The Annals of Statistics
2008, Vol. 36, No. 5, 2232–2260
DOI: 10.1214/07-AOS544
© Institute of Mathematical Statistics, 2008

PROFILE-KERNEL LIKELIHOOD INFERENCE WITH DIVERGING NUMBER OF PARAMETERS[1]

BY CLIFFORD LAM AND JIANQING FAN

Princeton University

The generalized varying coefficient partially linear model with a growing number of predictors arises in many contemporary scientific endeavor. In this paper we set foot on both theoretical and practical sides of profile likelihood estimation and inference. When the number of parameters grows with sample size, the existence and asymptotic normality of the profile likelihood estimator are established under some regularity conditions. Profile likelihood ratio inference for the growing number of parameters is proposed and Wilk's phenomenon is demonstrated. A new algorithm, called the accelerated profile-kernel algorithm, for computing profile-kernel estimator is proposed and investigated. Simulation studies show that the resulting estimates are as efficient as the fully iterative profile-kernel estimates. For moderate sample sizes, our proposed procedure saves much computational time over the fully iterative profile-kernel one and gives stabler estimates. A set of real data is analyzed using our proposed algorithm.

1. Introduction.

Semiparametric models with large number of predictors arise frequently in many contemporary statistical studies. Large data set and high-dimensionality characterize many contemporary scientific endeavors [5, 7]. Statistical models with many predictors are frequently employed to enhance the explanatory and predictive powers. At the same time, semiparametric modeling is frequently incorporated to balance between modeling biases and "curse of dimensionality." Profile likelihood techniques [21] are frequently applied to this kind of semiparametric model. When the number of predictors is large, it is more realistic to regard it growing with the sample size. Yet, few results are available for semiparametric profile inferences when the number of parameters diverges with sample size. This paper focuses on profile likelihood inferences with diverging number of parameters in the context of the generalized varying coefficient partially linear model (GVCPLM).

GVCPLM is an extension of the generalized linear model [19] and the generalized varying-coefficient model [4, 11]. It allows some coefficient functions to vary with certain covariates U, such as age [8], toxic exposure level or time variable in a

Received June 2007; revised June 2007.

[1] Supported in part by NSF Grants DMS-03-54223, DMS-07-04337 and NIH Grant R01-GM072611.

AMS 2000 subject classifications. Primary 62G08; secondary 62J12, 62F12.

Key words and phrases. Generalized linear models, varying coefficients, high dimensionality, asymptotic normality, profile likelihood, generalized likelihood ratio tests.

longitudinal data or survival analysis [20]. Therefore, general interactions, not just the linear interaction as in parametric models, between the variable U and these covariates are explored nonparametrically.

If Y is a response variable and $(U, \mathbf{X}, \mathbf{Z})$ is the associated covariates, then by letting $\mu(u, \mathbf{x}, \mathbf{z}) = E\{Y|(U, \mathbf{X}, \mathbf{Z}) = (u, \mathbf{x}, \mathbf{z})\}$, the GVCPLM takes the form

$$g\{\mu(u, \mathbf{x}, \mathbf{z})\} = \mathbf{x}^T \boldsymbol{\alpha}(u) + \mathbf{z}^T \boldsymbol{\beta}, \tag{1.1}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\beta}$ a vector of unknown regression coefficients and $\boldsymbol{\alpha}(\cdot)$ a vector of unknown regression functions. One of the advantages over the varying coefficient model is that GVCPLM allows more efficient estimation when some coefficient functions are not really varying with U, after adjustment of other genuine varying effects. It also allows a more interpretable model, where primary interest is focused on the parametric component.

1.1. *A motivating example.* We use a real data example to demonstrate the need for GVCPLM. The Fifth National Bank of Springfield faced a gender discrimination suit in which female employees received substantially smaller salaries than male employees. This example is based on a real case with data dated 1995. Only the bank's name is changed. See Example 11.3 of [2]. Among 208 employees, eight variables are collected. They include employee's salary; age; year hired; number of years of working experience at another bank; gender; PC Job, a dummy variable with value 1 if the employee's job is computer related; educational level, a categorical variable with categories 1 (finished school), 2 (finished some college courses), 3 (obtained a bachelor's degree), 4 (took some graduate courses), 5 (obtained a graduate degree); job grade, a categorical variable indicating the current job level, the possible levels being 1–6 (6 the highest).

Fan and Peng [8] has conducted such a salary analysis using an additive model with quadratic spline and does not find significant evidence of gender difference. However, salary is directly related to the job grade. With the adjustment for the job grade, the salary discrimination can not easily be seen. An important question then arises if female employees have lower probability getting promoted. In analyzing such probability, a common tool will be the logistic regression, a class of the generalized linear model (e.g., see [19]).

To this end, we create a binary response variable **HighGrade4**, indicating if Job Grade is greater than 4. The associated covariates are **Female** (1 for female employee and 0 otherwise), **Age**, **TotalYrsExp** (total years of working experience), **PCJob**, **Edu** (level of education). Clearly interactions between **Age** and **TotalYrsExp** have to be considered.

If interactions between different variables are considered, then the number of predictors will be large compared with the sample size $n = 208$. This motivates us to consider the setting $p_n \to \infty$ as $n \to \infty$ and to present general theories in Section 2, where such a setting will be faced by many modern statistical applications.

1.2. *Goals of the paper.* When the number of parameters $\boldsymbol{\beta}$ is fixed and the link g is identity, the model (1.1) has been considered by Li et al. [16], Zhang, Lee and Song [29] and Xia, Zhang and Tong [27] and Ahmad, Leelahanon and Li [1]. Fan and Huang [6] proposed a profile-kernel inference for such a varying coefficient partial linear model (VCPLM) and Li and Liang [17] considered a backfitting-based procedure for model selection in VCPLM. All of these papers rely critically on the explicit form of the estimation procedures and the techniques cannot easily be applied to the GVCPLM.

Modern statistical applications often involve estimation of a large number of parameters. It is of interest to derive asymptotic properties for the profile likelihood estimators under model (1.1) when the number of parameters diverges. Several fundamental questions arise naturally. Does the profile likelihood estimator [21] still possess efficient sampling properties? Does the profile likelihood ratio test for the parametric component possess Wilks type of phenomenon, namely, whether the asymptotic null distributions are independent of nuisance functions and parameters? And, does the usual sandwich formula provide a consistent estimator of the covariance matrix of the profile likelihood estimator? These questions are poorly understood and will be thoroughly investigated in Section 2. Pioneering work on statistical inference with diverging number of parameters include [8, 13, 22] and [9].

Another goal of this paper is to provide an efficient algorithm for computing profile likelihood estimates under the model (1.1). To this end, we propose a new algorithm, called the accelerated profile-kernel algorithm, based on an important modification of the Newton–Raphson iterations. Computational difficulties [18] of the profile-kernel approach are significantly reduced, while nice sampling properties of such an approach over the backfitting algorithm (e.g., [12]) are retained. This will be convincingly demonstrated in Section 4, where the Poisson and Logistic specifications are considered for simulations. A new difference-based estimate for the parametric component is proposed as an initial estimate of our proposed profile-kernel procedure. Our method expands significantly the idea used in [28] and [6] for the partial linear model.

2. Properties of profile likelihood inference.

Let $(Y_{ni}; \mathbf{X}_i, \mathbf{Z}_{ni}, U_i)$, where $1 \le i \le n$, be a random sample where Y_{ni} is a scalar response variable, U_i, $\mathbf{X}_i \in \mathbb{R}^q$ and $\mathbf{Z}_{ni} \in \mathbb{R}^{p_n}$ are vectors of explanatory variables. We consider model (1.1) with $\boldsymbol{\beta}_n$ and $\mathbf{Z}_n$ having dimensions $p_n \to \infty$ as $n \to \infty$. Like the distributions in the exponential family, we assume that the conditional variance depends on the conditional mean so that $\mathrm{Var}(Y|U, \mathbf{X}, \mathbf{Z}_n) = V(\mu(u, \mathbf{X}, \mathbf{Z}_n))$ for a given function V (our result is applicable even when V is multiplied by an unknown scale). Then, the conditional quasi-likelihood function is given by

$$Q(\mu, y) = \int_{\mu}^{y} \frac{s - y}{V(s)}\, ds.$$

As in [24], we denote by $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ the "least favorable curve" of the nonparametric function $\boldsymbol{\alpha}(u)$, which is defined as the one that maximizes

$$E_0\{Q(g^{-1}(\boldsymbol{\eta}^T\mathbf{X}+\boldsymbol{\beta}_n{}^T\mathbf{Z}_n),Y_n)|U=u\} \tag{2.1}$$

with respect to $\boldsymbol{\eta}$, where E_0 is the expectation taken under the true parameters $\boldsymbol{\alpha}_0(u)$ and $\boldsymbol{\beta}_{n0}$. As will be discussed in Section 2.1, through the use of least favorable curve, no undersmoothing of the nonparametric component is required to achieve asymptotic normality when p_n is diverging with n. Note that $\boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(u)=\boldsymbol{\alpha}_0(u)$. Under some mild conditions, it satisfies

$$\frac{\partial}{\partial\boldsymbol{\eta}}E_0\{Q(g^{-1}(\boldsymbol{\eta}^T\mathbf{X}+\boldsymbol{\beta}_n{}^T\mathbf{Z}_n),Y_n)|U=u\}|_{\boldsymbol{\eta}=\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}=0. \tag{2.2}$$

The profile-likelihood function for $\boldsymbol{\beta}_n$ is then

$$Q_n(\boldsymbol{\beta}_n)=\sum_{i=1}^n Q\{g^{-1}(\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i+\boldsymbol{\beta}_n^T\mathbf{Z}_{ni}),Y_{ni}\}, \tag{2.3}$$

if the least-favorable curve $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(\cdot)$ is known.

The least-favorable curve defined by (2.1) can be estimated by its sample version through a local polynomial regression approximation. For U in a neighborhood of u, approximate the jth component of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(\cdot)$ as

$$\begin{aligned}\alpha_j(U)&\approx\alpha_j(u)+\frac{\partial\alpha_j(u)}{\partial u}(U-u)+\cdots+\frac{\partial^p\alpha_j(u)}{\partial u^p}(U-u)^p/p!\\&\equiv a_{0j}+a_{1j}(U-u)+\cdots+a_{pj}(U-u)^p/p!.\end{aligned}$$

Denoting $\mathbf{a_r}=(a_{r1},\ldots,a_{rq})^T$ for $r=0,\ldots,p$, for each given $\boldsymbol{\beta}_n$, we then maximize the local likelihood

$$\sum_{i=1}^n Q\left\{g^{-1}\left(\sum_{r=0}^p\mathbf{a_r}{}^T\mathbf{X}_i(U_i-u)^r/r!+\boldsymbol{\beta}_n^T\mathbf{Z}_{ni}\right),Y_{ni}\right\}K_h(U_i-u) \tag{2.4}$$

with respect to $\mathbf{a_0},\ldots,\mathbf{a_p}$, where $K(\cdot)$ is a kernel function and $K_h(t)=K(t/h)/h$ is a re-scaling of K with bandwidth h. Thus, we get estimate $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)=\hat{\mathbf{a}}_{\mathbf{0}}(u)$.

Plugging our estimates into the profile-kernel likelihood function (2.3), we have

$$\hat{Q}_n(\boldsymbol{\beta}_n)=\sum_{i=1}^n Q\{g^{-1}(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i+\boldsymbol{\beta}_n^T\mathbf{Z}_{ni}),Y_{ni}\}, \tag{2.5}$$

maximizing $\hat{Q}_n(\boldsymbol{\beta}_n)$ with respect to $\boldsymbol{\beta}_n$ to get $\hat{\boldsymbol{\beta}}_n$. With $\hat{\boldsymbol{\beta}}_n$, the varying coefficient functions are estimated as $\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(u)$.

One property of the profile quasi-likelihood is that the first- and second-order Bartlett's identities continue to hold. In particular, with the definition given by (2.3), then for any $\boldsymbol{\beta}_n$, we have

$$\mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial Q_n}{\partial\boldsymbol{\beta}_n}\right)=0,\qquad\mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial Q_n}{\partial\boldsymbol{\beta}_n}\frac{\partial Q_n}{\partial\boldsymbol{\beta}_n^T}\right)=-\mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial^2 Q_n}{\partial\boldsymbol{\beta}_n\,\partial\boldsymbol{\beta}_n^T}\right). \tag{2.6}$$

See [24] for more details. These properties give rise to the asymptotic efficiency of the profile likelihood estimator.

2.1. *Consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_n$.* We need Regularity Conditions (A)–(G) in Section 5 for the following results.

THEOREM 1 (Existence of profile likelihood estimator). *Assume that Conditions* (A)–(G) *are satisfied. If $p_n^4/n \to 0$ as $n \to \infty$ and $h = O(n^{-a})$ with $(4(p+1))^{-1} < a < 1/2$, then there is a local maximizer $\hat{\boldsymbol{\beta}}_n \in \Omega_n$ of $\hat{Q}_n(\boldsymbol{\beta}_n)$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P(\sqrt{p_n/n})$.*

Note that the optimal bandwidth $h = O(n^{-1/(2p+3)})$ is included in Theorem 1. Hence, $\sqrt{n/p_n}$-consistency is achieved without the need of undersmoothing of the nonparametric component.

Define $I_n(\boldsymbol{\beta}_n) = n^{-1}\mathbf{E}_{\boldsymbol{\beta}_n}(\frac{\partial Q_n}{\partial \boldsymbol{\beta}_n}\frac{\partial Q_n}{\partial \boldsymbol{\beta}_n^T})$, which is an extension of the Fisher matrix. Since the dimensionality grows with sample size, we need to consider the arbitrary linear combination of the profile kernel estimator $\hat{\boldsymbol{\beta}}_n$ as stated in the following theorem.

THEOREM 2 (Asymptotic normality). *Under Conditions* (A)–(G), *if $p_n^5/n = o(1)$ and $h = O(n^{-a})$ for $3/(10(p+1)) < a < 2/5$, then the consistent estimator $\hat{\boldsymbol{\beta}}_n$ in Theorem 1 satisfies*

$$\sqrt{n}A_n I_n^{1/2}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \xrightarrow{\mathscr{D}} N(0, G),$$

where A_n is an $l \times p_n$ matrix such that $A_nA_n^T \to G$, and G is an $l \times l$ nonnegative symmetric matrix.

A remarkable technical achievement of our result is that it does not require undersmoothing of the nonparametric component, as in Theorem 1, thanks to the profile likelihood approach. The key lies in a special orthogonality property of the least favorable curve [see equation (2.2) and Lemma 2]. Asymptotic normality without undersmoothing is also proved in [26] for both backfitting and profiling methods.

Theorem 2 shows that profile likelihood produces a semi-parametric efficient estimate even when the number of parameters diverges. To see this more explicitly, let $p_n = r$ be a constant. Then, by taking $A_n = I_r$, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \xrightarrow{\mathscr{D}} N(0, I^{-1}(\boldsymbol{\beta}_{n0})).$$

The asymptotic variance of $\hat{\boldsymbol{\beta}}_n$ achieves the efficient lower bound given, for example, in [24].

2.2. *Profile likelihood ratio test.* After estimation of parameters, it is of interest to test the statistical significance of certain variables in the parametric component. Consider the problem of testing linear hypotheses:

$$H_0 : A_n \boldsymbol{\beta}_{n0} = 0 \longleftrightarrow H_1 : A_n \boldsymbol{\beta}_{n0} \neq 0,$$

where A_n is an $l \times p_n$ matrix and $A_n A_n^T = I_l$ for a fixed l. Note that both the null and the alternative hypotheses are semi-parametric, with nuisance functions $\boldsymbol{\alpha}(\cdot)$. The generalized likelihood ratio test (GLRT) is defined by

$$T_n = 2\left\{\sup_{\Omega_n} \hat{Q}_n(\boldsymbol{\beta}_n) - \sup_{\Omega_n; A_n\boldsymbol{\beta}_n = 0} \hat{Q}_n(\boldsymbol{\beta}_n)\right\}.$$

The following theorem shows that, even when the number of parameters diverges with sample size, T_n still follows a chi-square distribution asymptotically, without reference to any nuisance parameters and functions. This reveals the Wilk's phenomenon, as termed in [10].

THEOREM 3. *Assuming Conditions* (A)–(G), *under* H_0, *we have*

$$T_n \xrightarrow{\mathcal{D}} \chi_l^2,$$

provided that $p_n^5/n = o(1)$ *and* $h = O(n^{-a})$ *for* $3/(10(p+1)) < a < 2/5$.

2.3. *Consistency of the sandwich covariance formula.* The estimated covariance matrix for $\hat{\boldsymbol{\beta}}_n$ can be obtained by the sandwich formula

$$\hat{\Sigma}_n = n^2\{\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)\}^{-1}\widehat{\mathrm{cov}}\{\nabla \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)\}\{\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)\}^{-1},$$

where the middle matrix has (j, k) entry given by

$$(\widehat{\mathrm{cov}}\{\nabla \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)\})_{jk} = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}}\right\} - \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}}\right\}.$$

With the notation $\Sigma_n = I_n^{-1}(\boldsymbol{\beta}_{n0})$, we have the following consistency result for the sandwich formula.

THEOREM 4. *Assuming Conditions* (A)–(G), *if* $p_n^4/n = o(1)$ *and* $h = O(n^{-a})$ *with* $(4(p+1))^{-1} < a < 1/2$, *we have*

$$A_n \hat{\Sigma}_n A_n^T - A_n \Sigma_n A_n^T \xrightarrow{\mathbb{P}} 0 \qquad \text{as } n \to \infty$$

for any $l \times p_n$ *matrix* A_n *such that* $A_n A_n^T = G$.

This result provides a simple way to construct confidence intervals for $\boldsymbol{\beta}_n$. Simulation results show that this formula indeed provides a good estimate of the covariance of $\hat{\boldsymbol{\beta}}_n$ for a variety of practical sample sizes.

3. Computation of the estimates. Finding $\hat{\boldsymbol{\beta}}_n$ to maximize the profile likelihood (2.5) poses some interesting challenges, as the function $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ in (2.5) depends on $\boldsymbol{\beta}_n$ implicitly (except the least-square case). The full profile-kernel estimate is to directly employ the Newton–Raphson iterations

$$\boldsymbol{\beta}_n^{(k+1)} = \boldsymbol{\beta}_n^{(k)} - \{\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n^{(k)})\}^{-1} \nabla \hat{Q}_n(\boldsymbol{\beta}_n^{(k)}), \tag{3.1}$$

starting from the initial value $\boldsymbol{\beta}^{(0)}$. We will call the estimate $\boldsymbol{\beta}_n^{(k)}$ and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n^{(k)}}(u)$ the k-step estimate [3, 23].

The first two derivatives of $\nabla \hat{Q}_n(\boldsymbol{\beta}_n)$ are given by

$$\begin{aligned}
\nabla \hat{Q}_n(\boldsymbol{\beta}_n) &= \sum_{i=1}^n q_{1i}(\boldsymbol{\beta}_n)\big(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i\big), \\
\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n) &= \sum_{i=1}^n q_{2i}(\boldsymbol{\beta}_n)\big(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i\big)\big(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i\big)^T \\
&\quad + \sum_{i=1}^n \left\{ q_{1i}(\boldsymbol{\beta}_n) \sum_{r=1}^q \frac{\partial^2 \hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}(U_i)}{\partial \boldsymbol{\beta}_n \, \partial \boldsymbol{\beta}_n^T} X_{ir} \right\},
\end{aligned} \tag{3.2}$$

where $q_l(x, y) = \frac{\partial^l}{\partial x^l} Q(g^{-1}(x), y)$, $q_{ki}(\boldsymbol{\beta}_n) = q_k(\hat{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})$ $(k = 1, 2)$ with $\hat{m}_{ni}(\boldsymbol{\beta}_n) = \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \mathbf{Z}_{ni}^T \boldsymbol{\beta}_n$. In the above formulae, $\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u) = \frac{\partial \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}$ is a p_n by q matrix and $\alpha^{(r)}_{\boldsymbol{\beta}_n}(u)$ is the rth component of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$.

3.1. *Methodology.* As the first two derivatives of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ are hard to compute in (3.2), one can employ the backfitting algorithm, which iterates between (2.4) and (2.3). This is really the same as the fully iterated algorithm (3.1), but ignores the functional dependence of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ in (2.5) on $\boldsymbol{\beta}_n$; it uses the value of $\boldsymbol{\beta}_n$ in the previous step of the iteration as a proxy. More precisely, the backfitting algorithm treats the terms $\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u)$ and $\hat{\boldsymbol{\alpha}}''_{\boldsymbol{\beta}_n}(u)$ in (3.2) as zero. The maximization is thus much easier to carry out, but the convergence speed can be reduced. See [12] and [18] for more descriptions of the two methods and some closed-form solutions proposed for the partially linear models.

Between these two extreme choices is our modified algorithm, which ignores the computation of the second derivative of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ in (3.1), but keeps its first derivative in the iteration. Namely, the second term in (3.2) is treated as zero. Since the function $q_2(\cdot, \cdot) < 0$ by Regularity Condition (D), by ignoring the second term in (3.2), the modified $\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n)$ in equation (3.2) is still negative-definite. This ensures the Newton–Raphson update of the profile-kernel procedure can be carried out smoothly. The intuition behind the modification is that, for a neighborhood around the true parameter $\boldsymbol{\beta}_{n0}$, the least favorable curve $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ should be approximately linear in $\boldsymbol{\beta}_n$. It turns out that this algorithm improves significantly the computation with achieved accuracy. At the same time, it enhances dramatically the stability of the algorithm. We will term the algorithm as the accelerated profile-kernel algorithm. A theorem on the computation and property of $\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u)$ follows.

THEOREM 5. *Under Regularity Conditions* (A)–(G), *provided* $\sqrt{p_n}(h + c_n \log^{1/2}(1/h)) = o(1)$ *where* $c_n = (nh)^{-1/2}$, *for each* $\boldsymbol{\beta}_n \in \Omega_n$,

$$\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u) = -\left\{\sum_{i=1}^{n} q_{2i}(\boldsymbol{\beta}_n)\mathbf{Z}_{ni}\mathbf{X}_i^T K_h(U_i - u)\right\} \cdot \left\{\sum_{i=1}^{n} q_{2i}(\boldsymbol{\beta}_n)\mathbf{X}_i\mathbf{X}_i^T K_h(U_i - u)\right\}^{-1}$$

is a consistent estimator of $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$ *which holds uniformly in* $u \in \Omega$.

When the quasi-likelihood becomes a square loss, the accelerated profile-kernel algorithm is exactly the same as that used to compute the full profile likelihood estimate, since $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(\cdot)$ is linear in $\boldsymbol{\beta}_n$.

3.2. *Difference-based estimation.* We generalize the difference-based idea to obtain an initial estimate $\boldsymbol{\beta}_n^{(0)}$. The idea has been used in [28] and [6] to remove the nonparametric component in the partially linear model.

We first consider the specific case of the GVCPLM:

$$Y = \boldsymbol{\alpha}(U)^T\mathbf{X} + \boldsymbol{\beta}_n{}^T\mathbf{Z}_n + \varepsilon. \tag{3.3}$$

This is the varying-coefficient partially linear model studied by Zhang, Lee and Song [29] and Xia, Zhang and Tong [27]. Let $\{(U_i, \mathbf{X}_i^T, \mathbf{Z}_{ni}^T, Y_i)\}_{i=1}^n$ be a random sample from the model (3.3), with the data ordered according to the U_i's. Under mild conditions, the spacing $U_{i+j} - U_i$ is $O_P(1/n)$, so that

$$\boldsymbol{\alpha}(U_{i+j}) - \boldsymbol{\alpha}(U_i) \approx \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1(U_{i+j} - U_i), \qquad j = 1, \dots, q. \tag{3.4}$$

Indeed, it can be approximately zero; the linear term is used to reduce the approximation errors.

For given weights w_j (its dependence on i is suppressed), define

$$Y_i^* = \sum_{j=1}^{q+1} w_j Y_{i+j-1}, \qquad \mathbf{Z}_{ni}^* = \sum_{j=1}^{q+1} w_j \mathbf{Z}_{n(i+j-1)}, \qquad \varepsilon_i^* = \sum_{j=1}^{q+1} w_j \varepsilon_{i+j-1}.$$

If we choose the weights to satisfy $\sum_{j=1}^{q+1} w_j \mathbf{X}_{i+j-1} = \mathbf{0}$, then using (3.3) and (3.4), we have

$$Y_i^* \approx \boldsymbol{\gamma}_0{}^T\mathbf{X}_i w_1 + \boldsymbol{\gamma}_1{}^T \sum_{j=1}^{q+1} w_j U_{i+j-1}\mathbf{X}_{i+j-1} + \boldsymbol{\beta}_n^T\mathbf{Z}_{ni}^* + \varepsilon_i^*.$$

Ignoring the approximation, which is of order $O_P(n^{-1})$, the above is a multiple regression model with parameters $(\boldsymbol{\gamma}_{\mathbf{0}}, \boldsymbol{\gamma}_{\mathbf{1}}, \boldsymbol{\beta}_n)$. The parameters can be found by a weighted least square fit to the $(n-q)$ starred data. This yields a root-n consistent estimate of $\boldsymbol{\beta}_n$, as the above approximation for the finite q is of order $O_P(n^{-1})$.

To solve $\sum_{j=1}^{q+1} w_j \mathbf{X}_{i+j-1} = \mathbf{0}$, we need to find the rank of the matrix $(\mathbf{X}_i, \ldots, \mathbf{X}_{i+q})$, denoted by r. Fix $q+1-r$ of the w_j's and the rest can be determined uniquely by solving the system of linear equations for $\{w_j, j = 1, \ldots, q+1\}$. For random designs, with probability 1, $r = q$. Hence, the direction of the weights $\{w_j, j = 1, \ldots, q+1\}$ is uniquely determined. For example, in the partial linear model, $q = 1$ and $\mathbf{X}_i = 1$. Hence, $(w_1, w_2) = c(1, -1)$ and the constant c can be taken to have a norm one. This results in the difference-based estimator in [28] and [6].

To use the differencing idea to obtain an initial estimate of $\boldsymbol{\beta}_n$ for the GVCPLM, we apply the transformation of the data. If g is the link function, we use $g(Y_i)$ as the transformed data and proceed with the difference-based method as for the VCPLM. Note that for some models like the logistic regression with logit link and Poisson log-linear model, we need to make adjustments in transforming the data. We use $g(y) = \log(\frac{y+\delta}{1-y+\delta})$ for the logistic regression and $g(y) = \log(y+\delta)$ for the Poisson regression. Here, the parameter δ is treated as a smoothing parameter like h, and its choice will be discussed in Section 3.3.

3.3. *Choice of bandwidth.* The two-dimensional smoothing parameters (δ, h) mentioned in the previous section can be selected by a K-fold cross-validation, using the quasi-likelihood as a criterion function. As demonstrated in Section 4, the practical accuracy can be achieved in several iterations using the accelerated profile-kernel algorithm. Hence, the profile-kernel estimate can be computed rapidly. As a result, the K-fold cross-validation is not too computationally intensive, as long as K is not too large (e.g., $K = 5$ or 10).

4. Numerical properties. To evaluate the performance of estimator $\hat{\boldsymbol{\alpha}}(\cdot)$, we use the square-root of average errors (RASE)

$$\mathrm{RASE} = \left\{ n_{\mathrm{grid}}^{-1} \sum_{k=1}^{n_{\mathrm{grid}}} \|\hat{\boldsymbol{\alpha}}(u_k) - \boldsymbol{\alpha}(u_k)\|^2 \right\}^{1/2},$$

over $n_{\mathrm{grid}} = 200$ grid points $\{u_k\}$. The performance of the estimator $\hat{\boldsymbol{\beta}}_n$ is assessed by the generalized mean square error (GMSE)

$$\mathrm{GMSE} = (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T \mathbf{B} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}),$$

where $\mathbf{B} = E\mathbf{Z}_n\mathbf{Z}_n^T$.

Throughout our simulation studies, the dimensionality of a parametric component is taken as $p_n = \lfloor 1.8 n^{1/3} \rfloor$ and the nonparametric component as $q = 2$ in

TABLE 1
Computation time and accuracy for different computing algorithms

n	p_n	**Backfitting**	**Accelerated profile-kernel**	**Full profile-kernel**
		Median and SD_{mad} (in parentheses) of computing times in seconds		
200	10	0.6 (0.0)	0.7 (0.0)	77.2 (0.2)
400	13	0.8 (0.0)	1.4 (0.0)	463.2 (0.9)
		Median and SD_{mad} (in parentheses) of GMSE (multiplied by 10^4)		
200	10	10.72 (6.47)	5.45 (2.71)	9.74 (14.67)
400	13	5.63 (4.39)	2.78 (1.19)	5.26 (9.46)
		Median RASE relative to the oracle estimate		
200	10	0.848	0.970	0.895
400	13	0.856	0.986	0.882

which $X_1 = 1$ and $X_2 \sim N(0, 1)$. The rate $p_n = O_P(n^{1/3})$ is not the same as presented in the theorems in Section 2, but we use this to show the capability of handling a higher rate of parameters growth for the accelerated profile-kernel method. In addition, the covariates $(\mathbf{Z}_n^T, X_2)^T$ are a $(p_n + 1)$-dimensional normal random vector with mean zero and covariance matrix (σ_{ij}), where $\sigma_{ij} = 0.5^{|i-j|}$. Furthermore, we always take $U \sim U(0, 1)$ independent of the other covariates. Finally, we use SD_{mad} to denote the robust estimate of standard deviation, which is defined as the interquartile range divided by 1.349. The number of simulations is 400, except that in Table 1 (which is 50) due to the intensive computation of the fully iterated profile-kernel estimate.

Poisson model. We use the log-link for the response Y given $(U, \mathbf{X}, \mathbf{Z_n})$, with $\boldsymbol{\beta}_{n0} = (0.5, 0.3, -0.5, 1, 0.1, -0.25, 0, \ldots, 0)^T$, $\alpha_1(u) = 4 + \sin(2\pi u)$ and $\alpha_2(u) = 2u(1 - u)$.

Bernoulli model. We use the logit-link for the response Y given $(U, \mathbf{X}, \mathbf{Z_n})$, with $\boldsymbol{\beta}_{n0} = (3, 1, -2, 0.5, 2, -2, 0, \ldots, 0)^T$ and $\alpha_1(u) = 2(u^3 + 2u^2 - 2u)$ and $\alpha_2(u) = 2\cos(2\pi u)$.

Throughout our numerical studies, we use the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and the 5-fold cross-validation to choose a bandwidth h and δ. With the assistance of the 5-fold cross-validation, we chose $\delta = 0.1$ and $h = 0.1, 0.08, 0.075$ and 0.06 respectively for $n = 200, 400, 800$ and 1500 for the Poisson model. For the Bernoulli model, $\delta = 0.005$ and $h = 0.45, 0.4, 0.25$ and 0.18 were chosen respectively for $n = 200, 400, 800$ and 1500.

Note that X_2 and the Z_{ni}'s are not bounded r.v.s as needed in Condition (A) in Section 5. However, these still satisfy the moment conditions needed in the proofs, and Condition (A) is imposed merely to simplify these proofs. Condition (B) is satisfied mainly because the correlations between further Z_{ni}'s are weak, and condition (C) is satisfied because it involves products of standard normal r.v.s which are bounded in the first two moments.

4.1. *Comparisons of algorithms.* We first compare the computing times and the accuracies among three algorithms: 3-step backfitting, 3-step accelerated profile-kernel and fully-iterated profile-kernel algorithms. All of them use the difference-based estimate as the initial estimate. Table 1 summarizes the results based on the Poisson model with 50 samples.

With the same initial values, the backfitting algorithm is slightly faster than the accelerated profile-kernel algorithm, which is in turn by far faster than the full profile-kernel algorithm. Our experience shows that the backfitting algorithm needs more than 20 iterations to converge without improving too much the GMSE. In terms of the accuracy of estimating the parametric component, the accelerated profile-kernel algorithm is about twice as accurate as the backfitting algorithm and the full profile-kernel one. This demonstrates the advantage of keeping the curvature of the least-favorable function in the Newton–Raphson algorithm. For the nonparametric component, we compare RASEs of three algorithms with those based on the oracle estimator, which uses the true value of $\boldsymbol{\beta}_n$. The ratios of the RASEs based on the oracle estimator and those based on the three algorithms are reported in Table 1. It is clear that the accelerated profile-kernel estimate performs very well in estimating the nonparametric components, mimicking very well the oracle estimator. The second best is the backfitting algorithm.

We have also compared the three algorithms using the Bernoulli model. Our proposed accelerated profile-kernel estimate still performs the best in terms of accuracy, though the improvement is not as dramatic as those for the Poisson model. We speculate that the poor performance of the full profile-kernel estimate is due to its unstable implementation that is related to computing the second derivatives of the least-favorable curve.

We next demonstrate the accuracy of the three-step accelerated profile-kernel estimate (3S), compared with the fully-iterated accelerated profile-kernel estimate (AF) (iterating until convergence), and the difference-based estimate (DBE), which is our initial estimate. Table 2 reports the ratios of GMSE based on 400 simulations. It demonstrates convincingly that, with the DBE as the initial estimate, three iterations achieve the accuracy that is comparable with the fully iterated algorithm.

TABLE 2
Medians of the percentages of GMSE based on the accelerated profile-kernel estimates

		Poisson		**Bernoulli**	
n	p_n	**AF/DBE**	**AF/3S**	**AF/DBE**	**AF/3S**
200	10	8.2	99.9	64.1	101.7
400	13	6.0	100.2	52.7	104.7
800	16	5.0	100.1	50.9	102.6
1500	20	4.2	100.0	46.4	100.5

TABLE 3
One-step estimate of parametric components with different bandwidths

		Poisson				Bernoulli	
		Median and SD$_{\mathrm{mad}}$ of GMSE×10^5		Mean and SD of MSE ×10^4 for β_5		Median and SD$_{\mathrm{mad}}$ of GMSE×10	
n	p_n	h_{CV}	$1.5h_{\mathrm{CV}}$	$0.66h_{\mathrm{CV}}$	h_{CV}	$0.66h_{\mathrm{CV}}$	h_{CV}
200	10	5.9 (3.0)	6.4 (3.3)	993 (112)	995 (105)	8.2 (4.4)	8.4 (5.1)
400	13	3.1 (1.4)	3.0 (1.4)	1004 (67)	1001 (65)	4.8 (2.2)	5.4 (2.5)
800	16	1.7 (0.7)	1.7 (0.6)	999 (47)	999 (46)	2.7 (1.0)	2.7 (1.1)
1500	20	1.1 (0.3)	1.1 (0.4)	1000 (32)	1000 (32)	1.8 (0.7)	1.8 (0.6)

SD and SD$_{\mathrm{mad}}$ are shown in parentheses.

In fact, the one-step accelerated profile-kernel estimates improve dramatically (not shown here) our initial estimate (DBE). On the other hand, the DBE itself is not accurate enough for GCVPLM.

The effect of bandwidth choice on the estimation of the parametric component is summarized in Table 3. Denote by h_{CV} the bandwidth chosen by the cross-validation. We scaled the bandwidth up and down by using a factor of 1.5. For illustration, we use the one-step accelerated profile-kernel estimate. The results for the three-step profile-kernel estimate are similar. We evaluate the performance for all components using GMSE and for the specific component β_5 using MSE (the results for other components are similar). We do not report all the results here in order to save space. It is clear that the GMSE does not sensitively depend on the bandwidth, as long as it is reasonably close to h_{CV}. This is consistent with our asymptotic results.

4.2. *Accuracy of profile-likelihood inferences.* To test the accuracy of the sandwich formula for estimating standard errors, the standard deviations of the estimated coefficients (using the one-step accelerated profile-kernel estimate) are computed from the 400 simulations using h_{CV}. These can be regarded as the true standard errors (columns labeled SD). The 400 estimated standard errors are summarized by their median (columns SD$_m$) and its associated SD$_{\mathrm{mad}}$. Table 4 summarizes the results. Clearly, the sandwich formula does a good job, and accuracy gets better as n increases.

We now study the performance of GLRT in Section 2.2. To this end, we consider the following null hypothesis:

$$H_0 : \beta_7 = \beta_8 = \cdots = \beta_{p_n} = 0.$$

We examine the power of the test under a sequence of the alternative hypotheses indexed by a parameter γ as follows:

$$H_1 : \beta_7 = \beta_8 = \gamma, \qquad \beta_j = 0 \qquad \text{for } j > 8.$$

TABLE 4
Standard deviations and estimated standard errors

		Poisson, values x 1000				**Bernoulli, values x 10**			
		$\hat{\beta}_1$		$\hat{\beta}_3$		$\hat{\beta}_2$		$\hat{\beta}_4$	
n	p_n	**SD**	**SD$_\mathbf{m}$**	**SD**	**SD$_\mathbf{m}$**	**SD**	**SD$_\mathbf{m}$**	**SD**	**SD$_\mathbf{m}$**
200	10	9.1	8.5 (1.3)	9.9	9.4 (1.3)	3.6	2.9 (0.4)	3.2	2.8 (0.4)
400	13	6.0	5.6 (0.7)	6.5	6.1 (0.7)	2.3	2.1 (0.2)	2.2	2.0 (0.2)
800	16	3.7	3.8 (0.3)	4.1	4.2 (0.4)	1.7	1.6 (0.1)	1.5	1.5 (0.1)
1500	20	2.8	2.7 (0.2)	3.1	3.0 (0.2)	1.2	1.2 (0.1)	1.1	1.1 (0.1)

SD$_{\text{mad}}$ are shown in parentheses.

When $\gamma = 0$, the alternative hypothesis becomes the null hypothesis.

Under the null hypothesis, the GLRT statistics are computed for each of 400 simulations, using the one-step accelerated profile-kernel estimates. Their distribution is summarized by a kernel density estimate and can be regarded as the true null distribution. This is compared with the asymptotic null distribution $\chi^2_{p_n-6}$. Figures 1(a) and (c) show the results when $n = 400$. The finite sample null density is seen to be reasonably close to the asymptotic one, except for the Monte Carlo error.

The power of the GLR test is studied under a sequence of alternative models, progressively deviating from the null hypothesis, namely, as γ increases. Again, the one-step accelerated profile-kernel algorithm is employed. The power functions are calculated at three significance levels: 0.1, 0.05 and 0.01, using the asymptotic distribution. They are the proportion of rejection among the 400 simulations and are depicted in Figures 1(b) and (d). The power curves increase rapidly with γ, which shows the GLR test is powerful. The powers at $\gamma = 0$ are approximately the same as the significance level except the Monte Carlo error. This shows that the size of the test is reasonably accurate.

4.3. *A real data example.* This is the analysis of the data in Section 1.1, where details of data and variables are given.

To examine the nonlinear effect of age and its nonlinear interaction with the experience, we appeal to the following GVCPLM (interactions between age and covariates other than **TotalYrsExp** are considered but found to be insignificant):

$$\begin{aligned}\log\left(\frac{p_H}{1-p_H}\right) &= \alpha_1(\text{Age}) + \alpha_2(\text{Age})\text{TotalYrsExp} \\ &\quad + \beta_1\text{Female} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i,\end{aligned} \tag{4.1}$$

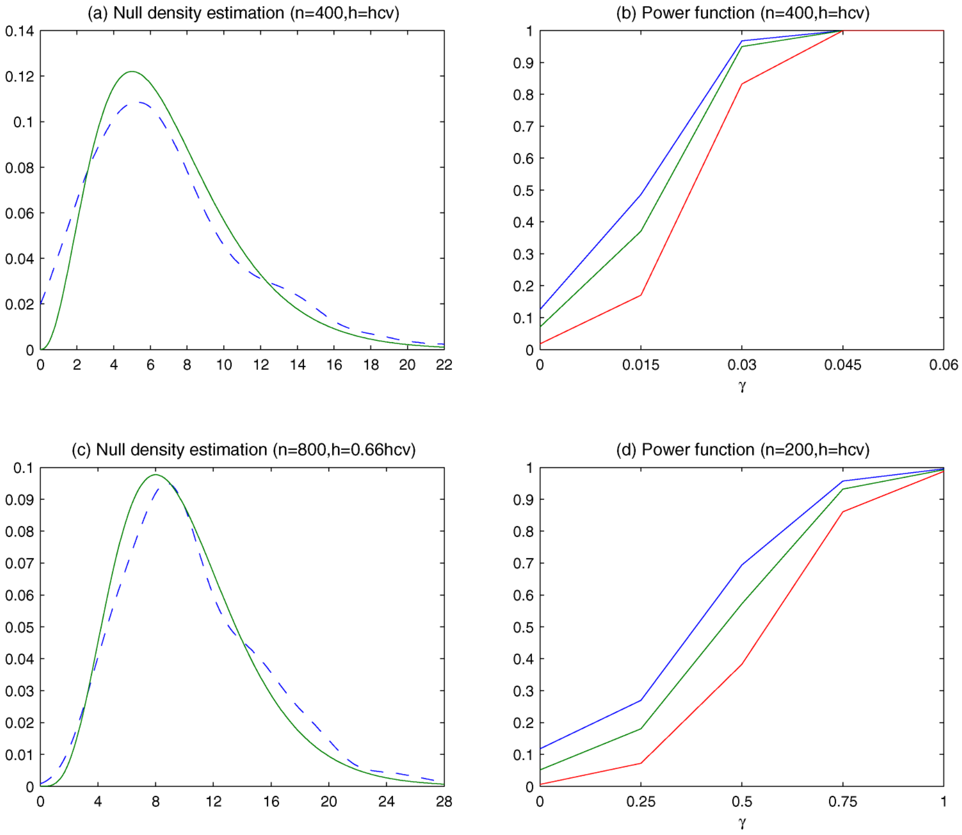

FIG. 1. (a) *Asymptotic null distribution* (*solid*) *and estimated true null distribution* (*dotted*) *for the Poisson model*. (b) *The power function at significant level* $\alpha = 0.01, 0.05$ *and* 0.1. *The captions for* (c) *and* (d) *are the same as those in* (a) *and* (b) *except that the Bernoulli model is now used.*

where p_H is the probability of having a high grade job. Formally, we are testing

$$H_0 : \beta_1 = 0 \longleftrightarrow H_1 : \beta_1 < 0. \tag{4.2}$$

A 20-fold CV is employed to select the bandwidth h and the parameter δ in the transformation of the data. This yields $h_{\mathrm{CV}} = 24.2$, $\delta_{\mathrm{CV}} = 0.1$. Table 5 shows the results of the fit using the three-step accelerated profile-kernel estimate. The

TABLE 5
Fitted coefficients (*sandwich SD*) *for model* (4.1)

Response	**Female**	**PCJob**	**Edu$_1$**	**Edu$_2$**	**Edu$_3$**	**Edu$_4$**
HighGrade4	−1.96 (0.57)	−0.02 (.076)	−5.14 (0.85)	−4.77 (0.98)	−2.72 (0.52)	−2.85 (0.96)
HighGrade5	−2.22 (0.59)	−1.96 (0.61)	−5.69 (0.67)	−5.95 (0.97)	−3.09 (0.72)	−1.26 (1.10)

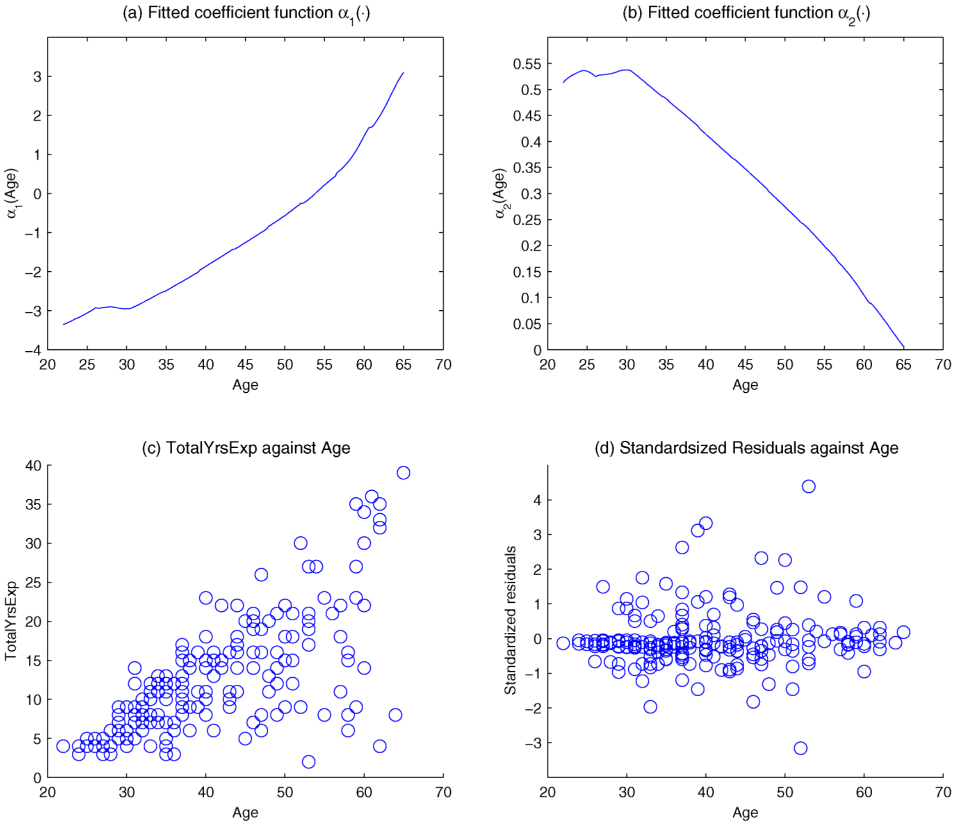

FIG. 2. (a) *Fitted coefficient function* $\alpha_1(\cdot)$. (b) *Fitted coefficient function* $\alpha_2(\cdot)$. (c) *The scatter plot "TotalYrsExp" against "Age."* (d) *Standardized residuals against the variable "Age."*

coefficient for **Female** is significantly negative. The education also plays an important role in getting a high grade job. All coefficients are negative, as they are contrasted with the highest education level. The **PCJob** does not seem to play any significant role in getting promotion. Figures 2(a) and (b) depict the estimated coefficient functions. They show that as age increases, one has a better chance of being in a higher job grade, and then the marginal effect of working experience is large when age is around 30 or less, but starts to fall as one gets older. However, the second result should be interpreted with caution, as the variables **Age** and **TotalYrsExp** are highly correlated [Figure 2(c)]. The standardized residuals $(y - \hat{p}_{\mathrm{H}})/\sqrt{\hat{p}_{\mathrm{H}}(1 - \hat{p}_{\mathrm{H}})}$ against **Age** is plotted in Figure 2(d). It shows that the fit seems reasonable. Other diagnostic plots also look reasonable, but they are not shown here.

We have conducted another fit using a binary variable **HighGrade5**, which is 0 only when job grade is less than 5. The coefficients are shown in Table 5 and the **Female** coefficient is close to the first fit.

We now employ the generalized likelihood ratio test to the problem (4.2). The GLR test statistic is 14.47 with one degree of freedom, resulting in a P-value of 0.0001. We have also conducted the same analysis using **HighGrade5** as the binary response. The GLR test statistic is now 13.76 and the associated P-value is 0.0002. The fitted coefficients are summarized in Table 5. The result provides stark evidence that even after adjusting for other confounding factors and variables, it is harder for female employees of the Fifth National Bank to get promoted to a high grade job.

Not shown in this paper, we have conducted the analysis again after deleting 6 data points corresponding to 5 male executives and 1 female employee having many years of working experience and high salaries. The test results are still similar.

5. Technical proofs. In this section the proofs of Theorems 1–4 will be given. We introduce some notation and regularity conditions for our results. In the following and thereafter, the symbol $\otimes$ represents the Kronecker product between matrices, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the minimum and maximum eigenvalues of a symmetric matrix A. We let $Q_{ni}(\boldsymbol{\beta}_n)$ be the ith summand of (2.3).

Denote the true linear parameter by $\boldsymbol{\beta}_{n0}$, with parameter space $\Omega_n \subset \mathbb{R}^{p_n}$. Let $\mu_k = \int_{-\infty}^{\infty} u^k K(u)\,du$ and $A_p(\mathbf{X}) = (\mu_{i+j})_{0\le i,j\le p} \otimes \mathbf{X}\mathbf{X}^T$. Set

$$\rho_l(t) = (dg^{-1}(t)/dt)^l / V(g^{-1}(t)), \qquad m_{ni}(\boldsymbol{\beta}_n) = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni},$$

$$\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u) = \frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}, \qquad \boldsymbol{\alpha}^{(r)\prime\prime}_{\boldsymbol{\beta}_n}(u) = \frac{\partial^2 \boldsymbol{\alpha}^{(r)}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n \, \partial \boldsymbol{\beta}_n^T}.$$

REGULARITY CONDITIONS.

(A) The covariates $\mathbf{Z}_n$ and $\mathbf{X}$ are bounded random variables.

(B) The smallest and the largest eigenvalues of the matrix $I_n(\boldsymbol{\beta}_{n0})$ are bounded away from zero and infinity for all n. In addition, the expectation $\mathbf{E}_0[\nabla^T Q_{n1}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})]^4 = O(p_n^4)$.

(C) $\mathbf{E}_{\boldsymbol{\beta}_n} |\frac{\partial^{l+j} Q_{n1}(\boldsymbol{\beta}_n)}{\partial^j \boldsymbol{\alpha}\, \partial\beta_{nk_1} \cdots \partial\beta_{nk_l}}|$ and $\mathbf{E}_{\boldsymbol{\beta}_n} |\frac{\partial^{l+j} Q_{n1}(\boldsymbol{\beta}_n)}{\partial^j \boldsymbol{\alpha}\, \partial\beta_{nk_1} \cdots \partial\beta_{nk_l}}|^2$ are bounded, with $l = 1, \ldots, 4$ and $j = 0, 1$.

(D) The function $q_2(x, y) < 0$ for $x \in \mathbb{R}$ and y in the range of the response variable, and $\mathbf{E}_0\{q_2(m_{n1}(\boldsymbol{\beta}_n), Y_{n1}) A_p(\mathbf{X}_1) | U = u\}$ is invertible.

(E) The functions $V''(\cdot)$ and $g'''(\cdot)$ are continuous. The least-favorable curve $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ is three times continuously differentiable in $\boldsymbol{\beta}_n$ and u.

(F) The random variable U has a compact support Ω. The density function $f_U(u)$ of U has a continuous second derivative and is uniformly bounded away from zero.

(G) The kernel K is a bounded symmetric density function with bounded support.

Note the above conditions are assumed to hold uniformly in $u \in \Omega$. Condition (A) is imposed just for the simplicity of proofs. The boundedness of covariates is imposed to ensure various products involving $q_l(\cdot,\cdot)$, $\mathbf{X}$ and $\mathbf{Z}_n$ have bounded first and second moments. Conditions (B) and (C) are uniformity conditions on higher-order moments of the likelihood functions. They are stronger than those of the usual asymptotic likelihood theory, but they facilitate technical proofs. Condition (G) is also imposed for simplicity of technical arguments. All of these conditions can be relaxed at the expense of longer proofs.

Before proving Theorem 1, we need two important lemmas. Lemma 1 concerns the order approximations to the least-favorable curve $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(\cdot)$, while Lemma 2 holds the key to showing why undersmoothing is not needed in Theorems 1 and 2. Let $c_n = (nh)^{-1/2}$ and $\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}, \dots, \hat{\mathbf{a}}_{p\boldsymbol{\beta}_n}$ maximize (2.4).

LEMMA 1. *Under Regularity Conditions* (A)–(G), *for each* $\boldsymbol{\beta}_n \in \Omega_n$, *the following holds uniformly in* $u \in \Omega$:

$$\|\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)\| = O_P\bigl(h^{p+1} + c_n \log^{1/2}(1/h)\bigr).$$

Likewise, the norm of the kth *derivative of the above with respect to any* β_{nj}*'s, for* $k = 1, \dots, 4$, *all have the same order uniformly in* $u \in \Omega$.

We omit the proof of Lemma 1. Please refer to the technical report [15] for a proof.

LEMMA 2. *Under Regularity Conditions* (A)–(G), *if* $p_n^s/n \to 0$ *for* $s > 5/4$, $h = O(n^{-a})$ *with* $(2s(p+1))^{-1} < a < 1 - s^{-1}$, *then for each* $\boldsymbol{\beta}_n \in \Omega_n$,

$$n^{-1/2}\|\nabla \hat{Q}_n(\boldsymbol{\beta}_n) - \nabla Q_n(\boldsymbol{\beta}_n)\| = o_P(1).$$

PROOF. Define

$$\mathbf{K}_1 = n^{-1/2}\sum_{i=1}^n q_2(m_{ni}(\boldsymbol{\beta}_n), Y_{ni})\bigl(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i\bigr)\bigl(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)\bigr)^T\mathbf{X}_i,$$

$$\mathbf{K}_2 = n^{-1/2}\sum_{i=1}^n q_1(m_{ni}(\boldsymbol{\beta}_n), Y_{ni})\bigl(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)\bigr)^T\mathbf{X}_i.$$

Then by Taylor's expansion, Lemma 1 and Condition (C),

$$n^{-1/2}\bigl(\nabla \hat{Q}_n(\boldsymbol{\beta}_n) - \nabla Q_n(\boldsymbol{\beta}_n)\bigr) = \mathbf{K}_1 + \mathbf{K}_2 + \text{smaller order terms}.$$

Define, for Ω as in Condition (F),

$$S = \{f \in C^2(\Omega) : \|f\|_\infty \le 1\},$$

equipped with a metric $\rho(f_1, f_2) = \|f_1 - f_2\|_\infty$, where $\|f\|_\infty = \sup_{u\in\Omega} |f(u)|$. We also let, for $r = 1, \dots, q$ and $l = 1, \dots, p_n$,

$$A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n) = q_2(\mathbf{X}^T \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u) + \mathbf{Z}_n^T \boldsymbol{\beta}_n, y) X_r \biggl(Z_{nl} + \mathbf{X}^T \frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial \beta_{nl}} \biggr),$$

$$B_r(y, u, \mathbf{X}, \mathbf{Z}_n) = q_1(\mathbf{X}^T \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u) + \mathbf{Z}_n^T \boldsymbol{\beta}_n, y) X_r.$$

By Lemma 1, for any positive sequences (δ_n) with $\delta_n \to 0$ as $n \to \infty$, we have $P_0(\lambda_r \in S) \to 1$ and $P_0(\gamma_{rl} \in S) \to 1$, where

$$\lambda_r = \delta_n \bigl(h^{p+1} + c_n \log^{1/2}(1/h)\bigr)^{-1} (\hat{\alpha}_{\boldsymbol{\beta}_n}^{(r)} - \alpha_{\boldsymbol{\beta}_n}^{(r)}),$$

$$\gamma_{rl} = \delta_n \bigl(h^{p+1} + c_n \log^{1/2}(1/h)\bigr)^{-1} \biggl(\frac{\partial \hat{\alpha}_{\boldsymbol{\beta}_n}^{(r)}}{\partial \beta_{nl}} - \frac{\partial \alpha_{\boldsymbol{\beta}_n}^{(r)}}{\partial \beta_{nl}} \biggr),$$

$r = 1, \dots, q$ and $l = 1, \dots, p_n$. Hence, for sufficiently large n, we have $\lambda_r, \gamma_{rl} \in S$. The following three points allow us to utilize [14] to prove our lemma:

I. For any $v \in S$, we will view the map $v \mapsto A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n) v(u)$ as an element of $C(S)$, the space of continuous functions on S equipped with the sup norm. For $v_1, v_2 \in S$, we have

$$\begin{aligned} &|A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n) v_1(u) - A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n) v_2(u)| \\ &\quad = |A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n)(v_1 - v_2)(u)| \le |A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n)| \|v_1 - v_2\|. \end{aligned}$$

A similar result holds for $B_r(y, u, \mathbf{X}, \mathbf{Z}_n)$.

II. Note that equation (2.2) is true for all $\boldsymbol{\beta}_n$, and by differentiating w.r.t. $\boldsymbol{\beta}_n$, we get the following formulas:

$$E_0\bigl(q_1(m_n(\boldsymbol{\beta}_n), Y_n)\mathbf{X} | U = u\bigr) = \mathbf{0},$$

$$E_0\bigl(q_2(m_n(\boldsymbol{\beta}_n), Y_n)\mathbf{X}\bigl(\mathbf{Z}_n + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U)\mathbf{X}\bigr)^T | U = u\bigr) = \mathbf{0}.$$

Thus, we can easily see that

$$E_0(A_{rl}(Y, U, \mathbf{X}, \mathbf{Z}_n)) = 0$$

for each $r = 1, \dots, q$ and $l = 1, \dots, p_n$. Also, we have

$$E_0(A_{rl}(Y, U, \mathbf{X}, \mathbf{Z}_n)^2) < \infty,$$

by Regularity Conditions (A) and (C). For $B_r(Y, U, \mathbf{X}, \mathbf{Z}_n)$, results hold similarly.

III. Let $H(\cdot, S)$ denote the metric entropy of the set S w.r.t. the metric ρ. Then

$$H(\epsilon, S) \le C_0 \epsilon^{-1}$$

for some constant C_0. Hence, $\int_0^1 H^{1/2}(\epsilon, S)\, d\epsilon < \infty$.

Conditions of Theorem 1 in [14] can be derived from the three notes above, so that we have

$$n^{-1/2}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})(\cdot),$$

where $A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})(\cdot)$, $i = 1, \dots, n$, being i.i.d. replicates of the function $A_{rl}(Y, U, \mathbf{X}, \mathbf{Z}_n)(\cdot)$ in $C(S)$, converges weakly to a Gaussian measure on $C(S)$. Hence, since $\lambda_r, \gamma_{rl} \in S$,

$$n^{-1/2}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})(\lambda_r) = O_P(1),$$

which implies that

$$n^{-1/2}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})\big(\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n} - \alpha^{(r)}_{\boldsymbol{\beta}_n}\big) = O_P\big(\delta_n^{-1}\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big).$$

Similarly, applying Theorem 1 of [14] again, we have

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} B_r(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})&\left(\frac{\partial \hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}}{\partial \beta_{nl}} - \frac{\partial \alpha^{(r)}_{\boldsymbol{\beta}_n}}{\partial \beta_{nl}}\right) \\ &= O_P\big(\delta_n^{-1}\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big). \end{aligned}$$

Then the column vector $\mathbf{K}_1$, which is p_n-dimensional has the lth component equal

$$\begin{aligned} \sum_{r=1}^{q}\left\{n^{-1/2}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})\big(\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n} - \alpha^{(r)}_{\boldsymbol{\beta}_n}\big)\right\} \\ = O_P\big(\delta_n^{-1}\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big), \end{aligned}$$

using the result just proved. Hence, we have shown

$$\|\mathbf{K}_1\| = O_P\big(\sqrt{p_n}\delta_n^{-1}\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big) = o_P(1),$$

since δ_n can be made arbitrarily slow in converging to 0. Similarly, we have $\|\mathbf{K}_2\| = o_P(1)$ as well. The conclusion of the lemma follows. □

LEMMA 3. *Assuming Regularity Conditions* (A)–(G), *we have for each* $\boldsymbol{\beta}_n \in \Omega_n$,

$$n^{-1}\|\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n) - \nabla^2 Q_n(\boldsymbol{\beta}_n)\| = O_P\big(p_n\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big).$$

LEMMA 4. *Under Regularity Conditions* (A)–(G) *and* $p_n^4/n = o(1)$,

$$\|n^{-1}\nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})\| = o_P(p_n^{-1}),$$

$$\|n^{-1}\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})\| = o_P(p_n^{-1}) + O_P\big(p_n\big(h^{p+1} + c_n \log^{1/2}(1/h)\big)\big).$$

We omit the proofs of the lemmas. Please refer to the technical report [15] for a proof.

PROOF OF THEOREM 1. Let $\gamma_n = \sqrt{p_n/n}$. Our aim is to show that, for a given $\epsilon > 0$,

$$\mathbb{P}\Big\{\sup_{\|\mathbf{v}\|=C} \hat{Q}_n(\boldsymbol{\beta}_{n0} + \gamma_n \mathbf{v}) < \hat{Q}_n(\boldsymbol{\beta}_{n0})\Big\} \geq 1 - \epsilon, \tag{5.1}$$

so that this implies, with probability tending to 1, there is a local maximum $\hat{\boldsymbol{\beta}}_n$ in the ball $\{\boldsymbol{\beta}_{n0} + \gamma_n \mathbf{v} : \|\mathbf{v}\| \leq C\}$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P(\gamma_n)$.

Define the terms $\hat{I}_1 = \gamma_n \nabla^T \hat{Q}_n(\boldsymbol{\beta}_{n0})\mathbf{v}$, $\hat{I}_2 = \frac{\gamma_n^2}{2}\mathbf{v}^T \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0})\mathbf{v}$ and $\hat{I}_3 = \frac{\gamma_n^3}{6}\nabla^T(\mathbf{v}^T \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n^*)\mathbf{v})\mathbf{v}$. By Taylor's expansion,

$$\hat{Q}_n(\boldsymbol{\beta}_{n0} + \gamma_n \mathbf{v}) - \hat{Q}_n(\boldsymbol{\beta}_{n0}) = \hat{I}_1 + \hat{I}_2 + \hat{I}_3,$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_{n0}$ and $\boldsymbol{\beta}_{n0} + \gamma_n \mathbf{v}$.

We further split $\hat{I}_1 = D_1 + D_2$, where

$$D_1 = \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\big(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_i)\mathbf{X}_i\big)^T \mathbf{v}\gamma_n,$$

$$D_2 = \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\mathbf{X}_i^T\big(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}}(U_i) - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_i)\big)^T \mathbf{v}\gamma_n,$$

with $\hat{m}_{ni}(\boldsymbol{\beta}_n) = \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}$. By Condition (A) and Lemma 1, D_2 has order smaller than D_1. Using Taylor's expansion, we have

$$D_1 = \gamma_n \mathbf{v}^T\left(\sum_{i=1}^n \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \boldsymbol{\beta}_n} + \sqrt{n}\mathbf{K}_1\right) + \text{smaller order terms},$$

where $\mathbf{K}_1$ is as defined in Lemma 2 so that within the lemma's proof we have $\|\mathbf{K}_1\| = o_P(1)$. Using equation (2.6), we have, by the mean–variance decomposition,

$$\left\|\mathbf{v}^T \sum_{i=1}^n \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \boldsymbol{\beta}_n}\right\| = O_P\big(\sqrt{n\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}}\big) = O_P(\sqrt{n})\|\mathbf{v}\|,$$

where the last inequality follows from Condition (B). Hence,

$$|\hat{I}_1| = O_P(\sqrt{n}\gamma_n)\|\mathbf{v}\|.$$

Next, consider $\hat{I}_2 = I_2 + (\hat{I}_2 - I_2)$, where

$$\begin{aligned} I_2 &= \frac{1}{2}\mathbf{v}^T \nabla^2 Q_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 \\ &= -\frac{n}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 + \frac{n}{2}\mathbf{v}^T\{n^{-1}\nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})\}\mathbf{v}\gamma_n^2 \\ &= -\frac{n}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 + o_P(n\gamma_n^2)\|\mathbf{v}\|^2, \end{aligned}$$

with the last line following from Lemma 4. Using Lemma 3,

$$\|\hat{I}_2 - I_2\| = o_P(n\gamma_n^2\|\mathbf{v}\|^2).$$

On the other hand, by Condition (B), we have

$$|n\gamma_n^2\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}| \geq O(n\gamma_n^2\lambda_{\min}(I_n(\boldsymbol{\beta}_{n0}))\|\mathbf{v}\|^2) = O(n\gamma_n^2\|\mathbf{v}\|^2).$$

Hence, $\hat{I}_2 - I_2$ has a smaller order than I_2.

Finally, consider $\hat{I}_3$. We suppress the dependence of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$ and its derivatives on U_i, and denote $q_{1i} = q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})$. Using Taylor's expansions, expanding $\hat{Q}_n(\boldsymbol{\beta}_n^*)$ at $\boldsymbol{\beta}_{n0}$ and then $\hat{Q}_n(\boldsymbol{\beta}_{n0})$ at $\boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}$, we can arrive at

$$\begin{aligned} \hat{Q}_n(\boldsymbol{\beta}_n^*) = Q_n(\boldsymbol{\beta}_{n0}) + \sum_{i=1}^n \{ & q_{1i}\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}) \\ & + q_{1i}(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)^T(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_{n0})\}(1 + o_P(1)). \end{aligned}$$

Substituting $\hat{Q}_n(\boldsymbol{\beta}_n^*)$ into $\hat{I}_3$ with the right-hand side above, by Condition (C) and Lemma 1, we have

$$\hat{I}_3 = \frac{1}{6}\sum_{i,j,k=1}^{p_n} \frac{\partial^3 Q_n(\boldsymbol{\beta}_{n0})}{\partial\boldsymbol{\beta}_{ni}\,\partial\boldsymbol{\beta}_{nj}\,\partial\boldsymbol{\beta}_{nk}} v_i v_j v_k \gamma_n^3 + \text{smaller order terms.}$$

Hence,

$$|\hat{I}_3| = O_P\big(np_n^{3/2}\gamma_n^3\|\mathbf{v}\|^3\big) = O_P\big(\sqrt{p_n^4/n}\|\mathbf{v}\|\big)n\gamma_n^2\|\mathbf{v}\|^2 = o_P(1)n\gamma_n^2\|\mathbf{v}\|^2.$$

Comparing, we find the order of $-n\gamma_n^2\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}$ dominates all other terms by allowing $\|\mathbf{v}\| = C$ to be large enough. This proves (5.1). □

PROOF OF THEOREM 2. Note that by Theorem 1, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P(\sqrt{p_n/n})$. Since $\nabla\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) = 0$, by Taylor's expansion,

$$\nabla\hat{Q}_n(\boldsymbol{\beta}_{n0}) + \nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + \mathcal{C} = 0, \tag{5.2}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_{n0}$ $\hat{\boldsymbol{\beta}}_n$ and $\mathcal{C} = \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T \nabla^2(\nabla \hat{Q}_n(\boldsymbol{\beta}_n^*))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}))$, which is understood as a vector of quadratic components.

Using a similar argument to approximating $\hat{I}_3$ in Theorem 1, by Lemma 1 and noting $\|\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_{n0}\| = o_P(1)$, we have $\|\nabla^2 \frac{\partial \hat{Q}_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}}\|^2 = O_P(n^2 p_n^2)$. Hence,

$$\begin{aligned}\|n^{-1}\mathcal{C}\|^2 &\le n^{-2}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|^4 \sum_{j=1}^{p_n} \left\| \nabla^2 \frac{\partial \hat{Q}_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}} \right\|^2 \\ &= O_P(p_n^5/n^2) = o_P(n^{-1}).\end{aligned} \tag{5.3}$$

At the same time, by Lemma 4 and the Cauchy–Schwarz inequality,

$$\begin{aligned}&\|n^{-1}\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})\| \\ &\quad = o_P((np_n)^{-1/2}) + O_P\bigl(\sqrt{p_n^3/n}\bigl(h^{p+1} + c_n \log^{1/2}(1/h)\bigr)\bigr) \\ &\quad = o_P(n^{-1/2}).\end{aligned} \tag{5.4}$$

Combining (5.2), (5.3) and (5.4), we have

$$\begin{aligned}I_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) &= n^{-1}\nabla \hat{Q}_n(\boldsymbol{\beta}_{n0}) + o_P(n^{-1/2}) \\ &= n^{-1}\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(n^{-1/2}),\end{aligned} \tag{5.5}$$

where the last line follows from Lemma 2. Consequently, using equation (5.5), we get

$$\begin{aligned}&\sqrt{n} A_n I_n^{1/2}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \\ &\quad = n^{-1/2} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0})) \\ &\quad = n^{-1/2} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(1),\end{aligned} \tag{5.6}$$

since $\|A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0})\| = O(1)$ by conditions of Theorem 2.

We now check the Lindeberg–Feller Central Limit Theorem (see, e.g., [25]) for the last term in (5.6). Let $B_{ni} = n^{-1/2} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_{ni}(\boldsymbol{\beta}_{n0})$, $i = 1, \dots, n$. Given $\epsilon > 0$,

$$\sum_{i=1}^n E_0 \|B_{ni}\|^2 1\{\|B_{ni}\| > \epsilon\} \le n\sqrt{E_0\|B_{n1}\|^4 \cdot \mathbb{P}(\|B_{n1}\| > \epsilon)}.$$

Using Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}(\|B_{n1}\| > \epsilon) &\le n^{-1}\epsilon^{-2} E\|A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})\|^2 \\ &= n^{-1}\epsilon^{-2}\,\mathrm{tr}(G) = O(n^{-1}),\end{aligned} \tag{5.7}$$

where tr(A) is the trace of square matrix A. Similarly, we can show that, using Condition (B),

$$\begin{aligned} E_0\|B_{n1}\|^4 &\le \sqrt{l}n^{-2}\lambda^2_{\min}(A_nA_n^T)\lambda^2_{\max}(I_n(\boldsymbol{\beta}_{n0})) \\ &\quad \times \sqrt{E_0(\nabla Q_{n1}(\boldsymbol{\beta}_{n0})^T\nabla Q_{n1}(\boldsymbol{\beta}_{n0}))^4} \\ &= O(p_n^2/n^2). \end{aligned} \tag{5.8}$$

Therefore, (5.7) and (5.8) together imply

$$\sum_{i=1}^n E_0\|B_{ni}\|^2 1\{\|B_{ni}\| > \epsilon\} = O\big(\sqrt{p_n^2/n}\big) = o(1).$$

Also,

$$\begin{aligned} \sum_{i=1}^n \mathrm{Var}_0(B_{ni}) &= \mathrm{Var}_0(A_nI_n^{-1/2}(\boldsymbol{\beta}_{n0})\nabla Q_{n1}(\boldsymbol{\beta}_{n0})) \\ &= A_nA_n^T \to G. \end{aligned}$$

Therefore, B_{ni} satisfies the conditions of the Lindeberg–Feller Central Limit Theorem. Consequently, using (5.6), it follows that

$$\sqrt{n}A_nI_n^{1/2}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \xrightarrow{\mathscr{D}} N(0, G),$$

and this completes the proof. □

Referring back to Section 2.2, let B_n be a $(p_n - l) \times p_n$ matrix satisfying $B_nB_n^T = I_{p_n-l}$ and $A_nB_n^T = 0$. Since $A_n\boldsymbol{\beta}_n = 0$ under H_0, rows of A_n are perpendicular to $\boldsymbol{\beta}_n$ and the orthogonal complement of rows of A_n is spanned by rows of B_n since $A_nB_n^T = 0$. Hence,

$$\boldsymbol{\beta}_n = B_n^T\boldsymbol{\gamma}$$

under H_0, where $\boldsymbol{\gamma}$ is a $(p_n - l) \times 1$ vector. Then, under H_0, the profile likelihood estimator is also the local maximizer $\hat{\boldsymbol{\gamma}}_n$ of the problem

$$\hat{Q}_n(B_n^T\hat{\boldsymbol{\gamma}}_n) = \max_{\boldsymbol{\gamma}_n} Q_n(B_n^T\boldsymbol{\gamma}_n).$$

LEMMA 5. *Assuming the conditions in Theorem 3 and under the null hypothesis H_0 as in the theorem,*

$$B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}) = \frac{1}{n}B_n^T\{B_nI_n(\boldsymbol{\beta}_{n0})B_n^T\}^{-1}B_n^T\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(n^{-1/2}).$$

We omit the proof of the lemma. Please refer to the technical report [15] for a proof.

PROOF OF THEOREM 3. By Taylor's expansion, expanding $\hat{Q}_n(B_n^T\hat{\boldsymbol{\gamma}}_n)$ at $\hat{\boldsymbol{\beta}}_n$ and noting that we have $\nabla^T\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) = 0$, then $\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol{\gamma}}_n) = T_1 + T_2$, where

$$T_1 = -\tfrac{1}{2}(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)^T\nabla^2\hat{Q}_n(\hat{\boldsymbol\beta}_n)(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n),$$
$$T_2 = \tfrac{1}{6}\nabla^T\{(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)^T\nabla^2\hat{Q}_n(\beta_n^*)(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)\}(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n).$$

Denote by $\Theta_n = I_n(\boldsymbol\beta_{n0})$ and $\boldsymbol\Phi_n = \frac{1}{n}\nabla Q_n(\boldsymbol\beta_{n0})$. Using equation (5.5) and noting that Θ_n has eigenvalues uniformly bounded away from 0 and infinity by Condition (B), we have $\hat{\boldsymbol\beta}_n - \boldsymbol\beta_{n0} = \Theta_n^{-1}\boldsymbol\Phi_n + o_P(n^{-1/2})$. Combining this with Lemma 5, under the null hypothesis H_0,

$$\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n = \Theta_n^{-1/2}\{I_{p_n} - \Theta_n^{1/2}B_n^T(B_n\Theta_nB_n^T)^{-1}B_n\Theta_n^{1/2}\}\Theta_n^{-1/2}\boldsymbol\Phi_n + o_P(n^{-1/2}). \tag{5.9}$$

Since $S_n = I_{p_n} - \Theta_n^{1/2}B_n^T(B_n\Theta_nB_n^T)^{-1}B_n\Theta_n^{1/2}$ is a $p_n \times p_n$ idempotent matrix with rank l, it follows by mean–variance decomposition of the term $\|\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n\|^2$ and Condition (B) that $\|\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n\| = O_P(n^{-1/2})$. Hence, using a similar argument as in the approximation of order for $|\hat{I}_3|$ in Theorem 1, we have

$$|T_2| = O_P(np_n^{3/2})\cdot\|\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n\|^3 = o_P(1).$$

Hence, $\hat{Q}_n(\hat{\boldsymbol\beta}_n) - \hat{Q}(B_n^T\hat{\boldsymbol\gamma}_n) = T_2 + o_P(1)$.

By Lemma 4 and an approximation to $n^{-1}\|\nabla^2\hat{Q}_n(\hat{\boldsymbol\beta}_n) - \nabla^2\hat{Q}_n(\boldsymbol\beta_{n0})\| = o_P(p_n^{-1/2})$ (the proof is similar to that for Lemma 6 with the proof of order for $|\hat{I}_3|$ in Theorem 1, and is omitted), we have

$$\begin{aligned}&\|\tfrac{1}{2}(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)^T\{\nabla^2\hat{Q}_n(\hat{\boldsymbol\beta}_n) + nI_n(\boldsymbol\beta_{n0})\}(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)\|\\&\quad = O_P(l/n)\cdot n\{o_P(p_n^{-1/2}) + O_P(p_n(h^{p+1} + c_n\log^{1/2}(1/h)))\} = o_p(1).\end{aligned}$$

Therefore,

$$\hat{Q}_n(\hat{\boldsymbol\beta}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol\gamma}_n) = \frac{n}{2}(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n)^TI_n(\boldsymbol\beta_{n0})(\hat{\boldsymbol\beta}_n - B_n^T\hat{\boldsymbol\gamma}_n) + o_P(1).$$

By (5.9), we have $\hat{Q}_n(\hat{\boldsymbol\beta}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol\gamma}_n) = \frac{n}{2}\boldsymbol\Phi_n^T\Theta_n^{-1/2}S_n\Theta_n^{-1/2}\boldsymbol\Phi_n + o_P(1)$. Since S_n is idempotent, it can be written as $S_n = D_n^TD_n$, where D_n is an $l \times p_n$ matrix satisfying $D_nD_n^T = I_l$. By Theorem 2, we have already shown that $\sqrt{n}D_n\Theta_n^{-1/2}\boldsymbol\Phi_n \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_l)$. Hence,

$$2\{\hat{Q}_n(\hat{\boldsymbol\beta}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol\gamma}_n)\} = n(D_n\Theta_n^{-1/2}\boldsymbol\Phi_n)^T(D_n\Theta_n^{-1/2}\boldsymbol\Phi_n) \xrightarrow{\mathcal{D}} \chi_l^2. \qquad \square$$

LEMMA 6. *Assuming Conditions* (A)–(G) *and* $p_n^4/n = o(1)$, *we have*

$$n^{-1}\|\nabla^2Q_n(\hat{\boldsymbol\beta}_n) - \nabla^2Q_n(\boldsymbol\beta_{n0})\| = o_P(1).$$

We omit the proof of the lemma. Please refer to the technical report [15] for a proof.

PROOF OF THEOREM 4. Let $\hat{\mathcal{A}}_n = -n^{-1}\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)$, $\hat{\mathcal{B}}_n = \widehat{\text{cov}}\{\nabla \hat{Q}_n(\hat{\boldsymbol{\beta}}_n)\}$ and $\mathcal{C} = I_n(\boldsymbol{\beta}_{n0})$. Write

$$I_1 = \hat{\mathcal{A}}_n^{-1}(\hat{\mathcal{B}}_n - \mathcal{C})\hat{\mathcal{A}}_n^{-1}, \qquad I_2 = \hat{\mathcal{A}}_n^{-1}(\mathcal{C} - \hat{\mathcal{A}}_n)\hat{\mathcal{A}}_n^{-1},$$
$$I_3 = \hat{\mathcal{A}}_n^{-1}(\mathcal{C} - \hat{\mathcal{A}}_n)\mathcal{C}^{-1}.$$

Then, $\hat{\Sigma}_n - \Sigma_n = I_1 + I_2 + I_3$. Our aim is to show that, for all $i = 1, \dots, p_n$,

$$\lambda_i(\hat{\Sigma}_n - \Sigma_n) = o_P(1),$$

so that $A_n(\hat{\Sigma}_n - \Sigma_n)A_n^T \xrightarrow{\mathbb{P}} 0$, where $\lambda_i(A)$ is the ith eigenvalue of a symmetric matrix A. Using the inequalities

$$\lambda_{\min}(I_1) + \lambda_{\min}(I_2) + \lambda_{\min}(I_3) \le \lambda_{\min}(I_1 + I_2 + I_3),$$
$$\lambda_{\max}(I_1 + I_2 + I_3) \le \lambda_{\max}(I_1) + \lambda_{\max}(I_2) + \lambda_{\max}(I_3),$$

it suffices to show that $\lambda_i(I_j) = o_P(1)$ for $j = 1, 2, 3$. From the definition of I_1, I_2 and I_3, it is clear that we only need to show $\lambda_i(\mathcal{C} - \hat{\mathcal{A}}_n) = o_P(1)$ and $\lambda_i(\hat{\mathcal{B}}_n - \mathcal{C}) = o_P(1)$. Let $K_1 = I_n(\boldsymbol{\beta}_{n0}) + n^{-1}\nabla^2 Q_n(\boldsymbol{\beta}_{n0})$, $K_2 = n^{-1}(\nabla^2 Q_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2 Q_n(\boldsymbol{\beta}_{n0}))$ and $K_3 = n^{-1}(\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2 Q_n(\hat{\boldsymbol{\beta}}_n))$. Then,

$$\mathcal{C} - \hat{\mathcal{A}}_n = K_1 + K_2 + K_3.$$

Applying Lemma 4 to K_1, Lemma 6 to K_2 and Lemma 3 to K_3, we have $\|\mathcal{C} - \hat{\mathcal{A}}\| = o_P(1)$. Thus, $\lambda_i(\mathcal{C} - \hat{\mathcal{A}}) = o_P(1)$. Hence, the only thing left to show is $\lambda_i(\hat{\mathcal{B}}_n - \mathcal{C}) = o_P(1)$.

To this end, consider the decomposition

$$\hat{\mathcal{B}}_n - \mathcal{C} = K_4 + K_5,$$

where

$$K_4 = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} \right\} - I_n(\boldsymbol{\beta}_{n0}),$$
$$K_5 = -\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} \right\}.$$

Our goal is to show that K_4 and K_5 are $o_P(1)$, which then implies $\lambda_i(\hat{\mathcal{B}}_n - \mathcal{C}) = o_P(1)$. We consider K_4 first, which can be further decomposed into $K_4 = K_6 + K_7$, where

$$K_6 = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\},$$

$$K_7 = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\} - I_n(\boldsymbol{\beta}_{n0}).$$

Observe that

$$\begin{aligned} K_6 = \Bigg\{ & \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \left\{ \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\} \\ & + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \left\{ \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \right\} \\ & + \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\} \left\{ \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \right\} \Bigg\}, \end{aligned}$$

and this suggests that an approximation of the order of $\frac{\partial}{\partial \beta_{nk}}(\hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\boldsymbol{\beta}_{n0}))$ for each $k = 1, \ldots, p_n$ and $i = 1, \ldots, n$ is rewarding. Define

$$a_{ik} = \frac{\partial}{\partial \beta_{nk}}(\hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\hat{\boldsymbol{\beta}}_n)) \quad \text{and} \quad b_{ik} = \frac{\partial}{\partial \beta_{nk}}(Q_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\boldsymbol{\beta}_{n0})),$$

then $\frac{\partial}{\partial \beta_{nk}}(\hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\boldsymbol{\beta}_{n0})) = a_{ik} + b_{ik}$. By Taylor's expansion, suppressing dependence of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$ and its derivatives on U_i,

$$a_{ik} = \left\{ \frac{\partial^2 Q_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}\, \partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}^T} (\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n} - \boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}) + \frac{\partial Q_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}^T} \left(\frac{\partial \hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}}{\partial \beta_{nk}} - \frac{\partial \boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}}{\partial \beta_{nk}} \right) \right\} (1 + o_P(1)).$$

Using Lemma 1, Condition (C), with an argument similar to the proof of Lemma 3, we then have

$$a_{ik} = O_P\big(h^{p+1} + c_n \log^{1/2}(1/h)\big).$$

Similarly, Taylor's expansion gives

$$b_{ik} = \frac{\partial^2 Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}\, \partial {\boldsymbol{\beta}_n}^T} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) (1 + o_P(1)),$$

which implies that, by Theorem 1 and Regularity Condition (C),

$$|b_{ik}| = O_P\big(\sqrt{p_n^2/n}\big).$$

Using the approximations of a_{ik} and b_{ik} above, by Condition (C),

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\left\{\frac{\partial\hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial\beta_{nk}}-\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nk}}\right\}\right| &\le\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\right|\cdot|a_{ik}+b_{ik}|\\ &=O_P\big(h^{p+1}+c_n\log^{1/2}(1/h)+n^{-1/2}p_n\big).\end{aligned}$$

This shows that

$$\|K_6\|=O_P\big(p_n\big(h^{p+1}+c_n\log^{1/2}(1/h)\big)+p_n^2n^{-1/2}\big)=o_P(1)$$

by the conditions of the theorem.

For K_7, note that $\|(K_7)\|=O_P(p_n^2/n)=o_P(1)$, since

$$E_0K_7=n^{-2}(np_n^2)E_0\left\{\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nk}}-E_0\left(\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nk}}\right)\right\}^2.$$

Hence, using $K_4=K_6+K_7$,

$$\|K_4\|=o_P(1)+O_P\big(p_n\big(h^{p+1}+c_n\log^{1/2}(1/h)\big)+\sqrt{p_n^4/n}\big)=o_P(1).$$

Finally, consider K_5. Defining $A_j=n^{-1}\sum_{i=1}^{n}(a_{ij}+b_{ij})+n^{-1}\sum_{i=1}^{n}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}$, where a_{ij} and b_{ij} are defined as before, we can then rewrite $K_5=\{A_jA_k\}$. Now

$$\begin{aligned}|A_j|&\le\sup_{i,j}|a_{ij}+b_{ij}|+\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\right|\\ &=O_P\big(h^{p+1}+c_n\log^{1/2}(1/h)+n^{-1/2}p_n\big)+O_P(n^{-1/2}),\end{aligned}$$

where the last line follows from the approximations for a_{ij} and b_{ij}, and mean–variance decomposition of the term $n^{-1}\sum_{i=1}^{n}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}$. Hence,

$$\|K_5\|=O_P\big(p_n\big(h^{p+1}+c_n\log^{1/2}(1/h)+n^{-1/2}p_n\big)^2\big)=o_P(1),$$

and this completes the proof. □

PROOF OF THEOREM 5. In expression (2.4), we set $p=0$, which effectively assumes $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)\approx\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ for U_i in a neighborhood of u. Then by definition, $\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)$ maximizes (2.4), which leads to $\sum_{i=1}^{n}q_1(\mathbf{X}_i^T\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)+\mathbf{Z}_{ni}^T\boldsymbol{\beta}_n)\mathbf{X}_iK_h(U_i-u)=0$. Differentiating this w.r.t. β_{nj}, we have

$$\sum_{i=1}^{n}q_2(\mathbf{X}_i^T\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)+\mathbf{Z}_{ni}^T\boldsymbol{\beta}_n,Y_{ni})\left(Z_{nij}+\left(\frac{\partial\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial\beta_{nj}}\right)^T\mathbf{X}_i\right)\mathbf{X}_iK_h(U_i-u)=0.$$

Solving for $\frac{\partial \hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}$ from the above equation, which is true for $j = 1, \ldots, p_n$, we get the same expression as given in the lemma.

Hence, it remains to show that $\frac{\partial \hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}$ is a consistent estimator of $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$. This is done by Lemma 1, where

$$\left\| \frac{\partial \hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n} - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u) \right\| = O_P\bigl(\sqrt{p_n}(h + c_n \log^{1/2}(1/h))\bigr) = o_P(1),$$

and the proof completes. □

REFERENCES

[1] AHMAD, I., LEELAHANON, S. and LI, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Ann. Statist.* **33** 258–283. MR2157803

[2] ALBRIGHT, S. C., WINSTON, W. L. and ZAPPE, C. J. (1999). *Data Analysis and Decision Making with Microsoft Excel*. Pacific Grove, Duxbury, CA. Available at http://www.alibris.com/booksearch.detailinvid=9470354547&browse=1&qwork=1492588&qsort=&page=1.

[3] BICKEL, P. J. (1975). One-step Huber estimates in linear models. *J. Amer. Statist. Assoc.* **70** 428–433. MR0386168

[4] CAI, Z., FAN, J. and LI, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446

[5] DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, 2000, to the American Mathematical Society on "Math Challenges of the 21st Century." Available at http://www.inma.ucl.ac.be/~francois/these/papers/entry-Donoho-2000.html.

[6] FAN, J. and HUANG, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

[7] FAN, J. and LI, R. (2006). Statistical challenges with high-dimensionality: Feature selection in knowledge discovery. *Proceedings of International Congress of Mathematicians* (M. Sanz-Solé, J. Soria, J. L. Varona and J. Verdera, eds.) **III** 595–622. MR2275698

[8] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[9] FAN, J., PENG, H. and HUANG, T. (2005). Semilinear high-dimensional model for normalization of microarray data: A theoretical analysis and partial consistency (with discussion). *J. Amer. Statist.* **100** 781–813. MR2201010

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

[11] HASTIE, T. J. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

[12] HU, Z., WANG, N. and CARROLL, R. J. (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* **91** 251–262. MR2081299

[13] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

[14] JAIN, N. and MARCUS, M. (1975). Central limit theorems for C(S)-valued random variables. *J. Funct. Anal.* **19** 216–231. MR0385994

[15] LAM, C. and FAN, J. (2007). Profile-kernel likelihood inference with diverging number of parameters. Available at http://arxiv.org/PS_cache/math/pdf/0701/0701004v2.pdf.

[16] LI, Q., HUANG, C. J., LI., D. and FU, T. T. (2002). Semiparametric smooth coefficient models. *J. Bus. Econom. Statist.* **20** 412–422. MR1939909

[17] LI, R. and LIANG, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36** 261–286.

[18] LIN, X. and CARROLL, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. Roy. Statist. Soc. Ser. B* **68** 69–88. MR2212575

[19] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London. MR0727836

[20] MURPHY, S. A. (1993). Testing for a time dependent coefficient in Cox's regression model. *Scand. J. Statist.* **20** 35–50. MR1221960

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

[22] PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

[23] ROBINSON, P. M. (1988). The stochastic difference between econometric and statistics. *Econometrica* **56** 531–547. MR0946120

[24] SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802. MR1193312

[25] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press. MR1652247

[26] VAN KEILEGOM, I. and CARROLL, R. J. (2007). Backfitting versus profiling in general criterion functions. *Statist. Sinica* **17** 797–816.

[27] XIA, Y., ZHANG, W. and TONG, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91** 661–681. MR2090629

[28] YATCHEW, A. (1997). An elementary estimator for the partially linear model. *Economics Lett.* **57** 135–143. MR1600233

[29] ZHANG, W., LEE, S. Y. and SONG, X. Y. (2002). Local polynomial fitting in semivarying coefficient model. *J. Multivariate Anal.* **82** 166–188. MR1918619

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: wlam@princeton.edu
jqfan@princeton.edu