| Statistica Sinica Preprint No: SS-2016-0459 | |
|---|---|
| **Title** | Two-sample Functional Linear Models |
| **Manuscript ID** | SS-2016-0459 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202016.0459 |
| **Complete List of Authors** | Wenchao Xu,<br>Riquan Zhang and<br>Hua Liang |
| **Corresponding Author** | Riquan Zhang |
| **E-mail** | zhangriquan@163.com |
| Notice: Accepted version subject to English editing. | |

Statistica Sinica

# TWO-SAMPLE FUNCTIONAL LINEAR MODELS

Wenchao Xu[1], Riquan Zhang[1] and Hua Liang[2]

[1]*East China Normal University and* [2]*George Washington University*

*Abstract:* In this paper we study two-sample functional linear regression with a scaling transformation of regression functions. We consider estimation of the intercept, the slope function and the scalar parameter based on the functional principal component analysis. We also establish the rates of convergence for the estimator of the slope function, which is shown to be optimal in a minimax sense under certain smoothness assumptions. We further investigate semiparametric efficiency for the estimation of the scalar parameter and hypothesis testing. We also extend the proposed method to sparsely and irregularly sampled functional data and establish the consistency for the estimators of the scalar and the slope function. We evaluate numerical performance of the proposed methods through simulation studies and illustrate their utility via analysis of an AIDS data set.

*Key words and phrases:* Functional linear regression, functional principal component analysis, hypothesis testing, minimax rate of convergence, semiparametric comparison, semiparametric efficiency.

## 1. Introduction

Functional data analysis (FDA) has become increasingly more important in the past two decades. See the monographs by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Hsing and Eubank (2015), the articles by Yao, Müller, and Wang (2005a,b), Müller (2005), Hall, Müller, and Wang (2006), Li and Hsing (2010), Li, Wang, and Carroll (2013), Cuevas (2014), Chen et al. (2017) and Wang, Chiou, and Müller (2016), and the references therein.

This paper studies the semiparametric comparison of regression models in FDA. Specifically speaking, consider

$$Y = (1-U)r(X) + U\theta r(X) + \varepsilon = (1-U+U\theta)r(X) + \varepsilon, \qquad (1.1)$$

where $U$ is a Bernoulli random variable with $\pi = E(U) = P(U = 1)$, $\theta \in (0, \infty)$ is an unknown parameter, $X(t)$ is a random function in the class $L_2(\mathcal{I})$ of the square-integrable functions on a compact interval $\mathcal{I}$ of $\mathbb{R}^1$, $r(\cdot)$ is a functional from $L_2(\mathcal{I})$ to $\mathbb{R}^1$ and $\varepsilon$ is a random error, independent of $(U, X)$, with mean zero and finite variance $\sigma^2$. Furthermore, we assume that $U$ and $X$ are independent.

Model (1.1) refers to a two-sample problem; i.e., in the first sample $(U = 0)$ the relationship between $Y$ and $X(t)$ is described by $r(X)$ and in the second sample $(U = 1)$ this relationship changes to $\theta r(X)$. For independent data, Schick (1993) treated $r(\cdot)$ as a nonparametric function and established semiparametric efficiency for estimating $\theta$.

There are many possible choices for $r(\cdot)$, for example, the fully nonparametric form (Ferraty and Vieu, 2006), single index functional form (Chen, Hall, and Müller, 2011). In this paper, we entertain a linear relationship between $r$ and $X(t)$:

$$r(X) = a + \int_{\mathcal{I}} X(t)b(t)\,dt$$

with an unknown intercept $a$ and a square integrable slope function $b(t)$. As a result, we formulate our two-sample functional linear regression as

$$E\{Y|X(t), U\} = (1 - U + U\theta)\{a + \int_{\mathcal{I}} X(t)b(t)dt\}. \qquad (1.2)$$

Let $\{(Y_i, X_i, U_i), i = 1, \ldots, n\}$ be independent and identically distributed data from model (1.1). The goal is to estimate $\theta, a$ and $b(t)$ based on the sample. Model (1.2) is also related to the functional mixture regression (FMR) of Yao, Fu, and Lee (2011), which is an extension of classical finite mixture regression models (DeSarbo and Cron, 1988). However, they are different, since the group label for each observation of FMR is unknown, while it is known in (1.2). If $\theta \equiv 1$, (1.2) reduces to a functional linear model: $Y = a + \int_{\mathcal{I}} X(t)b(t)\,dt + \varepsilon$. This model has been investigated intensively in the literature. The focus is generally on the estimation of $a$ and $b(t)$. See, for example, Cardot, Ferraty, and Sarda (2003), Ramsay and Silverman (2005), Cai and Hall (2006), Hall and Horowitz (2007), Li and Hsing

(2007), Crambes, Kneip, and Sarda (2009), Yuan and Cai (2010), and Cai and Yuan (2012). The most frequently used approach for estimating $b(t)$ was developed on the basis of functional principal components analysis (F-PCA) or reproducing kernel Hilbert space (RKHS) methods. Cai and Hall (2006) and Hall and Horowitz (2007) studied the prediction and estimation of the slope function $b(t)$ based on the FPCA method. Yuan and Cai (2010) and Cai and Yuan (2012) investigated the estimation of the slope function and adaptive prediction based on the RKHS method, while Cardot, Ferraty, and Sarda (2003) and Li and Hsing (2007) considered approximating $b(t)$ and $X(t)$ by B-spline and Fourier approximation, respectively. They established the rates of convergence for the resulting estimators or predictions under various assumptions. More recently, Lei (2014) studied global testing for $b(t)$ based on the FPCA approach, and Shang and Cheng (2015) studied statistical inference for (generalized) functional linear models under the RKHS framework.

In this paper, we adopt the FPCA method to estimate the unknown slope function $b(t)$. FPCA is essentially a dimension reduction procedure that has been well examined in the literature. See, for example, James, Hastie, and Sugar (2000), Yao, Müller, and Wang (2005a), Hall, Müller, and Wang (2006), and Li and Hsing (2010). We modify the method proposed by He, Müller, and Wang (2000) and Yao, Müller, and Wang (2005b) to fit our setting. First, we use the population least squares to obtain basis representations of $\theta$, $a$ and $b(t)$. Then, we replace the unknown quantities by their empirical versions with finite terms. We derive the optimal rate of convergence for the FPCA-based estimator of the slope function $b(t)$ under certain smoothness assumptions, establish the consistency and asymptotic normality for the estimator of $\theta$, and show that this naive FPCA-based estimator is not efficient in the sense of Bickel et al. (1998). We then construct an asymptotically efficient estimator of $\theta$ and propose a test statistic for $\theta$. In practice $X_i$'s may be sparsely observed at a set of discrete points with noise (Yao, Müller, and Wang, 2005a,b; Li and Hsing, 2010; Zhang and Wang, 2016). We further extend the FPCA-based estimation method to sparsely and irregularly sampled functional data and establish asymptotic consistency properties for the resulting estimators.

The rest of the paper is organized as follows. Section 2 discusses the identifiability, derives the estimators of $\theta, a$ and $b(t)$, and investigates the asymptotic properties for the proposed estimators, such as consistency and asymptotic normality for the estimator of the primary parameter $\theta$ and the rates of convergence and optimality for the estimator of the slope function $b(t)$. Section 3 derives the efficient influence function for estimating $\theta$, and constructs an efficient estimator. We propose a testing procedure for $\theta$ in Section 4. Section 5 extends the proposed estimator to sparsely and irregularly sampled functional data. Section 6 presents simulation studies for evaluating the finite-sample performance of the proposed procedures. Section 7 analyzes a dataset from an AIDS study. All proofs are relegated to the Supplementary Materials.

## 2. Identifiability and Estimation

In this section, we first explore the identifiability issue for the model (1.2). Then we use the population least squares to obtain basis representations of $\theta, a$ and $b(t)$ (He, Müller, and Wang, 2000; Yao, Müller, and Wang, 2005b). The proposed estimators are obtained by replacing the unknown quantities in the representations with their empirical versions. In what follows, we write $\int pq$ for $\int_{\mathcal{I}} p(t)q(t)\,dt$.

### 2.1 Model Identifiability

First, we show that the functional model (1.1) is identifiable under mild conditions on the distribution of $X$. Let the covariance function of $X(\cdot)$ be $K(s,t) = \mathrm{Cov}\{X(s), X(t)\}$. Its corresponding covariance operator $K : L_2(\mathcal{I}) \to L_2(\mathcal{I})$, is defined by the mapping $(Kf)(s) = \int_{\mathcal{I}} K(s,t)f(t)\,dt$ for any $f \in L_2(\mathcal{I})$. If $K$ is continuous and square integrable, we have the spectral decomposition from Mercer's theorem (Hsing and Eubank, 2015, pp 120): $K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and $\phi_1, \phi_2, \ldots$ are the orthonormal eigenfunctions of the operator $K$. The $\phi_j$'s are also known as the functional principal components. The operator $K$ is of full rank in $L_2(\mathcal{I})$ (Hall and Hooker, 2016) in the sense that all $\lambda_j's \neq 0$ and $\phi_1, \phi_2, \ldots$ are complete in $L_2(\mathcal{I})$.

**Proposition 1.** *Suppose $a \neq 0$ or $b(t) \neq 0$ almost everywhere on $\mathcal{I}$. If there exist an alternative model intercept $a_1$, a slope function $b_1(t)$ and a scalar parameter $\theta_1$, such that*

$$P\left\{(1-U+U\theta)(a+\int Xb) = (1-U+U\theta_1)(a_1+\int Xb_1)\right\} = 1, \quad (2.3)$$

*then $a = a_1, \theta = \theta_1$ and $b(t) = b_1(t)$ for almost all $t \in \mathcal{I}$.*

Throughout this paper, we assume that $K$ is of full rank and $a \neq 0$ or $b(t) \neq 0$ almost everywhere on $\mathcal{I}$.

## 2.2 Population Least Squares

Let $\Xi = (0, +\infty) \times \mathbb{R} \times L_2(\mathcal{I})$ and

$$S(\vartheta, \nu, \xi) = E\left\{Y - (1-U+U\vartheta)\left(\nu + \int_{\mathcal{I}} X(t)\xi(t)\,dt\right)\right\}^2.$$

It follows from the proof of Proposition 1 that $\theta, a$ and $b(t)$ are the unique minimum of $S(\vartheta, \nu, \xi)$ over $(\vartheta, \nu, \xi) \in \Xi$, that is

$$(\theta, a, b) = \arg \min_{(\vartheta,\nu,\xi)\in\Xi} S(\vartheta, \nu, \xi).$$

Recall that $U$ and $X$ are independent. It is clear that

$$a = \frac{(1-\pi)\mu_0 + \pi\theta\mu_1}{1-\pi+\pi\theta^2} - \int_{\mathcal{I}} \mu_X(t) b(t)\, dt, \tag{2.4}$$

where $\mu_j = E(Y|U=j), j = 0, 1$ and $\mu_X(t) = E\{X(t)\}$. It is easy to verify that $\mu_1 = \theta\mu_0$. Consequently, finding $\vartheta, \nu$ and $\xi(\cdot)$ to minimize $S(\vartheta, \nu, \xi)$ is equivalent to finding $\vartheta$ and $\xi(\cdot)$ to minimize

$$E\left[Y - \frac{1-U+U\vartheta}{1-\pi+\pi\vartheta^2}\{(1-\pi)\mu_0 + \pi\vartheta\mu_1\} \right.$$
$$\left. -(1-U+U\vartheta)\int_{\mathcal{I}} \{X(t) - \mu_X(t)\}\xi(t)\, dt\right]^2. \tag{2.5}$$

Define two cross-covariance functions:

$$g(t) = E\{(Y-\mu_Y)(X(t)-\mu_X(t))|U=1\},$$

$$h(t) = E\{(Y-\mu_Y)(X(t)-\mu_X(t))|U=0\},$$

where $\mu_Y = E(Y)$. Then $g(t) = \theta h(t)$ for all $t \in \mathcal{I}$.

Moreover, if we expand $b(t) = \sum_{j=1}^{\infty} b_j\phi_j(t)$, $g(t) = \sum_{j=1}^{\infty} g_j\phi_j(t)$ and $h(t) = \sum_{j=1}^{\infty} h_j\phi_j(t)$, with $b_j = \int b\phi_j, g_j = \int g\phi_j$ and $h_j = \int h\phi_j$, then by minimizing the objective function (2.5) subject to $\vartheta$ and $\xi(\cdot)$, we can get

$$b_j = \lambda_j^{-1}\frac{(1-\pi)h_j + \pi\theta g_j}{1-\pi+\pi\theta^2}, \tag{2.6}$$

and

$$\theta = \frac{\sum_{j=1}^{\infty} \lambda_j^{-1} g_j^2 + \mu_1^2}{\sum_{j=1}^{\infty} \lambda_j^{-1} g_j h_j + \mu_0 \mu_1}. \tag{2.7}$$

Furthermore $\sum_{j=1}^{\infty} \lambda_j^{-1} h_j^2 = E\{\int (X - \mu_X) b\}^2 \leq \int E(X - \mu_X)^2 \int b^2$ by using the Cauchy-Schwarz inequality. Recall that $\int E(X^2) < \infty$ and $g_j = \theta h_j$ and the assumption that $b(t)$ is square-integrable, then $\sum_{j=1}^{\infty} \lambda_j^{-1} g_j^2$, $\sum_{j=1}^{\infty} \lambda_j^{-1} g_j h_j$ and $\sum_{j=1}^{\infty} \lambda_j^{-1} h_j^2$ are all convergent. Hence the right-hand side of (2.7) is well-defined.

## 2.3 Estimation

Next, we describe the empirical versions of the basis representations of $\theta, a$ and $b(t)$. The conventional estimator $\widehat{K}$ of $K$ is defined as

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

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Mercer's theorem implies the spectral decomposition of the covariance functions $\widehat{K}$ as $\widehat{K}(s, t) = \sum_{j=1}^{\infty} \widehat{\lambda}_j \widehat{\phi}_j(s) \widehat{\phi}_j(t)$, where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq 0$ are eigenvalues and $\widehat{\phi}_1, \widehat{\phi}_2, \ldots$ are the corresponding orthonormal eigenfunctions. Notice that $\widehat{\lambda}_j$ vanish for $j \geq n + 1$, so the functions $\widehat{\phi}_{n+1}, \widehat{\phi}_{n+2}, \ldots$ may be chosen arbitrarily.

Define

$$\widehat{g}(t) = \frac{1}{n\widehat{\pi}} \sum_{i=1}^{n} Y_i U_i \{X_i(t) - \bar{X}(t)\},$$
$$\widehat{h}(t) = \frac{1}{n(1-\widehat{\pi})} \sum_{i=1}^{n} Y_i (1-U_i) \{X_i(t) - \bar{X}(t)\},$$

where $\widehat{\pi} = n^{-1}\sum_{i=1}^{n} U_i$. Note that $E[YU\{X(t) - \mu_X(t)\}] = \pi g(t)$. Therefore, we can treat $\widehat{g}(t)$ as an estimator of $g(t)$. Similarly, $\widehat{h}(t)$ is an estimator of $h(t)$. Note that we can represent $\widehat{g}(t) = \sum_{j=1}^{\infty} \widehat{g}_j \widehat{\phi}_j(t)$ and $\widehat{h}(t) = \sum_{j=1}^{\infty} \widehat{h}_j \widehat{\phi}_j(t)$ with $\widehat{g}_j = \int \widehat{g}\widehat{\phi}_j$ and $\widehat{h}_j = \int \widehat{h}\widehat{\phi}_j$.

(2.7) suggests an estimator of $\theta$:

$$\widehat{\theta} = \frac{\sum_{j=1}^{m_n} \widehat{\lambda}_j^{-1} \widehat{g}_j^2 + \widehat{\mu}_1^2}{\sum_{j=1}^{m_n} \widehat{\lambda}_j^{-1} \widehat{g}_j \widehat{h}_j + \widehat{\mu}_0 \widehat{\mu}_1}, \tag{2.8}$$

where $\widehat{\mu}_0 = \{n(1-\widehat{\pi})\}^{-1} \sum_{j=1}^{n} Y_j(1-U_j)$ and $\widehat{\mu}_1 = (n\widehat{\pi})^{-1}\sum_{j=1}^{n} Y_j U_j$ are the sample average of $\mu_0$ and $\mu_1$, respectively, while $m_n$ is a positive integer less than $n$. Assumptions on $m_n$ will be imposed later. In practice, $m_n$ can be chosen by cross-validation.

(2.6) motivates us an estimator of $b(t)$:

$$\widehat{b}(t) = \sum_{j=1}^{m_n} \widehat{b}_j \widehat{\phi}_j(t), \quad \text{where} \quad \widehat{b}_j = \frac{(1-\widehat{\pi})\widehat{h}_j + \widehat{\pi}\widehat{\theta}\widehat{g}_j}{\widehat{\lambda}_j(1-\widehat{\pi}+\widehat{\pi}\widehat{\theta}^2)}. \tag{2.9}$$

Finally, (2.4) suggests an estimator of $a$:

$$\widehat{a} = \frac{(1-\widehat{\pi})\widehat{\mu}_0 + \widehat{\pi}\widehat{\theta}\widehat{\mu}_1}{1-\widehat{\pi}+\widehat{\pi}\widehat{\theta}^2} - \int_{\mathcal{I}} \bar{X}(t)\widehat{b}(t)\, dt. \tag{2.10}$$

## 2.4 Asymptotic Properties

We now derive the asymptotic normality for the estimator $\widehat{\theta}$ and the rate of convergence for the estimator $\widehat{b}(t)$ under the $L_2$-norm, and show that the rate of convergence is optimal in the minimax sense.

The Karhunen-Loève expansion of the random function $X(t)$ is given by $X(t) = \mu_X(t) + \sum_{j=1}^{\infty} \xi_j \phi_j(t)$, where the random variables $\xi_j = \int (X - \mu_X)\phi_j$ are uncorrelated random variables with mean zero and variance $E(\xi_j^2) = \lambda_j$, known as functional principal component scores. Let $C > 1$ be a constant large enough. We make the assumptions.

(A1) $X(t)$ has finite fourth moment in the sense that $\int_{\mathcal{I}} E\{X^4(t)\}dt < \infty$; $E(\xi_j^4) \le C\lambda_j^2$ for all $j \ge 1$.

(A2) $C^{-1}j^{-\alpha} \le \lambda_j \le Cj^{-\alpha}$ and $\lambda_j - \lambda_{j+1} \ge C^{-1}j^{-\alpha-1}$ for some $\alpha > 1$ and all $j \ge 1$.

(A3) $|b_j| \le Cj^{-\beta}$ for some $\beta > \alpha/2 + 1$ and all $j \ge 1$.

(A4) $m_n \to \infty$ and $m_n^{2\alpha+2}/n \to 0$ as $n \to \infty$.

Assumptions (A1)–(A3) are standard conditions in the literature of functional linear regression if the FPCA approach is used. See, e.g. Cai and Hall (2006) and Hall and Horowitz (2007). In Assumption (A2), $\alpha$

measures the smoothness of the covariance function $K$, and also impacts on the rate of convergence in estimating the slope function $b(t)$ (Theorem 2 below). The second part of Assumption (A2) requires that the space among $\lambda_j$ are not too small to ensure that each individual $\phi_j$ is identifiable. Assumption (A3) implicates that $b(t)$ is sufficiently smooth given $\beta > \alpha/2+1$. See Hall and Horowitz (2007) for detailed discussions on these assumptions. Assumption (A4) is a technical condition for proving Theorems below. The same assumption has been made by Imaizumi and Kato (2018) for functional linear regression with functional responses. Note that if $m_n \asymp n^{1/(\alpha+2\beta)}$, it is easy to verify that Assumption (A4) holds, where for two positive sequences $r_n$ and $s_n$, $r_n \asymp s_n$ means that $r_n/s_n$ is bounded away from 0 and $\infty$ as $n \to \infty$.

**Theorem 1.** *Under Assumptions (A1)–(A4), $\widehat{\theta}$ is a consistent estimator of $\theta$. Furthermore, we have*

$$\sqrt{n}(\widehat{\theta}-\theta) = n^{-1/2}\sum_{i=1}^{n}\psi(\theta; Y_i, X_i, U_i) + o_p(1) \xrightarrow{d} N\left(0, \frac{u_4\theta^2+\sigma^2 u_2(1-\pi+\pi\theta^2)}{\pi(1-\pi)u_2^2}\right),$$

*where*

$$\psi(\theta; Y, X, U) = \left(\frac{U}{\pi}-\frac{1-U}{1-\pi}\right)\frac{r^2(X)}{u_2}\theta + \left(\frac{U}{\pi}-\theta\frac{1-U}{1-\pi}\right)\frac{r(X)}{u_2}\varepsilon$$

*is the influence function of* $\theta$ *and* $u_k = E\{r(X)\}^k = E\big(a + \int Xb\big)^k$ *for* $k = 2, 4$.

**Remark 1.** Assumption (A1) ensures that $u_2$ and $u_4$ are finite. The result of Theorem 1 implies that when $\pi$ gets close to 0 or 1, the asymptotic variance of $\widehat{\theta}$ can be very large. So the performance of the estimator $\widehat{\theta}$ may be poor when the sample size of one group is too small compared to that of another one.

Next, we establish the asymptotic property for $\widehat{b}(t)$. Let $\mathcal{F} = \mathcal{F}(C, \alpha, \beta)$ denote the set of all distributions $F$ of $(Y, X, U)$ compatible with Assumptions (A1)–(A3) for given values of $C, \alpha$ and $\beta$. Then, following Theorem 1 of Hall and Horowitz (2007), we obtain the same rate of convergence of $\widehat{b}(t)$ as Hall and Horowitz (2007) did.

**Theorem 2.** *Suppose that Assumptions (A1)–(A3) are satisfied. Take* $m_n \asymp n^{1/(\alpha+2\beta)}$. *Then, we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} P_F\left[\int_{\mathcal{I}} \{\widehat{b}(t) - b(t)\}^2\, dt > Mn^{-(2\beta-1)/(\alpha+2\beta)}\right] = 0. \quad (2.11)$$

*Furthermore,*

$$\liminf_{n\to\infty} n^{(2\beta-1)/(\alpha+2\beta)} \inf_{\bar{b}} \sup_{F\in\mathcal{F}} E_F \int_{\mathcal{I}} \{\bar{b}(t) - b(t)\}^2\, dt > 0, \quad (2.12)$$

*where* $\inf_{\bar{b}}$ *is taken over all possible estimators* $\bar{b}$.

(2.12) shows that the minimax lower bound of the convergence rate for estimating $b(t)$ is $n^{-(2\beta-1)/(\alpha+2\beta)}$ and (2.11) indicates that this rate is achieved with $m_n \asymp n^{1/(\alpha+2\beta)}$. Therefore, $\widehat{b}(t)$ with $m_n \asymp n^{1/(\alpha+2\beta)}$ is a rate-optimal estimator and $n^{-(2\beta-1)/(\alpha+2\beta)}$ is the minimax optimal rate of convergence under the $L_2$-risk, which is determined by the smoothness of the slope function and the decay rate of the eigenvalues of the covariance function.

## 3. Semiparametric Efficiency

The estimator $\widehat{\theta}$ of the parameter $\theta$ proposed in Section 2 was actually derived on a basis of the expression $g_j = \theta h_j$ and $\mu_1 = \theta\mu_0$. This hints us that there are many potential estimators of $\theta$, for example, $\widehat{g}_1/\widehat{h}_1$ or $(\widehat{g}_1 + 2\widehat{g}_2)/(\widehat{h}_1 + 2\widehat{h}_2)$. A natural question raised here is whether $\widehat{\theta}$ is optimal among all regular estimators of $\theta$. We now investigate semiparametric efficiency for the semiparametric model (1.1). We will demonstrate that $\widehat{\theta}$ is not semiparametrically efficient even when $\varepsilon$ is normally distributed, and derive the efficient score and propose an efficient estimator based on $\widehat{\theta}$ when $\varepsilon$ is normally distributed.

To achieve this goal, we first derive the efficient score and information bound. Similar derivation for general semiparametric models for independent data refers to Severini and Wong (1992); Bickel et al. (1998); Brown and Newey (1998).

Suppose $\varepsilon \sim N(0, \sigma^2)$. We show in Section S5 of the Supplementary Materials that for model (1.2), the efficient score for $\theta$ is

$$\dot{l}_\theta^* = \frac{U(1-\pi) - (1-U)\pi\theta}{(1-\pi+\pi\theta^2)\sigma^2} r(X)\varepsilon. \tag{3.13}$$

Then the semiparametric information bound for $\theta$ is

$$I(\theta) = E(\dot{l}_\theta^{*2}) = \frac{\pi(1-\pi)}{(1-\pi+\pi\theta^2)\sigma^2} u_2. \tag{3.14}$$

This means that the lower bound on the asymptotic variance of regular estimators of $\theta$ is $(1-\pi+\pi\theta^2)\sigma^2/\{\pi(1-\pi)u_2\}$. Theorem 1 indicates that $\widehat{\theta}$ can not achieve this bound, and $\widehat{\theta}$ is not semiparametrically efficient even if $\varepsilon$ is normally distributed.

Next, we construct a more efficient estimator of $\theta$ than $\widehat{\theta}$ using $\widehat{\theta}$ as a preliminary estimator, and demonstrate that the resultant estimator is semiparametrically efficient when $\varepsilon$ follows the normal distribution. On a basis of Bickel et al. (1998), the efficient influence function for $\theta$ is given by

$$\psi^*(\theta; Y, X, U) = I^{-1}(\theta)\dot{l}_\theta^* = \left(\frac{U}{\pi} - \theta\frac{1-U}{1-\pi}\right)\frac{r(X)}{u_2}\varepsilon.$$

Thus, we construct the estimator of $\theta$:

$$\widehat{\theta}^* = \widehat{\theta} + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{U_i}{\widehat{\pi}} - \widehat{\theta}\frac{1-U_i}{1-\widehat{\pi}}\right)\frac{\widehat{r}(X_i)}{\widehat{u}_2}\widehat{\varepsilon}_i, \tag{3.15}$$

where $\widehat{r}(X_i) = \widehat{a} + \int_{\mathcal{I}} X_i(t)\widehat{b}(t)\,dt, \widehat{u}_2 = n^{-1}\sum_{i=1}^n \widehat{r}^2(X_i)$ and $\widehat{\varepsilon}_i = Y_i - (1 - U_i + U_i\widehat{\theta})\widehat{r}(X_i)$. $\widehat{\theta}^*$ is derived as a one-step Newton-Raphson approximation essentially.

**Theorem 3.** *Under the assumptions of Theorem 2, the estimator $\widehat{\theta}^*$ is asymptotically normal, i.e.,*

$$\sqrt{n}(\widehat{\theta}^* - \theta) = n^{-1/2}\sum_{i=1}^n \psi^*(\theta; Y_i, X_i, U_i) + o_p(1) \xrightarrow{d} N(0, I^{-1}(\theta)).$$

*Furthermore, when $\varepsilon$ follows the normal distribution, $\widehat{\theta}^*$ is semiparametrically efficient.*

**Remark 2.** Note that when the density function of $\varepsilon$ is known but not normal or unknown, $\widehat{\theta}^*$ is not semiparametrically efficient. Schick (1993) constructed an efficient estimator for $\theta$ in model(1.1) using a discretized root-$n$ preliminary estimator when the error density function is unknown. It is also of interest to derive such an efficient estimator of $\theta$ in model (1.2) if the error density function is unknown. We leave this as a future topic.

Once the more efficient estimator $\widehat{\theta}^*$ is available, we can update the estimators of $a$ and $b(t)$ as follows.

$$\widehat{b}^*(t) = \sum_{j=1}^{m_n} \widehat{b}_j^*\widehat{\phi}_j(t) \quad \text{with} \quad \widehat{b}_j^* = \widehat{\lambda}_j^{-1}\frac{(1-\widehat{\pi})\widehat{h}_j + \widehat{\pi}\widehat{\theta}^*\widehat{g}_j}{1 - \widehat{\pi} + \widehat{\pi}\widehat{\theta}^{*2}},$$

$$\widehat{a}^* = \frac{(1-\widehat{\pi})\widehat{\mu}_0 + \widehat{\pi}\widehat{\theta}^*\widehat{\mu}_1}{1 - \widehat{\pi} + \widehat{\pi}\widehat{\theta}^{*2}} - \int_{\mathcal{I}} \bar{X}(t)\widehat{b}^*(t)\,dt.$$

From the proof of Theorem 2 in the Supplementary Materials, $\widehat{b}^*(t)$ with $m_n \asymp n^{1/(\alpha+2\beta)}$ is also a rate-optimal estimator. Theoretically, $\widehat{b}(t)$ and $\widehat{b}^*$ have the same rate of convergence, but we will see that $\widehat{b}^*$ has a better finite sample performance in Section 6.

## 4. Hypothesis Testing

The scalar parameter $\theta$ is sometimes of primary interest. For example, $\theta = 1$ means that the two curves of two groups are identical and indicates that the two corresponding treatments have similar effects. So, we may be interested in testing $\theta = 1$. In general, we can test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Theorem 2 implies that $\{nI(\theta)\}^{1/2}(\widehat{\theta}^* - \theta) \to N(0, 1)$ in distribution. This result can be used to derive a test statistic after we estimate the information bound $I(\theta)$ by substituting all unknown quantities with their estimates. We estimate $I(\theta)$ by $\widehat{I}(\theta)$:

$$\widehat{I}(\theta) = \frac{\widehat{\pi}(1-\widehat{\pi})}{(1-\widehat{\pi}+\widehat{\pi}\widehat{\theta}^{*2})\widehat{\sigma}^{*2}}\widehat{u}_2^*,$$

where

$$\widehat{u}_2^* = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{a}^* + \int_{\mathcal{I}} X_i(t)\widehat{b}^*(t)\,dt\right\}^2 \quad \text{and} \quad \widehat{\sigma}^{*2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (1-\widehat{\pi}+\widehat{\pi}\widehat{\theta}^{*2})\widehat{u}_2^*.$$

From the proof of Lemma 5 in Section S6 of the Supplementary Materials, we can also prove that $\widehat{u}_2^*$ converges to $u_2$ in probability. By Theorem 3, it is easy to verify that $\widehat{I}(\theta)$ is a consistent estimator of $I(\theta)$. Consequently, we propose a test statistic:

$$T_n^* = \{n\widehat{I}(\theta)\}^{1/2}(\widehat{\theta}^* - \theta_0).$$

This statistic is asymptotically normal under $H_0$ by using the Slutsky theorem. This suggests rejecting $H_0$ when $|T_n^*|$ is larger than $z_{1-\tau/2}$, where $z_\tau$ is the $\tau$-th quantile of the standard normal distribution. The procedure is equivalent to the Wald-type test.

## 5. Extension to Sparse and Irregular Data

The methodological and theoretical development in the previous sections was done on a basis of the assumption that predictor trajectory $X(t)$ is fully observed without noise, which may not be true in practice. In this section, we assume that $X_i(t)$ can only be realized at some discrete set of sampling points with additional measurement errors; i.e., we observe data

$$W_{ij} = X_i(T_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, N_i, \tag{5.16}$$

where $\epsilon_{ij}$'s are i.i.d. measurement errors with mean zero and finite variance $\sigma_\epsilon^2$, and each $N_i \geq 2$. Assume that the $X_i$'s, $T_{ij}$'s, and $\epsilon_{ij}$'s are all independent.

Most existing literature has classified functional data into sparse and dense according to the number of observations within each curve; see Li and Hsing (2010). For dense functional data, we can smooth each individual curve first to construct the curve $\widehat{X}_i$ from the data $\mathcal{D}_i = \{(T_{ij}, W_{ij}) : 1 \leq j \leq N_i\}$ (Ramsay and Silverman, 2005). It has been shown by Hall, Müller, and Wang (2006) that when the observations are dense enough, the smoothing errors are asymptotically negligible. Therefore, the methodology developed in the previous sections would be carried out as if $\widehat{X}_i$ were the true curve. For sparse functional data, however, such a pre-smoothing method is inadequate.

The proposed estimation procedure in Section 2 can be extended to the case of sparse and irregular designs. A key step is to estimate $\mu_X(t), K(s,t)$ and $g(t), h(t)$, based on sparsely observed longitudinal data $\mathcal{D} = \{(T_{ij}, W_{ij}) : 1 \leq i \leq n, 1 \leq j \leq N_i\}$. We adapt the idea of pooling sparse longitudinal data across subjects and apply the local linear smoother to the resulting scatter plots (Yao, Müller, and Wang, 2005a,b; Hall, Müller, and Wang, 2006; Li and Hsing, 2010; Zhang and Wang, 2016). Let $\kappa(\cdot)$ be a univariate kernel function. Then the mean function $\mu_X$, covariance function $K$, and cross-covariance functions $f, g$ are estimated as follows. By an abuse of notation, we use $c_0$ and $c_1$ for local linear regression in estimating these functions in this section.

Step 1 The local linear estimator of the mean function $\mu_X(t)$ is $\widetilde{\mu}_X(t) = \widehat{c}_0$, where

$$(\widehat{c}_0, \widehat{c}_1) = \arg\min_{c_0,c_1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \kappa\left(\frac{T_{ij}-t}{d_\mu}\right)\{W_{ij} - c_0 - c_1(T_{ij}-t)\}^2$$

with a bandwidth $d_\mu$.

Step 2 Let $G_i(T_{ij}, T_{il}) = \{W_{ij} - \widetilde{\mu}_X(T_{ij})\}\{W_{il} - \widetilde{\mu}_X(T_{il})\}$ for $1 \le j, l \le N_i$. The local linear estimator of the covariance function $K(s,t)$ is $\widetilde{K}(s,t) = \widehat{c}_0$, where

$$(\widehat{c}_0, \widehat{c}_1, \widehat{c}_2) = \arg\min_{c_0,c_1,c_2} \sum_{i=1}^{n} \sum_{1\le j\ne l\le N_i} \kappa\left(\frac{T_{ij}-s}{d_K}\right)\kappa\left(\frac{T_{il}-t}{d_K}\right)$$
$$\times \{G_i(T_{ij}, T_{il}) - c_0 - c_1(T_{ij}-s) - c_2(T_{il}-t)\}^2$$

with a bandwidth $d_K$.

Step 3 Let $C_i(T_{ij}) = Y_i\{W_{ij} - \widetilde{\mu}_X(T_{ij})\}$ for $1 \le j \le N_i$. The local linear estimators of the cross-covariance functions $g(t)$ and $h(t)$ are $\widetilde{g}(t) = \widehat{c}_0/\widehat{\pi}$ and $\widetilde{h}(t) = \widetilde{c}_0/(1-\widehat{\pi})$, respectively, where

$$(\widehat{c}_0, \widehat{c}_1) = \arg\min_{c_0,c_1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \kappa\left(\frac{T_{ij}-t}{d_g}\right)\{C_i(T_{ij})U_i - c_0 - c_1(T_{ij}-t)\}^2,$$

$$(\widetilde{c}_0, \widetilde{c}_1) = \arg\min_{c_0,c_1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \kappa\left(\frac{T_{ij}-t}{d_h}\right)\{C_i(T_{ij})(1-U_i) - c_0 - c_1(T_{ij}-t)\}^2$$

with bandwidths $d_g$ and $d_h$.

Bandwidths $d_\mu, d_K, d_g$ and $d_h$ for the above smoothing steps are selected by leave-one-curve-out cross-validation or generalized cross-validation. We denote the estimators of $\lambda_j$ and $\phi_j(t)$ by $\widetilde{\lambda}_j$ and $\widetilde{\phi}_j(t)$, respectively, which can be calculated from an eigenvalue decomposition of $\widetilde{K}(\cdot,\cdot)$ by discretization and matrix spectral decomposition (Yao, Müller, and Wang, 2005a). Therefore, motivated by population representations in Section 2, $\theta, a$, and $b(t)$ are estimated as follows.

$$\widetilde{\theta} = \frac{\sum_{j=1}^{m_n} \widetilde{\lambda}_j^{-1} \widetilde{g}_j^2 + \widehat{\mu}_1^2}{\sum_{j=1}^{m_n} \widetilde{\lambda}_j^{-1} \widetilde{g}_j \widetilde{h}_j + \widehat{\mu}_0 \widehat{\mu}_1}, \quad \text{and} \quad \widetilde{b}(t) = \sum_{j=1}^{m_n} \widetilde{b}_j \widetilde{\phi}_j(t)$$

with

$$\widetilde{b}_j = \widetilde{\lambda}_j^{-1} \frac{(1-\widehat{\pi})\widetilde{h}_j + \widehat{\pi}\widetilde{\theta}\widetilde{g}_j}{1-\widehat{\pi}+\widehat{\pi}\widetilde{\theta}^2}, \quad \text{and} \quad \widetilde{a} = \frac{(1-\widehat{\pi})\widehat{\mu}_0 + \widehat{\pi}\widetilde{\theta}\widehat{\mu}_1}{1-\widehat{\pi}+\widehat{\pi}\widetilde{\theta}^2} - \int_{\mathcal{I}} \widetilde{\mu}_X(t)\widetilde{b}(t)\, dt,$$

where $\widetilde{f}_j = \int_{\mathcal{I}} \widetilde{f}(t)\widetilde{\phi}_j(t)\, dt$ and $\widetilde{g}_j = \int_{\mathcal{I}} \widetilde{g}(t)\widetilde{\phi}_j(t)\, dt$.

We establish consistency of the proposed estimators for sparse and irregular functional data. Let $\rho_{n1} = d_g^2 + (nd_g)^{-1/2}, \rho_{n2} = d_h^2 + (nd_h)^{-1/2}$ and $\rho_{n3} = d_K^2 + (nd_K^2)^{-1/2}$. We make the following assumptions for Theorem 4.

(B1) $\kappa(\cdot)$ is a symmetric probability density function on $[-1,1]$ and is Lipschitz continuous: There exists $0 < L < \infty$ such that $|\kappa(s) - \kappa(t)| \le L|s-t|$ for any $s, t \in [0,1]$.

(B2) $T_{ij}$'s are i.i.d. copies of a random variable $T$ defined on $\mathcal{I}$ with density function $\varphi_T(\cdot)$ and there exists some constants $m_T > 0$ and $M_T < \infty$ such that $m_T \leq \varphi_T(t) \leq M_T$ for all $t \in \mathcal{I}$. Furthermore, the second derivative of $\varphi_T(\cdot)$ is bounded on $\mathcal{I}$.

(B3) The second derivatives of $\mu_X(\cdot), g(\cdot)$ and $h(\cdot)$ are bounded on $\mathcal{I}$; All second-order partial derivatives of $K(s,t)$ are bounded on $\mathcal{I}^2$.

(B4) $d_\mu \to 0$ and $\log(n)/(nd_\mu) \to 0$.

(B5) $d_K \to 0$ and $\log(n)/(nd_K^2) \to 0$; $\sup_{t\in\mathcal{I}} E|X(t) - \mu_X(t)|^4 < \infty$ and $E|\epsilon_{ij}|^4 < \infty$.

(B6) $d_g \to 0$ and $\log(n)/(nd_g) \to 0$; $d_h \to 0$ and $\log(n)/(nd_h) \to 0$.

(B7) $m_n \to \infty$, $m_n^{\alpha+1/2}\rho_{n1} \to 0$, $m_n^{\alpha+1/2}\rho_{n2} \to 0$ and $m_n^{2\alpha+3/2}\rho_{n3} \to 0$ as $n \to \infty$.

Assumptions (B1)–(B5) are adapted from Zhang and Wang (2016). Assumptions (B4) and (B5) are special cases of (C1b)–(C3b) and (D1b)–(D3b) in Zhang and Wang (2016) for sparse functional data, respectively. Assumption (B6) is similar to (B4) and is used to establish the $L_2$ rates of convergence of $\widetilde{g}(t)$ and $\widetilde{h}(t)$. Assumption (B7) is a technique condition for proving Theorem 4.

**Theorem 4.** *Suppose that Assumptions (A2)–(A3) and (B1)–(B7) hold. For sparse data:* $\max_{1\le i\le n} N_i \le N_0 < \infty$*,* $\widetilde{\theta}$ *is consistent and*

$$\int_{\mathcal{I}} \{\widetilde{b}(t) - b(t)\}^2 \, dt \xrightarrow{p} 0.$$

**Remark 3.** Whether $\widetilde{\theta}$ remains root-$n$ consistency for sparse data and what is the rate of convergence of $\widetilde{b}(t)$ are not clear. In addition, the impact of $N_i$ on the asymptotic properties of $\widetilde{\theta}$ and $\widetilde{b}(t)$ is unknown. These topics warrant future work.

## 6. Simulation Studies

We conduct three Monte Carlo simulation studies to evaluate the numerical performance of the proposed estimation and test procedures. In Section 6.1, we examine the finite sample performance of $\widehat{\theta}, \widehat{\theta}^*$ and $\widehat{b}(\cdot), \widehat{b}^*(\cdot)$ for different sample sizes, variances of the error, and the smoothness of covariance function $K$. In Section 6.2, we assess the type I error rate and power of the statistic $T_n^*$. In Section S1 of the Supplemental Material, we examine the finite sample performance of $\widetilde{\theta}$ and $\widetilde{b}(\cdot)$.

### 6.1 Estimation

For $r(X) = a + \int_{\mathcal{I}} X(t)b(t)\, dt$, we adopt a design similar to that of Hall and Horowitz (2007) and Yuan and Cai (2010); that is, $\mathcal{I} = [0, 1]$, $a = 0$, and

$b(t)$ is given by

$$b(t) = 0.3\phi_1(t) + \sum_{k=2}^{50} 4(-1)^{k+1}k^{-2}\phi_k(t),$$

where $\phi_1(t) = 1$ and $\phi_{k+1}(t) = 2^{1/2}\cos(k\pi t)$ for $k \geq 1$. The random function $X(t)$ was generated as $X(t) = \sum_{k=1}^{50} \gamma_k Z_k \phi_k(t)$, where $Z_k$ were independently sampled from the uniform distribution on $[-3^{1/2}, 3^{1/2}]$. It is clear that the eigenvalues of the covariance function of $X(t)$ are $\gamma_k^2$. There are two sets of the $\gamma_k$, "well-spaced" and "closely spaced" eigenvalues, used in Hall and Horowitz (2007) and Yuan and Cai (2010). We only consider the "well-spaced" eigenvalues, in that $\gamma_k = (-1)^{k+1}k^{-\alpha/2}$ with $\alpha = 1.1, 1.5, 2, 2.5$.

Let $\theta = 1.5$ and $U$ follow the binomial distribution with a success probability $\pi = 0.6$. The error $\varepsilon$ follows the normal $N(0, \sigma^2)$, where $\sigma = 0.5$ or $1.0$. In addition, $X(t), U$, and $\varepsilon$ are independently. We consider $n = 200, 350, 500$ and $800$.

For each configuration, we repeated $Q = 1000$ times, and chose $m_n$ by 10-fold cross-validation. Table 1 presents the averages and standard deviations of the estimated $\widehat{\theta}$ and $\widehat{\theta}^*$. For each combination of $\alpha, \sigma$, the average of $\widehat{\theta}$ gets closer to the true value and the standard deviation decreases with increasing $n$. Comparing the results of $\widehat{\theta}$ with $\widehat{\theta}^*$, $\widehat{\theta}^*$ has a smaller standard deviation than $\widehat{\theta}$. This observation concurs that $\widehat{\theta}^*$ is more efficient than $\widehat{\theta}$.

We use the mean integrated squared error (MISE) to evaluate the per-

Table 1: The results for simulation study (estimation). Average and standard deviation (in parentheses) of estimators (Est) $\widehat{\theta}$ and $\widehat{\theta}^*$ given $\theta = 1.5$.

| Est | $\sigma$ | $n$ | $\alpha = 1.1$ | $\alpha = 1.5$ | $\alpha = 2.0$ | $\alpha = 2.5$ |
|---|---|---|---|---|---|---|
| $\widehat{\theta}$ | 0.5 | 200 | 1.582(0.306) | 1.582(0.317) | 1.557(0.321) | 1.555(0.326) |
| | | 350 | 1.546(0.221) | 1.554(0.228) | 1.557(0.238) | 1.558(0.254) |
| | | 500 | 1.526(0.178) | 1.531(0.186) | 1.531(0.197) | 1.533(0.202) |
| | | 800 | 1.517(0.147) | 1.522(0.153) | 1.524(0.159) | 1.527(0.162) |
| | 1.0 | 200 | 1.633(0.399) | 1.649(0.464) | 1.621(0.504) | 1.653(0.610) |
| | | 350 | 1.573(0.289) | 1.572(0.298) | 1.581(0.330) | 1.600(0.393) |
| | | 500 | 1.548(0.236) | 1.561(0.264) | 1.567(0.282) | 1.562(0.291) |
| | | 800 | 1.525(0.175) | 1.530(0.195) | 1.548(0.212) | 1.542(0.229) |
| $\widehat{\theta}^*$ | 0.5 | 200 | 1.474(0.131) | 1.471(0.140) | 1.471(0.162) | 1.493(0.192) |
| | | 350 | 1.482(0.095) | 1.487(0.102) | 1.486(0.121) | 1.485(0.135) |
| | | 500 | 1.486(0.079) | 1.490(0.085) | 1.497(0.104) | 1.499(0.115) |
| | | 800 | 1.492(0.064) | 1.495(0.071) | 1.496(0.081) | 1.498(0.090) |
| | 1.0 | 200 | 1.498(0.265) | 1.499(0.305) | 1.543(0.397) | 1.518(0.442) |
| | | 350 | 1.508(0.196) | 1.505(0.216) | 1.510(0.256) | 1.523(0.300) |
| | | 500 | 1.499(0.162) | 1.509(0.192) | 1.510(0.207) | 1.513(0.224) |
| | | 800 | 1.496(0.117) | 1.496(0.137) | 1.511(0.161) | 1.501(0.182) |

formance of the estimator $\widehat{b}(t)$:

$$\mathrm{MISE}(\widehat{b}(t)) = Q^{-1} \sum_{q=1}^{Q} \int_0^1 \{\widehat{b}(t)^{[q]} - b(t)\}^2 \, dt,$$

where $\{\widehat{b}(t)^{[q]}, q = 1, \ldots, Q\}$ are estimators of $b(t)$ obtained from the $Q =$ 1000 datasets. $\mathrm{MISE}(\widehat{b}^*)$ is defined analogously. The MISE and associated standard deviation of the estimates $\widehat{b}(t)$ and $\widehat{b}^*(t)$ are displayed in Table 2. For each combination of $\alpha$ and $\sigma$, MISE and the standard deviation decrease as $n$ increases. MISE of $\widehat{b}^*(t)$ is consistently smaller than that of $\widehat{b}(t)$, and the MISE increases with $\sigma$ for given $n$ and $\alpha$. MISE of $\widehat{b}^*(t)$ or

$\widehat{b}(t)$ also shows an increasing trend with $\alpha$ for given $n$ and $\sigma$. Given $\sigma$, the standard derivation of MISE of $\widehat{b}^*(t)$ or $\widehat{b}(t)$ seems stable with $\alpha$ when $n = 800$, but increases with $\alpha$ when $n$ is less than 800. It is interesting that the standard deviation of MISE of $\widehat{b}^*(t)$ is consistently larger than that of $\widehat{b}(t)$.

We also compare the proposed method with the FMR method (Yao, Fu, and Lee, 2011) under the current simulation setting, where the number of groups is 2 for FMR. Because FMR is a nonparametric model, we can only compare the performance of the estimators of $b(t)$. Let $\widehat{b}^{\mathrm{FMR}}(t)$ be the FMR estimator of $b(t)$ proposed by Yao, Fu, and Lee (2011). The MISE and associated standard deviation of $\widehat{b}^{\mathrm{FMR}}(t)$ are displayed in Table 2. For each configuration, MISE and the standard deviation of $\widehat{b}^{\mathrm{FMR}}(t)$ are consistently larger than those of $\widehat{b}(t)$ and $\widehat{b}^*(t)$. This may indicate that the proposed estimators outperform the competitor $\widehat{b}^{\mathrm{FMR}}(t)$.

## 6.2 Testing

We examine the finite-sample performance of the statistic $T_n^*$ given in Section 4, and use the same setting for $r(X)$ and $U$ as in Subsection 6.1 but let $\theta = 1$ and $\alpha = 1.1$. Consider the hypothesis:

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta = c,$$

Table 2: The results for simulation study (estimation). MISE of the estimated slope functions $\widehat{b}(t)$, $\widehat{b}^*$ and $\widehat{b}^{\mathrm{FMR}}$. The corresponding standard deviations are given in the parentheses.

| | $\sigma$ | $n$ | $\alpha = 1.1$ | $\alpha = 1.5$ | $\alpha = 2.0$ | $\alpha = 2.5$ |
|---|---|---|---|---|---|---|
| $\widehat{b}(t)$ | 0.5 | 200 | 0.126(0.078) | 0.103(0.061) | 0.161(0.059) | 0.345(0.074) |
| | | 350 | 0.095(0.058) | 0.080(0.044) | 0.078(0.038) | 0.085(0.042) |
| | | 500 | 0.076(0.045) | 0.067(0.034) | 0.136(0.034) | 0.136(0.027) |
| | | 800 | 0.065(0.035) | 0.060(0.026) | 0.058(0.022) | 0.061(0.021) |
| | 1.0 | 200 | 0.164(0.105) | 0.155(0.099) | 0.221(0.111) | 0.247(0.134) |
| | | 350 | 0.184(0.084) | 0.214(0.098) | 0.177(0.063) | 0.195(0.078) |
| | | 500 | 0.093(0.052) | 0.089(0.050) | 0.099(0.053) | 0.167(0.051) |
| | | 800 | 0.074(0.039) | 0.071(0.033) | 0.078(0.035) | 0.095(0.048) |
| $\widehat{b}^*$ | 0.5 | 200 | 0.115(0.076) | 0.092(0.058) | 0.151(0.053) | 0.337(0.069) |
| | | 350 | 0.088(0.056) | 0.073(0.042) | 0.071(0.035) | 0.077(0.039) |
| | | 500 | 0.072(0.044) | 0.062(0.032) | 0.131(0.031) | 0.132(0.025) |
| | | 800 | 0.062(0.035) | 0.057(0.025) | 0.055(0.020) | 0.058(0.020) |
| | 1.0 | 200 | 0.157(0.103) | 0.145(0.094) | 0.215(0.145) | 0.242(0.135) |
| | | 350 | 0.178(0.084) | 0.166(0.066) | 0.173(0.062) | 0.189(0.074) |
| | | 500 | 0.088(0.049) | 0.084(0.046) | 0.093(0.050) | 0.164(0.049) |
| | | 800 | 0.071(0.038) | 0.068(0.030) | 0.076(0.032) | 0.092(0.048) |
| $\widehat{b}^{\mathrm{FMR}}$ | 0.5 | 200 | 0.479(0.353) | 0.531(0.424) | 0.696(0.706) | 0.907(0.887) |
| | | 350 | 0.359(0.298) | 0.464(0.396) | 0.597(0.563) | 0.760(0.821) |
| | | 500 | 0.299(0.274) | 0.399(0.377) | 0.535(0.550) | 0.734(0.846) |
| | | 800 | 0.241(0.257) | 0.350(0.348) | 0.467(0.530) | 0.608(0.648) |
| | 1.0 | 200 | 1.106(1.097) | 1.369(1.415) | 1.961(2.552) | 2.400(3.091) |
| | | 350 | 0.955(0.873) | 1.220(1.342) | 1.640(2.051) | 2.117(2.728) |
| | | 500 | 0.911(0.883) | 1.068(1.087) | 1.596(2.175) | 2.261(3.768) |
| | | 800 | 0.756(0.737) | 1.024(1.164) | 1.326(1.640) | 1.852(2.726) |

where $c$ ranges from 1 to 1.6 with increment 0.01. To show the effects of estimating $I(\theta)$ by $\widehat{I}(\theta)$, we also proceed $T_n = \{nI(\theta)\}^{1/2}(\widehat{\theta}^* - \theta)$ as if $I(\theta)$ were known, and compare it with $T_n^* = \{n\widehat{I}(\theta)\}^{1/2}(\widehat{\theta}^* - \theta)$. The exact value

$I(\theta)$ was calculated based on (3.14) with $\pi = 0.6, \theta = 1, \sigma = 0.5$ or 1.0 and $u_2$ was calculated as $u_2 = E\left(\int_0^1 Xb\right)^2 = 0.3^2 + 16\sum_{j=2}^{50} j^{-(4+\alpha)}$.

We set 0.05 as the nominal level, and generated 1000 datasets, each consisting of $n = 500$ or 800 random samples to calculate type I errors and power of $T_n$ and $T_n^*$. Figure 1 displays the power against $c$ for four different settings: $(\sigma, n) = (0.5, 500), (1.0, 500), (0.5, 800)$ and $(1.0, 800)$. In each plot, the solid and dashed lines denote the power functions of $T_n$ and $T_n^*$, respectively. These two curves close each other. This indicates good performance of $\widehat{I}(\theta)$ as an estimator of $I(\theta)$, and $T_n^*$ performs well. The type I errors (the power at $c = 1$) for the four settings are displayed in Table 3. They close to the nominal level 0.05. Moreover, we also observe that the empirical size of power increases to 1 as $c$ increases. The results demonstrate that the proposed $T_n^*$ is a useful test.

Table 3: The results for simulation study (testing). Type I error rates of $T_n^*$ and $T_n$ for the four different settings in respect to the nominal level 0.05.

| $(\sigma, n)$ | (0.5, 500) | (1.0, 500) | (0.5, 800) | (1.0, 800) |
|---|---|---|---|---|
| $T_n^*$ | 0.053 | 0.058 | 0.052 | 0.057 |
| $T_n$ | 0.055 | 0.061 | 0.052 | 0.048 |

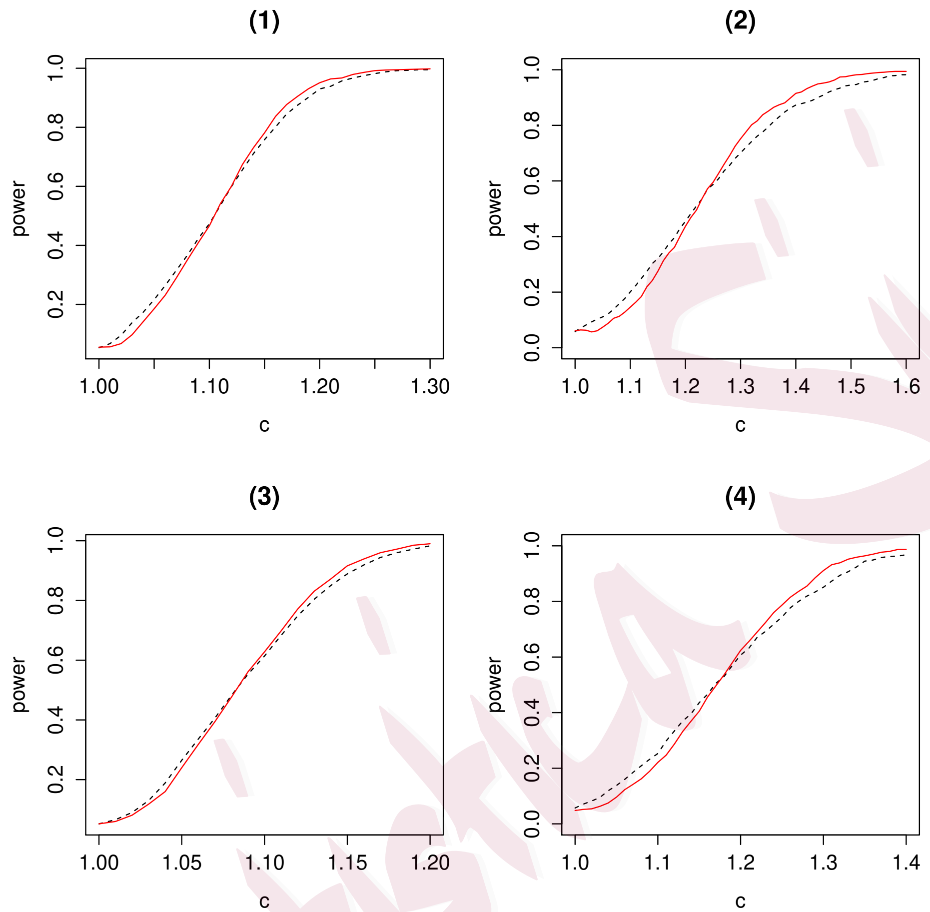


Figure 1: The results for simulation study (testing). The power functions of the test statistic $T_n^*$ (dashed line) and $T_n$ (solid line) for the four different settings (1)-(4) corresponding to $(\sigma, n) = (0.5, 500)$, $(1.0, 500), (0.5, 800)$ and $(1.0, 800)$ for (1)-(4), respectively.

## 7. Application to an AIDS Dataset

Now we illustrate the proposed procedures by analyzing a dataset from an AIDS study. CD4+ cells are targets of HIV and decline after an HIV infection. Thus, when antiviral therapies suppress viral load, CD4+ cell count may recover to a higher level (Lederman et al., 1998). It is believed that the virologic response (measured by viral load) and immunologic response (measured by CD4+ cell count) are negatively correlated during antiviral treatments. However, this relationship may not be constant during the whole period of treatment. In fact, the discordance between virologic and immunologic responses has been observed in several clinical studies (Mellors et al., 1996; Wu, Ding, and DeGruttola, 1998). Motivated by an ACTG study (Lederman et al., 1998), we use model (1.2) and apply the proposed procedures to analyze a data set from this study, in which 53 HIV-1 infected patients were divided into two arms (arms 1 and 2) and were treated with potent antiviral drugs. 361 observations of viral load and CD4+ cell count were obtained on days 0, 2, 7, 10, 14, 21, and 28.

The patterns of CD4 and viral load of the two arms show similarities (See Figure 2) and the combination of these two arms may be beneficial to evaluating the treatment and increasing the power. We apply model (1.2) and the proposed methods to analyze this dataset. $\theta$ reflects how close the effects of the viral load on the CD4 cell count in two arms are. In the initial analysis of this study, the observations from two arms were combined for the preliminary report. We now rigorously evaluate the difference by estimating $\theta$ and further investigate whether such a combination is proper. We average CD4 count over time and divide it by 1000, and treat it as the response variable $Y$, and use viral load as the functional predictor $X(t), t \in \mathcal{I}$, where we take $\mathcal{I} = [0, 29]$.

The smoothing parameter $m_n = 2$ was obtained by leave-one-out cross-validation. The estimated $\widehat{\theta}^* = 0.9591$. The estimated slope function and associated pointwise confidence interval are depicted in the left panel of Figure 3. The pattern shows that the CD4+ cell count increases as viral load decreases in the primary treatment period. This negative relationship lasts until day 15, and then changes to a slight positive trend. The right panel of Figure 3 plots the normal Q-Q plot of residuals, and suggests a reasonable fit of the data. We further consider whether the two treatment arms are significantly different; i.e., test $H_0 : \theta = 1$. The statistic $|T_n^*| = 0.1392 < 1.96$. This indicates that the difference of two treatments between two arms may be insignificant.

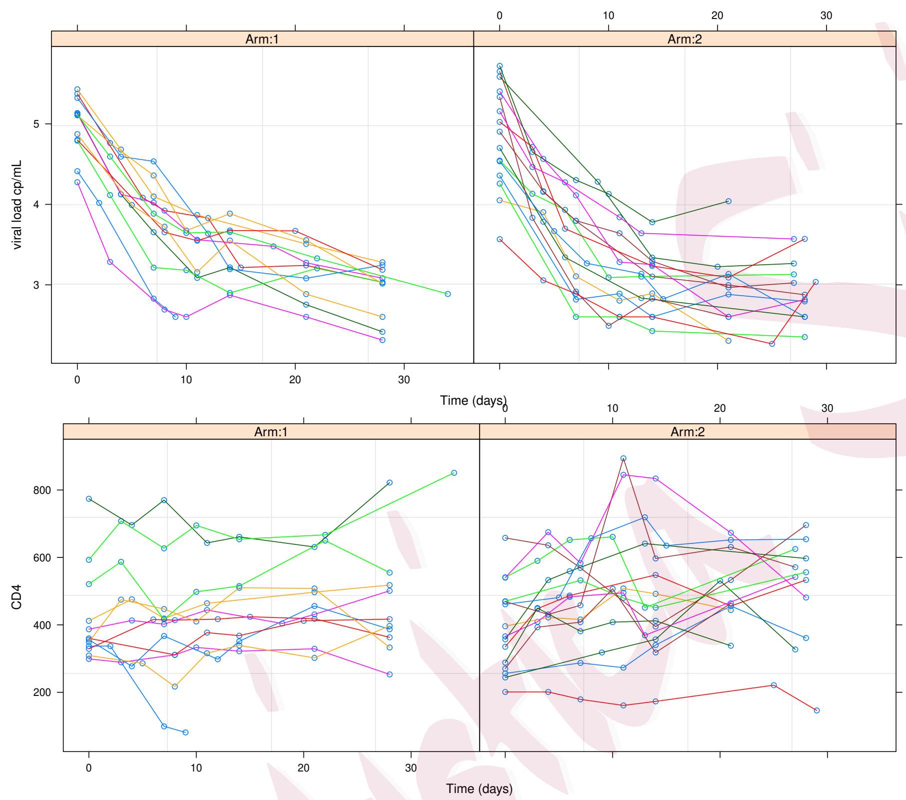


Figure 2: The scatter plots of viral load (upper panel) and CD4 cell count (lower panel) against treatment times for two arms.

## 8. Discussions

In this paper, we have studied two-sample functional linear models that combine two functional curves with similar patterns, and have developed estimation and testing procedures. Briefly, the proposed methods have

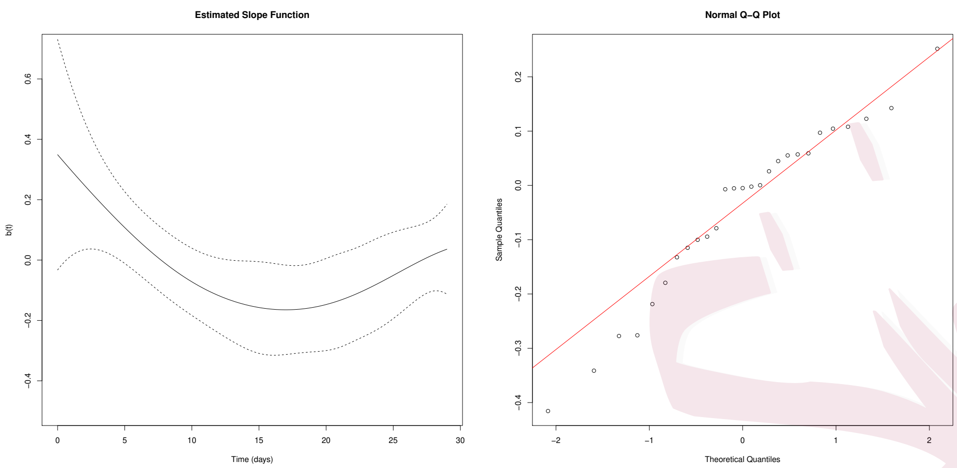


Figure 3: The results for the CD4 dataset. The estimated slope function $\widehat{b}(t)$ and 95% bootstrap pointwise confidence interval (left panel), and the Q-Q plot of the residuals (right panel).

the properties: (i) the estimators of the scalar parameter are asymptotically normal, and the estimators of the nonparametric functions have the optimal rates of convergence; (ii) the proposed methods show promising performance in finite sample situations; and (iii) the implemented algorithm is computationally efficient.

There are interesting possible extensions of two-sample FLM. Generally, our model implies that the two curves differ from each other by a constant $\theta$, which may not be true; i.e., $\theta$ could also be a function of time. It would be interesting to consider estimation of this function and $a$ and $b(t)$, find limiting distributions and discuss efficiency accordingly. However, there are considerable issues with identifiability and efficiency for estimation of $\theta(t)$.

In this article, we have focused on modeling with the linear relationship between $r(X)$ and $X(t)$. It is of interest to extend the methods to a nonparametric or semiparametric relationship. However, the theory and implementation of such an extension is much more complicated and warrants further studies. Semiparametric asymptotic efficiency for estimating $\theta$ when $\varepsilon$ is unknown is a much more complex problem, both technically and practically.

## Supplementary Materials

The Supplementary Materials present a simulation example continued from Section 6, the proofs of Proposition 1 and Theorems 1 to 4, the derivation of the efficient score given in (3.13).

### Acknowledgements

The authors thank the co-editors and three referees for their constructive suggestions and comments that have substantially improved an earlier version of this paper. Zhang's research was partially supported by National Natural Science Foundation of China Grants (11571112, 11701360), Doctoral Fund of Ministry of Education of China (20130076110004), Program of Shanghai Subject Chief Scientist (14XD1401600) and the 111 Project (B14019). Liang's research was partially supported by NSF grant DMS-1620898, and by Award Number 11529101, made by National Natural Science Foundation of China.

## References

Bickel, P. J., Klaasen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

Brown, B. W. and Newey, W. K. (1998). Efficient semiparametric estimation of expectations. *Econometrica* **66**, 453-464.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159-2179.

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

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13**, 571-591.

Chen, D., Hall, P. and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.* **39**, 1720-1747.

Chen, K., Zhang, X., Petersen, A. and Müller, H.-G. (2017). Quantifying infinite-dimensional data: Functional data analysis in action. *Statistics in Biosciences* **9**, 582-604.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plann. Inference* **147**, 1-23.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *J. Classif.* **5**, 249-282.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, New York.

Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *J. Roy. Statist. Soc. Ser. B* **78**, 637-653.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493-1517.

He, G., Müller, H.-G. and Wang, J.-L. (2000). Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability* (Edited by M. Puri), 197-210. VSP International Science Publishers, The Netherlands.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester.

Imaizumi, M. and Kato, K. (2018). PCA-based estimation for functional linear regression with functional responses. *J. Multivariate Anal.* **163**, 15-36.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587-602.

Lederman, M., Connick, E., Landay, A., Kuritzkes, D., Spritzkes, J., St Clair, M., Kotzin, B., Fox, L., Chiozzi, M., Leonard, J., Rousseau, F., Wade, M., Roe, J., Martinez, A. and Kessler, H. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS clinical trials group protocol 315. *J. Infect. Dis.* **178**, 70-79.

Lei, J. (2014). Adaptive global testing for functional linear models. *J. Amer. Statist. Assoc.* **109**, 624-634.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.* **98**, 1782-1804.

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

Li, Y., Wang, N. and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108**, 1284-1294.

Mellors, J., Rinaldo, C., Gupta, R. M., White, P., Todd, J. A. and Kingsley, L. A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272**, 1167-1170.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scand. J. Statist.* **32**, 223-240.

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

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486-1521.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768-1802.

Shang, Z. and Cheng, G. (2015). Nonparametric inference in generalized functional linear models. *Ann. Statist.* **43**, 1742-1773.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3**, 257-295.

Wu, H., Ding, A. and DeGruttola, V. (1998). Estimation of HIV dynamic parameters. *Statist. Medicine* **17**, 2463-2485.

Yao, F., Fu, Y. and Lee, T. C. (2011). Functional mixture regression. *Biostatistics* **12**, 341-353.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577-590.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-2903.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412-3444.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44**, 2281-2321.

School of Statistics, East China Normal University, Shanghai, 200241, China

E-mail: (wcx_stat@126.com)

School of Statistics, East China Normal University, Shanghai, 200241, China

E-mail: (zhangriquan@163.com)

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: (hliang@gwu.edu)