Biometrika (20xXX), **XX**, X, *pp*. 1–12
Printed in Great Britain

Functional Linear Regression for Discretely Observed Data: From Ideal to Reality

BY HANG ZHOU, FANG YAO

Department of Probability and Statistics, School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing 100871, China
h_zhou@pku.edu.cn *fyao@math.pku.edu.cn*

AND HUIMING ZHANG

Department of Mathematics, University of Macau, Macau 999078, China
huimingzhang@um.edu.mo

SUMMARY

Despite extensive study on functional linear regression, there exists a fundamental gap in theory between the ideal estimation from fully observed covariate functions and the reality that one can only observe functional covariates discretely with noise. The challenge arises when deriving a sharp perturbation bound for the estimated eigenfunctions in the latter case, which renders existing techniques for functional linear regression not applicable. We use a pooling method to attain the estimated eigenfunctions and propose a sample-splitting strategy to estimate the principal component scores, which facilitates the theoretical treatment for discretely observed data. The slope function is estimated by approximated least squares, and we show that the resulting estimator attains the optimal convergence rates for both estimation and prediction when the number of measurements per subject reaches a certain magnitude of the sample size. This phase transition phenomenon differs from the known results for the pooled mean and covariance estimation, and reveals the elevated difficulty in estimating the regression function. Numerical experiments, using simulated and real data examples, yield favorable results when compared with existing methods.

Some key words: Compact operator; Perturbation bound; Phase transition; Principal component analysis.

1. INTRODUCTION

Functional data analysis has been extensively developed over the past decades and widely applied in various fields, such as economics, chemometrics, image analysis and meteorology. For a comprehensive treatment on functional data, we recommend the monographs by Ramsay & Silverman (2005) and Hsing & Eubank (2015), which provide introductions and discussions on methodologies and theoretical results.

Among problems involving functional data, the regression model has received substantial attention in the literature. Functional principal component analysis has been widely adopted as a regularization tool in the (generalized) functional linear model (Yao et al., 2005b; Cai & Hall, 2006; Hall & Horowitz, 2007; Dou et al., 2012) and others. The seminal work, Hall & Horowitz (2007), established the minimax rate-optimal slope estimator for the function-on-scalar regression model via the plug-in method, that is, plugging in the estimated eigenpairs with regulariza-

tion to estimate the slope function. Cai & Hall (2006) showed that the prediction error for the plug-in estimator could reach root-n consistency by assuming the new observation is a suitably smooth fixed curve. The methodology and theory for functional generalized linear model were established by Dou et al. (2012), which adopted a truncated likelihood method and a change-of-measure argument to overcome the difficulty caused by the nonlinear link function. Another line of research on functional linear regression is under the reproducing kernel Hilbert space framework (Yuan & Cai, 2010; Cai & Yuan, 2012, among others). Under this framework, one may only bound the prediction error instead of the $\mathcal{L}^2$ convergence of the slope function. We stress that all the above works deal with fully observed functional covariates, and their theories are based on the cross-sectional sample covariance.

However, fully observed functional data are often viewed as an ideal scenario, while in practice we typically have n random curves observed at N discrete time points with noise for each subject. In this realistic scenario, there are two typical strategies in estimating the population quantities, pre-smoothing individual curves for dense observations versus pooling information together for sparse observations. The impact of dense and sparse designs on the mean and covariance estimation has been understood (Zhang & Chen, 2007; Zhang & Wang, 2016, among others), while theories for functional linear regression with discretely observed covariates are challenging due to the infinite-dimensionality nature of functional data. To suppress the estimation bias, the number of principal components used in estimating the slope function should grow with the sample size, n. Since the perturbation technique in Hall & Horowitz (2007) is no longer applicable without the cross-sectional sample covariance, there has not been meaningful progress on obtaining a sharp bound for eigenfunction estimation with diverging index based on discrete data. Therefore, it is mathematically challenging to study the plug-in method of Hall & Horowitz (2007), which technically depends on this sharp bound and such perturbation techniques, to obtain a minimax rate. Despite extensive studies on functional linear regression, there exists a fundamental gap in the theory between the idea estimation and reality observation, and remains as an unsolved problem over a decade.

Motivated by the aforementioned challenges, we present a novel approach for functional linear regression with discretely observed covariates. By adopting the pooling strategy with a kernel smoother, we obtain the estimated covariance function and eigenfunctions. A building block is using sample-splitting method to estimate the principal component scores, which eliminates the dependence between the covariates and the estimated eigenfunctions and makes the error quantification tractable. The slope function is estimated from the approximated least square equation, which is more feasible in theoretical analysis than the plug-in method. The main contributions of this paper are three-fold. First, we derive the expectation bound for the projections of the estimated covariance function onto eigenfunctions. Second, based on these quantities, an improved perturbation bound for a diverging number of estimated eigenfunctions is obtained. This bound is essential for nearly all methods involving inverse models in functional data analysis and has its own merits deserving further investigation. Third, we show that when the number of time-points, N, reaches a certain magnitude of the sample size, n, the proposed method attains the optimal convergence rate, i.e. the rate for the case where covariates are fully observed, for both estimation and prediction. This phase transition is distinct from those concerning mean and covariance estimation, and has not yet been revealed elsewhere for functional linear regression.

In the sequel, we use C stand for a positive constant which may vary from place to place. The relation $a_n \lesssim b_n$ indicates that $a_n \leq \mathrm{C}b_n$ for large n and the relation $\gtrsim$ is defined analogously. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a function $p(s) \in \mathcal{L}^2[0,1]$, we use $\|p\|^2$ to denote $\int_{[0,1]} p(s)^2 \mathrm{d}s$ and $\|p\|_\infty$ stands for $\sup_{s\in[0,1]} |p(s)|$. For a function $A(s,t) \in \mathcal{L}^2[0,1]^2$, define

$\|A\|_{\mathrm{HS}}^2 = \iint_{[0,1]^2} A(s,t)^2 \mathrm{d}s\mathrm{d}t$ and $\|A\|_j^2 = \int_{[0,1]} \{\int_{[0,1]} A(s,t)\phi_j(s)\mathrm{d}s\}^2 \mathrm{d}t$, where ϕ_j is the jth eigenfunction of the functional covariate.

2. Proposed methodology

2.1. Model setting and challenges

We assume that $X(t)$ is a square integrable stochastic process on $[0,1]$ with the mean function $\mu(t) = E\{X(t)\} = 0$ without loss of generality and covariance function $C(s,t) = \mathrm{cov}\{X(s), X(t)\}$. By the Mercer's Theorem, C admits the spectral decomposition

$$C(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t), \tag{1}$$

where $\lambda_1 > \lambda_2 > \cdots > 0$ are eigenvalues and $\phi_1, \phi_2, \ldots$ are the corresponding eigenfunctions.

Let $X_i(t)$ be independent and identically distributed copies of $X(t)$ for $i \in \{1, \ldots, n\}$. Assume, for each i, Y_i conditional on X_i follows the function-on-scalar linear model

$$Y_i = \int_{[0,1]} X_i(s)\beta(s)\mathrm{d}s + e_i,$$

where $\beta = \sum_{k=1}^{\infty} b_{0k}\phi_k$ is the unknown slope function and $\{e_i\}_{i=1}^n$ are independent and identically distributed copies of e with mean zero and variance σ_Y^2. The process, X_i, admits the so-called Karhunen-Lòeve expansion $X_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t)$, and the principal component scores, $\{\xi_{ik} = \int_{[0,1]} X_i(s)\phi_k(s)\mathrm{d}s\}_{k=1}^{\infty}$, are uncorrelated random variables with mean zero and variances $\{\lambda_k\}_{k=1}^{\infty}$.

In general, one cannot observe $X_i(t)$ for all $t \in [0,1]$. The actual observations of each X_i are

$$X_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \qquad j = 1, \ldots, N,$$

where ε_{ij}'s are independent and identically distributed copies of ε with mean zero and variance σ_X^2. We further assume that T_{ij}'s are independent and identically distributed copies of T, which follows the uniform distribution on $[0,1]$, and X, T, ε and e are mutually independent.

In mean and covariance estimation, there are two typical approaches to bridge the gap from the discrete observations to the fully observed curves. One method is to pre-smooth individual curves and applicable for sufficiently dense data. Particularly, when $N \gtrsim n^{5/4}$ and the tuning parameter is optimally chosen for each individual, the mean and covariance estimation based on pre-smoothed curves attain the parametric root-n convergence, referred to as the ultra-dense case (Ramsay & Silverman, 2005; Zhang & Chen, 2007). For sparse functional data, it is suggested to borrow information from all subjects by pooling observations together to estimate population quantities, for instance, the mean and covariance functions as well as eigenvalues and eigenfunctions (Yao et al., 2005a; Hall et al., 2006; Paul & Peng, 2009; Cai & Yuan, 2011). In Cai & Yuan (2011) and Zhang & Wang (2016), it has been shown that when $N \gtrsim n^{1/4}$, the pooling estimators for mean and covariance functions reach the optimal root-n consistency, which provides theoretical insight for the advantage of pooling estimation over the pre-smoothing method.

Before proceeding to our proposal, we shall give a comprehensive synopsis of the state-of-art methodologies and techniques in functional linear regression, and discuss the challenges caused by discrete observations. Assume that the estimated covariance function, $\hat{C}$, admits an empirical version of decomposition (1) with estimated eigenpairs $\{\hat{\lambda}_k, \hat{\phi}_k\}_{k=1}^{\infty}$. The following perturbation

series plays a key role in the theoretical analysis of plug-in method (Hall & Horowitz, 2007),

$$E(\|\hat{\phi}_k - \phi_k\|^2) \asymp \sum_{j\neq k} \frac{E\{\langle(\hat{C}-C)\phi_k, \phi_j\rangle^2\}}{(\lambda_j - \lambda_k)^2}. \tag{2}$$

This type of expansion can be found in Bosq (2000), Li & Hsing (2010) and Dou et al. (2012), see Chapter 5 in Hsing & Eubank (2015) for details. For the fully observed case with the cross-sectional sample covariance, $\hat{C} = n^{-1}\sum_{i=1}^n X_i \otimes X_i$, the numerator in each term of summation (2) can be reduced to the principal component score and $E\{\langle(\hat{C}-C)\phi_k, \phi_j\rangle^2\} \lesssim \lambda_j\lambda_k/n$. Assume that the eigenvalues decay in a polynomial rate, by Lemma 7 in Dou et al. (2012),

$$E(\|\hat{\phi}_k - \phi_k\|^2) \lesssim \frac{\lambda_k}{n}\sum_{j\neq k}\frac{\lambda_j}{(\lambda_j-\lambda_k)^2} \lesssim \frac{k^2}{n},$$

which is optimal in the minimax sense (Wahl, 2021).

When there is no fully observed sample covariance, an obstacle comes from quantifying the summation (2). For pre-smoothing methods, the reconstructed $\hat{X}_i$ achieves a root-n convergence in the $\mathcal{L}^2$ sense when $N \gtrsim n^{5/4}$ and thus, the estimated covariance function, $\hat{C}(s,t) = n^{-1}\sum_{i=1}^n \hat{X}_i(s)\hat{X}_i(t)$, admits $\|\hat{C}-C\|_{\mathrm{HS}} = O_p(n^{-1/2})$. This, however, does not yield the optimal convergence for a diverging number of eigenfunctions by summation (2). The expectation of the numerator in each term of (2) is no longer the principal component score and such a complicated form makes it mathematically challenging to quantify this infinite summation as $|\lambda_k - \lambda_j| \to 0$. For pooling method, it is also highly nontrivial to sum up all $E\{\langle(\hat{C}-C)\phi_k, \phi_j\rangle^2\}$ with respect to j, k, see comments below Theorem 1 for more detail. Thus, it is difficult to obtain a sharp bound for summation (2) and the plug-in method, which technically depends on such perturbation bounds, fails to attain a minimax rate no matter how large N is.

2.2. *Proposed Method*

Recall the discussion on the magnitude of N needed for mean and covariance estimation to reach optimal by pre-smoothing and pooling, i.e., $n^{5/4}$ versus $n^{1/4}$, we are inspired to use the more efficient pooling method. To avoid dependence in the subsequent analysis, we adopt a sample-splitting strategy which can be dated back to Larson (1931).

Let $\mathcal{X}_1 = \{(i,j)|i \le n/2; j \le N; i,j \in \mathbb{N}_+\}$ and $\mathcal{X}_2 = \{(i,j)|n/2 < i \le n; j \le N; i,j \in \mathbb{N}_+\}$ be the disjoint sample-splitting index sets. Denote $\hat{C}_{(1)}(s,t)$ and $\hat{C}_{(2)}(s,t)$ the estimated covariance functions based on $\mathcal{X}_1$ and $\mathcal{X}_2$ by the pooling method, respectively. We adopt the local constant smoother here to avoid distraction from tedious calculations. Denote $\delta_{ijl} = X_{ij}X_{il} = \{X_i(T_{ij}) + \varepsilon_{ij}\}\{X_i(T_{il}) + \varepsilon_{il}\}$ and for $r = 1, 2$,

$$\hat{C}_{(r)}(s,t) = \frac{1}{n}\sum_{i=1}^n \frac{1}{N(N-1)} \sum_{1\le j\neq l\le N} \frac{1}{h^2}\mathrm{K}\left(\frac{T_{ij}-s}{h}\right)\mathrm{K}\left(\frac{T_{il}-t}{h}\right)\delta_{ijl}\mathbf{1}_{\{(i,j),(i,l)\in\mathcal{X}_r\}},$$

where h is the bandwidth and K is a symmetric density kernel on $[-1,1]$ with $\int\{u^2\mathrm{K}(u) + \mathrm{K}^2(u)\}\mathrm{d}u < \infty$. The estimated covariance function, $\hat{C}_{(r)}$, admits an empirical version of decomposition (1) with estimated eigenparis $\{\hat{\lambda}_{(r),k}, \hat{\phi}_{(r),k}\}_{k=1}^{\infty}$. The sample-splitting estimator of

the principal component score is given by the Monte Carlo average,

$$\hat{\xi}_{ik} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} X_{ij}\hat{\phi}_{(2),k}(T_{ij}) & (i,j) \in \mathcal{X}_1 \\ \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} X_{ij}\hat{\phi}_{(1),k}(T_{ij}) & (i,j) \in \mathcal{X}_2. \end{cases}$$

Given a slope function $\beta = \sum_{k=1}^{\infty} b_k\phi_k \in \mathcal{L}^2[0,1]$, the actual likelihood equation $L_n(\beta) = n^{-1}\sum_{i=1}^{n}(\langle X_i,\beta\rangle Y_i - \langle X_i,\beta\rangle^2/2)$ can be approximated by $\tilde{L}_n(b) = n^{-1}\sum_{i=1}^{n}\{(\xi_i^{\mathrm{T}}b)Y_i - (\xi_i^{\mathrm{T}}b)^2/2\}$ for $b \in \mathbb{R}^m$, where $\xi_i = (\xi_{i1},\ldots,\xi_{im})^{\mathrm{T}}$ and $m \in \mathbb{N}_+$ is diverging with n to suppress the estimation bias. After substituting scores by their estimates, we obtain $\hat{\beta}(s) = \sum_{k=1}^{m}\hat{b}_k\hat{\phi}_k(s)$ with

$$\hat{b} = (\hat{b}_1,\ldots,\hat{b}_m)^{\mathrm{T}} = \underset{b\in\mathbb{R}^m}{\arg\max}\,\frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{\xi}_i^{\mathrm{T}}b)Y_i - \frac{(\hat{\xi}_i^{\mathrm{T}}b)^2}{2}\right\},$$

and $\hat{\phi}_k = (\hat{\phi}_{(1),k} + \hat{\phi}_{(2),k})/2$.

For prediction, the target is to recover the functional $\theta_0(X^*) = \int X^*(s)\beta(s)\mathrm{d}s$, where X^* is an independent and identically distributed copy of X. Denote $\tilde{\theta}_n(X^*) = \int X^*(s)\hat{\beta}(s)\mathrm{d}s$ as the estimator of $\theta_0(X^*)$ based on $\hat{\beta}$ and the fully observed X^*. When X^* is observed at some discrete time points with noise, i.e., $\{X_j^* = X^*(T_j^*) + \varepsilon_j^*, j = 1,\ldots,N\}$, where T_j^* and ε_j^* are random copies of T and ε, $\theta_0(X^*)$ is estimated by

$$\hat{\theta}_n(X^*) = \frac{1}{N}\sum_{j=1}^{N}\hat{\beta}(T_j^*)X_j^*. \tag{3}$$

Remark 1. In contrast to the plug-in method, the approximated least squared method does not require the sharp bounds for the eigenfunctions and perturbation series. Moreover, the inverse in plug-in method may be numerically unstable in practice. Thus, the approximated least square method is more appealing for the functional linear model with discretely observed covariates.

Remark 2. The conditional expectation method proposed by Yao et al. (2005a) is widely adopted for extremely sparse functional data, especially for finite N. However, the conditional expectation estimator is consistent only for $N \lesssim (n/\log n)^{1/4}$ (Dai et al., 2018), which narrows the sampling rate to a small range. Since we study the optimal convergence for functional linear regression over a wide range of N, the conditional expectation estimator is not a necessarily suitable choice. The Monte Carlo score estimator is attractive for theoretical analysis, while rigorous theory for functional linear regression based on conditional expectation estimators remains an open problem.

Remark 3. The sample-splitting method has a long history and is adopted in many problems, such as variable selection in high dimension, change point detection, testing and false discovery rate control (Wasserman & Roeder, 2009; Zou et al., 2020; Meinshausen et al., 2009; Du et al., 2021, among others). We adopt sample-splitting here to eliminate the dependence between X_i and $\hat{\phi}_k$, which makes the error quantification mathematically tractable. Similar asymptotic arguments via sample-splitting have been successfully applied to semi-parametric models, see Section 25 in van der Vaart (1998) for details. Besides theoretical considerations, the practical merits of our proposed method are demonstrated via numerical experiments in Section 4.

3. THEORETICAL RESULTS

In this section, we present the $\mathcal{L}^2$ convergence rate of $\hat{\beta}$ and theoretical properties of its corresponding prediction error. When N is larger than a certain magnitude of n that is relevant to the smoothness of β and $C(s,t)$, these rates are optimal in the minimax sense for the case where the covariates are fully observed. Before introducing our results, we impose the following regularity conditions.

Condition 1. $X(t)$ has finite fourth moment, $\int E\{X^4(t)\}\mathrm{d}t < \infty$ and $E(\xi_j^4) \lesssim \lambda_j^2$ for all j.

Condition 2. The covariance function has bounded second order derivatives and the eigenvalue is decreasing with $\mathrm{C}j^{-a} \geq \lambda_j \geq \lambda_{j+1} + \mathrm{C}^{-1}j^{-a-1}$ for each $j \in \mathbb{N}_+$ with a constant $a > 1$.

Condition 3. For each $j \in \mathbb{N}_+$, the eigenfunctions satisfy $|\phi_j(t)|_\infty = O(1)$ and

$$|\phi_j^{(k)}(t)|_\infty \lesssim j^{c/2}|\phi_j^{(k-1)}|_\infty \text{ for } k = 1, 2,$$

where c is a positive constant, and assume $\phi_j(0) = \phi_j(1)$ and $\phi_j^{(1)}(0) = \phi_j^{(1)}(1)$ without loss of generality.

Condition 4. The slope function, β, admits the generalized Fourier expansions $\beta(t) = \sum_{k=1}^{\infty} b_{0k}\phi_k(t)$ with the basis $\{\phi_k\}_{k=1}^{\infty}$, where $|b_{0k}| \lesssim k^{-b}$ for $k \geq 1$ and $b > a/2 + 2$.

Conditions 1–2 are standard conditions in functional principal component analysis (Hall & Hosseini-Nasab, 2006; Hall & Horowitz, 2007; Yuan & Cai, 2010; Dou et al., 2012). Condition 3 characterizes the frequency increment of each eigenfunction via the amplitude of its derivatives, and a typical example is the Fourier basis in which $c = 2$. The boundary assumption on $\phi_j, \phi_j^{(1)}$ eliminates the edge effect caused by the local constant smoother for technical convenience, and may be relaxed with more technicality. Condition 4 sets up a smoothness class of regression functions (Hall & Horowitz, 2007; Dou et al., 2012). The inequality $b > a/2 + 2$ quantifies the smoothness and alignment of the slope function relative to the functional covariate. It is expected that the elevated difficulty of quantifying error in estimated eigenfunctions (and the slope function) requires slightly higher degree of smoothness in β, compared to those in Hall & Horowitz (2007) and Dou et al. (2012).

Condition 5. For $n, N \to \infty$, assume that

(i) $m^{2a+4}/n \to 0$, $m^{a+2}/N \to 0$, $N/m^{(a+2b-1)} \to 0$;
(ii) $m^{2a+4}/(nNh) \to 0$, $m^{a+1}/(n^{1/2}Nh) \to 0$, $Nhm^{2a+3} \to \infty$ and
(iii) $h \max\left\{n^{(3a+2c+4)/(8a+16)}, N^{(3a+2c+4)/(4a+8)}\right\} \to 0$.

Condition 6. For $n, N \to \infty$, assume that

(i) $m \asymp n^{1/(a+2b)}$, $\bar{N} = \max\{n^{(2a+2)/(a+2b)}, n^{(a+b)/(a+2b)}\}$ and $N \gtrsim \bar{N}$;
(ii) $h \asymp \underline{N}^{1/4}n^{-(2a+b+c+1)/2(a+2b)}$ with $\underline{N} = \min\{n^{(2a+2)/(a+2b)}, n^{(a+b)/(a+2b)}\}$ and $c \leq 2a + 2$.

Condition 5 gives basic assumptions on the relations among m, N, h and n needed for consistence. In particular, the number of expansion components of $\hat{\beta}$, denoted by m, should grow with n to suppress the approximation bias of β. Nevertheless, m cannot be too large, so that the leading m components can be well estimated based on the pooled covariance. The estimation variance and bias are characterized by nNh (or $n^{1/2}Nh$) and h, respectively. Condition 6 states the range for N, m and h under which the convergence rate could reach optimality.

We begin with an expectation bound for $\langle\Delta_{(r)}\phi_k, \phi_j\rangle^2$ and $\|\Delta_{(r)}\|_j^2$, which is fundamental in the subsequent analysis and has its own merits deserving further investigation.

THEOREM 1. *Under Conditions 1–3 and 5, let* $\Delta_{(r)} = \hat{C}_{(r)} - C$, *for any* $1 \le j \ne k \le m$ *and* $r = 1, 2$,

$$E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2) \lesssim \frac{j^{-a}k^{-a}}{n} + h^4\left(k^{2c}j^{-2a} + k^{-2a}j^{2c}\right)$$

and

$$E(\|\Delta_{(r)}\|_j^2) \lesssim \frac{j^{-a}}{n}\left(1 + \frac{1}{Nh}\right) + h^4 j^{2c}.$$

When dealing with inverse problems involving functional covariates, the resolvent technique is widely adopted to obtain the optimal convergence (Cai & Hall, 2006; Hall & Horowitz, 2007; Dou et al., 2012; Hsing & Eubank, 2015). Theorem 1 makes the first attempt at obtaining such bounds for discretely observed functional data. As revealed in Zhang & Wang (2016), the convergence rate for $\|\Delta_{(r)}\|_{\mathrm{HS}}^2$ admits a two-dimensional kernel smoothing rate $n^{-1}\{1 + (Nh)^{-2}\} + h^4$. However, due to the added smoothness by integration, $E(\|\Delta_{(r)}\|_j^2)$ and $E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2)$ admit degenerated kernel smoothing rates with variances $n^{-1}\{1 + (Nh)^{-1}\}$ and n^{-1}, respectively. By Bessel's equality, $\sum_{k=1}^{\infty} E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2) = E(\|\Delta_{(r)}\|_j^2)$ and $\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2) = E(\|\Delta_{(r)}\|_{\mathrm{HS}}^2)$, which indicates that one cannot sum up all $E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2)$ with respect to j, k due to the smoothing bias. This is materially different from the case using the fully observed sample covariance, $\hat{C} = n^{-1}\sum_{i=1}^{n} X_i \otimes X_i$. Based on the second assertion of Theorem 1 and Theorem 5.1.8 in Hsing & Eubank (2015), the following theorem offers an improved bound for a diverging number of estimated eigenfunctions.

THEOREM 2. *Under Conditions 1–3 and 5, define* $\eta_j = (1/2)\min_{k\ne j}|\lambda_j - \lambda_k|$ *and* $\Omega_m(n, N) = \{\|\Delta_{(r)}\|_{\mathrm{HS}} < \eta_m/2\}$. *Then,* $\mathrm{pr}\{\Omega_m(n, N)\} \to 1$, *and on* $\Omega_m(n, N)$

$$E(\|\hat{\phi}_{(r),j} - \phi_j\|^2) \lesssim \frac{j^{a+2}}{n}\left(1 + \frac{1}{Nh}\right) + h^4 j^{2a+2c+2}, \quad j = 1, \ldots, m,$$

for $r = 1, 2$.

The event $\Omega_m(n, N)$ denotes the set of all realizations such that $\|\Delta_{(r)}\|_{\mathrm{HS}} < \eta_m/2$ for sample size n and sampling rate N (Hall & Horowitz, 2007). It is worth mentioning that "on $\Omega_m(n, N)$" is not in a sense relating to a conditioning argument in probability theory. It should be interpreted as stating that the obtained bounds are valid for all realizations for which $\|\Delta_{(r)}\|_{\mathrm{HS}} < \eta_m/2$. Alternatively, $\Omega_m(n, N)$ can be regarded as the index set containing those eigenfunctions that can be well estimated by the observed data. Theorem 2 is essential for most methods based on principal component analysis that needs to quantify the error of a diverging number of estimated eigenfunctions. Theorem 2 improves the bound $\|\hat{\phi}_{(r),j} - \phi_j\|^2 \lesssim \eta_j^{-2}\|\Delta_{(r)}\|_{\mathrm{HS}}^2$, which is the state-of-art result for eigenfunctions with diverging index from discretely observed data. The improvement is twofold. First, the result in Theorem 2 is a one-dimensional kernel smoothing rate compared to $\|\Delta_{(r)}\|_{\mathrm{HS}}^2$, since integration usually brings extra smoothness (Cai & Hall, 2006; Hall et al., 2006). Second, it reduces the magnitude of variance from j^{2a+2}/n to j^{a+2}/n. Although this bound may not be sharp, it suffices to derive optimal convergence for $\hat{\beta}$ in the subsequent analysis when coupled with the approximated least squared estimation.

The following theorem establishes our main result, the $\mathcal{L}^2$ convergence rate of $\hat{\beta}$.

THEOREM 3. *Under Conditions 1–5, on the high probability set* $\Omega_m(n, N)$,

$$\|\hat{\beta} - \beta\|^2 = O_p\left(\frac{m^{a+1}}{n} + \frac{1}{m^{2b-1}} + \delta_n\right), \tag{4}$$

where $\delta_n = m^a\mathcal{R} + (nNh)^{-1}$ *with*

$$\mathcal{R} = \left\{\frac{1}{N} + \frac{1}{n}\left(1 + \frac{1}{Nh}\right)\right\}\left\{\frac{m^{2a+3}}{n}\left(1 + \frac{1}{Nh}\right) + h^4 m^{3a+2c+3} + \frac{m^{a+1}}{N}\right\}. \quad (5)$$

In addition, if Condition 6 holds, $\hat{\beta}$ *reaches the optimal rate of convergence*

$$\|\hat{\beta} - \beta\|^2 = O_p\left(n^{-\frac{2b-1}{a+2b}}\right).$$

To appreciate this result, the first two terms in the right-hand side of (4) are the same as those for the fully observed case, while δ_n can be viewed as the contamination caused by the discrete observations and measurements error. Specifically, the terms in δ_n involving h^4 represent the estimation bias, and those involving $1/N$ are owing to the discrete approximation. The terms containing m and its positive powers in (4) and (5) arise from the increasing number of components in both eigen-decomposition and approximated least square, while the term $m^{-(2b-1)}$ reflects the approximation bias for the underlying β. The second assertion of Theorem 3 reveals that when $N \gtrsim \max\{n^{(2a+2)/(a+2b)}, n^{(a+b)/(a+2b)}\}$, this rate becomes optimal in the minimax sense (Hall & Horowitz, 2007), i.e., the phase transition occurs. The order of magnitude, $\max\{(2a+2)/(a+2b), (a+b)/(a+2b)\}$, which relies on the smoothness parameters, a and b, is always within $(1/2, 1)$. This differs from the phase transition of the pooled mean and covariance estimation that occurs at $O(n^{1/4})$ and reflects the elevated difficulties in estimating the slope function of the regression model; that is, more measurements are typically needed to reach the optimal rate of convergence.

Next we study the asymptotic behavior of the prediction error for the proposed estimator $\hat{\theta}_n(X^*)$. Let E_* be the expectation taken over X^*, T_j^* and ε_j^*, while E_{X^*} denotes the expectation taken over X^* only. The accuracy of prediction is naturally measured by the prediction error

$$\mathcal{E}(\hat{\theta}_n) = E_*[\{\hat{\theta}_n(X^*) - \theta_0(X^*)\}^2].$$

Recall that $\tilde{\theta}_n(X^*) = \int X^*(s)\hat{\beta}(s)\mathrm{d}s$ is the estimator of $\theta_0(X^*)$ based on the fully observed curve X^*. The following theorem states the asymptotic behaviors of $\mathcal{E}(\hat{\theta}_n)$ and $\mathcal{E}(\tilde{\theta}_n)$.

THEOREM 4. *Under Conditions 1–5, the predictor error* $\mathcal{E}(\hat{\theta}_n)$ *incurs a numerical approximation error of* $O_p(N^{-1})$ *that differs from that of the fully observed case, i.e.,*

$$\mathcal{E}(\hat{\theta}_n) = \mathcal{E}(\tilde{\theta}_n) + O_p\left(\frac{1}{N}\right),$$

and on the high probability set $\Omega_m(n, N)$,

$$\mathcal{E}(\tilde{\theta}_n) = O_p\left(\frac{m}{n} + \frac{1}{m^{a+2b-1}} + \delta_n'\right),$$

where $\delta_n' = \mathcal{R} + m^{1-a}/(nNh)$. *In addition, if Condition 6 holds,*

$$\mathcal{E}(\tilde{\theta}_n) = O_p\left(n^{-\frac{a+2b-1}{a+2b}}\right).$$

The first assertion of Theorem 4 indicates that $\mathcal{E}(\hat{\theta}_n)$ is N^{-1} larger than $\mathcal{E}(\tilde{\theta}_n)$, which is caused by the integral approximation and cannot be improved. We thus focus on the asymptotic behavior of $\mathcal{E}(\tilde{\theta}_n)$. Cai & Hall (2006) treated X^* as a fixed curve and obtained a root-n convergence rate for $\mathcal{E}(\tilde{\theta}_n)$. In contrast, Cai & Yuan (2012) regarded X^* as an independent copy of X and obtained the optimal nonparametric convergence for prediction error under the reproducing kernel Hilbert space framework. We adopt the principal component analysis framework

while treating the new observation as an independent copy of X. In this setting, the prediction error $\mathcal{E}(\tilde{\theta}_n) = E_{X^*}[\{\tilde{\theta}_n(X^*) - \theta_0(X^*)\}^2]$ can be reduced to $\sum_{j=1}^{\infty} \lambda_j \{\int (\hat{\beta} - \beta)\phi_j\}^2$ due to the Mercer's Theorem, and further calculations are based on the bounds in Theorem 1 and 2. In fact, the integration enables $\int X^* \hat{\beta}$ to be estimated at a faster rate which $\hat{\beta}$ itself could not reach. Our prediction admits a convergence rate faster than $n^{(1-2b)/(a+2b)}$, which is comparable to that in Cai & Yuan (2012) and is slightly slower than n^{-1} due to randomness of X^*.

Before introducing the minimax lower bound for the prediction error, a regularity condition on the regression noise e is needed. Let P_β be the conditional probability of Y given X and E_β denote the corresponding expectation. Based on Kullback-Leibler distance $K(\cdot \mid \cdot)$, assume that

Condition 7. $E\{K(P_{\beta_1} \mid P_{\beta_2})\} = E[E_{\beta_1}\{\log(dP_{\beta_1}/dP_{\beta_2})\}] \le K_{\sigma^2} E[\{\int X(\beta_1 - \beta_2)\}^2]$ for $\beta_1, \beta_2 \in \mathcal{L}^2[0,1]$, where K_{σ^2} is a positive variance-dependent constant.

Many examples including the exponential family ensure the existence of the constant K_{σ^2} in Condition 7, see Theorem 1 in Du & Wang (2014). When the regression noise is normal with mean zero and variance σ_Y^2, the inequality in Condition 7 becomes equality with $K_{\sigma^2} = (2\sigma_Y^2)^{-1}$. The next theorem, combining with Theorem 4, unveils that our proposed estimator is minimax optimal in the prediction sense.

THEOREM 5. *Suppose that the regression error e satisfies Condition 7, and denote $\mathcal{G}$ a functional class for X^*. Assume that*

(a.) $X^*(t)$ is a mean zero square integrable stochastic process on $[0,1]$ with the covariance function C that admits the decomposition $C(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, with eigenvalue satisfying $\lambda_j \lesssim j^{-a}$ for each $j \in \mathbb{N}_+$ and some positive constant $a > 1$;

(b.) $\beta(s) = \sum_{k=1}^{\infty} b_k \phi_k(t)$ with $|b_k| \lesssim k^{-b}$ for $k \in \mathbb{N}_+$.

Then

$$\liminf_{n\to\infty} n^{\frac{a+2b-1}{a+2b}} \inf_{\tilde{\beta}} \sup_{X^*, \beta \in \mathcal{G}} E\{\mathcal{E}(\tilde{\theta}_n)\} \ge \mathrm{C}, \quad \textit{for all alternative estimator } \tilde{\beta}.$$

We conclude this section by providing an outline of our theoretical developments for better appreciation. By introducing Condition 3 on frequency increments of eigenfunctions, after careful calculations on bias and variance terms, we establish the key result on expectation bounds of $E(\langle \Delta_{(r)}\phi_k, \phi_j\rangle^2)$ and $E(\|\Delta_{(r)}\|_j^2)$ for discretely observed functional covariates. Combing with Theorem 5.1.8 in Hsing & Eubank (2015), the improved bounds for the estimated eigenfunctions and scores are obtained. The sample-splitting strategy eliminates the dependence between X_i and $\hat{\phi}_k$, which makes the error quantification mathematically tractable. Applying these new bounds in the approximated least square equation, the main results on estimation and prediction error of $\hat{\beta}$ are consequently obtained.

4. NUMERICAL EXAMPLES

4.1. *Simulation study*

In this subsection, we first evaluate the numerical performance using simulated examples by contrasting with comparable methods, the plug-in method (Hall & Horowitz, 2007) and the approximated least squared method without sample-splitting. For the latter, we use two choices of scores, the integral approximation and the conditional expectation estimator (Yao et al., 2005a). For practical implementation, the truncation number m is chosen by the five-fold predictive cross-validation. The underlying trajectories are simulated as $X_i(t) = \sum_{j=1}^{50} \xi_{ij}\phi_j(t), i = 1, \ldots, n$, where the scores ξ_{ij}'s are generated from $N(0, \lambda_j)$ with $\lambda_j = j^{-2}$ and $\phi_1(t) = 1$, $\phi_j(t) = 2^{1/2}\cos\{(j-1)\pi t\}$ for $j \in \mathbb{N}_+$. The response, Y_i, is generated by

$\int_{[0,1]} \beta(s)X_i(s)\mathrm{d}s + e_i$, where the slope function $\beta(s) = \sum_{j=1}^{50} b_j\phi_j$ with $b_1 = 1$, $b_j = 4j^{-3}$ for $j > 1$ and $e_i \sim \mathcal{N}(0, 0.5^2)$. The corresponding observed data for each curve are generated by $X_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, 0.1^2)$ and $T_{ij} \sim \mathrm{Unif}[0, 1]$.

These methods are compared by the integrated mean squared error, $\{E(\|\hat{\beta} - \beta\|^2)\}^{1/2}$, and prediction error, $(2n)^{-1}\sum_{i=1}^{2n}(\hat{y_i^*} - y_i^*)^2$, where $X_{ij}^*, T_{ij}^*, y_i^*$ are random samples in the testing set and $\hat{y_i^*}$ is estimator of y_i^* based on $\hat{\beta}$ and X_{ij}^*. The sample size of the testing set is twice of the training set. We repeat the above procedure for 200 Monte Carlo runs and results are shown in Table 1. It is seen that the proposed method performs better than not only the plug-in method, but also the approximated least squared method with integrated scores and conditional expectation scores, respectively, in terms of both integrated mean squared error and prediction error. Furthermore, for a typical N, the prediction error decreases relatively slow as n grows and decreases faster as N grows for a fixed n. This explains the first assertion of Theorem 4, that is, when the approximation error $1/N$ dominates, the prediction error cannot be much improved in the discretely observed case.

Table 1: The Monte Carlo averages with standard errors in parentheses of integrated mean squared error and prediction error by five-fold cross validation based on 200 replications using both methods

	N	n	Plug-in	IN	PACE	Proposed
PE	10	100	0.3979(0.0852)	0.4281(0.0648)	0.4920(1.6318)	0.3624(0.0395)
		200	0.3902(0.0811)	0.4193(0.0461)	0.3634(0.0563)	0.3518(0.0310)
	30	100	0.3235(0.0535)	0.3312(0.0481)	0.3262(0.1026)	0.3104(0.0419)
		200	0.3089(0.0396)	0.3150(0.0299)	0.3053(0.0331)	0.2971(0.0268)
IMSE	10	100	0.4230(0.2018)	0.4167(0.1355)	0.4589(0.8767)	0.3799(0.1083)
		200	0.3924(0.2264)	0.3875(0.1405)	0.3502(0.1817)	0.3403(0.0950)
	30	100	0.3770(0.2209)	0.3930(0.1959)	0.3988(0.4032)	0.3283(0.1501)
		200	0.3125(0.2355)	0.3156(0.1616)	0.3091(0.2309)	0.2692(0.1230)

PE: prediction error; IMSE: integrated mean squared error; n: sample size; N: observations per subject; Plug-in: plug-in method in Hall & Horowitz (2007); IN: approximated least square method with integrated scores; PACE: approximated least square method with conditional expectation scores in Yao et al. (2005a); Proposed: our proposed method.

We conclude this subsection by an illustration on the convergence rates across different n and N. As our theory indicates that, when phase transition occurs, the logarithm of the integrated mean squared error and the logarithm of the sample size admit a linear relationship with the slope $(1 - 2b)/(a + 2b)$. As shown in the left panel of Figure 1, the pattern indeed tends to be linear of the slope $(1 - 2b)/(a + 2b)$ as N grows. For a fixed n ($\log n = 4$ for example), one sees that the slope shifts down across $N = 5, 10, 20, 30$, then tends to be stable for $N = 50, 80, 100$, indicating that the phase transition might occur around $N = 50$ (or slightly earlier).

4.2. *Real data example*

This subsection presents an illustration of our proposed method on a spectroscopic calibration dataset, the wheat data (Kalivas, 1997), which can be downloaded at https://github.com/nanxstats/OHPL/blob/d704bba6140437379d191bb80dcb158e92f99fde/data/wheat.rda. The aim is to predict the ingredient of interest given the near infrared spectrum of the sample. Meanwhile, the functional relationship is explored between near infrared trajectories and the constituents of the food.

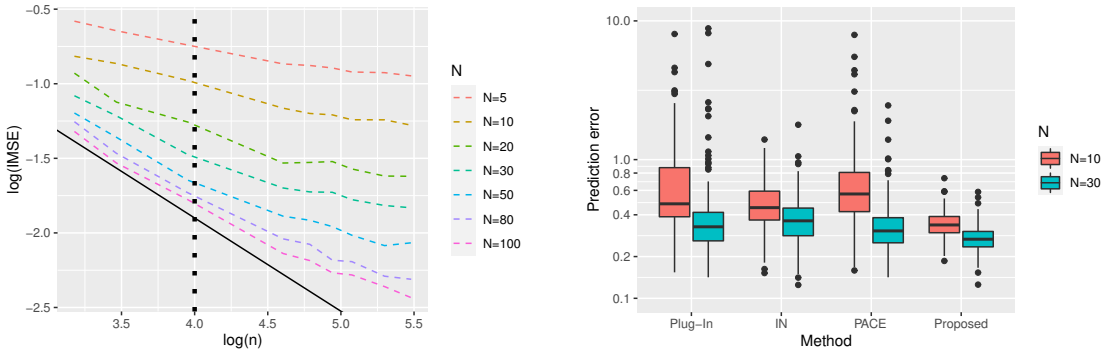

Fig. 1: Left: plot of $\log(\text{IMSE})$ for different N over $\log(n)$, the colored dashed lines correspond to different value of N, and the slope of the solid line represents the theoretical optimal value (noting the intercept is not relevant here). Right: box-plots for the prediction error on testing sets for different methods in wheat dataset.

The wheat data set contains 100 subjects, where the functional covariate is the near infrared spectrum measured at 701 equally spaced frequencies with a spacing of 2nm between 1100nm and 2500nm, and the response is the protein content. Here we use the first 80 samples as the training set and the last 20 samples as the testing set. The spectra scheme is typically (ultra) dense and for illustration, we randomly choose $N = 10$ and $N = 30$ measurements with equal probability to mimic two sparse and (moderately) dense designs, respectively. The proposed method is compared with three methods used in Section 4.1 and the tuning parameter, m, is selected by the five-fold cross validation. We repeat the random selection of N in the above procedure for 200 times and calculate the mean and standard error of the prediction error on the testing set, respectively, shown in Table 2. We find that the results are consistent with those in the simulation studies, that is, our proposed method performs better than the comparable methods. In addition, due to the numerically unstable inverse encountered in plug-in and conditional expectation methods, the Mento Carlo standard errors of these two methods are relatively large, which is also seen in the right panel of Figure 1.

Table 2: The Monte Carlo averages with standard errors in parentheses of prediction errors on the testing sets by five-fold cross validation based on 200 replications using both methods

N	Plug-in	IN	PACE	Proposed
10	0.7810(0.8404)	0.4993(0.1981)	1.5172(9.1116)	0.3453(0.0742)
30	0.5127(0.9252)	0.3982(0.1828)	0.3625(0.2443)	0.2778(0.0659)

N: observations per subject; Plug-in: plug-in method in Hall & Horowitz (2007); IN: approximated least square method with integrated scores; PACE: approximated least square method with conditional expectation scores in Yao et al. (2005a); Proposed: our proposed method.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes auxiliary lemmas and proofs.

ACKNOWLEDGEMENTS

Fang Yao is the corresponding author. This is supported by National Natural Science Foundation of China Grants 11931001 and 11871080, the National Key R&D Program of China Grant 2020YFE0204200, the LMAM, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education.

REFERENCES

BOSQ, D. (2000). *Linear Processes in Function Spaces*. New York: Springer.

CAI, T. T. & HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159–2179.

CAI, T. T. & YUAN, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39**, 2330–2355.

CAI, T. T. & YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Am. Statist. Assoc.* **107**, 1201–1216.

DAI, X., MÜLLER, H.-G. & TAO, W. (2018). Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statist. Sinica* **28**, 1583–1609.

DOU, W. W., POLLARD, D. & ZHOU, H. H. (2012). Estimation in functional regression for general exponential families. *Ann. Statist.* **40**, 2421–2451.

DU, L., GUO, X., SUN, W. & ZOU, C. (2021). False discovery rate control under general dependence by symmetrized data aggregation. *J. Am. Statist. Assoc.* , published online.

DU, P. & WANG, X. (2014). Penalized likelihood functional regression. *Statist. Sinica* **24**, 1017–1041.

HALL, P. & HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70–91.

HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components. *J. R. Statist. Soc.* B **68**, 109–126.

HALL, P., MÜLLER, H.-G. & WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–1517.

HSING, T. & EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Chichester: Wiley.

KALIVAS, J. H. (1997). Two data sets of near infrared spectra. *Chemomet. Intel. Lab. Syst.* **37**, 255–259.

LARSON, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.* **22**, 45–55.

LI, Y. & HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38**, 3321–3351.

MEINSHAUSEN, N., MEIER, L. & BÜHLMANN, P. (2009). P-values for high-dimensional regression. *J. Am. Statist. Assoc.* **104**, 1671–1681.

PAUL, D. & PENG, J. (2009). Consistency of restricted maximum likelihood estimators of principal components. *Ann. Statist.* **37**, 1229–1271.

RAMSAY, J. & SILVERMAN, B. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

WAHL, M. (2021). Lower bounds for invariant statistical models with applications to principal component analysis. *Ann. Inst. H. Poincaré Probab. Statist., to appear* .

WASSERMAN, L. & ROEDER, K. (2009). High dimensional variable selection. *Ann. Statist.* **37**, 2178–2201.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–590.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–2903.

YUAN, M. & CAI, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412–3444.

ZHANG, J.-T. & CHEN, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35**, 1052–1079.

ZHANG, X. & WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44**, 2281–2321.

ZOU, C., WANG, G. & LI, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Ann. Statist.* **48**, 413–439.