Biometrika (2015), **xx**, x, *pp*. 1–14

Printed in Great Britain

Diagnostic studies in sufficient dimension reduction

BY XIN CHEN

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546
stacx@nus.edu.sg

R. DENNIS COOK

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
dennis@stat.umn.edu

AND CHANGLIANG ZOU

Institute of Statistics, Nankai University, Tianjin 300071, China
nk.chlzou@gmail.com

SUMMARY

Sufficient dimension reduction in regression strives to reduce the predictor dimension by replacing the original predictors with some set of linear combinations without loss of information. Numerous dimension reduction methods have been invented based on this paradigm. However, little effort has been devoted to diagnostic studies within the context of dimension reduction. In this paper we introduce methods to check goodness-of-fit for a given dimension reduction subspace. The key idea is to extend the so-called distance correlation to measure the conditional dependence relationship between the covariates and response given a reduction subspace. Our methods require only minimal assumptions, which are usually much less restrictive than the conditions needed to justify the original methods themselves. Asymptotic properties of the test statistic are studied. Numerical examples demonstrate the effectiveness of the proposed approach.

Some key words: Asymptotic Normality; Central subspace; Conditional independence; Distance correlation; Kernel smoothing; Permutation reduction

1. INTRODUCTION

Let $X = (x_1, \ldots, x_p)^{\mathsf{T}} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. The general goal of a regression of Y on X is inference about the conditional distribution of Y given X. When the dimension of X is not small, it is usually desirable to reduce its dimensionality as a preliminary step in an analysis. Sufficient dimension reduction is important in both theory and practice (Cook, 1994, 1998). The basic idea is to replace the predictor vector with its projection onto a subspace of the predictor space without loss of information on the conditional distribution of Y given X. If a predictor subspace $\mathcal{S} \subseteq \mathbb{R}^p$ satisfies

$$Y \perp\!\!\!\perp X \mid \mathcal{P}_{\mathcal{S}} X, \tag{1}$$

where $\perp\!\!\!\perp$ stands for independence and $\mathcal{P}_{(\cdot)}$ represents the projection matrix with respect to the standard inner product, then $\mathcal{S}$ is called a dimension reduction subspace. The statement that Y is independent of X given $\mathcal{P}_{\mathcal{S}} X$ is equivalent to stipulating that $\mathcal{P}_{\mathcal{S}} X$ carries all of the regression information that X has about Y. The central subspace is defined as the intersection of all dimension reduction spaces, which is also a dimension reduction subspace under mild conditions (Cook, 1998).

Various dimension reduction methods have been developed, including sliced inverse regression (Li, 1991), sliced average variance estimation (Cook & Weisberg, 1991), principal Hessian directions (Li, 1992; Cook, 1998), minimum average variance estimator (Xia et al., 2002), contour regression (Li et al., 2005), inverse regression estimator (Cook & Ni, 2005), principal fitted components (Cook, 2007), directional regression (Li & Wang, 2007), likelihood acquired directions (Cook & Forzani, 2009), semiparametric dimension reduction methods (Ma & Zhu, 2012) and direction estimation via distance covariance (Sheng & Yin, 2013).

All current sufficient dimension reduction methods rely for their validity on various conditions that may or may not be satisfied in applications. For example, sliced inverse regression and other methods require the so-called linearity condition, which is essentially uncheckable in practice. A fundamental but unexplored question is how to measure the relative worth of dimension reduction methods in the analysis of a particular dataset. Given an estimate $\mathcal{S}$ of a dimension reduction space, we develop a technique for assessing the worth of $\mathcal{S}$ by quantifying deviations from the conditional dependence relationship (1). The ideal case is that Y and X are independent given $\mathcal{P}_{\mathcal{S}}X$. Typically it is not easy and not enough to use the conditional covariance to measure conditional dependence. Although some efforts have been made for model checking (Stute & Zhu, 2005; Xia, 2009), the challenges associated with designing a general approach to testing (1) are yet to be addressed well.

In this paper, we gauge the conditional dependence of Y and X given $\mathcal{P}_{\mathcal{S}}X$ by utilizing distance covariance (Székely et al., 2007; Székely & Rizzo, 2009), which has computationally straightforward empirical formulas. The asymptotic properties of our test statistic are derived accordingly, so we can then use it to compute an empirical p-value and to test the sufficiency of any dimension reduction subspace. Simulation studies and two real data analysis examples demonstrate the value of our approach.

Distance covariance was introduced into sufficient dimension reduction by Sheng & Yin (2013). Their methodology seeks directions $\eta \in \mathbb{R}^{p\times d}$ that maximizes the marginal distance covariance between Y and $\eta^{\mathsf{T}}X$. Their goal was to estimate the central subspace, while ours concerns diagnostic studies after dimension reduction.

2. THEORETICAL DEVELOPMENTS

2·1. *Preliminaries on distance covariance*

Székely et al. (2007) introduced distance covariance for measuring dependence between two random vectors. Let $\psi_X(t)$ and $\psi_Y(s)$ be the characteristic functions of the random vectors $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ with finite first moments, and let $\psi_{X,Y}(t,s)$ be the joint characteristic function of X and Y. The distance covariance between X and Y is defined as

$$\begin{aligned}\mathrm{dcov}^2(X,Y) &= \int_{\mathbb{R}^{d_X+d_Y}} \|\psi_{X,Y}(t,s) - \psi_X(t)\psi_Y(s)\|^2 \omega(t,s)dtds \\ &\equiv \|\psi_{X,Y}(t,s) - \psi_X(t)\psi_Y(s)\|_\omega^2,\end{aligned} \tag{2}$$

where $\omega(t,s)$ is a positive weight function. For a complex valued function ψ, $\|\psi\|^2 = \psi\bar{\psi}$ where $\bar{\psi}$ is the conjugate of ψ.

With a properly chosen function $\omega(t,s)$, Székely et al. (2007, Remark 3) stated that

$$\mathrm{dcov}^2(X,Y) = S_1 + S_2 - 2S_3,$$

where $S_1 = E(\|X_1 - X_2\|\|Y_1 - Y_2\|)$, $S_2 = E(\|X_1 - X_2\|)E(\|Y_1 - Y_2\|)$, $S_3 = E(\|X_1 - X_2\|\|Y_1 - Y_3\|)$, (X_i, Y_i) $(i = 1, 2, 3)$ are independent copies of (X, Y) and $\|\cdot\|$ denotes the Euclidean norm.

A property of the distance covariance is that $\mathrm{dcov}(X, Y) = 0$ if and only if X and Y are independent. A natural estimator of $\mathrm{dcov}^2(X, Y)$ is

$$\mathrm{dcov}_n^2(\mathcal{X}, \mathcal{Y}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3,$$

where quantities $(\mathcal{X}, \mathcal{Y}) = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denote the sample versions of their population counterparts and $\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$ are the usual moment estimators of S_1, S_2 and S_3. Székely et al. (2007) proposed to use $n\mathrm{dcov}_n^2(\mathcal{X}, \mathcal{Y})/\hat{S}_2$ as a test statistic for independence.

2·2. *Sufficiency test statistic*

Given a basis matrix $A \in \mathbb{R}^{p\times d}$, we propose to assess the conditional dependence between Y and X given $A^\mathsf{T}X$ by using the p-value from a test of the null hypothesis

$$H_0 : Y \perp\!\!\!\perp X \mid \mathcal{P}_{\mathcal{S}_A} X, \tag{3}$$

where $\mathcal{S}_A = \mathrm{span}(A)$. Extending the distance covariance to a corresponding conditional distance covariance will provide a tool for testing the conditional hypothesis (3).

Let $\psi^{(X)}_{A^\mathsf{T}X}(t \mid B)$ and $\psi^{(Y)}_{A^\mathsf{T}X}(s \mid B)$ be the characteristic functions of X and Y given $A^\mathsf{T}X = B$, and let $\psi^{(X,Y)}_{A^\mathsf{T}X}(t, s \mid B)$ be the joint characteristic function of X and Y given $A^\mathsf{T}X = B$. Using (2) as a guide, the conditional distance covariance between X and Y given $A^\mathsf{T}X = B$ can accordingly be defined as the square root of

$$V(A; B) = \|\psi^{(X,Y)}_{A^\mathsf{T}X}(t, s \mid B) - \psi^{(X)}_{A^\mathsf{T}X}(t \mid B)\psi^{(Y)}_{A^\mathsf{T}X}(s \mid B)\|^2_\omega. \tag{4}$$

Clearly, $V(A; B) = 0$ for all B if and only if X and Y are conditionally independent and hence under hypothesis (3), $V(A; B) = 0$ by (1). Following an argument of Székely et al. (2007, Remark 3), $V(A; B)$ can be written as

$$\begin{aligned} V(A; B) &= E(\|P_1 - P_2\||Q_1 - Q_2|) + E(\|P_1 - P_2\|)E(|Q_1 - Q_2|) - 2E(\|P_1 - P_2\||Q_1 - Q_3|) \\ &\equiv S_1(A; B) + S_2(A; B) - 2S_3(A; B), \end{aligned} \tag{5}$$

where (P_i, Q_i) $(i = 1, 2, 3)$ are independent and identically distributed as $\psi^{(X,Y)}_{A^\mathsf{T}X}(\cdot, \cdot \mid B)$, provided that $E(\|P_1\|^2) < \infty$, $E(Q_1^2) < \infty$ and $E(\|P_1\|Q_1|) < \infty$, which are assumed throughout this paper. The (P_i, Q_i)'s all depend on B, but we suppress this dependence for notational convenience. Similarly, the expectations in (5) are to be understood as conditional expectations given B. We next develop a sample version $V_n(A; B)$ of $V(A; B)$, which can be used to assess the sufficiency of A.

Define $V_n(A; B) = \|\psi^{(X,Y)}_{n,A^\mathsf{T}X}(t, s \mid B) - \psi^{(X)}_{n,A^\mathsf{T}X}(t \mid B)\psi^{(Y)}_{n,A^\mathsf{T}X}(s \mid B)\|^2_\omega$. It is often desirable to reduce the dimension of the predictors to no more than three, and a straightforward way to obtain these sample quantities is by using a kernel-type estimator. Any appropriate linear smoother should work well for this purpose. In an unpublished 2012 technical report, Póczos and Schneider derived a version of conditional distance covariance using the k-nearest neighbour method. They proved consistency, but the asymptotic distribution of their estimators is apparently unknown and no formal approach for testing conditional independence was offered.

Let $f_0(\cdot)$ be the density function of $A^\mathsf{T}X$. The kernel estimator of $f_0(B)$ is given by $\hat{f}_0(B) = n^{-1}\sum_{k=1}^n K_h(A^\mathsf{T}X_k - B)$, where $K_h(\cdot) = K(\cdot/h)/h^d$ denotes a d-dimensional kernel function. Then,

the kernel empirical characteristic functions can be defined as

$$\psi^{(X,Y)}_{nA^{\mathsf{T}}X}(t,s \mid B) = \frac{1}{n\hat{f}_0(B)} \sum_{k=1}^{n} K_h(A^{\mathsf{T}}X_k - B)\exp(it^{\mathsf{T}}X_k + isY_k),$$

$$\psi^{(X)}_{nA^{\mathsf{T}}X}(t \mid B) = \frac{1}{n\hat{f}_0(B)} \sum_{k=1}^{n} K_h(A^{\mathsf{T}}X_k - B)\exp(it^{\mathsf{T}}X_k),$$

$$\psi^{(Y)}_{nA^{\mathsf{T}}X}(t \mid B) = \frac{1}{n\hat{f}_0(B)} \sum_{k=1}^{n} K_h(A^{\mathsf{T}}X_k - B)\exp(isY_k).$$

By the arguments in the proof of Theorem 1 in Székely et al. (2007), the sample version of $V(A;B)$, $V_n(A;B)$, can be expressed as

$$V_n(A;B) = S_{n,1}(A;B) + S_{n,2}(A;B) - 2S_{n,3}(A;B),$$

where

$$S_{n,1}(A;B) = \frac{1}{n^2\hat{f}_0^2(B)} \sum_{k,l} K_h(A^{\mathsf{T}}X_k - B)K_h(A^{\mathsf{T}}X_l - B)\|X_k - X_l\||Y_k - Y_l|,$$

$$S_{n,2}(A;B) = \frac{1}{n^2\hat{f}_0^2(B)} \sum_{k,l} K_h(A^{\mathsf{T}}X_k - B)K_h(A^{\mathsf{T}}X_l - B)\|X_k - X_l\| \times \frac{1}{n^2\hat{f}_0^2(B)} \sum_{k,l} K_h(A^{\mathsf{T}}X_k - B)K_h(A^{\mathsf{T}}X_l - B)|Y_k - Y_l|,$$

$$S_{n,3}(A;B) = \frac{1}{n^3\hat{f}_0^3(B)} \sum_{k,l,m} K_h(A^{\mathsf{T}}X_k - B)K_h(A^{\mathsf{T}}X_l - B)K_h(A^{\mathsf{T}}X_m - B)\|X_k - X_l\||Y_k - Y_m|$$

can be calculated in $O(n^2)$ time. This is obviously true for $S_{n,1}(A;B)$ and $S_{n,2}(A;B)$. For $S_{n,3}$, its summation can be rewritten as a weighted average of the product of sample distances, $\sum_k K_h(A^{\mathsf{T}}X_k - B)\sum_l K_h(A^{\mathsf{T}}X_l - B)\|X_k - X_l\|\sum_m K_h(A^{\mathsf{T}}X_m - B)|Y_k - Y_m|$, from which we see $S_{n,3}$ can also be calculated in $O(n^2)$ time.

We use $V_n(A;B)$ to construct a statistic for testing (3) by first scaling to address asymptotic bias and then applying a Cramer–von Mises functional, which is known to produce effective tests. By Theorem 1 in Section 2·3, the asymptotic bias of $V_n(A;B)$ is proportional to $\mu_x(B)\mu_y(B)/f_0(B)$, so we stabilize $V_n(A;B)$ by normalizing by a consistent estimator of $\mu_x(B)\mu_y(B)/f_0(B)$. From Lemma 5 in the Appendix, $S_{n,2}(A;B) \to \mu_x(B)\mu_y(B)$ in probability as $n \to \infty$. Consequently, we normalize by $S_{n,2}(A;B)/\hat{f}_0(B)$, leading to the test statistic

$$T_n(A) = \int \frac{V_n(A;B)}{S_{n,2}(A;B)}\hat{f}_0(B)dB,$$

which can be approximated with $T_n(A) \approx n^{-1}\sum_{i=1}^{n} V_n(A;A^{\mathsf{T}}X_i)/S_{n,2}(A;A^{\mathsf{T}}X_i)$.

2·3. *Asymptotic results*

Let $f^{(X)}_{A^{\mathsf{T}}X}(\cdot \mid B)$, $f^{(Y)}_{A^{\mathsf{T}}X}(\cdot \mid B)$ and $f^{(X,Y)}_{A^{\mathsf{T}}X}(\cdot,\cdot \mid B)$ denote the conditional density functions of X and Y, and the joint density function of X and Y given $A^{\mathsf{T}}X = B$. For the asymptotic analysis, we need the following regularity conditions.

Condition 1. The density functions above are continuous and bounded away from 0. The support of $A^{\mathsf{T}}X$, Ω, is bounded and compact in $\mathbb{R}^d$.

Condition 2. The continuous kernel function $K(t)$ is Lipschitz on [-1,1] and, for some $q > d/2$,

$$\int K(t)dt = 1,\ \int t^i K(t)dt = 0, 1 \le i \le q-1,\ 0 \ne \int t^q K(t)dt < \infty.$$

Condition 3. The bandwidth $h \to 0$, $nh^{2d} \to \infty$ and $nh^{2q+d/2} \log n \to 0$.

Condition 4. $E(\|P\|^4) < \infty$, $E(Q^4) < \infty$ and $E(\|P\|^2 Q^2) < \infty$, where $(P, Q) \sim \psi_{A^\tau X}^{(X,Y)}(\cdot, \cdot \mid B)$.

Condition 5. Write $f_1(x, y, t) = f_{A^\tau X}^{(X,Y)}(x, y \mid t) f_0(t)$ which is q times differentiable with respect to t and its qth-order derivative is uniformly bounded by a constant C_0 which does not depend on t.

Conditions 1 and 5 require that the density functions are positive and sufficiently smooth. Condition 5 facilities the control of the remainders of Taylor expansions. We may consider relaxing this condition by imposing local Lipschitz properties of density functions, which are widely imposed in the literature (Li et al., 2011). Condition 2 implies that the kernel function is bounded from above, which holds for many well-known kernel functions. Condition 3 gives the condition on the bandwidth h which is relatively mild. Condition 4 requires some finite moments, which is a necessary condition for asymptotic normality.

We first have the following uniform consistency result for $V_n(A; B)$.

PROPOSITION 1. *Under Conditions 1–5,*

$$\sup_{B \in \mathbb{R}^d} |V_n(A; B) - V(A; B)| = O\{h^q + (nh^d)^{-1/2} \log n\},\ \textit{almost surely}.$$

By this proposition, $V_n(A; B)$ is an appealing quantity for assessing (3).

The following theorem establishes the weak convergence of $V_n(A; B)$. Let $D = \int K^2(t)dt$, $Z = (X, Y)$, $\nu_x(B) = E\left\{E^2(\|X - P\| \mid Z)\right\}$, $\nu_y(B) = E\left\{E^2(|Y - Q| \mid Z)\right\}$, $\mu_x(B) = E(\|P_1 - P_2\|)$ and $\mu_y(B) = E(|Q_1 - Q_2|)$.

THEOREM 1. *Suppose Conditions 1–4 hold. Then as $n \to \infty$ under (3),*

$$\frac{nh^d f_0(B)}{D} V_n(A; B) \to -2\mu_x\mu_y + \mu_x\mu_y\mathcal{N}_4^2 + 2\mathcal{N}_4(\mathcal{N}_3 - \mu_x\mathcal{N}_2 - \mu_y\mathcal{N}_1) + \mathcal{N}_1\mathcal{N}_2$$

in distribution, where $(\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4)$ is normally distributed with mean zero and covariance matrix

$$\Lambda = \begin{pmatrix} 4\nu_x & 4\mu_x\mu_y & 4\nu_x\mu_y + 2\mu_x^2\mu_y & 2\mu_x \\ 4\mu_x\mu_y & 4\nu_y & 4\nu_y\mu_x + 2\mu_y^2\mu_x & 2\mu_y \\ 4\nu_x\mu_y + 2\mu_x^2\mu_y & 4\nu_y\mu_x + 2\mu_y^2\mu_x & \lambda & 3\mu_x\mu_y \\ 2\mu_x & 2\mu_y & 3\mu_x\mu_y & 1 \end{pmatrix}, \tag{6}$$

with

$$\lambda = E\left\{E^2(\|X - P_1\||Y - Q_2| \mid Z)\right\} + E\left\{E^2(\|X - P_1\||Q_1 - Q_2| \mid Z)\right\}.$$

In particular, $\left\{nh^d f_0(B)/D\right\} E\left\{V_n(A; B)\right\} \to \mu_x\mu_y$ *as* $n \to \infty$.

The unconditional distance covariance $n\mathrm{dcov}_n^2(\mathcal{X}, \mathcal{Y})/\hat{S}_2$ converges in distribution to a weighted sum of independent chi-squared variables (Székely et al., 2007). Like its unconditional version, $V_n(A; B)$ is not asymptotically normal. Due to the use of kernel smoothing, the weak convergence rate of $V_n(A; B)$ is of $O(nh^d)$ instead of $O(n)$, which is expected due to the use of kernel estimation.

We next give the null distribution of $T_n(A)$. In preparation, let

$$L_1(Z_j, Z_k, Z_l, Z_m) = \frac{1}{6} \sum_{(i_1,i_2)\in(j,k,l,m)} \|X_{i_1} - X_{i_2}\||Y_{i_1} - Y_{i_2}|,$$

$$L_2(Z_j, Z_k, Z_l, Z_m) = \frac{1}{6} \sum_{(i_1,i_2,i_3,i_4)\in(j,k,l,m)} \|X_{i_1} - X_{i_2}\||Y_{i_3} - Y_{i_4}|,$$

$$L_3(Z_j, Z_k, Z_l, Z_m) = \frac{1}{24} \sum_{(i_1,i_2,i_3)\in(j,k,l,m)} \|X_{i_1} - X_{i_2}\||Y_{i_1} - Y_{i_3}|,$$

and $\gamma_h(Z_j, Z_k, Z_l, Z_m) = (L_1 + L_2 - 2L_3) \int \prod_{r=j,k,l,m} K_h(A^\mathsf{T} X_r - B) f_0(B)/S_2(A; B) dB$, which is the kernel function of a U-statistic related to $T_n(A)$.

THEOREM 2. *Suppose Conditions 1–5 hold. Then as $n \to \infty$ under (3),*

$$nh^{d/2} T_n(A) - h^{-d/2} D \to N(0, V_{T0}),$$

where $V_{T0} = 72 \lim_{n\to\infty} h^d E\{\gamma_{h2}^2(Z_1, Z_2)\}$ with $\gamma_{h2}(Z_1, Z_2) = E\{\gamma_h(Z_1, Z_2, Z_3, Z_4) \mid Z_1, Z_2\}$.

Theorem 2 suggests that the null hypothesis be rejected when $\left\{nh^{d/2} T_n(A) - h^{-d/2} D\right\} / \sqrt{V_{T0}} > z_\alpha$, where z_α is the upper α quantile of $N(0, 1)$. $E\{\gamma_{h2}^2(Z_1, Z_2)\}$ is proportional to h^{-d} and thus V_{T_0} is a positive constant depending on the joint distribution of X and Y as well as $K(\cdot)$. $T_n(A)$ is asymptotically normal with a faster rate $O(nh^{d/2})$ than that of $V_n(A; B)$. Under (3), both the asymptotic mean and variance of $nT_n(A)$ are of order $O(h^{-d})$, which is similar to those in lack-of-test statistics for nonparametric regression, such as the generalized likelihood ratio statistic in Fan et al. (2001).

Theorem 2 is proved in the Appendix using the following rationale. The statistic $T_n(A)$ is asymptotically equivalent to $\tilde{T}_n(A)$, a degenerate U-statistic of order four with kernel function $\gamma_h(Z_1, \ldots, Z_4)$. The limit distribution of a degenerate U-statistic when its kernel function is fixed is a linear combination of independent, centered χ_1^2 distributions, and cannot be derived using classical martingale methods. However, in certain cases in which the kernel function of the U-statistic depends on n just like $\gamma_h(Z_1, \ldots, Z_4)$, a normal distribution can be achieved. Theorem 1 is proved by applying the theorems in Hall (1984) to the projection of $\tilde{T}_n(A)$.

The following theorem investigates the asymptotic behavior of T_n without requiring (3).

THEOREM 3. *Suppose Conditions 1–5 hold.*

(i) Under a fixed alternative so that $\int V(A; B) f_0(B)/S_2(A; B) dB > 0$, $nh^d T_n(A) \to \infty$ in probability;

(ii) Let $\mathcal{P}_\mathcal{S}$ represent the projection matrix of a sufficient dimension reduction space. Under a local alternative $A = \mathcal{P}_\mathcal{S}^\mathsf{T} + (nh^{d/2})^{-1/2}\Delta$ for some $\Delta \in \mathbb{R}^{p\times d}$,

$$nh^{d/2} T_n(A) - h^{-d/2} D - \delta \to N(0, V_{T1}),$$

in distribution as $n \to \infty$, where $\delta = \int \left\{\int G_v^\mathsf{T}(B)\Delta^\mathsf{T} x f_{X|\mathcal{P}_\mathcal{S}X}(x|B)dx\right\}^2 dB$, $V_{T1} = 72 \lim_{n\to\infty} h^d E\{\gamma_{h2}^2(Z_1, Z_2)\}$, and $G_v(t)$ is the gradient of $f_{Y|\mathcal{P}_\mathcal{S}X}(v \mid t)$ with respect to t.

Theorem 3-(i) shows that under any conditionally dependent alternative, the probability that $\{nh^{d/2} T_n(A) - h^{-d/2} D\}/\sqrt{V_{T1}} > z_\alpha$ tends to one as $n \to \infty$. That is, the $T_n(A)$ test of conditional independence is consistent against all types of conditional dependence. Theorem 3-(ii) guarantees that the $T_n(A)$ test has nontrivial power against contiguous alternative of order $(nh^{d/2})^{-1/2}$. Together with Theorem 2, Theorem 3-(ii) establishes that the power of $T_n(A)$ is given by $\Phi\{-(V_{T0}/V_{T1})^{1/2} z_\alpha + \delta/V_{T1}^{1/2}\}$,

where Φ is the standard normal distribution function. This result also reveals that the $T_n(A)$ cannot distinguish alternatives of order smaller than $(nh^{d/2})^{-1/2}$ from the null, while the test would be consistent with the contiguous alternatives of order larger than $(nh^{d/2})^{-1/2}$.

3. PRACTICAL GUIDELINES

3·1. *Permutation test*

Similar to its unconditional counterpart in Székely et al. (2007), as shown in Theorem 2, the test statistic $T_n(A)$ is not asymptotically free of nuisance parameters under the null hypothesis (3). Its asymptotic variance depends on the conditional distributions of X and Y given $A^{\mathsf{T}}X$. To implement the proposed test for small samples, we obtain a reference distribution for $T_n(A)$ using a permutation procedure, the p-value being the fraction of replicates of $T_n(A)$ under random permutations of the indices of the sample of Y that are at least as large as the observed statistic. The effectiveness of this permutation procedure is evaluated in Section 4.

3·2. *Bandwidth choice*

Like many other smoothing-based tests, the performance of the proposed test depends upon the bandwidth h. It is widely acknowledged that the optimal h for nonparametric estimation is generally not optimal for testing (Hart, 1997). Selection of h for optimal power is an open problem (Kulasekera and Wang, 1997). Asymptotically, a range of bandwidths which satisfy Condition 3 could maintain the consistency of the test, while a specific bandwidth may maximize the power. The amount of smoothing applied will affect the power of the test, but we have observed in our simulations that the observed significance changes mildly over a wide range of values for h. In addition, we found that a larger bandwidth generally leads to better power. This can be understood from part (ii) of Theorem 3. The power function is given in Section 2·3 under the local alternative $A = \mathcal{P}_{\mathcal{S}}^{\mathsf{T}} + (nh^d/2)^{-1/2}\Delta$. With a larger bandwidth $h' > h$, the power becomes $\Phi\{-(V_{T0}/V_{T1})^{1/2}z_\alpha + (h/h')^{-d/2}\delta/V_{T1}^{1/2}\}$, resulting in an improvement. However, in practice, the condition $nh^{2q+d/2}\log n \to 0$ will be violated if h is too large. In that case, the conditional distance covariance tends to the unconditional one, which tests the marginal independence of X and Y rather than the conditional independence. In other words, an inappropriately large h will yield a much larger false alarm rate when X and Y are dependent.

Based on Condition 3 and our numerical experience, we recommend the empirical bandwidth $h = 0.5\sum_{i=1}^{d}\mathrm{sd}(U_i)n^{-1/(4+d/4)}$, where $U_i = \beta_i^{\mathsf{T}}X$, β_i is the ith column of A, and $\mathrm{sd}(U_i)$ is the sample standard deviation of U_i ($i = 1, \ldots, d$). This formula works well for a wide range of models and sample sizes as shown in Section 4. How to best utilize the data to select an optimal h for the proposed test warrants attention.

3·3. *Assessing fits*

An issue could be whether there is sufficient information in the data to contradict a particular estimate $\hat{A}$. The properties of our proposed method hold straightforwardly if we use cross validation, splitting the data randomly into two parts with one part used to determine $\hat{A}$ and the other part used to compute the statistic $T_n(\hat{A})$ and perform the permutation test.

Our asymptotic results are not strictly applicable when $\hat{A}$ is based on the full data because they do not account for the dependence introduced by using $\hat{A}$ instead of a fixed A. Nevertheless, Theorems 2 and 3 suggest that $T_n(\hat{A})$ could still be used to assess the conditional dependence between Y and X given $\hat{A}^{\mathsf{T}}X$ to a useful approximation. Theorem 3-(ii) indicates that if $\mathcal{S}_{\hat{A}}$ is in the $o\{(nh^{d/2})^{-1/2}\}$-neighborhood of $\mathcal{S}_A$, $T_n(\hat{A})$ would behave approximately as $T_n(A)$, and thus the p-value from $T_n(\hat{A})$ should still be informative. For example, sliced inverse regression provides a $\sqrt{n}$-consistent estimate of

the central subspace given that the linearity and coverage conditions hold (Cook, 1998). When $\hat{A}$ is not a consistent estimator, Theorem 3-(i) tells that roughly the p-value from $T_n(\hat{A})$ would be quite small. In this paper, we make no attempt to provide a formal analysis of $T_n(\hat{A})$, which deserves research. In Section 4 we provide numerical support for these indications when the permutation procedure is used to compute p-values.

We use a toy example to demonstrate the proposed test. Suppose that we are estimating the central subspace in the single-index model $Y = \exp(x_1 + x_2 + x_3) + \epsilon$, where $X = (x_1, x_2, x_3)^{\mathsf{T}} \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le 3$, and ϵ is a standard normal variate that is independent of X. The central subspace under this model is spanned by $(1, 1, 1)^T$. We generated one set of simulated data with sample size 100. Given the specific sample, there are various possible methods to get estimates of A. The estimate using directional regression, denoted as A_{DR}, and the estimate using one-component partial least squares, denoted as A_{PLS}, are $(0.55, 0.60, 0.58)^{\mathsf{T}}$ and $(0.36, 0.74, 0.57)^{\mathsf{T}}$. The angles between the estimates and the central subspace are 2.10 and 15.5 degrees, respectively. In practice we would not know these angles since we would not know the central subspace. The statistics based on A_{DR} and A_{PLS} are clearly dependent on the estimates, and thus not directly comparable. We calculated p-values – 0.15 for directional regression and 0.015 for partial least squares, based on 400 permutations, leading to the conclusion that there is no information in the data to contradict the directional regression estimate, but the data do cast doubt on the partial least squares estimate.

4. SIMULATION STUDIES

The first simulation results in this section are intended to support our contention that the permutation test based on $T_n(\hat{A})$ can be useful for comparing the performance of competing dimension reduction methods. For each of the following three simulation models (Chen et al., 2010), we ran $1,000$ replications and 400 permutation samples were used. We also used the Epanechnikov kernel with the empirical bandwidth given in Section 3.

We considered the following three models. The first model, Model I is $X = \Gamma(Y + Y^2) + \Psi^{1/2}\epsilon$, where $\epsilon \sim N(0, I_{10})$, $Y \sim N(0, 1)$, $\Psi_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le 10$, $\Gamma = (1, -1, \ldots, 1, -1)^{\mathsf{T}}/\sqrt{10}$, and $Y \perp\!\!\!\perp \epsilon$. The central subspace is the column space of $\Psi^{-1}\Gamma$. Model I is an instance of the principal fitted component model (Cook, 2007; Cook & Forzani, 2008).

The second model is, Model II, $Y = x_1/\{0.5 + (x_2 + 1.5)^2\} + 0.2\epsilon$, where $\epsilon \sim N(0, 1)$, $X = (x_1, \ldots, x_{10})^{\mathsf{T}} \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le 10$ and $X \perp\!\!\!\perp \epsilon$. The central subspace is spanned by the directions $\beta_1 = (1, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0, 1, \ldots, 0)^{\mathsf{T}}$.

The third model, Model III, is $Y = (X^{\mathsf{T}}\beta_1)^2 + |X^{\mathsf{T}}\beta_2| + 0.5\epsilon$, where $\epsilon \sim N(0, 1)$, $\beta_1 = (0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta_2 = (0.5, -0.5, 0.5, -0.5, 0, \ldots, 0)^{\mathsf{T}}$. The predictor vector $X = (x_1, \ldots, x_{10})^{\mathsf{T}}$ is independent of ϵ and defined as follows: The last nine components $(x_2, \ldots, x_{10})^{\mathsf{T}} \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ and the first component $x_1 = |x_2 + x_3| + \zeta$ where ζ is an independent standard normal variable. The central subspace is spanned by the vectors β_1 and β_2.

The simulated type I errors for the three models is given in Table 1, under various values of n, and the nominal type I errors. The empirical levels are close to the nominal level in most cases, which shows the effectiveness of the suggested permutation procedure. Next, we evaluate the power of the proposed method. In Model I we used the alternative $\Gamma' = \Gamma + \Psi\beta$, where $\beta = (1, \ldots, 1)^{\mathsf{T}}/\sqrt{10}$. The alternative values of β in Model II were $\beta'_1 = (1/\sqrt{10}, 0, 3/\sqrt{10}, \ldots, 0)^{\mathsf{T}}$ and $\beta'_2 = (0, 1/\sqrt{10}, 0, 3/\sqrt{10}, \ldots, 0)^{\mathsf{T}}$, while the alternative values of β in Model III were $\beta'_1 = (1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)^{\mathsf{T}}$ and $\beta'_2 = (0, 0, 1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)^{\mathsf{T}}$. Table 1 presents power results for Models I-III when $n = 100$, 200 and 400. The test has better efficiency with larger n as expected.

Table 1. *Empirical sizes and power (%) of the proposed test using the permutation procedure*

Model	n	1%		5%		10%	
		size	power	size	power	size	power
	100	0.8	17.0	6.0	38.9	9.9	50.1
I	200	0.9	49.4	6.0	74.3	10.0	82.1
	400	1.1	88.4	5.4	96.8	9.9	98.3
	100	1.1	7.1	5.6	21.2	11.0	32.8
II	200	1.1	18.9	5.6	44.4	9.6	57.4
	400	0.9	65.0	6.8	87.7	13.6	93.2
	100	0.7	7.0	5.6	21.8	9.2	32.9
III	200	0.7	20.0	4.3	45.2	8.6	57.7
	400	0.8	62.3	4.8	85.0	9.8	90.9

Table 2. *Rejection rates (%) for the proposed test using directional regression (DR), ordinary least squares (OLS), principal fitted components (PFC), partial least squares (PLS) and minimum average variance estimation (MAVE)*

n	Model I				Model II				Model III			
		1%	5%	10%		1%	5%	10%		1%	5%	10%
	DR	14.8	30.3	41.9	DR	1.8	10.0	18.3	DR	6.5	22.3	35.2
100	PFC	2.7	8.2	12.2	PFC	1.2	5.9	12.0	PFC	9.3	23.9	36.8
	OLS	81.6	91.0	93.6	PLS	6.1	18.7	31.0	MAVE	0.7	4.0	8.1
	DR	16.3	33.3	42.6	DR	4.1	14.3	24.5	DR	14.8	38.4	52.6
200	PFC	2.5	8.4	13.0	PFC	1.8	9.0	16.6	PFC	28.3	55.2	68.1
	OLS	91.4	97.2	98.6	PLS	22.9	47.6	61.0	MAVE	0.2	2.3	6.2

Next, we used the proposed test to compare a few dimension reduction methods under the three models considered. We compared the results of directional regression, principal fitted components and ordinary least squares in Model I. These methods are known to produce root-n consistent estimators of the central subspace under Model I. The average angles between the central subspace and the estimates from directional regression, principal fitted components and ordinary least squares in Model I are 43.6, 12.7 and 74.2 degrees for $n = 100$, and 33.3, 8.7 and 68.5 degrees for $n = 200$. The ordinary least squares method in Model I has the worst performance, the boxplots of Model I in Fig. 1 showing that the empirical p-values from ordinary least squares are very small for both sample sizes. This can be clearly seen from Table 2 as the rejection rates for ordinary least squares are all above 80% at 1%, 5% and 10% significance levels.

For Model II, we compared the first two components from directional regression, principal fitted components and partial least squares. Here, directional regression gives a root-n consistent estimator of the central subspace, but the corresponding asymptotic properties of principal fitted components and partial

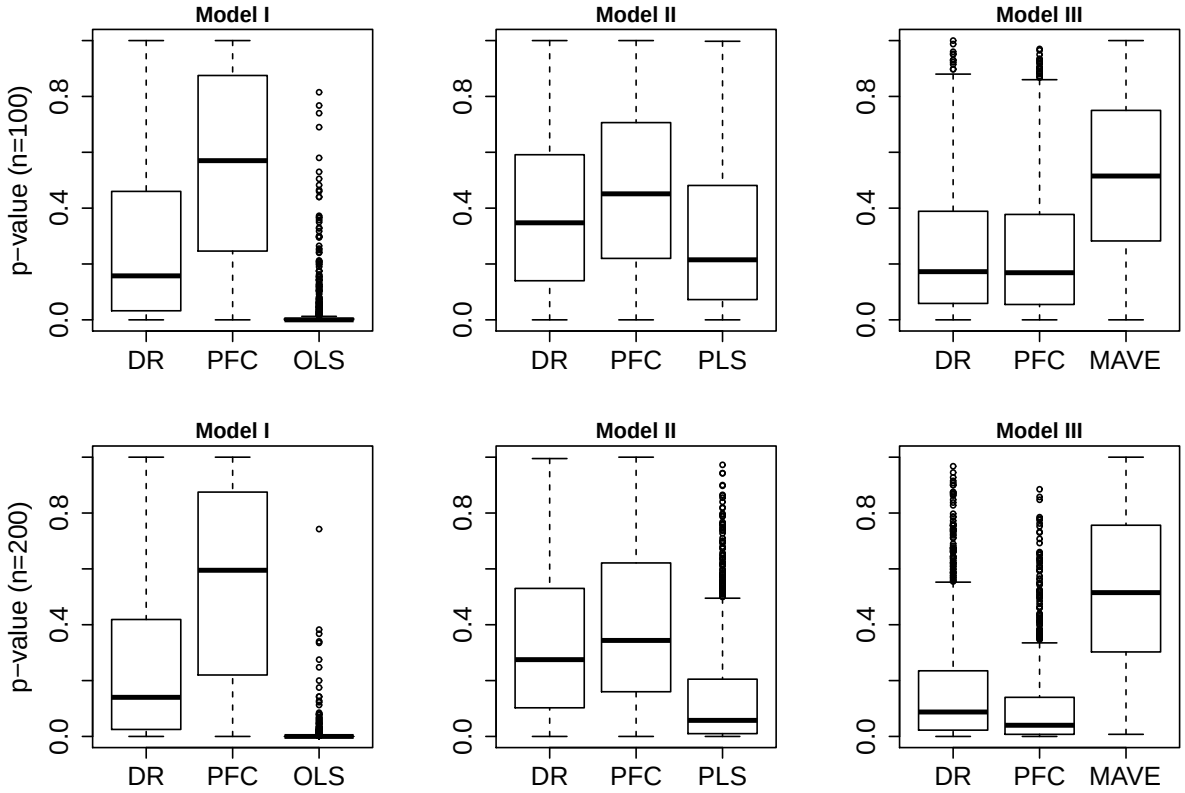

Fig. 1. Box-plots of p-values via the proposed test using directional regression (DR), ordinary least squares (OLS), principal fitted components (PFC), partial least squares (PLS) and minimum average variance estimation (MAVE)

least squares are unknown. The average values of the largest principal angles between the central subspace and the estimates using directional regression, principal fitted components and partial least squares in Model II are 49.5, 31.8 and 75.6 degrees for $n = 100$; 34.3, 22.4 and 75.0 degrees for $n = 200$. It appears that the partial least squares method in Model II is the worst performer. This is also the conclusion from the boxplots for Model II in Fig. 1 and the rejection rates for partial least squares in Table 2.

For Model III, we compared the first two components from directional regression, principal fitted components and minimum average variance estimation (Xia, 2007). The average values of the largest principal angles between the central subspace and the estimates from directional regression, principal fitted components and minimum average variance estimation in Model III are 52.7, 60.5 and 32.5 degrees for $n = 100$; 39.3, 54.1 and 14.4 for $n = 200$. In Model III, since x_1 is generated as $|x_2 + x_3| + \zeta$, the linearity condition does not hold. Therefore directional regression and principal fitted components might not work well here. The boxplots of Model III in Fig. 1 show that the empirical p-values of these two methods are small compared with the minimum average variance estimation method. Similarly, the rejection rates for minimum average variance estimation are much smaller than the other two methods as reflected in Table 2, demonstrating the superiority of minimum average variance estimation for such a model. In general, the results of our sufficiency test are consistent with theoretical analysis and can well reflect the relative goodness-of-fit of a given estimate.

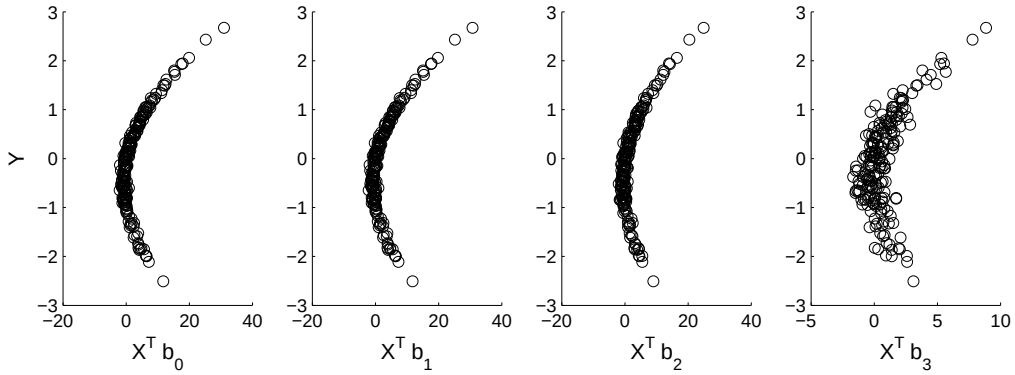

Fig. 2. Scatter plots of Y versus $X^{\mathsf{T}}b_i$, $i = 0, 1, 2, 3$

We also repeated the simulations leading to Fig. 1 using cross validation. On each replication the data were split randomly into two sets of $n/2$ observations each, with one set being used to determine $\hat{A}$ and the other being used for the permutation test. The boxplots of the p-values, which are available in the Supplement, are very much like those of Fig. 1.

To further demonstrate the effectiveness of our method, we simulated one set of data with size 200 from Model I and drew a 1×4 plot, with each subplot being of Y versus $X^{\mathsf{T}}b_i$, where b_i $(i = 0, 1, 2, 3)$, with Euclidean length one, denote the true direction of the central subspace, and the estimates from principal fitted components, directional regression and ordinary least squares, respectively. The empirical p-values were $0.985, 0.988, 0.27, 0$ respectively. In Fig. 2, both the plots of the principal fitted components and directional regression are quite close to that with the true direction. In contrast, the plot of the ordinary least squares solution is far away from the true one, which coincides with the fact that its corresponding p-value is 0.

5. Real-Data Analysis

5·1. *Salary data*

We used our method to assess various dimension reduction methods applied to a regression involving employee salary in the Fifth National Bank of Springfield. The response variable Y is an employee's annual salary. Six predictors are possibly associated with the salary: the employee's current job level; the number of years of the employment for a current employee; an employee's age; the the number of years for an employee working at another bank prior to the Fifth National; the employee's gender; and a binary variable indicating whether the employee's job is computer related. One obvious outlier was removed from this dataset, leaving 207 observations in the subsequent analysis. The aim of this study is to understand how an employee's salary associates with these six predictors.

Ma and Zhu (2012) studied this dataset thoroughly and we calculated p-values based on their selected methods. They constructed estimates using their semi-parametric method, directional regression and minimum average variance estimation. Minimum average variance estimation requires continuous predictors, however, the gender is a factor. The linearity condition and constant variance condition are not satisfied in this dataset as discussed by Ma and Zhu (2012). Thus minimum average variance estimation and directional regression might not work well.

Ma and Zhu (2012) use a quadratic fit to perform a cross-validation procedure to calculate the prediction error. They claimed that it fits the data well and the resulting prediction errors for semi-parametric method, minimum average variance estimation and directional regression are, respectively, 21.3, 23.4 and

47.0. We uses their estimates in combination with our sufficiency statistics to find the empirical p-values for these three methods with permutation size 400. The resulting p-values for semi-parametric method, minimum average variance estimation and directional regression are, respectively, 0.170, 0.394 and 0. We can see that our results are in agreement with Ma and Zhu's prediction errors in this dataset. There is information in the data to contradict the directional regression estimate since its p-value is essentially 0, but neither of the other estimates is ruled out.

5·2. *Application to linear regression*

Our approach is not limited to the sufficient dimension reduction framework. Consider the standard homoscedastic linear regression framework

$$Y = a + \beta^{\mathsf{T}} X + \epsilon,$$

where $X \perp\!\!\!\perp \epsilon$. Let $\hat{\beta}$ denote some estimator of β. Assessing the fit of this model can be considered in two stages. The first is to ask if there is information in the data to contradict the conditional independence statement $Y \perp\!\!\!\perp X \mid \hat{\beta}^{\mathsf{T}} X$. Finding no contradictory evidence, the second stage is to ask if the model itself holds. Our method can check the conditional independence of Y and X given $\hat{\beta}^{\mathsf{T}} X$. If that check is passed, a sufficient summary plot (Cook, 1998) of Y versus $\hat{\beta}^{\mathsf{T}} X$ may be adequate for the second stage.

We use ozone concentration data (Cleveland, 1993) to demonstrate how this checking method works. In this example, we study the relationship between ozone concentration Y and radiation level, temperature, and wind speed using 111 observations taken daily in New York from May to September 1973. The ordinary least squares estimate $\hat{\beta}_{\mathrm{OLS}}$ and the one-component partial least squares estimate $\hat{\beta}_{\mathrm{PLS}}$ were obtained. The empirical p-values based on the statistics $T_n(\hat{\beta}_{\mathrm{OLS}})$ and $T_n(\hat{\beta}_{\mathrm{PLS}})$ were computed as 0.102 and 0 respectively based on 400 permutations. We see that ordinary least squares estimate cannot be rejected based on the p-value, but there is sufficient information in the data to reject the partial least squares estimate. The plot of Y versus $\hat{\beta}_{\mathrm{OLS}}^{\mathsf{T}} X$ in left panel of Fig. 3 shows a clear curvature pattern, and could now be used to guide revision of the model. Partial least squares estimate should be rejected and the corresponding plot in the right panel of Fig. 3 does not show much useful information compared with the left one. These data are often studied with a single-index model, and it is known that ordinary least squares provides a model–robust estimator under mild conditions (Cook, 1998).

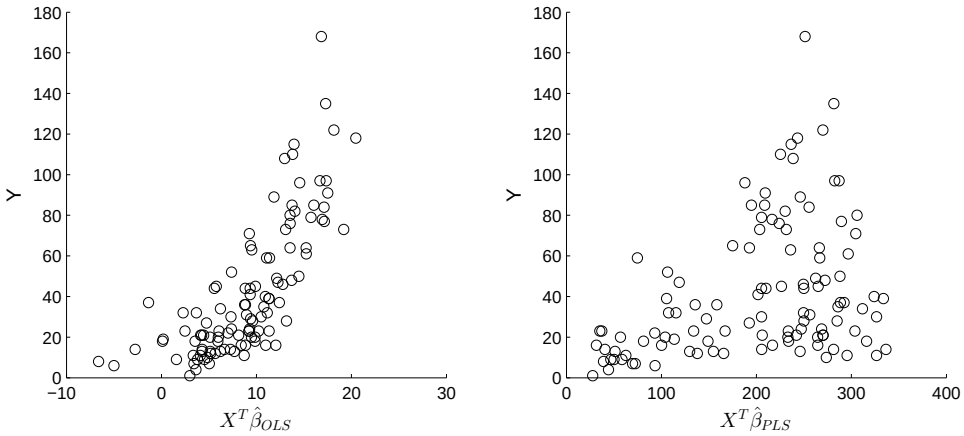

Fig. 3. Scatter plots of Y vs $X^{\mathsf{T}}\hat{\beta}$ for ordinary least squares and partial least squares

6. DISCUSSION

Our method addresses a crucial question in sufficient dimension reduction, providing a technique to assess the relative worth of dimension reduction methods regardless of the assumptions underlying them. Our method is not designed to compare the average performance of estimators, but should be useful for evaluating the sufficiency of estimates on a particular dataset.

Continuous density functions were required for the asymptotic results of Section 2·3, but the conditional distance covariance $V(A; B)$ does not require densities. Our test statistic $T_n(A)$ seems to work well even if some of the predictors are discrete. For instance, the regression study in Section 5·1 has discrete predictors and yet our results agree well with those from Ma & Zhu (2012). Nevertheless, it would be useful to develop a modified version of $V_n(A, B)$ that accommodates discrete predictors because the consistency of kernel-based estimators is not valid without the continuity assumption.

Our goodness-of-fit test is designed to detect any type of deviation from conditional independence (1), which can be a clear advantage in many regressions. Nevertheless, this general ability means that it may not be as powerful as tests tailored to detect specific types of conditional dependence. Our approach does not directly apply to dimension reduction for the conditional mean $E(Y \mid X)$ or conditional variance $\mathrm{var}(Y \mid X)$ which leads to the statements $Y \perp\!\!\!\perp E(Y \mid X) \mid \mathcal{P}_{\mathcal{S}}X$ or $Y \perp\!\!\!\perp \mathrm{var}(Y \mid X) \mid \mathcal{P}_{\mathcal{S}}X$. The general issue is to check the statement as $Y \perp\!\!\!\perp f(X) \mid \mathcal{P}_{\mathcal{S}}X$, where f is a measurable function of X. The challenge lies in that $f(X)$ may have to be estimated after dimension reduction, as pointed out by Li et al. (2003). We think this general issue deserves further study.

ACKNOWLEDGEMENT

The authors would like to thank the editor, an associate editor and two anonymous referees for their constructive comments that lead to substantial improvement for the paper. This research was supported by the National Natural Science Foundation and the Foundation for the Authors of National Excellent Doctoral Dissertations. Three authors contributed equally and Zou is the corresponding author.

Supplementary Material

The Supplementary Material contains the proofs of the technical results and the cross validation version of Fig. 1.

REFERENCES

CHEN, X., ZOU, C. & COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723.

CLEVELAND, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, New Jersey.

COOK, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc*. **89** 177–190.

COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics.* Wiley, New York.

COOK, R. D. & NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428.

COOK, R. D. (2007). Fisher Lecture: Dimension deduction in regression (with discussion). *Statist. Sci.* **22** 1-26.

COOK, R. D. & FORZANI, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23** 485–501.

COOK, R. D. & FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104** 197–208.

COOK, R. D. & WEISBERG, S. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc*. **86** 328–332.

FAN, J., ZHANG, C. & ZHANG, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193.

HALL, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric estimators. *J. Multivar. Anal.* **14**, 1–6.

HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer, New York.

KULASEKERA, K. B. & WANG, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *J. Amer. Statist. Assoc.* **92** 500–511.

LI, B., COOK, R. D. & CHIAROMONTE, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Ann. Statist.* **31** 1636–1668.

LI, B., ZHA, H. & CHIAROMONTE, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616.

LI, B. & WANG, S. (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** 316–327.

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.

LI, L., COOK, R. D. & NACHTSHEIM, C. J. (2005). Model-free variable selection. *J. Roy. Statist. Soc. B* **67** 285–299.

LI, L., ZHU, L. & ZHU, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. *J. R. Statist. Soc. B* **73** 59–80.

MA, Y. & ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107** 168–179.

SHENG, W. & YIN, X. (2013). Direction estimation in single-index models via distance covariance. *J. Multivariate Anal.* **122** 148–161.

STUTE, W. & ZHU, L. X. (2005). Nonparametric checks for single-index models. *Ann. Statist.* **33** 1048–1083.

SZÉKELY, G. J., RIZZO, M. L. & BAKIROV, N. K. (2007). Measuring and testing independence by correlation of distances. *Ann. Statist.* **35** 2769–2794.

SZÉKELY, G. J. & RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.* **4** 1236–1265.

XIA, Y., TONG, H., LI, W. K. & ZHU, L. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. Roy. Statist. Soc. Ser. B* **64** 363–410.

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

XIA, Y. (2009). Model checking in regression via dimension reduction. *Biometrika* **96** 133–148.