The Annals of Statistics
2015, Vol. 43, No. 3, 1351–1390
DOI: 10.1214/15-AOS1313

JOINT ASYMPTOTICS FOR SEMI-NONPARAMETRIC REGRESSION MODELS WITH PARTIALLY LINEAR STRUCTURE

BY GUANG CHENG[1] AND ZUOFENG SHANG

Purdue University

We consider a joint asymptotic framework for studying semi-nonparametric regression models where (finite-dimensional) Euclidean parameters and (infinite-dimensional) functional parameters are both of interest. The class of models in consideration share a partially linear structure and are estimated in two general contexts: (i) quasi-likelihood and (ii) true likelihood. We first show that the Euclidean estimator and (pointwise) functional estimator, which are re-scaled at different rates, jointly converge to a zero-mean Gaussian vector. This weak convergence result reveals a surprising *joint asymptotics phenomenon*: these two estimators are asymptotically independent. A major goal of this paper is to gain first-hand insights into the above phenomenon. Moreover, a likelihood ratio testing is proposed for a set of joint local hypotheses, where a new version of the Wilks phenomenon [*Ann. Math. Stat.* **9** (1938) 60–62; *Ann. Statist.* **1** (2001) 153–193] is unveiled. A novel technical tool, called a *joint Bahadur representation*, is developed for studying these joint asymptotics results.

1. Introduction. In the literature, a statistical model is called *semi-nonparametric* if it contains both finite-dimensional and infinite-dimensional unknown parameters of interest (e.g., [14]). An example is semi-nonparametric copula model that can be applied to address tail dependence among shocks to different financial series and also to recover the shape of the "news impact curve" for individual financial series. Another example is the semi-nonparametric binary regression models proposed by Banerjee, Mukherjee and Mishra [2] to define the conditional probability of attending primary school in Indian villages through an appropriate link function influenced by a set of covariates such as gender and household income. As a first step in exploring the joint asymptotics results, we focus on the semi-nonparametric regression models with a partial linear structure in this paper.

The existing semiparametric literature is concerned with asymptotic theories and inference procedures for the Euclidean parameter *only*. The functional parameter is profiled out as an infinite-dimensional nuisance parameter; see [3, 8–10, 25, 29]. In the special case where both parameters are estimable at the same root-n

Received May 2014; revised January 2015.

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

MSC2010 subject classifications. Primary 62G20, 62F12; secondary 62F03.

Key words and phrases. Joint asymptotics, joint Bahadur representation, local likelihood ratio test, semi-nonparametric models, smoothing spline.

rate (e.g., [19, 20]), we can combine them as an infinite-dimensional parameter and then apply the functional Z-estimation theorem (e.g., Theorem 3.3.1 in [34]), to study its joint asymptotic distribution. However, it is more common for the two parameters to be estimated at different parametric and nonparametric rates. In general, their radically different parameter dimensionality poses technical challenges for the construction of valid procedures for joint inference. In this paper, we develop a new technical tool, called a *joint Bahadur representation* (JBR), for studying the joint asymptotics results. As far as we are aware, our joint asymptotic theories and inference procedures are new. The only relevant reference of which we are aware is [27], which focuses on a fully parametric setting.

In this paper, we assume a partially linear structure for the conditional mean of the response, and then estimate the model in two general contexts: (i) quasi-likelihood and (ii) true likelihood. Within this framework, we derive a joint limit distribution for the Euclidean estimator and the (point-wise) functional estimator as a zero-mean Gaussian vector after they are re-scaled properly. One surprising result is that these two estimators are asymptotically independent. This asymptotic independence will prove to be useful in making joint inference. For example, it is now straightforward to construct the joint confidence interval based on two marginal ones. Under similar conditions, the marginal limit distribution for the Euclidean estimator coincides with that derived in [22]. On the other hand, we observe that the (pointwise) marginal asymptotic results for the nonparametric component are generally different from those derived in a purely nonparametric setup (without the Euclidean parameter) (i.e., [30]), even though the Euclidean parameter is estimated at a faster rate; see Remark 5.1. This conclusion is a bit counterintuitive.

We next propose likelihood ratio testing for a variety of joint local hypotheses such as $H_0 : \theta = \theta_0$ and $g(z_0) = w_0$ and $H_0 : x^T\theta + g(z_0) = \alpha$, where θ and g denote the parametric and nonparametric components, respectively. Conventional semiparametric testing only focuses on the parametric components; see [10, 25]. However, in practice, it is often of great interest to evaluate the nonparametric components at the same time. For example, we may test the joint effect of child gender θ and household income g on the probability of attending primary school in the Indian schooling model; see [2]. In particular, we show that the null limit distribution is a mixture of two *independent* Chi-square distributions that are contributed by the parametric and nonparametric components, respectively. Note that this independence property is implied by the joint asymptotics phenomenon, and is practically useful in finding the critical value. In the parametric framework, Wilks (1938) showed that the likelihood ratio test statistic (under $H_0 : \theta = \theta_0$) converges to a Chi-square distribution. Fan et al. (2001) call the above result the *Wilks phenomenon* due to the nice property that the asymptotic null distribution is free of nuisance parameters, and further generalize it to the nonparametric setting. Therefore, we unveil a new version of Wilks phenomenon that adapts to the semi-nonparametric context in this paper. As far as we are aware, this joint testing result is new. The only relevant paper of which we are aware is [2], where the authors consider two separate null hypotheses, that is, $H_{01} : \theta = \theta_0$ and $H_{02} : g(z_0) = w_0$, under the monotonicity constraint of $g(\cdot)$.

The class of semi-nonparametric regression models considered in this paper serves as a natural platform to deliver a new theoretical insight: joint asymptotics phenomenon. We also note that our results may be extended to the other models: (i) generalized additive partially linear models, (ii) partial functional linear regression models [32] and (iii) partially linear Cox proportional hazard models [15] by either modifying the JBR or the criterion function; see Section 6 for more elaborations. All the possible extensions mentioned above require a smoothness assumption on the nonparametric function. This assumption is crucially different from the shape-constraint assumption, which in general leads to the "nonstandard asymptotics" problems (e.g., [7, 17, 21]). Our framework cannot be easily adapted to handle these challenging problems, which are usually analyzed by rather different technical tools.

The rest of this paper is organized as follows. Section 2 introduces the model assumptions and builds a theoretical foundation. Sections 3 and 4 formally discuss the joint limit distribution and joint local hypothesis testing, respectively. In Section 5, we give three concrete examples with extensive simulations to illustrate our theory. Section 6 discusses some possible extensions. The proofs are postponed to the Appendix or online supplementary document [6].

2. Preliminaries.

This section introduces the model assumptions and establishes the theoretical foundation of our results in two layers: (i) the partially linear extension of reproducing kernel Hilbert space (RKHS) theory and (ii) the joint Bahadur representation. Both technical results are of independent interest.

2.1. *Notation and model assumptions.*

Suppose that $T_i = (Y_i, X_i, Z_i)$, $i = 1, \ldots, n$, are i.i.d. copies of $T = (Y, X, Z)$, where $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the response variable, $U = (X, Z) \in \mathcal{U} \equiv \mathbb{I}^p \times \mathbb{I}$ is the covariate variable, and $\mathbb{I} = [0, 1]$. Throughout the paper we assume that the density of Z, denoted by $\pi(z)$, has positive lower bound and finite upper bound for $z \in [0, 1]$. Consider a general class of semi-nonparametric regression models with the following partially linear structure:

$$\mu_0(U) \equiv E(Y|U) = F\bigl(X^T\theta_0 + g_0(Z)\bigr), \tag{2.1}$$

where $F(\cdot)$ is some known link function and $g_0(\cdot)$ is some unknown smooth function. This primary assumption covers two classes of statistical models. The first class is called *generalized partially linear models* [5]; here the data are modeled by $y|u \sim p(y; \mu_0(u))$ for a conditional distribution p. Instead of assuming the underlying distribution, the second class specifies only the relationship between the conditional mean and the conditional variance: $\mathrm{Var}(Y|U) = \mathcal{V}(\mu_0(U))$ for some known positive-valued function $\mathcal{V}$. The nonparametric estimation of g in the second situation uses the quasi-likelihood $Q(y; \mu) \equiv \int_y^{\mu} (y - s)/\mathcal{V}(s)\, ds$

with $\mu = F(x^T\theta + g(z))$ [37]. Despite the distinct modeling principles, these two classes have a large overlap under many common combinations of $(F, \mathcal{V})$, as summarized in Table 2.1 of [23]. From now on, we work with a general criterion function $\ell(y; a) : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$, which can represent either $\log p(y; F(a))$ or $Q(y; F(a))$.

Let the full parameter space for $f \equiv (\theta, g)$ be $\mathcal{H} \equiv \mathbb{R}^p \times H^m(\mathbb{I})$, where $H^m(\mathbb{I})$ is an mth order Sobolev space defined as

$$
\begin{aligned}
H^m(\mathbb{I}) \equiv \{g : \mathbb{I} \mapsto \mathbb{R} | g^{(j)} &\text{ is absolutely continuous} \\
&\text{for } j = 0, 1, \ldots, m-1 \text{ and } g^{(m)} \in L_2(\mathbb{I})\}.
\end{aligned}
$$

With some abuse of notation, $\mathcal{H}$ may also refer to $\mathbb{R}^p \times H_0^m(\mathbb{I})$, where $H_0^m(\mathbb{I})$ is a homogeneous subspace of $H^m(\mathbb{I})$. The space $H_0^m(\mathbb{I})$ is also known as the class of periodic functions such that a function $g \in H_0^m(\mathbb{I})$ has additional restrictions $g^{(j)}(0) = g^{(j)}(1)$ for $j = 0, 1, \ldots, m-1$. Throughout this paper we assume $m > 1/2$ to be known. Consider the penalized semi-nonparametric estimator

$$
\begin{aligned}
(\widehat{\theta}_{n,\lambda}, \widehat{g}_{n,\lambda}) &= \underset{(\theta,g)\in\mathcal{H}}{\arg\max}\, \ell_{n,\lambda}(f) \\
&= \underset{(\theta,g)\in\mathcal{H}}{\arg\max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i; X_i^T\theta + g(Z_i)\big) - (\lambda/2) J(g, g) \right\},
\end{aligned}
\tag{2.2}
$$

where $J(g, \tilde{g}) = \int_{\mathbb{I}} g^{(m)}(z)\tilde{g}^{(m)}(z)\, dz$ and $\lambda \to 0$ as $n \to \infty$. Here, we use $\lambda/2$ (rather than λ) to simplify future expressions. Write $\widehat{f}_{n,\lambda} = (\widehat{\theta}_{n,\lambda}, \widehat{g}_{n,\lambda})$. The existence of $\widehat{g}_{n,\lambda}$ is guaranteed by Theorem 2.9 of [13] when the null space $\mathcal{N}_m \equiv \{g \in H^m(\mathbb{I}) : J(g, g) = 0\}$ is finite-dimensional and $\ell(y; a)$ is concave and continuous w.r.t. a.

We next assume some basic model conditions. For simplicity, throughout the paper we do not distinguish $f = (\theta, g) \in \mathcal{H}$ from its associated function $f \in \mathcal{F} \equiv \{f(x, z) = x^T\theta + g(z) : (\theta, g) \in \mathcal{H}, (x, z) \in \mathcal{U}\}$. Let $\mathcal{I}_0$ be the range for the true function $f_0(x, z) \in \mathcal{F}$, that is, a compact interval. Denote the first-, second- and third-order derivatives of $\ell(y; a)$ (w.r.t. a) by $\dot{\ell}_a$, $\ddot{\ell}_a$ and ℓ_a'''.

ASSUMPTION A1. (a) $\ell(y; a)$ is three times continuously differentiable and concave w.r.t. a. There exists a bounded open interval $\mathcal{I} \supset \mathcal{I}_0$ and positive constants C_0 and C_1 s.t.

$$
E\Big\{\exp\Big(\sup_{a\in\mathcal{I}} \big|\ddot{\ell}_a(Y; a)\big|/C_0\Big)\big|U\Big\} \le C_1 \qquad \text{a.s.}
\tag{2.3}
$$

and

$$
E\Big\{\exp\Big(\sup_{a\in\mathcal{I}} \big|\ell_a'''(Y; a)\big|/C_0\Big)\big|U\Big\} \le C_1 \qquad \text{a.s.}
\tag{2.4}
$$

(b) There exists a positive constant C_2 s.t. $C_2^{-1} \le I(U) \equiv -E(\ddot{\ell}_a(Y; X^T\theta_0 + g_0(Z))|U) \le C_2$, a.s.

(c) $\epsilon \equiv \dot{\ell}_a(Y; X^T\theta_0 + g_0(Z))$ satisfies $E(\epsilon|U) = 0$, $E(\epsilon^2|U) = I(U)$, a.s., and $E\{\epsilon^4\} < \infty$.

A detailed discussion of the above model assumptions can be found in [30]. In particular, Assumption A1(a) is typically used in semiparametric quasi-likelihood models; see [22]. Three concrete examples showing the validity of Assumption A1 are presented in Section 5.

Hereinafter, if for positive sequences a_μ and b_μ we have that a_μ/b_μ tends to a strictly positive constant, we write $a_\mu \asymp b_\mu$. If that constant is one, we write $a_\mu \sim b_\mu$. Let $\sum_\nu$ denote the sum over $\nu \in \mathbb{N} = \{0, 1, 2, \dots\}$ for convenience. Let the sup-norm of $g \in H^m(\mathbb{I})$ be $\|g\|_{\sup} = \sup_{z\in\mathbb{I}} |g(z)|$. Let λ^* be the optimal smoothing parameter; $\lambda^* \asymp n^{-2m/(2m+1)}$. For simplicity, we write $\lambda^{1/(2m)}$ as h, and thus $h^* \asymp n^{-1/(2m+1)}$.

2.2. *A partially linear extension of RKHS theory.* In this section, we adapt the nonparametric RKHS framework to our semi-nonparametric setup.

We define the inner product for $\mathcal{H}$ to be, for any $(\theta, g), (\tilde{\theta}, \tilde{g}) \in \mathcal{H}$,

$$\langle(\theta, g), (\tilde{\theta}, \tilde{g})\rangle = E_U\{I(U)(X^T\theta + g(Z))(X^T\tilde{\theta} + \tilde{g}(Z))\} + \lambda J(g, \tilde{g}), \tag{2.5}$$

and we define the norm to be $\|(\theta, g)\|^2 = \langle(\theta, g), (\theta, g)\rangle$. The validity of such a norm is demonstrated in Section S.1 of the supplement document [6] (under Assumption A3 introduced later). Under this norm, we will construct two linear operators, $R_u \in \mathcal{H}$, for any $u \in \mathcal{U}$, and $P_\lambda : \mathcal{H} \mapsto \mathcal{H}$ satisfying

$$\langle R_u, f\rangle = x^T\theta + g(z) \qquad \text{for any } u \in \mathcal{U} \text{ and } f \in \mathcal{H} \tag{2.6}$$

and

$$\langle P_\lambda f, \tilde{f}\rangle = \lambda J(g, \tilde{g}) \qquad \text{for any } f = (\theta, g), \tilde{f} = (\tilde{\theta}, \tilde{g}) \in \mathcal{H}. \tag{2.7}$$

As will be seen, R_u and P_λ are two major building blocks of this enlarged RKHS framework. In particular, Propositions 2.1 and 2.2 show that these two operators are actually built upon their nonparametric counterparts K_z and W_λ defined below.

Let $K(z_1, z_2)$ be a (symmetric) reproducing kernel of $H^m(\mathbb{I})$ endowed with the inner product $\langle g, \tilde{g}\rangle_1 = E_Z\{B(Z)g(Z)\tilde{g}(Z)\} + \lambda J(g, \tilde{g})$ and norm $\|g\|_1^2 = \langle g, g\rangle_1$, where $B(Z) = E\{I(U)|Z\}$. Hence, $K_z(\cdot) \equiv K(z, \cdot)$ satisfies $\langle K_z, g\rangle_1 = g(z)$. We next specify a positive definite self-adjoint operator $W_\lambda : H^m(\mathbb{I}) \mapsto H^m(\mathbb{I})$ satisfying $\langle W_\lambda g, \tilde{g}\rangle_1 = \lambda J(g, \tilde{g})$ for any $g, \tilde{g} \in H^m(\mathbb{I})$. The existence of such W_λ is proved in Section S.2 of the supplement document [6]. Write $V(g, \tilde{g}) = E_Z\{B(Z)g(Z)\tilde{g}(Z)\}$. Hence, $\langle g, \tilde{g}\rangle_1 = V(g, \tilde{g}) + \langle W_\lambda g, \tilde{g}\rangle_1$, which implies

$$V(g, \tilde{g}) = \langle(id - W_\lambda)g, \tilde{g}\rangle_1, \tag{2.8}$$

where id denotes the identity operator. We next assume that there exists a sequence of basis functions in the space $H^m(\mathbb{I})$ that simultaneously diagonalizes the bilinear forms V and J. Such an eigensystem assumption is typical in the smoothing spline literature; see [13].

ASSUMPTION A2. There exists a sequence of real-valued functions $h_\nu \in H^m(\mathbb{I})$, $\nu \in \mathbb{N}$ satisfying $\sup_{\nu\in\mathbb{N}} \|h_\nu\|_{\sup} < \infty$ and a nondecreasing real sequence $\gamma_\nu \asymp \nu^{2m}$ such that $V(h_\mu, h_\nu) = \delta_{\mu\nu}$ and $J(h_\mu, h_\nu) = \gamma_\mu \delta_{\mu\nu}$ for any $\mu, \nu \in \mathbb{N}$, where $\delta_{\mu\nu}$ is the Kronecker's delta. Furthermore, any $g \in H^m(\mathbb{I})$ admits the Fourier expansion $g = \sum_\nu V(g, h_\nu)h_\nu$ under the $\|\cdot\|_1$-norm.

Under Assumption A2 and by $B(Z) = E\{I(U)|Z\}$, it can be seen that $E\{I(U)h_\nu(Z)h_\mu(Z)\} = V(h_\nu, h_\mu) = \delta_{\nu\mu}$. Then we can easily derive explicit expressions for $\|g\|_1$, $W_\lambda h_\nu(\cdot)$ and $K_z(\cdot)$ in terms of the h_ν and γ_ν as follows:

$$\begin{aligned} &\|g\|_1^2 = \sum_\nu \big|V(g, h_\nu)\big|^2(1 + \lambda\gamma_\nu), \qquad W_\lambda h_\nu(\cdot) = \frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu} h_\nu(\cdot) \quad \text{and} \\ &K_z(\cdot) = \sum_\nu \frac{h_\nu(z)}{1+\lambda\gamma_\nu} h_\nu(\cdot). \end{aligned} \tag{2.9}$$

Using similar arguments to those in Proposition 2.2 of [30], we know that Assumption A2 holds when Assumption A1 is satisfied and the h_νs are chosen as the (normalized) solutions of the following ODE problem:

$$\begin{aligned} &(-1)^m h_\nu^{(2m)}(\cdot) = \gamma_\nu B(\cdot)\pi(\cdot)h_\nu(\cdot), \\ &\qquad h_\nu^{(j)}(0) = h_\nu^{(j)}(1) = 0, \qquad j = m, m+1, \ldots, 2m-1. \end{aligned} \tag{2.10}$$

For example, the h_νs are constructed as an explicit trigonometric basis in case (I) of Example 5.1. As will be seen later, by employing the above ordinary differential equation (ODE) approach, we will reduce the challenging infinite-dimensional inference problems to simple exercises on finding the underlying eigensystem. We remark that proving the existence of the above eigensystem is nontrivial and relies substantially on the ODE techniques developed in [4, 33].

We next state a regularity Assumption A3 guaranteeing that R_u and P_λ are both well defined. Define $A_0(Z) = E\{I(U)X|Z\}$ and $G(Z) = A_0(Z)/B(Z)$. Note that $G = (G_1, \ldots, G_p)^T$ is a p-dimensional vector-valued function, for example, $G(Z) = E(X|Z)$ in the L_2 regression.

ASSUMPTION A3. $G_1, \ldots, G_p \in L_2(P_Z)$, that is, G_k has a finite second moment, and the $p \times p$ matrix $\Omega \equiv E\{I(U)(X - G(Z))(X - G(Z))^T\}$ is positive definite.

Under the assumption that $G_k \in L_2(P_Z)$, the linear functional $\mathcal{A}_k$ defined by $\mathcal{A}_k g = V(G_k, g)$ is bounded (or equivalently, continuous) for any $g \in H^m(\mathbb{I})$ because of the following inequality: $|\mathcal{A}_k g| \le V^{1/2}(G_k, G_k)V^{1/2}(g, g) \le V^{1/2}(G_k, G_k)\|g\|_1 < \infty$. Thus, by Riesz's representation theorem, there exists an $A_k \in H^m(\mathbb{I})$ such that $\mathcal{A}_k g = \langle A_k, g\rangle_1$ for any $g \in H^m(\mathbb{I})$. Thus if we write $A = (A_1, \dots, A_p)^T$, then

$$V(G, g) = \langle A, g\rangle_1. \tag{2.11}$$

We also note that $A = (id - W_\lambda)G$ when $G_1, \dots, G_p \in H^m(\mathbb{I})$ based on (2.8). Taking $g = K_z$ in (2.11) and applying (2.9), we find that

$$A(z) = \sum_\nu \frac{V(G, h_\nu)}{1 + \lambda\gamma_\nu} h_\nu(z) \quad \text{and} \quad (W_\lambda A)(z) = \sum_\nu \frac{V(G, h_\nu)\lambda\gamma_\nu}{(1 + \lambda\gamma_\nu)^2} h_\nu(z). \tag{2.12}$$

Now, we are ready to construct R_u and P_λ in Propositions 2.1 and 2.2, respectively. Define $\Sigma_\lambda = E_Z\{B(Z)G(Z)(G(Z) - A(Z))^T\}$ as a $p \times p$ matrix.

PROPOSITION 2.1. *R_u defined in (2.6) can be expressed as $R_u : u \mapsto (H_u, T_u) \in \mathcal{H}$, where*

$$\begin{aligned} H_u &= (\Omega + \Sigma_\lambda)^{-1}\big(x - A(z)\big) \quad \textit{and} \\ T_u &= K_z - A^T(\Omega + \Sigma_\lambda)^{-1}\big(x - A(z)\big). \end{aligned} \tag{2.13}$$

PROPOSITION 2.2. *P_λ defined in (2.7) can be expressed as $P_\lambda : (\theta, g) \mapsto (H_g^*, T_g^*) \in \mathcal{H}$, where*

$$\begin{cases} H_g^* = -(\Omega + \Sigma_\lambda)^{-1} E\{B(Z)G(Z)(W_\lambda g)(Z)\}, \\ T_g^* = E\{B(Z)G(Z)^T(W_\lambda g)(Z)\}(\Omega + \Sigma_\lambda)^{-1} A + W_\lambda g. \end{cases}$$

Note that $\lim_{\lambda\to 0} \Sigma_\lambda = 0$ according to (A.2) in the Appendix. Therefore, $(\Omega + \Sigma_\lambda)^{-1}$ above is well defined under Assumption A3. In addition, we note that P_λ is self-adjoint and bounded because of the following inequality:

$$\begin{aligned} \|P_\lambda f\| &= \sup_{\|\tilde f\|=1} \big|\langle P_\lambda f, \tilde f\rangle\big| \\ &= \sup_{\|\tilde f\|=1} \big|\lambda J(g, \tilde g)\big| \le \sqrt{\lambda J(g, g)} \sup_{\|\tilde f\|=1} \sqrt{\lambda J(\tilde g, \tilde g)} \le \|f\|. \end{aligned} \tag{2.14}$$

Finally, we derive the Fréchet derivatives of $\ell_{n,\lambda}(f)$ defined in (2.2). Let $\Delta f, \Delta f_j \in \mathcal{H}$ for $j = 1, 2, 3$. The Fréchet derivative of $\ell_{n,\lambda}(f)$ is given by

$$\begin{aligned} D\ell_{n,\lambda}(f)\Delta f &= \frac{1}{n}\sum_{i=1}^n \dot\ell_a\big(Y_i; X_i^T\theta + g(Z_i)\big)\langle R_{U_i}, \Delta f\rangle - \langle P_\lambda f, \Delta f\rangle \\ &\equiv \big\langle S_n(f), \Delta f\big\rangle - \langle P_\lambda f, \Delta f\rangle \equiv \big\langle S_{n,\lambda}(f), \Delta f\big\rangle. \end{aligned}$$

Note that $S_{n,\lambda}(\widehat{f}_{n,\lambda}) = 0$. In particular, $S_{n,\lambda}(f_0)$ is of interest, and it can be expressed as

$$S_{n,\lambda}(f_0) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i R_{U_i} - P_\lambda f_0. \tag{2.15}$$

The Frechét derivatives of $S_{n,\lambda}$ and $DS_{n,\lambda}$, denoted $DS_{n,\lambda}(f)\Delta f_1 \Delta f_2$ and $D^2 S_{n,\lambda}(f)\Delta f_1 \Delta f_2 \Delta f_3$, can be explicitly calculated as $(1/n)\sum_{i=1}^{n} \ddot{\ell}_a(Y_i; X_i^T\theta + g(Z_i))\langle R_{U_i}, \Delta f_1\rangle\langle R_{U_i}, \Delta f_2\rangle - \langle P_\lambda \Delta f_1, \Delta f_2\rangle$ and $(1/n)\sum_{i=1}^{n} \ell_a'''(Y_i; X_i^T\theta + g(Z_i))\langle R_{U_i}, \Delta f_1\rangle\langle R_{U_i}, \Delta f_2\rangle\langle R_{U_i}, \Delta f_3\rangle$, respectively. Define $S(f) = E\{S_n(f)\}$, $S_\lambda(f) = S(f) - P_\lambda f$ and $DS_\lambda(f) = DS(f) - P_\lambda$, where $DS(f)\Delta f_1 \Delta f_2 = E\{\ddot{\ell}_a(Y; X^T\theta + g(Z))\langle R_U, \Delta f_1\rangle\langle R_U, \Delta f_2\rangle\}$. Since $\langle DS_\lambda(f_0) f, \tilde{f}\rangle = -\langle f, \tilde{f}\rangle$ for any $f, \tilde{f} \in \mathcal{H}$, we have the following result:

PROPOSITION 2.3. *$DS_\lambda(f_0) = -id$, where id is the identity operator on $\mathcal{H}$.*

2.3. *Joint Bahadur representation.* This section presents the second layer of our theoretical foundation, the *joint Bahadur representation* (*JBR*). The JBR is developed based on empirical processes theory and will prove to be a powerful tool in the study of joint asymptotics.

We start with a useful lemma stating the relationship between $\|f\|$ and $\|f\|_{\sup}$, where the former $f = (\theta, g)$ and the latter $f = x^T\theta + g(z)$.

LEMMA 2.4. *There exists a constant $c_m > 0$ such that $\|R_u\| \le c_m h^{-1/2}$ and $|f(u)| \le c_m h^{-1/2}\|f\|$ for any $u \in \mathcal{U}$ and $(\theta, g) \in \mathcal{H}$. In particular, c_m does not depend on the choice of u and (θ, g). Hence $\|f\|_{\sup} \le c_m h^{-1/2}\|f\|$.*

An additional convergence-rate condition is needed to obtain JBR. Assumption A4 implies that $\widehat{f}_{n,\lambda}$ achieves the optimal rate of convergence, that is, $O_P(n^{-m/(2m+1)})$, when $\lambda = \lambda^*$.

ASSUMPTION A4. $\|\widehat{f}_{n,\lambda} - f_0\| = O_P((nh)^{-1/2} + h^m)$.

Interestingly, we show below that the above rate condition (Assumption A4) is valid for a broad range of h once Assumptions A1–A3 hold (by employing the contraction mapping idea).

PROPOSITION 2.5. *Suppose Assumptions A1–A3 are satisfied. Furthermore, as $n \to \infty$, $h = o(1)$ and $n^{-1/2}h^{-2}(\log n)(\log\log n)^{1/2} = o(1)$. Then $\|\widehat{f}_{n,\lambda} - f_0\| = O_P((nh)^{-1/2} + h^m)$.*

We remark that the optimal rate for the smoothing parameter, that is, $h^* \asymp n^{-1/(2m+1)}$, satisfies the rate conditions for h specified in Proposition 2.5 when $m > 3/2$.

The following *joint Bahadur representation* can be viewed as a nontrivial extension of the Bahadur representation [1] for parametric models by adding a functional component.

THEOREM 2.6 (Joint Bahadur representation). *Suppose that Assumptions A1 through A4 hold, $h = o(1)$ and $nh^2 \to \infty$. Recall that $S_{n,\lambda}(f_0)$ is defined in (2.15). Then we have*

$$\|\widehat{f}_{n,\lambda} - f_0 - S_{n,\lambda}(f_0)\| = O_P(a_n \log n), \tag{2.16}$$

where $a_n = n^{-1/2}((nh)^{-1/2} + h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2} + C_\ell h^{-1/2}((nh)^{-1} + h^{2m})/\log n$ and $C_\ell = \sup_{u\in\mathcal{U}} E\{\sup_{a\in\mathcal{I}} |\ell_a'''(Y;a)||U = u\}$.

The proof of Theorem 2.6 relies heavily on modern empirical process theory, and in particular a *concentration inequality* given in the supplementary material [6].

3. Joint limit distribution. As far as we are aware, the current semiparametric literature on the smoothing spline models mostly focus on the asymptotic normality of the parametric parts, and derive only rates of convergence (in estimation) for functional parts; see [13, 35, 36]. In this section, we demonstrate the joint asymptotic normality of both parametric and functional parts.

We start from a preliminary result that for any $z_0 \in \mathbb{I}$, $(\sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0^*), \sqrt{nh}(\widehat{g}_{n,\lambda} - g_0^*)(z_0))$ weakly converges to a zero-mean Gaussian vector. Unfortunately, the center $(\theta_0^*, g_0^*) \equiv f_0 - P_\lambda f_0$ is biased and the asymptotic variance is not diagonal; see Theorem A.1 in the Appendix for more technical details. Under a regularity condition on the least favorable direction [18], that is, (3.1), we can remove the estimation bias for θ; see Lemma A.2 in the Appendix. In this case, the parametric estimate $\widehat{\theta}_{n,\lambda}$ is semiparametric efficient when Y belongs to an exponential family; see Remark 3.1. However, what is more surprising is that $\widehat{\theta}_{n,\lambda}$ *and $\widehat{g}_{n,\lambda}(z_0)$ become asymptotically independent* after the bias removal procedure. We call this discovery the *joint asymptotics phenomenon*. This leads to the first main result of this paper, given in Theorem 3.1 below.

THEOREM 3.1 (Joint limit distribution). *Let Assumptions A1 through A4 be satisfied. Suppose there exists $b \in (1/(2m), 1]$ such that G_k satisfies*

$$\sum_\nu |V(G_k, h_\nu)|^2 \gamma_\nu^b < \infty \qquad \text{for any } k = 1, \ldots, p. \tag{3.1}$$

Furthermore, as $n \to \infty$, $h = o(1)$, $nh^2 \to \infty$, $a_n \log n = o(n^{-1/2}h^{1/2})$ [with a_n defined as in (2.16)], $hV(K_{z_0}, K_{z_0}) \to \sigma_{z_0}^2$ and $n^{1/2}h^{m(1+b)} = o(1)$. Then we have, for any $z_0 \in \mathbb{I}$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}\{\widehat{g}_{n,\lambda}(z_0) - g_0(z_0) + (W_\lambda g_0)(z_0)\} \end{pmatrix} \xrightarrow{d} N(0, \Psi), \tag{3.2}$$

where

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \sigma_{z_0}^2 \end{pmatrix}. \tag{3.3}$$

Furthermore, *if*

$$\lim_{n\to\infty}(nh)^{1/2}(W_\lambda g_0)(z_0) = -b_{z_0}, \tag{3.4}$$

then we have

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi\right). \tag{3.5}$$

We remark that Theorem 3.1 holds under the optimal smoothing parameter $h^* = n^{-1/(2m+1)}$. It follows from (2.9) and (2.12) that

$$\sigma_{z_0}^2 = \lim_{h\to 0}\sum_\nu \frac{h|h_\nu(z_0)|^2}{(1+\lambda\gamma_\nu)^2}. \tag{3.6}$$

It is worth pointing out that we obtain the above results without strengthening the regularity conditions used in the semiparametric literature, for example, those in [22]. We next discuss the key condition (3.1). When $b = 0$, condition (3.1) reduces to Assumption A3 that $G_k \in L_2(P_Z)$. However, we require $1/(2m) < b \le 1$ such that the Fourier coefficients $V(G_k, h_\nu)$ in (3.1) converge to zero at a faster rate than ν^{-mb} because $\gamma_\nu \asymp \nu^{2m}$; see Assumption A2. It is well known that a faster decaying rate of the Fourier coefficients $V(G_k, h_\nu)$ implies a more smooth G_k; see [11], page 1681. Therefore, condition (3.1) requires more smoothness of G_k. In fact, it follows from [11] that (3.1) is equivalent to $G_k \in H^{mb}(\mathbb{I})$ with $1/(2m) < b \le 1$. Hence, the condition $G_k \in H^m(\mathbb{I})$ assumed in the classical semiparametric work by Mammen and van de Geer [22] may actually be weakened.

We next discuss three important consequences of Theorem 3.1. First, the asymptotic independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$ greatly facilitates the construction of the joint CI for $(\theta_0, g_0(z_0))$ by directly building on the marginal CIs. Second, based on Theorem 3.1 and the Delta method, we can easily establish the prediction interval for a new response Y_{new} given future data $u_0 = (x_0, z_0)$ and the CI for some real-valued smooth function of $(\theta_0, g_0(z_0))$; see Section 5. Finally, the nonparametric estimation bias b_{z_0} can be further removed under an additional assumption; see Corollary 3.2.

In Remarks 3.1 and 3.2 below, we compare the marginal limit distributions implied by Theorem 3.1 with those derived in the semiparametric [22] and nonparametric [30] literature.

REMARK 3.1. Our parametric limit distribution is $\sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \xrightarrow{d} N(0, \Omega^{-1})$, where $\Omega = E\{I(U)(X - G(Z))(X - G(Z))^T\}$. We find that it is exactly the same as that obtained in [22]; see Section S.14 of supplementary document

[6]. Mammen and van de Geer [22] further showed that the parametric estimate is semiparametric efficient when Y belongs to an exponential family; see their Remark 4.1. For example, in the partially linear models under Gaussian errors, Ω reduces to the semiparametric efficiency bound $E(X - E(X|Z))^{\otimes 2}$; see [18]. Note the profile approach in [22] treats g as a nuisance parameter, and thus it cannot be adapted to obtain our joint limiting distribution.

REMARK 3.2. Our (pointwise) nonparametric limit distribution, that is, $\sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \xrightarrow{d} N(b_{z_0}, \sigma_{z_0}^2)$, is in general different from that obtained in the nonparametric smoothing spline setup (without θ) in terms of different values of b_{z_0} and $\sigma_{z_0}^2$; see [30]. This is mainly due to the eigensystem difference in the two setups; see Remark 5.1 for more illustrations. An exception is the L_2 regression in which the two eigensystems coincide. Our general finding gives a counter-example to the common intuition in the literature that the nonparametric limit distribution is not affected by the involvement of a parametric component that is estimated at a faster convergence rate.

To further illustrate Theorem 3.1, we consider the partial smoothing spline model with unit error variance (Example 5.1) and the shape-rate Gamma model with unit shape (Example 5.2), which share the same joint limit distribution with an explicit covariance matrix Ψ.

COROLLARY 3.2 (Joint limit distribution for partial smoothing spline model and shape-rate gamma model). *Let $m > 1 + \sqrt{3}/2 \approx 1.866$, and $h \asymp h^*$. Suppose that (3.1) holds for some $1 \geq b > 1/(2m)$, and $E(X - E(X|Z))^{\otimes 2}$ is positive definite. Furthermore, $g_0 \in H^m(\mathbb{I})$ satisfies $\sum_\nu |V(g_0, h_\nu)|\nu^m < \infty$. Then, as $n \to \infty$,*

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}\big(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)\big) \end{pmatrix} \xrightarrow{d} N(0, \Psi), \tag{3.7}$$

where

$$\Psi = \begin{pmatrix} \{E[X - E(X|Z)]^{\otimes 2}\}^{-1} & 0 \\ 0 & \dfrac{\int_0^\infty (1 + x^{2m})^{-2}\, dx}{\pi} \end{pmatrix}.$$

In Corollary 3.2, we notice that the nonparametric estimation bias asymptotically vanishes. This is due to the condition $\sum_\nu |V(g_0, h_\nu)|\nu^m < \infty$, which imposes additional smoothness on $g_0 \in H^m(\mathbb{I})$. Therefore, convergence rate $n^{-m/(2m+1)}$ for $\widehat{g}_{n,\lambda}(z_0)$ is actually sub-optimal given this additional smoothness (under $\lambda = \lambda^*$). In practice, we select the smoothing parameter based on CV or GCV; see [13].

4. Joint hypothesis testing. In this section, we propose likelihood ratio testing for a set of joint local hypotheses in a general form (4.1). Under very general conditions, the null limit distribution is proved to be a mixture of a Chi-square distribution with p degrees of freedom and a scaled noncentral Chi-square distribution with one degree of freedom. Obviously, these two Chi-square distributions are contributed by the parametric and nonparametric components, respectively. Hence, we reveal a new version of the Wilks phenomenon [12, 38] which adapts to the semi-nonparametric context. We further give more explicit null limit distributions for three commonly used joint hypotheses. A key technical tool used in this section is a *restricted* version of JBR.

Consider the following joint hypothesis:

$$H_0: M\theta + Qg(z_0) = \alpha \quad \text{vs.} \quad H_1: M\theta + Qg(z_0) \neq \alpha, \tag{4.1}$$

where $M = (M_1^T, \dots, M_k^T)^T$ is a $k \times p$ matrix with $k \leq p+1$, $Q = (q_1, \dots, q_k)^T$ and the α are k-vectors. Without loss of generality, we assume $N \equiv (M, Q)$ to have elements in $\mathbb{I} = [0, 1]$. We further assume that the matrix N has full rank. M, Q and α are all prespecified according to the testing needs. For example, when N is the identity matrix I_{p+1} and $\alpha = (\theta_0^T, w_0)^T$, H_0 reduces to $(\theta^T, g(z_0))^T = (\theta_0^T, w_0)^T$. See Corollary 4.6 for more examples. This provides another way to construct the joint CIs for $(\theta_0^T, g_0(z_0))^T$ without estimating Ω^{-1} or σ_{z_0}. The simultaneous testing of two marginal hypotheses, that is, $H_0^P: \theta = \theta_0$ and $H_0^N: g(z_0) = w_0$, can also be used for this purpose, but it requires the very conservative Bonferroni correction. Moreover, our joint hypothesis is more general, and the testing approach is more straightforward to implement.

To define the likelihood ratio statistic, we establish the constrained estimate under (4.1) in three steps: (i) arbitrarily choose $(\theta^\dagger, w^\dagger) \in \mathbb{R}^p \times \mathbb{R}$ satisfying $M\theta^\dagger + Qw^\dagger = \alpha$; (ii) define $\widehat{f}_{n,\lambda}^0 \equiv (\widehat{\theta}_{n,\lambda}^0, \widehat{g}_{n,\lambda}^0) = \arg\max_{f \in \mathcal{H}_0} L_{n,\lambda}(f)$, where $\mathcal{H}_0 \equiv \{(\theta, g) \in \mathcal{H} | M\theta + Qg(z_0) = 0\}$ and

$$L_{n,\lambda}(f) = n^{-1} \sum_{i=1}^n \ell(Y_i; X_i^T\theta + g(Z_i) + X_i^T\theta^\dagger + w^\dagger) - (1/2)\lambda J(g, g); \tag{4.2}$$

(iii) define the constrained estimate as $\widehat{f}_{n,\lambda}^{H_0} = (\widehat{\theta}_{n,\lambda}^0 + \theta^\dagger, \widehat{g}_{n,\lambda}^0 + w^\dagger)$. Then, the LRT statistic is $\mathrm{LRT}_{n,\lambda} = \ell_{n,\lambda}(\widehat{f}_{n,\lambda}^{H_0}) - \ell_{n,\lambda}(\widehat{f}_{n,\lambda})$.

Given the inner product $\langle \cdot, \cdot \rangle$, we note that $\mathcal{H}_0$ is a closed subset in $\mathcal{H}$ and thus a Hilbert space. Hence, we will construct the projections of the two operators R_u and P_λ (associated with $\mathcal{H}$) onto the subspace $\mathcal{H}_0$, denoting them R_u^0 and P_λ^0, respectively. Lemma 4.1 below provides a preliminary step for the construction. Its proof is similar to that of Proposition 2.1 and is thus omitted.

LEMMA 4.1. *For any $u = (x, z) \in \mathcal{U}$ and $q \in \mathbb{I}$, define*

$$H_{q,u} = (\Omega + \Sigma_\lambda)^{-1}(x - qA(z)) \quad \text{and} \quad T_{q,u} = qK_z - A^T H_{q,u}.$$

Let $R_{q,u} \equiv (H_{q,u}, T_{q,u}) \in \mathcal{H}$. Then, for any $f \in \mathcal{H}$ and $u \in \mathcal{U}$, we have $\langle R_{q,u}, f\rangle = x^T\theta + qg(z)$.

Obviously, $R_{q,u}$ is a generalization of R_u defined in Proposition 2.1, that is, $R_u = R_{1,u}$. Lemma 4.1 implies that the restricted parameter space $\mathcal{H}_0$ can be rewritten as

$$\mathcal{H}_0 = \{f = (\theta, g) \in \mathcal{H} | \langle R_{q_j,W_j}, f\rangle = 0, j = 1, \dots, k\}, \tag{4.3}$$

where $W_j = (M_j, z_0)$. Define $H(Q, W) = (H_{q_1,W_1}, \dots, H_{q_k,W_k})$, $T(Q, W) = (T_{q_1,W_1}(z_0), \dots, T_{q_k,W_k}(z_0))$ and $M_K = MH(Q, W) + QT(Q, W)$. Construct the projections

$$R_u^0 = R_u - \sum_{j=1}^{k} \rho_{u,j} R_{q_j,W_j} \quad \text{and} \quad P_\lambda^0 f = P_\lambda f - \sum_{j=1}^{k} \zeta_j(f) R_{q_j,W_j},$$

where $(\rho_{u,1}, \dots, \rho_{u,k})^T = M_K^{-1}(MH_u + QT_u(z_0))$ and $(\zeta_1(f), \dots, \zeta_k(f))^T = M_K^{-1}(MH_g^* + QT_g^*(z_0))$. Recall that $R_u : u \mapsto (H_u, T_u)$ and $P_\lambda : (\theta, g) \mapsto (H_g^*, T_g^*)$ in Proposition 2.1. The invertibility of M_K is given in the proof of Proposition 4.2 below.

Proposition 4.2 below says that R_u^0 and P_λ^0 defined above are indeed what we need.

PROPOSITION 4.2. *Let $f = (\theta, g)$ and $\tilde{f} = (\tilde{\theta}, \tilde{g})$. For any $u = (x, z) \in \mathbb{I}^p \times \mathbb{I}$, $f, \tilde{f} \in \mathcal{H}_0$, we have $\langle R_u^0, f\rangle = x^T\theta + g(z)$ and $\langle P_\lambda^0 f, \tilde{f}\rangle = \lambda J(g, \tilde{g})$.*

Based on Proposition 4.2, we can write down the Fréchet derivatives of $L_{n,\lambda}$ defined in (4.2) under $\mathcal{H}_0$ by modifying those of $\ell_{n,\lambda}$ as follows: replace θ, g, R_U and P_λ by $\theta + \theta^\dagger$, $g + w^\dagger$, R_U^0 and P_λ^0. For example,

$$\begin{aligned} &DL_{n,\lambda}(f)\Delta f \\ &\quad = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_a(Y_i; X_i^T\theta + g(Z_i) + X_i^T\theta^\dagger + w^\dagger)\langle R_{U_i}^0, \Delta f\rangle - \langle P_\lambda^0 f, \Delta f\rangle \\ &\quad \equiv \langle S_n^0(f), \Delta f\rangle - \langle P_\lambda^0 f, \Delta f\rangle = \langle S_{n,\lambda}^0(f), \Delta f\rangle. \end{aligned}$$

Similarly, we have $S_{n,\lambda}^0(\widehat{f}_{n,\lambda}^0) = 0$. Also define $S^0(f) = E\{S_n^0(f)\}$ and $S_\lambda^0(f) = S^0(f) - P_\lambda^0(f)$. For the second derivative, we have $DS_{n,\lambda}^0(f)\Delta f_1\Delta f_2 = D^2L_{n,\lambda}(f)\Delta f_1\Delta f_2$ and $DS_\lambda^0(f)\Delta f_1\Delta f_2 = DS^0(f)\Delta f_1\Delta f_2 - \langle P_\lambda^0\Delta f_1, \Delta f_2\rangle$, where

$$DS^0(f)\Delta f_1\Delta f_2 = E\{\ddot{\ell}_a(Y; X^T\theta + g(Z) + X^T\theta^\dagger + w^\dagger)\langle R_U^0, \Delta f_1\rangle\langle R_U^0, \Delta f_2\rangle\}.$$

In Theorem 4.3 below, we present a new version of JBR that is restricted to the subspace $\mathcal{H}_0$. We need an additional Assumption A5 here. Let $f_0^0 \equiv (\theta_0 - \theta^\dagger, g_0 -$

$w^\dagger$), which belongs to $\mathcal{H}_0$ under H_0. Assumption A5 holds under mild conditions similar to those specified in Proposition 2.5. The proof can be similarly conducted by replacing the space $\mathcal{H}$ by $\mathcal{H}_0$, and thus, is omitted.

ASSUMPTION A5. Under H_0 specified in (4.1),

$$\|\widehat{f}^0_{n,\lambda} - f^0_0\| = O_P\big((nh)^{-1/2} + h^m\big).$$

THEOREM 4.3 (Restricted joint Bahadur representation). *Suppose that Assumptions A1, A2, A3 and A5 hold and that $h = o(1)$ and $nh^2 \to \infty$ as $n \to \infty$. Under H_0 specified in (4.1), we have $\|\widehat{f}^0_{n,\lambda} - f^0_0 - S^0_{n,\lambda}(f^0_0)\| = O_P(a_n \log n)$, where a_n is defined as in (2.16).*

Given the above preparatory results, we are ready to present general results for the null limit distribution of $-2n \cdot \mathrm{LRT}_{n,\lambda}$ in Theorem 4.4. Define $r_n = (nh)^{-1/2} + h^m$, and let

$$\Phi_\lambda = \Lambda N^T M_K^{-1} N \Lambda^T,$$

where

$$\Lambda = \begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0, z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix}.$$

THEOREM 4.4 (Joint local testing). *Suppose that Assumptions A1 through A5 are satisfied, there exists $b \in (1/(2m), 1]$ such that G_k satisfies (3.1), and $h = o(1)$, $nh^2 \to \infty$, $n^{1/2}h^{m(1+b)} = o(1)$, $r_n^2 h^{-1/2} = o(a_n)$ and $a_n = o(\min\{r_n, n^{-1}r_n^{-1}(\log n)^{-1}, n^{-1/2}h^{1/2}(\log n)^{-1}\})$, where a_n is defined as in (2.16). Furthermore, for any $z_0 \in [0, 1]$, $\lim_{h\to 0} \sqrt{n}(W_\lambda g_0)(z_0)/\sqrt{K(z_0, z_0)} = c_{z_0}$, $\lim_{h\to 0} \Phi_\lambda = \Phi_0$, where Φ_0 is a fixed $(p+1) \times (p+1)$ positive semidefinite matrix, and*

$$\lim_{h\to 0} hV(K_{z_0}, K_{z_0}) \to \sigma^2_{z_0} > 0, \tag{4.4}$$

$$\lim_{h\to 0} E_Z\{B(Z)|K_{z_0}(Z)|^2\}/K(z_0, z_0) \equiv c_0 \in (0, 1]. \tag{4.5}$$

Under H_0 specified in (4.1), we obtain: (i) $\|\widehat{f}_{n,\lambda} - \widehat{f}^{H_0}_{n,\lambda}\| = O_P(n^{-1/2})$; (ii) $-2n \times \mathrm{LRT}_{n,\lambda} = n\|\widehat{f}_{n,\lambda} - \widehat{f}^{H_0}_{n,\lambda}\|^2 + o_P(1)$;

$$\text{(iii)} -2n \cdot \mathrm{LRT}_{n,\lambda} \xrightarrow{d} \upsilon^T \Phi_0 \upsilon, \tag{4.6}$$

where $\upsilon \sim N\big(\big(\begin{smallmatrix} 0 \\ c_{z_0} \end{smallmatrix}\big), \big(\begin{smallmatrix} I_p & 0 \\ 0 & c_0 \end{smallmatrix}\big)\big)$.

The *parametric* convergence-rate result proved in (i) of Theorem 4.4 is reasonable since the null hypothesis imposes only a finite-dimensional constraint. By (2.9), it can be explicitly shown that

$$c_0 = \lim_{\lambda\to 0} \frac{Q_2(\lambda, z_0)}{Q_1(\lambda, z_0)} \qquad \text{where } Q_l(\lambda, z) \equiv \sum_{\nu\in\mathbb{N}} \frac{|h_\nu(z)|^2}{(1+\lambda\gamma_\nu)^l} \text{ for } l = 1, 2. \tag{4.7}$$

It is well known that the reproducing kernel K is uniquely determined for any Hilbert space if it exists; see [28], page 38. This implies that c_0 defined in (4.5) is also uniquely determined. Therefore, different choices of (h_ν, γ_ν) in (4.7) will give exactly the same value of c_0, although a particular choice may facilitate the computation of c_0. For example, in case (I) of Example 5.1, we can explicitly calculate c_0 as 0.75 (0.83) when $m = 2$ (3) by choosing the trigonometric basis (5.2).

The null limit distribution derived in Theorem 4.4 cannot be directly used for inference because of the nontrivial estimation of c_{z_0}. Hence, in Corollary 4.5, we present a set of conditions under which the estimation bias of $\widehat{g}_{n,\lambda}$ can be removed, and thus $c_{z_0} = 0$.

COROLLARY 4.5. *Suppose that Assumptions A1 through A5 are satisfied, and hypothesis H_0 holds. Let $m > 1 + \sqrt{3}/2 \approx 1.866$ and $G_1, \dots, G_p$ satisfy (3.1) with $1/(2m) < b \le 1$. Also assume that the Fourier coefficients $\{V(g_0, h_\nu)\}_{\nu\in\mathbb{N}}$ of g_0 satisfy $\sum_\nu |V(g_0, h_\nu)|\gamma_\nu^{1/2} < \infty$. Furthermore, if Φ_λ converges to some fixed $(p+1)\times(p+1)$ positive semidefinite matrix, that is, Φ_0, and (4.4) and (4.5) are both satisfied for any $z_0 \in [0, 1]$, then (4.6) holds with $c_{z_0} = 0$ given that $h = h^* \asymp n^{-1/(2m+1)}$.*

Combining Theorem 4.4 with Corollary 4.5, we immediately obtain Corollary 4.6, which gives null limit distributions of the three commonly assumed joint hypotheses.

COROLLARY 4.6. *Suppose that the conditions in Corollary 4.5 hold. We have:*

(I) $H_0 : \theta = \theta_0$ *and* $g(z_0) = w_0$:

$$-2n \cdot \mathrm{LRT}_{n,\lambda} \xrightarrow{d} \chi_p^2 + c_0\chi_1^2,$$

where the two Chi-square distributions are independent. In this case, $N = I_{p+1}$, $\alpha = (\theta_0^T, w_0)^T$ *and* $\Phi_\lambda = \Phi_0 = I_{p+1}$.

(II) $H_0 : D\theta = \theta_0'$ *and* $g(z_0) = w_0$ [*D is an $r \times p$ matrix with $0 < r \le p$ and* $\mathrm{rank}(D) = r$, *θ_0' is an r-vector with $0 < r < p$*]:

$$-2n \cdot \mathrm{LRT}_{n,\lambda} \xrightarrow{d} \chi_r^2 + c_0\chi_1^2,$$

where the two Chi-square distributions are independent. In this case, $N = \begin{pmatrix} D & 0_r \\ 0_p^T & 1 \end{pmatrix}$, $\alpha = (\theta_0'^T, w_0)^T$ *and* $\Phi_0 = \begin{pmatrix} \mathcal{P}_r & 0_p \\ 0_p^T & 1 \end{pmatrix}$ *with the projection matrix (of rank r)* $\mathcal{P}_r = \Omega^{-1/2}D^T(D\Omega^{-1}D^T)^{-1}D\Omega^{-1/2}$.

(III) $H_0 : x_0^T\theta + g(z_0) = \alpha$ (*α, x_0 and z_0 are given*):

$$-2n \cdot \mathrm{LRT}_{n,\lambda} \xrightarrow{d} c_0\chi_1^2.$$

In this case, $N = (x_0^T, 1)$ *and* $\Phi_0 = \begin{pmatrix} 0_{p\times p} & 0_p \\ 0_p^T & 1 \end{pmatrix}$.

The independence between the two Chi-square distributions in (I) and (II) follows from the joint asymptotics phenomenon that $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$ are asymptotically independent. In comparison with (I) and (II), we note that the null limit distribution in (III) is dominated by the effect from $g(z_0)$ because of its nonparametric nature, that is, its slower convergence rate.

As far as we are aware, Corollary 4.6 is a new version of the Wilks phenomenon [12, 38] that adapts to the semi-nonparametric context. Note that the value of c_0 converges to one as $m \to \infty$. Therefore, this new type of Wilks phenomenon reverts to the classical version in the parametric setup as $m \to \infty$ by further consideration of the independence of the two Chi-squares. For example, the null limit distribution in (I) of Corollary 4.6 becomes χ^2_{p+1} as $m \to \infty$.

In the end of this section, we apply Theorem 4.4 to partial smoothing spline models (Example 5.1) and shape-rate gamma models (Example 5.2). For simplicity, let $0 < z_0 < 1$. Corollary 4.7 directly follows from Corollary 4.5 and equivalent kernel theory [24, 26].

COROLLARY 4.7 (Joint local testing for partial smoothing spline model and shape-rate gamma model). *Suppose that the hypothesis H_0 specified in* (I) [(II) *or* (III)] *in Corollary* 4.6 *holds. Let $m > 1 + \sqrt{3}/2 \approx 1.866$, $G_1, \dots, G_p$ satisfy* (3.1) *with $1/(2m) < b \le 1$ and $h \asymp h^*$. Also assume that $g_0 \in H^m(\mathbb{I})$ satisfies $\sum_\nu |V(g_0, h_\nu)|\nu^m < \infty$, and $E(X - E(X|Z))^{\otimes 2}$ is positive definite. Then, as $n \to \infty$, the conclusion of* (I) [(II) *or* (III)] *in Corollary* 4.6 *holds with $c_0 = \frac{\pi(z_0)\int_{\mathbb{R}} \omega_0(t)^2\,dt}{\omega_0(0)}$, where the equivalent kernel function ω_0 is specified in* [24], *page* 184. *In particular, when $m = 2$ (3) and the design is uniform, $c_0 = 0.75$* (0.83).

As for the logistic regression model (Example 5.3), we need to numerically approximate the value of c_0 due to the implicit forms of the eigenfunctions and eigenvalues; see more detailed discussions in Section S.15 of the supplementary file [6].

5. Examples. In this section, we present three concrete examples together with simulations. In all the examples, the G_ks are sufficiently smooth for Theorem 3.1 and Corollary 4.6 to apply. Detailed assumption verifications for three examples can be found in Sections S.9, S.13 and S.15 of [6].

EXAMPLE 5.1 (Partial smoothing spline). Consider a partially linear regression model

$$Y = X^T\theta + g(Z) + \epsilon, \tag{5.1}$$

where $\epsilon \sim N(0, \sigma^2)$ with an unknown σ^2. Hence, $B(Z) = \sigma^{-2}$. For simplicity, Z is assumed to be uniformly distributed over $\mathbb{I}$. In this case, $V(g, \tilde{g})$ becomes the usual L^2-norm. The function *ssr*() in the R package *assist* was used to select the smoothing parameter λ based on CV or GCV; see [16]. The unknown error variance can be consistently estimated by $\widehat{\sigma}^2 = n^{-1}\sum_i (Y_i - X_i^T\widehat{\theta}_{n,\lambda} - \widehat{g}_{n,\lambda}(Z_i))^2/(n-\mathrm{trace}(A(\lambda)))$, where $A(\lambda)$ denotes the smoothing matrix; see [35].

We next consider two separate cases: (I) $g \in H_0^m(\mathbb{I})$ and (II) $g \in H^m(\mathbb{I})$.

Case (I) $g \in H_0^m(\mathbb{I})$: We choose the following trigonometric eigensystem for $H_0^m(\mathbb{I})$:

$$h_\mu(z) = \begin{cases} \sigma, & \mu = 0, \\ \sqrt{2}\sigma\cos(2\pi kz), & \mu = 2k, k = 1, 2, \dots, \\ \sqrt{2}\sigma\sin(2\pi kz), & \mu = 2k-1, k = 1, 2, \dots, \end{cases} \tag{5.2}$$

with the corresponding γ_ν specified as $\gamma_0 = 0$ and $\gamma_{2k-1} = \gamma_{2k} = \sigma^2(2\pi k)^{2m}$ for $k \geq 1$.

It follows from (3.6) and (5.2) that the asymptotic variance of $\widehat{g}_{n,\lambda}(z_0)$ is expressed as

$$\sigma_{z_0}^2 = \lim_{h\to 0}\left\{\sigma^2 h\left(1 + 2\sum_{k=1}^{\infty}\frac{1}{(1+(2\pi h\sigma^{1/m}k)^{2m})^2}\right)\right\}.$$

Lemma 6.1 in [30] leads to, for $l = 1, 2$,

$$\sum_{k=1}^{\infty}\frac{1}{(1+(2\pi h\sigma^{1/m}k)^{2m})^l} \sim \frac{I_l}{2\pi h\sigma^{1/m}}, \tag{5.3}$$

where $I_l = \int_0^\infty (1+x^{2m})^{-l}\,dx$. Therefore, we have $\sigma_{z_0}^2 = (I_2\sigma^{2-1/m})/\pi$. According to Corollary 3.2, the 95% prediction interval for Y at a new observed covariate $u_0 = (x_0, z_0)$ is

$$\widehat{Y} \pm 1.96\sqrt{\widehat{\sigma}^{2-1/m} I_2/(\pi nh) + \widehat{\sigma}^2}, \tag{5.4}$$

where $\widehat{Y} = x_0^T\widehat{\theta}_{n,\lambda} + \widehat{g}_{n,\lambda}(z_0)$ is the predicted response. We next calculate c_0 based on (4.7). It follows from (5.2) and (5.3) that

$$\begin{aligned} Q_l(\lambda, z_0) &= \sigma^2 + \sum_{k\geq 1}\left\{\frac{|h_{2k}(z_0)|^2}{(1+\lambda\sigma^2(2\pi k)^{2m})^l} + \frac{|h_{2k-1}(z_0)|^2}{(1+\lambda\sigma^2(2\pi k)^{2m})^l}\right\} \\ &= \sigma^2 + 2\sigma^2\sum_{k\geq 1}\frac{1}{(1+\lambda\sigma^2(2\pi k)^{2m})^l} \\ &= \sigma^2 + 2\sigma^2\sum_{k\geq 1}\frac{1}{(1+(2\pi h\sigma^{1/m}k)^{2m})^l} \sim \frac{I_l}{\pi h\sigma^{1/m}} \end{aligned}$$

for $l = 1, 2$. Hence we obtain

$$c_0 = I_2/I_1. \tag{5.5}$$

Further calculations reveal that $c_0 = 0.75$ (0.83) when $m = 2$ (3).

In the simulations, we first verify the joint asymptotics phenomenon, that is, (3.5), by investigating the (asymptotic) independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$. Let $\theta_0 = (8, -8)^T$ and $g_0(z) = 0.6\beta_{30,17}(z) + 0.4\beta_{3,11}(z)$, where $\beta_{a,b}$ is the density function for Beta(a, b). We estimate the nonparametric function g_0, which has many peaks and troughs, using periodic splines with $m = 2$; σ is set to one. To allow the linear and nonlinear covariates (X, Z) to be dependent, we generate them as follows: generate $U, V, Z \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$, and set $X_1 = (U + 0.2Z)/1.2$, $X_2 = (V + 0.2Z)/1.2$. This leads to $\text{corr}(X_1, Z) = \text{corr}(X_2, Z) \approx 0.20$, where corr denotes the correlation coefficient. The dependence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z)$ is evaluated through the absolute values of the sample correlation coefficients (ACC) between $\widehat{\theta}_{n,\lambda} = (\widehat{\theta}_{n,\lambda,1}, \widehat{\theta}_{n,\lambda,2})^T$ and $\widehat{g}_{n,\lambda}(z)$ at ten evenly spaced z points in $[0, 1]$ based on 500 replicated data sets. The results are summarized in Figure 1 for sample sizes $n = 100, 300, 1000$. As n increases, it is easy to see that the ACC curves become uniformly closer to zero, which strongly indicates the desired asymptotic independence.

To examine the performance of the 95% prediction intervals (5.4), we calculate the proportions of the prediction intervals covering the future response Y generated from model (5.1), that is, the coverage proportion. The simulation setup is the same as before, except that we assume a one-dimensional $\theta_0 = 4$ for simplicity. The new covariates are (x_0, z_0) with $x_0 = 1/4, 2/4, 3/4$ and z_0 being thirty evenly spaced points in $[0, 1]$. The coverage proportions are calculated based on

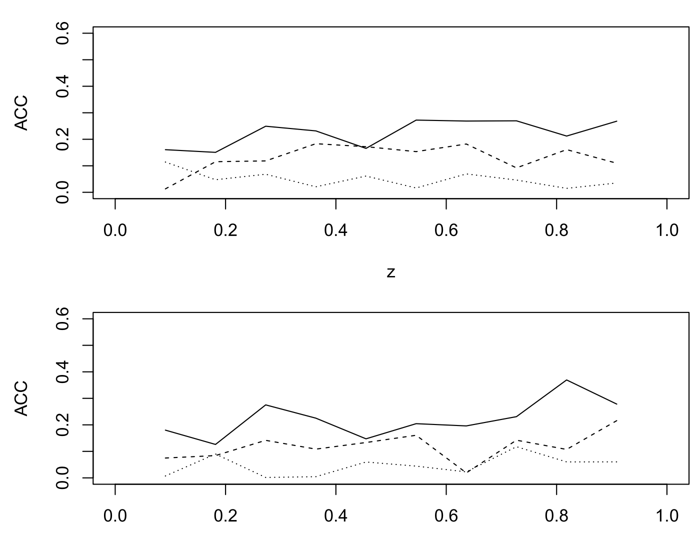

FIG. 1. *Absolute values of correlation coefficients* (*ACC*) *between* $\widehat{\theta}_{n,\lambda,1}$ *and* $\widehat{g}_{n,\lambda}(z)$ (*the upper plot*), *and* $\widehat{\theta}_{n,\lambda,2}$ *and* $\widehat{g}_{n,\lambda}(z)$ (*the lower plot*), *at ten evenly spaced nonlinear covariates in case* (I) *of Example* 5.1. *The three lines correspond to three sample sizes*: $n = 100$ (*solid*), $n = 300$ (*dashed*), $n = 1000$ (*dotted*).

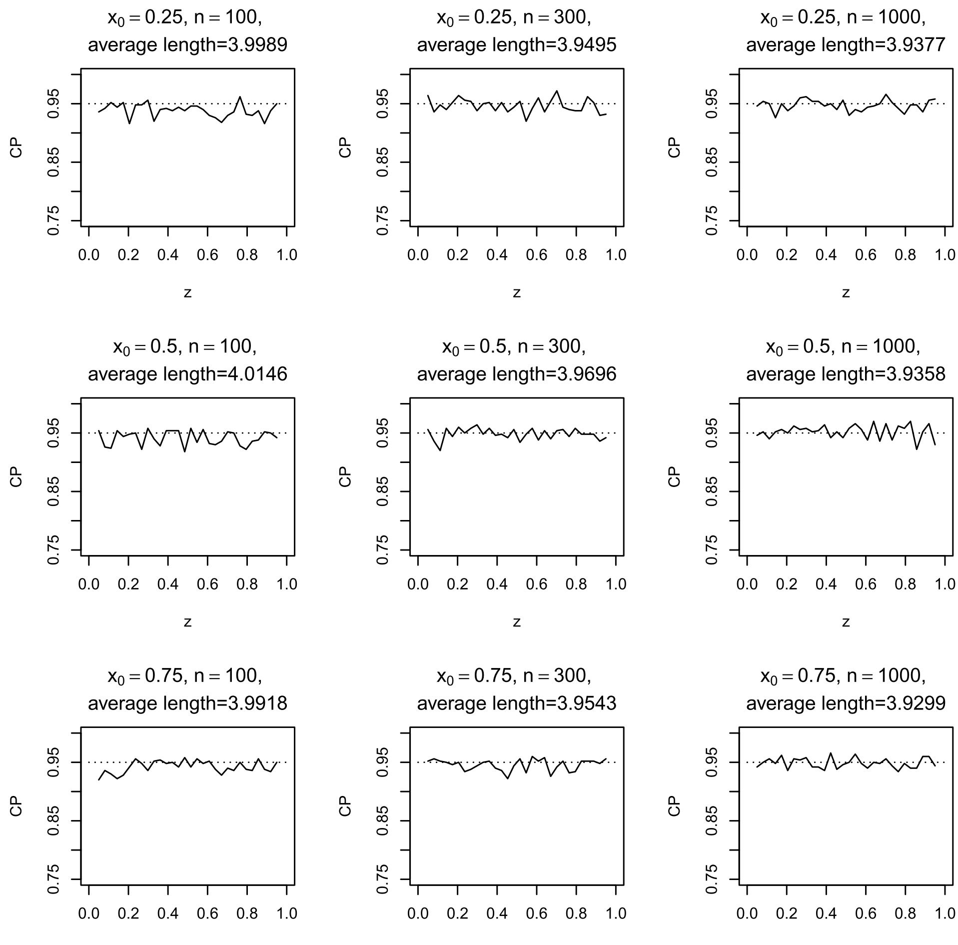

FIG. 2. *Coverage proportion of* 95% *prediction intervals in case* (I) *of Example* 5.1.

500 replications. We summarize our simulation results in Figure 2 for sample sizes $n = 100, 300, 1000$. As n grows, all the coverage proportions approach the nominal level, 95%. In addition, the prediction interval lengths approach the theoretical value indicated in formula (5.4), that is, $2 \times 1.96 = 3.92$.

Finally, we test $H_0 : x_0\theta + g(z_0) = 0$. The true parameters are chosen as $\theta_0 = -4$, $g_0(z) = \sin(\pi z)$ and $\sigma = 1$. The performance is demonstrated by calculating the powers for the nine combinations of $x_0 = 1/4, 2/4, 3/4$ and $z_0 = 1/4, 2/4, 3/4$ through 500 replicated data sets. In particular, H_0 is true when $x_0 = 1/4$ and $z_0 = 2/4$, and H_0 is false at the other values of (x_0, z_0). The results are summarized in Table 1 for sample sizes $n = 50, 100, 300, 500, 1000, 1500$. We observe that when $x_0 = 1/4$ and $z_0 = 2/4$, the power approaches the correct size 5%, while at the other values of (x_0, z_0), where H_0 does not hold, the power approaches one. This

TABLE 1

$100\times$ *power of the local LRT test for nine combinations of* x_0 *and* z_0 *for case* (I) *of Example* 5.1

	$n = 50$	$n = 100$	$n = 300$	$n = 500$	$n = 1000$	$n = 1500$
$x_0 = 1/4$						
$z_0 = 1/4$	43.00	56.60	77.60	90.40	97.80	98.60
$z_0 = 2/4$	20.60	13.00	7.20	7.00	5.60	5.10
$z_0 = 3/4$	42.00	50.00	77.60	89.60	97.80	99.20
$x_0 = 2/4$						
$z_0 = 1/4$	98.60	99.80	100	100	100	100
$z_0 = 2/4$	96.80	99.00	100	100	100	100
$z_0 = 3/4$	98.80	99.80	100	100	100	100
$x_0 = 3/4$						
$z_0 = 1/4$	99.80	100	100	100	100	100
$z_0 = 2/4$	99.60	100	100	100	100	100
$z_0 = 3/4$	99.60	100	100	100	100	100

shows the validity of our local LRT test. The detailed computational algorithm for the constrained estimate under H_0 is given in Section S.16 of the supplementary document [6].

Case (II) $g \in H^m(\mathbb{I})$: For this larger parameter space, we first construct an effective eigensystem that satisfies (2.10). Let $\tilde{h}_\nu$s and $\tilde{\gamma}_\nu$s be the normalized (with respect to the usual L_2-norm) eigenfunctions and eigenvalues of the boundary value problem $(-1)^m \tilde{h}_\nu^{(2m)} = \tilde{\gamma}_\nu \tilde{h}_\nu$, $\tilde{h}_\nu^{(j)}(0) = \tilde{h}_\nu^{(j)}(1) = 0$, $j = m, m+1, \ldots, 2m-1$. Then we can construct $h_\nu = \sigma \tilde{h}_\nu$ and $\gamma_\nu = \sigma^2 \tilde{\gamma}_\nu$. Consequently,

$$
\begin{aligned}
Q_l(\lambda, z) &= \sum_\nu \frac{|h_\nu(z)|^2}{(1+\lambda\gamma_\nu)^l} \\
&= \sigma^{2-1/m} h^{-1} \sum_\nu \frac{h\sigma^{1/m} |\tilde{h}_\nu(z)|^2}{(1+(h\sigma^{1/m})^{2m}\tilde{\gamma}_\nu)^l} \sim \sigma^{2-1/m} h^{-1} c_l(z),
\end{aligned}
\tag{5.6}
$$

where $c_l(z) = \lim_{h^\dagger \to 0} \sum_\nu \frac{h^\dagger |\tilde{h}_\nu(z)|^2}{(1+(h^\dagger)^{2m}\tilde{\gamma}_\nu)^l}$ and $h^\dagger = h\sigma^{1/m}$, for $l = 1, 2$. Hence, by (4.7), we have $c_0 = c_2(z_0)/c_1(z_0)$. In addition, by (3.6), we obtain the asymptotic variance of $\widehat{g}_{n,\lambda}(z_0)$ as $\sigma^{2-1/m} c_2(z_0)$, implying the following 95% prediction interval:

$$
\widehat{Y} \pm 1.96 \sqrt{\widehat{\sigma}^{2-1/m} c_2(z_0)/(nh) + \widehat{\sigma}^2}.
$$

The above discussion applies to general m. However, when $m = 2$, we can avoid estimating the $c_l(z_0)$s required in the inference by applying the equivalent kernel approach. Following the discussion in [30], we can actually obtain the same values of c_0 and $\sigma_{z_0}^2$ as in case (I). The simulation setup is the same as before

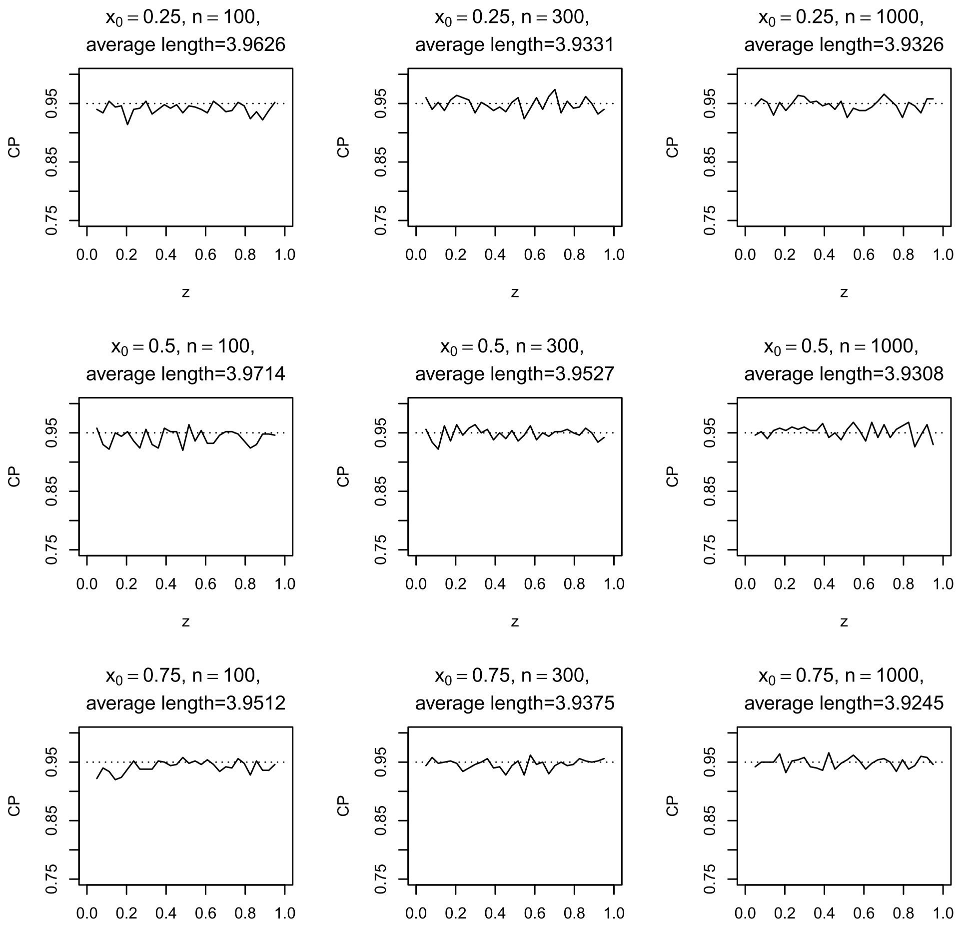

FIG. 3. *Coverage proportion of* 95% *prediction intervals in case* (II) *of Example* 5.1.

except that a different (nonperiodic) $g_0(z) = \sin(2.8\pi z)$ is used. Figure 3 displays the coverage proportion of the 95% prediction intervals for three sample sizes $n = 100, 300, 1000$. As n grows, all the coverage proportions approach the 95% nominal level, and the prediction interval lengths approach the theoretical value 3.92.

EXAMPLE 5.2 (Semiparametric gamma model). Consider a two-parameter exponential model

$$Y|X, Z \sim \text{Gamma}\big(\alpha, \exp\big(X^T\theta_0 + g_0(Z)\big)\big),$$

where $\alpha > 0$ is known, $g_0 \in H_0^m(\mathbb{I})$ and $Z \sim \text{Unif}[0, 1]$. It can be easily shown that $I(U) = \alpha$, and thus $B(Z) = \alpha$ in this model. Consequently, we can construct the basis functions h_ν as those defined in (5.2) with $\sigma = \alpha^{-1/2}$, and the eigenvalues as $\gamma_0 = 0$ and $\gamma_{2k-1} = \gamma_{2k} = \alpha^{-1}(2\pi k)^{2m}$ for $k \geq 1$. The remaining analysis is similar to case (I) of Example 5.1; for example, c_0 is given in (5.5).

EXAMPLE 5.3 (Semiparametric logistic regression). For the binary response $Y \in \{0, 1\}$, we consider the following semiparametric logistic model:

$$P(Y = 1|X = x, Z = z) = \frac{\exp(x^T\theta_0 + g_0(z))}{1 + \exp(x^T\theta_0 + g_0(z))}, \tag{5.7}$$

where $g_0 \in H^m(\mathbb{I})$. It can be shown that, in reasonable situations, all the conditions in Theorems 3.1 and 4.4 are satisfied; see Section S.15 in [6] for more details.

The solutions γ_ν and h_ν to the problem (2.10) are useful to calculate the quantities in the limit distribution (such as $\sigma_{z_0}^2$ and c_0 in Theorems A.1 and 4.4). However, in this model, due to the intractable forms of these solutions, we need to use consistent estimators of $B(\cdot)$ and $\pi(\cdot)$ to find the approximated solutions; for example, $\widehat{B}(\cdot)$ is a plug-in estimator and $\widehat{\pi}(\cdot)$ is a kernel density estimator.

Given the length of this paper, we conduct simulations only for the CIs of the conditional mean defined in (5.7) at a number of (x_0, z_0) values, that is, $x_0 = 1/4, 2/4, 3/4$ and thirty evenly spaced z_0 over $[0, 1]$. The true parameters are $\theta_0 = -0.5$ and $g_0(z) = 0.3(10^6)(1-z)^6 + (10^4)(1-z)^{10} - 2$. For simplicity, we generate $X, Z \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$. Based on 500 replicated data sets, we construct the 95% CIs and calculate their coverage proportions. The results are summarized in Figure 4 for various sample sizes $n = 400, 500, 700$. We observe that, as n increases, the coverage proportions approach the desired level, 95%, and the lengths of the CI approach zero.

REMARK 5.1. We use this logistic regression model to illustrate the eigensystem difference between the semi-nonparametric context and the nonparametric context, which leads to different inference for the nonparametric components [except under some strong conditions, e.g., (5.8) below]. This is slightly counterintuitive given that the parametric component can be estimated at a faster rate. As discussed above, the eigensystem for the semiparametric logistic model relies on $B(z)$ defined in (S.18) of [6]. According to Shang and Cheng [30], the eigensystem for the nonparametric logistic model relies on $I'(z)$ defined as $\exp(g_0(z))/(1 + \exp(g_0(z)))^2$. Therefore, the equivalence of the two eigensystems holds if and only if $B(z) = I'(z)$, that is,

$$E\left\{\frac{\exp(X^T\theta_0)}{(1 + \exp(X^T\theta_0 + g_0(z)))^2}\middle|Z = z\right\} = \frac{1}{(1 + \exp(g_0(z)))^2}. \tag{5.8}$$

If $\theta_0 = 0$, it is clear that (5.8) is true. However, we argue that in general (5.8) may not hold. For instance, it does not hold when $g_0(z) = 0$ for some $z \in [0, 1]$ because the above equation then simplifies to $E\{\frac{\exp(X^T\theta_0)}{(1+\exp(X^T\theta_0))^2}\} = 1$. This is not possible since $\frac{\exp(X^T\theta_0)}{(1+\exp(X^T\theta_0))^2} < 1$ almost surely. This concludes our argument. $\square$

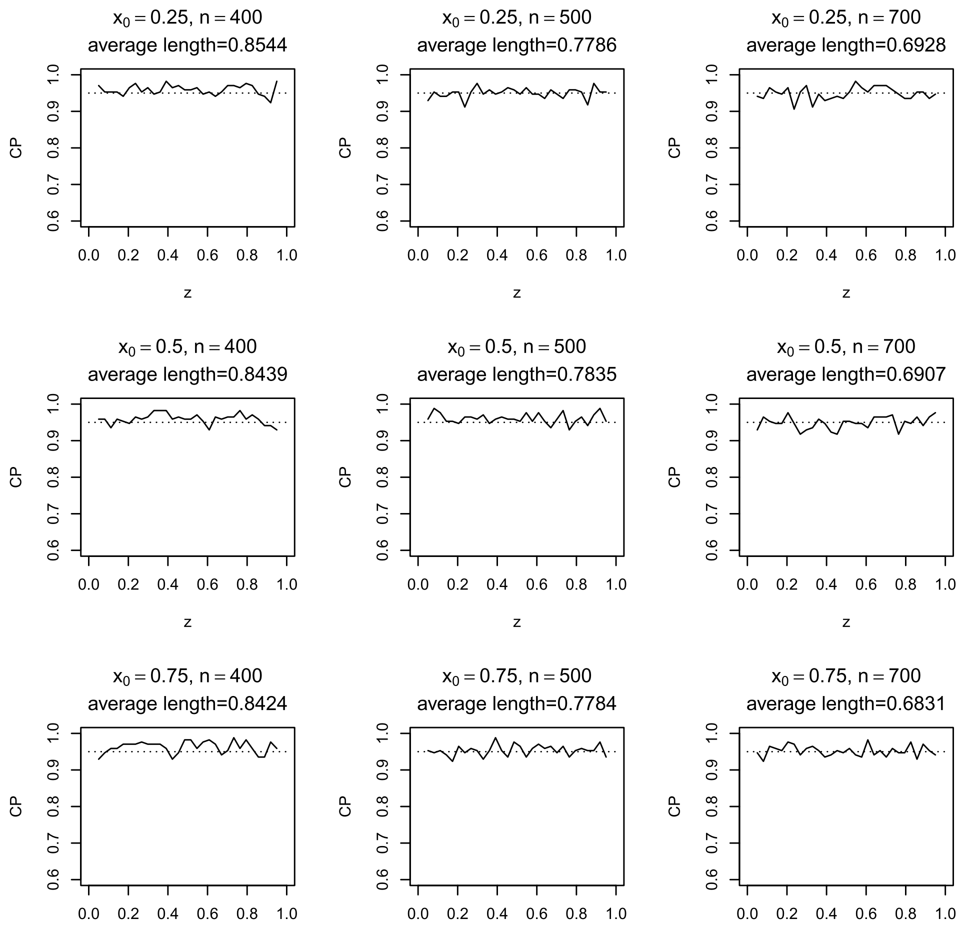

FIG. 4. *Coverage proportion of* 95% *CIs for the conditional mean constructed at a variety of* (x, z) *values*.

6. Future work. The general framework in this paper covers a wide range of commonly used models. In this section, we discuss some possible extensions using heuristic arguments, while omitting all the technical details due to the length of this paper. The first possible extension is to the class of generalized additive partial linear models in which $E(Y|X, Z) = F(X^T\theta_0 + \sum_{j=1}^J g_{j0}(Z_j))$. Our techniques are expected to handle this more general class by modifying the joint Bahadur representation in Theorem 2.6, that is, to replace f therein by $(\theta, g_1, \dots, g_J)$. The second possible extension is to deal with the functional data. In [31], we develop nonparametric inference for the (generalized) functional linear models, that is, $E(Y|Z) = F(\int_0^1 Z(t)\beta_0(t)\,dt)$, by penalizing the slope function $\beta(\cdot)$. By incorporating the techniques in [31] into our paper, we believe that it is feasible to do the joint asymptotic study of the (generalized) partial functional linear regression models [32], that is, $E(Y|X,Z) = F(X^T\theta_0 + \int_0^1 Z(t)\beta_0(t)\,dt)$. The third possible extension is from the above regression models to survival models. Specifically, our results may be extended to the partially linear Cox proportional hazard models (under right censored data) (i.e., [15]), by replacing our criterion function by their partial likelihood. This extension seems technically feasible given the quadratic structure of the profile likelihood (a generalization of partial likelihood) proven in [25].

APPENDIX

In this section, proofs of the main results are provided. In Section A.1, a preliminary lemma used for main results is provided. In Section A.2, an initial result about the joint limit distribution of the parametric and nonparametric estimators with biased center is given. Section A.3 includes the proof of Theorem 3.1. In Section A.4, the proof of Theorem 4.4 on the null limit distribution of likelihood ratio testing is provided.

For any $f = (\theta, g) \in \mathcal{H}$, we treat f as a "partly linear" function, that is, $f : (x, z) \mapsto x^T\theta + g(z)$, where $(x, z) \in \mathcal{U}$. Thus (θ, g) can be viewed as a bivariate function defined on $\mathcal{U}$. Throughout the Appendix, we will not distinguish (θ, g) and its associated function f. For instance, we use $(\theta, g) \in \mathcal{G}_0$ to mean $f \in \mathcal{G}_0$, some set of functions defined over $\mathcal{U}$.

A.1. An important lemma.

LEMMA A.1.

$$\lim_{\lambda\to 0} E_Z\{B(Z)(G(Z) - A(Z))(G(Z) - A(Z))^T\} = 0. \tag{A.1}$$

$$\lim_{\lambda\to 0} E_Z\{B(Z)G(Z)(G(Z) - A(Z))^T\} = 0. \tag{A.2}$$

PROOF. The proofs of (A.1) and (A.2) are similar, so we only show that (A.2) holds. Considering (2.11) and taking $g = h_\nu$, one has

$$V(G_k, h_\nu) = \langle A_k, h_\nu\rangle_1 = \left\langle \sum_\mu V(A_k, h_\mu)h_\mu, h_\nu\right\rangle_1 = (1 + \lambda\gamma_\nu)V(A_k, h_\nu), \tag{A.3}$$

and, taking $g = K_z$, one has $V(G_k, K_z) = A_k(z)$. By (A.3), $A_k = \sum_\nu \frac{V(G_k,h_\nu)}{1+\lambda\gamma_\nu} h_\nu$ holds in $L_2(\mathbb{I})$. For any $k, j = 1, \ldots, p$, by a straightforward calculation, we have

$$E_Z\{B(Z)G_j(Z)(G_k(Z) - A_k(Z))\} = \sum_\nu V(G_j, h_\nu)V(G_k, h_\nu)\frac{\lambda\gamma_\nu}{1 + \lambda\gamma_\nu}.$$

By square summability of $\{V(G_k, h_\nu)\}_{\nu\in\mathbb{N}}$ and dominated convergence theorem, the above sum converges to zero as $\lambda \to 0$. $\square$

A.2. An initial result on joint asymptotic distribution with biased center.

THEOREM A.1. *Let Assumptions A1 through A4 be satisfied. Suppose that as* $n \to \infty$, $h = o(1)$, $nh^2 \to \infty$ *and* $a_n \log n = o(n^{-1/2}h^{1/2})$, *where* a_n *is defined as in* (2.16). *Furthermore, assume that, as* $n \to \infty$,

$$
\begin{aligned}
&hV(K_{z_0}, K_{z_0}) \to \sigma_{z_0}^2, \qquad h^{1/2}(W_\lambda A)(z_0) \to \alpha_{z_0} \in \mathbb{R}^p \quad \text{and} \\
&h^{1/2} A(z_0) \to -\beta_{z_0} \in \mathbb{R}^p,
\end{aligned}
\tag{A.4}
$$

where A *is the Riesz representer defined in* (2.11). *Then we have, for any* $z_0 \in \mathbb{I}$,

$$
\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0^*) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0^*(z_0)) \end{pmatrix} \xrightarrow{d} N(0, \Psi^*),
\tag{A.5}
$$

where

$$
\Psi^* = \begin{pmatrix} \Omega^{-1} & \Omega^{-1}(\alpha_{z_0} + \beta_{z_0}) \\ (\alpha_{z_0} + \beta_{z_0})^T \Omega^{-1} & \sigma_{z_0}^2 + 2\beta_{z_0}^T \Omega^{-1} \alpha_{z_0} + \beta_{z_0}^T \Omega^{-1} \beta_{z_0} \end{pmatrix}.
\tag{A.6}
$$

Note that Ω^{-1} is well defined under Assumption A3. It follows from (2.9) and (2.12) that

$$
\alpha_{z_0} = \lim_{h \to 0} h^{1/2} \sum_\nu \frac{V(G, h_\nu)\lambda\gamma_\nu}{(1 + \lambda\gamma_\nu)^2} h_\nu(z_0),
$$

$$
\beta_{z_0} = -\lim_{h \to 0} h^{1/2} \sum_\nu \frac{V(G, h_\nu)}{1 + \lambda\gamma_\nu} h_\nu(z_0).
$$

PROOF OF THEOREM A.1. Define

$$
\widehat{f}_{n,\lambda}^h = (\widehat{\theta}_{n,\lambda}, h^{1/2}\widehat{g}_{n,\lambda}), \qquad f_0^{*h} = (\theta_0^*, h^{1/2} g_0^*), \qquad R_u^h = (H_u, h^{1/2} T_u),
$$

where we recall $f_0^* = (id - P_\lambda) f_0$, H_u, T_u were defined by (2.13), and P_λ is specified in Proposition 2.2. By Theorem 2.6,

$$
\mathrm{Rem}_n = \widehat{f}_{n,\lambda} - f_0^* - \frac{1}{n} \sum_{i=1}^n \epsilon_i R_{U_i}
$$

satisfies $\|\mathrm{Rem}_n\| = O_P(a_n \log n)$, which will imply by Assumption A1(b) that

$$
\left\| \widehat{\theta}_{n,\lambda} - \theta_0^* - \frac{1}{n} \sum_{i=1}^n \epsilon_i H_{U_i} \right\|_{l_2} = O_P(a_n \log n).
\tag{A.7}
$$

Define $\mathrm{Rem}_n^h = \widehat{f}_{n,\lambda}^h - f_0^{*h} - \frac{1}{n} \sum_{i=1}^n \epsilon_i R_{U_i}^h$, then it is easy to see that

$$
\mathrm{Rem}_n^h - h^{1/2}\mathrm{Rem}_n = \left((1 - h^{1/2}) \left(\widehat{\theta}_{n,\lambda} - \theta_0^* - \frac{1}{n} \sum_{i=1}^n \epsilon_i H_{U_i} \right), 0 \right).
$$

Thus, by (A.7),

$$\|\mathrm{Rem}_n^h - h^{1/2}\mathrm{Rem}_n\| \le (1-h^{1/2}) \cdot O\left(\left\|\widehat{\theta}_{n,\lambda} - \theta_0^* - \frac{1}{n}\sum_{i=1}^n \epsilon_i H_{U_i}\right\|_{l_2}\right)$$
$$= O_P(a_n \log n).$$

Since by assumption $a_n \log n = o(n^{-1/2})$, $\|\mathrm{Rem}_n^h\| = o_P(n^{-1/2})$. Next we will use Rem_n^h to obtain the target joint limiting distribution.

The idea is to employ the Cramér–Wald device. For any $x \in \mathbb{I}^p$, we will obtain the limiting distribution of $n^{1/2}x^T(\widehat{\theta}_{n,\lambda} - \theta_0^*) + (nh)^{1/2}(\widehat{g}_{n,\lambda}(z_0) - g_0^*(z_0))$. Note that this is equal to $n^{1/2}\langle R_u, \widehat{f}_{n,\lambda}^h - f_0^{*h}\rangle$ with $u = (x, z_0)$. Using the fact that

$$\left|n^{1/2}\left\langle R_u, \widehat{f}_{n,\lambda}^h - f_0^{*h} - \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i}^h\right\rangle\right|$$
$$\le n^{1/2}\|R_u\| \cdot \|\mathrm{Rem}_n^h\|$$
$$= O_P(n^{1/2}h^{-1/2}a_n \log n) = o_P(1),$$

we just need to find the limiting distribution of $n^{1/2}\langle R_u, \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i}^h\rangle$, which is equal to

$$n^{1/2}\left\langle R_u, \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i}^h\right\rangle = n^{-1/2}\sum_{i=1}^n \epsilon_i (x^T H_{U_i} + h^{1/2}T_{U_i}(z_0)).$$

Next we will use CLT to find its limiting distribution. By Assumption A1(c), that is, $E\{\epsilon^2|U\} = I(U)$, we have that

$$s_n^2 \equiv \mathrm{Var}\left(\sum_{i=1}^n \epsilon_i (x^T H_{U_i} + h^{1/2}T_{U_i}(z_0))\right)$$
$$= nE\{\epsilon^2|x^T H_U + h^{1/2}T_U(z_0)|^2\}$$
$$= nE\{E\{\epsilon^2|U\}|x^T H_U + h^{1/2}T_U(z_0)|^2\}$$
$$= nE\{I(U)|x^T H_U + h^{1/2}T_U(z_0)|^2\}.$$

A direct examination from (2.13) shows that

$$\begin{aligned} x^T H_U + h^{1/2}T_U(z) &\\ &= x^T(\Omega + \Sigma_\lambda)^{-1}(X - A(Z)) + h^{1/2}K_Z(z_0) \\ &\quad - h^{1/2}A(z_0)^T(\Omega + \Sigma_\lambda)^{-1}(X - A(Z)) \\ &= h^{1/2}K_Z(z_0) + (x - h^{1/2}A(z_0))^T(\Omega + \Sigma_\lambda)^{-1}(X - A(Z)). \end{aligned} \tag{A.8}$$

It follows by the proof of Lemma 2.4 that $|K_Z(z_0)| = O(h^{-1})$. On the other hand, for any $z \in \mathbb{I}$ and $j = 1, \dots, p$,

$$
\begin{aligned}
|A_k(z)| &= \left|\sum_{\nu=1}^{\infty} \frac{V(G_k, h_\nu)h_\nu}{1+\lambda\gamma_\nu}\right| \\
&\le \left(\sum_\nu |V(G_k, h_\nu)|^2 h_\nu(z)^2\right)^{1/2} \left(\sum_\nu \frac{1}{(1+\lambda\gamma_\nu)^2}\right)^{1/2} \le C_k' h^{-1/2},
\end{aligned}
$$

where C_k' is free of z. Thus, by (A.8), there exists a constant c' s.t. $|x^T H_U + h^{1/2} T_U(z)| \le c' h^{-1/2}$, a.s.

Thus

$$
\begin{aligned}
&E\{I(U)|x^T H_U + h^{1/2} T_U(z_0)|^2\} \\
&\quad = hE\{I(U)|K_Z(z_0)|^2\} \\
&\qquad + 2h^{1/2}(x - h^{1/2}A(z_0))^T (\Omega + \Sigma_\lambda)^{-1} E\{I(U)K_Z(z_0)(X - A(Z))\} \\
&\qquad + (x - h^{1/2}A(z_0))^T E\{I(U)H_U H_U^T\}(x - h^{1/2}A(z_0)).
\end{aligned}
\tag{A.9}
$$

Lemma A.1 tells us, as $\lambda \to 0$, $\Sigma_\lambda = E_Z\{B(Z)G(Z)(G(Z) - A(Z))^T\} \to 0$. It can be verified that

$$
\begin{aligned}
&E_U\{I(U)H_U H_U^T\} \\
&\quad = (\Omega + \Sigma_\lambda)^{-1} E\{I(U)(X - A(Z))(X - A(Z))^T\}(\Omega + \Sigma_\lambda)^{-1} \\
&\quad = (\Omega + \Sigma_\lambda)^{-1} E\{I(U)(X - G(Z) + G(Z) - A(Z)) \\
&\qquad\qquad \times (X - G(Z) + G(Z) - A(Z))^T\}(\Omega + \Sigma_\lambda)^{-1} \\
&\quad = (\Omega + \Sigma_\lambda)^{-1}(E\{I(U)(X - G(Z))(X - G(Z))^T\} \\
&\qquad\qquad + E\{I(U)(G(Z) - A(Z))(G(Z) - A(Z))^T\})(\Omega + \Sigma_\lambda)^{-1} \\
&\quad \to \Omega^{-1},
\end{aligned}
$$

where the last limit follows by Lemma A.1. By assumption, as $\lambda \to 0$, $hE\{I(U)|K_Z(z_0)|^2\} = hV(K_{z_0}, K_{z_0}) \to \sigma_{z_0}^2$, $h^{1/2}A(z_0) \to -\beta_{z_0}$ and

$$
\begin{aligned}
h^{1/2}&E\{I(U)K_Z(z_0)(X - A(Z))\} \\
&= h^{1/2}E\{B(Z)K_{z_0}(Z)(G(Z) - A(Z))\} \\
&= h^{1/2}(V(G, K_{z_0}) - V(A, K_{z_0})) \\
&= h^{1/2}(A(z_0) - V(A, K_{z_0})) \\
&= h^{1/2}(W_\lambda A)(z_0) \to \alpha_{z_0}.
\end{aligned}
$$

Thus, as λ approaches zero, the limit of (A.9) is

$$
\sigma_{z_0}^2 + 2(x + \beta_{z_0})^T \Omega^{-1} \alpha_{z_0} + (x + \beta_{z_0})^T \Omega^{-1}(x + \beta_{z_0}) = (x^T, 1)\Psi^*(x^T, 1)^T,
$$

where Ψ^* is defined in (A.6). So $s_n^2 \asymp n$. Then it can be shown that, for any $\varepsilon > 0$,

$$\begin{aligned}
&E\{|\epsilon(x^T H_U + h^{1/2} T_U(z_0))|^2 I|\epsilon(x^T H_U + h^{1/2} T_U(z_0))| \geq \varepsilon s_n\} \\
&\quad \leq (c' h^{-1/2})^2 E\{\epsilon^2 I(|\epsilon| \geq \varepsilon s_n h^{1/2}/c')\} \\
&\quad \leq (c' h^{-1/2})^2 E\{\epsilon^4\}^{1/2} P(|\varepsilon| \geq \varepsilon s_n h^{1/2}/c')^{1/2} \\
&\quad \leq (c' h^{-1/2}) E\{\epsilon^4\}^{1/2} (\varepsilon^4 s_n^4 h^2)^{-1/2} E\{\varepsilon^4\}^{1/2} \\
&\quad = \frac{(c')^2 E\{\varepsilon^4\}}{\varepsilon^2 s_n^2 h^2} \to 0,
\end{aligned}$$

where the last limit follows by $s_n^2 \asymp n$ and the assumption $nh^2 \to \infty$. Then as n approaches infinity,

$$\begin{aligned}
&\frac{1}{s_n^2} \sum_{i=1}^n E\{|\epsilon_i(x^T H_{U_i} + h^{1/2} T_{U_i}(z_0))|^2 I|\epsilon_i(x^T H_{U_i} + h^{1/2} T_{U_i}(z_0))| \geq \varepsilon s_n\} \\
&\quad = \frac{n}{s_n^2} E\{|\epsilon(x^T H_U + h^{1/2} T_U(z_0))|^2 I|\epsilon(x^T H_U + h^{1/2} T_U(z_0))| \geq \varepsilon s_n\} \to 0.
\end{aligned}$$

So Lindeberg's condition holds. The desired result follows immediately by central limit theorem. This completes the proof. $\square$

A.3. Proof of Theorem 3.1. The proof of Theorem 3.1 directly follows Theorem A.1 and the following lemma.

LEMMA A.2. *Suppose that there exists $b \in (1/(2m), 1]$ such that G_k satisfies* (3.1). *Then we have, for any $z_0 \in \mathbb{I}$, $h^{1/2} A(z_0) = o(1)$, $h^{1/2}(W_\lambda A)(z_0) = o(1)$. Furthermore, if $n^{1/2} h^{m(1+b)} = o(1)$, then as $n \to \infty$,*

$$\begin{pmatrix} \sqrt{n}(\theta_0^* - \theta_0) \\ \sqrt{nh}(g_0^*(z_0) - g_0(z_0) + (W_\lambda g_0)(z_0)) \end{pmatrix} \longrightarrow 0. \tag{A.10}$$

PROOF. We will show (A.10) in three steps:

(i) Show $\|V(G, W_\lambda g_0)\|_{l_2} = o(n^{-1/2})$. By (2.9),

$$V(G_k, W_\lambda g_0) = \sum_{\mu \in \mathbb{Z}} V(G_k, h_\mu) V(g_0, h_\mu) \frac{\lambda \gamma_\mu}{1 + \lambda \gamma_\mu},$$

for any $k = 1, \ldots, p$. Then by Cauchy's inequality, we have

$$\begin{aligned}
&|V(G_k, W_\lambda g_0)|^2 \\
&\quad \leq \sum_\mu |V(G_k, h_\mu)|^2 \frac{\lambda \gamma_\mu}{1 + \lambda \gamma_\mu} \sum_\mu |V(g_0, h_\mu)|^2 \frac{\lambda \gamma_\mu}{1 + \lambda \gamma_\mu}
\end{aligned}$$

$$\begin{aligned}
&\le \text{const}\cdot\lambda\sum_{\mu}\bigl|V(G_k,h_\mu)\bigr|^2\frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu}\\
&=\text{const}\cdot\lambda\sum_{\mu}\bigl|V(G_k,h_\mu)\bigr|^2\gamma_\mu^b\biggl(\frac{\lambda\gamma_\mu^{1-b}}{1+\lambda\gamma_\mu}\biggr)\\
&\le \text{const}\cdot\lambda^{1+b}.
\end{aligned}$$

Thus, when $n^{1/2}\lambda^{(1+b)/2}=n^{1/2}h^{m(1+b)}=o(1)$, $\|V(G,W_\lambda g_0)\|_{l_2}=o(n^{-1/2})$.

(ii) Show $\|A_k\|_{\sup}=O(1)$, for any $k=1,\dots,p$. Note for any $z\in\mathbb{I}$, by (2.11),

$$\begin{aligned}
A_k(z)&=\langle A_k,K_z\rangle_1=V(G_k,K_z)\\
&=\sum_{\mu\in\mathbb{N}}\frac{V(G_k,h_\mu)}{1+\lambda\gamma_\mu}h_\mu(z).
\end{aligned}$$

By boundedness of h_νs (Assumption A3) and by Cauchy's inequality, uniformly for $z\in\mathbb{I}$,

$$\begin{aligned}
\bigl|A_k(z)\bigr|^2&\le\sum_{\mu}\bigl|V(G_k,h_\mu)\bigr|^2(1+\gamma_\mu)^b\bigl|h_\mu(z)\bigr|^2\cdot\sum_{\mu}\frac{1}{(1+\gamma_\mu)^b(1+\lambda\gamma_\mu)^2}\\
&=O\biggl(\sum_{\mu}\frac{1}{(1+\gamma_\mu)^b}\biggr)=O(1),
\end{aligned}$$

where the last equality follows by $\gamma_\mu\asymp\mu^{2m}$ and $2mb>1$. This shows $\|A_k\|_{\sup}=O(1)$, implying $h^{1/2}A(z_0)=o(1)$. By (2.12), $(W_\lambda A)(z)=A(z)-\sum_\mu\frac{V(G,h_\mu)}{(1+\lambda\gamma_\mu)^2}\times h_\mu(z)$. Using the above derivations we can show that uniformly for $z\in\mathbb{I}$, $|\sum_\mu\frac{V(G,h_\mu)}{(1+\lambda\gamma_\mu)^2}h_\mu(z)|^2=O(\sum_\mu\frac{1}{(1+\gamma_\mu)^b})=O(1)$, implying $h^{1/2}(W_\lambda A)(z_0)=o(1)$.

(iii) By (i) and (ii), (A.10) follows by, as $n\to\infty$,

$$\begin{aligned}
&\begin{pmatrix} n^{1/2}(\theta_0^*-\theta_0)\\ (nh)^{1/2}\bigl(g_0^*(z)-g_0(z)+(W_\lambda g_0)(z)\bigr)\end{pmatrix}\\
&\quad=\begin{pmatrix} n^{1/2}(\Omega+\Sigma_\lambda)^{-1}V(G,W_\lambda g_0)\\ -(nh)^{1/2}V\bigl(G^T,W_\lambda g_0\bigr)(\Omega+\Sigma_\lambda)^{-1}A(z)\end{pmatrix}\to 0. \qquad\square
\end{aligned}$$

A.4. Proof of Theorem 4.4. For notational convenience, denote $\widehat{f}=\widehat{f}_{n,\lambda}$, $\widehat{f}^0=\widehat{f}_{n,\lambda}^{H_0}$, the constrained estimate of f under H_0, and $f=\widehat{f}^0-\widehat{f}=(\theta,g)$. By Assumptions A4 and A5, with large probability, $\|f\|\le r_n$, where $r_n=M((nh)^{-1/2}+h^m)$ for some large M. By Assumption A1(a), for some large constant $C>0$, the event $B_n\equiv B_{n1}\cap B_{n2}$ has large probability, where $B_{n1}=\{\max_{1\le i\le n}\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)|\le C\log n\}$ and $B_{n2}=\{\max_{1\le i\le n}\sup_{a\in\mathcal{I}}|\ell_a'''(Y_i;a)|\le C\log n\}$. Let a_n be defined as in (2.16).

By Taylor's expansion,

$$
\begin{aligned}
\mathrm{LRT}_{n,\lambda} &= \ell_{n,\lambda}(\widehat{f}^0) - \ell_{n,\lambda}(\widehat{f}) \\
&= S_{n,\lambda}(\widehat{f})f + \int_0^1\int_0^1 sDS_{n,\lambda}(\widehat{f}+ss'f)ff\,ds\,ds' \\
&= \int_0^1\int_0^1 sDS_{n,\lambda}(\widehat{f}+ss'f)ff\,da\,ds' \qquad \text{(A.11)}\\
&= \int_0^1\int_0^1 s\{DS_{n,\lambda}(\widehat{f}+ss'f)ff - DS_{n,\lambda}(f_0)ff\}\,ds\,ds' \\
&\quad + \frac{1}{2}(DS_{n,\lambda}(f_0)ff - E\{DS_{n,\lambda}(f_0)ff\}) + \frac{1}{2}E\{DS_{n,\lambda}(f_0)ff\}.
\end{aligned}
$$

Denote the above three sums by I_1, I_2 and I_3. Next we will study the asymptotic behavior of these sums. Denote $\tilde{f} = \widehat{f} + ss'f - f_0 = (\tilde{\theta}, \tilde{g})$, for any $0 \le s, s' \le 1$. So $\|\tilde{f}\| = O_P(r_n)$.

By calculations of the Frechét derivatives, we have

$$
\begin{aligned}
&DS_{n,\lambda}(\widehat{f}+ss'f)ff \\
&\quad = DS_{n,\lambda}(\tilde{f}+f_0)ff \\
&\quad = \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i) + X_i^T\tilde{\theta} + \tilde{g}(Z_i))(X_i^T\theta + g(Z_i))^2 - \langle P_\lambda f, f\rangle,
\end{aligned}
$$

and

$$
DS_{n,\lambda}(f_0)ff = \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i))(X_i^T\theta + g(Z_i))^2 - \langle P_\lambda f, f\rangle.
$$

On B_n,

$$
\begin{aligned}
&|DS_{n,\lambda}(\widehat{f}+ss'f)ff - DS_{n,\lambda}(f_0)ff| \\
&\quad \le \frac{1}{n}C(\log n)\|\tilde{f}\|_{\sup}\sum_{i=1}^n (X_i^T\theta + g(Z_i))^2 \\
&\quad = C(\log n)\|\tilde{f}\|_{\sup}\left\langle \frac{1}{n}\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i}, f\right\rangle \qquad \text{(A.12)}\\
&\quad = C(\log n)\|\tilde{f}\|_{\sup}\left\langle \frac{1}{n}\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i} \right. \\
&\qquad\qquad \left. - E_T\{(X^T\theta + g(Z))R_U\}, f\right\rangle \\
&\qquad + C(\log n)\|\tilde{f}\|_{\sup}E_T\{(X^T\theta + g(Z))^2\}.
\end{aligned}
$$

Now we study $\frac{1}{n}\|\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i} - E_T\{(X^T\theta + g(Z))R_U\}\|$. Let $d_n = c_m h^{-1/2} r_n$ and $\bar{f} = d_n^{-1} f/2 = (d_n^{-2}\theta/2, d_n^{-1} g/2) \equiv (\bar{\theta}, \bar{g})$. Consider $\psi(T; f) = X^T\theta + g(Z)$ and $\psi_n(T; \bar{f}) = (1/2)c_m^{-1} h^{1/2} d_n^{-1}\psi(T; 2d_n\bar{f})$. It is easy to see that $\psi_n(T; \bar{f})$, as a function of $\bar{f}$, satisfies the Lipschitz continuity condition (S.6) in the online supplementary.

Since $h = o(1)$ and $nh^2 \to \infty$, $d_n = o(1)$. Then by Lemma 2.4, on B_n, $\|\bar{f}\|_{\sup} \le 1/2$, which implies that for any $(x, z) \in \mathcal{U}$, $|x^T\bar{\theta} + \bar{g}(z)| \le 1/2$. Letting x approach zero, one gets that $|\bar{g}(z)| \le 1/2$, and thus, $\|\bar{g}\|_{\sup} \le 1/2$, which further implies that $|x^T\bar{\theta}| \le \|\bar{g}\|_{\sup} + \|\bar{f}\|_{\sup} \le 1$ for any $x \in \mathbb{I}^p$. Also note that

$$\begin{aligned}
J(\bar{g}, \bar{g}) &= d_n^{-2}\lambda^{-1}\big(\lambda J(g, g)\big)/4 \\
&\le d_n^{-2}\lambda^{-1}\|f\|^2/4 \\
&\le d_n^{-2}\lambda^{-1} r_n^2/4 \\
&< c_m^{-2} h\lambda^{-1}.
\end{aligned}$$

Thus, when event B_n holds, $\bar{f}$ is an element in $\mathcal{G}$. Then by Lemma S.3 (in the supplementary material [6]), with large probability

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \big[\big(X_i^T\theta + g(Z_i)\big)R_{U_i} - E_T\big\{\big(X^T\theta + g(Z)\big)R_U\big\}\big]\right\| \\
&\quad= \frac{c_m h^{-1/2} d_n}{n}\left\|\sum_{i=1}^n \big[\psi_n(T_i; \bar{f})R_{U_i} - E_T\{\psi_n(T; \bar{f})R_U\}\big]\right\| \qquad \text{(A.13)} \\
&\quad= O_P(a_n'),
\end{aligned}$$

where $a_n' = n^{-1/2}((nh)^{-1/2} + h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}$. So by $a_n' = o(r_n)$,

$$\begin{aligned}
&\big|DS_{n,\lambda}(\widehat{f} + ss' f)ff - DS_{n,\lambda}(f_0)ff\big| \\
&\quad= \|\tilde{f}\|_{\sup}\big(O_P(a_n' r_n \log n) + O_P(r_n^2 \log n)\big) \\
&\quad= h^{-1/2} r_n O_P(r_n^2 \log n) \qquad \text{(A.14)} \\
&\quad= O_P(r_n^3 h^{-1/2}\log n).
\end{aligned}$$

Thus $|I_1| = O_P(r_n^3 h^{-1/2}\log n)$.

Next we approximate I_2. Define $\psi(T; f) = \ddot{\ell}_a(Y; X^T\theta_0 + g_0(Z))(X^T\theta + g(Z))$. Then by calculation of the Fréchet derivative (Section 2.2),

$$\begin{aligned}
&DS_{n,\lambda}(f_0)ff - E\{DS_{n,\lambda}(f_0)ff\} \\
&\quad= \left\langle \frac{1}{n}\sum_{i=1}^n \big[\psi(T_i; f)R_{U_i} - E_T\{\psi(T; f)R_U\}\big], f\right\rangle.
\end{aligned}$$

Thus $2|I_2| \le \frac{1}{n}\|\sum_{i=1}^n[\psi(T_i; f)R_{U_i} - E_T\{\psi(T; f)R_U\}]\| \cdot \|f\|$. So it is sufficient to approximate $\|\sum_{i=1}^n[\psi(T_i; f)R_{U_i} - E_T\{\psi(T; f)R_U\}]\|$. Let $\tilde{\psi}_n(T; \bar{f}) = (1/2)C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\psi(T; 2d_n\bar{f})$ and $\psi_n(T_i; \bar{f}) = \tilde{\psi}_n(T_i; \bar{f})I_{A_i}$, where $\bar{f} = d_n^{-1}f/2$ and $A_i = \{\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i; a)| \le C\log n\}$ for $i = 1, \dots, n$. By similar derivations as the ones below (A.12), it can be shown that on B_n, $\bar{f} \in \mathcal{G}$. Observe that B_n implies $\bigcap_i A_i$. A direct examination shows that ψ_n satisfies (S.6). By Lemma S.3, with large probability,

$$\begin{aligned}&\left\|\sum_{i=1}^n[\psi_n(T_i; \bar{f})R_{U_i} - E_T\{\psi_n(T; \bar{f})R_U\}]\right\|\\ &\quad\le (n^{1/2}h^{-(2m-1)/(4m)} + 1)(5\log\log n)^{1/2}.\end{aligned}\tag{A.15}$$

On the other hand, by Chebyshev's inequality

$$\begin{aligned}P(A_i^c) &= \exp(-(C/C_0)\log n)E\left\{\exp\left(\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i; a)|/C_0\right)\right\}\\ &\le C_1 n^{-C/C_0}.\end{aligned}$$

Since $h = o(1)$ and $nh^2 \to \infty$, we may choose C to be large so that $(\log n)^{-1} \times n^{-C/(2C_0)} = o(a_n' h^{1/2}d_n^{-1})$, where

$$a_n' = n^{-1/2}((nh)^{-1/2} + h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}.$$

By (2.3), which implies $E\{\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i; a)||U_i\} \le 2C_1C_0^2$, we have, on B_n, $E_T\{|\psi(T; 2d_n\bar{f})|^2\} \le 2C_1C_0^2d_n^2$. So when n is large, on B_n, by Chebyshev's inequality,

$$\begin{aligned}&\|E_T\{\psi_n(T_i; \bar{f})R_{U_i}\} - E_T\{\tilde{\psi}_n(T_i; \bar{f})R_{U_i}\}\|\\ &\quad= \|E_T\{\tilde{\psi}_n(T_i; \bar{f})R_{U_i} \cdot I_{A_i^c}\}\|\\ &\quad\le (1/2)C^{-1}(\log n)^{-1}d_n^{-1}(E_T\{|\psi(T; 2d_n\bar{f})|^2\})^{1/2}P(A_i^c)^{1/2}\\ &\quad\le (1/2)2^{1/2}C^{-1}C_0C_1(\log n)^{-1}n^{-C/(2C_0)}\\ &\quad= o(a_n' h^{1/2}d_n^{-1}).\end{aligned}\tag{A.16}$$

Therefore, by (A.15) and (A.16), on B_n with large probability,

$$\begin{aligned}&\frac{1}{n}\left\|\sum_{i=1}^n[\psi(T_i; f)R_{U_i} - E_T\{\psi(T; f)R_U\}]\right\|\\ &\quad= \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\left\|\sum_{i=1}^n[\tilde{\psi}_n(T_i; \bar{f})R_{U_i} - E_T\{\tilde{\psi}_n(T; \bar{f})R_U\}]\right\|\\ &\quad\le \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\end{aligned}\tag{A.17}$$

$$
\begin{aligned}
&\times \left(\left\| \sum_{i=1}^{n} [\psi_n(T_i; \bar{f}) R_{U_i} - E_T\{\psi_n(T; \bar{f}) R_U\}] \right\| \right. \\
&\qquad \left. + n \| E_T\{\psi_n(T_i; \bar{f}) R_{U_i}\} - E_T\{\tilde{\psi}_n(T_i; \bar{f}) R_{U_i}\} \| \right) \\
&\le \frac{2C c_m (\log n) h^{-1/2} d_n}{n} \\
&\quad \times \left[\left(n^{1/2} h^{-(2m-1)/(4m)} + 1 \right) (5 \log\log n)^{1/2} + o\left(n a_n' h^{1/2} d_n^{-1} \right) \right] \\
&\le C' a_n' \log n,
\end{aligned}
$$

for some large constant $C' > 0$. Thus $|I_2| = O_P(a_n' r_n \log n)$.

Note that $I_3 = -\|f\|^2/2$. Therefore,

$$
\begin{aligned}
-2n \cdot \mathrm{LRT}_{n,\lambda} &= n \| \widehat{f}^0 - \widehat{f} \|^2 + O_P\left(n r_n a_n' \log n + n r_n^3 h^{-1/2} \log n \right) \\
&= n \| \widehat{f}^0 - \widehat{f} \|^2 + O_P\left(n r_n a_n \log n + n r_n^3 h^{-1/2} \log n \right).
\end{aligned}
$$

By $r_n^2 h^{-1/2} = o(a_n)$ and $n r_n a_n = o((\log n)^{-1})$, we have that $O_P(n r_n a_n \log n + n r_n^3 h^{-1/2} \log n) = o_P(1)$. This shows $-2n \cdot \mathrm{LRT}_{n,\lambda} = n \| \widehat{f}^0 - \widehat{f} \|^2 + o_P(1)$. So we only focus on $n \| \widehat{f}^0 - \widehat{f} \|^2$. By Theorems 2.6 and 4.3,

$$
n^{1/2} \| \widehat{f}^0 - \widehat{f} - S_{n,\lambda}^0(f_0^0) + S_{n,\lambda}(f_0) \| = O_P\left(n^{1/2} a_n \log n \right) = o_P(1), \tag{A.18}
$$

so we just have to focus on $n^{1/2}\{S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0)\}$. Recall that under H_0, $f_0^0 = (\theta_0^0, g_0^0) \in \mathcal{H}_0$, so

$$
\begin{aligned}
S_{n,\lambda}^0(f_0^0) &= \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_a\left(Y_i; X_i^T \theta_0^0 + g_0^0(Z_i) + X_i^T \theta^\dagger + w^\dagger \right) R_{U_i}^0 - P_\lambda^0 f_0^0 \\
&= \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_a\left(Y_i; X_i^T \theta_0 + g_0(Z_i) \right) R_{U_i}^0 - P_\lambda^0 f_0^0 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i R_{U_i}^0 - P_\lambda^0 f_0^0,
\end{aligned}
$$

where $\epsilon_i = \dot{\ell}_a(Y_i; X_i^T \theta_0 + g_0(Z_i))$, R_U^0 and $P_\lambda^0 f_0^0$ are defined in Section 4, and

$$
S_{n,\lambda}(f_0) = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i R_{U_i} - P_\lambda f_0.
$$

Consequently,

$$
\begin{aligned}
&S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(R_{U_i}^0 - R_{U_i} \right) - \left(P_\lambda^0 f_0^0 - P_\lambda f_0 \right)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(\sum_{j=1}^{k}\rho_{U_i,j}R_{q_j,W_j}\right)+\left(\sum_{j=1}^{k}\zeta_j R_{q_j,W_j}\right)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\big(H(Q,W)\rho_{U_i},\big(Q^T K_{z_0}-A^T H(Q,W)\big)\rho_{U_i}\big)\\
&\quad + \big(H(Q,W)\zeta,\big(Q^T K_{z_0}-A^T H(Q,W)\big)\zeta\big)\\
&= (\xi,\beta)+\big(H(Q,W)\zeta,\big(Q^T K_{z_0}-A^T H(Q,W)\big)\zeta\big),
\end{aligned}
$$

where

$$
\beta = -\delta K_{z_0} - A^T\xi,\ \xi = -(1/n)\sum_{i=1}^{n}\epsilon_i H(Q,W)\rho_{U_i}
$$

and

$$
\delta = (1/n)\sum_{i=1}^{n}\epsilon_i Q^T\rho_{U_i}.
$$

Therefore,

$$
\begin{aligned}
&\|S_{n,\lambda}^0(f_0^0)-S_{n,\lambda}(f_0)\|^2\\
&\quad = \|(\xi,\beta)\|^2 + 2\big\langle(\xi,\beta),\big(H(Q,W)\zeta,\big(Q^T K_{z_0}-A^T H(Q,W)\big)\zeta\big)\big\rangle\\
&\qquad + \big\|\big(H(Q,W)\zeta,\big(Q^T K_{z_0}-A^T H(Q,W)\big)\zeta\big)\big\|^2.
\end{aligned}
$$

We next evaluate the three items on the right-hand side of the above equation. Denote $\Sigma_\lambda = E_U\{I(U)(G(Z)-A(Z))(G(Z)-A(Z))^T\}$. Note $E_Z\{B(Z)(G(Z)-A(Z))K_{z_0}(Z)\} = V(G,K_{z_0}) - V(A,K_{z_0}) = \langle A, K_{z_0}\rangle_1 - V(A,K_{z_0}) = \langle W_\lambda A, K_{z_0}\rangle_1 = (W_\lambda A)(z_0)$. First,

$$
\begin{aligned}
&\|(\xi,\beta)\|^2\\
&\quad = E_U\{I(U)(X^T\xi+\beta(Z))^2\}+\lambda J(\beta,\beta)\\
&\quad = E_U\{I(U)[(X-A(Z))^T\xi-\delta K_{z_0}(Z)]^2\}+\lambda J(\beta,\beta)\\
&\quad = \xi^T E_U\{I(U)(X-A(Z))(X-A(Z))^T\}\xi\\
&\qquad - 2\xi^T E_U\{I(U)(X-A(Z))K_{z_0}(Z)\}\delta\\
&\qquad + \delta^2 E_Z B(Z)|K_{z_0}(Z)|^2 + \langle W_\lambda(\delta K_{z_0}+A^T\xi),\delta K_{z_0}+A^T\xi\rangle_1 \qquad \text{(A.19)}\\
&\quad = \xi^T(\Omega+\Sigma_\lambda)\xi - 2\xi^T E_Z\{B(Z)(G(Z)-A(Z))K_{z_0}(Z)\}\delta\\
&\qquad + \delta^2 V(K_{z_0},K_{z_0}) + \delta^2\langle W_\lambda K_{z_0},K_{z_0}\rangle_1\\
&\qquad + 2\delta\xi^T\langle W_\lambda A, K_{z_0}\rangle_1 + \xi^T\langle W_\lambda A, A^T\rangle_1\xi
\end{aligned}
$$

$$
\begin{aligned}
&= \xi^T \Gamma_\lambda \xi - 2\xi^T (W_\lambda A)(z_0)\delta + \delta^2 K(z_0, z_0) + 2\delta\xi^T (W_\lambda A)(z_0) \\
&= \xi^T \Gamma_\lambda \xi + \delta^2 K(z_0, z_0),
\end{aligned}
$$

where $\Gamma_\lambda = \Omega + \Sigma_\lambda + \langle W_\lambda A, A^T\rangle_1$ and $\Sigma_\lambda = E_Z\{B(Z)(G(Z) - A(Z))(G(Z) - A(Z))^T\}$. Second,

$$
\begin{aligned}
&\langle(\xi, \beta), \big(H(Q, W)\zeta, \big(Q^T K_{z_0} - A^T H(Q, W)\big)\zeta\big)\rangle \\
&\quad = E_U\{I(U)\big[\big(X - A(Z)\big)^T \xi - \delta K_{z_0}(Z)\big] \\
&\qquad\qquad \times \big[\big(X - A(Z)\big)^T H(Q, W)\zeta + Q^T \zeta K_{z_0}(Z)\big]\} \\
&\qquad + \langle W_\lambda \beta, Q^T \zeta K_{z_0} - A^T H(Q, W)\zeta\rangle_1 \\
&\quad = \xi^T E_U\{I(U)\big(X - A(Z)\big)\big(X - A(Z)\big)^T\} H(Q, W)\zeta \\
&\qquad + \xi^T E_U\{I(U)\big(X - A(Z)\big)K_{z_0}(Z)\} Q^T \zeta \\
&\qquad - \delta E_U\{I(U)K_{z_0}(Z)\big(X - A(Z)\big)^T\} H(Q, W)\zeta \\
&\qquad - \delta Q^T \zeta V(K_{z_0}, K_{z_0}) - \delta Q^T \zeta \langle W_\lambda K_{z_0}, K_{z_0}\rangle_1 \\
&\qquad + \delta\big(H(Q, W)\zeta\big)^T (W_\lambda A)(z_0) - Q^T \zeta \xi^T (W_\lambda A)(z_0) \\
&\qquad + \xi^T \langle W_\lambda A, A^T\rangle_1 H(Q, W)\zeta \\
&\quad = \xi^T \Gamma_\lambda H(Q, W)\zeta - \delta Q^T \zeta K(z_0, z_0).
\end{aligned}
\tag{A.20}
$$

Third, similar to the calculations in (A.19) and (A.20), we have

$$
\begin{aligned}
&\langle\big(H(Q, W)\zeta, \big(Q^T K_{z_0} - A^T H(Q, W)\big)\zeta\big), \\
&\quad \big(H(Q, W)\zeta, \big(Q^T K_{z_0} - A^T H(Q, W)\big)\zeta\big)\rangle \\
&= E_U\{I(U)\big[\big(X - A(Z)\big)^T H(Q, W)\zeta + Q^T \zeta K_{z_0}(Z)\big]^2\} \\
&\quad + \langle W_\lambda\big(Q^T \zeta K_{z_0} - A^T H(Q, W)\zeta\big), Q^T \zeta K_{z_0} - A^T H(Q, W)\zeta\rangle_1 \\
&= \zeta^T H(Q, W)^T \Gamma_\lambda H(Q, W)\zeta + \big(Q^T \zeta\big)^2 K(z_0, z_0).
\end{aligned}
\tag{A.21}
$$

It follows from (A.19) to (A.21) that

$$
\begin{aligned}
&\|S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0)\|^2 \\
&= \big(\xi + H(Q, W)\zeta\big)^T \Gamma_\lambda \big(\xi + H(Q, W)\zeta\big) + \big(\delta - Q^T \zeta\big)^2 K(z_0, z_0) \\
&= \begin{pmatrix} \xi + H(Q, W)\zeta \\ \delta - Q^T \zeta \end{pmatrix}^T \begin{pmatrix} \Gamma_\lambda & 0 \\ 0 & K(z_0, z_0) \end{pmatrix} \begin{pmatrix} \xi + H(Q, W)\zeta \\ \delta - Q^T \zeta \end{pmatrix}.
\end{aligned}
\tag{A.22}
$$

Next we find the limiting distribution of $n\|S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0)\|^2$, which leads to the limiting distribution of $-2n \cdot \mathrm{LRT}_{n,\lambda}$ in view of (A.18). By definition of ξ and the expressions of $H(Q,W)$, $T(Q,W)$, ρ_{U_i} and ζ in Section 4, we have

$$\begin{aligned}
&\xi + H(Q,W)\zeta \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i H(Q,W)M_K^{-1}\big(MH_{U_i} + QT_{U_i}(z_0)\big) \\
&\qquad + H(Q,W)M_K^{-1}\big(MH_{g_0}^* + QT_{g_0}^*(z_0)\big) \\
&\quad = H(Q,W)M_K^{-1}N\left(-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\begin{pmatrix} H_{U_i} \\ T_{U_i}(z_0)\end{pmatrix} + \begin{pmatrix} H_{g_0}^* \\ T_{g_0}^*(z_0)\end{pmatrix}\right) \\
&\quad = H(Q,W)M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix} \\
&\qquad \times\left(-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} + \begin{pmatrix} H_{g_0}^* \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\delta - Q^T\zeta \\
&\quad = Q^T M_K^{-1}N\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\begin{pmatrix} H_{U_i} \\ T_{U_i}(z_0)\end{pmatrix} - \begin{pmatrix} H_{g_0}^* \\ T_{g_0}^*(z_0)\end{pmatrix}\right) \\
&\quad = Q^T M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} - \begin{pmatrix} H_{g_0}^* \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\begin{pmatrix} \xi + H(Q,W)\zeta \\ \delta - Q^T\zeta\end{pmatrix} \\
&\quad = \begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix} M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix} \\
&\qquad \times\left(-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} + \begin{pmatrix} H_{g_0}^* \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).
\end{aligned} \tag{A.23}$$

Define $\tilde{M}_K = \binom{H(Q,W)}{-Q^T}^T \begin{pmatrix}\Gamma_\lambda & 0 \\ 0 & K(z_0,z_0)\end{pmatrix}\binom{H(Q,W)}{-Q^T}$, where we recall that $\Gamma_\lambda = \Omega + \Sigma_\lambda + \langle W_\lambda A, A^T\rangle_1$. Since for any $1 \le j,k \le p$, $\langle W_\lambda A_k, A_j\rangle_1 = \lambda\sum_\nu V(A_j, h_\nu)V(A_k, h_n u)\gamma_\nu = O(\lambda) = o(1)$, we have as $\lambda \to 0$, $\langle W_\lambda A, A^T\rangle_1 \to 0$, a $p \times p$ zero matrix. Define λ_1 as the maximum eigenvalue of $\langle W_\lambda A, A^T\rangle_1$, and λ_2 as the minimum eigenvalue of $\Omega + \Sigma_\lambda$. Thus $\lambda_1 = o(1)$. By equation (A.1) in Lemma A.1, λ_2 is asymptotically finitely upper bounded, and is lower bounded from zero. Note that

$$
\begin{aligned}
\tilde{M}_K - M_K &= \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix}^T \begin{pmatrix} \langle W_\lambda A, A^T\rangle_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix} \\
&\le \frac{\lambda_1}{\lambda_2} \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix}^T \begin{pmatrix} \Omega + \Sigma_\lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix} \\
&\le \frac{\lambda_1}{\lambda_2} \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix}^T \begin{pmatrix} \Omega + \Sigma_\lambda & 0 \\ 0 & K(z_0,z_0) \end{pmatrix} \begin{pmatrix} H(Q,W) \\ -Q^T \end{pmatrix} \\
&= \frac{\lambda_1}{\lambda_2} M_K.
\end{aligned} \tag{A.24}
$$

Define

$$
\begin{aligned}
\Psi_\lambda = &\begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix} N^T M_K^{-1} \tilde{M}_K M_K^{-1} N \\
&\times \begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1 \end{pmatrix} \begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix}.
\end{aligned}
$$

Therefore, by (A.24),

$$
\begin{aligned}
0 &\le \Psi_\lambda - \Phi_\lambda \\
&\le \frac{\lambda_1}{\lambda_2} \begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix} N^T M_K^{-1} N \\
&\quad \times \begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1 \end{pmatrix} \begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} = \frac{\lambda_1}{\lambda_2} \Phi_\lambda.
\end{aligned}
$$

Since $\Phi_\lambda \le \mathrm{trace}(\Phi_\lambda) I_{p+1} = k I_{p+1}$, $\Psi_\lambda - \Phi_\lambda = o(1) I_{p+1}$. Thus, as $n \to \infty$, Ψ_λ approaches Φ_0.

Next we will complete the proof by demonstrating the asymptotic distribution. It follows by Lemma A.2 that $n^{1/2} H_{g_0}^* = o(1)$. Denote $N_U = (-H_U^T, K_Z(z_0)/K(z_0, z_0)^{1/2})^T$. By Assumption A1(c),

$$
\begin{aligned}
&E\{\epsilon^2 N_U N_U^T\} \\
&\quad = E\left\{ I(U) \begin{pmatrix} H_U H_U^T & -H_U K_Z(z_0)/K(z_0,z_0)^{1/2} \\ -H_U^T K_Z(z_0)/K(z_0,z_0)^{1/2} & |K_Z(z_0)|^2/K(z_0,z_0) \end{pmatrix} \right\}.
\end{aligned}
$$

To find the limit of this matrix, note that as $\lambda \to 0$, the following limits hold:

- by Lemma A.1,

$$
\begin{aligned}
&E\{I(U) H_U H_U^T\} \\
&\quad = (\Omega + \Sigma_\lambda)^{-1} E\{I(U)(X - A(Z))(X - A(Z))^T\}(\Omega + \Sigma_\lambda)^{-1} \\
&\quad = (\Omega + \Sigma_\lambda)^{-1}(\Omega + E_Z\{B(Z)(G(Z) - A(Z))(G(Z) - A(Z))^T) \\
&\qquad \times (\Omega + \Sigma_\lambda)^{-1} \\
&\quad \to \Omega^{-1};
\end{aligned}
$$

- by $h^{1/2}(W_\lambda A)(z_0) \to 0$ (see Lemma A.2) and $hK(z_0, z_0) \to \sigma_{z_0}^2/c_0$ [by assumption (4.4)],

$$
\begin{aligned}
&E\{I(U)H_U K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\
&\quad = E\{I(U)(\Omega+\Sigma_\lambda)^{-1}(X - A(Z))K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\
&\quad = E\{B(Z)(G(Z) - A(Z))K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\
&\quad = (W_\lambda A)(z_0)/K(z_0,z_0)^{1/2} \to 0;
\end{aligned}
$$

- by assumption, $E\{B(Z)|K_Z(z_0)|^2\}/K(z_0,z_0) \to c_0$.

Thus, as $\lambda \to 0$, $E\{\epsilon^2 N_U N_U^T\} \to \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & c_0 \end{pmatrix}$. So as $n \to \infty$,

$$
\begin{aligned}
&n^{1/2}\begin{pmatrix} (\Omega+\Sigma_\lambda)^{1/2} & 0 \\ 0 & 1 \end{pmatrix}\left(-\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ \frac{K_{z_0}(Z_i)}{\sqrt{K(z_0,z_0)}} \end{pmatrix} + \begin{pmatrix} H_{g_0}^* \\ \frac{(W_\lambda g_0)(z_0)}{\sqrt{K(z_0,z_0)}} \end{pmatrix}\right) \\
&\quad \xrightarrow{d} \upsilon,
\end{aligned} \tag{A.25}
$$

where $\upsilon \sim N\left(\begin{pmatrix} 0 \\ c_{z_0} \end{pmatrix}, \begin{pmatrix} I_p & 0 \\ 0 & c_0 \end{pmatrix}\right)$. Therefore, it follows by (A.22), (A.23) and (A.25) that, as $n \to \infty$, $n\|S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0)\|^2 \xrightarrow{d} \upsilon^T \Phi_0 \upsilon$. It immediately follows that $\|\widehat{f}^0 - \widehat{f}\| = O_P(n^{-1/2})$. Besides, when $n \to \infty$, $-2n \cdot \mathrm{LRT}_{n,\lambda} \xrightarrow{d} \upsilon^T \Phi_0 \upsilon$.

SUPPLEMENTARY MATERIAL

Supplement to "Joint asymptotics for semi-nonparametric regression models with partially linear structure" (DOI: 10.1214/15-AOS1313SUPP; .pdf). Additional proofs are provided.

REFERENCES

[1] BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580. MR0189095

[2] BANERJEE, M., MUKHERJEE, D. and MISHRA, S. (2009). Semiparametric binary regression models under shape constraints with an application to Indian schooling data. *J. Econometrics* **149** 101–117. MR2518501

[3] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

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

[5] BOENTE, G., HE, X. and ZHOU, J. (2006). Robust estimates in generalized partially linear models. *Ann. Statist.* **34** 2856–2878. MR2329470

[6] CHENG, G. and SHANG, Z. (2015). Supplement to "Joint asymptotics for semi-nonparametric regression models with partially linear structure." DOI:10.1214/15-AOS1313SUPP.

[7] CHENG, G. (2009). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference* **139** 1980–1991. MR2497554

[8] CHENG, G. and HUANG, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.* **38** 2884–2915. MR2722459

[9] CHENG, G. and KOSOROK, M. R. (2008). General frequentist properties of the posterior profile distribution. *Ann. Statist.* **36** 1819–1853. MR2435457

[10] CHENG, G. and KOSOROK, M. R. (2009). The penalized profile sampler. *J. Multivariate Anal.* **100** 345–362. MR2483424

[11] COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695. MR1074429

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

[13] GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. MR1876599

[14] HECKMAN, J. and LEAMER, E. E., eds. (2007). *Handbook of Econometrics*, *Vol*. 6. Elsevier, Amsterdam.

[15] HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499

[16] KE, C. WANG, C. W. (2002). ASSIST: A suite of S-plus functions implementing spline smoothing techniques. Preprint.

[17] KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219. MR1041391

[18] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York. MR2724368

[19] KOSOROK, M. R., LEE, B. L. and FINE, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32** 1448–1491. MR2089130

[20] LI, Y., PRENTICE, R. L. and LIN, X. (2008). Semiparametric maximum likelihood estimation in normal transformation models for bivariate survival data. *Biometrika* **95** 947–960. MR2461222

[21] LOPUHAÄ, H. P. and NANE, G. F. (2013). Shape constrained non-parametric estimators of the baseline distribution in Cox proportional hazards model. *Scand. J. Stat.* **40** 619–646. MR3091700

[22] MAMMEN, E. and VAN DE GEER, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035. MR1447739

[23] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. MR3223057

[24] MESSER, K. and GOLDSTEIN, L. (1993). A new class of kernels for nonparametric curve estimation. *Ann. Statist.* **21** 179–195. MR1212172

[25] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168

[26] NYCHKA, D. (1995). Splines as local smoothers. *Ann. Statist.* **23** 1175–1197. MR1353501

[27] RADCHENKO, P. (2008). Mixed-rates asymptotics. *Ann. Statist.* **36** 287–309. MR2387972

[28] SAITOH, S. (1997). *Integral Transforms*, *Reproducing Kernels and Their Applications*. *Pitman Research Notes in Mathematics Series* **369**. Longman, Harlow. MR1478165

[29] SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511. MR1294076

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

[31] SHANG, Z. and CHENG, G. (2014). Nonparametric inference in generalized functional linear models. Purdue Technical Report.

[32] SHIN, H. (2009). Partial functional linear regression. *J. Statist. Plann. Inference* **139** 3405–3418. MR2549090

[33] STONE, M. H. (1926). A comparison of the series of Fourier and Birkhoff. *Trans. Amer. Math. Soc.* **28** 695–761. MR1501372

[34] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[35] WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

[36] WANG, Y. (2011). *Smoothing Splines*: *Methods and Applications*. *Monographs on Statistics and Applied Probability* **121**. CRC Press, Boca Raton, FL. MR2814838

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

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

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
250 N. UNIVERSITY STREET
WEST LAFAYETTE, INDIANA 47907
USA
E-MAIL: chengg@purdue.edu
shang9@purdue.edu