Title
Statistical Estimation in Varying-Coefficient Models

Authors
Fan, Jianqing
Zhang, Wenyang

Publication Date
1998

Statistical Estimation in Varying-Coefficient Models

Jianqing Fan*
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Wenyang Zhang
Department of Statistics
The Chinese University of Hong Kong
Shatin, Hong Kong

Abstract

Varying-coefficient models are a useful extension of the classical linear models. 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 can not 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 errors 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 dataset.

KEY WORDS: Varying-coefficient models, local linear fit, optimal rate of convergence, mean squared errors.
SHORT TITLE: Varying Coefficients Models.

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 an 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 so-called "curse of dimensionality". Examples includes additive modeling (Breiman and Friedman, 1995; Hastie and Tibshirani 1990), low-dimensional interaction modeling (Friedman 1991, Gu and Wahba, 1992, Stone *et al.*1997), multiple-index models (Härdle and Stoker 1990, Li 1991), and partially linear models (Wahba 1984; Green and Silverman 1994), and their hybrids (Carroll *et al.*1997, Fan *et al.*1997, Heckman *et al.*(1997)), among others. Different models explore different aspects of high-dimensional data and incorporate different prior knowledge into

*Partially supported by NSF Grant DMS-9503135 and an NSA Grant 96-1-0015.

modeling and approximation. They together form useful tool kits for processing high-dimensional data.

A useful extension of the classical linear model is the varying-coefficient models. This idea is scattered around text books. See for example page 245 of Shumway (1988). However, the potential of such a modeling techniques did not get fully explored until the seminal work of Cleveland *et al.*(1991) and Hastie and Tibshirani (1993). The varying-coefficient models assume that the following conditional linear model:

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

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

$$E(\varepsilon|U, X_1, \cdots, X_p) = 0$$

and

$$\text{var}(\varepsilon|U, X_1, \cdots, X_p) = \sigma^2(U).$$

By regarding $X_1 \equiv 1$, (1.1) allows varying intercept term in the model. The appeal of this model is that via allowing coefficients $a_1, \cdots, a_p$ to depend on U, the modeling bias can significantly be reduced and "curse of dimensionality" can be avoided. Another advantage of this model is its interpretability. This is particularly the case in the longitudinal study where it is reasonable to assume that the coefficients change over time t. See Hoover *et al* (1997) 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 $\{(U_i, X_{i1}, ..., X_{ip}, Y_i)\}_{i=1}^n$ from model (1.1). One simple approach to estimate the functions $a_j(\cdot)(j = 1, \cdots, p)$ 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 with bandwidth h, where $K_h(\cdot) = K(\cdot/h)/h$. The idea is due to Cleveland *et al.*(1991). While this idea is very simple and useful, it is implicitly assumed that functions $a_j(\cdot)$ possess about the same degrees of smoothness. If the functions process different degrees of smoothness, suboptimal estimators are obtained via using method (1.3).

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

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

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

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

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

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

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

with respect to a_p, b_p, c_p, d_p, where h_2 is the bandwidth in second step. In this way, we obtain a two-step estimator of $\hat{a}_{p,TS}(u_0)$ of $a_p(u_0)$. We will show that the bias of the two-step estimator is

of $O(h_2^4)$ and the variance of $O\{(nh_2)^{-1}\}$, provided that

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

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 moment of covariates. It plays little role in our understanding of the two-step estimation procedure. If X_i is assumed to have higher moments, the condition can be relaxed 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. The practical implications of this is that the two-step estimation method is not sensitive to the initial bandwidth h_0. This makes practical implementation much easier.

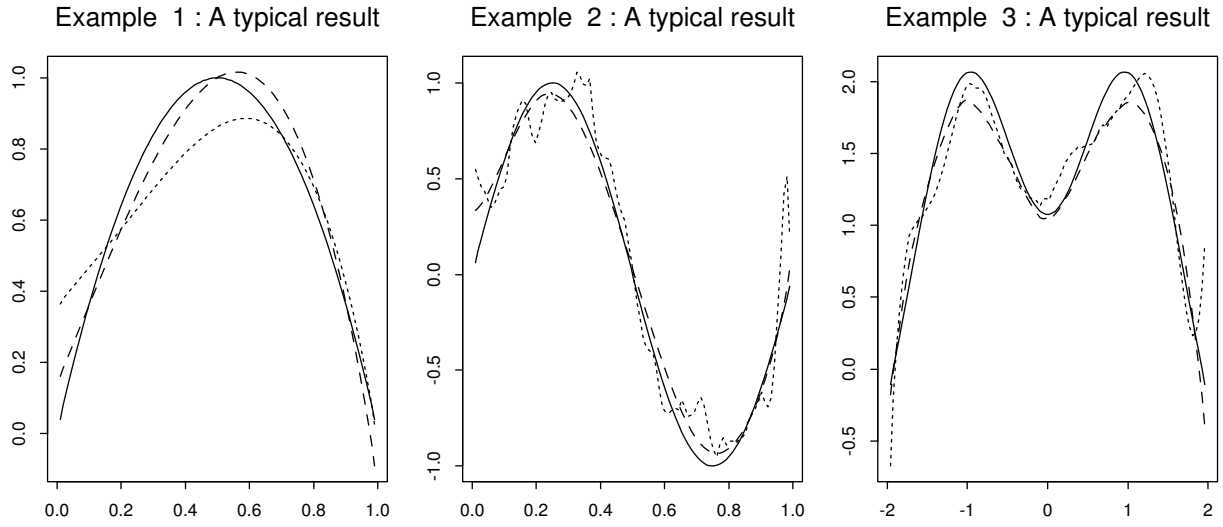

Figure 1: Comparisons the performance between the one-step and two-step estimator. Solid curve – true function; short-dashed curve – estimate based on the one-step procedure; long-dashed curve – estimate based on the two-step procedure.

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 $\lambda_1, \cdots, \lambda_p$. While this idea is powerful, there are a number of potential problems. First of all, there are p-smoothing parameters to choose simultaneously. This is quite a task in practice. Secondly, the computation can be quite a challenge. An iterative scheme was proposed in Hastie and Tibshirani (1993). Thirdly, the 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 examples to illustrate our methodology. The sample size $n = 500$ and $p = 2$. Figure 1 depicts a typical estimate of the one-step and two-step method 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. 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 be approximated well by a linear function in such a large neighborhood. While 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.

In real applications, we don't know in advance if a_p is really smoother than the rest of functions. The above discussion reveals the 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 degrees of the smoothness as the rest of the functions, we will demonstrate that the two-step estimation procedure achieves the same convergent rate as the one-step approach. Therefore, the two-step strategy is always more reliable than the one-step one. Details of implementing the two-step strategy 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 estimator is given in Section 3. Section 4 studies the asymptotic properties of the one-step and two-step estimators. In Section 5, we study the finite sample properties of the one-step and the two-step estimators via some simulated examples. The two-step techniques are further illustrated by an application to an environment data set. Technical proofs are given in Section 6.

2 Practical implementation of two-step estimators

As discussed in the introduction, one-step procedure is not optimal when coefficient functions admit different degrees of smoothness. However, we don't 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 $\hat{a}_{1,0}(U_i), \cdots, \hat{a}_{p,0}(U_i)$ for $i = 1, \cdots, n$. 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).

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

As we discussed in the introduction, the choice of initial bandwidth is not very sensitive to the two-step estimation as long as it is small enough so that the bias in the first step is not too large. This suggests the following simple automatic rule. Use the cross-validation or Generalized cross-validation (see e.g. Hoover *et al* 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 the univariate bandwidth selection procedures such as cross-validation (Stone, 1974), pre-asymptotic 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 in the second step. As we 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 for the two-step procedure: Bandwidth selection problems become relatively easy.

3 Formulae for the proposed estimators

The solutions 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 subscript "0", "1" and "2" respectively to indicate the variables related to the initial, one-step and two-step estimators.

Let

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

$$Y = (Y_1, \cdots, Y_n)^T, \quad \text{and} \quad W_0 = \text{diag}\,(K_{h_0}(U_1 - u_0), \cdots, K_{h_0}(U_n - u_0)).$$

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

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

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

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

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

$$\mathbf{X}_1 = (\mathbf{X}_3, \mathbf{X}_2) \quad W_1 = \text{diag}\,(K_{h_1}(U_1 - u_0), \cdots, K_{h_1}(U_n - u_0)).$$

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

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

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

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

where

$$W_2 = \text{diag}\,(K_{h_2}(U_1 - u_0), \cdots, K_{h_2}(U_n - u_0))$$

and $V = (V_1, \cdots, V_n)^T$ 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 $Y_1, \cdots, Y_n$. 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

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

Set

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

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 $s > 2$, $j = 1, \cdots, p$.

(2) $a_j''(\cdot)$ is continuous in a neighborhood of u_0, for $j = 1, \cdots, p$. Further, assume $a_j''(u_0) \neq 0$, for $j = 1, \cdots, p$.

(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 $i, j = 1, \cdots, p$, where $r_{ij}(u) = E(X_i X_j | U = u)$.

(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 $\mathcal{D}$ be the observed covariates vector, namely

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

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

$$\Psi = \text{diag}\left(\sigma^2(U_1), \cdots, \sigma^2(U_n)\right),$$

$$\alpha_j(u) = (r_{1j}(u), \cdots, r_{(p-1)j}(u))^T, \qquad \alpha_j = \alpha_j(u_0) \qquad \text{for} \qquad j = 1, \cdots, p$$

and

$$\Omega_i(u) = E\{(X_1, \cdots, X_i)^T (X_1, \cdots, X_i) | U = u\} \qquad \Omega_i = \Omega_i(u_0) \qquad \text{for} \qquad i = 1, \cdots, p.$$

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 $\hat{a}_{p,OS}(u_0)$ is given by*

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

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

$$var(\hat{a}_{p,OS}(u_0)|\mathcal{D}) = \frac{\sigma^2(u_0)(\lambda_2 r_{pp} + \lambda_3 \alpha_p^T \Omega_{p-1}^{-1} \alpha_p)}{n h_1 f(u_0) \lambda_1 r_{pp} (r_{pp} - \alpha_p^T \Omega_{p-1}^{-1} \alpha_p)} (1 + o_p(1)),$$

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 proof of Theorem 1 and other theorems are given in Section 6. It is clear that the conditional MSE of the one-step estimator $\hat{a}_{p,OS}(u_0)$ is only of order $O_P\{h_1^4 + (nh_1)^{-1}\}$ 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 $a_1, \cdots, a_{p-1}$ are transmitted to the bias of estimating a_p. Thus, the one-step estimator for a_p inherits non-negligible approximation errors and is not optimal.

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

Theorem 2 *If Conditions (1)–(6) and (7) hold, then the asymptotic conditional bias of $\hat{a}_{p,TS}(u_0)$ can be expressed as*

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

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

$$var(\hat{a}_{p,TS}(u_0)|\mathcal{D}) = \frac{(\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4)\sigma^2(u_0)}{nh_2 f(u_0) r_{pp}^2(\mu_4 - \mu_2^2)^2} \left(r_{pp} + r_{pp}^2 e_{p,p}^T \Omega_p^{-1} e_{p,p} - (\alpha_p^T, r_{pp})\Omega_p^{-1} \begin{pmatrix} \alpha_p \\ r_{pp} \end{pmatrix} \right) \{1 + o_P(1)\}$$

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 becomes

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

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 $a_1, \cdots, a_{p-1}$ 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 estimator.

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 estimator as $\hat{a}_{p,OS}$ and $\hat{a}_{p,TS}$ respectively.

Our technical results are to establish that the two-step estimator does not lose its statistical efficiency. 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) and (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*

$$bias(\hat{a}_{p,OS}(u_0)|\mathcal{D}) = \frac{h_1^2\mu_2}{2}a_p''(u_0)(1+o_P(1))$$

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

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

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

Theorem 4 *Suppose that Conditions (1)–(2), (4)–(6) and (7) hold. Then, we have the asymptotic conditional bias*

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

and the asymptotic variance

$$\begin{aligned} var(\hat{a}_{p,TS}(u_0)|\mathcal{D}) &= \frac{\nu_0\sigma^2(u_0)}{nh_2 f(u_0)r_{pp}^2} \\ &\quad \left(r_{pp} + r_{pp}^2 e_{p,p}^T\Omega_p^{-1}e_{p,p} - (\alpha_p^T, r_{pp})\Omega_p^{-1}\begin{pmatrix}\alpha_p \\ r_{pp}\end{pmatrix}\right)\{1+o_P(1)\}. \end{aligned}$$

Remark 2 *Consider the specific case where we have only two covariates $p = 2$. The asymptotic bias of the one-step estimator is simplified as*

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

and the asymptotic variance is given by

$$\frac{\sigma^2(u_0)\nu_0 r_{11}}{nh_1 f(u_0)(r_{11}r_{22}-r_{12}^2)}(1+o_P(1)).$$

For the two-step estimator, by taking initial bandwidth $h_0 = o(h_2)$, we obtain the same bias as the one-step estimator. 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 estimator is asymptotically identical.

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

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

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. Indeed, the two-step estimator shares the same asymptotic bias as that of the ideal estimator. However, the variance of the ideal estimator is typically small. This can easily be seen for the case $p = 2$. *Unless* $r_{12} = 0$, *namely* X_1 *and* X_2 *is uncorrelated given* $U = u_0$, *the asymptotic variance of the ideal estimator is always smaller.*

5 Simulations and Applications

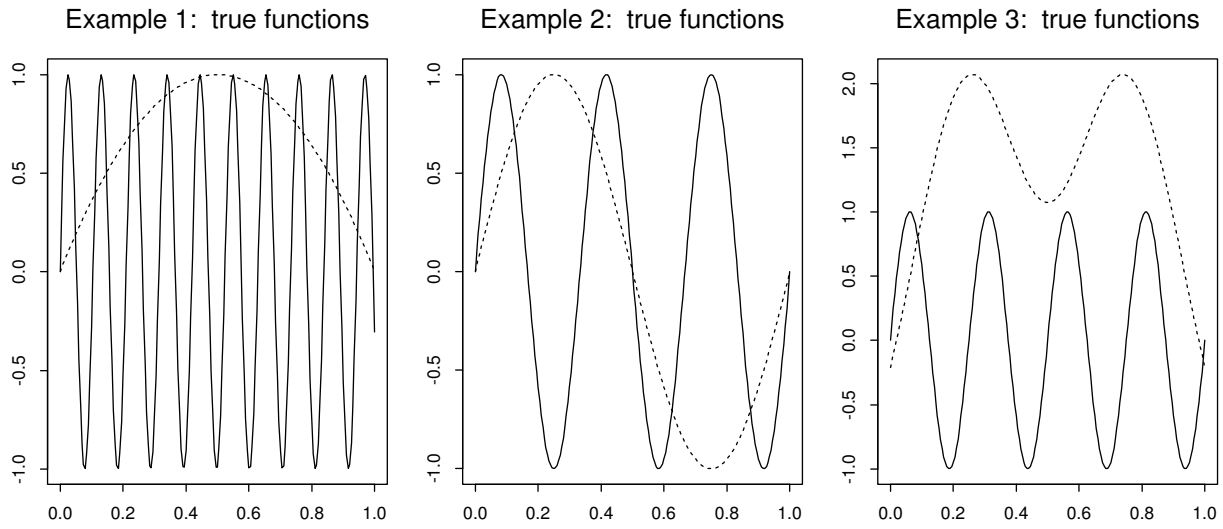

Figure 2: Varying-coefficient functions. Solid curve is $a_1(\cdot)$ and dashed curve is $a_2(\cdot)$.

We use the following three examples to illustrate the performance of our method:

Example 1: $Y = \sin(60U)X_1 + 4U(1-U) + \varepsilon$

Example 2: $Y = \sin(6\pi U)X_1 + \sin(2\pi U)X_2 + \varepsilon$

Example 3: $Y = \sin(8\pi(U-0.5))X_1 + \Big(3.5[\exp\{-(4U-1)^2\} + \exp\{-(4U-3)^2\}] - 1.5\Big)X_2 + \varepsilon,$

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 distribution of X_1 and X_2 is 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

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

Figure 2 gives the varying-coefficient functions a_1 and a_2.

For each of the above examples, we conducted 100 simulations with sample size $n = 250, 500, 1000$. The kernel function is taken to be $K(t) = (1 - t^2)_+$. The 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 chose $h_0 = 0.03, 0.04$ 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 sample such that the two-step estimator has the median performance. Once the sample is selected, the one-step estimate and the two-step estimate are computed. Figure 1 depicts the resulting estimate based on $n = 500$.

Figure 3 depicts the MISE as a function of bandwidth. The MISE curve for the two-step method is always below that 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 in turn suggests 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.

We now illustrate the methodology via an application to an environmental data set. The data set used here consist 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). Of interest is to study the association between levels of pollutants and number of daily total hospital admissions for circulation and respiration and to examine the extent to which the association varies time. We consider the relation among the number of daily hospital admission (Y) and level of pollutant Sulphur Dioxide X_2 (in $\mu g/m^3$), level of pollutant Nitrogen Dioxide X_3 (in $\mu g/m^3$), level of dust X_3 (in $\mu g/m^3$). We took $X_1 = 1$ – the intercept term, and $U = t =$ time. The model

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

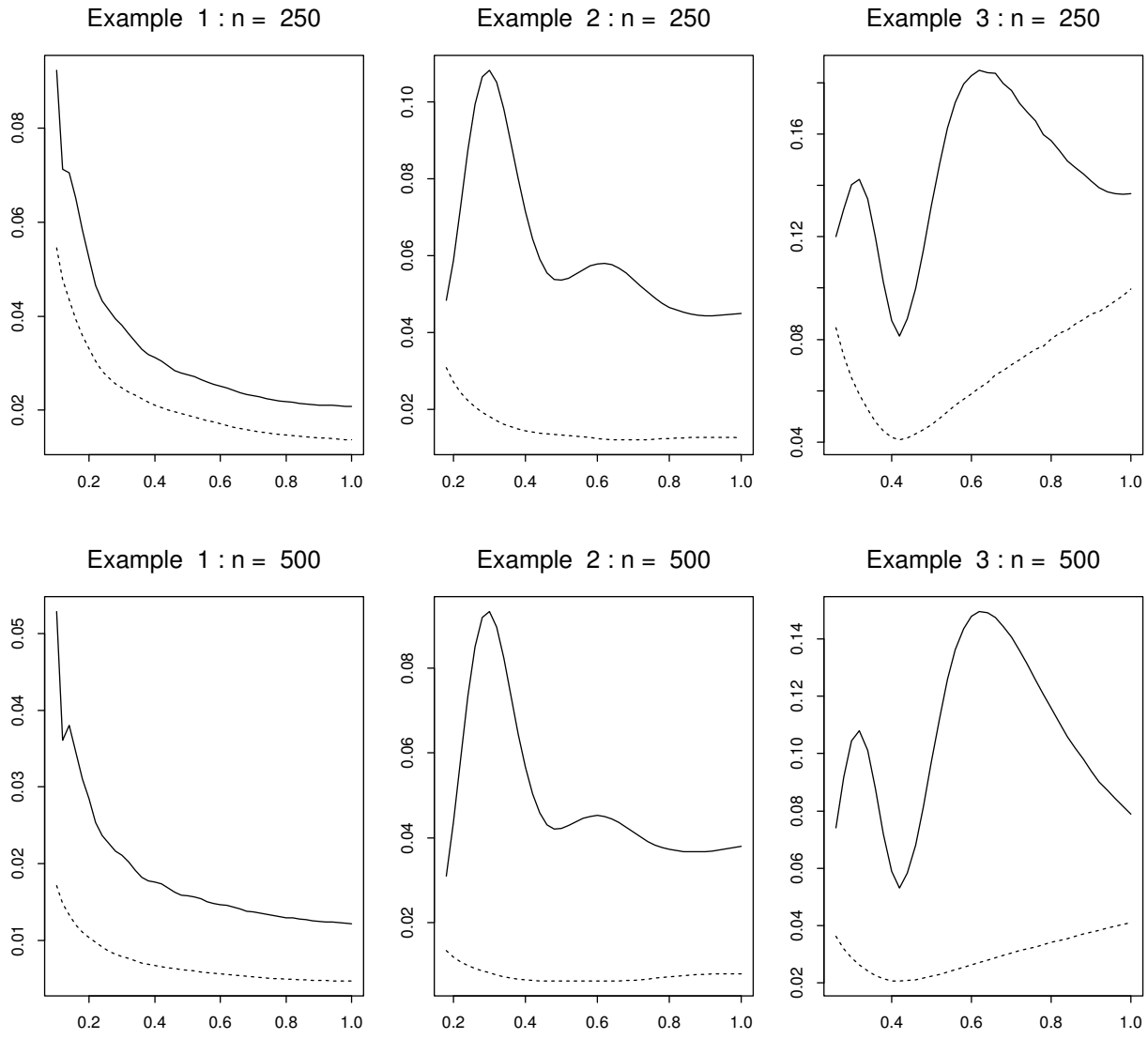

Figure 3: MISE as a function of bandwidth. Solid curve – one-step procedure; dashed curve – two-step procedure

is used to fit the given data. The two-step method is used to estimate the coefficient function $a_i(\cdot)$. An initial bandwidth $h_0 = 0.06 * 729$ (six percents of the whole interval) was chosen. An anticipated, the results do not alters much with different choices of bandwidths. The second stage bandwidths h_2 are chosen respectively 20%, 20%, 25% and 25% of the interval length for functions $a_1, \cdots, a_4$. Figure 4 depicts the estimated coefficient functions. It describes the extent to which the coefficients vary with time. The two dashes curves indicate the 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

the local polynomial regression. The figure indicates that there is some strong time effect. It is quite surprising to see that the time trend $a_1(t)$ is increasing instead of seasonal.

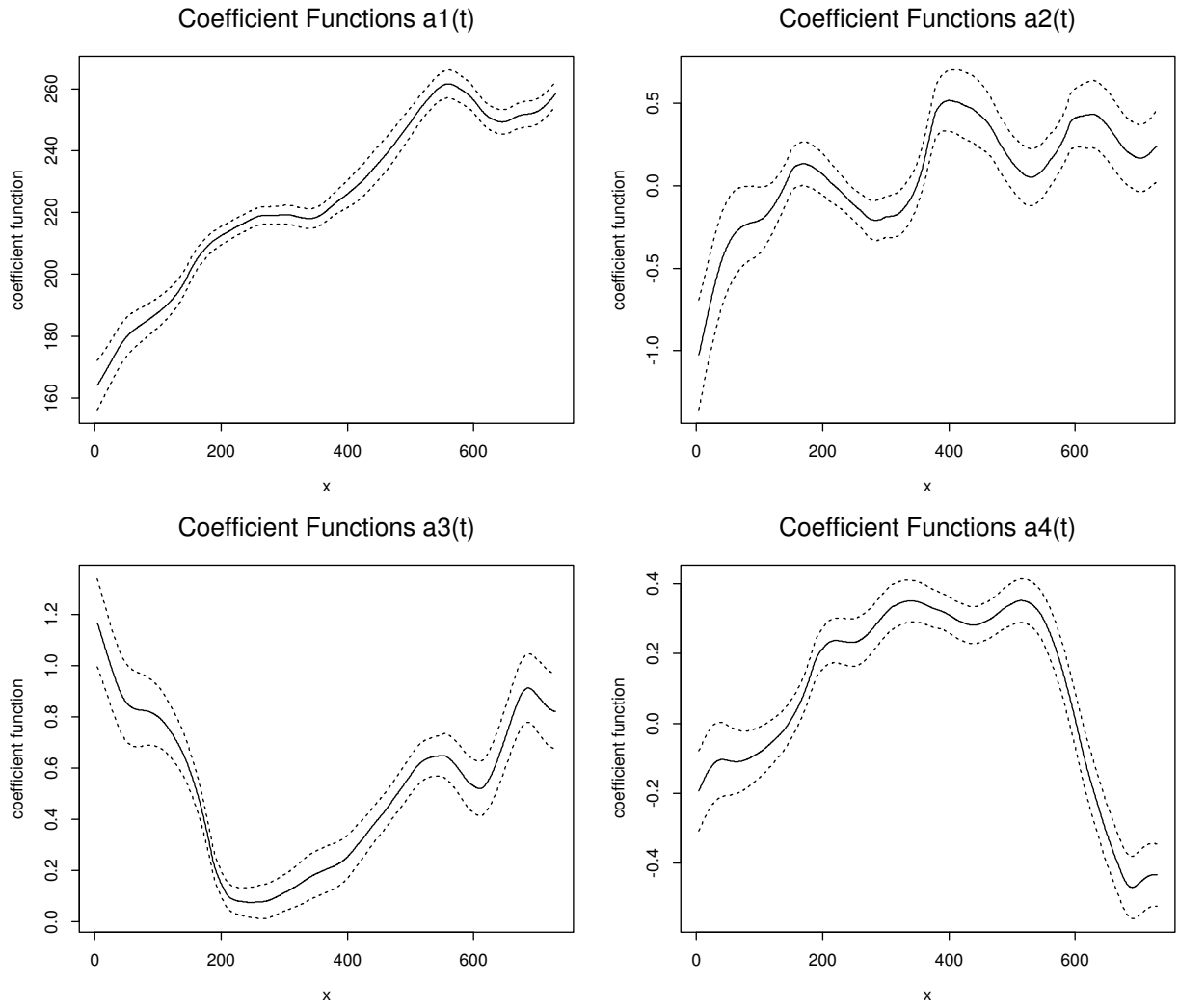

Figure 4: The estimated coefficient functions.

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 $\hat{a}_{p,TS}(u_0)$, the following lemma on the uniform convergence will be need.

Lemma 1 *Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d random vectors, where the Y_i's are scalar random variables. Assume further that $E|y|^s < \infty$ 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*

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

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

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

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

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

$$\beta_{(i)}^T = \sum_{j=1}^{p} a_j''(U_i)\mu_2(\alpha_j^T(U_i), r_{pj}(U_i))\otimes(1,0), \qquad \alpha^{*T} = (\alpha_p^T, r_{pp})\otimes(1,0).$$

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

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

We are now ready to prove our results.

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

$$\begin{aligned} Y = \quad & \mathbf{X}_1(a_1(u_0), a_1'(u_0), \cdots, a_{p-1}(u_0), a_{p-1}'(u_0), a_p(u_0), a_p'(u_0), \tfrac{1}{2}a_p''(u_0), \tfrac{1}{3!}a_p'''(u_0))^T \\ & + \tfrac{1}{2}\sum_{j=1}^{p-1} \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix} + \tfrac{1}{4!} \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1-u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n-u_0)^4 X_{np} \end{pmatrix} + \vec{\varepsilon} \end{aligned}$$

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

$$\begin{aligned} \hat{a}_{p,1}(u_0) = \quad & a_p(u_0) + \tfrac{1}{2}\sum_{j=1}^{p-1} e_{2p-1,2p+2}^T (\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix} \\ & + \tfrac{1}{4!} e_{2p-1,2p+2}^T (\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T W_1 \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1-u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n-u_0)^4 X_{np} \end{pmatrix} \\ & + e_{2p-1,2p+2}^T (\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T W_1 \vec{\varepsilon}. \end{aligned}$$

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

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

and

$$\begin{aligned} \mathbf{X}_3^T W_1 \mathbf{X}_2 &= nf(u_0) A S_{12} D (1 + o_P(1)) \\ \mathbf{X}_2^T W_1 \mathbf{X}_2 &= nf(u_0) D S_{22} D (1 + o_P(1)). \end{aligned} \tag{6.1}$$

Combination of the last three asymptotic expressions leads to

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

Similarly, we have

$$\mathbf{X}_3^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix} = nf(u_0) h_1^2 a_j''(u_0) A \left(\alpha_j \otimes (1,0)^T\right) \mu_2 (1 + o_P(1))$$

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,

$$\begin{aligned} & \mathbf{X}_1^T W_1 \begin{pmatrix} a_j''(\xi_{1j})(U_1-u_0)^2 X_{1j} \\ \vdots \\ a_j''(\xi_{nj})(U_n-u_0)^2 X_{nj} \end{pmatrix} \\ = \;& nf(u_0)h_1^2 a_j''(u_0)\mathrm{diag}(A,D)(\alpha_j^T \otimes (1,0)\mu_2, r_{pj}\mu_2, 0, r_{pj}\mu_4, 0)^T(1+o_P(1)). \end{aligned}$$

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

$$\begin{aligned} & \mathrm{bias}(\hat{a}_{p,1}(u_0)|\mathcal{D}) \\ = \;& \tfrac{1}{2}h_1^2 \sum_{j=1}^{p-1} a_j''(u_0) e_{2p-1,2p+2}^T S^{-1}(\alpha_j^T \otimes (1,0)\mu_2, r_{pj}\mu_2, 0, r_{pj}\mu_4, 0)^T(1+o_P(1)). \end{aligned}$$

Using the properties of the Kronecker product we have

$$\begin{aligned} & \mathrm{bias}(\hat{a}_{p,1}(u_0)|\mathcal{D}) \\ = \;& \tfrac{h_1^2\mu_2}{2(r_{pp}-\alpha_p^T\Omega_{p-1}^{-1}\alpha_p)r_{pp}} \sum_{j=1}^{p-1} (r_{pj}\alpha_p^T\Omega_{p-1}^{-1}\alpha_p - r_{pp}\alpha_p^T\Omega_{p-1}^{-1}\alpha_j) a_j''(u_0)(1+o_P(1)) \\ = \;& -\tfrac{h_1^2\mu_2}{2r_{pp}} \sum_{j=1}^{p-1} r_{pj} a_j''(u_0) + o_P(h_1^2). \end{aligned}$$

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

$$\begin{aligned} & \mathrm{var}(\hat{a}_{p,1}(u_0)|\mathcal{D}) \\ = \;& e_{2p-1,2p+2}^T (\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T W_1 \Psi W_1 \mathbf{X}_1 (\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1} e_{2p-1,2p+2} \\ = \;& \tfrac{\sigma^2(u_0)}{nh_1 f(u_0)} e_{2p-1,2p+2}^T S^{-1}\tilde{S}S^{-1} e_{2p-1,2p+2}(1+o_P(1)). \end{aligned}$$

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

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

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

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

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

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

By Lemma 1, we have

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

and

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

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

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

Since

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

it follows from (3.3) we get

$$\begin{aligned} \hat{a}_{p,2}(u_0) &= a_p(u_0) + \frac{1}{4!}(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} a_p^{(4)}(\eta_1)(U_1-u_0)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(\eta_n)(U_n-u_0)^4 X_{np} \end{pmatrix} \\ &\quad + (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} \sum_{j=1}^{p-1}(a_j(U_1) - \hat{a}_{j,0}(U_1))X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1}(a_j(U_n) - \hat{a}_{j,0}(U_n))X_{nj} \end{pmatrix} \\ &\quad + (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \vec{\varepsilon} \\ &\equiv a_p(u_0) + \frac{1}{4!}\tilde{J}_1 + \tilde{J}_2 + (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \vec{\varepsilon}. \end{aligned} \tag{6.5}$$

By simple calculation we have

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

By (6.4), we have

$$
\begin{aligned}
E(\tilde{J}_2|\mathcal{D}) &= (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} \sum_{j=1}^{p-1} E((a_j(U_1)-\hat{a}_{j,0}(U_1))|\mathcal{D})X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1} E((a_j(U_n)-\hat{a}_{j,0}(U_n))|\mathcal{D})X_{nj} \end{pmatrix} \\
&= -\tfrac{1}{2}h_0^2(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \begin{pmatrix} \sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(1)}^{*-1}\beta_{(1)}X_{1j} \\ \vdots \\ \sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(n)}^{*-1}\beta_{(n)}X_{nj} \end{pmatrix}(1+o_P(1)) \\
&= -\tfrac{h_0^2}{2r_{pp}}(\tfrac{\mu_4}{\mu_4-\mu_2^2},0,-\tfrac{\mu_2}{\mu_4-\mu_2^2},0) \begin{pmatrix} \sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj} \\ 0 \\ \sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj}\mu_2 \\ 0 \end{pmatrix}(1+o_P(1)) \\
&= -\tfrac{h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj}(1+o_P(1)).
\end{aligned}
$$

Therefore, by (6.5) we obtain

$$
\begin{aligned}
&\mathrm{bias}(\hat{a}_{p,2}(u_0)|\mathcal{D}) \\
&= \left(-\tfrac{h_0^2}{2r_{pp}}\sum_{j=1}^{p-1} e_{2j-1,2p}^T S_{(0)}^{*-1}\beta_{(0)}r_{pj} + \tfrac{\mu_4^2-\mu_2\mu_6}{4!(\mu_4-\mu_2^2)}a_p^{(4)}(u_0)h_2^4\right)(1+o_P(1)).
\end{aligned}
$$

By using the properties of the Kronecker product, we have

$$
\begin{aligned}
&\mathrm{bias}(\hat{a}_{p,2}(u_0)|\mathcal{D}) \\
&= \left(\tfrac{1}{4!}\tfrac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 - \tfrac{\mu_2 h_0^2}{2r_{pp}}\sum_{j=1}^{p} a_j''(u_0)(\alpha_p^T,0)\Omega_p^{-1}\begin{pmatrix}\alpha_j \\ r_{pj}\end{pmatrix}\right)(1+o_P(1)) \\
&= \tfrac{1}{4!}\tfrac{\mu_4^2-\mu_6\mu_2}{\mu_4-\mu_2^2}a_p^{(4)}(u_0)h_2^4 - \tfrac{\mu_2 h_0^2}{2r_{pp}}\textstyle\sum_{j=1}^{p-1} a_j''(u_0)r_{pj} + o_P(h_2^4+h_0^2).
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{var}(\hat{a}_{p,2}(u_0)|\mathcal{D}) \\
&= (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 \Psi W_2 \mathbf{X}_2 (\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}(1,0,0,0)^T \\
&\quad -2(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 H \Psi W_2 \mathbf{X}_2 (\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}(1,0,0,0)^T \\
&\quad +(1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 H \Psi H^T W_2 \mathbf{X}_2 (\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}(1,0,0,0)^T. \qquad (6.6)
\end{aligned}
$$

Using similar arguments as before, we can show that

$$\begin{aligned}
& (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1} \mathbf{X}_2^T W_2 \Psi W_2 \mathbf{X}_2 (\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1} (1,0,0,0)^T \\
= \quad & \frac{\mu_4^2 \nu_0 - 2\mu_4 \mu_2 \nu_2 + \mu_2^2 \nu_4}{n h_2 f(u_0) r_{pp} (\mu_4 - \mu_2^2)^2} \sigma^2(u_0)(1 + o_P(1)) \qquad (6.7)
\end{aligned}$$

Since

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

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

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

for $k = 1, \cdots, p$

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

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

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

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

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

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

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

This and (6.1) together yield

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

Let

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

and

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

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

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

Using Lemma 1 and tedious calculation, we obtain

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

Combination of this and (6.1) gives

$$\begin{aligned} & (1,0,0,0)(\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}\mathbf{X}_2^T W_2 H \Psi H^T W_2 \mathbf{X}_2 (\mathbf{X}_2^T W_2 \mathbf{X}_2)^{-1}(1,0,0,0)^T \\ = \; & \frac{\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4}{nh_2 f(u_0) r_{pp}^2(\mu_4 - \mu_2^2)^2} \sum_{j=1}^{p-1}\sum_{m=1}^{p-1} r_{pj} r_{pm} e_{2j-1,2p}^T S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2m-1,2p} \sigma^2(u_0)(1 + o_P(1)). \end{aligned} \tag{6.9}$$

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

$$\begin{aligned}\mathrm{var}(\hat{a}_{p,2}(u_0)|\mathcal{D}) = \; & \frac{\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4}{nh_2 f(u_0) r_{pp}^2(\mu_4 - \mu_2^2)^2} \\ & \left(r_{pp} + \sum_{j=1}^{p-1}\sum_{m=1}^{p-1} r_{pj} r_{pm} e_{2j-1,2p}^T S_{(0)}^{*-1} Q S_{(0)}^{*-1} e_{2m-1,2p} \right. \\ & \left. -2\sum_{j=1}^{p-1} r_{pj} e_{2j-1,2p}^T S_{(0)}^{*-1}\alpha^* \right) \sigma^2(u_0)(1 + o_P(1)).\end{aligned}$$

Using the properties of the Kronecker product we get

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

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. *Jour. Ameri. Statist. Assoc.*, to appear.

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

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991). Local regression models. In *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309–376. Wadsworth & Brooks, Pacific Grove.

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

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. 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. (1997). Direct estimation of additive and linear components for high dimensional data. Accepted by *The Annals of Statistics.*

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 (1993), 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 Tibshirani, R. J. (1993). Varying-coefficient models. *Jour. Roy. Statist. Soc. B.*, **55**, 757-796.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1995). Nonparametric characterization of selection bias using experimental data: a study of adult males in JTPA. *Manuscript.*

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*, to appear.

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. Wahrscheinlichkeitstheorie verw. Gebiete*, **61**, 405–415.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Jour. Ameri. 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). *Applied Statistical Time Series Analysis.* Prentice-Hall.

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

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.

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.