The Annals of Statistics
1999, Vol. 27, No. 5, 1491–1518

STATISTICAL ESTIMATION IN VARYING COEFFICIENT MODELS

BY JIANQING FAN[1] AND WENYANG ZHANG

University of North Carolina and Chinese University of Hong Kong

Varying coefficient models are a useful extension of classical linear models. They arise naturally when one wishes to examine how regression coefficients change over different groups characterized by certain covariates such as age. The appeal of these models is that the coefficient functions can easily be estimated via a simple local regression. This yields a simple one-step estimation procedure. We show that such a one-step method cannot be optimal when different coefficient functions admit different degrees of smoothness. This drawback can be repaired by using our proposed two-step estimation procedure. The asymptotic mean-squared error for the two-step procedure is obtained and is shown to achieve the optimal rate of convergence. A few simulation studies show that the gain by the two-step procedure can be quite substantial. The methodology is illustrated by an application to an environmental data set.

1. Introduction.

1.1. *Background.* Driven by many sophisticated applications and fueled by modern computing power, many useful data-analytic modeling techniques have been proposed to relax traditional parametric models and to exploit possible hidden structure. For introduction to these techniques, see the books by Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijbels (1996), among others. In dealing with high-dimensional data, many powerful approaches have been incorporated to avoid the so-called "curse of dimensionality." Examples include additive models [Breiman and Friedman (1995), Hastie and Tibshirani (1990)], low-dimensional interaction models, [Friedman (1991), Gu and Wahba (1993), Stone, Hansen, Kooperberg and Truong (1997)], multiple-index models [Härdle and Stoker (1990), Li (1991)], partially linear models [Wahba (1984), Green and Silverman (1994)], and their hybrids [Carroll, Fan, Gijbels and Wand (1997), Fan, Härdle and Mammen (1998), Heckman, Ichimura, Smith and Todd (1998)], among others. Different models explore different aspects of high-dimensional data and incorporate different prior knowledge into modeling and approximation. They together form useful tool kits for processing high-dimensional data.

A useful extension of classical linear models is varying coefficient models. This idea is scattered around in text books. See, for example, page 245 of Shumway (1988). However, the potential of such a modeling technique did not

Received October 1997; revised March 1999.
[1]Supported in part by NSF Grant DMS-98-03200 and NSA Grant 96-1-0015.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62J12.
Key words and phrases. Varying coefficient models, local linear fit, optimal rate of convergence, mean-squared errors.

get fully explored until the seminal work of Cleveland, Grosse and Shyu (1991) and Hastie and Tibshirani (1993). The varying coefficient models assume the following conditional linear structure:

$$Y = \sum_{j=1}^{p} a_j(U)X_j + \varepsilon \tag{1.1}$$

for given covariates $(U, X_1, \dots, X_p)'$ and response variable Y with

$$E(\varepsilon \mid U, X_1, \dots, X_p) = 0$$

and

$$\operatorname{var}(\varepsilon \mid U, X_1, \dots, X_p) = \sigma^2(U).$$

By regarding $X_1 \equiv 1$, (1.1) allows a varying intercept term in the model. The appeal of this model is that, via allowing coefficients $a_1, \dots, a_p$ to depend on U, the modeling bias can significantly be reduced and the "curse of dimensionality" can be avoided. Another advantage of this model is its interpretability. It arises naturally when one is interested in exploring how regression coefficients change over different groups such as age. It is particularly appealing in longitudinal studies where it allows one to examine the extent to which covariates affect responses over time. See Hoover, Rice, Wu and Yang (1997) and Fan and Zhang (2000) for details on novel applications of varying coefficient models to longitudinal data. For nonlinear time series applications, see Chen and Tsay (1993) where functional coefficient AR models are proposed and studied.

1.2. *Estimation methods.* Suppose that we have a random sample $\{(U_i, X_{i1}, \dots, X_{ip}, Y_i)\}_{i=1}^n$ from model (1.1). One simple approach to estimate the coefficient functions $a_j(\cdot)(j = 1, \dots, p)$ is to use local linear modeling. For each given point u_0, approximate the function locally as

$$a_j(u) \approx a_j + b_j(u - u_0), \tag{1.2}$$

for u in a neighborhood of u_0. This leads to the following local least-squares problem: minimize

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\{a_j + b_j(U_i - u_0)\}X_{ij}\right]^2 K_h(U_i - u_0) \tag{1.3}$$

for a given kernel function K and bandwidth h, where $K_h(\cdot) = K(\cdot/h)/h$. The idea is due to Cleveland, Grosse and Shyu (1991). While this idea is very simple and useful, it is implicitly assumed that the functions $a_j(\cdot)$ possess about the same degrees of smoothness and hence they can be approximated equally well in the same interval. If the functions possess different degrees of smoothness, suboptimal estimators are obtained via using the least-squares method (1.3).

To formulate the above intuition in a mathematical framework, let us assume that $a_p(\cdot)$ is smoother than the rest of the functions. For concreteness, we assume that a_p possesses a bounded fourth derivative so that the function can locally be approximated by a cubic function,

$$a_p(u) \approx a_p + b_p(u - u_0) + c_p(u - u_0)^2 + d_p(u - u_0)^3, \tag{1.4}$$

for u in a neighborhood of u_0. This naturally leads to the following weighted least-squares problem:

$$\begin{aligned} \sum_{i=1}^{n} \Bigg[Y_i - \sum_{j=1}^{p-1} \{a_j + b_j(U_i - u_0)\} X_{ij} \\ - \big\{a_p + b_p(U_i - u_0) + c_p(U_i - u_0)^2 + d_p(U_i - u_0)^3\big\} X_{ip} \Bigg]^2 \\ \times K_{h_1}(U_i - u_0). \end{aligned} \tag{1.5}$$

Let $\hat{a}_{j,1}, \hat{b}_{j,1}$ $(j = 1, \dots, p-1)$ and $\hat{a}_{p,1}, \hat{b}_{p,1}, \hat{c}_{p,1}, \hat{d}_{p,1}$ minimize (1.5). The resulting estimator $\hat{a}_{p,\mathrm{OS}}(u_0) = \hat{a}_{p,1}$ is called a one-step estimator. We will show that the bias of the one-step estimator is $O(h_1^2)$ and the variance of the one-step estimator is $O((nh_1)^{-1})$. Therefore, using the one-step estimator $\hat{a}_{p,\mathrm{OS}}(u_0)$, the optimal rate $O(n^{-8/9})$ cannot be achieved.

To achieve the optimal rate, a two-step procedure has to be used. The first step involves getting an initial estimate of $a_1(\cdot), \dots, a_{p-1}(\cdot)$. Such an initial estimate is usually undersmoothed so that the bias of the initial estimator is small. Then, in the second step, a local least-squares regression is fitted again via substituting the initial estimate into the local least-squares problem. More precisely, we use the local linear regression to obtain a preliminary estimate by minimizing

$$\sum_{k=1}^{n} \left(Y_k - \sum_{j=1}^{p} \{a_j + b_j(U_k - u_0)\} X_{kj} \right)^2 K_{h_0}(U_k - u_0) \tag{1.6}$$

for a given initial bandwidth h_0 and kernel K. Let $\hat{a}_{1,0}(u_0), \dots, \hat{a}_{p,0}(u_0)$ denote the initial estimate of $a_1(u_0), \dots, a_p(u_0)$. In the second step, we substitute the preliminary estimates $\hat{a}_{1,0}(\cdot), \dots, \hat{a}_{p-1,0}(\cdot)$ and use a local cubic fit to estimate $a_p(u_0)$, namely, minimize

$$\begin{aligned} \sum_{i=1}^{n} \Bigg(Y_i - \sum_{j=1}^{p-1} \hat{a}_{j,0}(U_i) X_{ij} \\ - \big\{a_p + b_p(U_i - u_0) + c_p(U_i - u_0)^2 + d_p(U_i - u_0)^3\big\} X_{ip} \Bigg)^2 \\ \times K_{h_2}(U_i - u_0) \end{aligned} \tag{1.7}$$

with respect to a_p, b_p, c_p, d_p, where h_2 is the bandwidth in the second step. In this way, a two-step estimator of $\hat{a}_{p,\mathrm{TS}}(u_0)$ of $a_p(u_0)$ is obtained. We will show that the bias of the two-step estimator is of $O(h_2^4)$ and the variance $O\{(nh_2)^{-1}\}$, provided that

$$h_0 = o(h_2^2), \qquad nh_0/\log h_0 \to \infty,$$

and $nh_0^3 \to \infty$. This means that when the optimal bandwidth $h_2 \sim n^{-1/9}$ is used, and the preliminary bandwidth h_0 is between the rates $O(n^{-1/3})$ and $O(n^{-2/9})$, the optimal rates of convergence $O(n^{-8/9})$ for estimating a_2 can be achieved.

Note that the condition $nh_0^3 \to \infty$ is only a convenient technical condition based on the assumption of the sixth bounded moments of the covariates. It plays little role in understanding the two-step procedure. If X_i is assumed to have higher moments, the condition can be relaxed to be as weak as $nh^{1+\delta} \to \infty$ for some small $\delta > 0$. See Condition (7) in Section 4 for details. Therefore, the requirement on h_0 is very minimal. A practical implication of this is that the two-step estimation method is not sensitive to the initial bandwidth h_0. This makes practical implementation much easier.

Another possible way to conduct variable smoothing for coefficient functions is to use the following smoothing spline approach proposed by Hastie and Tibshirani (1993):

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p} a_j(U_i)X_{ij}\right]^2 + \sum_{j=1}^{p} \lambda_j \int \{a_j''(u)\}^2\,du$$

for some smoothing parameters $\lambda_1, \dots, \lambda_p$. While this idea is powerful, there are a number of potential problems. First, there are p-smoothing parameters to choose simultaneously. This is quite a task in practice. Second, computation can be a challenge. An iterative scheme was proposed in Hastie and Tibshirani (1993). Third, sampling properties are somewhat difficult to obtain. It is not clear if the resulting method can achieve the same optimal rate of convergence as the one-step procedure.

The above theoretical work is not purely academic. It has important practical implications. To validate our asymptotic claims, we use three simulated example to illustrate our methodology. The sample size is $n = 500$ and $p = 2$. Figure 1 depicts typical estimates of the one-step and two-step methods, both using the optimal bandwidth for estimating $a_2(\cdot)$ (For the two-step estimator, we do not optimize simultaneously the bandwidths h_0 and h_2; rather, we only optimize the bandwidth h_2 for a given small bandwidth h_0). Details of simulations can be found in Section 5.2. In the first example, the bias of the one-step estimate is too large since the optimal bandwidth h_1 for a_2 is so large that a_1 can no longer to approximated well by a linear function in such a large neighborhood. In the second example the estimated curve is clearly undersmoothed by using the one-step estimate, since the optimal bandwidth for a_2 has to be very small in order to compromise for the bias arising from approximating a_1. The one-step estimator works reasonably well in the third example, though the two-step estimator still improves somewhat the quality of the one-step estimate.

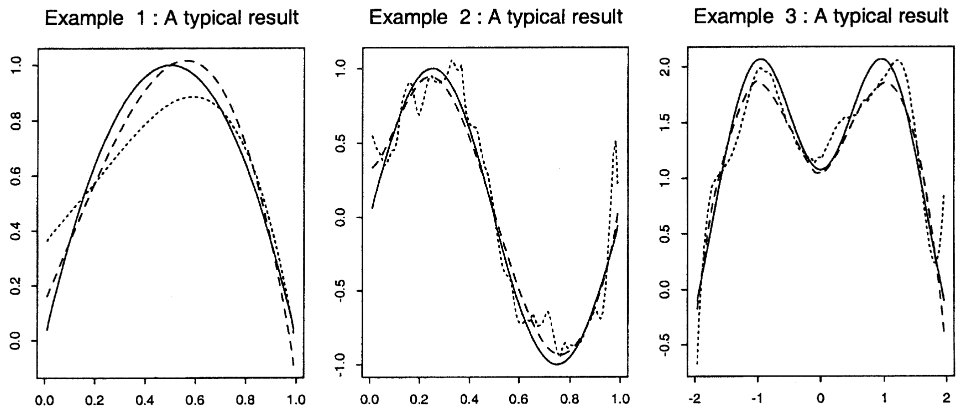

FIG. 1. *Comparisons of the performance between the one-step and two-step estimator. Solid curves: true functions; short-dashed curves: estimates based on the one-step procedure; long-dashed curves: estimates based on the two-step procedure.*

In real applications, we do not know in advance if a_p is really smoother than the rest of the functions. The above discussion reveals that the two-step procedure can lead to significant gain when a_p is smoother than the rest of the functions. When a_p has the same degree of smoothness as the rest of the functions, we will demonstrate that the two-step estimation procedure has the same performance as the one-step approach. Therefore, the two-step scheme is always more reliable than the one-step approach. Details of implementing the two-step method will be outlined in Section 2.

1.3. *Outline of the paper.* Section 2 gives strategies for implementing the two-step estimators. The explicit formulas for our proposed estimators are given in Section 3. Section 4 studies asymptotic properties of the one-step and two-step estimators. In Section 5, we study finite sample properties of the one-step and two-step estimators via some simulated examples. Two-step techniques are further illustrated by an application to an environmental data set. Technical proofs are given in Section 6.

2. Practical implementation of two-step estimators. As discussed in the introduction, a one-step procedure is not optimal when coefficient functions admit different degrees of smoothness. However, we do not know in advance which function is not smooth. To implement the two-step strategy, one minimizes (1.6) with a small bandwidth h_0 to obtain preliminary estimates $\hat{a}_{1,0}(U_i), \ldots, \hat{a}_{p,0}(U_i)$ for $i = 1, \ldots, n$. With these preliminary estimates, one can now estimate the coefficient functions $a_j(u_0)$ by using an equation that is similar to (1.7). Other techniques such as smoothing splines can also be used in the second stage of fitting.

In practical implementation, it usually suffices to use local linear fits instead of local cubic fits in the second step. This would result in computational savings. Our experience with local polynomial fits show that for practical purpose the local linear fit with optimally chosen bandwidth performs comparably with the local cubic fit with optimal bandwidth.

As discussed in the introduction, the two-step estimator is not very sensitive to the choice of initial bandwidth as long as it is small enough so that the bias in the first step smoothing is negligible. This suggests the following simple automatic rule: use cross-validation or generalized cross-validation [see, e.g., Hoover, Rice, Wu and Yang (1997)] to select the bandwidth $\hat{h}$ for the one-step fit. Then, use $h_0 = 0.5\,\hat{h}$ (say) as the initial bandwidth.

An advantage of the two-step procedure is that in the second step, the problem is really a univariate smoothing problem. Therefore, one can apply univariate bandwidth selection procedures such as cross-validation [Stone, (1974)], preasymptotic substitution method [Fan and Gijbels (1995)], plug-in bandwidth selector [Ruppert, Sheather and Wand (1995)] and empirical bias method [Ruppert (1997)] to select the smoothing parameter. As discussed before, the preliminary bandwidth h_0 is not very crucial to our final estimates, since for a wide range of bandwidth h_0 the two-step method will achieve the optimal rate. This is another benefit of the two-step procedure: bandwidth selection problems become relatively easy.

3. Formulas for the proposed estimators. The solution to the least squares problems (1.5)–(1.7) can easily be obtained. We take this opportunity to introduce necessary notation. In the notation below, we use subscripts “0”, “1” and “2,” respectively, to indicate the variables related to the initial, one-step and two-step estimators. Let

$$\mathbf{X}_0 = \begin{pmatrix} X_{11} & X_{11}(U_1 - u_0) & \cdots & X_{1p} & X_{1p}(U_1 - u_0) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n1}(U_n - u_0) & \cdots & X_{np} & X_{np}(U_n - u_0) \end{pmatrix},$$

$$Y = (Y_1, \dots, Y_n)^T \quad \text{and}$$

$$W_0 = \mathrm{diag}\big(K_{h_0}(U_1 - u_0), \dots, K_{h_0}(U_n - u_0)\big).$$

Then the solution to the least-squares problem (1.6) can be expressed as

$$\hat{a}_{j,0}(u_0) = e^T_{2j-1,2p}(\mathbf{X}_0^T W_0 \mathbf{X}_0)^{-1}\mathbf{X}_0^T W_0 Y, \qquad j = 1, \dots, p. \tag{3.1}$$

Here and hereafter, we will always use notation $e_{k,m}$ to denote the unit vector of length m with 1 at the kth position.

The solution to problem (1.5) can be expressed as follows. Let

$$\mathbf{X}_2 = \begin{pmatrix} X_{1p} & X_{1p}(U_1 - u_0) & X_{1p}(U_1 - u_0)^2 & X_{1p}(U_1 - u_0)^3 \\ \vdots & \vdots & \vdots & \vdots \\ X_{np} & X_{np}(U_n - u_0) & X_{np}(U_n - u_0)^2 & X_{np}(U_n - u_0)^3 \end{pmatrix}$$

and

$$\mathbf{X}_3 = \begin{pmatrix} X_{11} & X_{11}(U_1 - u_0) & \cdots & X_{1(p-1)} & X_{1(p-1)}(U_1 - u_0) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n1}(U_n - u_0) & \cdots & X_{n(p-1)} & X_{n(p-1)}(U_n - u_0) \end{pmatrix},$$

$$\mathbf{X}_1 = (\mathbf{X}_3, \mathbf{X}_2), \qquad W_1 = \operatorname{diag}\big(K_{h_1}(U_1 - u_0), \ldots, K_{h_1}(U_n - u_0)\big).$$

Then the solution to the least-squares problem (1.5) is given by

$$\hat{a}_{p,1}(u_0) = e^T_{2p-1,2p+2}(\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1}\mathbf{X}_1^T W_1 Y. \tag{3.2}$$

Using the notation introduced above, we can express the two-step estimator as

$$\hat{a}_{p,2}(u_0) = (1, 0, 0, 0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 (Y - V), \tag{3.3}$$

where

$$W_2 = \operatorname{diag}\big(K_{h_2}(U_1 - u_0), \ldots, K_{h_2}(U_n - u_0)\big)$$

and $V = (V_1, \ldots, V_n)^T$ with $V_i = \sum_{j=1}^{p-1} \hat{a}_{j,0}(U_i) X_{ij}$. Note that the two-step estimator $\hat{a}_{p,2}$ is a linear estimator for given bandwidths h_0 and h_2, since it is a weighted average of observations $Y_1, \ldots, Y_n$. The weights are somewhat complicated. To obtain these weights, let $\mathbf{X}_{(i)}$ be the matrix $\mathbf{X}_0$ with $u_0 = U_i$ and $W_{(i)}$ be the matrix W_0 with $u_0 = U_i$. Then

$$V_i = \sum_{j=1}^{p-1} X_{ij} e^T_{2j-1,2p}\big(\mathbf{X}_{(i)}^T W_{(i)} \mathbf{X}_{(i)}\big)^{-1}\mathbf{X}_{(i)}^T W_{(i)} Y.$$

Set

$$B_n = I_n - \sum_{j=1}^{p-1} \begin{pmatrix} X_{1j} e^T_{2j-1,2p}\big(\mathbf{X}_{(1)}^T W_{(1)} \mathbf{X}_{(1)}\big)^{-1} X_{(1)}^T W_{(1)} \\ \vdots \\ X_{nj} e^T_{2j-1,2p}\big(\mathbf{X}_{(n)}^T W_{(n)} \mathbf{X}_{(n)}\big)^{-1}\mathbf{X}_{(n)}^T W_{(n)} \end{pmatrix}.$$

Then

$$\hat{a}_{p,2}(u_0) = (1, 0, 0, 0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 B_n Y. \tag{3.4}$$

4. Main results. We impose the following technical conditions:

1. $EX_j^{2s} < \infty$, for some $s > 2$, $j = 1, \ldots, p$.
2. $a''_j(\cdot)$ is continuous in a neighborhood of u_0, for $j = 1, \ldots, p$. Further, assume $a''_j(u_0) \neq 0$, for $j = 1, \ldots, p$.
3. The function a_p has a continuous fourth derivative in a neighborhood of u_0.
4. $r''_{ij}(\cdot)$ is continuous in a neighborhood of u_0 and $r''_{ij}(u_0) \neq 0$, for $i, j = 1, \ldots, p$, where $r_{ij}(u) = E(X_i X_j \mid U = u)$.

5. The marginal density of U has a continuous second derivative in some neighborhood of u_0 and $f(u_0) \neq 0$.
6. The function $K(t)$ is a symmetric density function with a compact support.
7. $h_0/h_2 \to 0$ and $h_2 \to 0, nh_0^{\gamma}/\log h_0 \to \infty$, for any $\gamma > s/(s-2)$ with s given in condition 1.

Throughout this paper, we will use the following notation. Let

$$\mu_i = \int t^i K(t)\,dt \quad \text{and} \quad \nu_i = \int t^i K^2(t)\,dt$$

and $\mathscr{D}$ be the observed covariates vector, namely,

$$\mathscr{D} = (U_1, \dots, U_n, X_{11}, \dots, X_{1n}, \dots, X_{p1}, \dots, X_{pn})^T.$$

Set $r_{ij} = r_{ij}(u_0) = E(X_i X_j \mid U = u_0)$, for $i, j = 1, \dots, p$. Put

$$\Psi = \operatorname{diag}(\sigma^2(U_1), \dots, \sigma^2(U_n)),$$
$$\alpha_j(u) = (r_{1j}(u), \dots, r_{(p-1)j}(u))^T,$$
$$\alpha_j = \alpha_j(u_0) \quad \text{for } j = 1, \dots, p$$

and

$$\Omega_i(u) = E\{(X_1, \dots, X_i)^T (X_1, \dots, X_i) \mid U = u\},$$
$$\Omega_i = \Omega_i(u_0) \quad \text{for } i = 1, \dots, p.$$

For the one step-estimator, we have the following asymptotic bias and variance.

THEOREM 1. *Under conditions* 1–6, *if $h_1 \to 0$ in such a way that $nh_1 \to \infty$, then the asymptotic conditional bias of $\hat{a}_{p,\mathrm{OS}}(u_0)$ is given by*

$$\operatorname{bias}(\hat{a}_{p,\mathrm{OS}}(u_0) \mid \mathscr{D}) = -\frac{h_1^2 \mu_2}{2r_{pp}} \sum_{j=1}^{p-1} r_{pj} a_j''(u_0) + o_P(h_1^2)$$

and the asymptotic conditional variance of $\hat{a}_{p,\mathrm{OS}}(u_0)$ is

$$\operatorname{var}(\hat{a}_{p,\mathrm{OS}}(u_0) \mid \mathscr{D}) = \frac{\sigma^2(u_0)(\lambda_2 r_{pp} + \lambda_3 \alpha_p^T \Omega_{p-1}^{-1} \alpha_p)}{nh_1 f(u_0) \lambda_1 r_{pp} (r_{pp} - \alpha_p^T \Omega_{p-1}^{-1} \alpha_p)} (1 + o_p(1)),$$

where $\lambda_1 = (\mu_4 - \mu_2^2)^2$ $\lambda_2 = \nu_0 \mu_4^2 - 2\nu_2 \mu_2 \mu_4 + \mu_2^2 \nu_4$ *and* $\lambda_3 = 2\mu_2 \nu_2 \mu_4 - 2\nu_0 \mu_2^2 \mu_4 - \mu_2^2 \nu_4 + \nu_0 \mu_2^4$.

The proofs of Theorem 1 and other theorems are given in Section 6. It is clear that the conditional MSE of the one-step estimator $\hat{a}_{p,\mathrm{OS}}(u_0)$ is only $O_P\{h_1^4 + (nh_1)^{-1}\}$ which achieves the rate $O_P(n^{-4/5})$ when the bandwidth $h_1 = O(n^{-1/5})$ is used. The bias expression above indicates clearly that the approximation errors of functions $a_1, \dots, a_{p-1}$ are transmitted to the bias of estimating a_p. Thus, the one-step estimator for a_p inherits nonnegligible approximation errors and is not optimal. Note that Theorem 1 continues to hold if condition (3) is dropped. See also Theorem 3.

We now consider the asymptotic MSE for the two-step estimator.

THEOREM 2. *If conditions* 1–7 *hold, then the asymptotic conditional bias of* $\hat{a}_{p,\mathrm{TS}}(u_0)$ *can be expressed as*

$$\begin{aligned}&\operatorname{bias}(\hat{a}_{p,\mathrm{TS}}(u_0) \mid \mathscr{D})\\&\quad = \frac{1}{4!}\frac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 - \frac{\mu_2 h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} a_j''(u_0)r_{pj} + o_P(h_2^4+h_0^2)\end{aligned}$$

and the asymptotic conditional variance of $\hat{a}_{p,\mathrm{TS}}(u_0)$ *is given by*

$$\operatorname{var}(\hat{a}_{p,\mathrm{TS}}(u_0) \mid \mathscr{D}) = \frac{(\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2+\mu_2^2\nu_4)\sigma^2(u_0)}{nh_2 f(u_0)(\mu_4-\mu_2^2)^2} e_{p,p}^T\Omega_p^{-1}e_{p,p}\{1+o_P(1)\}.$$

By Theorem 2, the asymptotic variance of the two-step estimator is independent of the initial bandwidth as long as $nh_0^\gamma \to \infty$, where γ is given in condition 7. Thus, the initial bandwidth h_0 should be chosen as small as possible subject to the constraint that $nh_0^\gamma \to \infty$. In particular, when $h_0 = o(h_2^2)$, the bias from the initial estimator becomes negligible and the bias expression for the two-step estimator is

$$\frac{1}{4!}\frac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 + o_P(h_2^4).$$

Hence, via taking the optimal bandwidth h_2 of order $n^{-1/9}$, the conditional MSE of the two-step estimator achieves the optimal rate of convergence $O_P(n^{-8/9})$.

REMARK 1. Consider the ideal situation where $a_1, \dots, a_{p-1}$ are known. Then, one can simply run a local cubic estimator to estimate a_p. The resulting estimator has the asymptotic bias

$$\frac{1}{4!}\frac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 + o_P(h_2^4)$$

and asymptotic variance

$$\frac{\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2+\mu_2^2\nu_4}{nh_2 f(u_0)r_{pp}(\mu_4-\mu_2^2)^2}\sigma^2(u_0) + o_P\{(nh_2)^{-1}\}.$$

This ideal estimator has the same asymptotic bias as the two-step estimator. Further, this ideal estimator has the same order of variance as the two-step estimator. In other words, the two-step estimator enjoys the same optimal rate of convergence as the ideal one.

We now consider the case that a_p is as smooth as the rest of functions. In technical terms, we assume that a_p has only continuous second derivative. For this case, a local linear approximation is used for the function a_p in both the one-step and two-step procedure. With some abuse of notation, we still denote the resulting one-step and two-step estimators as $\hat{a}_{p,\mathrm{OS}}$ and $\hat{a}_{p,\mathrm{TS}}$, respectively.

Our technical results are to establish that the two-step estimator does not lose its statistical efficiency. Indeed, it has the same performance as the one-step procedure. Since it gains the efficiency when a_p is smoother, we conclude that the two-step estimator is preferable. These results give theoretical endorsement of the proposed two-step method in Section 2.

THEOREM 3. *Under conditions* 1, 2, 4–6, *if* $h_1 \to 0$ *and* $nh_1 \to \infty$, *then the asymptotic conditional bias of the one-step estimator is given by*

$$\operatorname{bias}\big(\hat{a}_{p,\mathrm{OS}}(u_0) \mid \mathscr{D}\big) = \frac{h_1^2\mu_2}{2} a_p''(u_0)(1 + o_P(1))$$

and the asymptotic conditional variance of $\hat{a}_{p,\mathrm{OS}}(u_0)$ *is given by*

$$\operatorname{var}\big(\hat{a}_{p,\mathrm{OS}}(u_0) \mid \mathscr{D}\big) = \frac{\sigma^2(u_0)\nu_0}{nh_1 f(u_0)} e_{p,p}^T \Omega_p^{-1} e_{p,p}\{1 + o_P(1)\}.$$

We now consider the asymptotic behavior for the two-step estimator.

THEOREM 4. *Suppose that conditions* 1, 2, 4–7 *hold. Then we have the asymptotic conditional bias*

$$\operatorname{bias}\big(\hat{a}_{p,\mathrm{TS}}(u_0) \mid \mathscr{D}\big) = \left(\frac{1}{2} a_p''(u_0)\mu_2 h_2^2 - \frac{\mu_2 h_0^2}{2r_{pp}} \sum_{j=1}^{p-1} a_j''(u_0) r_{pj}\right)(1 + o_P(1))$$

and the asymptotic variance

$$\operatorname{var}\big(\hat{a}_{p,\mathrm{TS}}(u_0) \mid \mathscr{D}\big) = \frac{\nu_0\sigma^2(u_0)}{nh_2 f(u_0)} e_{p,p}^T \Omega_p^{-1} e_{p,p}\{1 + o_P(1)\}.$$

REMARK 2. The asymptotic bias of the two-step estimator is simplified as

$$\tfrac{1}{2} a_2''(u_0)\mu_2 h_1^2(1 + o_P(1)),$$

by taking initial bandwidth $h_0 = o(h_2)$. Moreover, it has the same asymptotic variance as that of the one-step estimator. In other words, the performance of the one-step and two-step estimators is asymptotically identical.

REMARK 3. When $a_1(t), \ldots, a_{p-1}(t)$ are known, we can use the local linear fit to find an estimate of a_p. Such an ideal estimator possesses the bias

$$\tfrac{1}{2} a_p''(u_0)\mu_2 h_2^2\{1 + o_P(1)\}$$

and variance

$$\frac{\sigma^2(u_0)\nu_0}{nh_2 f(u_0) r_{pp}}\{1 + o_P(1)\}.$$

So, both one-step and two-step estimators have the same order of MSE as the ideal estimator. However, the variance of the ideal estimator is typically small. Unless $(X_1, \dots, X_{p-1})$ and X_p are uncorrelated given $U = u_0$, the asymptotic variance of the ideal estimator is always smaller.

5. Simulations and applications. In this section, we illustrate our methodology via an application to an environmental data set and via simulations. Throughout this section, we use the Epanechnikov kernal $K(t) = 0.75(1 - t^2)_+$.

5.1. *Applications to an environmental data set.* We now illustrate the methodology via an application to an environmental data set. The data set used here consists of a collection of daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995 (courtesy of Professor T. S. Lau). Three pollutants, sulphur dioxide (in $\mu g/m^3$), nitrogen dioxide (in $\mu g/m^3$) and dust (in $\mu g/m^3$), are considered here. Table 1 summarizes their correlation coefficients. The correlation between the dust level and NO_2 is quite high. Figure 2 depicts the marginal distributions of the level of pollutants in the summer (April 1–September 30) and winter seasons (October 1–March 31). The level of pollutants in the summer season (raining very often) tends to be lower and has smaller variation.

An objective of the study is to understand the association between the level of the pollutants and the number of daily total hospital admissions for circulatory and respiratory problems and to examine the extent to which the association varies over time. We consider the relationship among the number of daily hospital admissions (Y) and level of pollutants SO_2, NO_2 and dust, which are denoted by X_2, X_3 and X_4, respectively. We took $X_1 = 1$ as the intercept term and $U = t =$ time. The varying coefficient model

$$Y = a_1(t) + a_2(t)X_2 + a_3(t)X_3 + a_4(t)X_4 + \varepsilon \tag{5.1}$$

TABLE 1
Correlation coefficients among pollutants

	Sulphur dioxide	**Nitrogen dioxide**	**Dust**
Sulphur dioxide	1.000000	0.402452	0.281008
Nitrogen dioxide		1.000000	0.781975
Dust			1.000000

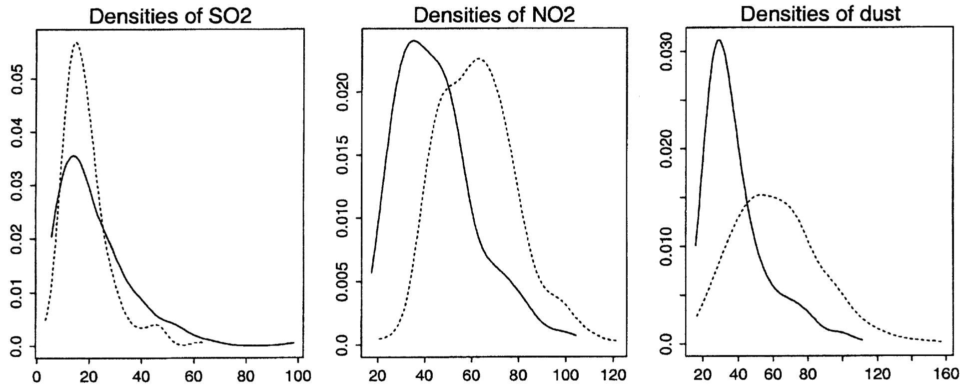

FIG. 2. *Density estimation for the distributions of pollutants. Solid curves are for the summer season and dashed curves are for the winter season.*

is fitted to the given data. The two-step method is employed. An initial bandwidth $h_0 = 0.06 * 729$ (six percent of the whole interval) was chosen. As anticipated, the results do not alter much with different choices of initial bandwidths. The second-stage bandwidths h_2 were chosen, respectively, 25%, 25%, 30% and 30% of the interval length for the functions $a_1, \dots, a_4$. Figure 3 depicts the estimated coefficient functions. They describe the extent to which the coefficients vary with time. Two short-dashed curves indicate pointwise 95% confidence intervals with bias ignored. The standard errors are computed from the second-stage local cubic regression. See Section 4.3 of Fan and Gijbels (1996) on how to compute the estimated standard errors from local polynomial regression. The figure shows that there is strong time effect on the coefficient functions. For comparison purposes, in Figure 3 we also superimpose the estimates (long-dashed curves) using the one-step procedure with bandwidths 25%, 25%, 30% and 30% of the time interval for $a_1, \dots, a_4$, respectively.

To compare the performance between the one-step and two-step methods, we define the relative efficiency between the one-step and the two-step methods via computing

$$\{\mathrm{RSS}_h(\text{one-step}) - \mathrm{RSS}_h(\text{two-step})\}/\mathrm{RSS}_h(\text{two-step}),$$

where RSS_h (one-step) and RSS_h (two-step) are the residual sum of squares for the one-step procedure using the bandwidth h and the two-step method using the same bandwidth h in the second stage, respectively. Figure 4a shows that the two-step method has smaller RSS than that of the one-step method. The gain is more pronounced as the bandwidth increases. This can be intuitively explained as follows. As bandwidth increases, at least one of the components would have nonnegligible biases.

Pollutants may not have an immediate effect on circulatory and respiratory systems. A natural question arises if there is any time lag in the response variable. To study this question, we fit model (5.1) for each time lag τ using

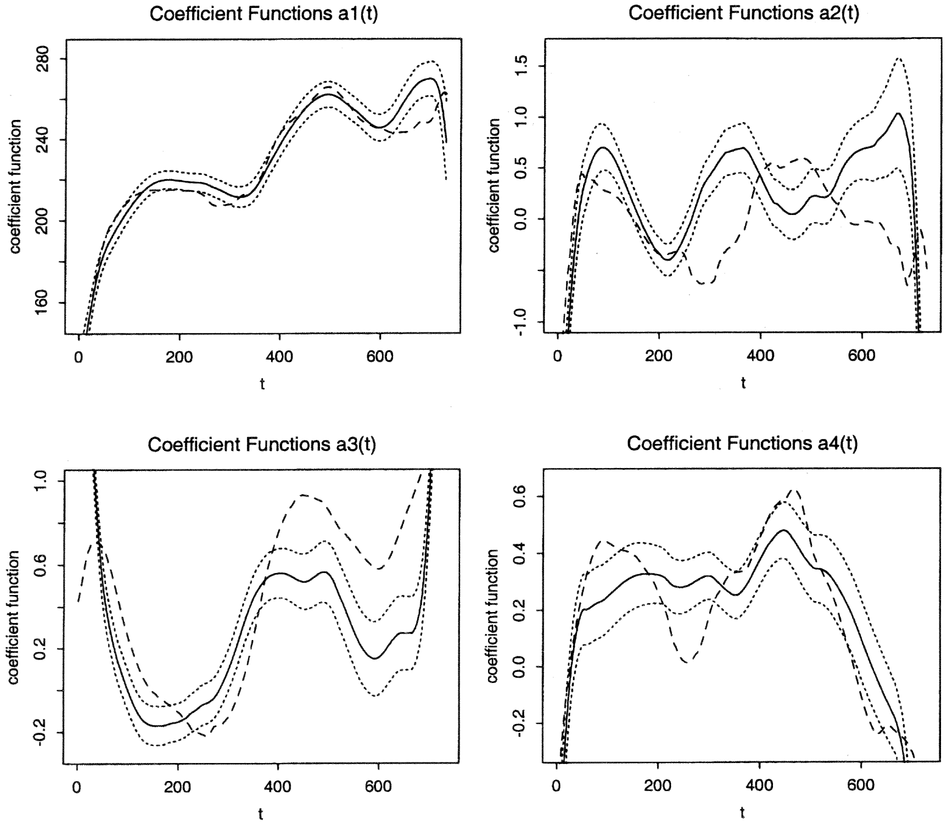

FIG. 3. *The estimated coefficient functions. The solid- and long-dashed curves are for the two-step and one-step methods, respectively. Two short-dashed curves indicate pointwise* 95% *confidence intervals with bias ignored.*

the data

$$\{Y(t+\tau), X_2(t), X_3(t), X_4(t), t = 1, 2, \dots, \}.$$

Figure 4b presents the resulting residuals sum of squares for each time lag. As τ gets larger, so does the residuals sum of squares. This in turn suggests no evidence for time delay in the response variable. We now examine how the expected number of hospital admissions changes, over time, when pollutants levels are set at their averages. Namely, we plot the function

$$\hat{Y}(t) = \hat{a}_1(t) + \hat{a}_2(t)\bar{X}_2 + \hat{a}_3(t)\bar{X}_3 + \hat{a}_4(t)\bar{X}_4$$

against t, where the estimated coefficient functions were obtained by the two-step approach. Figure 4c presents the result. It indicates an overall increasing trend in the number of hospital admissions for respiratory and circulatory problems. A seasonal pattern can also be seen. These features are not available in the usual parametric least-squares models.

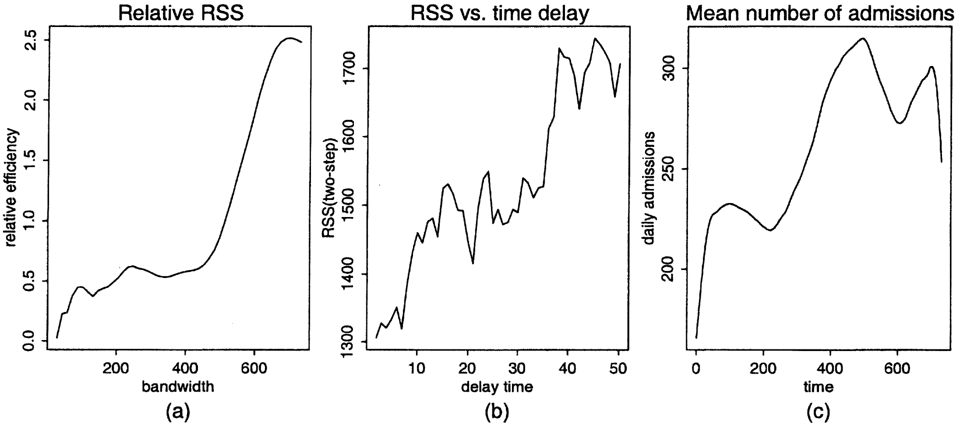

FIG. 4. (a) *Comparing the relative efficiency between the one-step and the two-step method.* (b) *Testing if there is any time delay in the response variable.* (c) *The expected number of hospital admissions over time when pollutant levels are set at their averages.*

5.2. *Simulations.* We use the following three examples to illustrate the performance of our method:

Example 1. $Y = \sin(60U)X_1 + 4U(1-U)X_2 + \varepsilon.$
Example 2. $Y = \sin(6\pi U)X_1 + \sin(2\pi U)X_2 + \varepsilon.$
Example 3. $Y = \sin\big(8\pi(U-0.5)\big)X_1$

$$+\big(3.5\big[\exp\{-(4U-1)^2\} + \exp\{-(4U-3)^2\}\big] - 1.5\big)X_2 + \varepsilon,$$

where U follows a uniform distribution on [0, 1] and X_1 and X_2 are normally distributed with correlation coefficient $2^{-1/2}$. Further, the marginal distributions of X_1 and X_2 are the standard normal and ε, U and (X_1, X_2) are independent. The random variable ε follows a normal distribution with mean zero and variance σ^2. The σ^2 is chosen so that the signal-to-noise ratio is about $5:1$, namely,

$$\sigma^2 = 0.2\,\mathrm{var}\big\{m(U, X_1, X_2)\big\} \quad \text{with } m(U, X_1, X_2) = E(Y \mid U, X_1, X_2).$$

Figure 5 shows the varying coefficient functions a_1 and a_2 for Examples 1–3.

For each of the above examples, we conducted 100 simulations with sample size $n = 250$, 500, 1000. Mean integrated squared errors for estimating a_2 are recorded. For the one-step procedure, we plot the MISE against h_1 and hence the optimal bandwidth can be chosen. For the two-step procedure, we choose some small initial bandwidth h_0 and then compute the MISE for the two-step estimator as a function of h_2. Specifically, we choose $h_0 = 0.03$, 0.04 and 0.05, respectively, for Examples 1, 2 and 3. The optimal bandwidths h_1 and h_2 were used to compute the resulting estimators presented in Figure 1. Among 100 samples, we select the one such that the two-step estimator attains the median performance. Once the sample is selected, the one-step estimate and

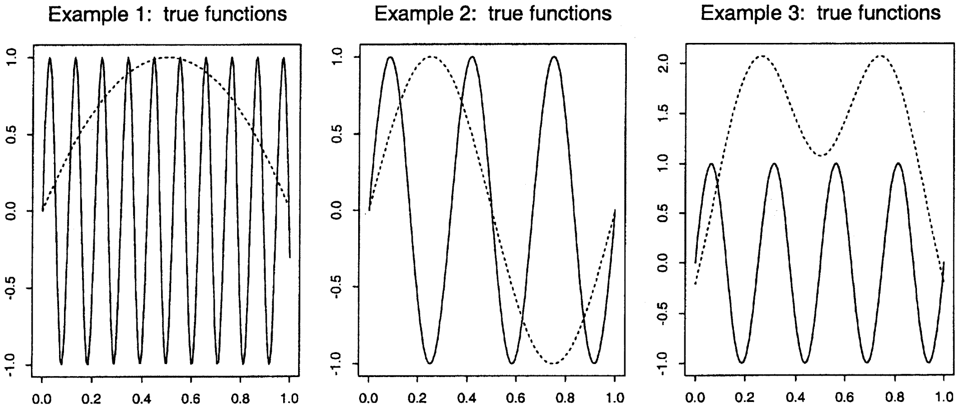

FIG. 5. *Varying coefficient functions. Solid curves are for $a_1(\cdot)$ and dashed curves are for $a_2(\cdot)$.*

the two-step estimate are computed. Figure 1 depicts the resulting estimate based on $n = 500$.

Figure 6 depicts the MISE as a function of bandwidth. The MISE curves for the two-step method are always smaller than those for the one-step approach for the three examples that we tested. This is in line with our asymptotic theory that the two-step approach outperforms the one-step procedure if the initial bandwidth is correctly chosen. The improvement of the two-step estimator is quite substantial if the optimal bandwidth is used (in comparison with the one-step approach using the optimal bandwidth). Further, for the two-step estimator, the MISE curve is flatter than that for the one-step method. This is turn reveals that the bandwidth for the two-step estimator is less crucial than that for the one-step procedure. This is an extra benefit of the two-step procedure.

6. Proofs. The proof of Theorem 3 (and Theorem 4) is similar to that of Theorem 1 (and Theorem 2). Thus, we only prove Theorems 1 and 2. When the asymptotic conditional bias and variance are calculated for the two-step procedure $\hat{a}_{p,\mathrm{TS}}(u_0)$, the following lemma on the uniform convergence will be used.

LEMMA 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d random vectors, where the Y_i's are scalar random variables. Assume further that $E|y|^3 < \infty$ and $\sup_x \int |y|^s f(x, y)\,dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_{x\in D}\left|n^{-1}\sum_{i=1}^{n}\{K_h(X_i - x)Y_i - E[K_h(X_i - x)Y_i]\}\right| = O_P[\{nh/\log(1/h)\}^{-1/2}]$$

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

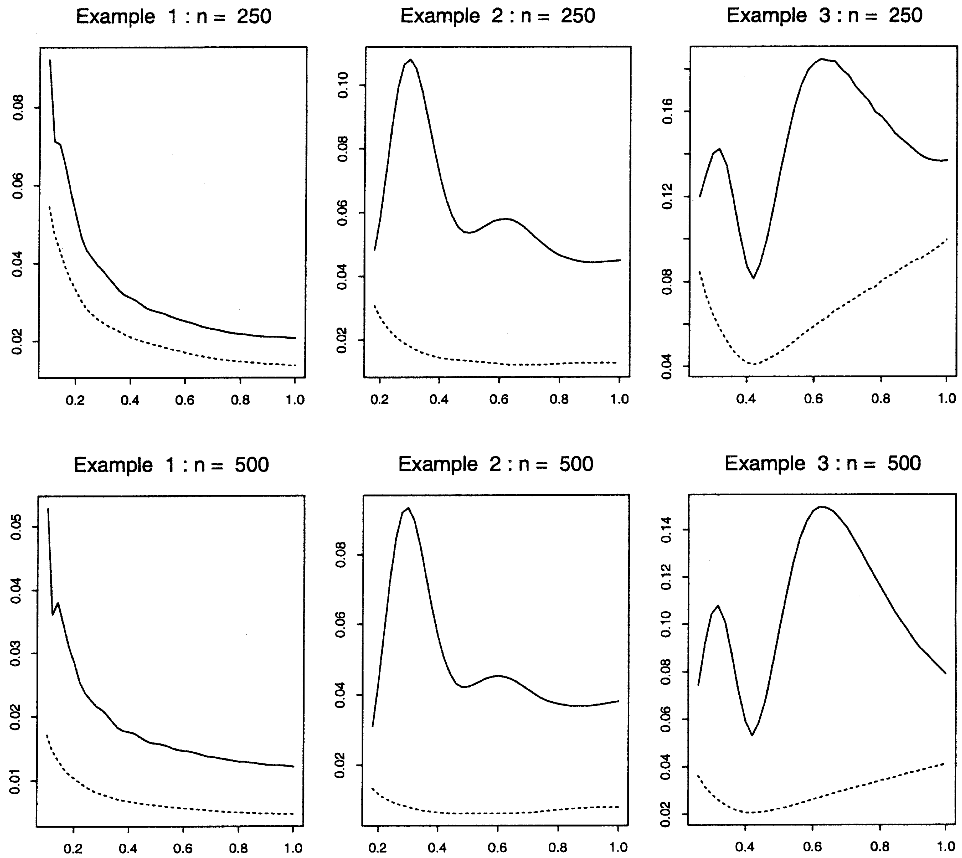

FIG. 6. *MISE as a function of bandwidth. Solid curve: one-step procedure; dashed curve: two-step procedure.*

PROOF. This follows immediately from the result obtained by Mack and Silverman (1982).

The following notation will be used in the proof of the theorems. Let

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix}$$

with

$$S_{11} = \Omega_{p-1} \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix}, \qquad S_{12} = \alpha_p \otimes \begin{pmatrix} \mu & 0 & \mu_2 & 0 \\ 0 & \mu_2 & 0 & \mu_4 \end{pmatrix}$$

and

$$S_{22} = r_{pp} \begin{pmatrix} \mu_0 & 0 & \mu_2 & 0 \\ 0 & \mu_2 & 0 & \mu_4 \\ \mu_2 & 0 & \mu_4 & 0 \\ 0 & \mu_4 & 0 & \mu_6 \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product. Let $\tilde{S}$ be the matrix similar to S except replacing μ_i by ν_i. Set

$$S^*_{(i)} = \Omega_p(U_i) \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix}, \qquad S^*_{(0)} = S^*_{(i)}|_{U_i=u_0}, \qquad Q = \Omega_p \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\beta^T_{(i)} = \sum_{j=1}^{p} a''_j(U_i)\mu_2(\alpha^T_j(U_i), r_{pj}(U_i)) \otimes (1, 0), \qquad \alpha^{*T} = (\alpha^T_p, r_{pp}) \otimes (1, 0).$$

Put

$$A = I_{p-1} \otimes \begin{pmatrix} 1 & 0 \\ 0 & h_1 \end{pmatrix}, \qquad G = I_p \otimes \begin{pmatrix} 1 & 0 \\ 0 & h_0 \end{pmatrix}$$

and

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & h_1 & 0 & 0 \\ 0 & 0 & h_1^2 & 0 \\ 0 & 0 & 0 & h_1^3 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & h_2 & 0 & 0 \\ 0 & 0 & h_2^2 & 0 \\ 0 & 0 & 0 & h_1^3 \end{pmatrix}.$$

We are now ready to prove our results.

PROOF OF THEOREM 1. First, let us calculate the asymptotic conditional bias of $\hat{a}_{p,1}(u_0)$. Note that by Taylor's expansion, we have

$$\begin{aligned} Y = \mathbf{X}_1\Big(&a_1(u_0), a'_1(u_0), \ldots, a_{p-1}(u_0), a'_{p-1}(u_0), a_p(u_0), a'_p(u_0), \\ &\frac{1}{2}a''_p(u_0), \frac{1}{3!}a'''_p(u_0)\Big)^T \\ &+ \frac{1}{2}\sum_{j=1}^{p-1} \begin{pmatrix} a''_j(\xi_{1j})(U_1 - u_0)^2 X_{1j} \\ \vdots \\ a''_j(\xi_{nj})(U_n - u_0)^2 X_{nj} \end{pmatrix} + \frac{1}{4!} \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1 - u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n - u_0)^4 X_{np} \end{pmatrix} + \boldsymbol{\varepsilon} \end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_n)^T$, ξ_{ij} and η_i are between U_i and u_0 for $i = 1, \ldots n$, $j = 1, \ldots, p-1$. Thus,

$$\begin{aligned} \hat{a}_{p,1}(u_0) = a_p(u_0) \\ + \tfrac{1}{2}\sum_{j=1}^{p-1} e^T_{2p-1,\,2p+2}(\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1}\mathbf{X}_1^T W_1 \begin{pmatrix} a''_j(\xi_{1j})(U_1 - u_0)^2 X_{1j} \\ \vdots \\ a''_j(\xi_{nj})(U_n - u_0)^2 X_{nj} \end{pmatrix} \end{aligned}$$

$$+\ \tfrac{1}{4!}e^T_{2p-1,\,2p+2}\big(\mathbf{X}_1^T W_1 \mathbf{X}_1\big)^{-1}\mathbf{X}_1^T W_1 \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1-u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n-u_0)^4 X_{np} \end{pmatrix}$$

$$+\ e^T_{2p-1,\,2p+2}\big(\mathbf{X}_1^T W_1 \mathbf{X}_1\big)^{-1}\mathbf{X}_1^T W_1 \boldsymbol{\varepsilon}.$$

Obviously,

$$\mathbf{X}_1^T W_1 \mathbf{X}_1 = \begin{pmatrix} \mathbf{X}_3^T W_1 \mathbf{X}_3 & \mathbf{X}_3^T W_1 \mathbf{X}_2 \\ \mathbf{X}_2^T W_1 \mathbf{X}_3 & \mathbf{X}_2^T W_1 \mathbf{X}_2 \end{pmatrix}.$$

By calculating the mean and variance, one can easily get

$$\mathbf{X}_3^T W_1 \mathbf{X}_3 = nf(u_0) A S_{11} A (1+o_P(1)),$$

$$\mathbf{X}_3^T W_1 \mathbf{X}_2 = nf(u_0) A S_{12} D (1+o_P(1))$$

and

$$\mathbf{X}_2^T W_1 \mathbf{X}_2 = nf(u_0) D S_{22} D (1+o_P(1)). \tag{6.1}$$

Combining the last three asymptotic expressions leads to

$$\mathbf{X}_1^T W_1 \mathbf{X}_1 = nf(u_0)\operatorname{diag}(A, D) S \operatorname{diag}(A, D)(1+o_P(1)).$$

Similarly, we have

$$\mathbf{X}_3^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix}$$

$$= nf(u_0) h_1^2 a_j''(u_0) A\big(\alpha_j \otimes (1,0)^T\big)\mu_2 (1+o_P(1))$$

and

$$\mathbf{X}_2^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix} = nf(u_0) h_1^2 a_j''(u_0) D \begin{pmatrix} r_{pj}\mu_2 \\ 0 \\ r_{pj}\mu_4 \\ 0 \end{pmatrix} (1+o_P(1)).$$

Thus,

$$\mathbf{X}_1^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix}$$

$$= nf(u_0) h_1^2 a_j''(u_0)\operatorname{diag}(A, D)$$

$$\times \big(\alpha_j^T \otimes (1,0)\mu_2, r_{pj}\mu_2, 0, r_{pj}\mu_4, 0\big)^T (1+o_P(1)).$$

So the asymptotic conditional bais of $\hat{a}_{p,1}(u_0)$ is given by

$$\operatorname{bias}\big(\hat{a}_{p,1}(u_0) \mid \mathscr{D}\big)$$
$$= \tfrac{1}{2}h_1^2 \sum_{j=1}^{p-1} a_j''(u_0) e_{2p-1,\,2p+2}^T S^{-1}(\alpha_j^T \otimes (1,0)\mu_2, r_{pj}\mu_2, 0, r_{pj}\mu_4, 0)^T(1+o_P(1)).$$

Using the properties of the Kronecker product we have

$$\begin{aligned}
&\operatorname{bias}(\hat{a}_{p,1}(u_0) \mid \mathscr{D})\\
&\quad= \frac{h_1^2\mu_2}{2(r_{pp} - \alpha_p^T\Omega_{p-1}^{-1}\alpha_p)r_{pp}}\\
&\qquad \times \sum_{j=1}^{p-1}\big(r_{pj}\alpha_p^T\Omega_{p-1}^{-1}\alpha_p - r_{pp}\alpha_p^T\Omega_{p-1}^{-1}\alpha_j\big)a_j''(u_0)(1+o_P(1))\\
&\quad= -\frac{h_1^2\mu_2}{2r_{pp}}\sum_{j=1}^{p-1} r_{pj}a_j''(u_0) + o_P(h_1^2).
\end{aligned}$$

We now calculate the asymptotic variance. Using an asymptotic argument similar to the above, it is easy to calculate that the asymptotic conditional variance of $\hat{a}_{p,1}(u_0)$ is given by

$$\begin{aligned}
&\operatorname{var}\big(\hat{a}_{p,1}(u_0) \mid \mathscr{D}\big)\\
&\quad= e_{2p-1,\,2p+2}^T(\mathbf{X}_1^T W_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T W_1 \Psi W_1 \mathbf{X}_1(\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} e_{2p-1,\,2p+2}\\
&\quad= \frac{\sigma^2(u_0)}{nh_1 f(u_0)} e_{2p-1,\,2p+2}^T S^{-1}\tilde{S}S^{-1} e_{2p-1,\,2p+2}(1+o_P(1)).
\end{aligned}$$

By using the properties of the Kronecker product, it follows that

$$\operatorname{var}\big(\hat{a}_{p,1}(u_0) \mid \mathscr{D}\big) = \frac{\sigma^2(u_0)(\lambda_2 r_{pp} + \lambda_3\alpha_p^T\Omega_{p-1}^{-1}\alpha_p)}{nh_1 f(u_0)\lambda_1 r_{pp}(r_{pp} - \alpha_p^T\Omega_{p-1}^{-1}\alpha_p)}(1+o_P(1)).$$

where $\lambda_1 = (\mu_4 - \mu_2^2)^2$, $\lambda_2 = \nu_0\mu_4^2 - 2\nu_2\mu_2\mu_4 + \mu_2^2\nu_4$, $\lambda_3 = 2\mu_2\nu_2\mu_4 - 2\nu_0\mu_2^2\mu_4 - \mu_2^2\nu_4 + \nu_0\mu_2^4$. This establishes the result in Theorem 1. □

PROOF OF THEOREM 2. We first compute the asymptotic conditional bias. Note that by Taylor's expansion, one obtains

$$\begin{aligned}
Y &= \mathbf{X}_{(i)}(a_1(U_i), a_1'(U_i), \ldots, a_p(U_i), a_p'(U_i))^T\\
&\quad + \tfrac{1}{2}\sum_{j=1}^{p}\begin{pmatrix} a_j''(\xi_{1j})(U_1 - U_i)^2 X_{1j}\\ \vdots \\ a_j''(\xi_{nj})(U_n - U_i)^2 X_{nj}\end{pmatrix} + \boldsymbol{\varepsilon}\\
&= \mathbf{X}_{(i)}(a_1(U_i), a_1'(U_i), \ldots, a_p(U_i), a_p'(U_i))^T
\end{aligned}$$

$$
+\frac{1}{2}\sum_{j=1}^{p}\begin{pmatrix} a_j''(U_i)(U_1-U_i)^2X_{1j} \\ \vdots \\ a_j''(U_i)(U_n-U_i)^2X_{nj} \end{pmatrix}
$$

$$
+\frac{1}{2}\sum_{j=1}^{p}\begin{pmatrix} (a_j''(\xi_{1j})-a_j''(U_i))(U_1-U_i)^2X_{1j} \\ \vdots \\ (a_j''(\xi_{nj})-a_j''(U_i))(U_n-U_i)^2X_{nj} \end{pmatrix}+\boldsymbol{\varepsilon},
$$

where ξ_{kj} is between U_i and U_k. Thus, for $l=1,\ldots,p-1$,

$$
\begin{aligned}
\hat{a}_{l,0}(U_i) &= a_l(U_i) \\
&\quad + \frac{1}{2}e_{2l-1,2p}^T(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T W_{(i)} \sum_{j=1}^{p}\begin{pmatrix} a_j''(U_i)(U_1-U_i)^2X_{1j} \\ \vdots \\ a_j''(U_i)(U_n-U_i)^2X_{nj} \end{pmatrix} \\
&\quad + \frac{1}{2}e_{2l-1,2p}^T(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T W_{(i)} \\
&\quad \times \sum_{j=1}^{p}\begin{pmatrix} (a_j''(\xi_{1j})-a_j''(U_i))(U_1-U_i)^2X_{1j} \\ \vdots \\ (a_j''(\xi_{nj})-a_j''(U_i))(U_n-U_i)^2X_{nj} \end{pmatrix} \\
&\quad + e_{2l-1,2p}^T(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T W_{(i)}\boldsymbol{\varepsilon}.
\end{aligned}
$$

By Lemma 1, we have

$$
\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)} = nf(U_i)GS_{(i)}^*G(1+o_P(1)) \tag{6.2}
$$

and

$$
\mathbf{X}_{(i)}^T W_{(i)}\sum_{j=1}^{p}\begin{pmatrix} a_j''(U_i)(U_1-U_i)^2X_{1j} \\ \vdots \\ a_j''(U_i)(U_n-U_i)^2X_{nj} \end{pmatrix} = nf(U_i)h_0^2G\beta_{(i)}(1+o_P(1)). \tag{6.3}
$$

Note that in our applications below, we only consider those U_i's which are in a neighborhood of u_0. By the continuity assumption, the term $o_P(1)$ holds uniformly in i such that U_i falls in the neighborhood of u_0. Combining (6.2) and (6.3), we have

$$\frac{1}{2}e_{2l-1,\,2p}^T(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T W_{(i)} \sum_{j=1}^{p} \begin{pmatrix} a_j''(U_i)(U_1-U_i)^2 X_{1j} \\ \vdots \\ a_j''(U_i)(U_n-U_i)^2 X_{nj} \end{pmatrix}$$
$$= \frac{1}{2}h_0^2 e_{2l-1,\,2p}^T S_{(i)}^{*-1}\beta_{(i)}(1+o_P(1)).$$

Note that K has a bounded support. From the last expression and the uniform continuity of functions $a_j''(\cdot)$ in a neighborhood of u_0, it follows that

$$E(\hat{a}_{l,\,0}(U_i) - a_l(U_i) \mid \mathscr{D}) = \frac{1}{2}h_0^2 e_{2l-1,\,2p}^T S_{(i)}^{*-1}\beta_{(i)}(1+o_P(1)). \tag{6.4}$$

Since

$$\begin{pmatrix} Y_1 - \sum_{j=1}^{p-1}\hat{a}_{j,\,0}(U_1)X_{1j} \\ \vdots \\ Y_n - \sum_{j=1}^{p-1}\hat{a}_{j,\,0}(U_n)X_{nj} \end{pmatrix} = \begin{pmatrix} a_p(U_1)X_{1p} \\ \vdots \\ a_p(U_n)X_{np} \end{pmatrix}$$
$$+ \begin{pmatrix} \sum_{j=1}^{p-1}(a_j(U_1) - \hat{a}_{j,\,0}(U_1))X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1}(a_j(U_n) - \hat{a}_{j,\,0}(U_n))X_{nj} \end{pmatrix} + \boldsymbol{\varepsilon},$$

it follows from (3.3) that

$$\hat{a}_{p,\,2}(u_0) = a_p(u_0) + \frac{1}{4!}(1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2$$
$$\times \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1-u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n-u_0)^4 X_{np} \end{pmatrix}$$

$$
\begin{aligned}
&+(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} \sum_{j=1}^{p-1}\big(a_j(U_1)-\hat{a}_{j,0}(U_1)\big)X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1}\big(a_j(U_n)-\hat{a}_{j,0}(U_n)\big)X_{nj} \end{pmatrix} \\
&+(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \varepsilon \\
\equiv{}& a_p(u_0)+\tfrac{1}{4!}\tilde{J}_1+\tilde{J}_2+(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \varepsilon.
\end{aligned}
$$

By simple calculation we have

$$
\begin{aligned}
E(\tilde{J}_1 \mid \mathscr{D}) &= h_2^4 a_p^{(4)}(u_0)(1,0,0,0)S_{22}^{-1}\begin{pmatrix} r_{pp}\mu_4 \\ 0 \\ r_{pp}\mu_6 \\ 0 \end{pmatrix}(1+o_P(1)) \\
&= h_2^4 a_p^{(4)}(u_0)\left(\frac{\mu_4}{\mu_4-\mu_2^2}, 0, -\frac{\mu_2}{\mu_4-\mu_2^2}, 0\right)\begin{pmatrix} \mu_4 \\ 0 \\ \mu_6 \\ 0 \end{pmatrix}(1+o_P(1)) \\
&= \frac{\mu_4^2-\mu_2\mu_6}{\mu_4-\mu_2^2} h_2^4 a_p^{(4)}(u_0)(1+o_P(1)).
\end{aligned}
$$

By (6.4), we have

$$
\begin{aligned}
E(\tilde{J}_2 \mid \mathscr{D}) ={}& (1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 \\
&\times \begin{pmatrix} \sum_{j=1}^{p-1} E\big((a_j(U_1)-\hat{a}_{j,0}(U_1)) \mid \mathscr{D}\big)X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1} E((a_j(U_n)-\hat{a}_{j,0}(U_n)) \mid \mathscr{D})X_{nj} \end{pmatrix}
\end{aligned}
$$

$$= -\tfrac{1}{2}h_0^2(1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(1)}^{*-1}\beta_{(1)}X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(n)}^{*-1}\beta_{(n)}X_{nj} \end{pmatrix}$$

$$\times (1+o_P(1))$$

$$= -\frac{h_0^2}{2r_{pp}}\left(\frac{\mu_4}{\mu_4-\mu_2^2}, 0, -\frac{\mu_2}{\mu_4-\mu_2^2}, 0\right)\begin{pmatrix} \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj} \\ 0 \\ \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj}\mu_2 \end{pmatrix}$$

$$\times (1+o_P(1))$$

$$= -\frac{h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj}(1+o_P(1)).$$

Therefore, by (6.5) we obtain

$$\operatorname{bias}\big(\hat{a}_{p,2}(u_0)\mid \mathscr{D}\big) = \left(-\frac{h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj} + \frac{\mu_4^2-\mu_2\mu_6}{4!(\mu_4-\mu_2^2)}a_p^{(4)}(u_0)h_2^4\right)(1+o_P(1)).$$

By using the properties of the Kronecker product, we have

$$\operatorname{bias}\big(\hat{a}_{p,2}(u_0)\mid \mathscr{D}\big) = \left(\frac{1}{4!}\frac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 - \frac{\mu_2 h_0^2}{2r_{pp}}\sum_{j=1}^{p} a_j''(u_0)(\alpha_p^T, 0)\Omega_p^{-1}\begin{pmatrix}\alpha_j \\ r_{pj}\end{pmatrix}\right)$$

$$\times (1+o_P(1))$$

$$= \frac{1}{4!}\frac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 - \frac{\mu_2 h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} a_j''(u_0)r_{pj} + o_P(h_2^4+h_0^2).$$

This proves the bias expression in Theorem 2.

We now calculate the asymptotic variance. Recall B_n defined at the end of Section 3. Denote by $H = I - B_n$. By (3.4), we have

$$
\begin{aligned}
&\operatorname{var}\big(\hat{a}_{p,2}(u_0) \mid \mathscr{D}\big) \\
&\quad = (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \Psi W_2 \mathbf{X}_2 \big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1} \\
&\qquad \times (1,0,0,0)^T \\
&\qquad - 2(1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 H \Psi W_2 \mathbf{X}_2 \big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1} \qquad (6.5)\\
&\qquad \times (1,0,0,0)^T \\
&\qquad + (1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 H \Psi H^T W_2 \mathbf{X}_2 \big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1} \\
&\qquad \times (1,0,0,0)^T.
\end{aligned}
$$

Using similar arguments as before, we can show that

$$
\begin{aligned}
&(1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 \Psi W_2 \mathbf{X}_2 \big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1}(1,0,0,0)^T \\
&\quad = \frac{\mu_4^2 \nu_0 - 2\mu_4 \mu_2 \nu_2 + \mu_2^2 \nu_4}{n h_2 f(u_0) r_{pp} \big(\mu_4 - \mu_2^2\big)^2} \sigma^2(u_0)(1 + o_P(1)). \qquad (6.6)
\end{aligned}
$$

Since

$$
H \Psi W_2 \mathbf{X}_2 = \sum_{j=1}^{p-1} \begin{pmatrix} X_{1j} e_{2j-1,2p}^T \big(\mathbf{X}_{(1)}^T W_{(1)} \mathbf{X}_{(1)}\big)^{-1} \mathbf{X}_{(1)}^T W_{(1)} \Psi W_2 \mathbf{X}_2 \\ \vdots \\ X_{nj} e_{2j-1,2p}^T \big(\mathbf{X}_{(n)}^T W_{(n)} \mathbf{X}_{(n)}\big)^{-1} \mathbf{X}_{(n)}^T W_{(n)} \Psi W_2 \mathbf{X}_2 \end{pmatrix}
$$

by Lemma 1, we have

$$
\mathbf{X}_{(i)}^T W_{(i)} \Psi W_2 \mathbf{X}_2 = n f(U_i) \sigma^2(U_i) K_{h_2}(U_i - u_0) G T_{2p\times 4,(i)} D_2 (1 + o_P(1)),
$$

where

$$
T_{2p\times 4,(i)} = \big(\tilde{u}_{k,l,(i)}\big)_{2p\times 4}, \qquad 1 \le k \le 2p,\ 0 \le l \le 3
$$

for $k = 1, \dots, p$,

$$
\begin{aligned}
\tilde{u}_{2k-1,0,(i)} &= r_{kp}(U_i), \qquad \tilde{u}_{2k-1,1,(i)} = r_{kp}(U_i)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1), \\
\tilde{u}_{2k-1,2,(i)} &= r_{kp}(U_i)\left(\frac{U_i - u_0}{h_2}\right)^2 + o_P(1)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1), \\
\tilde{u}_{2k-1,3,(i)} &= r_{kp}(U_i), \left(\frac{U_i - u_0}{h_2}\right)^3 + o_P(1)\left(\frac{U_i - u_0}{h_2}\right)^2 \\
&\qquad + o_P(1)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1),
\end{aligned}
$$

$$\tilde{u}_{2k,\,0,\,(i)} = o_P(1), \qquad \tilde{u}_{2k,\,1,\,(i)} = o_P(1)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1),$$

$$\tilde{u}_{2k,\,2,\,(i)} = o_P(1)\left(\frac{U_i - u_0}{h_2}\right)^2 + o_P(1)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1)$$

and

$$\tilde{u}_{2k,\,3,\,(i)} = o_P(1)\left(\frac{U_i - u_0}{h_2}\right)^3 + o_P(1)\left(\frac{U_i - u_0}{h_2}\right)^2 + o_P(1)\left(\frac{U_i - u_0}{h_2}\right) + o_P(1).$$

Thus, we obtain

$$\mathbf{X}_2^T W_2 H \Psi W_2 \mathbf{X}_2 = \frac{nf(u_0)}{h_2} \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1} \alpha^* r_{pj} \sigma^2(u_0) D_2 \times \begin{pmatrix} \nu_0 & 0 & \nu_2 & 0 \\ 0 & \nu_2 & 0 & \nu_4 \\ \nu_2 & 0 & \nu_4 & 0 \\ 0 & \nu_4 & 0 & \nu_6 \end{pmatrix} D_2(1 + o_P(1)).$$

This and (6.1) together yield

$$\begin{aligned} &(1,0,0,0)\big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1} \mathbf{X}_2^T W_2 H \Psi W_2 \mathbf{X}_2 \big(\mathbf{X}_2^T W_2 \mathbf{X}_2\big)^{-1} (1,0,0,0)^T \\ &\quad = \frac{\mu_4^2 \nu_0 - 2\mu_4 \mu_2 \nu_2 + \mu_2^2 \nu_4}{n h_2 f(u_0) r_{pp}^2 \big(\mu_4 - \mu_2^2\big)^2} \\ &\qquad \times \sum_{j=1}^{p-1} e_{2j-1,\,2p}^T S_{(0)}^{*-1} \alpha^* r_{pj} \sigma^2(u_0)(1 + o_P(1)). \end{aligned} \tag{6.7}$$

Let

$$X_{k(i)} = \big(X_{k1}, X_{k1}(U_k - U_i), \ldots, X_{kp}, X_{kp}(U_k - U_i)\big)^T$$

and

$$\mathbf{X}_{(i)}^T = \big(X_{1(i)}, X_{2(i)}, \ldots, X_{n(i)}\big).$$

Then we have

$$\mathbf{X}_2^T W_2 H \Psi H^T W_2 \mathbf{X}_2 = (v_{rs})_{4\times 4}, \qquad 0 \le r, s \le 3,$$

where

$$
\begin{aligned}
v_{rs} &= \sum_{i=1}^{n}\sum_{l=1}^{n}\bigg\{X_{ip}X_{lp}(U_i-u_0)^r(U_l-u_0)^s K_{h_2}(U_i-u_0)K_{h_2}(U_l-u_0)\\
&\qquad\times\sum_{j=1}^{p-1}X_{ij}e_{2j-1,\,2p}^T\big(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)}\big)^{-1}\mathbf{X}_{(i)}^T W_{(i)}\Psi\\
&\qquad\times\bigg(\sum_{m=1}^{p-1}X_{lm}e_{2m-1,\,2p}^T\big(\mathbf{X}_{(l)}^T W_{(l)}\mathbf{X}_{(l)}\big)^{-1}\mathbf{X}_{(l)}^T W_{(l)}\bigg)^T\bigg\}\\
&= \sum_{j=1}^{p-1}\sum_{m=1}^{p-1}\sum_{k=1}^{n}\bigg\{\sum_{i=1}^{n}X_{ip}X_{ij}(U_i-u_0)^r e_{2j-1,\,2p}^T\big(\mathbf{X}_{(i)}^T W_{(i)}\mathbf{X}_{(i)}\big)^{-1}X_{k(i)}K_{h_2}\\
&\qquad\qquad\times(U_i-u_0)K_{h_0}(U_k-U_i)\sigma^2(U_k)\bigg\}\\
&\quad\times\bigg\{\sum_{l=1}^{n}X_{lp}X_{lm}(U_l-u_0)^s X_{k(l)}^T\big(\mathbf{X}_{(l)}^T W_{(l)}\mathbf{X}_{(l)}\big)^{-1}e_{2m-1,\,2p}K_{h_2}\\
&\qquad\qquad\times(U_l-u_0)K_{h_0}(U_k-U_l)\bigg\}.
\end{aligned}
$$

Using Lemma 1 and tedious calculation, we obtain

$$
\begin{aligned}
&\mathbf{X}_2^T W_2 H\Psi H^T W_2\mathbf{X}_2\\
&\quad=\frac{nf(u_0)\sigma^2(u_0)}{h_2}\sum_{j=1}^{p-1}\sum_{m=1}^{p-1}r_{pj}r_{pm}e_{2j-1,\,2p}^T S_{(0)}^{*-1}QS_{(0)}^{*-1}e_{2m-1,\,2p}D_2\\
&\qquad\times\begin{pmatrix}\nu_0 & 0 & \nu_2 & 0\\ 0 & \nu_2 & 0 & \nu_4\\ \nu_2 & 0 & \nu_4 & 0\\ 0 & \nu_4 & 0 & \nu_6\end{pmatrix}D_2(1+o_P(1)).
\end{aligned}
$$

The combination of this and (6.1) gives

$$
\begin{aligned}
&(1,0,0,0)\big(\mathbf{X}_2^T W_2\mathbf{X}_2\big)^{-1}\mathbf{X}_2^T W_2 H\Psi H^T W_2\mathbf{X}_2\big(\mathbf{X}_2^T W_2\mathbf{X}_2\big)^{-1}(1,0,0,0)^T\\
&\quad=\frac{\mu_4^2\nu_0-2\mu_4\mu_2\nu_2+\mu_2^2\nu_4}{nh_2 f(u_0)r_{pp}^2(\mu_4-\mu_2^2)^2}\\
&\qquad\times\sum_{j=1}^{p-1}\sum_{m=1}^{p-1}r_{pj}r_{pm}e_{2j-1,\,2p}^T S_{(0)}^{*-1}QS_{(0)}^{*-1}e_{2m-1,\,2p}\sigma^2(u_0)(1+o_P(1)).
\end{aligned}
\tag{6.8}
$$

Substituting (6.7)–(6.9) into (6.6), we have

$$\begin{aligned}\operatorname{var}\big(\hat{a}_{p,2}(u_0)\mid \mathscr{D}\big) &= \frac{\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4}{nh_2 f(u_0) r_{pp}^2(\mu_4-\mu_2^2)^2}\\ &\quad\times\Bigg(r_{pp} + \sum_{j=1}^{p-1}\sum_{m=1}^{p-1} r_{pj}r_{pm} e_{2j-1,2p}^T S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2m-1,2p}\\ &\qquad\qquad - 2\sum_{j=1}^{p-1} r_{pj} e_{2j-1,2p}^T S_{(0)}^{*-1}\alpha^*\Bigg)\\ &\quad\times \sigma^2(u_0)(1+o_P(1)).\end{aligned}$$

Using the properties of the Kronecker product we get

$$\begin{aligned}&\operatorname{var}\big(\hat{a}_{p,2}(u_0)\mid \mathscr{D}\big)\\ &\quad= \frac{(\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4)\sigma^2(u_0)}{nh_2 f(u_0) r_{pp}^2(\mu_4-\mu_2^2)^2}\\ &\qquad\times\left(r_{pp} + r_{pp}^2 e_{p,p}^T\Omega_p^{-1}e_{p,p} - (\alpha_p^T, r_{pp})\Omega_p^{-1}\begin{pmatrix}\alpha_p\\ r_{pp}\end{pmatrix}\right)(1+o_P(1)).\end{aligned}$$

Note that $\Omega_p^{-1}(\alpha_p^T, r_{pp})^T = e_{p,p}$. This results in Theorem 2. □

Acknowledgment. The authors thank an Associate Editor for helpful comments that led to improvement of the presentation of the paper.

REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

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

CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1991). Local regression models. In *Statistical Models in S* (J. M. Chambers, and T. J. Hastie, eds.) 309–376. Wadsworth / Brooks-Cole, Pacific Grove, CA.

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

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1998). Direct estimation of additive and linear components for high dimensional data. *Ann. Statist.* **26** 943–971.

FAN, J. and ZHANG, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**. To appear.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

GU, C. and WAHBA, G. (1993). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals." *J. Comput. Graph. Statist*. **2** 97–117.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc*. **84** 986–995.

HASTIE, T. J. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

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

HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, **66** 1017–1098.

HOOVER, D. R., RICE, J. A., WU, C. O. and YANG, L. P. (1997). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc*. **86** 316–342.

MACK, Y. P., SILVERMAN, B. W. (1982). Weak and Strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc*. **92** 1049–1062.

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270.

SHUMWAY, R. H. (1988). *Apllied Staistical Time Series Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

STONE, C. J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist*. **25** 1371–1470.

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

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

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
E-MAIL: jfan@stat.ucla.edu

DEPARTMENT OF STATISTICS
CHINESE UNIVERSITY OF HONG KONG
SHATIN
HONG KONG