K_y = K_{0y} + K_{1y}$ and $K_x = K_{0x} + K_{1x}$. Here K_* is a reproducing kernel on $\mathcal{H}_*$ with $*$ running through the index set $\{0y, 1y, 0x, 1x\}$.

Upon piecing the marginal decomposition parts back to the tensor product space, we obtain $\mathcal{H}_0 = \mathcal{H}_{0y} \otimes \mathcal{H}_{0x}$ and $\mathcal{H}_1 = (\mathcal{H}_{0y} \otimes \mathcal{H}_{1x}) \oplus (\mathcal{H}_{1y} \otimes \mathcal{H}_{0x}) \oplus (\mathcal{H}_{1y} \otimes \mathcal{H}_{1x})$. Correspondingly, the reproducing kernels satisfy that

$$\begin{aligned}
K_0((t_1,s_1),(t_2,s_2)) &= K_{0y}(t_1,t_2)K_{0x}(s_1,s_2),\\
K_1((t_1,s_1),(t_2,s_2)) &= K_{0y}(t_1,t_2)K_{1x}(s_1,s_2) + K_{1y}(t_1,t_2)K_{0x}\\
&\quad \times (s_1,s_2) + K_{1y}(t_1,t_2)K_{1x}(s_1,s_2).
\end{aligned}$$

Let $N_y = \dim(\mathcal{H}_{0y})$ and $N_x = \dim(\mathcal{H}_{0x})$. Denote by $\{\psi_{k,y} : k = 1,\ldots,N_y\}$ and $\{\psi_{l,x} : l = 1,\ldots,N_x\}$, respectively, the basis functions of $\mathcal{H}_{0y}$ and $\mathcal{H}_{0x}$. With some abuse of notation, define $(K_{1y}g)(\cdot) = \int_{I_y} K_{1y}(\cdot,t)g(t)dt$ and $(K_{1x}f)(\cdot) = \int_{I_x} K_{1x}(\cdot,s)f(s)ds$. Now we can state the Representer Theorem as follows with its proof collected in the supplementary material.

Theorem 1. Let $\hat{\beta}_n$ be the minimizer of (3) in $\mathcal{H}(K)$. Then $\hat{\beta}_n$ resides in the subspace of functions of the form

$$\begin{aligned}
\beta(t,s) &= \left\{\sum_{k=1}^{N_y} d_{k,\beta_y}\psi_{k,y}(t) + \sum_{i=1}^{n} c_{i,\beta_y}(K_{1y}Y_i)(t)\right\}\\
&\quad \times \left\{\sum_{l=1}^{N_x} d_{l,\beta_x}\psi_{l,x}(s) + \sum_{j=1}^{n} c_{j,\beta_x}(K_{1x}X_j)(s)\right\}\\
&= \left\{d^\top_{\beta_y}\psi_y(t) + c^\top_{\beta_y}(K_{1y}Y)(t)\right\}\left\{d^\top_{\beta_x}\psi_x(s) + c^\top_{\beta_x}(K_{1x}X)(s)\right\},
\end{aligned} \tag{4}$$

where $d_{\beta_y} = (d_{1,\beta_y},\ldots,d_{N_y,\beta_y})^\top$, $c_{\beta_y} = (c_{1,\beta_y},\ldots,c_{n,\beta_y})^\top$, $d_{\beta_x} = (d_{1,\beta_x},\ldots,d_{N_x,\beta_x})^\top$, and $c_{\beta_x} = (c_{1,\beta_x},\ldots,c_{n,\beta_x})^\top$ are some coefficient vectors, and ψ_x, ψ_y, $K_{1y}Y$ and $K_{1x}X$ are vectors of functions.

For the purpose of illustration, we give a detailed example below.

Example 3. Consider the case of tensor product cubic splines with $I_y = I_x = [0,1]$. The marginal spaces $\mathcal{H}_y(K_y) = \mathcal{H}_x(K_x) = \{g : \int_0^1 (g'')^2 < \infty\}$ with the inner product

$$\langle f,g\rangle_{\mathcal{H}_y} = \left(\int_0^1 f\int_0^1 g + \int_0^1 f'\int_0^1 g'\right) + \int_0^1 f''g''dt.$$

The marginal space $\mathcal{H}_y(K_y)$ can be further decomposed into the tensor sum of $\mathcal{H}_{0y} = \{g : g'' = 0\}$ and $\mathcal{H}_{1y} = \{g : \int_0^1 g = \int_0^1 g' = 0, \int_0^1 (g'')^2 < \infty\}$. The reproducing kernel K_y is the orthogonal sum of $K_{0y}(t_1,t_2) = 1 + r_1(t_1)r_1(t_2)$ and $K_{1y}(t_1,t_2) = r_2(t_1)r_2(t_2) - r_4(|t_1 - t_2|)$, where $r_\nu(t) = B_\nu(t)/\nu!$ is a scaled version of the Bernoulli polynomial B_ν. The space $\mathcal{H}_{0y}$ has a dimension of $N_y = 2$ and a set of basis functions $\{1, r_1(t)\}$.

The function space $\mathcal{H}(K)$ is defined as $\mathcal{H}(K) = \{\beta : J(\beta) < \infty\}$ with the reproducing kernel $K(t, s) = K_y(t)K_x(s)$ and the penalty functional

$$\begin{aligned} J(\beta) = &\int_0^1 \left[\left\{\int_0^1 \frac{\partial^2}{\partial s^2}\beta(t, s)dt\right\}^2 + \left\{\int_0^1 \frac{\partial^3}{\partial t\partial s^2}\beta(t, s)dt\right\}^2\right]ds \\ &+ \int_0^1 \left[\left\{\int_0^1 \frac{\partial^2}{\partial t^2}\beta(t, s)ds\right\}^2 + \left\{\int_0^1 \frac{\partial^3}{\partial t^2\partial s}\beta(t, s)ds\right\}^2\right]dt \\ &+ \int_0^1\int_0^1 \left\{\frac{\partial^4}{\partial t^2\partial s^2}\beta(t, s)\right\}^2 dtds. \end{aligned}$$

Intuitively, these five integrals represent, respectively, the deviations of the function β from being linear in s, being linear in s and constant in t, being linear in t, being constant in s, and linear in t, and being linear in both s and t. And we have $\mathcal{H}(K) = \mathcal{H}_y(K_y) \otimes \mathcal{H}_x(K_x)$ and $K = K_yK_x$; see, for example, chap. 2 of Gu (2013).

2.2. Estimation Algorithm

To introduce the computational algorithm, we first need some simplification of notation. Let $N = N_yN_x$ and $L = n(N_y + N_x + n)$. We rewrite the functions spanning the subspace in Theorem 1 as $\psi_1(t, s) = \psi_{1,y}(t)\psi_{1,x}(s), \dots,$ $\psi_N(t, s) = \psi_{N_y,y}(t)\psi_{N_x,x}(s)$ and $\xi_1(t, s) = \psi_{1,y}(t)(K_{1x}X_1)(s)$ $, \dots, \xi_L(t, s) = (K_{1y}Y_n)(t)(K_{1x}X_n)(s)$. Thus, a function in this subspace has the form $\beta(t, s) = \mathbf{d}^T\psi(t, s) + \mathbf{c}^T\xi(t, s)$ for some coefficient vectors $\mathbf{d}$, $\mathbf{c}$ and vectors of functions $\psi(t, s)$, $\xi(t, s)$. To solve (3), we choose Gaussian quadrature with the Gauss–Legendre rule to calculate the integrals. Consider the Gaussian quadrature evaluation of an integral on I_y with knots $\{t_1, \dots, t_T\}$ and weights $\{\alpha_1, \dots, \alpha_T\}$ such that $\int_{I_y} f(t)dt = \sum_{j=1}^T \alpha_j f(t_j)$. Let W be the diagonal matrix with $\alpha_1, \dots, \alpha_T$ repeating n times on the diagonal. Then the estimation of β in (3) reduces to the minimization of

$$(Y_w - S_w\mathbf{d} - R_w\mathbf{c})^T(Y_w - S_w\mathbf{d} - R_w\mathbf{c}) + n\lambda\mathbf{c}^TQ\mathbf{c} \qquad (5)$$

with respect to $\mathbf{d}$ and $\mathbf{c}$, where $Y_w = W^{1/2}Y$ with $Y = (Y_1(t_1), \dots, Y_1(t_T), \dots, Y_n(t_1), \dots, Y_n(t_T))^\top$, $S_w = W^{1/2}S$ with S being an $nT \times N$ matrix with the $((i-1)T + j, \nu)$th entry $\int_{I_x} \psi_\nu(t_j, s)X_i(s)ds$, $R_w = W^{1/2}R$ with R being an $nT \times L$ matrix with the $((i-1)T + j, k)$th entry $\int_{I_x} \xi_k(t_j, s)X_i(s)ds$, and Q is a $L \times L$ matrix with the (i, j)th entry $\langle\xi_i, \xi_j\rangle_{\mathcal{H}_1}$. Let $Q_x = [\int_0^1\int_0^1 X_i(u)K(u, v)X_j(v)dudv]_{i,j=1}^n$, $Q_y = [\int_0^1\int_0^1 Y_i(u)K(u, v)Y_j(v)dudv]_{i,j=1}^n$, and $Q_{xy} = Q_x \bigotimes Q_y$, where $\bigotimes$ denotes the Kronecker product of two matrices. Then we have $Q = \text{diag}(Q_x, Q_x, Q_y, Q_y, Q_{xy})$.

We then use standard numerical linear algebra procedures, such as the Cholesky decomposition with pivoting and forward and back substitutions, to calculate $\mathbf{c}$ and $\mathbf{d}$ in (5) (Gu 2013, sec. 3.5). To choose the smoothing parameter λ in (5), a modified generalized cross-validation (GCV) score (Wahba and Craven 1979),

$$V(\lambda) = \frac{(nT)^{-1}Y_w^T(I - A(\lambda))^2Y_w}{\{(nT)^{-1}tr(I - \alpha A(\lambda))\}^2} \qquad (6)$$

is implemented, where $\alpha > 1$ is a fudge factor curbing undersmoothing (Kim and Gu 2004) and $A(\lambda)$ is the smoothing matrix bridging the prediction $\hat{Y}_w$ and the observation Y_w as $\hat{Y}_w = A(\lambda)Y_w$, similar to the hat matrix in a general linear model.

3. Optimal Mean Prediction Risk

We are interested in the estimation of coefficient function β and mean prediction, that is, to recover the functional $\eta_\beta(X, \cdot) = \int_{I_x}\beta(\cdot, s)X(s)ds$ based on the training sample (X_i, Y_i), $i = 1, \dots, n$. Let $\hat{\beta}_n(t, s)$ be an estimate of $\beta(t, s)$. Suppose (X_{n+1}, Y_{n+1}) is a new observation that has the same distribution as and is also independent of (X_i, Y_i), $i = 1, \dots, n$. Then the prediction accuracy can be naturally measured by the excess risk

$$\begin{aligned} &\mathfrak{R}_n(\hat{\beta}_n) = \\ &\int_{I_y}\left[\mathbb{E}^*\left\{Y_{n+1}(t) - \int_{I_x}\hat{\beta}_n(t, s)X_{n+1}(s)ds\right\}^2\right. \\ &\qquad \left. - \mathbb{E}^*\left\{Y_{n+1}(t) - \int_{I_x}\beta(t, s)X_{n+1}(s)ds\right\}^2\right]dt \\ &= \int_{I_y}\mathbb{E}^*\left\{\eta_{\hat{\beta}_n}(X_{n+1}, t) - \eta_\beta(X_{n+1}, t)\right\}^2 dt, \end{aligned}$$

where $\mathbb{E}^*$ represents the expectation taken over (X_{n+1}, Y_{n+1}) only. We shall study the convergence rate of $\mathfrak{R}_n$ as the sample size n increases.

This section collects two theorems whose combination indicates that our estimator achieves the optimal minimax convergence rate in mean prediction. We first establish the minimax lower bound for the convergence rate of the excess risk $\mathfrak{R}_n$. There is a one-to-one relationship between K and $\mathcal{H}(K)$ which is a linear functional space endowed with an inner product $\langle\cdot, \cdot\rangle_{\mathcal{H}(K)}$ such that

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

The kernel K can also be treated as an integral operator such that

$$K(\beta)(\cdot) = \langle K((t, s), \cdot), \beta\rangle_{L_2} = \int\int_I K((t, s), \cdot)\beta(t, s)dtds.$$

It follows from the spectral theorem that there exist a set of orthonormal eigenfunctions $\{\zeta_k : k \geq 1\}$ and a sequence of eigenvalues $\kappa_1 \geq \kappa_2 \geq \cdots > 0$ such that

$$K((t_1, s_1), (t_2, s_2)) = \sum_{k=1}^\infty \kappa_k\zeta_k(t_1, s_1)\zeta_k(t_2, s_2),$$
$$K(\zeta_k) = \kappa_k\zeta_k, k = 1, 2, \dots.$$

Denote $K^{1/2}\big((t_1, s_1), (t_2, s_2)\big) = \sum_{k=1}^\infty \kappa_k^{1/2}\zeta_k(t_1, s_1)\zeta_k(t_2, s_2)$. Let $C(t, s) = \text{cov}\big(X(t), X(s)\big)$ be the covariance kernel of X. Define a new kernel Π such that

$$\begin{aligned} \Pi((t_1, s_1), (t_2, s_2)) = \int\int\int_{I_x\times I_x\times I_y} &K^{1/2}((t_1, s_1), (z, u)) \\ &\times C(u, v)K^{1/2}((t_2, s_2), (z, v))dudvdz. \end{aligned} \qquad (7)$$

Let $\rho_1 \geq \rho_2 \geq \cdots > 0$ be the eigenvalues of Π and $\{\phi_j : j \geq 1\}$ be the corresponding eigenfunctions. Therefore,

$$\Pi\big((t_1, s_1), (t_2, s_2)\big) = \sum_{k=1}^{\infty} \rho_k \phi_k(t_1, s_1)\phi_k(t_2, s_2), \quad \forall\, (t_1, s_1), (t_2, s_2) \in I_y \times I_x.$$

Theorem 2. Assume that for any $\beta \in L_2([0, 1]^2)$

$$\int \mathbb{E}\Big(\int \beta(t, s)X(s)dt\Big)^4 dt \leq c \int \Big(\mathbb{E}\Big(\int \beta(t, s)X(s)ds\Big)^2\Big)^2 dt \tag{8}$$

for a positive constant c. Suppose that the eigenvalues $\{\rho_k : k \geq 1\}$ of the kernel Π in (7) satisfy $\rho_k \asymp k^{-2r}$ for some constant $0 < r < \infty$. Then,

$$\lim_{A\to\infty} \lim_{n\to\infty} \sup_{\beta\in\mathcal{H}(K)} \mathbb{P}\{\mathfrak{R}_n \geq An^{-\frac{2r}{2r+1}}\} = 0, \tag{9}$$

when λ is of order $n^{-2r/(2r+1)}$.

Theorem 2 indicates that the convergence rate is determined by the decay rate of the eigenvalues of this new operator Π, which is jointly determined by both reproducing kernel K and the covariance kernel C as well as the alignment between K and C in a complicated way. This result has not been reported in the literature before. A closely related result is from Yuan and Cai (2010) who studied an optimal prediction risk for functional linear models, where the optimal rate depends on the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$. It is interesting to see, on the other hand, whether the convergence rate of $\hat{\beta}_n$ in Theorem 2 is optimal. In the following, we derive a minimax lower bound for the risk.

Theorem 3. Let r be as in Theorem 2. Then the excess prediction risk satisfies

$$\lim_{c\to 0} \lim_{n\to\infty} \inf_{\tilde{\eta}} \sup_{\beta\in\mathcal{H}(K)} \mathbb{P}(\mathfrak{R}_n \geq cn^{-\frac{2r}{2r+1}}) = 1, \tag{10}$$

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

Theorem 3 shows that the minimax lower bound of the convergence rate for the prediction risk is $n^{-2r/2r+1}$, which is determined by r and the decay rate of the eigenvalues of Π. We have shown that this rate is achieved by our penalized estimator, and therefore our estimator is rate-optimal.

4. Numerical Experiments

We compared the proposed optimal penalized function-on-function regression (OPFFR) method with existing function-on-function linear regression models under two different designs. In a dense design, each curve was densely sampled at regularly spaced common time points. We compared the OPFFR with two existing models. In a sparse design, each curve was irregularly and sparsely sampled at possibly different time points. We extended the OPFFR to this design by adding an extra presmoothing step and compared it with the FPCA model. In the first model (Ramsay and Silverman 2005) for comparison, the coefficient function is estimated by penalizing its B-spline basis function expansion. This approach does not have the optimal mean prediction property and partially implemented in the `fda` package of R (`linmod` function) for the case of a fixed smoothing parameter. We shall add a search on the grid $10^{(-2:0.4:2)}$ for smoothing parameter selection to their implementation and denote this augmented approach by FDA. The coefficient function is represented in terms of 10 basis functions each for the t and s directions. The second model for comparison was the functional principal component analysis (hence denoted by FPCA) approach proposed by Yao, Müller, and Wang (2005b). The coefficient function is represented in terms of the leading functional principal components. This is implemented in the MatLab package `PACE` (`FPCreg` function) maintained by the UC-Davis research group. The Akaike information criterion (AIC) and fraction of variance explained (FVE) criterion were used to select the number of principal components for predictor and response, respectively. The cutoff value for FVE was 0.9. The "regular" parameter was set to 2 for the dense design and 0 for the sparse design. No binning was performed.

4.1. Simulation Study

4.1.1. Dense Design

We simulated data according to model (2) with three scenarios.

- Scenario 1: The predictor functions are $X_i(s) = \sum_{k=1}^{50}(-1)^{(k+1)}k^{-1}Z_{ik}\vartheta_1(s, k)$, where Z_{ik} is from the uniform distribution $U(-\sqrt{3}, \sqrt{3})$, and $\vartheta_1(s, k) = 1$ if $k = 1$ and $\sqrt{2}\cos((k - 1)\pi s)$ otherwise. The coefficient function $\beta(t, s) = e^{-(t+s)}$ is the exponential function of t and s.
- Scenario 2: The predictor functions $X_i(s)$ are the same as those in Scenario 1 and the coefficient function $\beta(t, s) = 4\sum_{k=1}^{50}(-1)^{(k+1)}k^{-2}\vartheta_1(t, k)\vartheta_1(s, k)$.
- Scenario 3: The predictor functions $X_i(s)$ are generated as $X_i(s) = \sum_{k=1}^{3}(-1)^{(k+1)}k^{-1}Z_{ik}\vartheta_2(s, k)$, where $\vartheta_2(s, k) = 1$ if $k = 3$ and $\sqrt{2}\cos(k\pi s)$ otherwise. The coefficient function $\beta(t, s) = 4\sum_{k=1}^{3}(-1)^{(k+1)}k^{-2}\vartheta_2(t, k)\vartheta_2(s, k)$.

For each simulation scenario, we generated $n = 30$ samples, each with 20 time points on the interval (0, 1). The random errors $\epsilon(t)$ were from a normal distribution with a constant variance σ^2. The value of σ was adjusted to deliver three levels of signal-to-noise ratio (SNR = 0.5, 5, and 10) in each scenario. To assess the mean prediction accuracy, we generated an additional $n^* = 30$ predictor curves $\tilde{X}$ and computed the mean integrated squared error MISE $= 1/n^* \sum_{i=1}^{n^*} \int_0^1 (\eta_{\hat{\beta}}(\tilde{X}_i, t) - \eta_\beta(\tilde{X}_i, t))^2 dt$, where $\hat{\beta}$ was the estimator obtained from the training data. We had 100 runs for each combination of scenario and SNR.

We applied the OPFFR, FDA, and FPCA methods to the simulated datasets. Figure 3 displayed the perspective plots of the true coefficient functions in the three scenarios as well as their respective estimates for a single run with SNR= 10. In the first two scenarios, both OPFFR and FDA did a decent job in recovering the true coefficient function although the FDA estimates were slightly oversmoothed. In both scenarios the FPCA estimates clearly suffered since the true coefficient function could not be effectively represented by the eigen-functions of the predictor processes.

Figure 4 gave the summary reports of performances in terms of MISEs based on 100 runs. When the signal-to-noise ratio is

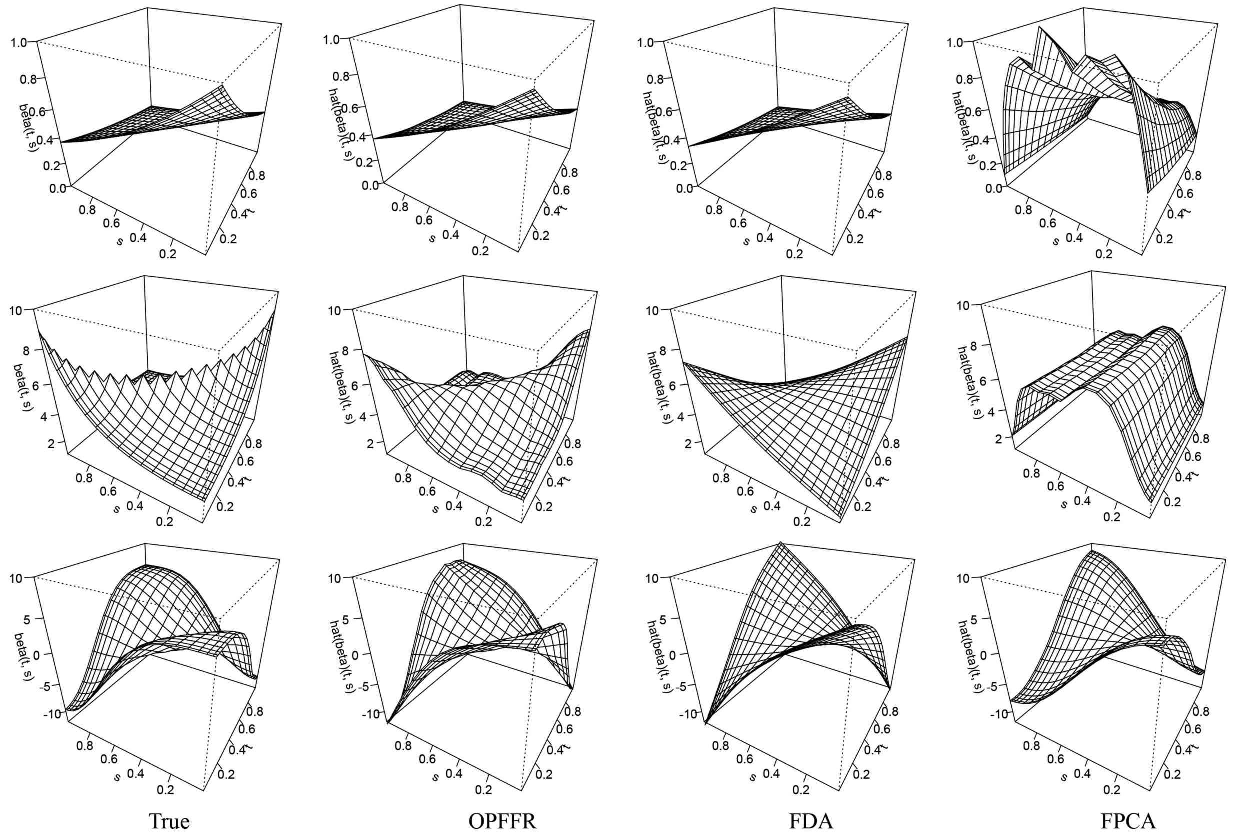


Figure 3. Perspective plots of the true $\beta(t, s)$ in three scenarios, and their respective estimates by the OPFFR, FDA, and FPCA methods when SNR = 10.

low, the OPFFR and FDA approaches had comparable performances. But when the signal-to-noise ratio increases, OPFFR showed clear advantage against FDA. The FPCA method failed to deliver competitive performance against the other two methods in all the settings due to its restrictive requirement of the effective representation of the coefficient function.

4.1.2. *Sparse Design*

In this section, we compared the performance of the proposed OPFFR method and the FPCA method regarding prediction error on sparsely, irregularly, and noisily observed functional data. To extend our method to sparsely and noisily observed data, we first applied the principal-component-analysis-through-conditional-expectation (PACE) method in Yao, Müller, and Wang (2005a) to the sparse functional data. Then we obtained a dense version of functional data by computing the PACE-fitted response and predictor functions at 50 selected time points for each curve. We applied the OPFFR method to these densely generated data and called this sparse extension to the OPFFR by the OPFFR-S method. The original OPFFR method, FPCA and OPFFR-S methods were all applied to the simulated data for comparison.

We first generated $n = 200$ samples for both response and predictor functions in Scenario 3, each with 50 time points on interval $(0, 1)$. To obtain different sparsity levels, we then randomly chose 5, 10, and 15 time points from the 50 ones for each curve independently. Normally distributed random errors were added to functional response and predictor with the SNR set to 10 in generating each pair of noisy response and predictor. The mean integrated squared error (MISE) was calculated based on additional $n^* = 50$ predictor curves without random noises.

Figure 5 displayed the perspective plots of the true coefficient functions in the sparse scenario as well as their respective estimates for a single run with 10 sampled time points per curve. The OPFFR-S method and FPCA performed well in estimating the coefficient function. The estimate recovered by the original OPFFR method was a little oversmoothed. In Figure 6, the performance in terms of MISEs based on 100 runs was compared. The OPFFR-S method always had the best prediction performances at all the three sparsity levels. When the sparsity level was high (5 time points per curve), the original OPFFR method had a worse prediction performance than the FPCA. However, its prediction performance quickly picked up as the data became denser. When the sparsity level was 15 time points per curve, it actually delivered a better prediction performance than the FPCA. Such an interesting phenomenon was referred to as the “phase transition” (Cai and Yuan 2011; Wang, Chiou, and Mller 2016).

5. Real Data Examples

We analyzed two real example in this section. We showed that our method had the numerical advantage over other approaches in terms of prediction accuracy in the analysis of the Canadian weather and histone regulation data. The results in the Canadian weather data, a dense design case, and the histone regulation data, a sparse design case, echoed with our findings in the simulation study. The smoothing parameters used in FDA for Canadian weather data were taken from the example codes in Ramsay,

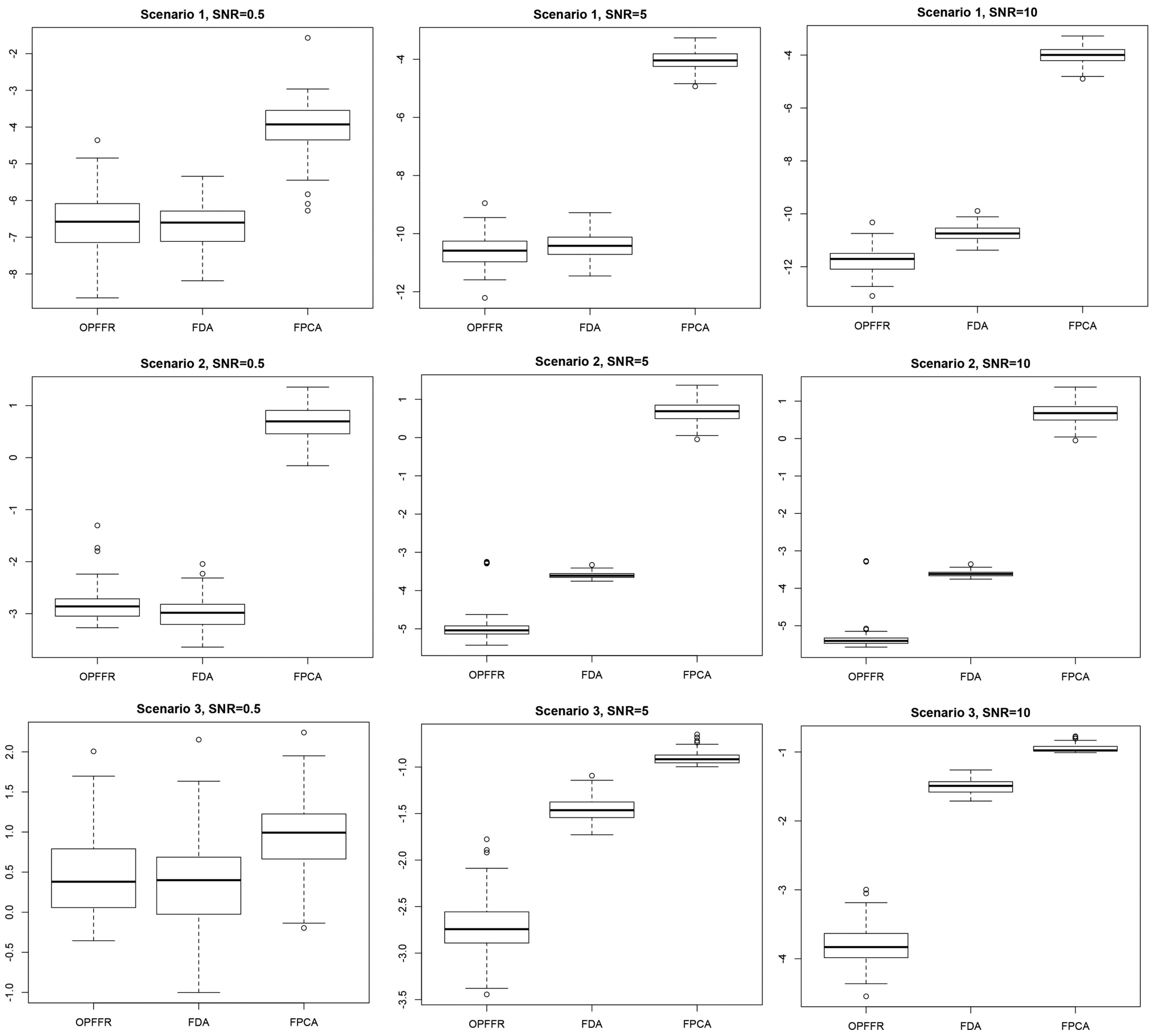
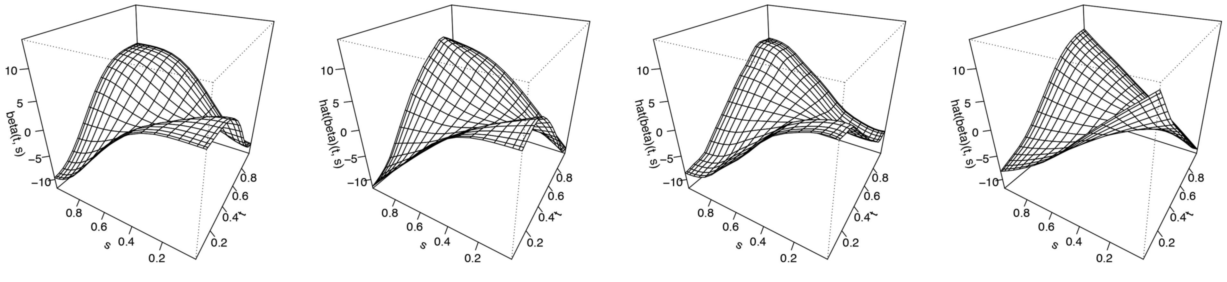


Figure 4. Boxplots of $\log_2$(MISE) for three scenarios under three signal-to-noise ratios (SNR = 0.5, 5, 10), based on 100 simulation runs. OPFFR is the proposed approach.

Hooker, and Graves (2009) and seven basis functions were used for the t and s directions, respectively. In the histone regulation data, we selected the smoothing parameter for FDA by a grid search on $10^{(-5:1:5)}$ and used six basis functions each for the t and s directions. For the FPCA method, the "regular" parameter was set to 2 for the Canadian weather data and 0 for the histone regulation data. The other parameters for FDA and FPCA approaches were the same as those used in the simulation study.

5.1. Canadian Weather Data

We first look at the Canadian weather data (Ramsay and Silverman 2005), a benchmark dataset in functional data analysis. The main goal is to predict the log daily precipitation profile based on the daily temperature profile for a geographic location in Canada. The daily temperature and precipitation data averaged over 1960 to 1994 were recorded at 35 locations in Canada. We compared OPFFR with FDA and FPCA in terms



True OPFFR-S FPCA OPFFR

Figure 5. Perspective plots of the true $\beta(t, s)$ in the sparse scenario, and their respective estimates by the OPFFR-S, FPCA, and OPFFR methods when the number of randomly selected time points is 10.

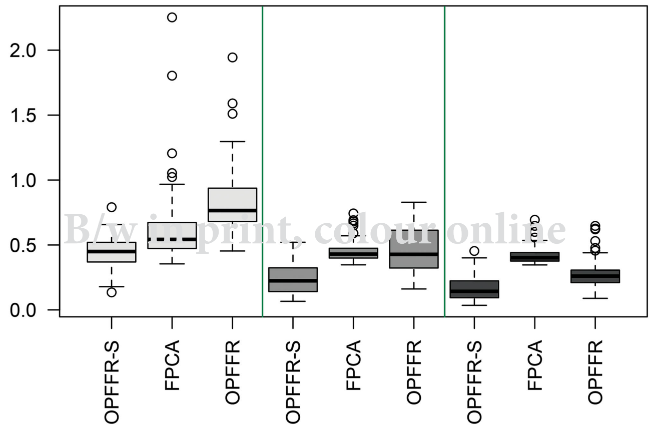


Figure 6. Boxplots of MISEs for the sparse scenario under three different sparsity levels, based on 100 simulation runs. The boxplots with different grayscale shades from left to right, respectively, represent the sparsity levels of 5, 10, and 15 time points per curve.

of prediction performance defined by integrated squared error (ISE) $\int_0^{365}(Y_i(t)-\eta_{\hat{\beta}_{-i}}(X_i,t))^2dt$, where $i=1,\ldots,35$ and $\hat{\beta}_{-i}$ was estimated by the dataset without the ith observation. For the convenience of calculation, we computed $\|Y_i(t)-\eta_{\hat{\beta}_{-i}}(X_i,t)\|_2^2$ at a grid of values t as the surrogate of ISE. Since the findings through the coefficient function estimates were similar to those in Ramsay and Silverman (2005), we only focused on the comparison of prediction performance. The summary in Table 1 clearly showed the numerical advantage of the proposed OPFFR method over the FDA and FPCA methods.

5.2. *Histone Regulation Data*

Nucleosomes, the basic units of DNA packaging in eukaryotic cells, consist of eight histone protein cores including two copies of H2A, H2B, H3, and H4. Besides the role as DNA scaffold, histones provide a complex regulatory platform for regulating gene activity (Wollmann et al. 2012). Focused study of the interaction between histones and gene activity may reveal how the organisms respond to the environmental changes. There are multiple sequence variants of histone proteins, which have some amino acid changes compared to their primary sequence, coexist in the same nucleus. For instance, in both plants and animals, there exist three variants of H3, the H3.1, the H3.3, and the centromere-specific CENP-A (CENH3) (Deal and Henikoff 2011). Each variant shows distinct regulatory mechanisms over gene expression.

In this article, an ultra-high throughput time course study was conducted to explore the interaction mechanism between the gene activity and histone variant, H3.3, during heat stress in *Arabidopsis thaliana*. In this study, the 12-day-old *Arabidopsis* seedlings that had been grown at 22°C were subject to heat stress of 38°C, and plants were harvested at 7 different time points within 24 hr for RNA sequencing (RNA-seq; Wang, Gerstein, and Snyder 2009) and ChIP sequencing (ChIP-seq; Johnson et al. 2007) experiments. We were interested in the genes responding to the heat shock, therefore 160 genes in response to heat (GO:0006951) pathway (Ashburner et al. 2000) were chosen. We selected 55 genes with the fold change above 0.5 at at least two consecutive time points in RNA-seq data. In ChIP-seq experiments, we calculated the mean of normalized read counts by taking the average of normalized read counts over seven time points for the region of 350 base pairs (bp) in the downstream of transcription start sites (TSS) of selected 55 genes. The normalized read counts over 350 bp from ChIP-seq and the normalized fragments per kilobase of transcript per million mapped reads (FPKM; Trapnell et al. 2010) over seven time points from RNA-seq were used to measure the histone levels and gene expression levels, respectively.

Table 1. The mean, standard deviation, and three quartiles of ISEs for the three approaches. The best result on each metric is in boldface.

Method	Median	Mean	Standard deviation	1st Qu.	3rd Qu.
OPFFR	**21.6400**	**40.2800**	**45.7631**	**13.8000**	**36.1700**
FDA	25.9000	44.1600	56.9544	18.7400	40.6100
FPCA	30.7752	45.5065	45.7763	20.5031	52.1827

Table 2. The mean, standard deviation, and three quartiles of ISEs for the four approaches. The best result on each metric is in boldface.

Method	Median	Mean	Standard deviation	1st Qu.	3rd Qu.
OPFFR	1.5700	**7.7120**	18.9180	**0.5077**	**5.1900**
OPFFR-S	**1.4070**	7.7150	18.6037	0.6972	5.5820
FDA	2.2060	7.9770	18.7004	0.5461	6.2750
FPCA	2.0170	8.4720	**18.3978**	0.9126	6.1790

We applied the OPFFR, FDA, and FPCA methods to histone regulation data in example 2. Since the gene expression levels were sparsely observed, we also applied the OPFFR-S method to the data. The comparison of the four methods is shown in Table 2. In the table, the standard deviation of ISEs was the only measure that neither the OPFFR nor the OPFFR-S was the most optimal. This was caused by a few observations where all the methods failed to make a good prediction and the OPFFR methods happened to have larger ISEs. In terms of all the other measures, the proposed OPFFR and OPFFR-S methods clearly showed the advantage in prediction accuracy again. Since the results from the OPFFR and OPFFR-S were comparable to each other, we chose to present all the following results based on the OPFFR analysis.

Figure 7 is the plot of the fitted coefficient function generated from our OPFFR method. For region between 300 bp and 350 bp, there was a strong negative influence of H3.3 on genes

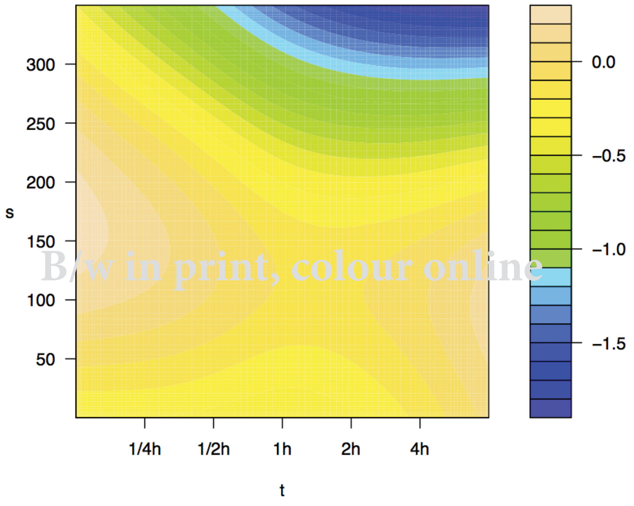


Figure 7. The estimated coefficient function $\beta(t,s)$ for the histone regulation study. The y-axis label represents the positions on genomes and x-axis label represents seven time points.

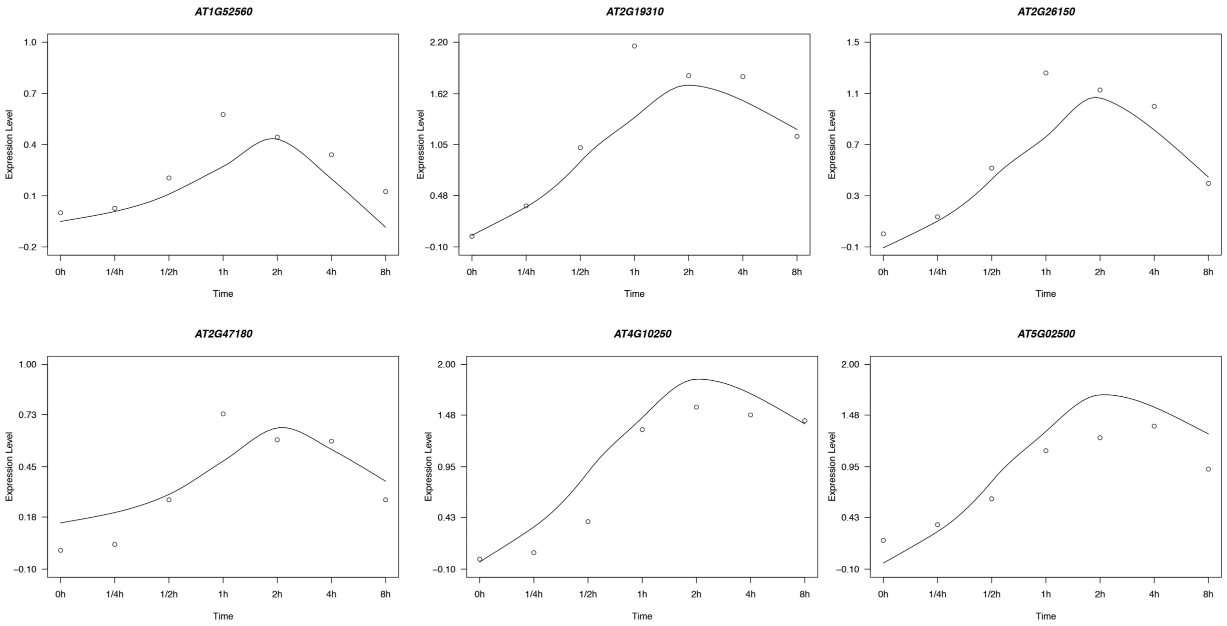


Figure 8. The fitted response functions for six genes in the histone regulation study. The y-axis stands for the normalized expression levels and x-axis label represents seven time points. The curve fitted using OPFFR is in the solid line, with the data in circles.

activity from half hour to 8 hr. It indicted that the loss of H3.3 might have the biological influence on the up-regulation of heat-induced genes. This negative correlation phenomenon was also observed after 30 min on the region of 250 bp to 300 bp between H3.3 and gene activity. In addition, the region from 50 bp to 150 bp had a positive effect on genes activity over time domain from 0 hr to half hour and 4 hr to 8 hr. Therefore, we provided a numerical evidence that heat-shock-induced transcription of genes in response to heat stress might be regulated via the epigenetic changes of H3.3, especially on the downstream region of TSS. The sample plots in Figure 8 showed a nice match of the predicted gene expression curves with the observed values.

6. Conclusion

In this article, we have presented a new analysis tool for modeling the relationship of a functional response against a functional predictor. The proposed method is more flexible and generally delivers a better numerical performance than the FPCA approach since it does not have the restrictive structural dependence assumption on the coefficient function. When compared with the penalized B-splines method, the proposed method has the theoretical advantage of possessing the optimal rate for mean prediction as well as some numerical advantage as shown in the numerical studies. Moreover, the Representer Theorem guarantees an exact solution to the penalized least squares, a property that is not shared by the existing penalized function-on-function regression models. The application of our method to a histone regulation study provided numerical evidence that the changes in H3.3 might regulate some genes through transcription regulations. Although such a finding sheds light on the relationship between histone variant H3.3 and gene activity, the details of the regulation process are still unknown and merit further investigations. For instance, we may investigate how the H3.3 organizes the chromatins to up-regulate those active genes. Such investigations would call for more collaborations between statisticians and biologists.

When the regression model has a scalar response against one or more functional predictors, methods other than the roughness penalty approach are available to overcome the inefficient basis representation drawback in the FPCA method. For example, Delaigle et al. (2012) considered a partial least-square (PLS) based approach. Ferré and Yao (2003) and Yao, Lei, and Wu (2015) translated the idea of sufficient dimension reduction (SDR) into the setting of functional regression models. Intuitively, these methods might be more efficient in their selection of the principal component basis functions since they incorporate the response information into consideration. However, our experiments with a functional response version of the functional PLS (Preda and Saporta 2005), not shown here due to space limit, did not look so promising. Therefore, further investigation in this direction is surely needed.

In some applications, the response functions may show a different level of smoothness from the predictor functions. Then it is reasonable to require different levels of smoothness for the coefficient function $\beta(t, s)$ in the directions of t and s. This can be effectively implemented through introducing five new smoothing parameters θ_k, $k = 1, \dots, 5$ in Example 3 such that θ_k^{-1} precedes the kth integral in the penalty J. Then one essentially selects λ/θ_k, $k = 1, \dots, 5$ to determine the appropriate level of smoothness for each component. Note that this would also require the incorporation of θ_k into the RKs. Numerically, these parameters can be tuned by first fixing all θ_k to some constant and selecting an optimal λ to pin down the overall smoothness level, and then fix λ and fine tune θ_k to determine the component-wise smoothness levels.

Q1

Supplementary Materials

Acknowledgments

The authors are grateful to Dr. Xiaoyu Zhang who kindly provided the histone regulation data and pertinent explanations of the experiments. The authors thank the associate editor and the reviewers for their insightful comments that have significantly improved the article.

Funding

Sun and Ma's research was supported by U.S. National Science Foundation under grants DMS-1440037, DMS-1440038, and DMS-1438957 and by U.S. National Institute of Health under grants R01GM122080 and R01GM113242. Du's research was supported by U.S. National Science Foundation under grant DMS-1620945. Wang's research was supported by U.S. National Science Foundation under grant DMS-1613060.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000), "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, 25, 25–29. [9] Q2

Ausió, J. (2006), "Histone Variants: The Structure Behind the Function," *Briefings in Functional Genomics & Proteomics*, 5, 228–243. [1]

Benatia, D., Carrasco, M., and Florens, J.-P. (2015), "Functional linear regression with functional response," Technical Report, Département de Sciences Économiques, Université de Montréal, Montréal, Canada. [2] Q3

Cai, T. T., and Yuan, M. (2011), "Optimal Estimation of the Mean Function Based on Discretely Sampled Functional Data: Phase Transition," *The Annals of Statistics*, 39, 2330–2355. [7]

——— (2012), "Minimax and Adaptive Prediction for Functional Linear Regression," *Journal of the American Statistical Association*, 107, 1201–1216. [1,3]

Crambes, C., and Mas, A. (2013), "Asymptotics of Prediction in Functional Linear Regression With Functional Outputs," *Bernoulli*, 19, 2627–2651. [2]

Cuevas, A., Febrero, M., and Fraiman, R. (2002), "Linear Functional Regression: The Case of Fixed Design and Functional Response," *Canadian Journal of Statistics*, 30, 285–300. [2]

Deal, R. B., and Henikoff, S. (2011), "Histone Variants and Modifications in Plant Gene Regulation," *Current Opinion in Plant Biology*, 14, 116–122. [9]

Delaigle, A., Hall, P., (2012), "Methodology and Theory for Partial Least Squares Applied to Functional Data," *The Annals of Statistics*, 40, 322–352. [10] Q4

Du, P., and Wang, X. (2014), "Penalized Likelihood Functional Regression," *Statistica Sinica*, 24, 1017–1041. [1,3]

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2011), "Kernel Regression with Functional Response," *Electronic Journal of Statistics*, 5, 159–171. [2]

Ferré, L., and Yao, A.-F. (2003), "Functional Sliced Inverse Regression Analysis," *Statistics*, 37, 475–488. [10]

Gu, C. (2013), *Smoothing Spline ANOVA Models* (2nd ed.), New York: Springer-Verlag. [5]

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2015), "Penalized Function-On-Function Regression," *Computational Statistics*, 30, 539–568. [3]

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007), "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *Science*, 316, 1497–1502. [1,9]

Kim, Y.-J., and Gu, C. (2004), "Smoothing Spline Gaussian Regression: More Scalable Computation via Efficient Approximation," *Journal of the Royal Statistical Society, Series B*, 66, 337–356. [5]

Lian, H. (2015), "Minimax Prediction for Functional Linear Regression with Functional Responses in Reproducing Kernel Hilbert Spaces," *Journal of Multivariate Analysis*, 140, 395–402. [3]

Malfait, N., and Ramsay, J. O. (2003), "The Historical Functional Linear Model," *Canadian Journal of Statistics*, 31, 115–128. [2]

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015), "Bayesian Function-On-Function Regression for Multilevel Functional Data," *Biometrics*, 71, 563–574. [2]

Preda, C., and Saporta, G. (2005), "PLS Regression on a Stochastic Process," *Computational Statistics & Data Analysis*, 48, 149–158. [10]

Ramsay, J. O., Hooker, G., and Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer Science & Business Media. [8] Q5

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer-Verlag. [1,3,6,8]

Scheipl, F., and Greven, S. (2016), "Identifiability in Penalized Function-on-Function Regression Models," *Electronic Journal of Statistics*, 10, 495–526. [3]

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010), "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation," *Nature Biotechnology*, 28, 511–515. [9]

Wahba, G. (1990), *Spline Models for Observational Data* (*CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59), Philadelphia, PA: SIAM. [3]

Wahba, G., and Craven, P. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403. [5]

Wang, J.-L., Chiou, J.-M., and Mller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [7]

Wang, Z., Gerstein, M., and Snyder, M. (2009), "RNA-Seq: A Revolutionary Tool for Transcriptomics," *Nature Reviews Genetics*, 10, 57–63. [1,9]

Wollmann, H., Holec, S., Alden, K., Clarke, N. D., Jacques, P.-E., and Berger, F. (2012), "Dynamic Deposition of Histone Variant H3.3 Accompanies Developmental Remodeling of the Arabidopsis transcriptome," *PLoS Genetics*, 8, e1002658. [9]

Yao, F., Lei, E., and Wu, Y. (2015), "Effective Dimension Reduction for Sparse Functional Data," *Biometrika*, 102, 421–437. [10]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [3,7]

——— (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903. [1,2,3,6]

Yuan, M., and Cai, T. T. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444. [1,3,4,6]