Journal of Multivariate Analysis **82**, 189–209 (2002)
doi:10.1006/jmva.2001.2013

Two-Step Likelihood Estimation Procedure for Varying-Coefficient Models

Zongwu Cai[1]

University of North Carolina at Charlotte
E-mail: zcai@uncc.edu

Received October 11, 1999; published online November 13, 2001

One of the advantages for the varying-coefficient model is to allow the coefficients to vary as smooth functions of other variables and the model can be estimated easily through a simple local quasi-likelihood method. This leads to a simple one-step estimation procedure. We show that such a one-step method cannot be optimal when some coefficient functions possess different degrees of smoothness. This drawback can be attenuated by using a two-step estimation approach. The asymptotic normality and mean-squared errors of the two-step method are obtained and it is also shown that the two-step estimation not only achieves the optimal convergent rate but also shares the same optimality as the ideal case where the other coefficient functions were known. A numerical study is carried out to illustrate the two-step method.

AMS 1991 subject classifications: 62G07; 62J12.

Key words and phrases: asymptotic normality; generalized linear model; local polynomial fitting; mean squared errors; optimal convergent rate; varying-coefficient model.

1. INTRODUCTION

In recent years, great progress has been made towards increasing the flexibility of generalized linear models. Of importance is the varying-coefficient (VC) model, which has gained considerable attention due to its various applications in many areas, such as biomedical study, finance, econometrics, environmental study, and political science. We refer to the articles by Hoover *et al.* (1998), Brumback and Rice (1998), and Fan and Zhang (1998, 2000) for details on novel applications of the VC model to longitudinal data; Chen and Tsay (1993), Cai *et al.* (2000), and Xia and Li (1999) for statistical inferences on the functional-coefficient nonlinear time series models; Cai and Tiwari (2000) for environmental study; Hong and Lee (1999) for applications in finance and econometrics; and Cederman and Penubarti (1999) for the study of international relationship conflict in political sciences. For more references, see Fan *et al.* (2000).

[1] This research was supported, in part, by the National Science Foundation Grant DMS 0072400 and funds provided by the University of North Carolina at Charlotte.

0047-259X/01 $35.00

The VC model, proposed by Hastie and Tibshirani (1993), has the form

$$g\{m(u, \mathbf{x})\} = \sum_{j=1}^{p} a_j(u)\, x_j, \tag{1}$$

where $g(\cdot)$ is a known link function, $m(u, \mathbf{x})$ is the mean function of the response variable Y given covariates $U = u$ and $\mathbf{X} = \mathbf{x}$. The appeal of the VC model is that by allowing the coefficients $\{a_j(\cdot)\}$ to depend on certain covariate U, the modeling bias can be significantly reduced and that the *curse of dimensionality* can be avoided. For the identity link function and the Gaussian errors, (1) was thoroughly studied by Cleveland *et al.* (1992), Hastie and Tibshirani (1993), Fan and Zhang (1999), Cai *et al.* (2000), Fan *et al.* (2000), among others. As pointed out by Fan and Zhang (1999), when the degrees of smoothness of $\{a_j(\cdot)\}$ are different, the local least square estimator is suboptimal under their asymptotic formulation. To achieve the optimal convergent rate, they proposed a two-step method, and they also derived the asymptotic bias and variance of the two-step estimator. Fan *et al.* (2000) explored the VC model by searching for the smoothing variable U as a linear combination of other covariates and they proposed some efficient algorithms to search for the unknown index and to estimate the coefficient functions. For the known link function and the exponential family, an intensive study on the VC model was carried out by Cai *et al.* (2000) that proposed using a local (quasi-) likelihood technique to estimate the coefficient functions and established the asymptotic normality of resulting estimators. They derived the standard error formulas for the estimated coefficient functions and proposed a goodness-of-fit test technique, based on a nonparametric maximum likelihood ratio type of test, to detect whether some coefficient functions are really varying or whether any covariates are statistically significant. In particular, they proposed an efficient modeling algorithm to make the VC model practically applicable.

When some coefficient functions process different degrees of smoothness, we show that the estimators based on the local (quasi-) likelihood method cannot achieve the optimal convergent rate. The intuition is clear: a smooth component demands a large bandwidth to reduce the variance, but a rough component needs a small bandwidth to reduce the bias. This problem cannot be solved by simply using a large bandwidth to estimating a smooth component only; see Fan and Zhang (1999). To attenuate this drawback, we use the two-step approach proposed in Fan and Zhang (1999). Assume without loss of generality that $a_p(\cdot)$ is smoother than $\{a_j(\cdot)\}_{j=1}^{p-1}$. In the first step, an initial estimate of $\{a_j(\cdot)\}_{j=1}^{p-1}$ is obtained. In the second step, to estimate $a_p(\cdot)$, the initial estimate of $\{a_j(\cdot)\}_{j=1}^{p-1}$ is used in lieu of $\{a_j(\cdot)\}_{j=1}^{p-1}$ in a local polynomial fitting. In such a way, we show that the two-step estimator not only achieves the optimal convergent rate but also shares the same optimality for the case that the coefficient functions $\{a_j(\cdot)\}_{j=1}^{p-1}$ were known. Moreover, it makes the implementations much easier. Finally, when $a_p(\cdot)$ is as smooth as the rest of functions, we show that the two-step estimator and the one-step estimator share the same asymptotic properties.

The article is organized as follows. We discuss in Section 2 estimation methods—the one-step procedure and the two-step approach. The asymptotic properties of the resulting estimators are established in Section 3. In Section 4, a small simulation is carried out to demonstrate the performance of the two-step estimator through two simulated examples. Finally, the technical proofs in Appendix conclude the article.

2. ESTIMATION METHODS

The primary goal is to estimate efficiently the coefficient functions $\{a_j(\cdot)\}$ by using a nonparametric method—local likelihood. The method is directly applicable to the situation that a conditional log-likelihood function $\ell(s, y)$ can be unspecified but the relationship between the mean and the variance function can be modeled via $\operatorname{var}(Y \mid U=u, \mathbf{X}=\mathbf{x}) = V\{m(u, \mathbf{x})\}$ for a known variance function $V(\cdot)$. In this case, a log-likelihood is replaced by a quasi-likelihood $Q(\mu, y)$, defined by $(\partial/\partial\mu)\, Q(\mu, y) = (y-\mu)/V(\mu)$, so that local likelihood becomes local quasi-likelihood; see Fan and Gijbels (1996, pp. 194), Carroll *et al.* (1997), Carroll *et al.* (1998), and Cai *et al.* (2000). For expositional purpose, we focus only on the canonical exponential family, so that the conditional log-likelihood function $\ell(s, y)$ is linear in y for fixed s.

For a given grid point u_0, the coefficient function $a_j(u)$ is approximated by $a_j + b_j(u-u_0)$ for u in a neighborhood of u_0. Note that a_j and b_j depend on the point u_0. Based on the random sample $\{(U_i, \mathbf{X}_i, Y_i)\}_{i=1}^{n}$, the weighted local log-likelihood is

$$\ell_n(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \ell\left[g^{-1}\left\{ \sum_{j=1}^{p} (a_j + b_j(U_i - u_0))\, X_{ij} \right\}, Y_i \right] K_h(U_i - u_0), \quad (2)$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, $h = h_n$ is a bandwidth, $\mathbf{a} = (a_1, \ldots, a_p)^T$, and $\mathbf{b} = (b_1, \ldots, b_p)^T$. Clearly, the local maximum likelihood estimates $\hat{\mathbf{a}}(u_0)$ and $\hat{\mathbf{b}}(u_0)$ are obtained by maximizing local log-likelihood function $\ell_n(\mathbf{a}, \mathbf{b})$. The components in $\hat{\mathbf{a}}(u_0)$ give the estimate of $a_1(u_0), \ldots, a_p(u_0)$. Note that the asymptotic properties of $\hat{\mathbf{a}}(\cdot)$ can be found in Cai *et al.* (2000). Clearly, this idea is simple and useful, but it is implicitly assumed that all functions $\{a_j(\cdot)\}$ possess the same degrees of smoothness. If some coefficient functions process different degrees of smoothness, it is shown in Theorem 1 that the estimator obtained by (2) is not optimal.

To formulate the foregoing idea, we assume that $a_p(\cdot)$ is smoother than the other functions and that it has a bounded fourth derivative, so that $a_p(u) \approx a_p + b_p(u-u_0) + c_p(u-u_0)^2 + d_p(u-u_0)^3$ for u in a neighborhood of u_0. This leads naturally to the locally weighted least-squares problem

$$\sum_{i=1}^{n} \ell[g^{-1}\{\mathbf{a}^T \mathbf{X}_i + \mathbf{b}^T(U_i-u_0)\,\mathbf{X}_i + c_p(U_i-u_0)^2\, X_{ip} + d_p(U_i-u_0)^3\, X_{ip}\}, Y_i]\, K_{h_1}(U_i-u_0). \tag{3}$$

Let $\hat{\mathbf{a}}_1$, $\hat{\mathbf{b}}_1$, $\hat{c}_{p,1}$ and $\hat{d}_{p,1}$ be the maximizers of (3). The resulting estimator $\hat{a}_{j,1}(u_0)$ of $a_j(u_0)$ is called one-step estimator. We show in Theorem 1 (below) that under some regularity conditions, the bias and variance of the one-step estimator are of the order h_1^2 and $(n\,h_1)^{-1}$. Therefore, the convergent rate for the asymptotic mean squared errors (MSE) is of the order $n^{-4/5}$ if $h_1 \sim n^{-1/5}$ but not optimal rate $n^{-8/9}$.

To achieve the optimal rate, we use the two-step procedure, described as follows. The first step is to get an initial estimate of $\{a_j(\cdot)\}_{j=1}^{p-1}$, denoted by $\{\hat{a}_{j,0}(u_0)\}_{j=1}^{p-1}$, which can be obtained by maximizing $\ell_n(\mathbf{a}, \mathbf{b})$ in (2) with the initial bandwidth h_0. Such an initial estimate, in general, is undersmoothed so that the bias is small. Then, in the second step, using the initial estimate $\{\hat{a}_{j,0}(\cdot)\}_{j=1}^{p-1}$ in lieu of $\{a_j(\cdot)\}_{j=1}^{p-1}$, we apply a local cubic fitting to estimate $a_p(u_0)$ by maximizing

$$\sum_{i=1}^{n} \ell[g^{-1}\{V_i + (a_p + b_p(U_i-u_0) + c_p(U_i-u_0)^2 + d_p(U_i-u_0)^3)\, X_{ip}\}, Y_i]\, K_{h_2}(U_i-u_0) \tag{4}$$

with respect to a_p, b_p, c_p and d_p, where $V_i = \sum_{j=1}^{p-1} \hat{a}_{j,0}(U_i)\, X_{ij}$ and h_2 is a bandwidth in the second step. In such a way, we obtain the two-step estimator $\hat{a}_{p,2}(u_0)$ of $a_p(u_0)$. We show in Theorem 2 that the two-step estimator can achieve the optimal convergent rate $n^{-8/9}$.

In support of the methodology, two simulated examples are used with sample size $n=400$ and $p=2$. Figure 1 depicts the estimation based on the one-step and two-step methods by using the optimal bandwidth for estimating $a_2(\cdot)$. For the two-step estimator, we optimize bandwidth h_2 for a given small bandwidth h_0; see details in Section 4.

In practice, we do not know in advance whether $a_p(\cdot)$ is really smoother than the rest of functions. The foregoing discussion reveals that the two-step procedure can lead to a significant gain when $a_p(\cdot)$ is smoother than the rest of functions. A question naturally arises is how the performance of

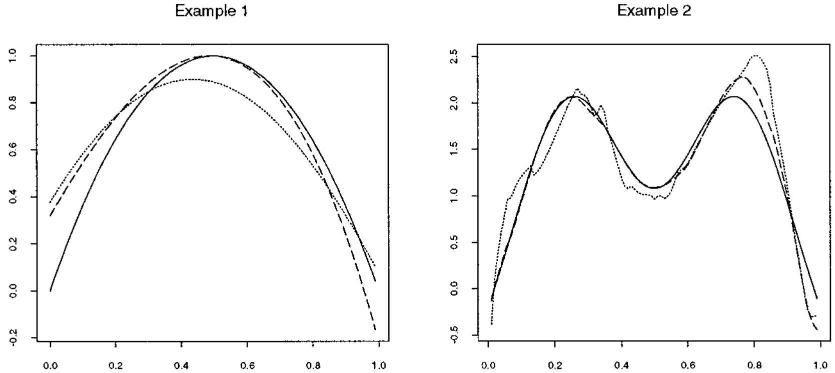

FIG. 1. Comparison of performance between the one-step and two-step estimators. Solid curve, true function; dotted curve, one-step procedure; dashed curve, two-step method.

the two-step procedure is when $a_p(\cdot)$ is as smooth as the rest of functions. To answer this question, we assume without loss of generality that $a_p(\cdot)$ has only continuous second derivative. For this case, the one-step estimator of $a_p(u_0)$ is the maximizer of (2) with bandwidth h_3, denoted by $\hat{a}_{p,3}(u_0)$. The two-step estimator of $a_p(u_0)$ is obtained by maximizing

$$\sum_{i=1}^{n} \ell[g^{-1}\{V_i + a_p X_{ip} + b_p(U_i - u_0)\, X_{ip}\}, Y_i]\, K_{h_3}(U_i - u_0)$$

with respect to a_p and b_p, denoted by $\hat{a}_{p,4}(u_0)$. Cai *et al.* (2000) showed that under conditions C1–C5 and C9 stated in Appendix, the respective asymptotic bias and variance of $\hat{a}_{p,3}(u_0)$ are

$$\frac{\mu_2 a_p''(u_0)}{2}\, h_3^2 \qquad \text{and} \qquad \frac{\mathbf{e}_{p,p}^T \boldsymbol{\Gamma}^{-1}(u_0)\, \mathbf{e}_{p,p}}{f(u_0)\, n\, h_3}, \tag{5}$$

where $\boldsymbol{\Gamma}(\cdot)$ is defined in (7) and $\mathbf{e}_{j,p}$ is the $p \times 1$ unit vector with 1 at the jth position. It is shown in Theorem 3 that the two-step estimator $\hat{a}_{p,4}(u_0)$ has the exact same asymptotic properties as $\hat{a}_{p,3}(u_0)$ provided that the initial bandwidth h_0 is small enough. Therefore, the two-step approach achieves the same convergent rate as the one-step procedure.

3. ASYMPTOTIC PROPERTIES

To study the asymptotic properties of $\hat{a}_{p,1}(u_0)$, $\hat{a}_{p,2}(u_0)$ and $\hat{a}_{p,4}(u_0)$, we introduce some notation. Denote by $\mu_k = \int u^k K(u)\, du$ and $\nu_k = \int u^k K^2(u)\, du$. Let $f(\cdot)$ be the marginal density of U and let

$$\rho(u, \mathbf{x}) = [g_1\{m(u, \mathbf{x})\}]^2\, V\{m(u, \mathbf{x})\}, \tag{6}$$

and

$$\mathbf{\Gamma}=\mathbf{\Gamma}(u)=E\{\rho(U,\mathbf{X})\,\mathbf{X}\,\mathbf{X}^T\,|\,U=u\}\equiv(\gamma_{ij}(u))_{p\times p}\equiv(\mathbf{\Gamma}_1,\ldots,\mathbf{\Gamma}_p),\qquad(7)$$

where $g_1(s)=g_0'(s)/g'(s)$ and $g_0(\cdot)$ is the canonical link function. Note that $\rho(u,\mathbf{x})=V\{m(u,\mathbf{x})\}$ for the canonical link function. The theorems are stated here but their proofs are relegated to the Appendix.

THEOREM 1. *Under conditions* C1–C6 *stated in the Appendix, we have*

$$\sqrt{n\,h_1}\left\{\hat{a}_{p,1}(u_0)-a_p(u_0)-\frac{h_1^2\mu_2}{2\gamma_{pp}(u_0)}\sum_{j=1}^{p-1}a_j''(u_0)\,\gamma_{jp}(u_0)+o_p(h_1^2)\right\}$$
$$\xrightarrow{\mathscr{D}}N(0,\sigma_{p,1}^2(u_0)),\qquad(8)$$

where

$$\sigma_{p,1}^2(u_0)=\frac{1}{f(u_0)}\left[\nu_0\mathbf{e}_{p,p}^T\mathbf{\Gamma}^{-1}(u_0)\,\mathbf{e}_{p,p}-\frac{2\mu_2\mu_4\nu_2-2\mu_2^2\mu_4\nu_0-\mu_2^2\nu_4+\mu_2^4\nu_0}{(\mu_4-\mu_2^2)^2\,\gamma_{pp}(u_0)}\right].$$

Remark 1. It is clear that the MSE of the one-step estimator $\hat{a}_{p,1}(u_0)$ is only of order $h_1^4+(n\,h_1)^{-1}$ which achieves the convergent rate $n^{-4/5}$ when bandwidth $h_1\sim n^{-1/5}$ is used. The bias expression above indicates clearly that the approximation errors of functions $\{a_j(\cdot)\}_{j=1}^{p-1}$ are transmitted to the bias of estimating $a_p(\cdot)$. Thus, the one-step estimator for $a_p(\cdot)$ inherits non-negligible approximation errors and is not optimal.

THEOREM 2. *Under conditions* C1–C8 *stated in the Appendix, then*

$$\sqrt{n\,h_2}\,\{\hat{a}_{p,2}(u_0)-a_p(u_0)-bias+o_p(h_2^4+h_0^2)\}\xrightarrow{\mathscr{D}}N(0,\sigma_{p,2}^2(u_0)),\qquad(9)$$

where the asymptotic bias is

$$bias=\frac{h_2^4a_p^{(4)}(u_0)(\mu_4^2-\mu_6\mu_2)}{24(\mu_4-\mu_2^2)}-\frac{h_0^2\mu_2}{2\gamma_{pp}(u_0)}\sum_{j=1}^{p-1}a_j''(u_0)\,\gamma_{jp}(u_0),$$

and the asymptotic variance is

$$\sigma_{p,2}^2(u_0)=\frac{\mu_4^2\nu_0-2\mu_2\mu_4\nu_2+\mu_2^2\nu_4}{f(u_0)(\mu_4-\mu_2^2)^2}\,\mathbf{e}_{p,p}^T\mathbf{\Gamma}^{-1}(u_0)\,\mathbf{e}_{p,p}.$$

By Theorem 2, the asymptotic variance of the two-step estimator is independent of the initial bandwidth as long as $n\,h_0^{\alpha}/\log h_0\to\infty$, where α is given in condition C8. Therefore, the initial bandwidth h_0 should be chosen as small as possible such that the constraint $n\,h_0^{\alpha}/\log h_0\to\infty$ is satisfied.

Especially, when the initial bandwidth h_0 is chosen as such that $h_0 = o(h_2^2)$, then the bias from the initial estimator becomes negligible and the bias expression for the two-step estimator becomes

$$\frac{a_p^{(4)}(u_0)(\mu_4^2 - \mu_6\mu_2)}{24(\mu_4 - \mu_2^2)} h_2^4 + o_p(h_2^4)$$

and the variance is

$$\frac{\mu_4^2 v_0 - 2\mu_2\mu_4 v_2 + \mu_2^2 v_4}{f(u_0)(\mu_4 - \mu_2^2)^2} \, \frac{\mathbf{e}_{p,p}^T \mathbf{\Gamma}^{-1}(u_0)\, \mathbf{e}_{p,p}}{n\, h_2} \{1 + o(1)\}.$$

Hence, by taking the optimal bandwidth $h_2 \sim n^{-1/9}$, the MSE of the two-step estimator achieves the optimal convergent rate $n^{-8/9}$.

As mentioned above, the choice of the initial bandwidth is not very sensitive to the two-step estimation as long as it is small enough so that the bias in the first step is not too large. This gives us a rule of thumb to choose h_0: Use the cross-validation or generalized cross-validation criterion (see, *e.g.*, Cai *et al.*, 2000) to select the bandwidth $\hat{h}_1$ for the one-step fitting. Then, use $h_0 = A_0\hat{h}_1$ ($A_0 = 1/2$, say, or smaller) or choose a very small h_0 as the initial bandwidth. Alternatively, as suggested by the referee, A_0 can be taken to be $A_0 = n^{-\alpha_1}$ with $\alpha_1 = 2/45$ or larger.

One of the advantages for the two-step procedure is that in the second step, the choice of bandwidth becomes really a univariate problem. Therefore, it may be easy to apply some univariate bandwidth selectors, such as cross-validation (Stone, 1974), pre-asymptotic substitution method (Fan and Gijbels, 1995), plug-in bandwidth selector (Rupert *et al.* 1995), and empirical bias method (Ruppert, 1997), to select the smoothing parameter in the second step. From the foregoing discussion, the initial bandwidth h_0 is not very crucial to the final estimate because for a wide range of bandwidth h_0, the two-step method achieves the optimal rate. This is another benefit of using the two-step procedure: the bandwidth selection problem is relatively easy.

Remark 2. Consider the ideal situation that $\{a_j(\cdot)\}_{j=1}^{p-1}$ are known. Then, one can simply run a local cubic fitting to estimate $a_p(\cdot)$. The resulting estimator has the following asymptotic bias

$$\frac{a_p^{(4)}(u_0)(\mu_4^2 - \mu_6\mu_2)}{24(\mu_4 - \mu_2^2)} h_2^4 + o_p(h_2^4)$$

and variance

$$\frac{\mu_4^2 v_0 - 2\mu_2\mu_4 v_2 + \mu_2^2 v_4}{f(u_0)(\mu_4 - \mu_2^2)^2} \, \frac{1}{n\, h_2 \gamma_{pp}(u_0)} \{1 + o(1)\}.$$

Therefore, the asymptotic bias for the ideal estimator is same as that for the two-step estimator and both have the same order of variance. In other words, the two-step estimator enjoys the same optimal convergent rate as the ideal estimator.

THEOREM 3. *Under conditions* C1–C5 *and* C7–C9 *stated in the Appendix, then*

$$\sqrt{n\,h_3}\,\{\hat{a}_{p,4}(u_0)-a_p(u_0)-bias+o_p(h_3^2+h_0^2)\}\xrightarrow{\mathscr{D}} N(0,\sigma_{p,4}^2(u_0))$$

with the asymptotic bias

$$bias=\frac{h_3^2 a_p''(u_0)\,\mu_2}{2}-\frac{h_0^2\mu_2}{2\gamma_{pp}(u_0)}\sum_{j=1}^{p-1}a_j''(u_0)\,\gamma_{jp}(u_0) \tag{10}$$

and variance

$$\sigma_{p,4}^2(u_0)=\frac{\nu_0\mathbf{e}_{p,p}^T\mathbf{\Gamma}^{-1}(u_0)\,\mathbf{e}_{p,p}}{f(u_0)}.$$

Remark 3. From Theorem 3, we conclude that, when all coefficient functions have the same degrees of smoothness, the one-step and two-step estimators share the same asymptotic variance, and, by taking the initial bandwidth $h_0=o(h_3)$, both have the same asymptotic bias by comparing (10) with (5). This leads to the conclusion that both estimators have the same performance. Hence the two-step method is always more reliable than the one-step method.

4. NUMERICAL EXAMPLES

We conduct a small simulation study on two models: a Bernoulli model and a Poisson model, to illustrate the performance of the two-step method and to compare it with the one-step procedure. The Epanechnikov kernel $K(u)=0.75(1-u^2)_+$ is employed. The covariates are taken as follows. (X_1, X_2) is generated from a bivariate normal with correlation coefficient $2^{-1/2}$, U is a uniform random variable on $[0, 1]$, and (X_1, X_2) is independent of U. For each example, with sample size $n=400$, the mean integrated squared errors (MISE) for estimating $a_2(\cdot)$ are recorded. For the one-step procedure, we plot MISE against h_1 and hence the optimal bandwidth $\hat{h}_1$ can be chosen. For the two-step procedure, first, we choose 20% of the optimal bandwidth $\hat{h}_1$ as initial bandwidth h_0 and then compute the MISE for the two-step estimator as a function of h_2. Therefore, the optimal

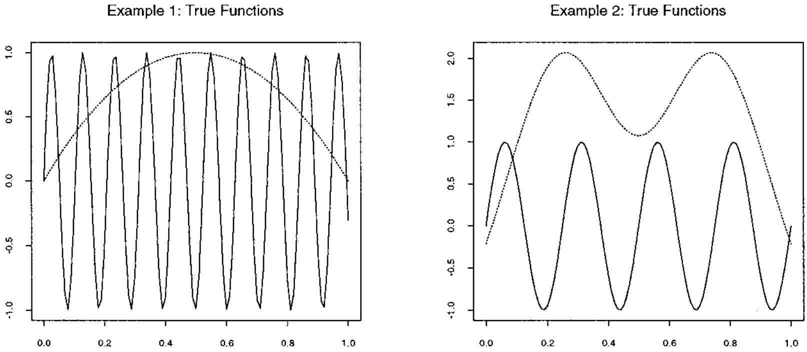

FIG. 2. Coefficient functions. Solid curve is $a_1(\cdot)$ and dotted curve is $a_2(\cdot)$.

bandwidth $\hat{h}_2$ for the two-step is determined. Figure 1 depicts the estimated curves of $a_2(\cdot)$ based on both the one-step and two-step methods and it shows clearly that the two-step approach outperforms the one-step method.

Example 1. The conditional probability of binary response variable $Y = 1$, given $U = u$, $X_1 = x_1$, and $X_2 = x_2$, is given by $\text{logit}\{P(Y = 1 \mid U = u, X_1 = x_1, X_2 = x_2)\} = a_1(u)\, x_1 + a_2(u)\, x_2$, where the coefficient functions $a_1(u) = \sin(60u)$ and $a_2(u) = 4u(1-u)$.

Example 2. The conditional distribution of Y, given that $U = u$, $X_1 = x_1$, and $X_2 = x_2$, is Poisson with mean $m(u, x_1, x_2)$ given by $\log\{m(u, x_1, x_2)\} = a_1(u)\, x_1 + a_2(u)\, x_2$, where the coefficient functions $a_1(u) = \sin(8\pi(u-0.5))$ and $a_2(u) = 3.5\exp(-(4\,u-1)^2) + 3.5\exp(-(4\,u-3)^2) - 1.5$.

Figure 2 displays the coefficient functions $a_1(\cdot)$ and $a_2(\cdot)$ for the above examples. Figure 3 represents the MISE as a function of bandwidth. The

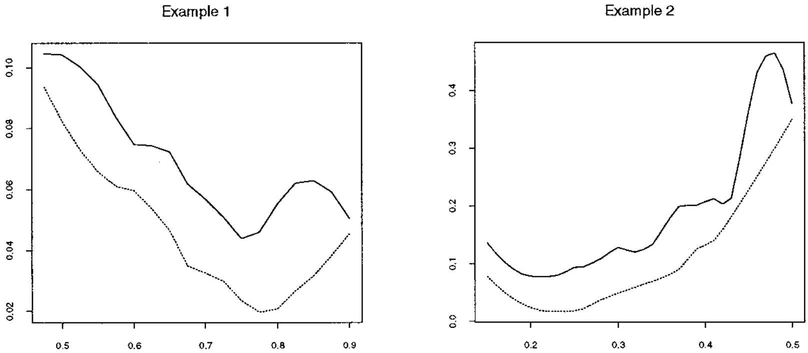

FIG. 3. MISE as a function of bandwidth. Solid curve, one-step procedure; dotted curve, two-step method.

MISE curve for the two-step method is always below that for the one-step approach for all examples. This is in line with the asymptotic theory that the two-step approach outperforms the one-step procedure if the initial bandwidth is correctly chosen.

APPENDIX: PROOFS

We first impose some regularity conditions. To this end, let $q_k(s, y) = (\partial^k/\partial s^k)\, \ell(s, y)$. Then, $q_k(s, y)$ is linear in y for fixed s,

$$q_1\{\eta(u, \mathbf{x}), m(u, \mathbf{x})\} = 0, \qquad \text{and} \qquad q_2\{\eta(u, \mathbf{x}), m(u, \mathbf{x})\} = -\rho(u, \mathbf{x}), \tag{A.1}$$

where $\rho(u, \mathbf{x})$ is defined in (6) and $\eta(u, \mathbf{x}) = g\{m(u, \mathbf{x})\}$. Note that we use the same notation as in Sections 2 and 3.

Conditions.

C1. The function $q_2(s, y) < 0$ for $s \in \Re$ and y in the range of the response variable.

C2. The functions $f(u)$, $V(m(u, \mathbf{x}))$, $V'(m(u, \mathbf{x}))$ and $g'''(m(u, \mathbf{x}))$ are continuous at the point $u = u_0$. Further, assume that $f(u_0) > 0$ and $\boldsymbol{\Gamma}(u_0) > 0$.

C3. $K(\cdot)$ is a symmetric and bounded density function with a bounded support, satisfying a Lipschitz condition.

C4. $E(|\mathbf{X}|^3 \mid U = u)$ is continuous at the point $u = u_0$.

C5. $E(Y^4 \mid U = u, \mathbf{X} = \mathbf{x})$ is bounded in a neighborhood of $u = u_0$.

C6. The function $a_p(\cdot)$ has a continuous fourth derivative in a neighborhood of u_0. Further, assume that $a_j''(\cdot)$ is continuous in a neighborhood of u_0 for $j = 1, ..., p-1$.

C7. $E\, |q_2(\eta(U, \mathbf{X}), Y)\, \mathbf{X}\mathbf{X}^T U^2|^{\gamma} < \infty$ for some $\gamma > 2$.

C8. $h_0 \to 0$ in such a way that $n\, h_0^{\alpha}/\log h_0 \to \infty$ for any $\alpha > \gamma/(\gamma-2)$ with γ given in condition C7.

C9. Assume that $a_j''(\cdot)$ is continuous in a neighborhood of u_0 for $j = 1, ..., p$.

Remark 4. Condition C1 guarantees that the sequence of maximizers of (3) and (4) lies in a compact set. Note that condition C2 implies that $q_1(\cdot, \cdot)$, $q_2(\cdot, \cdot)$, $q_3(\cdot, \cdot)$, $\rho'(\cdot, \cdot)$ and $m'(\cdot, \cdot)$ are continuous.

First, we present the detailed proof of Theorem 2, noting that a proof of Theorem 3 is similar but simpler, and then, we only give the outline of the proof of Theorem 1 since it is close to that of Theorem 2. To prove Theorem 2, we need the following lemma, due to Mack and Silverman (1982).

LEMMA 1. *Let* $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ *be iid random vectors, where the* Y*'s are scalar random variables. Assume further that* $E\,|Y|^s < \infty$ *and* $\sup_{\mathbf{x}} \int |y|^s f(\mathbf{x}, y)\,dy < \infty$, *where* $f(\cdot, \cdot)$ *denotes the joint density of* $(\mathbf{X}, Y)$. *Let* $K(\cdot)$ *be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_{x \in D} \left| \frac{1}{n} \sum_{i=1}^{n} \{Y_i K_h(\mathbf{X}_i - x) - E[Y_i K_h(\mathbf{X}_i - x)]\} \right| = O_p((n\,h/\log h)^{-1/2})$$

provided that $n^{2\varepsilon - 1}\,h \to \infty$ *for some* $\varepsilon < 1 - s^{-1}$.

Proof of Theorem 2. Let $\beta = (a_p, b_p, c_p, d_p)^T$,

$$\bar{\eta}(u) = a_p(u_0) + a_p'(u_0)(u - u_0) + \frac{a_p''(u_0)}{2}(u - u_0)^2 + \frac{a_p'''(u_0)}{6}(u - u_0)^3,$$

and

$$\beta^* = \gamma_n^{-1}(\beta_1 - a_p(u_0), h_2(\beta_2 - a_p'(u_0)), h_2^2(\beta_3 - a_p''(u_0)/2), h_2^3(\beta_4 - a_p'''(u_0)/6))^T,$$

where $\gamma_n = (n\,h_2)^{-1/2}$. It can be easily seen that

$$\sum_{j=1}^{p-1} \hat{a}_{j,0}(u_0)\,X_{ij} + \{a_p + b_p(U_i - u_0) + c_p(U_i - u_0)^2 + d_p(U_i - u_0)^3\}\,X_{ip}$$
$$= V_i + \bar{\eta}(U_i)\,X_{ip} + \gamma_n \beta^{*T} \mathbf{Z}_i X_{ip},$$

where $V_i = \sum_{j=1}^{p-1} \hat{a}_{j,0}(U_i)\,X_{ij}$ and $\mathbf{Z}_i = (1, (U_i - u_0)/h_2, (U_i - u_0)^2/h_2^2, (U_i - u_0)^3/h_2^3)^T$. Let

$$\hat{\beta}^* = \gamma_n^{-1}(\hat{\beta}_1 - a_p(u_0), h_2(\hat{\beta}_2 - a_p'(u_0)), h_2^2(\hat{\beta}_3 - a_p''(u_0)/2), h_2^3(\hat{\beta}_4 - a_p'''(u_0)/6))^T.$$

Then, $\hat{\beta}^*$ maximizes the normalized log-likelihood function

$$\ell_n^*(\beta^*) \equiv \sum_{i=1}^{n} (\ell[g^{-1}\{\tilde{\eta}_i + \gamma_n \beta^{*T} \mathbf{Z}_i X_{ip}\}, Y_i] - \ell[g^{-1}\{\tilde{\eta}_i\}, Y_i])\,K\{(U_i - u_0)/h_2\},$$

where $\tilde{\eta}_i = V_i + \bar{\eta}_i(U_i)\,X_{ip}$. We remark that condition C1 implies by the convexity lemma (see, e.g., Fan and Gijbels, 1996, p. 209) that $\ell_n^*(\cdot)$ is concave in β^*. Using the Taylor expansion of $\ell\{g^{-1}(\cdot), y\}$, we have

$$\ell_n^*(\beta^*) = \mathbf{W}_n^T \beta^* + \frac{1}{2} \beta^{*T} \Delta_n \beta^* + \frac{\gamma_n^3}{6} \sum_{i=1}^{n} q_3\{\eta_i^*, Y_i\}(\beta^{*T}\mathbf{Z}_i)^3\,X_{ip}^3 K\{(U_i - u_0)/h_2\}, \tag{A.2}$$

where

$$\mathbf{W}_n = \gamma_n \sum_{i=1}^{n} q_1\{\tilde{\eta}_i, Y_i\} \, \mathbf{Z}_i X_{ip} K\{(U_i - u_0)/h_2\},$$

$$\Delta_n = \frac{\gamma_n^2}{2} \sum_{i=1}^{n} q_2\{\tilde{\eta}_i, Y_i\} \, \mathbf{Z}_i \mathbf{Z}_i^T X_{ip}^2 K\{(U_i - u_0)/h_2\},$$

and η_i^* is between $\tilde{\eta}_i$ and $\tilde{\eta}_i + \gamma_n \beta^{*T} \, \mathbf{Z}_i X_{ip}$. It can be easily seen that

$$\tilde{\eta}_i = \eta(U_i, \mathbf{X}_i) + s_{i,n},$$

where

$$s_{i,n} = \sum_{j=1}^{p-1} [\hat{a}_{j,0}(U_i) - a_j(U_i)] \, X_{ij} + [\bar{\eta}(U_i) - a_p(U_i)] \, X_{ip}.$$

By Theorem 1 of Cai *et al.* (2000), Lemma 1 and the condition that $a_p(\cdot)$ has the bounded fourth derivative, we have

$$s_{i,n} = t_{i,n} + o_p(h_2^4 + h_0^2), \tag{A.3}$$

where

$$t_{i,n} = \frac{1}{n\,h_0} \sum_{j=1}^{p-1} \frac{\mathbf{e}_{j,p}^T \boldsymbol{\Gamma}^{-1}(U_i) \, X_{ij}}{f(U_i)} \sum_{k=1}^{n} q_1\{\eta(U_k, \mathbf{X}_k), Y_k\} \, \mathbf{X}_k K\{(U_k - U_i)/h_0\}$$
$$+ \frac{h_0^2 \mu_2}{2} \sum_{j=1}^{p-1} a_j''(U_i) \, X_{ij} - \frac{a_p^{(4)}(u_0)}{4!} (U_i - u_0)^4 \, X_{ip},$$

and $o_p(\cdot)$ in (A.3) holds uniformly in i such that U_i falls in the neighborhood of u_0 by the continuity assumptions. Then, $\ell_n^*(\beta^*)$ in (A.2) becomes

$$\mathbf{W}_n^T \beta^* + \frac{1}{2} \beta^{*T} \Delta_n \beta^* + \frac{\gamma_n^3}{6} \sum_{i=1}^{n} q_3\{\eta(U_i, \mathbf{X}_i), Y_i\} (\beta^{*T} \mathbf{Z}_i X_{ip})^3 \, K\{(U_i - u_0)/h_2\}$$
$$+ o_p(1) \tag{A.4}$$

by the continuity assumption. Since $K(\cdot)$ is bounded, $q_3(\cdot, \cdot)$ is linear in Y and $E(|Y| \,|\, U, \mathbf{X}) < \infty$, then the expected value of the absolute value of the last term in (A.4) is bounded by

$$O(n\gamma_n^3 E \, |q_3(\eta(U, \mathbf{X}), Y) \, X_p^3 K\{(U - u_0)/h_2\}|) = O(\gamma_n) \tag{A.5}$$

by condition C4. Therefore, the last term in (A.4) is of order $O_p(\gamma_n)$. It follows from the Taylor expansion, (A.3), and the same arguments in the proof of (A.5) that

$$\mathbf{\Delta}_n = \mathbf{\Delta}_n^* + o_p(1),$$

where

$$\mathbf{\Delta}_n^* = \frac{\gamma_n^2}{2} \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\ \mathbf{Z}_i \mathbf{Z}_i^T X_{ip}^2 K\{(U_i - u_0)/h_2\}.$$

By the fact that $q_2(s, y)$ is linear in y and using the second result of (A.1), we obtain

$$E(\mathbf{\Delta}_n^*) = h_2^{-1} E[q_2\{\eta(U, \mathbf{X}), m(U, \mathbf{X})\}\ K\{(U - u_0)/h_2\}\ \mathbf{Z}\mathbf{Z}^T X_p^2] \to -\mathbf{\Delta},$$

where $\Delta = f(u_0)\, \gamma_{pp} \mathbf{\Omega}$, and

$$\mathbf{\Omega} = \begin{pmatrix} 1 & 0 & \mu_2 & 0 \\ 0 & \mu_2 & 0 & \mu_4 \\ \mu_2 & 0 & \mu_4 & 0 \\ 0 & \mu_4 & 0 & \mu_6 \end{pmatrix}.$$

Similar arguments show that $\mathrm{var}(\mathbf{\Delta}_n^*) = O\{(n\, h_2)^{-1}\}$. Therefore,

$$\mathbf{\Delta}_n = -\mathbf{\Delta} + o_p(1).$$

This, in conjunction with (A.4) and (A.5), implies that

$$\ell_n^*(\beta^*) = \mathbf{W}_n^T \beta^* - \tfrac{1}{2}\, \beta^{*T}(\mathbf{\Delta} + o_p(1))\, \beta^* + o_p(1).$$

Using the quadratic approximation lemma (see, e.g., Fan and Gijbels, 1996, p. 210), we obtain

$$\hat{\beta}^* = \mathbf{\Delta}^{-1} \mathbf{W}_n + o_p(1), \tag{A.6}$$

if $\mathbf{W}_n$ is a sequence of stochastically bounded random vectors. The asymptotic normality of $\hat{\beta}^*$ follows from that of $\mathbf{W}_n$. Hence, it suffices to establish the asymptotic normality of $\mathbf{W}_n$. By the Taylor expansion of $q_1(s, y)$ with respect to s, we have

$$\begin{aligned}\mathbf{W}_n = {} & \gamma_n \sum_{i=1}^{n} q_1\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip} K\{(U_i - u_0)/h_2\} \\ & + \gamma_n \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip} t_{i,n} K\{(U_i - u_0)/h_2\} \\ & + \gamma_n \sum_{i=1}^{n} q_3\{\eta_i^{**}, Y_i\}\, \mathbf{Z}_i X_{ip} s_{i,n}^2 K\{(U_i - u_0)/h_2\} \\ & + \gamma_n \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip} [s_{i,n} - t_{i,n}]\, K\{(U_i - u_0)/h_2\}, \qquad \text{(A.7)}\end{aligned}$$

where η_i^{**} is between $\eta(U_i, \mathbf{X}_i)$ and $\eta(U_i, \mathbf{X}_i) + s_{i,n}$. Similar to the proof of (A.5), the third term in (A.7) is $o_p(1)$. It is easy to see that

$$\gamma_n \sum_{i=1}^{n} |q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip}|\, K\{(U_i - u_0)/h_2\} = O_p(\gamma_n^{-1}),$$

which, in conjunction with (A.3), implies that the last term in (A.7) becomes $o_p(\gamma_n^{-1}(h_2^4 + h_0^2))$. Therefore,

$$\begin{aligned}\mathbf{W}_n = {} & \gamma_n \sum_{i=1}^{n} q_1\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip} K\{(U_i - u_0)/h_2\} \\ & + \gamma_n \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i X_{ip} t_{i,n} K\left(\frac{U_i - u_0}{h_2}\right) + o_p(1 + \gamma_n^{-1}(h_2^4 + h_0^2)). \\ & \qquad \text{(A.8)}\end{aligned}$$

Let

$$\mathbf{W}_{n,2}^* = \frac{\gamma_n h_0^2 \mu_2}{2} \sum_{j=1}^{p-1} \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i a_j''(U_i)\, X_{ij} X_{ip} K\{(U_i - u_0)/h_2\}$$

and

$$\mathbf{W}_{n,3}^* = -\frac{\gamma_n a_p^{(4)}(u_0)}{4!} \sum_{i=1}^{n} q_2\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i (U_i - u_0)^4\, X_{ip}^2 K\{(U_i - u_0)/h_2\}.$$

By Lemma 1 and tedious calculations, the second term in (A.8) becomes

$$\begin{aligned}& \gamma_n \sum_{i=1}^{n} \left[\sum_{j=1}^{p-1} \mathbf{e}_{j,p}^T \boldsymbol{\Gamma}^{-1}(U_i)\, \mathbf{X}_i \gamma_{jp}(U_i) \right] q_1\{\eta(U_i, \mathbf{X}_i), Y_i\}\, \mathbf{Z}_i K\{(U_i - u_0)/h_2\} \\ & \quad + \mathbf{W}_{n,2}^* + \mathbf{W}_{n,3}^* + o_p(1).\end{aligned}$$

Therefore, (A.8) becomes

$$\mathbf{W}_n = \mathbf{W}_{n,1}^* + \mathbf{W}_{n,2}^* + \mathbf{W}_{n,3}^* + o_p(1+\gamma_n^{-1}(h_2^4+h_0^2)) = \mathbf{W}_n^* + o_p(1+\gamma_n^{-1}(h_2^4+h_0^2)), \tag{A.9}$$

where

$$\mathbf{W}_{n,1}^* = \gamma_n \sum_{i=1}^{n} \left[\sum_{j=1}^{p-1} \mathbf{e}_{j,p}^T \mathbf{\Gamma}^{-1}(U_i)\, \mathbf{X}_i \gamma_{jp}(U_i) + X_{ip} \right] q_1\{\eta(U_i, \mathbf{X}_i), Y_i\} \times \mathbf{Z}_i K\{(U_i - u_0)/h_2\}.$$

Then, by (A.6)–(A.9),

$$\hat{\beta}^* = \mathbf{\Delta}^{-1}\mathbf{W}_n^* + o_p(1+\gamma_n^{-1}(h_2^4+h_0^2)). \tag{A.10}$$

Note that $\mathbf{W}_n^*$ is a sum of iid random vectors. In order to establish its asymptotic normality, it suffices to compute the mean and covariance matrix of $\mathbf{W}_n^*$ by Lyapounov condition. To this effect, we have that $E(\mathbf{W}_{n,1}^*) = \mathbf{0}$ by the first result in (A.1), and

$$\begin{aligned} E(\mathbf{W}_{n,2}^*) &= \frac{n\gamma_n h_0^2 \mu_2}{2} \sum_{j=1}^{p-1} E[q_2\{\eta(U,\mathbf{X}), m(U,\mathbf{X})\} \\ &\quad \times \mathbf{Z} a_j''(U)\, X_j X_p K\{(U-u_0)/h_2\}] \\ &= -\frac{h_0^2\mu_2}{2\gamma_n} f(u_0) \sum_{j=1}^{p-1} a_j''(u_0)\, \gamma_{jp} \begin{pmatrix} 1 \\ 0 \\ \mu_2 \\ 0 \end{pmatrix} \{1+o(1)\}. \end{aligned} \tag{A.11}$$

Likewise,

$$\begin{aligned} E(\mathbf{W}_{n,3}^*) &= -\frac{n\gamma_n a_p^{(4)}(u_0)}{24} E[q_2\{\eta(U,\mathbf{X}), m(U,\mathbf{X})\} \\ &\quad \times \mathbf{Z}(U-u_0)^4\, X_p^2 K\{(U-u_0)/h_2\}] \\ &= \frac{h_2^4 a_p^{(4)}(u_0)}{24\gamma_n} f(u_0)\, \gamma_{pp} \begin{pmatrix} \mu_4 \\ 0 \\ \mu_6 \\ 0 \end{pmatrix} \{1+o(1)\}. \end{aligned} \tag{A.12}$$

Similarly,

$$\operatorname{var}(\mathbf{W}_{n,1}^*) = n\gamma_n^2 \operatorname{var}\left[\left\{\sum_{j=1}^{p-1} \mathbf{e}_{j,p}^T \mathbf{\Gamma}^{-1}(U)\,\mathbf{X}\gamma_{jp}(U) + X_p\right\} q_1\{\eta(U,\mathbf{X}), Y)\}\right.$$
$$\left.\times\, \mathbf{Z}K\{(U-u_0)/h_2\}\right]$$
$$= f(u_0)\,\gamma_{pp}^2 \mathbf{e}_{p,p}^T \mathbf{\Gamma}^{-1}(u_0)\,\mathbf{e}_{p,p}\mathbf{\Psi}\{1+o(1)\} \equiv \mathbf{\Lambda} + o(1), \qquad \text{(A.13)}$$

where

$$\mathbf{\Psi} = \begin{pmatrix} \nu_0 & 0 & \nu_2 & 0 \\ 0 & \nu_2 & 0 & \nu_4 \\ \nu_2 & 0 & \nu_4 & 0 \\ 0 & \nu_4 & 0 & \nu_6 \end{pmatrix},$$

$$\operatorname{var}(\mathbf{W}_{n,2}^*) = o(1), \qquad \text{and} \qquad \operatorname{var}(\mathbf{W}_{n,3}^*) = o(1). \qquad \text{(A.14)}$$

Therefore,

$$\operatorname{var}(\mathbf{W}_n^*) = \mathbf{\Lambda} + o(1). \qquad \text{(A.15)}$$

We now use the Cramér–Wold device to derive the asymptotic normality of $\mathbf{W}_n^*$. For any unit vector $\mathbf{d} \in \Re^4$, if

$$\{\mathbf{d}^T \operatorname{var}(\mathbf{W}_n^*)\,\mathbf{d}\}^{-1/2}\{\mathbf{d}^T\,\mathbf{W}_n^* - \mathbf{d}^T\,E(\mathbf{W}_n^*)\} \xrightarrow{\mathscr{D}} N(0, 1), \qquad \text{(A.16)}$$

then

$$\{\operatorname{var}(\mathbf{W}_n^*)\}^{-1/2}\,(\mathbf{W}_n^* - E(\mathbf{W}_n^*)) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \mathbf{I}_4). \qquad \text{(A.17)}$$

Combining (A.10)–(A.17), we obtain

$$\hat{\beta}^* + \frac{h_0^2\mu_2}{2\gamma_n} f(u_0)\,\mathbf{\Delta}^{-1}\sum_{j=1}^{p-1} a_j''(u_0)\,\gamma_{jp}\begin{pmatrix} 1 \\ 0 \\ \mu_2 \\ 0 \end{pmatrix}\{1+o(1)\}$$
$$-\frac{h_2^4 a_p^{(4)}(u_0)}{24\gamma_n} f(u_0)\,\mathbf{\Delta}^{-1}\gamma_{pp}\begin{pmatrix} \mu_4 \\ 0 \\ \mu_6 \\ 0 \end{pmatrix}\{1+o(1)\} \xrightarrow{\mathscr{D}} N(\mathbf{0}, \mathbf{\Delta}^{-1}\mathbf{\Lambda}\mathbf{\Delta}^{-1}).$$

Some algebraic computations yield

$$\boldsymbol{\Omega}^{-1}=\begin{pmatrix} \dfrac{\mu_4}{\mu_4-\mu_2^2} & 0 & -\dfrac{\mu_2}{\mu_4-\mu_2^2} & 0 \\[2ex] 0 & \dfrac{\mu_6}{\mu_6\mu_2-\mu_4^2} & 0 & -\dfrac{\mu_4}{\mu_6\mu_2-\mu_4^2} \\[2ex] -\dfrac{\mu_2}{\mu_4-\mu_2^2} & 0 & \dfrac{1}{\mu_4-\mu_2^2} & 0 \\[2ex] 0 & -\dfrac{\mu_4}{\mu_6\mu_2-\mu_4^2} & 0 & \dfrac{\mu_2}{\mu_6\mu_2-\mu_4^2} \end{pmatrix},$$

so that

$$\boldsymbol{\Omega}^{-1}\begin{pmatrix}1\\0\\\mu_2\\0\end{pmatrix}=\mathbf{e}_{1,4}, \qquad \text{and} \qquad \boldsymbol{\Omega}^{-1}\begin{pmatrix}\mu_4\\0\\\mu_6\\0\end{pmatrix}=\begin{pmatrix}\dfrac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}\\[2ex] 0\\[1ex] \dfrac{\mu_6-\mu_4\mu_2}{\mu_4-\mu_2^2}\\[2ex] 0\end{pmatrix}.$$

Therefore, the assertion in (9) holds. In order to prove (A.16), we need only to check Lyapounov's condition for that sequence, which can be easily verified. This completes the proof of the theorem.

Proof of Theorem 1. Let $\beta=(\mathbf{a}^T, \mathbf{b}^T, c_p, d_p)^T$,

$$\bar{\eta}(u, \mathbf{x})=\sum_{j=1}^{p}\{a_j(u_0)+a_j'(u_0)(u-u_0)\}\,x_j+\frac{a_p''(u_0)}{2}(u-u_0)^2\,x_p$$
$$+\frac{a_p'''(u_0)}{6}(u-u_0)^3\,x_p,$$

and

$$\beta^*=\gamma_n^{-1}\biggl(\beta_1-a_1(u_0), \ldots, \beta_p-a_p(u_0), h_1\{\beta_{p+1}-a_1'(u_0)\}, \ldots, h_1\{\beta_{2p}-a_p'(u_0)\},$$
$$h_1^2\left\{\beta_{2p+1}-\frac{a_p''(u_0)}{2}\right\}, h_1^3\left\{\beta_{2p+2}-\frac{a_p'''(u_0)}{3}\right\}\biggr)^T,$$

where $\gamma_n = (n\,h_1)^{-1/2}$. It can be easily seen that

$$\sum_{j=1}^{p} \{a_j + b_j(U_i - u_0)\}\, X_{ij} + c_p(U_i - u_0)^2\, X_{ip} + d_p(U_i - u_0)^3\, X_{ip}$$
$$= \bar{\eta}(U_i, \mathbf{X}_i) + \gamma_n \beta^{*T}\tilde{\mathbf{Z}}_i,$$

where $\tilde{\mathbf{Z}}_i = (\mathbf{X}_i^T, (U_i - u_0)/h_1\mathbf{X}_i^T, (U_i - u_0)^2/h_1^2 X_{ip}, (U_i - u_0)^3/h_1^3 X_{ip})^T$. Let

$$\hat{\beta}^* = \gamma_n^{-1}\Bigg(\hat{\beta}_1 - a_1(u_0), \ldots, \hat{\beta}_p - a_p(u_0), h_1\{\hat{\beta}_{p+1} - a_1'(u_0)\}, \ldots, h_1\{\hat{\beta}_{2p} - a_p'(u_0)\},$$
$$h_1^2\left\{\hat{\beta}_{2p+1} - \frac{a_p''(u_0)}{2}\right\}, h_1^3\left\{\hat{\beta}_{2p+2} - \frac{a_p'''(u_0)}{3}\right\}\Bigg)^T,$$

then, $\hat{\beta}^*$ maximizes

$$\sum_{i=1}^{n} [\ell\{g^{-1}(\bar{\eta}_i + \gamma_n\beta^{*T}\tilde{\mathbf{Z}}_i), Y_i\} - \ell\{g^{-1}(\bar{\eta}_i), Y_i\}]\, K\{(U_i - u_0)/h_1\},$$

where $\bar{\eta}_i = \bar{\eta}(U_i, \mathbf{X}_i)$. Following the same line as in the proof of Theorem 2, one obtains

$$\hat{\beta}^* = f^{-1}(u_0)\, \mathbf{\Omega}_1^{-1}\mathbf{W}_n + o_p(1), \tag{A.18}$$

where

$$\mathbf{\Omega}_1 = \mathbf{\Omega}_1(u_0) = \begin{pmatrix} \mathbf{\Gamma} & \mathbf{0} & \mu_2\mathbf{\Gamma}_p & \mathbf{0} \\ \mathbf{0} & \mu_2\mathbf{\Gamma} & \mathbf{0} & \mu_4\mathbf{\Gamma}_p \\ \mu_2\mathbf{\Gamma}_p^T & \mathbf{0}^T & \mu_4\gamma_{pp} & 0 \\ \mathbf{0}^T & \mu_4\mathbf{\Gamma}_p^T & 0 & \mu_6\gamma_{pp} \end{pmatrix}$$

and

$$\mathbf{W}_n = \gamma_n \sum_{i=1}^{n} q_1\{\bar{\eta}_i, Y_i\}\, \tilde{\mathbf{Z}}_i K\{(U_i - u_0)/h_1\}.$$

The asymptotic normality of $\hat{\beta}^*$ follows from that of $\mathbf{W}_n$. Hence it suffices to establish the asymptotic normality of $\mathbf{W}_n$. To this effect, it suffices to compute the mean and covariance matrix of $\mathbf{W}_n$ by Lyapounov condition because $\mathbf{W}_n$ is a sum of iid random vectors. By a Taylor series expansion and the first result in (A.1), we have

$$\eta(u_0 + h_1 u, \mathbf{x}) = \bar{\eta}(u_0 + h_1 u, \mathbf{x}) + \frac{h_1^2 u^2}{2}\sum_{j=1}^{p} a_j''(u_0)\, x_j + o(h_1^2),$$

and

$$q_1\{\bar{\eta}(u_0+h_1u,\, \mathbf{x}),\, m(u_0+h_1u,\, \mathbf{x})\} = \rho(u_0,\, \mathbf{x})\,\frac{h_1^2u^2}{2}\sum_{j=1}^{p} a_j''(u_0)\, x_j + o(h_1^2).$$

Use the above expression to obtain

$$E(\mathbf{W}_n) = \frac{h_1^2}{2\gamma_n}\, f(u_0)\begin{pmatrix} \mu_2\mathbf{\Gamma} \\ \mathbf{0}^T \\ \mu_4\mathbf{\Gamma}_p^T \\ 0 \end{pmatrix}\mathbf{a}''(u_0)\,\{1+o(1)\}. \tag{A.19}$$

Similarly,

$$\operatorname{var}(\mathbf{W}_n) = f(u_0)\,\mathbf{\Psi}_1\{1+o(1)\}, \tag{A.20}$$

where

$$\mathbf{\Psi}_1 = \mathbf{\Psi}_1(u_0) = \begin{pmatrix} \nu_0\mathbf{\Gamma} & \mathbf{0} & \nu_2\mathbf{\Gamma}_p & \mathbf{0} \\ \mathbf{0} & \nu_2\mathbf{\Gamma} & \mathbf{0} & \nu_4\mathbf{\Gamma}_p \\ \nu_2\mathbf{\Gamma}_p^T & \mathbf{0}^T & \nu_4\gamma_{pp} & 0 \\ \mathbf{0}^T & \nu_4\mathbf{\Gamma}_p^T & 0 & \nu_6\gamma_{pp} \end{pmatrix}.$$

By using the Cramér–Wold device, checking the Lyapounov's condition, and combining (A.18), (A.19) and (A.20), we obtain

$$\hat{\beta}^* - \frac{(n\,h_1^5)^{1/2}}{2}\,\mathbf{\Omega}_1^{-1}\begin{pmatrix} \mu_2\mathbf{\Gamma} \\ \mathbf{0}^T \\ \mu_4\mathbf{\Gamma}_p^T \\ 0 \end{pmatrix}\mathbf{a}''(u_0)\{1+o(1)\} \xrightarrow{\mathscr{D}} N(\mathbf{0},\, f^{-1}(u_0)\,\mathbf{\Omega}_1^{-1}\mathbf{\Psi}_1\mathbf{\Omega}_1^{-1}).$$

It is easily seen that

$$\mathbf{\Omega}_1^{-1} = \begin{pmatrix} \mathbf{\Gamma}^{-1} + \mu_2^2\mathbf{e}_{p,\,p}\mathbf{e}_{p,\,p}^T\zeta & \mathbf{0} & -\mu_2\mathbf{e}_{p,\,p}\zeta & \mathbf{0} \\ \mathbf{0} & \mu_2^{-1}\mathbf{\Gamma}^{-1} + \mu_2^{-1}\mu_4^2\mathbf{e}_{p,\,p}\mathbf{e}_{p,\,p}^T\lambda & \mathbf{0} & -\mu_4\mathbf{e}_{p,\,p}\lambda \\ -\mu_2\mathbf{e}_{p,\,p}^T\zeta & \mathbf{0}^T & \zeta & 0 \\ \mathbf{0}^T & -\mu_4\mathbf{e}_{p,\,p}^T\lambda & 0 & \mu_2\lambda \end{pmatrix},$$

where $\zeta = \gamma_{pp}^{-1}(\mu_4 - \mu_2^2)^{-1}$ and $\lambda = \gamma_{pp}^{-1}(\mu_6\mu_2 - \mu_4^2)^{-1}$, and that

$$\mathbf{\Omega}_1^{-1}\begin{pmatrix} \mu_2\mathbf{\Gamma} \\ \mathbf{0}^T \\ \mu_4\mathbf{\Gamma}_p^T \\ 0 \end{pmatrix} = \begin{pmatrix} \mu_2[\mathbf{I} - \gamma_{pp}^{-1}\mathbf{e}_{p,p}\mathbf{\Gamma}_p^T] \\ \mathbf{0}^T \\ \gamma_{pp}^{-1}\mathbf{\Gamma}_p^T \\ 0 \end{pmatrix}.$$

Also,

$$(\mathbf{I}_p, \mathbf{0})\,\mathbf{\Omega}_1^{-1}\mathbf{\Psi}_1\mathbf{\Omega}_1^{-1}\begin{pmatrix} \mathbf{I}_p \\ \mathbf{0} \end{pmatrix} = \nu_0\mathbf{\Gamma}^{-1}.$$

It follows in an obvious manner that (8) holds. This completes the proof of the theorem.

ACKNOWLEDGMENT

The author is indebted to Professor Jianqing Fan for bringing the author's attention to the problem and the editor and two referees for their constructive comments which helped to improve an earlier version of this article.

REFERENCES

B. Brumback and J. Rice, Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion), *J. Amer. Statist. Assoc.* **93** (1998), 961–976.

Z. Cai, J. Fan, and R. Li, Efficient estimation and inferences for varying-coefficient models, *J. Amer. Statist. Assoc.* **95** (2000), 888–902.

Z. Cai, J. Fan, and Q. Yao, Functional-coefficient regression model for nonlinear time series, *J. Amer. Statist. Assoc.* **95** (2000), 941–956.

Z. Cai and R. C. Tiwari, Application of a local linear autoregressive model to BOD time series, *Environmetrics* **11** (2000), 341–350.

R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand, Generalized partially linear single-index models, *J. Amer. Statist. Assoc.* **92** (1997), 477–489.

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

L. E. Cederman and M. Penubarti, Evolutionary liberalism: Exploring the dynamics of interstate conflict, manuscript submitted for publication, 1999.

R. Chen and R. S. Tsay, Functional-coefficient autoregressive models, *J. Amer. Statist. Assoc.* **88** (1993), 298–308.

W. S. Cleveland, E. Grosse, and W. M. Shyu, Local regression models, *in* "Statistical Models in S" (J. M. Chambers and T. J. Hastie, Eds), Wadsworth & Brooks, Pacific Grove, 1992.

J. Fan and I. Gijbels, Data driven bandwidth selection in local polynomial fitting: Variable bandwidth spatial adaptation, *J. Roy. Statist. Soc. Ser. B* **57** (1995), 371–394.

J. Fan and I. Gijbels, "Local Polynomial Modelling and Its Applications," Chapman and Hall, London, 1996.

J. Fan, Q. Yao, and Z. Cai, Adaptive varying-coefficient linear models, *J. Roy. Statist. Soc. Ser. B*, in press.

J. Fan and J. Zhang, Comments on "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves", *J. Amer. Statist. Assoc.* **93** (1998), 980–983.

J. Fan and W. Zhang, Statistical estimation in varying-coefficient models, *The Annals of Statistics* **27** (1999), 1491–1518.

J. Fan and J. Zhang, Functional linear models for longitudinal data, *Journal of the Royal Statistical Society, Series B* **62** (2000), 303–322.

T. J. Hastie and R. J. Tibshirani, Varying-coefficient models (with discussion), *J. Roy. Statist. Soc. Ser. B* **55** (1993), 757–796.

Y. Hong and T.-H. Lee, Inference and forecast of exchange rates via generalized spectrum and nonlinear times series models, manuscript submitted for publication, 1999.

D. R. Hoover, J. A. Rice, C. O. Wu, and L. P. Yang, Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* **85** (1998), 809–822.

Y. P. Mack and B. W. Silverman, Weak and strong uniform consistency of kernel regression estimates, *Z. Wahr. Gebiete* **61** (1982), 405–415.

D. Ruppert, Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *J. Amer. Statist. Assoc.* **92** (1997), 1057–1062.

D. Ruppert, S. J. Sheather, and M. P. Wand, An effective bandwidth selection for local least squares regression, *J. Amer. Statist. Assoc.* **90** (1995), 1257–1270.

M. Stone, Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. Ser. B* **36** (1974), 111–147.

Y. Xia and W. K. Li, On the estimation and testing of functional-coefficient linear models, *Statist. Sinica* **9** (1999), 735–757.