# STATISTICAL ESTIMATION IN VARYING COEFFICIENT MODELS 

By Jianqing Fan ${ }^{1}$ and Wenyang Zhang<br>University of North Carolina and Chinese University of Hong Kong<br>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)], lowdimensional 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

[^0]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:
\[

$$
\begin{equation*}
Y=\sum_{j=1}^{p} a_{j}(U) X_{j}+\varepsilon \tag{1.1}
\end{equation*}
$$

\]

for given covariates $\left(U, X_{1}, \ldots, X_{p}\right)^{\prime}$ and response variable $Y$ with

$$
E\left(\varepsilon \mid U, X_{1}, \ldots, X_{p}\right)=0
$$

and

$$
\operatorname{var}\left(\varepsilon \mid U, X_{1}, \ldots, X_{p}\right)=\sigma^{2}(U)
$$

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 $a_{1}, \ldots, a_{p}$ 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 $\left\{\left(U_{i}, X_{i 1}, \ldots, X_{i p}, Y_{i}\right)\right\}_{i=1}^{n}$ from model (1.1). One simple approach to estimate the coefficient functions $a_{j}(\cdot)(j=1, \ldots, p)$ is to use local linear modeling. For each given point $u_{0}$, approximate the function locally as

$$
\begin{equation*}
a_{j}(u) \approx a_{j}+b_{j}\left(u-u_{0}\right) \tag{1.2}
\end{equation*}
$$

for $u$ in a neighborhood of $u_{0}$. This leads to the following local least-squares problem: minimize

$$
\begin{equation*}
\sum_{i=1}^{n}\left[Y_{i}-\sum_{j=1}^{p}\left\{a_{j}+b_{j}\left(U_{i}-u_{0}\right)\right\} X_{i j}\right]^{2} K_{h}\left(U_{i}-u_{0}\right) \tag{1.3}
\end{equation*}
$$

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,

$$
\begin{equation*}
a_{p}(u) \approx a_{p}+b_{p}\left(u-u_{0}\right)+c_{p}\left(u-u_{0}\right)^{2}+d_{p}\left(u-u_{0}\right)^{3}, \tag{1.4}
\end{equation*}
$$

for $u$ in a neighborhood of $u_{0}$. This naturally leads to the following weighted least-squares problem:

$$
\begin{aligned}
\sum_{i=1}^{n} & {\left[Y_{i}-\sum_{j=1}^{p-1}\left\{a_{j}+b_{j}\left(U_{i}-u_{0}\right)\right\} X_{i j}\right.} \\
& \left.-\left\{a_{p}+b_{p}\left(U_{i}-u_{0}\right)+c_{p}\left(U_{i}-u_{0}\right)^{2}+d_{p}\left(U_{i}-u_{0}\right)^{3}\right\} X_{i p}\right]^{2} \\
& \times K_{h_{1}}\left(U_{i}-u_{0}\right)
\end{aligned}
$$

Let $\hat{a}_{j, 1}, \hat{b}_{j, 1}(j=1, \ldots, p-1)$ and $\hat{a}_{p, 1}, \hat{b}_{p, 1}, \hat{c}_{p, 1}, \hat{d}_{p, 1}$ minimize (1.5). The resulting estimator $\hat{a}_{p \text {, os }}\left(u_{0}\right)=\hat{a}_{p, 1}$ is called a one-step estimator. We will show that the bias of the one-step estimator is $O\left(h_{1}^{2}\right)$ and the variance of the one-step estimator is $O\left(\left(n h_{1}\right)^{-1}\right)$. Therefore, using the one-step estimator $\hat{a}_{p, \text { OS }}\left(u_{0}\right)$, the optimal rate $O\left(n^{-8 / 9}\right)$ 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 $a_{1}(\cdot), \ldots, a_{p-1}(\cdot)$. 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

$$
\begin{equation*}
\sum_{k=1}^{n}\left(Y_{k}-\sum_{j=1}^{p}\left\{a_{j}+b_{j}\left(U_{k}-u_{0}\right)\right\} X_{k j}\right)^{2} K_{h_{0}}\left(U_{k}-u_{0}\right) \tag{1.6}
\end{equation*}
$$

for a given initial bandwidth $h_{0}$ and kernel $K$. Let $\hat{a}_{1,0}\left(u_{0}\right), \ldots, \hat{a}_{p, 0}\left(u_{0}\right)$ denote the initial estimate of $a_{1}\left(u_{0}\right), \ldots, a_{p}\left(u_{0}\right)$. In the second step, we substitute the preliminary estimates $\hat{a}_{1,0}(\cdot), \ldots, \hat{a}_{p-1,0}(\cdot)$ and use a local cubic fit to estimate $a_{p}\left(u_{0}\right)$, namely, minimize

$$
\begin{align*}
& \sum_{i=1}^{n}\left(Y_{i}-\sum_{j=1}^{p-1} \hat{a}_{j, 0}\left(U_{i}\right) X_{i j}\right. \\
& \left.\quad-\left\{a_{p}+b_{p}\left(U_{i}-u_{0}\right)+c_{p}\left(U_{i}-u_{0}\right)^{2}+d_{p}\left(U_{i}-u_{0}\right)^{3}\right\} X_{i p}\right)^{2}  \tag{1.7}\\
& \quad \times K_{h_{2}}\left(U_{i}-u_{0}\right)
\end{align*}
$$

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 $\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right)$ of $a_{p}\left(u_{0}\right)$ is obtained. We will
show that the bias of the two-step estimator is of $O\left(h_{2}^{4}\right)$ and the variance $O\left\{\left(n h_{2}\right)^{-1}\right\}$, provided that

$$
h_{0}=o\left(h_{2}^{2}\right), \quad n h_{0} / \log h_{0} \rightarrow \infty
$$

and $n h_{0}^{3} \rightarrow \infty$. 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 $O\left(n^{-1 / 3}\right)$ and $O\left(n^{-2 / 9}\right)$, the optimal rates of convergence $O\left(n^{-8 / 9}\right)$ for estimating $a_{2}$ can be achieved.

Note that the condition $n h_{0}^{3} \rightarrow \infty$ 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 $n h^{1+\delta} \rightarrow$ $\infty$ 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):

$$
\sum_{i=1}^{n}\left[Y_{i}-\sum_{j=1}^{p} a_{j}\left(U_{i}\right) X_{i j}\right]^{2}+\sum_{j=1}^{p} \lambda_{j} \int\left\{a_{j}^{\prime \prime}(u)\right\}^{2} d u
$$

for some smoothing parameters $\lambda_{1}, \ldots, \lambda_{p}$. 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 $\hat{a}_{1,0}\left(U_{i}\right), \ldots, \hat{a}_{p, 0}\left(U_{i}\right)$ for $i=1, \ldots, n$. With these preliminary estimates, one can now estimate the coefficient functions $a_{j}\left(u_{0}\right)$ 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)], plugin 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, onestep and two-step estimators. Let

$$
\begin{aligned}
\mathbf{X}_{0} & =\left(\begin{array}{ccccc}
X_{11} & X_{11}\left(U_{1}-u_{0}\right) & \cdots & X_{1 p} & X_{1 p}\left(U_{1}-u_{0}\right) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
X_{n 1} & X_{n 1}\left(U_{n}-u_{0}\right) & \cdots & X_{n p} & X_{n p}\left(U_{n}-u_{0}\right)
\end{array}\right), \\
Y & =\left(Y_{1}, \ldots, Y_{n}\right)^{T} \quad \text { and } \\
W_{0} & =\operatorname{diag}\left(K_{h_{0}}\left(U_{1}-u_{0}\right), \ldots, K_{h_{0}}\left(U_{n}-u_{0}\right)\right) .
\end{aligned}
$$

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

$$
\begin{equation*}
\hat{a}_{j, 0}\left(u_{0}\right)=e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{0}^{T} W_{0} \mathbf{X}_{0}\right)^{-1} \mathbf{X}_{0}^{T} W_{0} Y, \quad j=1, \ldots, p . \tag{3.1}
\end{equation*}
$$

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

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

$$
\mathbf{X}_{2}=\left(\begin{array}{cccc}
X_{1 p} & X_{1 p}\left(U_{1}-u_{0}\right) & X_{1 p}\left(U_{1}-u_{0}\right)^{2} & X_{1 p}\left(U_{1}-u_{0}\right)^{3} \\
\vdots & \vdots & \vdots & \vdots \\
X_{n p} & X_{n p}\left(U_{n}-u_{0}\right) & X_{n p}\left(U_{n}-u_{0}\right)^{2} & X_{n p}\left(U_{n}-u_{0}\right)^{3}
\end{array}\right)
$$

and

$$
\begin{aligned}
& \mathbf{X}_{3}=\left(\begin{array}{ccccc}
X_{11} & X_{11}\left(U_{1}-u_{0}\right) & \cdots & X_{1(p-1)} & X_{1(p-1)}\left(U_{1}-u_{0}\right) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
X_{n 1} & X_{n 1}\left(U_{n}-u_{0}\right) & \cdots & X_{n(p-1)} & X_{n(p-1)}\left(U_{n}-u_{0}\right)
\end{array}\right) \\
& \mathbf{X}_{1}=\left(\mathbf{X}_{3}, \mathbf{X}_{2}\right), \quad W_{1}=\operatorname{diag}\left(K_{h_{1}}\left(U_{1}-u_{0}\right), \ldots, K_{h_{1}}\left(U_{n}-u_{0}\right)\right) .
\end{aligned}
$$

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

$$
\begin{equation*}
\hat{a}_{p, 1}\left(u_{0}\right)=e_{2 p-1,2 p+2}^{T}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} \mathbf{X}_{1}^{T} W_{1} Y \tag{3.2}
\end{equation*}
$$

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

$$
\begin{equation*}
\hat{a}_{p, 2}\left(u_{0}\right)=(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2}(Y-V) \tag{3.3}
\end{equation*}
$$

where

$$
W_{2}=\operatorname{diag}\left(K_{h_{2}}\left(U_{1}-u_{0}\right), \ldots, K_{h_{2}}\left(U_{n}-u_{0}\right)\right)
$$

and $V=\left(V_{1}, \ldots, V_{n}\right)^{T}$ with $V_{i}=\sum_{j=1}^{p-1} \hat{a}_{j, 0}\left(U_{i}\right) X_{i j}$. 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

$$
V_{i}=\sum_{j=1}^{p-1} X_{i j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} Y
$$

Set

$$
B_{n}=I_{n}-\sum_{j=1}^{p-1}\left(\begin{array}{c}
X_{1 j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(1)}^{T} W_{(1)} \mathbf{X}_{(1)}\right)^{-1} X_{(1)}^{T} W_{(1)} \\
\vdots \\
X_{n j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(n)}^{T} W_{(n)} \mathbf{X}_{(n)}\right)^{-1} \mathbf{X}_{(n)}^{T} W_{(n)}
\end{array}\right)
$$

Then

$$
\begin{equation*}
\hat{a}_{p, 2}\left(u_{0}\right)=(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} B_{n} Y \tag{3.4}
\end{equation*}
$$

4. Main results. We impose the following technical conditions:
5. $E X_{j}^{2 s}<\infty$, for some $s>2, j=1, \ldots, p$.
6. $a_{j}^{\prime \prime}(\cdot)$ is continuous in a neighborhood of $u_{0}$, for $j=1, \ldots, p$. Further, assume $\alpha_{j}^{\prime \prime}\left(u_{0}\right) \neq 0$, for $j=1, \ldots, p$.
7. The function $a_{p}$ has a continuous fourth derivative in a neighborhood of $u_{0}$.
8. $r_{i j}^{\prime \prime}(\cdot)$ is continuous in a neighborhood of $u_{0}$ and $r_{i j}^{\prime \prime}\left(u_{0}\right) \neq 0$, for $i, j=$ $1, \ldots, p$, where $r_{i j}(u)=E\left(X_{i} X_{j} \mid U=u\right)$.
9. The marginal density of $U$ has a continuous second derivative in some neighborhood of $u_{0}$ and $f\left(u_{0}\right) \neq 0$.
10. The function $K(t)$ is a symmetric density function with a compact support.
11. $h_{0} / h_{2} \rightarrow 0$ and $h_{2} \rightarrow 0, n h_{0}^{\gamma} / \log h_{0} \rightarrow \infty$, for any $\gamma>s /(s-2)$ with $s$ given in condition 1 .

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

$$
\mu_{i}=\int t^{i} K(t) d t \quad \text { and } \quad \nu_{i}=\int t^{i} K^{2}(t) d t
$$

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

$$
\mathscr{D}=\left(U_{1}, \ldots, U_{n}, X_{11}, \ldots, X_{1 n}, \ldots, X_{p 1}, \ldots, X_{p n}\right)^{T}
$$

Set $r_{i j}=r_{i j}\left(u_{0}\right)=E\left(X_{i} X_{j} \mid U=u_{0}\right)$, for $i, j=1, \ldots, p$. Put

$$
\begin{aligned}
\Psi & =\operatorname{diag}\left(\sigma^{2}\left(U_{1}\right), \ldots, \sigma^{2}\left(U_{n}\right)\right), \\
\alpha_{j}(u) & =\left(r_{1 j}(u), \ldots, r_{(p-1) j}(u)\right)^{T}, \\
\alpha_{j} & =\alpha_{j}\left(u_{0}\right) \text { for } j=1, \ldots, p
\end{aligned}
$$

and

$$
\begin{aligned}
\Omega_{i}(u) & =E\left\{\left(X_{1}, \ldots, X_{i}\right)^{T}\left(X_{1}, \ldots, X_{i}\right) \mid U=u\right\} \\
\Omega_{i} & =\Omega_{i}\left(u_{0}\right) \text { for } i=1, \ldots, p
\end{aligned}
$$

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

THEOREM 1. Under conditions $1-6$, if $h_{1} \rightarrow 0$ in such a way that $n h_{1} \rightarrow \infty$, then the asymptotic conditional bias of $\hat{a}_{p, \mathrm{OS}}\left(u_{0}\right)$ is given by

$$
\operatorname{bias}\left(\hat{a}_{p, \mathrm{OS}}\left(u_{0}\right) \mid \mathscr{D}\right)=-\frac{h_{1}^{2} \mu_{2}}{2 r_{p p}} \sum_{j=1}^{p-1} r_{p j} a_{j}^{\prime \prime}\left(u_{0}\right)+o_{P}\left(h_{1}^{2}\right)
$$

and the asymptotic conditional variance of $\hat{a}_{p, \mathrm{OS}}\left(u_{0}\right)$ is

$$
\operatorname{var}\left(\hat{a}_{p, \mathrm{OS}}\left(u_{0}\right) \mid \mathscr{D}\right)=\frac{\sigma^{2}\left(u_{0}\right)\left(\lambda_{2} r_{p p}+\lambda_{3} \alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}\right)}{n h_{1} f\left(u_{0}\right) \lambda_{1} r_{p p}\left(r_{p p}-\alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}\right)}\left(1+o_{p}(1)\right),
$$

where $\lambda_{1}=\left(\mu_{4}-\mu_{2}^{2}\right)^{2} \lambda_{2}=\nu_{0} \mu_{4}^{2}-2 \nu_{2} \mu_{2} \mu_{4}+\mu_{2}^{2} \nu_{4}$ 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 $\hat{a}_{p, \text { OS }}\left(u_{0}\right)$ is only $O_{P}\left\{h_{1}^{4}+\left(n h_{1}\right)^{-1}\right\}$ which achieves the rate $O_{P}\left(n^{-4 / 5}\right)$ when the bandwidth $h_{1}=O\left(n^{-1 / 5}\right)$ is used. The bias expression above indicates clearly that the approximation errors of functions $a_{1}, \ldots, a_{p-1}$ 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 $\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right)$ can be expressed as

$$
\begin{aligned}
& \operatorname{bias}\left(\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=\frac{1}{4!} \frac{\mu_{4}^{2}-\mu_{6} \mu_{2}}{\mu_{4}-\mu_{2}^{2}} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}-\frac{\mu_{2} h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p-1} a_{j}^{\prime \prime}\left(u_{0}\right) r_{p j}+o_{P}\left(h_{2}^{4}+h_{0}^{2}\right)
\end{aligned}
$$

and the asymptotic conditional variance of $\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right)$ is given by

$$
\operatorname{var}\left(\hat{\alpha}_{p, \mathrm{TS}}\left(u_{0}\right) \mid \mathscr{D}\right)=\frac{\left(\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}\right) \sigma^{2}\left(u_{0}\right)}{n h_{2} f\left(u_{0}\right)\left(\mu_{4}-\mu_{2}^{2}\right)^{2}} e_{p, p}^{T} \Omega_{p}^{-1} e_{p, p}\left\{1+o_{P}(1)\right\}
$$

By Theorem 2, the asymptotic variance of the two-step estimator is independent of the initial bandwidth as long as $n h_{0}^{\gamma} \rightarrow \infty$, where $\gamma$ is given in condition 7. Thus, the initial bandwidth $h_{0}$ should be chosen as small as possible subject to the constraint that $n h_{0}^{\gamma} \rightarrow \infty$. In particular, when $h_{0}=o\left(h_{2}^{2}\right)$, the bias from the initial estimator becomes negligible and the bias expression for the two-step estimator is

$$
\frac{1}{4!} \frac{\mu_{4}^{2}-\mu_{6} \mu_{2}}{\mu_{4}-\mu_{2}^{2}} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}+o_{P}\left(h_{2}^{4}\right) .
$$

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 $O_{P}\left(n^{-8 / 9}\right)$.

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

$$
\frac{1}{4!} \frac{\mu_{4}^{2}-\mu_{6} \mu_{2}}{\mu_{4}-\mu_{2}^{2}} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}+o_{P}\left(h_{2}^{4}\right)
$$

and asymptotic variance

$$
\frac{\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}}{n h_{2} f\left(u_{0}\right) r_{p p}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}} \sigma^{2}\left(u_{0}\right)+o_{P}\left\{\left(n h_{2}\right)^{-1}\right\}
$$

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 $\hat{\alpha}_{p, \text { os }}$ and $\hat{a}_{p, \text { TS }}$, 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 onestep 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 $h_{1} \rightarrow 0$ and $n h_{1} \rightarrow \infty$, then the asymptotic conditional bias of the one-step estimator is given by

$$
\operatorname{bias}\left(\hat{a}_{p, \text { OS }}\left(u_{0}\right) \mid \mathscr{D}\right)=\frac{h_{1}^{2} \mu_{2}}{2} a_{p}^{\prime \prime}\left(u_{0}\right)\left(1+o_{P}(1)\right)
$$

and the asymptotic conditional variance of $\hat{\alpha}_{p, \mathrm{OS}}\left(u_{0}\right)$ is given by

$$
\operatorname{var}\left(\hat{a}_{p, \mathrm{OS}}\left(u_{0}\right) \mid \mathscr{D}\right)=\frac{\sigma^{2}\left(u_{0}\right) \nu_{0}}{n h_{1} f\left(u_{0}\right)} e_{p, p}^{T} \Omega_{p}^{-1} e_{p, p}\left\{1+o_{P}(1)\right\} .
$$

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

$$
\operatorname{bias}\left(\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right) \mid \mathscr{D}\right)=\left(\frac{1}{2} a_{p}^{\prime \prime}\left(u_{0}\right) \mu_{2} h_{2}^{2}-\frac{\mu_{2} h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p-1} a_{j}^{\prime \prime}\left(u_{0}\right) r_{p j}\right)\left(1+o_{P}(1)\right)
$$

and the asymptotic variance

$$
\operatorname{var}\left(\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right) \mid \mathscr{O}\right)=\frac{\nu_{0} \sigma^{2}\left(u_{0}\right)}{n h_{2} f\left(u_{0}\right)} e_{p, p}^{T} \Omega_{p}^{-1} e_{p, p}\left\{1+o_{P}(1)\right\} .
$$

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

$$
\frac{1}{2} a_{2}^{\prime \prime}\left(u_{0}\right) \mu_{2} h_{1}^{2}\left(1+o_{P}(1)\right),
$$

by taking initial bandwidth $h_{0}=o\left(h_{2}\right)$. 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

$$
\frac{1}{2} a_{p}^{\prime \prime}\left(u_{0}\right) \mu_{2} h_{2}^{2}\left\{1+o_{P}(1)\right\}
$$

and variance

$$
\frac{\sigma^{2}\left(u_{0}\right) \nu_{0}}{n h_{2} f\left(u_{0}\right) r_{p p}}\left\{1+o_{P}(1)\right\} .
$$

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 ( $X_{1}, \ldots, X_{p-1}$ ) 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 $K(t)=$ $0.75\left(1-t^{2}\right)_{+}$.
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 $\mu \mathrm{g} / \mathrm{m}^{3}$ ), nitrogen dioxide (in $\mu \mathrm{g} / \mathrm{m}^{3}$ ) and dust (in $\mu \mathrm{g} / \mathrm{m}^{3}$ ), are considered here. Table 1 summarizes their correlation coefficients. The correlation between the dust level and $\mathrm{NO}_{2}$ 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 $\mathrm{SO}_{2}, \mathrm{NO}_{2}$ 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

$$
\begin{equation*}
Y=a_{1}(t)+a_{2}(t) X_{2}+a_{3}(t) X_{3}+a_{4}(t) X_{4}+\varepsilon \tag{5.1}
\end{equation*}
$$

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 $a_{1}, \ldots, a_{4}$. 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 $a_{1}, \ldots, a_{4}$, 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

$$
\left\{\mathrm{RSS}_{h}(\text { one-step })-\mathrm{RSS}_{h}(\text { two-step })\right\} / \mathrm{RSS}_{h}(\text { two-step }),
$$

where $\mathrm{RSS}_{h}$ (one-step) and $\mathrm{RSS}_{h}$ (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 $\tau$ 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

$$
\left\{Y(t+\tau), X_{2}(t), X_{3}(t), X_{4}(t), t=1,2, \ldots,\right\} .
$$

Figure 4 b presents the resulting residuals sum of squares for each time lag. As $\tau$ 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 twostep 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. $Y=\sin (60 U) X_{1}+4 U(1-U) X_{2}+\varepsilon$.
Example 2. $Y=\sin (6 \pi U) X_{1}+\sin (2 \pi U) X_{2}+\varepsilon$.
Example 3. $Y=\sin (8 \pi(U-0.5)) X_{1}$

$$
+\left(3.5\left[\exp \left\{-(4 U-1)^{2}\right\}+\exp \left\{-(4 U-3)^{2}\right\}\right]-1.5\right) X_{2}+\varepsilon
$$

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 $\varepsilon, U$ and ( $X_{1}, X_{2}$ ) are independent. The random variable $\varepsilon$ follows a normal distribution with mean zero and variance $\sigma^{2}$. The $\sigma^{2}$ is chosen so that the signal-to-noise ratio is about 5 : 1, namely,

$$
\sigma^{2}=0.2 \operatorname{var}\left\{m\left(U, X_{1}, X_{2}\right)\right\} \quad \text { with } m\left(U, X_{1}, X_{2}\right)=E\left(Y \mid U, X_{1}, X_{2}\right)
$$

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 twostep 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 $\hat{a}_{p, \mathrm{TS}}\left(u_{0}\right)$, the following lemma on the uniform convergence will be used.

Lemma 1. Let $\left(X_{1}, Y_{1}\right), \ldots,\left(X_{n}, Y_{n}\right)$ be i.i.d random vectors, where the $Y_{i}$ 's are scalar random variables. Assume further that $E|y|^{3}<\infty$ and $\sup _{x} \int|y|^{s} f(x, y) d y<\infty$, 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

$$
\sup _{x \in D}\left|n^{-1} \sum_{i=1}^{n}\left\{K_{h}\left(X_{i}-x\right) Y_{i}-E\left[K_{h}\left(X_{i}-x\right) Y_{i}\right]\right\}\right|=O_{P}\left[\{n h / \log (1 / h)\}^{-1 / 2}\right]
$$

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

$$
S=\left(\begin{array}{ll}
S_{11} & S_{12} \\
S_{12}^{T} & S_{22}
\end{array}\right)
$$

with

$$
S_{11}=\Omega_{p-1} \otimes\left(\begin{array}{cc}
\mu_{0} & 0 \\
0 & \mu_{2}
\end{array}\right), \quad S_{12}=\alpha_{p} \otimes\left(\begin{array}{cccc}
\mu & 0 & \mu_{2} & 0 \\
0 & \mu_{2} & 0 & \mu_{4}
\end{array}\right)
$$

and

$$
S_{22}=r_{p p}\left(\begin{array}{cccc}
\mu_{0} & 0 & \mu_{2} & 0 \\
0 & \mu_{2} & 0 & \mu_{4} \\
\mu_{2} & 0 & \mu_{4} & 0 \\
0 & \mu_{4} & 0 & \mu_{6}
\end{array}\right),
$$

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

$$
S_{(i)}^{*}=\Omega_{p}\left(U_{i}\right) \otimes\left(\begin{array}{cc}
\mu_{0} & 0 \\
0 & \mu_{2}
\end{array}\right), \quad S_{(0)}^{*}=\left.S_{(i)}^{*}\right|_{U_{i}=u_{0}}, \quad Q=\Omega_{p} \otimes\left(\begin{array}{cc}
1 & 0 \\
0 & 0
\end{array}\right)
$$

and

$$
\beta_{(i)}^{T}=\sum_{j=1}^{p} a_{j}^{\prime \prime}\left(U_{i}\right) \mu_{2}\left(\alpha_{j}^{T}\left(U_{i}\right), r_{p j}\left(U_{i}\right)\right) \otimes(1,0), \quad \alpha^{* T}=\left(\alpha_{p}^{T}, r_{p p}\right) \otimes(1,0)
$$

Put

$$
A=I_{p-1} \otimes\left(\begin{array}{cc}
1 & 0 \\
0 & h_{1}
\end{array}\right), \quad G=I_{p} \otimes\left(\begin{array}{cc}
1 & 0 \\
0 & h_{0}
\end{array}\right)
$$

and

$$
D=\left(\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & h_{1} & 0 & 0 \\
0 & 0 & h_{1}^{2} & 0 \\
0 & 0 & 0 & h_{1}^{3}
\end{array}\right), \quad D_{2}=\left(\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & h_{2} & 0 & 0 \\
0 & 0 & h_{2}^{2} & 0 \\
0 & 0 & 0 & h_{1}^{3}
\end{array}\right)
$$

We are now ready to prove our results.
Proof of Theorem 1. First, let us calculate the asymptotic conditional bias of $\hat{\alpha}_{p, 1}\left(u_{0}\right)$. Note that by Taylor's expansion, we have

$$
\begin{aligned}
Y= & \mathbf{X}_{1}\left(a_{1}\left(u_{0}\right), a_{1}^{\prime}\left(u_{0}\right), \ldots, a_{p-1}\left(u_{0}\right), a_{p-1}^{\prime}\left(u_{0}\right), a_{p}\left(u_{0}\right), a_{p}^{\prime}\left(u_{0}\right)\right. \\
& \left.\frac{1}{2} a_{p}^{\prime \prime}\left(u_{0}\right), \frac{1}{3!} a_{p}^{\prime \prime \prime}\left(u_{0}\right)\right)^{T} \\
& +\frac{1}{2} \sum_{j=1}^{p-1}\left(\begin{array}{c}
a_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-u_{0}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-u_{0}\right)^{2} X_{n j}
\end{array}\right)+\frac{1}{4!}\left(\begin{array}{c}
a_{p}^{(4)}\left(\eta_{1}\right)\left(U_{1}-u_{0}\right)^{4} X_{1 p} \\
\vdots \\
a_{p}^{(4)}\left(\eta_{n}\right)\left(U_{n}-u_{0}\right)^{4} X_{n p}
\end{array}\right)+\boldsymbol{\varepsilon}
\end{aligned}
$$

where $\boldsymbol{\varepsilon}=\left(\varepsilon_{1}, \ldots \varepsilon_{n}\right)^{T}, \xi_{i j}$ and $\eta_{i}$ are between $U_{i}$ and $u_{0}$ for $i=1, \ldots n, j=$ $1, \ldots, p-1$. Thus,

$$
\begin{aligned}
\hat{a}_{p, 1}\left(u_{0}\right)= & a_{p}\left(u_{0}\right) \\
& +\frac{1}{2} \sum_{j=1}^{p-1} e_{2 p-1,2 p+2}^{T}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} \mathbf{X}_{1}^{T} W_{1}\left(\begin{array}{c}
\alpha_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-u_{0}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-u_{0}\right)^{2} X_{n j}
\end{array}\right)
\end{aligned}
$$

$$
\begin{aligned}
& +\frac{1}{4!} e_{2 p-1,2 p+2}^{T}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} \mathbf{X}_{1}^{T} W_{1}\left(\begin{array}{c}
a_{p}^{(4)}\left(\eta_{1}\right)\left(U_{1}-u_{0}\right)^{4} X_{1 p} \\
\vdots \\
a_{p}^{(4)}\left(\eta_{n}\right)\left(U_{n}-u_{0}\right)^{4} X_{n p}
\end{array}\right) \\
& +e_{2 p-1,2 p+2}^{T}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} \mathbf{X}_{1}^{T} W_{1} \boldsymbol{\varepsilon} .
\end{aligned}
$$

Obviously,

$$
\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}=\left(\begin{array}{ll}
\mathbf{X}_{3}^{T} W_{1} \mathbf{X}_{3} & \mathbf{X}_{3}^{T} W_{1} \mathbf{X}_{2} \\
\mathbf{X}_{2}^{T} W_{1} \mathbf{X}_{3} & \mathbf{X}_{2}^{T} W_{1} \mathbf{X}_{2}
\end{array}\right)
$$

By calculating the mean and variance, one can easily get

$$
\begin{aligned}
& \mathbf{X}_{3}^{T} W_{1} \mathbf{X}_{3}=n f\left(u_{0}\right) A S_{11} A\left(1+o_{P}(1)\right), \\
& \mathbf{X}_{3}^{T} W_{1} \mathbf{X}_{2}=n f\left(u_{0}\right) A S_{12} D\left(1+o_{P}(1)\right)
\end{aligned}
$$

and

$$
\begin{equation*}
\mathbf{X}_{2}^{T} W_{1} \mathbf{X}_{2}=n f\left(u_{0}\right) D S_{22} D\left(1+o_{P}(1)\right) \tag{6.1}
\end{equation*}
$$

Combining the last three asymptotic expressions leads to

$$
\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}=n f\left(u_{0}\right) \operatorname{diag}(A, D) S \operatorname{diag}(A, D)\left(1+o_{P}(1)\right)
$$

Similarly, we have

$$
\begin{aligned}
& \mathbf{X}_{3}^{T} W_{1}\left(\begin{array}{c}
a_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-u_{0}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-u_{0}\right)^{2} X_{n j}
\end{array}\right) \\
& \quad=n f\left(u_{0}\right) h_{1}^{2} \alpha_{j}^{\prime \prime}\left(u_{0}\right) A\left(\alpha_{j} \otimes(1,0)^{T}\right) \mu_{2}\left(1+o_{P}(1)\right)
\end{aligned}
$$

and

$$
\mathbf{X}_{2}^{T} W_{1}\left(\begin{array}{c}
\alpha_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-u_{0}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-u_{0}\right)^{2} X_{n j}
\end{array}\right)=n f\left(u_{0}\right) h_{1}^{2} \alpha_{j}^{\prime \prime}\left(u_{0}\right) D\left(\begin{array}{c}
r_{p j} \mu_{2} \\
0 \\
r_{p j} \mu_{4} \\
0
\end{array}\right)\left(1+o_{P}(1)\right)
$$

Thus,

$$
\begin{aligned}
& \mathbf{X}_{1}^{T} W_{1}\left(\begin{array}{c}
\alpha_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-u_{0}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-u_{0}\right)^{2} X_{n j}
\end{array}\right) \\
&= n f\left(u_{0}\right) h_{1}^{2} \alpha_{j}^{\prime \prime}\left(u_{0}\right) \operatorname{diag}(A, D) \\
& \times\left(\alpha_{j}^{T} \otimes(1,0) \mu_{2}, r_{p j} \mu_{2}, 0, r_{p j} \mu_{4}, 0\right)^{T}\left(1+o_{P}(1)\right)
\end{aligned}
$$

So the asymptotic conditional bais of $\hat{a}_{p, 1}\left(u_{0}\right)$ is given by

$$
\begin{aligned}
& \operatorname{bias}\left(\hat{a}_{p, 1}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=\frac{1}{2} h_{1}^{2} \sum_{j=1}^{p-1} a_{j}^{\prime \prime}\left(u_{0}\right) e_{2 p-1,2 p+2}^{T} S^{-1}\left(\alpha_{j}^{T} \otimes(1,0) \mu_{2}, r_{p j} \mu_{2}, 0, r_{p j} \mu_{4}, 0\right)^{T}\left(1+o_{P}(1)\right)
\end{aligned}
$$

Using the properties of the Kronecker product we have

$$
\begin{aligned}
& \operatorname{bias}\left(\hat{a}_{p, 1}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=\frac{h_{1}^{2} \mu_{2}}{2\left(r_{p p}-\alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}\right) r_{p p}} \\
& \quad \times \sum_{j=1}^{p-1}\left(r_{p j} \alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}-r_{p p} \alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{j}\right) a_{j}^{\prime \prime}\left(u_{0}\right)\left(1+o_{P}(1)\right) \\
& = \\
& -\frac{h_{1}^{2} \mu_{2}}{2 r_{p p}} \sum_{j=1}^{p-1} r_{p j} \alpha_{j}^{\prime \prime}\left(u_{0}\right)+o_{P}\left(h_{1}^{2}\right)
\end{aligned}
$$

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 $\hat{a}_{p, 1}\left(u_{0}\right)$ is given by

$$
\begin{aligned}
& \operatorname{var}\left(\hat{\alpha}_{p, 1}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=e_{2 p-1,2 p+2}^{T}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} \mathbf{X}_{1}^{T} W_{1} \Psi W_{1} \mathbf{X}_{1}\left(\mathbf{X}_{1}^{T} W_{1} \mathbf{X}_{1}\right)^{-1} e_{2 p-1,2 p+2} \\
& \quad=\frac{\sigma^{2}\left(u_{0}\right)}{n h_{1} f\left(u_{0}\right)} e_{2 p-1,2 p+2}^{T} S^{-1} \tilde{S} S^{-1} e_{2 p-1,2 p+2}\left(1+o_{P}(1)\right)
\end{aligned}
$$

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

$$
\operatorname{var}\left(\hat{\alpha}_{p, 1}\left(u_{0}\right) \mid \mathscr{D}\right)=\frac{\sigma^{2}\left(u_{0}\right)\left(\lambda_{2} r_{p p}+\lambda_{3} \alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}\right)}{n h_{1} f\left(u_{0}\right) \lambda_{1} r_{p p}\left(r_{p p}-\alpha_{p}^{T} \Omega_{p-1}^{-1} \alpha_{p}\right)}\left(1+o_{P}(1)\right)
$$

where $\lambda_{1}=\left(\mu_{4}-\mu_{2}^{2}\right)^{2}, \quad \lambda_{2}=\nu_{0} \mu_{4}^{2}-2 \nu_{2} \mu_{2} \mu_{4}+\mu_{2}^{2} \nu_{4}, \quad \lambda_{3}=2 \mu_{2} \nu_{2} \mu_{4}-2 \nu_{0} \mu_{2}^{2} \mu_{4}-$ $\mu_{2}^{2} \nu_{4}+\nu_{0} \mu_{2}^{4}$. 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

$$
\begin{aligned}
Y= & \mathbf{X}_{(i)}\left(a_{1}\left(U_{i}\right), a_{1}^{\prime}\left(U_{i}\right), \ldots, a_{p}\left(U_{i}\right), a_{p}^{\prime}\left(U_{i}\right)\right)^{T} \\
& +\frac{1}{2} \sum_{j=1}^{p}\left(\begin{array}{c}
a_{j}^{\prime \prime}\left(\xi_{1 j}\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(\xi_{n j}\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right)+\boldsymbol{\varepsilon} \\
= & \mathbf{X}_{(i)}\left(a_{1}\left(U_{i}\right), a_{1}^{\prime}\left(U_{i}\right), \ldots, a_{p}\left(U_{i}\right), a_{p}^{\prime}\left(U_{i}\right)\right)^{T}
\end{aligned}
$$

$$
\begin{aligned}
& +\frac{1}{2} \sum_{j=1}^{p}\left(\begin{array}{c}
a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right) \\
& +\frac{1}{2} \sum_{j=1}^{p}\left(\begin{array}{c}
\left(a_{j}^{\prime \prime}\left(\xi_{1 j}\right)-a_{j}^{\prime \prime}\left(U_{i}\right)\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
\left(a_{j}^{\prime \prime}\left(\xi_{n j}\right)-a_{j}^{\prime \prime}\left(U_{i}\right)\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right)+\boldsymbol{\varepsilon}
\end{aligned}
$$

where $\xi_{k j}$ is between $U_{i}$ and $U_{k}$. Thus, for $l=1, \ldots, p-1$,

$$
\begin{aligned}
\hat{a}_{l, 0}\left(U_{i}\right)= & a_{l}\left(U_{i}\right) \\
& +\frac{1}{2} e_{2 l-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} \sum_{j=1}^{p}\left(\begin{array}{c}
a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right) \\
& +\frac{1}{2} e_{2 l-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} \\
& \times \sum_{j=1}^{p}\left(\begin{array}{c}
\left(a_{j}^{\prime \prime}\left(\xi_{1 j}\right)-a_{j}^{\prime \prime}\left(U_{i}\right)\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
\left(a_{j}^{\prime \prime}\left(\xi_{n j}\right)-a_{j}^{\prime \prime}\left(U_{i}\right)\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right) \\
& +e_{2 l-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} \boldsymbol{\varepsilon} .
\end{aligned}
$$

By Lemma 1, we have

$$
\begin{equation*}
\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}=n f\left(U_{i}\right) G S_{(i)}^{*} G\left(1+o_{P}(1)\right) \tag{6.2}
\end{equation*}
$$

and
(6.3) $\quad \mathbf{X}_{(i)}^{T} W_{(i)} \sum_{j=1}^{p}\left(\begin{array}{c}a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\ \vdots \\ a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}\end{array}\right)=n f\left(U_{i}\right) h_{0}^{2} G \beta_{(i)}\left(1+o_{P}(1)\right)$.

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

$$
\begin{aligned}
& \frac{1}{2} e_{2 l-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} \sum_{j=1}^{p}\left(\begin{array}{c}
\alpha_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{1}-U_{i}\right)^{2} X_{1 j} \\
\vdots \\
a_{j}^{\prime \prime}\left(U_{i}\right)\left(U_{n}-U_{i}\right)^{2} X_{n j}
\end{array}\right) \\
& \quad=\frac{1}{2} h_{0}^{2} e_{2 l-1,2 p}^{T} S_{(i)}^{*-1} \beta_{(i)}\left(1+o_{P}(1)\right)
\end{aligned}
$$

Note that $K$ has a bounded support. From the last expression and the uniform continuity of functions $\alpha_{j}^{\prime \prime}(\cdot)$ in a neighborhood of $u_{0}$, it follows that

$$
\begin{equation*}
E\left(\hat{a}_{l, 0}\left(U_{i}\right)-a_{l}\left(U_{i}\right) \mid \mathscr{D}\right)=\frac{1}{2} h_{0}^{2} e_{2 l-1,2 p}^{T} S_{(i)}^{*-1} \beta_{(i)}\left(1+o_{P}(1)\right) . \tag{6.4}
\end{equation*}
$$

Since

$$
\begin{aligned}
\left(\begin{array}{c}
Y_{1}-\sum_{j=1}^{p-1} \hat{a}_{j, 0}\left(U_{1}\right) X_{1 j} \\
\vdots \\
Y_{n}-\sum_{j=1}^{p-1} \hat{a}_{j, 0}\left(U_{n}\right) X_{n j}
\end{array}\right)= & \left(\begin{array}{c}
a_{p}\left(U_{1}\right) X_{1 p} \\
\vdots \\
a_{p}\left(U_{n}\right) X_{n p}
\end{array}\right) \\
& +\left(\begin{array}{c}
\sum_{j=1}^{p-1}\left(a_{j}\left(U_{1}\right)-\hat{a}_{j, 0}\left(U_{1}\right)\right) X_{1 j} \\
\vdots \\
\sum_{j=1}^{p-1}\left(a_{j}\left(U_{n}\right)-\hat{a}_{j, 0}\left(U_{n}\right)\right) X_{n j}
\end{array}\right)+\boldsymbol{\varepsilon}
\end{aligned}
$$

it follows from (3.3) that

$$
\begin{aligned}
\hat{a}_{p, 2}\left(u_{0}\right)= & a_{p}\left(u_{0}\right)+\frac{1}{4!}(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \\
& \times\left(\begin{array}{c}
a_{p}^{(4)}\left(\eta_{1}\right)\left(U_{1}-u_{0}\right)^{4} X_{1 p} \\
\vdots \\
a_{p}^{(4)}\left(\eta_{n}\right)\left(U_{n}-u_{0}\right)^{4} X_{n p}
\end{array}\right)
\end{aligned}
$$

$$
\begin{aligned}
& +(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2}\left(\begin{array}{l}
\sum_{j=1}^{p-1}\left(a_{j}\left(U_{1}\right)-\hat{a}_{j, 0}\left(U_{1}\right)\right) X_{1 j} \\
\vdots \\
\sum_{j=1}^{p-1}\left(a_{j}\left(U_{n}\right)-\hat{a}_{j, 0}\left(U_{n}\right)\right) X_{n j}
\end{array}\right) \\
& +(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \varepsilon \\
& \equiv a_{p}\left(u_{0}\right)+\frac{1}{4!} \tilde{J}_{1}+\tilde{J}_{2}+(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \varepsilon .
\end{aligned}
$$

By simple calculation we have

$$
\begin{aligned}
E\left(\tilde{J}_{1} \mid \mathscr{D}\right) & =h_{2}^{4} a_{p}^{(4)}\left(u_{0}\right)(1,0,0,0) S_{22}^{-1}\left(\begin{array}{c}
r_{p p} \mu_{4} \\
0 \\
r_{p p} \mu_{6} \\
0
\end{array}\right)\left(1+o_{P}(1)\right) \\
& =h_{2}^{4} a_{p}^{(4)}\left(u_{0}\right)\left(\frac{\mu_{4}}{\mu_{4}-\mu_{2}^{2}}, 0,-\frac{\mu_{2}}{\mu_{4}-\mu_{2}^{2}}, 0\right)\left(\begin{array}{c}
\mu_{4} \\
0 \\
\mu_{6} \\
0
\end{array}\right)\left(1+o_{P}(1)\right) \\
& =\frac{\mu_{4}^{2}-\mu_{2} \mu_{6}}{\mu_{4}-\mu_{2}^{2}} h_{2}^{4} a_{p}^{(4)}\left(u_{0}\right)\left(1+o_{P}(1)\right) .
\end{aligned}
$$

By (6.4), we have

$$
\begin{aligned}
E\left(\tilde{J}_{2} \mid \mathscr{D}\right)= & (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \\
& \times\left(\begin{array}{l}
\sum_{j=1}^{p-1} E\left(\left(a_{j}\left(U_{1}\right)-\hat{a}_{j, 0}\left(U_{1}\right)\right) \mid \mathscr{D}\right) X_{1 j} \\
\vdots \\
\sum_{j=1}^{p-1} E\left(\left(a_{j}\left(U_{n}\right)-\hat{a}_{j, 0}\left(U_{n}\right)\right) \mid \mathscr{D}\right) X_{n j}
\end{array}\right)
\end{aligned}
$$

$$
\begin{aligned}
= & -\frac{1}{2} h_{0}^{2}(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2}\left(\begin{array}{c}
\sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(1)}^{*-1} \beta_{(1)} X_{1 j} \\
\vdots \\
\sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(n)}^{*-1} \beta_{(n)} X_{n j}
\end{array}\right) \\
& \times\left(1+o_{P}(1)\right) \\
= & -\frac{h_{0}^{2}}{2 r_{p p}}\left(\frac{\mu_{4}}{\mu_{4}-\mu_{2}^{2}}, 0,-\frac{\mu_{2}}{\mu_{4}-\mu_{2}^{2}}, 0\right)\left(\begin{array}{l}
\sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \beta_{(0)} r_{p j} \\
0 \\
\left.\sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \beta_{(0)} r_{p j} \mu_{2}\right) \\
\\
\end{array}\right) \times\left(1+o_{P}(1)\right) \\
= & \frac{h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \beta_{(0)} r_{p j}\left(1+o_{P}(1)\right) .
\end{aligned}
$$

Therefore, by (6.5) we obtain

$$
\begin{aligned}
& \operatorname{bias}\left(\hat{a}_{p, 2}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=\left(-\frac{h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \beta_{(0)} r_{p j}+\frac{\mu_{4}^{2}-\mu_{2} \mu_{6}}{4!\left(\mu_{4}-\mu_{2}^{2}\right)} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}\right)\left(1+o_{P}(1)\right)
\end{aligned}
$$

By using the properties of the Kronecker product, we have

$$
\begin{aligned}
& \operatorname{bias}\left(\hat{\alpha}_{p, 2}\left(u_{0}\right) \mid \mathscr{D}\right) \\
&=\left(\frac{1}{4!} \frac{\mu_{4}^{2}-\mu_{6} \mu_{2}}{\mu_{4}-\mu_{2}^{2}} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}-\frac{\mu_{2} h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p} a_{j}^{\prime \prime}\left(u_{0}\right)\left(\alpha_{p}^{T}, 0\right) \Omega_{p}^{-1}\binom{\alpha_{j}}{r_{p j}}\right) \\
& \times\left(1+o_{P}(1)\right) \\
&= \frac{1}{4!} \frac{\mu_{4}^{2}-\mu_{6} \mu_{2}}{\mu_{4}-\mu_{2}^{2}} a_{p}^{(4)}\left(u_{0}\right) h_{2}^{4}-\frac{\mu_{2} h_{0}^{2}}{2 r_{p p}} \sum_{j=1}^{p-1} a_{j}^{\prime \prime}\left(u_{0}\right) r_{p j}+o_{P}\left(h_{2}^{4}+h_{0}^{2}\right) .
\end{aligned}
$$

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

$$
\begin{aligned}
\operatorname{var}\left(\hat{a}_{p, 2}\right. & \left.\left(u_{0}\right) \mid \mathscr{D}\right) \\
= & (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \Psi W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \\
& \times(1,0,0,0)^{T} \\
- & 2(1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} H \Psi W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \\
& \times(1,0,0,0)^{T} \\
+ & (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} H \Psi H^{T} W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \\
& \times(1,0,0,0)^{T} .
\end{aligned}
$$

Using similar arguments as before, we can show that

$$
\begin{align*}
& (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} \Psi W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1}(1,0,0,0)^{T} \\
& \quad=\frac{\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}}{n h_{2} f\left(u_{0}\right) r_{p p}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}} \sigma^{2}\left(u_{0}\right)\left(1+o_{P}(1)\right) \tag{6.6}
\end{align*}
$$

Since

$$
H \Psi W_{2} \mathbf{X}_{2}=\sum_{j=1}^{p-1}\left(\begin{array}{c}
X_{1 j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(1)}^{T} W_{(1)} \mathbf{X}_{(1)}\right)^{-1} \mathbf{X}_{(1)}^{T} W_{(1)} \Psi W_{2} \mathbf{X}_{2} \\
\vdots \\
X_{n j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(n)}^{T} W_{(n)} \mathbf{X}_{(n)}\right)^{-1} \mathbf{X}_{(n)}^{T} W_{(n)} \Psi W_{2} \mathbf{X}_{2}
\end{array}\right)
$$

by Lemma 1, we have

$$
\mathbf{X}_{(i)}^{T} W_{(i)} \Psi W_{2} \mathbf{X}_{2}=n f\left(U_{i}\right) \sigma^{2}\left(U_{i}\right) K_{h_{2}}\left(U_{i}-u_{0}\right) G T_{2 p \times 4,(i)} D_{2}\left(1+o_{P}(1)\right),
$$

where

$$
T_{2 p \times 4,(i)}=\left(\tilde{u}_{k, l,(i)}\right)_{2 p \times 4}, \quad 1 \leq k \leq 2 p, 0 \leq l \leq 3
$$

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

$$
\begin{aligned}
\tilde{u}_{2 k-1,0,(i)}= & r_{k p}\left(U_{i}\right), \quad \tilde{u}_{2 k-1,1,(i)}=r_{k p}\left(U_{i}\right)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1) \\
\tilde{u}_{2 k-1,2,(i)}= & r_{k p}\left(U_{i}\right)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{2}+o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1) \\
\tilde{u}_{2 k-1,3,(i)}= & r_{k p}\left(U_{i}\right),\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{3}+o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{2} \\
& +o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1)
\end{aligned}
$$

$$
\begin{aligned}
& \tilde{u}_{2 k, 0,(i)}=o_{P}(1), \quad \tilde{u}_{2 k, 1,(i)}=o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1), \\
& \tilde{u}_{2 k, 2,(i)}=o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{2}+o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1)
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{u}_{2 k, 3,(i)}= & o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{3}+o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)^{2} \\
& +o_{P}(1)\left(\frac{U_{i}-u_{0}}{h_{2}}\right)+o_{P}(1)
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
\mathbf{X}_{2}^{T} W_{2} H \Psi W_{2} \mathbf{X}_{2}= & \frac{n f\left(u_{0}\right)}{h_{2}} \sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \alpha^{*} r_{p j} \sigma^{2}\left(u_{0}\right) D_{2} \\
& \times\left(\begin{array}{cccc}
\nu_{0} & 0 & \nu_{2} & 0 \\
0 & \nu_{2} & 0 & \nu_{4} \\
\nu_{2} & 0 & \nu_{4} & 0 \\
0 & \nu_{4} & 0 & \nu_{6}
\end{array}\right) D_{2}\left(1+o_{P}(1)\right) .
\end{aligned}
$$

This and (6.1) together yield

$$
\begin{align*}
& (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} H \Psi W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1}(1,0,0,0)^{T} \\
& \quad=\frac{\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}}{n h_{2} f\left(u_{0}\right) r_{p p}^{2}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}}  \tag{6.7}\\
& \quad \times \sum_{j=1}^{p-1} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} \alpha^{*} r_{p j} \sigma^{2}\left(u_{0}\right)\left(1+o_{P}(1)\right)
\end{align*}
$$

Let

$$
X_{k(i)}=\left(X_{k 1}, X_{k 1}\left(U_{k}-U_{i}\right), \ldots, X_{k p}, X_{k p}\left(U_{k}-U_{i}\right)\right)^{T}
$$

and

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

Then we have

$$
\mathbf{X}_{2}^{T} W_{2} H \Psi H^{T} W_{2} \mathbf{X}_{2}=\left(v_{r s}\right)_{4 \times 4}, \quad 0 \leq r, s \leq 3
$$

where

$$
\begin{aligned}
& v_{r s}=\sum_{i=1}^{n} \sum_{l=1}^{n}\{ X_{i p} X_{l p}\left(U_{i}-u_{0}\right)^{r}\left(U_{l}-u_{0}\right)^{s} K_{h_{2}}\left(U_{i}-u_{0}\right) K_{h_{2}}\left(U_{l}-u_{0}\right) \\
& \times \sum_{j=1}^{p-1} X_{i j} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^{T} W_{(i)} \Psi \\
&\left.\times\left(\sum_{m=1}^{p-1} X_{l m} e_{2 m-1,2 p}^{T}\left(\mathbf{X}_{(l)}^{T} W_{(l)} \mathbf{X}_{(l)}\right)^{-1} \mathbf{X}_{(l)}^{T} W_{(l)}\right)^{T}\right\} \\
&=\sum_{j=1}^{p-1} \sum_{m=1}^{p-1} \sum_{k=1}^{n}\left\{\sum_{i=1}^{n} X_{i p} X_{i j}\left(U_{i}-u_{0}\right)^{r} e_{2 j-1,2 p}^{T}\left(\mathbf{X}_{(i)}^{T} W_{(i)} \mathbf{X}_{(i)}\right)^{-1} X_{k(i)} K_{h_{2}}\right. \\
&\left.\times\left(U_{i}-u_{0}\right) K_{h_{0}}\left(U_{k}-U_{i}\right) \sigma^{2}\left(U_{k}\right)\right\} \\
& \times\left\{\sum_{l=1}^{n} X_{l p} X_{l m}\left(U_{l}-u_{0}\right)^{s} X_{k(l)}^{T}\left(\mathbf{X}_{(l)}^{T} W_{(l)} \mathbf{X}_{(l)}\right)^{-1} e_{2 m-1,2 p} K_{h_{2}}\right. \\
&\left.\times\left(U_{l}-u_{0}\right) K_{h_{0}}\left(U_{k}-U_{l}\right)\right\}
\end{aligned}
$$

Using Lemma 1 and tedious calculation, we obtain

$$
\begin{aligned}
& \mathbf{X}_{2}^{T} W_{2} H \Psi H^{T} W_{2} \mathbf{X}_{2} \\
&=\frac{n f\left(u_{0}\right) \sigma^{2}\left(u_{0}\right)}{h_{2}} \sum_{j=1}^{p-1} \sum_{m=1}^{p-1} r_{p j} r_{p m} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2 m-1,2 p} D_{2} \\
& \times\left(\begin{array}{cccc}
\nu_{0} & 0 & \nu_{2} & 0 \\
0 & \nu_{2} & 0 & \nu_{4} \\
\nu_{2} & 0 & \nu_{4} & 0 \\
0 & \nu_{4} & 0 & \nu_{6}
\end{array}\right) D_{2}\left(1+o_{P}(1)\right)
\end{aligned}
$$

The combination of this and (6.1) gives

$$
\begin{align*}
& (1,0,0,0)\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1} \mathbf{X}_{2}^{T} W_{2} H \Psi H^{T} W_{2} \mathbf{X}_{2}\left(\mathbf{X}_{2}^{T} W_{2} \mathbf{X}_{2}\right)^{-1}(1,0,0,0)^{T} \\
& \quad=\frac{\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}}{n h_{2} f\left(u_{0}\right) r_{p p}^{2}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}}  \tag{6.8}\\
& \quad \times \sum_{j=1}^{p-1} \sum_{m=1}^{p-1} r_{p j} r_{p m} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2 m-1,2 p} \sigma^{2}\left(u_{0}\right)\left(1+o_{P}(1)\right)
\end{align*}
$$

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

$$
\begin{aligned}
\operatorname{var}\left(\hat{a}_{p, 2}\left(u_{0}\right) \mid \mathscr{D}\right)= & \frac{\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}}{n h_{2} f\left(u_{0}\right) r_{p p}^{2}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}} \\
& \times\left(r_{p p}+\sum_{j=1}^{p-1} \sum_{m=1}^{p-1} r_{p j} r_{p m} e_{2 j-1,2 p}^{T} S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2 m-1,2 p}\right. \\
& \times \sigma^{2}\left(u_{0}\right)\left(1+o_{P}(1)\right) .
\end{aligned}
$$

Using the properties of the Kronecker product we get

$$
\begin{aligned}
& \operatorname{var}\left(\hat{\alpha}_{p, 2}\left(u_{0}\right) \mid \mathscr{D}\right) \\
& \quad=\frac{\left(\mu_{4}^{2} \nu_{0}-2 \mu_{4} \mu_{2} \nu_{2}+\mu_{2}^{2} \nu_{4}\right) \sigma^{2}\left(u_{0}\right)}{n h_{2} f\left(u_{0}\right) r_{p p}^{2}\left(\mu_{4}-\mu_{2}^{2}\right)^{2}} \\
& \quad \times\left(r_{p p}+r_{p p}^{2} e_{p, p}^{T} \Omega_{p}^{-1} e_{p, p}-\left(\alpha_{p}^{T}, r_{p p}\right) \Omega_{p}^{-1}\binom{\alpha_{p}}{r_{p p}}\right)\left(1+o_{P}(1)\right)
\end{aligned}
$$

Note that $\Omega_{p}^{-1}\left(\alpha_{p}^{T}, r_{p p}\right)^{T}=e_{p, p}$. 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 singleindex 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


[^0]:    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.

