Submitted to the Brazilian Journal of Probability and Statistics
arXiv: arXiv:0000.0000

Dimension Reduction Based on Conditional Multiple Index Density Function

Jun Zhang[a], **Baohua He**[b] **Tao Lu**[c] **and Songqiao Wen**[d]

[a]*Shenzhen University*
[b]*Sun Yat-Sen University*
[c]*University of Nevada*
[d]*Shenzhen University*

Abstract. In this paper, a dimension reduction method is proposed by using the first derivative of the conditional density function of response given predictors. To estimate the central subspace, we propose a direct methodology by taking expectation of the product of predictor and kernel function about response, which helps to capture the directions in the conditional density function. The consistency and asymptotic normality of the proposed estimation methodology are investigated. Furthermore, we conduct some simulations to evaluate the performance of our proposed method and compare with existing methods, and a real data set is analyzed for illustration.

1 Introduction

Accompanying the advancement of sciences and technologies in various fields such as biology, economics and finance, etc, scientific data has the tendency of growing in both size and complexity. Analysis of high-dimensional data calls for new statistical theories and methodologies. A natural way to analyze high-dimensional data is to first reduce the dimensionality of the original data without losing vital information. To reduce the problem of many covariates to one with a few covariates, sufficient dimension reduction Cook (1998) aims at finding low-dimensional linear combinations of the predictors without loss of information on $Y|X$. Suppose X is a random vector in $\mathbb{R}^p$ and Y is a univariate random variable, we need seek a $p \times d$ matrix B with $d \leq p$ satisfying

$$Y \perp\!\!\!\perp X|B^\tau X, \tag{1.1}$$

where τ is the transpose operator on a vector or a matrix and $\perp\!\!\!\perp$ indicates independence throughout this paper. That is, given $B^\tau X$, Y and X are independent. The space spanned by the column of B, which is denoted as $\boldsymbol{S}(B)$,

MSC 2010 subject classifications: Primary 62G05; 62G08; 62G20
Keywords and phrases. Central subspace, Conditional density function, Dimensional reduction, Kernel function.

is defined as a dimension reduction subspace (Cook, 1994, 1998). In fact, we are not concerned about the specific form of B since any orthogonal transformation of B from right does not affect the conditional independent property of (1.1). If all the other dimension reduction space include $\boldsymbol{S}(B)$ as their subspace, then $\boldsymbol{S}(B)$ is a so-called central dimension reduction subspace (CS), and we denote the central dimension reduction subspace (CS) as $\boldsymbol{S}_{Y|X}$. Cook (1998) gives some mild conditions which guarantee the unique existence of $\boldsymbol{S}_{Y|X}$. Without notational confusion, we assume that $\boldsymbol{S}_{Y|X}$ coincides with $\boldsymbol{S}(B)$ throughout this paper, namely, $\boldsymbol{S}_{Y|X} = \boldsymbol{S}(B)$. The dimension of $\boldsymbol{S}_{Y|X}$ is called the structural dimension and it is denoted as d in this paper.

To estimate the central subspace $\boldsymbol{S}_{Y|X}$, there are many useful methods available in the literature which make use of effective tools to reduce high-dimensional variables to equivalent ones comprising only some linear combinations of the original variables. For instance, ordinary least squares (Li and Duan, 1989, OLS), sliced inverse regression (Li, 1991, SIR), principal hessian directions (Li, 1992, pHd), sliced average variance estimation (Cook and Weisberg, 1991, SAVE), directional regression (Li and Wang, 2007, DR), minimum average variance estimation (Xia et al., 2002, MAVE), and outer product of gradient based on conditional density functions (Samarov, 1993, dOPG), density-MAVE (Xia, 2007, dMAVE), score dimension reduction (Wang and Zhu, 2013) etc. These dimension reduction methods are widely used for a large dataset and preserve $\sqrt{n}$ consistency of the associated estimators, furthermore, the computational cost is not expensive. Once the central subspace is identified, subsequent analysis with the low-dimensional $B^\tau X$ will help to construct another regression models or other statistical models based on $B^\tau X$. Specially, when the structural dimension is 1, model (1.1) reduces to the popular single-index models (Liang et al., 2010; Li et al., 2014; Chen et al., 2010; Peng and Huang, 2011; Cui et al., 2009). Moreover, if the estimation structural dimension is 1,2 or 3, the graphical visualization will gain comprehensive insights about the data, see Cook (1998) with a deep discussion about the graphical methodology. So, the dimension reduction method is widely applied in practice. See, Zhu et al. (2016,?); Lansangan and Barrios (2017); Luo et al. (2017); Yoshida (2017); Deng and Wang (2017); Sheng and Yin (2016); Zhou and Zhu (2016).

The central subspace satisfying model (1.1) is equivalent to the conditional density function of $Y|X$ being the same as that of $Y|B^\tau X$ for all possible value of X and Y if the conditional density function of Y given X exists, i.e.,

$$f_{Y|X}(y|x) = f_{Y|B^\tau X}(y|B^\tau x). \tag{1.2}$$

Model (1.2) indicates that all the directions can be captured in the conditional density function $f_{Y|X}(y|x)$. Then, an estimator of the conditional density function $f_{Y|X}(y|x)$ is needed to be proposed in the primary step to find the central subspace. Xia (2007) used the "double-kernel" local linear smoothing method (Fan et al., 1996) to estimate the conditional density function $f_{Y|X}(y|x)$. They used the fact that the conditional density function $f_{Y|X}(y|x)$ is asymptotically equivalent to the conditional regression mean function, and the directions defined in model (1.1) will be all captured in this conditional regression mean function.

In this paper, we introduce a simple methodology for dimension reduction that based on the linear condition on X (Li, 1991) and the existence of the conditional density function of Y given X. The linear condition on X is widely used in the dimension reduction literature, and the existence of conditional density function of Y given X is also a mild assumption. The methodology proposed in this paper is to take expectation of the product of X and the kernel function of Y. Specifically, we use the idea of "double-kernel" local linear smoothing method (Fan et al., 1996) to estimate the conditional density function $f_{Y|X}(y|x)$ in the first step. Due to the asymptotically equivalent conditional density function $f_{Y|X}(y|x)$ and the conditional regression mean function (Fan et al., 1996; Xia, 2007). Together with model assumption (1.2), the conditional regression mean function will be used to capture all the direction of central subspace. Next, the proposed method in this paper is easy to implement by using a spectral decomposition on a kernel matrix. As the bandwidth converges to zero, the kernel matrix will eventually recover the central subspace in the population level. In this paper, we also provide the asymptotic properties of the estimators of the estimated kernel matrix and associated eigenvalues and eigenvectors.

The reminder of the paper is organized as follows. In Section 2, we introduce the rationale of the dimension reduction method at population level, and illustrate its theoretical result. At the sample level, we introduce a direct estimation approach to estimating $\boldsymbol{S}_{Y|X}$, and establish the asymptotic properties of the resultant estimators. We demonstrate the methodologies through simulations and an analysis of a real data in Section 3. An analysis of a real data is presented in Section 4. We conclude this paper with a brief discussion in Section 5. All proofs are given in the Appendix.

2 Methodology Development

2.1 The Population level

We introduce the proposed dimension reduction method in the population level. When the conditional density function of Y given X exists, model (1) is equivalent to the conditional density function of $Y|X$ being the same as that of $Y|B^\tau X$ for all possible values of (x, y) over the support of (X, Y), ie, $f_{Y|X}(y|x) = f_{Y|B^\tau X}(y|B^\tau x)$. It implies that

$$\partial f_{Y|X}(y|x)/\partial x = B\big[\partial f_{Y|B^\tau X}(y|B^\tau x)/\partial(B^\tau x)\big]. \tag{2.1}$$

In other words, all the basis directions in $\boldsymbol{S}_{Y|X}$ can be captured by the first derivation of the conditional density function. Thus, to recover $\boldsymbol{S}_{Y|X}$, we need to estimate $\partial f_{Y|X}(y|x)/\partial x$ or $\boldsymbol{E}[\partial f_{Y|X}(y|X)/\partial X]$. However, we could not estimate this derivative directly by nonparametric smoothing when the dimensions of X are high. To avoid using nonparametric smoothing method in high-dimensional case, if X is further assumed temporarily to be normally distributed with mean zero and identical covariance matrix, a direct application of Stein (1981)'s Lemma 4 and combining with equations (1.2) and (2.1) entail that

$$\begin{aligned}&\boldsymbol{E}\big[X f_{Y|X}(y|X)\big] = \boldsymbol{E}\big[\partial f_{Y|X}(y|X)/\partial X\big]\\ &= B\boldsymbol{E}\big[\partial f_{Y|B^\tau X}(y|B^\tau X)/\partial(B^\tau X)\big].\end{aligned} \tag{2.2}$$

It is seen that $\boldsymbol{E}\big[X f_{Y|X}(y|X)\big]$ can be used to seek the directions of $\boldsymbol{S}_{Y|X}$. However, in the argument (2.2), the density $f_{Y|X}(y|x)$ is still unknown and needs to be estimated. Xia (2007) used the idea of "double-kernel" smoothing method studied in Fan et al. (1996) to construct an estimator of conditional density function $f_{Y|X}(y|x)$. We denote $K(\cdot)$ as a symmetric density function and h is the bandwidth, $h > 0$ and $K_h(x) = h^{-1}K(x/h)$. If $h \to 0$ and $n \to \infty$, we have

$$\begin{aligned}\boldsymbol{E}\big(K_h(Y-y)|X=x\big) \quad &= \quad \boldsymbol{E}\big(K_h(Y-y)|B^\tau X = B^\tau x\big)\\ &\longrightarrow \quad f_{Y|B^\tau X}(y|B^\tau x)\end{aligned} \tag{2.3}$$

Combining with (2.2) and (2.3), if X follows a normal distribution with identical covariance matrix, as $h \to 0$ and $n \to \infty$, we have

$$\begin{aligned}&\boldsymbol{E}\big(XK_h(Y-y)\big) = \boldsymbol{E}\Big[X\boldsymbol{E}\big(K_h(Y-y)|X\big)\Big]\\ &\longrightarrow \boldsymbol{E}\big(X f_{Y|B^\tau X}(y|B^\tau x)\big) = B\big[\partial f_{Y|B^\tau X}(y|B^\tau x)/\partial(B^\tau x)\big].\end{aligned} \tag{2.4}$$

Equation (2.4) indicates all the directions can be captured by the $\boldsymbol{E}\big[XK_h(Y-y)\big]$ under normality assumption when $h \to 0$ as $n \to \infty$. In fact, we avoid to estimate $f_{Y|X}(y|x)$, $\partial f_{Y|X}(y|x)/\partial x$ and $\boldsymbol{E}[\partial f_{Y|X}(y|X)/\partial X]$ by using non-parametric smoothing methods, while a simple moment-based estimator of $\boldsymbol{E}\big(XK_h(Y-y)\big)$ helps to infer the central subspace $\boldsymbol{S}_{Y|X}$. However, the normality assumption for X is relatively restrictive. Without loss of generality, we assume $\boldsymbol{E}X = 0$ and relax the normality assumption of X to the widely used linearity condition (Li, 1991) in the following

$$\boldsymbol{E}\big[X|B^\tau X\big] = P_B^\tau(\Sigma_X)X, \tag{2.5}$$

where $P_B(\Sigma_X) = B(B^\tau\Sigma_X B)^{-1}B^\tau\Sigma_X$, $\Sigma_X = \mathrm{Var}(X)$. As a consequence, under the linearity condition of X, $\boldsymbol{E}\big[XK_h(Y-y)\big]$ still could be used to seek $\boldsymbol{S}_{Y|X}$ when $h \to 0$ and $n \to \infty$. Define

$$\begin{aligned} \boldsymbol{D}_h(y) &= \boldsymbol{E}\big[XK_h(Y-y)\big], &(2.6)\\ \boldsymbol{D}_B(y) &= P_B^\tau\boldsymbol{E}\big[Xf_{Y|B^\tau X}(y|B^\tau X)\big], &(2.7)\end{aligned}$$

First of all, we list some conditions for our asymptotic results.

(A1) The conditional density functions $f_{Y|X}(y|x)$ and $f_{Y|B^\tau X}(y|B^\tau x)$ have continuous and bounded second order derivatives with respect to x. Furthermore, $\boldsymbol{D}_B(y) < \infty$ for all possible values of y on the support of Y, and $\boldsymbol{D}_B < \infty$.

(A2) $\boldsymbol{E}X^\tau X < \infty$, $\boldsymbol{D}_h(y) < \infty$, $\boldsymbol{D}_h < \infty$ for all possible values of y on the support of Y as $h \to 0$.

(A3) The kernel function $K(\cdot)$ is symmetric about 0 and has a compact support. Moreover,

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2K(u) < \infty.$$

(A4) The density function $g(\cdot)$ of Y has bounded second order derivatives. Moreover, $\boldsymbol{E}(X^\tau X)^4 < \infty$, $\boldsymbol{E}\big[\boldsymbol{D}_B(Y)g(Y)\big] < \infty$.

Theorem 2.1. *If $\boldsymbol{E}X = 0$ and Σ_X is a positive definite matrix, moreover, X satisfies the linearity condition (2.5). Under the conditions (A1)-(A4), as $h \to 0$, $n \to \infty$, we have*

(1) $\Sigma_X^{-1}\boldsymbol{D}_h(y) \to \Sigma_X^{-1}\boldsymbol{D}_B(y) \subseteq \boldsymbol{S}_{Y|X}$.

(2) *Let $\boldsymbol{D}_h = \boldsymbol{E}\big[\boldsymbol{D}_h(\tilde{Y})\boldsymbol{D}_h^\tau(\tilde{Y})\big]$, $\boldsymbol{D}_B = \boldsymbol{E}\big[\boldsymbol{D}_B(\tilde{Y})\boldsymbol{D}_B^\tau(\tilde{Y})\big]$, where $\tilde{Y}$ is an independent copy of Y. The kernel matrix $\boldsymbol{V}_h$, defined as $\boldsymbol{V}_h = \Sigma_X^{-1}\boldsymbol{D}_h\Sigma_X^{-1}$, satisfies $\boldsymbol{V}_h \to \boldsymbol{V}_B = \Sigma_X^{-1}\boldsymbol{D}_B\Sigma_X^{-1} \subseteq \boldsymbol{S}_{Y|X}$.*

Theorem 2.1 indicates that $\boldsymbol{V}_B$ estimates the central subspace $\boldsymbol{S}_{Y|X}$ in the population level. If we apply a spectral decomposition on the kernel matrix $\boldsymbol{V}_B$ to get $\{\beta_1, \ldots, \beta_k\}$, which are the eigenvectors of kernel matrix $\boldsymbol{V}_B$ corresponding to its largest k nonzero eigenvalues, then the space $S(\beta_1, \ldots, \beta_k)$ spanned by $\{\beta_1, \ldots, \beta_k\}$ will recover $\boldsymbol{S}_{Y|X}$. Theorem 2.1 entails a direct estimation approach by using the kernel matrix $\boldsymbol{V}_h$ to estimate $\boldsymbol{S}_{Y|X}$ when $h \to 0$ as $n \to \infty$. In next subsection, we introduce the estimation procedures of $\boldsymbol{D}_h(y)$ and $\boldsymbol{V}_h$ and present the asymptotic properties of these estimators.

2.2 Estimation Procedures and Asymptotic Results

Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d sample from the model (1.2). Denote $K_h(\cdot) = h^{-1}K(\cdot/h)$. The estimators of $\boldsymbol{D}_h(y)$ and Σ_X are defined as

$$\hat{\boldsymbol{D}}_h(y) = \tfrac{1}{n}\textstyle\sum_{i=1}^n (X_i - \bar{X})K_h(Y_i - y), \tag{2.8}$$

$$\hat{\Sigma}_X = \tfrac{1}{n}\textstyle\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\tau, \quad \bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i. \tag{2.9}$$

Using (2.8) and (2.9), the moment estimators of $\boldsymbol{D}_h$ and $\boldsymbol{V}_h$ in Theorem 2.1 are proposed as

$$\hat{\boldsymbol{D}}_h = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{D}}_h(Y_i)\hat{\boldsymbol{D}}_h^\tau(Y_i), \tag{2.10}$$

$$\hat{\boldsymbol{V}}_h = \hat{\Sigma}_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1}. \tag{2.11}$$

Theorem 2.2. *Under the conditions of Theorem 2.1, if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have*

$$\hat{\boldsymbol{V}}_h \xrightarrow{P} \boldsymbol{V}_B.$$

The consistency of $\hat{\boldsymbol{V}}_h$ to $\boldsymbol{V}_B$ enables us to estimate $\boldsymbol{S}_{Y|X}$ by using the first k eigenvectors of the estimated kernel matrix $\hat{\boldsymbol{V}}_h$ associated with its k largest nonzero eigenvalues.

Let $\mathrm{Vech}(A) = (a_{11}, \ldots, a_{p1}, a_{22}, \ldots, a_{p2}, a_{33}, \ldots, a_{pp})^\tau$ be a $p(p+1)/2$ dimension vector for any symmetric $p \times p$ dimensional matrix $A = (a_{ij})_{p\times p}$. In the following, we use notation $A^{\otimes 2} = AA^\tau$ for any matrix or vector A.

Define

$$\begin{aligned}\boldsymbol{T}_h(x, y) \;=\; & (x^{\otimes 2} - \Sigma_X)\Sigma_X^{-1}\boldsymbol{D}_B + \boldsymbol{D}_B\Sigma_X^{-1}(x^{\otimes 2} - \Sigma_X) \\ & + \boldsymbol{E}\big\{\boldsymbol{D}_B(y)g(y)\big\}x^\tau + x\boldsymbol{E}\big\{\boldsymbol{D}_B^\tau(y)g(y)\big\} \\ & + \boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0 | (X_1, Y_1) = (x, y)\big) - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0,\end{aligned} \tag{2.12}$$

where $\hat{\boldsymbol{D}}_h^0$ is defined as (A.2) in Appendix. Next, we present the asymptotic distribution of the estimated kernel matrix $\hat{\boldsymbol{V}}_h$.

Theorem 2.3. *Under the conditions of Theorem 2.1, $nh^4 \to 0$ and $nh^2 \to \infty$ as $n \to \infty$, moreover, $a^\tau \mathrm{Cov}\big(\mathrm{Vech}(\Sigma_X^{-1}\boldsymbol{T}_h(X,Y)\Sigma_X^{-1})\big)a \to \sigma_a^2 > 0$ for any $a \in \mathbb{R}^{\frac{p(p+1)}{2}}$ and $a \neq 0$, we have*

$$\sqrt{n}(\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B) \xrightarrow{L} \mathcal{H}. \tag{2.13}$$

where $a^\tau \mathrm{Vech}(\mathcal{H})$ follows a normal distribution $N(0, \sigma_a^2)$ for any $a \in \mathbb{R}^{\frac{p(p+1)}{2}}$ and $a \neq 0$.

Let $\lambda(A)$ stands for the vector of ordered eigenvalues of A, i.e, denoted as $\lambda(A) = (\lambda_1(A), \ldots, \lambda_p(A))^\tau$ satisfying $\lambda_1(A) \geq \ldots \geq \lambda_p(A)$. Denote $\lambda_1(\boldsymbol{V}_B) > \ldots > \lambda_l(\boldsymbol{V}_B)$ are the distinct eigenvalues of $\boldsymbol{V}_B$ with the multiplicity of $\lambda_i(\boldsymbol{V}_B)$ being m_i, $i = 1, \ldots, l$, and $m_1 + m_2 + \ldots + m_l = p$.

The orthogonal matrices Q makes

$$Q^\tau \boldsymbol{V}_B Q = \begin{bmatrix} \lambda_1(\boldsymbol{V}_B)I_{m_1} & 0 & \ldots & 0 \\ 0 & \lambda_2(\boldsymbol{V}_B)I_{m_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_l(\boldsymbol{V}_B)I_{m_l} \end{bmatrix},$$

where $\tilde{\mathcal{H}} = (\tilde{\mathcal{H}}_{i,j})$ denotes the partitioning of $Q^\tau \mathcal{H} Q$ in blocks of order $m_i \times m_j$. Theorem 2.3 entails to derive the asymptotic distribution of the eigenvalues of the estimated kernel matrix $\hat{\boldsymbol{V}}_h$. By application of Theorem 3.1 and Theorem 3.2 in Eaton and Tyler (1991), we have the following asymptotic result.

Theorem 2.4. *Under the conditions of Theorem 2.3, $\lambda_1(\boldsymbol{V}_B) > \lambda_2(\boldsymbol{V}_B) > \ldots > \lambda_l(\boldsymbol{V}_B)$ is the distinct eigenvalues of $\boldsymbol{V}_B$ with the multiplicity of $\lambda_i(\boldsymbol{V}_B)$ being m_i, $i = 1, \ldots, l$ and $m_1 + m_2 + \ldots + m_l = p$, we have*

$$\sqrt{n}\big(\lambda(\hat{\boldsymbol{V}}_h) - \lambda(\boldsymbol{V}_B)\big) \xrightarrow{L} \begin{bmatrix} \lambda(\tilde{\mathcal{H}}_{11}) \\ \lambda(\tilde{\mathcal{H}}_{22}) \\ \vdots \\ \lambda(\tilde{\mathcal{H}}_{ll}) \end{bmatrix}.$$

From Theorem 2.4, if the first kth of the kernel matrix $\boldsymbol{V}_B$ are nonzero, the first kth nonzero eigenvalues $\{\hat{b}_1, \ldots, \hat{b}_k\}$ of the estimated kernel matrix $\hat{\boldsymbol{V}}_h$ are root-n consistent, and the first dth eigenvectors with $d \leq k$ corresponding to nonzero eigenvalues helps to infer the central subspace $\boldsymbol{S}_{Y|X}$.

Theorem 2.5. *Let e be any unit length vector which is orthogonal to $\boldsymbol{S}_{Y|X}$, and suppose that $e^\tau \mathrm{Cov}\big(\Sigma_X^{-1}\boldsymbol{T}_h(X,Y)\Sigma_X^{-1}b_j\big)e \to e^\tau W_j e > 0$ as $h \to 0$, $j = 1,\ldots,k$. Under the conditions of Theorem 2.3, we have*

$$\sqrt{n}e^\tau \hat{b}_j \xrightarrow{L} N(0, e^\tau W_j e).$$

3 Simulations Studies

In this section, we conduct some simulations to evaluate the performance of our proposed method. To evaluate the estimation accuracy, we use a measure between two subspace of $\mathbb{R}^p$. Let B is a $p\times k$ matrix spanning $\boldsymbol{S}_{Y|X}$, and $\hat{B}$ is a $p \times k$ matrix to estimate B. The measure between $\boldsymbol{S}_{Y|X}$ and its estimator $\widehat{\boldsymbol{S}}_{Y|X}$ is defined as

$$\mathrm{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X}) = \|P_B - P_{\hat{B}}\|,$$

where P_B and $P_{\hat{B}}$ is the projection operator in the standard inner product of B and $\hat{B}$, and $\|\cdot\|$ is the Euclidean matrix norm. The smaller value of $\mathrm{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X})$, the better performance of $\widehat{\boldsymbol{S}}_{Y|X}$. A more detailed discussion of the distance measure can be referred to Li et al. (2005).

We compare our proposed method with some useful methods, SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), pHd (Li, 1992), DR (Li and Wang, 2007), rMAVE (Xia et al., 2002), dMAVE (Xia, 2007). For SIR, SAVE, and DR, we consider two cases with the slice number $H = 5$ and $H = 10$. For our method, we use the kernel function $K(t) = (15/16)(1-t^2)^2 I(t^2 < 1)$ and bandwidth $h = n^{-1/3}$ to satisfy the conditions in Theorems. The dimension of predictor X is choosen as $p = 5, 10, 20$ for Example 1 and Example 2.

Example 1 The following three models are used:

$$Y = \frac{1}{2}\exp(X_1 - 1)\varepsilon; \tag{3.1}$$

$$Y = \sin\left\{\frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 + \varepsilon\right\}; \tag{3.2}$$

$$Y = \cos\left\{\exp(\frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 + \varepsilon)\right\}; \tag{3.3}$$

In these models, the structure dimension of $\boldsymbol{S}_{Y|X}$ is one. The variables X_{ij} independently follows $t(8)$ ($t-$distribution with 8 freedom-degree) for $i = 1,\ldots,n$, $j = 1,\ldots,p$, and the error $\varepsilon \sim N(0,1)$ satisfying $\varepsilon \perp\!\!\!\perp X$. Among these dimension reduction methods, we present estimation error for each model with 300 replications. In Table 1, we present the mean of distance

measure $\text{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X}) = \|P_B - P_{\hat{B}}\|$ between these methods and the true central subspace and also the standard error.

From Table 1, we see that our method outperforms SAVE, pHd and r-MAVE in this three models. As the numbers of predictor p increases, our method still could detect the underlying true dimension reduction subspace and has better performance than SAVE and rMAVR. Both SIR with slice number $H = 5$, $H = 10$ and dMAVE have comparable performance with our method, and the latter has a slightly better performance than DR in the three models.

Example 2 The following two models are used:

$$Y = \text{sign}(\beta_1^\tau X + \varepsilon_1) \times \log(|6 + \beta_2^\tau X + \varepsilon_2|), \tag{3.4}$$

$$Y = \text{sign}(2\beta_1^\tau X + \varepsilon_1) \times \log(|4 + 2\beta_2^\tau X + \varepsilon_2|), \tag{3.5}$$

where $\text{sign}(\cdot)$ is the sign function. In this example, it is seenthat that $\boldsymbol{S}_{Y|X}$ is spanned by (β_1, β_2).

For model (3.4), the first five elements of β_1 are $(1, 1, 1, 1, 0)/2$ and the rest are all zeros. The first five elements of β_2 are $(-1, 2, -2, 1, 1)/\sqrt{11}$ and the rest are all zeros. The distribution of X is $N(0, I_p)$, where I_p is a $p \times p$ identical matrix, and unobservable noise ε_1 and ε_2 follow from $N(0, 0.5^2)$. Moreover, X is independent with $(\varepsilon_1, \varepsilon_2)$, and ε_1 is independent with ε_2.

For model (3.5), the first five elements of β_1 are $(1, 1, 1, 1, 0)/2$ and the rest are all zeros. The first five elements of β_2 are $(1, -1, 1, -1, 0)/2$, and the rest are all zeros too. The distribution of X is also $N(0, I_p)$ and independent with $(\varepsilon_1, \varepsilon_2)$, and unobservable noise ε_1 and ε_2 are independent and both are $N(0, 1)$. The models investigated here are similar to Chen and Li (1998) and Xia (2007). The model (3.5) is the same as the model in Example 4.1 in Xia (2007). We consider the models (3.4)-(3.5) with different dimension reduction methods with sample size $n = 300$ in this example. The slice numbers of SIR, SAVE and DR are also choosen as $H = 5$ and $H = 10$. The simulation results of mean and standard error with 300 replications are reported in Table 2.

From Table 2, the mean of $\text{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X})$ by our method is better than SAVE, pHd, rMAVE. As the numbers of predictor p increase, our method still has comparable performance with SIR and DR with the slice number $H = 5$, $H = 10$. For model (3.5), when we change different parameter values of (β_1, β_2) and reduce the variance of $(\varepsilon_1, \varepsilon_2)$, the performance of our method is better than dMAVE. For model (3.4), dMAVE could not detect the underlying dimension reduction subspace when the number of predictors is relatively small $p = 5$, while our method has stable performance in both cases even when the number of predictors p increases to 20.

Table 1 *The mean ± standard error (sd) of* $\text{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X})$ *for Example 1.*

		SIR($H=5,10$)	SAVE($H=5,10$)	DR($H=5,10$)	pHd	rMAVE	dMAVE	EPPK
			$p=5$, $n=300$					
Model (3.1)	mean	0.14 0.12	0.26 0.25	019 0.14	0.77	0.57	0.13	0.13
	sd	0.05 0.04	0.10 0.13	0.06 0.06	0.19	0.24	0.05	0.05
Model (3.2)	mean	0.16 0.19	0.51 0.65	0.29 0.34	0.90	0.23	0.16	0.18
	sd	0.06 0.06	0.25 0.24	0.14 0.16	0.15	0.14	0.05	0.07
Model (3.3)	mean	0.17 0.17	0.45 0.60	0.23 0.24	0.79	0.34	0.19	0.19
	sd	0.07 0.06	0.22 0.23	0.09 0.09	0.21	0.21	0.07	0.07
			$p=10$, $n=300$					
Model (3.1)	mean	0.22 0.19	0.63 0.80	0.28 0.25	0.84	0.72	0.19	0.21
	sd	0.05 0.05	0.25 0.19	0.08 0.06	0.12	0.17	0.04	0.05
Model (3.2)	mean	0.29 0.28	0.85 0.93	0.50 0.56	0.96	0.32	0.25	0.30
	sd	0.07 0.07	0.17 0.09	0.18 0.19	0.06	0.13	0.06	0.08
Model (3.3)	mean	0.26 0.27	0.89 0.95	0.36 0.37	0.89	0.44	0.28	0.28
	sd	0.08 0.07	0.11 0.07	0.12 0.11	0.11	0.18	0.07	0.07
			$p=20$, $n=300$					
Model (3.1)	mean	0.32 0.28	0.97 0.99	0.43 0.36	0.89	0.86	0.28	0.30
	sd	0.06 0.05	0.13 0.02	0.09 0.06	0.08	0.13	0.05	0.05
Model (3.2)	mean	0.40 0.41	0.97 0.98	0.69 0.73	0.98	0.44	0.37	0.43
	sd	0.07 0.08	0.04 0.02	0.17 0.16	0.03	0.14	0.06	0.07
Model (3.3)	mean	0.39 0.38	0.97 0.99	0.57 0.57	0.97	0.59	0.42	0.41
	sd	0.06 0.06	0.04 0.02	0.15 0.14	0.04	0.16	0.09	0.07

Table 2 *The mean ± standard error (sd) of* $\text{dist}(\boldsymbol{S}_{Y|X}, \widehat{\boldsymbol{S}}_{Y|X})$ *for Example 2.*

		SIR($H=5,10$)	SAVE($H=5,10$)	DR($H=5,10$)	pHd	rMAVE	dMAVE	EPPK
			$p=5$, $n=300$					
Model (3.4)	mean	0.13 0.11	0.18 0.21	0.15 0.14	0.95	0.85	0.71	0.16
	sd	0.04 0.03	0.07 0.11	0.04 0.04	0.11	0.14	0.24	0.05
Model (3.5)	mean	0.14 0.14	0.20 0.32	0.15 0.17	0.99	0.46	0.12	0.19
	sd	0.05 0.05	0.08 0.16	0.05 0.06	0.02	0.22	0.04	0.06
			$p=10$, $n=300$					
Model (3.4)	mean	0.19 0.18	0.42 0.76	0.23 0.22	0.99	0.93	0.91	0.24
	sd	0.03 0.04	0.17 0.20	0.04 0.05	0.02	0.09	0.11	0.05
Model (3.5)	mean	0.24 0.22	0.58 0.87	0.28 0.27	0.99	0.67	0.20	0.30
	sd	0.05 0.05	0.20 0.15	0.06 0.06	0.02	0.20	0.04	0.06
			$p=20$, $n=300$					
Model (3.4)	mean	0.30 0.27	0.93 0.99	0.36 0.34	0.99	0.97	0.97	0.36
	sd	0.05 0.04	0.10 0.01	0.06 0.05	0.01	0.04	0.04	0.05
Model (3.5)	mean	0.34 0.34	0.96 0.99	0.40 0.41	0.99	0.79	0.29	0.45
	sd	0.05 0.05	0.06 0.02	0.06 0.07	0.01	0.15	0.05	0.08

4 Real data analysis: Cars data

In this section, we apply our dimension reduction method to the Cars data. This data is about the 1983 ASA Data Exposition of Statistical Graphics

Technology. There are 406 observations with eight variables: Y-miles per gallon, X_1-number of cylinders, X_2-engine displacement (cu. inches), X_3-horsepower, X_4-vehicle weight (lbs.), X_5-time to accelerate from 0 to 60 mph (sec.), X_6-model year (modulo 100), and origin of car (1=American, 2=European, 3=Japanese). We transform the variable origin of car to the pairwise variable (X_7, X_8), where $(X_7, X_8) = (1,0), (0,1), (0,0)$ corresponding to American cars, European cars and Japanese cars. Before applying our dimension reduction method, we standardize all the covariates separately. The bandwidth is taken as three cases, $h = n^{-1/3}$, $h = n^{-1/2}$ and $h = n^{-4/5}$.

Denote the estimated central subspace by our method as $\widehat{\boldsymbol{S}}_{Y|X}$ in this dataset. To determine the number of dimension d, we use bootstrap to select the dimension based on Ye and Weiss (2003), and BIC type criterion proposed by Zhu et al. (2006). We re-sample $n = 300$ dataset without replacement and the standardized covariates, and we obtained a number of 500 bootstrap estimates $\widehat{\boldsymbol{S}}^b_{Y|X}$, $b = 1, \ldots, 500$ based on our method. To measure the distance between the data estimator $\widehat{\boldsymbol{S}}_{Y|X}$ and bootstrap estimator $\widehat{\boldsymbol{S}}^b_{Y|X}$, we adopt the vector correction coefficient q (Hotelling, 1936) and use $\arccos(q)$ as a measure, see more details in Ye and Weiss (2003). The mean of bootstrap distance measure $\arccos(\hat{q}^b)$ is reported in Table 3.

Table 3 *Bootstrap mean of* $\arccos(\hat{q}^b)$ *between* $\widehat{\boldsymbol{S}}^b_{Y|X}$ *and* $\widehat{\boldsymbol{S}}_{Y|X}$.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
$h=n^{-1/3}$	21.9695	14.1481	20.3936	14.9362	15.5745	20.5963	30.1306
$h=n^{-1/2}$	22.8674	14.2178	21.6424	14.9306	15.4797	20.3341	28.8257
$h=n^{-4/5}$	24.8663	14.1931	21.8925	14.7187	15.6498	21.5143	27.2937

From Table 3, the case of $d = 2$ has the smallest value and then we suggest the dimension in this dataset is 2. The BIC type criterion proposed by Zhu et al. (2006) also suggests that the dimension is 2. Now we use our method to this dataset and use the bandwidth $h = n^{-1/2}$ here, the two directions are $\hat{\beta}_1 = (-0.11, 0.73, -0.59, 0.14, -0.15, -0.20, -0.09, -0.07)^\tau$ and $\hat{\beta}_2 = (0.01, -0.55, -0.43, 0.68, -0.13, -0.15, 0.05, -0.08)^\tau$. The first two estimated direction indicate that the three predictors X_2-engine displacement (cu. inches), X_3-horsepower and X_4-vehicle weight (lbs.) are of dominating effects. The plots of Y against $\hat{\beta}_1^\tau X$ and $\hat{\beta}_2^\tau X$ are shown in Figure 1. Figure 1 presents the scatter plot with the response Y-miles per gallon and the first two estimated directions. In the left panel plot, the response Y has a linear trend in the first direction for American and Japanese cars. In the right panel plot, the response Y has a slight linear trend in the second direction

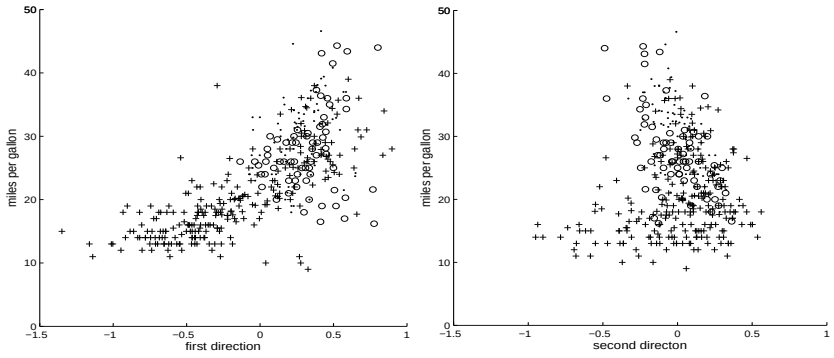

Figure 1 *Cars data. The left panel, the scatter plots of Y against the first direction. The right panel, the scatter plots of Y against the second direction. The origins of cars are denoted by "+" for American cars, " ∘ " for European cars and "·" for Japanese cars.*

for European and Japanese cars and has no trend for American cars.

5 Discussion and Further research

This paper gives a topic of the estimation procedure for dimension reduction by using the conditional density function. We present the asymptotic results of the proposed estimators and investigate the numerical performance. We can study the proposed methods in this paper to consider the estimation in the divergent parameters (Wu and Li, 2014; Zhu et al., 2006) and variable selection problems (Chen et al., 2010). One can also use the proposed methods in this paper to consider the measurement errors data (Li and Yin, 2007), longitudinal data (Bi and Qu, 2015; Li and Yin, 2009), missing data (Ding and Wang, 2011; Guo et al., 2014) in a future work. The research for this topic is ongoing.

Appendix A: Appendix section

In this section, we give the proofs of our main results. Without loss of generality, in the following we assumes $\boldsymbol{E}X = 0$.

A.1 Proof of Theorem 2.1

Proof. Invoking the assumption of that the conditional density function of $Y|X$ being the same as that of $Y|B^\tau X$, i.e, $f_{Y|X}(y|x) = f_{Y|B^\tau X}(y|B^\tau X)$,

when the linear condition (2.5) of X holds, we have

$$\begin{aligned}
&\boldsymbol{E}\big[X f_{Y|X}(y|X)\big] = \boldsymbol{E}\big[X f_{Y|B^\tau X}(y|B^\tau X)\big] \\
&= \boldsymbol{E}\big\{\boldsymbol{E}\big[X f_{Y|B^\tau X}(y|B^\tau X)|B^\tau X\big]\big\} = \boldsymbol{E}\big\{f_{Y|B^\tau X}(y|B^\tau X)\boldsymbol{E}\big[X|B^\tau X\big]\big\} \\
&= \boldsymbol{E}\big[f_{Y|B^\tau X}(y|B^\tau X)P_B^\tau X\big] = P_B^\tau \boldsymbol{E}\big[X f_{Y|B^\tau X}(y|B^\tau X)\big], \qquad \text{(A.1)}
\end{aligned}$$

where $P_B^\tau(\Sigma_X) = B(B^\tau\Sigma_X B)^{-1}B^\tau\Sigma_X$. From the expression (A.1), we see that $\Sigma_X^{-1}\boldsymbol{E}\big[X f_{Y|X}(y|X)\big] \subseteq \boldsymbol{S}_{Y|X}$. Fan et al. (1996) shows that $\boldsymbol{E}\big(K_h(Y-y)|X=x\big) \to f_{Y|X}(y|x)$ as $h\to 0$, $n\to\infty$. Using this results, we have

$$\begin{aligned}
\boldsymbol{D}_h(y) &= \boldsymbol{E}\big[XK_h(Y-y)\big] = \boldsymbol{E}\big\{X\boldsymbol{E}\big[K_h(Y-y)|X\big]\big\} \\
&\to \boldsymbol{E}\big\{X f_{Y|X}(y|X)\big\} = \boldsymbol{E}\big[X f_{Y|B^\tau X}(y|B^\tau X)\big] \\
&= P_B^\tau \boldsymbol{E}\big[X f_{Y|B^\tau X}(y|B^\tau X)\big] = \boldsymbol{D}_B(y),
\end{aligned}$$

i.e, $\Sigma_X^{-1}\boldsymbol{D}_h(y) \to \Sigma_X^{-1}\boldsymbol{D}_B(y) \subseteq \boldsymbol{S}_{Y|X}$ as $h\to 0$, $n\to\infty$. If we take $\tilde{Y}$ as an independent copy of Y, then we obtain that the kernel matrix $\boldsymbol{V}_h = \Sigma_X^{-1}\boldsymbol{D}_h\Sigma_X^{-1} = \Sigma_X^{-1}\boldsymbol{E}\big[\boldsymbol{D}_h^{\otimes 2}(\tilde{Y})\big]\Sigma_X^{-1} \to \Sigma_X^{-1}\boldsymbol{E}\big[\boldsymbol{D}_B^{\otimes 2}(\tilde{Y})\big]\Sigma_X^{-1} = \boldsymbol{V}_B \subseteq \boldsymbol{S}_{Y|X}$. So that $\boldsymbol{V}_h$ will seek out the central subspace $\boldsymbol{S}_{Y|X}$ as $h\to 0$, $n\to\infty$. □

A.2 Proof of Theorems 2.2

Proof. **Step 2.1** In this step, we prove the consistency of the estimator $\hat{\boldsymbol{D}}_h$.

$$\begin{aligned}
\hat{\boldsymbol{D}}_h &= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n (X_i-\bar{X})(X_j-\bar{X})^\tau K_h(Y_i-Y_s)K_h(Y_j-Y_s) \\
&= \frac{6a_n}{n(n-1)(n-2)}\sum_{1\le i<j<s\le n}^n u_h\big((X_i,Y_i),(X_j,Y_j),(X_s,Y_s)\big) + R_n \\
&= a_n\hat{\boldsymbol{D}}_h^0 + R_n \qquad \text{(A.2)}
\end{aligned}$$

where $a_n = \frac{(n-1)(n-2)}{n^2}$ and

$$\begin{aligned}
&u_h\big((X_i,Y_i),(X_j,Y_j),(X_s,Y_s)\big) \\
&= \frac{1}{6}\big[(X_iX_j^\tau + X_jX_i^\tau)K_h(Y_i-Y_s)K_h(Y_j-Y_s) \\
&\quad + (X_iX_s^\tau + X_sX_i^\tau)K_h(Y_i-Y_j)K_h(Y_s-Y_j) \\
&\quad + (X_jX_s^\tau + X_sX_j^\tau)K_h(Y_j-Y_i)K_h(Y_s-Y_i)\big]. \qquad \text{(A.3)}
\end{aligned}$$

From expressions (A.2) and (A.3), we see that $\hat{\boldsymbol{D}}_h^0$ is a $U-$statistics with symmetric kernel $u_h(\cdot)$. Furthermore, as $h \to 0$, $n \to \infty$,

$$
\begin{aligned}
&\boldsymbol{E}u_h\big((X_1,Y_1),(X_2,Y_2),(X_3,Y_3)\big)\\
&= \boldsymbol{E}\big\{\boldsymbol{E}u_h\big((X_1,Y_1),(X_2,Y_2),(X_3,Y_3)\big)\big|(X_3,Y_3)\big\}\\
&= \boldsymbol{E}[\boldsymbol{D}_h^{\otimes 2}(Y_3)] = \boldsymbol{E}[\boldsymbol{D}_h^{\otimes 2}(\tilde{Y})]\\
&\to \boldsymbol{E}_{\tilde{Y}}\big\{\boldsymbol{E}_X\big[X f_{Y|B^\tau X}(\tilde{Y}|B^\tau X)\big]\boldsymbol{E}_X\big[X^\tau f_{Y|B^\tau X}(\tilde{Y}|B^\tau X)\big]|\tilde{Y}\big\}\\
&= P_B^\tau \boldsymbol{E}_{\tilde{Y}}\big\{\boldsymbol{E}_X\big[X f_{Y|B^\tau X}(\tilde{Y}|B^\tau X)\big]\boldsymbol{E}_X\big[X^\tau f_{Y|B^\tau X}(\tilde{Y}|B^\tau X)\big]|\tilde{Y}\big\}P_B\\
&= \boldsymbol{E}[\boldsymbol{D}_B^{\otimes 2}(\tilde{Y})] = \boldsymbol{D}_B,
\end{aligned}
$$

where $\tilde{Y}$ is an independent copy of Y, $P_B = B(B^\tau\Sigma_X B)^{-1}B^\tau\Sigma_X$, and $\boldsymbol{E}_X(\cdot)$ stands for taking expectation about X and $\boldsymbol{E}_{\tilde{Y}}(\cdot)$ stands for taking expectation about $\tilde{Y}$. Since $\boldsymbol{E}u_h\big((X_1,Y_1),(X_2,Y_2),(X_3,Y_3)\big) \to \boldsymbol{D}_B$, thus, $\boldsymbol{E}\hat{\boldsymbol{D}}_h^0 \to \boldsymbol{D}_B$. By the convergence of the U-statistics, we have $\hat{\boldsymbol{D}}_h^0 \xrightarrow{P} \boldsymbol{D}_B$. Moreover, $a_n \to 1$, then $a_n\hat{\boldsymbol{D}}_h^0 \xrightarrow{P} \boldsymbol{D}_B$.

Next, we prove $R_n = o_P(1)$ in the following.

$$
\begin{aligned}
R_n &= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n\big[(\bar{X}^{\otimes 2} - X_i\bar{X}^\tau - \bar{X}X_j^\tau)K_h(Y_i-Y_s)K_h(Y_j-Y_s)\big]\\
&\quad + \frac{1}{n^3}\sum_{i=1}^n\sum_{j\neq i}\big[(X_i-\bar{X})(X_j-\bar{X})^\tau K_h(0)K_h(Y_j-Y_i)\big]\\
&\quad + \frac{1}{n^3}\sum_{j=1}^n\sum_{i\neq j}\big[(X_i-\bar{X})(X_j-\bar{X})^\tau K_h(0)K_h(Y_i-Y_j)\big]\\
&\quad + \frac{1}{n^3}\sum_{i=1}^n\sum_{s\neq i}\big[(X_i-\bar{X})^{\otimes 2}K_h^2(Y_i-Y_s)\big]\\
&\quad + \frac{1}{n^3}\sum_{i=1}^n\big[(X_i-\bar{X})^{\otimes 2}K_h^2(0)\big]\\
&= R_{1n}+R_{2n}+R_{3n}+R_{4n}+R_{5n}. \qquad \text{(A.4)}
\end{aligned}
$$

Note that $\boldsymbol{E}X = 0$, then $\sqrt{n}\bar{X} = O_P(1)$. Similar to proof of the consistency

of $\hat{\boldsymbol{D}}_h^0$, we have

$$\begin{aligned} R_{1n} &= \bar{X}^{\otimes 2}\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{n}K_h(Y_i-Y_s)K_h(Y_j-Y_s) \\ &\quad +\Big[\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{n}X_iK_h(Y_i-Y_s)K_h(Y_j-Y_s)\Big]\bar{X}^{\tau} \\ &\quad +\bar{X}\Big[\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{n}X_j^{\tau}K_h(Y_i-Y_s)K_h(Y_j-Y_s)\Big] \qquad \text{(A.5)} \\ &= O_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

It follows that

$$\begin{aligned} R_{2n} &= \frac{K(0)}{nh}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}(X_iX_j^{\tau}-X_i\bar{X}^{\tau}-\bar{X}X_j^{\tau}-\bar{X}^{\otimes 2})K_h(Y_i-Y_j)\Big] \\ &= O_P\left(\frac{1}{nh}\right). \qquad \text{(A.6)} \end{aligned}$$

The analysis of R_{3n}, R_{4n} are similar to R_{2n} and will be $O_P\left(\frac{1}{nh}\right)$. For R_{5n},

$$R_{5n} = \frac{K(0)^2}{(nh)^2}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^{\otimes 2} = O_P\left(\frac{1}{n^2h^2}\right). \qquad \text{(A.7)}$$

Based on (A.5)-(A.7), provided $h\to 0$ and $nh\to\infty$ as $n\to\infty$, $R_n = o_P(1)$.

Step 2.2 Note that $\hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^{\otimes 2} \xrightarrow{P} \Sigma_X$, then $\hat{\Sigma}_X - \Sigma_X = o_P(1)$. We have proved that $\hat{\boldsymbol{D}}_h - \boldsymbol{D}_B = o_P(1)$, then

$$\begin{aligned} \hat{\boldsymbol{V}}_h - \boldsymbol{V}_B &= \hat{\Sigma}_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1} - \Sigma_X^{-1}\boldsymbol{D}_B\Sigma_X^{-1} \qquad \text{(A.8)} \\ &= \hat{\Sigma}_X^{-1}(\Sigma_X-\hat{\Sigma}_X)\Sigma_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1} + \Sigma_X^{-1}(\hat{\boldsymbol{D}}_h-\boldsymbol{D}_B)\hat{\Sigma}_X^{-1} \\ &\quad +\Sigma_X^{-1}\boldsymbol{D}_B\hat{\Sigma}_X^{-1}(\Sigma_X-\hat{\Sigma}_X)\Sigma_X^{-1} = o_P(1), \end{aligned}$$

which indicates that $\hat{\boldsymbol{V}}_h \xrightarrow{P} \boldsymbol{V}_B \subseteq \boldsymbol{S}_{Y|X}$, and we have completed the proof of Theorem 2.2. □

A.3 Proof of Theorem 2.3

Proof. In this section, we drive the asymptotic distribution of $\hat{\boldsymbol{V}}_h$. Based on (A.8),

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B)\\
&= \sqrt{n}(\hat{\Sigma}_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1} - \Sigma_X^{-1}\boldsymbol{D}_B\Sigma_X^{-1})\\
&= \hat{\Sigma}_X^{-1}\sqrt{n}(\Sigma_X - \hat{\Sigma}_X)\Sigma_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1} + \Sigma_X^{-1}\sqrt{n}(\hat{\boldsymbol{D}}_h - \boldsymbol{D}_B)\hat{\Sigma}_X^{-1}\\
&\quad + \Sigma_X^{-1}\boldsymbol{D}_B\hat{\Sigma}_X^{-1}\sqrt{n}(\Sigma_X - \hat{\Sigma}_X)\Sigma_X^{-1}.
\end{aligned}$$

First, we derive $\sqrt{n}(\hat{\boldsymbol{D}}_h - \boldsymbol{D}_B)$ as a sum of i.i.d random variable and an asymptotic negligible part. According to (A.2),(A.4) and (A.6)-(A.7), if $\sqrt{n}h^2 \to 0$ and $\sqrt{n}h \to \infty$, we have

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{D}}_h - \boldsymbol{D}_B) \qquad\qquad (A.9)\\
=\ & \sqrt{n}(a_n\hat{\boldsymbol{D}}_h^0 + R_{1n} - \boldsymbol{D}_B) + \sqrt{n}(R_{2n} + R_{3n} + R_{4n} + R_{5n})\\
=\ & \sqrt{n}\big(a_n(\hat{\boldsymbol{D}}_h^0 - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0) + R_{1n}\big) + \sqrt{n}(a_n\boldsymbol{E}\hat{\boldsymbol{D}}_h^0 - \boldsymbol{D}_B) + O_P(\frac{1}{\sqrt{n}h})\\
=\ & \sqrt{n}\big(a_n(\hat{\boldsymbol{D}}_h^0 - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0) + R_{1n}\big) + O(a_n\sqrt{n}h^2 + \sqrt{n}(a_n - 1)) + O_P(\frac{1}{\sqrt{n}h})\\
=\ & a_n\sqrt{n}(\hat{\boldsymbol{D}}_h^0 - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0) + \sqrt{n}R_{1n} + o_P(1).
\end{aligned}$$

Note that $\hat{\boldsymbol{D}}_h^0$ is a U-statistic, and by Hajek projection (Serfling, 1980), we have $\sqrt{n}(\hat{\boldsymbol{D}}_h^0 - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\big[\boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0|(X_i,Y_i)\big) - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0\big] + o_P(1)$, which implies

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{D}}_h - \boldsymbol{D}_B)\\
=\ & \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big[\boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0|(X_i,Y_i)\big) - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0\Big] + O_P\big(\sqrt{n}(a_n - 1)\big) + o_P(1)\\
=\ & \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big[\boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0|(X_i,Y_i)\big) - \boldsymbol{E}\hat{\boldsymbol{D}}_h^0\Big] + o_P(1). \qquad (A.10)
\end{aligned}$$

We calculate the second part $\sqrt{n}R_{1n}$ in (A.5). Based on (A.5),

$$\begin{aligned}
\sqrt{n}R_{1n} \ =\ & \Big[\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n X_iK_h(Y_i - Y_s)K_h(Y_j - Y_s)\Big]\sqrt{n}\bar{X}^\tau\\
&+\sqrt{n}\bar{X}\Big[\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n X_j^\tau K_h(Y_i - Y_s)K_h(Y_j - Y_s)\Big] + O_P(\frac{1}{\sqrt{n}}).
\end{aligned}$$

The analysis of $\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^n X_iK_h(Y_i-Y_s)K_h(Y_j-Y_s)$ is similar to $\hat{\boldsymbol{D}}_h$ and is expressed as a sum of U-statistic and a negligible part. Note that

$$\begin{aligned}&\boldsymbol{E}\left[X_1K_h(Y_1-Y_3)K_h(Y_2-Y_3)\right]\\&=P_B^\tau\boldsymbol{E}_{\tilde{Y}}\big\{\boldsymbol{E}_X\big[Xf_{Y|B^\tau X}(\tilde{Y}|B^\tau X)\big]g(\tilde{Y})\big\}+O(h^2)\\&=\boldsymbol{E}\big\{\boldsymbol{D}_B(\tilde{Y})g(\tilde{Y})\big\}+O(h^2)\end{aligned}$$

where $g(\cdot)$ is the density function of Y. Then, we have

$$\begin{aligned}\sqrt{n}R_{1n}\;&=\;\boldsymbol{E}\big\{\boldsymbol{D}_B(\tilde{Y})g(\tilde{Y})\big\}\sqrt{n}\bar{X}^\tau+\sqrt{n}\bar{X}\boldsymbol{E}\big\{\boldsymbol{D}_B^\tau(\tilde{Y})g(\tilde{Y})\big\}\\&\quad+o_P(1).\end{aligned}\tag{A.11}$$

Together with (A.10) and (A.11), we have

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol{D}}_h-\boldsymbol{D}_B)\\=\;&\frac{1}{\sqrt{n}}\sum_{i=1}^n\big[\boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0|(X_i,Y_i)\big)-\boldsymbol{E}\hat{\boldsymbol{D}}_h^0+\boldsymbol{E}\big\{\boldsymbol{D}_B(\tilde{Y})g(\tilde{Y})\big\}X_i^\tau\\&+X_i\boldsymbol{E}\big\{\boldsymbol{D}_B^\tau(\tilde{Y})g(\tilde{Y})\big\}\big]+o_P(1).\end{aligned}\tag{A.12}$$

Second, the $\sqrt{n}(\hat{\Sigma}_X-\Sigma_X)=\frac{1}{\sqrt{n}}\sum_{i=1}^n(X_i^{\otimes2}-\Sigma_X)+o_P(\frac{1}{\sqrt{n}})$. Note that $\hat{\boldsymbol{D}}_h\xrightarrow{P}\boldsymbol{D}_B$ and $\hat{\Sigma}_X\xrightarrow{P}\Sigma_X$, then

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{V}}_h-\boldsymbol{V}_B)\;&=\;\sqrt{n}(\hat{\Sigma}_X^{-1}\hat{\boldsymbol{D}}_h\hat{\Sigma}_X^{-1}-\Sigma_X^{-1}\boldsymbol{D}_B\Sigma_X^{-1})\\&=\;\frac{1}{\sqrt{n}}\sum_{i=1}^n\Sigma_X^{-1}\boldsymbol{T}_h(X_i,Y_i)\Sigma_X^{-1}+o_P(1),\end{aligned}$$

where

$$\begin{aligned}&\boldsymbol{T}_h(X_i,Y_i)\\&=(X_i^{\otimes2}-\Sigma_X)\Sigma_X^{-1}\boldsymbol{D}_B+\boldsymbol{D}_B\Sigma_X^{-1}(X_i^{\otimes2}-\Sigma_X)\\&+\boldsymbol{E}\big(\hat{\boldsymbol{D}}_h^0|(X_i,Y_i)\big)-\boldsymbol{E}\hat{\boldsymbol{D}}_h^0+\boldsymbol{E}\big\{\boldsymbol{D}_B(\tilde{Y})g(\tilde{Y})\big\}X_i^\tau+X_i\boldsymbol{E}\big\{\boldsymbol{D}_B^\tau(\tilde{Y})g(\tilde{Y})\big\}.\end{aligned}\tag{A.13}$$

Note that $\{\boldsymbol{T}_h(X_i,Y_i),i=1,\ldots,n\}$ are i.i.d random variables. For any $a\in\mathbb{R}^{\frac{p(p+1)}{2}}$ and $a\neq0$, if $a^\tau\mathrm{Cov}\big(\mathrm{vech}(\Sigma_X^{-1}\boldsymbol{T}_h(X_1,Y_1)\Sigma_X^{-1})\big)a\to\sigma_a^2>0$ as $h\to0$, $n\to\infty$, the CLT theorem entails that $\sqrt{n}a^\tau\big(\mathrm{Vech}(\Sigma_X^{-1}\boldsymbol{T}_h(X_i,Y_i)\Sigma_X^{-1})-\mathrm{Vech}(\Sigma_X^{-1}\boldsymbol{E}\boldsymbol{T}_h(X_i,Y_i)\Sigma_X^{-1})\big)\xrightarrow{L}N(0,\sigma_a^2)$, and we write it as

$$\sqrt{n}(\hat{\boldsymbol{V}}_h-\boldsymbol{V}_B)\xrightarrow{L}\mathcal{H}.\tag{A.14}$$

We have completed the proof of Theorem 2.3. □

A.4 Proof of Theorem 2.4

Proof. In this section, we get the asymptotic distribution of the nonzero eigenvalues of $\hat{\boldsymbol{V}}_h$. Denote $\lambda_1 > \lambda_2 > \ldots > \lambda_l$ is the distinct eigenvalues of positive semi-definite matrix $\boldsymbol{V}_B$ with the multiplicity of λ_i being m_i, $i = 1, \ldots, l$ and $m_1 + m_2 + \ldots + m_l = p$. There exists orthogonal matrices Q such that

$$Q^\tau \boldsymbol{V}_B Q = \begin{bmatrix} \lambda_1 I_{m_1} & 0 & \ldots & 0 \\ 0 & \lambda_2 I_{m_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_l I_{m_l} \end{bmatrix}$$

From the results (A.14), we have $\sqrt{n}(Q^\tau \hat{\boldsymbol{V}}_h Q - Q^\tau \boldsymbol{V}_B Q) \xrightarrow{L} Q^\tau \mathcal{H} Q$, and we partition $Q^\tau \hat{\boldsymbol{V}}_h Q$ in the similar way of $Q^\tau \boldsymbol{V}_B Q$ as

$$Q^\tau \hat{\boldsymbol{V}}_h Q = \begin{bmatrix} V_{n,11} & V_{n,12} & \ldots & V_{n,1l} \\ V_{n,21} & V_{n,22} & \ldots & V_{n,2l} \\ \vdots & \vdots & \ddots & \vdots \\ V_{n,l1} & V_{n,l2} & \ldots & V_{n,ll} \end{bmatrix}$$

We define that $\lambda(A)$ stands for the vector of ordered eigenvalues of A, i.e, denoted as $\lambda(A) = (\lambda_1(A), \lambda_2(A), \ldots, \lambda_k(A))^\tau$, and let $1_{m_i} \in \mathbb{R}^{m_i}$ stand for the vector of ones for $i = 1, \ldots, l$. Applying the result of Theorem 3.1. in Eaton and Tyler (1991), we have

$$\sqrt{n}(\lambda(Q^\tau \hat{\boldsymbol{V}}_h Q) - \lambda(Q^\tau \boldsymbol{V}_B Q) = \sqrt{n} \begin{bmatrix} \lambda(V_{n,11}) & - & \lambda_1 1_{m_1} \\ \lambda(V_{n,22}) & - & \lambda_2 1_{m_2} \\ & \vdots & \\ \lambda(V_{n,ll}) & - & \lambda_l 1_{m_l} \end{bmatrix} + o_P(1)$$

Since $\sqrt{n}(Q^\tau \hat{\boldsymbol{V}}_h Q - Q^\tau \boldsymbol{V}_B Q) \xrightarrow{L} Q^\tau \mathcal{H} Q$, we have

$$\sqrt{n} \begin{bmatrix} V_{n,11} & - & \lambda_1 I_{m_1} \\ V_{n,22} & - & \lambda_2 I_{m_2} \\ & \vdots & \\ V_{n,ll} & - & \lambda_l I_{m_l} \end{bmatrix} \xrightarrow{L} \tilde{\mathcal{H}} = \begin{bmatrix} \tilde{\mathcal{H}}_{11} \\ \tilde{\mathcal{H}}_{22} \\ \vdots \\ \tilde{\mathcal{H}}_{ll} \end{bmatrix},$$

where $\tilde{\mathcal{H}} = (\tilde{\mathcal{H}}_{i,j})$ is the partitioning of $Q^\tau \mathcal{H} Q$ in blocks of order $m_i \times m_j$. Moreover, the eigenvalue vector $\lambda(A)$ is a continuous function about matrix

A, and Q is an orthogonal matrix, then $\lambda(Q^\tau \hat{\boldsymbol{V}}_h Q) = \lambda(\hat{\boldsymbol{V}}_h)$, $\lambda(Q^\tau \boldsymbol{V}_B Q) = \lambda(\boldsymbol{V}_B)$. We have

$$\sqrt{n}\big(\lambda(\hat{\boldsymbol{V}}_h) - \lambda(\boldsymbol{V}_B)\big) \xrightarrow{L} \begin{bmatrix} \lambda(\tilde{\mathcal{H}}_{11}) \\ \lambda(\tilde{\mathcal{H}}_{22}) \\ \vdots \\ \lambda(\tilde{\mathcal{H}}_{ll}) \end{bmatrix}.$$

We have completed the proof of Theorem 2.4. □

A.5 Proof of Theorem 2.5

Proof. For any vector e is orthogonal to $\boldsymbol{S}_{Y|X}$, satisfying $e^\tau e = 1$. Note that $\hat{\boldsymbol{V}}_h \hat{b}_j = \hat{\lambda}_j \hat{b}_j$, $j = 1, \ldots, k$, then

$$\begin{aligned}
\sqrt{n} e^\tau \hat{b}_j &= \frac{\sqrt{n}}{\hat{\lambda}_j} e^\tau \hat{\boldsymbol{V}}_h \hat{b}_j \\
&= \frac{1}{\lambda_j} e^\tau \sqrt{n}\{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\} b_j + \frac{\lambda_j - \hat{\lambda}_j}{\lambda_j \hat{\lambda}_j} e^\tau \sqrt{n}\{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\} b_j \\
&\quad + \frac{1}{\lambda_j} e^\tau \sqrt{n}\{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\}\{\hat{b}_j - b_j\} + \frac{1}{\hat{\lambda}_j} \sqrt{n} e^\tau \boldsymbol{V}_B \hat{b}_j \\
&= I_{n,1} + I_{n,2} + I_{n,3} + I_{n,4}.
\end{aligned}$$

Using the proof of Theorem 2.3, we have

$$\sqrt{n}\{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_X^{-1} \boldsymbol{T}_h(X_i, Y_i) \Sigma_X^{-1} + o_P(1),$$

where $\boldsymbol{T}_h(X_i, Y_i)$ is defined in (A.13), then

$$I_{n,1} = \sqrt{n} e^\tau \{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\} b_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e^\tau \Sigma_X^{-1} \boldsymbol{T}_h(X_i, Y_i) \Sigma_X^{-1} b_j + o_P(1).$$

Here, $\{e^\tau \Sigma_X^{-1} \boldsymbol{T}_h(X_i, Y_i) \Sigma_X^{-1} b_j\}_{i=1}^n$ are i.i.d random variables, and if the limit of variance $e^\tau \mathrm{Cov}\big(\Sigma_X^{-1} \boldsymbol{T}_h(X_i, Y_i) \Sigma_X^{-1} b_j\big) e \to e^\tau W_j e > 0$ as $h \to 0$, $n \to \infty$. The CLT theorem entails that

$$I_{n,1} \xrightarrow{L} N(0, e^\tau W_j e). \tag{A.15}$$

Next, we show $I_{n,2}, I_{n,3}, I_{n,4}$ are $o_P(1)$. Theorem 2.4 entails that $\hat{\lambda}_j = \lambda_j + O_p(\frac{1}{\sqrt{n}})$, and using $I_{n,1} = O_P(1)$, we have

$$I_{n,2} = \frac{\lambda_j - \hat{\lambda}_j}{\lambda_j \hat{\lambda}_j} I_{n,1} = \frac{O_P(\frac{1}{\sqrt{n}})}{\lambda_j(\lambda_j + O_P(\frac{1}{\sqrt{n}}))} O_P(1) = o_P(1). \tag{A.16}$$

By the perturbation theory and $\hat{\boldsymbol{V}}_h = \boldsymbol{V}_B + o_P(1)$, we obtain $\hat{b}_j = b_j + o_P(1)$, and then

$$I_{n,3} = \frac{1}{\lambda_j} e^\tau \sqrt{n}\{\hat{\boldsymbol{V}}_h - \boldsymbol{V}_B\}\{\hat{b}_j - b_j\} = O_P(1)o_P(1) = o_P(1). \quad \text{(A.17)}$$

As e is orthogonal to $\boldsymbol{S}_{Y|X}$, and $\boldsymbol{V}_B \subseteq \boldsymbol{S}_{Y|X}$, then $e^\tau \boldsymbol{V}_B = 0$, and also

$$I_{n,4} = \frac{1}{\hat{\lambda}_j} \sqrt{n} e^\tau \boldsymbol{V}_B \hat{b}_j = o_P(1). \quad \text{(A.18)}$$

Together with the result (A.15)-(A.18), we have

$$\sqrt{n} e^\tau \hat{b}_j = I_1 + I_2 + I_3 + I_4 = I_1 + o_P(1) \xrightarrow{L} N(0, e^\tau W_j e).$$

We have completed the proof of Theorem 2.5. □

Acknowledgements

The authors thank the editor, the associate editor and a referee for their constructive suggestions that helped us to improve the early manuscript. The research work of Zhang Jun is supported by the National Natural Sciences Foundation of China (Grant No. 11401391).

References

Chen, C. H. and Li, K. C. (1998) Can SIR be as popular as multiple linear regression? *Statist. Sinica*, **8**, 289–316.

Cook, R. D. (1994) On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177–189.

Cook, R. D. (1998) *Regression Graphics. Ideas for studying regressions through graphics.* New York: Wiley.

Cook, R. D. and Weisberg, S. (1991) Discussion of "Sliced inverse regression for dimension reduction". by K.-C.Li. *J. Amer. Statist. Assoc.*, **86**, 328–332.

Eaton, M. and Tyler. D. (1991) On Wielandt's Inequality and Its Applications to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix. *Ann. Statist.* **19**, 260–271.

Fan, J., Yao, Q. and Tong, H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika.* **83**, 189–196.

Hotelling H. (1936) Relations between two sets of variates. *Biometrika.* **28**, 321–377.

Li, B., Zha, H. and Chiaromonte, F. (2005) Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33**, 1580–1616.

Li, B., Wang S. L. (2007) On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997–1008.

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

Li, K. C. (1992) On principal Hessian directions for data visuallization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025–1039.

Li, K. C. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.* **17**, 1009–1052.

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

Stein, C. (1981) Estimation the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968–979.

Samarov, A. M. (1993) Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88**, 836–847.

Zhu, L. X. Miao, B. Q. and Peng, H. (2006) Sliced inverse regression with large dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630–643.

Wang, T. and Zhu, L. X. (2013) Sparse sufficient dimension reduction using optimal scoring. *Comput. Statist. Data Anal.* **57**, 223–232.

Xia, Y. (2007) A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–2690.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction (with discussion). *J. R. Statist. Soc. B.* **64**, 363–410.

Zhu, X., Chen, F., Guo, X. and Zhu, L. X. (2016) Heteroscedasticity testing for regression models: a dimension reduction-based model adaptive approach. *Comput. Statist. Data Anal.* **103**, 263–283.

Zhu, X., Chen, F., Guo, X. and Zhu, L. X. (2016) An adaptive-to-model test for partially parametric single-index models. *Stat. Compu.*

Lansangan, Joseph Ryan G. and Barrios, Erniel B. (2017) Simultaneous dimension reduction and variable selection in modeling high dimensional data. *Comput. Statist. Data Anal.* **112**, 242–256.

Luo, Wei and Zhu, Yeying and Ghosh, Debashis. (2017) On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika* **104**, 51–65.

Yoshida, Takuma. (2017) Nonlinear surface regression with dimension reduction method. *AStA Adv. Stat. Anal.* **101**, 29–50.

Deng, Jianqiu and Wang, Qihua. (2017) Dimension reduction estimation for probability density with data missing at random when covariables are present. *J. Statist. Plann. Inference.* **181**, 11–29.

Sheng, Wenhui and Yin, Xiangrong. (2016) Sufficient dimension reduction via distance covariance. *J. Comput. Graph. Statist.* **25**, 91–104.

Zhou, Jingke and Zhu, Lixing. (2016) Principal minimax support vector machine for sufficient dimension reduction with contaminated data. *Comput. Statist. Data Anal.* **94**, 33–48.

Cui, Xia and Härdle, Wolfgang Karl and Zhu, Lixing. (2009) The EFM approach for single-index models. *Ann. Statist.* **39**, 1658–1688.

Peng, Heng and Huang, Tao. (2011) Penalized least squares for single index models. *J. Statist. Plann. Inference.* **141**, 1362–1379.

Chen, Xin and Zou, Changliang and Cook, Dennis R. (2010) Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696–3723.

Li, Gaorong and Peng, Heng and Dong, Kai and Tong, Tiejun. (2014) Simultaneous confidence bands and hypothesis testing for single-index models *Statist. Sinica*, **24**, 937–955.

Liang, H. and Liu, X. and Li, R. and Tsai, C. L. (2010) Estimation and testing for partially linear single-index models. *Ann. Statist.* **38**, 3811-3836.

Li, Bing and Yin, Xiangrong. (2007) On surrogate dimension reduction for measurement error regression: an invariance law. *Ann. Statist.* **35**, 2143–2172.

Wu, Yichao and Li, Lexin. (2011) Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist. Sinica*, **21**, 707–730.

Ding, Xiaobo and Wang, Qihua. (2011) Fusion-refinement procedure for dimension reduction with missing response at random. *J. Amer. Statist. Assoc.* **106**, 1193–1207.

Guo, Xu and Wang, Tao and Xu, Wangli and Zhu, Lixing. (2014) Dimension reduction with missing response at random. *Comput. Statist. Data Anal.* **69**, 228–242.

Bi, Xuan and Qu, Annie. (2015) Sufficient dimension reduction for longitudinal data. *Statist. Sinica*, **251**, 787–807.

Li, Lexin and Yin, Xiangrong. (2009) Longitudinal data analysis using sufficient dimension reduction method. *Comput. Statist. Data Anal.* **53**, 4106–4115.

J. Zhang
College of Mathematics and Statistics
Institute of Statistical Sciences
Shenzhen-Hong Kong Joint Research
Center for Applied Statistical Sciences
Shenzhen
China

B.He
School of Mathematics and Computational Science
Sun Yat-Sen University
Guangzhou
China
E-mail: mrhephd@163.com

T.Lu
Department of Mathematics and Statistics
University of Nevada
Reno, NV
USA

S.Wen
College of Mathematics and Statistics
Shenzhen University
Shenzhen
China