Adaptive Global Testing for Functional Linear Models

Jing Lei

Carnegie Mellon University

August 5, 2013

Author's Footnote:

Jing Lei is Assistant Professor, Department of Statistics, Carnegie Mellon University. Mailing address: 132 Baker Hall, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (email: jinglei@andrew.cmu.edu). This work is supported by National Science Foundation Grant BCS-0941518.

Abstract

This paper studies global testing of the slope function in functional linear regression models. A major challenge in functional global testing is to choose the dimension of projection when approximating the functional regression model by a finite dimensional multivariate linear regression model. We develop a new method that simultaneously tests the slope vectors in a sequence of functional principal components regression models. The sequence of models being tested is determined by the sample size and is an integral part of the testing procedure. Our theoretical analysis shows that the proposed method is uniformly powerful over a class of smooth alternatives when the signal to noise ratio exceeds the detection boundary. The methods and results reflect the deep connection between the functional linear regression model and the Gaussian sequence model. We also present an extensive simulation study and a real data example to illustrate the finite sample performance of our method.

Keywords: Functional linear regression, detection boundary, adaptive testing, functional principal components

1. INTRODUCTION

In the functional linear regression model, the data consists of i.i.d pairs $(Y_i, X_i)_{i=1}^n$ satisfying

$$Y_i = b_0 + \langle X_i, \theta \rangle + \sigma Z_i, \quad i = 1, ..., n, \tag{1}$$

where X_i is a random sample from a zero-mean stochastic process indexed on $[0,1]$ with sample paths in $L_2[0,1]$; $\theta \in L_2[0,1]$ is the slope function; and Z_i's are independent standard Gaussian noise. Here $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $L_2[0,1]$: $\langle f, g \rangle = \int fg$.

The functional linear model is an important component of functional data analysis and there has been a vast literature on estimation and prediction for functional linear models. Methods of estimating the slope function are studied in, among others, Cardot, Ferraty & Sarda (2003), Yao, Müller & Wang (2005), Crambes, Kneip & Sarda (2009), Cardot & Johannes (2010). Minimax rates of estimation are established by Hall & Horowitz (2007), Cai & Zhou (2008), Meister (2011), using functional principal components regression, and by Yuan & Cai (2010) using a Reproducing Kernel Hilbert Space approach. The problem of prediction is studied by Cai & Hall (2006).

The focus of this paper is hypothesis testing for the functional linear regression model (1). Specifically, we are interested in testing the presence of a global linear effect of X on Y:

$$H_0 : \theta = 0 \qquad \text{against} \qquad H_a : \theta \neq 0.$$

Despite the advances in estimation and prediction, there has been relatively less work on hypothesis testing for functional linear models. Cardot, Ferraty, Mas & Sarda (2003) proposed a method that reduces the problem to testing the linear effect of the first m coordinates in a basis expansion (for example, functional principal components) of X and θ. González-Manteiga & Martínez-Calvo (2011) developed a bootstrap approach to construct pointwise confidence intervals for the slope function $\theta(t)$. Other work on related topics include testing linear effects of scalar covariates on functional response (Shen & Faraway (2004), Zhang & Chen (2007)), nonparametric regression effects of scalar covariates on functional responeses (Cardot, Prchal & Sarda (2007)), and functional two-sample tests (Hall & van Keilegom (2007), Zhang & Chen (2007)).

In this paper, we develop a new approach to test functional global linear effects that parallels the theory and methodology for minimax estimation in functional linear models (Hall & Horowitz (2007), Cai & Zhou (2008)). Our method follows the functional principal components regression approach, but simultaneously tests the slope vectors in an increasing sequence of finite dimensional regression models. A simple version of this procedure can be described as follows. Given two positive integers $k_{n,\min} \leq k_{n,\max}$, we test the functional principal components regression models corresponding to the first $m_k = 2^k$ principal components, for all $k_{n,\min} \leq k \leq k_{n,\max}$. Our theory suggests that $k_{n,\min}$ grows slowly as n (say, for example, $\log\log n$), and $k_{n,\max}$ grows at the rate of $\log n$. Such a choice ensures that the search range is large enough so that a wide collection of smooth alternatives can be successfully detected. On the other hand, the number of simultaneous tests grows slowly with n so that the power is not severely reduced by multiple testing.

In our method, choosing the number of principal components is an integral part of the testing procedure, while many existing methods need to first specify a dimension of projection and then perform a finite dimensional test. Under suitable smoothness conditions detailed in Section 3, our testing procedure can be justified through the following detection boundary framework.

P1. If the signal to noise ratio, $\|\theta\|_2^2/\sigma^2$, exceeds the rate $(r_n^*)^2$, then our test statistic can consistently separate the null and all smooth alternatives, with

$$r_n^* = \left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{2\beta}{4\beta+2\gamma+1}},$$

where, roughly speaking, γ corresponds to the decay rate of the eigenvalues of the covariance operator of X in the sense that the jth eigenvalue is of order $j^{-\gamma}$; and β corresponds to the smoothness of slope function θ, in the sense that its jth coefficient in an orthogonal basis expansion is bounded by $j^{-\beta-1/2}$. Here $\|\theta\|_2^2 = \int_0^1 \theta^2(t)dt$ is the usual squared L_2 norm on $L_2[0,1]$.

P2. If the signal to noise ratio, $\|\theta\|_2^2/\sigma^2$, decays faster than the rate $(r_n^*)^2$, then no test can consistently separate the null and all possible alternatives.

A detailed definition of these terms, the exact meaning of β and γ, as well as exact conditions and rigorous statements, are given in Section 3.

Property **P1** means that the test procedure is consistent whenever the signal to noise ratio exceeds a certain threshold, specified by $(r_n^*)^2$. This is a uniform power guarantee over a class of alternatives, for which only smoothness is assumed. Property **P2** indicates that such a critical rate cannot be improved. Thus our method is indeed asymptotically minimax consistent. Such a framework of high dimensional and nonparametric testing has been considered in the literature by Ingster (1982), Spokoiny (1996), Donoho & Jin (2004), Ingster, Tsybakov & Verzelen (2010), Ingster, Sapatinas & Suslina (2012), Arias-Castro, Candès & Plan (2011).

The critical separation rate r_n^* given above is slightly smaller than the well-known corresponding minimax error rate of estimation, which is $n^{-\beta/(2\beta+\gamma+1)}$ (see Hall & Horowitz (2007), Cai & Zhou (2008), and also Meister (2011)). In other words, the testing problem is more subtle than estimation since it is still possible to detect the effect even when the signal to noise ratio is below the optimal estimation error rate. This is because testing only tries to tell the presence of a linear effect rather than to recover it. Similar phenomenon has been observed in other high dimensional or nonparametric inference problems such as Ingster et al. (2012), Donoho & Jin (2004).

The $\log\log n$ factor in the critical rate r_n^* is the price for adaptivity. In fact, if we know that the covariate functions and the alternative hypothesis belong to specific classes of smooth functions (that is, we know the values of β and γ), then one can find an optimal number of principal components for the test without searching for it, and therefore avoid the $\log\log n$ term. In the case of unknown smoothness, we need an extra $\log\log n$ factor to search over a range of dimensions of projection on the leading principal components. This reflects another difference between testing and estimation in functional linear models: For estimation it is possible to obtain adaptive estimators that achieve the same non-adaptive minimax optimal rate (Cai & Zhou (2008)).

The minimaxity of our test is illustrated through an extensive simulation study. In comparison with methods using a single dimension of projection, our method always performs at least as well as the competitors. A single dimension of projection may be suitable for some alternatives, but there always exist other alternatives for which it performs poorly. Because when the selected dimension is too high the test will be too noisy, and when the dimension is too low, it may miss the signal. The minimaxity of our method ensures that it always has good power against smooth alternatives in the detectable region. In our real data example, we consider a situation where two different projection dimensions lead to different results. Our method can be used to support the functional principal component regression approach and reconcile the test results obtained from different finite dimensional tests.

The methods and results in this paper are built on top of the deep connection between the functional linear regression model and the Gaussian white noise model. Let Γ be the covariance operator of X. It is shown by Meister (2011) that if Γ and σ^2 are known, then (1) is asymptotically equivalent to the Gaussian white noise model in Le Cam's sense (Le Cam (1986), Le Cam & Yang (2000))

$$dY(t) = \left[\Gamma^{1/2}\theta\right](t)dt + n^{-1/2}\sigma dW(t), \tag{2}$$

where $(Y(t) : 0 \leq t \leq 1)$ is an Itō process; $W(t)$ is a standard Wiener process; and the differentiation shall be interpreted in terms of the Itō calculus. Moreover, one can project both sides of (2) on the eigenfunctions of Γ, obtaining the following equivalent Gaussian sequence model.

$$\eta_j = \theta_j + (n\kappa_j)^{-1/2}\sigma Z_j, \quad j \geq 1, \tag{3}$$

where $\eta_j = \kappa_j^{-1/2}\int_0^1 \phi_j(t)dY(t)$, $\theta_j = \int_0^1 \phi_j(t)\theta(t)dt$, $Z_j = \int_0^1 \phi_j(t)dW(t)$ (i.i.d standard normal), and $(\kappa_j, \phi_j)_{j\geq 1}$ are the eigenvalues and eigenfunctions of Γ (see Section 2.1). The asymptotic equivalence theory suggests that, when Γ and σ^2 are known, the asymptotic error rate of an inference procedure in models (2) and (3) can usually be carried over to a corresponding inference problem in model (1). In the functional linear regression model Γ and σ^2 often need to be estimated from the data. One of the main efforts in this paper is to show that, under standard regularity conditions, when the unknown quantities are substituted by their empirical estimates, the test developed for the Gaussian sequence model with known covariance remains consistent in the detectable region as defined in Section 3.2.

In Section 2 we present some preliminaries including functional principal components regression and describe our test procedure. In Section 3 we derive the asymptotic properties of our procedure under a detection boundary framework. The finite sample behavior of the proposed method is presented in Section 4 through both simulation and real data examples. Section 5 gives some concluding remarks. Technical proofs are postponed to the Appendix. Some lengthy proofs and additional simulation results are included in the Supplementary Material.

2. PROBLEM FORMULATION AND METHODOLOGY

Let $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ be the $n \times 1$ vector of the response variable, the collection of observed covariate functions, and the vector of unobserved additive noise, respectively. For presentation simplicity assume that $\mathbb{E}Y = b_0 = 0$ (see Remark 2.2 below). We also assume that $\mathbb{P}_X$, the marginal distribution of X, has zero measure on any finite dimensional subspaces.

The methodological and theoretical development of this paper will also be based on the assumption that the X_i curves are fully observed without noise. In practice X_i's are usually observed at discrete locations with noises. A standard approach is to first estimate each X_i using smoothing techniques (kernel, spline, or local polynomial). Let N be the number of observations on each curve, and h be the smoothing bandwidth. It can be shown (Hall, Müller & Wang 2006; Zhang & Chen 2007) that when the observation is dense enough ($Nhn^{-\delta_1} \to \infty$, $N^{1-\delta_2}h \to \infty$, with $h = O(n^{-1/4})$ and positive constants δ_1, δ_2), the resulting estimators of individual curves and covariance operator are $\sqrt{n}$-consistent under standard smoothness conditions on X. In other words, they are as good as using the true curves X_i for estimating the covariance operator. The proofs of this paper will go through for such pre-smoothed densely observed data because we only require the covariance operator to be estimated with $O(1/\sqrt{n})$ accuracy.

2.1 Functional principal components regression

We study functional linear regression using the functional principal components approach and establish its connection with the Gaussian sequence model. The following general discussion can be found in the literature (see Hall & Horowitz (2007), Meister (2011), for example).

For any $s, t \in [0,1]$, let $\Gamma(s,t) = \mathrm{Cov}(X(s), X(t))$ be the covariance of $X(s)$ and $X(t)$. Then Γ defines a symmetric function $[0,1]^2 \mapsto \mathbb{R}$. Let $\widehat{\Gamma}(s,t) = n^{-1}\sum_{i=1}^n X_i(s)X_i(t)$ be the sample covariance. Γ and $\widehat{\Gamma}$ can be written in the eigen-decomposition (also known as the Karhunen-Loève expansion)

$$\Gamma(s,t) = \sum_{j=1}^{\infty} \kappa_j \phi_j(s)\phi_j(t), \qquad \widehat{\Gamma}(s,t) = \sum_{j=1}^{\infty} \widehat{\kappa}_j \widehat{\phi}_j(s)\widehat{\phi}_j(t), \tag{4}$$

where the non-increasing sequences $(\kappa_j : j \geq 1)$ and $(\widehat{\kappa}_j : j \geq 1)$ are the population and sample eigenvalues and $(\phi_j : j \geq 1)$, $(\widehat{\phi}_j : j \geq 1)$ are the corresponding eigenfunctions, each forming an orthonormal basis of $L_2[0,1]$. By the linear independence of X_i's, we have $\widehat{\kappa}_n > \widehat{\kappa}_{n+1} = 0$. The covariance functions Γ, $\widehat{\Gamma}$ can be viewed as linear operators from $L_2[0,1]$ to $L_2[0,1]$ as follows.

$$(\Gamma f)(t) = \int_0^1 \Gamma(s,t) f(s) ds = \sum_{j=1}^{\infty} \kappa_j \langle f, \phi_j \rangle \phi_j(t), \quad \forall\ f \in L_2[0,1]. \tag{5}$$

Using the idea of principal components regression, we first represent the regression coefficient function θ in these two bases:

$$\theta = \sum_{j \geq 1} \theta_j \phi_j = \sum_{j \geq 1} \widehat{\theta}_j^* \widehat{\phi}_j, \quad \text{where} \ \ \theta_j = \langle \phi_j, \theta \rangle, \ \widehat{\theta}_j^* = \langle \widehat{\phi}_j, \theta \rangle.$$

In view of equation (5) and the fact that $\mathbb{E} YX(t) = (\Gamma\theta)(t)$, we have the following estimate of θ.

$$\widehat{\theta} = \sum_{j=1}^{n} \widehat{\theta}_j \widehat{\phi}_j \, , \tag{6}$$

where

$$\widehat{\theta}_j = \widehat{\kappa}_j^{-1} \left\langle n^{-1} \sum_{i=1}^{n} Y_i X_i, \widehat{\phi}_j \right\rangle .$$

The next lemma relates $\widehat{\theta}_j$ to $\widehat{\theta}_j^*$, which is useful for the development of our test and theoretical analysis. It is proved in Appendix A.

Lemma 2.1. *Let $\widehat{\theta}$ be defined as in (6), then*

$$\widehat{\theta}_j = \langle \widehat{\phi}_j, \widehat{\theta} \rangle = \widehat{\theta}_j^* + \frac{\sigma}{\sqrt{n\widehat{\kappa}_j}} Z_j^*, \quad j = 1, ..., n, \tag{7}$$

where $(Z_j^* : 1 \leq j \leq n)$ is a sequence of independent standard Gaussian random variables.

Lemma 2.1 provides a starting point for the development of our methods. It relates functional linear regression to the Gaussian sequence model in (3) with obvious correspondence. The Gaussian sequence model (3) can be viewed as a population version of (7). This is clearly an ill-posed inverse problem because $(n\kappa_j)^{-1/2} \to \infty$, as $j \to \infty$. Minimax testing problem for model (3) has been studied by Ingster et al. (2012). Inspired by the result in Meister (2011), our strategy is to make use of such a similarity between (7) and (3), showing that the tests developed for the latter can be used to solve the former.

Remark 2.2. *When Y (and possibly X) is not centered, one can re-center the data by removing the sample mean from each data point. In this case, Lemma 2.1 holds in exactly the same manner. Moreover, it is known (Hall et al. 2006) that the estimated covariance operator is still $\sqrt{n}$-consistent. As a result, the estimation error induced in re-centering does not affect the methods and results presented below.*

2.2 The Exponential Scan Test

Under the null hypothesis $\theta = 0$ and hence $\widehat{\theta}_j^* = \langle \theta, \widehat{\phi}_j \rangle = 0$ for all j. Eq. (7) suggests that $\sqrt{n\widehat{\kappa}_j}\sigma^{-1}\widehat{\theta}_j = Z_j^*$ are independent standard Gaussian random variables. For all $1 \le m \le n$, define

$$S_{n,m} = \frac{n}{\widehat{\sigma}^2} \sum_{j=1}^{m} \widehat{\kappa}_j \widehat{\theta}_j^2, \quad \text{and} \quad T_{n,m} = \frac{1}{\sqrt{2m}}(S_{n,m} - m), \tag{8}$$

where $\widehat{\sigma}^2$ is an estimate of σ^2 and will be discussed later. If $\widehat{\sigma}^2$ is an accurate estimate of σ^2, then the statistic $S_{n,m}$ has approximately a χ^2 distribution with m degrees of freedom under the null hypothesis, while $T_{n,m}$, a centered and scaled version of $S_{n,m}$, converges weakly to a standard normal when m and n are large (see Cardot, Ferraty, Mas & Sarda (2003), Ingster et al. (2012)).

For a fixed value of m, one can easily derive a level α test using $S_{n,m}$ or $T_{n,m}$. However, each m leads to a different test, whose power depends on the particular simple alternative hypothesis. We propose to scan over different values of m so that the test can detect a wide range of alternative hypotheses. Specifically, let $m_0 = m_0(n)$ be an integer depending on n such that $m_0(n)/\log n \to 0$ and $m_0(n) \to \infty$. For example, one can choose $m_0(n) = \lfloor \sqrt{\log n} \rfloor$. Then we define $m_k = m_0 2^k$ for $k = 0, 1, ..., k_{n,\max} := \lceil \log_2(n^{1/3}/m_0) \rceil$. The Exponential Scan Test is given by

$$\text{Reject } H_0, \quad \text{if} \quad \psi_{\text{ES}}(\mathbf{Y}, \mathbf{X}) = 1,$$

where

$$\psi_{\text{ES}}(\mathbf{Y}, \mathbf{X}) = \mathbb{1}\left\{\exists 0 \le k \le k_{n,\max}, \text{ s.t. } T_{n,m_k} \ge b(m_k)\right\}, \tag{9}$$

where $T_{n,m}$ is defined as in (8), and $b(m)$ is a function to be determined by the user. We call it the Exponential Scan Test since it searches sub-models with exponentially increasing number of principal components. It is inspired by similar methods developed for the Gaussian sequence model (Ingster et al. (2012)). To apply this procedure, we need to specify two components: (1) the function $b(m)$, and (2) the estimator $\widehat{\sigma}^2$. Here we give some brief comments on their choices for practical concerns. Some theoretical discussions on $b(m)$ and $\widehat{\sigma}^2$ are given in Section 3.

Choosing the threshold $b(m)$ for a specific level α. Suppose we want to construct a level α test for a given $\alpha \in (0, 1)$. One choice of $b(m)$ can be given by Bonferroni correction:

$$b(m) = \frac{1}{\sqrt{2m}} \left[t(\alpha/(k_{n,\max} + 1), m) - m \right], \tag{10}$$

where $t(a, m)$ is the upper a-quantile of a χ^2 random variable with m degrees of freedom. More generally, one can use $b(m_k) = [t(\alpha_k, m_k) - m_k] / \sqrt{2m}$, with $\sum_{k=0}^{k_{n,\max}} \alpha_k = \alpha$.

Estimating the noise variance. A consistent estimator of σ^2 can be obtained by the residual mean square in the linear regression of Y on the first m_n estimated principal components, provided that m_n grows slowly to infinity as n increases (Cai & Zhou (2008)). In our implementation, $m_n = \lfloor \sqrt{n} \rfloor$ gives reasonable estimates for small and moderate values of n. For theoretical concerns, we actually need $\widehat{\sigma}^2$ to be accurate enough with a specific rate of convergence. Further discussion is given in Section 3.2.

3. THEORETICAL PROPERTIES

In this section we discuss the asymptotic properties of our method in a detection boundary framework. Specifically, we rigorously state and prove properties **P1** and **P2** listed in Section 1. In the following discussion, all limits are considered as $n \to \infty$. For two positive sequences (a_n) and (b_n), $a_n = o(b_n)$ means $a_n/b_n \to 0$, and $a_n = \omega(b_n)$ means $a_n/b_n \to \infty$. The "big O" notation is defined as usual: $a_n = O(b_n)$ means $\limsup a_n/b_n < \infty$. Also, $a_n \asymp b_n$ means $c_1 \le \liminf a_n/b_n \le \limsup a_n/b_n \le c_2$ for some positive constants c_1, c_2. Unless otherwise noted, the notation $\sum_j$ means summing over all positive integers j.

3.1 The Detection Boundary Framework and Critical Separation Rates

The function space $L_2[0, 1]$ is infinite dimensional. Therefore, inferences for functional linear models are typically carried out under some smoothness conditions:

$$c_1 j^{-\gamma} \le \kappa_j \le c_2 j^{-\gamma},\ \forall\ j \ge 1; \quad \theta \in \Theta(\beta, L) := \left\{ \theta : \sum_{j \ge 1} j^{2\beta} \theta_j^2 \le L^2 \sigma^2 \right\}, \tag{11}$$

for some positive constants c_1, c_2, γ, β, and L. These conditions imply the smoothness of X and θ respectively, indicating that the higher order terms (θ_j and κ_j with large values of j) in (7) and (3) can be safely ignored. Similar conditions are considered in estimating θ, such as Cavalier & Tsybakov (2002) for the white noise and Gaussian sequence models, and Hall & Horowitz (2007), Meister (2011) for functional linear regression. It is also considered in hypothesis testing for white noise and Gaussian sequence models by Ingster et al. (2012).

For any test $\psi = \psi(\mathbf{Y}, \mathbf{X}) : (\mathbb{R} \otimes L_2[0,1])^n \mapsto [0,1]$, we define its type I error in the usual sense:

$$\alpha_n(\psi) = \mathbb{E}_{\theta=0}\psi(\mathbf{Y}, \mathbf{X}),$$

and the type II error at $\theta \neq 0$:

$$\lambda_n(\psi, \theta) = \mathbb{E}_\theta(1 - \psi(\mathbf{Y}, \mathbf{X})).$$

Let Θ be a class of alternatives, define the worst case type II error as

$$\lambda_n(\psi, \Theta) = \sup_{\theta \in \Theta} \lambda_n(\psi, \theta).$$

A conservative goal in designing a test is to control the worst case total error $\alpha_n(\psi) + \lambda_n(\psi, \Theta)$. This corresponds to the minimax criterion. However, for any test at a given level α, we can always find a $\theta \in \Theta(\beta, L)\backslash\{0\}$ close enough to zero such that the test has power only slightly larger than α. To exclude this case, we consider a modified testing problem that provides some separation between the null and alternative. For any $r > 0$, define

$$\Theta(\beta, L, r) = \{\theta : \theta \in \Theta(\beta, L), \|\theta\|_2^2 \geq r^2\sigma^2\}.$$

Given $r_n > 0$, we consider the following testing problem:

$$H_0 : \theta = 0 \quad \text{against} \quad H_a : \theta \in \Theta(\beta, L, r_n). \tag{12}$$

Then it is natural to ask how the worst case total error changes with r_n. Following this idea, the critical separation rate for the functional linear model is defined as follows.

Definition 3.1 (Critical Separation Rate)**.** *For testing problem* (12) *under model (1) and condition (11), a sequence $r_n^* > 0$ is called the critical separation rate if the following holds.*

1. *If $r_n = o(r_n^*)$, then there exists a distribution P on (Y, X) such that $\alpha_n(\psi) + \lambda_n(\psi, \Theta(\beta, L, r_n)) \to 1$ as $n \to \infty$ for any ψ;*

2. *If $r_n = \omega(r_n^*)$, then there exists a test ψ such that $\alpha_n(\psi) + \lambda_n(\psi, \Theta(\beta, L, r_n)) \to 0$ as $n \to \infty$.*

In the next subsection we shall see that the critical separation rate for the functional linear global testing problem is

$$r_n^* = \left(\frac{\sqrt{\log\log n}}{n}\right)^{\frac{2\beta}{4\beta+2\gamma+1}}, \tag{13}$$

and ψ_{ES} is a consistent test whenever the signal to noise ratio exceeds this rate.

Remark: The sequence of critical rate r_n^*, if it exists, is not unique. For example, if r_n^* is such a sequence, then any sequence $r_n' \asymp r_n^*$ also satisfies the definition. Therefore, the definition ignores constants and focuses on rates of convergence/growth. When $r_n \asymp r_n^*$, the so-called "sharp asymptotic result" (Meister (2011), Ingster et al. (2012)) suggests that the worst case total error stays at a constant level and bounded away from both zero and one. Such a framework (also known as the detection boundary problem) has been studied in high-dimensional and nonparametric testing problems such as Ingster (1982), Donoho & Jin (2004), Ingster et al. (2010), Ingster et al. (2012) Arias-Castro et al. (2011).

3.2 Asymptotic Results

In this section we present our main consistency results. The idea and argument resembles that for the Gaussian sequence model given in Ingster et al. (2012), but requires a careful control of the estimation error in $\widehat{\theta}_j$, $\widehat{\Gamma}$, and $\widehat{\sigma}^2$. Formally, we need the following conditions with some positive constants c_3 and C.

$$\int \mathbb{E}X(t)^4 dt < \infty; \qquad \mathbb{E}\langle X, \phi_j\rangle^4 \le C\kappa_j^2, \ \ \forall\, j \ge 1. \tag{14}$$

$$\kappa_j - \kappa_{j+1} \ge c_3 j^{-\gamma-1}, \ \ \forall\, j \ge 1; \qquad 1 < \gamma < \beta - 3/2. \tag{15}$$

Equation (14) ensures that X has light tails so that the empirical covariance operator has $\sqrt{n}$ consistency. The first condition in (15) requires a gap between the eigenvalues of Γ. It ensures the accuracy of the estimated eigenfunctions $\widehat{\phi}_j$. The second condition in (15) requires that the slope function θ is smoother than the covariate function X, so that the errors in $\widehat{\theta}_j^* = \langle \widehat{\phi}_j, \theta\rangle$ do not accumulate. These conditions are standard ones used in the minimax estimation and prediction for functional linear models (Meister (2011), Hall & Horowitz (2007), Cai & Hall (2006)).

The following theorem, proved in Appendix A.2, says that under the above smoothness and eigen-gap conditions, the Exponential Scan Test is uniformly consistent over $\Theta(\beta, L, r_n)$ whenever r_n exceeds the rate $(\sqrt{\log\log n}/n)^{2\beta/(4\beta+2\gamma+1)}$.

Theorem 3.2. *Consider testing problem* (12) *under model* (1) *and conditions* (11)*, (14), and (15). For ψ_{ES} given in (9) with $b(m) = 4\sqrt{\log\log m}$ and $\widehat{\sigma}^2$ such that $\widehat{\sigma}^2 - \sigma^2 = o_P(n^{-1/5})$ uniformly* *over* $\Theta(\beta, L)$*, we have,*

$$\lim_{n\to\infty} \alpha_n(\psi_{\mathrm{ES}}) + \lambda_n(\psi_{\mathrm{ES}}, \Theta(\beta, L, r_n)) = 0.$$

whenever $r_n = \omega(r_n^*)$ *where* r_n^* *is defined in* (13)

In our proof of Theorem 3.2, it is clearly indicated that the $\sqrt{\log\log n}$ factor in the critical separation rate comes from the need to search for an optimal dimension of projection onto principal components. If β is known, one can construct test with a projection dimension determined by the sample size and β and achieve the non-adaptive critical rate $n^{-2\beta/(4\beta+2\gamma+1)}$.

In order to apply Theorem 3.2, we need an estimator $\widehat{\sigma}^2$ with convergence rate faster than $n^{-1/5}$ uniformly over all smooth θ. Now we show that such an estimator exists under the condition of the theorem. From Eq. (7), we see that for large values of j, $\sqrt{n\widehat{\kappa}_j}\widehat{\theta}_j \approx \sigma Z_j^*$, because $\sqrt{n\widehat{\kappa}_j}\widehat{\theta}_j^*$ goes to 0 as j becomes large. Note that Z_j^*'s are independent standard Gaussian, so we can estimate $\widehat{\sigma}^2$ using $\widehat{\theta}_j$'s with large values of j. In particular, we consider

$$\widehat{\sigma}^2 = \frac{2}{n} \sum_{\frac{n}{2}<j\le n} n\widehat{\kappa}_j\widehat{\theta}_j^2. \tag{16}$$

The following lemma indicates that $\widehat{\sigma}^2$ given in (16) is a qualified estimator to apply Theorem 3.2. The proof is elementary and omitted.

Lemma 3.3. *Let* $\widehat{\sigma}^2$ *be given in (16), then under conditions (11), (14), and (15), we have, uniformly over* $\Theta(\beta, L)$*,*

$$\widehat{\sigma}^2 - \sigma^2 = o_P(n^{-1/5}).$$

Remark. In practice it is more convenient to estimate $\widehat{\sigma}^2$ by residual mean square of principal components regression with m_n dimensions, where m_n can be chosen to grow slowly with n (for example, $\lfloor\sqrt{n}\rfloor$). We found it highly non-trivial to derive a rigorous rate of convergence of such a simple estimator without making stronger assumptions, although it gives very good empirical performance.

Next we state a lower bound result, saying that the rate in Theorem 3.2 cannot be improved and is the critical separation rate for the testing problem. It is a consequence of the lower bound results for the Gaussian sequence model established in Ingster et al. (2012), and the asymptotic equivalence between the Gaussian sequence model and the functional linear model (Meister (2011)).

Theorem 3.4. *Consider testing problem* (12) *under model* (1) *and condition (11). All tests* $\psi(\mathbf{Y},\mathbf{X}) : (\mathbb{R} \otimes L_2[0,1])^n \mapsto [0,1]$ *satisfy*

$$\lim_{n\to\infty} \alpha_n(\psi) + \lambda_n(\psi, \Theta(\beta, L, r_n)) = 1,$$

whenever $r_n = o(r_n^*)$ *with* r_n^* *defined in* (13).

We discuss the proof of Theorem 3.4 in Appendix A.3. We also give a direct argument for a slightly weaker result ($r_n = o(n^{-2\beta/(4\beta+2\gamma+1)})$) by explicitly constructing a least favorable alternative, without invoking the asymptotic equivalence machinery.

4. NUMERICAL EXAMPLES

In this section we present simulation studies and a real data example. The noise variance is estimated by residual mean square estimators discussed in Section 2.2. The threshold $b(m)$ is determined by equation (10) for a given level α.

4.1 Fixed Simple Alternative

We first look at a simple setting where the alternative is fixed. Let $\theta_j = r\bar{\theta}_j/\|\bar{\theta}\|_2$, with $\bar{\theta}_1 = 0.3$, $\bar{\theta}_j = 4(-1)^j j^2$ for $j \geq 2$. The covariance operator has eigenvalues $\kappa_j = j^{-1.1}$, and eigenfunctions $\phi_1(t) = 1$, $\phi_j(t) = \sqrt{2}\cos((j-1)\pi t)$ for $j \geq 2$. The covariate curves are generated as $X_i(t) = \sum_{j=1}^{100} \sqrt{\kappa_j} X_{ij}\phi_j(t)$, for $i = 1, ..., n$, where X_{ij} are independent standard Gaussian; and $Y_i = \langle X_i, \theta\rangle + Z_i$, with $\theta(t) = \sum_{i=1}^{100} \theta_j\phi_j(t)$ and Z_i being independent standard Gaussian. We use this setting to demonstrate the small sample ($n = 50$), moderate sample ($n = 100$), and large sample ($n = 500$) behavior of our test. This setting corresponds to $\alpha = 1.1$ and $\beta < 1.5$. Although it does not satisfy the second part of Eq. (15), our simulation results suggest that our test is still consistent when the signal is strong enough. A similar setting is used in the estimation literature (Hall & Horowitz (2007)). We considered four different values of our signal strength $r^2 = 0, 0.1, 0.2, 0.5$, where $r^2 = 0$ corresponds to the null hypothesis. We report the percentage of rejections in 500 independently generated samples for all values of n. The result is summarized in Table 1. It can be seen from Table 1 that the test controls type I error at the nominal level and its power increases as the sample size and signal strength increase.

Table 1: Simulation results for a fixed simple alternative under Gaussian design over 500 repetitions. Reported numbers are percentage of rejections.

		$\|\theta\|_2^2 = 0$	0.1	0.2	0.5
level = 5%	$n = 50$	4.4	17.4	34.8	74.2
	100	4.8	30.0	54.2	95.0
	500	3.0	97.0	100	100
1%	$n = 50$	1.4	5.8	18.4	54.4
	100	1.6	14.8	35.6	87.8
	500	0.8	90.2	100	100

4.2 Sparse Signals with Varying Location

In this subsection we demonstrate the minimaxity of the proposed test. We compare our method in terms of the power for a random alternative at targeted type I error level $\alpha = 0.05$, with three competing methods, namely, FVE80, FVE85 and FVE90. The FVE80 method takes $\widehat{m}$ leading principal components such that they explain at least 80 percent of the total variance. Mathematically, $\widehat{m} = \inf_m \frac{\sum_{1\le j\le m}\widehat{\kappa}_j}{\sum_{j\ge 1}\widehat{\kappa}_j} \ge 0.8$ (and likewise for FVE85 and FVE90). Then the test is given by an F-test applied to the principal component regression of dimension $\widehat{m}$.

To avoid using the same alternative throughout the simulation, we randomly generate the regression function $\theta = \sum_{j=1}^{100} \theta_j \phi_j$, where ϕ_j's are the same eigenfunctions as in the previous setting. In the first model, we fix $\theta_j = 0$ for $j > 2$, and let $\theta_j = b_j \cdot I_j$ for $j = 1, 2$, where $b_j's$ are independent Unif$(0, 1)$ random variables and (I_1, I_2) is drawn from a multinomial distribution Mult$(1; 0.5, 0.5)$. That is, a signal of random size appears randomly in one of the first two principal components. Then θ function is scaled by a factor of r for $r^2 = 0, 0., 0.2, 0.5, 1.5$. We denote this model as $M(2, 1)$. The second model is similar, where two signals appear randomly in two of the first nine principal components. We denote this model as $M(9, 2)$. The data is generated similarly as in Section 4.1, expect that we set $\kappa_j = j^{-1.7}$ in generating $X(t)$. This makes the covariate curves $X(t)$ smoother, with fewer principal components selected in FVE methods. In our simulation FVE80 uses 4 or 5 principal components; FVE85 typically uses 6 or 7 PC's; and FVE90 uses 10-12 PC's. Table 2 reports the proportion of rejections over 500 repetitions at level 0.05.

Taking into account of random fluctuations in the simulation, a difference in rejection percentage of 4 can be considered significant. From Table 2 we see that, in all settings the test ψ_{ES} performs at least as well as other methods. In Model (2,1), where the signals concentrate in the first two PC's, the method FVE80 uses four or five leading principal components and hence performs better than the other two FVE methods. The ψ_{ES} test gives similar (or even slightly better, especially when the sample size is small) performance as the FVE80 method. On the other hand, in Model (9,2), the signal is more spread out and the FVE80 method does not consistently outperform the other FVE methods. Actually, in this setting, the signal "randomly" favors one of the three FVE methods. When the signal is strong and sample size is large, FVE85 and FVE90 outperform FVE80, which concentrates only on the first few principal components and will not capture any signal for a significant proportion of the random samples. Again, the ψ_{ES} test is comparable to the best FVE method. A remarkable example is $n = 500$ and $r^2 = 1.5$, where the power of FVE80 is much lower than FVE90, but ψ_{ES} remains powerful.

In the Supplementary Material, we report additional results when the X process is generated from a Gaussian mixture and when the additive error variable is generated from a non-Gaussian distribution. The results agree with those reported in Tables 1 and 2.

4.3 θ generated by spline basis

Now we generate the function θ from B-splines with d degrees of freedom. Two values of d are investigated, where $d = 4$ corresponds to very smooth functions, and $d = 16$ corresponds to less smooth functions. The B-spline coefficients are generated randomly from independent standard Gaussian. When the coefficients are generated, the θ function is then computed and re-scaled so that $\|\theta\|_2^2 = r^2$ for $r^2 = 0, 0.1, 0.2, 0.5, 1$. The results are summarized in Table 3. Because θ is generated from a different basis, its representation on the trigonometric basis (the eigenfunctions of X process) is not sparse and is typically evenly spread in the first few eigenfunctions. This makes FVE80 more preferable than FVE85 and FVE90, because it always captures a good proportion of the signal while has small variability. However, the ψ_{ES} test is competitive in all situations, and is preferable when the sample size is small and the alternative is smooth.

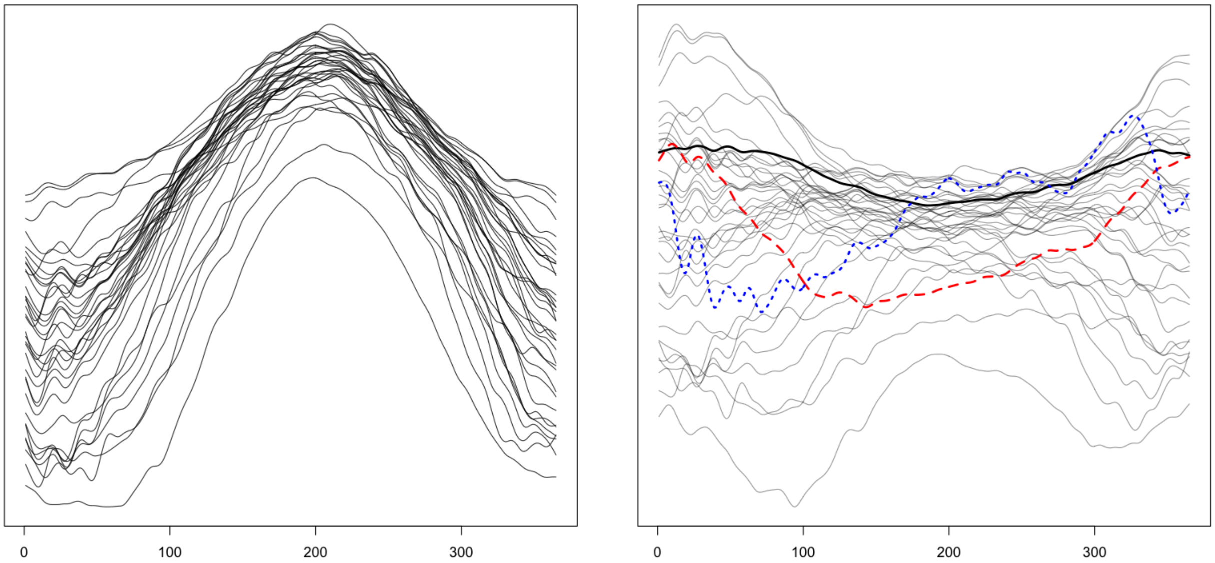
Table 2: Simulation results for randomized signals under Gaussian design over 500 repetitions. Level = 0.05. Reported numbers are percentage of rejections.

			$\|\theta\|_2^2 = 0$	0.1	0.2	0.5	1.5
Model (2,1)	$n = 50$	ψ_{ES}	5.2	16.2	26.4	54.2	80.8
		FVE80	5.4	13.0	21.4	47.8	76.2
		FVE85	3.6	11.2	17.4	41.8	72.4
		FVE90	4.2	9.6	16.4	36.8	65.4
	$n = 100$	ψ_{ES}	4.6	25.8	42.2	68.2	90.4
		FVE80	6.0	24.4	42.6	67.0	89.0
		FVE85	5.4	22.2	37.2	61.8	87.8
		FVE90	3.8	18.4	28.6	54.2	86.4
	$n = 500$	ψ_{ES}	5.8	67.2	84.6	94.4	97.2
		FVE80	7.2	65.0	83.2	94.6	96.8
		FVE85	6.2	62.8	82.4	93.4	96.6
		FVE90	7.0	57.6	79.4	93.6	96.2
Model (9,2)	$n = 50$	ψ_{ES}	5.6	9.0	14.0	29.6	43.4
		FVE80	5.6	9.2	10.8	26.4	43.2
		FVE85	5.8	7.6	10.4	26.0	42.5
		FVE90	5.6	7.2	8.4	22.2	41.6
	$n = 100$	ψ_{ES}	5.6	13.4	27.8	39.8	65.8
		FVE80	6.2	13.2	28.4	38.6	61.8
		FVE85	5.4	11.6	24.2	41.8	67.0
		FVE90	7.6	10.8	23.6	36.8	66.6
	$n = 500$	ψ_{ES}	4.4	42.4	47.8	72.4	93.4
		FVE80	5.2	45.2	50.8	68.6	80.8
		FVE85	5.2	45.6	52.4	75.2	92.8
		FVE90	5.0	40.8	48.2	76.8	95.6

Table 3: Simulation results for randomized signals using spline basis under Gaussian design over 500 repetitions. Level = 0.05. Reported numbers are percentage of rejections.

			$\|\theta\|_2^2 = 0$	0.1	0.2	0.5	1
df = 4	$n = 50$	ψ_{ES}	7.0	26.2	43.8	75.0	91.8
		FVE80	6.8	20.8	36.4	69.4	89.4
		FVE85	6.4	17.4	32.8	65.6	87.6
		FVE90	4.6	12.2	27.4	56.2	83.6
	$n = 100$	ψ_{ES}	5.4	39.8	62.6	90.8	97.8
		FVE80	7.4	37.0	62.0	90.8	98.0
		FVE85	6.0	32.6	53.8	90.2	98.0
		FVE90	6.0	26.6	47.4	86.0	97.4
	$n = 500$	ψ_{ES}	4.8	93.4	96.8	100	100
		FVE80	6.0	94.2	98.2	100	100
		FVE85	4.0	91.8	98.2	100	100
		FVE90	5.0	87.2	97.4	99.8	100
df = 16	$n = 50$	ψ_{ES}	5.6	13.2	17.4	36.0	60.8
		FVE80	4.8	9.0	14.6	30.4	57.0
		FVE85	4.4	9.2	14.8	29.2	52.0
		FVE90	6.2	7.8	11.4	23.6	46.6
	$n = 100$	ψ_{ES}	5.0	15.2	27.4	55.4	75.0
		FVE80	5.2	14.6	24.2	55.4	75.8
		FVE85	4.6	13.2	23.8	52.2	75.6
		FVE90	4.6	12.6	20.0	45.4	71.0
	$n = 500$	ψ_{ES}	4.0	54.0	78.8	96.0	99.0
		FVE80	5.8	56.2	81.4	96.4	98.6
		FVE85	6.2	52.4	79.0	96.6	99.2
		FVE90	5.6	47.4	75.2	95.8	99.0

Figure 1: Daily temperature curves for 35 Canadian weather stations. Left panel: smoothed curves. Right panel, re-centered curves and first three principal components (re-scaled for better visualization). Thick solid curve: FPC1; Dashed curve: FPC2; Dotted curve: FPC3.



4.4 Application to the Canadian Weather Data.

Now we apply our method to the Canadian Weather Data. The data set contains daily temperature together with the annual total precipitation at 35 weather stations. The data is available in the `fda` package in `R`. It has been studied in various works on functional data analysis, such as functional ANOVA (Ramsay & Silverman (2005) Chapter 13), classification and clustering of functional data (Clarkson, Fraley, Gu & Ramsey (2005), Giraldo, Delicado & Mateu (2012)), functional linear regression (Ramsay & Silverman (2005) Chapter 15, James, Wang & Zhu (2009)), and two/multi-sample mean function testing (Zhang & Chen (2007)).

We consider a functional linear regression problem of predicting the (log) annual precipitation using the daily temperature. The daily temperature vector can be treated as a dense, regular, and possibly noisey observation from a smooth curve. As discussed in Section 2.1 (see also the references above), a common practice is to pre-smooth the curves and then treat them as independent observations of a process $X(t)$. Here we use `R` function `smooth.spline` with parameter `spar` $= 0.4$. The 35 smoothed curves are plotted in the left panel of Figure 1.

The centered curves and first three principal components are plotted in the right panel of

Figure 1. The first functional principal component (FPC) is relatively flat, roughly representing the overall temperature during the year (especially the winter temperature). It explains 88% of the total variance of the smoothed temperature curve. In order to make the testing problem more subtle, we remove the linear effect of the first FPC and test if the remaining FPC's have a significant linear relationship with the response variable. More specifically, we consider new response variable $Y_i' = Y_i - \widehat{Y}_i(PC1)$, where $\widehat{Y}_i(PC1)$ is the fitted value when regressing Y on the first FPC. The new covariate curves are $X_i' = X_i - \langle X_i, \widehat{\phi}_1 \rangle \widehat{\phi}_1$. In this case, if we apply functional principal component regression of Y' on X' and test the significance of linear relationship at level 0.05, the result would be different when different numbers of FPC's are used. When the first remaining FPC is used, the p-value is 0.0589, which is slightly higher than the nominal level. However, the p-value becomes 7.5×10^{-5} if the first two remaining FPC's are used. Although it seems straightforward that the linear relationship exists, the test result can go either way, depending on whether or not more than one remaining FPC is used. On the other hand, the ψ_{ES} test checks simultaneously three functional principal components regression models with one, two and four leading FPC's. The resulting p-values are 8.18×10^{-3}, 2.63×10^{-7}, and 5.37×10^{-8}, respectively. The simultaneous test would reject the null hypothesis at both 0.05 and 0.01 level, and hence confirm the existence of a linear relationship between daily temperature curve and log-precipitation after removing the effect of the first FPC. Thus, our method can be used to further support the use of functional principal component regression in global testing, reconciling the p-values when different numbers of FPC's are considered.

5. DISCUSSION

We have derived the critical separation rates and consistent adaptive tests for the functional linear model, by exploiting its connection with the Gaussian sequence model. Our results indicate that the signal strength required by a consistent test is less than the minimax error of estimating θ under the same regularity conditions. Such a fundamental difference between testing and estimation can also be seen from the construction of the test statistics. Usually the test statistics approximate some functions of the parameter θ, rather than θ itself. This observation suggests the possibility of avoiding the eigen-gap and extra smoothness conditions in eq. (15). The consistency result developed in this paper would still hold if one can establish a good lower bound on the norm of the projection of θ on the estimated principal subspace, which reduces to showing the quality of estimated principal subspaces instead of fast convergence rate for each individual eigenfunction. Furthermore, if one can avoid the eigen-gap condition, it is then possible to extend the results to other classes of covariance and smoothness structures, for example, when κ_j and θ_j decay exponentially rather than polynomially. The exponential decay of κ_j corresponds to the "severely ill-posed" inverse problems (compared to the "mildly ill-posed" inverse problems considered here) and the exponential decay of θ_j corresponds to classes of analytic functions (compared to Sobolev functions in this paper).

Hilgert, Mas & Verzelen (2012) independently consider the global testing under the functional linear model with a different alternative hypothesis $\|\Gamma^{1/2}\theta\|_2 \geq r_n'$ (instead of $\|\theta\|_2 \geq r_n$ considered in this paper). They propose simultaneously testing a sequence of FPC regression with geometrically increasing dimension based on the F-statistic. Due to the different alternative class, the asymptotic results are not directly comparable. We tried to investigate the minimax rates of both methods under some special cases. When $\kappa_j = j^{-\gamma}$, one can find a point θ that is on the detection boundary suggested by both methods, and hence no method is dominated by the other (ignoring log terms). We note that our formulation is directly aligned with the majority of estimation and prediction literature (for example, Cai & Hall (2006), Hall & Horowitz (2007), Meister (2011), Cai & Yuan (2012)). The technical tools used in this paper are also different and reflect the deep connection between functional linear model and other popular nonparametric function estimation models.

The results and methods developed in this paper are applicable to functional data where the X process is smooth and observed on a dense grid with noise. The situation for sparsely and noisily observed X process is important and poses further challenges for statistical inferences. It is known that the sparsity of observation may have an effect on estimating the covariance and mean function (Hall et al. (2006), Cai & Yuan (2011)). Intuitively, it shall also affect the global testing. The separation rate and optimal testing procedure for sparsely observed functional data is an interesting and important problem for future study. Another interesting problem for future work is to develop inference methods and theory for other related models, such as functional regression with multiplicative errors.

APPENDIX A. TECHNICAL PROOFS

A.1 Proof of Lemma 2.1

Proof of Lemma 2.1. The proof is direct calculation with linear algebra.

Let $\widehat{X}_{ij} = \widehat{\kappa}_j^{-1/2}\langle X_i, \widehat{\phi}_j\rangle$, for $1 \le i, j \le n$. Denote $\widehat{\mathbf{X}} = (\widehat{X}_{ij})$ be the $n \times n$ matrix whose (i,j)th entry is $\widehat{X}_{ij}$, so that $\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} = n\mathbf{I}$. Let $\widehat{\theta}^* = (\langle\theta, \widehat{\phi}_1\rangle, ..., \langle\theta, \widehat{\phi}_n\rangle)^T$, then

$$\begin{aligned}
\widehat{\theta}_j =& \widehat{\kappa}_j^{-1}\left\langle n^{-1}\sum_{i=1}^n Y_i X_i, \widehat{\phi}_j\right\rangle = \widehat{\kappa}_j^{-1}\left\langle n^{-1}\sum_{i=1}^n \left(\langle X_i, \theta\rangle + \sigma Z_i\right) X_i, \widehat{\phi}_j\right\rangle \\
=& \widehat{\kappa}_j^{-1}\left\langle n^{-1}\sum_{i=1}^n\left(\sum_{k=1}^n \sqrt{\widehat{\kappa}_k}\widehat{X}_{ik}\widehat{\theta}_k^* + \sigma Z_i\right)\sum_{k=1}^n\sqrt{\widehat{\kappa}_k}\widehat{X}_{ik}\widehat{\phi}_k, \widehat{\phi}_j\right\rangle \\
=& \widehat{\kappa}_j^{-1} n^{-1}\sum_{i=1}^n\left(\sum_{k=1}^n\sqrt{\widehat{\kappa}_k}\widehat{X}_{ik}\widehat{\theta}_k^* + \sigma Z_i\right)\sqrt{\widehat{\kappa}_j}\widehat{X}_{ij} \\
=& \widehat{\kappa}_j^{-1/2} n^{-1}\sum_{i=1}^n\left(\sum_{k=1}^n\sqrt{\widehat{\kappa}_k}\widehat{X}_{ik}\widehat{X}_{ij}\widehat{\theta}_k^* + \sigma Z_i\widehat{X}_{ij}\right) \\
=& \widehat{\kappa}_j^{-1/2} n^{-1}\sum_{k=1}^n\widehat{\theta}_k^*\sqrt{\widehat{\kappa}_k}\sum_{i=1}^n\widehat{X}_{ik}\widehat{X}_{ij} + \frac{\sigma}{\sqrt{n\widehat{\kappa}_j}}\sum_{i=1}^n n^{-1/2}\widehat{X}_{ij}Z_i \\
=& \widehat{\theta}_j^* + \frac{\sigma}{\sqrt{n\widehat{\kappa}_j}}Z_j^*,
\end{aligned}$$

where the last inequality uses the orthogonality of $\widehat{\mathbf{X}}$ and defines $Z_j^* = \sum_{i=1}^n n^{-1/2}\widehat{X}_{ij}Z_i$. Again the orthogonality of $\widehat{\mathbf{X}}$ implies that $\mathbf{Z}^*$ is n-dimensional standard Gaussian and independent of $\mathbf{X}$. □

A.2 Proof of Consistency (Theorem 3.2)

Proof of Theorem 3.2. First we consider the type I error. Let $T_{n,m}^* = (2m)^{-1/2}\sum_{j=1}^m(\sigma^{-2}n\widehat{\kappa}_j\widehat{\theta}_j^2 - 1)$. That is, $T_{n,m}^*$ is the "ideal" version of $T_{n,m}$, with $\widehat{\sigma}^2$ replaced by σ^2. When $\theta = 0$, we have $\widehat{\theta}_j^* = 0$ for all j and hence

$$\begin{aligned}
&T_{n,m}^* = \frac{1}{\sqrt{2m}}\left(\sum_{j=1}^m Z_j^{*2} - m\right), \\
&\sup_{k=0}^{k_{n,\max}}|T_{n,m_k} - T_{n,m_k}^*| \le \frac{1}{\sqrt{2m_{k_{n,\max}}}}\sum_{j=1}^{m_{k_{n,\max}}} Z_j^{*2}\left|1 - \frac{\sigma^2}{\widehat{\sigma}^2}\right| \le o_P(n^{-1/5})\sqrt{m_{k_{n,\max}}} = o_P(1),
\end{aligned}$$

because $\sum_{j=1}^m Z_j^{*2} = O_P(m)$ and $m_{k_{n,\max}} = O(n^{1/3})$.

Then $\alpha_n(\psi_{\mathrm{ES}}) = \mathbb{P}_0(\psi_{\mathrm{ES}} = 1)$ and can be bounded by using Lemma 1 of (Laurent & Massart 2000) with $x_k = 2\log\log m_k$ and $a = 1/\sqrt{2m_k}$.

$$
\begin{aligned}
&\mathbb{P}_0(\psi_{\mathrm{SE}}(\mathbf{Y}, \mathbf{X}) = 1)\\
\le&\mathbb{P}_0\left(\exists k:\ T_{n,m_k} \ge 4\sqrt{\log\log m_k}\right)\\
\le&\sum_{k=0}^{k_{n,\max}} \mathbb{P}_0\left(T^*_{n,m_k} \ge 3\sqrt{\log\log m_k}\right) + \mathbb{P}_0\left(\sup_{k=0}^{k_{n,\max}} |T_{n,m_k} - T^*_{n,m_k}| \ge \sqrt{\log\log m_0}\right)\\
\le&\sum_{k=0}^{k_{n,\max}} \mathbb{P}_0\left(T_{n,m_k} \ge 2\sqrt{x_k} + \sqrt{\frac{2}{m_k}x_k}\right) + o(1)\\
\le&\sum_{k\ge 0}(\log m_k)^{-2} + o(1) = \sum_{k\ge 0}(\log m_0 + k\log 2)^{-2} + o(1)\\
\le&\frac{1}{\log 2(\log m_0 - \log 2)} + o(1) = o(1).
\end{aligned}
\tag{A.1}
$$

For the type II error, we expand the test statistic $T_{n,m}$ ($m_0 \le m \le n^{1/3}$) as follows.

$$
\begin{aligned}
T_{n,m} =&\frac{1}{\sqrt{2m}}\sum_{j=1}^{m}\left[\widehat{\sigma}^{-2} n\widehat{\kappa}_j(\widehat{\theta}^*_j)^2 + 2\widehat{\sigma}^{-2}\sigma\sqrt{n\widehat{\kappa}_j}\widehat{\theta}^*_j Z^*_j + \frac{\sigma^2}{\widehat{\sigma}^2}\left((Z^*_j)^2 - 1\right)\right]\\
=&\widehat{\sigma}^{-2}Q_{1,m} + \widehat{\sigma}^{-2}\sigma Q_{2,m} + \frac{\sigma^2}{\widehat{\sigma}^2}Q_{3,m} + Q_{4,m},
\end{aligned}
\tag{A.2}
$$

where

$$
\begin{aligned}
Q_{1,m} &= \frac{1}{\sqrt{2m}}\sum_{j=1}^{m} n\widehat{\kappa}_j(\widehat{\theta}^*_j)^2, \quad Q_{2,m} = \frac{2}{\sqrt{2m}}\sum_{j=1}^{m}\sqrt{n\widehat{\kappa}_j}\widehat{\theta}^*_j Z^*_j,\\
Q_{3,m} &= \frac{1}{\sqrt{2m}}\sum_{j=1}^{m}\left((Z^*_j)^2 - 1\right), \quad Q_{4,m} = \sqrt{\frac{m}{2}}\left(\frac{\sigma^2}{\widehat{\sigma}^2} - 1\right).
\end{aligned}
$$

Now consider a fixed $m \in [m_0, n^{1/3}]$. Because $\widehat{\sigma}^2 - \sigma^2 = o_P(n^{-1/5})$ and $m \le n^{1/3}$, we have $Q_{4,m} = o_P(1)$ and hence $\sigma^2\widehat{\sigma}^{-2}Q_{3,m} + Q_{4,m} = O_P(1)$ uniformly over $\Theta(\beta, L)$. On the other hand, we observe that,

$$
Q_{2,m} = \frac{2}{(2m)^{1/4}}Q_{1,m}^{\frac{1}{2}}\widetilde{Z}_m,
$$

where

$$
\widetilde{Z}_m = \sum_{j=1}^{m}\frac{\sqrt{n\widehat{\kappa}_j}\widehat{\theta}^*_j}{(\sqrt{2m}Q_{1,m})^{1/2}}Z^*_j
$$

is a standard Gaussian because Z^*_j are independent standard Gaussian as shown in Lemma 2.1.

Define $\bar{m} = \lfloor (2L/r_n)^{1/\beta} \rfloor$. If $r_n = o((\log n)^{-\beta/2})$, then $\bar{m} \in [m_0, n^{1/3}]$ for n large enough. By our construction of the sequence $m_0, m_1, ..., m_{k_{n,\max}}$, there exists a unique k such that $\bar{m} \leq m_k < 2\bar{m}$. Denote this m_k by m^*. For all $\theta \in \Theta(\beta, L, r_n)$, we have the following control of type II error.

$$\begin{aligned}
&\mathbb{P}_\theta \left[\psi_{\mathrm{ES}} = 0\right] \leq \mathbb{P}_\theta \left[T_{n,m^*} < 4\sqrt{\log\log m^*}\right] \\
\leq &\mathbb{P}_\theta \left[\widehat{\sigma}^{-2} Q_{1,m^*} + \widehat{\sigma}^{-2}\sigma Q_{2,m^*} \leq 5\sqrt{\log\log m^*}\right] + \mathbb{P}_\theta \left[\frac{\sigma^2}{\widehat{\sigma}^2} Q_{3,m^*} + Q_{4,m^*} \leq -\sqrt{\log\log m^*}\right] \\
= &\mathbb{P}_\theta \left[\widehat{\sigma}^{-2} Q_{1,m^*} + \frac{2}{(2m^*)^{1/4}} \sqrt{Q_{1,m^*}}\widetilde{Z}_{m^*} \leq 5\sqrt{\log\log m^*}\right] + o(1) \\
= &\mathbb{P}_\theta \left[\widehat{\sigma}^{-2} + \frac{2}{(2m^*)^{1/4}} Q_{1,m^*}^{-\frac{1}{2}} \widetilde{Z}_{m^*} \leq \frac{5\sqrt{\log\log m^*}}{Q_{1,m^*}}\right] + o(1).
\end{aligned}$$

In Lemma A.1 we show that $\sqrt{\log\log m^*}/Q_{1,m^*} = o_P(1)$, uniformly for all $\theta \in \Theta(\beta, L, r_n)$. As a consequence, the term $(\sqrt{m^*}Q_{1,m^*})^{-1/2}\widetilde{Z}_{m^*} = o_P(1)$ uniformly over θ. Therefore we have $\mathbb{P}_\theta[\psi_{\mathrm{ES}} = 0] = o(1)$ uniformly over $\Theta(\beta, L, r_n)$ and hence $\lambda_n(\psi_{\mathrm{ES}}, \Theta(\beta, L, r_n)) = o(1)$.

The case $(\log n)^{-\beta/2} = O(r_n)$ can be dealt with by taking $m^* = m_0$, followed by the same argument presented above. □

Lemma A.1. *Let m^* be the value of m_k such that $\bar{m} \leq m_k \leq 2\bar{m}$, where $\bar{m} = \lfloor (2L/r_n)^{1/\beta} \rfloor$ (define $m^* = m_0$ if $\bar{m} < m_0$). Let $U_n(r_n) = nr_n^{(4\beta+2\gamma+1)/2\beta}$. If $r_n = \omega((\sqrt{\log\log n}/n)^{2\beta/(4\beta+2\gamma+1)})$, then there exists a constant $c_{\beta,L} > 0$ such that*

$$\lim_{n\to\infty} \inf_{\theta\in\Theta(\beta,L,r_n)} \mathbb{P}_\theta \left[Q_{1,m^*} \geq c_{\beta,L} U_n(r_n)\right] = 1.$$

Lemma A.1 implies that with overwhelming probability, Q_{1,m^*} grows at least as fast as $U_n(r_n)$. When r_n exceeds the rate $(\sqrt{\log\log n}/n)^{2\beta/(4\beta+2\gamma+1)}$ we have $U_n(r_n) = \omega(\sqrt{\log\log n})$, which is precisely what we need in the proof of Theorem 3.2.

Proof. First consider the case $r_n = o((\log n)^{-\beta/2})$. Then for n large enough we have $m^* \in [\bar{m}, 2\bar{m})$.

We proceed with a decomposition of Q_{1,m^*}:

$$\begin{aligned}
Q_{1,m^*} &= \frac{n}{\sqrt{2m^*}} \sum_{j=1}^{m^*} (\kappa_j + (\widehat{\kappa}_j - \kappa_j))(\theta_j + (\widehat{\theta}_j^* - \theta_j))^2 \\
&= \frac{n}{\sqrt{2m^*}} \sum_{j=1}^{m^*} \kappa_j \theta_j^2 + \frac{n}{\sqrt{2m^*}} \sum_{j=1}^{m^*} (\widehat{\kappa}_j - \kappa_j)\theta_j^2 +
\end{aligned}$$

$$
\begin{aligned}
&2\frac{n}{\sqrt{2m^*}}\sum_{j=1}^{m^*}\kappa_j\theta_j(\widehat{\theta}^*_j-\theta_j)+2\frac{n}{\sqrt{2m^*}}\sum_{j=1}^{m^*}(\widehat{\kappa}_j-\kappa_j)\theta_j(\widehat{\theta}^*_j-\theta_j)+\\
&\frac{n}{\sqrt{2m^*}}\sum_{j=1}^{m^*}\kappa_j(\widehat{\theta}^*_j-\theta_j)^2+\frac{n}{\sqrt{2m^*}}\sum_{j=1}^{m^*}(\widehat{\kappa}_j-\kappa_j)(\widehat{\theta}^*_j-\theta_j)^2\\
:=\ &J_{1,m^*}+J_{2,m^*}+J_{3,m^*}+J_{4,m^*}+J_{5,m^*}+J_{6,m^*} \qquad \text{(A.3)}
\end{aligned}
$$

The plan is to show that, when $r_n=\omega(n^{-2\beta/(4\beta+2\gamma+1)})$ we have

$$
J_{1,m^*}\geq c(\beta,L)U_n(r_n)=\omega(1),
$$

with some constant $c(\beta,L)$ independent of θ, while each of the last five terms is $o_p\left(J_{1,m^*}\right)$.

The desired lower bound of J_{1,m^*} can be derived as follows (see also Ingster et al. (2012)).

$$
\begin{aligned}
J_{1,m^*}=&\frac{n}{\sqrt{2m^*}}\sum_{j=1}^{m^*}\kappa_j\theta_j^2\\
\geq&\frac{n\kappa_{m^*}}{\sqrt{2m^*}}\sum_{j=1}^{m^*}\theta_j^2\geq\frac{n\kappa_{m^*}}{\sqrt{2m^*}}\left(r_n^2-\sum_{j=m^*+1}^{\infty}\theta_j^2\right)\\
=&\frac{n\kappa_{m^*}}{\sqrt{2m^*}}\left(r_n^2-\sum_{j=m^*+1}^{\infty}j^{-2\beta}j^{2\beta}\theta_j^2\right)\geq\frac{n\kappa_{m^*}}{\sqrt{2m^*}}\left(r_n^2-(m^*+1)^{-2\beta}L^2\right)\\
\geq&2^{-3/2}c_1n(m^*)^{-\gamma-1/2}r_n^2\geq 2^{-3/2}c_1(2L)^{-(2\gamma+1)/2\beta}nr_n^{\frac{4\beta+2\gamma+1}{2\beta}}=c(\beta,L)U_n(r_n),
\end{aligned}
$$

where $c(\beta,L)=2^{-3/2}c_1(2L)^{(2\gamma+1)/2\beta}$, and c_1 is the constant in condition (11).

To control the other five terms, the key is to show that the approximation errors $\widehat{\kappa}_j-\kappa_j$ and $\widehat{\theta}^*_j-\theta_j$ are small enough. By Equation (5.4) in Meister (2011) we have, under assumption (15), for all $\theta\in\Theta(\beta,L)$,

$$
\mathbb{E}\sum_j(\widehat{\theta}^*_j-\theta_j)^2\leq C\mathbb{E}\|\widehat{\Gamma}-\Gamma\|_{\text{HS}}^2,
$$

where the constant C depends only on L, and $\|\cdot\|_{\text{HS}}$ is the Hilbert-Schmidt norm. It is standard to show that under assumption (14) we have $\mathbb{E}\|\widehat{\Gamma}-\Gamma\|_{\text{HS}}^2=O(n^{-1})$ (Hall & Horowitz (2007), Section 5.3).

Consider event

$$
E_1:=\left\{\|\widehat{\Gamma}-\Gamma\|_{\text{HS}}^2\leq\frac{\log n}{n}\right\},
$$

then assumption (14) ensures that $\mathbb{P}(E_1) \to 1$ as $n \to \infty$ because by Markov's inequality,

$$\mathbb{P}\left(\|\widehat{\Gamma} - \Gamma\|_{\text{HS}}^2 \geq \frac{\log n}{n} \right) \leq \frac{n}{\log n} \mathbb{E}\|\widehat{\Gamma} - \Gamma\|_{\text{HS}}^2 = o(1).$$

Note that on E_1 we have $\sup_j |\widehat{\theta}_j^* - \theta_j| \leq \sqrt{C}\sqrt{\frac{\log n}{n}}$, and $\sup_j |\widehat{\kappa}_j - \kappa_j| \leq \sqrt{\frac{\log n}{n}}$ (Weyl's inequality).

The arguments used to control the last five terms in the decomposition of Q_{1,m^*} in (A.3) are rather similar. We shall just give the detail for J_{3,m^*}, which is the most complicated one. We also focus the "hardest" case that $\|\theta\|_2 = r_n \sigma$. Then on E_1 we have, by Cauchy-Schwartz,

$$\begin{aligned}
\frac{|J_{3,m^*}|}{J_{1,m^*}} &= \frac{\frac{n}{\sqrt{2m^*}} \left| \sum_{j=1}^{m^*} \kappa_j \theta_j (\widehat{\theta}_j^* - \theta_j) \right|}{\frac{n}{\sqrt{2m^*}} \sum_{j=1}^{m^*} \kappa_j \theta_j^2} \\
&\leq \frac{\frac{n}{\sqrt{2m^*}} \left(\sum_{j=1}^{m^*} \kappa_j \theta_j^2 \right)^{1/2} \left(\sum_{j=1}^{m^*} \kappa_j (\widehat{\theta}_j^* - \theta_j)^2 \right)^{1/2}}{\frac{n}{\sqrt{2m^*}} \sum_{j=1}^{m^*} \kappa_j \theta_j^2} \\
&= \frac{\sqrt{n} \left(\sum_{j=1}^{m^*} \kappa_j (\widehat{\theta}_j^* - \theta_j)^2 \right)^{1/2}}{\left(n \sum_{j=1}^{m^*} \kappa_j \theta_j^2 \right)^{1/2}} \leq \frac{2c_2 (2m^*)^{-1/4} \sqrt{\log n}}{[U_n(r_n)]^{1/2}} \\
&\asymp \frac{\sqrt{\log n} r_n^{\frac{1}{4\beta}}}{n^{1/2} r_n^{\frac{4\beta+2\gamma+1}{4\beta}}} = \frac{\sqrt{\log n}}{n^{1/2} r_n^{\frac{4\beta+2\gamma}{4\beta}}} = o\left(\sqrt{\log n} n^{-\frac{1}{8\beta+4\gamma+2}} \right) = o(1).
\end{aligned}$$

It is straightforward to check that the convergence does not depend on θ and hence is uniform over $\Theta(\beta, L, r_n)$.

In the case $(\log n)^{-\beta/2} = O(r_n)$, then for n large enough we have $\bar{m} < m_0$. It is straightforward to check that $J_{1,m_0} \geq c(\beta, L) n \cdot \text{Poly}(\log n)$, where $\text{Poly}(\log n)$ is a polynomial of $\log n$ and hence $J_{1,m_0} = \omega(n^{1-\delta})$ for all $\delta > 0$. The rest of the proof follows the previous case and is omitted. □

A.3 Proof of Lower Bound (Theorem 3.4)

Proof of Theorem 3.4. The proof of the lower bound result follows from the combination of two existing results, both established recently in the literature.

The first result is the lower bound of detection boundary for the Gaussian sequence model in Eq. (3). As derived in Lemma 2.1, there is an obvious and natural correspondence between the Gaussian sequence model and the functional linear model. As a result, the same testing problem can be studied under both models. Ingster et al. (2012) have shown that the detection boundary for the testing problem in Eq. (12) under the Gaussian sequence model is $r_n^* = (\sqrt{\log\log n}/n)^{2\beta/(4\beta+2\alpha+1)}$. It implies that if $r_n = o(r_n^*)$ then all tests must satisfy $\alpha_n(\psi) + \lambda_n(\psi, \Theta(\beta, L)) \to 1$.

The second result is the asymptotic equivalence between the functional linear model (1) and the Gaussian sequence model (3). Specifically, Meister (2011) has shown that, under regularity conditions (14), (15), the two models are asymptotically equivalent in Le Cam's sense (Le Cam (1986)), provided that the covariance operator Γ and noise variance σ^2 are known in the functional linear model. Roughly speaking, an important consequence of asymptotic equivalence is that any inference procedure for one model can be transformed (without involving unknown parameters) to form an inference procedure for the other one, with the same asymptotic risk for all bounded loss functions.

In this case, the two testing problems shall have exactly the same detection boundary r_n^*. However, for our testing problem the covariance operator and the noise variance are not known and hence the problem can only become harder. Thus the detection boundary for our testing problem is at least as large as r_n^*. □

In the Supplementary Material, we give a direct and constructive proof of a slightly weaker version of Theorem 3.4. That is, we drop the $\log\log n$ term and prove that no test can be consistent when $r_n = o(n^{-2\beta/(4\beta+2\beta+1)})$, with an additional assumption of $4\beta + 2\gamma \geq 1$. The idea of the proof does not involve the asymptotic equivalence machinery, but is based another general framework of Le Cam's (Le Cam (1986), Le Cam & Yang (2000)). The key is to construct a least favorable prior μ of θ supported on $\Theta(\beta, L, r_n)$ such that the induced marginal distribution of $(\mathbf{Y}, \mathbf{X})$ is close to that under the null. Such a least favorable prior also provides insights to the upper bound construction.

REFERENCES

Arias-Castro, E., Candès, E. J., & Plan, Y. (2011), "Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism," *The Annals of Statistics*, 39, 2533–2556.

Cai, T., & Hall, P. (2006), "Prediction in functional linear regression," *The Annals of Statistics*, 34, 2159–2179.

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

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

Cai, T., & Zhou, H. (2008), "Adaptive functional linear regression,", Manuscript.

Cardot, H., Ferraty, F., Mas, A., & Sarda, P. (2003), "Testing hypotheses in the functional linear model," *Scandinavian Journal of Statistics*, 30, 241–255.

Cardot, H., Ferraty, F., & Sarda, P. (2003), "Spline estimators for the functional linear model," *Statistica Sinica*, 13, 571–591.

Cardot, H., & Johannes, J. (2010), "Thresholding projection estimators in functional linear models," *Journal of Multivariate Analysis*, 101, 395–408.

Cardot, H., Prchal, L., & Sarda, P. (2007), "No effect and lack-of-fit permutation tests for functional regression," *Computational Statistics*, 22, 371–390.

Cavalier, L., & Tsybakov, A. (2002), "Sharp adaptation for inverse problems with random noise," *Probability Theory and Related Fields*, 123, 323–354.

Clarkson, D., Fraley, C., Gu, C., & Ramsey, J. (2005), *S+ Functional Data Analysis* Springer.

Crambes, C., Kneip, A., & Sarda, P. (2009), "Smoothing splines estimators for functional linear regression," *The Annals of Statistics*, 37, 35–72.

Donoho, D., & Jin, J. (2004), "Higher criticism for detecting sparse heterogeneous mixtures," *The Annals of Statistics*, 32, 962–994.

Giraldo, R., Delicado, P., & Mateu, J. (2012), "Hierarchical clustering of spatially correlated functional data," *Statistica Neerlandica*, 66, 403–421.

González-Manteiga, W., & Martínez-Calvo, A. (2011), "Bootstrap in functional linear regression," *Journal of Statistical Planning and Research*, 141, 453–461.

Hall, P., & Horowitz, J. (2007), "Methodology and Convergence Rates for Functional Linear Regression," *The Annals of Statistics*, 35, 70–91.

Hall, P., Müller, H., & Wang, J. (2006), "Properties of principal component methods for functional and longitudinal data analysis," *The Annals of Statistics*, 34, 1493–1517.

Hall, P., & van Keilegom, I. (2007), "Two-sample tests in functional data analysis starting from discrete data," *Statistica Sinica*, 17, 1511–1531.

Hilgert, N., Mas, A., & Verzelen, N. (2012), "Minimax adaptive tests for the Functional Linear model,", http://arxiv.org/abs/1206.1194.

Ingster, Y. I. (1982), "Minimax nonparametric detection of signals in Gaussian white noise," *Problems in Information Transmission*, 18, 130–140.

Ingster, Y. I., Sapatinas, T., & Suslina, I. A. (2012), "Minimax signal detection in ill-posed inverse problems," *The Annals of Statistics*, 40(3), 1524–1549.

Ingster, Y. I., Tsybakov, A. B., & Verzelen, N. (2010), "Detection boundary in sparse regression," *Electronic Journal of Statistics*, 4, 1476–1526.

Ingster, Y., & Suslina, I. (2003), *Nonparametric Goodness-of-Fit Testing Under Gaussian Models* Springer.

James, G. M., Wang, J., & Zhu, J. (2009), "Functional linear regression that's interpretable," *The Annals of Statistics*, 37, 2083–2108.

Laurent, B., & Massart, P. (2000), "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, 28, 1302–1338.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory* Springer-Verlag.

Le Cam, L., & Yang, G. (2000), *Asymptotics in Statistics: Some Basic Concepts* Springer.

Meister, A. (2011), "Asymptotic Equivalence of Functional Linear Regression and a White Noise Inverse Problem," *The Annals of Statistics*, 39, 1471–1495.

Ramsay, J. O., & Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn Springer.

Rao, M. M. (2000), *Stochastic Processes: Inference Theory* Kluwer Academic Publishers.

Shen, Q., & Faraway, J. (2004), "An F test for linear models with functional responses," *Statistica Sinica*, 14, 1239–1257.

Spokoiny, V. G. (1996), "Adaptive hypothesis testing using wavelets," *The Annals of Statistics*, 24, 2477–2498.

Verzelen, N., & Villers, F. (2010), "Goodness-of-fit tests for high-dimensional Gaussian linear models," *The Annals of Statistics*, 38, 704–752.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.

Yuan, M., & Cai, T. (2010), "A reproducing kernel Hilbert space approach to functional linear regression," *The Annals of Statistics*, 38, 3412–3444.

Zhang, J.-T., & Chen, J. (2007), "Statistical inferences for functional data," *The Annals of Statistics*, 35, 1052–1079.

Supplementary Material

APPENDIX B. A CONSTRUCTIVE LOWER BOUND PROOF

We state and give a constructive proof of the non-adaptive lower bound.

Theorem B.1. *Consider testing problem* (12) *under model* (1) *and condition (11). Assume further that* $4\beta + 2\gamma > 1$ *then all tests* $\psi(\mathbf{Y}, \mathbf{X}) : (\mathbb{R} \otimes L_2[0,1])^n \mapsto [0,1]$ *satisfy*

$$\lim_{n\to\infty} \alpha_n(\psi) + \lambda_n(\psi, \Theta(\beta, L, r_n)) = 1,$$

whenever $r_n = o(n^{-(2\beta)/(4\beta+2\gamma+1)})$.

Proof. For any distribution μ supported on $\Theta(\beta, L, r_n)$, let $\mathbb{P}_\mu(\mathbf{Y}, \mathbf{X})$ be the joint distribution of the n-fold data set $(Y_1, X_1), ..., (Y_n, X_n)$ when $\theta \sim \mu$, and $\mathbb{P}_0(\mathbf{Y}, \mathbf{X})$ be the corresponding distribution when $\theta = 0$. By Le Cam's theory, if $d_{\mathrm{tv}}(\mathbb{P}_\mu, \mathbb{P}_0) \to 0$, then for any test the worst case total error tends to 1. Here $d_{\mathrm{tv}}(\cdot, \cdot)$ denotes the usual total variation distance between two probability measures. If $\mathbb{P}_\mu$ and $\mathbb{P}_0$ have densities f_μ and f_0 then $d_{\mathrm{tv}}(\mathbb{P}_\mu, \mathbb{P}_0) = 2^{-1}\int |f_\mu - f_0|$. In our problem X is a random function and density does not exist. However, the likelihood ratio $d\mathbb{P}_\mu/d\mathbb{P}_0$ is still well-defined (Rao (2000), Chapter V) and can be used to control the total variation distance (Ingster & Suslina (2003)):

$$d_{\mathrm{tv}}(\mathbb{P}_0, \mathbb{P}_\mu) \le \frac{1}{2}\sqrt{\mathbb{E}_0\left(\frac{d\mathbb{P}_\mu}{d\mathbb{P}_0}\right)^2 - 1}.$$

Thus, in order to derive the lower bound, it suffices to find a μ supported on $\Theta(\beta, L, r_n)$ such that $\mathbb{E}_0(d\mathbb{P}_\mu/d\mathbb{P}_0)^2 \to 1$ when $r_n n^{2\beta/(4\beta+2\gamma+1)} \to 0$. If we choose X to be a Gaussian process with covariance Γ, then the analysis reduces to bounding the likelihood ratio for an infinite dimensional Gaussian random design regression model, which is related to a similar argument in (Verzelen & Villers 2010).

The key is to find a distribution μ supported on $\Theta(\beta, L, r_n)$ (note that μ must depend on r_n) such that

$$\limsup_{n\to\infty} \mathbb{E}_0\left(\frac{d\mathbb{P}_\mu}{d\mathbb{P}_0}\right)^2 = 1.$$

To proceed, pick an arbitrary orthonormal basis $\{\phi_j : j \ge 1\}$ of $L_2[0,1]$ (for example, the trigonometric basis). Let $X_i = \sum_{j\ge1}\sqrt{\kappa_j}X_{ij}\phi_j$, for $1 \le i \le n$ and $j \ge 1$, where X_{ij}'s are independent standard Gaussian variables, and $\kappa_j > 0$ satisfies (11) and (15). Now define $\theta = \sum_{j\geq 1} \theta_j \phi_j$ with $\theta_j = \zeta_j \tau_j \sigma$, where $\{\tau_j : j \geq 1\} \in \Theta(\beta, L, r_n)$ is to be chosen later and ζ_j's are independent Rademacher random variables. Let μ be the corresponding distribution of θ.

For convenience, we let, with a slight abuse of notation, $\mathbf{X}$ be the $n \times \infty$ matrix whose (i,j)th entry is $\sqrt{\kappa_j} X_{ij}$.

Now calculate the n-tuple likelihood ratio with a given θ from model (1):

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}(\mathbf{Y}, \mathbf{X}) = \exp\left\{\sigma^{-2}\mathbf{Y}^T\mathbf{X}\theta - \frac{1}{2\sigma^2}\|\mathbf{X}\theta\|_2^2\right\}.$$

Let θ and θ' be two independent copies of $\theta \sim \mu$ generated from ζ and ζ', respectively, then

$$\begin{aligned}
\mathbb{E}_0\left[\frac{d\mathbb{P}_\mu(\mathbf{Y},\mathbf{X})}{d\mathbb{P}_0(\mathbf{Y},\mathbf{X})}\right]^2 &= \mathbb{E}_0\left[\mathbb{E}_{\theta\sim\mu}\exp\left\{\sigma^{-2}\mathbf{Y}^T\mathbf{X}\theta - \frac{1}{2\sigma^2}\|\mathbf{X}\theta\|_2^2\right\}\right]^2 \\
&= \mathbb{E}_0\mathbb{E}_{\theta,\theta'\sim\mu}\exp\left\{\sigma^{-2}\mathbf{Y}^T\mathbf{X}(\theta+\theta') - \frac{1}{2\sigma^2}\|\mathbf{X}\theta\|_2^2 - \frac{1}{2\sigma^2}\|\mathbf{X}\theta'\|_2^2\right\} \\
&= \mathbb{E}_{\theta,\theta'\sim\mu}\mathbb{E}_{\mathbf{X}}\exp\{\sigma^{-2}\theta^T\mathbf{X}^T\mathbf{X}\theta'\} \\
&= \mathbb{E}_{\zeta,\zeta'}\left[\mathbb{E}_{X_1}\exp\left\{\sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j\tau_j \sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j'\tau_j\right\}\right]^n.
\end{aligned}$$

Now let $W = \sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j\tau_j$ and $W' = \sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j'\tau_j$. Then, conditioning on (ζ, ζ'), W and W' are jointly Gaussian, with common marginal distribution $N(0, s^2)$ where $s^2 = \sum_j \kappa_j\tau_j^2$. Note that $\rho = \mathrm{Cov}(W, W') = \sum_j \kappa_j\tau_j^2\zeta_j\zeta_j' \overset{d}{=} \sum_j \kappa_j\tau_j^2\zeta_j$ since ζ and ζ' are independent Rademacher random variables.

Let $W'' = W' - \rho s^{-2}W$, then $W'' \sim N(0, s^2 - \rho^2 s^{-2})$ and is independent of W. Then we have, conditioning on ζ, ζ',

$$\begin{aligned}
&\mathbb{E}_{X_1}\exp\left\{\sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j\tau_j \sum_{j=1}^\infty \sqrt{\kappa_j}X_{1j}\zeta_j'\tau_j\right\} \\
=&\mathbb{E}_{W,W'}\exp(WW') = \mathbb{E}_{W,W''}\exp\left\{W''W + \rho s^{-2}W^2\right\} \\
=&\mathbb{E}_W\left[\exp\left(\rho s^{-2}W^2\right)\mathbb{E}_{W''}\left(\exp\left\{W''W\right\}\middle| W\right)\right] \\
=&\mathbb{E}_W\exp\left(\rho s^{-2}W^2 + \frac{1}{2}(s^2 - \rho^2 s^{-2})W^2\right) = \sqrt{\frac{1}{1-(2\rho + s^4 - \rho^2)}}.
\end{aligned}$$

Then, for s^2 small enough using the fact $\log(1-x)^{-1} \leq x + x^2$ for $x \leq 1/2$,

$$\mathbb{E}_0\left[\frac{d\mathbb{P}_\mu(\mathbf{Y},\mathbf{X})}{d\mathbb{P}_0(\mathbf{Y},\mathbf{X})}\right]^2 = \mathbb{E}_\zeta\exp\left\{\frac{n}{2}\log\frac{1}{1-(2\rho+s^4-\rho^2)}\right\}$$

$$
\begin{aligned}
&\leq \mathbb{E}_\zeta \exp\left\{\frac{n}{2}\left[2\rho + s^4 - \rho^2 + (2\rho + s^4 - \rho^2)^2\right]\right\} \\
&\leq \exp(5ns^4)\mathbb{E}_\zeta \exp\{n\rho\} = \exp(5ns^4)\prod_{j=1}^{\infty}\cosh(n\kappa_j\tau_j^2) \\
&\leq \exp\left(5ns^4 + \frac{n^2}{2}\sum_{j=1}^{\infty}\kappa_j^2\tau_j^4\right). \qquad \text{(A.4)}
\end{aligned}
$$

Consider the sequence: $\tau_j^2 = A_n\kappa_j^{-1}\left(1 - (j/M_n)^{2\beta}\right)_+$, where A_n and M_n are chosen such that $\sum_j \tau_j^2 = r_n^2\sigma^2$, $\sum_j j^{2\beta}\tau_j^2 = L^2\sigma^2$. This choice of τ is motivated by minimizing $\sum_j \kappa_j^2\tau_j^4$ subject to $\sum_j \tau_j^2 \geq r_n^2\sigma^2$ and $\sum_j j^{2\beta}\tau_j^2 \leq L^2\sigma^2$. Its derivation and existence will be discussed in detail in Equation (A.4).

Then, by Riemann sum approximation,

$$
\begin{aligned}
r_n^2\sigma^2 = \sum_j \tau_j^2 &\asymp A_n \sum_{j=1}^{[M_n]} j^\gamma \left(1 - \left(\frac{j}{M_n}\right)^{2\beta}\right) \asymp A_n M_n^{\gamma+1}\int_0^1 t^\gamma(1 - t^{2\beta})dt \\
&\asymp A_n M_n^{\gamma+1}.
\end{aligned}
$$

Similarly from $\sum_j j^{2\beta}\tau_j^2 = L^2\sigma^2$ we derive $A_n M_n^{2\beta+\gamma+1} \asymp 1$. It is then straightforward to verify that

$$
A_n \asymp r_n^{\frac{2\beta+\gamma+1}{\beta}}, \qquad M_n \asymp r_n^{-\frac{1}{\beta}}.
$$

Then $\sum_j \kappa_j^2\tau_j^4 \asymp r_n^{\frac{4\beta+2\gamma+1}{\beta}}$, and $s^2 = \sum_j \kappa_j\tau_j^2 \asymp r_n^{\frac{2\beta+\gamma}{\beta}}$. Therefore, when $r_n = o(n^{-\frac{2\beta}{4\beta+2\gamma+1}})$, we have $ns^4 = o(1)$ (because $4\beta + 2\gamma \geq 1$) as well as $n^2\sum_j \kappa_j^2\tau_j^4 = o(1)$. □

More on the construction of the least favorable prior

The choice of the sequence τ_j in the proof of Theorem B.1 may seem mysterious at first. Here we give a full explanation for the motivation and detailed derivation.

Recall that in the proof of Theorem 3.4, we need to show that there exists a $\tau \in \Theta(\beta, L, r_n)$ such that the right hand side of Equation (A.4) is $o(1)$ whenever $r_n = o(n^{-2\beta/(4\beta+2\gamma+1)})$. Observe that there are two terms in the exponent: $5n\gamma^4$ and $2^{-1}n^2\sum_j \kappa_j^2\tau_j^4$. The second term has an extra factor of n and shall be dominating. The strategy is to minimize the second term over all $\tau \in \Theta(\beta, L, r_n)$, and show that the value of Equation (A.4) is $o(1)$ at this minimum point. Such a minimization problem has also been used to give a lower bound for the Gaussian sequence model in (Ingster et al. 2012) but the details are omitted there. Here we work out the full details. In this subsection we shall work with $\sigma^2 = 1$ with out loss of generality.

Consider a optimization problem:

$$\max_{\tau} \sum_j \kappa_j^2 \tau_j^4, \qquad \text{s.t.} \sum_j \tau_j^2 \geq r_n^2, \ \sum_j j^{2\beta} \tau_j^2 \leq L^2. \tag{A.5}$$

Let $x_j = \kappa_j \tau_j^2$, $\sigma_j^2 = \kappa_j^{-1}$, $a_j = L^{-1/\beta} j^\beta$, the optimization problem can be written equivalently as

$$\min_x \sum_j x_j^2, \qquad \text{s.t.} \sum_j \sigma_j^2 x_j \geq r_n^2, \quad \sum_j a_j^2 \sigma_j^2 x_j \leq 1, \quad x_j \geq 0, \ \forall\ j.$$

Consider Lagrangian, with $u > 0$, $v > 0$ and $w > 0$,

$$\begin{aligned} L(x; u, v, w) &= \sum_j x_j^2 + u \left(r_n^2 - \sum_j \sigma_j^2 x_j \right) + v \left(\sum_j a_j^2 \sigma_j^2 x_j - 1 \right) - \sum_j w_j x_j \\ &= \sum_j \left[x_j^2 - (u\sigma_j^2 - v a_j^2 \sigma_j^2 + w_j) x_j \right] + r_n^2 u - v. \end{aligned}$$

The dual function

$$g(u, v, w) = \min_x L(x; u, v, w) = -\frac{1}{4} \sum_j (u\sigma_j^2 - v a_j^2 \sigma_j^2 + w_j)^2 + r_n^2 u - v.$$

The dual problem is

$$\max_{u,v,w} \ g(u, v, w), \quad \text{s.t. } u \geq 0, \ v \geq 0, w_j \geq 0, \ \forall\ j.$$

Observe that $g(u, v, w)$ must be maximized by taking $\widetilde{w}_j = (v a_j^2 \sigma_j^2 - u \sigma_j^2)_+$.

Let

$$g(u, v) = \max_w g(u, v, w) = -\frac{1}{4} \sum_j (u\sigma_j^2 - v a_j^2 \sigma_j^2)_+^2 + r_n^2 u - v.$$

Assume the optimal solution for u is not 0. Let $v = Bu$. Then the dual problem becomes

$$\max_{u,B} \ -\frac{1}{4} \left[\sum_j \sigma_j^4 (1 - B a_j^2)_+^2 \right] u^2 + (r_n^2 - B) u, \quad \text{s.t. } u \geq 0, \ B \geq 0.$$

The above maximization problem is maximized by taking

$$\widetilde{u} = \frac{2(r_n^2 - B)_+}{\sum_j \sigma_j^4 (1 - B a_j^2)_+^2}.$$

Let $g(B) = g(\widetilde{u}, B)$. We have

$$g(B) = \begin{cases} \frac{(r_n^2 - B)^2}{\sum_j \sigma_j^4 (1 - Ba_j^2)_+^2}, & \text{if } 0 \leq B \leq r_n^2, \\ 0, & B \geq r_n^2. \end{cases}$$

Note that $g(0) = g(r_n^2) = 0$ because $\sum_j \sigma_j^4 = \infty$. Therefore the maximizer of B must be in $(0, r_n^2)$ and hence $\widetilde{u} > 0$, $\widetilde{v} > 0$. By complimentary slackness we must have $\sum_j \sigma_j^2 \widetilde{x}_j = r_n^2$, $\sum_j a_j^2 \sigma_j^2 \widetilde{x}_j = 1$, and $\widetilde{x}_j = 0$, for all $1 - Ba_j^2 < 0$.

Note that

$$\widetilde{x}_j = \frac{1}{2}(\widetilde{u}\sigma_j^2 - \widetilde{v}a_j^2\sigma_j^2 + \widetilde{w}_j).$$

Plugging in,

$$\begin{aligned} \widetilde{x}_j &= \frac{1}{2}(\widetilde{u}\sigma_j^2 - \widetilde{v}a_j^2\sigma_j^2)_+ = \frac{1}{2}\widetilde{u}\sigma_j^2(1 - Ba_j^2)_+ \\ &= \frac{(r_n^2 - B)}{\sum_j \sigma_j^4 (1 - Ba_j^2)_+^2} \sigma_j^2 (1 - Ba_j^2)_+ = A_n \sigma_j^2 (1 - Ba_j^2)_+, \end{aligned}$$

with

$$A_n = \frac{(r_n^2 - B)}{\sum_j \sigma_j^4 (1 - Ba_j^2)_+^2}.$$

APPENDIX C. ADDITIONAL SIMULATION RESULTS

C.1 Gaussian mixture design

Now we provide simulation results that parallels Tables 1 and 2 using a Gaussian mixture design. The results are qualitatively similar to those in the Gaussian design.

Table 4: Simulation results for a fixed simple alternative under Gaussian mixture design over 500 repetitions. Reported numbers are percentage of rejections. For $\langle X, \phi_j \rangle$ the mixture distributions are $N(-0.5\sqrt{\kappa_j}, 0.75\kappa_j)$ and $N(0.5\sqrt{\kappa_j}, 0.75\kappa_j)$ with equal proportion.

		$r^2 = 0$	0.1	0.2	0.5
level = 5%	$n = 50$	4.8	16.8	34.4	71.4
	100	5.4	24.8	55.4	92.2
	500	4.8	96.2	100	100
1%	$n = 50$	1.0	8.6	18.4	51.8
	100	1.2	12.6	35.6	81.8
	500	0.8	89.8	100	100

C.2 Non-Gaussian error

Here we give an example of non-Gaussian error distribution. The simulation set up is analogous to those in Table 1 and Table 4, but the error distribution is changed from a standard Gaussian to a Student-t distribution with six degrees of freedom (normalized to have standard deviation 1). The simulation results are similar to those for the case of Gaussian error.

Table 5: Simulation results for randomized signals under Gaussian mixture design over 500 repetitions. Level = 0.05. Reported numbers are percentage of rejections. For $\langle X, \phi_j \rangle$ the mixture distributions are $N(-0.5\sqrt{\kappa_j}, 0.75\kappa_j)$ and $N(0.5\sqrt{\kappa_j}, 0.75\kappa_j)$ with equal proportion.

			$\|\theta\|_2^2 = 0$	0.1	0.2	0.5	1.5
Model (2,1)	$n = 50$	ψ_{ES}	4.4	16.2	29.8	52.0	76.6
		FVE80	5.0	13.8	26.2	48.0	75.2
		FVE85	4.2	13.0	23.8	44.8	71.8
		FVE90	3.8	11.4	21.0	40.4	70.6
	$n = 100$	ψ_{ES}	4.2	24.0	42.0	69.0	88.6
		FVE80	4.2	27.4	41.0	69.6	89.2
		FVE85	4.6	24.6	39.0	65.6	88.0
		FVE90	5.4	19.6	31.6	60.8	86.6
	$n = 500$	ψ_{ES}	3.2	66.6	82.0	94.2	99.0
		FVE80	4.0	67.2	82.0	95.2	99.0
		FVE85	4.6	65.4	80.4	93.8	99.2
		FVE90	4.2	58.6	76.4	92.4	98.8
Model (9,2)	$n = 50$	ψ_{ES}	4.4	12.4	20.6	32.2	56.2
		FVE80	5.6	11.0	19.0	30.8	55.6
		FVE85	3.6	9.4	17.0	28.4	55.4
		FVE90	5.0	9.2	16.6	26.6	52.4
	$n = 100$	ψ_{ES}	5.4	17.2	21.0	43.2	69.0
		FVE80	4.6	17.6	22.4	43.0	64.6
		FVE85	4.2	17.4	21.4	43.0	68.4
		FVE90	4.8	15.6	19.0	42.2	68.2
	$n = 500$	ψ_{ES}	4.4	41.6	56.4	79.6	93.4
		FVE80	5.2	45.2	59.6	75.8	90.6
		FVE85	4.6	44.2	60.6	77.0	91.4
		FVE90	6.2	41.0	59.6	82.6	94.2

Table 6: Simulation results for a fixed simple alternative under Gaussian design over 500 repetitions. The error distribution is Student-t with six degrees of freedom. Reported numbers are percentage of rejections.

		$r^2 = 0$	0.1	0.2	0.5
level = 5%	$n = 50$	4.4	17.4	34.8	74.2
	100	4.8	30.0	54.2	95.0
	500	3.0	97.0	100	100
1%	$n = 50$	1.4	5.8	18.4	54.4
	100	1.6	14.8	35.6	87.8
	500	0.8	90.2	100	100