DIMENSION-REDUCTION FOR CONDITIONAL VARIANCE IN REGRESSIONS

LI-PING ZHU and LI-XING ZHU

East China Normal University and Hong Kong Baptist University

Supplementary Material

This note contains the regularity conditions and the details for proving the results in our main context. Since the proofs for Theorems 1, 2, 4 and 6 are more of interest to our theoretical study, they will be demonstrated in detail. The proofs of Theorems 3 and 5 will be outlined in brief because they are similar to Theorem 2 in Zhu and Fang (1996), and Lemma 1 in Xia, Tong, Li and Zhu (2002), respectively.

Appendix A: Some conditions

Some regularity conditions are assumed as follows.

(C1) The kernel function $K_w(\cdot)$ is a continuous density function having bounded support.

(C2) The density function of X satisfies: $0 < \inf_t f_X(t) \le \sup_t f_X(t) < \infty$, and its second derivative $f_X^{(2)}(t)$ satisfies the local Liptschitz condition over the support $\mathcal{T}$ of X, namely, there exist a constant c such that $|f_X^{(2)}(t+v) - f_X^{(2)}(t)| \le c|t|$ for any t in a neighborhood of zero.

(C3) Let $\phi(x) =: E(Y|X=x)$. The $(r+3)-$th derivative of $\phi(x)$, written by $\phi^{(r+3)}(\cdot)$, exists and is continuous over $\mathcal{T}$.

(C4) The variance function $\sigma^2(x) = E[(Y-\phi(X))^2|X=x]$ has a bounded second derivative over $\mathcal{T}$.

(C5) The kernel function $K_l(u)$ is bounded and symmetric, and is Lipschitz continuous on $\mathcal{T}$; Moreover, it satisfies $\int_{\mathcal{T}} K_l(u) = 1$; $\int_{\mathcal{T}} u^i K_l(u) = 0, i = 1, \ldots, d-1$; $\int_{\mathcal{T}} u^d K_l(u) = M_K \neq 0$, $d \ge 2$.

(C6) The bandwidth h_l satisfies $nh_w^2 h_l^{2d_M+2} \to \infty$ as $n \to \infty$.

(C7) The link function $\phi(x)$ and the variance function $\sigma^2(x)$ are bounded on $\mathcal{T}$.

(C8) The bandwidth h_w satisfies $\sqrt{n}h_w^{p+1} \to \infty$ and $\sqrt{n}h_w^{r+1} \to 0$.

(C9) $E[(Y-\phi(X))^4|X=x]$ has a bounded second derivative over $\mathcal{T}$.

(C10) The density function $f(x^T\beta)$ of $X^T\beta$, the density function $f_1(x^T\alpha)$ of $X^T\alpha$, $\phi(x) = E(Y|X=x)$ and $\sigma^2(x) = E(e^2|X=x)$ are $d-$times differentiable on $\mathcal{T}$ and their derivatives satisfy the following condition: Let $H(\cdot)$ stand for $f(x^T\beta)$, $f_1(x^T\alpha)$, $\phi(x)$ and $\sigma^2(x)$ respectively, there exists a neighborhood of the origin, say U, and a constant $c>0$ such that, for any $u \in U$,

$$H^{(d-1)}(t+u) - H^{(d-1)}(t) \le c|u|,$$

where $H^{(d-1)}(t)$ denotes the $(d-1)$th derivative of the function $H(\cdot)$.

Appendix B: Some Lemmas

We first present some Lemmas to facilitate the proof of the theorems.

Recall that a three-step procedure is proposed to remove the effect of the CMS and then to obtain "pure" residuals. The first step is to implement the OPG method to fully identify and to estimate the vectors in the CMS.

Suppose that $\{(x_i, y_i), i = 1, \ldots, n\}$ is a random sample. We consider local $r-$th order polynomial fitting in the form of the following minimization problem

$$\min_{a_t, b_t, c_{ti_1\ldots i_p}} \sum_{i=1}^n \Big[y_i - a_t - b_t^T(x_i - t) - \sum_{1<k\le r}\sum_{i_1+\ldots+i_p=k} c_{ti_1\ldots i_p}\{x_i-t\}_1^{i_1}\ldots \{x_i-t\}_p^{i_p}\Big]^2 K_w\{(x_i-t)/h_w\}, \quad (B.1)$$

where $\{x_i - t\}_k$ denotes the kth element of vector $x_i - t$, and $K_w\{(x_i-t)/h_w\}$ is a $p-$variate kernel.

For ease of illustration, let $\{(x_i-t)_{(k)}^T, i = 1, \ldots, n\}$ denote all distinct columns $\{x_i-t\}_1^{i_1}\ldots\{x_i-t\}_p^{i_p}$ satisfying $i_1+\ldots+i_p = k$, $Y_n = (y_1, \ldots, y_n)^T$ is a vector, and W_{nt} is a diagonal matrix of weights, with entries $K_w\{(x_i-t)/h_w\}$. Denote by X_{nt} the predictor matrix whose (l,j)-block is $(x_l-t)_{(j)}^T$ for $l = 1, \ldots, n$, and $j = 0, \ldots, r$. When $j = 0$, $(x_l - t)_{(0)} = 1$ for all t and l. We re-organize the minimization problem (B.1) as follows:

$$\min_{\beta_{0t},\ldots,\beta_{rt}} \sum_{i=1}^n \Big[y_i - \beta_{0t} - \beta_{1t}^T(x_i-t)_{(1)} - \ldots - \beta_{rt}^T(x_i-t)_{(r)}\Big]^2 K_w\Big(\frac{x_i-t}{h_w}\Big). \quad (B.2)$$

Under the weighted least squares measure (B.2), we have $\widehat{\beta}_t =: (\widehat{\beta}_{0t}, \ldots, \widehat{\beta}_{rt})^T = (X_{nt}^T W_{nt} X_{nt})^{-1}(X_{nt}^T W_{nt} Y_n)$. Clearly, the (i,j)th block of $S_{nt} =: (X_{nt}^T W_{nt} X_{nt})$ is $s_{nt,ij} = \sum_{v=1}^n (x_v - t)_{(i)} K_w\{(x_v - t)/h_w\}(x_v - t)_{(j)}^T$ and the jth element of $S_y =: (X_{nt}^T W_{nt} Y_n)$ is $s_{yj} =: \sum_{v=1}^n (x_v - t)_{(j)} K_w\{(x_v - t)/h_w\} y_v$.

Let $e =: Y_n - \phi(X_n)$ and $\phi(X_n) = (\phi(x_1), \ldots, \phi(x_n))^T$. Hence,

$$\begin{aligned}
\widehat{\beta}_t &= S_{nt}^{-1}(X_{nt}^T W_{nt} Y_n) \\
&= \beta_t + S_{nt}^{-1}[X_{nt}^T W_{nt}(\phi(X_n) - X_{nt}\beta_t)] + S_{nt}^{-1}(X_{nt}^T W_{nt} e) \\
&=: \beta_t + I_{n1}(t) + I_{n2}(t). \qquad \text{(B.3)}
\end{aligned}$$

Therefore, to study the consistency of $\widehat{\beta}_t$, we need to study the last two terms in the RHS of (B.3). Clearly, the second term is for the bias, and the third term for the variance. The following lemma is to provide bounds to these two terms.

Let $\mu_j = \int u^j K_w(u)du$ and $\nu_j = \int u^j K_w^2(u)du$, and denote

$$S_r = (\mu_{i+j-2})_{1 \le i,j \le r+1}, \qquad \tilde{S}_r = (\mu_{i+j-1})_{1 \le i,j \le r+1},$$
$$S_r^* = (\nu_{i+j-2})_{1 \le i,j \le r+1}, \qquad \tilde{S}_r^* = (\nu_{i+j-1})_{1 \le i,j \le r+1}.$$

Lemma 1 *Assume that conditions C1– C4 hold. Then we have that uniformly for t over the support of X, as $n \to \infty$,*

$$\begin{aligned}
I_{n1}(t) &= (X_{nt}^T W_{nt} X_{nt})^{-1}[X_{nt}^T W_{nt}(\phi(X_n) - X_{nt}\beta_t)] \\
&= h_w^{r+1} H^{-1}\left[\frac{S_r^{-1}}{f_X(t)} - h_w \frac{S_r^{-1}}{f_X(t)} f^{(1)}(t)\tilde{S}_r \frac{S_r^{-1}}{f_X(t)}\right] \times \Big[f_X(t)c_r\beta_{r+1} \\
&\quad + h_w[f^{(1)}(t)\tilde{c}_r\beta_{r+1} + f(t)\tilde{c}_r\beta_{r+2}] + O_P(\delta_n^2)\Big], \qquad \text{(B.4)}
\end{aligned}$$

and

$$\begin{aligned}
I_{n2}(t) &= (nh_w^p)^{-1} H^{-1}\left[\frac{S_r^{-1}}{f_X(t)} - h_w \frac{S_r^{-1}}{f_X(t)} f^{(1)}(t)\tilde{S}_r \frac{S_r^{-1}}{f_X(t)} + O_P(\delta_n^2)\right] \times \\
&\quad \left[[\sigma^2(t) f_X(t) S_r^* + h_w[\sigma^2(t) f_X(t)]^{(1)}\tilde{S}_r^* + O_P(\delta_n^2)]\right] \times \\
&\quad \left[\frac{S_r^{-1}}{f_X(t)} - h_w \frac{S_r^{-1}}{f_X(t)} f^{(1)}(t)\tilde{S}_r \frac{S_r^{-1}}{f_X(t)} + O_P(\delta_n^2)\right] H^{-1}, \qquad \text{(B.5)}
\end{aligned}$$

holds uniformly for t, where $H = diag(1, h_w, \ldots, h_w^r)$ is a diagonal matrix with h_w^i corresponding to the column indices of $(x_k - t)_{(i)}$ defined below (B.1), and $\delta_n^2 = h_w^2 + (nh_w^p)^{-1/2}$.

Proof of Lemma 1. We first deal with the bias term $I_{n1}(t)$. Towards this end, we separately consider S_{nt} and $S'_{nt} = [X_{nt}^T W_{nt}(\phi(X_n) - X_{nt}\beta_t)]$.

We now study S_{nt}. Following Fan et al (1996), it suffices to show that the (i,j)th entry $s_{nt,ij}$ of S_{nt} admits the following uniform convergence:

$$s_{nt,ij} = nh_w^{p+i+j}[f_X(t)\mu_{i+j} + h_w f_X^{(1)}(t)h_w\mu_{i+j+1} + O_P(\delta_n^2)], \qquad \text{(B.6)}$$

We then prove that the bias term of $s_{nt,ij}$ is of order $O(nh_w^{p+i+j+2})$. To this end, its expectation is computed as

$$\begin{aligned}
E(s_{nt,ij}) &= nE[(X-t)_{(i)}K_w\{(X-t)/h_w\}(X-t)_{(j)}^T] \\
&= n\int (x-t)_{(i)}K_w\{(x-t)/h_w\}(x-t)_{(j)}^T f_X(x)dx \\
&= n\int h_w^{p+i+j}u^{i+j}K_w(u)f_X(t+h_wu)du \\
&= n\int h_w^{p+i+j}u^{i+j}K_w(u)[f_X(t) + f_X^{(1)}(t)h_wu + f_X^{(2)}(t^*)(h_wu)^2]du \\
&= n\int h_w^{p+i+j}u^{i+j}K_w(u)[f_X(t) + f_X^{(1)}(t)h_wu + f_X^{(2)}(t)(h_wu)^2]du \\
&\quad + \int h_w^{p+i+j}u^{i+j}K_w(u)(f_X^{(2)}(t^*) - f_X^{(2)}(t))(h_wu)^2]du \\
&= nh_w^{p+i+j}[f_X(t)\mu_{i+j} + h_w f^{(1)}(t)\mu_{i+j+1} + O(h_w^2)],
\end{aligned}$$

where t^* lies in the interval $[t, t+h_wu]$. The last equation holds because

$$\int (h_wu)^{i+j+2}K_w(u)|f_X^{(2)}(t^*) - f_X^{(2)}(t)|du \le c\int (h_wu)^{i+j+3}K_w(u)du = O(h_w^{i+j+3}),$$

under the local Liptschitz condition of condition (C2). Therefore, to obtain the uniform convergence of (B.6), it remains to show that the variance term is of order $O(nh_w^p h_w^{2i+2j})$. Note that $\text{Var}(s_{nt,ij}) \le E(s_{nt,ij})^2$, and again by the local Liptschitz condition of condition (C2),

$$\begin{aligned}
E(s_{nt,ij})^2 &= nE[(X-t)_{(i)}K_w\{(X-t)/h_w\}(X-t)_{(j)}^T]^2 \\
&= n\int [(x-t)_{(i)}K_w\{(x-t)/h_w\}(x-t)_{(j)}^T]^2 f_X(x)dx \\
&= n\int h_w^p h_w^{2i+2j}u^{2i+2j}K_w^2(u)f_X(t+h_wu)du \\
&= n\int h_w^p h_w^{2i+2j}u^{2i+2j}K_w^2(u)[f_X(t) + f_X^{(1)}(t)h_wu + f_X^{(2)}(t^*)(h_wu)^2]du \\
&= O(nh_w^p h_w^{2i+2j}).
\end{aligned}$$

Consequently, a direct application of Theorem 37 of Pollard (1984, page 34) yields the desired result of (B.6). This implies that

$$S_{nt} = nh_w^p H[f_X(t)S_r + h_w f_X^{(1)}(t)\tilde{S}_r + O_P(\delta_n^2)]H, \tag{B.7}$$

Now we turn to study the convergence of $S'_{nt} = X_{nt}^T W_{nt}(\phi(X_n) - X_{nt}\beta_t)$ of $I_{n1}(t)$. Denote its jth element of S'_{nt} by $s'_{nt,j}$. Note that the ith row of $\phi(X_n) - X_{nt}\beta_t$ is $\phi(x_i) - \beta_{0t} - \sum_{v=1}^{r}(x_i - t)_{(v)}^T\beta_{vt} = (x_i - t)_{(r+1)}^T\beta_{r+1,t} + (x_i - t)_{(r+2)}^T\beta_{r+2,t} + (x_i - t)_{(r+3)}^T\beta_{r+3,t}^*$ by using conditions (C2) and (C3) and Taylor expansion. Following similar arguments for proving (B.6), we have

$$\begin{aligned} s'_{mt,j} &= nh_w^{p+r+j}\Big[[f_X(t)\mu_{r+j} + f_X^{(1)}(t)\mu_{r+j+1}]\beta_{r+1,t} + h_w[f_X(t)\mu_{r+j+1} \\ &\qquad + f_X^{(1)}(t)\mu_{r+j+2}]\beta_{r+2,t} + O_P(\delta_n^2)\Big], \end{aligned} \tag{B.8}$$

holds uniformly for t over the support of X. The details are omitted.

Further, set $c_r = (\mu_{r+1}, \ldots, \mu_{2r+1})^T$ and $\tilde{c}_r = (\mu_{r+2}, \ldots, \mu_{2r+2})^T$. By using the facts that $(AB)^{-1} = B^{-1}A^{-1}$ and $(A+hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$, it follows from (B.7) and (B.8) that

$$\begin{aligned} I_{n1}(t) = S_{nt}^{-1}S'_{nt} &= (X_{nt}^T W_{nt} X_{nt})^{-1}[X_{nt}^T W_{nt}(\phi(X_n) - X_{nt}\beta_t)] \\ &= h_w^{r+1}H^{-1}\left[\frac{S_r^{-1}}{f_X(t)} - h_w\frac{S_r^{-1}}{f_X(t)}f^{(1)}(t)\tilde{S}_r\frac{S_r^{-1}}{f_X(t)}\right] \times \Big[f_X(t)c_r\beta_{r+1} \\ &\qquad + h_w[f^{(1)}(t)\tilde{c}_r\beta_{r+1} + f(t)\tilde{c}_r\beta_{r+2}] + O_P(\delta_n^2)\Big], \end{aligned} \tag{B.9}$$

holds uniformly for t.

It is the position to bound the variance term $I_{n2}(t) = S_{nt}^{-1}(X_{nt}^T W_{nt} e)$ of (B.5). Note that $\mathrm{Var}[S_{nt}^{-1}(X_{nt}^T W_{nt} e)] = E[\mathrm{Var}(S_{nt}^{-1}(X_{nt}^T W_{nt} e)|X_n)]$. We then first study the conditional variance. Let $\sigma^2(X) =: E[(Y - E(Y|X))^2|X] = E(e^2|X)$, and $\sigma^2(X_n) = (\sigma^2(x_1), \ldots, \sigma^2(x_n))^T$. Hence,

$$\mathrm{Var}[S_{nt}^{-1}(X_{nt}^T W_{nt} e)|X_n] = S_{nt}^{-1}S_{nt}^* S_{nt}^{-1}, \tag{B.10}$$

where $S_{nt}^* = X_{nt}^T W_{nt}\mathrm{Var}(e|X_n)W_{nt}X_{nt}$, with its $(i,j)-$th entry $s_{nt,ij}^* = \sum_{v=1}^{n}(x_v - t)_i K_w^2\{(x_v - t)/h_w\}(x_v - t)_j^T\sigma^2(x_v)$. From conditions (C2), (C4) and similar arguments as used for proving (B.6), we derive that

$$s_{nt,ij}^* = nh_w^{p+i+j}[\sigma^2(t)f_X(t)\nu_{i+j} + h_w[\sigma^2(t)f_X(t)]^{(1)}\nu_{i+j+1} + O_P(\delta_n^2)]. \tag{B.11}$$

From (B.6), (B.7) and (B.11), we obtain that

$$\begin{aligned} S_{nt}^{-1}S_{nt}^{*}S_{nt}^{-1} &= \mathrm{Var}[S_{nt}^{-1}(X_{nt}^{T}W_{nt}e)|X_n] \\ &= (nh_w^p)^{-1}H^{-1}\left[\frac{S_r^{-1}}{f_X(t)} - h_w\frac{S_r^{-1}}{f_X(t)}f^{(1)}(t)\tilde{S}_r\frac{S_r^{-1}}{f_X(t)} + O_P(\delta_n^2)\right] \times \\ &\quad \left[[\sigma^2(t)f_X(t)S_r^{*} + h_w[\sigma^2(t)f_X(t)]^{(1)}\tilde{S}_r^{*} + O_P(\delta_n^2)]\right] \times \\ &\quad \left[\frac{S_r^{-1}}{f_X(t)} - h_w\frac{S_r^{-1}}{f_X(t)}f^{(1)}(t)\tilde{S}_r\frac{S_r^{-1}}{f_X(t)} + O_P(\delta_n^2)\right]H^{-1}. \quad (B.12) \end{aligned}$$

Notice that $\mathrm{Var}(I_{n2}) = E[S_{nt}^{-1}S_{nt}^{*}S_{nt}^{-1}]$. Therefore, (B.5) holds. #

The convergence of $\phi(t)$ and its first derivative $\phi^{(1)}(t)$ follows directly from (B.3), (B.9), (B.10) and (B.5), which is stated in the following Lemma.

LEMMA 2 *Assume that the conditions C1-C4 hold. We have*

$$\begin{aligned} \sup_t |\widehat{\phi}(t) - \phi(t)| &= O_P\{h_w^{r+1} + (nh_w^p)^{-1/2}\log n\}, \quad (B.13) \\ \sup_t |\widehat{\phi}^{(1)}(t) - \phi^{(1)}(t)| &= O_P\{h_w^{r} + (nh_w^{p+2})^{-1/2}\log n\}. \quad (B.14) \end{aligned}$$

The third step is to approximate the link function $\phi(\cdot)$ by a different kernel function $K_{lh_l}(\cdot)$ and then obtain a "pure" residual without impact from the CMS, as described in the beginning of Section 5. Specifically, when β is a $p \times d_M$ matrix which spans the CMS, we have, $\phi(x) = E(Y|X) = E(Y|X^T\beta)$. Thus $\phi(x)$ can be estimated in the following way,

$$\begin{aligned} \hat{\phi}(x) &= \hat{g}(x^T\beta)/\hat{f}(x^T\beta) \\ &= \frac{1}{nh_l^{d_M}}\sum_{i=1}^{n} y_i K_l\left(\frac{x_i^T\beta - x^T\beta}{h_l}\right) \Bigg/ \frac{1}{nh_l^{d_M}}\sum_{i=1}^{n} K_l\left(\frac{x_i^T\beta - x^T\beta}{h_l}\right), \end{aligned}$$

where $K_l(\cdot)$ is a $\mathcal{R}^{d_M} \to \mathcal{R}$ kernel function with the bandwidth h_l, and the true structural dimension, d_M, of the CMS. In the following we write $K_{lh_l}(\cdot) = K_l(\cdot/h_l)/h_l^{d_M}$ for notational clarity.

After we remove the effect of the CMS exhaustively through the three-step procedure, we proposed the e^2-based OPG method to target the CVS. The following Lemma 3 serves for proving its asymptotic normality.

Before we present Lemma 3, we define

$$\xi_j(x, \beta^*) = K_{lh_l}(x_j^T\beta^* - x^T\beta^*)y_j, \text{ and}$$

$$\alpha_n(x,\beta^*) \;\; = \;\; \frac{1}{n}\sum_{j=1}^{n}\Big(\xi_j(x,\beta^*) - E[\xi_j(x,\beta^*)]\Big).$$

Obviously, $\hat{g}(x^T\beta^*) = \frac{1}{n}\sum_{j=1}^{n}\xi_j(x,\beta^*)$.

LEMMA 3 *Under conditions (C1), (C2), and (C5)-(C7). Then, we have, as* $n \to \infty$,

$$\sup_{x\in R^p}\;\sup_{\|\beta^*-\beta\|=O\{n^{-1/2}h_w^{-1}\}} |\alpha_n(x,\beta^*) - \alpha_n(x,\beta)| = O_P\{\log n\cdot n^{-1}h_w^{-1}h_l^{-d_M-1}\}.$$

Proof of Lemma 3: From the definition of $\alpha_n(x,\beta^*)$, we have

$$\alpha_n(x,\beta^*) - \alpha_n(x,\beta) = \frac{1}{n}\sum_{j=1}^{n}\bigg(\Big(\xi_j(x,\beta^*) - \xi_j(x,\beta)\Big) - E\Big(\xi_j(x,\beta^*) - \xi_j(x,\beta)\Big)\bigg).$$

Let $\epsilon > 0$ be given. In order to use chain lemma (see, e.g. Pollard, 1984), we show that

$$P\Big(\Big|\xi_j(x,\beta^*) - \xi_j(x,\beta)\Big| > \frac{\epsilon}{2}\Big) \le \frac{1}{2}, \text{ for } x \in \mathcal{R}^p, \|\beta^*-\beta\| = O(n^{-\frac{1}{2}}). \quad \text{(B.15)}$$

By Chebyschev's inequality, the LHS of (B.15) is less than or equal to $\frac{4}{\epsilon^2}E\Big(\Big|\xi_n(x,\beta^*) - \xi_n(x,\beta)\Big|\Big)^2$. Note that

$$\begin{aligned}
& \frac{4}{\epsilon^2}E\Big(\xi(x,\beta^*) - \xi(x,\beta)\Big)^2 \\
= \;& \frac{4}{\epsilon^2}E\Big(K_{lh_l}(\beta^*X - \beta^*x)Y - K_{lh_l}(\beta^TX - \beta^Tx)Y\Big)^2 \\
= \;& \frac{4}{\epsilon^2}E\Big(Y^2[K_{lh_l}(\beta^*X - \beta^*x) - K_{lh_l}(\beta^TX - \beta^Tx)]^2\Big) \\
= \;& \frac{4}{\epsilon^2}E\Big([\phi(X) + \sigma(X)]^2[K_{lh_l}(\beta^*X - \beta^*x) - K_{lh_l}(\beta^TX - \beta^Tx)]^2\Big) \\
\le \;& \frac{8}{\epsilon^2}E\Big(\phi^2(X)[K_{lh_l}(\beta^*X - \beta^*x) - K_{lh_l}(\beta^TX - \beta^Tx)]^2\Big). \\
& + \frac{8}{\epsilon^2}E\Big(\sigma^2(X)[K_{lh_l}(\beta^*X - \beta^*x) - K_{lh_l}(\beta^TX - \beta^Tx)]^2\Big). \qquad \text{(B.16)}
\end{aligned}$$

Let $T = \Big\{t \in \mathcal{R}^p : K_l\Big((t-x)\beta^*/h_l\Big) > 0 \text{ or } K_l\Big((t-x)\beta/h_l\Big) > 0\Big\}$ and $U = \Big\{u \in \mathcal{R}^p : x + h_l u \in T\Big\} = \Big\{u \in \mathcal{R}^p : K_l(u^T\beta^*) > 0 \text{ or } K_l(u^T\beta) > 0\Big\}$.

It is easy to prove that the second term of (B.16) has the same order as the first term because its structure is essentially the same as the first one. Therefore, we only deal with the first term of (B.16). Note that

$$
\begin{aligned}
& \frac{8}{\epsilon^2} E\Big(\phi^2(X)\Big(K_{lh_l}(\beta^* X - \beta^* x) - K_{lh_l}(\beta^T X - \beta^T x)\Big)^2\Big) \\
= & \frac{8}{\epsilon^2} \int_{t\in \mathcal{T}} \Big(\phi^2(t)\Big(K_{lh_l}(\beta^* t - \beta^* x) - K_{lh_l}(\beta^T t - \beta^T x)\Big)^2\Big) f_X(t)dt \\
= & \frac{8}{\epsilon^2 h_l^{2d_M}} \int_{t\in \mathcal{T}} \Big(\phi^2(t)\Big(K_l\Big(\frac{\beta^* t - \beta^* x}{h_l}\Big) - K_l\Big(\frac{\beta^T t - \beta^T x}{h_l}\Big)\Big)^2\Big) f_X(t)dt \\
= & \frac{8h_l^p}{\epsilon^2 h_l^{2d_M}} \int_U \Big(\phi^2(x+h_l u)\Big(K_l(\beta^* u) - K_l(\beta^T u)\Big)^2\Big) f_X(x+h_l u)du \\
\leq & \frac{8h_l^p}{\epsilon^2 h_l^{2d_M}} C M_\phi^2 M_f \int_U \Big(u^T(\beta^* - \beta)\Big)^2 du \\
\leq & \frac{8h_l^p}{\epsilon^2 h_l^{2d_M}} C M_\phi^2 M_f O(h_l^{-p}) O(h_l^{-2} n^{-1} h_w^{-2}) = O\Big(\frac{1}{n h_w^2 h_l^{2d_M+2}}\Big)
\end{aligned}
$$

where M_ϕ and M_f are the upper bounds of $\phi(X)$ and f_X, respectively. Together with condition C6 that $n h_w^2 h_l^{2d_M+2} \to \infty$, we obtain that the LHS of (B.15) is less than or equal to $\frac{1}{2}$ when n is large enough. We have the local result of (B.15).

By setting $\delta_n^2 = O\Big(\frac{1}{n h_w^2 h_l^{2d_M+2}}\Big)$ and $\alpha_n^2 = \log^2 n \cdot h_w^2 \cdot h_l^{2d_M+2}$, the remaining proof is a modification of the proof of Theorem 37 in Pollard (1984, page 34). The details are omitted. □

Appendix C: Proof of Theorems

Proof of Theorem 1: Because it is easy to show that part (a) of Theorem 1 implies part (b); part (c) implies part (a) and part (b); part (d) implies part (b), then the proof of Theorem 1 can be concluded by proving (1): (b) implies (c); and (2): (c) implies (d).

(1): Expanding the LHS of (b), we obtain

$$
\begin{aligned}
& E[e^2 \mathrm{Var}(Y|X)|\alpha^T X] = E\Big([Y - E(Y|X)]^2 \mathrm{Var}(Y|X)|\alpha^T X\Big) \\
= & E\Big(E[(Y - E(Y|X)]^2 \mathrm{Var}(Y|X)|X|\alpha^T X\Big) = E[\mathrm{Var}^2(Y|X)|\alpha^T X],
\end{aligned}
$$

and

$$
\begin{aligned}
& E(e^2|\alpha^T X) E[\mathrm{Var}(Y|X)|\alpha^T X] \\
= & E\Big(E[[Y - E(Y|X)]^2|X]|\alpha^T X\Big) E[\mathrm{Var}(Y|X)|\alpha^T X] = E^2[\mathrm{Var}(Y|X)|\alpha^T X].
\end{aligned}
$$

By (b), we have

$$\mathrm{Var}[\mathrm{Var}(Y|X)|\alpha^T X] = E[\mathrm{Var}^2(Y|X)|\alpha^T X] - E^2[\mathrm{Var}(Y|X)|\alpha^T X] = 0.$$

Therefore, $\mathrm{Var}(Y|X)$ is a measurable function of $\alpha^T X$.
(2): Now we prove that (c) implies (d). Note that

$$\begin{aligned}
& \mathrm{Cov}\Big([Y - E(Y|X)]^2, l(X)|\alpha^T X\Big) \\
= \ & E\Big([Y - E(Y|X)]^2 l(X)|\alpha^T X\Big) - E\Big([Y - E(Y|X)]^2|\alpha^T X\Big) E[l(X)|\alpha^T X] \\
= \ & E[\mathrm{Var}(Y|X) l(X)|\alpha^T X] - E[\mathrm{Var}(Y|X)|\alpha^T X] E[l(X)|\alpha^T X] \\
= \ & \mathrm{Var}(Y|X) E[l(X)|\alpha^T X] - \mathrm{Var}(Y|X) E[l(X)\Big|\alpha^T X] = 0.
\end{aligned}$$

This leads to the desired result. □

Proof of Theorem 2: For notational convenience, we write $b_j = \phi^{(1)}(x_j)$ and its corresponding estimator $\hat{b}_j = \widehat{\phi}^{(1)}(x_j)$. We expand $\hat{b}_j \hat{b}_j^T$ into three parts to obtain

$$\begin{aligned}
\frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j \hat{b}_j^T &= \frac{h_w}{\sqrt{n}} \sum_{j=1}^n (\hat{b}_j - b_j + b_j)(\hat{b}_j - b_j + b_j)^T \\
= \ & \frac{h_w}{\sqrt{n}} \sum_{j=1}^n [(\hat{b}_j - b_j)(\hat{b}_j - b_j)^T + 2(\hat{b}_j - b_j) b_j^T + b_j b_j^T] = I_1 + I_2 + I_3.
\end{aligned}$$

By invoking Lemma 2 and condition C8, we have

$$I_1 = o_P(1). \tag{C.1}$$

Now we deal with I_2. For notational clarity, we introduce a block matrix $v_1 = (\mathbf{0}, I_p, \mathbf{0}, \ldots, \mathbf{0})$ with the $p \times p$ identity matrix I_p corresponding to the column indices of $\{(x_i - t)_{(1)}\}$ in X_{nt}. Therefore, $b_j = v_1 \beta_j$ and $\hat{b}_j = v_1 \hat{\beta}_j$ with a slight abuse of notation. Thus we analyze the convergence of $\frac{1}{\sqrt{n}} \sum_{j=1}^n H(\hat{\beta}_j - \beta_j)\beta_j^T$. Let $\varepsilon_n = Y_n - \phi(X_n)$. After simple algebra calculation, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{j=1}^n H(\hat{\beta}_j - \beta_j)\beta_j^T &= \frac{1}{\sqrt{n}} \sum_{j=1}^n H[(X_{nj}^T W_{nj} X_{nj})^{-1}(X_{nj}^T W_{nj} Y_n) - \beta_j]\beta_j^T \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^n H(X_{nj}^T W_{nj} X_{nj})^{-1} X_{nj}^T W_{nj}[\phi(X_n) - X_{nj}\beta_j]\beta_j^T \\
&\quad + \frac{1}{\sqrt{n}} \sum_{j=1}^n H(X_{nj}^T W_{nj} X_{nj})^{-1} X_{nj}^T W_{nj} \varepsilon_n \beta_j^T =: I_{21} + I_{22}.
\end{aligned}$$

Equation (B.9) entails that $\frac{1}{\sqrt{n}}\sum_{j=1}^n H(X_{nj}^T W_{nj} X_{nj})^{-1} X_{nj}^T W_{nj}[\phi(X_n) - X_{nj}\beta_j] = O(\sqrt{n}h_w^{r+1})$ almost surely. Recalling that $\sqrt{n}h_w^{r+1} \to 0$, we can have $I_{21} = o_P(1)$.

We turn to deal with I_{22} now. By using the result of (B.5), we have

$$\begin{aligned} I_{22} &= \frac{1}{\sqrt{n}nh_w^p} S_r^{-1} H^{-1} \sum_{j=1}^n \frac{1}{f_X(x_j)} X_{nj}^T W_{nj} \varepsilon_n \beta_j^T \\ &\quad + \frac{1}{\sqrt{n}} \sum_{j=1}^n H\Big[(X_{nj}^T W_{nj} X_{nj})^{-1} - H^{-1} S_r^{-1} H^{-1}/(nh_w^p f_X(x_j))\Big] X_{nj}^T W_{nj} \varepsilon_n \beta_j^T \\ &=: I_{221} + I_{222}. \end{aligned}$$

Following standard arguments of $U-$statistic theory (Serfling 1980), we can find that

$$I_{221} = \frac{1}{\sqrt{n}} \sum_{j=1}^n S_r^{-1}(\mu_1, \ldots, \mu_r)^T \varepsilon_j \beta_j^T + o_P(1). \tag{C.2}$$

By (B.5) and (C.2), we can easily have $I_{222} = o_P(1)$. Therefore,

$$I_2 = \frac{1}{\sqrt{n}} \sum_{j=1}^n S_r^{-1}(\mu_1, \ldots, \mu_r)^T [y_j - \phi(x_j)] \beta_j^T + o_P(1). \tag{C.3}$$

Some elementary calculation yields

$$I_3 = \frac{h_w}{\sqrt{n}} \sum_{j=1}^n (b_j b_j^T - \Delta) = O_P(h_w) = o_P(1). \tag{C.4}$$

The desired conclusion follows from (C.1), (C.3) and (C.4). □

Proof of Theorem 3: The proof is almost identical to that for Theorem 2 in Zhu and Fang (1996). Hence we omit the details. □

Proof of Theorem 4: First note that for any eigenvalue $\lambda_l(\Omega_n)$, $1 \le l \le p$, $[\log \lambda_i(\Omega_n) + 1 - \lambda_i(\Omega_n)] \le 0$. If $k < K_1$, then $\min(k, \tau) = k$ for large n. Let

$$\log L_{K_1} - \log L_k = -\frac{n}{2} \sum_{i=k+1}^{K_1} [\log \lambda_i(\Omega_n) + 1 - \lambda_i(\Omega_n)] =: \frac{n}{2} W_{n1}(K_1, k).$$

Hence,

$$\begin{aligned} G(K_1) - G(k) &= \log L_{K_1} - \log L_k - C_n(K_1 - k)(2p - k - K_1 + 1)/2 \\ &= [nW_{n1}(K_1, k) - C_n(K_1 - k)(2p - k - K_1 + 1)]/2, \end{aligned}$$

and then, Theorem 2 yields that when $k < K_1$, $W_{n1}(K_1, k) > c$ for a constant, and

$$P\Big(G(K_1) - G(k) < 0\Big) \to 0, \text{ as } \tfrac{C_n}{n} \to 0. \qquad (C.5)$$

If $k > K_1$, then $\min(k, \tau) = K_1$. Similarly

$$\log L_{K_1} - \log L_k = -\frac{n}{2} \sum_{i=K_1+1}^{k} \Big(\log \lambda_i(\Omega_n) + 1 - \lambda_i(\Omega_n) \Big) =: \frac{n}{2} W_{n2}(K_1, k).$$

Thus, when $k > K_1$,

$$\begin{aligned} & G(K_1) - G(k) = \log L_{K_1} - \log L_k - C_n(K_1 - k)(2p - k - K_1 + 1)/2 \\ = & [nW_{n2}(K_1, k) - C_n(K_1 - k)(2p - k - K_1 + 1)]/2 \\ = & -nh_w^2 \sum_{i=K_1+1}^{k} [\lambda_i(\Omega_n) - 1]^2 [1 + o_P(1)]/(2h_w^2) - C_n(K_1 - k)(2p - k - K_1 + 1)/2. \end{aligned}$$

Therefore, when $k > K_1$, we employ Theorem 2 to get

$$P\Big(G(K_1) - G(k) < 0\Big) \to 0, \text{ as } C_n h_w^2 \to \infty. \qquad (C.6)$$

Conclusively, it follows from (C.5) and (C.6) that $\widehat{K_1} \to K_1$ in probability. The proof is concluded. □

Proof of Theorem 5: The proof is essentially the same as that for Lemma 1 in Xia, Tong, Li and Zhu (2002). Hence the details are omitted. □

Proof of Theorem 6: Without notational confusion, we write the $p \times \widehat{K}$ matrix $\widehat{\beta}$ as $\widehat{\beta}_{\widehat{K}}$, and $p \times K$ matrix $\widehat{\beta}$ as $\widehat{\beta}_K$. First, we show that $\widehat{\beta}_{\widehat{K}}$ in $\frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^* b_j^{*,T}$ can be replaced by $\widehat{\beta}_K$, which does not affect the asymptotic distribution of $\frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^* b_j^{*,T}$. This claim can be verified as follows. On one hand,

$$\begin{aligned} & P\Big\{ \frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^{*2} \le t \Big\} \\ = & P\Big\{ \frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^{*2} \le t, \widehat{K} = K \Big\} + P\Big\{ \frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^{*2} \le t, \widehat{K} \ne K \Big\} \\ = & P\Big\{ \frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^{*2} \le t \Big| \widehat{K} = K \Big\} + o(1). \end{aligned} \qquad (C.7)$$

The last equation holds because $\widehat{K} \to K$ in probability (this is proven in Theorem 4), the RHS in the above equation has the same limit as $P\Big\{ \frac{h_w}{\sqrt{n}} \sum_{j=1}^n \hat{b}_j^{*2} \le t \Big| \widehat{K} =$

$K\Big\}$ as $n \to \infty$. On the other hand,

$$P\Big\{\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*2} \le t\Big\} \ge P\Big\{\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*2} \le t\Big|\widehat{K}=K\Big\}P\Big\{\widehat{K}=K\Big\}$$
$$= P\Big\{\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*2} \le t\Big|\widehat{K}=K\Big\} - o(1). \qquad \text{(C.8)}$$

By (C.7) and (C.8) together, we show that $\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*}\hat{b}_j^{*,T}$ based on $\widehat{\beta}_{\widehat{K}}$ has the same asymptotic distribution as that based on $\widehat{\beta}_K$. Therefore, in the following, we can treat the involved $\widehat{\beta}_{\widehat{K}}$ as $\widehat{\beta}_K$ throughout the proof.

Expand $\hat{b}_j^{*}\hat{b}_j^{*,T}$ into three parts:

$$\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*}\hat{b}_j^{*,T} = \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}(\hat{b}_j^{*} - b_j^{*} + b_j^{*})(\hat{b}_j^{*} - b_j^{*} + b_j^{*})^T$$
$$= \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\Big((\hat{b}_j^{*} - b_j^{*})(\hat{b}_j^{*} - b_j^{*})^T + 2(\hat{b}_j^{*} - b_j^{*})b_j^{*,T} + b_j^{*}b_j^{*,T}\Big) = I_4 + I_5 + I_6.$$

Deal with I_5 first. I_4 can be handled in a similar way.

After simple algebraic calculation, we have

$$\begin{aligned}
\frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}\hat{b}_j^{*}b_j^{*,T} &= \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}v_1 \times [X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} \hat{e}_n^2 b_j^{*,T} + o_P(1) \\
&= \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}v_1 [X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} e_n^2 b_j^{*,T} \\
&\quad + \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}v_1 [X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} (\hat{e}_n^2 - e_n^2) b_j^{*,T} \\
&= I_{51} + I_{52}.
\end{aligned}$$

Similar to the proof of (C.3) in Theorem 2,

$$I_{51} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}S_r^{-1}(\mu_1,\ldots,\mu_r)^T[e_j^2 - \sigma^2(x_j)]\beta_j^T + o_P(1) \qquad \text{(C.9)}$$

follows from the condition (C9).

Decompose I_{52} into two terms: $I_{52} = \frac{h_w}{\sqrt{n}}\sum_{j=1}^{n}v_1[X_{nj}^T W_{nj} X_{nj}]^{-1}X_{nj}W_{nj}[(\hat{e}_n - e_n)^2 + 2e_n(\hat{e}_n - e_n)]b_j^{*,T} =: I_{521} + I_{522}$, where $e_n = \{y_1 - \phi(\beta^T x_1), \ldots, y_n - \phi(\beta^T x_n)\}^T$ and $\hat{e}_n = \{y_1 - \hat{\phi}(\hat{\beta}^T x_1), \ldots, y_n - \hat{\phi}(\hat{\beta}^T x_n)\}^T$. A direct application of Lemma 3 shows that $e_i - \hat{e}_i = [\phi(x_i^T\hat{\beta}) - \phi(x_i^T\beta)] + [\hat{\phi}(x_i^T\beta) - \phi(x_i^T\beta)] + O_P[(\log n)^{1/2} \cdot n^{-1}h_w^{-1}h_l^{-d_M-1} + h_l^d]$ holds almost surely for all x_i's. We now prove that I_{522} is $o_P(1)$ and hence $I_{521} = o_P(1)$ can be obtained from the convergence of $\hat{\phi}(\cdot)$ and $\hat{\beta}$. Specifically,

$$\begin{aligned} I_{522} &= \frac{h_w}{\sqrt{n}}\sum_{j=1}^n v_1[X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} e_n[\hat{\phi}(X_n^T\hat{\beta}) - \hat{\phi}(X_n^T\beta)]b_j^{*,T} \\ &\quad -\frac{h_w}{\sqrt{n}}\sum_{j=1}^n v_1[X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} e_n[\hat{\phi}(X_n^T\beta) - \phi(X_n^T\beta)]b_j^{*,T} \\ &= \frac{h_w}{\sqrt{n}}\sum_{j=1}^n v_1[X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} e_n[\phi(X_n^T\hat{\beta}) - \phi(X_n^T\beta)]b_j^{*,T} \\ &\quad -\frac{h_w}{\sqrt{n}}\sum_{j=1}^n v_1[X_{nj}^T W_{nj} X_{nj}]^{-1} X_{nj} W_{nj} e_n[\hat{\phi}(X_n^T\beta) - \phi(X_n^T\beta)]b_j^{*,T} + o_P(1) \\ &=: I_{5221} + I_{5222}. \end{aligned}$$

Using Taylor expansion and Slutsky Theorem, we can derive that $I_{5221} = o_P(1)$. By using the fact that

$$\sup_{x_i^T\beta} |\hat{\phi}(x_i^T\beta) - \phi(x_i^T\beta)| = O\{h_l^d + (nh_l^{d_M})^{-1/2}\log n\}, \tag{C.10}$$

and similar arguments for proving (B.11), we can have $I_{5222} = o_P(1)$, and hence $I_{5221} = o_P(1)$. Consequently,

$$I_5 = \frac{1}{\sqrt{n}}\sum_{j=1}^n S_r^{-1}(\mu_1, \ldots, \mu_r)^T[e_j^2 - \sigma^2(x_j)]\beta_j^T + o_P(1). \tag{C.11}$$

Similarly, we can have $I_4 = o_P(1)$ and $I_6 = o_P(1)$. The result follows. □

Acknowledgment The first author was supported by a NSF grant from National Natural Science Foundation of China (No10701035), ChenGuang project of Shanghai Education Development Foundation (No2007CG33), a special fund for young teachers in Shanghai universities (No79001320), and Science and Technology Commission of Shanghai Municipality Pujiang Project (No. 07PJ14037). The second author was supported by a grant (HKU 7058/05P) from the Research Grants Council of Hong Kong, and a FRG grant from Hong Kong Baptist University, Hong Kong. The authors are grateful to the editor, an associate editor, and the two referees for their generous help and their constructive comments and suggestions, which led to a great improvement of our earlier draft.

References

Fan J. Q, Gijbels, I. Hu, T. C. and Huang, L. S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statist. Sinica* **6,** 113–127.

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

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* John Wiley & Sons, New York.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.X. (2002). An adaptive estimation of optimal regression subspace, *J. Roy. Statist. Soc. B* **64,** 363–410.

LI-PING ZHU: School of finance and statistics, East China Normal University, No. 500 Dong Chuang Road, Shanghai, China.

E-mail: lpzhu@stat.ecnu.edu.cn

LI-XING ZHU: The corresponding author. Fax: (852)-3411-5811.

Department of mathematics, Hong Kong Baptist University, Hong Kong, China.

E-mail: lzhu@hkbu.edu.hk