The Annals of Statistics
2007, Vol. 35, No. 6, 2654–2690
DOI: 10.1214/009053607000000352

A CONSTRUCTIVE APPROACH TO THE ESTIMATION OF DIMENSION REDUCTION DIRECTIONS[1]

BY YINGCUN XIA

National University of Singapore

In this paper we propose two new methods to estimate the dimension-reduction directions of the central subspace (CS) by constructing a regression model such that the directions are all captured in the regression mean. Compared with the inverse regression estimation methods [e.g., *J. Amer. Statist. Assoc.* **86** (1991) 328–332, *J. Amer. Statist. Assoc.* **86** (1991) 316–342, *J. Amer. Statist. Assoc.* **87** (1992) 1025–1039], the new methods require no strong assumptions on the design of covariates or the functional relation between regressors and the response variable, and have better performance than the inverse regression estimation methods for finite samples. Compared with the direct regression estimation methods [e.g., *J. Amer. Statist. Assoc.* **84** (1989) 986–995, *Ann. Statist.* **29** (2001) 1537–1566, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** (2002) 363–410], which can only estimate the directions of CS in the regression mean, the new methods can detect the directions of CS exhaustively. Consistency of the estimators and the convergence of corresponding algorithms are proved.

1. Introduction. Suppose X is a random vector in $\mathbb{R}^p$ and Y is a univariate random variable. Let $B_0 = (\beta_{01}, \dots, \beta_{0q})$ denote a $p \times q$ orthogonal matrix with $q \le p$, that is, $B_0^\top B_0 = I_q$, where I_q is a $q \times q$ identity matrix. Given $B_0^\top X$, if Y and X are independent, that is, $Y \perp\!\!\!\perp X | B_0^\top X$, then the space spanned by the column vectors $\beta_{01}, \beta_{02}, \dots, \beta_{0q}$, $\mathcal{S}(B_0)$, is called the dimension reduction space. If all the other dimension reduction spaces include $\mathcal{S}(B_0)$ as their subspace, then $\mathcal{S}(B_0)$ is the so-called central dimension reduction subspace (CS); see Cook [6]. The column vectors $\beta_{01}, \beta_{02}, \dots, \beta_{0q}$ are called the CS directions. Dimension reduction is a fundamental statistical problem both in theory and in practice. See Li [22, 23] and Cook [6] for more discussion. If the conditional density function of Y given X exists, then the definition is equivalent to the conditional density function of $Y|X$ being the same as that of $Y|B_0^\top X$ for all possible values of X and Y, that is,

$$f_{Y|X}(y|x) = f_{Y|B_0^\top X}(y|B_0^\top x). \tag{1.1}$$

Other alternative definitions for the dimension reduction space can be found in the literature.

Received April 2005; revised January 2006.

[1]Supported by NUS FRG R-155-000-048-112.

AMS 2000 subject classifications. Primary 62G08; secondary 62G09, 62H05.

Key words and phrases. Conditional density function, convergence of algorithm, double-kernel smoothing, efficient dimension reduction, root-n consistency.

In the last decade or so, a series of papers (e.g., Härdle and Stoker [15], Li [22], Cook and Weisberg [8], Samarov [26], Hristache, Juditsky, Polzehl and Spokoiny [19], Yin and Cook [34], Xia, Tong, Li and Zhu [33], Cook and Li [7], Li, Zha and Chiaromonte [21] and Lue [24]) have considered issues related to the dimension reduction problem, with the aim of estimating the dimension reduction space and relevant functions. The estimation methods in the literature can be classified into two groups: inverse regression estimation methods (e.g., SIR, Li [22] and SAVE, Cook and Weisberg [8]) and direct regression estimation methods (e.g., ADE, Härdle and Stoker [15] and MAVE of Xia, Tong, Li and Zhu [33]). The inverse regression estimation methods are computationally easy and are widely used as an initial step in data mining, especially for large data sets. However, these methods have poor performance in finite samples and need strong assumptions on the design of covariates. The direct regression estimation methods have much better performance for finite samples than the inverse regression estimations. They need no strong requirements on the design of covariates or on the response variable. However, the direct regression estimation methods cannot find the directions in CS exhaustively, such as those in the conditional variance.

None of the methods mentioned above uses the definitions directly in searching for the central space. As a consequence, they fail in one way or another to estimate CS efficiently. A straightforward approach in using definition (1.1) is to look for B_0 in order to minimize the difference between those two conditional density functions. The conditional density functions can be estimated using nonparametric smoothers. Obviously, this approach is not efficient in theory due to the "curse of dimensionality" in nonparametric smoothing. In calculations, the minimization problem is difficult to implement. People have observed that the CS in the regression mean function, that is, the central mean space (CMS), can be estimated much more efficiently than the general CS. See, for example, Yin and Cook [34], Cook and Li [7] and Xia, Tong, Li and Zhu [33]. Motivated by this observation, one can construct a regression model such that the CS coincides with the CMS space in order to reduce the difficulty of estimation. In this paper we first construct a regression model in which the conditional density function $f_{Y|X}(y|x)$ is asymptotically equal to the conditional mean function. Then, we apply the methods of searching for the CMS to the constructed model. Based on the discussion above, this constructive approach is expected to be more efficient than the inverse regression estimation methods for finite samples, and can detect the CS directions exhaustively.

In the estimation of dimension reduction space, most methods need in one way or another to deal with nonparametric estimation. In terms of nonparametric estimation, the inverse regression estimation methods employ a nonparametric regression of X on Y while the direct regression estimation methods employ a nonparametric regression of Y on X. In contrast to existing methods, the methods in this paper search for CS from both sides by investigating conditional density functions. A similar idea appeared in Yin and Cook [35] for a general single-index model. To overcome the difficulties of calculation, we propose two algorithms in this paper using an idea similar to that of Xia, Tong, Li and Zhu [33]. The algorithm solves the minimization problem in the method by treating it as two separate quadratic programming problems, which have simple analytic solutions and can be calculated quite efficiently. The convergence of the algorithm can be proved. Our constructive approach can overcome the disadvantage in inverse regression estimation, requiring a symmetric design for explanatory variables, and also the disadvantage in direct regression estimation, of not finding the CS directions exhaustively. Simulations suggest that the proposed methods have very good performance for finite samples and are able to estimate the CS directions in very complicated structures. Applying the proposed methods to two real data sets, some useful patterns have been observed, based on the estimations.

To estimate the CS, we need to estimate the directions B_0 as well as the dimension q of the space. In this paper, however, we focus on the estimation of the directions by assuming that q is known.

2. Estimation methods. As discussed above, the direct regression estimation has good performance for finite samples. However, it cannot detect exhaustively the CS directions in complicated structures. Motivated by these facts, our strategy is to construct a semiparametric regression model such that all the CS directions are captured in the regression mean function. As we can see from (1.1), all the directions can be captured in the conditional density function. Thus, we will construct a regression model such that the conditional density function is asymptotically equal to the regression mean function.

The primary step is thus to construct an estimate for the conditional density function. Here, we use the idea of the "double-kernel" local linear smoothing method studied in Fan, Yao and Tong [13] for the estimation. Consider $H_b(Y-y)$ with y running through all possible values, where $H(v)$ is a symmetric density function, $b>0$ is a bandwidth and $H_b(v)=b^{-1}H(v/b)$. If $b\to 0$ as $n\to\infty$, then from (1.1) we have

$$\begin{aligned} m_b(x,y) &\stackrel{\text{def}}{=} E\big(H_b(Y-y)|X=x\big)\\ &= E\big(H_b(Y-y)|B_0^\top X=B_0^\top x\big)\\ &\to f_{Y|B_0^\top X}(y|B_0^\top x). \end{aligned}$$

See Fan, Yao and Tong [13]. The above equation indicates that all the directions can be captured by the conditional mean function $m_b(x,y)$ of $H_b(Y-y)$ on $X=x$ with x and y running through all possible values. Now, consider a regression model nominally for $H_b(Y-y)$ as

$$H_b(Y-y)=m_b(X,y)+\varepsilon_b(y|X),$$

where $\varepsilon_b(y|X) = H_b(Y - y) - E(H_b(Y - y)|X)$ with $E\varepsilon_b(y|X) = 0$. Let $g_b(B_0^\top x, y) = E(H_b(Y - y)|B_0^\top X = B_0^\top x)$. If (1.1) holds, then $m_b(x, y) = g_b(B_0^\top x, y)$. The model can be written as

$$H_b(Y - y) = g_b(B_0^\top X, y) + \varepsilon_b(y|X). \tag{2.1}$$

As $b \to 0$, we have $g_b(B_0^\top x, y) \to f_{Y|B_0^\top X}(y|B_0^\top x)$. Thus, the directions B_0 defined in (1.1) are all captured in the regression mean function in model (2.1) if y runs through all possible values.

Based on model (2.1), we propose two methods to estimate the directions. One of the methods is a combination of the outer product of gradients (OPG) estimation method (Härdle [16], Samarov [26] and Xia, Tong, Li and Zhu [33]) with the "double-kernel" local linear smoothing method (Fan, Yao and Tong [13]). The other one is a combination of the minimum average (conditional) variance estimation (MAVE) method (Xia, Tong, Li and Zhu [33]) with the "double-kernel" local linear smoothing method. The structure adaptive weights in Hristache, Juditsky and Spokoiny [20] and Hristache, Juditsky, Polzehl and Spokoiny [19] are used in the estimation.

2.1. *Estimation based on outer products of gradients.* Consider the gradient of the conditional mean function $m_b(x, y)$ with respect to x. If (1.1) holds, then it follows that

$$\frac{\partial m_b(x, y)}{\partial x} = \frac{\partial g_b(B_0^\top x, y)}{\partial x} = B_0 \bigtriangledown g_b(B_0^\top x, y), \tag{2.2}$$

where $\bigtriangledown g_b(v_1, \dots, v_q, y) = (\bigtriangledown_1 g_b(v_1, \dots, v_q, y), \dots, \bigtriangledown_q g_b(v_1, \dots, v_q, y))^\top$ with

$$\bigtriangledown_k g_b(v_1, \dots, v_q, y) = \frac{\partial}{\partial v_k} g_b(v_1, \dots, v_q, y), \qquad k = 1, 2, \dots, q.$$

Thus, the directions B_0 are contained in the gradients of the regression mean function in model (2.1). One way to estimate B_0 is by considering the expectation of the outer product of the gradients

$$\begin{aligned} &E\left\{\left(\frac{\partial m_b(X, Y)}{\partial x}\right)\left(\frac{\partial m_b(X, Y)}{\partial x}\right)^\top\right\} \\ &\qquad = B_0 E\{\bigtriangledown g_b(B_0^\top X, Y) \bigtriangledown^\top g_b(B_0^\top X, Y)\} B_0^\top. \end{aligned}$$

It is easy to see that B_0 is in the space spanned by the first q eigenvectors of the expectation of the outer products.

Suppose that $\{(X_i, Y_i), i = 1, 2, \dots, n\}$ is a random sample from (X, Y). To estimate the gradient $\partial m_b(x, y)/\partial x$, we can use nonparametric kernel smoothing methods. For simplicity, we adopt the following notation scheme. Let $K_0(v^2)$ be a univariate symmetric density function and define $K(v_1, \dots, v_d) = K_0(v_1^2 + \cdots + v_d^2)$ for any integer d and $K_h(u) = h^{-d}K(u/h)$, where d is the dimension of u and $h > 0$ is a bandwidth. Let $H_{b,i}(y) = H_b(Y_i - y)$, where $H(\cdot)$ and b are defined above. For any (x, y), the principle of the local linear smoother suggests minimizing

$$n^{-1}\sum_{i=1}^{n}\{H_{b,i}(y) - a - b^\top(X_i - x)\}^2 K_h(X_{ix}) \tag{2.3}$$

with respect to a and b to estimate $m_b(x, y)$ and $\partial m_b(x, y)/\partial x$, respectively, where $X_{ix} = X_i - x$. See Fan and Gijbels [11] for more details. For each pair of (X_j, Y_k), we consider the minimization problem

$$(\hat{a}_{jk}, \hat{b}_{jk}) = \arg\min_{a_{jk}, b_{jk}} \sum_{i=1}^{n}[H_{b,i}(Y_k) - a_{jk} - b_{jk}^\top X_{ij}]^2 w_{ij}, \tag{2.4}$$

where $X_{ij} = X_i - X_j$ and $w_{ij} = K_h(X_{ij})$. We consider an average of their outer products,

$$\hat{\Sigma} = n^{-2}\sum_{k=1}^{n}\sum_{j=1}^{n}\hat{\rho}_{jk}\hat{b}_{jk}\hat{b}_{jk}^\top,$$

where $\hat{\rho}_{jk}$ is a trimming function introduced for technical purposes to handle the notorious boundary points. In this paper we adopt the following trimming scheme. For any given point (x, y), we use all observations to estimate its function value and its gradient as in (2.3). We then consider the estimates in a compact region of (x, y). Moreover, for those points with too few observations around, their estimates might be unreliable. They should not be used in the estimation of the CS directions and should be trimmed off. Let $\rho(\cdot)$ be any bounded function with bounded second order derivatives on $\mathbb{R}$ such that $\rho(v) > 0$ if $v > \omega_0$; $\rho(v) = 0$ if $v \le \omega_0$ for some small $\omega_0 > 0$. We take $\hat{\rho}_{jk} = \rho(\hat{f}(X_j))\rho(\hat{f}_Y(Y_k))$, where $\hat{f}(x)$ and $\hat{f}_Y(y)$ are estimators of the density functions of X and Y, respectively. The CS directions can be estimated by the first q eigenvectors of $\hat{\Sigma}$.

To allow the estimation to be adaptive to the structure of the dependence of Y on X, we may follow the idea of Hristache, Juditsky, Polzehl and Spokoiny [19] and replace w_{ij} in (2.4) by

$$w_{ij} = K_h(\hat{\Sigma}^{1/2}X_{ij}),$$

where $\hat{\Sigma}^{1/2}$ is a symmetric matrix with $(\hat{\Sigma}^{1/2})^2 = \hat{\Sigma}$. Repeat the above procedure until convergence. We call this procedure the method of outer product of gradient based on the conditional density functions (dOPG). To implement the estimation procedure, we suggest the following dOPG algorithm.

Step 0. Set $\hat{\Sigma}_{(0)} = I_p$ and $t = 0$.

Step 1. With $w_{ij} = K_h(\hat{\Sigma}_{(t)}^{1/2} X_{ij})$, calculate the solution to (2.4),

$$\begin{pmatrix} a_{jk}^{(t)} \\ b_{jk}^{(t)} \end{pmatrix} = \left\{ \sum_{i=1}^{n} K_{h_t}(\hat{\Sigma}_{(t)}^{1/2} X_{ij}) \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} \right\}^{-1} \times \sum_{i=1}^{n} K_{h_t}(\hat{\Sigma}_{(t)}^{1/2} X_{ij}) \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} H_{b_t,i}(Y_k),$$

where h_t and b_t are bandwidths [details are given in (2.6) and (2.7) below].

Step 2. Define $\rho_{jk}^{(t)} = \rho(\tilde{f}^{(t)}(X_j))\rho(\tilde{f}_Y^{(t)}(Y_k))$ with

$$\tilde{f}_Y^{(t)}(y) = n^{-1} \sum_{i=1}^{n} H_{b_t,i}(y),$$

$$\tilde{f}^{(t)}(x) = (n\tilde{\mu})^{-1} h_t^p \prod_{\lambda_k^{(t)} > h_t} \frac{\lambda_k^{(t)}}{h_t} \sum_{i=1}^{n} K_{h_t}(\hat{\Sigma}_{(t)}^{1/2} X_{ix}),$$

where $\lambda_k^{(t)}$, $k = 1, \ldots, p$, are the eigenvalues of $\hat{\Sigma}_{(t)}^{1/2}$ and $\tilde{\mu} = \int K_0(\sum_{\lambda_k^{(t)} > h_t} v_k^2) \prod_{\lambda_k^{(t)} > h_t} dv_k$. Calculate the average of outer products,

$$\hat{\Sigma}_{(t+1)} = n^{-2} \sum_{j,k=1}^{n} \rho_{jk}^{(t)} b_{jk}^{(t)} (b_{jk}^{(t)})^{\top}.$$

Step 3. Set $t := t + 1$. Repeat Steps 1 and 2 until convergence. Denote the final value of $\hat{\Sigma}_{(t)}$ by $\Sigma_{(\infty)}$. Suppose the eigenvalue decomposition of $\Sigma_{(\infty)}$ is $\Gamma diag(\lambda_1, \ldots, \lambda_p)\Gamma^{\top}$, where $\lambda_1 \geq \cdots \geq \lambda_p$. Then the estimated directions are the first q columns of Γ, denoted by $\hat{B}_{\mathrm{dOPG}}$.

In the dOPG algorithm, $\tilde{f}_Y^{(t)}(y)$ and $\tilde{f}^{(t)}(x)$, $t > 0$, are the estimators of the density functions of Y and $B_0^{\top} X$, respectively. A justification is given in the proof of Theorem 3.1 in Section 6.2. In calculations, the usual stopping criterion can be used. For example, if the largest singular value of $\hat{\Sigma}_{(t)} - \hat{\Sigma}_{(t+1)}$ is smaller than 10^{-6}, then we stop the iteration and take $\hat{\Sigma}_{(t+1)}$ as the final estimator. The eigenvalues of $\Sigma_{(\infty)}$ can be used to determine the dimension of the CS. However, we will not go into the details on this issue in this paper. In practice, we may need to standardize $X_i = (X_{i1}, \ldots, X_{ip})^{\top}$ by setting $X_i := S_X^{-1/2}(X_i - \bar{X})$ and standardize Y_i by setting $Y_i := (Y_i - \bar{Y})/\sqrt{s_Y}$, where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $S_X = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\top}$, $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and $s_Y = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Then the estimated CS directions are the first q columns of $\Gamma S_X^{-1/2}$.

2.2. *MAVE based on conditional density function.* Note that if (1.1) holds, then the gradients $\partial m_b(x,y)/\partial x$ at all (x,y) are in a common q-dimensional subspace as shown in equation (2.2). To use this observation, we can replace b in (2.3), which is an estimate of the gradient, by $Bd(x,y)$ and have the local linear approximation

$$n^{-1}\sum_{i=1}^{n}\{H_{b,i}(y)-a-d^{\top}B^{\top}(X_i-x)\}^2K_h(X_{ix}),$$

where $d=d(x,y)$ is introduced to take the role of $\bigtriangledown g_b(B_0^{\top}x,y)$ in (2.2). Note that the above weighted mean of squares uses the local approximation errors of $H_{b,i}(y)$ by a hyperplane with the normal vectors in a common space spanned by B. Since B is common for all x and y, it should be estimated with aims to minimize the approximation errors for all possible X_j and Y_k. As a consequence, we propose to estimate B_0 by minimizing

$$n^{-3}\sum_{k=1}^{n}\sum_{j=1}^{n}\hat{\rho}_{jk}\sum_{i=1}^{n}\{H_{b,i}(Y_k)-a_{jk}-d_{jk}^{\top}B^{\top}X_{ij}\}^2w_{ij} \tag{2.5}$$

with respect to $a_{jk}, d_{jk}=(d_{jk1},\dots,d_{jkq})^{\top}$, $j,k=1,\dots,n$ and $B:B^{\top}B=I_q$, where $\hat{\rho}_{jk}$ is defined above. This estimation procedure is similar to the minimum average (conditional) variance estimation method (Xia, Tong, Li and Zhu [33]). Because the method is based on the conditional density functions, we call it the minimum average (conditional) variance estimation based on the conditional density functions (dMAVE).

The minimization problem in (2.5) can be solved by fixing (a_{jk},d_{jk}), $j,k=1,\dots,n$, and B alternately. As a consequence, we need to solve two quadratic programming problems which have simple analytic solutions. For any matrix $B=(\beta_1,\dots,\beta_d)$, we define operators $\ell(\cdot)$ and $\mathcal{M}(\cdot)$, respectively, as

$$\ell(B)=(\beta_1^{\top},\dots,\beta_d^{\top})^{\top} \quad\text{and}\quad \mathcal{M}(\ell(B))=B.$$

We propose the following dMAVE algorithm to implement the estimation.

Step 0. Let $B_{(1)}$ be an initial estimator of the CS directions. Set $t=1$.

Step 1. Let $B=B_{(t)}$, calculate the solutions of (a_{jk},d_{jk}), $j,k=1,\dots,n$, to the minimization problem in (2.5):

$$\begin{pmatrix}a_{jk}^{(t)}\\ d_{jk}^{(t)}\end{pmatrix}=\left\{\sum_{i=1}^{n}K_{h_t}(B_{(t)}^{\top}X_{ij})\begin{pmatrix}1\\ B_{(t)}^{\top}X_{ij}\end{pmatrix}\begin{pmatrix}1\\ B_{(t)}^{\top}X_{ij}\end{pmatrix}^{\top}\right\}^{-1}$$
$$\times\sum_{i=1}^{n}K_{h_t}(B_{(t)}^{\top}X_{ij})\begin{pmatrix}1\\ B_{(t)}^{\top}X_{ij}\end{pmatrix}H_{b_t,i}(Y_k),$$

where h_t and b_t are two bandwidths (details are discussed below).

Step 2. Let $\rho_{jk}^{(t)} = \rho(\hat{f}_{B_{(t)}}(X_j))\rho(\hat{f}_Y^{(t)}(Y_k))$ with $\hat{f}_Y^{(t)}(y) = n^{-1}\sum_{i=1}^n H_{b_t,i}(y)$ and $\hat{f}_{B_{(t)}}(x) = n^{-1}\sum_{i=1}^n K_{h_t}(B_{(t)}^\top X_{ix})$. Fixing $a_{jk} = a_{jk}^{(t)}$ and $d_{jk} = d_{jk}^{(t)}$, calculate the solution of B or $\ell(B)$ to (2.5):

$$\mathbf{b}^{(t+1)} = \left\{\sum_{k,j,i=1}^n \rho_{jk}^{(t)} K_{h_t}(B_{(t)}^\top X_{ij}) X_{ijk}^{(t)} (X_{ijk}^{(t)})^\top\right\}^{-1} \times \sum_{k,j,i=1}^n \rho_{jk}^{(t)} K_{h_t}(B_{(t)}^\top X_{ij}) X_{ijk}^{(t)} \{H_{b_t,i}(Y_k) - a_{jk}^{(t)}\},$$

where $X_{ijk}^{(t)} = d_{jk}^{(t)} \otimes X_{ij}$.

Step 3. Calculate $\Lambda_{(t+1)} = \{\mathcal{M}(\mathbf{b}^{(t+1)})\}^\top \mathcal{M}(\mathbf{b}^{(t+1)})$ and $B_{(t+1)} = \mathcal{M}(\mathbf{b}^{(t+1)}) \times \Lambda_{(t+1)}^{-1/2}$. Set $t := t+1$ and go to Step 1.

Step 4. Repeat Steps 1–3 until convergence. Let $B_{(\infty)}$ be the final value of $B_{(t)}$. Then our estimators of the directions are the columns in $B_{(\infty)}$, denoted by $\hat{B}_{\mathrm{dMAVE}}$.

The dMAVE algorithm needs a consistent initial estimator in Step 0 to guarantee its theoretical justification. In the following, we use the first iteration estimator of dOPG, the first q eigenvectors of $\hat{\Sigma}_{(1)}$, as the initial value. Actually, any initial estimator that satisfies (6.6) can be used and Theorem 3.2 will hold. Similar to dOPG, the standardization procedure can be carried out for dMAVE in practice. The stopping criterion for dOPG can also be used here.

Note that the estimation in the procedure is related to nonparametric estimation of conditional density functions. Several bandwidth selection methods are available for the estimation. See, for example, Silverman [28], Scott [27] and Fan, Yao and Tong [13]. Our theoretical verification of the convergence for the algorithms requires some constraints on the bandwidths, although we believe these constraints can be removed with more complicated technical proofs. To ensure the requirements on bandwidths can be satisfied, after standardizing the variables we use the following bandwidths in our calculations. In the first iteration, we use slightly larger bandwidths than the optimal ones in terms of MISE as

$$h_0 = c_0 n^{-1/(p_0+6)}, \qquad b_0 = c_0 n^{-1/(p_0+5)}, \tag{2.6}$$

where $p_0 = \max(p, 3)$. Then we reduce the bandwidths in each iteration as

$$\begin{aligned} h_{t+1} &= \max\{r_n h_t, c_0 n^{-1/(q+4)}\}, \\ b_{t+1} &= \max\{r_n b_t, c_0 n^{-1/(q+3)}, c_0 n^{-1/5}\} \end{aligned} \tag{2.7}$$

for $t \ge 0$, where $r_n = n^{-1/(2(p_0+6))}$, $c_0 = 2.34$ as suggested by Silverman [28] if the Epanechnikov kernel is used. Here, the bandwidth b is selected smaller than h based on simulation comparisons.

Fan and Yao ([12], page 337) proposed a method, called profile least-squares estimation, for the single-index model and its variants by solving a similar minimization problem as in (2.5). The method can also be used here for the estimation of B_0 in (2.1).

3. Asymptotic results. To exclude the trivial cases, we assume that $p > 1$ and $q \ge 1$. Let $f_0(y|v_1, \dots, v_q)$, $f_0(v_1, \dots, v_q)$ and $f_Y(y)$ be the (conditional) density functions of $Y|B_0^\top X$, $B_0^\top X$ and Y, respectively. Let $\rho_0(x, y) = \rho(f_0(B_0^\top x)) \times \rho(f_Y(y))$, $\triangledown f_0(y|v_1, \dots, v_q) = (\partial f_0(y|v_1, \dots, v_q)/\partial v_1, \dots, \partial f_0(y|v_1, \dots, v_q)/\partial v_q)^\top$, $\mu_B(u) = E(X|B^\top X = u)$ and $w_B(u) = E\{XX^\top|B^\top X = u\}$. For any matrix A, let $|A|$ denote its largest singular value, which is the same as the Euclidean norm if A is a vector. Let $\tilde{B}_0 : p \times (p - q)$ be such that $(B_0, \tilde{B}_0)^\top(B_0, \tilde{B}_0) = I_p$. We need the following conditions for (1.1) to prove our theoretical results.

(C1) [*Design of* X.] The density function $f(x)$ of X has bounded second order derivatives on $\mathbb{R}^p$; $E|X|^r < \infty$ for some $r > 8$; the functions $\mu_B(u)$ and $w_B(u)$ have bounded derivatives with respect to u and B for B in a small neighborhood of $B_0 : |B - B_0| \le \delta$ for some $\delta > 0$.

(C2) [*Conditional density function*.] The conditional density functions $f_{Y|X}(y|x)$ and $f_{Y|B^\top X}(y|u)$ have bounded fourth order derivatives with respect to x, u and B for B in a small neighborhood of B_0; the conditional density functions of $f_{\tilde{B}_0^\top X, Y|B_0^\top X}(u, y|v)$ and $\int |\triangledown f_0(y|u)|\, dy$ are bounded for all u, y and v.

(C3) [*Efficient dimension*.] The matrix $M_0 = \int \rho_0(x, y) \triangledown f_0(y|B_0^\top x) \triangledown^\top f_0(y|B_0^\top x) \times f(x) f_Y(y)\, dx\, dy$ has full rank q.

(C4) [*Kernel functions*.] $K_0(v^2)$ and $H(v)$ are two symmetric univariate density functions with bounded second order derivatives and compact supports.

(C5) [*Bandwidths for consistency*.] Bandwidths $h_0 = c_1 n^{-r_h}$ and $b_0 = c_2 n^{-r_b}$, where $0 < r_h, r_b \le 1/(p_0 + 6)$, $p_0 = \max\{p, 3\}$. For $t \ge 1$, $h_t = \max\{r_n h_{t-1}, \hbar\}$ and $b_t = \max\{r_n b_{t-1}, \bar{b}\}$, where $r_n = n^{-r_h/2}$, $\hbar = c_3 n^{-r'_h}$, $\bar{b} = c_4 n^{-r'_b}$ with $0 < r'_h, r'_b \le 1/(q + 3)$, and c_1, c_2, c_3, c_4 are constants.

In (C1), the finite moment requirement for $|X|$ can be removed if we adopt the trimming scheme of Härdle, Hall and Ichimura [14]. However, as noticed in Delecroix, Hristache and Patilea [10], this scheme causes some technical problems in the proofs. Based on assumptions (C2) and (C4), the smoothness of $g_b(u, y)$ is implied. A lower order of smoothness is sufficient if we are only interested in the estimation consistency. The second order differentiable requirement in (C4) can ensure the Fourier transformations of the kernel functions being absolutely integrable; see Chung [5], page 166. Popular kernel functions such as the Epanechnikov kernel and the quadratic kernel are included in (C4). The Gaussian kernel can be used with some modifications to the proofs. Condition (C3) indicates that the dimension q cannot be further reduced. For ease of exposition, we further assume that $\mu_{0H} = \int H(v)\, dv = 1$,

$\mu_{2H} = \int v^2 H(v)\,dv = 1$, $\mu_{0q} = \int K(v_1, \dots, v_q)\,dv_1 \cdots dv_q = 1$ and $\mu_{2q} = \int K(v_1, \dots, v_q) v_1^2\,dv_1 \cdots dv_q = 1$; otherwise, we take $H(v) := H(v/\tau_{2H}^{1/2})/\tau_{2H}^{1/2}$ and $K(v_1, \dots, v_q) = \mu_{0q}^{-1} K(v_1/\sqrt{\mu_{2q}}, \dots, v_q/\sqrt{\mu_{2q}})/\sqrt{\mu_{2q}}$. The bandwidths satisfying (C5) can be found easily. For example, the bandwidths given in (2.6) and (2.7) satisfy the requirements. Actually, a wider range of bandwidths can be used; see the proofs. Let $\nu_B(x) = \mu_B(B^\top x) - x$, $\bar{w}_B(x) = w_B(B^\top x) - \mu_B(B^\top x)\mu_B^\top(B^\top x)$ and $f_0(x) = f_0(B_0^\top x)$. For any square matrix A, A^{-1} and A^+ denote the inverse (if it exists) and the Moore–Penrose inverse matrices, respectively.

THEOREM 3.1. *Suppose conditions* (C1)–(C5) *hold. Then we have*

$$|\hat{B}_{\mathrm{dOPG}}\hat{B}_{\mathrm{dOPG}}^\top - B_0 B_0^\top| = O(\hbar^4 + \delta_{q\hbar\bar{b}}^2 + \delta_{q\hbar\bar{b}}\bar{b}^4 + \delta_n^2/\bar{b}^2 + n^{-1/2})$$

in probability as $n \to \infty$, where $\delta_{q\hbar\bar{b}} = (n\hbar^q\bar{b}/\log n)^{-1/2}$ and $\delta_n = (\log n/n)^{1/2}$. If $\hbar^4 + \delta_{q\hbar\bar{b}}^2 + \delta_{q\hbar\bar{b}}\bar{b}^4 + \delta_n^2/\bar{b}^2 = o(n^{-1/2})$ can be satisfied, then

$$\sqrt{n}\{\ell(\hat{B}_{\mathrm{dOPG}}\hat{B}_{\mathrm{dOPG}}^\top B_0) - \ell(B_0)\} \xrightarrow{D} N(0, W_0),$$

where

$$W_0 = \mathrm{Var}\big[\rho_0(X,Y)M_0^{-1}\big(\triangledown f_0(Y|B_0^\top X) f_Y(Y) - E\{\triangledown f_0(Y|B_0^\top X) f_Y(Y)|X\}\big) \otimes (\bar{w}_{B_0}^+(X)\nu_{B_0}(X))\big].$$

The first part of Theorem 3.1 indicates that $\hat{B}_{\mathrm{dOPG}}$ is a consistent estimator of an orthogonal basis, $B_0 Q$ with $Q = B_0^\top \hat{B}_{\mathrm{dOPG}}$, in CS and $|\hat{B}_{\mathrm{dOPG}} - B_0 Q| = O(\hbar^4 + \delta_{q\hbar\bar{b}}^2 + \delta_{q\hbar\bar{b}}\bar{b}^4 + \delta_n^2/\bar{b}^2 + n^{-1/2})$ in probability. See Bai, Miao and Rao [2] and Xia, Tong, Li and Zhu [33] for alternative presentations of the asymptotic results. If the bandwidths in (2.7) are used, then the consistency rate is $O(n^{-4/(q+4)+1/(q+3)}\log n + n^{-1/2})$ in probability. A faster consistency rate can be obtained by adjusting the bandwidths. The convergence of the corresponding algorithm is also implied in the proof in Section 6. If $q \le 3$, then the condition for the normality can be satisfied by taking

$$1 > r_h' > \tfrac{1}{8}, \qquad \tfrac{2}{7}r_h' < r_b' < \tfrac{1}{2} - q r_h'.$$

If we use higher order polynomial smoothing, it is possible to show that root-n consistency can be achieved for any dimension q. See, for example, Härdle and Stoker [15] and Samarov [26], where the higher order kernel, a counterpart of the higher order polynomial smoother, was used. However, using higher order polynomial smoothers increases the difficulty of calculations while the improvement of finite sample performance is not substantial.

THEOREM 3.2. *If conditions* (C1)–(C5) *holds*, *then*

$$|\hat{B}_{\mathrm{dMAVE}}\hat{B}_{\mathrm{dMAVE}}^{\top} - B_0B_0^{\top}| = O\{\hbar^4 + \delta_{q\hbar\hbar}^2 + \delta_{q\hbar\hbar}\hbar^4 + \delta_n^2/\hbar^2 + n^{-1/2}\}$$

in probability as $n \to \infty$. *If* $\hbar^4 + \delta_{q\hbar\hbar}^2 + \delta_{q\hbar\hbar}\hbar^4 + \delta_n^2/\hbar^2 = o(n^{-1/2})$ *can be satisfied*, *then*

$$\sqrt{n}\{\ell(\hat{B}_{\mathrm{dMAVE}}\hat{B}_{\mathrm{dMAVE}}^{\top}B_0) - \ell(B_0)\} \xrightarrow{D} N(0, D_0^{+}\Sigma_0 D_0^{+}),$$

where $D_0 = \int \rho_0(x, y) \bigtriangledown f_0(y|B_0^{\top}x) \bigtriangledown^{\top} f_0(y|B_0^{\top}x) \otimes \{\nu_{B_0}(x)\nu_{B_0}^{\top}(x)\} f_0(x) \times f_Y(y)\,dx\,dy$ *and*

$$\Sigma_0 = \mathrm{Var}\big[\rho_0(X, Y)\big(\bigtriangledown f_0(Y|B_0^{\top}X) f_Y(Y) - E\{\bigtriangledown f_0(Y|B_0^{\top}X) f_Y(Y)|X\}\big) \otimes \nu_{B_0}(X)\big].$$

The proof of Theorem 3.2 is given in Section 6. The convergence of the dMAVE algorithm is implied in the proof. Similar remarks on dOPG are applicable to dMAVE. Moreover, $\hat{B}_{\mathrm{dMAVE}}$ converges to $B_0\tilde{Q}$, where $\tilde{Q}$ is determined by the initial consistent estimator of the directions. For example, $\tilde{Q} = \hat{B}_{(1)}^{\top}B_0$ if $B_{(1)}$ is used as the initial estimator. Similarly, root-n consistency holds for $q \le 3$. It is possible that root-n consistency holds for $q > 3$ if the higher order local polynomial smoothing method is used. In spite of the equivalence in terms of consistency rate for both dOPG and dMAVE, our simulations suggest that dMAVE has better performance than dOPG in finite samples. Theoretical comparison of efficiencies between the two methods is not clear. In a very special case when $q = 1$ and the CS is in the regression mean, Xia [30] proved that dMAVE is more efficient than dOPG.

We here give some discussion about the requirements on the distributions of X and Y. If Y is discrete, we can consider the conditional cumulative distribution functions and have $F_{Y|X}(y|x) = F_{Y|B_0^{\top}X}(y|B_0^{\top}x)$ when $Y \perp\!\!\!\perp X|B_0^{\top}X$ holds. Similar to (2.1), we can consider a regression model

$$I(Y < y) = G(B_0^{\top}X, y) + e(y|X),$$

where $G(B_0^{\top}x, y) = E\{I(Y < y)|X = x\} = E\{I(Y < y)|B_0^{\top}X = B_0^{\top}x\}$ and $e(y|X) = I(Y < y) - G(B_0^{\top}X, y)$. Similar theoretical consistency results are possible to be obtained following the same techniques developed here. If some covariates in X are discrete, our algorithms in searching for a consistent initial estimator will fail. However, if a consistent initial estimator can be found by, for example, the methods in Horowitz and Härdle [18] and Hristache, Juditsky, Polzehl and Spokoiny [19] and $B^{\top}X$ has a continuous density function for all B in a neighborhood around B_0, then our theoretical results in the above theorems still hold.

4. Simulations. We now demonstrate the performance of the proposed estimation methods by simulations. We will compare them with some existing methods including SIR (Li [22]), SAVE (Cook and Weisberg [8]), PHD (Li [23]) and rMAVE (Xia, Tong, Li and Zhu [33]). The computer codes used here can be obtained from www.jstatsoft.org/v07/i01 for the SIR, SAVE and PHD methods (courtesy of Professor S. Weisberg) and www.stat.nus.edu.sg/~staxyc for rMAVE, dOPG and dMAVE. In the following calculations, we use the quadratic kernel $H(v) = K_0(v^2) = (15/16)(1-v^2)^2 I(v^2 < 1)$ and $\omega_0 = 0.01$. The bandwidths in (2.6) and (2.7) are used. For the inverse regression methods, the number of slices is chosen to be between 5 and 30 and most close to $n/(2p)$. We define an overall estimation error for the estimator $\hat{B} : \hat{B}^\top \hat{B} = I_q$ by the maximum singular value of $B_0 B_0^\top - \hat{B}\hat{B}^\top$; see Li, Zha and Chiaromonte [21].

EXAMPLE 4.1. Consider the model

$$Y = \operatorname{sign}(2X^\top \beta_1 + \varepsilon_1)\log(|2X^\top \beta_2 + 4 + \varepsilon_2|), \tag{4.1}$$

where sign$(\cdot)$ is the sign function. The coordinates $X \sim N(0, I_p)$ and unobservable noises $\varepsilon_1 \sim N(0,1)$ and $\varepsilon_2 \sim N(0,1)$ are independent. For β_1, the first four elements are all 0.5 and the others are zero. For β_2, the first four elements are 0.5, -0.5, 0.5, -0.5, respectively, and all the others are zero. A similar model was investigated by Chen and Li [3]. In order to show the effect on the estimation performance of the number of covariates, we vary p in the simulation. With different sample sizes, 200 replications are drawn from the model. The calculation results are listed in Table 1. To get some intuition about the size of estimation errors, Figure 1 shows a typical sample of size $n = 200$ and its estimate with estimation error 0.21. The structure can be estimated quite well in the sample.

TABLE 1
Mean (and standard deviation) of estimation errors for Example 4.1

n	**p**	**dOPG**	**dMAVE**	**rMAVE**	**SIR**	**SAVE**	**PHD**
	5	0.25 (0.09)	0.22 (0.08)	0.43 (0.19)	0.29 (0.09)	0.87 (0.19)	0.72 (0.22)
100	10	0.55 (0.19)	0.35 (0.07)	0.64 (0.19)	0.46 (0.10)	0.94 (0.06)	0.90 (0.13)
	20	0.81 (0.13)	0.54 (0.10)	0.88 (0.12)	0.64 (0.11)	0.96 (0.06)	0.93 (0.07)
	5	0.17 (0.05)	0.14 (0.04)	0.27 (0.13)	0.19 (0.05)	0.55 (0.26)	0.47 (0.15)
200	10	0.32 (0.09)	0.24 (0.06)	0.46 (0.17)	0.30 (0.06)	0.96 (0.08)	0.73 (0.16)
	20	0.62 (0.15)	0.36 (0.06)	0.66 (0.16)	0.43 (0.06)	0.93 (0.04)	0.94 (0.08)
	5	0.13 (0.04)	0.13 (0.04)	0.19 (0.07)	0.16 (0.05)	0.32 (0.18)	0.37 (0.12)
300	10	0.24 (0.06)	0.18 (0.04)	0.36 (0.16)	0.24 (0.05)	0.85 (0.17)	0.59 (0.15)
	20	0.48 (0.13)	0.28 (0.05)	0.55 (0.16)	0.35 (0.05)	0.92 (0.03)	0.84 (0.12)
	5	0.11 (0.04)	0.11 (0.04)	0.21 (0.12)	0.14 (0.04)	0.22 (0.11)	0.31 (0.10)
400	10	0.21 (0.04)	0.16 (0.04)	0.31 (0.11)	0.21 (0.05)	0.66 (0.22)	0.51 (0.13)
	20	0.31 (0.06)	0.25 (0.04)	0.49 (0.15)	0.29 (0.04)	0.98 (0.04)	0.76 (0.14)

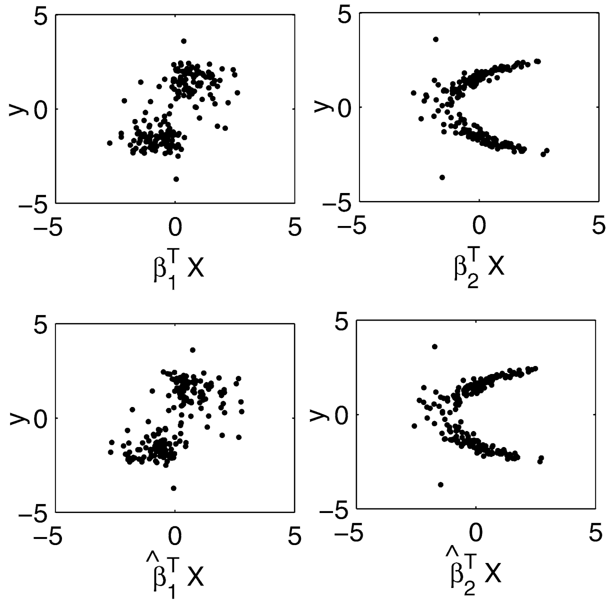

FIG. 1. *A typical data set of size* 200 *from Example* 4.1 *with* $p = 10$ *to show the size of estimation error and its graphic performance. The upper two panels are plots of y against the true CS directions, the lower two panels y against the estimated directions using dMAVE. The estimated directions are respectively* $\hat{\beta}_1 = (0.42, 0.64, 0.44, 0.45, -0.01, -0.07, 0.02, -0.00, -0.08, 0.07)^\top$ *and* $\hat{\beta}_2 = (-0.54, 0.43, -0.57, 0.43, 0.01, -0.04, -0.01, 0.07, -0.05, 0.07)^\top$ *with estimation error* 0.21.

In model (4.1) the CS directions are hidden in a complicated structure and are not easy to detect directly by the conditional regression mean function. When the sample size is large (≥ 200) and p is not large ($= 5$), all the methods give accurate estimates. As p increases, rMAVE performs not so well because the second direction is not captured in the regression mean function; SAVE and PHD also fail to give accurate estimates. SIR performs much better in all the situations than SAVE and PHD. dOPG has about the same performance as SIR. dMAVE is the best in all situations among all the methods.

EXAMPLE 4.2. Now, consider the CS in conditional mean as well as the conditional variance as in the model

$$Y = 2(X^\top \beta_1)^d + 2\exp(X^\top \beta_2)\varepsilon, \tag{4.2}$$

where $X = (x_1, \dots, x_{10})^\top$ with $x_1, \dots, x_{10} \sim \text{Uniform}(-\sqrt{3}, \sqrt{3})$ and $\varepsilon \sim N(0, 1)$ are independent, $\beta_1 = (1, 2, 0, 0, 0, 0, 0, 0, 0, 2)^\top/3$ and $\beta_2 = (0, 0, 3, 4, 0, 0, 0, 0, 0, 0)^\top/5$. For model (4.2), one CS direction is contained in the regression mean and the other in the conditional variance. One typical data set with size 200 is shown in Figure 2. Table 2 lists the calculation results for 200 replications.

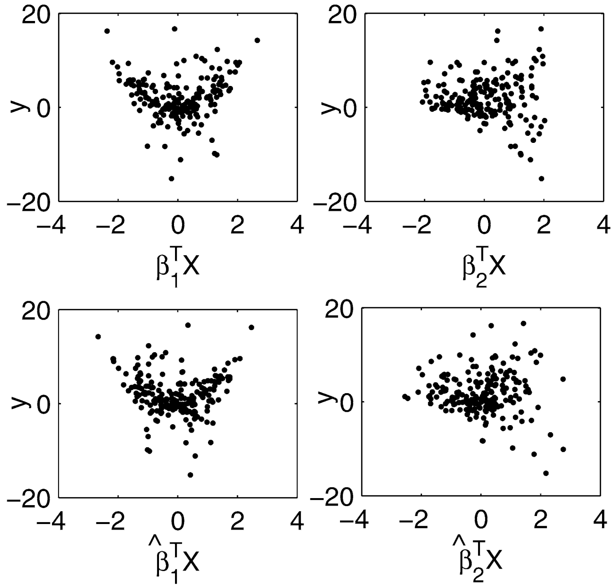

FIG. 2. *A typical data set with* $n = 200$ *from Example* 4.2 *and its dMAVE estimation. The upper two panels are plots of* y *against the true CS directions, the lower two panels* y *against the estimated directions with estimation error* 0.31.

Because rMAVE cannot detect the CS directions hidden in the conditional variance directly, it has very poor overall estimation performance as listed in Table 2. If $d = 1$, that is, the regression mean function is monotonic, SIR works reasonably well; if $d = 2$, the regression mean function is symmetric and SIR fails to find the direction hidden in the regression mean. As a consequence, its performance is very poor. The performances of SAVE and PHD are also far from satisfactory, though they are applicable to the model theoretically. The proposed dOPG and dMAVE perform very well and are better than the existing methods listed in Table 2.

TABLE 2
Mean (and standard deviation) of estimation errors for Example 4.2

d	**n**	**dOPG**	**dMAVE**	**rMAVE**	**SIR**	**SAVE**	**PHD**
	100	0.57 (0.15)	0.44 (0.12)	0.85 (0.13)	0.63 (0.15)	0.93 (0.08)	0.99 (0.08)
1	200	0.36 (0.08)	0.28 (0.06)	0.76 (0.16)	0.42 (0.09)	0.91 (0.12)	0.98 (0.07)
	400	0.24 (0.05)	0.18 (0.04)	0.68 (0.15)	0.29 (0.06)	0.64 (0.16)	0.97 (0.07)
	100	0.63 (0.19)	0.46 (0.16)	0.85 (0.16)	0.96 (0.09)	0.90 (0.06)	0.91 (0.11)
2	200	0.33 (0.10)	0.28 (0.06)	0.70 (0.18)	0.95 (0.07)	0.87 (0.11)	0.88 (0.11)
	400	0.22 (0.05)	0.19 (0.04)	0.66 (0.19)	0.95 (0.09)	0.85 (0.12)	0.89 (0.11)

EXAMPLE 4.3. In this example we demonstrate the consistency rates of the estimation methods by checking how the estimation errors change with sample size n. Consider the model

$$Y = \frac{x_1}{0.5 + (1.5 + x_2)^2} + x_3(x_3 + x_4 + 1) + 0.1\varepsilon, \tag{4.3}$$

where $\varepsilon \sim N(0, 1)$ and $X \sim N(0, I_{10})$ are independent. Model (4.3) is a combination of the two examples in Li [22]. For this model, all the theoretical requirements for the methods are satisfied. Therefore, it is fair to use the model to check their consistency rates.

In the top panel in Figure 3 the proposed methods have much smaller estimation errors than the inverse regression estimations. Because all the directions are hidden in the regression mean function, it is not surprising that rMAVE has the best performance. Multiplied by root-n, the errors should stay at a constant level if the theoretical root-n consistency is applicable to the range of sample size. The bottom panel suggests that the estimation errors of SIR and SAVE do not start to show a root-n decreasing rate for sample sizes up to 1000, while PHD, rMAVE, dOPG and dMAVE demonstrate a clear root-n consistency rate.

EXAMPLE 4.4. In our last simulation example, we consider a model with a very complicated structure. Suppose $(X_i, Y_i), i = 1, 2, \dots, n$, are drawn independently from the model $Y = \beta_1^\top X/2 + \varepsilon(1 - |\beta_1^\top X|^2)^{1/2}$, where (X, ε) satisfies $\{X \sim N(0, I_{10}), \varepsilon \sim N(0, 1) : |\beta_1^\top X| \le 1, |\beta_2^\top X| \le 1, 0.5 < (\beta_1^\top X)^2(1 - \varepsilon^2) + \varepsilon^2 \le 1\}$, and β_1 and β_2 are defined in Example 4.1. The calculation results based on 200 replications are listed in Table 3. Because of the complicated structure as shown in Figure 4, the CS directions are not easy to estimate and observe directly. However, with moderate sample size, the proposed methods can still estimate the directions accurately. It is interesting to see that SAVE also works in this example.

Based on the simulations, we have the following observations. (1) The existing methods (rMAVE, PHD, SIR and SAVE) fail in one way or another to estimate the CS directions efficiently, while dOPG and dMAVE are efficient for all the examples. (2) dOPG and dMAVE demonstrate very good finite sample performance, even a root-n rate of estimation efficiency, while some of the existing methods do not show a clear root-n rate in the range of sample sizes investigated. (3) dOPG and dMAVE are less sensitive to the number of covariates than PHD, SAVE and SIR. Simulations not reported here also suggest that the asymmetric design of X has less effect on dOPG and dMAVE than on the inverse regression estimation methods. (4) If the CS directions are all hidden in the regression mean function, rMAVE is the best and should be used. Otherwise, dOPG and dMAVE are recommended.

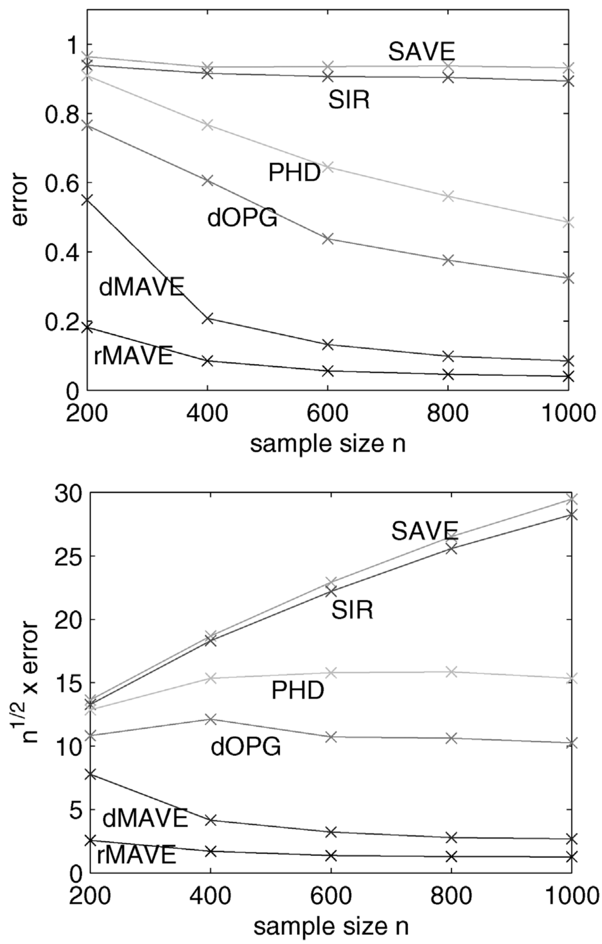

FIG. 3. *The calculation results for Example* 4.3 *using different estimation methods. The lines are the mean of estimation errors with different sample sizes and* 200 *replications. The top panel is the plot of the errors against sample size*; *the bottom panel is the errors multiplied by root-n against sample size.*

5. Real data analysis.

EXAMPLE 5.1 (*Cars data*). This data was used by the American Statistical Association in its second (1983) exposition of statistical graphics technology. The data set is available at lib.stat.cmu.edu/datasets/cars.data. There are 406 observations on eight variables: miles per gallon (Y), number of cylinders (X_1), engine displacement (X_2), horsepower (X_3), vehicle weight (X_4), time to accelerate from 0 to 60 mph (X_5), model year (X_6) and origin of the car ($1 =$ American, $2 =$ European, $3 =$ Japanese).

TABLE 3
Mean (and standard deviation) of estimation errors for Example 4.4

n	**dOPG**	**dMAVE**	**rMAVE**	**SIR**	**SAVE**	**PHD**
200	0.5909 (0.29)	0.5089 (0.30)	0.9411 (0.07)	0.8770 (0.12)	0.9242 (0.19)	0.9833 (0.05)
400	0.2117 (0.19)	0.1498 (0.10)	0.9573 (0.05)	0.8783 (0.13)	0.7677 (0.18)	0.9789 (0.03)
600	0.1148 (0.04)	0.1059 (0.03)	0.9725 (0.03)	0.8758 (0.13)	0.5357 (0.21)	0.9799 (0.03)
800	0.0876 (0.03)	0.0862 (0.02)	0.9744 (0.03)	0.8737 (0.14)	0.3657 (0.13)	0.9757 (0.04)
1000	0.0782 (0.02)	0.0779 (0.02)	0.9671 (0.04)	0.8819 (0.13)	0.2604 (0.06)	0.9789 (0.04)

Now we investigate the relation between the response variable Y and the covariates $X = (X_1, \dots, X_8)^\top$, where $X_1, \dots, X_6$ are defined above, $X_7 = 1$ if a car is from America and 0 otherwise and $X_8 = 1$ if it is from Europe and 0 otherwise. Thus, $(X_7, X_8) = (1, 0)$, $(0, 1)$ and $(0, 0)$ correspond to American cars, European cars and Japanese cars, respectively. For ease of explanation, all covariates are standardized separately. When applying dOPG to the data, the first four largest eigenvalues are 21.1573, 1.6077, 0.2791 and 0.2447. Thus, we consider CS with dimension 2. Based on dMAVE, the two directions (coefficients of X) are estimated as $\hat{\beta}_1 = (-0.33, -0.45, -0.45, -0.53, 0.14, 0.42, 0.00, -0.02)^\top$ and $\hat{\beta}_2 =$

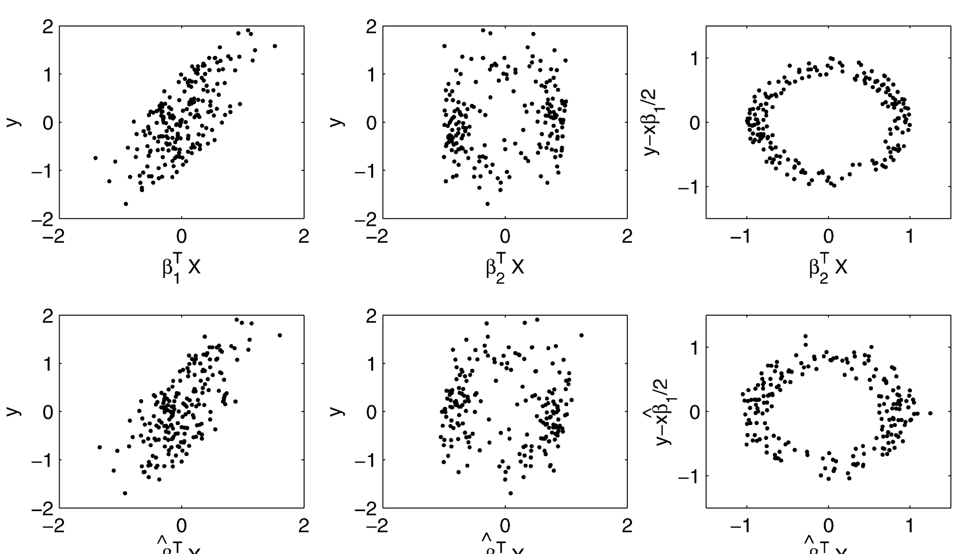

FIG. 4. *A typical data set from Example 4.4 with $n = 200$ and its dMAVE estimation. The upper three panels are plots of y against the true CS directions and $y - x^\top \beta_1/2$ against the second direction; the lower three panels are plots of y against the estimated CS directions (with estimation error 0.32) and $y - x^\top \hat{\beta}_1/2$ against the second estimated direction.*

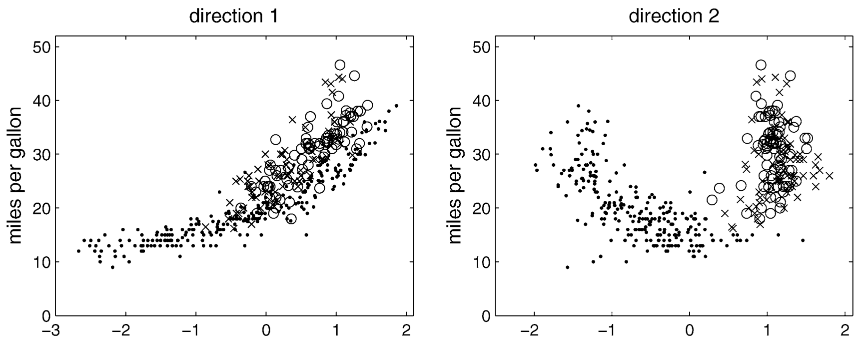

FIG. 5. *The estimation results for Example 5.1 using dMAVE. The two panels are plots of Y against the two estimated CS directions. The origins of cars are denoted by* "·" *for American cars*, "×" *for European cars and* "∘" *for Japanese cars*.

$(0.00, 0.15, -0.10, -0.23, -0.12, -0.17, -0.88, 0.29)^\top$. The plots of Y against $\hat{\beta}_1^\top X$ and $\hat{\beta}_2^\top X$ are shown in Figure 5.

Based on the estimated CS directions and Figure 5, we have the following insights to the data. The first direction highlights the common structure for cars of all origins: miles per gallon (Y) decreases with number of cylinders (X_1), engine displacement (X_2), horsepower (X_3) and vehicle weight (X_4), and increases with the time to accelerate (X_5) and model year (X_6). The second direction indicates the difference between American cars and European or Japanese cars.

EXAMPLE 5.2 (*Ground level ozone*). Air pollution has a serious impact on the health of plants and animals (including humans); see the report of the World Health Organization (WHO) [29]. Substances not naturally found in the air or at greater concentrations than usual are referred to as "pollutants." The main pollutants include nitrogen dioxide (NO_2), carbon dioxide (CO), sulphur dioxide (SO_2), respirable particulates, ground-level ozone (O_3) and others. Pollutants can be classified as either primary pollutants or secondary pollutants. Primary pollutants are substances directly produced by a process, such as ash from a volcanic eruption or the carbon monoxide gas from a motor vehicle exhaust. Secondary pollutants are products of reactions among primary pollutants and other gases. They are not directly emitted and thus cannot be controlled directly. The main secondary pollutant is ozone.

Next, we investigate the statistical relation between the level of ground-level ozone and the levels of primary pollutants and weather conditions by applying our method to pollution data observed in Hong Kong [(1994–1997) www.stat.nus.edu.sg/%7Estaxyc/HongKongAirpollution.xls] and Chicago [(1995–2000) www.ihapss.jhsph.edu/data/data.htm]. This investigation is of interest in understanding how the secondary pollutant ozone is generated from the primary pollutants and weather conditions. Let Y, N, S, P, T and H be the weekly average levels of ozone, nitrogen dioxide (NO_2), sulphur dioxide (SO_2), respirable

TABLE 4
The estimated CS directions in Example 5.2

City	Direction	***N***	***S***	***P***	***T***	***H***	***N*∗*S***	***N*∗*P***	***N*∗*T***
	β_1	0.10	−0.13	−0.06	−0.00	−0.00	0.06	0.29	0.19
Chicago	β_2	−0.10	−0.11	0.39	−0.25	−0.07	0.12	−0.15	0.09
	β_1	0.32	−0.15	0.23	0.10	−0.41	−0.07	0.20	0.42
Hong Kong	β_2	−0.04	−0.08	−0.12	0.18	0.19	−0.21	0.35	0.17
		N*∗*H	***S*∗*P***	***S*∗*T***	***S*∗*H***	***P*∗*T***	***P*∗*H***	***T*∗*H***	
	β_1	0.04	−0.18	0.27	−0.01	−0.06	0.36	0.77	
Chicago	β_2	−0.51	0.46	−0.20	−0.21	−0.15	−0.16	0.32	
	β_1	0.10	0.01	−0.05	−0.31	0.53	0.12	−0.14	
Hong Kong	β_2	−0.52	−0.26	−0.18	0.42	0.22	−0.29	−0.19	

particulates, temperature and humidity, respectively. To include the interaction between primary pollutants and weather conditions into the model directly, we further consider their cross-products resulting in 15 covariates altogether, denoted by X. For ease of explanation, all covariates are standardized separately. For all possible working dimensions, only the first two dimensions show clear relations with Y. We further calculate the eigenvalues in dOPG. The largest four eigenvalues are 10.78, 2.93, 2.11, 1.70 for Chicago, and 6.89, 1.24, 0.69, 0.52 for Hong Kong. Now we consider dimension reduction with efficient dimension 2, although the estimation of the number of dimensions needs further investigation. The estimates for the first two directions are given in Table 4.

The plots of Y against the two estimated directions are shown in Figure 6. The plots show strong similar patterns in the two separated cities. If we check the estimated coefficients (directions), NO_2 and particulates (or their interaction) are the most important pollutants that affect the level of ozone. Temperature and humidity and their interaction are the other important factors. The interactions of weather conditions with NO_2 and particulates also contribute to the variation of ozone levels. These statistical conclusions give support to the chemical claim that ozone is formed by chemical reactions between reactive organic gases and oxides of nitrogen in the presence of sunlight; see the WHO report [29].

6. Proofs.

6.1. *Basic ideas of the proofs.* The basic idea to prove the theorems is based on the convergence of the algorithms and that the true dimension reduction space is the attractor of the algorithms. We here give a more detailed outline for the proof of Theorem 3.2. Suppose the estimate of B_0 in an iteration of the dMAVE algorithm

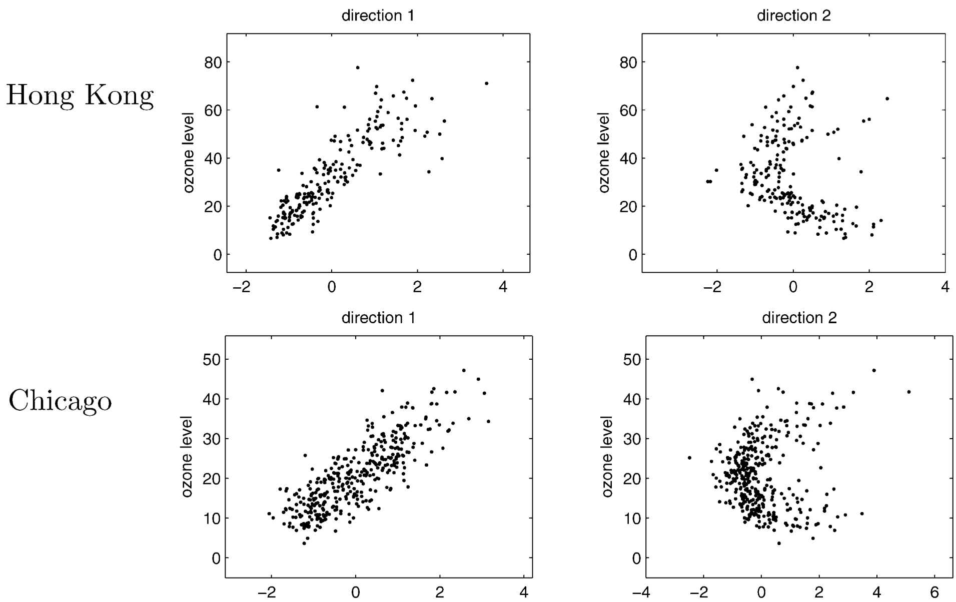

FIG. 6. *The estimation results for Example 5.2 using dMAVE. The upper two panels are the levels of ozone against the first two estimated CS directions in Hong Kong, the lower two panels are those in Chicago.*

is $B_{(t)}$. It follows from Step 2 that

$$\mathbf{b}^{(t+1)} = \ell(B_0) + \left\{ \sum_{k,j,i=1}^{n} \rho_{jk}^{(t)} K_{h_t}(B_{(t)}^{\top} X_{ij}) X_{ijk}^{(t)} (X_{ijk}^{(t)})^{\top} \right\}^{-1}$$

$$\times \sum_{k,j,i=1}^{n} \rho_{jk}^{(t)} K_{h_t}(B_{(t)}^{\top} X_{ij}) X_{ijk}^{(t)} \tag{6.1}$$

$$\times \{ H_{b,i}(Y_k) - a_{jk}^{(t)} - \ell(B_0)^{\top} X_{ijk}^{(t)} \},$$

where $X_{ijk}^{(t)}$ is defined in the algorithm. By the decomposition in Step 3, we obtain the estimate $B_{(t+1)}$ in the next iteration. If the initial value $B_{(1)}$ is a consistent estimator of B_0, by Lemmas 6.3, 6.4 and 6.5 below, we will obtain a recurring relation for the iterations as

$$\ell(B_{(t+1)}) - \ell(B_0) = \Theta_t \{ \ell(B_{(t)}) - \ell(B_0) \} + \Gamma_{n,t},$$

with $|\Theta_t| < 1$ and $|\Gamma_{n,t}| = o(1)$ almost surely when $t \geq 1$. Therefore, the dimension reduction space is an attractor in the algorithm. This recurring relation is then used to prove the convergence of the algorithm and the consistency of the final estimator. To ensure the convergence of the algorithm, we need to consider consistency with probability 1.

The details of the proofs are organized as follows. In Section 6.2 we first list a series of lemmas, Lemmas 6.1–6.5. Based on these lemmas, the theorems are then proved. The proofs of Lemmas 6.1–6.5 are algebraic albeit complex calculations based on Lemmas 6.6 and 6.7. They can be found in Xia [31] and are available upon request. Lemmas 6.6 and 6.7 are two basic results used in the proof dealing with uniform consistency. Their proofs are given in Section 6.3.

6.2. *Proofs of the theorems.* We first introduce notation. Let $\varepsilon_{b,i}(y) = H_b(Y_i - y) - E(H_b(Y_i - y)|X_i)$, $\mathcal{D}_Y \subset \mathbb{R}$ be a compact interior support of Y; that is, for any $v \in \mathcal{D}_Y$, there exists $\delta > 0$ such that $\inf_{y:|y-v|<\delta} f_Y(y) > 0$. Similarly, we can define a compact interior support $\mathcal{D}_X$ for X. For $\mathcal{B} \subset \{B : B^\top B = I_q\}$, define $\delta_B = \max\{|B - B_0| : B \in \mathcal{B}\}$. For any index set $\mathcal{Z}$ and random matrix $A_n(z)$, we say $A_n(z) = \mathcal{O}(a_n|z \in \mathcal{Z})$, or $A_n(z) = \mathcal{O}(a_n)$ for simplicity, if $\sup_{z\in\mathcal{Z}} |A_n(z)|/a_n = O(1)$ almost surely. As usual, $A_n = O_P(a_n)$ indicates that every term in A_n is $O(a_n)$ in probability as $n \to \infty$. Recall that $B_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0q})$ and $B = (\beta_1, \beta_2, \ldots, \beta_q)$. Let $H_{b,i}^{1,B}(x) = g_b(B_0^\top x, y) + \triangledown^\top g_b(B_0^\top x, y)B^\top X_{ix}$, $H_{b,i}^{2,B}(x) = \sum_{\iota,\kappa=1}^q \triangledown_{\iota,\kappa}^2 g_b(B_0^\top x, y)(\beta_\iota^\top X_{ix})(\beta_\kappa^\top X_{ix})/2$ and $H_{b,i}^{3,B}(x) = \sum_{\iota,\kappa,\tau=1}^q \{\triangledown_{\iota,\kappa,\tau}^3 g_b(B_0^\top x, y)\ (\beta_\iota^\top X_{ix})(\beta_\kappa^\top X_{ix})(\beta_\tau^\top X_{ix})\}/6$, where $X_{ix} = X_i - x$, $\triangledown g_b(v_1, \ldots, v_q, y)$ is defined in Section 2 and

$$\triangledown_{\iota,\kappa}^2 g_b(v_1, \ldots, v_q, y) = \frac{\partial^2}{\partial v_\iota\, \partial v_\kappa} g_b(v_1, \ldots, v_q, y) \qquad \text{for } \iota, \kappa = 1, 2, \ldots, q,$$

and $\triangledown_{\kappa,\tau,\iota}^3 g_b$ is defined naturally. By the Taylor expansion of $g_b(B_0^\top X_i, y)$ at $B_0^\top x$, it follows from model (2.1) that

$$H_{b,i}(y) = H_{b,i}^{1,B_0}(x) + H_{b,i}^{2,B_0}(x) + H_{b,i}^{3,B_0}(x) + \varepsilon_{b,i}(y) + O(|B_0^\top X_{ix}|^4) \tag{6.2}$$

almost surely. Let $\delta_{mh} = (nh^m/\log n)^{-1/2}$, $\delta_{mhb} = (nh^m b/\log n)^{-1/2}$ for any integer m, $\delta_b = (nb/\log n)^{-1/2}$, $\delta_n = (\log n/n)^{1/2}$ and $r_{mhb} = h^2 + b^4 + \delta_b + \delta_{mh}$. Let f_B, f and f_Y be the density functions of $B^\top X$, X and Y, respectively. Again, for simplicity, we write $f_B(x)$, $\mu_B(x)$ and $w_B(x)$ for $f_B(B^\top x)$, $\mu_B(B^\top x)$ and $w_B(B^\top x)$, respectively; see also the definitions in Section 3. Let $c, c_0, c_1, \ldots,$ be positive constants, where c may have different values at different places.

LEMMA 6.1 (Kernel smoother in the first iteration). *Let*

$$\begin{pmatrix} a_{xy} \\ b_{xy}h \end{pmatrix} = \left\{\sum_{i=1}^n K_h(X_{ix}) \begin{pmatrix} 1 \\ X_{ix}/h \end{pmatrix} \begin{pmatrix} 1 \\ X_{ix}/h \end{pmatrix}^\top\right\}^{-1} \sum_{i=1}^n K_h(X_{ix}) \begin{pmatrix} 1 \\ X_{ix}/h \end{pmatrix} H_{b,i}(y).$$

Under assumptions (C1), (C2) *and* (C4), *if* $h \to 0, b \to 0$ *and* $nh^{p+2}b/\log n \to \infty$, *then we have*

$$a_{xy} = g_b(B_0^\top x, y) + \tfrac{1}{2}\sum_{\kappa=1}^q \triangledown_{\kappa,\kappa}^2 g_b(B_0^\top x, y)h^2 + \mathcal{O}(h^3 + \delta_{phb}|x \in \mathcal{D}_X, y \in \mathcal{D}_Y),$$

$$b_{xy} = B_0 \bigtriangledown g_b(B_0^\top x, y) + \{\mu_{2p} n h^2 f(x)\}^{-1} \sum_{i=1}^{n} K_h(X_{ix}) X_{ix} \varepsilon_{b,i}(y)$$
$$+ \mathcal{O}(h^2 + \delta_{phb} | x \in \mathcal{D}_X, y \in \mathcal{D}_Y).$$

LEMMA 6.2 (Kernel smoother in dOPG). *Define* $\mathcal{D}_q = \{D = B\,\mathrm{diag}(\lambda_1, \dots, \lambda_q) B^\top + \tilde{B}\,\mathrm{diag}(\lambda_{q+1}, \dots, \lambda_p) \tilde{B}^\top : (B, \tilde{B})^\top (B, \tilde{B}) = I_p,\ c_1 > \min(\lambda_1, \dots, \lambda_q) \ge c_0 > 0,\ B \in \mathcal{B}$ *and* $\max(\lambda_{q+1}, \dots, \lambda_p)/h^2 \le e_n\}$. *Let*

$$S_n^D(x) = n^{-1} \sum_{i=1}^{n} K_h(D^{1/2} X_{ix}) \begin{pmatrix} 1 \\ X_{ix} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ix} \end{pmatrix}^\top$$

and

$$\begin{pmatrix} a_{xy}^D \\ b_{xy}^D \end{pmatrix} = \{n S_n^D(x)\}^{-1} \sum_{i=1}^{n} K_h(D^{1/2} X_{ix}) \begin{pmatrix} 1 \\ X_{ix} \end{pmatrix} H_{b,i}(y).$$

Under assumptions (C1), (C2) *and* (C4), *if* $nh^{q+2}b/\log n \to \infty$, $b \to 0, h \to 0$, $\delta_B/h \to 0$ *and* $e_n \to 0$, *then we have*

$$a_{xy}^D = g_b(B_0^\top x, y) + \tfrac{1}{2} \sum_{\kappa=1}^{q} \bigtriangledown_{\kappa,\kappa}^2 g_b(B_0^\top x, y) h^2$$
$$+ \mathcal{O}(h^3 + \delta_{qhb} | x \in \mathcal{D}_X, y \in \mathcal{D}_Y, D \in \mathcal{D}_q),$$
$$b_{xy}^D = B_0\{\bigtriangledown g_b(B_0^\top x, y) + \mathcal{O}(h^2 + \delta_{qh} + e_n)\} + \mathcal{E}_{n,0}^D(x, y)$$
$$+ \mathcal{O}(\epsilon_{qhb} | x \in \mathcal{D}_X, y \in \mathcal{D}_Y, D \in \mathcal{D}_q),$$

where $\epsilon_{qhb} = h^4 + (h^2 + \delta_{qh})\delta_{qhb} + (h^2 + \delta_{qhb})e_n + (h + \delta_{qhb}/h)\delta_B$ *and*

$$\mathcal{E}_{n,0}^D(x, y) = h^{p-q}\{n f_B(x)\}^{-1}$$
$$\times \prod_{\tau=1}^{q} \lambda_\tau^{1/2} \bar{w}_B^+(x) \sum_{i=1}^{n} K_h(D^{1/2} X_{ix}) \{\mu_B(x) - X_i\} \varepsilon_{b,i}(y).$$

LEMMA 6.3 (Kernel smoother in dMAVE). *Let*

$$\Sigma_n^B(x) = n^{-1} \sum_{i=1}^{n} K_h(B^\top X_{ix}) \begin{pmatrix} 1 \\ B^\top X_{ix}/h \end{pmatrix} \begin{pmatrix} 1 \\ B^\top X_{ix}/h \end{pmatrix}^\top$$

and

$$\begin{pmatrix} a_{xy}^B \\ d_{xy}^B h \end{pmatrix} = \{n \Sigma_n^B(x)\}^{-1} \sum_{i=1}^{n} K_h(B^\top X_{ix}) \begin{pmatrix} 1 \\ B^\top X_{ix}/h \end{pmatrix} H_{b,i}(y).$$

Under assumptions (C1), (C2) *and* (C4), *if* $nh^q b/\log n \to \infty$, $b \to 0$, $h \to 0$ *and* $\delta_B/h \to 0$, *then*

$$a_{xy}^B = g_b(B_0^\top x, y) + \nabla^\top g_b(B_0^\top x, y)(B_0 - B)^\top \nu_B(x) + \frac{1}{2}\sum_{\kappa=1}^{q} \nabla_{\kappa,\kappa}^2 g_b(B_0^\top x, y)h^2$$
$$+ \mathcal{V}_{1n}^B(x, y) + \mathcal{O}(h^4 + \delta_{qh}\delta_{qhb} + h\delta_B + \delta_B^2 | x \in \mathcal{D}_X, y \in \mathcal{D}_Y, B \in \mathcal{B}),$$
$$d_{xy}^B h = \nabla g_b(B_0^\top x, y)h + M_{1n}^B(x, y)h^3 + \mathcal{V}_{2n}^B(x, y)$$
$$+ \mathcal{O}(h^4 + \delta_{qh}\delta_{qhb} + h\delta_B + \delta_B^2 | x \in \mathcal{D}_X, y \in \mathcal{D}_Y, B \in \mathcal{B}),$$

where

$$\mathcal{V}_{1n}^B(x, y) = \{1 + M_{2n}^B(x, h)h\}\mathcal{E}_{n,1}^B(x, y) + M_{3n}^B(x, h)h\mathcal{E}_{n,2}^B(x, y),$$
$$\mathcal{V}_{2n}^B(x, y) = M_{4n}^B(x)h\mathcal{E}_{n,1}^B(x, y) + \{1 + M_{5n}^B(x, h)h\}\mathcal{E}_{n,2}^B(x, y),$$

$M_{kn}^B(x), k = 1, 2, \dots, 5$, *are bounded continuous functions* (*details can be found in the proofs*) *and*

$$\mathcal{E}_{n,1}^B(x, y) = \{nf_B(x)\}^{-1}\sum_{i=1}^{n} K_h(B^\top X_{ix})\varepsilon_{b,i}(y),$$
$$\mathcal{E}_{n,2}^B(x, y) = \{nhf_B(x)\}^{-1}\sum_{i=1}^{n} K_h(B^\top X_{ix})B^\top X_{ix}\varepsilon_{b,i}(y).$$

LEMMA 6.4 (Denominator of dMAVE). *Let* $\hat\rho_{jk}^B = \rho(\hat f_B(X_j))\rho(\hat f_Y(Y_k))$, *where*

$$\hat f_B(x) = n^{-1}\sum_{i=1}^{n} K_h(B^\top X_{ix}), \qquad \hat f_Y(y) = n^{-1}\sum_{i=1}^{n} H_b(Y_i - y).$$

Let $X_{ijk}^B = d_{jk}^B \otimes X_{ij}$, *where* $d_{jk}^B = d_{X_j Y_k}^B$. *Suppose* (C1)–(C4) *hold and* $nh^{q+2}b/\log n \to \infty$, $nb^2/\log n \to \infty$, $b \to 0$, $h \to 0$ *and* $\delta_B/h \to 0$. *We have*

$$\left\{n^{-3}\sum_{k,j,i=1}^{n} \hat\rho_{jk}^B K_h(B^\top X_{ij})X_{ijk}^B (X_{ijk}^B)^\top\right\}^{-1}$$
$$= (I_q \otimes B)L_1^B(I_q \otimes B^\top)h^{-2} + (I_q \otimes B)L_2$$
$$+ L_3(I_q \otimes B^\top) + \frac{1}{2}D_B^+ + \mathcal{O}\{(r_{qhb} + \delta_{qhb})/h | B \in \mathcal{B}\},$$

where L_1, L_2 *and* L_3 *are constant matrices* (*details can be found in the proof*) *and* $D_B = \int \rho(f_B(x))\rho(f_Y(y)) \nabla g_b(B_0^\top x, y) \nabla^\top g_b(B_0^\top x, y) \otimes \{\nu_B(x)\nu_B^\top(x)\} f(x) \times f(y)\, dx\, dy$.

LEMMA 6.5 (Numerator of dMAVE). *Suppose conditions* (C1)–(C4) *hold. If* $b \to 0$, $h \to 0$, $nh^q b/\log n \to \infty$, $nb^2/\log n \to \infty$ *and* $\delta_B/h \to 0$, *then*

$$\begin{aligned}
n^{-3} & \sum_{k,j,i=1}^{n} \hat{\rho}_{jk}^{B} K_h(B^\top X_{ij}) X_{ijk}^{B}\{H_{b,i}(Y_k) - a_{jk}^{B} - \ell(B_0)^\top X_{ijk}^{B}\} \\
&= D_B\big(\ell(B) - \ell(B_0)\big) + \Phi_n(B_0) \\
&\quad + \mathcal{O}\{h^4 + r_{qhb}\delta_{qhb} + \delta_{qhb}^2 + \delta_n^2/b^2 + (\delta_{qhb}/h + h)\delta_B | B \in \mathcal{B}\},
\end{aligned}$$

where $a_{jk}^B = a_{X_j Y_k}^B$, $\Phi_n(B_0) = O(\delta_n + \delta_{qhb}^2/h)$ *almost surely and* $\Phi_n(B_0) = O_P(n^{-1/2})$ *with* $(I_q \otimes B_0^\top)\Phi_n(B_0) = 0$ *and* $\sqrt{n}\Phi_n(B_0) \xrightarrow{D} N(0, \Sigma_0)$, *where* Σ_0 *is given in Theorem* 3.2.

PROOF OF THEOREM 3.1. By Lemma 6.1, write

$$\begin{aligned}
b_{xy} = B_0 c_n(x, y) &+ \{\mu_{2p} n h_0^2 f(x)\}^{-1} \\
&\times \sum_{i=1}^{n} K_{h_0}(X_{ix}) X_{ix} \varepsilon_{b_0,i}(y) + \tilde{B}_0 \mathcal{O}(h_0^2 + \delta_{ph_0b_0}),
\end{aligned}$$

where $(B_0, \tilde{B}_0)$ is a $p \times p$ orthogonal matrix and $c_n(x, y) = \triangledown g_b(B_0^\top x, y) + \mathcal{O}(h_0^2 + \delta_{ph_0b_0})$. By Lemma 6.6, the second term on the right-hand side above is $\mathcal{O}(\delta_{ph_0b_0}/h_0)$. It follows from Step 2 in the dOPG algorithm that

$$\begin{aligned}
\hat{\Sigma}_{(1)} = (B_0, \tilde{B}_0) C_n (B_0, \tilde{B}_0)^\top &+ n^{-3} \sum_{i,j,k=1}^{n} (S_{ijk} + S_{ijk}^\top) \\
&+ \mathcal{O}\{(h_0^2 + \delta_{ph_0b_0})\delta_{ph_0b_0}/h_0\},
\end{aligned} \tag{6.3}$$

where $\hat{\Sigma}_{(1)}$ and $\rho_{jk}^{(0)}$ are defined in the algorithm, $S_{ijk} = \rho_{jk}^{(0)}\{\mu_{2p} h_0^2 f(X_j)\}^{-1} B_0 \triangledown g_{b_0}(B_0^\top X_j, Y_k) K_{h_0}(X_{ij}) X_{ij}^\top \varepsilon_{b_0,i}(Y_k)$ and

$$\begin{aligned}
C_n &= n^{-2} \sum_{j,k=1}^{n} \rho_{jk}^{(0)} \begin{pmatrix} c_n(X_j, Y_k) \\ \mathcal{O}(h_0^2 + \delta_{ph_0b_0}) \end{pmatrix} \begin{pmatrix} c_n(X_j, Y_k) \\ \mathcal{O}(h_0^2 + \delta_{ph_0b_0}) \end{pmatrix}^\top \\
&= \begin{pmatrix} \Lambda_n^{(1)} & \mathcal{O}(h_0^2 + \delta_{ph_0b_0}) \\ \mathcal{O}(h_0^2 + \delta_{ph_0b_0}) & \mathcal{O}(h_0^4 + \delta_{ph_0b_0}^2) \end{pmatrix},
\end{aligned}$$

where $\Lambda_n^{(1)} = n^{-2} \sum_{j,k=1}^{n} \rho_{jk}^{(0)} c_n(X_j, Y_k) c_n^\top(X_j, Y_k)$. By Lemma 6.6, we have $\tilde{f}_Y^{(0)}(y) = f_Y(y) + f_Y''(y) b_0^2/2 + \mathcal{O}(b_0^4 + \delta_{b_0} | y \in \mathcal{D}_Y)$, $\tilde{f}^{(0)}(x) = f(x) + \mathcal{O}(h_0^2 + \delta_{ph_0} | x \in \mathcal{D}_X)$. By the definition of $\rho(\cdot)$, we have $\rho_{xy}^{(0)} = \rho(f(x))\tilde{\rho}_{b_0}(f_Y(y)) + \mathcal{O}(r_{ph_0b_0} | x \in \mathbb{R}^p, y \in \mathbb{R})$, where $\tilde{\rho}_{b_0}(f_Y(y)) = \rho(f_Y(y)) + \rho'(f_Y(y)) f_Y''(y) b_0^2/2$.

Let

$$\tilde{S}_{ijk} = \rho(f(X_j))\tilde{\rho}_{b_0}(f_Y(Y_k))B_0 \nabla g_{b_0}(B_0^\top X_j, Y_k) \\ \times \{\mu_{2p}h_0^2 f(X_j)\}^{-1} K_{h_0}(X_{ij})X_{ij}^\top \varepsilon_{b_0,i}(Y_k).$$

By (C5) and Lemma 6.7, we have $n^{-3}\sum_{i,j,k=1}^n \tilde{S}_{ijk} = \mathcal{O}\{(\delta_n + \delta_{ph_0b}^2 + \delta_n^2/b_0^2)/h_0\}$. Thus,

$$n^{-3}\sum_{i,j,k=1}^n S_{ijk} = n^{-3}\sum_{i,j,k=1}^n \tilde{S}_{ijk} + \mathcal{O}\{r_{ph_0b_0}\delta_{ph_0b_0}h_0^{-1}\} = \mathcal{O}(\tilde{\lambda}_n^{(1)}), \tag{6.4}$$

where

$$\tilde{\lambda}_n^{(1)} = \delta_n/h_0 + \delta_{ph_0b}^2/h_0 + \delta_n^2/(b_0^2h_0) + h_0^4 + r_{ph_0b_0}\delta_{ph_0b_0}h_0^{-1}.$$

By (C3) and the strong law of large numbers for U-statistics (cf. Hoeffding [17]), $\Lambda_n^{(1)} = \int \rho(f(x))\rho(f_Y(y)) \nabla g_{b_0}(B_0^\top x, y) \nabla^\top g_{b_0}(B_0^\top x, y)\} f(x)\ f_Y(y)\,dx\,dy + o(1)$ almost surely, which is of full rank asymptotically. Thus, its eigenvalues are greater than a positive constant asymptotically. On the other hand, the eigenvalues of the lower right principal submatrix in C_n are of order $\tilde{\lambda}_n^{(1)}$. Let $\lambda_1^{(1)} \ge \cdots \ge \lambda_p^{(1)}$ be the eigenvalues of $\hat{\Sigma}_{(1)}$ and $\beta_1^{(1)}, \ldots, \beta_p^{(1)}$ be the corresponding eigenvectors. By the interlacing theorem (cf. Ando [1]), we have $\min\{\lambda_1^{(1)}, \ldots, \lambda_q^{(1)}\} > c$ and $\max\{\lambda_{q+1}^{(1)}, \ldots, \lambda_p^{(1)}\} = \mathcal{O}(\tilde{\lambda}_n^{(1)})$. By (6.3) and (6.4), we have

$$\hat{\Sigma}_{(1)} = B_0\Lambda_n^{(1)}B_0^\top + \mathcal{O}(\delta_B^{(1)}), \tag{6.5}$$

where $\delta_B^{(1)} = r_{ph_0b_0} + \delta_{ph_0b_0} + \delta_{ph_0}^2/h_0^2 + \delta_n/h_0 + \delta_n^2/(b_0^2h_0)$. Let $B_{(1)} = (\beta_1^{(1)}, \ldots, \beta_1^{(q)})$. By Lemma 3.1 of Bai, Miao and Rao [2], we have

$$B_{(1)}B_{(1)}^\top - B_0B_0^\top = \mathcal{O}(\delta_B^{(1)}). \tag{6.6}$$

Let $t = 1$. Consider the $(t+1)$st iteration. Let $\varepsilon_{n,0}^{(t)}(x, y) = \varepsilon_{n,0}^{\hat{\Sigma}_{(t)}}(x, y)$ as defined in Lemma 6.2. By the conditions on bandwidths in (C5), we have $e_n^{(1)} \overset{\text{def}}{=} \tilde{\lambda}_n^{(1)}/h_1^2 \to 0$ and $\delta_B^{(1)}/h_1 \to 0$. By Lemma 6.2, similar to (6.3), we have from the algorithm

$$\begin{aligned}\hat{\Sigma}_{(t+1)} &= (B_0, \tilde{B}_0)C_n^{(t)}(B_0, \tilde{B}_0)^\top \\ &\quad + n^{-2}\sum_{j,k=1}^n \{S_{jk}^{(t)} + (S_{jk}^{(t)})^\top\} + \mathcal{O}(\epsilon_{qh_tb_t}\delta_{qh_tb_t}),\end{aligned} \tag{6.7}$$

where $S_{jk}^{(t)} = \rho_{jk}^{(t)} B_0\{\nabla g_{b_t}(B_0^\top X_j, Y_k) + \mathcal{O}(h^2 + \delta_{qh_t} + e_n^{(t)})\}\{\varepsilon_{n,0}^{(t)}(X_j, Y_k)\}^\top$ and

$$C_n^{(t)} = \begin{pmatrix} \Lambda_n^{(t)} & \mathcal{O}(\epsilon_{qhb}) \\ \mathcal{O}(\epsilon_{qhb}) & \mathcal{O}(\epsilon_{qhb}^2) \end{pmatrix},$$

where $\Lambda_n^{(t)} = n^{-2}\sum_{j,k=1}^n \rho_{jk}^{(t)} \nabla g_{b_t}(B_0^\top X_j, Y_k) \nabla^\top g_{b_t}(B_0^\top X_j, Y_k) + \mathcal{O}\{h_t^2 + \delta_{qh_t} + e_n^{(t)}\}$. Note that $B_{(t)}^\top \mathcal{E}_{n,0}^{(t)}(X_j, Y_k) = 0$, $\mathcal{E}_{n,0}^{(t)}(X_j, Y_k) = \mathcal{O}(\delta_{qh_tb_t})$ and $B_0^\top \mathcal{E}_{n,0}^{(t)}(X_j, Y_k) = \mathcal{O}(\delta_{qh_tb_t}\delta_B^{(t)})$. It follows that

$$
\begin{aligned}
& n^{-2}\sum_{j,k=1}^n \{S_{jk}^{(t)} + (S_{jk}^{(t)})^\top\} \\
&= (B_0, \tilde{B}_0)\left[(B_0, \tilde{B}_0)^\top n^{-3}\sum_{j,k=1}^n \{S_{jk}^{(t)} + (S_{jk}^{(t)})^\top\}(B_0, \tilde{B}_0)\right](B_0, \tilde{B}_0)^\top \qquad (6.8)\\
&= (B_0, \tilde{B}_0)\begin{pmatrix} 0 & C_{12,n}^{(t)} \\ (C_{12,n}^{(t)})^\top & 0 \end{pmatrix}(B_0, \tilde{B}_0)^\top + \mathcal{O}(\delta_{qh_tb_t}\delta_B^{(t)}),
\end{aligned}
$$

where $C_{12,n}^{(t)} = n^{-2}\sum_{j,k=1}^n \rho_{jk}^{(t)}\{\nabla g_{b_t}(B_0^\top X_j, Y_k) + \mathcal{O}(h_t^2 + \delta_{qh_t} + e_n^{(t)})\} \times \{\mathcal{E}_{n,0}^{(t)}(X_j, Y_k)\}^\top \tilde{B}_0$. Similar to $\rho_{xy}^{(0)}$, we have $\rho_{jk}^{(t)} = \tilde{\rho}_{jk}^{(t)} + \mathcal{O}(r_{qh_tb_t})$, where $\tilde{\rho}_{jk}^{(t)} = \rho(f_{B_0}(X_j))\{\rho(f_Y(Y_k)) + \rho'(f_Y(Y_k))f_Y''(Y_k)b_t^2/2\}$. By (C5) and Lemma 6.7, we have

$$
\begin{aligned}
C_{12,n}^{(t)} &= n^{-2}\sum_{j,k=1}^n \tilde{\rho}_{jk}^{(t)} \nabla g_{b_t}(B_0^\top X_j, Y_k)\{\mathcal{E}_{n,0}^{(t)}(X_j, Y_k)\}^\top \tilde{B}_0 \\
&\quad + \mathcal{O}(r_{qh_tb_t}\delta_{qh_tb_t} + e_n^{(t)}\delta_{qh_tb_t}) \qquad (6.9)\\
&= \mathcal{O}(\delta_n + \delta_{qh_tb_t}^2 + \delta_n^2 b_t^{-2} + r_{qh_tb_t}\delta_{qh_tb_t} + e_n^{(t)}\delta_{qh_tb_t}).
\end{aligned}
$$

By the strong law of large numbers for U-statistics, it follows that $\Lambda_n^{(t)} = M_0 + o(1)$ almost surely, where M_0 is defined in (C3). Let $\lambda_1^{(t+1)} \ge \cdots \ge \lambda_p^{(t+1)}$ be the eigenvalues of $\hat{\Sigma}_{(t+1)}$ and $B_{(t+1)}$ the first q eigenvectors. By the same arguments as for $\tilde{\lambda}_n^{(1)}$, it follows from (6.7), (6.8) and (6.9) that $\min\{\lambda_1^{(t+1)}, \dots, \lambda_q^{(t+1)}\} > c$ and $\max\{\lambda_{q+1}^{(t+1)}, \dots, \lambda_p^{(t+1)}\} = \mathcal{O}\{\tilde{\lambda}_n^{(t+1)}\}$, where $\tilde{\lambda}_n^{(t+1)} = \epsilon_{qh_tb_t}\delta_{qh_tb_t} + \epsilon_{qh_tb_t}^2 + \delta_{qh_tb_t}\delta_B^{(t)}$. Considering $e_n^{(t+1)}h_{t+1} \overset{\text{def}}{=} \tilde{\lambda}_n^{(t+1)}/h_{t+1}$, there exists a constant c_1 which does not depend on t such that

$$
e_n^{(t+1)}h_{t+1} \le c_1\{\chi_{0,n}^{(t)} + \chi_{1,n}^{(t)}e_n^{(t)}h_t + \chi_{2,n}^{(t)}\delta_B^{(t)}\}, \qquad (6.10)
$$

where $\chi_{0,n}^{(t)} = (h_t^4 + h_t^2\delta_{qh_tb_t} + \delta_{qh_tb_t}\delta_{qh_t})\delta_{qh_tb_t}/h_{t+1}$, $\chi_{1,n}^{(t)} = (h_t^2 + \delta_{qh_tb_t})\delta_{qh_tb_t}/(h_th_{t+1})$ and $\chi_{2,n}^{(t)} = \delta_{qh_tb_t}/h_{t+1}$. By (6.7) and (6.8), we write

$$
\begin{aligned}
\hat{\Sigma}_{(t+1)} &= B_0\Lambda_n^{(t)}B_0 + B_0\tilde{C}_{12,n}^{(t)}\tilde{B}_0^\top \\
&\quad + \tilde{B}_0(\tilde{C}_{12,n}^{(t)}B_0)^\top + \mathcal{O}\{\epsilon_{qh_tb_t} + \delta_{qh_tb_t}\delta_B^{(t)}\}, \qquad (6.11)
\end{aligned}
$$

where $\tilde{C}^{(t)}_{12,n}$ is the first term on the right-hand side of the first equation in (6.9). By the same arguments as for (6.6), we have $B_{(t+1)}B_{(t+1)}^\top - B_0B_0^\top = \mathcal{O}\{\delta_{qh_tb_t}(\delta_{qh_tb_t}+r_{qh_tb_t})+(h_t^2+r_{qh_tb_t})e_n^{(t)}+(h+\delta_{qhb}/h)\delta_B^{(t)}+\delta_n+\delta_n^2/b_t^2+h_t^4\}$. That is,

$$\delta_B^{(t+1)} \le c_2\{\chi_{3,n}^{(t)}+\chi_{4,n}^{(t)}e_n^{(t)}h_t+\chi_{5,n}^{(t)}\delta_B^{(t)}\} \tag{6.12}$$

for a constant c_2 independent of t, where $\chi_{3,n}^{(t)} = \delta_{qh_tb_t}(\delta_{qh_tb_t}+r_{qh_tb_t})+h_t^4+\delta_n^2/b_t^2+\delta_n$, $\chi_{4,n}^{(t)} = (h_t^2+r_{qh_tb_t})/h_t$ and $\chi_{5,n}^{(t)} = h_t+\delta_{qh_tb_t}/h_t$. Note that h_t and b_t decreasing with t, by (C5), we have $\delta_{qh_tb_t}/h_{t+1} \le \delta_{q\hbar\bar{b}}/\hbar \to 0$. It follows that $e_n^{(t+1)} = \lambda_n^{(t+1)}/h_{t+1}^2 \to 0$, $\delta_B^{(t+1)} = O(r_{qh_tb_t})$ and $\delta_B^{(t+1)}/h_{t+1} \to 0$. Iterating (6.10) and (6.12), it follows that

$$\delta_B^{(\infty)} = \mathcal{O}\{\chi_{3,n}^{(\infty)}+\chi_{4,n}^{(\infty)}\chi_{0,n}^{(\infty)}\} = \mathcal{O}\{\hbar^4+\delta_{q\hbar\bar{b}}(\delta_{q\hbar\bar{b}}+\hbar^2+\bar{b}^4)+\delta_n^2/\bar{b}^2+\delta_n\}$$

and $e_n^{(\infty)} = \mathcal{O}(\delta_{q\hbar\bar{b}})$. This is the first part of Theorem 3.1. By (6.11) and the equations above, write

$$\hat{\Sigma}_{(\infty)} = \{B_0+\eta_n\}\Lambda_n^{(\infty)}\{B_0+\eta_n\}^\top + \mathcal{O}\{\hbar^4+\delta_{q\hbar\bar{b}}(\delta_{q\hbar\bar{b}}+\bar{b}^4)+\delta_n^2/\bar{b}^2\},$$

where $\eta_n = \tilde{C}_{12,n}^{(\infty)}(\Lambda_n^{(\infty)})^{-1} = \mathcal{O}\{\hbar^4+\delta_{q\hbar\bar{b}}(\delta_{q\hbar\bar{b}}+\bar{b}^4)+\delta_n^2/\bar{b}^2+\delta_n\}$. Note that $B_{(\infty)}^\top \bar{w}^+_{B_{(\infty)}}(x) = 0$ and, thus, $B_{(\infty)}^\top\eta_n = 0$. We have $\tilde{\Lambda}_n \overset{\text{def}}{=} (B_0+\eta_n)^\top(B_0+\eta_n) = I_q+\mathcal{O}(\delta_n^2)$. Let $\tilde{\eta}_n = \{B_0+\eta_n\}\tilde{\Lambda}_n^{-1/2}$. It follows that

$$\hat{\Sigma}_{(\infty)} = \tilde{\eta}_n\Lambda_n^{(\infty)}\tilde{\eta}_n^\top + \mathcal{O}\{\hbar^4+\delta_{q\hbar\bar{b}}(\delta_{q\hbar\bar{b}}+\bar{b}^4)+\delta_n^2/\bar{b}^2\}.$$

Let $\hat{B}_{\text{dOPG}}$ be the first q eigenvectors of $\hat{\Sigma}_{(\infty)}$. By Lemma 3.1 of Bai, Miao and Rao [2], we have

$$\begin{aligned}\hat{B}_{\text{dOPG}}\hat{B}_{\text{dOPG}}^\top - B_0B_0^\top \\ = B_0\eta_n^\top+\eta_nB_0^\top+O\{\hbar^4+\delta_{q\hbar\bar{b}}(\delta_{q\hbar\bar{b}}+\bar{b}^4)+\delta_n^2/\bar{b}^2\}.\end{aligned} \tag{6.13}$$

By Lemma 6.7 and (C5), we have

$$\begin{aligned}\eta_n &= n^{-2}\sum_{j,k=1}^n \rho(f_{B_0}(X_j))\rho(f_Y(Y_k))\mathcal{E}_{n,0}^{(\infty)}(X_j,Y_k)\,\nabla^\top g_b(B_0^\top X_j,Y_k)\big(\Lambda_n^{(\infty)}\big)^{-1} \\ &\quad + \mathcal{O}\{r_{q\hbar\bar{b}}\delta_{q\hbar\bar{b}}\} \\ &= n^{-1}\sum_{i=1}^n \rho(f_{B_0}(X_i))\rho(f_Y(Y_i))\bar{w}^+_{B_0}(X_i)\nu_{B_0}(X_i)\zeta_i^\top\big(\Lambda_n^{(\infty)}\big)^{-1} + \mathcal{O}\{r_{q\hbar\bar{b}}\delta_{q\hbar\bar{b}}\},\end{aligned}$$

where $\zeta_i = \nabla g_{\bar{b}}(B_0^\top X_i,Y_i)f_Y(Y_i) - E\{\nabla g_{\bar{b}}(B_0^\top X_i,Y_i)f_Y(Y_i)|B_0^\top X_i\}$. Let $\tilde{\zeta}_i = \nabla f(Y_i|B_0^\top X_i)f_Y(Y_i) - E\{\nabla f(Y_i|B_0^\top X_i)f_Y(Y_i)|B_0^\top X_i\}$. As $b \to 0$, we have

$\Lambda_n^{(\infty)} \to M_0$ almost surely, where M_0 is defined in (C3). By calculating the mean and covariance matrix, we have

$$n^{-1}\sum_{i=1}^{n}\rho(f_{B_0}(X_i))\rho(f_Y(Y_i))\bar{w}_{B_0}^{+}(X_i)\nu_{B_0}(X_i)(\tilde{\zeta}_i^{\top}-\zeta_i^{\top})=o_P(n^{-1/2}).$$

It follows from the two equations above and the conditions in the theorem for the bandwidths that

$$\eta_n = n^{-1}\sum_{i=1}^{n}\rho(f_{B_0}(X_i))\rho(f_Y(Y_i))\bar{w}_{B_0}^{+}(X_i)\nu_{B_0}(X_i)\tilde{\zeta}_i^{\top}M_0^{-1}+o_P(n^{-1/2}). \tag{6.14}$$

After vectorizing η_n, the second part of Theorem 3.1 follows from (6.13), (6.14) and the central limit theorem. $\square$

PROOF OF THEOREM 3.2. Consider the initial estimator $B_{(1)}$ in (6.6). Let $\tilde{Q}=B_{(1)}^{\top}B_0$. For simplicity, we assume $\lim_{n\to\infty}\tilde{Q}=I_q$; otherwise, we may use the basis $B_0\tilde{Q}$ and consider the expansion in Lemmas 6.3, 6.4 and 6.5 at $(B_0\tilde{Q})^{\top}x$. Let $\tilde{\delta}_B^{(t)}$ be the consistency rate of the estimator in the t'th iteration. Write $\ell(B_0)=(I_q\otimes B_0)\ell(I_q)$. By the definition of D_B in Lemma 6.4, it follows that

$$\begin{aligned}(I_q\otimes B)^{\top}D_B&=0,\\ I_q\otimes B&=I_q\otimes B_0+O(\delta_B),\\ (I_q\otimes B_0)^{\top}\Phi_n(B_0)&=0.\end{aligned}\tag{6.15}$$

By the definition of the Moore–Penrose inverse, we have $D_B^{+}D_B=I_q\otimes(\tilde{B}\tilde{B}^{\top})$, where $(B,\tilde{B})$ is a $p\times p$ orthogonal matrix. By Lemmas 6.4, 6.5 and (6.1), for every $B_{(t)}$ in $\mathcal{B}=\{B:|B-B_0|\le\tilde{\delta}_B^{(t)}\}$, if $\tilde{\delta}_B^{(t)}/h_t\to 0$ we have

$$\begin{aligned}\mathbf{b}^{(t+1)}&=(I_q\otimes B_0)\{\ell(I_q)+O(c_n^{(t)})\}\\ &\quad+\tfrac{1}{2}\Psi_{(t)}\{\ell(B_{(t)})-\ell(B_0)\}+\tfrac{1}{2}D_{(t)}^{+}\Phi_n(B_0)\\ &\quad+\mathcal{O}\{\Delta_t+(h_t+\delta_{qh_tb_t}/h_t)\tilde{\delta}_B^{(t)}\},\end{aligned}\tag{6.16}$$

where $\Delta_t=h_t^4+(h_t^2+b_t^4+\delta_{qh_tb_t})\delta_{qh_tb_t}+\delta_n^2/b_t^2$, $c_n^{(t)}=\{\Delta_t+(\delta_{qh_tb_t}/h_t+h_t)\tilde{\delta}_B^{(t)}\}/h_t^2$, $D_{(t)}=D_{B_{(t)}}$ and $\Psi_{(t)}=I_q\otimes(\tilde{B}_{(t)}\tilde{B}_{(t)}^{\top})=\Psi+\tilde{\delta}_B^{(t)}$, where $\Psi=I_q\otimes(\tilde{B}_0\tilde{B}_0^{\top})$ is a projection matrix and $(B_0,\tilde{B}_0)$ is a $p\times p$ orthogonal matrix. We have

$$\begin{aligned}\mathcal{M}(\mathbf{b}^{(t+1)})&=B_0\Lambda_n^{(t)}+\tfrac{1}{2}\mathcal{M}(\Psi\{\ell(B_{(t)})-\ell(B_0)\})+\tfrac{1}{2}\mathcal{M}(D_{(t)}^{+}\Phi_n(B_0))\\ &\quad+\mathcal{O}\{\Delta_t+(h_t+\delta_{qh_tb_t}/h_t)\tilde{\delta}_B^{(t)}\},\end{aligned}$$

where $\Lambda_n^{(t)} = I_q + O(c_n^{(t)})$ and $\mathcal{M}(\cdot)$ is defined in Section 2.2. Note that

$$\begin{aligned}\tilde{\Lambda}_n^{(t+1)} &\stackrel{\text{def}}{=} \{\mathcal{M}(\mathbf{b}^{(t+1)})\}^\top \mathcal{M}(\mathbf{b}^{(t+1)}) \\ &= (\Lambda_n^{(t)})^2 + \mathcal{O}\{\delta_B^{(t)} + \tilde{\delta}_n + \Delta_t + (h_t + \delta_{qh_tb_t}/h_t)\tilde{\delta}_B^{(t)}\},\end{aligned}$$

where $\tilde{\delta}_n = \delta_n + \delta_{qh_tb_t}^2/h_t$. If $c_n^{(t)} = o(1)$ almost surely, then by Step 3,

$$\begin{aligned}B_{(t+1)} &= B_0 + \tfrac{1}{2}\mathcal{M}(\Psi\{\ell(B_{(t)}) - \ell(B_0)\}) + \tfrac{1}{2}\mathcal{M}(D_{(t)}^+\Phi_n(B_0)) \\ &\quad + \mathcal{O}\{\Delta_t + (h_t + \delta_{qh_tb_t}/h_t)\tilde{\delta}_B^{(t)}\} \\ &= B_0 + \tfrac{1}{2}\mathcal{M}(\Psi\{\ell(B_{(t)}) - \ell(B_0)\}) \\ &\quad + \mathcal{O}\{\tilde{\delta}_n + \Delta_t + (h_t + \delta_{qh_tb_t}/h_t)\tilde{\delta}_B^{(t)}\}.\end{aligned} \tag{6.17}$$

By (C5) and (6.6), we have $\delta_{qh_tb_t}/h_t^2 \le \delta_{q\hbar\bar{b}}/\hbar^2 \to 0$, $\delta_B^{(1)}/h_1 \to 0$ and $c_n^{(1)} \to 0$ almost surely. Thus, (6.17) holds for $t = 1$. By assumption (C5), it follows that $\tilde{\delta}_B^{(2)}/h_2 = o(1)$ and $c_n^{(2)} = o(1)$ almost surely. Thus, (6.17) holds for $t = 2$. Iterating the formula, we have

$$\tilde{\delta}_B^{(\infty)} = \mathcal{O}(\Delta_\infty + \tilde{\delta}_n) = \mathcal{O}\{\hbar^4 + (\hbar^2 + \bar{b}^4 + \delta_{q\hbar\bar{b}})\delta_{q\hbar\bar{b}} + \tilde{\delta}_n\}.$$

A more detailed derivation was given in Xia, Tong and Li [32]. Therefore, the first part of Theorem 3.2 follows immediately. By the first equation of (6.17) with $t = \infty$ and Lemma 6.5, we have

$$\begin{aligned}B_{(\infty)} - B_0 &= \tfrac{1}{2}\mathcal{M}(\Psi\{\ell(B_{(\infty)}) - \ell(B_0)\}) + \tfrac{1}{2}\mathcal{M}(D_{(\infty)}^+\Phi_n(B_0)) \\ &\quad + O_P\{\hbar^4 + (\hbar^2 + \bar{b}^4 + \delta_{q\hbar\bar{b}})\delta_{q\hbar\bar{b}}\}.\end{aligned}$$

Multiplying both sides by $B_0^\top$, by (6.15), we have

$$B_0^\top B_{(\infty)} - I = O_P\{\hbar^4 + (\hbar^2 + \bar{b}^4 + \delta_{q\hbar\bar{b}})\delta_{q\hbar\bar{b}}\}.$$

It follows that

$$\begin{aligned}B_{(\infty)}B_{(\infty)}^\top B_0 - B_0 &= \tfrac{1}{2}\mathcal{M}(\Psi\{\ell(B_{(\infty)}) - \ell(B_0)\}) + \tfrac{1}{2}\mathcal{M}(D_{(\infty)}^+\Phi_n(B_0)) \\ &\quad + O_P\{\hbar^4 + (\hbar^2 + \bar{b}^4 + \delta_{q\hbar\bar{b}})\delta_{q\hbar\bar{b}}\}.\end{aligned}$$

Note that $\Psi D_{(\infty)}^+ = D_{(\infty)}^+ + O_P(\tilde{\delta}_B^{(\infty)})$. We have

$$\ell(B_{(\infty)}B_{(\infty)}^\top B_0) - \ell(B_0) = D_{(\infty)}^+\Phi_n(B_0) + O_P\{\hbar^4 + (\hbar^2 + \bar{b}^4 + \delta_{q\hbar\bar{b}})\delta_{q\hbar\bar{b}}\}.$$

This is the second part of Theorem 3.2. □

6.3. *Auxiliaries.*

LEMMA 6.6. *Suppose $m_n(\chi, Z), n = 1, 2, \ldots,$ are measurable functions of Z with index $\chi \in \mathbb{R}^d$, where d is an integer, such that* (I) $|m_n(\chi, Z)| \le M(Z)$ *with $E(M^r(Z)) < \infty$ for some $r > 2$;* (II) $\sup_\chi E|m_n(\chi, Z)|^2 < a_n$; *and* (III) $|m_n(\chi, Z) - m_n(\chi', Z)| \le |\chi - \chi'|^{\alpha_1} n^{\alpha_2} G(Z)$ *with some $\alpha_1, \alpha_2 > 0$ and $E|G(Z)| < \infty$. Suppose $\{Z_i, i = 1, \ldots, n\}$ is a random sample from Z. If $a_n = cn^{-\delta}$ with $0 \le \delta < 1 - 2/r$ and $c > 0$, then for any positive α_0 we have*

$$\sup_{|\chi| \le n^{\alpha_0}} \left| n^{-1} \sum_{i=1}^n \{m_n(\chi, Z_i) - Em_n(\chi, Z_i)\} \right| = O\{(a_n \log n/n)^{1/2}\}$$

almost surely.

PROOF. The "continuity argument" approach is used here. See, for example, Mack and Silverman [25] and Härdle, Hall and Ichimura [14]. Note that $\mathcal{D}_n \stackrel{\text{def}}{=} \{|\chi| \le n^{\alpha_0}\}$ is bounded and its Borel measure is less than $c_1 n^{\alpha_0 d}$ for some constant c_1. There are n^{α_4} ($\alpha_4 > \alpha_0 d + (1 + \alpha_2)d/\alpha_1$) balls B_{n_k} centered at χ_{n_k}, $1 \le k \le n^{\alpha_4}$, with diameter less than $c_2 n^{-(1+\alpha_2)/\alpha_1}$, such that $\mathcal{D}_n \subset \bigcup_{1 \le k \le n^{\alpha_4}} B_{n_k}$. It follows that

$$\begin{aligned} \sup_{\chi \in \mathcal{D}_n} &\left| \frac{1}{n} \sum_{i=1}^n \{m_n(\chi, Z_i) - Em_n(\chi, Z_i)\} \right| \\ &\le \max_{1 \le k \le n^{\alpha_4}} \left| \frac{1}{n} \sum_{i=1}^n \{m_n(\chi_{n_k}, Z_i) - Em_n(\chi_{n_k}, Z_i)\} \right| \\ &\quad + \max_{1 \le k \le n^{\alpha_4}} \sup_{\chi \in B_{n_k}} \left| \frac{1}{n} \sum_{i=1}^n [\{m_n(\chi, Z_i) - m_n(\chi_{n_k}, Z_i)\} \right. \\ &\qquad\qquad \left. - E\{m_n(\chi, Z_i) - m_n(\chi_{n_k}, Z_i)\}] \right| \\ &\stackrel{\text{def}}{=} \max_{1 \le k \le n^{\alpha_4}} |R_{n,k,1}| + \max_{1 \le k \le n^{\alpha_4}} \sup_{\chi \in B_{n_k}} |R_{n,k,2}|. \end{aligned} \tag{6.18}$$

By condition (III) and the definition of B_{n_k}, we have

$$\begin{aligned} \max_{1 \le k \le n^{\alpha_4}} \sup_{\chi \in B_{n_k}} |m_n(\chi, Z_i) - m_n(\chi_{n_k}, Z_i)| &\le \max_{1 \le k \le n^{\alpha_4}} \sup_{\chi \in B_{n_k}} n^{\alpha_2} |\chi - \chi_{n_k}|^{\alpha_1} G(Z_i) \\ &\le c_3 n^{-1} G(Z_i). \end{aligned}$$

By the strong law of large numbers, we have

$$\max_{1 \le k \le n^{\alpha_4}} \sup_{\chi \in B_{n_k}} |R_{n,k,2}| \le c_4 n^{-2} \sum_{i=1}^n \{G(Z_i) + EG(Z_i)\} = O(n^{-1}) \tag{6.19}$$

almost surely. Let $T_n = (na_n/\log n)^{1/2}$, $m_n^o(\chi_{n_k}, Z_i) = m_n(\chi_{n_k}, Z_i)I\{|M(Z_i)| \geq T_n\}$ and $m_n^I(\chi_{n_k}, Z_i) = m_n(\chi_{n_k}, Z_i) - m_n^o(\chi_{n_k}, Z_i)$. Write

$$R_{n,k,1} = \frac{1}{n}\sum_{i=1}^{n}[m_n^o(\chi_{n_k}, Z_i) - E\{m_n^o(\chi_{n_k}, Z_i)\}] + \frac{1}{n}\sum_{i=1}^{n}\xi_{n_k,i}, \tag{6.20}$$

where $\xi_{n_k,i} = m_n^I(\chi_{n_k}, Z_i) - E\{m_n^I(\chi_{n_k}, Z_i)\}$. By truncation, it follows that

$$E|m_n^o(\chi_{n_k}, Z_i)| \leq T_n^{-r+1}E|M(Z_i)|^r.$$

If $a_n = cn^{-\delta}$ with $0 \leq \delta < 1 - 2/r$, we have

$$n^{-1}\left|\sum_{i=1}^{n} Em_n^o(\chi_{n_k}, Z_i)\right| \leq E|M(Z_1)|^r T_n^{-r+1} = o(\{a_n \log(n)/n\}^{1/2}). \tag{6.21}$$

Again by truncation, we have

$$\begin{aligned}\sum_{i=1}^{n}|m_n^o(\chi_{n_k}, Z_i)| &\leq \sum_{i=1}^{n}|M(Z_i)|I\bigl(|M(Z_i)| \geq T_n\bigr)\\ &\leq T_n^{-r+1}\sum_{i=1}^{n}|M(Z_i)|^r I\bigl(|M(Z_i)| \geq T_n\bigr).\end{aligned}$$

For fixed T, by the strong law of large numbers, we have

$$n^{-1}\sum_{i=1}^{n}|M(Z_i)|^r I\bigl(|M(Z_i)| \geq T\bigr) \to E\bigl\{|M(Z_1)|^r I\bigl(|M(Z_1)| \geq T\bigr)\bigr\}$$

almost surely. The right-hand side above is dominated by $E\{|M(Z_i)|^r\}$ and $\to 0$ as $T \to \infty$. Note that T_n increases to ∞ with n. For large n such that $T_n > T$, we have

$$C_n \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^{n}|M(Z_i)|^r I\bigl(|M(Z_i)| \geq T_n\bigr) \leq n^{-1}\sum_{i=1}^{n}|M(Z_i)|^r I\bigl(|M(Z_i)| \geq T\bigr) \to 0$$

almost surely as $T \to \infty$. It follows that

$$\max_{1\leq k\leq n^{\alpha_4}} n^{-1}\left|\sum_{i=1}^{n} m_n^o(\chi_{n_k}, Z_i)\right| \leq C_n T_n^{-r+1} = o\{(a_n \log n/n)^{1/2}\} \tag{6.22}$$

almost surely. By condition (II), we have

$$\begin{aligned}\max_{1\leq k\leq n^{\alpha_4}} \operatorname{Var}\left(\sum_{i=1}^{n}\xi_{n_k,i}\right) &\leq n \max_{1\leq k\leq n^{\alpha_4}} E\{m_n^I(\chi_{n_k}, Z_1)\}^2\\ &\leq n \max_{1\leq k\leq n^{\alpha_4}} E\{m_n(\chi_{n_k}, Z_1)\}^2\\ &\leq c_5 n a_n \stackrel{\text{def}}{=} N_1.\end{aligned} \tag{6.23}$$

By the condition on a_n and the definition of $\xi_{n_k,i}$, we have

$$\max_{1\le k\le n^\alpha} |\xi_{n_k,i}| \le c_6 T_n = c_6(na_n/\log n)^{1/2} \overset{\text{def}}{=} N_2. \tag{6.24}$$

Let $N_3 = c_7(na_n \log n)^{1/2}$ with $c_7^2 > 2(\alpha_4+2)(c_5+c_6c_7)$. By Bernstein's inequality (cf. de la Peña [9]), we have from (6.23) and (6.24) that

$$\begin{aligned} P\left(\left|\sum_{i=1}^n \xi_{n_k,i}\right| > N_3\right) &\le 2\exp\left(\frac{-N_3^2}{2(N_1+N_2N_3)}\right) \\ &\le 2\exp\{-c_7^2 \log n/(2c_5+2c_6c_7)\} \\ &\le c_8 n^{-\alpha_4-2}. \end{aligned}$$

It follows that

$$\begin{aligned} &\sum_{n=1}^\infty \Pr\left(\max_{1\le k\le n^{\alpha_4}}\left|\sum_{i=1}^n \xi_{n_k,i}\right| \ge N_3\right) \\ &\qquad \le \sum_{n=1}^\infty n^{\alpha_4} \max_{1\le k\le n^{\alpha_4}} \Pr\left(\left|\sum_{i=1}^n \xi_{n_k,i}\right| \ge N_3\right) < \infty. \end{aligned} \tag{6.25}$$

By the Borel–Cantelli lemma (cf. Chow and Teicher [4], page 60), we have

$$\max_{1\le k\le n^{\alpha_4}}\left|\sum_{i=1}^n \xi_{n_k,i}\right| = O(N_3) \tag{6.26}$$

almost surely. Combining (6.20), (6.21), (6.22) and (6.26), we have

$$\max_{1\le k\le n^{\alpha_4}} |R_{n,k,1}| = O\{(a_n \log(n)/n)^{1/2}\} \tag{6.27}$$

almost surely. Lemma 6.6 follows from (6.18), (6.19) and (6.27). $\square$

For any function $G(X_i, Y_i, X_j, Y_j, X_k, Y_k)$ [or $G(X_j, Y_j, X_k, Y_k)$], we introduce a projection operator E_k as follows:

$$E_k G(X_i, Y_i, X_j, Y_j, X_k, Y_k) = E\{G(X_i, Y_i, X_j, Y_j, X_k, Y_k)|X_i, Y_i, X_j, Y_j\}.$$

LEMMA 6.7. *Let $\mathcal{A} = \{A : A^\top A = I_\kappa\}$ with $1 \le \kappa \le p$. Suppose $g_0(y)$, $g_1(x)$, $g_2(x)$ are bounded continuous functions. If conditions* (C2) *and* (C4) *hold with B replaced by A for all $A \in \mathcal{A}$, then*

$$\begin{aligned} &n^{-3}\sum_{i,j,k=1}^n K_h(A^\top X_{ij}) g_1(X_i) g_2(X_j) g_0(Y_k) \bigtriangledown g_b(B_0^\top X_j, Y_k)\varepsilon_{b,i}(Y_k) \\ &\quad = n^{-1}\sum_{i=1}^n E_j E_k\{K_h(A^\top X_{ij}) \bigtriangledown g_b(B_0^\top X_j, Y_k)\varepsilon_{b,i}(Y_k)\} + \mathcal{O}(\varsigma_{\kappa hb}|A\in\mathcal{A}), \end{aligned}$$

where $\varsigma_{\kappa hb} = \delta_n^3 h^{-\kappa} b^{-2} + \delta_{\kappa hb}^2 + \delta_n^2 b^{-2}$ *and the first term on the right-hand side is* $\mathcal{O}(\delta_n)$.

PROOF. For easy exposition, we consider $g_k \equiv 1, k = 0, 1, 2$ only. Let $\Delta_n(A)$ be the left-hand side of the equation in the lemma. Let $\varphi_K(s) = (2\pi)^{-\kappa} \times \int \exp(\imath s^\top u) K(u)\, du$ and $\varphi_H(t) = (2\pi)^{-1} \int \exp(\imath t v) H(v)\, dv$ be the Fourier transformations, where $\imath$ is the imaginary unit. It follows from the inverse Fourier transformation that $g_b(u, y) = b^{-1} \int \varphi_H(t') e^{-\imath t' y/b} E\{e^{\imath t' Y/b} |\ B_0^\top X = u\}\, dt'$. Thus,

$$\nabla g_b(B_0^\top X_j, Y_k) = b^{-1} \int \varphi_H(t') \nabla \tilde{g}_b(B_0^\top X_j) e^{-\imath t' Y_k/b}\, dt', \tag{6.28}$$

where $\nabla \tilde{g}_b(u) = \partial E(e^{\imath t' Y/b} | B_0^\top X = u)/\partial u$. We have

$$\begin{aligned}
\Delta_n(A) = {} & \frac{1}{n^3 b} \int \varphi_H(t') \sum_{i,j,k=1}^n \{K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j) \\
& \qquad - E_j[K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j)]\} \\
& \qquad \times \{\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b} - E_k[\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b}]\}\, dt' \\
& + \frac{1}{n^2 b} \int \varphi_H(t') \sum_{i,k=1}^n E_j[K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j)] \\
& \qquad \times \{\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b} - E_k[\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b}]\}\, dt' \\
& + \frac{1}{n^2 b} \int \varphi_H(t') \sum_{i,j=1}^n E_k[\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b}]\{K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j) \\
& \qquad - E_j[K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j)]\}\, dt' \\
& + \frac{1}{nb} \int \varphi_H(t') \sum_{i=1}^n E_j[K_h(A^\top X_{ij}) \nabla \tilde{g}_b(B_0^\top X_j)] \\
& \qquad \times E_k[\varepsilon_{b,i}(Y_k) e^{-\imath t' Y_k/b}]\, dt' \\
\stackrel{\text{def}}{=} {} & \Delta_{n,1}(A) + \Delta_{n,2}(A) + \Delta_{n,3}(A) + \Delta_{n,4}(A).
\end{aligned} \tag{6.29}$$

By inverse Fourier transformation, it follows that $K_h(A^\top X_{ij}) = h^{-\kappa} \int \varphi_K(s) \times e^{-\imath s^\top A^\top X_{ij}/h}\, ds$ and $H_b(Y_i - Y_k) = b^{-1} \int \varphi_H(t) e^{-\imath t (Y_i - Y_k)/b}\, dt$. Thus,

$$\Delta_{n,1}(A) = \frac{1}{n^3 h^\kappa b^2} \int \prod_{\ell=1}^3 \sum_{i=1}^n m_{\ell,n}(A, s, t, t', X_i, Y_i) \varphi_K(s) \varphi_H(t) \varphi_H(t')\, ds\, dt\, dt',$$

where

$$m_{1,n}(A,s,t,t',X_i,Y_i) = e^{-\imath s^\top A^\top X_i/h} \bigtriangledown \tilde{g}_b(B_0^\top X_i) - E[e^{-\imath s^\top A^\top X_i/h} \bigtriangledown \tilde{g}_b(B_0^\top X_i)],$$

$$m_{2,n}(A,s,t,t',X_i,Y_i) = e^{\imath(t-t')Y_i/b} - E\big[e^{\imath(t-t')Y_i/b}\big]$$

and

$$m_{3,n}(A,s,t,t',X_i,Y_i) = e^{-\imath tY_i/b} - E(e^{-\imath tY_i/b}|X_i).$$

By (C2), we have that $|\bigtriangledown \tilde{g}_b(u)| \le \int |\bigtriangledown f_0(y|u)|\,dy$ is bounded. For any $r > 2$, it follows that $\sup_{t'} E\{\bigtriangledown \tilde{g}_b(B_0^\top X_i)\}^r \le c$ and that

$$\sup_{A,s,t,t'} E|m_{\ell,n}(A,s,t,t',X_i,Y_i)|^r \le c, \qquad \ell = 1,2,3,$$

where c is a finite constant. For any $\alpha_0 > 0$, let $\mathcal{D}_n' = \{(t,t',s): |t| \le n^{\alpha_0}, |t'| \le n^{\alpha_0}, |s| \le n^{\alpha_0}\}$. By taking $\chi = (A,t,t',s)$ and $a_n = c$, we have from Lemma 6.6,

$$\sup_{A\in\mathcal{A},(t,t',s)\in\mathcal{D}_n'} n^{-1}\left|\sum_{i=1}^n m_{\ell,n}(A,s,t,t',X_i,Y_i)\right| = O(\delta_n), \qquad \ell = 1,2,3, \tag{6.30}$$

almost surely. On the other hand, $|m_{\ell,n}(A,s,t,t',X_i,Y_i)|$ is bounded. Thus,

$$\sup_{A\in\mathcal{A},(t,t',s)} n^{-1}\left|\sum_{i=1}^n m_{\ell,n}(A,s,t,t',X_i,Y_i)\right| = O(1), \qquad \ell = 1,2,3. \tag{6.31}$$

By (C4), the Fourier transformation functions $\varphi_K(\cdot)$ and $\varphi_H(\cdot)$ are absolutely integrable; see Chung [5], page 166. We can choose α_0 such that

$$\int_{|s|>n^{\alpha_0}} |\varphi_K(s)|\,ds = O(\delta_n^3), \qquad \int_{|t|>n^{\alpha_0}} |\varphi_H(t)|\,dt < O(\delta_n^3). \tag{6.32}$$

Partitioning the integration region in $\Delta_{n,1}(A)$ into two parts, we have from (6.30)–(6.32) that

$$\begin{aligned}
\sup_{A\in\mathcal{A}} |\Delta_{n,1}(A)| &\le \frac{1}{n^3h^\kappa b^2}\int_{(s,t,t')\in\mathcal{D}_n'} \prod_{\ell=1}^3 \sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n m_{\ell,n}(A,s,t,t',X_i,Y_i)\right| \\
&\qquad\qquad \times |\varphi_K(s)\varphi_H(t)\varphi_H(t')|\,ds\,dt\,dt' \\
&\quad + \frac{1}{n^3h^\kappa b^2}\int_{(s,t,t')\notin\mathcal{D}_n'} \prod_{\ell=1}^3 \sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n m_{\ell,n}(A,s,t,t',X_i,Y_i)\right| \\
&\qquad\qquad \times |\varphi_K(s)\varphi_H(t)\varphi_H(t')|\,ds\,dt\,dt' \\
&= (h^\kappa b^2)^{-1} O(\delta_n^3) \int |\varphi_K(s)\varphi_H(t)\varphi_H(t')|\,ds\,dt\,dt' \qquad (6.33)
\end{aligned}$$

$$+ (h^\kappa b^2)^{-1} O(1) \int_{(s,t,t')\notin \mathcal{D}'_n} |\varphi_K(s)\varphi_H(t)\varphi_H(t')|\, ds\, dt\, dt'$$
$$= O(\delta_n^3 h^{-\kappa} b^{-2})$$

almost surely. Let $\tilde{g}(X_i) = E_j[K_h(A^\top X_{ij}) \triangledown \tilde{g}_b(B_0^\top X_j)]$. It is easy to see that $\tilde{g}(X_i) = O(1)$ almost surely. Applying the inverse Fourier transformation to $\varepsilon_{b,i}(Y_k)$ and using similar arguments leading to (6.33), we have

$$\sup_{A\in\mathcal{A}} |\Delta_{n,2}(A)| = O(\delta_n^2 b^{-2}) \tag{6.34}$$

almost surely. Applying the inverse Fourier transformation to $K_h(A^\top X_{ij})$, similar to (6.33), we have

$$\sup_{A\in\mathcal{A}} |\Delta_{n,3}(A)| = O(\delta_n^2 h^{-\kappa} b^{-1}) \tag{6.35}$$

almost surely. By (6.28), we have

$$\Delta_{n,4}(A) = n^{-1} \sum_{i=1}^n E_j E_k\{K_h(A^\top X_{ij}) \triangledown g_b(B_0^\top X_j, Y_k)\varepsilon_{b,i}(Y_k)\}.$$

By Lemma 6.6, we have

$$\sup_{A\in\mathcal{A}} \Delta_{n,4}(A) = O(\delta_n) \tag{6.36}$$

almost surely. Finally, Lemma 6.7 follows from (6.33)–(6.36) and (6.29). □

Acknowledgments. Two referees, an Associate Editor, Professor Z. D. Bai and Professor B. Brown provided very valuable comments and suggestions for the paper.

REFERENCES

[1] ANDO, T. (1987). Totally positive matrices. *Linear Algebra Appl.* **90** 165–219. MR0884118

[2] BAI, Z. D., MIAO, B. Q. and RAO, C. R. (1991). Estimation of directions of arrival of signals: Asymptotic results. In *Advances in Spectrum Analysis and Array Processing* (S. Haykin, ed.) **1** 327–347. Prentice Hall, Englewood Cliffs, NJ.

[3] CHEN, C.-H. and LI, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8** 289–316. MR1624402

[4] CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory. Independence, Interchangeability, Martingales*. Springer, New York. MR0513230

[5] CHUNG, K. L. (1968). *A Course in Probability Theory*. Harcourt, Brace and World, New York. MR0229268

[6] COOK, R. D. (1998). *Regression Graphics*. Wiley, New York. MR1645673

[7] COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. MR1902895

[8] COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

[9] DE LA PEÑA, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153

[10] DELECROIX, M., HRISTACHE, M. and PATILEA, V. (2005). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975

[11] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587

[12] FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455

[13] FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. MR1399164

[14] HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

[15] HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

[16] HÄRDLE, W. and TSYBAKOV, A. B. (1991). Comment on "Sliced inverse regression for dimension reduction," by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 333–335.

[17] HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. Mimeo Report No. 302, Inst. Statist., Univ. North Carolina.

[18] HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104

[19] HRISTACHE, M., JUDITSKY, A., POLZEHL, J. and SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537–1566. MR1891738

[20] HRISTACHE, M., JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333

[21] LI, B., ZHA, H. and CHIAROMONTE, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616. MR2166556

[22] LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

[23] LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039. MR1209564

[24] LUE, H.-H. (2004). Principal Hessian directions for regression with measurement error. *Biometrika* **91** 409–423. MR2081310

[25] MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415. MR0679685

[26] SAMAROV, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88** 836–847. MR1242934

[27] SCOTT, D. W. (1992). *Multivariate Density Estimation*: *Theory*, *Practice and Visualization*. Wiley, New York. MR1191168

[28] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. MR0848134

[29] WORLD HEALTH ORGANIZATION (2003). Reports on a WHO/HEI working group. Bonn, Germany.

[30] XIA, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

[31] XIA, Y. (2006). A constructive approach to the estimation of dimension reduction directions. Technical report, Dept. Statistics and Applied Probability, National Univ. Singapore.

[32] XIA, Y., TONG, H. and LI, W. K. (2002). Single-index volatility models and estimation. *Statist. Sinica* **12** 785–799. MR1929964

[33] XIA, Y., TONG, H., LI, W. K. and ZHU, L. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297

[34] YIN, X. and COOK, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 159–175. MR1904698

[35] YIN, X. and COOK, R. D. (2005). Direction estimation in single-index regressions. *Biometrika* **92** 371–384. MR2201365

DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
REPUBLIC OF SINGAPORE
E-MAIL: staxyc@stat.nus.edu.sg