Bernoulli **9**(4), 2003, 579–605

Local polynomial fitting based on empirical likelihood

JIAN ZHANG[1,2] and ANNA LIU[3]

[1]*Institute of Mathematics and Statistics, University of Kent at Canterbury, Kent CT2 7NF, UK*
E-mail: J.Zhang@ukc.ac.uk
[2]*Chinese Academy of sciences, Beijing 100080, China*
[3]*Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106-3110, USA*

A new nonparametric regression technique is proposed which involves the extension of local polynomial fitting to the empirical likelihood context, where the distribution of the stochastic error is not fully specified. The aim of this extension is to reduce the possible modelling bias of parametric likelihood and to allow one to use the auxiliary information about the stochastic error in the local polynomial fitting. The asymptotic bias and variance, consistency and asymptotic distribution of the proposed estimators are established. The proposed estimators are shown to inherit the main advantage of the local polynomial estimator based on the parametric likelihood over the Nadaraya–Watson kernel estimator near the boundaries. Moreover, the proposed estimators can be more flexible and efficient than the parametric likelihood based local polynomial estimator when the distribution of the stochastic error is misspecified. The new method is illustrated with applications to some simulated and real data sets.

Keywords: empirical likelihood; local polynomial; nonparametric regression.

1. Introduction

The method of empirical likelihood, introduced by Owen (1988), is commonly employed to deal with the possible modelling bias of parametric likelihood. In the present paper, a new estimator for a nonparametric function is developed by incorporating this method into the framework of local polynomial modelling. By local polynomial expansion we reduce the nonparametric function estimation problem to several parametric estimation problems. Then the empirical likelihood approach can be applied to each parametric problem. Unlike the parametric likelihood based estimators (here 'parametric likelihood' means the likelihood based on the parametric model of the stochastic error in the regression case; see, for example, Fan and Gijbels 1996), the new estimator only requires one to specify some conditional estimating equations rather than the full probabilistic mechanism for the observations. It thus allows one to relax not only the assumptions imposed on the form of a regression function but also those imposed on the stochastic error.

By way of illustration, we consider the regression model

$$Y = \theta(X) + \varepsilon$$

with response Y, covariate X, regression function θ, and stochastic error ε. Given X, ε is assumed to be symmetrically distributed, that is, $\theta(X)$ is the centre of symmetry of Y. This model is just the symmetric location model when θ is restricted to a finite-dimensional parametric space, which has been well studied (see Bickel *et al*. 1993, pp. 75 and 400–405). Here we consider the nonparametric case where θ is a nonparametric function from [0, 1] to $\mathbb{R}^1$ with $p+1$ continuous derivatives. To use the information about ε, we let $0 = s_0 < s_1 < \ldots < s_{k_0}$ and $S_k = [s_{k-1}, s_k)$, $1 \leqslant k \leqslant k_0$. Set $H_k(y, \theta(x)) = I(y - \theta(x) \in S_k) - I(y - \theta(x) \in -S_k)$, $1 \leqslant k \leqslant k_0$, where $I(\cdot)$ is the indicator of a set. Let $H = (H_1, \ldots, H_{k_0})^{\mathrm{T}}$. Then we have the conditional equations

$$\mathrm{E}\{H_k(Y, \theta(X))|X\} = 0, \qquad 1 \leqslant k \leqslant k_0, \tag{1.1}$$

for θ. Note that as $\max_{1 \leqslant k \leqslant k_0}(s_k - s_{k-1}) \to 0$, $k_0 \to \infty$, these equations are asymptotically equivalent to the assumption that ε is symmetric. These kinds of constraints were introduced in Zhang and Gijbels (1999; 2003).

Let (x_i, y_i), $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) observations from the above model. Given $x_0 \in (0, 1)$, if we have n i.i.d. observations y_i, $i = 1, \ldots, n$, with the same covariate x_0, then the conditional nonparametric likelihood at $\theta(x_0)$ is of the form $\prod_{i=1}^n p_i$, where p_i is the mass we place at point (x_0, y_i). In practice, observations with the same covariate x_0 are rare. This problem can be solved by the local modelling technique (see Fan and Gijbels 1996): take all (x_i, y_i), weight the logarithm of the nonparametric likelihood in such a way that it places more emphasis on those observations with covariates close to x_0, and at the same time approximate $\theta(x)$ in (1.1) by its pth-order Taylor expansion at x_0. More specifically, let $K(\cdot)$ be a bounded symmetric density function with support $[-1, 1]$. Set $K_h(\cdot) = K(\cdot/h)/h$ and $\underline{X}(t) = (1, t, \ldots, t^p)^{\mathrm{T}}$. Then the profile local polynomial empirical likelihood function at x_0 is defined as follows:

$$l(\beta) = \sup\left\{\sum_{i=1}^n K_h(x_i - x_0)\log p_i | p_i \geqslant 0,\, 1 \leqslant i \leqslant n,\, \sum_{i=1}^n p_i = 1,\, \sum_{i=1}^n p_i \underline{H}(y_i, x_i, x_0, \beta) = 0\right\}, \tag{1.2}$$

where $\otimes$ is the Kronecker product, $\beta = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$, and

$$\underline{H}(y_i, x_i, x_0, \beta) = H\left(y_i, \underline{X}\left(\frac{x_i - x_0}{h}\right)^{\mathrm{T}}\beta\right) \otimes \underline{X}\left(\frac{x_i - x_0}{h}\right).$$

It is easily shown by the Lagrange multiplier method that

$$\begin{aligned} l(\beta) = &\sum_{i=1}^n K_h(x_i - x_0)\log\left[K_h(x_i - x_0)\Big/\sum_{j=1}^n K_h(x_j - x_0)\right] \\ &- \sum_{i=1}^n K_h(x_i - x_0)\log(1 + \alpha_n(x_0, \beta)^{\mathrm{T}}\underline{H}(y_i, x_i, x_0, \beta)), \end{aligned}$$

where $\alpha_n(x_0, \beta)$ satisfies

$$\sum_{i=1}^{n} K_h(x_i - x_0) \frac{\underline{H}(y_i, x_i, x_0, \beta)}{1 + \alpha_n(x_0, \beta)^{\mathrm{T}} \underline{H}(y_i, x_i, x_0, \beta)} = 0. \tag{1.3}$$

Choose an appropriate space, say Θ_0, for β. Let $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$ be the maximum estimator over Θ_0 based on $l(\beta)$. Then the local polynomial empirical likelihood estimator of $\theta(x_0)$ is given by $\widehat{\theta}(x_0) = \widehat{\beta}_0$. Through the coefficients of the higher-order terms in the polynomial fit, $\widehat{\beta}$ provides an estimator for the higher-order derivative $\theta^{(r)}(x_0)$, namely, $\widehat{\theta}_r(x_0) = r!\widehat{\beta}_r/h^r$. $\widehat{\beta}$ also provides an estimator for the conditional distribution of Y given $X = x_0$, say $\widehat{F}_{Y|X=x_0}$, with

$$\widehat{F}_{Y|X=x_0}(\{y_i\}) = \frac{K_h(x_i - x_0)}{1 + \alpha_n(x_0, \widehat{\beta})^{\mathrm{T}} \underline{H}(y_i, x_i, x_0, \widehat{\beta})}, \qquad i = 1, \ldots, n. \tag{1.4}$$

In this paper, we study these kinds of estimators under a more general set of conditional equations, which includes the conditional symmetric model as a special case. Under some regularity conditions, the above estimators are proved to be consistent and asymptotically normal. The asymptotic bias and variance are also derived, which have the same performance as the parametric likelihood based local polynomial estimator near the boundaries. It is shown that the new estimators can be more flexible and efficient than the parametric likelihood based local polynomial estimator. In particular, in the setting of the symmetric location model, the new estimators are nearly adaptive with respect to the unknown density function of ε. That is, when the number of equations in (1.1) tends to infinity, we can estimate the regression function asymptotically equally well whether or not we know the density of ε. This implies that the least squares based local polynomial estimator may be inefficient when the stochastic error is not normal. Note that the least squares based local polynomial estimator can be used under the assumption that the second moment of the stochastic error exists.

The idea of using the local polynomial fit to the parametric likelihood based regression models appeared, for example, in Stone (1977), Cleveland (1979), Tibshirani and Hastie (1987) and Fan and Gijbels (1996). Carroll *et al*. (1998) developed an alternative method called the local moment method. It is known that the empirical likelihood has the advantage over the moment method that a weaker restriction is imposed on the model (see Hanfelt and Liang 1995; Kitamura 1997; and Qin and Lawless 1994). In a similar setting, Zhang and Gijbels (1999; 2003) introduced an approximate empirical likelihood for a nonparametric function and gave the global convergence rate of the corresponding maximum estimator. Unlike the above estimators, our new one are based on local weighting of logarithms of empirical likelihoods.

The rest of this paper is organised as follows. In Section 2 we investigate the asymptotic properties of the proposed estimators. Applications to both simulated and real data sets are presented in Section 3. The technical conditions are given in Section 4. The proofs of the main results can be found in the Appendix.

2. Asymptotic theory

In what follows, we consider a general nonparametric regression model with response Y, covariate X, and a general constraint function $G = (G_1, \ldots, G_{k_0})^{\mathrm{T}}$. Assume that the regression function $\theta(\cdot)$ has $p+1$ continuous derivatives. Adopt the same notation $\underline{\beta}$, $\theta(x_0)$, $\theta^{(r)}(x_0)$, $\alpha_n(x_0, \beta)$, $\underline{l}(\beta)$, and Θ_0 as in Section 1, and the associated estimators $\widehat{\underline{\beta}}$, $\widehat{\theta}(x_0)$, $\hat{\theta}_r(x_0)$ and $\alpha_n(x_0, \widehat{\beta})$, but replace H (and $\underline{H}$) by the general function G (and $\underline{G}$), which satisfies

$$\mathrm{E}[G_k(Y, \theta(X))|X] = 0, \qquad k = 1, 2, \ldots, k_0. \tag{2.1}$$

Note that the general nonparametric regression model reduces to the ordinary nonparametric mean regression model and the median regression model if we set $G = Y - \theta(X)$ and $G = I(Y \leqslant \theta(X)) - \frac{1}{2}$, respectively. To keep our proofs simple, we assume that G has a continuous derivative with respect to θ.

2.1. Bias, covariance and normality

Let $\lambda_0 = (\theta(x_0), h\theta^{(1)}(x_0), \ldots, h^p\theta^{(p)}(x_0)/p!)$ be an inner point of Θ_0. We begin by showing in the following theorem that $\widehat{\theta}(x_0)$ is weakly consistent.

Theorem 1. *Under Conditions* A1–A8 *in Section 4, for* $2 < \alpha_1 \leqslant \alpha_0$ *(with* α_0 *defined in Condition* A2*), as* $h = h_n \to 0$, $hn^{1-2/\alpha_1}/\log n \to \infty$ *and* $h^{p+1}n^{1/\alpha_1} \to 0$, *we have*

$$\widehat{\beta} - \lambda_0 = o_p(n^{-1/\alpha_1}), \qquad \alpha_n(x_0, \widehat{\beta}) = o_p(n^{-1/\alpha_1}).$$

For the next theorem, we set

$$\mu_{j+l} = \int t^{j+l}K(t)\mathrm{d}t, \qquad \nu_{j+l} = \int t^{j+l}K^2(t)\mathrm{d}t,$$
$$S = (\mu_{j+l})_{0\leqslant j,l\leqslant p}, \qquad S^* = (\nu_{j+l})_{0\leqslant j,l\leqslant p},$$
$$V_G(x_0) = \mathrm{E}[G(Y, \theta(x_0))G^{\mathrm{T}}(Y, \theta(x_0))|X = x_0],$$
$$D_G(x_0) = \mathrm{E}[\partial G(Y, \theta(x_0))/\partial\theta|X = x_0].$$

Let f be the density of X, and define

$$V_{\alpha G}(x_0) = \frac{1}{f(x_0)}\Big[V_G(x_0)^{-1} - V_G(x_0)^{-1}D_G(x_0)\big(D_G(x_0)^{\mathrm{T}}V_G(x_0)^{-1}D_G(x_0)\big)^{-1}$$
$$\times\ D_G(x_0)^{\mathrm{T}}V_G(x_0)^{-1}\Big] \otimes S^{-1}S^*S^{-1},$$

$$V_{\beta G}(x_0) = \frac{1}{f(x_0)}[D_G(x_0)^{\mathrm{T}}V_G(x_0)^{-1}D_G(x_0)]^{-1}S^{-1}S^*S^{-1}.$$

If f and $\theta^{(p+1)}$ have continuous derivatives, then we define

$$\begin{aligned}\text{bias} = {} & h^{p+1}S^{-1}(\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}\frac{\theta^{(p+1)}(x_0)}{(p+1)!} + h^{p+2}S^{-1}(\mu_{p+2}, \ldots, \mu_{2p+2})^{\mathrm{T}} \\ & \times \left\{\frac{\theta^{(p+2)}(x_0)}{(p+2)!} + \frac{\theta^{(p+1)}(x_0)}{(p+1)!}\frac{f'(x_0)}{f(x_0)}\right\}.\end{aligned} \tag{2.2}$$

Let $\xrightarrow{\mathcal{L}}$ stand for convergence in distribution. The next theorem establishes the asymptotic normality of $\widehat{\beta}$ and $\alpha_n(x_0, \widehat{\beta})$.

Theorem 2. *Suppose that Conditions* A1–A8, B1 *and* B2 *in Section 4 hold. Suppose that f and $\theta^{(p+1)}$ have continuous derivatives. Then as $h = h_n \to 0$, $hn^{1-2/\alpha_0}/\log n \to \infty$ with α_0 defined in Condition* A2 *and $h^{p+1}n^{1/\alpha_0} \to 0$,*

$$\sqrt{nh}V_{\beta G}(x_0)^{-1/2}\{\widehat{\beta} - \lambda_0 - \text{bias}(1 + o(1))\} \xrightarrow{\mathcal{L}} \mathrm{N}(0, I_{p+1}).$$

Furthermore, if $nh^{2p+3} \to 0$, then

$$\sqrt{nh}V_{\alpha G}(x_0)^{-1/2}\alpha_n(x_0, \hat{\beta}) \xrightarrow{\mathcal{L}} N(0, I_{k_0(p+1)}),$$

where I_{p+1} and $I_{k_0(p+1)}$ are the $p \times p$ and $k_0(p+1) \times k_0(p+1)$ unit matrices, and $N(0, I_{p+1})$ and $N(0, I_{k_0(p+1)})$ are normal distributions.

Remark 1. $\alpha_n(x_0, \widehat{\beta})$ is useful in developing some asymptotic theories for the estimated conditional distribution $\hat{F}_{Y|X=x_0}$ as shown in (1.4) and for the nonparametric likelihood ratio statistics investigated in Fan and Zhang (2000).

Remark 2. The requirement that G is differentiable in θ can be relaxed by imposing some entropy condition on G (see Condition A4′ in Section 4). Then Theorems 1 and 2 can cover the special example in (1.1). For example, suppose G is bounded. Then, under Conditions A1, A4′, and A5–A8, as $h = h_n \to 0$, $nh/\log n \to \infty$,

$$\widehat{\beta} - \lambda_0 = o_p(1), \qquad \alpha_n(x_0, \widehat{\beta}) = o_p(1).$$

Furthermore, asymptotic normality still holds if we impose second-order differentiability on $\mathrm{E}\{G(Y, t)|X\}$ with respect to t. Here $D_G(x_0)$ should be defined as $\partial \mathrm{E}[G(Y, \theta(x_0))|X = x_0]/\partial t$. A rigorous justification of the statement is tedious but very similar to that given in Zhang and Gijbels (1999) and so not pursued here.

Remark 3. Let e_{r+1} denote the unit vector with 1 in the $(r+1)$th position. Then, from Theorem 2, we obtain the asymptotic value of $\text{bias}\{\hat{\theta}_r(x_0)\}$ (defined as the leading term of the bias of $\hat{\theta}_r(x_0)$): for odd $p - r$,

$$\text{bias}\{\widehat{\theta}_r(x_0)\} = e_{r+1}^{\mathrm{T}}S^{-1}(\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}\frac{r!}{(p+1)!}\theta^{(p+1)}(x_0)h^{p+1-r}(1 + o(1)); \tag{2.3}$$

for even $p-r$,

$$\text{bias}\{\widehat{\theta}_r(x_0)\} = e_{r+1}^{\mathrm{T}} S^{-1}(\mu_{p+2}, \ldots, \mu_{2p+2})^{\mathrm{T}} \frac{r!}{(p+2)!} \Big[\theta^{(p+2)}(x_0) + (p+2)\theta^{(p+1)}(x_0)\frac{f'(x_0)}{f(x_0)}\Big] h^{p+2-r}(1+o(1)). \tag{2.4}$$

This is exactly the same as in the case of the standard local polynomial fit (Fan and Gijbels, 1996, p. 62), where there is a theoretical difference between the cases $p-r$ odd and $p-r$ even. For $p-r$ even, the leading term $O(h^{p+1})$ in the bias expression (2.2) is zero due to the symmetry of the kernel K and thus the second-order term is presented in (2.4). For $p-r$ odd, the asymptotic bias has a simpler structure and does not involve $f'(x_0)$, a factor appearing in the asymptotic bias when $p-r$ is even.

We also have the asymptotic variance of $\hat{\theta}$:

$$\text{var}\{\widehat{\theta}_r(x_0)\} = e_{r+1}^{\mathrm{T}} S^{-1} S^* S^{-1} e_{r+1} \frac{(r!)^2[D_G(x_0)^{\mathrm{T}} V_G(x_0)^{-1} D_G(x_0)]^{-1}}{f(x_0) n h^{2r+1}}(1+o(1)). \tag{2.5}$$

As a result of Theorem 2, we obtain

$$\text{var}\{\hat{\theta}_r(x_0)\}^{-1/2}\Big\{\hat{\theta}_r(x_0) - \theta_r(x_0) - \text{bias}\{\hat{\theta}_r(x_0)\}(1+o(1))\Big\} \xrightarrow{\mathcal{L}} \mathrm{N}(0, 1).$$

Remark 4. Note that the kernel function K has support $[-1, 1]$. Then for each x_0, the local neighbourhood of x_0 in our procedure is $[x_0-h, x_0+h]$. Since we assume for simplicity that the covariate X has support $[0, 1]$, this neighbourhood can lie outside $[0, 1]$ as x_0 is close to the boundary. When this happens, x_0 is called a boundary point. More specifically, x_0 is referred to a left (or right) boundary point if $x_0 - h < 0$ (or $x_0 + h > 1$). We consider only the left boundary points of the form $x_0 = ch$ and the right boundary points of the form $x_0 = 1 - ch$, with $c > 0$. For $c < 1$, the ch and $1-ch$ are the boundary points, whereas for $c > 1$ they are interior points. Like Fan and Gijbels (1996, p. 69), we wish to address the question of whether our procedure suffers boundary effects, that is, whether the orders of the asymptotic biases and variances of our estimators are different at the boundary and in the interior. For this purpose we derived the following asymptotic bias and variance formulae for our estimators. These formulae are derived in a way analogous to those for a fixed interior point. Let

$$\mu_{j,c} = \int_{-c}^{1} u^j K(u)\mathrm{d}u, \qquad S = (\mu_{j+l,c})_{0\leqslant j,l\leqslant p}, \qquad K_{r,c}^*(t) = e_{r+1}^{\mathrm{T}} S_c^{-1}(1, t, \ldots, t^p)^{\mathrm{T}} K(t).$$

Then for $x_0 = ch$,

$$\text{bias}\{\widehat{\theta}_r(x_0)\} = \left\{\int_{-c}^{1} t^{p+1} K_{r,c}^*(t)\mathrm{d}t\right\} \frac{r!}{(p+1)!}\theta^{(p+1)}(0+)h^{p+1-r}(1+o(1)). \tag{2.6}$$

and

$$\mathrm{var}\{\widehat{\theta}_r(x_0)\} = \int_{-c}^{1} K^{*2}_{r,c}(t)\mathrm{d}t \frac{(r!)^2[D(0+)^{\mathrm{T}} V(0+)^{-1} D(0+)]^{-1}}{f(0+)nh^{2r+1}}(1 + o(1)). \tag{2.7}$$

For right boundary points $x_0 = 1 - ch$, the asymptotic bias and variance expressions are similar to those provided in (2.6) and (2.7), but with the integral interval $[-c, 1]$ replaced by $[-1, c]$ and $0+$ by $1-$. For even $p - r$, a comparison of (2.4) and (2.5) with (2.6) and (2.7) shows that the order of the asymptotic bias is different at the boundary and in the interior. In contrast, for odd $p - r$, the asymptotic bias and variance are of the same order at the boundary and in the interior, and are continuous in c. This means that our procedure adapts automatically to estimation at the boundaries if we choose odd $p - r$.

Remark 5. The odd-degree fit is better than the even-degree fit. The reason is that for even $p - r$ not only the unknown derivative $\theta^{(p+1)}(x_0)$ but also unknown $f'(x_0)$ and $\theta^{(p+1)}(x_0)$ are involved in the asymptotic bias. Moreover, the proposed procedure will suffer boundary effects. In contrast, for odd $p - r$ only $\theta^{(p+1)}(x_0)$ is unknown in the asymptotic bias. Furthermore, the proposed procedure does not suffer boundary effects, as shown in Remark 4.

2.2. Efficiency and adaptiveness

Note that it follows from Theorem 2 that the bias of $\hat{\beta}$ is asymptotically free of the constraint function G. This leads to a simple criterion, the asymptotic covariance $V_{\beta G}(x_0)$, for the comparison of the efficiencies of the above local estimators derived from a class of constraint functions which satisfy the regularity conditions A1–A8, B1 and B2 in Section 4. Let $l(z, x_0) = \partial \log f_{\varepsilon|X=x_0}/\partial z$, where $f_{\varepsilon|X=x_0}$ is the conditional density of ε given $X = x_0$. It follows directly from Bhapkar (1991) that

$$D_G(x_0)^{\mathrm{T}} V_G(x_0)^{-1} D_G(x_0) \leqslant \mathrm{E}\, l(z, x_0)^2$$

for any estimating function G such that

$$\mathrm{E}[G(Y, \theta(X))|X = x_0] = 0, \qquad V_G(x_0) < \infty,$$

and $D_G(x_0)$ exists. Thus

$$V_{\beta G}(x_0) \geqslant V_{\beta l(z,x_0)}.$$

When equality holds, $\widehat{\beta}$ is efficient.

As pointed out by a referee, the usual notion of adaptiveness appears to be limited to semi-parametric models. For instance, if θ is a parametric function parametrized by a finite-dimensional vector, say γ, and if the model is $Y = \theta(X, \gamma) + \varepsilon$, where ε is symmetric conditional on X, then we know that adaptive estimation of γ is possible, as shown in Bickel *et al*. (1993). It is also possible to generalize this notion to local polynomial models by defining a local efficiency bound. We conjecture that in the parametric case this bound can be achieved by appropriately choosing a sequence of functions for G. For example, in the setting of the symmetric location model mentioned in Section 1, we let $G = H$ defined in (1.1) and k_0 converge to ∞ with increasing sample size. A more detailed description of this notion and the proofs of the related results are beyond the scope of this paper.

3. Numerical examples

3.1. Bandwidth selection

When we apply the local polynomial empirical likelihood estimator to a finite sample, we must first select the bandwidth. This smoothing parameter plays a very important role in the trade-off between reducing bias and variance, so we need to choose it carefully instead of randomly. There are different kinds of bandwidth selection method (Fan and Gijbels 1996, Chapter 4). We follow Carroll *et al*. (1998), viewing the mean square error (MSE) as a function of h. Ideally, we should choose the optimal bandwidth by minimizing the MSE function with respect to h, where

$$\mathrm{MSE}(x_0, h) = \mathrm{var}(x_0, h) + \mathrm{bias}^2(x_0, h)$$

with $\mathrm{var}(x_0, h)$ and $\mathrm{bias}(x_0, h)$ being the variance and bias of $\widehat{\theta}(x_0)$, respectively. In practice, the MSE is unknown and estimated by the empirical bias bandwidth selection (EBBS) method and the sandwich method.

The basic idea behind EBBS is as follows. For fixed x_0 and h_0, according to the asymptotic results in our asymptotic theories, $\mathrm{bias}(x_0, h_0)$ should be of the form $\mathrm{bias}(x_0, h_0) = f(h_0, \gamma) = \gamma_1 h_0^{p+1} + \ldots + \gamma_t h_0^{p+t}$, where $t \geqslant 1$ and $\gamma = (\gamma_1, \ldots, \gamma_t)$ is unknown. The local polynomial estimator $\widehat{\theta}(x_0, h_0)$ should be well described by $\gamma_0 + f(h_0, \gamma) + o_p(h_0^{p+t})$, where $\gamma_0 = \theta(x_0)$ in the limit. Then let $(\widehat{\gamma}_0, \widehat{\gamma})$ minimize $\sum_{k=1}^{K}\{\widehat{\theta}(x_0, h_k) - (\widehat{\gamma}_0 + f(h_k, \widehat{\gamma}))\}^2$, in which $\{h_1, \ldots, h_K\}$ is a grid of bandwidths in a neighbourhood, H_0, of h_0 with $K \geqslant t+1$. It is obvious that if H_0 is small enough, the bias should be well estimated at h_0 by $f(h_0, \widehat{\gamma})$. In practice, we need to choose K and t. See Carroll *et al*. (1998) for specific selection techniques. In our simulation and real data fitting, we take $t = 1$ and $K = 3$. We are most attracted by the EBBS property of avoiding the direct estimation of the higher-order derivatives arising in the asymptotic bias formulae, which might limit the range of applications because of its complications.

The sandwich formula for the asymptotic covariance matrix of β is analogous to that in Carroll *et al*. (1998), that is,

$$\{\{\hat{D}(x_0)\}\{\hat{V}(x_0)\}^{-1}\{\hat{D}(x_0)\}\}^{-1},$$

where

$$\hat{D}(x_0) = \sum_{i=1}^{n} K_h(x_i - x_0)\left[\frac{\partial G(y_i, \underline{X}((x_i - x_0)/h)^{\mathrm{T}}\widehat{\beta})}{\partial\theta} \otimes \underline{X}((x_i - x_0)/h)\underline{X}^{\mathrm{T}}((x_i - x_0)/h)\right]$$

and

$$\hat{V}(x_0) = \sum_{i=1}^{n} K_h^2(x_i - x_0)[G(y_i, \underline{X}((x_i - x_0)/h)^{\mathrm{T}}\widehat{\beta})G^{\mathrm{T}}(y_i, \underline{X}((x_i - x_0)/h)^{\mathrm{T}}\widehat{\beta}) \otimes \underline{X}((x_i - x_0)/h)\underline{X}^{\mathrm{T}}((x_i - x_0)/h)].$$

It is easily seen from our asymptotic results that the sandwich formula provides consistent variance estimators.

3.2. Curve construction

In practice, to construct the estimator of the curve θ, we often begin by estimating the values of θ at $x_0 = t_j$, $j = 1, \ldots, m$, as shown in Section 1, where $0 = t_1 < \ldots < t_m = 1$ are equispaced grid points. Denote these estimators by $\hat{\theta}(t_j)$, $j = 1, \ldots, m$. Then, a naive approach for constructing a curve estimator is to simply connect these point estimators by lines. Unfortunately the resulting curve might not be smooth, especially when $m \geqslant n$, where n is the sample size. We can use the moving average technique to improve the smoothness of the resulting curve. The basic idea behind this technique is that for any i_0 with a prespecified constant m_w, we have $\theta(t_j) \approx \theta(t_{i_0})$, provided $|j - i_0| \leqslant m_w$. For those j, we can write $\widehat{\theta}(t_j) = \theta(t_{i_0}) + e_j$, where e_j is a random error. This leads to using $\sum_{|j-i_0|\leqslant m_w} \widehat{\theta}(t_j)/(2m_w + 1)$ rather than $\widehat{\theta}(t_{i_0})$ to estimate $\theta(t_{i_0})$. We choose $m_w = 1$ and $m = n$ in our examples below.

3.3. Simulation

In the following examples the x_i were generated from the uniform distribution on [0, 1]. Local linear empirical likelihood fitting (i.e., $p = 1$) is used to estimate the regression functions.

Example 1. The regression model is

$$Y = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon.$$

Given X, ε follows the t distribution with 3 degrees of freedom and the constraint function is

$$G(y, \theta(x)) = y - \theta(x).$$

Generate a sample of size 200. Figure 1 shows the performance of the local linear empirical likelihood fit when ε has heavy tails.

Example 2. Use the same notation as in Example 1, except that we now assume that, given X, ε follows the normal distribution $N(0, \sigma(X)^2)$, $\sigma(X)^2 = 1 + X^2$. Generate a sample of size 200. Figure 2 shows the performance of the local linear empirical likelihood fit when ε is heteroscedastic.

3.4. Application

Example 3. *Great Barrier Reef data*. In a survey of the fauna on the sea bed in an area lying between the coast of northern Queensland and the Great Barrier Reef, a sample of size 155 was collected from a number of locations. In view of the large numbers of types of species

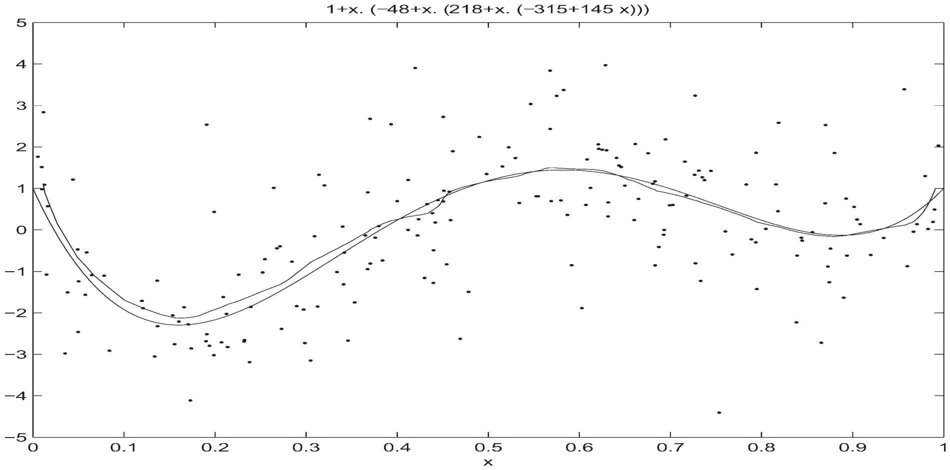

Figure 1. The smoother curve denotes the underlying regression function, while the other is the estimated curve derived from the local linear empirical likelihood fit under the first moment constraint. The data plotted, are sampled from the conditional model that given X, $\varepsilon \sim t_3$, the t distribution with 3 degrees of freedom, where X is uniformly distributed on [0, 1].

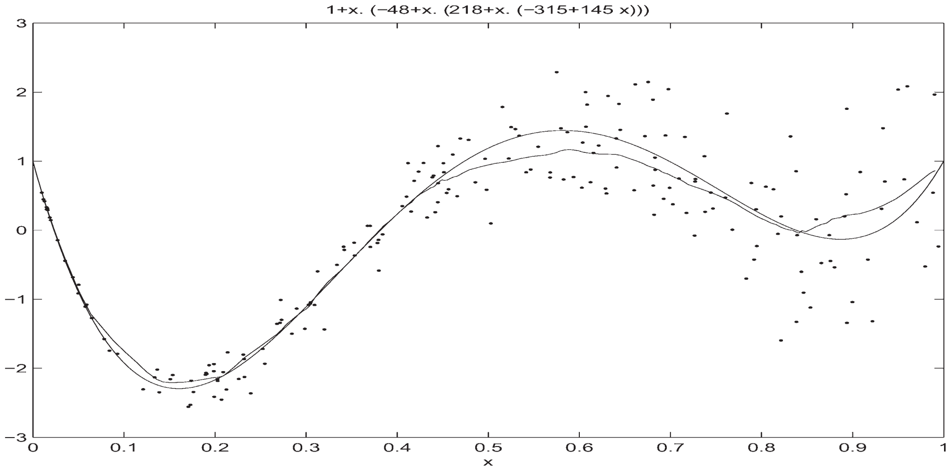

Figure 2. The smoother curve denotes the underlying regression function, while the other is the estimated curve derived from the local linear empirical likelihood fit under the first moment constraint. The data plotted, are sampled from $\varepsilon \sim N(0, \sigma(X)^2)$ with $\sigma(X)^2 = 1 + X^2$, where X is uniformly distributed on [0, 1].

captured in the survey the response variable is expressed as a score, on a log-weight scale, which combines information across species. The relationship between the catch score and the spatial coordinates (i.e., latitude, longitude and depth) was analysed in Bowman and Azzalini (1997, pp. 53–55 and p. 81) via ordinary nonparametric regression. Here we use our proposed method to analyse these data. We let $p = 1$, $\varepsilon(x) = y - \theta(x)$, and either

$$G(y, \theta(x)) = \varepsilon(x) \tag{3.1}$$

or

$$G(y, \theta(x)) = (\varepsilon(x), \varepsilon(x)^3)^{\mathrm{T}}. \tag{3.2}$$

In Figures 3, 4 and 5, we present the fitted results for the relationship between the catch score and the spatial coordinates (i.e., latitude, longitude and depth). In these figures, the solid, dashed and dotted curves stand for the results based on the least squares local linear fit, the local linear empirical likelihood fits with the restriction function in (3.1) and with the restriction function in (3.2), respectively. They show that there is little evidence of change with latitude, whereas there are marked changes in the catch score with longitude and depth. Note that in all figures, the corresponding curves differ from one another. It is natural to ask which one is better for interpreting the data. With this aim in mind, some goodness-of-fit tests for these restrictions are needed. The details can be found in Fan and Zhang (2000).

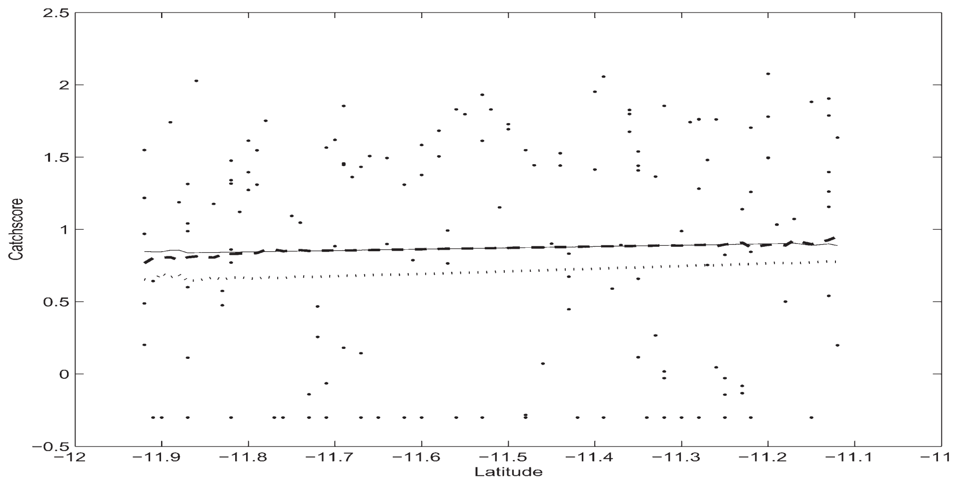

Figure 3. Relationship between catch score and latitude. The solid, dashed and dotted curves are, respectively, the local linear least squares fit, the local linear empirical likelihood fit under the first moment constraint, and the local linear empirical likelihood fit under the first and third moment constraints.

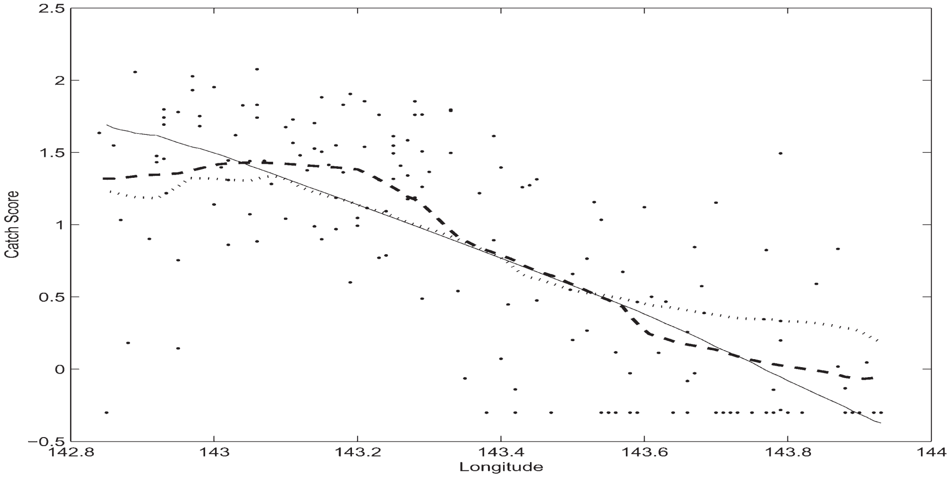

Figure 4. Relationship between catch score and longitude. The solid, dashed and dotted curves are, respectively, the local linear least squares fit, the local linear empirical likelihood fit under the first moment constraint, and the local linear empirical likelihood fit under the first and third moment constraints.

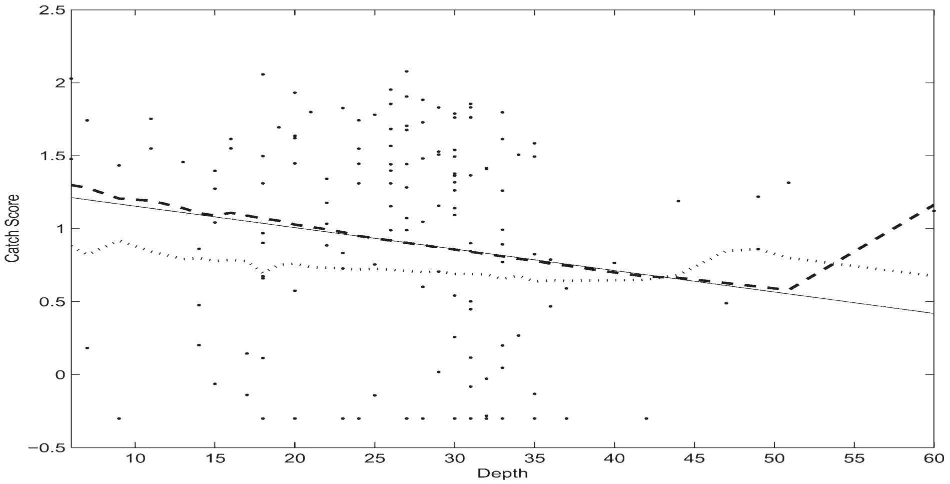

Figure 5. Relationship between catch score and depth. The solid, dashed and dotted curves are, respectively, the local linear least squares fit, the local linear empirical likelihood fit under the first moment constraint, and the local linear empirical likelihood fit under the first and third moment constraints.

4. Technical conditions

We begin with some notation. Suppose there exists $Z(y, x)$ (independent of h) such that

$$Z(y, x) \geqslant \sup_{\beta \in \Theta_0} \left\| G\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta \right) \right\| I(|x - x_0| \leqslant h).$$

Let

$$W_n(x_0, \beta) = \frac{1}{n} \sum_{i=1}^{n} K_h(x_i - x_0) \underline{G}(y_i, x_i, x_0, \beta) \underline{G}^{\mathrm{T}}(y_i, x_i, x_0, \beta).$$

To establish the consistency of $\widehat{\beta}$, we impose the following eight regularity conditions when $x_0 \in (0, 1)$:

Condition A1. *There exists a constant* c_0 *such that, for* $x \in [0, 1]$ *and* $x + \Delta \in [0, 1]$,

$$|f(x + \Delta) - f(x)| \leqslant c_0 |\Delta|.$$

Condition A2. *For some* $2 < \alpha_0 \leqslant \infty$,

$$\sup_{x \in [0,1]} \mathrm{E}\{Z(Y, X)^{\alpha_0} | X = x\} < \infty.$$

Here $\alpha_0 = \infty$ *means* $Z(Y, X)$ *is bounded by some constant.*

Condition A3. *For* $1 \leqslant j \leqslant k_0$, *as* $h = h_n \to 0$, *uniformly for* $\beta \in \Theta_0$ *and* $|t| \leqslant 1$,

$$\mathrm{E}\left\{ G_j^2\left(Y, \underline{X}\left(\frac{X - x_0}{h}\right)^{\mathrm{T}} \beta \right) | X = x_0 + th \right\} = O(1).$$

Condition A4. *There exists* $\psi_{h1}(y, x)$ *such that, for* $\beta_j \in \Theta_0$, $j = 1, 2$,

$$\mathrm{E}\psi_{h1}(Y, X) K\left(\frac{X - x_0}{h}\right) = O(1), \qquad \mathrm{E} Z(Y, X) \psi_{h1}(Y, X) K\left(\frac{X - x_0}{h}\right) = O(1),$$

and for $|x - x_0| \leqslant h$,

$$\left\| G\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta_1 \right) - G\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta_2 \right) \right\| \leqslant \psi_{h1}(y, x) \|\beta_1 - \beta_2\|.$$

Condition A5. *The function* θ *has a* $(p + 1)$th *continuous derivative and there exists* $\psi_{h2}(x)$ *such that*

$$\mathrm{E}\left\{K_h(X-x_0)\psi_{h2}(X)\left\|\underline{X}\left(\frac{X-x_0}{h}\right)\right\|\right\}=O(1),$$

$$\left\|\mathrm{E}\left\{\left[G\left(Y,\underline{X}\left(\frac{X-x_0}{h}\right)^{\mathrm{T}}\beta\right)-G(Y,\theta(X))\right]|X=x\right\}\right\|\leqslant\psi_{h2}(x)(\|\beta-\lambda_0\|+\|\beta(x,x_0)-\theta(x)\|),$$

for $\beta\in\Theta_0,\ |x-x_0|\leqslant h$, *where*

$$\beta(x,x_0)=\underline{X}\left(\frac{x-x_0}{h}\right)^{\mathrm{T}}\lambda_0.$$

Condition A6. *As* $n\to\infty,\ h=h_n\to 0$,

$$P\{W_n(x_0,\beta)>0,\beta\in\Theta_0\}\to 1,$$

where $W_n(x_0,\beta)>0$ *means that* $W_n(x_0,\beta)$ *is positive definite.*

Condition A7. *For* $1\leqslant k_1,j_1\leqslant k_0$, *as* $h=h_n\to 0$,

$$\mathrm{E}\{G^2_{k_1}(Y,\beta(X,x_0))G^2_{j_1}(Y,\beta(X,x_0))|X=x_0+th\}=O(1),$$

uniformly for $\beta\in\Theta_0$ *and* $|t|\leqslant 1$. *As* $\delta\to 0$ *and* $h=h_n\to 0$, *uniformly for* $\|\beta-\lambda_0\|\leqslant\delta$,

$$\left\|\mathrm{E}\left\{G\left(Y,\underline{X}\left(\frac{X-x_0}{h}\right)^{\mathrm{T}}\beta\right)G^{\mathrm{T}}\left(Y,\underline{X}\left(\frac{X-x_0}{h}\right)^{\mathrm{T}}\beta\right)|X=x_0+th\right\}-V_G(x_0)\right\|$$

is of order $o(1)$ *Moreover, we suppose* $V_G(x_0)$ *and* S *are positive definite.*

Condition A8. *For any fixed constant* $\rho>0$ *there exists a positive constant* $c(\rho)$ *such that, as* $h=h_n\to 0$,

$$\inf_{\|\beta-\lambda_0\|\geqslant\rho}\|\mathrm{E}K_h(X-x_0)\underline{G}(Y,X,x_0,\beta)\|\geqslant c(\rho).$$

In addition, there exists a fixed positive constant c *such that as* $\|\beta-\lambda_0\|+h\to 0,\beta\in\Theta_0$, $\|\mathrm{E}K_h(X-x_0)\underline{G}(Y,X,x_0,\beta)\|$ *is bounded below by* $c\|\beta-\lambda_0\|+O(h^{p+1})$, *where* c *and* $O(h^{p+1})$ *are independent of* β.

When G is not smooth, we need to replace Condition A4 by the following condition:

Condition A4′. *Set*

$$g(y, x, \beta) = K\left(\frac{x - x_0}{h}\right) G_{i_1}\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta\right)\left(\frac{x - x_0}{h}\right)^{j_1}, \mathcal{F}(i_1, j_1) = \{g(\cdot, \cdot, \beta): \beta \in \Theta_0\},$$

$$g_1(y, x, \beta) = K\left(\frac{x - x_0}{h}\right) G_{i_1}\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta\right)\left(\frac{x - x_0}{h}\right)^{j_1}$$

$$\times G_{k_1}\left(y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}} \beta\right)\left(\frac{x - x_0}{h}\right)^{s_1}, \mathcal{F}(i_1, j_1, k_1, s_1) = \{g_1(\cdot, \cdot, \beta): \beta \in \Theta_0\}.$$

There exist positive constants c_1, c_2, w_1 *and* w_2 *such that*

$$N(\delta, L_2(P_n), \mathcal{F}(i_1, j_1)) \leqslant c_1 \delta^{-w_1},$$

$$N(\delta, L_2(P_n), \mathcal{F}(i_1, j_1, k_1, s_1)) \leqslant c_2 \delta^{-w_2}$$

where P_n *is the empirical distribution of* (x_i, y_i), $i = 1, \ldots, n$, *and* $N(d, L_2(P_n), \Upsilon)$ *is called the covering number of* Υ, *which is defined in Pollard (1984).*

To obtain asymptotic normality, we need two additional conditions.

Condition B1. *For small* $\delta_0 > 0$, *there exists a function* $U_1(y, x)$ *satisfying*

$$\mathrm{E} K_h(X - x_0) U_1(Y, X) = O(1),$$

$$\mathrm{E} K_h(X - x_0) Z(Y, X) U_1(Y, X) = O(1),$$

$$\sup_{\|\beta - \lambda_0\| \leqslant \delta} \left\| \frac{\partial \underline{G}(y, x, x_0, \beta)}{\partial \beta} \right\| I(|x - x_0| \leqslant h) \leqslant U_1(y, x).$$

There exists a function $U_2(y, x)$ *satisfying*

$$\mathrm{E} K_h(X - x_0) U_2(Y, X) = O(1),$$

$$\left\| \frac{\partial \underline{G}(y, x, x_0, \beta)}{\partial \beta} - \frac{\partial \underline{G}(y, x, x_0, \lambda_0)}{\partial \beta} \right\| \leqslant U_2(y, x) \|\beta - \lambda_0\|.$$

Furthermore,

$$\mathrm{E} K_h(X - x_0) \left\| \frac{\partial \underline{G}(Y, X, x_0, \lambda_0)}{\partial \beta^{\mathrm{T}}} \right\|^2 = O(1),$$

$$\mathrm{E} K_h(X - x_0) \frac{\partial \underline{G}(Y, X, x_0, \lambda_0)}{\partial \beta^{\mathrm{T}}} = f(x_0) D_G(x_0) \otimes S + o(1).$$

Condition B2. *For some small* $\delta_0 > 0$, *there exists a function* $U_3(y, x)$ *such that*

$$\mathrm{E}K_h(X - x_0)U_3(Y, X) = O(1),$$

$$\sup_{\|\beta-\lambda_0\|\leqslant\delta_0}\left\|\frac{\partial^2 \underline{G}(y, x, x_0, \beta)}{\partial\beta\partial\beta^{\mathrm{T}}}\right\| \leqslant U_3(y, x).$$

For $U_1(y, x)$ *defined in Condition* B1,

$$\mathrm{E}K_h(X - x_0)U_1(Y, X)^2 = O(1).$$

When G is not smooth, we need to impose some conditions similar to those in Zhang and Gijbels (1999). The details are not pursued here.

For $x_0 = 0$ or 1, conditions similar to A1–A8, B1 and B2 can be imposed by restricting the value of t (or $(x - x_0)/h$)) to [0, 1] or [−1, 0] in the above.

Appendix: Proofs of theorems

We first introduce three lemmas which will be used in the proof of Theorem 1. Let $Z_i = Z(y_i, x_i)$, $1 \leqslant i \leqslant n$. Denote

$$A_{n1}(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\underline{G}(y_i, x_i, x_0, \beta)I(Z_i \leqslant n^{1/\alpha_1}), \tag{A.1}$$

$$A_n(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\underline{G}(y_i, x_i, x_0, \beta).$$

Lemma A.1. *Under Conditions* A1–A4, *for* $2 < \alpha_1 \leqslant \alpha_0$, *as* $h = h_n \to 0$ *and* $hn^{1-2/\alpha_1}/\log n \to \infty$, *there exists a sequence of constants* $(d_{n1})_{n=1}^{\infty}$, $0 < d_{n1} \to 0$, *such that, uniformly for* $\beta \in \Theta_0$,

$$A_n(x_0, \beta) = \mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \beta) + o_p(n^{-1/\alpha_1})d_{n1}, \tag{A.2}$$

$$A_{n1}(x_0, \beta) = \mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \beta) + o_p(n^{-1/\alpha_1})d_{n1}. \tag{A.3}$$

Furthermore, under Condition A5,

$$\mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \beta) = O(h^{p+1} + \|\beta - \lambda_0\|). \tag{A.4}$$

Proof. Without loss of generality, we assume $x_0 \in (0, 1)$. Write $A_n(x_0, \beta)$ as

$$A_n(x_0, \beta) = A_{n1}(x_0, \beta) + A_{n2}(x_0, \beta),$$

with

$$A_{n2}(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\underline{G}(y_i, x_i, x_0, \beta)I(Z_i > n^{1/\alpha_1}).$$

It follows from Conditions A1 and A2 that, for $2 < \alpha_1 \leqslant \alpha_0$,

$$\mathrm{E}\sup_{\beta\in\Theta_0} \|A_{n2}(x_0, \beta)\| \leqslant \mathrm{E}K_h(X - x_0)Z(Y, X)I(Z(Y, X) > n^{1/\alpha_1})\sqrt{p+1}f(x_0 + th)\mathrm{d}t$$
$$= o(n^{-1/\alpha_1}),$$

which implies

$$A_n(x_0, \beta) = \mathrm{E}K_h(X - x_0)G(Y, X, x_0, \beta) + \frac{1}{nh}\sum_{i=1}^{n}\{f_{ni}(\beta) - \mathrm{E}f_{ni}(\beta)\} + o_p(n^{-1/\alpha_1}), \tag{A.5}$$

where

$$f_{ni}(\beta) = K\left(\frac{x_i - x_0}{h}\right)\underline{G}(y_i, x_i, x_0, \beta)I(Z_i \leqslant n^{1/\alpha_1}).$$

Set

$$g(y, x, \beta) = n^{-1/\alpha_1}K\left(\frac{x - x_0}{h}\right)G_{i_1}\left(Y, \underline{X}\left(\frac{x - x_0}{h}\right)^{\mathrm{T}}\beta\right) \times \left(\frac{x - x_0}{h}\right)^{j_1} I(Z(y, x) \leqslant n^{1/\alpha_1}),$$

$$\mathcal{F}(i_1, j_1) = \{g(\cdot, \cdot, \beta): \beta \in \Theta_0\}.$$

Then, by Conditions A1 and A3, we have

$$\sup_{\beta\in\Theta_0} \mathrm{E}g^2(Y, X, \beta) = O(hn^{-2/\alpha_1}).$$

For $g(\beta_j) = g(y, x, \beta_j) \in \mathcal{F}(i_1, j_1)$, $j = 1, 2$, by Condition A4, we have

$$|g(\beta_1) - g(\beta_2)| \leqslant n^{-1/\alpha_1}K\left(\frac{x - x_0}{h}\right)\left(\frac{x - x_0}{h}\right)^{j_1}\psi_{h1}(y, x)\|\beta_1 - \beta_2\|.$$

Let $u_n = hn^{1-2/\alpha_1}$ and $d_{n1}^2 = (\log n/u_n)^{1/2}$. By Lemma 7.2 in Zhang and Gijbels (2003), there exist positive constants c_j, $1 \leqslant j \leqslant 4$, and w_0, such that, for any positive constant M_0,

$$P\left\{\sup_{\beta\in\Theta_0}\left|\frac{1}{nh}\sum_{i=1}^{n}[g(y_i, x_i, \beta) - \mathrm{E}g(Y, X, \beta)]\right| \geqslant M_0 n^{-1/\alpha_1}d_{n1}\right\}$$
$$\leqslant c_1(n^{1/\alpha_1}h^{-1}d_{n1})^{w_0}\exp\left\{-\frac{M_0^2 nh^2 n^{-4/\alpha_1}d_{n1}^2}{c_3 hn^{-2/\alpha_1}}\right\} + c_2(hn^{-2/\alpha_1})^{-w_0}\exp\{-c_4 nhn^{-2/\alpha_1}\}. \tag{A.6}$$

As $h = h_n \to 0$, $u_n/\log n \to \infty$, we have

$$\log u_n + \log n = o\{(u_n \log n)^{1/2}\} = o(u_n d_{n1}^2),$$

thus (A.6) tends to zero. This, together with (A.5), completes the proofs for (A.2) and (A.3).

Finally, (A.4) follows from Condition A5 and the equality

$$\|\mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \beta)\| = \left\|\mathrm{E}\left\{K_h(X - x_0)\Psi(\beta) \otimes \underline{X}\left(\frac{X - x_0}{h}\right)\right\}\right\|,$$

where

$$\Psi(\beta) = \mathrm{E}\left[G\left(Y, \underline{X}\left(\frac{X - x_0}{h}\right)^{\mathrm{T}}\beta\right) - G(Y, \theta(X))|X\right].$$

□

Write

$$W_{n1}(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\underline{G}(y_i, x_i, x_0, \beta)\underline{G}^{\mathrm{T}}(y_i, x_i, x_0, \beta)I(Z_i \leqslant n^{1/\alpha_1}).$$

Lemma A.2. *Under Conditions* A1 *and* A2, *as* $h = h_n \to 0$,

$$\sup_{\beta\in\Theta_0} \|W_{n1}(x_0, \beta)\| = O_p(1). \tag{A.7}$$

Under Conditions A1, A2, A4 *and* A7, *as* $h = h_n \to 0$ *and* $nh \to \infty$,

$$W_n(x_0, \beta) = f(x_0)V_G(x_0) \otimes S + o_p(1) \tag{A.8}$$

uniformly for $\|\beta - \lambda_0\| \leqslant \delta \to 0$.

Proof. Equation (A.7) follows from the fact that, under Conditions A1 and A2,

$$\mathrm{E}\sup_{\beta\in\Theta_0} \|W_{n1}(x_0, \beta)\| \leqslant (p + 1)\mathrm{E}K_h(X - x_0)Z(Y, X)^2 = O(1).$$

Note that, by Condition A4,

$$\begin{aligned}\|W_n(x_0, \beta) - W_n(x_0, \lambda_0)\| &\leqslant \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)Z_i\psi_{h1}(y_i, x_i)\|\beta - \lambda_0\| \\ &= O_p(1)\|\beta - \lambda_0\|.\end{aligned}$$

In order to prove (A.8), it suffices to show that

$$W_n(x_0, \lambda_0) = f(x_0)V_G(x_0) \otimes S + o_p(1). \tag{A.9}$$

To this end, we calculate the mean and covariance of $W_n(x_0, \lambda_0)$. It is easily seen that, under Conditions A1 and A7,

$$\mathrm{E}W_n(x_0, \beta) = f(x_0)V_G(x_0) \otimes S + o_p(1). \tag{A.10}$$

For $k = (p+1)(k_1 - 1) + k_2$ and $j = (p+1)(j_1 - 1) + j_2$ with $1 \leqslant k_1, j_1 \leqslant k_0$, $1 \leqslant k_2, j_2 \leqslant p+1$, we obtain that the variance of the (k, j)th element of $W_n(x_0, \lambda_0)$ is smaller than or equal to

$$\frac{1}{n}\mathrm{E}K_h(X - x_0)^2 G_{k_1}^2(Y, \beta(X, x_0))G_{k_2}^2(Y, \beta(X, x_0))\left(\frac{X - x_0}{h}\right)^{2(k_2+j_2-2)}$$

$$= O\left(\frac{f(x_0)}{nh}\right) = o(1),$$

by Condition A7. This, together with (A.10), leads to (A.9). □

Lemma A.3. *Under Conditions* A1–A7, *if both* $V_G(x_0)$ *and* S *are positive definite, then, for any* $2 < \alpha_1 \leqslant \alpha_0$, *as* $h = h_n \to 0$, $hn^{1-2/\alpha_1}/\log n \to \infty$, $h^{p+1}n^{1/\alpha_0} = o(1)$, *there exists a sequence of constants* $(d_{n1})_{n=1}^{\infty}$, $0 < d_{n1} \to 0$, *such that*

$$\alpha_n(x_0, \beta) = o_p(n^{-1/\alpha_1})d_{n1} + O(h^{p+1} + \|\beta - \lambda_0\|)$$

uniformly for $\|\beta - \lambda_0\| \leqslant O(n^{-1/\alpha_0})$.

Proof. Without loss of generality, we assume $x_0 \in (0, 1)$. By Condition A2, we have

$$\max_{1 \leqslant i \leqslant n} Z_i = o_p(n^{1/\alpha_0}).$$

It follows from Lemma A.1 that, as $h = h_n \to 0$, $hn^{1-2/\alpha_1}/\log n \to \infty$,

$$\|A_n(x_0, \beta)\| = o_p(n^{-1/\alpha_1})d_{n1} + O(h^{p+1} + \|\beta - \lambda_0\|) \tag{A.11}$$

for some $0 < d_{n1} \to 0$ and uniformly for $\|\beta - \lambda_0\| \leqslant O(n^{-1/\alpha_0})$. Thus, we have

$$\|A_n(x_0, \beta) \max_{1 \leqslant i \leqslant n} Z_i \sqrt{p+1}\| = o_p(1) \tag{A.12}$$

uniformly for $\|\beta - \lambda_0\| \leqslant O(n^{-1/\alpha_0})$. It follows from Lemma A.2 that there exists a positive constant c such that, as $h = h_n \to 0$, $hn \to \infty$ and $\delta \to 0$,

$$\sup\{\rho_n(x_0, \beta) \colon \|\beta - \lambda_0\| \leqslant \delta\} \geqslant c, \tag{A.13}$$

where $\rho_n(x_0, \beta)$ is the minimum eigenvalue of $W_n(x_0, \beta)$. Finally, by (A.11)–(A.13) and by using the technique of Owen (1988), we have

$$\|\alpha_n(x_0, \beta)\| \leqslant \frac{\|A_n(x_0, \beta)\|}{\rho_n(x_0, \beta) - \|A_n(x_0, \beta)\| \max_{1 \leqslant i \leqslant n} Z_i \sqrt{p+1}}$$

$$= O_p(\|A_n(x_0, \beta)\|)$$

$$= o_p(n^{-1/\alpha_1})d_{n1} + O(h^{p+1} + \|\beta - \lambda_0\|)$$

uniformly for $\|\beta - \lambda_0\| \leqslant O(n^{-1/\alpha_0})$. □

Proof of Theorem 1. Without loss of generality, we assume $x_0 \in (0, 1)$. We first establish some facts. Let $2 < \alpha_1 \leqslant \alpha_0$, $d_{n2} \geqslant 0$, $d_{n2}^2 = h^{p+1} n^{1/\alpha_1}$, and $d_n = \max\{d_{n1}, d_{n2}\}$, where d_{n1} is defined in Lemma A.3. Then, by Lemmas A.1 and A.3, we have

$$\alpha_n(x_0, \lambda_0) = o_p(n^{-1/\alpha_1}) d_n, \qquad A_n(x_0, \lambda_0) = o_p(n^{-1/\alpha_1}) d_n.$$

Our first fact is that

$$\begin{aligned} 0 &\geqslant -\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\log(1 + \alpha_n(x_0, \lambda_0)^{\mathrm{T}} \underline{G}(y_i, x_i, x_0, \lambda_0)) \\ &\geqslant -\alpha_n(x_0, \lambda_0)^{\mathrm{T}} A_n(x_0, \lambda_0) \\ &= -|o_p(n^{-2/\alpha_1})| d_n^2. \end{aligned} \tag{A.14}$$

Let $u_0 = u_0(\beta) \in \mathbb{R}^{k_0(p+1)}$, $\|u_0\| = 1$, satisfying

$$u_0 \|\mathrm{E} K_h(X - x_0) \underline{G}(Y, X, x_0, \beta)\| = \mathrm{E} K_h(X - x_0) \underline{G}(Y, X, x_0, \beta).$$

Denote

$$T_{n1} = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\log\left(1 + n^{-1/\alpha_1} d_n u_0^{\mathrm{T}} \underline{G}(y_i, x_i, x_0, \beta)\right) I(Z_i \leqslant n^{1/\alpha_1}).$$

Then we have

$$T_{n1} = n^{-1/\alpha_1} d_n u_0^{\mathrm{T}} A_{n1}(x_0, \beta) - n^{-2/\alpha_1} d_n^2 W_{n1}^*(x_0, \beta). \tag{A.15}$$

Here $A_{n1}(x_0, \beta)$ is as given in (A.1), and

$$W_{n1}^*(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{1}{2(1 + t_i)^2}(u_0^{\mathrm{T}} \underline{G}(y_i, x_i, x_0, \beta))^2 I(Z_i \leqslant n^{1/\alpha_1}),$$

where, for $1 \leqslant i \leqslant n$, t_i lies between 0 and $n^{-1/\alpha_1} d_n u_0^{\mathrm{T}} \underline{G}(y_i, x_i, x_0, \beta)$. When $\max_i Z_i \leqslant n^{1/\alpha_1}$, $\max_i |t_i| \leqslant \sqrt{p+1} d_n$ uniformly in β. This leads to

$$W_{n1}^*(x_0, \beta) \leqslant \frac{1}{2(1 - \sqrt{p+1} d_n)^2} u_0^{\mathrm{T}} W_{n1} u_0.$$

By Lemma A.2, we obtain that $W_{n1}^*(x_0, \beta)$ is uniformly bounded in $\beta \in \Theta_0$. This, in conjunction with (A.15) and Lemma A.1, leads to our second fact, namely that, uniformly for $\beta \in \Theta_0$,

$$T_{n1} = n^{-1/\alpha_1} d_n u_0^{\mathrm{T}} \mathrm{E}\{K_h(X - x_0) \underline{G}(Y, X, x_0, \beta)\} + O_p(n^{-1/\alpha_1}) d_n^2. \tag{A.16}$$

Furthermore,

$$P\Big(\max_i Z_{hi} > n^{1/\alpha_1}\Big) = o(1), \tag{A.17}$$

our third fact.

Denote

$$T_n(x_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\log(1 + \alpha_n(x_0, \beta)^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \beta)),$$

$$\Xi = \{\alpha\colon 1 + \alpha^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \beta) > 0,\ 1 \leqslant i \leqslant n\}.$$

Then, when

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\underline{G}(y_i, x_i, x_0, \beta)\underline{G}^{\mathrm{T}}(y_i, x_i, x_0, \beta) > 0,$$

we have our final fact,

$$-T_n(x_0, \beta) = \min_{\alpha\in\Xi} -\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\log(1 + \alpha^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \beta)). \tag{A.18}$$

Now combining the facts (A.14), (A.16), (A.17), (A.18) and Condition A8, we obtain that, for any fixed positive constant ρ, as $n \to \infty$,

$$\begin{aligned}
&P\left\{\sup_{\|\beta-\lambda_0\|\geqslant\rho}(-T_n(x_0, \beta)) > -T_n(x_0, \lambda_0)\right\}\\
&\qquad\leqslant P\left\{\sup_{\|\beta-\lambda_0\|\geqslant\rho}(-T_n(x_0, \beta)) > -|o_p(n^{-2/\alpha_1})|d_n^2\right\}\\
&\qquad\leqslant P\left\{\sup_{\|\beta-\lambda_0\|\geqslant\rho}(-T_{n1}) > -|o_p(n^{-2/\alpha_1})|d_n^2\right\} + P\{\max_i Z_i > n^{1/\alpha_1}\} + o(1)\\
&\qquad\leqslant P\left\{c\inf_{\|\beta-\lambda_0\|\geqslant\rho}\|\mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \beta)\| \leqslant |O_p(n^{-1/\alpha_1})d_n|\right\} + o(1),
\end{aligned}$$

which implies

$$\|\widehat{\beta} - \lambda_0\| = o_p(1). \tag{A.19}$$

Similarly, for any constants $0 < \rho_n \to 0$ and δ small enough, we have

$$P\{\delta \geqslant \|\widehat{\beta} - \lambda_0\| \geqslant \rho_n\} \leqslant P\{c\inf_{\delta\geqslant\|\beta-\lambda_0\|\geqslant\rho_n}\|\beta - \lambda_0\| + O(h^{p+1}) \leqslant |O_p(n^{-1/\alpha_1}d_n|\} + o(1). \tag{A.20}$$

It follows from (A.19) and (A.20) that

$$\begin{aligned}
\hat{\beta} - \lambda_0 &= O_p(n^{-1/\alpha_1})d_n + O(h^{p+1})\\
&= o_p(n^{-1/\alpha_1}).
\end{aligned}$$

Using Lemma 3.1 again, we obtain $\alpha_n(x_0, \widehat{\beta}) = o_p(n^{-1/\alpha_1})$. □

We now turn to some technical lemmas for the proof of Theorem 2. For this purpose, we first introduce some additional notation. Let

$$B_{n1}(\beta, \alpha) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{\underline{G}(y_i, x_i, x_0, \beta)}{1 + \alpha^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \beta)},$$

$$B_{n2}(\beta, \alpha) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{\alpha^{\mathrm{T}}\partial\underline{G}(y_i, x_i, x_0, \beta)/\partial\beta^{\mathrm{T}}}{1 + \alpha^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \beta)};$$

$$C_{n11}(\beta, \alpha) = \frac{\partial B_{n1}(\beta, \alpha)}{\partial\alpha^{\mathrm{T}}}, \qquad C_{n12}(\beta, \alpha) = \frac{\partial B_{n1}(\beta, \alpha)}{\partial\beta^{\mathrm{T}}},$$

$$C_{n21}(\beta, \alpha) = \frac{\partial B_{n2}(\beta, \alpha)}{\partial\alpha^{\mathrm{T}}}, \qquad C_{n22}(\beta, \alpha) = \frac{\partial B_{n2}(\beta, \alpha)}{\partial\beta^{\mathrm{T}}}.$$

Lemma A.4. *Under Conditions* A1, A2, A4 *and* A7, *as* $h = h_n \to 0$, *and* $nh \to \infty$, *for any random vectors* $\zeta_1 = \lambda_0 + o_p(1)$ *and* $\alpha_1 = o_p(1)$, *we have*

$$C_{n11}(\zeta_1, \alpha_1) = -f(x_0)V_G(x_0) \otimes S + o_p(1).$$

Proof. Note that

$$C_{n11}(\zeta_1, \alpha_1) = -W_n(x_0, \zeta_1) + R_{n11},$$

where

$$R_{n1} = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{\alpha_1^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \zeta_1)(2 + \alpha_1^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \zeta_1))}{(1 + \alpha_1^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \zeta_1))^2}$$

$$\times\, \underline{G}(y_i, x_i, x_0, \zeta_1)\underline{G}^{\mathrm{T}}(y_i, x_i, x_0, \zeta_1).$$

Note that, under Condition A2,

$$\max_i Z_i = O_p(n^{-1/\alpha_0}),$$

which implies

$$\max_i \|\alpha_1^{\mathrm{T}}\underline{G}(y_i, x_i, x_0, \zeta_1)\| = o_p(1)$$

by the assumption that $\alpha_1 = o_p(n^{-1/\alpha_0})$. Therefore,

$$\|R_{n11}\| \leqslant \frac{(p+1)|o_p(1)|(2 + |o_p(1)|)}{(1 - |o_p(1)|)^2}\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)Z_i^2$$

$$= o_p(1).$$

Now Lemma A.2 and the assumption that $\zeta_1 = \lambda_0 + o_p(1)$ complete the proof. □

Lemma A.5. *Under Conditions* A1*,* A2 *and* B1*, as* $h = h_n \to 0$ *and* $nh \to \infty$*, for any random vectors* $\zeta_1 = \lambda_0 + o_p(1)$ *and* $\alpha_1 = o_p(n^{-1/\alpha_0})$*, we have*

$$C_{n12}(\zeta_1, \alpha_1) = f(x_0)D_G(x_0) \otimes S + o_p(1),$$

$$C_{n21}(\zeta_1, \alpha_1)^{\mathrm{T}} = f(x_0)D_G(x_0) \otimes S + o_p(1).$$

Proof. We only need to consider $C_{n12}(\zeta_1, \alpha_1)$ because $C_{n12}(\zeta_1, \alpha_1) = C_{n21}(\zeta_1, \alpha_1)^{\mathrm{T}}$. For simplicity, we write $\underline{G}(y_i, x_i, x_0, \zeta_1)$ as $\underline{G}_i$. Note that

$$C_{n12}(\zeta_1, \alpha_1) = D_n(\zeta_1) + R_{n12},$$

where

$$D_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{\partial \underline{G}(y_i, x_i, x_0, \beta)}{\partial \beta},$$

$$R_{n12} = -\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)\frac{\alpha_1^{\mathrm{T}}\underline{G}_i}{1 + \alpha_1^{\mathrm{T}}\underline{G}_i}\frac{\partial \underline{G}_i}{\partial \beta^{\mathrm{T}}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\underline{G}_i\alpha_1^{\mathrm{T}}\partial \underline{G}_i/\partial \beta^{\mathrm{T}}}{(1 + \alpha_1^{\mathrm{T}}\underline{G}_i)^2}.$$

By Condition B1, we have, as $h = h_n \to 0$ and $nh \to \infty$,

$$\begin{aligned}\|D_n(\zeta_1) - D_n(\lambda_0)\| &\leqslant \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)U_2(y_i, x_i)\|\zeta_1 - \lambda_0\| \\ &= O_p(\|\zeta_1 - \lambda_0\|) = o_p(1) \end{aligned} \tag{A.21}$$

and

$$D_n(\lambda_0) = f(x_0)D_G(x_0) \otimes S + o_p(1). \tag{A.22}$$

Observe that under Condition A2 and the assumption that $\alpha_1 = o_p(n^{-1/\alpha_0})$, we have

$$\max_i |\alpha_1^{\mathrm{T}}\underline{G}_i| = o_p(1)$$

which, with Condition B1, implies

$$\begin{aligned}\|R_{n12}\| &\leqslant \frac{|o_p(1)|}{1 - |o_p(1)|}\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x_0)U_1(y_i, x_i) + \frac{|o_p(1)|}{(1 - |o_p(1)|)^2}\frac{1}{n}\sum_{i=1}^{n} Z_i U_1(y_i, x_i) \\ &= o_p(1). \end{aligned} \tag{A.23}$$

Now combining (A.21), (A.22) and (A.23), we obtain the desired result. □

Lemma A.6. *Under Conditions* A1*,* A2 *and* B2*, as* $h = h_n \to 0$*, for any random vectors* $\zeta_1 = \lambda_0 + o_p(1)$ *and* $\alpha_1 = o_p(n^{-1/\alpha_0})$*, we have*

$$C_{n22}(\zeta_1, \alpha_1) = o_p(1).$$

The proof of this lemma is similar to the proof of Lemma 4.5 and thus omitted.

Denote

$$C=\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \qquad C_{22.1}=C_{22}-C_{21}C_{11}^{-1}C_{12},$$

where

$$C_{11}=-f(x_0)V_G(x_0)\otimes S, \qquad C_{22}=0, \qquad C_{12}=C_{21}^{\mathrm{T}}=f(x_0)D_G(x_0)\otimes S.$$

Lemma A.7. *Suppose Conditions* A1, A4, A7, B1 *and* B2 *hold. Then, as* $h=h_n\to 0$, *we have*

$$nh\,\mathrm{var}(B_{n1}(\lambda_0,0))=f(x_0)V_G(x_0)\otimes S^*+\mathrm{o}(1).$$

If $\theta(x)$ *has a* $(p+1)$*th continuous derivative* $\theta^{(p+1)}(x)$, *then*

$$C_{22.1}^{-1}C_{21}C_{11}^{-1}\mathrm{E}B_{n1}(\lambda_0,0)=\mathrm{bias}^*(1+o(1))$$

with

$$\mathrm{bias}^*=h^{p+1}S^{-1}(\mu_{p+1},\ldots,\mu_{2p+1})^{\mathrm{T}}\theta^{(p+1)}(x_0)/(p+1)!.$$

In addition, if f *and* $\theta^{(p+1)}(x)$ *have continuous derivatives, then*

$$C_{22.1}^{-1}C_{21}C_{11}^{-1}\mathrm{E}B_{n1}(\lambda_0,0)=\mathrm{bias}(1+o(1)),$$

where bias *is defined in Section 2*.

Proof. Note that

$$\begin{aligned}
&\mathrm{E}B_{n1}(\lambda_0,0)\\
&=\mathrm{E}K_h(X-x_0)\mathrm{E}\left[G\left(Y,\theta(X)-(X-x_0)^{p+1}\frac{\theta^{(p+1)}(x_0)}{(p+1)!}-o_p(h^{p+1})\right)|X\right]\otimes\underline{X}\left(\frac{X-x_0}{h}\right)\\
&=-\mathrm{E}K_h(X-x_0)\mathrm{E}\left\{\frac{\partial G(Y,\theta(X))}{\partial\theta}|X\right\}(X-x_0)^{p+1}\frac{\theta^{(p+1)}(x_0)}{(p+1)!}\otimes\underline{X}\left(\frac{X-x_0}{h}\right)+o_p(h^{p+1})\\
&=-f(x_0)D_G(x_0)h^{p+1}\otimes(\mu_{p+1},\ldots,\mu_{2p+1})^{\mathrm{T}}\frac{\theta^{(p+1)}(x_0)}{(p+1)!}(1+o_p(1)).
\end{aligned}$$

Note that K is symmetric and the $(r+1)$th element of $S^{-1}(\mu_{p+1},\ldots,\mu_{2p+1})^{\mathrm{T}}$ is zero. To obtain the non-zero bias when $p-r$ is even, we expand $\mathrm{E}B_{n1}(\lambda_0,0)$ up to order h^{p+2}:

$$
\begin{aligned}
\mathrm{E}B_{n1}(\lambda_0, 0) = \mathrm{E}K_h(X - x_0)\mathrm{E}\Bigg[G\Bigg(& Y, \theta(X) - (X - x_0)^{p+1}\frac{\theta^{(p+1)}(x_0)}{(p+1)!} \\
& - (X - x_0)^{p+2}\frac{\theta^{(p+2)}(x_0)}{(p+2)!} - o_p(h^{p+2})\Bigg)\Big|X\Bigg] \otimes \underline{X}\left(\frac{X - x_0}{h}\right) \\
= - f(x_0)D_G(x_0) \otimes & \Bigg[(\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}\frac{\theta^{(p+1)}(x_0)}{(p+1)!}h^{p+1} \\
& + \Bigg[(\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}\frac{\theta^{(p+1)}(x_0)}{(p+1)!}\frac{f'(x_0)}{f(x_0)} \\
& + (\mu_{p+2}, \ldots, \mu_{2p+2})^{\mathrm{T}}\frac{\theta^{(p+2)}(x_0)}{(p+2)!}\Bigg]h^{p+2}\Bigg](1 + o_p(1)).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\mathrm{cov}(B_{n1}(\lambda_0, 0)) = \; & \frac{1}{n}[\mathrm{E}K_h^2(X - x_0)\underline{G}(Y, X, x_0, \lambda_0)\underline{G}^{\mathrm{T}}(Y, X, x_0, \lambda_0)) \\
& - \mathrm{E}K_h(X - x_0)\underline{G}(Y, X, x_0, \lambda_0) \\
& \times \mathrm{E}K_h(X - x_0)\underline{G}^{\mathrm{T}}(Y, X, x_0, \lambda_0)] \\
= \; & \frac{f(x_0)}{nh}\{V_G(x_0) \otimes S^* + O(h^{2p+3})\}.
\end{aligned}
$$

□

Proof of Theorem 2. Write $\hat{\alpha} = \alpha_n(x_0, \widehat{\beta})$. Then, applying Theorem 1, we have

$$\widehat{\beta} - \lambda_0 = o_p(1), \qquad \hat{\alpha} = o_p(1),$$

which, by the assumption, implies that as n is large, $\widehat{\beta}$ is an inner point of Θ_0. Since $\widehat{\beta}$ is the maximum estimator, we have

$$B_{n1}(\widehat{\beta}, \hat{\alpha}) = 0, \qquad B_{n2}(\widehat{\beta}, \hat{\alpha}) = 0.$$

By virtue of a Taylor expansion, these become

$$0 = B_{n1}(\lambda_0, 0) + C_{n11}(\zeta_1, \alpha_1)\hat{\alpha} + C_{n12}(\zeta_1, \alpha_1)(\widehat{\beta} - \lambda_0), \tag{A.24}$$

$$0 = B_{n2}(\lambda_0, 0) + C_{n21}(\zeta_1, \alpha_1)\hat{\alpha} + C_{n22}(\zeta_1, \alpha_1)(\widehat{\beta} - \lambda_0), \tag{A.25}$$

where the (ζ_j, α_j), $j = 1, 2$, are between $(\widehat{\beta}, \hat{\alpha})$ and $(\lambda_0, 0)$. Write

$$C_n = \begin{pmatrix} C_{n11}(\zeta_1, \alpha_1) & C_{n12}(\zeta_1, \alpha_1) \\ C_{n21}(\zeta_1, \alpha_1) & C_{n22}(\zeta_1, \alpha_1) \end{pmatrix}.$$

Applying Lemmas A.4–A.6, we have

$$C_n(\zeta_1, \alpha_1) = C + o_p(1),$$

which, in conjunction with (A.24) and (A.25), implies that

$$\begin{aligned}\begin{pmatrix} \hat{\alpha} \\ \widehat{\beta} - \lambda_0 \end{pmatrix} &= -C_n^{-1}\begin{pmatrix} B_{n1}(\lambda_0, 0) \\ 0 \end{pmatrix} \\ &= -C^{-1}\begin{pmatrix} B_{n1}(\lambda_0, 0) \\ 0 \end{pmatrix}(1 + o_p(1)).\end{aligned}$$

Combining this with (A.24) and (A.25), we have

$$\sqrt{nh}(\widehat{\beta} - \lambda_0) = C_{22.1}^{-1} C_{21} C_{11}^{-1} \sqrt{nh} B_{n1}(\lambda_0, 0)(1 + o_p(1)),$$

$$\sqrt{nh}\hat{\alpha} = -(C_{11}^{-1} + C_{11}^{-1} C_{12} C_{22.1}^{-1} C_{21} C_{11}^{-1})\sqrt{nh} B_{n1}(\lambda_0, 0)(1 + o_p(1)).$$

Finally, according to the Cramér–Wold device and Lemma A.7, to establish the asymptotic normality of $\widehat{\beta}$, it suffices to check Lyapunov's condition for any one-dimensional projection of $C_{22.1}^{-1} C_{21} C_{11}^{-1} \sqrt{nh} \times B_{n1}(\lambda_0, 0)$, which is straight forward.

We can prove the result for $\hat{\alpha}$ analogously. □

Acknowledgements

J. Zhang's research is partially supported by the National Natural Science Foundation of China and a grant from the EURANDOM research programme. The authors thank the referees, the Editor, an associate editor and Professors A. J. Lenstra and R. Gill for their comments that helped to improve the presentation.

References

Bhapkar, V.P. (1991) Sufficiency, ancillarity, and information in estimating functions. In V.P. Godambe (ed.), *Estimating Functions*, pp. 241–254. Oxford: Oxford University Press.

Bickel, P.J., Klaassen, A.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Bowman, A.W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis.* Oxford: Oxford University Press.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998) Local estimating equations. *J. Amer. Statist. Assoc.*, **93**, 214–227.

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fan, J. and Zhang, J. (2000) Sieve empirical likelihood ratio tests for nonparametric functions. EURANDOM Report 2000-046.

Hanfelt, J.J. and Liang, K. (1995) Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461–477.

Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, **25**, 2084–2102.

Owen, A. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Pollard, D. (1984). *Convergence of Stochastic Processes.* New York: Springer-Verlag.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Stone, C.J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

Tibshirani, R. and Hastie, T. (1987) Local likelihood estimation. *J. Amer. Statist. Assoc.*, **82**, 559–567.

Zhang, J. and Gijbels, I. (1999) Sieve empirical likelihood and extensions of the generalized least squares. Discussion Paper 9911, Institut de Statistique, Université Catholique de Louvain.

Zhang, J. and Gijbels, I. (2003) Sieve empirical likelihood and extensions of the generalized least squares. *Scand. J. Statist.*, **30**, 1–24.

Received October 2000 and revised February 2003