Clifford Lam

Profile-kernel likelihood inference with diverging number of parameters

Article (Accepted version) (Refereed)

Original citation:
Lam, Clifford and Fan, Jianqing (2008) *Profile-kernel likelihood inference with diverging number of parameters. The annals of statistics*, 36 (5). pp. 2232-2260. ISSN 0090-5364

DOI: 10.1214/07-AOS544

A Profile-Kernel Estimation with Diverging Number of Linear Parameters

By Clifford Lam and Jiangqing Fan
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ, 08544

May 26, 2006

Abstract

A generalization to the varying coefficient model, the generalized varying coefficient partially linear model (GVCPLM) has gained significant attention because of its generality and incorporated predictive and explanatory power. Since modern statistical problems usually deal with data of vast dimensionality, a large model is usually unavoidable for predictive purpose. In this paper we set foot on both theoretical and practical sides of profile likelihood estimation when the number of linear parameters in the model grows with sample size. Existence of profile likelihood estimator and asymptotic normality for the linear parameters are established under regularity conditions. Profile likelihood ratio statistic for the linear parameters is discussed and Wilk's phenomenon demonstrated as proposed by Fan, Zhang and Zhang (2001). We propose a profile-kernel based algorithm for evaluating the varying coefficients and the linear parameters. Simulation study shows that the resulting estimates are as efficient as the fully iterative profile-kernel estimates. For moderate sample size, our proposed procedure saves much computational time over the fully iterative profile-kernel one and gives stabler estimates. A set of real data has been analyzed using the GVCPLM with our proposed algorithm.

1 Introduction

The generalized varying-coefficient models, proposed by Hastie and Tibshirani (1993), has attracted more attention over the last decade. It is a form of semiparametric regression which extends the generalized linear model (e.g. McCullagh and Nelder (1989)) naturally so that the linear parameters become nonparametric functions of a covariate U, e.g. time variable in a longitudinal data analysis. For instance, see Cai, Fan and Li (2000) for a detailed account on statistical inferences on such models and references therein. A further generalization to the generalized varying coefficient model is to allow for an additive parametric part, resulting in the generalized varying coefficient partially linear model (GVCPLM). 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

$$g\{\mu(u, \mathbf{x}, \mathbf{z})\} = \mathbf{x}^T \boldsymbol{\alpha}(u) + \mathbf{z}^T \boldsymbol{\beta}, \tag{1}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\beta}$ an unknown regression coefficient and $\boldsymbol{\alpha}(\cdot)$ an unknown regression function. One of the advantages over the varying coefficient model is that GVCPLM allows for estimation of effects more efficiently when they are not really varying with U, after adjustment of other genuine varying effects. It also allows for more interpretable model, where primary interest is focused on the parametric component. This model is relatively new in the literature. Instead, a special case called the partially linear model (PLM) is studied more extensively, where the vector $\mathbf{x}$ is set to the scalar 1. See, for example, Engle,*et al.* (1986), Wahba (1984) and Speckman (1988). Severini and Wong (1992) established theories in generalized profile likelihood approach for efficient estimation of the parametric component without the need of undersmoothing, and Severini and Staniswalis (1994) proposed an iterative procedure for this profile likelihood estimation. Carroll *et al.* (1997) studied the generalized partially linear single-index model. More references can be found in Härdle, Liang and Gao (2000).

The goals of this paper are two-fold: to establish theories in statistical inferences when the dimension of the parametric component diverges with the sample size, and to compute the estimates efficiently without sacrificing accuracy.

For the estimation aspect, Zhang, Lee and Song (2002), Li, Huang, Ki and Fu (2002) and Xia, Zhang and Tong (2004) considered the varying coefficient partially linear model (VCPLM, g being the identity link) and proposed different methods of estimation. Ahmad, Leelahanon and Li (2005) considered a series approximation approach for estimating the nonparametric component in the VCPLM, while Fan and Huang (2005) proposed a profile-kernel approach for the VCPLM which has closed form solutions. Li and Liang (2005) considered a backfitting-based procedure for estimating a GVCPLM (a general link g).

In this paper we propose a profile-kernel procedure for the GVCPLM in (1) based on Newton-Raphson iterations. Computational difficulties (e.g. Lin and Carroll (2006)) of the profile-kernel approach is overcome by introducing modifications to updating of the parametric component. For moderate sample size the computational expenses are then greatly reduced while nice properties of profile-kernel approach over backfitting (e.g. Hu *et al.* (2004)) are retained. This will be further demonstrated in section 4, where Poisson and Logistic GVCPLM are considered for simulations. We also introduce a difference-based estimation for the parametric component of the GVCPLM, which serves well for an initial estimate of our proposed profile-kernel procedure. Such an idea for estimation is used, for example, in Yatchew (1997) for the partial linear model.

For estimation with diverging number of parameters, early of such works include Huber (1973) (more of his work can be found in Huber (1981)) which gave related theories on M-estimators, and Portnoy (1988) which analyzed a regular exponential family under the same setting. Fan and Peng (2004) analyzed a general parametric model using the penalized likelihood approach under such setting. Donoho (2000) gave a full introduction on how high dimensional data affects the trend of data analysis, with examples in various fields of applications. Fan and Li (2006) proposed the penalized likelihood method to achieve both estimation and variable selection simultaneously in various fields involving high dimensional data analysis. We give two examples where a large number of parameters is to be estimated relative to the sample size.

Example 1 (*Framingham Heart Study (FHS)*). In this classical study initiated in 1948, the FHS follows a representative sample of 5,209 adults and their offspring aged 28-62 years in Framingham, Massachusetts. One goal of the study is to identify major risk factors associated with heart disease, stroke and other diseases. The study lasted for more than half a century, with original participants' adult children and their spouses also participated in the study. There are around $p = 100$ variables for the study and so the number of parameters is large relative to the sample size. For more information on this study, see the website of National Heart, Lung and Blood Institute (http://www.nhlbi.nih.gov/about/framingham).

Example 2 (*Computational Biology*). DNA microarrays monitor the mRNA expressions of thousands of genes in many areas of biomedical research. The cDNA microarrays measures the abundance of mRNA expressions by mixing mRNAs of treatment and control cells or tissues. However, systematic biases due to experimental variations have to be removed first before the expression data can be used for further analysis. Example of such biases include efficiency of dye incorporation, intensity effect and print-tip block effect, among others. The process of removing these experimental biases is called normalization, and is critical to multiple array comparison.

Let Y_g be the log-ratio of the intensity of gene g of the treatment sample over that of the control sample. Denote A_g the average log-intensities of gene g at the treatment and control samples, r_g and c_g the row and column of the block where the cDNA of gene g resides. Fan *et al.*(2004) proposed the following model to estimate the intensity and block effect:

$$Y_g = \alpha_g + \beta_{r_g} + \gamma_{c_g} + f(A_g) + \epsilon_g,\ g = 1, \cdots, N$$

where α_g is the treatment effect of gene g, β_{r_g} and γ_{r_g} are block effects decomposed into row and column components, $f(A_g)$ represents the intensity effect, and N is the total number of genes. Even with replications of genes, we can see that the above model has number of parameters $p = O(N)$. However the number of significant genes is relatively small, so that α_g has a sparse structure. The goal is to find genes g with α_g statistically significantly different from 0.

The outline of the paper is as follows. In section 2 we briefly review the profile likelihood estimation with local polynomial modelling, as well as presenting asymptotic results in sections 2.1-2.3. Section 3 turns to the computational aspect, and sections 3.1-3.4 discuss the elements of our proposed profile-kernel procedure, as well as how to choose smoothing parameters. A simulation study is given in section 4, as well as an analysis of a real data set using the proposed methodology. The proofs of our results is given in section 5, and technical details in the appendix.

2 Properties of profile likelihood estimation

Let $(Y_{ni}; \mathbf{X}_i, \mathbf{Z}_{ni}, U_i)_{1\le i\le n}$ be a random sample where Y_{ni} is a scalar response variable, U_i is a scalar variable, $\mathbf{X}_i \in \mathbb{R}^q$ and $\mathbf{Z}_{ni} \in \mathbb{R}^{p_n}$ are vectors of explanatory variables. Note that $\mathbf{Y}_{ni}$ and $\mathbf{Z}_{ni}$ depends on n, and $p_n \to \infty$ as $n \to \infty$.

The model we consider for the data is the generalized varying coefficient partially linear model(GVCPLM), as in model (1), with $\boldsymbol{\beta}_n$ and $\mathbf{Z}_n$ having dimensions depending on n now. The quasi-likelihood function for the response Y is

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

where $V(\cdot)$ is the variance function for Y. As in Severini and Wong (1992), we denote by $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ the 'least favorable curve' of the nonparametric function $\boldsymbol{\alpha}(u)$ when we fix the linear parameter to be $\boldsymbol{\beta}_n$ for estimation purpose. It can be defined such that

$$(2) \qquad \frac{\partial}{\partial \boldsymbol{\eta}} E_0 \left\{ Q(g^{-1}(\boldsymbol{\eta}^T \mathbf{X} + {\boldsymbol{\beta}_n}^T \mathbf{Z}_n), Y_n) | U = u \right\} |_{\boldsymbol{\eta} = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)} = 0,$$

where E_0 means expectation is taken under the true parameters $\boldsymbol{\alpha}_0(u)$ and $\boldsymbol{\beta}_{n0}$. Note that $\boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(u) = \boldsymbol{\alpha}_0(u)$. The global likelihood function for the data is then

$$(3) \qquad Q_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} Q\{g^{-1}(\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}), Y_{ni}\}.$$

To estimate the parameters in (3), we first treat $\boldsymbol{\beta}_n$ as a constant. The model then becomes purely nonparametric and estimation of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$ is done through a local polynomial regression of order p for the j^{th} component of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$, which approximate

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

for U in a neighborhood of u. Denote $\mathbf{a_r} = (a_{r1}, \cdots, a_{rq})^T$ for $r = 0, \ldots, p$, noting that they depend on $\boldsymbol{\beta}_n$. We then maximize the local likelihood

$$(4) \qquad \sum_{i=1}^{n} Q\{g^{-1}(\sum_{r=0}^{p} \mathbf{a_r}^T \mathbf{X}_i (U_i - u)^r / r! + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}), Y_{ni}\} K_h(U_i - u)$$

with respect to $\mathbf{a_0}, \cdots, \mathbf{a_p}$. $K(\cdot)$ is a kernel function, and $K_h(t) = K(t/h)/h$ is a rescaling of K with bandwidth h. So we get estimate $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_j) = \hat{\mathbf{a}}_{\mathbf{0}}(U_j)$ for $j = 1, \ldots, n$.

Plugging our estimates into the global likelihood function (3), we have

$$(5) \qquad \hat{Q}_n(\boldsymbol{\beta}_n) := \sum_{i=1}^{n} Q\{g^{-1}(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}), Y_{ni}\}.$$

This is now a pure parametric model with parameter $\boldsymbol{\beta}_n$. Maximizing $\hat{Q}_n(\boldsymbol{\beta}_n)$ with respect to $\boldsymbol{\beta}_n$ to get $\hat{\boldsymbol{\beta}}_n$, which amounts to solving $\nabla \hat{Q}_n(\boldsymbol{\beta}_n) = 0$. With $\hat{\boldsymbol{\beta}}_n$, we estimate our varying coefficients as $\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(u)$.

One property of the quasi-likelihood is that the first and second order Bartlett's identities hold. In particular, if we define the marginal global likelihood for $\boldsymbol{\beta}_n$ as in (3), then

$$(6) \qquad \mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial Q_n}{\partial \boldsymbol{\beta}_n}\right) = 0, \; nI_n(\boldsymbol{\beta}_n) = \mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial Q_n}{\partial \boldsymbol{\beta}_n}\frac{\partial Q_n}{\partial \boldsymbol{\beta}_n^T}\right) = -\mathbf{E}_{\boldsymbol{\beta}_n}\left(\frac{\partial^2 Q_n}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T}\right),$$

where $I_n(\boldsymbol{\beta}_n)$ is the marginal Fisher Information of a single observation for $\boldsymbol{\beta}_n$ (See Severini and Wong (1992) for more details).

Note that equation (2) is true for all $\boldsymbol{\beta}_n$, and so by differentiating w.r.t. $\boldsymbol{\beta}_n$ we get the following important formulas:

$$\begin{aligned} &E_0(q_1(m_n(\boldsymbol{\beta}_n), Y_n)\mathbf{X}|U=u) = \mathbf{0}, \\ &E_0(q_2(m_n(\boldsymbol{\beta}_n), Y_n)\mathbf{X}(\mathbf{Z}_n + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U)\mathbf{X})^T|U=u) = \mathbf{0}, \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u) = \frac{\partial \boldsymbol{\alpha}_{\beta_n}(u)}{\partial \boldsymbol{\beta}_n}$ and $q_l(x,y) = \frac{d^l}{dx^l}Q(g^{-1}(x), y)$.

In the subsequent sections we need some regularity conditions, which are presented in section 5, for our results to hold.

2.1 Asymptotic normality and consistency of $\hat{\boldsymbol{\beta}}_n$

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 $nh^{2p+2} = O(1)$ with $nh^{p+2} \to \infty$, then there is a local maximizer $\hat{\boldsymbol{\beta}}_n \in \Omega_n$ of $\hat{Q}_n(\boldsymbol{\beta}_n)$ such that $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\right\| = O_P(\sqrt{p_n/n})$.*

This consistent rate is the same as the result of the M-estimator that was studied by Huber (1973), in which the number of parameters diverges. This rate of convergence is also obtained by Zhang, Lee and Song (2002) for p_n a constant. They also assumed $nh^{2p+2} = O(1)$.

Since the usual optimal bandwidth for minimizing conditional MSE or weighted MISE is $h = O(n^{-1/(2p+3)})$(Fan and Gijbels (1996)), it does not satisfy the assumption $nh^{2p+2} = O(1)$. However, note that under the optimal bandwidth, we have $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\right\| = O_P(\sqrt{p_n/n^{(2p+2)/(2p+3)}})$ (follow the same lines of proof in theorem 1 to get this). This rate is worse than $\sqrt{n/p_n}$, but with somewhat stronger assumption $p_n^5/n^{(2p+1)/(2p+3)} = o(1)$, a form of $\sqrt{n}$-consistency can be recovered as in theorem 2. In particular, if $\sup_n p_n < \infty$, this stronger assumption is automatically satisfied, showing that $\sqrt{n}$-consistency can be achieved under optimal bandwidth. This is in line with the results, for instance, by Severini and Staniswalis(1994) or Carroll et al. (1997).

Theorem 2 (Asymptotic normality). *Under Conditions (A) - (G), if $p_n^5/n \to 0$ as $n \to \infty$, then the $\sqrt{n/p_n}$-consistent local maximizer $\hat{\boldsymbol{\beta}}_n$ in theorem 1 satisfies*

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

where A_n is an $l \times p_n$ matrix such that $A_n A_n^T \to G$, and G is a $l \times l$ nonnegative symmetric matrix. Furthermore, if $p_n^5/n^{(2p+1)/(2p+3)} \to 0$, then the local maximizer $\hat{\boldsymbol{\beta}}_n$

in theorem 1, estimated under the optimal bandwidth $h = O(n^{-1/(2p+3)})$, still satisfies the above asymptotic normality.

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

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

which shows that the variance of $\hat{\boldsymbol{\beta}}_n$ achieves the efficient lower bound (See for example Carroll et al. (1997)). This also agrees with the result by Fan and Huang(2005), who studied the same type of model under the usual linear regression setting with p_n a constant. The result presented here can be considered a further generalization of theirs.

2.2 Hypothesis testing

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:

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

where A_n is a $l \times p_n$ matrix and $A_nA_n^T = I_l$ for a fixed l. Both the null and the alternative hypotheses are semi-parametric, with nuisance functions $\boldsymbol{\alpha}(\cdot)$. The generalized likelihood ratio test (GLRT) has statistic of the form

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

where $\hat{Q}_n(\boldsymbol{\beta}_n)$ is as defined in (5). It turns out 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. This reveals the Wilk's phenomenon, as termed in Fan et al (2001). Hence under a semi-parametric model with increasing number of parameters, traditional likelihood ratio theory continues to apply and testing of linear hypotheses becomes easy.

Theorem 3 *Assuming conditions (A) - (G), under H_0, we have*

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

provided that $p_n^5/n \to 0$ when $nh^{2p+2} = O(1)$, or $p_n^5/n^{(2p+1)/(2p+3)} \to 0$ when $h = O(n^{-1/(2p+3)})$.

2.3 Consistency of the sandwich covariance formula

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

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

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

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

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

Theorem 4 *Assuming conditions (A) - (G). If $p_n^5/n \to 0$ when $nh^{2p+2} = O(1)$ and $nh^2 \to \infty$ as $n \to \infty$, we have*

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

for any $l \times p_n$ (l is a fixed integer) matrix A_n such that $A_n A_n^T = G$. The same conclusion holds if $p_n^5/n^{(2p+2)/(2p+3)} = o(1)$ when $h = O(n^{-(2p+3)})$.

This result provides a way for constructing confidence intervals for $\boldsymbol{\beta}_n$. However we stress the independence of such estimate in testing hypothesis as in section 2.2. Simulation results show that this formula indeed provide good estimates of the variances for $\hat{\boldsymbol{\beta}}_n$. For more details on sandwich covariance formula, see Kauermann and Carroll (2001).

The theorems presented so far have assumptions $p_n^4/n = o(1)$ or $p_n^5/n = o(1)$ which are somewhat strong. However we will use $p_n^3/n = O(1)$ in our simulation in section 4 to demonstrate a wider applicability of our theories in models like the generalized linear models.

3 Computation of the estimates

A profile-kernel approach for estimating $\boldsymbol{\beta}_n$ in (3) is to find $\hat{\boldsymbol{\beta}}_n$ maximizing (5). Backfitting algorithm, on the other hand, does not assume $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ in (5) to depend on $\boldsymbol{\beta}_n$, and the maximization w.r.t. $\boldsymbol{\beta}_n$ is thus much easier to carry out. The updated $\boldsymbol{\beta}_n$ is then substituted into (4) to find $\hat{\boldsymbol{\alpha}}(u)$ again, and the iterations repeated until convergence.

See Lin and Carroll (2006), Hu *et al* (2004) for more descriptions of the two methods and some closed-form solutions proposed for the partially linear models.

In general, the profile-kernel estimation can be carried out through the use of the Newton-Raphson algorithm on updating both $\boldsymbol{\beta}_n$ and $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ alternately. We will describe modifications and implementations of the following steps in subsequent sections:

Unmodified profile-kernel updating procedure

Step 0 (Initialization). Find $\boldsymbol{\beta}_n^{(0)}$, an initial estimate for $\boldsymbol{\beta}_n$. Set $k = 0$.

Step 1. Compute $b_i = \mathbf{Z}_{ni}^T \boldsymbol{\beta}_n^{(k)}$. Replaces $\mathbf{Z}_{ni}^T \boldsymbol{\beta}_n$ in (3) by b_i and the problem becomes purely nonparametric (generalized varying coefficient model). Efficient estimation for $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n^{(k)}}(u)$ is available, for instance, in Cai, Fan and Li(2000).

Step 2. Replaces $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ in (3) by $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ and the problem becomes purely parametric. Perform a Newton-Raphson iteration

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

Here $\hat{Q}_n(\boldsymbol{\beta}_n)$ is as defined in (5). Derivative is taken with respect to $\boldsymbol{\beta}_n$, noting that $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ depends on $\boldsymbol{\beta}_n$ as well. Set k to $k+1$.

Step 3. Iterate steps 1 and 2 until convergence.

Section 3.1 gives a detail account of obtaining an initial estimate for $\boldsymbol{\beta}_n$.

For modifications, we introduce a quick implementation of step 2 in section 3.3, which not only helps save vast amount of computational time for moderate sample size, but also is much stabler comparing with the full procedure.

The idea behind the foregoing algorithm is to estimate a least favorable curve $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ for $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ at $\boldsymbol{\beta}_n = \boldsymbol{\beta}_n^{(k)}$ in light of lemma 6, which then allow us to update $\boldsymbol{\beta}_n^{(k)}$ to $\boldsymbol{\beta}_n^{(k+1)}$ as in step 2. Step 1 involves nonparametric estimation and is discussed in section 3.2.

In step 3 we need to iterate steps 1 and 2 until convergence. In practice, as is demonstrated in simulation study in section 4, only several iterations are needed for practical accuracy. We name the estimates by doing step 0 and step 1 the one-cycle estimates, and those obtained by iterating steps 2 and step 1 $(m-1)$ more times as the ***m*-cycles estimates**.

3.1 Difference-based estimation for VCPLM

The idea of differencing to remove nonparametric part in a partially linear model (PLM) has been applied, with different usages, in Yatchew (1997) and Fan and Huang (2005). We generalize this idea and apply on the varying coefficient partially linear model (VCPLM).

Consider the VCPLM with the structure

$$Y = \boldsymbol{\alpha}(U)^T\mathbf{X} + {\boldsymbol{\beta}_n}^T\mathbf{Z}_n + \varepsilon, \tag{8}$$

where Y is a response variable and $(U, \mathbf{X}^T, {\mathbf{Z_n}}^T)$ is the vector of associated covariates, with $\mathbf{X}$ being a q dimensional and $\mathbf{Z}_n$ being a p_n dimensional vector. The error term ε has mean 0 and unknown variance σ^2. This is a special case of the GVCPLM where in equation (3), g is the identity link and Q is the log-likelihood of normal density. However it is only used to motivate our procedure.

Let $\{(U_i, \mathbf{X}_i^T, \mathbf{Z}_{ni}^T, Y_i)\}_{i=1}^n$ be a random sample from (8) above, with the data ordered according to the U_i's. Under mild conditions, the spacing $U_{i+1} - U_i$ is $O_P(1/n)$, so that $\boldsymbol{\alpha}(U_{i+1}) - \boldsymbol{\alpha}(U_i) \approx \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1}(U_{i+1} - U_i)$. Using model (8),

$$\sum_{j=1}^{q+1} w_j Y_{i+j-1} = \sum_{j=1}^{q+1} w_j \boldsymbol{\alpha}(U_{i+j-1})^T \mathbf{X}_{i+j-1} + \boldsymbol{\beta}_n^T \sum_{j=1}^{q+1} w_j \mathbf{Z}_{n(i+j-1)} + \sum_{j=1}^{q+1} w_j \varepsilon_{i+j-1}.$$

Here w_j depends on i as well, but we drop this subscript for simplicity. If we define $Y_i^* = \sum_{j=1}^{q+1} w_j Y_{i+j-1}$, $\mathbf{Z}_{ni}^* = \sum_{j=1}^{q+1} w_j \mathbf{Z}_{n(i+j-1)}$, $\varepsilon_i^* = \sum_{j=1}^{q+1} w_j \varepsilon_{i+j-1}$ and impose the constraint $\sum_{j=1}^{q+1} w_j \mathbf{X}_{i+j-1} = \mathbf{0}$, then we can re-write the above equation as

$$Y_i^* \approx {\boldsymbol{\gamma_0}}^T \sum_{r=2}^{q+1} \sum_{j=r}^{q+1} w_j \mathbf{X}_{i+j-1} + {\boldsymbol{\gamma_1}}^T \sum_{r=2}^{q+1} \sum_{j=r}^{q+1} w_j \mathbf{X}_{i+j-1}(U_{i+r-1} - U_{i+r-2}) + \boldsymbol{\beta}_n^T \mathbf{Z}_i^* + \varepsilon_i^*,$$

which is a linear model with parameter $(\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}, \boldsymbol{\beta}_n)$. In our simulation study in section 4, we choose i to be $1, 2, \cdots, n-q$ so that we have exactly $(n-q)$ 'starred' data points and the ε_i^*'s are dependent in general, but with known dependence structure. So we can perform a weighted least square fit to the starred data to find $(\hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\beta}}_n)$. To solve $\sum_{j=1}^{q+1} w_j \mathbf{X}_{i+j-1} = \mathbf{0}$, we need to find the rank r of the matrix $(\mathbf{X}_i, \cdots, \mathbf{X}_{i+q})$, and then fix $q+1-r$ of the w_j's so that the rest can be determined uniquely by just solving a system of linear equations.

One concern of the above approximation is the sparsity of the U_i's, especially in the tail regions. Then $O_P(1/n)$ spacing is not achievable in the tails. In this case we may want to remove these sparse data points first before aggregating with w_j's to avoid deterioration of quality for the estimate $\hat{\boldsymbol{\beta}}_n$. In section 4, we take U to be uniformly distributed over $(0, 1)$ so that sparsity problem can be avoided for the ease of our demonstration.

To use the differencing idea to obtain an initial estimate of $\boldsymbol{\beta}_n$ for GVCPLM, we apply 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, adjustments needed to be made in transforming the data. We use $g(y) = \log\left(\frac{y+\delta}{1-y+\delta}\right)$ for the logistic regression and $g(y) = \log(y + \delta)$ for the Poisson regression. Here δ is treated as a smoothing parameter like h in estimating varying coefficients, and the choice of which are discussed in section 3.4.

3.2 One-step estimation for the nonparametric component

Given $\boldsymbol{\beta}_n = \boldsymbol{\beta}_n^{(k)}$, model (3) becomes purely nonparametric and we estimate the varying coefficients $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ by using the one-step local MLE. The one-step estimates are as efficient as the fully iterative ones but save considerable computational time. For more theoretical properties, see for example Cai, Fan and Li (2000). We briefly describe the method here.

The local likelihood is as defined in (4), denoted by $l_{\boldsymbol{\beta}_n}(\boldsymbol{\gamma}, u)$, where $\boldsymbol{\gamma} = \left(\mathbf{a_0}^T, \cdots, \mathbf{a_p}^T\right)^T$. Given an initial estimator $\hat{\boldsymbol{\gamma}}_0 = \hat{\boldsymbol{\gamma}}_0(u_0) = \left(\hat{\mathbf{a}}_\mathbf{0}(u_0)^T, \cdots, \hat{\mathbf{a}}_\mathbf{p}(u_0)^T\right)^T$, one step of the Newton-Raphson algorithm produces the updated estimator

$$\hat{\boldsymbol{\gamma}}_{\text{OS}} = \hat{\boldsymbol{\gamma}}_0 - \{\nabla^2 l_{\boldsymbol{\beta}_n}(\hat{\boldsymbol{\gamma}}_0, u_0)\}^{-1} \nabla l_{\boldsymbol{\beta}_n}(\hat{\boldsymbol{\gamma}}_0, u_0),$$

where derivatives are taken with respect to $\boldsymbol{\gamma}$. In univariate generalized linear models, the least-squares estimate serves a natural candidate as an initial estimator. We adapt a variation as described in Cai, Fan and Li (2000), where we first find a sub-grid points of all the U_i's and obtain local MLE $\hat{\boldsymbol{\gamma}}$ on the sub-grid points. Then use these estimates as initial values for carrying out the one-step local MLE procedure on the rest of the U_i's.

The matrix $\nabla^2 l_{\boldsymbol{\beta}_n}(\boldsymbol{\gamma}, u)$ can be nearly singular for certain U_i, due to possible data sparsity in certain local regions, or when bandwidth is too small. We adapt the ridge regression approach to overcome this problem. We omit the details here.

3.3 Fast updating of $\boldsymbol{\beta}_n^{(k)}$

The profile-kernel approach essentially treats $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ from step 1 as a function of both u and $\boldsymbol{\beta}_n$ (Lin and Carroll (2006)). Updating of $\boldsymbol{\beta}_n^{(k)}$ in step 2 needs the first and second derivatives of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)$ with respect to $\boldsymbol{\beta}_n$, which can be computationally intensive to calculate. More precisely, denote $\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u) = \frac{\partial \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}$ which is a p_n by q matrix, $\alpha^{(r)}_{\boldsymbol{\beta}_n}(u)$ the r^{th} component of $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ and $\hat{m}_{ni}(\boldsymbol{\beta}_n) = \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$, we need to calculate

$$\begin{aligned}
&\nabla \hat{Q}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i),\\
(9)\qquad &\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^n q_2(\hat{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i)(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i)^T\\
&\qquad + \sum_{i=1}^n \left\{ q_1(\hat{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni}) \sum_{r=1}^q \frac{\partial^2 \hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}(U_i)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T} X_{ir} \right\}.
\end{aligned}$$

The following lemma shows how to construct a consistent estimator of $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$. The proof is in the Appendix.

Lemma 5 *Under regularity conditions (A)-(G), provided $\sqrt{p_n}\left(h + \frac{1}{\sqrt{nh}}\right) = o(1)$, we have for each $\boldsymbol{\beta}_n \in \Omega_n$,*

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u) \overset{def}{=} -&\left\{ \sum_{i=1}^n q_2(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n, Y_{ni})\mathbf{Z}_{ni}\mathbf{X}_i^T K_h(U_i - u) \right\}\\
&\cdot \left\{ \sum_{i=1}^n q_2(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(u)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n, Y_{ni})\mathbf{X}_i\mathbf{X}_i^T K_h(U_i - u) \right\}^{-1}
\end{aligned}$$

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

In implementing step 2 of the profile-kernel procedure, the first and second derivatives of $\hat{\boldsymbol{\alpha}}$ w.r.t. $\boldsymbol{\beta}_n$ are to be calculated at each U_i, which post a computational challenge to the profile-kernel procedure. On the other hand, ***the backfitting algorithm set all such derivatives to zero in equation (9)***, thus reducing vastly the amount of computations of each update. See Hu *et al* (2004) for a comparison of the two methods.

We propose a profile-kernel procedure which is 'in between' the full profile-kernel procedure and backfitting, with two major modifications to the full profile-kernel one:

Modifications of step 2 in the proposed profile-kernel procedure

(I) The second derivatives $\frac{\partial^2 \hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T}$ are set to $\mathbf{0}$ in equation (9).

(II) The first derivatives $\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u)$ are calculated on a sub-grid points of the U_i's and those on the rest of the U_i's are approximated by interpolation.

Since the function $q_2(\cdot,\cdot) < 0$ by regularity condition (D), we see that the modified $\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_n)$ in equation (9) is negative-definite. This ensures the Newton-Raphson update in step 2 of the profile-kernel procedure can be carried out without trouble.

The idea behind modification (I) is that, for a neighborhood around the true parameter $\boldsymbol{\beta}_{n0}$ which is small enough, the least favorable curve $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ should be approximately linear in $\boldsymbol{\beta}_n$. In fact, to estimate such second derivatives, same amount of local data around u is needed which has served to estimate the first derivative $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$ already, so variability of the resulting estimates of $\boldsymbol{\beta}_n$ may increase by incorporating the second derivatives into the updating procedure.

For modification (II), the idea is that $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$ is approximately linear in a small neighborhood of u. The bandwidth h in estimating $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ is a natural parameter to define what is a 'small' neighborhood around u. In this paper where a constant bandwidth h is used (see section 3.4), we calculate $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$ at the minimum and maximum values of U_i's from the data (assuming sparsity of the tail regions is avoided, see section 3.1), as well as calculating such on a grid of values of u with grid width approximately equals to h. Then $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)$ for data point U_i is found by interpolating the nearest two points on the grid. If variable type of bandwidth is used then the grid points can be defined also according to how h varies.

With these modifications, the update of $\boldsymbol{\beta}_n^{(k)}$ is much faster than the original profile-kernel procedure.

3.4 Choice of bandwidth

As usual the optimal bandwidth h_{opt} for estimating $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ given $\boldsymbol{\beta}_n$ is of order $n^{-1/(2p+3)}$, which can be seen immediately from equation (18). The equation also gives the order of the MSE to be $n^{-(2p+2)/(2p+3)}$ when such an optimal bandwidth is used. This optimal bandwidth order can be used without affecting the asymptotic properties of our estimator $\hat{\boldsymbol{\beta}}_n$, as shown in Theorems 1 and 2. We do not derive explicit expressions for the theoretical optimal bandwidth and MSE here.

As mentioned at the end of section 3.1, we have an extra smoothing parameter δ to be determined due to adjustments to transformation of the response Y_{ni}. This two dimensional smoothing parameter (δ, h) can be found by doing a K-fold cross-validation. Since we have suggested a quick profile-kernel procedure and practical accuracy can be achieved in several iterations as demonstrated in section 4, for K not too large (e.g. K=5 or 10) the cross-validation procedure is not too computationally intensive.

4 Simulation Study

In this section we first demonstrate how our proposed iterative procedure saves computational time as well as being stabler over the fully iterative procedure. Then using our iterative procedure, we demonstrate the finite sample performance of our estimates and augment our theoretical results.

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

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

where $\{u_k, k = 1, \cdots, n_{\text{grid}}\}$ are the grid points at which the function $\hat{\boldsymbol{\alpha}}(\cdot)$ is evaluated. The Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and $n_{\text{grid}} = 200$ are used in our simulation. For assessing the performance of the estimator $\hat{\boldsymbol{\beta}}_n$, we use the generalized mean square error (GMSE)

$$\text{GMSE} = (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T E\mathbf{Z}^*\mathbf{Z}^{*T}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}),$$

where $\mathbf{Z}^*$ is a new realization of the random variable $\mathbf{Z}$.

Simulation 1. In this simulation, we consider a semi-varying Poisson regression model. The response Y, given $(U, \mathbf{X}, \mathbf{Z_n})$, has a Poisson distribution with mean function $\mu(U, \mathbf{X}, \mathbf{Z}_n)$ where

$$\log(\mu(U, \mathbf{X}, \mathbf{Z}_n)) = \mathbf{X}^T\boldsymbol{\alpha}(U) + \mathbf{Z}_n^T\boldsymbol{\beta}_n.$$

We simulate 50 samples of sizes 200 and 400 with $p_n = \lfloor 1.8n^{1/3} \rfloor$ from the above model, meaning $p_{200} = 10$ and $p_{400} = 13$. For the covariates, we take $U \sim U(0,1), \mathbf{X} = (X_1, X_2)^T$ with $X_1 \equiv 1$ and $X_2 \sim N(0,1)$ such that $(\mathbf{Z}_n^T, X_2)^T$ is a $(p_n + 1)-$dimensional normal distribution with mean zero and covariance matrix (σ_{ij}), where $\sigma_{ij} = 0.5^{|i-j|}$. For the parameters of the model, $\boldsymbol{\beta}_{n0} = (0.5, 0.3, -0.5, 1, 0.1, -0.25, 0, \cdots, 0)^T$ which is p_n-dimensional, $\boldsymbol{\alpha}(u) = (\alpha_1(u), \alpha_2(u))^T$ where

$$\alpha_1(u) = 4 + \sin(2\pi u), \text{ and } \alpha_2(u) = 2u(1-u).$$

Using a 5-fold cross-validation (CV), we calculate 4-cycles estimates using our proposed profile-kernel procedure in order to obtain the CV value. We finally chose $\delta = 0.1$ and $h = 0.1, 0.08$ for $n = 200, 400$ respectively.

The median GMSE and respective computing times of $\hat{\boldsymbol{\beta}}_n$ among the 4-cycles estimators of backfitting, the proposed and full profile-kernel procedures are summarized in table 1. The SD_{mad} is a robust estimate of standard deviation and is defined by

Table 1: Simulation results of different fitting schemes for Poisson model

		Median(SD_{mad}) GMSE (multiplied by 10000)		
		backfitting	profile-kernel,	profile-kernel,
n	p_n		proposed	full
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(SD_{mad}) 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)
		Relative Median RASE (%)		
200	10	84.8	97.0	89.5
400	13	85.6	98.6	88.2

interquartile range divided by 1.349. We see that the proposed profile-kernel procedure has the smallest GMSE. The full profile-kernel procedure performs only slightly better than backfitting, but with much greater variability in the GMSE. In terms of computing times, backfitting wins against our proposed procedure slightly, but at the price of doubling the GMSE on average. Hence the proposed profile-kernel procedure gains the best trade-off between computational cost and accuracy. Comparing with the full profile-kernel procedure, it saves a vast amount of computations as well on average, and the savings grows as n increases. We also know (not shown in the table) that on average backfitting needs more than 20 iterations to converge without improving the GMSE too much. For a logistic data simulation (not shown here), our proposed procedure is still better than backfitting in terms of accuracy, but not as large an improvement as in the Poisson case.

The relative median RASE in table 1 is defined as $\text{RASE}_0/\text{RASE}_1$, where RASE_0 is the RASE calculated from the fit with true value of $\boldsymbol{\beta}_n$ known in advance (oracle estimate), and RASE_1 is the RASE calculated from different procedures. Clearly our proposed procedure is closest to the oracle estimate on average.

Simulation 2. In this simulation 400 samples of sizes 200, 400, 800 and 1500 with $p_n = \lfloor 1.8n^{1/3} \rfloor$ are drawn from the Poisson model introduced in simulation 1. Estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(u)$ are obtained by the proposed profile-kernel procedure, but with variants:

OS Our proposed profile-kernel procedure, iterated until convergence.

FS Same, except that we don't use the One-step procedure as in Cai, Fan and Li (2000) to estimate the nonparametric component, but by iterating Newton-Raphson algorithm until convergence.

DBE The difference-based estimation, same as one-cycle estimate.

4C The four-cycles estimate.

We compare median GMSE of the above procedures in table 2. The OS, 4C and FS procedures perform as good as each other, meaning that the one-step updating of nonparametric component works well and our proposed procedure converges early. In fact (not shown in the table) the two-cycles estimates improve the DBE dramatically already.

We summarized the effect of bandwidth choice and practical accuracy of estimated parameters (two-cycles) in table 3. We denote h_{CV} the choice of our bandwidth for the nonparametric component. It is clear that the GMSE does not sensitively depends on the bandwidth on average, as long as it is close to h_{CV}. The right column of the table shows the estimate for β_5. Being close to the true parameter value at different bandwidth choices with small variability (estimates of other β_i's are performing well similarly, and are not shown), the two-cycles estimate works well.

To test the accuracy of the sandwich covariance formula, the standard deviations of the estimated coefficients (two-cycles esimates) are computed among the 400 simulations at h_{CV}. These can be regarded as the true standard errors (columns labeled SD in table 4), and the 400 estimated standard errors are summarized by their median (columns SD_m) and the associated SD_{mad} (interquartile range divided by 1.349). Note that we have multiplied all values by 1000 for compact presentation. Clearly the sandwich formula does a good job, and accuracy gets better as n increases.

Finally we want to examine if the GLRT in section 2.2 performs well in testing a linear hypothesis on $\boldsymbol{\beta}_n$. To this end, we consider the following null hypothesis:

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

where we still have $p_n = \lfloor 1.8n^{1/3} \rfloor$. The alternative hypothesis is indexed by a parameter

Table 2: Simulation results for variants of profile-kernel procedures

		Relative Median GMSE (%)					
		Poisson			*Logistic*		
n	p_n	FS/OS	FS/DBE	FS/4C	FS/OS	FS/DBE	FS/4C
200	10	100.0	8.2	99.9	99.8	64.1	101.7
400	13	100.2	6.0	100.2	99.9	52.7	104.7
800	16	100.1	5.0	100.1	100.0	50.9	102.6
1500	20	100.0	4.2	100.0	100.0	46.4	100.5

Table 3: Summary statistics of two-cycles estimate

		Poisson				*Logistic*			
		Median(SD_{mad})		$\hat{\beta}_5$		*Median*(SD_{mad})		$\hat{\beta}_5$	
		GMSE$\times 10^5$		*mean(SD)*$\times 10^4$		*GMSE*$\times 10$		*mean(SD)*	
n	p_n	h_{CV}	$1.5h_{\text{CV}}$	$0.66h_{\text{CV}}$	h_{CV}	$0.66h_{\text{CV}}$	h_{CV}	h_{CV}	$1.5h_{\text{CV}}$
200	10	5.9(3.0)	6.4(3.3)	993(112)	995(105)	8.2(4.4)	8.4(5.1)	1.78(.40)	1.59(.37)
400	13	3.1(1.4)	3.0(1.4)	1004(67)	1001(65)	4.8(2.2)	5.4(2.5)	1.81(.26)	1.64(.27)
800	16	1.7(0.7)	1.7(0.6)	999(47)	999(46)	2.7(1.0)	2.7(1.1)	1.94(.20)	1.85(.19)
1500	20	1.1(0.3)	1.1(0.4)	1000(32)	1000(32)	1.8(0.7)	1.8(0.6)	1.97(.15)	1.91(.14)

δ as follows:

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

When $\delta = 0$, the alternative collapses to the null hypothesis. The GLRT statistic is computed for each simulation using the two-cycles estimates. Corresponding to $\delta = 0$, the kernel density estimate of the finite sample null distribution of these statistics is compared to the proposed asymptotic chi-squared density with d.f.$= p_n - 6$. Figure 1(a) shows the comparison when $n = 400$. The finite sample null density is seen to be close to the theoretical asymptotic chi-squared density.

To see the power of the test, we increases δ in the alternative H_1 and calculate the GLRT statistic in each simulation based on two-cycles estimates again. Three power functions are calculated corresponding to three different significance levels: 0.1, 0.05 and 0.01, using the theoretical chi-squared distribution to find the corresponding critical region. The proportion of rejection among the 400 statistics is the simulated power. We see from figure 1(b) that the upper two power curves are of slightly higher significance levels (corresponds to $\delta = 0$) than the theoretical significance levels 0.1 and 0.05. This suggests slightly thicker tail regions in the null density as seen also in figure 1(a). The power curves increase rapidly with δ, showing that the GLRT performs well.

Table 4: Standard deviations and estimated standard errors

		Poisson, values$\times$*1000*				*Logistic, values*$\times$*10*			
		$\hat{\beta}_1$		$\hat{\beta}_3$		$\hat{\beta}_2$		$\hat{\beta}_4$	
			SD$_m$		SD$_m$		SD$_m$		SD$_m$
n	p_n	SD	(SD$_{\text{mad}}$)	SD	(SD$_{\text{mad}}$)	SD	(SD$_{\text{mad}}$)	SD	(SD$_{\text{mad}}$)
200	10	9.1	8.5(1.3)	9.9	9.4(1.3)	3.6	2.9(.4)	3.2	2.8(.4)
400	13	6.0	5.6(0.7)	6.5	6.1(0.7)	2.3	2.1(.2)	2.2	2.0(.2)
800	16	3.7	3.8(0.3)	4.1	4.2(0.4)	1.7	1.6(.1)	1.5	1.5(.1)
1500	20	2.8	2.7(0.2)	3.1	3.0(0.2)	1.2	1.2(.1)	1.1	1.1(.1)

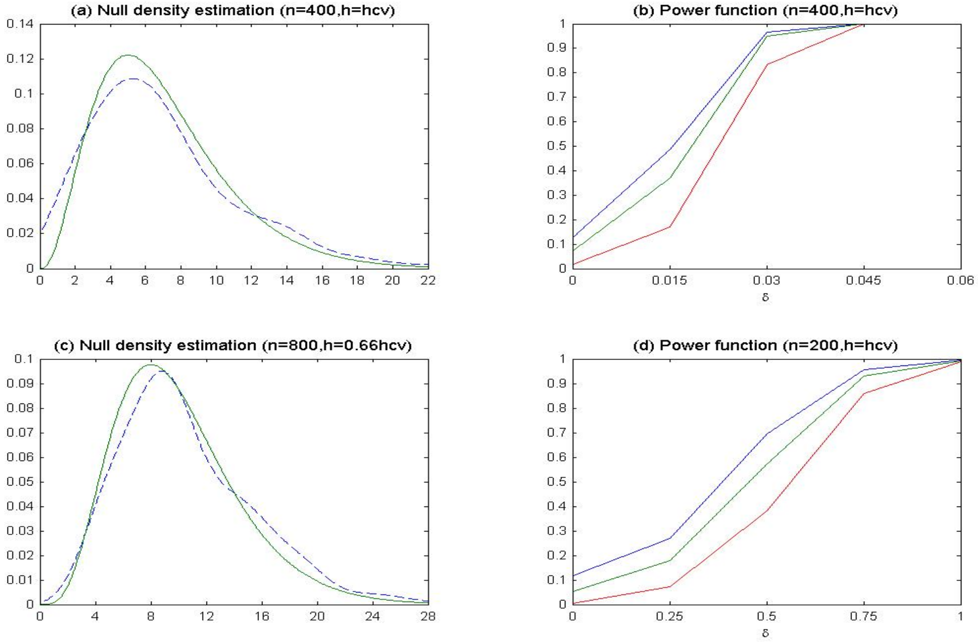

Figure 1: *Plots for simulation 2 and 3. (a) and (b) are plots for the Poisson GVCPLM while (c) and (d) are plots for the Logistic GVCPLM. In (a) and (c), dotted lines are the estimated null densities and the solid lines are χ^2-densities with d.f.$=p_n - 6$. (7 and* 10 *resp.) (b) and (d) are power functions of GLRT.*

Simulation 3. In this simulation, we consider a semi-varying Logistic regression model. The response Y, given $(U, \mathbf{X}, \mathbf{Z_n})$, has a Bernoulli distribution with success probability $p(U, \mathbf{X}, \mathbf{Z}_n)$ where

$$p(U, \mathbf{X}, \mathbf{Z}_n)) = \exp\{\mathbf{X}^T \boldsymbol{\alpha}(U) + \mathbf{Z}_n^T \boldsymbol{\beta}_n\}/[1 + \exp\{\mathbf{X}^T \boldsymbol{\alpha}(U) + \mathbf{Z}_n^T \boldsymbol{\beta}_n\}].$$

Same as simulation 1, we simulate 400 samples of sizes 200, 400, 800 and 1500 with $p_n = \lfloor 1.8 n^{1/3} \rfloor$ from the above model. For the covariates, we take $U \sim U(0,1), \mathbf{X} = (X_1, X_2)^T$ with $X_1 \equiv 1$ and $X_2 \sim N(0,1)$, and $\mathbf{Z}_n$ is a p_n-dimensional normal distribution with mean zero and covariance matrix (σ_{ij}), where $\sigma_{ij} = 0.5^{|i-j|}$. For the parameters of the model, $\boldsymbol{\beta}_{n0} = (3, 1, -2, 0.5, 2, -2, 0, \cdots, 0)^T$ which is p_n-dimensional, $\boldsymbol{\alpha}(u) = (\alpha_1(u), \alpha_2(u))^T$ where

$$\alpha_1(u) = 2(u^3 + 2u^2 - 2u), \text{ and } \alpha_2(u) = 2\cos(2\pi u).$$

Bandwidth (δ, h) is chosen by a 5-fold CV, where δ appears in the transformation of data $y \to \log\left(\frac{y+\delta}{1-y+\delta}\right)$. We finally chose $\delta = 0.005$ and $h = 0.45, 0.4, 0.25$ and 0.18, corresponding to $n = 200, 400, 800$ and 1500.

We compare median GMSE of the above procedures on the right of the table 2. The OS and FS procedures perform similar to each other, meaning that the one-step updating of nonparametric component works fine. The FS/DBE column shows that, unlike in the Poisson regression case, one update of the initial estimate $\boldsymbol{\beta}_n^{(0)}$ does not decrease the GMSE by a very large proportion.

Similar to the Poisson case, the right side of table 3 shows that sensitivity of estimates to bandwidth choice is not high. We also see a good accuracy of the sandwich covariance formula from table 4.

To examine the performance of the GLRT for the Logistic GVCPLM we use the same null and alternative hypotheses as defined in simulation 2. The estimated null density is close to the theoretical χ^2 density in figure 1(c) and the GLRT works well as seen from figure 1(d).

Real data example. We used Example 11.3 and the accompanying data set of Albright, Winston and Zappe (1999), where the Fifth National Bank of Springfield faced a gender discrimination suit in which female 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.) Fan and Peng (2004) has done such a salary analysis using an additive model with quadratic spline, and did not find a significant evidence of gender discrimination. We focus on another question: whether it was harder for female employees to be promoted.

The data set consists of 208 employees which include the following variables:

- EduLev: 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).
- JobGrade: a categorical variable indicating the current job level, the possible levels being 1–6 (6 highest).
- YrHired: year that an employee was hired.
- YrBorn: year that an employee was born.
- Gender: a categorical variable with values 'Female' and 'Male'.
- YrsPrior: number of years of working experience at another bank prior to working at the Fifth National Bank.
- PCJob: a dummy variable with value 1 if the employee's current job is computer related and value 0 otherwise.

Table 5: Fitted coefficients (sandwich SD) for model (10)

Response	Female	PCJob	Edu$_1$	Edu$_2$	Edu$_3$	Edu$_4$
HighGrade4	-1.66(.50)	-0.11(.71)	-4.32(.68)	-4.12(.80)	-2.33(.45)	-2.44(.89)
HighGrade5	-1.66(.50)	-1.25(.50)	-3.86(.52)	-3.92(.59)	-2.41(.59)	-0.95(.98)

We use **JobGrade** as the response variable and **Gender** as one of the covariates. The aim is to find if the **Gender** variable, after controlling for other factors such as educational level and years of prior experience, is significant in explaining **JobGrade**. We want to fit as large a model as possible to reduce modelling bias, and our theories allow us to interpret the model as usual. To simplify analysis, we create a response variable **HighGrade4** which is 0 if **JobGrade** is less than 4 and 1 otherwise. We can then fit a logistic regression or a logistic GVCPLM to the data and then carry out a GLRT to test the gender effect. From figure 2(a), the correlation between **Age** and **TotalYrsExp** (the total years of relevant working experience, calculated from **YrHired** and **YrsPrior**) is high, we use the following logistic GVCPLM

$$\begin{aligned}\log\left(\frac{p_H}{1-p_H}\right) &= \alpha_1(\text{Age}) + \alpha_2(\text{Age})\text{TotalYrsExp} \\ &\quad + \beta_1\text{Female} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i\end{aligned} \tag{10}$$

to reduce modelling bias, where p_H is the probability of having a job grade 4 or above. Interaction terms such as that between **Female** and **Edu**$_i$ are considered, but tested non-significant with GLRT so that we do not include those terms in the model above. (Including interaction terms increases the number of linear parameters, but theorem 3 still applies.) We use a 20-fold CV and find $h_{\text{CV}} = 23.5$, $\delta_{\text{CV}} = 0.1$.

Table 5 shows the results of the fit. (Two-cycles estimates using our proposed profile-kernel procedure.) It has a negative coefficient for **Female** and appears statistically significant since the estimated sandwich SD is small. Figure 2(b) shows the standardized residuals $(y - \hat{p}_{\text{H}})/\sqrt{\hat{p}_{\text{H}}(1-\hat{p}_{\text{H}})}$ against **Age** and the fit seems reasonable. (Other diagnostic plots are not shown.) From figure 2(c), we see that as age increases one has a better chance of being in a higher job grade. Figure 2(d) shows that the marginal effect of working experience is large when age is around 30 or less, but start to fall as one gets older.

We have done another fit using a binary variable **HighGrade5** which is similar to **HighGrade4** but is 0 only when job grade is less than 5. The coefficients are shown in table 5 and the **Female** coefficient is very close to the first fit.

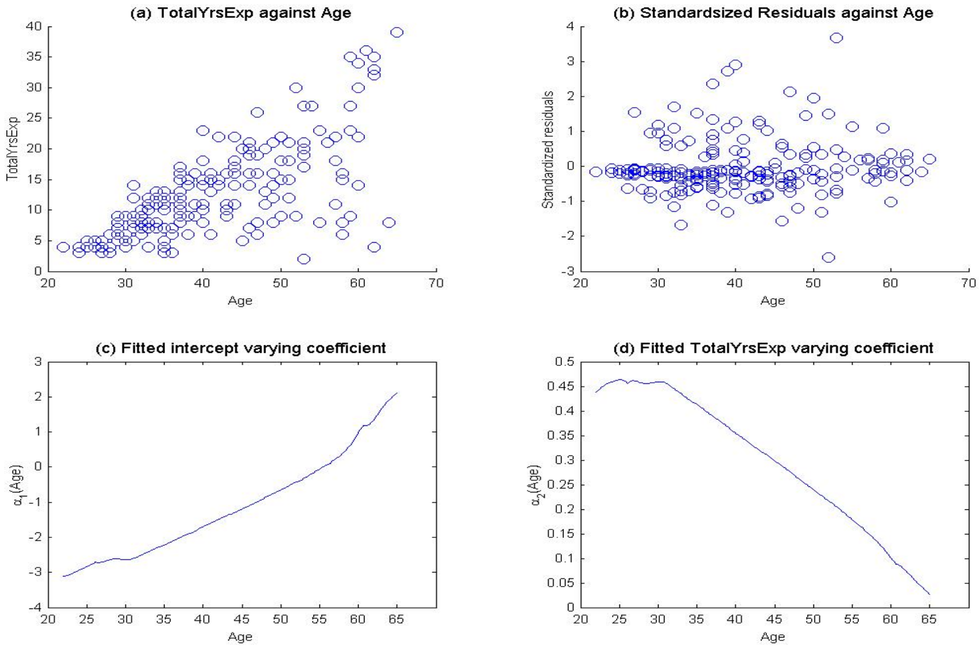

Figure 2: *(a) and (b):*TotalYrsExp*and standardized residuals against Age. (c) and (d): Varying coefficients for the logistic GVCPLM for the data.*

Formally, we are testing

$$H_0 : \beta_1 = 0 \longleftrightarrow H_1 : \beta_1 < 0.$$

Table 6 shows significant test results no matter we are using **HighGrade4** or **HighGrade5** as the response. Not shown in this paper, we have done the test again after deleting 6 data points corresponding to 5 male executives and 1 female having many years of working experience and high salaries. The test results are still similar. In fact from the raw data, female staffs are usually having a lower job grade than male with similar profile of educational level, working experience and age, even their salaries difference may not be apparent. The test results support that female staff of the Fifth National Bank of Springfield is harder to be promoted to a higher job grade than male.

Table 6: Generalized likelihood ratio test for $\beta_1 = 0$

Response	χ^2-statistic	P-value
HighGrade4	13.8095	0.0002
HighGrade5	11.3544	0.0008

5 Technical Proofs.

In this section proofs of Theorems 1-4 will be given. We introduce some notations and regularity conditions for our results to hold. In the following and thereafter, the symbol $\otimes$ represents the Kronecker product between matrices, and $\lambda_{\min}(A), \lambda_{\max}(A)$ denotes respectively the minimum and maximum eigenvalues of the matrix A.

Denote the true linear parameter by $\boldsymbol{\beta}_{n0}$, with parameter space $\Omega_n \subset \mathbb{R}^{p_n}$. Let $\rho_l(t) = \left(dg^{-1}(t)/dt\right)^l / V(g^{-1}(t)$, $\quad m_{ni}(\boldsymbol{\beta}_n) = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)^T \mathbf{X}_i + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}$, $\mu_k = \int u^k K(u)du$, $\quad A_p(\mathbf{X}) = (\mu_{i+j})_{0 \le i,j \le p} \otimes \mathbf{X}\mathbf{X}^T$, $\quad \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u) = \frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n}$, $\boldsymbol{\alpha}^{(r)\prime\prime}_{\boldsymbol{\beta}_n}(u) = \frac{\partial^2 \boldsymbol{\alpha}^{(r)}_{\boldsymbol{\beta}_n}(u)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T}$ and $q_l(x,y) = \frac{d^l}{dx^l} Q(g^{-1}(x), y)$ for $l = 1, \cdots, 4$.

Regularity Conditions:

(A) $|(\mathbf{Z}_n)_j|$, $\|\mathbf{X}\|$, are $O_P(1)$ and $\left\|\frac{\partial \boldsymbol{\alpha}_{\beta_n}(u)}{\partial \beta_{nj}}\right\|$, $\left\|\frac{\partial^2 \boldsymbol{\alpha}_{\beta_n}(u)}{\partial \beta_{nj} \partial \beta_{nk}}\right\|$ and $\left\|\frac{\partial^3 \boldsymbol{\alpha}_{\beta_n}(u)}{\partial \beta_{nj} \partial \beta_{nk} \partial \beta_{nl}}\right\|$ are finite, $j, k, l = 1, \cdots, p_n$.

(B) $I_n(\boldsymbol{\beta}_{n0}) = \mathbf{E}_0\left[\nabla Q_{n1}(\boldsymbol{\beta}_{n0}) \nabla^T Q_{n1}(\boldsymbol{\beta}_{n0})\right]$
$= \mathbf{E}_0 \left\{ q_1^2(m_{n1}(\boldsymbol{\beta}_{n0}), Y_{n1})(\mathbf{Z}_{n1} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_1)\mathbf{X}_1)(\mathbf{Z}_{n1} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_1)\mathbf{X}_1)^T \right\}$

satisfies the condition

$$0 < C_1 < \lambda_{\min}\{I_n(\boldsymbol{\beta}_{n0})\} \le \lambda_{\max}\{I_n(\boldsymbol{\beta}_{n0})\} < C_2 < \infty \text{ for all } n.$$

(C) $\mathbf{E}_{\boldsymbol{\beta}_n}\left|\frac{\partial^{l+j} Q_{ni}(\boldsymbol{\beta}_n)}{\partial^j \boldsymbol{\alpha} \partial \beta_{nk_1} \cdots \partial \beta_{nk_l}}\right| \le C_l < \infty$, $\mathbf{E}_{\boldsymbol{\beta}_n}\left|\frac{\partial^{l+j} Q_{ni}(\boldsymbol{\beta}_n)}{\partial^j \boldsymbol{\alpha} \partial \beta_{nk_1} \cdots \partial \beta_{nk_l}}\right|^2 \le \tilde{C}_l < \infty$ for some constants $C_l, \tilde{C}_l$ and for all n, with $l = 1, \cdots, 4$ 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 varying coefficient $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)$ 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 (D) ensures a unique solution in the local likelihood (4). Condition (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 imposed just for the simplicity of proofs. It can be relaxed at the expense of longer proofs.

Before proving Theorem 1, we need two important lemmas concerning order approximations to the varying coefficients. Let $c_n = (nh)^{-1/2}$, $\boldsymbol{\alpha}^{(p)}_{u\boldsymbol{\beta}_n}(u) = \frac{\partial^p \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial u^p}$. Define the following:

$$\begin{aligned}
\bar{\boldsymbol{\alpha}}_{ni}(u) &= \mathbf{X}_i^T \left(\sum_{k=0}^{p} \frac{(U_i - u)^k}{k!} \boldsymbol{\alpha}^{(k)}_{u\boldsymbol{\beta}_n}(u) \right) + \boldsymbol{\beta}_n^T \mathbf{Z}_{ni}, \\
\hat{\boldsymbol{\beta}}^* &= c_n^{-1} \left((\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n} - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u))^T, h(\hat{\mathbf{a}}_{1\boldsymbol{\beta}_n} - \boldsymbol{\alpha}'_{u\boldsymbol{\beta}_n}(u))^T, \cdots, \frac{h^p}{p!}(\hat{\mathbf{a}}_{p\boldsymbol{\beta}_n} - \boldsymbol{\alpha}^{(p)}_{u\boldsymbol{\beta}_n}(u))^T \right)^T, \\
\mathbf{X}_i^* &= \left(1, \frac{U_i - u}{h}, \cdots, \left(\frac{U_i - u}{h} \right)^p \right)^T \otimes \mathbf{X}_i.
\end{aligned}$$

Lemma 6 *Under regularity conditions (A) - (G), for each $\boldsymbol{\beta}_n \in \Omega_n$, the following holds uniformly in $u \in \Omega$:*

$$\left\| \hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u) \right\| = O_P(h^{p+1} + \frac{1}{\sqrt{nh}}).$$

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

Proof of lemma 6. Our first step is to show that, uniform in $u \in \Omega$,

$$\hat{\boldsymbol{\beta}}^* = \tilde{\mathbf{A}}_n^{-1} \mathbf{W}_n + O_P(h^{p+1} + c_n \log^{1/2}(1/h)),$$

where

$$\begin{aligned}
\tilde{\mathbf{A}}_n &= f_U(u) E_0 \left\{ \rho_2(\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U)^T \mathbf{X} + \mathbf{Z}_n^T \boldsymbol{\beta}_n) A_p(\mathbf{X}) | U = u \right\}, \\
\mathbf{W}_n &= h c_n \sum_{i=1}^{n} q_1(\bar{\boldsymbol{\alpha}}_{ni}, Y_{ni}) \mathbf{X}_i^* K_h(U_i - u), \\
\mathbf{A}_n &= h c_n^2 \sum_{i=1}^{n} q_2(\bar{\boldsymbol{\alpha}}_{ni}, Y_{ni}) \mathbf{X}_i^* \mathbf{X}_i^{*T} K_h(U_i - u).
\end{aligned}$$

Since expression (4) is maximized at $(\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}, \cdots, \hat{\mathbf{a}}_{p\boldsymbol{\beta}_n})^T$, $\hat{\boldsymbol{\beta}}^*$ maximizes

$$\begin{aligned} l_n(\boldsymbol{\beta}^*) &= h\sum_{i=1}^{n}\left\{Q(g^{-1}(c_n\mathbf{X}_i^{*T}\boldsymbol{\beta}^*+\bar{\boldsymbol{\alpha}}_{ni}),Y_{ni})-Q(g^{-1}(\bar{\boldsymbol{\alpha}}_{ni}),Y_{ni})\right\} \\ &= \mathbf{W}_n^T\boldsymbol{\beta}^*+\frac{1}{2}\boldsymbol{\beta}^{*T}\mathbf{A}_n\boldsymbol{\beta}^*+\frac{hc_n^3}{6}\sum_{i=1}^{n}q_3(\eta_i,y_{ni})(\mathbf{X}_i^{*T}\boldsymbol{\beta}^*)^3K_h(U_i-u), \end{aligned}$$

where η_i lies between $\bar{\boldsymbol{\alpha}}_{ni}$ and $\bar{\boldsymbol{\alpha}}_{ni}+c_n\mathbf{X}_i^{*T}\boldsymbol{\beta}^*$. The concavity of $l_n(\boldsymbol{\beta}^*)$ is ensured by condition (D). Note that $K(\cdot)$ is bounded, so under condition (C) the third term on the right hand side is bounded by

$$O_P(nhc_n^3E|q_3(\eta_1,Y_{n1})\|\mathbf{X}_1\|^3K_h(U_1-u)|=O_P(c_n)=o_P(1).$$

Direct calculation yields

$$\begin{aligned} E_0\mathbf{A}_n &= -\tilde{\mathbf{A}}_n+o(1), \\ \mathrm{Var}_0((\mathbf{A}_n)_{ij}) &= O((nh)^{-1}), \end{aligned}$$

so that mean-variance decomposition yields

$$\mathbf{A}_n=-\tilde{\mathbf{A}}_n+o_P(1).$$

Hence we have

$$l_n(\boldsymbol{\beta}^*)=\mathbf{W}_n^T\boldsymbol{\beta}^*-\frac{1}{2}\boldsymbol{\beta}^{*T}\tilde{\mathbf{A}}_n\boldsymbol{\beta}^*+o_P(1). \tag{11}$$

Note that $\mathbf{A}_n$ is a sum of i.i.d. random variables of kernel form, by lemma (A.2),

$$\mathbf{A}_n=-\tilde{\mathbf{A}}_n+o_P(1)+O_P\left\{h^{p+1}+c_n\log^{1/2}(1/h)\right\} \tag{12}$$

uniformly in $u\in\Omega$. Hence by the Convexity lemma (Pollard, 1991), equation (11) also holds uniformly in $\boldsymbol{\beta}^*\in C$ for any compact set C. Lemma A.1 then yields

$$\sup_{u\in\Omega}|\hat{\boldsymbol{\beta}}^*-\tilde{\mathbf{A}}_n^{-1}\mathbf{W}_n|\xrightarrow{\mathbb{P}}0. \tag{13}$$

Furthermore, by the definition of $\hat{\boldsymbol{\beta}}^*$,

$$\frac{\partial}{\partial\boldsymbol{\beta}^*}l_n(\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}^*=\hat{\boldsymbol{\beta}}^*}=hc_n\sum_{i=1}^{n}q_1(\bar{\boldsymbol{\alpha}}_{ni}+c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*,Y_{ni})\mathbf{X}_i^*K_h(U_i-u)=0. \tag{14}$$

Expanding $q_1(\bar{\boldsymbol{\alpha}}_{ni} + c_n \mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*, \cdot)$ at $\bar{\boldsymbol{\alpha}}_{ni}$,

$$\mathbf{W}_n + \mathbf{A}_n\hat{\boldsymbol{\beta}}^* + \frac{hc_n^3}{2}\sum_{i=1}^{n} q_3(\bar{\boldsymbol{\alpha}}_{ni} + \hat{\zeta}_i, Y_{ni})\mathbf{X}_i^*(\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*)^2 K_h(U_i - u) = 0 \tag{15}$$

where $\hat{\zeta}_i$ lies between 0 and $c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*$. Using condition (C), the last term has order $O_P(c_n^3 hn\|\hat{\boldsymbol{\beta}}^*\|^2) = O_P(c_n\|\hat{\boldsymbol{\beta}}^*\|^2)$. By (13), we know that $\|\hat{\boldsymbol{\beta}}^*\| \le o_P(1) + \|\tilde{\mathbf{A}}_n^{-1}\mathbf{W}_n\| \le o_P(1) + O(1)\cdot\|\mathbf{W}_n\|$. Note that by direct calculation,

(16)

$$\begin{aligned} E_0\mathbf{W}_n &= \frac{\sqrt{nh}h^{p+1}}{(p+1)!}\boldsymbol{\alpha}_{u\boldsymbol{\beta}_n}^{(p+1)}(u)^T \\ &\quad \times E_0\left\{\rho_2(\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U)^T\mathbf{X} + \mathbf{Z}_n^T\boldsymbol{\beta}_n)\mathbf{X}(\mu_{p+1},\cdots,\mu_{2p+1})^T \otimes \mathbf{X}|U=u\right\} + o(c_n^{-1}h^{p+1}), \\ \mathrm{Var}_0\mathbf{W}_n &= O(1), \end{aligned}$$

and hence $\|\mathbf{W}_n\| = O_P(1 + c_n^{-1}h^{p+1})$ which implies $O_P(c_n\|\hat{\boldsymbol{\beta}}^*\|^2) = o_P(1)$. With this, combining (12) and (15), we obtain

$$\mathbf{W}_n - \tilde{\mathbf{A}}_n\hat{\boldsymbol{\beta}}^*\left[1 + O_P\left\{h^{p+1} + c_n\mathrm{log}^{1/2}(1/h)\right\}\right] + o_P(1) = 0.$$

Hence,

$$\hat{\boldsymbol{\beta}}^* = \tilde{\mathbf{A}}_n^{-1}\mathbf{W}_n + O_P(h^{p+1} + c_n\log^{1/2}(1/h)) \tag{17}$$

holds uniformly for $u \in \Omega$ by (13). As a direct consequence, by using (16),

$$\left\|\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)\right\| = O_P(h^{p+1} + \frac{1}{\sqrt{nh}}) \tag{18}$$

which holds uniformly for $u \in \Omega$.

Differentiate both sides of (14) w.r.t. β_{nj},

$$hc_n\sum_{i=1}^{n} q_2(\bar{\boldsymbol{\alpha}}_{ni} + c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*, Y_{ni})\left(\frac{\partial\bar{\boldsymbol{\alpha}}_{ni}}{\partial\beta_{nj}} + c_n\left(\frac{\partial\hat{\boldsymbol{\beta}}^*}{\partial\beta_{nj}}\right)^T\mathbf{X}_i^*\right)\mathbf{X}_i^*K_h(U_i - u) = 0, \tag{19}$$

which holds for all $u \in \Omega$. By Taylor's expansion and similar treatments to (15),

$$\mathbf{W}_n^1 + \mathbf{W}_n^2 + (\mathbf{A}_n + \mathbf{B}_n^1 + \mathbf{B}_n^2)\frac{\partial\hat{\boldsymbol{\beta}}^*}{\partial\beta_{nj}} + O_P(c_n\|\hat{\boldsymbol{\beta}}^*\|^2),$$

where

$$\begin{aligned}\mathbf{W}_n^1 &= hc_n\sum_{i=1}^n q_2(\bar{\boldsymbol{\alpha}}_{ni},Y_{ni})\frac{\partial\bar{\boldsymbol{\alpha}}_{ni}}{\partial\beta_{nj}}\mathbf{X}_i^*K_h(U_i-u),\\ \mathbf{W}_n^2 &= hc_n\sum_{i=1}^n q_3(\bar{\boldsymbol{\alpha}}_{ni},Y_{ni})c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*\frac{\partial\bar{\boldsymbol{\alpha}}_{ni}}{\partial\beta_{nj}}\mathbf{X}_i^*K_h(U_i-u),\\ \mathbf{B}_n^1 &= hc_n^2\sum_{i=1}^n q_3(\bar{\boldsymbol{\alpha}}_{ni},Y_{ni})c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*\mathbf{X}_i^*\mathbf{X}_i^{*T}K_h(U_i-u),\\ \mathbf{B}_n^2 &= \frac{1}{2}hc_n^2\sum_{i=1}^n q_4(\bar{\boldsymbol{\alpha}}_{ni}+\hat{\zeta}_i,Y_{ni})(c_n^2\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*)^2\mathbf{X}_i^*\mathbf{X}_i^{*T}K_h(U_i-u),\end{aligned}$$

with $\hat{\zeta}_i$ lies between 0 and $c_n\mathbf{X}_i^{*T}\hat{\boldsymbol{\beta}}^*$. The equation holds for all $u\in\Omega$. Note that $O_P(c_n\|\hat{\boldsymbol{\beta}}^*\|^2)=o_P(1)$ uniformly for $u\in\Omega$ by (13). The order of $\mathbf{W}_n^2$ is smaller than that of $\mathbf{W}_n^1$, and the order of $\mathbf{B}_n^1$ and $\mathbf{B}_n^2$ are smaller than that of $\mathbf{A}_n$. Hence

$$\frac{\partial\hat{\boldsymbol{\beta}}^*}{\partial\beta_{nj}}=\tilde{\mathbf{A}}_n^{-1}\mathbf{W}_n^1+o_P(1+c_n^{-1}h^{p+1})$$

uniformly in $u\in\Omega$, by noting that

$$\begin{aligned}E_0\mathbf{W}_n^1 &= \frac{\partial}{\partial\beta_{nj}}E_0\mathbf{W}_n+o(c_n^{-1}h^{p+1}),\\ \mathrm{Var}_0\mathbf{W}_n^1 &= O(1).\end{aligned}$$

From this, for $j=1,\cdots,p_n$, we have

$$\left\|\frac{\partial\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial\beta_{nj}}-\frac{\partial\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial\beta_{nj}}\right\|=O_P(h^{p+1}+\frac{1}{\sqrt{nh}}). \tag{20}$$

uniformly in $u\in\Omega$. Differentiating (14) again w.r.t. β_{nk} and so on, and follow similar arguments as above, we get results for higher order derivatives. □

Lemma 7 *Under regularity conditions (A) - (G), the following holds uniformly in $u\in\Omega$:*

$$\left\|\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)\right\|=O_P(1).$$

Likewise, the norm of the k^{th} derivative of the above with respect to any β_{nj}'s, $k=1,\cdots,4$, all have order $O(1)$ uniformly in $u\in\Omega$.

Proof of lemma 7. It follows immediately from lemma 6 and condition (A). □

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

$$\mathbb{P}\left\{\sup_{\|\mathbf{v}\|=C}\hat{Q}_n(\boldsymbol{\beta}_{n0}+\gamma_n\mathbf{v})<\hat{Q}_n(\boldsymbol{\beta}_{n0})\right\}\geq 1-\epsilon, \tag{21}$$

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 $\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_{n0}\|=O_P(\gamma_n)$.

By Taylor's expansion,

$$\begin{aligned}
D_n(\mathbf{v}) &:= \hat{Q}_n(\boldsymbol{\beta}_{n0}+\gamma_n\mathbf{v})-\hat{Q}_n(\boldsymbol{\beta}_{n0})\\
&= \nabla^T\hat{Q}_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n+\frac{1}{2}\mathbf{v}^T\nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2+\frac{1}{6}\nabla^T(\mathbf{v}^T\nabla^2\hat{Q}_n(\boldsymbol{\beta}_n^*)\mathbf{v})\mathbf{v}\gamma_n^3\\
&:= \hat{I}_1+\hat{I}_2+\hat{I}_3,
\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_{n0}$ and $\boldsymbol{\beta}_{n0}+\gamma_n\mathbf{v}$, and $\|\mathbf{v}\|=C$ with C a large constant.

Consider

$$\begin{aligned}
\hat{I}_1 &= \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}),Y_{ni})(\mathbf{Z}_{ni}+\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}}(U_i)\mathbf{X}_i)^T\mathbf{v}\gamma_n\\
&= \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}),Y_{ni})(\mathbf{Z}_{ni}+\boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_i)\mathbf{X}_i)^T\mathbf{v}\gamma_n\\
&\quad +\sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}),Y_{ni})\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}}(U_i)-\boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_i))^T\mathbf{v}\gamma_n,\\
&:= D_1+D_2
\end{aligned}$$

where $\hat{m}_{ni}(\boldsymbol{\beta}_n)=\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i+\boldsymbol{\beta}_n^T\mathbf{Z}_{ni}$. D_2 has order smaller than D_1 by condition (A) and lemma 6. Using Taylor's expansion,

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

where K_1 is as defined in lemma 8 so that within the lemma's proof we have $\|\mathbf{K}_1\|=o_P(1)$. Using equation (6), we have by mean-variance decomposition

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

where last inequality follows from Cauchy-Schwarz and condition (B).

Hence

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

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

$$\begin{aligned}
I_2 &= \frac{1}{2}\mathbf{v}^T \nabla^2 Q_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 \\
&= -\frac{n}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 + \frac{n}{2}\mathbf{v}^T\left\{n^{-1}\nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})\right\}\mathbf{v}\gamma_n^2 \\
&\overset{\text{lemma 16}}{=} -\frac{n}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2 + o_P(1)n\gamma_n^2\|\mathbf{v}\|^2.
\end{aligned}$$

We want to show that $\hat{I}_2 - I_2$ has order smaller than $\frac{n}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\gamma_n^2$.

By Taylor's expansion,

$$\begin{aligned}
\hat{I}_2 - I_2 &= \frac{1}{2}\mathbf{v}^T\left\{\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0}) - \nabla^2 Q_n(\boldsymbol{\beta}_{n0})\right\}\mathbf{v}\gamma_n^2 \\
&= \frac{1}{2}\mathbf{v}^T\nabla^2\left\{\sum_{i=1}^n q_1(\tilde{m}_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i))\right\}\mathbf{v}\gamma_n^2 \\
&= \frac{1}{2}\mathbf{v}^T B_n\mathbf{v}\gamma_n^2 + \text{smaller order terms}
\end{aligned}$$

where $\tilde{m}_{ni}(\boldsymbol{\beta}_n) = \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}}(U_i)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$ with $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$ lies between $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$ and $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$. Denote $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i) = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}$ and so on. We have used condition (C) together with lemma 6 and 7 to arrive at the last equality, where

$$\begin{aligned}
B_n = \sum_{i=1}^n \{ & q_3(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}})^T\mathbf{X}_i \\
&+ q_2(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\sum_{r=1}^q X_{ir}\boldsymbol{\alpha}^{(r)\prime\prime}_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}) \\
&+ q_2(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}})^T \\
&+ q_2(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}})\mathbf{X}_i(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)^T \\
&+ q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\sum_{r=1}^q X_{ir}(\hat{\boldsymbol{\alpha}}^{(r)\prime\prime}_{\boldsymbol{\beta}_{n0}} - \boldsymbol{\alpha}^{(r)\prime\prime}_{\boldsymbol{\beta}_{n0}})\},
\end{aligned}$$

with $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n} = \frac{\partial\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}}{\partial\boldsymbol{\beta}_n}$ and $\boldsymbol{\alpha}^{(r)\prime\prime}_{\boldsymbol{\beta}_n} = \frac{\partial^2\boldsymbol{\alpha}^{(r)}_{\boldsymbol{\beta}_n}}{\partial\boldsymbol{\beta}_n\partial\boldsymbol{\beta}_n^T}$, $r = 1, \cdots, q$. Using Cauchy-Schwarz inequality, conditions (A), (B), lemma 6 and 7,

$$\begin{aligned}
|\mathbf{v}^T B_n\mathbf{v}\gamma_n^2| &\le O_P(p_n(h^{p+1} + \frac{1}{\sqrt{nh}}))\cdot O_P(n\gamma_n^2\|\mathbf{v}\|^2) \\
&= o_P(n\gamma_n^2\|\mathbf{v}\|^2).
\end{aligned}$$

By condition (B), we have

$$\begin{aligned}
\left|\frac{n\gamma_n^2}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}\right| &\ge O(n\gamma_n^2\lambda_{\min}(I_n(\boldsymbol{\beta}_{n0}))\|\mathbf{v}\|^2) \\
&= O(n\gamma_n^2\|\mathbf{v}\|^2).
\end{aligned}$$

Finally consider $\hat{I}_3$. Note that

$$\hat{Q}_n(\boldsymbol{\beta}_n^*) \le Q_n(\boldsymbol{\beta}_{n0}) + \{\sum_{i=1}^n q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)) \\ + \sum_{i=1}^n q_1(\hat{m}_{ni}(\boldsymbol{\beta}_{n0}), Y_n i)(\mathbf{Z}_{ni} + \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_{n0}}\mathbf{X}_i)\gamma_n \mathbf{v}\}(1 + o_P(1)),$$

and by condition (C), lemma 6 and 7 again, 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,

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

Comparing, we find the order of $-\frac{n\gamma_n^2}{2}\mathbf{v}^T I_n(\boldsymbol{\beta}_{n0})\mathbf{v}$, which is negative, dominates all other terms by allowing $\|\mathbf{v}\| = C$ to be large enough. This proves (21). □

Before proving Theorem 2, we need another lemma.

Lemma 8 *Under regularity conditions (A) - (G), if $p_n^3/n \to 0$ with $nh^{p+2} \to \infty$ and $nh^{2p+3} = O(1)$, then for each $\boldsymbol{\beta}_n \in \Omega_n$,*

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

Proof of lemma 8. Define

$$\mathbf{K}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n q_2(m_{ni}(\boldsymbol{\beta}_n), Y_{ni})(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i)(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i))^T\mathbf{X}_i,$$
$$\mathbf{K}_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n q_1(m_{ni}(\boldsymbol{\beta}_n), Y_{ni})(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i))\mathbf{X}_i,$$

then by Taylor's expansion, lemma 6 and condition (C),

$$\frac{1}{\sqrt{n}}(\nabla\hat{Q}_n(\boldsymbol{\beta}_n) - \nabla Q_n(\boldsymbol{\beta}_n)) = \mathbf{K}_1 + \mathbf{K}_2 + \text{smaller order terms},$$

where $m_{ni}(\boldsymbol{\beta}_n) = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$. Define, for Ω as in condition (F),

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

equipped with a metric

$$\rho(f_1, f_2) = \|f_1 - f_2\|_\infty,$$

with $\|f\|_\infty = \sup_{u\in\Omega} |f(u)|$. We also let, for $r = 1, \cdots, q$ and $l = 1, \cdots, p_n$,

$$\begin{aligned} A_{rl}(y, u, \mathbf{X}, \mathbf{Z}_n) &= q_2(\mathbf{X}^T \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u) + \mathbf{Z}_n^T \boldsymbol{\beta}_n, y) X_r \left(Z_{nl} + \mathbf{X}^T \frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)}{\partial \beta_{nl}} \right), \\ B_r(y, u, \mathbf{X}, \mathbf{Z}_n) &= q_1(\mathbf{X}^T \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u) + \mathbf{Z}_n^T \boldsymbol{\beta}_n, y) X_r. \end{aligned}$$

By lemma 6, for any $\delta > 0$ and as $n \to \infty$, we have

$$P_0\Bigg(\underbrace{n^{-\delta} \left(h^{p+1} + \frac{1}{\sqrt{nh}} \right)^{-1} (\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n} - \alpha^{(r)}_{\boldsymbol{\beta}_n})}_{:=\lambda_r} \in S \Bigg) \to 1,$$

$$P_0\Bigg(\underbrace{n^{-\delta} \left(h^{p+1} + \frac{1}{\sqrt{nh}} \right)^{-1} \left(\frac{\partial \hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}}{\partial \beta_{nl}} - \frac{\alpha^{(r)}_{\boldsymbol{\beta}_n}}{\partial \beta_{nl}} \right)}_{:=\gamma_{rl}} \in S \Bigg) \to 1,$$

where $r = 1, \cdots, q$ and $l = 1, \cdots, p_n$. Hence for sufficiently large n, we have $\lambda_r, \gamma_{rl} \in S$. The following three points allow us to utilize Jain and Marcus (1975) 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

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

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

II. By equation (7), we can easily see that

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

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

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

by regularity conditions (A) and (C). Similar results hold for $B_r(Y, U, \mathbf{X}, \mathbf{Z}_n)$.

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

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

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

Conditions of Theorem 1 in Jain and Marcus(1975) can be derived from the three notes above, so that we have

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

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

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

which implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})(\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n} - \alpha^{(r)}_{\boldsymbol{\beta}_n}) = O_P\left(n^{\delta}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right).$$

Similarly, apply Theorem 1 of Jain and Marcus(1975) again, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} B_r(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})\left(\frac{\partial\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n}}{\partial\beta_{nl}} - \frac{\alpha^{(r)}_{\boldsymbol{\beta}_n}}{\partial\beta_{nl}}\right) = O_P\left(n^{\delta}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right).$$

Then the column vector $\mathbf{K}_1$ which is p_n-dimensional, has the l^{th} component equals

$$\sum_{r=1}^{q}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_{rl}(Y_i, U_i, \mathbf{X}_i, \mathbf{Z}_{ni})(\hat{\alpha}^{(r)}_{\boldsymbol{\beta}_n} - \alpha^{(r)}_{\boldsymbol{\beta}_n})\right\} = O_P\left(n^{\delta}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right),$$

using the result just proved. Hence we have shown

$$\|\mathbf{K}_1\| = O_P\left(\sqrt{p_n}n^{\delta}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right) = o_P(1),$$

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

Proof of Theorem 2. We first assume $nh^{2p+2} = O(1)$ and $nh^{p+2} \to \infty$ as in Theorem 1, so that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P(\sqrt{p_n/n})$. Since $\nabla\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) = 0$, by Taylor's expansion,

$$\nabla\hat{Q}_n(\boldsymbol{\beta}_{n0}) + \nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T\nabla^2(\nabla\hat{Q}_n(\boldsymbol{\beta}^*_n))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) = 0,$$

where $\boldsymbol{\beta}^*_n$ lies between $\boldsymbol{\beta}_{n0}$ and $\hat{\boldsymbol{\beta}}_n$. This implies

$$\begin{aligned}\frac{1}{n}\nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) = &-\frac{1}{n}(\nabla\hat{Q}_n(\boldsymbol{\beta}_{n0}) \\ &+ \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T\nabla^2(\nabla\hat{Q}_n(\boldsymbol{\beta}^*_n))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})).\end{aligned} \tag{22}$$

Define $\mathcal{C} = \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T \nabla^2(\nabla \hat{Q}_n(\boldsymbol{\beta}_n^*))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}))$. Using similar argument to approximating $\hat{I}_3$ in Theorem 1, using lemma 6 and lemma 7, and noting $\|\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_{n0}\| = o_P(1)$, we have $\left\|\nabla^2 \frac{\partial^2 \hat{Q}_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}}\right\|^2 = O_P(n^2 p_n^2)$. Hence

$$
\begin{aligned}
\|n^{-1}\mathcal{C}\|^2 &\leq \frac{1}{2n^2}\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^4 \left\|\nabla^2 \frac{\partial^2 \hat{Q}_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}}\right\|^2 \\
&\leq \frac{1}{2n^2} O_P\left(\frac{p_n^2}{n^2}\right) \sum_{j=1}^{p_n} O_P(n^2 p_n^2) \\
&= O_P\left(\frac{p_n^5}{n^2}\right) = o_P\left(\frac{1}{n}\right).
\end{aligned} \tag{23}
$$

At the same time, by lemma 16 and Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\left\|\frac{1}{n}\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})\right\| \\
&\leq o_P\left(\frac{1}{\sqrt{np_n}}\right) + O_P\left(\sqrt{\frac{p_n^3}{n}}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right) \\
&\leq o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\frac{1}{\sqrt{n}} \cdot \left(\sqrt{\frac{p_n^3}{n}} + \sqrt{\frac{p_n^3}{n^{(p+1)/(p+2)}}}\right)\right) \\
&= o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{24}
$$

where the second last line used $nh^{2p+2} = O(1)$ and $nh^{p+2} \to \infty$, and the last line used assumption $p_n^5/n \to 0$.

Combining (22),(23) and (24), we have

$$
\begin{aligned}
I_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) &= \frac{1}{n}\nabla \hat{Q}_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{1}{n}\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{25}
$$

where the last line follows from lemma 8. Consequently, using equation (25), we get

$$
\begin{aligned}
&\sqrt{n} A_n I_n^{1/2}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \\
&\quad = \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0})) \\
&\quad = \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P(1),
\end{aligned} \tag{26}
$$

where the last equality holds since by condition of Theorem 2, $\|A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0})\|$ is of order $O(1)$.

Let $B_{ni} = \frac{1}{\sqrt{n}} A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_{ni}(\boldsymbol{\beta}_{n0})$, where $Q_{ni}(\boldsymbol{\beta}_n) = Q(g^{-1}(m_{ni}(\boldsymbol{\beta}_n)), Y_{ni})$, $i = 1, \cdots, n$. Given $\epsilon > 0$,

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

Using Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}(\|B_{n1}\| > \epsilon) &\le \frac{E_0 \|B_{n1}\|^2}{\epsilon^2} \\ &= \frac{1}{n\epsilon^2} E \|A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})\|^2 \\ &= \frac{1}{n\epsilon^2} tr\{I_n^{-1/2}(\boldsymbol{\beta}_{n0}) A_n^T A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) E_0(\nabla Q_{n1}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})^T)\} \\ &= \frac{1}{n\epsilon^2} tr\{I_n^{-1/2}(\boldsymbol{\beta}_{n0}) A_n^T A_n I_n^{1/2}(\boldsymbol{\beta}_{n0})\} \\ &= \frac{1}{n\epsilon^2} tr(G) = O\left(\frac{1}{n}\right), \end{aligned} \tag{27}$$

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

$$\begin{aligned} E_0 \|B_{n1}\|^4 &\le \frac{\sqrt{l}}{n^2} \lambda_{\max}^2(A_n A_n^T) \lambda_{\max}^2(I_n^{-1}(\boldsymbol{\beta}_{n0})) \sqrt{E_0 \nabla Q_{n1}(\boldsymbol{\beta}_{n0})^T \nabla Q_{n1}(\boldsymbol{\beta}_{n0})} \\ &= O\left(\frac{p_n^2}{n^2}\right). \end{aligned} \tag{28}$$

Therefore (27) and (28) together implies

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

Also,

$$\begin{aligned} \sum_{i=1}^{n} \mathrm{Var}_0(B_{ni}) &= n \mathrm{Var}_0(B_{n1}) = \mathrm{Var}_0(A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})) \\ &= A_n I_n^{-1/2}(\boldsymbol{\beta}_{n0}) E_0 \nabla Q_{n1}(\boldsymbol{\beta}_{n0}) \nabla Q_{n1}(\boldsymbol{\beta}_{n0})^T I_n^{-1/2}(\boldsymbol{\beta}_{n0}) A_n^T \\ &= A_n A_n^T \to G. \end{aligned}$$

Therefore B_{ni} satisfies the conditions of the Lindeberg-Feller central limit Theorem (see for example, Van der Vaart(1998)). Consequently, asymptotic normality of $\sum_{i=1}^{n} B_{ni}$ follows. Using (26), it means

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

For the optimal bandwidth $h = O(n^{-1/(2p+3)})$, we can follow same lines of proof in Theorem 1 to arrive at $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\right\| = O_P(\sqrt{p_n/n^{(2p+2)/(2p+3)}})$. Note that the proof of Theorem 2 is affected only in (23) and (24). With the condition $p_n^5/n^{(2p+1)/(2p+3)} \to 0$, (23) becomes

$$\begin{aligned}\|n^{-1}\mathcal{C}\|^2 &\leq \frac{1}{2n^2} O_P\left(\frac{p_n^2}{n^2} \cdot n^{2/(2p+3)}\right) \sum_{j=1}^{p_n} O_P(n^2 p_n^2) \\ &= O_P\left(\frac{p_n^5}{n^2} \cdot n^{2/(2p+3)}\right) = O_P(p_n^5/n^{(2p+1)/(2p+3)} \cdot \frac{1}{n}) \\ &= o_P\left(\frac{1}{n}\right).\end{aligned}$$

For (24), since $p_n^5/n \to 0$, $p_n^4/n^{(2p+2)/(2p+3)} \to 0$ is automatically satisfied and so by lemma 16,

$$\begin{aligned}&\left\|\frac{1}{n}\nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})\right\| \\ &= o_P\left(\frac{n^{1/(4p+6)}}{p_n n^{1/(4p+6)}} \cdot \sqrt{\frac{p_n}{n}}\right) + O_P\left(2p_n n^{-(p+1)/(2p+3)} \cdot \sqrt{\frac{p_n}{n}} \cdot n^{1/(4p+6)}\right) \\ &= o_P\left(\frac{1}{\sqrt{np_n}}\right) + O_P\left(\frac{1}{\sqrt{n}} \cdot \sqrt{\frac{p_n^3}{n^{(2p+1)/(2p+3)}}}\right) \\ &= o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Hence conclusion of Theorem 2 still follows. □

Refer back to section 2.2, let B_n be a $(p_n - l) \times p_n$ matrix satisfying $B_n B_n^T = I_{p_n - l}$ and $A_n B_n^T = 0$. 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 by $A_n B_n^T = 0$. Hence

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

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

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

To prove Theorem 3 we need the following lemmas, the proofs of which are given in the appendix.

Lemma 9 *Assuming regularity conditions (A) - (G). Under the null hypothesis H_0 as in Theorem 3, if $nh^{2p+2} = O(1)$, then under $p_n^5/n = o(1)$,*

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

Moreover, if $h = O(n^{-1/(2p+3)})$, then under $p_n^5/n^{(2p+1)/(2p+3)} = o(1)$, the same conclusion still holds.

Lemma 10 *Under regularity conditions (A) - (G) and $p_n^5/n = o(1)$, we have*

$$\frac{1}{n}\|\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0})\| = o_P\left(\frac{1}{\sqrt{p_n}}\right)$$

if $nh^{2p+2} = O(1)$. Moreover if $h = O(n^{-1/(2p+3)})$, then assuming further $p_n^5/n^{(2p+2)/(2p+3)} = o(1)$, the same conclusion still holds.

Lemma 11 *Assuming the conditions of Theorem 3, under the null hypothesis H_0, we have*

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

Proof of Theorem 3. Adapting the notation in lemma 11, substituting equation (30) into its conclusion we get

$$\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol{\gamma}}_n) = \frac{n}{2}\boldsymbol{\Phi}_n^T \Theta_n^{-1/2} S_n \Theta_n^{-1/2} \boldsymbol{\Phi}_n + o_P(1),$$

where $\Theta_n = I_n(\boldsymbol{\beta}_{n0})$, $\boldsymbol{\Phi}_n = \frac{1}{n}\nabla Q_n(\boldsymbol{\beta}_{n0})$ and $S_n = I_n - \Theta_n^{1/2} B_n^T (B_n \Theta_n B_n^T)^{-1} B_n \Theta_n^{1/2}$. Since S_n is idempotent, it can be written as $S_n = D_n^T D_n$ where D_n is a $l \times p_n$ matrix satisfying $D_n D_n^T = I_l$.

By the proof of Theorem 2, substituting A_n there with D_n, using equation (26), 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

$$2\{\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \hat{Q}_n(\boldsymbol{\beta}_{n0})\} = n(D_n \Theta_n^{-1/2} \boldsymbol{\Phi}_n)^T (D_n \Theta_n^{-1/2} \boldsymbol{\Phi}_n) \xrightarrow{\mathcal{D}} \chi_l^2. \ \square$$

To prove Theorem 4, we need two lemmas. The proofs are given in the appendix.

Lemma 12 *Assuming the conditions of Theorem 4, we have*

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

Lemma 13 *Assuming the conditions of Theorem 4, we have for each* $\boldsymbol{\beta}_n \in \Omega_n$*,*

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

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

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

then we can rewrite

$$\hat{\Sigma}_n - \Sigma_n = I_1 + I_2 + I_3.$$

Our aim is to show that, for all $i = 1, \cdots, p_n$,

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

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

$$\begin{aligned}\lambda_{\min}(I_1) + \lambda_{\min}(I_2) + \lambda_{\min}(I_3) &\le \lambda_{\min}(I_1 + I_2 + I_3)\\ &\le \lambda_{\max}(I_1 + I_2 + I_3) \le \lambda_{\max}(I_1) + \lambda_{\max}(I_2) + \lambda_{\max}(I_3),\end{aligned}$$

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

$$\begin{aligned}K_1 &= I_n(\boldsymbol{\beta}_{n0}) + n^{-1}\nabla^2 Q_n(\boldsymbol{\beta}_{n0}),\\ K_2 &= n^{-1}(\nabla^2 Q_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2 Q_n(\boldsymbol{\beta}_{n0})),\\ K_3 &= n^{-1}(\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2 Q_n(\hat{\boldsymbol{\beta}}_n)),\end{aligned}$$

then

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

Applying lemma 16 on K_1, lemma 12 on K_2 and lemma 13 on K_3, we have $\|\mathcal{C} - \hat{\mathcal{A}}\| = o_P(1)$, and so $\lambda_i(\mathcal{C} - \hat{\mathcal{A}}) = o_P(1)$. 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

$$\lambda_i(\hat{\mathcal{B}}_n - \mathcal{C}) = K_4 + K_5$$

where

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

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

$$K_4 = K_6 + K_7,$$

where

$$\begin{aligned} K_6 &= \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\}, \\ K_7 &= \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \right\} - I_n(\boldsymbol{\beta}_{n0}). \end{aligned}$$

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 $k = 1, \cdots, p_n$ and $i = 1, \cdots, n$ is rewarding. Define

$$\begin{aligned} a_{ik} &= \frac{\partial}{\partial \beta_{nk}}(\hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\hat{\boldsymbol{\beta}}_n)), \\ b_{ik} &= \frac{\partial}{\partial \beta_{nk}}(Q_{ni}(\hat{\boldsymbol{\beta}}_n) - Q_{ni}(\boldsymbol{\beta}_{n0})), \end{aligned}$$

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,

$$\begin{aligned} a_{ik} &= \frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{\partial Q_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} \\ &= \frac{\partial}{\partial \beta_{nk}}(q_1(\tilde{m}_{ni}(\hat{\boldsymbol{\beta}}_n), Y_{ni})(\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i) - \boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}(U_i))^T \mathbf{X}_i) \\ &= q_2(\tilde{m}_{ni}(\hat{\boldsymbol{\beta}}_n), Y_{ni}) \left(Z_{nik} + \frac{\partial \tilde{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i)}{\partial \beta_{nk}}^T \mathbf{X}_i \right) (\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i) - \boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}(U_i))^T \mathbf{X}_i \\ &\quad + q_1(\tilde{m}_{ni}(\hat{\boldsymbol{\beta}}_n), Y_{ni}) \left(\frac{\partial \hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i)}{\partial \beta_{nk}} - \frac{\partial \boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}(U_i)}{\partial \beta_{nk}} \right)^T \mathbf{X}_i, \end{aligned}$$

where $\tilde{m}_{ni}(\hat{\boldsymbol{\beta}}_n) = \tilde{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i)$ lies between $\boldsymbol{\alpha}_{\hat{\boldsymbol{\beta}}_n}(U_i)$ and $\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}_n}(U_i)$. Using lemma 6, 7 and conditions (A) and (C), with argument similar to the proof of lemma 13, we then have

$$|a_{ik}| \le O_P\bigg(h^{p+1} + \frac{1}{\sqrt{nh}}\bigg).$$

Similarly, using Taylor's expansion and lemma 6, 7, regularity conditions (A) and (C),

$$\begin{aligned} b_{ik} &= \frac{\partial Q_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}} \\ &= \bigg\{ q_2(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})(\mathbf{Z}_{ni} + \boldsymbol{\alpha}'_{\boldsymbol{\beta}_{n0}}(U_i)\mathbf{X}_i)^T(Z_{nik} + \mathbf{X}_i^T\frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)}{\partial \beta_{nk}}) \\ &\quad + q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\left(\mathbf{X}_i^T\frac{\partial^2 \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)}{\partial \beta_{nk}\partial \boldsymbol{\beta}_n^T}\right)\bigg\}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + \text{smaller order terms}, \end{aligned}$$

which implies that, by Cauchy-Schwarz inequality, together with Theorem 1 and regularity conditions (A) and (C) again,

$$|b_{ik}| \le O_P\left(\frac{p_n}{\sqrt{n^{d_h}}}\right), \text{ where } d_h = \begin{cases} 1, \text{ if } nh^{2p+2} = O(1). \\ \frac{2p+2}{2p+3}, \text{ if } nh^{2p+3} = O(1). \end{cases}$$

Hence using the approximations of a_{ik} and b_{ik} above,

$$\begin{aligned} &\bigg|\frac{1}{n}\sum_{i=1}^n \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}}\bigg\{\frac{\partial \hat{Q}_{ni}(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nk}} - \frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial \beta_{nk}}\bigg\}\bigg| \\ &\le \frac{1}{n}\sum_{i=1}^n \bigg|q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\bigg(Z_{nij} + \mathbf{X}_i^T\frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)}{\partial \beta_{nj}}\bigg)\bigg| \cdot |a_{ik} + b_{ik}| \\ &\le \sup_{1\le k\le p_n, 1\le i\le n} |a_{ik} + b_{ik}| \cdot \bigg\{E_0\bigg(|q_1(m_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})|\bigg|Z_{nij} + \mathbf{X}_i^T\frac{\partial \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)}{\partial \beta_{nj}}\bigg|\bigg) + o_P(1)\bigg\} \\ &\le O_P\bigg(h^{p+1} + \frac{1}{\sqrt{nh}} + \frac{p_n}{\sqrt{n^{d_h}}}\bigg), \end{aligned}$$

where the second last line follows from mean variance decomposition and conditions (A) and (C). This shows that

$$\|K_6\| \le O_P\bigg(p_n\bigg(h^{p+1} + \frac{1}{\sqrt{nh}}\bigg) + \sqrt{\frac{p_n^4}{n^{d_h}}}\bigg) = o_P(1)$$

by the conditions of the Theorem.

For K_7, note that

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

$$\leq\frac{1}{n^2\epsilon^2}E_0\sum_{j,k=1}^{p_n}\sum_{i=1}^{n}\left\{\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nk}}-E_0\left(\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nk}}\right)\right\}^2$$

$$=O\left(\frac{np_n^2}{n^2\epsilon^2}\right)=O\left(\frac{p_n^2}{n}\right)=o(1),$$

which implies that $\|K_7\|=o_P(1)$. Hence using $K_4=K_6+K_7$,

$$\|K_4\|\leq o_P(1)+O_P\left(p_n\left(h^{p+1}+\frac{1}{\sqrt{nh}}\right)+\sqrt{\frac{p_n^4}{n^{d_h}}}\right)=o_P(1).$$

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

$$|A_j|\leq\sup_{i,j}|a_{ij}+b_{ij}|+\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Q_{ni}(\boldsymbol{\beta}_{n0})}{\partial\beta_{nj}}\right|$$

$$\leq O_P\left(h^{p+1}+\frac{1}{\sqrt{nh}}+\sqrt{\frac{p_n^4}{n^{d_h}}}\right)+O_P(n^{-1/2}),$$

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

$$\|K_5\|\leq O_P\left(p_n\left(h^{p+1}+\frac{1}{\sqrt{nh}}+\sqrt{\frac{p_n^4}{n^{d_h}}}\right)^2\right)=o_P(1),$$

and the proof completes. □

6 Appendix

Lemma 14 (Lemma A.1) *Let C and D be respectively compact sets in R^d and R^p and $f(\mathbf{x},\boldsymbol{\theta})$ is a continuous function in $\boldsymbol{\theta}\in C$ and $\mathbf{x}\in D$. Assume that $\hat{\boldsymbol{\theta}}(\mathbf{x})\in C$ is continuous in $\mathbf{x}\in D$, and is the unique maximizer of $f(\mathbf{x},\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_n(\mathbf{x})\in C$ be a maximizer of $f_n(\mathbf{x},\boldsymbol{\theta})$. If*

$$\sup_{\theta\in C,\mathbf{x}\in D}|f_n(\mathbf{x},\boldsymbol{\theta})-f(\mathbf{x},\boldsymbol{\theta})|\to 0,\ \textit{then}\ \sup_{\mathbf{x}\in D}|\hat{\boldsymbol{\theta}}_n(\mathbf{x})-\hat{\boldsymbol{\theta}}(\mathbf{x})|\to 0,\ \textit{as}\ n\to\infty.$$

Proof: This is Lemma A.1 of Carroll et al. (1997).

Lemma 15 (Lemma A.2) *Let* $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ *be i.i.d. random vectors, where the* Y_i*'s are scalar random variables. Assume further that* $E|Y|^r < \infty$ *and* $\sup_{\mathbf{x}} \int |y|^r f(\mathbf{x}, y) dy < \infty$ *where* f *denotes the joint density of* $(\mathbf{X}, Y)$*. Let* K *be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then,*

$$\sup_{\mathbf{x} \in D} \left| n^{-1} \sum_{i=1}^{n} \{K_h(\mathbf{X}_i - \mathbf{x})Y_i - E[K_h(\mathbf{X}_i - \mathbf{x})Y_i]\} \right| = O_P\left(\sqrt{\frac{\log(1/h)}{nh}}\right),$$

provided that $n^{2\epsilon - 1}h \to \infty$ *for some* $\epsilon < 1 - r^{-1}$.

Proof: This is a direct result of Mack and Silverman (1982).

Lemma 16 *Under conditions of Theorem 1, when* $nh^{2p+2} = O(1)$,

$$\begin{aligned} \left\| \frac{1}{n} \nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| &= o_P\left(\frac{1}{p_n}\right), \\ \left\| \frac{1}{n} \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| &= o_P\left(\frac{1}{p_n}\right) + O_P\left(p_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right). \end{aligned}$$

Moreover, if $h = O(n^{-1/(2p+3)})$*, then assuming* $p_n^4/n^{(2p+2)/(2p+3)} = o(1)$,

$$\begin{aligned} \left\| \frac{1}{n} \nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| &= o_P\left(\frac{1}{p_n n^{1/(4p+6)}}\right), \\ \left\| \frac{1}{n} \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| &= o_P\left(\frac{1}{p_n n^{1/(4p+6)}}\right) + O_P\left(p_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right). \end{aligned}$$

Proof of lemma 16. First we assume $p_n^4/n \to 0$ and $nh^{2p+2} = O(1)$. Given $\epsilon > 0$, by Chebyshev's inequality,

$$\begin{aligned} &\mathbb{P}\left(p_n \left\| \frac{1}{n} \nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| \geq \epsilon\right) \\ &\leq \frac{p_n^2}{n^2\epsilon^2} E_0 \sum_{i,j=1}^{p_n} \left\{ \frac{\partial^2 Q_n(\boldsymbol{\beta}_{n0})}{\partial \beta_{ni} \partial \beta_{nj}} - E_0 \frac{\partial^2 Q_n(\boldsymbol{\beta}_{n0})}{\partial \beta_{ni} \partial \beta_{nj}} \right\}^2 \\ &= O\left(\frac{np_n^4}{n^2\epsilon^2}\right) = O\left(\frac{p_n^4}{n}\right) = o(1) \end{aligned}$$

which proves the first equation in the lemma. From this, triangle inequality immediately gives

$$\left\| \frac{1}{n} \nabla^2 \hat{Q}_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0}) \right\| = o_P\left(\frac{1}{p_n}\right) + \| n^{-1} \nabla^2 (\hat{Q}_n(\boldsymbol{\beta}_{n0}) - Q_n(\boldsymbol{\beta}_{n0})) \|.$$

Note that by Taylor's expansion,

$$\nabla^2(\hat{Q}_n(\boldsymbol{\beta}_{n0}) - Q_n(\boldsymbol{\beta}_{n0})) = \sum_{i=1}^{n} \nabla^2 q_1(\tilde{m}_{ni}(\boldsymbol{\beta}_{n0}), Y_{ni})\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_{n0}}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_{n0}}(U_i)),$$

where $\tilde{m}_{ni}(\boldsymbol{\beta}_n) = \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$, with $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$ lies between $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$ and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$. Expanding the above (details omitted), using lemma 2 and 3, Cauchy-Schwarz inequality and condition (C), we can obtain

$$\|n^{-1}\nabla^2(\hat{Q}_n(\boldsymbol{\beta}_{n0}) - Q_n(\boldsymbol{\beta}_{n0}))\| \leq O_P\left(p_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right),$$

and this yields the second equation in the lemma.

Now assume $h = O(n^{-1/(2p+3)})$ and $p_n^4/n^{(2p+2)/(2p+3)}$. Given $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(p_n n^{1/(4p+6)}\left\|\frac{1}{n}\nabla^2 Q_n(\boldsymbol{\beta}_{n0}) + I_n(\boldsymbol{\beta}_{n0})\right\| \geq \epsilon\right) \\
&\leq \frac{p_n^2 n^{1/(2p+3)}}{n^2\epsilon^2} E_0 \sum_{i,j=1}^{p_n}\left\{\frac{\partial^2 Q_n(\boldsymbol{\beta}_{n0})}{\partial\beta_{ni}\partial\beta_{nj}} - E_0\frac{\partial^2 Q_n(\boldsymbol{\beta}_{n0})}{\partial\beta_{ni}\partial\beta_{nj}}\right\}^2 \\
&= O\left(\frac{p_n^4}{n^{(2p+2)/(2p+3)}}\right) = o(1)
\end{aligned}$$

which proves the third equation. The fourth one follows from similar arguments as before. □

Proof of lemma 16. In expression (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. Using the same notation as in the proof of lemma 6, we have $\bar{\boldsymbol{\alpha}}_{ni}(u) = \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$, $\hat{\boldsymbol{\beta}}^* = c_n^{-1}(\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(u))$ and $\mathbf{X}_i^* = \mathbf{X}_i$. Following the proof of lemma 6, we arrive at equation (19), which in this case is reduced to

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

Solving for $\frac{\partial\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial\boldsymbol{\beta}_n}$ from the above equation, which is true for $j = 1, \cdots, p_n$, 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 $\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(u)$. However this is done by the proof of lemma 6 already, where equation (20) becomes

$$\left\|\frac{\partial\hat{\mathbf{a}}_{0\boldsymbol{\beta}_n}(u)}{\partial\boldsymbol{\beta}_n} - \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(u)\right\| = O_P\left(\sqrt{p_n}\left(h + \frac{1}{\sqrt{nh}}\right)\right) = o_P(1)$$

and the proof completes. □

Proof of lemma 9. Since $B_n B_n^T = I_{p_n-l}$, for each $\mathbf{v} \in \mathbb{R}^{p_n-l}$, we have

$$\|B_n^T \mathbf{v}\| \leq \|\mathbf{v}\|. \tag{29}$$

Following the proof of Theorem 1, we still have $\|B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n)\| = O_P\left(\sqrt{\frac{p_n}{n}}\right)$ when $nh^{2p+2} = O(1)$ (resp. $\|B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n)\| = O_P\left(\sqrt{\frac{p_n}{n}} \cdot n^{1/(4p+6)}\right)$ when $h = O(n^{-1/(2p+3)})$). Hence under $p_n^5/n \to 0$ (resp. $p_n^5/n^{(2p+1)/(2p+3)}$), following the proof of Theorem 2,

$$\begin{aligned}
&I_n(\boldsymbol{\beta}_{n0})B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}) = \frac{1}{n}\nabla \hat{Q}_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
\overset{\text{lemma8}}{\Rightarrow}\ &I_n(\boldsymbol{\beta}_{n0})B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}) = \frac{1}{n}\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
\overset{\text{Eqn.(29)}}{\Rightarrow}\ &B_n I_n(\boldsymbol{\beta}_{n0})B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}) = \frac{1}{n}B_n\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
\Rightarrow\ &B_n^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}) = \frac{1}{n}B_n^T(B_n I_n(\boldsymbol{\beta}_{n0})B_n^T)^{-1}B_n\nabla Q_n(\boldsymbol{\beta}_{n0}) + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where the last line is true since $B_n I_n(\boldsymbol{\beta}_{n0})B_n^T$ has eigenvalues uniformly bounded away from 0 and infinity, like $I_n(\boldsymbol{\beta}_{n0})$ does. □

Proof of lemma 10. First we assume $nh^{2p+2} = O(1)$. By Taylor's expansion and Cauchy-Schwarz inequality,

$$\begin{aligned}
\frac{1}{n^2}\|\nabla^2\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})\|^2 &\leq \frac{1}{n^2}\left\|\nabla^T(\nabla^2\hat{Q}_n(\beta_n^*))\right\|^2 \cdot \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|^2 \\
&= \frac{1}{n^2}O_P(n^2p_n^3)\cdot O_P\left(\frac{p_n}{n}\right) \\
&= O_P\left(\frac{p_n^4}{n}\right) = o_P\left(\frac{1}{p_n}\right),
\end{aligned}$$

where β_n^* lies between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_{n0}$. The second line follows from the result of Theorem 1 and the proof of order for $|\hat{I}_3|$ in the Theorem.

If $h = O(n^{-1/(2p+3)})$, then

$$\begin{aligned}
\frac{1}{n^2}\|\nabla^2\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \nabla^2\hat{Q}_n(\boldsymbol{\beta}_{n0})\|^2 &\leq \frac{1}{n^2}\left\|\nabla^T(\nabla^2\hat{Q}_n(\beta_n^*))\right\|^2 \cdot \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|^2 \\
&= \frac{1}{n^2}O_P(n^2p_n^3)\cdot O_P\left(\frac{p_n}{n}\cdot n^{1/(2p+3)}\right) \\
&= O_P\left(\frac{p_n^4}{n}\cdot n^{1/(2p+3)}\right) = o_P\left(\frac{1}{p_n}\right),
\end{aligned}$$

where the second line follows from the proof of Theorem 1 again. The last line holds since we assumed $p_n^5/n^{(2p+2)/(2p+3)} \to 0$. □

Proof of lemma 11. By Taylor's expansion, expanding $\hat{Q}(B_n^T\hat{\boldsymbol{\gamma}}_n)$ at $\hat{\boldsymbol{\beta}}_n$,

$$\begin{aligned}\hat{Q}_n(\hat{\boldsymbol{\beta}}_n) - \hat{Q}_n(B_n^T\hat{\boldsymbol{\gamma}}_n) &= \nabla^T\hat{Q}_n(\hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n) \\ &- \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n)^T\nabla^2\hat{Q}_n(\hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n) \\ &+ \frac{1}{6}\nabla\{(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n)^T\nabla^2\hat{Q}_n(\beta_n^*)(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n)\}(\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n) \\ &:= T_1 + T_2 + T_3.\end{aligned}$$

Note $T_1 = 0$ by definition of $\hat{\boldsymbol{\beta}}_n$. Denote $\Theta_n = I_n(\boldsymbol{\beta}_{n0})$ and $\boldsymbol{\Phi}_n = \frac{1}{n}\nabla Q_n(\boldsymbol{\beta}_{n0})$. Using equation (25) and noting that Θ_n has eigenvalues uniformly bounded away from 0 and infinity (condition (B)), we have

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0} = \Theta_n^{-1}\boldsymbol{\Phi}_n + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Combining this with lemma 9, under the null hypothesis H_0,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n &= \Theta_n^{-1/2}\{I_n - \Theta_n^{1/2}B_n^T(B_n\Theta_nB_n^T)^{-1}B_n\Theta_n^{1/2}\}\Theta_n^{-1/2}\boldsymbol{\Phi}_n \\ &+ o_P(n^{-1/2}).\end{aligned} \tag{30}$$

But $S_n := I_n - \Theta_n^{1/2}B_n^T(B_n\Theta_nB_n^T)^{-1}B_n\Theta_n^{1/2}$ is a $p_n \times p_n$ idempotent matrix with rank $p_n - (p_n - l) = l$, it follows by s standard argument that

$$\|\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n\| = O_P\left(\sqrt{\frac{l}{n}}\right).$$

Hence using similar argument as in the approximation of order for $|\hat{I}_3|$ in Theorem 1, we have

$$\begin{aligned}|T_3| &= O_P(np_n^{3/2}) \cdot \|\hat{\boldsymbol{\beta}}_n - B_n^T\hat{\boldsymbol{\gamma}}_n\|^3 \\ &= O_P\left(np_n^{3/2} \cdot \frac{l^{3/2}}{n^{3/2}}\right) = O_P\left(\frac{p_n^{3/2}l^{3/2}}{\sqrt{n}}\right) \\ &= o_P(1).\end{aligned}$$

Hence

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

Finally by lemma 16 and 10, we have

$$\begin{aligned}
&\left\| \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - B_n^T \hat{\boldsymbol{\gamma}}_n)\{\nabla^2 \hat{Q}_n(\hat{\boldsymbol{\beta}}_n) + nI_n(\boldsymbol{\beta}_{n0})\}(\hat{\boldsymbol{\beta}}_n - B_n^T \hat{\boldsymbol{\gamma}}_n) \right\| \\
&\le O_P\left(\frac{l}{n}\right) \cdot n \left\{ o_P\left(\frac{1}{\sqrt{p_n}}\right) + O_P\left(p_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right)\right\} \\
&= o_P\left(\frac{l}{\sqrt{p_n}}\right) + O_P\left(lp_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right) = o_p(1),
\end{aligned}$$

and the conclusion of the lemma follows. □

Proof of lemma 12. Consider

$$\begin{aligned}
n^{-1}\|\nabla^2 Q_n(\boldsymbol{\beta}_n) - \nabla^2 Q_n(\boldsymbol{\beta}_{n0})\|^2 &= \frac{1}{n^2}\sum_{i,j=1}^{p_n}\left(\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_n)}{\partial\beta_{ni}\partial\beta_{nj}} - \frac{\partial^2 Q_n(\boldsymbol{\beta}_{n0})}{\partial\beta_{ni}\partial\beta_{nj}}\right)^2 \\
&= \frac{1}{n^2}\sum_{i,j=1}^{p_n}\left(\sum_{k=1}^{p_n}\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_{ni}\partial\beta_{nj}\partial\beta_{nk}}(\hat{\beta}_{nk} - \beta_{0k})\right)^2 \\
&\le \frac{1}{n^2}\sum_{i,j=1}^{p_n}\sum_{k=1}^{p_n}\left(\frac{\partial^3 Q_n(\boldsymbol{\beta}^*)}{\partial\beta_{ni}\partial\beta_{nj}\partial\beta_{nk}}\right)^2\|\hat{\beta}_{nk} - \beta_{0k}\|^2,
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_{n0}$. Similar to approximating the order of $\hat{I}_3$ in the proof of Theorem 1, the last line of the above equation is less than or equal to

$$\frac{1}{n^2}O_p(n^2p_n^3)\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|^2. \tag{31}$$

If $nh^{2p+2} = O(1)$, then by Theorem 1, we have $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P\left(\sqrt{\frac{p_n}{n}}\right)$. Hence

$$(31) = \frac{1}{n^2}O_P(n^2p_n^3)O_P\left(\frac{p_n}{n}\right) = O_P\left(\frac{p_n^4}{n}\right) = o_P(1).$$

If $h = O(n^{-(2p+3)}$, using similar arguments in the proof of Theorem 1 we have $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| \le O_P\left(\sqrt{p_n/n^{(2p+2)/(2p+3)}}\right)$. Hence

$$(31) = \frac{1}{n^2}O_P(n^2p_n^3)O_P\left(p_n/n^{(2p+2)/(2p+3)}\right) = O_P(p_n^4/n^{(2p+2)/(2p+3)}) = o_P(1). \ \square$$

Proof of lemma 13. By Taylor's expansion,

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial}{\partial\beta_{nk}}(\nabla\hat{Q}_n(\boldsymbol{\beta}_n) - \nabla Q_n(\boldsymbol{\beta}_n))\\
&= \frac{1}{n}\sum_{i=1}^{n} q_3(\tilde{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})(Z_{nik} + \left(\frac{\partial\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right)^T \mathbf{X}_i)(\mathbf{Z}_{ni} + \tilde{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i)\\
&\qquad\qquad \times \mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i))\\
&+ \frac{1}{n}\sum_{i=1}^{n} q_2(\tilde{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})\left(\frac{\partial\tilde{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right)\mathbf{X}_i\mathbf{X}_i^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i))\\
&+ \frac{1}{n}\sum_{i=1}^{n} q_2(\tilde{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})(\mathbf{Z}_{ni} + \tilde{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\mathbf{X}_i)\mathbf{X}_i^T\left(\frac{\partial\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}} - \frac{\partial\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right)\\
&+ \frac{1}{n}\sum_{i=1}^{n} q_2(\tilde{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})(Z_{nik} + \left(\frac{\partial\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right)^T \mathbf{X}_i)(\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i) - \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i))\mathbf{X}_i\\
&+ \frac{1}{n}\sum_{i=1}^{n} q_1(\tilde{m}_{ni}(\boldsymbol{\beta}_n), Y_{ni})\left(\frac{\partial\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}} - \frac{\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right)\mathbf{X}_i,
\end{aligned}
$$

where $\tilde{m}_{ni}(\boldsymbol{\beta}_n) = \tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)^T\mathbf{X}_i + \mathbf{Z}_{ni}^T\boldsymbol{\beta}_n$, with $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$ lies between $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)$ and $\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)$. By lemmas 6 and 7, the main order of the above sum comes from the non-tilde version of individual terms in the sum. Together with regularity conditions (A) and (C),

$$
\begin{aligned}
&\left\|\frac{1}{n}\frac{\partial}{\partial\beta_{nk}}(\nabla\hat{Q}_n(\boldsymbol{\beta}_n) - \nabla Q_n(\boldsymbol{\beta}_n))\right\|\\
&\quad \le O(1)\cdot\Bigg(\sup_i\|\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i) - \boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)\| + \sup_i\left\|\frac{\partial\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}} - \frac{\partial\boldsymbol{\alpha}_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right\|\\
&\quad + \sup_i\|\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i) - \hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)\| + \sup_i\left\|\frac{\partial\hat{\boldsymbol{\alpha}}'_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}} - \frac{\boldsymbol{\alpha}'_{\boldsymbol{\beta}_n}(U_i)}{\partial\beta_{nk}}\right\|\Bigg)\\
&\quad \le O(1)o_P\left(\sqrt{p_n}\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right),
\end{aligned}
$$

where the last line follows from lemma 6. Hence

$$
n^{-1}\|\nabla^2\hat{Q}_n(\boldsymbol{\beta}_n) - \nabla^2 Q_n(\boldsymbol{\beta}_n)\| \le o_P\left(p_n\left(h^{p+1} + \frac{1}{\sqrt{nh}}\right)\right) = o_P(1)
$$

which follows from conditions on h in the lemma. □

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.

[2] Albright, S.C., Winston, W.L. and Zappe, C.J. (1999). *Data Analysis and Decision Making with Microsoft Excel.* Duxbury, Pacific Grove, CA.

[3] Cai, Z., Fan, J. and Li, R. (2000). Efficient Estimation and Inferences for Varying-Coefficient Models. *J. Amer. Statist. Assoc.*, **95**, 888–902.

[4] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized Partially Linear Single-Index Models. *J. Amer. Statist. Assoc.*, **92**, 477–489.

[5] Donoho, D.L., High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Lecture on August 8, 2000, to the American Mathematical Society "Math Challenges of the 21st Century".

[6] Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310–320.

[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* New York: Chapman and Hall.

[8] Fan, J. and Huang, T. (2005). Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models. *Bernoulli.*, **11**, 1031–1057.

[9] Fan, J. and Jiang, J. (1999). Variable bandwidth and One-step Local M-Estimator. *Science in China*, **29**, 1–15; (English series) 2000, **35**, 65 – 80.

[10] Fan, J. and Li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the Madrid International Congress of Mathematicians 2006.* To appear.

[11] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**, 928–961.

[12] Fan, J., Tam, P., Vande Woude, G. and Ren, Y. (2004) Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl. Acad. Sci. USA*, **101**, 1135–1140.

[13] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Ann. Statist.*, **29**, 153–193.

[14] Härdle, W., Liang, H. and Gao, J.T. (2000). *Partially Linear Models.* Springer-Verlag, New York.

[15] Hastie, T.J. and Tibshirani, R. (1993). Varying-coefficient models. *J. R. Statist. Soc.* B, **55**, 757–796.

[16] 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.

[17] Huber, P.J. (1981). *Robust Statistics.* New York: John Wiley & Sons.

[18] Huber, P.J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

[19] Jain, N. and Marcus, M. (1975). Central Limit Theorems for C(S)-valued Random Variables. *J. Funct. Anal.*, **19**, 216–231.

[20] Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.*, **96**, 1387–1396.

[21] Li, Q., Huang, C.J., Li., D. and Fu, T.T. (2002). Semiparametric smooth coefficient models. *J. Bus. Econom. Statist.*, **20**, 412–422.

[22] Li, R. and Liang, H. (2005). Variable Selection in Semiparametric Regression Modeling. To appear.

[23] Lin, X. and Carroll, R.J. (2006). Semiparametric estimation in general repeated measures problems. *J. R. Statist. Soc.* B, **68**, Part 1, 69–88.

[24] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

[25] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econ. Theory*, **7**, 186–199.

[26] Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity. *Ann. Statist.*, **16**, 356–366.

[27] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *J. Amer. Statist. Assoc.*, **89**, 501–511.

[28] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B, **50**, 413–436.

[29] Van Der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Univ. Press.

[30] Wahba, G. (1984) Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329. Institute of Statistical Mathematics, Tokyo.

[31] Wong, W.H. and Severini, T.A. (1992). Profile Likelihood and Conditionally Parametric Models. *Ann. Statist.*, **20**, 1768–1802.

[32] Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika*, **91**, 661–681.

[33] Yatchew, A. (1997). An elementary estimator for the partially linear model. *Economics Letters*, **57**, 135–143.

[34] Zhang, W., Lee, S.Y., and Song, X.Y. (2002). Local Polynomial fitting in semivarying coefficient model. *J. Mult. Anal.*, **82**, 166–188.