Statistica Sinica **31** (2021), 2179-2194
doi:https://doi.org/10.5705/ss.202019.0095

TESTING THE LINEAR MEAN AND CONSTANT VARIANCE CONDITIONS IN SUFFICIENT DIMENSION REDUCTION

Tingyou Zhou[1], Yuexiao Dong[2] and Liping Zhu[3,4]

[1]*Zhejiang University of Finance and Economics,* [2]*Temple University,*
[3]*Renmin University of China and* [4]*Zhejiang Gongshang University*

Abstract: Sufficient dimension-reduction (SDR) methods characterize the relationship between a response Y and the covariates $\mathbf{x}$ using a few linear combinations of the covariates. Extensive SDR techniques have been developed, among which, the inverse regression-based methods are perhaps the most appealing in practice, because they do not involve multi-dimensional smoothing and are easy to implement. However, these methods require two distributional assumptions on the covariates. In particular, the first-order methods, such as the sliced inverse regression, require the linear conditional mean (LCM) assumption, while the second-order methods, such as the sliced average variance estimation, also require the constant conditional variance (CCV) assumption. We check the validity of the LCM and CCV conditions using mean independence tests, which are facilitated by the martingale difference divergence. We propose a consistent bootstrap procedure to decide the critical values of the test. Monte Carlo simulations and an application to a horse mussels data set demonstrate the finite-sample performance of the proposed method.

Key words and phrases: Constant variance, dimension reduction, inverse regression, linear mean, mean independence.

1. Introduction

Sufficient dimension reduction (SDR) (Li (1991); Cook (1998)) has received considerable attention in the past two decades. As a useful tool to reduce dimensionality, SDR can be combined with many other multivariate analysis methods to build regression models. SDR methods have also been widely used for exploratory data analysis and data visualization. Let Y be the response variable, and let $\mathbf{x}$ be the p-dimensional predictor. When the goal is an inference about the conditional distribution of Y given $\mathbf{x}$, SDR aims to find $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with $d < p$, such that

$$Y \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \tag{1.1}$$

Corresponding author: Liping Zhu, Center for Applied Statistics, Renmin University of China, Beijing 100872, China. E-mail: zhu.liping@ruc.edu.cn.

where "$\perp\!\!\!\perp$" means statistical independence. Under (1.1), the conditional distribution of Y given $\mathbf{x}$ is the same as that of Y given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. The column space of $\boldsymbol{\beta}$ is referred to as the dimension reduction space. If the intersection of all dimension reduction spaces exists and satisfies (1.1), this minimum subspace of $\mathbb{R}^p$ is called the central space (Cook (1998); Chiaromonte and Cook (2002)). When the goal is an inference about the conditional mean $E(Y \mid \mathbf{x})$, SDR considers

$$Y \perp\!\!\!\perp E(Y \mid \mathbf{x}) \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}. \tag{1.2}$$

The column space of $\boldsymbol{\alpha}$ is referred to as the mean dimension reduction space. The smallest mean dimension reduction space that satisfies (1.2) is called the central mean space (Cook and Li (2002)).

Many inverse regression-based methods exist in the literature to estimate the central space and the central mean space. Estimators of the central space include, among others, the sliced inverse regression (SIR) (Li (1991)), sliced average variance estimation (SAVE) (Cook and Weisberg (1991)), directional regression (Li and Wang (2007)), and cumulative slicing estimation (Zhu, Zhu and Feng (2010)). The ordinary least squares (OLS), principal Hessian directions (PHD) (Li (1992)), and contour regression (Li, Zha and Chiaromonte (2005)) methods are perhaps the most popular for estimating the central mean space. The aforementioned methods fall into two categories. In the first category, the SIR and OLS involve linear functions of $\mathbf{x}$, such as $E(\mathbf{x}Y)$ and $E(\mathbf{x} \mid Y)$, and are called first-order methods. In the second category, the SAVE, PHD, directional, and contour regression methods involve quadratic functions of $\mathbf{x}$, such as $E(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})$ and $E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|Y)$, and are called second-order methods. Unlike other nonparametric and semiparametric methods, the inverse regression-based SDR methods do not involve multi-dimensional smoothing, regardless of p. This feature, together with the fact that they are easy to implement, makes these methods very appealing in practice.

Two assumptions about the distribution of $\mathbf{x}$ are required for the inverse regression-based SDR methods to properly recover the central space or the central mean space. To ease subsequent presentation, we use $\mathbf{B} \in \mathbb{R}^{p\times d}$ to denote the basis of the central space or that of the central mean space. The first-order methods require that

$$E(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \text{ is a linear function of } \mathbf{B}^{\mathrm{T}}\mathbf{x}, \tag{1.3}$$

which is referred to as the linear conditional mean (LCM) condition. In addition to the LCM, the second-order methods require that

$$\mathrm{var}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \text{ is a constant matrix,} \tag{1.4}$$

which is known as the constant conditional variance (CCV) condition. When (1.3) holds for all possible $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{x}$ must have an elliptically contoured distribution. When both (1.3) and (1.4) hold for all possible $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{x}$ has to be multivariate normal.

The LCM and CCV conditions have motivated many important developments in the SDR literature. To achieve these conditions, Cook and Nachtsheim (1994) proposed elliptically contoured reweighting, and Cook (1998) suggested marginal predictor transformations. To relax these conditions, Xia et al. (2002) proposed a minimum average variance estimation based on semiparametric models, Fukumizu, Bach and Jordan (2009) developed a contrast function by using operators on reproducing kernel Hilbert spaces to estimate the subspaces, and Li and Dong (2009) and Dong and Li (2010) introduced the concept of a central solution space and modeled $E(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ parametrically. More recently, Ma and Zhu (2012) proposed a semiparametric approach where $E(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ and $\mathrm{var}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ are estimated using nonparametric smoothing. These methods avoid the common assumptions of a linear mean and a constant variance on the covariates, but are computationally expensive compared with the classical SIR and SAVE.

We provide a novel treatment of the LCM and CCV conditions. Based on a root-n-consistent estimator of $\mathbf{B}$, we formally test the validity of the LCM and CCV conditions using hypothesis testing. It turns out that (1.3) and (1.4) are equivalent to statements about mean independence. Thus, testing the validity of the two conditions becomes equivalent to testing for mean independence. There is an extensive body of literature on consistently testing the correct specification of a particular regression model, which involves testing for mean independence. Most of these approaches can be divided into two classes: local smoothing approaches (Zheng (1996); Li (1999); Koul and Ni (2004); Guo, Wang and Zhu (2016)), and global smoothing approaches (Stute (1997); Li, Hsiao and Zinn (2003); Escanciano (2006)). The local approach requires nonparametric smoothing and, thus, its finite-sample performance depends heavily on the choice of the bandwidth. In general, the global approach turns the mean independence into an infinite number of unconditional constraints.

To formally measure the departure of the mean independence between two random variables U and V, Shao and Zhang (2014) extended the distance correlation proposed by Székely, Rizzo and Bakirov (2007) and Székely and Rizzo

(2009), introducing a novel metric called the martingale difference divergence (MDD). They found that the MDD of V given U is always nonnegative, and is equal to zero if and only if the conditional mean of V given U is independent of U. We observe that testing the LCM and CCV conditions is equivalent to testing for mean independence. Therefore, our test procedure is facilitated by the MDD originally proposed in Shao and Zhang (2014).

The rest of this paper is organized as follows. In Section 2, we explain the rationale of our test for the LCM and CCV conditions. Then, we investigate the sample-level properties of our proposal in Section 3. An extension to the high-dimensional case is discussed in Section 4. Numerical studies are conducted in Section 5 with Monte Carlo simulations and an application to a horse mussels data set. All technical proofs are collected in the online Supplementary Material.

2. The Principle of Testing LCM and CCV

To simplify the discussion in this section, without loss of generality, we assume that $E(\mathbf{x}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{x}) = \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix. This is a valid assumption, owing to the invariance property (Cook (1998)) of the central space and the central mean space. Let "$\otimes$" be the Kronecker product, and denote $\mathbf{P_B} = \mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}$ as the projection matrix onto the column space of $\mathbf{B} \in \mathbb{R}^{p\times d}$. We have the following key observation.

Proposition 1. *Suppose* $E(\mathbf{x}) = \mathbf{0}$ *and* $\mathrm{var}(\mathbf{x}) = \mathbf{I}_p$. *Then,*

1. *The LCM condition* (1.3) *holds if and only if* $E(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\varepsilon})$ *almost surely, where* $\boldsymbol{\varepsilon} \overset{\mathrm{def}}{=} \mathbf{x} - \mathbf{P_B}\mathbf{x}$.

2. *Suppose the LCM condition is true. Then, the CCV condition* (1.4) *holds if and only if* $E(\boldsymbol{\varepsilon}\otimes\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\varepsilon}\otimes\boldsymbol{\varepsilon})$, *almost surely.*

3. *The LCM condition* (1.3) *and the CCV condition* (1.4) *hold simultaneously if and only if* $E(\boldsymbol{\zeta} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\zeta})$ *almost surely, where* $\boldsymbol{\zeta} \overset{\mathrm{def}}{=} \{\boldsymbol{\varepsilon}^{\mathrm{T}}, (\boldsymbol{\varepsilon}\otimes\boldsymbol{\varepsilon})^{\mathrm{T}}\}^{\mathrm{T}}$.

Consider two random vectors $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \mathbb{R}^t$. Proposition 1 suggests that the LCM and CCV conditions have the same form as $E(\mathbf{v} \mid \mathbf{u}) = E(\mathbf{v})$, almost surely. This motivates us to consider testing $E(\mathbf{v} \mid \mathbf{u}) = E(\mathbf{v})$ almost surely, for any $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \mathbb{R}^t$, which can then be used to facilitate the tests for the LCM and CCV conditions.

Note that $E(\mathbf{v} \mid \mathbf{u}) = E(\mathbf{v})$ means that the conditional mean of $\mathbf{v}$ given $\mathbf{u}$ is independent of $\mathbf{u}$. We refer to this property as the mean independence, which measures the relationship between two random vectors $\mathbf{v}$ and $\mathbf{u}$, and lies between independence and uncorrelatedness. Specifically, $\mathbf{v} \perp\!\!\!\perp \mathbf{u}$ implies $E(\mathbf{v} \mid \mathbf{u}) = E(\mathbf{v})$, almost surely, which implies $\operatorname{cov}(\mathbf{v}, \mathbf{u}) = \mathbf{0}$. Therefore, to measure the mean independence, we can use the MDD (Shao and Zhang (2014)). Although Shao and Zhang (2014) only consider the case of a scalar response $\mathbf{v} \in \mathbb{R}$, the definition of the MDD can be generalized to the case with a vector response $\mathbf{v} \in \mathbb{R}^t$.

Let $|\mathbf{c}|_q \overset{\text{def}}{=} (\mathbf{c}^{\mathrm{T}}\mathbf{c})^{1/2}$ be the Euclidean norm of $\mathbf{c} \in \mathbb{R}^q$. For $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \mathbb{R}^t$, denote $(\widetilde{\mathbf{v}}, \widetilde{\mathbf{u}})$ as an independent copy of $(\mathbf{v}, \mathbf{u})$. From part (1) of Theorem 1 in Shao and Zhang (2014), the square of the MDD is equivalent to $m(\mathbf{v} \mid \mathbf{u})$, which is defined as

$$m(\mathbf{v} \mid \mathbf{u}) \overset{\text{def}}{=} -E\left[\{\mathbf{v} - E(\mathbf{v})\}^{\mathrm{T}}\{\widetilde{\mathbf{v}} - E(\widetilde{\mathbf{v}})\}|\mathbf{u} - \widetilde{\mathbf{u}}|_q\right]. \tag{2.1}$$

The next result is similar to Theorem 1 of Shao and Zhang (2014).

Proposition 2. *If $E(|\mathbf{u}|_q^2 + |\mathbf{v}|_t^2) < \infty$, then $m(\mathbf{v} \mid \mathbf{u}) \geq 0$, and the equality holds if and only if $E(\mathbf{v} \mid \mathbf{u}) = E(\mathbf{v})$, almost surely.*

Proposition 1 and Proposition 2 together provide the basic principle for testing the LCM and CCV conditions in this study. For first-order methods such as the OLS and SIR, only the LCM condition is required. Motivated by part 1 of Proposition 1, we consider the following hypotheses:

$$\begin{aligned} &H_0 : E(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\varepsilon}) \text{ a.s. } \text{ for some } \mathbf{B} \in \mathbb{R}^{p\times d} \text{ vs.} \\ &H_1 : E(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \neq E(\boldsymbol{\varepsilon}) \text{ a.s. } \text{ for all } \mathbf{B} \in \mathbb{R}^{p\times d}. \end{aligned} \tag{2.2}$$

where "a.s." means almost surely. The hypotheses in (2.2) test the mean independence between $\boldsymbol{\varepsilon}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{x}$, and are referred to as the LCM hypotheses. To test the hypotheses in (2.2), Proposition 2 suggests that we consider the following pivotal quantity:

$$m(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \overset{\text{def}}{=} -E\left[\{\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\}^{\mathrm{T}}\{\widetilde{\boldsymbol{\varepsilon}} - E(\widetilde{\boldsymbol{\varepsilon}})\}|\mathbf{B}^{\mathrm{T}}(\mathbf{x} - \widetilde{\mathbf{x}})|_d\right], \tag{2.3}$$

where $\widetilde{\mathbf{x}}$ is an independent copy of $\mathbf{x}$, $\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{P}_{\mathbf{B}}\mathbf{x}$, and $\widetilde{\boldsymbol{\varepsilon}} \overset{\text{def}}{=} \widetilde{\mathbf{x}} - \mathbf{P}_{\mathbf{B}}\widetilde{\mathbf{x}}$.

For second-order methods, such as the SAVE, PHD, and directional regression, both conditions are required. Motivated by part 3 of Proposition 1, we consider the following hypotheses:

$$\begin{aligned} &H_0 : E(\boldsymbol{\zeta} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\zeta}) \text{ a.s. } \text{ for some } \mathbf{B} \in \mathbb{R}^{p\times d} \text{ vs.} \\ &H_1 : E(\boldsymbol{\zeta} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \neq E(\boldsymbol{\zeta}) \text{ a.s. } \text{ for all } \mathbf{B} \in \mathbb{R}^{p\times d}. \end{aligned} \tag{2.4}$$

The hypotheses in (2.4) test the conditional mean independence between the response $\boldsymbol{\zeta}$ and the predictor $\mathbf{B}^{\mathrm{T}}\mathbf{x}$, and are referred to here as the joint hypotheses. To test the hypotheses in (2.4), Proposition 2 suggests that we consider

$$m(\boldsymbol{\zeta} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \overset{\mathrm{def}}{=} -E\left[\{\boldsymbol{\zeta} - E(\boldsymbol{\zeta})\}^{\mathrm{T}}\{\widetilde{\boldsymbol{\zeta}} - E(\widetilde{\boldsymbol{\zeta}})\}|\mathbf{B}^{\mathrm{T}}(\mathbf{x} - \widetilde{\mathbf{x}})|_d\right], \tag{2.5}$$

where $\widetilde{\boldsymbol{\zeta}} \overset{\mathrm{def}}{=} \{\widetilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}, (\widetilde{\boldsymbol{\varepsilon}} \otimes \widetilde{\boldsymbol{\varepsilon}})^{\mathrm{T}}\}^{\mathrm{T}}$ is an independent copy of $\boldsymbol{\zeta} = \{\boldsymbol{\varepsilon}^{\mathrm{T}}, (\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon})^{\mathrm{T}}\}^{\mathrm{T}}$.

3. The Sample-Level Properties

We focus on testing the LCM hypotheses in (2.2) in this section. The properties of the sample-level test for the joint hypotheses in (2.4) are similar, and are thus omitted for ease of presentation. Let $\{(\mathbf{x}_j, Y_j) : j = 1, \ldots, n\}$ be an independent and identically distributed (i.i.d.) sample of $(\mathbf{x}, Y)$. Our main idea is to test (2.2) using the sample estimator of $m(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$. Let $\widehat{\mathbf{B}}$ be a sample estimator of $\mathbf{B}$ that depends on $\mathbf{x}_j$ and Y_j, for $j = 1, \ldots, n$. Let $\mathbf{P}_{\widehat{\mathbf{B}}} \overset{\mathrm{def}}{=} \widehat{\mathbf{B}}\big(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}}\big)^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$, $\mathbf{Q}_{\widehat{\mathbf{B}}} \overset{\mathrm{def}}{=} \mathbf{I}_p - \mathbf{P}_{\widehat{\mathbf{B}}}$, $\widehat{\boldsymbol{\varepsilon}}_j \overset{\mathrm{def}}{=} \mathbf{Q}_{\widehat{\mathbf{B}}}\mathbf{x}_j$, and

$$\overline{\boldsymbol{\varepsilon}} \overset{\mathrm{def}}{=} n^{-1}\sum_{j=1}^{n}\widehat{\boldsymbol{\varepsilon}}_j.$$

The sample estimator of $m(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ becomes

$$\widehat{\omega}_n \overset{\mathrm{def}}{=} -n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n}\left\{(\widehat{\boldsymbol{\varepsilon}}_j - \overline{\boldsymbol{\varepsilon}})^{\mathrm{T}}(\widehat{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}})|\widehat{\mathbf{B}}^{\mathrm{T}}(\mathbf{x}_j - \mathbf{x}_k)|_d\right\}. \tag{3.1}$$

We follow Ma and Zhu (2013) to ensure the identifiability of $\mathbf{B}$. Specifically, for an arbitrary basis matrix $\mathbf{B}_t \in \mathbb{R}^{p\times d}$ of the central (mean) space, we write $\mathbf{B}_t \overset{\mathrm{def}}{=} (\mathbf{B}_u^{\mathrm{T}}, \mathbf{B}_l^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{B}_u$ is a $d \times d$ upper submatrix, $\mathbf{B}_l$ is a $(p - d) \times d$ lower submatrix, and the subscript "t" stands for total. We assume, without loss of generality, that $\mathbf{B}_u$ is invertible. If $\mathbf{B}_u$ is not invertible, we can always rotate the order of $\mathbf{x}$ to ensure that it is invertible, because the rank of $\mathbf{B}_t$ is d. As long as $\mathbf{B}_u$ is invertible, the column spaces of $\mathbf{B}_t$ and $\mathbf{B}_t\mathbf{B}_u^{-1}$ are identical. We define $\mathbf{B} \overset{\mathrm{def}}{=} \mathbf{B}_t\mathbf{B}_u^{-1}$, where the upper $d \times d$ submatrix of $\mathbf{B}$ is an identity matrix. This uniquely defines the true parameter. At the sample level, we apply a certain SDR method to estimate $\mathbf{B}$. The resultant estimate is denoted as $\widehat{\mathbf{B}}_t$, which is of the form $\widehat{\mathbf{B}}_t \overset{\mathrm{def}}{=} (\widehat{\mathbf{B}}_u^{\mathrm{T}}, \widehat{\mathbf{B}}_l^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\mathbf{B}}_u$ is a $d\times d$ upper submatrix, and $\widehat{\mathbf{B}}_l$ is a $(p-d)\times d$ lower submatrix. We then define $\widehat{\mathbf{B}} \overset{\mathrm{def}}{=} \widehat{\mathbf{B}}_t\widehat{\mathbf{B}}_u^{-1}$ as the sample estimator of $\mathbf{B}$.

Some notation is needed before we state the main theorem. Let $i = (-1)^{1/2}$ be the imaginary unit. Let $c_p \overset{\mathrm{def}}{=} \pi^{(1+p)/2}/\Gamma\{(1 + p)/2\}$, where $\Gamma(\cdot)$ is the gamma function. For a complex-valued function $\boldsymbol{\gamma}$: $\mathbb{R}^q \to \mathbb{C}^p$, we define its norm as

$$\|\boldsymbol{\gamma}(\mathbf{s})\|^2 \overset{\text{def}}{=} \int_{\mathbb{R}^q} |\boldsymbol{\gamma}(\mathbf{s})|_p^2 (c_q |\mathbf{s}|_q^{1+q})^{-1} d\mathbf{s}, \text{ where } |\boldsymbol{\gamma}(\mathbf{s})|_p^2 \overset{\text{def}}{=} \sum_{j=1}^{p} \nu_j(\mathbf{s})\overline{\nu_j(\mathbf{s})},$$

where $\nu_j(\mathbf{s}) \in \mathbb{C}$ is the jth element of $\boldsymbol{\gamma}(\mathbf{s}) \in \mathbb{C}^p$, and $\overline{\nu_j(\mathbf{s})}$ is the conjugate of $\nu_j(\mathbf{s})$, for $j = 1, \ldots, p$. Similar notation is introduced in Shao and Zhang (2014). Let "$\xrightarrow{d}$" stand for "convergence in distribution," and let "$\xrightarrow{p}$" stand for "converge in probability." The following technical condition about $\widehat{\mathbf{B}}$ is needed for the main result.

(C1). Suppose

$$\widehat{\mathbf{B}} - \mathbf{B} = n^{-1} \sum_{j=1}^{n} \boldsymbol{\ell}_1(\mathbf{x}_j, Y_j) + o_p(n^{-1/2}), \text{ and}$$

$$\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}} = n^{-1} \sum_{j=1}^{n} \boldsymbol{\ell}_2(\mathbf{x}_j, Y_j) + o_p(n^{-1/2}).$$

Assume $E\{\boldsymbol{\ell}_k(\mathbf{x}, Y)\} = \mathbf{0}$ and the elements of $\mathrm{var}\{\mathrm{vec}(\boldsymbol{\ell}_k(\mathbf{x}, Y))\}$ are bounded, for $k = 1, 2$, where $\mathrm{vec}(\mathbf{M})$ is the vector formed by concatenating the columns of matrix $\mathbf{M}$.

Theorem 1. *Suppose $E(\mathbf{x}) = \mathbf{0}$, $\mathrm{var}(\mathbf{x}) = \mathbf{I}_p$, and condition* (C1) *holds. Let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{C}^p$ be a complex-valued zero-mean Gaussian process with covariance function*

$$\begin{aligned}\mathrm{cov}_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{s}_0) \overset{\text{def}}{=} \; & E\Big[\big\{\boldsymbol{\varepsilon} \exp(i\mathbf{s}^T\mathbf{B}^T\mathbf{x}) - \boldsymbol{\ell}_2(\mathbf{x}, Y)\mathbf{g}(\mathbf{s}) + \mathbf{h}(\mathbf{s})\boldsymbol{\ell}_1(\mathbf{x}, Y)\mathbf{s}\big\} \\ & \big\{\boldsymbol{\varepsilon} \exp(-i\mathbf{s}_0^T\mathbf{B}^T\mathbf{x}) - \boldsymbol{\ell}_2(\mathbf{x}, Y)\mathbf{g}(-\mathbf{s}_0) - \mathbf{h}(-\mathbf{s}_0)\boldsymbol{\ell}_1(\mathbf{x}, Y)\mathbf{s}_0\big\}^T\Big],\end{aligned} \tag{3.2}$$

where $\mathbf{g}(\mathbf{s}) \overset{\text{def}}{=} E\{\mathbf{x}\exp(i\mathbf{s}^T\mathbf{B}^T\mathbf{x})\}$ *and* $\mathbf{h}(\mathbf{s}) \overset{\text{def}}{=} E[\boldsymbol{\varepsilon}\{i\cos(\mathbf{s}^T\mathbf{B}^T\mathbf{x}) - \sin(\mathbf{s}^T\mathbf{B}^T\mathbf{x})\}\mathbf{x}^T]$.

1. *Under $H_0 : E(\boldsymbol{\varepsilon} \mid \mathbf{B}^T\mathbf{x}) = E(\boldsymbol{\varepsilon})$ a.s., we have $n\widehat{\omega}_n \xrightarrow{d} \|\boldsymbol{\phi}(\mathbf{s})\|^2$ as n goes to infinity.*

2. *Under $H_1 : E(\boldsymbol{\varepsilon} \mid \mathbf{B}^T\mathbf{x}) \neq E(\boldsymbol{\varepsilon})$ a.s., we have $n\widehat{\omega}_n \xrightarrow{p} \infty$ as n goes to infinity.*

Theorem 1 is similar to Theorem 5 of Székely, Rizzo and Bakirov (2007) and Theorem 4 of Shao and Zhang (2014). We reject H_0 in (2.2) when $\widehat{\omega}_n$ is sufficiently large. The exact form of $\|\boldsymbol{\phi}(\mathbf{s})\|^2$ is very complicated and difficult to use in practice. To approximate the asymptotic distribution of $\widehat{\omega}_n$, we propose the following bootstrap procedure.

S0. Based on an i.i.d. sample $\{(\mathbf{x}_j, Y_j) : j = 1, \ldots, n\}$, use a chosen SDR method to estimate $\mathbf{B} \in \mathbb{R}^{p\times d}$ as $\widehat{\mathbf{B}}$. Compute $\mathbf{P}_{\widehat{\mathbf{B}}} = \widehat{\mathbf{B}}\big(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}}\big)^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$, $\mathbf{Q}_{\widehat{\mathbf{B}}} = \mathbf{I}_p - \mathbf{P}_{\widehat{\mathbf{B}}}$, and $\widehat{\boldsymbol{\varepsilon}}_j = \mathbf{Q}_{\widehat{\mathbf{B}}}\mathbf{x}_j$, for $j = 1, \ldots, n$. Calculate the test statistic $\widehat{\omega}_n$ in (3.1).

S1. In the (t)th iteration, let $\{W_j^{(t)} : j = 1, \ldots, n\}$ be i.i.d. Bernoulli random variables, such that $\Pr(W_j^{(t)} = 1) = \Pr(W_j^{(t)} = -1) = 0.5$. Set $\mathbf{x}_j^{(t)} \overset{\text{def}}{=} \mathbf{P}_{\widehat{\mathbf{B}}}\mathbf{x}_j + W_j^{(t)}\widehat{\boldsymbol{\varepsilon}}_j$, for $j = 1, \ldots, n$.

S2. Based on $\{(\mathbf{x}_j^{(t)}, Y_j) : j = 1, \ldots, n\}$, use the same SDR method as in step S0 to estimate $\mathbf{B}$. Denote the corresponding estimator as $\widehat{\mathbf{B}}^{(t)}$.

S3. Compute $\mathbf{P}_{\widehat{\mathbf{B}}^{(t)}} \overset{\text{def}}{=} \widehat{\mathbf{B}}^{(t)}\big\{(\widehat{\mathbf{B}}^{(t)})^{\mathrm{T}}\widehat{\mathbf{B}}^{(t)}\big\}^{-1}(\widehat{\mathbf{B}}^{(t)})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\varepsilon}}_j^{(t)} \overset{\text{def}}{=} \mathbf{x}_j^{(t)} - \mathbf{P}_{\widehat{\mathbf{B}}^{(t)}}\mathbf{x}_j^{(t)}$, for $j = 1, \ldots, n$. Let

$$\overline{\boldsymbol{\varepsilon}}^{(t)} \overset{\text{def}}{=} n^{-1}\sum_{j=1}^{n}\widehat{\boldsymbol{\varepsilon}}_j^{(t)},$$

and calculate

$$\widehat{\omega}_n^{(t)} \overset{\text{def}}{=} -n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n}\Big\{(\widehat{\boldsymbol{\varepsilon}}_j^{(t)} - \overline{\boldsymbol{\varepsilon}}^{(t)})^{\mathrm{T}}(\widehat{\boldsymbol{\varepsilon}}_k^{(t)} - \overline{\boldsymbol{\varepsilon}}^{(t)}) \mid (\widehat{\mathbf{B}}^{(t)})^{\mathrm{T}}(\mathbf{x}_j^{(t)} - \mathbf{x}_k^{(t)}) \mid_d\Big\}.$$

S4. Repeat S1–S3 T times. Calculate the p-value, defined as

$$T^{-1}\sum_{t=1}^{T}\mathbf{1}(\widehat{\omega}_n < \widehat{\omega}_n^{(t)}),$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. For a given significance level α, reject $H_0 : E(\boldsymbol{\varepsilon} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = E(\boldsymbol{\varepsilon})$ if the p-value is less than α.

The validity of the bootstrap procedure is guaranteed by the next theorem. Define $\mathbf{x}^* \overset{\text{def}}{=} \mathbf{P}_{\widehat{\mathbf{B}}}\mathbf{x} + W^*\mathbf{Q}_{\widehat{\mathbf{B}}}\mathbf{x}$, where W^* is a Bernoulli random variable, such that $\Pr(W^* = 1) = \Pr(W^* = -1) = 0.5$. It follows that $\{(\mathbf{x}_j^{(t)}, Y_j) : j = 1, \ldots, n\}$ is an i.i.d. sample of $(\mathbf{x}^*, Y)$. The following technical conditions are needed before we state the main result.

(C2). Suppose

$$\widehat{\mathbf{B}}^{(t)} - \widehat{\mathbf{B}} = n^{-1}\sum_{j=1}^{n}\boldsymbol{\ell}_1(\mathbf{x}_j^{(t)}, Y_j) + o_p(n^{-1/2}) \text{ and}$$

$$\mathbf{P}_{\widehat{\mathbf{B}}^{(t)}} - \mathbf{P}_{\widehat{\mathbf{B}}} = n^{-1}\sum_{j=1}^{n}\boldsymbol{\ell}_2(\mathbf{x}_j^{(t)}, Y_j) + o_p(n^{-1/2}).$$

Assume $E\{\boldsymbol{\ell}_k(\mathbf{x}^*, Y)\} = \mathbf{0}$ and the elements of $\mathrm{var}\{\mathrm{vec}(\boldsymbol{\ell}_k(\mathbf{x}^*, Y))\}$ are bounded, for $k = 1, 2$.

(C3). Let $\boldsymbol{\phi}^* : \mathbb{R}^d \to \mathbb{C}^p$ be a complex-valued zero-mean Gaussian process with covariance function

$$\begin{aligned}&\mathrm{cov}_{\boldsymbol{\phi}^*}(\mathbf{s}, \mathbf{s}_0)\\ &\overset{\mathrm{def}}{=} E\Big[\big\{\boldsymbol{\varepsilon}\exp(i\mathbf{s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}^*) - \boldsymbol{\ell}_2(\mathbf{x}^*, Y)\mathbf{g}^*(\mathbf{s}) + \mathbf{h}^*(\mathbf{s})\boldsymbol{\ell}_1(\mathbf{x}^*, Y)\mathbf{s}\big\}\\ &\qquad \big\{\boldsymbol{\varepsilon}\exp(-i\mathbf{s}_0^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}^*) - \boldsymbol{\ell}_2(\mathbf{x}^*, Y)\mathbf{g}^*(-\mathbf{s}_0) - \mathbf{h}^*(-\mathbf{s}_0)\boldsymbol{\ell}_1(\mathbf{x}^*, Y)\mathbf{s}_0\big\}^{\mathrm{T}}\Big],\end{aligned}$$

where $\mathbf{g}^*(\mathbf{s}) \overset{\mathrm{def}}{=} E\{\mathbf{x}^*\exp(i\mathbf{s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}^*)\}$ and $\mathbf{h}^*(\mathbf{s}) \overset{\mathrm{def}}{=} E[\boldsymbol{\varepsilon}\{i\cos(\mathbf{s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}^*) - \sin(\mathbf{s}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}^*)\}(\mathbf{x}^*)^{\mathrm{T}}]$. Suppose $\mathrm{cov}_{\boldsymbol{\phi}^*}(\mathbf{s}, \mathbf{s}_0)$ is equal to $\mathrm{cov}_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{s}_0)$ defined in (3.2), as long as $E(\mathbf{x}^*) = E(\mathbf{x})$ and $\mathrm{var}(\mathbf{x}^*) = \mathrm{var}(\mathbf{x})$.

(C4) Assume that $\psi(\mathbf{B}) \overset{\mathrm{def}}{=} E(\mathbf{Q}_{\mathbf{B}}\boldsymbol{\varepsilon}^* \mid \mathbf{P}_{\mathbf{B}}\mathbf{x}^*)$ is Lipschitz continuous.

Theorem 2. *Suppose* $E(\mathbf{x}) = \mathbf{0}$, $\mathrm{var}(\mathbf{x}) = \mathbf{I}_p$, *and conditions* (C1)–(C4) *hold. Then,* $\widehat{\omega}_n^{(t)}$ *has the same asymptotic null distribution as* $\widehat{\omega}_n$. *Specifically,* $n\widehat{\omega}_n^{(t)} \xrightarrow{d} \|\boldsymbol{\phi}(\mathbf{s})\|^2$ *as* n *goes to infinity.*

4. An Extension

If the predictor dimension p is very large, we assume, under the principle of sparsity, that $Y \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}_{\mathcal{A}_1}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}_1}$ when the central space is considered, and $Y \perp\!\!\!\perp E(Y \mid \mathbf{x}) \mid \boldsymbol{\alpha}_{\mathcal{A}_2}^{\mathrm{T}}\mathbf{x}_{\mathcal{A}_2}$ when the central mean space is considered, where

$$\begin{aligned}\mathcal{A}_1 &\overset{\mathrm{def}}{=} \{k \mid F(y \mid \mathbf{x}) \text{ relies functionally on } X_k \ \text{ for } \ y \in \mathbb{R},\, k = 1, \ldots, p\},\\ \mathcal{A}_2 &\overset{\mathrm{def}}{=} \{k \mid E(y \mid \mathbf{x}) \text{ relies functionally on } X_k \ \text{ for } \ y \in \mathbb{R},\, k = 1, \ldots, p\},\end{aligned}$$

and $F(y \mid \mathbf{x})$ and $E(y \mid \mathbf{x})$ are the respective conditional distribution and conditional mean functions of Y, given $\mathbf{x}$. To ease subsequent presentation, we use $\mathcal{A}$ to denote either $\mathcal{A}_1$ or $\mathcal{A}_2$, and $\mathbf{B}_{\mathcal{A}}$ to denote either $\boldsymbol{\beta}_{\mathcal{A}_1}$ or $\boldsymbol{\alpha}_{\mathcal{A}_2}$.

When p is moderately large, we can first apply sparse SDR methods, such as those of Li (2007), Bondell and Li (2009), and Chen, Zou and Cook, (2010), to simultaneously select the variables (i.e., estimate the active index set $\mathcal{A}$) and reduce the dimension (i.e., estimate the basis matrix $\mathbf{B}_{\mathcal{A}}$).

When p is extremely large, we recommend using a model-free screening approach, such as the SIRS of Zhu et al. (2011), DC-SIS of Li, Zhong and Zhu (2012), or MDC-SIS of Shao and Zhang (2014), to exclude as many inactive predictors as possible before using SDR methods to further reduce the predictor dimension. Once the number of active predictors is reduced to a moderate scale, we implement sparse SDR methods to obtain consistent estimators of $\mathcal{A}$ and $\mathbf{B}_{\mathcal{A}}$. The proposed test procedure is based on $\mathbf{x}_{\widehat{\mathcal{A}}}$ and $\widehat{\mathbf{B}}_{\mathcal{A}}$, and remains valid as long as the estimate of $\mathbf{B}_{\mathcal{A}}$ is consistent.

We advocate a two-stage test procedure in the high-dimensional case. We randomly split the full sample $\mathcal{D}$ into two equal halves, $\mathcal{D}_1$ and $\mathcal{D}_2$. First, we implement DC-SIS (Li, Zhong and Zhu (2012)) on data set $\mathcal{D}_1$, and retain the top-ranked covariates as the active ones. Next, we implement the sparse SDR method of Li (2007) on data set $\mathcal{D}_2$ to estimate $\mathcal{A}$ and $\mathbf{B}_{\mathcal{A}}$. We conduct our test procedure based on $\mathbf{x}_{\widehat{\mathcal{A}}}$ and $\widehat{\mathbf{B}}_{\mathcal{A}}$. We adopt a data-splitting strategy to avoid inflating type-I error rates in our test procedure. When some inactive covariates are retained in the screening stage, directly implementing our testing procedure without random splitting leads to inflated type-I error rates (Fan, Guo and Hao (2012)).

5. Numerical Studies

Example 1. We conduct simulations to demonstrate the performance of our proposed test. We fix the sample size at $n = 200$. We evaluate the predictor dimension $p = 8$ for the low-dimensional case, and $p =$ 1,000 for the high-dimensional case. We consider two models.

(I): In the first model, the central space is spanned by $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$ and $Y = X_1 + \delta$. Thus, $d = 1$. The predictors $\mathbf{x} = (X_1, \ldots, X_p)$ are generated as follows: $X_1, X_3, \ldots, X_p$ are drawn independently from a standard normal distribution, and $X_2 = X_1 + c_1(X_1^2 - 1) + |c_2 X_1 + 1|\epsilon$.

(II): In the second model, the central space is spanned by $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$ and $(0, 1, 0, \ldots, 0)^{\mathrm{T}}$, and $Y = 5X_1/\{0.5 + (X_2 + 1.5)^2\} + \delta$. Thus, $d = 2$. The predictors $\mathbf{x} = (X_1, \ldots, X_p)$ are generated as follows: $X_1, X_2, X_4, \ldots, X_p$ are drawn independently from a standard normal distribution, and $X_3 = X_1 + X_2 + c_1(X_1^2 - 1) + |c_2 X_2 + 1|\epsilon$.

In both models, we generate ϵ and δ independently from a standard normal distribution.

We first evaluate the performance of testing the LCM condition. We fix $c_2 = 0$ and evaluate $c_1 = 0, 0.1, \ldots, 0.5$. To illustrate the performance of our proposed method, we evaluate two test statistics: (a) based on the observed data set $(\mathbf{x}_i, Y_i)_{i=1}^n$, we estimate $\mathbf{B}$ using the SIR method to obtain $\widehat{\mathbf{B}}^{sir}$, and then construct the test statistics by replacing $\widehat{\mathbf{B}}$ with $\widehat{\mathbf{B}}^{sir}$ in (3.1); (b) we suppose the true $\mathbf{B}$ matrix is known as a prior, and then construct the test statistics by replacing $\widehat{\mathbf{B}}$ with $\mathbf{B}$ in (3.1), which acts as a benchmark.

When evaluating the performance of the joint test (i.e., simultaneously testing the LCM and CCV conditions), we consider $c_1 = c_2 = 0, 0.1, \ldots, 0.5$. We also evaluate two test statistics: (a) test statistics based on $\widehat{\mathbf{B}}^{save}$, where $\widehat{\mathbf{B}}^{save}$ is obtained using the SAVE method; (b) test statistics based on the true $\mathbf{B}$ matrix by assuming that it is known as a prior.

To put our test procedure into practice, we can choose estimates of $\mathbf{B}$ that do not rely on the linear mean or constant variance conditions, such as those of Xia et al. (2002), Xia (2007), Fukumizu, Bach and Jordan (2009), Li and Dong (2009), Dong and Li (2010), and Ma and Zhu (2012), among others. However, these estimates are usually computationally expensive compared with the classical SIR and SAVE methods. Therefore, we simply suggest estimating $\mathbf{B}$ using a method that can be easily computed. Here, we implement SIR or SAVE to obtain $\widehat{\mathbf{B}}$, and then use our proposed test based on this $\widehat{\mathbf{B}}$ to check whether the LCM condition or the joint conditions hold. If the null hypothesis is not rejected, then we are confident that $\widehat{\mathbf{B}}$ is valid. If the null hypothesis is rejected, we can choose other methods that avoid the LCM condition or the joint conditions in order to re-estimate $\mathbf{B}$. However, in doing so, we may lose power because $E(\boldsymbol{\varepsilon} \mid \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x})$ may be very close to $E(\boldsymbol{\varepsilon})$ for a lousy estimate $\widehat{\mathbf{B}}$ obtained under the alternative hypothesis.

We decide whether to reject the null hypothesis using the bootstrap procedures with $T = 500$. We repeat each experiment 500 times and study the size and the power of the our tests separately.

We first evaluate the size of the test. Note that the LCM condition holds if and only if $c_1 = 0$, and the joint condition holds if and only if $c_1 = c_2 = 0$. We thus fix $c_1 = c_2 = 0$ to study the size of all tests. We investigate different significance levels, with $\alpha = 0.01, 0.02, 0.05$. The empirical sizes based on 500 repetitions are summarized in Table 1, which indicates that tests based on $\widehat{\mathbf{B}}$ behave similarly to those based on the true $\mathbf{B}$ matrix, and the empirical sizes are close to the nominal level α.

We then study the power performance of the test procedures. We fix $c_2 = 0$ and evaluate $c_1 = 0.1, 0.2, \ldots, 0.5$ when testing the LCM condition, and $c_1 = c_2 = 0.1, 0.2, \ldots, 0.5$ when testing the joint condition LCM+CCV. We fix the significance level $\alpha = 0.05$; the results are summarized in Table 2.

Table 2 indicates that our proposed method performs satisfactorily. In general, the power values of the tests gradually increase to one when c_1 goes up from 0.1 to 0.5. In the low-dimensional case $p = 8$, the power values of the LCM test

Table 1. The empirical sizes of the test procedures when $c_1 = c_2 = 0$.

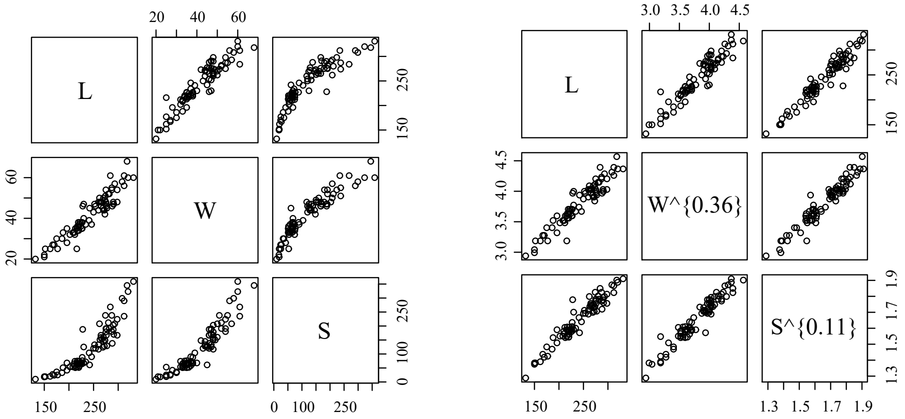
p	Test	Method	$d = 1$			$d = 2$		
			α			α		
			0.01	0.02	0.05	0.01	0.02	0.05
$p = 8$	LCM	$\widehat{\mathbf{B}}^{sir}$	0.02	0.03	0.04	0.01	0.02	0.05
		$\mathbf{B}$	0.01	0.03	0.05	0.01	0.01	0.04
	LCM+CCV	$\widehat{\mathbf{B}}^{save}$	0.01	0.03	0.05	0.01	0.02	0.04
		$\mathbf{B}$	0.01	0.02	0.05	0.01	0.02	0.04
$p = 1{,}000$	LCM	$\widehat{\mathbf{B}}^{sir}$	0.01	0.02	0.05	0.01	0.02	0.05
		$\mathbf{B}$	0.01	0.02	0.05	0.01	0.02	0.06
	LCM+CCV	$\widehat{\mathbf{B}}^{save}$	0.01	0.01	0.04	0.01	0.02	0.04
		$\mathbf{B}$	0.01	0.02	0.05	0.01	0.02	0.04

Table 2. The empirical power of the test procedures with $\alpha = 0.05$.

p	Test		$d = 1$					$d = 2$				
8	LCM	c_1	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
		$\widehat{\mathbf{B}}^{sir}$	0.13	0.54	0.87	0.99	1.00	0.10	0.49	0.87	0.98	1.00
		$\mathbf{B}$	0.17	0.65	0.95	1.00	1.00	0.14	0.49	0.88	0.99	1.00
	LCM+CCV	$c_1 = c_2$	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
		$\widehat{\mathbf{B}}^{save}$	0.12	0.49	0.85	0.98	1.00	0.07	0.28	0.59	0.78	0.88
		$\mathbf{B}$	0.15	0.51	0.87	0.99	1.00	0.10	0.40	0.80	0.98	1.00
1,000	LCM	c_1	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
		$\widehat{\mathbf{B}}^{sir}$	0.09	0.46	0.85	0.97	1.00	0.12	0.45	0.86	0.97	1.00
		$\mathbf{B}$	0.16	0.59	0.94	1.00	1.00	0.14	0.49	0.87	0.98	1.00
	LCM+CCV	$c_1 = c_2$	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
		$\widehat{\mathbf{B}}^{save}$	0.09	0.43	0.80	0.97	1.00	0.06	0.25	0.48	0.69	0.80
		$\mathbf{B}$	0.12	0.45	0.89	0.99	1.00	0.09	0.36	0.78	0.97	1.00

and the joint test LCM+CCV both exceed 0.85 when the signal intensity parameter c_1 increases to 0.3 in Model (I), where the structure dimension $d = 1$, and finally reaches one when $c_1 = 0.5$. The results for Model (II), where $d = 2$, are quite similar, though a little inferior to those for Model (I). This is reasonable because it is a more complicated problem in SDR when the structure dimension increases. The results are similar in the ultrahigh-dimensional case when $p = 1{,}000$.

Example 2. We apply our proposed method to a horse mussels data set, provided by Mike Camden, Wellington Polytechnic, Wellington, New Zealand. The response variable Y is the mussels' muscle mass M, the edible portion of the mussel, which is measured in grams. The covariates $\mathbf{x}$ include the shell length L in millimeters, shell width W in millimeters, and shell mass S in grams. The



(a) $\mathbf{x} = (L, W, S)^{\mathrm{T}}$ (b) $\mathbf{x}^* = (L, W^{0.36}, S^{0.11})^{\mathrm{T}}$

Figure 1. Scatter plot matrices for covariate vectors $\mathbf{x}$ and $\mathbf{x}^*$.

sample size is 82.

We first visually evaluate whether the LCM condition holds. The scatter plot matrix of the shell length L, shell width W, and shell mass S is presented in Figure 1 (A). From Figure 1, the curvature between L and S, together with that between W and S raises doubts about the LCM condition required by the SIR method. Thus, Cook (1998) used power transformations of the covariates to make the LCM condition hold approximately. By applying the maximum likelihood estimates, the shell width W is transformed to $W^{0.36}$, the shell mass S is transformed to $S^{0.11}$, and the shell length L is not transformed. The scatter plot matrix after the transformations is shown in Figure 1 (B). It seems that the LCM condition holds after the transformation.

Now, we apply our proposed method to test whether the LCM condition holds. Recall that $\mathbf{x} = (L, W, S)^{\mathrm{T}}$ is the covariate vector before the transformation. We set $\mathbf{x}^* = (L, W^{0.36}, S^{0.11})^{\mathrm{T}}$ as the covariate vector after the transformation. Then, we need to test the LCM condition for data sets $(\mathbf{x}_i, Y_i)_{i=1}^{82}$ and $(\mathbf{x}_i^*, Y_i)_{i=1}^{82}$. Given the structure dimension $d = 1$, our proposed method can be applied directly . Specifically, we first apply the SIR method to the original data set $(\mathbf{x}_i, Y_i)_{i=1}^{82}$ to estimate $\mathbf{B}$, where $\widehat{\mathbf{B}}$ denotes the corresponding estimator. We then carry out the test procedure based on $\mathbf{x}_i$ and $\widehat{\mathbf{B}}$, obtaining the p-value 0.000. Then, we test the LCM condition on the transformed data set $(\mathbf{x}_i^*, Y_i)_{i=1}^{82}$. We estimate $\mathbf{B}$ using the SIR method to obtain $\widehat{\mathbf{B}}^*$, and then conduct the test based on $\mathbf{x}_i^*$ and $\widehat{\mathbf{B}}^*$. The p-value of the test is 0.908.

From the above tests, we can confidently reject the null hypothesis that the LCM condition holds for dataset $(\mathbf{x}_i, Y_i)_{i=1}^{82}$, but accept that it holds for the data set $(\mathbf{x}_i^*, Y_i)_{i=1}^{82}$. That is, the LCM condition is violated on the original data set, but holds after the power transformation proposed by Cook (1998). Such results are in accordance with those shown in Figure 1, thus proving the validity of our proposed method.

To determine whether the power transformation helps to obtain more accurate estimators, we conduct a simple bootstrap procedure, as follows. For the original data $(\mathbf{x}_i, Y_i)_{i=1}^{82}$, we estimate $\mathbf{B}$ using the SIR method to obtain $\widehat{\mathbf{B}}$, which we treat as the true $\mathbf{B}$. Then, we bootstrap from the original data 500 times, obtaining $\widehat{\mathbf{B}}^{(t)}$ using the SIR, where $t = 1, \ldots, 500$. To assess the distance between $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}^{(t)}$, we adopt the trace correlation proposed in Ferré (1998) to obtain $r(d)^{(t)}$, for $t = 1, \ldots, 500$. Based on 500 repetitions, the average $r(d)^{(t)}$ is 0.74, and the standard deviation is 0.21. Similarly, for the transformed data $(\mathbf{x}_i^*, Y_i)_{i=1}^{82}$, we get $r(d)^{*(t)}$, for $t = 1, \ldots, 500$, with an average of 0.95 and a standard deviation of 0.06. According to Ferré (1998), a trace correlation $r(d) \in [0, 1]$ and a larger value indicates that the two subspaces are closer together. Thus, the power transformation results in more accurate estimators when the LCM fails.

Supplementary Material

The online supplementary material contains the proofs of Propositions 1 and 2 and Theorems 1 and 2.

Acknowledgments

Zhou's work was supported by the National Natural Science Foundation of China (11801501). Zhu is the corresponding author, and his work was supported by the Beijing Natural Science Foundation (Z19J00009) and the National Natural Science Foundation of China (11731011, 11931014).

References

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 287–299.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**, 3696–3723.

Chiaromonte, F. and Cook, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics* **54**, 768–795.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association* **89**, 592–599.

Cook, R. D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 28–33.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika* **97**, 279–294.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory* **22**, 1030–1051.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of The Royal Statistical Society: Series B (Statistical Methodology)* **74**, 37–65.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association* **93**, 132–140.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics* **37**, 1871–1905.

Guo, X., Wang, T. and Zhu, L. (2016). Model checking for parametric single-index models: A dimension reduction model-adaptive approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 1013–1035.

Koul, H. L. and Ni, P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference* **119**, 109–141.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**, 603–613.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics* **37**, 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025–1039.

Li, Q. (1999). Consistent model specification tests for time series econometric models. *Journal of Econometrics* **92**, 101–147.

Li, Q., Hsiao, C. and Zinn, J. (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics* **112**, 295–325.

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

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Ma, Y. and Zhu, L. P. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *Annals of statistics* **41**, 250–258.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association* **109**, 1302–1318.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**, 613–641.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariances. *The Annals of Statistics* **3**, 1236–1265.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

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

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space, with discussion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 363–410.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* **75**, 263–289.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Zhu, L. P., Zhu, L. X. and Feng, Z. F. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* **105**, 1455–1466.

Tingyou Zhou

School of Data Sciences, Zhejiang University of Finance and Economics, 18 Xueyuan Road, Xiasha High Education District, Hangzhou 310018, China.

E-mail: zhoutingyou@zufe.edu.cn

Yuexiao Dong

Department of Statistical Science, Fox School of Business, Temple University, 1801 Liacouras Walk, Philadelphia, PA 19122, USA.

E-mail: ydong@temple.edu

Liping Zhu

Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

School of Statistics and Mathmatics, Zhejiang Gongshang University, Hangzhou 310018, China.

E-mail: zhu.liping@ruc.edu.cn

(Received March 2019; accepted May 2020)