The Annals of Statistics
2015, Vol. 43, No. 4, 1742–1773
DOI: 10.1214/15-AOS1322
© Institute of Mathematical Statistics, 2015

NONPARAMETRIC INFERENCE IN GENERALIZED FUNCTIONAL LINEAR MODELS

BY ZUOFENG SHANG[1] AND GUANG CHENG

Purdue University

We propose a roughness regularization approach in making nonparametric inference for generalized functional linear models. In a reproducing kernel Hilbert space framework, we construct asymptotically valid confidence intervals for regression mean, prediction intervals for future response and various statistical procedures for hypothesis testing. In particular, one procedure for testing global behaviors of the slope function is adaptive to the smoothness of the slope function and to the structure of the predictors. As a by-product, a new type of Wilks phenomenon [*Ann. Math. Stat.* **9** (1938) 60–62; *Ann. Statist.* **29** (2001) 153–193] is discovered when testing the functional linear models. Despite the generality, our inference procedures are easy to implement. Numerical examples are provided to demonstrate the empirical advantages over the competing methods. A collection of technical tools such as integro-differential equation techniques [*Trans. Amer. Math. Soc.* (1927) **29** 755–800; *Trans. Amer. Math. Soc.* (1928) **30** 453–471; *Trans. Amer. Math. Soc.* (1930) **32** 860–868], Stein's method [*Ann. Statist.* **41** (2013) 2786–2819] [Stein, *Approximate Computation of Expectations* (1986) IMS] and functional Bahadur representation [*Ann. Statist.* **41** (2013) 2608–2638] are employed in this paper.

1. Introduction.

Rapid development in technology makes it possible to collect measurements intensively over an entire time domain. This forms the so-called *sample curve*. In functional data analysis, one may regress the response variable on the sample curve using (generalized) functional linear models, as in, for example, [8, 20]. Functional principle component analysis (FPCA) is commonly used for analyzing such models; see, for instance, [2, 3, 13–15, 17, 34]. For example, Müller and Stadtmüller [20] proposed a set of FPCA-based inference procedures, while Dou et al. [8] established minimax estimation rates in a similar framework. The success of these FPCA-based approaches hinges on the availability of a good estimate of the functional principal components for the slope function; see [4]. On the other hand, the truncation parameter in the FPCA changes in a discrete manner, which may yield an imprecise control on the model complexity, as pointed

Received May 2014; revised February 2015.

[1] Supported by NSF CAREER Award DMS-11-51692, DMS-14-18042, Simons Foundation 305266.

MSC2010 subject classifications. Primary 62G20, 62F25; secondary 62F15, 62F12.

Key words and phrases. Generalized functional linear models, minimax adaptive test, nonparametric inference, reproducing kernel Hilbert space, roughness regularization.

out in [21]. Recently, Crambes et al. [6], Yuan and Cai [35] and Cai and Yuan [4], among others, have proposed roughness regularization methods that circumvent the aforementioned perfect alignment requirement and allow one to regularize the model complexity in a continuous manner. As far as we are aware, these works focus mostly on the *estimation or prediction* in the functional linear models. An exception is the prediction intervals obtained in [6] under the restrictive Gaussian errors; see (5.3) therein. However, it is yet unknown how to handle a broader range of inference problems such as (adaptive) hypothesis testing for generalized functional linear models in the above roughness regularization framework.

The major goal of this paper is to systematically conduct asymptotic inference in the class of generalized functional linear models, which cover ℓ_2 regression, logistic regression and exponential family models. Specifically, we construct confidence intervals for regression mean, prediction intervals for future response and various statistical procedures for hypothesis testing. As far as we are aware, all these inference results are new. In particular, these inference procedures maintain the modeling and computation flexibility by taking advantage of the roughness regularization. However, this practical superiority comes at the price of a much harder theoretical investigation. A key technical tool we develop in this paper is the *Bahadur representation for functional data*, which provides a unified treatment for various inference problems. Due to the involvement of a covariance operator, we note that this new Bahadur representation is dramatically different from that recently established in the nonparametric regression framework [26]. In addition, we employ the integro-differential equation techniques [29–31] to explicitly characterize the underlying eigen-system that leads to more transparent inference procedures; see Proposition 2.2. As a side remark, our general theory does not require the Sacks–Ylvisaker (SY) conditions as in [35], although assuming a pseudo version of SY conditions (given in Section S.2) can facilitate the implementation.

To be more specific, we show that the proposed confidence/prediction intervals asymptotically achieve the desirable coverage probability. We also propose a procedure for testing functional contrast and show the null limit distribution as a standard normal distribution. As for testing global behaviors of the slope function, we propose a penalized likelihood ratio test (PLRT) that achieves the minimax rate of testing established in [15]. In the particular case of functional linear models, we observe a new version of the Wilks phenomenon [11, 33] arising from PLRT, by which we mean that the null limit distribution, which is derived as a Chi-square distribution with diverging degrees of freedom, is free of the true model parameters. A major advantage of the Wilks type of results is that we can directly simulate the null limit distribution (without resorting to bootstrap) in practice. In PLRT, we also point out that the class of functions in the alternative hypothesis is allowed to be infinite-dimensional in contrast to the parametric class considered in [15]. Besides, the rejection region of PLRT is based on the asymptotic distribution, which makes the procedure more applicable in general modeling setup, that is, in *generalized* functional linear models.

In reality, the smoothness of the slope function and the structure of the predictors are typically unknown. To address this issue, we modify the above PLRT in an "adaptive" fashion. Explicitly, we conduct a sequence of standardized PLRT procedures over multiple smoothness levels, and then use the maximal one as the new test (after a second standardization). This new testing method does not rely on prior knowledge of the above two crucial quantities, and is shown to achieve the minimax rate of testing (up to logarithm term) established in [15]. In fact, our adaptive procedures can be viewed as a generalization of the adaptive Neyman test studied in [9, 10] to functional data. Due to the distinct model structure and test construction, the Darling–Erdős theorem used in [9, 10] is no longer applicable. Instead, we adapt the impressive and powerful Gaussian approximation tool recently proposed in [5] to show that in both Gaussian and sub-Gaussian settings, the null limit is a type of extreme value distribution. Our adaptive testing procedures differ from the FPCA-based tests such as those considered in [15, 17] in two ways: (i) our tests work for non-Gaussian models; (ii) our tests provide an asymptotic null limit distribution, from which the correct test size can be achieved. Besides, our tests do not require the "eigen-gap" condition in the FPCA literature, as in, for example, [17]. Simulation results demonstrate the advantages of our methods in terms of desirable sizes and powers. In particular, we observe that PLRT is more powerful than the adaptive testing procedures. This is reasonable since PLRT incorporates prior knowledge on smoothness of the covariance and reproducing kernels. However, their difference quickly vanishes when the sample size is large or the signal strength is strong.

The rest of this paper is organized in the following way. In Section 2, basic assumptions on model and parameter space are given. Section 3 presents the key technical device of this paper: Bahadur representation for functional data. In Section 4, asymptotically valid confidence intervals for regression mean and prediction intervals for future response are constructed. In Section 5, a procedure for testing functional contrast and a global testing for the slope function, that is, PLRT, are established. Theoretical properties are also demonstrated. Section 6 contains two adaptive testing procedures for either Gaussian or sub-Gaussian errors. Their null limit distributions and minimax properties are carefully examined. A simulation study is provided in Section 7. The generalized cross validation (GCV) is used to select the roughness penalty parameter in the simulations. Section 8 discusses the technical connection between our work and [26]. All technical proofs are deferred to the Supplementary Material [27].

2. Preliminaries.

2.1. *Model assumptions.*

Suppose the data $(Y_i, X_i(t))$, $i = 1, \dots, n$, are i.i.d. copies of $(Y, X(t))$, where Y is a univariate response variable taking values in $\mathcal{Y}$, a subset of real numbers, and $X(t)$ is a real-valued random predictor process over $\mathbb{I} = [0, 1]$. Consider the following generalized functional linear model:

$$\mu_0(X) \equiv E\{Y|X\} = F\left(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt\right), \tag{2.1}$$

where F is a known link function, α_0 is a scalar and $\beta_0(\cdot)$ is a real-valued function. The conditional mean w.r.t. $X = X(\cdot)$ can be understood as a function of a collection of random variables $\{X(t): 0 \leq t \leq 1\}$ throughout the paper. Let $\beta \in H^m(\mathbb{I})$, the m-order Sobolev space defined by

$$\begin{aligned} H^m(\mathbb{I}) = \{\beta : \mathbb{I} \mapsto \mathbb{R} | \beta^{(j)}, j = 0, \ldots, m-1, \\ \text{are absolutely continuous, and } \beta^{(m)} \in L^2(\mathbb{I})\}. \end{aligned}$$

Therefore, the unknown parameter $\theta \equiv (\alpha, \beta)$ belongs to $\mathcal{H} \equiv \mathbb{R}^1 \times H^m(\mathbb{I})$. We further assume $m > 1/2$ such that $H^m(\mathbb{I})$ is a reproducing kernel Hilbert space.

In this paper, we consider a general loss function $\ell(y; a)$ defined over $y \in \mathcal{Y}$ and $a \in \mathbb{R}$, which covers two important classes of statistical models: (i) $\ell(y; a) = \log p(y; F(a))$, where $y|x \sim p(y; \mu_0(x))$ for a conditional distribution p; (ii) $\ell(y; a) = Q(y; F(a))$, where $Q(y; \mu) \equiv \int_y^\mu (y - s)/\mathcal{V}(s)\,ds$ is a quasi-likelihood with some known positive-valued function $\mathcal{V}$ satisfying $\mathcal{V}(\mu_0(X)) = \mathrm{Var}(Y|X)$; see [32]. Note that these two criterion functions coincide under some choices of $\mathcal{V}$. The regularized estimator is given by

$$\begin{aligned} &(\widehat{\alpha}_{n,\lambda}, \widehat{\beta}_{n,\lambda}) \\ &\quad = \arg \sup_{(\alpha,\beta)\in\mathcal{H}} \ell_{n,\lambda}(\theta) \\ &\quad \equiv \arg \sup_{(\alpha,\beta)\in\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell\left(Y_i; \alpha + \int_0^1 X_i(t)\beta(t)\,dt\right) - (\lambda/2) J(\beta, \beta) \right\}, \end{aligned} \tag{2.2}$$

where $J(\beta, \widetilde{\beta}) = \int_0^1 \beta^{(m)}(t)\widetilde{\beta}^{(m)}(t)\,dt$ is a roughness penalty. Here, we use $\lambda/2$ to simplify future expressions. In the special ℓ_2-regression, Yuan and Cai [35] study the minimax optimal estimation and prediction by assuming the same roughness penalty.

We next assume the following smoothness and tail conditions on ℓ. Denote the first-, second- and third-order derivatives of $\ell(y; a)$ w.r.t. a by $\dot{\ell}_a(y; a)$, $\ddot{\ell}_a(y; a)$ and $\ell_a'''(y; a)$, respectively.

ASSUMPTION A1. (a) $\ell(y; a)$ is three times continuously differentiable and strictly concave w.r.t a. There exist positive constants C_0 and C_1 s.t.,

$$\begin{aligned} &E\left\{\exp\left(\sup_{a\in\mathbb{R}} |\ddot{\ell}_a(Y; a)|/C_0\right)|X\right\} \leq C_1, \\ &E\left\{\exp\left(\sup_{a\in\mathbb{R}} |\ell_a'''(Y; a)|/C_0\right)|X\right\} \leq C_1, \qquad \text{a.s.} \end{aligned} \tag{2.3}$$

(b) There exists a positive constant C_2 s.t.,

$$C_2^{-1} \le B(X) \equiv -E\Big\{\ddot{\ell}_a\Big(Y;\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt\Big)\Big|X\Big\} \le C_2 \qquad \text{a.s.}$$

In addition, X is weighted-centered in the sense that $E\{B(X)X(t)\} = 0$ for any $t \in \mathbb{I}$.

(c) $\epsilon \equiv \dot{\ell}_a(Y;\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt)$ satisfies $E\{\epsilon|X\} = 0$ and $E\{\epsilon^2|X\} = B(X)$, a.s.

The weighted center condition in Assumption A1(b) is only used to simplify our technical arguments. Actually, it always holds after a simple data transformation; see the Supplementary Material [27], Section S.1. Next, we give three examples to illustrate the validity of Assumption A1.

EXAMPLE 2.1 (Gaussian model). In the functional linear models under Gaussian errors, that is, $Y = \alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt + v$ and $v|X \sim N(0,\sigma^2)$, we can easily verify Assumption A1 with $B(X) = \sigma^{-2}$ and $\epsilon = v/\sigma^2$ given that $E\{X(t)\} = 0$.

EXAMPLE 2.2 (Logistic model). In the logistic regression, we assume $P(Y = 1|X) = 1 - P(Y = 0|X) = \exp(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt)/(1 + \exp(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt))$. It is easy to see that $\ell(y;a) = ay - \log(1 + \exp(a))$ and $B(X) = \exp(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt)/(1 + \exp(\alpha_0 + \int_0^1 X(t)\beta_0(t)))^2 \le 1$. Assumption A1(a) follows from simple algebra. Assumption A1(b) follows from data transformation and the following L^2 bounded condition: $\int_0^1 X^2(t)\,dt \le c$ a.s. The latter condition implies that the range of $\mu_0(X)$ is finite, and thus $B(X)$ is bounded away from zero. Since $\epsilon = Y - \exp(X^T\theta_0 + g_0(Z))/(1 + \exp(X^T\theta_0 + g_0(Z)))$, Assumption A1(c) can be verified by direct calculations.

EXAMPLE 2.3 (Exponential family). Let (Y, X) follow the one-parameter exponential family

$$Y|X \sim \exp\Big\{Y\Big(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt\Big) + A(Y) - G\Big(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt\Big)\Big\},$$

where $A(\cdot)$ and $G(\cdot)$ are known, and $\dot{G} = F$ [recall that F is the link function satisfying (2.1)]. We assume that G has bounded second- and third-order derivatives, and $\ddot{G} \ge \delta$ for some constant $\delta > 0$; see similar conditions on page 738 of [20]. It is easy to see that $\ell(y;a) = ya + A(y) - G(a)$, and hence, $\dot{\ell}_a(y;a) = y - \dot{G}(a)$, $\ddot{\ell}_a(y;a) = -\ddot{G}(a)$ and $\ell_a'''(y;a) = -\dddot{G}(a)$. Clearly, $\ddot{\ell}_a$ and ℓ_a''' are both bounded, and hence Assumption A1(a) holds. Furthermore, $B(X) = \ddot{G}(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt)$ satisfies Assumption A1(b). Since $\epsilon = Y - \dot{G}(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt) = Y - \mu_0(X)$, it is easy to see that $E\{\epsilon|X\} = E\{Y|X\} - \mu_0(X) = 0$, and $E\{\epsilon^2|X\} = \mathrm{Var}(Y|X) = \ddot{G}(\alpha_0 + \int_0^1 X(t)\beta_0(t)\,dt)$ (see [19]), and therefore, Assumption A1(c) holds.

2.2. *Reproducing kernel Hilbert space.* We introduce an inner product in $H^m(\mathbb{I})$, that is,

$$\langle \beta, \widetilde{\beta}\rangle_1 = V(\beta, \widetilde{\beta}) + \lambda J(\beta, \widetilde{\beta}), \tag{2.4}$$

where $V(\beta, \widetilde{\beta}) \equiv \int_0^1 \int_0^1 C(s,t)\beta(t)\widetilde{\beta}(s)\,ds\,dt$ and $C(s,t) \equiv E\{B(X)X(t)X(s)\}$ is a weighted covariance function. Denote the corresponding norm as $\|\cdot\|_1$. Define a linear bounded operator $C(\cdot)$ from $L^2(\mathbb{I})$ to $L^2(\mathbb{I})$: $(C\beta)(t) = \int_0^1 C(s,t)\beta(s)\,ds$. Below we assume a regularity condition on $C\beta$, which implies the positive definiteness of V, such that the above inner product (2.4) is well defined.

ASSUMPTION A2. $C(s,t)$ is continuous on $\mathbb{I} \times \mathbb{I}$. Furthermore, for any $\beta \in L^2(\mathbb{I})$ satisfying $C\beta = 0$, we have $\beta = 0$.

Suppose that C is continuous over $\mathbb{I} \times \mathbb{I}$. By Mercer's theorem, C admits the spectral decomposition $C(s,t) = \sum_{\nu=1}^{\infty} \zeta_\nu \psi_\nu(s)\psi_\nu(t)$, where $\{\psi_\nu(\cdot), \zeta_\nu \geq 0\}_{\nu \geq 1}$ forms an orthonormal basis in $L^2(\mathbb{I})$ under the usual L^2-norm. Therefore, for any $\beta \in L^2(\mathbb{I})$, we have $\beta(\cdot) = \sum_{\nu=1}^{\infty} b_\nu \psi_\nu(\cdot)$ and $V(\beta, \beta) = \sum_{\nu=1}^{\infty} \zeta_\nu b_\nu^2$ for a sequence of square summable b_ν's. Assumption A2 directly implies that all the eigenvalues of C are positive, that is, $\zeta_\nu > 0$ for all $\nu \geq 1$. Therefore, if $V(\beta, \beta) = 0$, that is, $\sum_{\nu=1}^{\infty} \zeta_\nu b_\nu^2 = 0$, we can easily show that $\beta = \sum_{\nu=1}^{\infty} b_\nu \psi_\nu = 0$. Hence $\langle\cdot,\cdot\rangle_1$ is well defined. Moreover, together with Proposition 2 of [35], Assumption A2 implies that $H^m(\mathbb{I})$ is indeed a reproducing kernel Hilbert space (RKHS) under $\langle\cdot,\cdot\rangle_1$. We denote its reproducing kernel function as $K(s,t)$.

As for the joint parameter space $\mathcal{H}$, we also need to assume a proper inner product under which it is a well-defined Hilbert space. Define, for any $\theta = (\alpha, \beta)$, $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta}) \in \mathcal{H}$,

$$\begin{aligned}\langle \theta, \widetilde{\theta}\rangle \equiv E\biggl\{B(X)\biggl(\alpha + \int_0^1 X(t)\beta(t)\,dt\biggr)\biggl(\widetilde{\alpha} + \int_0^1 X(t)\widetilde{\beta}(t)\,dt\biggr)\biggr\} \\ + \lambda J(\widetilde{\beta}, \beta).\end{aligned} \tag{2.5}$$

By rewriting $\langle \theta, \widetilde{\theta}\rangle = E\{B(X)\}\alpha\widetilde{\alpha} + \langle \beta, \widetilde{\beta}\rangle_1$, we note that $\langle\cdot,\cdot\rangle$ is a well-defined inner product under Assumptions A1(b) and A2. The corresponding norm is denoted as $\|\cdot\|$. Given the above relation between $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle_1$, it is easy to show that $\mathcal{H}$ inherits the completeness of $H^m(\mathbb{I})$. This means $\mathcal{H}$ is indeed a Hilbert space as described in Proposition 2.1 below.

PROPOSITION 2.1. *Under $\langle\cdot,\cdot\rangle$, $\mathcal{H}$ is a Hilbert space.*

In the literature, the estimation/prediction rate results in the (generalized) functional linear models are mostly expressed in terms of L^2-norm; see [13, 14, 20, 35]. We remark that our norm $\|\cdot\|$ is stronger than the L^2-norm used in the above literature under Assumption A1(b).

We next assume a sequence of basis functions in $H^m(\mathbb{I})$ which can simultaneously diagonalize V and J. Let $\|\cdot\|_{L^2}$ and $\|\cdot\|_{\sup}$ denote the L^2 and supremum norms, respectively. Denote $a_n \asymp b_n$ if and only if there exist positive constants c_1, c_2 such that $c_1 \le a_\nu/b_\nu \le c_2$ for all ν.

ASSUMPTION A3. There exists a sequence of functions $\{\varphi_\nu\}_{\nu\ge 1} \subset H^m(\mathbb{I})$ such that $\|\varphi_\nu\|_{L^2} \le C_\varphi \nu^a$ for each $\nu \ge 1$, some constants $a \ge 0$, $C_\varphi > 0$ and

$$V(\varphi_\nu, \varphi_\mu) = \delta_{\nu\mu}, \qquad J(\varphi_\nu, \varphi_\mu) = \rho_\nu \delta_{\nu\mu} \qquad \text{for any } \nu, \mu \ge 1, \tag{2.6}$$

where $\delta_{\nu\mu}$ is Kronecker's notation, and ρ_ν is a nondecreasing nonnegative sequence satisfying $\rho_\nu \asymp \nu^{2k}$ for some constant $k > a + 1/2$. Furthermore, any $\beta \in H^m(\mathbb{I})$ admits the Fourier expansion $\beta = \sum_{\nu=1}^\infty V(\beta, \varphi_\nu)\varphi_\nu$ with convergence in $H^m(\mathbb{I})$ under $\langle\cdot,\cdot\rangle_1$.

We remark that Assumption A3 is the price we need to pay for making valid statistical inference in addition to those required for minimax estimation, as in, for example, [4, 35].

Assumption A3 can be directly implied by the pseudo Sacks–Ylvisaker (SY) conditions, which are slightly different from the conventional SY conditions proposed in [22–25]; see Section S.2 in the Supplementary Material [27] for more details. Proposition 2.2 below discusses the construction of an eigen-system satisfying Assumption A3 under this condition.

PROPOSITION 2.2 (Eigen-system construction). *Suppose the covariance function C satisfies Assumption* A2 *and the pseudo SY conditions of order $r \ge 0$ specified in Section* S.2. *Furthermore, the boundary value problem* (S.2) *in Section* S.3 *is regular in the sense of* [1]. *Consider the following integro-differential equations*:

$$\begin{cases} (-1)^m y_\nu^{(2m)}(t) = \rho_\nu \int_0^1 C(s,t) y_\nu(s)\, ds, \\ y_\nu^{(j)}(0) = y_\nu^{(j)}(1) = 0, \qquad j = m, \dots, 2m-1. \end{cases} \tag{2.7}$$

Let (ρ_ν, y_ν) be the corresponding eigenvalues and eigenfunctions of problem (2.7), *and let $\varphi_\nu = y_\nu/\sqrt{V(y_\nu, y_\nu)}$. Then (ρ_ν, φ_ν) satisfy Assumption* A3 *with $k = m + r + 1$ and $a = r + 1$ if one of the following additional assumptions is satisfied*:

(i) $r = 0$;
(ii) $r \ge 1$, *and for $j = 0, 1, \dots, r-1$, $C^{(j,0)}(0,t) = 0$ for any $0 \le t \le 1$, where $C^{(j,0)}(s,t)$ is the jth-order partial derivative with respect to s.*

The proof of Proposition 2.2 relies on a nontrivial application of the general integro-differential equation theory developed in [29–31]. In particular, the order of ρ_ν in problem (2.7) is, in general, equivalent to the order of eigenvalues in an ordinary differential problem; see [29], Theorem 7. More explicitly, $\rho_\nu \approx (c\pi\nu)^{2k}$ as $\nu \to \infty$ for some constant $c > 0$; see [31], equation (20).

In the Gaussian model with unit variance (see Example 2.1), it can be shown with arguments similar to those in [23–25] that the covariance function C satisfies the pseudo SY conditions of order $r = 0$ when $X(t)$ is Brownian motion with $C(s,t) = \min\{s,t\}$. We also note that the boundary condition in Proposition 2.2(ii) was also assumed in [22] when X is Gaussian of order $r > 0$. The integro-differential equations (2.7) can be translated into easily computable differential equations. More specifically, we rewrite $y_\nu(t)$ in (2.7) as $\ddot{g}_\nu(t)$, and thus obtain that

$$(2.8)\qquad \begin{cases} (-1)^{m+1} g_\nu^{(2m+2)}(t) = \rho_\nu g_\nu(t), \\ g_\nu^{(j)}(0) = g_\nu^{(j)}(1) = 0, \qquad j = m+2, \dots, 2m+1, \\ g_\nu(0) = \dot{g}_\nu(1) = 0. \end{cases}$$

Note that (2.7) and (2.8) share the same eigenvalues. Numerical examinations show that $\rho_\nu \approx (\pi\nu)^{2(m+1)}$. The function g_ν's have closed forms

$$g_\nu(t) = \mathrm{Re}\left(\sum_{j=1}^{2(m+1)} a_{\nu,j} \exp\bigl(\rho_\nu^{1/(2(m+1))} z_j t\bigr) \right), \qquad \nu = 1, 2, \dots,$$

where $\mathrm{Re}(\cdot)$ means the real part of a complex number, $z_1, \dots, z_{2(m+1)}$ are the complex (distinct) roots of $z^{2(m+1)} = (-1)^{m+1}$ and $a_{\nu,1}, \dots, a_{\nu,2(m+1)}$ are complex constant coefficients determined by the boundary value conditions in (2.8). It follows by Proposition 2.2 that the resultant ρ_ν and the corresponding scaled functions $\varphi_\nu = y_\nu / \sqrt{V(y_\nu, y_\nu)}$ satisfy Assumption A3, where recall that y_ν is the second-order derivative of g_ν.

In the logistic regression (Example 2.2) or exponential family models (Example 2.3), the approach given in Section S.5 (Supplementary Material [27]) can be used to find (ρ_ν, φ_ν) without verifying the pseudo SY conditions. To do so, we need to replace the kernel function C by its sample version $C_n(s,t) \equiv n^{-1} \sum_{i=1}^n \widehat{B}(X_i) X_i(s) X_i(t)$, where $\widehat{B}(X)$ is the plug-in estimate of $B(X)$.

Recall that K is the reproducing kernel function for $H^m(\mathbb{I})$ under $\langle \cdot, \cdot \rangle_1$. For any $t \in \mathbb{I}$, define $K_t(\cdot) = K(t, \cdot) \in H^m(\mathbb{I})$. Under Assumption A3, we may write $K_t = \sum_{\nu \ge 1} a_\nu \varphi_\nu$ for a real sequence a_ν. Clearly, $\varphi_\nu(t) = \langle K_t, \varphi_\nu \rangle_1 = a_\nu(1 + \lambda\rho_\nu)$, for all $\nu \ge 1$. So $K_t = \sum_{\nu \ge 1} \frac{\varphi_\nu(t)}{1+\lambda\rho_\nu} \varphi_\nu$. Define W_λ as an operator from $H^m(\mathbb{I})$ to $H^m(\mathbb{I})$ satisfying $\langle W_\lambda \beta, \widetilde{\beta} \rangle_1 = \lambda J(\beta, \widetilde{\beta})$, for all $\beta, \widetilde{\beta} \in H^m(\mathbb{I})$. Hence W_λ is linear, nonnegative definite and self-adjoint. For any $\nu \ge 1$, write $W_\lambda \varphi_\nu = \sum_\mu b_\mu \varphi_\mu$. Then by Assumption A3, for any $\mu \ge 1$, $\lambda\rho_\nu \delta_{\nu\mu} = \lambda J(\varphi_\nu, \varphi_\mu) = \langle W_\lambda \varphi_\nu, \varphi_\mu \rangle_1 = b_\mu(1 + \lambda\rho_\mu)$. Therefore, $b_\nu = \lambda\rho_\nu/(1 + \lambda\rho_\nu)$ and $b_\mu = 0$ if $\mu \ne \nu$, which implies $W_\lambda \varphi_\nu = \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu} \varphi_\nu$. Thus we have shown the following result.

PROPOSITION 2.3. *Suppose Assumption A3 holds. For any $t \in \mathbb{I}$,*

$$K_t(\cdot) = \sum_\nu \frac{\varphi_\nu(t)}{1+\lambda\rho_\nu}\varphi_\nu(\cdot),$$

and for any $\nu \geq 1$,

$$(W_\lambda\varphi_\nu)(\cdot) = \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\varphi_\nu(\cdot).$$

Propositions 2.4 and 2.5 below define two operators, R_x and P_λ, that will be used in the Fréchet derivatives of the criterion function $\ell_{n,\lambda}$. We first define $\tau(x)$ as follows. For any L^2 integrable function $x = x(t)$ and $\beta \in H^m(\mathbb{I})$, $\mathcal{L}_x(\beta) \equiv \int_0^1 x(t)\beta(t)\,dt$ defines a linear bounded functional. Then by the Riesz representation theorem, there exists an element in $H^m(\mathbb{I})$, denoted as $\tau(x)$, such that $\mathcal{L}_x(\beta) = \langle \tau(x), \beta\rangle_1$ for all $\beta \in H^m(\mathbb{I})$. If we denote $\tau(x) = \sum_{\nu=1}^\infty x_\nu^*\varphi_\nu$, then $x_\nu^*(1+\lambda\rho_\nu) = \langle \tau(x), \varphi_\nu\rangle_1 = \int_0^1 x(t)\varphi_\nu(t)\,dt \equiv x_\nu$ for any $\nu \geq 1$. Thus $\tau(x) = \sum_{\nu=1}^\infty \frac{x_\nu}{1+\lambda\rho_\nu}\varphi_\nu$.

PROPOSITION 2.4. *For any $x \in L^2(\mathbb{I})$, define $R_x = (E\{B(X)\}^{-1}, \tau(x))$. Then $R_x \in \mathcal{H}$ and $\langle R_x, \theta\rangle = \alpha + \int_0^1 x(t)\beta(t)\,dt$ for any $\theta = (\alpha, \beta) \in \mathcal{H}$.*

It should be noted that R_x depends on h according to the definition of $\tau(x)$.

PROPOSITION 2.5. *For any $\theta = (\alpha, \beta) \in \mathcal{H}$, define $P_\lambda\theta = (0, W_\lambda\beta)$. Then $P_\lambda\theta \in \mathcal{H}$ and $\langle P_\lambda\theta, \widetilde{\theta}\rangle = \langle W_\lambda\beta, \widetilde{\beta}\rangle_1$ for any $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta}) \in \mathcal{H}$.*

For notational convenience, denote $\Delta\theta = (\Delta\alpha, \Delta\beta)$ and $\Delta\theta_j = (\Delta\alpha_j, \Delta\beta_j)$ for $j = 1, 2, 3$. The Fréchet derivative of $\ell_{n,\lambda}(\theta)$ w.r.t. θ is given by

$$S_{n,\lambda}(\theta)\Delta\theta \equiv D\ell_{n,\lambda}(\theta)\Delta\theta = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_a\bigl(Y_i; \langle R_{X_i}, \theta\rangle\bigr)\langle R_{X_i}, \Delta\theta\rangle - \langle P_\lambda\theta, \Delta\theta\rangle.$$

The second- and third-order Fréchet derivatives of $\ell_{n,\lambda}(\theta)$ can be shown to be, respectively,

$$\begin{aligned} &DS_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2 \\ &\quad\equiv D^2\ell_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2 \\ &\quad= \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_a\bigl(Y_i; \langle R_{X_i}, \theta\rangle\bigr)\langle R_{X_i}, \Delta\theta_1\rangle\langle R_{X_i}, \Delta\theta_2\rangle - \langle P_\lambda\Delta\theta_1, \Delta\theta_2\rangle \end{aligned}$$

and

$$\begin{aligned} &D^2S_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2\Delta\theta_3 \\ &\quad\equiv D^3\ell_{n,\lambda}(\theta)\Delta\theta_1\Delta\theta_2\Delta\theta_3 \\ &\quad= \frac{1}{n}\sum_{i=1}^n \ell_a'''\bigl(Y_i; \langle R_{X_i}, \theta\rangle\bigr)\langle R_{X_i}, \Delta\theta_1\rangle\langle R_{X_i}, \Delta\theta_2\rangle\langle R_{X_i}, \Delta\theta_3\rangle. \end{aligned}$$

Define $S_n(\theta) = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_a(Y_i; \langle R_{X_i}, \theta\rangle) R_{X_i}$, $S(\theta) = E\{S_n(\theta)\}$ and $S_\lambda(\theta) = E\{S_{n,\lambda}(\theta)\}$ with expectations taken under the true model.

3. Bahadur representation for functional data.

In this section, we extend the functional Bahadur representation originally established in the nonparametric regression framework [26] to the generalized functional linear models. This new technical tool is fundamentally important in the sense that it provides a unified treatment for various inference problems.

Denote $h = \lambda^{1/(2k)}$, where k is specified in Assumption A3. An auxiliary norm is introduced for technical purpose: $\|\theta\|_2 = |\alpha| + \|\beta\|_{L^2}$ for any $\theta = (\alpha, \beta) \in \mathcal{H}$. The following result gives a useful relationship between the two norms $\|\cdot\|_2$ and $\|\cdot\|$. Recall that a is defined in Assumption A3.

LEMMA 3.1. *There exists a constant* $\kappa > 0$ *such that for any* $\theta \in \mathcal{H}$, $\|\theta\|_2 \le \kappa h^{-(2a+1)/2}\|\theta\|$.

To obtain an appropriate Bahadur representation for the functional data, we need the following regularity conditions on X. Recall that $\|X\|_{L^2}^2 = \int_0^1 X^2(t)\,dt$.

ASSUMPTION A4. There exists a constant $s \in (0, 1)$ such that

$$E\{\exp(s\|X\|_{L^2})\} < \infty. \tag{3.1}$$

Moreover, suppose that there exists a constant $M_0 > 0$ such that for any $\beta \in H^m(\mathbb{I})$,

$$E\left\{\left|\int_0^1 X(t)\beta(t)\,dt\right|^4\right\} \le M_0\left[E\left\{\left|\int_0^1 X(t)\beta(t)\,dt\right|^2\right\}\right]^2. \tag{3.2}$$

It is easy to see that (3.1) holds for any bounded stochastic process X, that is, $\|X\|_{L^2} \le c$ a.s. for some constant $c > 0$. This applies to Example 2.2 which usually requires X to be almost surely bounded in terms of L^2-norm. Equation (3.1) also holds for the Gaussian process as described in Proposition 3.2 below. The result applies to Examples 2.1 and 2.3 where X can be Gaussian.

PROPOSITION 3.2. *If X is a Gaussian process with square-integrable mean function, then* (3.1) *holds for any* $s \in (0, 1/4)$.

The fourth moment condition (3.2) is valid for $M_0 = 3$ when X is a Gaussian process; see [35] for more discussions. The following result shows that (3.2) actually holds in more general settings.

PROPOSITION 3.3. *Suppose* $X(t) = u(t) + \sum_{\nu=1}^\infty \xi_\nu\omega_\nu\psi_\nu(t)$, *where* $u(\cdot) \in L^2(\mathbb{I})$ *is nonrandom,* ψ_ν *is orthonormal* $L^2(\mathbb{I})$*-basis,* ω_ν *is a real square-summable sequence and* ξ_ν *are independent random variables drawn from some* *symmetric distribution with finite fourth-order moment. Then for any* $\beta(t) = \sum_{\nu=1}^{\infty} b_\nu \psi_\nu(t)$ *with* b_ν *being real square-summable*, (3.2) *holds with* $M_0 = \max\{E\{\xi_\nu^4\}/E\{\xi_\nu^2\}^2, 3\}$.

Lemma 3.4 below proves a concentration inequality as a preliminary step in obtaining the Bahadur representation. Denote $T = (Y, X) \in \mathcal{T}$ as the data variable. Let $\psi_n(T; \theta)$ be a function over $\mathcal{T} \times \mathcal{H}$, which might depend on n. Define

$$H_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\psi_n(T_i; \theta) R_{X_i} - E_T\{\psi_n(T; \theta) R_X\}],$$

where $E_T\{\cdot\}$ means the expectation w.r.t. T. Define $\mathcal{F}_{p_n} = \{\theta = (\alpha, \beta) \in \mathcal{H} : |\alpha| \leq 1, \|\beta\|_{L^2} \leq 1, J(\beta, \beta) \leq p_n\}$, where $p_n \geq 1$.

LEMMA 3.4. *Suppose Assumptions* A1 *to* A4 *hold. In addition*, $\psi_n(T_i; 0) = 0$ *a.s.*, *there exists a constant* $C_\psi > 0$ *s.t.*, *and the following Lipschitz continuity holds*:

$$|\psi_n(T; \theta) - \psi_n(T; \widetilde{\theta})| \leq C_\psi \|\theta - \widetilde{\theta}\|_2 \qquad \text{for any } \theta, \widetilde{\theta} \in \mathcal{F}_{p_n}. \tag{3.3}$$

Then as $n \to \infty$,

$$\sup_{\theta \in \mathcal{F}_{p_n}} \frac{\|H_n(\theta)\|}{p_n^{1/(4m)} \|\theta\|_2^\gamma + n^{-1/2}} = O_P((h^{-1} \log\log n)^{1/2}),$$

where $\gamma = 1 - 1/(2m)$.

Our last assumption is concerned with the convergence rate of $\widehat{\theta}_{n,\lambda}$. Define $r_n = (nh)^{-1/2} + h^k$. Recall that k is specified in Assumption A3.

ASSUMPTION A5. $\|\widehat{\theta}_{n,\lambda} - \theta_0\| = O_P(r_n)$.

Proposition 3.5 states that Assumptions A1 to A4 are actually sufficient to imply the above rate of convergence if the smoothing parameter is properly chosen. Note that no estimation consistency is required in Proposition 3.5.

PROPOSITION 3.5. *Suppose that Assumptions* A1 *to* A4 *hold*, *and that the following rate conditions on* h (*or equivalently*, λ) *are satisfied*:

$$\begin{aligned} h &= o(1), \\ n^{-1/2} h^{-1} &= o(1), \\ n^{-1/2} h^{-(a+1)-((2k-2a-1)/(4m))} (\log n)^2 (\log\log n)^{1/2} &= o(1). \end{aligned} \tag{3.4}$$

Then Assumption A5 *is satisfied.*

It follows by Proposition 3.3 that if $X(t) = u(t) + \sum_{\nu=1}^{\infty} \xi_\nu \omega_\nu \psi_\nu(t)$ with ξ_ν being independent random variables following symmetric distribution with bounded support, say, $[-N, N]$, then (3.2) holds. In this case, X is almost surely L^2 bounded since $\|X\|_{L^2} \leq \|u\|_{L^2} + \sqrt{\sum_\nu \xi_\nu^2 \omega_\nu^2} \leq \|u\|_{L^2} + N\sqrt{\sum_\nu \omega_\nu^2}$, a.s. Then Proposition 3.5 states that if Assumptions A1 to A4 hold and the smoothing parameter is tuned to satisfy (3.4), then Assumption A5 holds for the above L^2 bounded X.

Condition (3.4) is satisfied for a suitable range of h. To illustrate this point, we consider the following simple but representative case. Under the setup of Proposition 2.2, we have $a = r + 1$ and $k = m + r + 1$. Suppose $r = 0$, that is, X corresponds to zero-order covariance. Thus $a = 1$ and $k = m + 1$. Denote $h^* \asymp n^{-1/(2k+1)}$, $h^{**} \asymp n^{-2/(4k+1)}$ and $h^{***} \asymp n^{-1/(2k)}$. It can be shown that when $m > (3 + \sqrt{5})/4$, h^*, h^{**} and h^{***} all satisfy the conditions of (3.4). It should be mentioned that h^* yields the optimal estimation rate $n^{-k/(2k+1)}$ [35], h^{**} yields the optimal testing rate $n^{-2k/(4k+1)}$ as will be shown in later sections, and h^{***} yields the optimal prediction rate [3].

Now we are ready to present the Bahadur representation based on the functional data.

THEOREM 3.6 (Bahadur representation for functional data). *Suppose that Assumptions A1–A5 hold, and as $n \to \infty$, $h = o(1)$ and $\log(h^{-1}) = O(\log n)$. Furthermore, (3.2) holds. Then, as $n \to \infty$, $\|\widehat{\theta}_{n,\lambda} - \theta_0 - S_{n,\lambda}(\theta_0)\| = O_P(a_n)$, where*

$$a_n = n^{-1/2} h^{-(4ma+6m-1)/(4m)} r_n (\log n)^2 (\log\log n)^{1/2} + C_\ell h^{-1/2} r_n^2,$$

and

$$C_\ell \equiv \sup_{x \in L^2(\mathbb{I})} E\Big\{\sup_{a\in\mathbb{R}} |\ell_a'''(Y; a)| \big| X = x\Big\}.$$

We next give an example rate of a_n in Theorem 3.6 when ℓ is quadratic. In this case, we have $C_\ell = 0$. Suppose $a = 1$ and $k = m + 1$; see the discussions below Proposition 3.5. Direct examinations show that a_n is of the order $o(n^{-1/2})$ when $m > 1 + \sqrt{3}/2 \approx 1.866$ and $h = h^*$, h^{**} and h^{***}.

An immediate consequence of Bahadur representation is the following point-wise limit distribution of the slope function estimate. This local result is new and of independent interest, for example, point-wise CI.

COROLLARY 3.7. *Suppose that the conditions of Theorem 3.6 are satisfied, $\sup_{\nu\geq 1} \|\varphi_\nu\|_{\sup} \leq C_\varphi \nu^a$ for $\nu \geq 1$ and that $E\{\exp(s|\epsilon|)\} < \infty$, for some constant $s > 0$. Furthermore, as $n \to \infty$, $nh^{2a+1}(\log(1/h))^{-4} \to \infty$, $n^{1/2} a_n = o(1)$ and $\sum_{\nu=1}^{\infty} \frac{|\varphi_\nu(z)|^2}{(1+\lambda\rho_\nu)^2} \asymp h^{-(2a+1)}$. Then we have for any $z \in \mathbb{I}$,*

$$\frac{\sqrt{n}(\widehat{\beta}_{n,\lambda}(z) - \beta_0(z) + (W_\lambda \beta_0)(z))}{\sqrt{\sum_{\nu=1}^{\infty} (|\varphi_\nu(z)|^2/(1 + \lambda\rho_\nu)^2)}} \xrightarrow{d} N(0, 1).$$

In addition, if $\sqrt{n}(W_\lambda\beta_0)(z)/\sqrt{\sum_{\nu=1}^\infty \frac{|\varphi_\nu(z)|^2}{(1+\lambda\rho_\nu)^2}} = o(1)$, *then*

$$\frac{\sqrt{n}(\widehat{\beta}_{n,\lambda}(z) - \beta_0(z))}{\sqrt{\sum_{\nu=1}^\infty |\varphi_\nu(z)|^2/(1+\lambda\rho_\nu)^2}} \xrightarrow{d} N(0,1).$$

Corollary 3.7 applies to any point $z \in \mathbb{I}$ satisfying $\sum_{\nu=1}^\infty \frac{|\varphi_\nu(z)|^2}{(1+\lambda\rho_\nu)^2} \asymp h^{-(2a+1)}$. Validity of this condition is discussed in Section S.14 of the Supplementary Material [27].

The condition $\sqrt{n}(W_\lambda\beta_0)(z)/\sqrt{\sum_{\nu=1}^\infty \frac{|\varphi_\nu(z)|^2}{(1+\lambda\rho_\nu)^2}} = o(1)$ holds if $nh^{4k} = o(1)$ and the true slope function $\beta_0 = \sum_\nu b_\nu\varphi_\nu$ satisfies the condition (**U**): $\sum_\nu b_\nu^2\rho_\nu^2 < \infty$. To see this, observe that

$$\begin{aligned}\left|(W_\lambda\beta_0)(z)\right| &= \left|\sum_\nu b_\nu \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\varphi_\nu(z)\right| \\ &\le C_\varphi\lambda\sum_\nu |b_\nu|\frac{\rho_\nu \nu^a}{1+\lambda\rho_\nu} \\ &\le C_\varphi\lambda\sqrt{\sum_\nu b_\nu^2\rho_\nu^2}\sqrt{\sum_\nu \frac{\nu^{2a}}{(1+\lambda\rho_\nu)^2}},\end{aligned}$$

where the last term is of the order $O(\lambda h^{-(2a+1)/2})$. Hence, it leads to

$$\sqrt{n}(W_\lambda\beta_0)(z)\Big/\sqrt{\sum_{\nu=1}^\infty \frac{|\varphi_\nu(z)|^2}{(1+\lambda\rho_\nu)^2}} \asymp \sqrt{nh^{2a+1}}(W_\lambda\beta_0)(z) = o(1).$$

REMARK 3.1 (Convergence rate). Corollary 3.7 derives the convergence rate of the local estimate $\widehat{\beta}_{n,\lambda}(z)$ as $\sqrt{nh^{2a+1}}$. The factor a (defined in Assumption A3) generically reflects the impact of the covariance operator on the convergence rate. For example, Proposition 2.2 shows that $a = r+1$ with r being the order of the covariance function under the pseudo SY condition. The above observation coincides with the arguments in [13], that the covariance effect in general influences the (global) rate expressions. When the eigenfunctions are uniformly bounded, that is, $a = 0$, the above rate becomes $\sqrt{nh}$, which is exactly the rate derived in the general nonparametric regression setup; see Theorem 3.5 in [26].

REMARK 3.2. (Undersmoothing) Assumption A3 implies that $\beta_0 \in H^m(\mathbb{I})$ has the property that $\sum_\nu b_\nu^2\rho_\nu < \infty$. However, the condition (**U**) imposes a faster decay rate on the generalized Fourier coefficients b_ν, which in turn requires more smoothness of β_0. Since we still employ the mth order penalty in (2.2), condition (**U**) can be treated as a type of undersmoothing condition. More generally, similar conditions will be implicitly imposed in the inference procedures to be presented later.

4. Confidence/prediction interval.

In this section, we consider two interconnected inference procedures: (i) confidence interval for the conditional mean and (ii) prediction interval for a future response.

4.1. *Confidence interval for conditional mean.* For any (nonrandom) $x_0 \in L^2(\mathbb{I})$, we construct a confidence interval $\mu_0(x_0) = E\{Y|X = x_0\}$ by centering around the plug-in estimate $\widehat{Y}_0 \equiv F(\widehat{\alpha}_{n,\lambda} + \int_0^1 x_0(t)\widehat{\beta}_{n,\lambda}(t)\,dt)$. Define $\mu_0'(x_0) = \dot{F}(\alpha_0 + \int_0^1 x_0(t)\beta_0(t)\,dt)$, $\sigma_n^2 = E\{B(X)\}^{-1} + \sum_{\nu=1}^{\infty} \frac{|x_\nu^0|^2}{(1+\lambda\rho_\nu)^2}$, where $x_\nu^0 = \int_0^1 x_0(t)\varphi_\nu(t)\,dt$.

THEOREM 4.1 (Confidence interval construction). *Let Assumptions A1 through A5 be satisfied for the true parameter $\theta_0 = (\alpha_0, \beta_0)$, and $\mu_0'(x_0) \neq 0$. Furthermore, assume (3.2) and $E\{\exp(s|\epsilon|)\} < \infty$ for some $s > 0$. If $h = o(1)$, $\log(h^{-1}) = O(\log n)$, $nh^{2a+1}(\log n)^{-4} \to \infty$, $na_n^2 = o(1)$ and $\|R_{x_0}\| \asymp \sigma_n$, then as $n \to \infty$,*

$$\frac{\sqrt{n}}{\sigma_n}\biggl(\widehat{\alpha}_{n,\lambda} + \int_0^1 x_0(t)\widehat{\beta}_{n,\lambda}(t)\,dt - \alpha_0 - \int_0^1 x_0(t)\beta_0(t)\,dt - \int_0^1 x_0(t)(W_\lambda\beta_0)(t)\,dt\biggr) \xrightarrow{d} N(0,1).$$

Furthermore, if $\beta_0 = \sum_\nu b_\nu\varphi_\nu$ with $\sum_\nu b_\nu^2\rho_\nu^2 < \infty$ and $nh^{4k} = o(1)$, then $\frac{\sqrt{n}}{\sigma_n}\int_0^1 x_0(t)(W_\lambda\beta_0)(t)\,dt = o(1)$ so that we have

$$\frac{\sqrt{n}}{\sigma_n\mu_0'(x_0)}(\widehat{Y}_0 - \mu_0(x_0)) \xrightarrow{d} N(0,1). \tag{4.1}$$

Hence the $100(1-\widetilde{\alpha})\%$ confidence interval for $\mu_0(x_0)$ is

$$[\widehat{Y}_0 \pm n^{-1/2} z_{\widetilde{\alpha}/2}\sigma_n\widehat{\mu}_0'(x_0)], \tag{4.2}$$

where $z_{\widetilde{\alpha}/2}$ is the $(1-\widetilde{\alpha}/2)$-quantile of $N(0,1)$ and $\widehat{\mu}_0'(x_0) \equiv \dot{F}(\widehat{\alpha}_{n,\lambda} + \int_0^1 x_0(t) \times \widehat{\beta}_{n,\lambda}(t)\,dt)$.

In the Gaussian model (Example 2.1) with $B(X) \equiv 1$, if X is Brownian motion with $C(s,t) = \min\{s,t\}$, then σ_n^2 has an explicit form with $\rho_\nu \approx (2\pi\nu)^{2(m+1)}$ and φ_ν solved by (2.8). As for Examples 2.2 and 2.3, one can obtain σ_n^2 by following the approach outlined in Section S.5 (Supplementary Material [27]).

A direct byproduct of Theorem 4.1 is the prediction rate $\sigma_n/\sqrt{n}$. Proposition 4.2 further characterizes this rate in various situations. Suppose that $|x_\nu^0| \asymp \nu^{a-d}$ for some constant d. A larger d usually yields a smoother function x_0.

PROPOSITION 4.2. *The prediction rate in Theorem 4.1 satisfies*

$$\sigma_n/\sqrt{n} = \begin{cases} n^{-1/2}, & \text{if } d-a > 1/2, \\ n^{-1/2}\big(\log(1/h)\big)^{1/2}, & \text{if } d-a = 1/2, \\ n^{-1/2}h^{d-a-1/2}, & \text{if } d-a < 1/2. \end{cases}$$

In particular, if $d-a<1/2$ and $h=h^{***} \asymp n^{-1/(2(m+a))}$, then $\sigma_n/\sqrt{n} = n^{-(d+m-1/2)/(2(m+a))}$. Furthermore, if $k=m+a$ as in the setting of Proposition 2.2, then $\sigma_n/\sqrt{n}$ is minimax optimal when $h=h^{***}$.

Proposition 4.2 states that when $d-a>1/2$, that is, the process x_0 is sufficiently smooth, then the prediction can be conducted in terms of root-n rate regardless of the choice of h. This result coincides with [3] in the special FPCA setting. Moreover, when $d-a<1/2$ and $h=h^{***}$, the rate becomes optimal. Again, this is consistent with [3] in the setting that the true slope function belongs to a Sobolev rectangle. Interestingly, it can be checked that $h=h^{***}$ satisfies the rate conditions in Theorem 4.1 if $a=1$, $k=m+1$ and $m>1+\sqrt{3}/2$; see the discussions below Theorem 3.6.

4.2. *Prediction interval for future response.* Following Theorem 4.1, we can establish the prediction interval for the future response Y_0 conditional on $X=x_0$. Write $Y_0-\widehat{Y}_0=\xi_n+\epsilon_0$, where $\xi_n=\mu_0(x_0)-\widehat{Y}_0$ and $\epsilon_0=\dot{\ell}_a(Y_0;\alpha_0+\int_0^1 x_0(t)\beta_0(t)\,dt)$. Since ϵ_0 is independent of ξ_n depending on all the past data $\{Y_i,X_i\}_{i=1}^n$, we can easily incorporate the additional randomness from ϵ_0 into the construction of the prediction interval. This leads to a nonvanishing interval length as sample size increases. This is crucially different from that of confidence interval.

Let F_{ξ_n} and F_{ϵ_0} be the distribution functions of ξ_n and ϵ_0, respectively. Denote the distribution function of $\xi_n+\epsilon_0$ as $G\equiv F_{\xi_n}*F_{\epsilon_0}$, and $(l_{\widetilde{\alpha}},u_{\widetilde{\alpha}})$ as its $(\widetilde{\alpha}/2)$th and $(1-\widetilde{\alpha}/2)$th quantiles, respectively. Then the $100(1-\widetilde{\alpha})\%$ prediction interval for Y_0 is given as

$$[\widehat{Y}_0+l_{\widetilde{\alpha}},\widehat{Y}_0+u_{\widetilde{\alpha}}].$$

Theorem 4.1 directly implies that $\xi_n \overset{a}{\sim} N(0,(n^{-1/2}\sigma_n\widehat{\mu}_0'(x_0))^2)$, where $\overset{a}{\sim}$ means *approximately distributed*. If we further assume that $\epsilon_0\sim N(0,B^{-1}(x_0))$ [see Assumption A1(c)], that is, $B(x_0)$ is the reciprocal error variance for the L^2 loss, the above general formula reduces to

$$\Big[\widehat{Y}_0 \pm z_{\widetilde{\alpha}/2}\sqrt{B(x_0)+\big(n^{-1/2}\sigma_n\widehat{\mu}_0'(x_0)\big)^2}\Big]. \tag{4.3}$$

The unknown quantities in (4.2) and (4.3) can be estimated by plug-in approach.

5. Hypothesis testing.

We consider two types of testing for the generalized functional linear models: (i) testing the *functional contrast* defined as $\int_0^1 w(t)\beta(t)\,dt$ for some given weight function $w(\cdot)$, for example, $w=X$ and (ii) testing the intercept value and the global behavior of the slope function, for example, $\alpha=0$ and β is a linear function.

5.1. *Testing functional contrast.* In practice, it is often of interest to test the *functional contrast*. For example, we may test single frequency or frequency contrast of the slope function; see Examples 5.1 and 5.2. In general, we test $H_0^{CT}: \int_0^1 w(t)\beta(t)\,dt = c$ for some known $w(\cdot)$ and c.

Consider the following test statistic:

$$CT_{n,\lambda} = \frac{\sqrt{n}(\int_0^1 w(t)\widehat{\beta}_{n,\lambda}(t)\,dt - c)}{\sqrt{\sum_{\nu=1}^{\infty}(w_\nu^2/(1+\lambda\rho_\nu)^2)}}, \tag{5.1}$$

where $w_\nu = \int_0^1 w(t)\varphi_\nu(t)\,dt$. Recall that (φ_ν, ρ_ν) is the eigensystem satisfying Assumption A3. Let $w \in L^2(\mathbb{I})$ and $\tau(w) \in H^m(\mathbb{I})$ be such that $\langle \tau(w), \beta\rangle_1 = \int_0^1 w(t)\beta(t)\,dt$, for any $\beta \in H^m(\mathbb{I})$. We can verify that $\tau(w) = \sum_{\nu=1}^{\infty} \frac{w_\nu}{1+\lambda\rho_\nu}\varphi_\nu$. Then, under H_0^{CT}, $CT_{n,\lambda}$ can be rewritten as

$$\frac{\sqrt{n}\langle \tau(w), (\widehat{\beta}_{n,\lambda} - \beta)\rangle_1}{\|\tau(w)\|_1}. \tag{5.2}$$

It follows from Theorem 4.1 that (5.2) converges weakly to a standard normal distribution. This is summarized in the following theorem. Define $M_a = \sum_{\nu=1}^{\infty} \frac{w_\nu^2}{(1+\lambda\rho_\nu)^a}$ for $a = 1, 2$.

THEOREM 5.1 (Functional contrast testing). *Suppose that Assumptions A1 through A5 hold. Furthermore, let $\beta_0 = \sum_\nu b_\nu\varphi_\nu$ with $\sum_\nu b_\nu^2\rho_\nu^2 < \infty$, and assume (3.2), $E\{\exp(s|\epsilon|)\} < \infty$ for some $s > 0$, and as $n \to \infty$, $h = o(1)$, $\log(h^{-1}) = O(\log n)$, $nh^{2a+1}(\log(1/h))^{-4} \to \infty$, $M_1 \asymp M_2$, $na_n^2 = o(1)$, $nh^{4k} = o(1)$. Then, under H_0^{CT}, we have $CT_{n,\lambda} \xrightarrow{d} N(0,1)$ as $n \to \infty$.*

EXAMPLE 5.1 (Testing single frequency). Suppose that the slope function has an expansion $\beta = \sum_{\nu=1}^{\infty} b_\nu\varphi_\nu$, and we want to test whether $b_{\nu^*} = 0$ for some $\nu^* \geq 1$. In other words, we are interested in knowing whether the ν^*-level frequency of β vanishes. Let $w(t) = (C\varphi_{\nu^*})(t)$. Then it is easy to see that $\int_0^1 w(t) \times \beta(t)\,dt = b_{\nu^*}$. That is, the problem reduces to testing $H_0^{CT}: \int_0^1 w(t)\beta(t)\,dt = 0$. It can be shown directly that $M_a = (1+\lambda\rho_{\nu^*})^{-a} \asymp 1$ for $a = 1, 2$. If $r = 0$ (see Proposition 3.2 for validity), then it can be shown that when $m > (3+\sqrt{5})/4 \approx 1.309$, the rate conditions in Theorem 5.1 are satisfied for $h = h^*$. This means that H_0^{CT} is rejected at level 0.05 if $|CT_{n,\lambda}| > 1.96$.

EXAMPLE 5.2 (Testing frequency contrast). Following Example 5.1, we now test whether $\sum_{\nu=1}^{\infty} \mathfrak{c}_\nu b_\nu = 0$, for some real sequence $\mathfrak{c}_\nu$ satisfying $0 < \inf_{\nu\geq 1}|\mathfrak{c}_\nu| \leq \sup_{\nu\geq 1}|\mathfrak{c}_\nu| < \infty$. Suppose that the covariance function $C(\cdot,\cdot)$ satisfies the conditions in Proposition 2.2 with order $r > 0$. It follows from Proposition 2.2 and its proof that the eigenfunction φ_ν can be managed so that $\|\varphi_\nu^{(2m)}\|_{\sup} \leq C_\varphi \nu^{2m+r+1}$

and $\|C\varphi_\nu\|_{\sup} \le \rho_\nu^{-1}\|\varphi_\nu^{(2m)}\|_{\sup} \asymp \nu^{-(r+1)}$, for all $\nu \ge 1$. So the function $w(t) = \sum_{\nu=1}^\infty \mathfrak{c}_\nu (C\varphi_\nu)(t)$ is well defined since the series is absolutely convergent on [0, 1]. It is easy to see that $\int_0^1 w(t)\beta(t)\,dt = \sum_\nu \mathfrak{c}_\nu b_\nu$ and $\mathfrak{c}_\nu = \int_0^1 w(t)\varphi_\nu(t)\,dt = w_\nu$. So the problem reduces to testing $H_0^{CT} : \int_0^1 w(t)\beta(t)\,dt = 0$. For $a = 1, 2$

$$M_a = \sum_{\nu=1}^\infty \frac{w_\nu^2}{(1+\lambda\rho_\nu)^a} = \sum_{\nu=1}^\infty \frac{\mathfrak{c}_\nu^2}{(1+\lambda\rho_\nu)^a} \asymp \sum_{\nu=1}^\infty \frac{1}{(1+\lambda\rho_\nu)^a} \asymp h^{-1}.$$

It can be shown that when $m > (3+\sqrt{5})$, the rate conditions in Theorem 5.1 are satisfied for $h = h^*$. We reject H_0^{CT} at level 0.05 if $|CT_{n,\lambda}| > 1.96$.

5.2. *Likelihood ratio testing*. Consider the following simple hypothesis:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \in \mathcal{H} - \{\theta_0\}, \tag{5.3}$$

where $\theta_0 \in \mathcal{H}$. The penalized likelihood ratio test statistic is defined as

$$\mathrm{PLRT} = \ell_{n,\lambda}(\theta_0) - \ell_{n,\lambda}(\widehat{\theta}_{n,\lambda}). \tag{5.4}$$

Recall that $\widehat{\theta}_{n,\lambda}$ is the maximizer of $\ell_{n,\lambda}(\theta)$ over $\mathcal{H}$. The proposed likelihood ratio testing also applies to the composite hypothesis; that is, θ belongs to a certain class. See Remark 5.1 for more details.

Theorem 5.2 below derives the null limiting distribution of $\mathrm{PLRT}_{n,\lambda}$.

THEOREM 5.2 (Likelihood ratio testing). *Suppose that H_0 holds, and Assumptions A1 through A5 are satisfied for the hypothesized value θ_0. Let h satisfy the following rate conditions: as $n \to \infty$, $nh^{2k+1} = O(1)$, $nh \to \infty$, $n^{1/2}a_n = o(1)$, $nr_n^3 = o(1)$, $n^{1/2}h^{-(a+1/2+(2k-2a-1)/(4m))}r_n^2(\log n)^2(\log\log n)^{1/2} = o(1)$ and $n^{1/2}h^{-(2a+1+(2k-2a-1)/(4m))} \times r_n^3(\log n)^3(\log\log n)^{1/2} = o(1)$. Furthermore, there exists a constant $M_4 > 0$ s.t. $E\{\epsilon^4|X\} \le M_4$, a.s. Then as $n \to \infty$,*

$$-(2u_n)^{-1/2}\bigl(2n\sigma^2 \cdot \mathrm{PLRT} + u_n + n\sigma^2\|W_\lambda\beta_0\|_1^2\bigr) \xrightarrow{d} N(0,1), \tag{5.5}$$

where $u_n = h^{-1}\sigma_1^4/\sigma_2^2$, $\sigma^2 = \sigma_1^2/\sigma_2^2$ and $\sigma_l^2 = h\sum_\nu (1+\lambda\rho_\nu)^{-l}$ for $l = 1, 2$.

By carefully examining the proof of Theorem 5.2, it can be shown that $n\|W_\lambda\beta_0\|_1^2 = o(n\lambda) = o(u_n)$. Therefore, $-2n\sigma^2 \cdot \mathrm{PLRT}$ is asymptotically $N(u_n, 2u_n)$ which is nearly $\chi^2_{u_n}$ as $n \to \infty$. Hence we claim the null limit distribution as being approximately $\chi^2_{u_n}$, denoted as

$$-2n\sigma^2 \cdot \mathrm{PLRT} \overset{a}{\sim} \chi^2_{u_n}, \tag{5.6}$$

where $\overset{a}{\sim}$ means *approximately distributed*; see [11]. If C satisfies the conditions of Proposition 2.2 with order $r \ge 0$, then $\rho_\nu \approx (c\nu)^{2k}$ (see the comments below Proposition 2.2), where $k = m + r + 1$ and $c > 0$ is constant. It is easy to see that $\sigma_l^2 \approx c^{-1} \int_0^\infty (1 + x^{2k})^{-l}\, dx$ for $l = 1, 2$. In Example 2.1, since the covariance function is free of the model parameters α_0, β_0, we can see that c is also free of the model parameters. In particular, when $B(X) \equiv 1$, $m = 2$ and $X(t)$ is Brownian motion with $C(s,t) = \min\{s,t\}$ and $r = 0$, we have $k = 3$ and $c = \pi$ [by solving (2.8)]. This yields $\sigma_l^2 \approx 0.2876697, 0.2662496$ for $l = 1, 2$, respectively. In the end, we obtain $\sigma^2 \approx 1.080451$ and $u_n \approx 0.3108129/h$ in (5.6). As seen above, the null limiting distribution has the nice property that it is free of the unknown model parameters, that is, so-called Wilks phenomenon [11, 33]. Hence we unveil a new version of Wilks phenomenon that applies to the functional data. This Wilks type of result enables us to simulate the null limit distribution directly without resorting to bootstrap or other resampling methods.

The quantities σ_1^2, σ_2^2 in Theorem 5.2 depend on the population eigenpairs. However, it is possible to replace these quantities by suitable estimators so that the results become more applicable. In Section S.18 (Supplementary Material [27]), we discuss the validity of this "plug-in" approach for Theorems 4.1, 5.1 and 5.2.

REMARK 5.1 (Composite hypothesis). By examining the proof of Theorem 5.2, we find that the null limiting distribution derived therein remains the same even when the hypothesized value θ_0 is unknown. An important consequence is that the proposed likelihood ratio approach can also be used to test a composite hypothesis such as $H_0 : \alpha = \alpha_0$ and $\beta \in \mathcal{P}_j$, where $\mathcal{P}_j$ represents the class of the jth-order polynomials. Under H_0, β is of the form $\beta(t) = \sum_{l=0}^{j} b_l t^l$ for some unknown vector $\mathbf{b} = (b_0, b_1, \ldots, b_j)^T$. In this case, the slope function and intercept can be estimated through the following "parametric" optimization:

$$(5.7)\quad \begin{aligned}(\widehat{\alpha}^0, \widehat{\mathbf{b}}^0) = \arg \max_{\alpha, b_0, \ldots, b_j \in \mathbb{R}} & \; n^{-1} \sum_{i=1}^{n} \ell\left(Y_i; \alpha + \sum_{l=0}^{j} b_l \int_0^1 X_i(t) t^l\, dt\right) \\ & - (\lambda/2) \mathbf{b}^T D \mathbf{b},\end{aligned}$$

where $D = [D_{l_1 L^2}]_{l_1, l_2 = 0, \ldots, j}$ is a $(j+1) \times (j+1)$ matrix with $D_{l_1 l_2} = J(t^{l_1}, t^{l_2})$. The corresponding slope function estimate is $\widehat{\beta}^0(t) = \sum_{l=0}^{j} \widehat{b}_l^0 t^l$. The test statistic for this composite hypothesis is defined as $\mathrm{PLRT} = \ell_{n,\lambda}(\widehat{\alpha}^0, \widehat{\beta}^0) - \ell_{n,\lambda}(\widehat{\alpha}_{n,\lambda}, \widehat{\beta}_{n,\lambda})$. Let $\theta_0 = (\alpha_0, \beta_0)$ be the unknown true model parameter under H_0, where β_0 can be represented as $\sum_{l=0}^{j} b_l^0 t^l$. Hence, we can further decompose the above PLRT as $\mathrm{PLRT}_1 - \mathrm{PLRT}_2$, where $\mathrm{PLRT}_1 = \ell_{n,\lambda}(\theta_0) - \ell_{n,\lambda}(\widehat{\alpha}_{n,\lambda}, \widehat{\beta}_{n,\lambda})$, $\mathrm{PLRT}_2 = \ell_{n,\lambda}(\theta_0) - \ell_{n,\lambda}(\widehat{\alpha}^0, \widehat{\beta}^0)$. Note that PLRT_1 is the test statistic for the simple hypothesis $\theta = \theta_0$ versus $\theta \neq \theta_0$, and PLRT_2 for the parametric hypothesis $(\alpha, \mathbf{b}) = (\alpha_0, \mathbf{b}^0)$ versus $(\alpha, \mathbf{b}) \neq (\alpha_0, \mathbf{b}^0)$, where $\mathbf{b}^0 = (b_0^0, \ldots, b_j^0)^T$. Conventional theory on parametric likelihood ratio testing leads to $-2n \cdot \mathrm{PLRT}_2 = O_P(1)$. On the other hand, Theorem 5.2 shows that $-2n\sigma^2 \cdot \mathrm{PLRT}_1 \overset{a}{\sim} \chi_{u_n}^2$. Therefore, we conclude that the null limit distribution for testing the composite hypothesis also follows $\chi_{u_n}^2$.

In the end of this section, we show that the proposed PLRT is optimal in the minimax sense [16] when $h = h^{**}$. To derive the minimax rate of testing (also called as minimum separation rate), we consider a local alternative written as $H_{1n}: \theta = \theta_{n0}$, where the alternative value is assumed to deviate from the null value by an amount of θ_n, that is, $\theta_{n0} = \theta_0 + \theta_n$. For simplicity, we assume $\theta_0 = 0$, and thus $\theta_{n0} = \theta_n$. Define the alternative value set $\theta_n \in \Theta_b \equiv \{(\alpha, \beta) \in \mathcal{H}: |\alpha| \le b, \|\beta\|_{L^2} \le b, J(\beta, \beta) \le b\}$ for some fixed constant $b > 0$.

THEOREM 5.3. *Let Assumptions A1–A5 be satisfied uniformly under $\theta = \theta_{n0} \in \Theta_b$. Let h satisfy $nh^{3/2} \to \infty$, as $n \to \infty$, and also the rate conditions specified in Theorem 5.2. Furthermore*, $\inf_{y \in \mathcal{Y}, a \in \mathbb{R}}(-\ddot{\ell}_a(y; a)) > 0$, *and there is a constant $M_4 > 0$ s.t. for $\theta_n = (\alpha_n, \beta_n) \in \Theta_b$, $\epsilon_n \equiv \dot{\ell}_a(Y; \alpha_n + \int_0^1 X(t)\beta_n(t)\,dt)$ satisfies $E\{\epsilon_n^4 | X\} \le M_4$, a.s. Then for any $\varepsilon > 0$, there exist positive constants N_ε and C_ε s.t. when $n \ge N_\varepsilon$,*

$$\inf_{\theta_n \in \Theta_b : \|\theta_n\| \ge C_\varepsilon \eta_n} P_{\theta_n}(\textit{reject } H_0) \ge 1 - \varepsilon,$$

where $\eta_n \asymp \sqrt{(nh^{1/2})^{-1} + \lambda}$.

The model assumption $\inf_{y \in \mathcal{Y}, a \in \mathbb{R}}(-\ddot{\ell}_a(y; a)) > 0$ trivially holds for Gaussian regression and exponential family considered in Examples 2.1 and 2.3. As for the logistic model with L^2 bounded X, this condition can be replaced by $\inf_{y \in \mathcal{Y}, a \in \mathcal{I}}(-\ddot{\ell}_a(y; a)) > 0$ under which the same conclusion as in Theorem 5.3 holds, where $\mathcal{I}$ is some bounded open interval including the range of $\langle R_X, \theta_{n0}\rangle$ for every $\theta_{n0} \in \Theta_b$.

Theorem 5.3 states that the PLRT is able to detect any local alternative with separation rate no faster than η_n. In particular, the minimum separation rate, that is, $n^{-2k/(4k+1)}$, is achieved when $h = h^{**}$. Note that h^{**} satisfies the rate conditions required by Theorem 5.3. For example, when $k = m + r + 1$, $a = r + 1$, $r = 0$ (see the discussions below Proposition 3.5), we can verify this fact for $m > (7 + \sqrt{33})/8 \approx 1.593$ by direct calculations. In the specific ℓ_2 regression, Corollary 4.6 of [15] proves that the above minimax rate, that is, $n^{-2k/(4k+1)}$, is optimal but under the perfect alignment condition. Therefore, we prove that the proposed PLRT can achieve the minimax optimality under more general settings.

The likelihood ratio testing procedure developed in this section requires prior knowledge on the smoothness of the true slope function and covariance kernel function, which might not be available in practice. This motivates us to propose two adaptive testing procedures in the next section.

6. Adaptive testing construction. In this section, we develop two adaptive testing procedures for $H_0: \beta = \beta_0$ without knowing m and r, that is, the smoothness of the true slope function and covariance kernel function. One works for Gaussian errors, and another works for sub-Gaussian errors. The test statistics for both cases are maximizers over a sequence of standardized PLRTs. We derive the null limit distribution as an extreme value distribution using Stein's method [5, 28]. Their minimax properties will also be carefully studied. To the best of our knowledge, our adaptive testing procedures are the first ones developed in the roughness regularization framework, which forms an interesting complement to those based on FPCA techniques [15, 17].

In this section, we focus on the ℓ_2 regression with two types of error: Gaussian error (Section 6.1) and sub-Gaussian error (Section 6.2). For simplicity, we assume $\beta_0 = 0$, $\alpha = 0$, and the errors to be of unit standard deviations. In addition, we assume that the covariate process $X(t)$ has zero mean and is independent of the error term. We remark that it is possible to extend our results in this section to the general loss functions, but with extremely tedious technical arguments.

Our test statistic is built upon a modified estimator $\widetilde{\beta}_{n,\lambda}$ that is constructed in the following three steps. The first step is to find a sequence of empirical eigenfunctions $\widehat{\varphi}_\nu$ that satisfy $\widehat{V}(\widehat{\varphi}_\nu, \widehat{\varphi}_\mu) = \delta_{\nu\mu}$ for all $\nu, \mu \geq 1$, where $\widehat{V}(\beta, \widetilde{\beta}) = \int_0^1 \int_0^1 \widehat{C}(s,t)\beta(s)\widetilde{\beta}(t)\,ds\,dt$ and $\widehat{C}(s,t) = n^{-1}\sum_{i=1}^n X_i(s)X_i(t)$. We offer two methods for finding $\widehat{\varphi}_\nu$. The first method conducts a spectral decomposition, $\widehat{C}(s,t) = \sum_{\nu=1}^\infty \widehat{\zeta}_\nu \widehat{\psi}_\nu(s)\widehat{\psi}_\nu(t)$, with some nonincreasing positive sequence $\widehat{\zeta}_\nu$ and orthonormal functions $\widehat{\psi}_\nu$ in the usual L^2-norm. Construct $\widehat{\varphi}_\nu = \widehat{\psi}_\nu / \sqrt{\widehat{\zeta}_\nu}$. This method is easy to implement, but implicitly assumes the perfect alignment condition. Our second method is more generally applicable, but requires more tedious implementation. Specifically, we apply similar construction techniques as in Section S.5 (Supplementary Material [27]) by using the sample versions of $\widetilde{K}$, C and T therein. In particular, we choose $m = 1, 2$ such that the true slope function is more possible to be covered.

The second step is to define a data-dependent parameter space. Note that $H^m(\mathbb{I})$ can be alternatively defined as $\{\sum_{\nu=1}^\infty b_\nu \varphi_\nu : \sum_{\nu=1}^\infty b_\nu^2 \nu^{2k} < \infty\}$, where k depends on m in an implicit manner. An approximate parameter space is $\mathcal{B}_k = \{\sum_{\nu=1}^\infty b_\nu \widehat{\varphi}_\nu : \sum_{\nu=1}^\infty b_\nu^2 \nu^{2k} < \infty\}$. The consistency of the sample eigenfunctions implies that $\mathcal{B}_k$ is a reasonable approximation of $H^m(\mathbb{I})$; see [14]. The data-dependent parameter space is thus defined as

$$\mathcal{B}_{kn} \equiv \left\{ \sum_{\nu=1}^n b_\nu \widehat{\varphi}_\nu \,\middle|\, b_1, \ldots, b_n \in \mathbb{R} \right\}.$$

In $\mathcal{B}_{kn}$, we can actually use the first $K_n \to \infty$ ($K_n \ll n$) eigenfunctions as the basis. However, such a general choice would give rise to unnecessary tuning of K_n in practice.

In the last step, we obtain the desirable estimator as $\widetilde{\beta}_{n,\lambda} = \arg\sup_{\beta \in \mathcal{B}_{kn}} \ell_{n,\lambda}(\beta)$, where

$$\ell_{n,\lambda}(\beta) = -\frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{\nu=1}^n b_\nu \omega_{i\nu} \right)^2 \Big/ 2 - (\lambda/2)\sum_{\nu=1}^n b_\nu^2 \nu^{2k}, \tag{6.1}$$

and $\omega_{i\nu} = \int_0^1 X_i(t)\widehat{\varphi}_\nu(t)\,dt$ for $i = 1, \ldots, n$ and $\nu \geq 1$. The smoothing parameter λ depends on both n and k, denoted as λ_k. In particular, we choose λ_k as $c_0^{2k} n^{-4k/(4k+1)}(\log\log n)^{2k/(4k+1)}$ for some constant $c_0 > 0$ irrelevant to k. As will be seen in later theorems, this choice yields the minimax optimality of the adaptive testing. Define $Y = (Y_1, \ldots, Y_n)^T$, $b = (b_1, \ldots, b_n)^T$, $\Lambda_k = \operatorname{diag}(1^{2k}, 2^{2k}, \ldots, n^{2k})$, $\Omega_i = (\omega_{i1}, \ldots, \omega_{in})$ and $\Omega = (\Omega_1^T, \ldots, \Omega_n^T)^T$. Hence we can rewrite $-2n\ell_{n,\lambda}(\beta)$ as

$$(Y - \Omega b)^T(Y - \Omega b) + n\lambda_k b^T \Lambda_k b,$$

whose minimizer [equivalently, the maximizer of $\ell_{n,\lambda}(\beta)$] is $\widehat{b}_{n,k} = (\Omega^T\Omega + n\lambda_k\Lambda_k)^{-1}\Omega^T Y$. Note that $\Omega^T\Omega = nI_n$ by the following analysis: for any $\nu, \mu \geq 1$,

$$\begin{aligned}\sum_{i=1}^n \omega_{i\nu}\omega_{i\mu} &= \sum_{i=1}^n \int_0^1 X_i(t)\widehat{\varphi}_\nu(t)\,dt \int_0^1 X_i(s)\widehat{\varphi}_\mu(s)\,ds \\ &= \int_0^1\int_0^1 \sum_{i=1}^n X_i(s)X_i(t)\widehat{\varphi}_\nu(s)\widehat{\varphi}_\mu(t)\,ds\,dt \\ &= n\int_0^1\int_0^1 \widehat{C}(s,t)\widehat{\varphi}_\nu(s)\widehat{\varphi}_\mu(t)\,ds\,dt = n\delta_{\nu\mu}.\end{aligned}$$

Therefore, $\widehat{b}_{n,k} = (nI_n + n\lambda_k\Lambda_k)^{-1}\Omega^T Y$ and $\widetilde{\beta}_{n,\lambda} = (\widehat{\varphi}_1, \ldots, \widehat{\varphi}_n)\widehat{b}_{n,k}$.

In the above analysis, we implicitly assume k to be known. However, the value of k is usually unavailable in practice. To resolve this issue, we will conduct our testing procedure over a sequence of integer k, that is, $\{1, 2, \ldots, k_n\}$, as will be seen in the next two subsections. The full adaptivity of testing procedure is achieved when we allow $k_n \to \infty$ so that the unknown k can eventually be captured by this sequence.

6.1. *Gaussian error.* In this subsection, we denote the PLRT as $\mathrm{PLRT}_k \equiv \ell_{n,\lambda}(0) - \ell_{n,\lambda}(\widetilde{\beta}_{n,\lambda})$ due to its dependence on k. By plugging in the above form of $\widetilde{\beta}_{n,\lambda}$, we obtain

$$\mathrm{PLRT}_k = -\frac{1}{2n} Y^T\Omega(nI_n + n\lambda_k\Lambda_k)^{-1}\Omega^T Y. \tag{6.2}$$

We next derive a standardized version of PLRT_k under H_0. Define $d_\nu(k) = 1/(1 + \lambda_k\rho_\nu(k))$, where $\rho_\nu(k) = \nu^{2k}$, for any $\nu, k \geq 1$. Under H_0, we have $Y = \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, and thus $-2n\mathrm{PLRT}_k = \sum_{\nu=1}^n d_\nu(k)\eta_\nu^2$ for $\eta_1, \ldots, \eta_n \overset{\text{i.i.d.}}{\sim} N(0,1)$ by straightforward calculation. Hence, we have $E\{-2n\mathrm{PLRT}_k\} = \sum_{\nu=1}^n d_\nu(k)$ and $\operatorname{Var}(-2n\mathrm{PLRT}_k) = 2\sum_{\nu=1}^n d_\nu^2(k)$. The standardized version of PLRT_k can be written as

$$\tau_k = \frac{-2n \cdot \mathrm{PLRT}_k - \sum_{\nu=1}^n d_\nu(k)}{(2\sum_{\nu=1}^n d_\nu(k)^2)^{1/2}}. \tag{6.3}$$

Inspired by Theorem 5.1, τ_k is presumably of standard normal distribution for any particular k. However, k is often unavailable in practice. As discussed previously, we shall construct the adaptive testing based on a sequence of τ_k as follows: (i) define $AT_n^* = \max_{1\le k\le k_n} \tau_k$, and (ii) standardize AT_n^* as

$$AT_n = B_n\bigl(AT_n^* - B_n\bigr),$$

where B_n satisfies $2\pi B_n^2 \exp(B_n^2) = k_n^2$; see [12]. By Cramér [7], $B_n = \sqrt{2\log k_n} - \frac{1}{2}(\log\log k_n + \log 4\pi)/\sqrt{2\log k_n} + O(1/\log k_n) \asymp \sqrt{2\log k_n}$ as n becomes sufficiently large.

THEOREM 6.1. *Suppose $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$. Then for any $\bar{\alpha} \in (0, 1)$, we have under $H_0 : \beta = 0$,*

$$P(AT_n \le c_{\bar{\alpha}}) \to 1 - \bar{\alpha} \qquad \textit{as } n \to \infty,$$

where $c_{\bar{\alpha}} = -\log(-\log(1 - \bar{\alpha}))$.

The proof of Theorem 6.1 is mainly based on Stein's leave-one-out method [28] since under H_0, τ_k can be written as a sum of independent random variables, that is, $\tau_k = \sum_{\nu=1}^n [d_\nu(k)/s_{n,k}](\eta_\nu^2 - 1)$, where $s_{n,k}^2 = 2\sum_{\nu=1}^n d_\nu(k)^2$.

In the end, we investigate the optimality of the proposed adaptive testing procedure. Consider the local alternative $H_{1n} : \beta \in \mathcal{B}_{k,1}$, where

$$\mathcal{B}_{k,1} \equiv \left\{\sum_{\nu=1}^\infty b_\nu \widehat{\varphi}_\nu : \sum_{\nu=1}^\infty b_\nu^2 \nu^{2k} \le 1\right\},$$

for some fixed but *unknown* integer $k \ge 1$. For any real sequence $\mathfrak{b} = \{b_\nu\}$ satisfying $\sum_{\nu=1}^\infty b_\nu^2 \nu^{2k} \le 1$, let $\beta_{\mathfrak{b}} = \sum_{\nu=1}^\infty b_\nu \widehat{\varphi}_\nu$ be the alternative function value, and let $P_{\mathfrak{b}}$ be the corresponding probability measure. The following result shows that the adaptive test AT_n achieves the optimal minimax rate (up to an logarithmic order), that is, $\delta(n, k) \equiv n^{-2k/(4k+1)}(\log\log n)^{k/(4k+1)}$, for testing the hypothesis $H_0 : \beta = 0$, with the alternative set being certain Sobolev ellipsoid $\mathcal{B}_{k,1}$; see [15].

Define $\|\mathfrak{b}\|_{\ell^2}^2 = \sum_{\nu=1}^\infty b_\nu^2$ and $\|\mathfrak{b}\|_{k,\ell^2}^2 = \sum_{\nu=1}^\infty b_\nu^2 \rho_\nu(k)$.

THEOREM 6.2. *Suppose $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$. Then, for any $\varepsilon \in (0, 1)$, there exist positive constants N_ε and C_ε such that for any $n \ge N_\varepsilon$,*

$$\inf_{\substack{\|\mathfrak{b}\|_{\ell^2} \ge C_\varepsilon \delta(n,k)\\ \|\mathfrak{b}\|_{k,\ell^2} \le 1}} P_{\mathfrak{b}}(\textit{reject } H_0) \ge 1 - \varepsilon.$$

In Gaussian white noise models, Fan [9] and Fan and Lin [10] proposed an adaptive Neyman test based on multiple standardized test, and derived the null limit distribution using the Darling–Erdős theorem. Theorems 6.1 and 6.2 can be viewed as extensions of such results to functional data under Gaussian errors.

However, the Darling–Erdős theorem is no longer applicable in our setup due to the difference in modeling and test construction. Instead, we employ the Stein leave-one-out method. More interestingly, Stein's method can even be applied to handle sub-Gaussian errors, as will be seen in Section 6.2.

6.2. *Sub-Gaussian error.* In this subsection, we consider models with sub-Gaussian errors; that is, there exists a positive constant C_ϵ such that $E\{\exp(t\epsilon)\} \le \exp(C_\epsilon t^2)$ for all $t \in \mathbb{R}$. Further relaxation to the error term with finite fourth moment is straightforward, but requires more stringent conditions on the design. For simplicity, we assume deterministic design, and suppose that X_i's satisfy the following moment condition:

$$\max_{1\le\nu\le n}\sum_{i=1}^{n}\omega_{i\nu}^4 = o\bigl(n^{8/5}(\log\log n)^{-14/5}\bigr). \tag{6.4}$$

Recall that $\omega_{i\nu} = \int_0^1 X_i(t)\widehat{\varphi}_\nu(t)\,dt$ and is nonrandom under deterministic design. Condition (6.4) implies that for any $\nu = 1,\dots,n$, the magnitudes of $\omega_{1\nu},\dots,\omega_{n\nu}$ should be comparable given the restriction that $\sum_{i=1}^n \omega_{i\nu}^2 = n$. It rules out the situation that the sequence $\omega_{i\nu}$ is spiked at $i=\nu$, that is, $\omega_{\nu\nu}^2 = n$ and $\omega_{i\nu} = 0$, for any $i \ne \nu$. This special situation essentially gives rise to $\Omega = \sqrt{n}I$ such that PLRT_k defined in (6.2) can be written as a scaled sum of independent centered squares of the errors. The leave-one-out method employed in Theorem 6.1 can handle this special case.

We first standardize PLRT_k. The non-Gaussian assumption yields a substantially different design matrix. Hence, the scale factor is chosen to be different from the one used in Section 6.1, as described below. The standardized version is defined as

$$\tilde{\tau}_k = \frac{-2n\cdot\mathrm{PLRT}_k - \sum_{\nu=1}^n d_\nu(k)}{(2\sum_{i\ne j} a_{ij}^2(k))^{1/2}},$$

where $a_{ij}(k)$ is the (i,j)th entry of $A_k \equiv n^{-1}\Omega(I_n + \lambda_k\Lambda_k)^{-1}\Omega^T$ for $1 \le i, j \le n$. Note that the scale factor in $\tilde{\tau}_k$, that is, the term $(2\sum_{i\ne j} a_{ij}(k)^2)^{1/2}$, differs from the one in τ_k. Technically, this new scale factor will facilitate the asymptotic theory developed later in this section. Let $AT_n^* = \max_{1\le k\le k_n}\tilde{\tau}_k$, and $AT_n = B_n(AT_n^* - B_n)$, where B_n satisfies $2\pi B_n^2\exp(B_n^2) = k_n^2$.

THEOREM 6.3. *Suppose $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$. Furthermore, ϵ is sub-Gaussian, and (6.4) holds. Then for any $\bar{\alpha} \in (0,1)$, we have under $H_0 : \beta = 0$,*

$$P(AT_n \le c_{\bar{\alpha}}) \to 1 - \bar{\alpha} \qquad \text{as } n \to \infty,$$

where $c_{\bar{\alpha}} = -\log(-\log(1-\bar{\alpha}))$.

The proof of Theorem 6.3 is mainly based on Stein's exchangeable pair method; see [28].

We conclude this subsection by showing that the proposed adaptive test can still achieve the optimal minimax rate (up to a logarithmic order) specified in [15], that is, $\delta(n,k)$, even under non-Gaussian errors. Recall that $\delta(n,k)$, $\|\mathfrak{b}\|_{\ell^2}$, $\|\mathfrak{b}\|_{k,\ell^2}$ and $P_{\mathfrak{b}}$ are defined in Section 6.1.

THEOREM 6.4. *Suppose $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$. Furthermore, ϵ is sub-Gaussian, and* (6.4) *holds. Then, for any $\varepsilon \in (0,1)$, there exist positive constants N_ε and C_ε such that for any $n \geq N_\varepsilon$,*

$$\inf_{\substack{\|\mathfrak{b}\|_{\ell^2} \geq C_\varepsilon \delta(n,k) \\ \|\mathfrak{b}\|_{k,\ell^2} \leq 1}} P_{\mathfrak{b}}(\textit{reject } H_0) \geq 1 - \varepsilon.$$

7. Simulation study. In this section, we investigate the numerical performance of the proposed procedures for inference. We consider four different simulation settings. The settings in Sections 7.1–7.3 are exactly the same as those in Hilgert et al. [15] and Lei [17] so that we can fairly compare our testing results with theirs. We focus on models with Gaussian error and choose $m = 2$, that is, cubic spline. Confidence interval in Section 4, penalized likelihood ratio test in Section 5.2 and adaptive testing procedure in Section 6.1 are examined. The setting in Section 7.4 is about functional linear logistic regression. Size and power of the PLRT test are examined.

7.1. *Setting* 1. Data were generated in the same way as in Hilgert et al. [15]. Consider the functional linear model $Y_i = \int_0^1 X_i(t)\beta_0(t)\,dt + \epsilon_i$, with ϵ_i being independent standard normal for $i = 1, \dots, n$. Let $\lambda_j = (j - 0.5)^{-2}\pi^{-2}$ and $V_j(t) = \sqrt{2}\sin((j - 0.5)\pi t)$, $t \in [0,1]$, $j = 1, 2, \dots, 100$. The covariate curve $X_i(t)$ was Brownian motion simulated as $X_i(t) = \sum_{j=1}^{100} \sqrt{\lambda_j}\eta_{ij} V_j(t)$, where η_{ij}'s are independent standard normal for $i = 1, \dots, n$ and $j = 1, \dots, 100$. Each $X_i(t)$ was observed at 1000 evenly spaced points over [0, 1]. The true slope function was chosen as

$$\beta_0^{B,\xi}(t) = \frac{B}{\sqrt{\sum_{k=1}^{\infty} k^{-2\xi-1}}} \sum_{j=1}^{100} j^{-\xi-0.5} V_j(t).$$

Figure 1 displays β_0. Four different signal strengths $B = (0, 0.1, 0.5, 1)$ and three smoothness parameters $\xi = (0.1, 0.5, 1)$ were considered. Note that $B = 0$ implies $\beta_0 = 0$.

For each case study, we considered sample sizes $n = 100$ and $n = 500$ respectively, and ran 10,000 trials to investigate the Monte Carlo performance of our methods.

Case study 1: 95% *confidence interval for conditional mean.* In this study, we set $\mu_0(x_0) = E\{Y|X_0 = x_0\} = \int_0^1 x_0(t)\beta_0(t)\,dt$ with $B = 1, \xi = 1$, where x_0 is

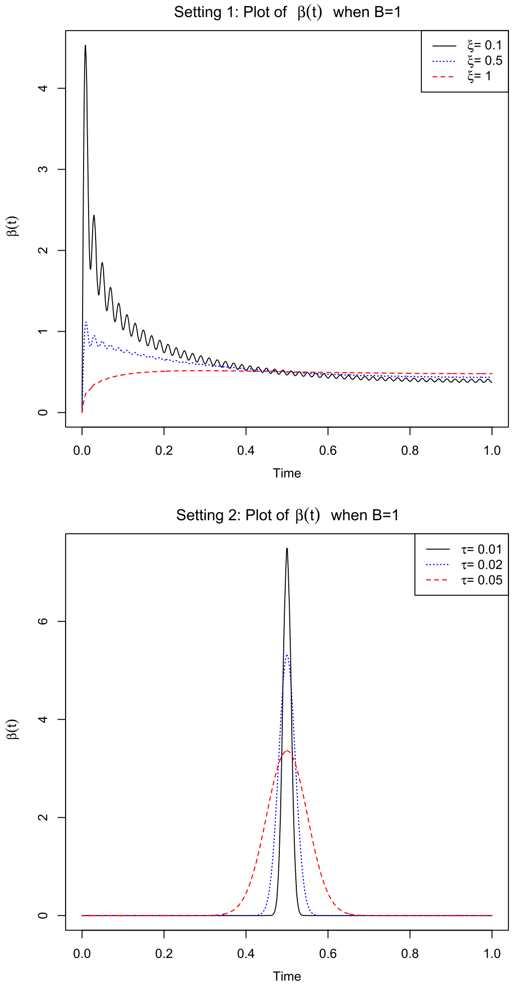

FIG. 1. *Plots of $\beta_0(t)$ in settings* 1 *and* 2.

independent of $X_1, \dots, X_n$ and randomly generated from the same distribution as X_1. From (4.2), the 95% confidence interval for $\mu_0(x_0)$ is

$$[\widehat{Y}_0 - n^{-1/2} z_{0.025}\sigma_n, \widehat{Y}_0 + n^{-1/2} z_{0.025}\sigma_n],$$

where $\sigma_n^2 = 1 + \sum_{\nu=1}^{\infty} x_\nu^2/(1+\lambda\rho_\nu)$, $x_\nu = \int_0^1 x_0(t)\varphi_\nu(t)\,dt$. Here φ_ν and ρ_ν are both obtained through (2.8).

With 10,000 replications, percentages of the conditional mean $\mu_0(x_0)$ beyond the scope of CI and the average lengths of the confidence intervals are summarized in Table 1.

Case study 2: *Size of the tests*. Denote the testing methods proposed by Hilgert et al. [15] as HMV13$^{(1)}$ and HMV13$^{(2)}$. Under $H_0 : \beta = 0$, we calculated the sizes

TABLE 1
Case study 1: *Percentage of* $\mu_0(x_0)$ *outside the* 95% *confidence intervals* ± *standard deviation* (*average length of the* 95% *confidence intervals*)

$n = 100$	$n = 500$
4.89 ± 0.42 (0.56)	5.01 ± 0.19 (0.39)

of PLRT and AT (adaptive testing), that is, the percentages of rejecting H_0, and then compared them with HMV13$^{(1)}$ and HMV13$^{(2)}$ (directly cited from [15]) in Table 2. Numerically, we found that AT converges to the Gumbel distribution very slowly, which is a common phenomenon in the extreme value literature; see [9, 10]. Following an idea similar to [10], finite sample distributions of AT based on one million replications were instead used. Obviously, from Table 2, the proposed PLRT and AT are both valid test statistics achieving desirable sizes.

Case study 3: *Power comparison.* In this study, we generated β_0 under different signal strengths $B = (0.1, 0.5, 1)$ and smoothing parameters $\xi = (0.1, 0.5, 1)$. Tables 3 and 4 summarize the powers of four different testing methods, that is, the percentages of rejecting $H_0 : \beta = 0$ at 95% significance level, under $n = 100$ and $n = 500$. From $n = 100$ to $n = 500$, the powers of all tests increase. In particular, PLRT generally performs better than AT since PLRT incorporates known information from the model, that is, $r = 0$ (smoothness of the covariance kernel) and $m = 2$ (smoothness of the functional parameter), while AT is adaptive on these quantities. The power loss is the price paid for adaptiveness. We also note that for weaker signals $B = 0.1$, the powers of PLRT and AT improve those of HMV13$^{(1)}$, HMV13$^{(2)}$, while for stronger signals $B = 0.5$, 1, the powers of all tests are comparable.

7.2. *Setting* 2.

Let the true slope function be

$$\beta_0^{B,\tau}(t) = B\exp\left\{-\frac{(t-0.5)^2}{2\tau^2}\right\}\left\{\int_0^1 \exp\left\{-\frac{(x-0.5)^2}{\tau^2}\right\}dx\right\}^{-1/2},$$

TABLE 2
Case study 2: *Sizes of the tests*

	$n = 100$	$n = 500$
HMV13$^{(1)}$	3.47 (±0.36)	2.61 (±0.14)
HMV13$^{(2)}$	4.97 (±0.43)	5.26 (±0.20)
AT	5.13 (±0.43)	5.04 (±0.19)
PLRT	5.45 (±0.45)	5.19 (±0.20)

TABLE 3
Case study 3: $n = 100$. *Powers*

	Test	$\boldsymbol{B = 0.1}$	$\boldsymbol{B = 0.5}$	$\boldsymbol{B = 1}$
$\xi = 0.1$	HMV13$^{(1)}$	3.88 (±0.38)	21.41 (±0.8)	77.24 (±0.82)
	HMV13$^{(2)}$	5.80 (±0.46)	26.38 (±0.86)	81.78 (±0.76)
	AT	6.12 (±0.47)	30.77 (±0.90)	81.56 (±0.76)
	PLRT	21.27 (±0.80)	42.34 (±0.97)	84.20 (±0.71)
$\xi = 0.5$	HMV13$^{(1)}$	4.74 (±0.42)	46.47 (±0.98)	98.68 (±0.22)
	HMV13$^{(2)}$	6.65 (±0.49)	52.79 (±0.98)	99.06 (±0.19)
	AT	8.28 (±0.54)	71.08 (±0.89)	99.86 (±0.07)
	PLRT	23.13 (±0.83)	74.74 (±0.85)	99.70 (±0.11)
$\xi = 1$	HMV13$^{(1)}$	4.8 (±0.42)	62.67 (±0.95)	99.75 (±0.10)
	HMV13$^{(2)}$	7.07 (±0.5)	68.30 (±0.91)	99.84 (±0.08)
	AT	9.47 (±0.57)	83.20 (±0.73)	99.98 (±0.03)
	PLRT	23.95 (±0.84)	84.03 (±0.72)	99.98 (±0.03)

where $B = (0.5, 1, 2)$ and $\tau = (0.01, 0.02, 0.05)$. The processes $X_i(t)$ and the samples were generated in the same way as in Setting 1.

The powers in Setting 2 are summarized in Tables 5 and 6. We observe similar phenomena as in Setting 1, that under weaker signals, say $\tau = 0.01$, $B = 0.5$, PLRT and AT demonstrate larger powers, while the powers of all procedures become comparable under stronger signals. Again, PLRT generally has larger powers than the adaptive procedure AT. All the powers increase as sample size becomes larger.

TABLE 4
Case study 3: $n = 500$. *Powers*

	Test	$\boldsymbol{B = 0.1}$	$\boldsymbol{B = 0.5}$	$\boldsymbol{B = 1}$
$\xi = 0.1$	HMV13$^{(1)}$	5.17 (±0.19)	86.98 (±0.29)	100 (±0)
	HMV13$^{(2)}$	8.48 (±0.24)	90.89 (±0.25)	100 (±0)
	AT	9.57 (±0.26)	89.14 (±0.27)	100 (±0)
	PLRT	20.00 (±0.35)	88.19 (±0.28)	100 (±0)
$\xi = 0.5$	HMV13$^{(1)}$	8.81 (±0.25)	99.85 (±0.03)	100 (±0)
	HMV13$^{(2)}$	13.07 (±0.30)	99.88 (±0.03)	100 (±0)
	AT	20.20 (±0.35)	100 (±0)	100 (±0)
	PLRT	29.47 (±0.40)	99.90 (±0.03)	100 (±0)
$\xi = 1$	HMV13$^{(1)}$	11.38 (±0.28)	99.99 (±0.01)	100 (±0)
	HMV13$^{(2)}$	16.13 (±0.32)	100 (±0)	100 (±0)
	AT	26.51 (±0.39)	100 (±0)	100 (±0)
	PLRT	34.08 (±0.42)	100 (±0)	100 (±0)

TABLE 5
Setting 2: $n = 100$. *Powers*

	Test	$B = 0.5$	$B = 1$	$B = 2$
$\tau = 0.01$	HMV13$^{(1)}$	4.94 (±0.42)	11.85 (±0.63)	46.69 (±0.98)
	HMV13$^{(2)}$	7.25 (±0.51)	15.49 (±0.71)	53.56 (±0.98)
	AT	9.88 (±0.58)	23.86 (±0.84)	69.46 (±0.90)
	PLRT	17.9 (±0.75)	33.25 (±0.92)	81.04 (±0.77)
$\tau = 0.02$	HMV13$^{(1)}$	7.33 (±0.51)	23.09 (±0.83)	80.26 (±0.78)
	HMV13$^{(2)}$	10 (±0.59)	28.54 (±0.89)	84.04 (±0.72)
	AT	14.58 (±0.69)	42.21 (±0.97)	93.54 (±0.48)
	PLRT	22.87 (±0.82)	53.21 (±0.98)	97.83 (±0.29)
$\tau = 0.05$	HMV13$^{(1)}$	13.85 (±0.68)	56.51 (±0.97)	99.48 (±0.14)
	HMV13$^{(2)}$	18.13 (±0.50)	63.09 (±0.95)	99.65 (±0.12)
	AT	28.31 (±0.88)	78.52 (±0.80)	99.96 (±0.04)
	PLRT	37.54 (±0.95)	87.63 (±0.65)	100 (±0)

7.3. *Setting* 3. In this setting, data were generated in the same way as in Section 4.2 of [17]. Hence we will compare our PLRT and AT with the testing procedure in [17], denoted as L13. Specifically, the covariance operator has eigenvalues $\kappa_j = j^{-1.7}$ and eigenfunctions $\phi_1(t) = 1, \phi_j(t) = \sqrt{2}\cos((j-1)\pi t)$ for $j \geq 2$. The covariate processes are $X_i(t) = \sum_{j=1}^{100} \sqrt{\kappa_j}\eta_j\phi_j(t)$, where η_j's are independent standard normal. Each $X_i(t)$ was observed on 1000 evenly spaced points over [0, 1].

In the first case denoted as Model(2, 1), let $\theta_j = \bar{\theta}_j/\|\bar{\theta}\|_2$, where $\bar{\theta}_j = 0$ for $j > 2$, $\bar{\theta}_j = b_j \cdot I_j$ for $j = 1, 2$, b_1 and b_2 are independent Unif(0, 1), and

TABLE 6
Setting 2: $n = 500$. *Powers*

	Test	$B = 0.5$	$B = 1$	$B = 2$
$\tau = 0.01$	HMV13$^{(1)}$	12.41 (±0.42)	54.6 (±0.63)	99.75 (±0.98)
	HMV13$^{(2)}$	17.99 (±0.51)	63.16 (±0.71)	99.98 (±0.98)
	AT	28.93 (±0.40)	79.75 (±0.35)	100 (±0)
	PLRT	34.77 (±0.42)	86.08 (±0.30)	100 (±0)
$\tau = 0.02$	HMV13$^{(1)}$	26.11 (±0.51)	88.91 (±0.83)	100 (±0)
	HMV13$^{(2)}$	33.95 (±0.59)	92.62 (±0.89)	100 (±0)
	AT	50.25 (±0.44)	97.03 (±0.15)	100 (±0)
	PLRT	56.57 (±0.43)	99.20 (±0.08)	100 (±0)
$\tau = 0.05$	HMV13$^{(1)}$	65.38 (±0.68)	99.95 (±0.97)	100 (±0)
	HMV13$^{(2)}$	72.74 (±0.50)	99.99 (±0.95)	100 (±0)
	AT	86.92 (±0.30)	100 (±0)	100 (±0)
	PLRT	92.07 (±0.24)	100 (±0)	100 (±0)

TABLE 7
Setting 3: *Powers*

	Sample size	**Test**	$r^2 = 0.1$	$r^2 = 0.2$	$r^2 = 0.5$	$r^2 = 1.5$
Model(2, 1)	$n = 50$	L13	16.20 (±1.02)	26.40 (±1.22)	54.20 (±1.38)	80.80 (±1.09)
		AT	47.81 (±1.38)	64.94 (±1.32)	84.75 (±1.00)	99.13 (±0.26)
		PLRT	57.76 (±1.37)	72.70 (±1.23)	90.18 (±0.82)	99.52 (±0.19)
	$n = 100$	L13	25.80 (±0.86)	42.20 (±0.97)	68.20 (±0.91)	90.40 (±0.58)
		AT	65.53 (±0.93)	79.75 (±0.79)	96.97 (±0.34)	99.99 (±0.02)
		PLRT	74.04 (±0.86)	87.98 (±0.64)	98.22 (±0.26)	100 (±0)
	$n = 500$	L13	67.20 (±0.41)	84.60 (±0.32)	94.40 (±0.20)	97.20 (±0.14)
		AT	97.81 (±0.13)	100 (±0)	100 (±0)	100 (±0)
		PLRT	98.5 (±0.11)	99.94 (±0.02)	100 (±0)	100 (±0)
Model(9, 2)	$n = 50$	L13	9.00 (±0.79)	14.00 (±0.96)	29.60 (±1.27)	43.40 (±1.37)
		AT	21.72 (±1.14)	27.57 (±1.24)	37.67 (±1.34)	53.33 (±1.38)
		PLRT	39.54 (±1.36)	46.22 (±1.38)	56.92 (±1.37)	73.42 (±1.22)
	$n = 100$	L13	13.4 (±0.67)	27.8 (±0.88)	39.8 (±0.96)	65.8 (±0.93)
		AT	27.86 (±0.88)	21.63 (±0.81)	47.61 (±0.98)	65.69 (±0.93)
		PLRT	45.80 (±0.98)	53.72 (±0.98)	67.12 (±0.92)	83.88 (±0.72)
	$n = 500$	L13	42.40 (±0.43)	47.8 (±0.44)	72.4 (±0.39)	93.4 (±0.22)
		AT	49.40 (±0.44)	58.29 (±0.43)	80.21 (±0.35)	91.23 (±0.25)
		PLRT	69.44 (±0.40)	80.00 (±0.35)	91.70 (±0.24)	99.22 (±0.08)

(I_1, I_2) follows a multinomial distribution Mult(1; 0.5, 0.5). Let the true function be $\beta_0(t) = r\sum_{j=1}^{100}\theta_j\phi_j(t)$, where $r^2 = (0, 1, 0.2, 0.5, 1.5)$.

In the second case denoted as Model(9, 2), a different choice of θ_j was considered. Specifically, $\theta_j = \bar{\theta}_j/\|\bar{\theta}\|_2$, where $\bar{\theta}_j = 0$ for $j > 9$, $\bar{\theta}_j = b_j \cdot I_j$ for $j = 1, \ldots, 9$, $b_1, \ldots, b_9$ are independent Unif(0, 1), and $(I_1, \ldots, I_9)$ follows a multinomial distribution Mult(2; 1/9, …, 1/9).

In both cases, the samples were drawn from $Y_i = \int_0^1 X_i(t)\beta(t)\,dt + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i are independent standard Gaussian. 5000 Monte Carlo trials were conducted in each case under different sample sizes $n = 50, 100$ and 500.

Results are summarized in Table 7, from which we can see that the powers of AT and PLRT improve those of L13, especially when $r^2 = 0.1, 0.2$ (weaker signals). As n increases, the power of L13 becomes more comparable to those of PLRT and AT especially when $r^2 = 1.5$ (stronger signal). Again, PLRT generally has larger powers than adaptive methods.

7.4. *Setting* 4. Let $Y \in \{0, 1\}$ be a binary variable generated from the following functional logistic regression model:

$$P(Y = 1|X) = \frac{\exp(\int_0^1 X(t)\beta_0(t)\,dt)}{1 + \exp(\int_0^1 X(t)\beta_0(t)\,dt)}.$$

TABLE 8
Setting 4: *Size and power*

	$n = 100$	$n = 500$
Size	0.054	0.046
Power	0.387	0.985

The predictor process X_i was simulated as $X_i(t) = \sum_{j=1}^{100} \sqrt{\lambda_j}\eta_{ij} V_j(t)$, where λ_j and $V_j(t)$ are exactly the same as in Setting 1, η_{ij}'s are independent truncated normals, that is, $\eta_{ij} = \xi_{ij} I_{\{|\xi_{ij}| \le 0.5\}} + 0.5 I_{\{\xi_{ij} > 0.5\}} - 0.5 I_{\{\xi_{ij} < -0.5\}}$, with ξ_{ij} being a standard normal random variable. Each $X_i(t)$ was observed at 1000 evenly spaced points over [0, 1]. We intend to test $H_0 : \beta = 0$. To examine the power, data were generated under $\beta_0(t) = 3 * 10^5(t^{11}(1-t)^6)$ for $t \in [0, 1]$.

We examined two sample sizes: $n = 100$ and $n = 500$. Results (summarized in Table 8) were based on 10,000 independent trials. It can be seen that when $n = 100$ and 500, the test achieves the desired sizes. The power at $n = 100$ is small, but the power at $n = 500$ approaches one, demonstrating the asymptotic property of the test.

8. Discussion.

The current paper and our previous work on nonparametric regression models [26] are both built upon the RKHS framework and theory. Hence it seems necessary for us to comment their technical connections and differences to facilitate the reading. Compared to [26], the RKHS considered in the current paper has a substantially different structure that involves a covariance function of the predictor process. This immediately causes a difference in building the eigensystems: [26] relies on an ODE system, but the current paper relies on an integro-differential system. Hence the methods of analyzing both systems are crucially different. Meanwhile, the asymptotic analysis on the statistical inference such as the penalized likelihood ratio test are also different. For example, [26] only considers the reproducing kernel, while the current work requires a delicate interaction between the reproducing kernel and the covariance kernel. More importantly, the relaxation of perfect alignment between both kernels poses more technical challenges.

Besides, Assumption A3 requires $\|\varphi_\nu\|_{L^2} \le C_\varphi \nu^a$ for $\nu \ge 1$ and a constant $a \ge 0$. The introduction of factor a in Assumption A3 is helpful in simplifying our proofs. However, it is interesting to investigate how to avoid imposing this seemingly "redundant" a. As indicated by Proposition 2.2, that a relates to C (and hence V), one possible strategy is to avoid the use of V. Instead, one may use its empirical version, namely V_n, as suggested by one referee. This would require a delicate analysis of the convergence of V_n, which may be handled by techniques in [18]. We leave this as a future exploration.

Acknowledgments. We thank Pang Du for providing us his R code for functional logistic regression, and thank Ph.D. student Meimei Liu at Purdue for help with the simulation study. We also thank Co-editor Runze Li, an Associate Editor, and two referees for helpful comments that lead to important improvements on the paper.

Guang Cheng was on sabbatical at Princeton while the revision of this work was carried out; he would like to thank the Princeton ORFE department for its hospitality and support.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric inference in generalized functional linear models" (DOI: 10.1214/15-AOS1322SUPP; .pdf). Proofs are provided.

REFERENCES

[1] BIRKHOFF, G. D. (1908). Boundary value and expansion problems of ordinary linear differential equations. *Trans. Amer. Math. Soc.* **9** 373–395. MR1500818

[2] BUNEA, F., IVANESCU, A. E. and WEGKAMP, M. H. (2011). Adaptive inference for the mean of a Gaussian process in functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 531–558. MR2853729

[3] CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

[4] CAI, T. T. and YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216. MR3010906

[5] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[6] CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

[7] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, NJ. MR0016588

[8] DOU, W. W., POLLARD, D. and ZHOU, H. H. (2012). Estimation in functional regression for general exponential families. *Ann. Statist.* **40** 2421–2451. MR3097608

[9] FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688. MR1395735

[10] FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. MR1649196

[11] FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193. MR1833962

[12] HALL, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16** 433–439. MR0531778

[13] HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

[14] 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. MR2278365

[15] HILGERT, N., MAS, A. and VERZELEN, N. (2013). Minimax adaptive tests for the functional linear model. *Ann. Statist.* **41** 838–869. MR3099123

[16] INGSTER, YU. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I–III. *Math. Methods Statist.* **2** 85–114; **3** 171–189; **4** 249–268. MR1259685

[17] LEI, J. (2014). Adaptive global testing for functional linear models. *J. Amer. Statist. Assoc.* **109** 624–634. MR3223738

[18] MENDELSON, S. (2010). Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.* **20** 988–1027. MR2729283

[19] MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80. MR0642719

[20] MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

[21] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

[22] RITTER, K., WASILKOWSKI, G. W. and WOŹNIAKOWSKI, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks–Ylvisaker conditions. *Ann. Appl. Probab.* **5** 518–540. MR1336881

[23] SACKS, J. and YLVISAKER, D. (1968). Designs for regression problems with correlated errors; many parameters. *Ann. Math. Stat.* **39** 49–69. MR0220424

[24] SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors. III. *Ann. Math. Stat.* **41** 2057–2074. MR0270530

[25] SACKS, J. and YLVISAKER, N. D. (1966). Designs for regression problems with correlated errors. *Ann. Math. Stat.* **37** 66–89. MR0192601

[26] SHANG, Z. and CHENG, G. (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist.* **41** 2608–2638. MR3161439

[27] SHANG, Z. and CHENG, G. (2015). Supplement to "Nonparametric inference in generalized functional linear models." DOI:10.1214/15-AOS1322SUPP.

[28] STEIN, C. (1986). *Approximate Computation of Expectations. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **7**. IMS, Hayward, CA. MR0882007

[29] TAMARKIN, J. D. (1927). The notion of the Green's function in the theory of integro-differential equations. *Trans. Amer. Math. Soc.* **29** 755–800. MR1501413

[30] TAMARKIN, J. D. (1930). The notion of the Green's function in the theory of integro-differential equations. II. *Trans. Amer. Math. Soc.* **32** 860–868. MR1501566

[31] TAMARKIN, J. D. and LANGER, R. E. (1928). On integral equations with discontinuous kernels. *Trans. Amer. Math. Soc.* **30** 453–471. MR1501439

[32] WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** 439–447. MR0375592

[33] WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9** 60–62.

[34] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

[35] YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
250 N. UNIVERSITY STREET
WEST LAFAYETTE, INDIANA 47906
USA
E-MAIL: shang9@purdue.edu
chengg@purdue.edu