Statistica Sinica **31** (2021), 1055-1079
doi:https://doi.org/10.5705/ss.202019.0311

OMNIBUS MODEL CHECKS OF LINEAR ASSUMPTIONS THROUGH DISTANCE COVARIANCE

Kai Xu and Daojiang He

Anhui Normal University

Abstract: Although the adequacy of linearity is well researched in the statistical literature, few studies examine this topic from the viewpoint of a measure of association. Inspired by the well-known distance covariance (dCov), we propose two omnibus tests for the goodness-of-fit of linearity. Methodologically, our tests do not include any tuning parameters and are conveniently implemented. The theoretical details are of independent interest, mainly because the kernel induced by the dCov is not smooth. We investigate the convergence of our tests under null, fixed, and local alternative hypotheses, and devise a bootstrap scheme to approximate their null distributions, showing that its consistency is justified. Numerical studies demonstrate the effectiveness of our proposed tests relative to that of several existing tests.

Key words and phrases: Bootstrap, distance covariance, goodness-of-fit test, linearity.

1. Introduction

Let $(\mathbf{x}, Y)$ be a random vector in $\mathbb{R}^{d+1}$ such that Y has a finite expectation. Denote $m(\mathbf{x}) = E(Y|\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$. In a regression analysis, the relationship between Y and $\mathbf{x}$ can always be written as

$$Y = m(\mathbf{x}) + \eta, \tag{1.1}$$

where $m(\cdot)$ is the regression function, $\mathbf{x}$ is a d-dimensional predictor, and $\eta = Y - m(\mathbf{x})$ is the error, which has conditional mean zero, given $\mathbf{x}$. To motivate our procedure, we further assume that η is independent of $\mathbf{x}$. If there is no priori preference of the independence between η and $\mathbf{x}$, the method proposed in this paper can be used to simultaneously test independence and goodness-of-fit in the regression model (1.1)

Within a parametric framework, one often assumes that $m(\cdot)$ belongs to the following parametric class:

Corresponding author: Kai Xu, School of Mathematics and Statistics, Anhui Normal University, Wuhu 241002, China. E-mail: tjxxukai@163.com.

$$\mathcal{M}_{\boldsymbol{\beta}} = \{\mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^d\}, \tag{1.2}$$

where $\mathbf{g}(\mathbf{x}) = \{g_1(\mathbf{x}), \ldots, g_d(\mathbf{x})\}^{\mathrm{T}}$ is the vector of known predictor functions, and $\boldsymbol{\beta}$ is the unknown parameter vector. Linearity provides a simple parametric approach to directly model the effects of a set of covariates on a response. Often we need to determine the adequacy of such parametric fits. The null hypothesis is

$$H_0 : m(\mathbf{x}) \in \mathcal{M}_{\boldsymbol{\beta}}, \tag{1.3}$$

whereas the alternative to be tested is

$$H_1 : m(\mathbf{x}) \notin \mathcal{M}_{\boldsymbol{\beta}}. \tag{1.4}$$

Much of the recent statistical literature has focused on checking the adequacy of parametric models. These methods include, but are not limited to, those of Bierens (1982), Zheng (1996), Stute (1997), Fan and Huang (2001), Khmaladze and Koul (2004), Koul and Ni (2004), Guerre and Lavergne (2005), Escanciano (2006), Stute, Xu and Zhu (2008), and Christensen and Sun (2010). Refer to González-Manteiga and Crujeiras (2013) for an excellent recent review. However, few methods examine this problem from the viewpoint of a measure of association. To the best of our knowledge, the recent work by Sen and Sen (2014) was the first to employ the Hilbert-Schmidt independence criterion (Gretton et al. (2008, HSIC)) to construct an omnibus goodness-of-fit test for linear models when η is independent of $\mathbf{x}$. Specifically, their proposed method uses the HSIC to test for independence between the predictor and the residual obtained from the parametric fit. Let $\|\cdot\|$ denote the Euclidean norm. A commonly used kernel associated with the HSIC is the Gaussian kernel $k(\mathbf{w}, \mathbf{w}') = \exp\{-\|\mathbf{w}-\mathbf{w}'\|^2/(2\gamma^2)\}$, which is related to a bandwith parameter γ. As a dependence metric, choosing a good bandwidth parameter remains an important open problem; see Yao, Zhang and Shao (2018) for related discussions in other contexts.

We propose two omnibus tests for the adequacy of linearity. Both use the popular distance covariance (Székely, Rizzo and Bakirov (2007, dCov)), which measures the distance between the joint characteristic function of two random vectors of arbitrary dimensions and the product of their marginal characteristic functions in terms of a weighted $\mathcal{L}^2$-norm. Refer to Li, Zhong and Zhu (2012), Matteson and Tsay (2017), and Yao, Zhang and Shao (2018) for other applications that use a dCov-based dependence measure. The proposed tests differ from that of Sen and Sen (2014) in three major respects. First, our tests do not include bandwidth parameters, such as the choice of γ, and are easy to implement.

Second, the techniques developed by Sen and Sen (2014) are not directly extendable to our framework without serious technical work, largely because the kernel induced by the dCov is not differentiable. As such, we cannot apply techniques such as the Taylor expansion, which are essential for establishing the asymptotic theory of the HSIC-based lack-of-fit test. Third, we develop a distribution theory for our test statistics based on the theory of U-statistics indexed by parameters (Sherman (1994)). In particular, we study the limiting distribution of our test statistics under local alternative hypotheses, which was not considered in Sen and Sen (2014).

The rest of the paper is organized as follows. In Section 2, we review the distance covariance and its sample estimates. In Section 3, we discuss the test statistics and their asymptotic properties. Here we also propose a bootstrap procedure that approximates the asymptotic critical values of the test statistics, and prove its consistency. In Section 4, we examine the finite-sample performance of our proposed tests using Monte Carlo simulations. Section 5 concludes the paper. Several technical proofs are presented in Section 6. The remaining technical proofs, as well as additional numerical results, are gathered in the online Supplementary Material.

2. Preliminaries

The dCov introduced by Székely, Rizzo and Bakirov (2007) is a multivariate measure of independence between d_1- and d_2-dimensional random vectors $\mathbf{u}$ and $\mathbf{v}$, with $E(\|\mathbf{u}\| + \|\mathbf{v}\|) < \infty$, where d_1 and d_2 can be arbitrary. The nonnegative number $dCov(\mathbf{u}, \mathbf{v})$ is defined as the positive square root of

$$dCov(\mathbf{u}, \mathbf{v})^2 = \frac{1}{c_{d_1} c_{d_2}} \int_{\mathbb{R}^{d_1+d_2}} \frac{|\phi_{\mathbf{u},\mathbf{v}}(\boldsymbol{t}, \boldsymbol{s}) - \phi_{\mathbf{u}}(\boldsymbol{t})\phi_{\mathbf{v}}(\boldsymbol{s})|^2}{\|\boldsymbol{t}\|^{1+d_1} \|\boldsymbol{s}\|^{1+d_2}} d\boldsymbol{t} d\boldsymbol{s},$$

where $\phi_{\mathbf{u}}$ and $\phi_{\mathbf{v}}$ are the individual characteristic functions of $\mathbf{u}$ and $\mathbf{v}$, respectively, $\phi_{\mathbf{u},\mathbf{v}}$ is the joint characteristic function of $\mathbf{u}$ and $\mathbf{v}$, $c_d = \pi^{(1+d)/2}/\Gamma\{(1 + d)/2\}$, and $\Gamma(\cdot)$ is a gamma function. The dCov enjoys many desirable properties (Székely, Rizzo and Bakirov (2007); Székely and Rizzo (2012, 2013)). In particular, $dCov(\mathbf{u}, \mathbf{v})^2 = 0$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are independent.

To obtain a practical estimator for the squared dCov, Székely and Rizzo (2009) state that

$$dCov(\mathbf{u}, \mathbf{v})^2 = E\{U(\mathbf{u}, \mathbf{u}')V(\mathbf{v}, \mathbf{v}')\},$$

where $U(\mathbf{u}, \mathbf{u}') = \|\mathbf{u} - \mathbf{u}'\| - E(\|\mathbf{u} - \mathbf{u}'\| \mid \mathbf{u}) - E(\|\mathbf{u} - \mathbf{u}'\| \mid \mathbf{u}') + E(\|\mathbf{u} - \mathbf{u}'\|)$,

$V(\mathbf{v},\mathbf{v}') = \|\mathbf{v}-\mathbf{v}'\| - E(\|\mathbf{v}-\mathbf{v}'\| \mid \mathbf{v}) - E(\|\mathbf{v}-\mathbf{v}'\| \mid \mathbf{v}') + E(\|\mathbf{v}-\mathbf{v}'\|)$, and $(\mathbf{u}',\mathbf{v}')$ is an independent and identically distributed (i.i.d.) copy of $(\mathbf{u},\mathbf{v})$. Suppose that $(\mathbf{u}_i,\mathbf{v}_i)$, for $i=1,\ldots,n$, is a random sample from the population $(\mathbf{u},\mathbf{v})$. Define

$$A_{ij} = \|\mathbf{u}_i-\mathbf{u}_j\| - n^{-1}\sum_{k=1}^{n}\|\mathbf{u}_i-\mathbf{u}_k\| - n^{-1}\sum_{l=1}^{n}\|\mathbf{u}_j-\mathbf{u}_l\| + n^{-2}\sum_{k,l=1}^{n}\|\mathbf{u}_k-\mathbf{u}_l\|,$$

$$B_{ij} = \|\mathbf{v}_i-\mathbf{v}_j\| - n^{-1}\sum_{k=1}^{n}\|\mathbf{v}_i-\mathbf{v}_k\| - n^{-1}\sum_{l=1}^{n}\|\mathbf{v}_j-\mathbf{v}_l\| + n^{-2}\sum_{k,l=1}^{n}\|\mathbf{v}_k-\mathbf{v}_l\|.$$

Then a natural estimator for $dCov(\mathbf{u},\mathbf{v})^2$ can be defined as

$$dCov_{1,n}(\mathbf{u},\mathbf{v})^2 = n^{-2}\sum_{i,j=1}^{n} A_{ij}B_{ij}. \tag{2.1}$$

Székely, Rizzo and Bakirov (2007) showed that $dCov_{1,n}(\mathbf{u},\mathbf{v})^2$ is a V-type statistic. That is, $dCov_{1,n}(\mathbf{u},\mathbf{v})^2$ is a biased estimate of $dCov(\mathbf{u},\mathbf{v})^2$. In practice, researchers are sometimes interested in U-type statistics. Székely and Rizzo (2014) constructed an unbiased estimator defined as

$$dCov_{2,n}(\mathbf{u},\mathbf{v})^2 = \{n(n-3)\}^{-1}\sum_{i\neq j}^{n}\mathcal{A}_{ij}\mathcal{B}_{ij}, \tag{2.2}$$

where

$$\begin{aligned}\mathcal{A}_{ij} &= \|\mathbf{u}_i-\mathbf{u}_j\| - (n-2)^{-1}\sum_{k=1}^{n}\|\mathbf{u}_i-\mathbf{u}_k\| - (n-2)^{-1}\sum_{l=1}^{n}\|\mathbf{u}_j-\mathbf{u}_l\| \\ &\quad + \{(n-1)(n-2)\}^{-1}\sum_{k,l=1}^{n}\|\mathbf{u}_k-\mathbf{u}_l\|, \\ \mathcal{B}_{ij} &= \|\mathbf{v}_i-\mathbf{v}_j\| - (n-2)^{-1}\sum_{k=1}^{n}\|\mathbf{v}_i-\mathbf{v}_k\| - (n-2)^{-1}\sum_{l=1}^{n}\|\mathbf{v}_j-\mathbf{v}_l\| \\ &\quad + \{(n-1)(n-2)\}^{-1}\sum_{k,l=1}^{n}\|\mathbf{v}_k-\mathbf{v}_l\|.\end{aligned}$$

Based on (2.2) and (2.1), we construct two statistics for testing (1.3).

3. Method

3.1. The statistics

Consider the regression model given in (1.1), with $E\{m^2(\mathbf{x})\} < \infty$ and $E(\eta^2) < \infty$. Define

$$\boldsymbol{\beta}_0 = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} E\{Y - \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}\}^2. \tag{3.1}$$

Supposing $E\|\mathbf{g}(\mathbf{x})\|^2 < \infty$, we obtain $\boldsymbol{\beta}_0 = [E\{\mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^{\mathrm{T}}\}]^{-1}E\{\mathbf{g}(\mathbf{x})m(\mathbf{x})\}$. Apparently, under H_0, $m(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}_0$. Assume that we have a random sample $(\mathbf{x}_i, Y_i)$, for $i = 1, \ldots, n$, from model (1.1), and define the unobserved errors $\eta_i = Y_i - m(\mathbf{x}_i)$, for $i = 1, \ldots, n$. We estimate $\boldsymbol{\beta}_0$ using the sample analogue of (3.1), and obtain the following least squares estimator of $\boldsymbol{\beta}_0$:

$$\boldsymbol{\beta}_n = \left\{ n^{-1}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}} \right\}^{-1} n^{-1}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)Y_i, \tag{3.2}$$

when $n^{-1}\sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}$ is invertible. Let $\eta_{in} = Y_i - \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta}_n$ be the resulting residuals. Assume that $\widehat{A}_{ij}$ and $\widehat{B}_{ij}$ and $\widehat{\mathcal{A}}_{ij}$ and $\widehat{\mathcal{B}}_{ij}$ are defined as in (2.1) and (2.2), respectively, except that $\|\mathbf{u}_i - \mathbf{u}_j\|$ and $\|\mathbf{v}_i - \mathbf{v}_j\|$ are replaced by $\|\mathbf{x}_i - \mathbf{x}_j\|$ and the observed $|\eta_{in} - \eta_{jn}|$, respectively. Following the argument presented in Section 2, we propose the following V-type and U-type test statistics:

$$V_n = n^{-2}\sum_{i,j=1}^{n} \widehat{A}_{ij}\widehat{B}_{ij}, \tag{3.3}$$

and

$$U_n = \{n(n-3)\}^{-1}\sum_{i\neq j} \widehat{\mathcal{A}}_{ij}\widehat{\mathcal{B}}_{ij}, \tag{3.4}$$

respectively. In essence, our methods use the dCov to test for independence between the predictor and the residual obtained from the parametric fit. From Corollary 2 of Székely, Rizzo and Bakirov (2007), it is easy to see that under the independence of $\mathbf{x}$ and η, as $n \to \infty$, $n\{dCov_{1,n}(\mathbf{x}, \eta)\}^2$ converges in distribution to

$$\sum_{i=1}^{\infty} \lambda_i \mathcal{Z}_i^2, \tag{3.5}$$

satisfying $\sum_{i=1}^{\infty} \lambda_i = E(|\eta_1 - \eta_2|)E(\|\mathbf{x}_1 - \mathbf{x}_2\|)$, where $\mathcal{Z}_i$ denotes an independent standard normal random variable, and λ_i is a nonnegative constant that depends on the distribution of $(\eta, \mathbf{x})$ and is defined in (3.6). The relations between U and V statistics (Serfling (1980)) show that $n\{dCov_{2,n}(\mathbf{x}, \eta)\}^2$ converges in distribution to $\sum_{i=1}^{\infty} \lambda_i(\mathcal{Z}_i^2 - 1)$. However, η_i is unobserved. In contrast to $\{dCov_{1,n}(\mathbf{x}, \eta)\}^2$ and $\{dCov_{2,n}(\mathbf{x}, \eta)\}^2$, the asymptotic theory of U_n and V_n is not trivial.

3.2. Convergence of U_{n} and V_{n} under null, fixed, and local alternative hypotheses

To obtain asymptotic distributions under the null (1.3) , we make the following mild conditions.

Condition A. Let $f_{\eta^{(2)}_{(1)}}(\cdot)$ denote the density function of $\eta^{(2)}_{(1)} = \eta_1 - \eta_2$. Here, $f_{\eta^{(2)}_{(1)}}(\cdot)$ is bounded in a neighborhood around zero, with $f_{\eta^{(2)}_{(1)}}(0) > 0$ and $f_{\eta^{(2)}_{(1)}}(y) - f_{\eta^{(2)}_{(1)}}(0) = o(1)$ as $y \to 0$.

Condition B. The matrix $\boldsymbol{\Sigma} = E\{\mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^{\mathrm{T}}\}$ is positive definite.

Condition C.

1. $E(\eta^2) < \infty$ and $E\{m^2(\mathbf{x})\} < \infty$;
2. There exists some $\gamma_1 > 0$ such that $E(1+\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2)\{1+\|\mathbf{g}(\mathbf{x}_k)\|^{2+\gamma_1} + \|\mathbf{g}(\mathbf{x}_l)\|^{2+\gamma_1}\} < \infty$, for $1 \le i, j, k, l \le 4$.

To derive the asymptotic distributions under the fixed alternative (1.4), we impose the following additional conditions. Define the error under a model misspecification as $\varepsilon = Y - \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}_0 = m(\mathbf{x}) - \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta$.

Condition D.

1. $f_{\varepsilon^{(2)}_{(1)}}(\cdot)$ is bounded in a neighborhood around zero, where $f_{\varepsilon^{(2)}_{(1)}}(\cdot)$ is the density function of $\varepsilon^{(2)}_{(1)} = \varepsilon_1 - \varepsilon_2$.
2. $E(\varepsilon^2) < \infty$.

These moment assumptions in Conditions A-D are slightly stronger than those stated in Székely, Rizzo and Bakirov (2007). This is the price we pay for dealing with the regression effect on the limiting behavior of dCov. In addition, our conditions are formally related to, but essentially different from those of Sen and Sen (2014). For example, their Conditions 3(d), 4c(ii), and 4c(iv) are not applicable under our framework because the kernel induced by dCov is not differentiable.

Let $\{\phi_i(\cdot)\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$ denote the orthonormal eigenfunctions and the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, respectively, defined in relation to the kernel $h_0^{(2)}(\mathbf{z}_1, \mathbf{z}_2)$, where

$$h_0^{(2)}(\mathbf{z}_1, \mathbf{z}_2) = C_\eta(\eta_1, \eta_2)C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2),$$

with $\mathbf{z}_i = (\eta_i, \mathbf{x}_i)$, for $i = 1, 2$, $C_\eta(\eta_1, \eta_2) = |\eta_1 - \eta_2| - E(|\eta_1 - \eta_2| \mid \eta_1) - E(|\eta_1 - \eta_2| \mid \eta_2) + E(|\eta_1 - \eta_2|)$, and $C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\| - E(\|\mathbf{x}_1 - \mathbf{x}_2\| \mid \mathbf{x}_1) - E(\|\mathbf{x}_1 - \mathbf{x}_2\| \mid \mathbf{x}_2) + E(\|\mathbf{x}_1 - \mathbf{x}_2\|)$. Further, define

$$\boldsymbol{\Lambda} = -E\left[\{\mathbf{g}(\mathbf{x}_1) - E\mathbf{g}(\mathbf{x}_1)\}\{\mathbf{g}(\mathbf{x}_2) - E\mathbf{g}(\mathbf{x}_2)\}^{\mathrm{T}}\|\mathbf{x}_1 - \mathbf{x}_2\|\right], \quad \text{and}$$
$$h_1^{(1)}(\mathbf{z}_1) = 4\{1 - 2F_\eta(\eta_1)\}E[\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_1],$$

where $\boldsymbol{\Lambda}$ is a positive-definite matrix, because it is simply the martingale difference divergence matrix (Lee and Shao (2018)) between $\mathbf{g}(\mathbf{x})$ and $\mathbf{x}$, and F_η is the distribution function of η.

Theorem 1.

1. *Assume that the null hypothesis* (1.3) *holds. Under Conditions A, B, and C, as* $n \to \infty$, nU_n *converges in distribution to*

$$\sum_{i=1}^{\infty} \lambda_i(\mathcal{Z}_i^2 - 1) + \mathcal{N}^{\tau}\boldsymbol{\Sigma}^{-1}\mathcal{W} + 2f_{\eta_{(1)}^{(2)}}(0)\mathcal{W}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\mathcal{W}, \tag{3.6}$$

where each $\mathcal{Z}_i$ *is i.i.d.* $N(0,1)$, *and* $(\mathcal{Z}_i, \mathcal{N}, \mathcal{W}) \in \mathbb{R}^{2d+1}$ *are jointly Gaussian random variables with* $E(\mathcal{Z}_i, \mathcal{N}, \mathcal{W}) = \mathbf{0}_{2d+1}$, $var(\mathcal{Z}_i) = 1$, $var(\mathcal{N}) = var\{h_1^{(1)}(\mathbf{z}_1)\}$, $var(\mathcal{W}) = var\{\mathbf{g}(\mathbf{x}_1)\eta_1\}$, $cov(\mathcal{Z}_i, \mathcal{N}) = cov\{\phi_i(\mathbf{z}_1), h_1^{(1)}(\mathbf{z}_1)\})$, $cov(\mathcal{Z}_i, \mathcal{W}) = cov\{\phi_i(\mathbf{z}_1), \mathbf{g}(\mathbf{x}_1)\eta_1\}$, *and* $cov(\mathcal{N}, \mathcal{W}) = cov\{h_1^{(1)}(\mathbf{z}_1\}, \mathbf{g}(\mathbf{x}_1)\eta_1\}$. *In addition,* nV_n *converges in distribution to*

$$\sum_{i=1}^{\infty} \lambda_i \mathcal{Z}_i^2 + \mathcal{N}^{\tau}\boldsymbol{\Sigma}^{-1}\mathcal{W} + 2f_{\eta_{(1)}^{(2)}}(0)\mathcal{W}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\mathcal{W}. \tag{3.7}$$

2. *Assume that the fixed alternative* (1.4) *holds. Under Conditions A, B, C2, and D, as* $n \to \infty$, *both* $n^{1/2}\{U_n - dCov^2(\varepsilon_1, \mathbf{x}_1)\}$ *and* $n^{1/2}\{V_n - dCov^2(\varepsilon_1, \mathbf{x}_1)\}$ *converge in distribution to a Gaussian random variable with mean zero and variance* $4var[E\{C_\varepsilon(\varepsilon_1, \varepsilon_2)C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)|\varepsilon_1, \mathbf{x}_1\} + \varrho_1^{\tau}\boldsymbol{\Sigma}^{-1}\mathbf{g}(\mathbf{x}_1)\varepsilon_1]$, *with* $\varrho_1 = -E[\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)I(\varepsilon_1 > \varepsilon_2)]$.

The second statement in Theorem 1 asserts that both U_n and V_n converge in probability to $dCov^2(\varepsilon_1, \mathbf{x}_1) > 0$ under the alternative (1.4). Therefore, both nU_n and nV_n converge in probability to ∞ under the alternative (1.4). The first statement in Theorem 1 asserts that $nU_n = O_p(1)$ and $nV_n = O_p(1)$ under the null hypothesis (1.3). That is, our tests are consistent against any fixed alternative.

In addition to testing the null (1.3) against the fixed alternative (1.4), we test for local departures from the null. The results can be used to evaluate the asymptotic power of the statistics U_n and V_n. Consider the local alternatives

$$H_{1n} : m(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\beta}_0 + n^{-1/2}\ell(\mathbf{x}), \tag{3.8}$$

where $\ell(\mathbf{x}) \notin \mathcal{M}_{\boldsymbol{\beta}}$ and $E\{\mid \ell(\mathbf{x}) \mid\} < \infty$. To obtain the next assertion, we impose the following regularity assumptions.

Condition E. There exists some $\gamma_2 > 0$ such that $E\{|\ell(\mathbf{x})|^{2+\gamma_2}\} < \infty$.

Write $\varrho_{0\ell} = E\{\mathbf{g}(\mathbf{x})\ell(\mathbf{x})\} \in \mathbb{R}^d$, $\varrho_{1\ell} \stackrel{\text{def}}{=} -E\{\mathbf{g}(\mathbf{x}_1) - E\mathbf{g}(\mathbf{x}_1)\}\{\ell(\mathbf{x}_2) - E\ell(\mathbf{x}_2)\}\|\mathbf{x}_1 - \mathbf{x}_2\| \in \mathbb{R}^d$, and $\varrho_{2\ell} \stackrel{\text{def}}{=} -E\{\ell(\mathbf{x}_1) - E\ell(\mathbf{x}_1)\}\{\ell(\mathbf{x}_2) - E\ell(\mathbf{x}_2)\}\|\mathbf{x}_1 - \mathbf{x}_2\| \in \mathbb{R}$.

Theorem 2. *Suppose that Conditions A, B, C, and E hold. Under the local alternative* (3.8)*, as $n \to \infty$, nU_n converges in distribution to*

$$\begin{aligned}&\sum_{i=1}^{\infty} \lambda_i(\mathcal{Z}_i^2 - 1) + \{\mathcal{N} - 4\varrho_{1\ell} f_{\eta_{(1)}^{(2)}}(0)\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathcal{W} + \varrho_{0\ell}) \\ &+2f_{\eta_{(1)}^{(2)}}(0)(\mathcal{W} + \varrho_{0\ell})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}(\mathcal{W} + \varrho_{0\ell}) + 2f_{\eta_{(1)}^{(2)}}(0)\varrho_{2\ell},\end{aligned} \tag{3.9}$$

and nV_n converges in distribution to

$$\begin{aligned}&\sum_{i=1}^{\infty} \lambda_i\mathcal{Z}_i^2 + \{\mathcal{N} - 4\varrho_{1\ell} f_{\eta_{(1)}^{(2)}}(0)\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathcal{W} + \varrho_{0\ell}) \\ &+2f_{\eta_{(1)}^{(2)}}(0)(\mathcal{W} + \varrho_{0\ell})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}(\mathcal{W} + \varrho_{0\ell}) + 2f_{\eta_{(1)}^{(2)}}(0)\varrho_{2\ell}.\end{aligned} \tag{3.10}$$

By direct calculation, we have

$$\begin{aligned}&2f_{\eta_{(1)}^{(2)}}(0)\varrho_{2\ell} + 2f_{\eta_{(1)}^{(2)}}(0)\varrho_{0\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\varrho_{0\ell} - 4f_{\eta_{(1)}^{(2)}}(0)\varrho_{1\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\varrho_{0\ell} \\ &= -2f_{\eta_{(1)}^{(2)}}(0)E[\{\varrho_{0\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{g}(\mathbf{x}_1) - \ell(\mathbf{x}_1)\}\{\varrho_{0\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{g}(\mathbf{x}_2) - \ell(\mathbf{x}_2)\}C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)] \\ &\stackrel{\text{def}}{=} 2f_{\eta_{(1)}^{(2)}}(0)\Delta_\ell.\end{aligned}$$

Obviously, Δ_ℓ is the squared martingale difference divergence (Shao and Zhang (2014)) between $\{\varrho_{0\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{g}(\mathbf{x}) - \ell(\mathbf{x})\}$ and $\mathbf{x}$. According to Theorem 1 of Shao and Zhang (2014), it follows that $\Delta_\ell \geq 0$. Because $\ell(\cdot)$ is a nonlinear function, that is, $\mathrm{pr}\{\ell(\mathbf{x}_1) = \varrho_{0\ell}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{g}(\mathbf{x}_1)\} < 1$, we have $\Delta_\ell > 0$. In addition, observing $\Delta_{a\ell} = a^2\Delta_\ell$ for an arbitrary constant a, we can show that our proposed tests have power tending to one under local alternatives of order less than $n^{-1/2}$.

3.3. The bootstrap and its consistency

Part I of Theorem 1 describes the asymptotic distributions of U_n and V_n; their consistency can be deduced from Part II of Theorem 1. However, the asymptotic distributions are not practical for determining the critical values of the test statistics because they depend on the infinitely many nuisance parameters λ_i in a complicated way. Therefore, we propose computing the critical value using a bootstrap method. The proposed bootstrap procedure is closely related to the work of Leucht and Neumann (2009), who provide a bootstrap approximation for degenerate U-type and V-type statistics of degree two in an i.i.d. setup. Nevertheless, the i.i.d. situation is violated in regression models. Thus we follow the idea devised in Sen and Sen (2014) for non-i.i.d. cases:

Step 1. Calculate the residuals $\eta_{in} = Y_i - \mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\beta}_n$ $(i = 1, \ldots, n)$ and generate an i.i.d. bootstrap sample $\{\widehat{\eta}^*_{in}, \mathbf{x}_{in}\}_{i=1}^n$ of size n from the measure $\mathrm{pr}_n = \mathrm{pr}_{n,\eta_n} \times \mathrm{pr}_{n,\mathbf{x}}$, where pr_{n,η_n} and $\mathrm{pr}_{n,\mathbf{x}}$ are the empirical distributions of η_{in} and $\mathbf{x}_i$, respectively.

Step 2. Denote $Y_{in} = \mathbf{g}(\mathbf{x}_{in})^{\mathrm{T}}\boldsymbol{\beta}_n + \eta^*_{in}$ and compute the bootstrapped least-squares estimator $\boldsymbol{\beta}^*_n$ using the bootstrap sample $(Y_{in}, \mathbf{x}_{in})$. Further, compute the bootstrap residuals $\eta^{**}_{in} = Y_{in} - \mathbf{g}(\mathbf{x}_{in})^{\mathrm{T}}\boldsymbol{\beta}^*_n$.

Step 3. Compute the bootstrap test statistics U^*_n and V^*_n, with $(\eta_{in}, \mathbf{x}_i)$ replaced by $(\eta^{**}_{in}, \mathbf{x}_{in})$, for $i = 1, \ldots, n$. Given the data, we approximate the distributions of nU_n and nV_n using the conditional distributions of nU^*_n and nV^*_n, respectively.

We now impose slightly stronger conditions than those applied to Theorem 1. Higher-order moments are used to verify the uniform integrability of some class of random variables, which is needed in the proof. To demonstrate the consistency of the bootstrap, we first assume that $m(\cdot)$ does not necessarily belong to $\mathcal{M}_{\boldsymbol{\beta}}$.

Condition F. There exists some $\gamma_3 > 0$ such that $E(|\varepsilon|^{2+\gamma_3}) < \infty$ and $E\{\|\mathbf{g}(\mathbf{x})\|^{4+\gamma_3}\} < \infty$.

Theorem 3. *Suppose the conditions for Theorem* 1 *hold and that Condition F is satisfied. Conditional on the observed data almost surely, regardless of the model misspecification, nU^*_n converges in distribution to*

$$\sum_{i=1}^{\infty} \widetilde{\lambda}_i(\widetilde{\mathcal{Z}}_i^2 - 1) + \widetilde{\mathcal{N}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\widetilde{\mathcal{W}} + 2f_{\varepsilon^{(2)}_{(1)}}(0)\widetilde{\mathcal{W}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\widetilde{\mathcal{W}}, \tag{3.11}$$

where $\widetilde{\mathcal{Z}}_i$ *are i.i.d.* $N(0,1)$. *Furthermore,* $(\widetilde{\mathcal{Z}}_i, \widetilde{\mathcal{N}}, \widetilde{\mathcal{W}}) \in \mathbb{R}^{2d+1}$ *are jointly Gaussian random variables with mean zero and covariance defined as in* $(\mathcal{Z}_i, \mathcal{N}, \mathcal{W})$ *of Theorem* 1, *where* $\mathbf{z}_i = (\eta_i, \mathbf{x}_i)$ *is replaced with* $\widetilde{\mathbf{z}}_i = (\widetilde{\varepsilon}_i, \mathbf{x}_i)$. *Here,* $\widetilde{\varepsilon}_i$ *is independent of* $\mathbf{x}_i$ *and has the same distribution as* $\varepsilon_i = m(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_i)^T\boldsymbol{\beta}_0 + \eta_i$. *In addition,* nV_n^* *converges in distribution to*

$$\sum_{i=1}^{\infty} \widetilde{\lambda}_i \widetilde{\mathcal{Z}}_i^2 + \widetilde{\mathcal{N}}^T \boldsymbol{\Sigma}^{-1} \widetilde{\mathcal{W}} + 2 f_{\varepsilon_{(1)}^{(2)}}(0) \widetilde{\mathcal{W}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \widetilde{\mathcal{W}}. \tag{3.12}$$

From Theorem 3, it follows that under the null hypothesis (1.3), $\varepsilon_i = \eta_i$. That is, if (1.3) holds, nU_n^* and nV_n^* converge in distribution to $\sum_{i=1}^{\infty} \lambda_i(\mathcal{Z}_i^2 - 1) + \mathcal{N}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathcal{W} + 2 f_{\eta_{(1)}^{(2)}}(0) \mathcal{W}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\ \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \mathcal{W}$ and $\sum_{i=1}^{\infty} \lambda_i \mathcal{Z}_i^2 + \mathcal{N}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathcal{W} + 2 f_{\eta_{(1)}^{(2)}}(0) \mathcal{W}^{\mathrm{T}}\ \boldsymbol{\Sigma}^{-1}\ \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \mathcal{W}$, respectively, conditional on the observed data almost surely. Recall that we reject the null hypothesis for large values of U_n and V_n. Therefore, from part I of Theorem 1, the p-values of the tests based on nU_n and nV_n, denoted by

$$\begin{aligned} \text{pvalue}^U &= \text{pr}\{nU_n^* > nU_n \mid \mathbf{x}_i, Y_i, i = 1, \ldots, n\}, \text{and} \\ \text{pvalue}^V &= \text{pr}\{nV_n^* > nV_n \mid \mathbf{x}_i, Y_i, i = 1, \ldots, n\}, \end{aligned}$$

respectively, asymptotically follow the uniform distribution on $(0, 1)$ in distribution under the null hypothesis (1.3). Consequently, the proposed bootstrap tests based on nU_n^* and nV_n^* have a correct asymptotic level.

On the other hand, Theorem 3 also indicates that under the fixed alternative (1.4) or the local alternative (3.8), we still have $nU_n^* = O_p^*(1)$ and $nV_n^* = O_p^*(1)$, where the probability is taken with respect to the bootstrapped space. From part II of Theorem 1 and Theorem 2, it follows that under the fixed alternative (1.4) or the local alternative (3.8), which decays at an order slower than $n^{-1/2}$, $nU_n \to \infty$ and $nV_n \to \infty$ in probability. That is, the corresponding p-values, that is, pvalue^U and pvalue^V, converge in probability to zero. Therefore, the bootstrap scheme is consistent and can detect alternatives that tend to the null at the parametric rate $n^{-1/2}$.

4. Simulations

In this section, we examine the finite-sample performance of the proposed testing statistics V_n and U_n, as defined in (3.3) and (3.4). Denote the V- and U-types of test statistic by VT and UT, respectively. For comparison purposes, we

consider four typical methods: the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) tests (Stute (1997)), the adaptive Neyman (AN) test (Fan and Huang (2001)), and the HSIC-based test (Sen and Sen (2014)). The simulation study is conducted using R. Specifically, we implement the KS and CvM tests using the "IntRegGOF" library. We consider the AN test described in (2.1) of (Fan and Huang (2001)) when noise is normal. The code required to implement the HSIC-based test of Sen and Sen (2014) is available on their home page. Following the suggestion of Sen and Sen (2014), we implement their method using standardized variables and taking Gaussian kernels with unit bandwidths. We also examine the sensitivity of the HSIC-based test with respect to the choice of γ, which is denoted by HSIC(γ). HSIC(1) corresponds to the test proposed by Sen and Sen (2014). Following the suggestion of a reviewer, we also consider an HSIC test in which the parameter is chosen as the median of the pairwise sample distances (mHSIC). Computations are based on 1,000 samples. In the ith sample, 500 bootstrap samples were generated to compute the empirical p-value, p_i. The empirical size and power are computed as $1000^{-1}\sum_{i=1}^{1000} I(p_i \le \alpha)$ at a significance level of α.

Three data-generating models are considered. The first two models test for multiple linear models, and the third model focuses on the univariate linear model.

Model 1. In our first example, the data are generated from the quadratic regression model

$$Y = X_1 + aX_2^2 + 2X_4 + \eta, \tag{4.1}$$

with the predictor vector $\mathbf{x} = (X_1, X_2, X_3, X_4)$, where the predictors X_1, X_2, and X_3 are normally distributed with mean zero and variance one. The pairwise correlation between these three random variables are 0.5. The predictor X_4 is binary, independent of X_1, X_2, and X_3, and satisfies $\mathrm{pr}(X_4 = 1) = 0.4$ and $\mathrm{pr}(X_4 = 0) = 0.6$. In addition, η follows a standardized normal distribution. This model is adapted from Example 4 of Fan and Huang (2001). In this example, the sample size is 100, the dimension of the predictor is four, and $a = 0$ corresponds to the null hypothesis.

Table 1 clearly shows that under the null, the nominal levels $\alpha = 0.01, 0.05$, 0.10 are reasonably approximated for the VT, UT, KS, CvM, mHSIC, and HSIC(1) tests. Furthermore, the empirical sizes of the AN test tend to be larger than the nominal levels. We further observe that the HSIC($1/\sqrt{2}$) and HSIC($1/\sqrt{3}$) tests tend to give zero rejection rates and have almost no power

Table 1. Empirical size and power of the VT, UT, KS, CvM, AN, and HSIC tests under model (4.1) with different a and α when $n = 100$.

level	test	$a = 0$	$a = 0.15$	$a = 0.25$	$a = 0.35$	$a = 0.45$	$a = 0.55$
$\alpha = 0.01$	VT	0.008	0.012	0.169	0.414	0.656	0.822
	UT	0.007	0.010	0.143	0.327	0.557	0.750
	HSIC(1)	0.013	0.033	0.105	0.275	0.483	0.669
	HSIC($1/\sqrt{2}$)	0.002	0.004	0.010	0.018	0.041	0.072
	HSIC($1/\sqrt{3}$)	0.000	0.000	0.000	0.000	0.000	0.001
	mHSIC	0.008	0.032	0.061	0.124	0.217	0.318
	KS	0.006	0.017	0.048	0.101	0.193	0.327
	CvM	0.008	0.033	0.081	0.185	0.319	0.469
	AN	0.033	0.031	0.051	0.134	0.295	0.551
$\alpha = 0.05$	VT	0.053	0.157	0.424	0.699	0.879	0.958
	UT	0.042	0.119	0.321	0.581	0.793	0.911
	HSIC(1)	0.060	0.130	0.330	0.560	0.766	0.887
	HSIC($1/\sqrt{2}$)	0.029	0.044	0.073	0.127	0.198	0.287
	HSIC($1/\sqrt{3}$)	0.002	0.005	0.010	0.018	0.027	0.038
	mHSIC	0.054	0.095	0.178	0.308	0.473	0.610
	KS	0.046	0.092	0.163	0.276	0.455	0.637
	CvM	0.049	0.119	0.250	0.410	0.589	0.770
	AN	0.079	0.090	0.164	0.298	0.514	0.757
$\alpha = 0.10$	VT	0.096	0.284	0.570	0.816	0.945	0.988
	UT	0.079	0.212	0.470	0.714	0.884	0.956
	HSIC(1)	0.113	0.257	0.467	0.712	0.868	0.953
	HSIC($1/\sqrt{2}$)	0.085	0.117	0.166	0.249	0.360	0.482
	HSIC($1/\sqrt{3}$)	0.036	0.038	0.052	0.073	0.100	0.124
	mHSIC	0.105	0.180	0.289	0.457	0.640	0.765
	KS	0.107	0.162	0.260	0.435	0.620	0.773
	CvM	0.099	0.212	0.369	0.560	0.749	0.876
	AN	0.123	0.159	0.244	0.397	0.634	0.837

in all cases. These findings suggest that the performance of the HSIC method is adversely affected by some tuning parameters. Of course, the HSIC(1) test works well and outperforms the mHSIC test in this example. As anticipated, as a becomes large, the power of each test increases monotonically to one in this quadratic regression model. In particular, the proposed tests VT and UT do significantly better than the KS, CvM, mHSIC, and HSIC(1) tests. It is expected that the KS and CvM tests underperform when the dimension of the predictor is larger than one. This is because the indicators $I(\mathbf{x}_j \leq \mathbf{x}_i)$, involved in the computation of the KS and CvM tests, are zero when the dimension of the predictor vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is large.

Table 2. Empirical size and power of the VT, UT, KS, CvM, AN, and HSIC tests under model (4.2) with different a and α when $n = 100$ and $d = 4$.

level	test	$a = 0$	$a = 1.5$	$a = 3.5$	$a = 5.5$	$a = 7.5$	$a = 9.5$
$\alpha = 0.01$	VT	0.007	0.037	0.373	0.805	0.972	0.995
	UT	0.006	0.034	0.357	0.791	0.965	0.996
	HSIC(1)	0.006	0.008	0.078	0.344	0.733	0.930
	HSIC($1/\sqrt{2}$)	0.000	0.000	0.001	0.002	0.009	0.061
	HSIC($1/\sqrt{3}$)	0.000	0.000	0.000	0.000	0.000	0.000
	mHSIC	0.014	0.037	0.206	0.588	0.841	0.950
	KS	0.012	0.016	0.027	0.056	0.114	0.181
	CvM	0.009	0.012	0.029	0.077	0.153	0.243
	AN	0.026	0.031	0.050	0.103	0.310	0.668
$\alpha = 0.05$	VT	0.046	0.140	0.601	0.942	0.996	1.000
	UT	0.043	0.127	0.576	0.922	0.992	1.000
	HSIC(1)	0.045	0.068	0.268	0.674	0.916	0.985
	HSIC($1/\sqrt{2}$)	0.005	0.005	0.009	0.059	0.184	0.383
	HSIC($1/\sqrt{3}$)	0.000	0.000	0.000	0.000	0.001	0.001
	mHSIC	0.059	0.107	0.438	0.800	0.962	0.989
	KS	0.061	0.063	0.108	0.190	0.299	0.442
	CvM	0.051	0.056	0.119	0.237	0.408	0.554
	AN	0.091	0.097	0.125	0.270	0.565	0.873
$\alpha = 0.10$	VT	0.091	0.237	0.728	0.970	0.998	1.000
	UT	0.087	0.209	0.691	0.958	0.997	1.000
	HSIC(1)	0.092	0.159	0.429	0.805	0.964	0.995
	HSIC($1/\sqrt{2}$)	0.024	0.031	0.075	0.202	0.424	0.677
	HSIC($1/\sqrt{3}$)	0.000	0.001	0.002	0.004	0.013	0.030
	mHSIC	0.108	0.176	0.577	0.884	0.978	0.994
	KS	0.109	0.123	0.194	0.302	0.466	0.592
	CvM	0.107	0.123	0.218	0.387	0.570	0.703
	AN	0.139	0.140	0.198	0.393	0.683	0.923

Model 2. In our second example, we consider the interactive regression model

$$Y = 5X_1 - X_2^2 + aX_1X_2 + \eta, \tag{4.2}$$

with the predictor vector $\mathbf{x} = (X_1, \ldots, X_d)$ $(d \geq 2)$, where $\mathbf{x}$ has independent components, and each component is taken from a uniform distribution on the unit interval. Here, η is drawn from normal distribution with mean zero and variance one. We take $n = 100$ and $d = 4, 2$. This model is adapted from Model 1 of Sen and Sen (2014), where we omit the intercept term. The null hypothesis holds when $a = 0$.

The simulation results are summarized in Tables 2 and 3. Our proposed

Table 3. Empirical size and power of the VT, UT, KS, CvM, AN, and HSIC tests under model (4.2) with different a and α when $n = 100$ and $d = 2$.

level	test	$a = 0$	$a = 1.5$	$a = 3.5$	$a = 5.5$	$a = 7.5$	$a = 9.5$
$\alpha = 0.01$	VT	0.010	0.078	0.744	0.992	1.000	1.000
	UT	0.009	0.077	0.726	0.991	1.000	1.000
	HSIC(1)	0.006	0.045	0.342	0.813	0.982	1.000
	HSIC($1/\sqrt{2}$)	0.009	0.013	0.117	0.468	0.855	0.978
	HSIC($1/\sqrt{3}$)	0.005	0.006	0.035	0.198	0.551	0.854
	mHSIC	0.007	0.050	0.523	0.953	0.997	1.000
	KS	0.007	0.015	0.098	0.314	0.565	0.779
	CvM	0.008	0.028	0.236	0.696	0.957	0.993
	AN	0.027	0.029	0.046	0.118	0.318	0.694
$\alpha = 0.05$	VT	0.047	0.247	0.903	0.999	1.000	1.000
	UT	0.045	0.231	0.889	0.998	1.000	1.000
	HSIC(1)	0.048	0.131	0.566	0.946	0.997	1.000
	HSIC($1/\sqrt{2}$)	0.037	0.070	0.291	0.723	0.961	0.995
	HSIC($1/\sqrt{3}$)	0.032	0.044	0.152	0.447	0.802	0.957
	mHSIC	0.048	0.180	0.753	0.987	0.999	1.000
	KS	0.045	0.090	0.311	0.632	0.878	0.969
	CvM	0.049	0.132	0.506	0.901	0.992	1.000
	AN	0.079	0.082	0.140	0.279	0.580	0.901
$\alpha = 0.10$	VT	0.089	0.382	0.937	1.000	1.000	1.000
	UT	0.087	0.362	0.928	1.000	1.000	1.000
	HSIC(1)	0.112	0.233	0.690	0.975	0.999	1.000
	HSIC($1/\sqrt{2}$)	0.087	0.150	0.441	0.834	0.972	0.998
	HSIC($1/\sqrt{3}$)	0.077	0.096	0.272	0.615	0.896	0.978
	mHSIC	0.105	0.294	0.855	0.992	0.999	1.000
	KS	0.109	0.173	0.477	0.785	0.947	0.991
	CvM	0.107	0.212	0.647	0.961	0.997	1.000
	AN	0.130	0.152	0.214	0.391	0.700	0.952

tests clearly outperform the competing choices. The difference between VT and UT is relatively small. For the VT, UT, KS, CvM, mHSIC, and HSIC(1) tests, the empirical type-I error rates are very close to the nominal test levels. However, for the HSIC($1/\sqrt{2}$) and HSIC($1/\sqrt{3}$) tests (AN test), the empirical sizes tend to be less (slightly larger) than the nominal levels. In this example, the mHSIC test however performs slightly better than the HSIC(1) test does. These observations indicate that the performance of the HSIC-based test depends significantly on some tuning parameters. It is anticipated that the KS and CvM tests behave poorly as the dimension of the predictor increases in all simulation settings considered here.

Table 4. Empirical size and power of the VT, UT, KS, CvM, AN, and HSIC tests under model (4.3) with different a and α when $n = 50$.

level	test	$a=0$	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$
$\alpha = 0.01$	VT	0.007	0.028	0.071	0.198	0.386	0.613
	UT	0.006	0.029	0.069	0.193	0.380	0.611
	HSIC(1)	0.013	0.017	0.044	0.106	0.234	0.423
	HSIC($1/\sqrt{2}$)	0.013	0.016	0.028	0.067	0.143	0.267
	HSIC($1/\sqrt{3}$)	0.008	0.010	0.016	0.040	0.087	0.163
	mHSIC	0.007	0.011	0.035	0.095	0.191	0.349
	KS	0.009	0.017	0.031	0.065	0.134	0.228
	CvM	0.012	0.017	0.038	0.088	0.207	0.325
	AN	0.025	0.028	0.032	0.034	0.041	0.057
$\alpha = 0.05$	VT	0.048	0.081	0.208	0.417	0.658	0.838
	UT	0.048	0.086	0.197	0.405	0.644	0.830
	HSIC(1)	0.063	0.081	0.136	0.275	0.488	0.668
	HSIC($1/\sqrt{2}$)	0.052	0.060	0.108	0.196	0.337	0.515
	HSIC($1/\sqrt{3}$)	0.047	0.059	0.081	0.135	0.237	0.380
	mHSIC	0.049	0.069	0.118	0.238	0.416	0.605
	KS	0.052	0.072	0.122	0.218	0.355	0.510
	CvM	0.065	0.080	0.149	0.283	0.429	0.622
	AN	0.076	0.079	0.083	0.097	0.112	0.157
$\alpha = 0.10$	VT	0.109	0.147	0.318	0.559	0.775	0.911
	UT	0.105	0.141	0.307	0.539	0.761	0.901
	HSIC(1)	0.102	0.135	0.230	0.423	0.604	0.771
	HSIC($1/\sqrt{2}$)	0.108	0.122	0.192	0.294	0.456	0.626
	HSIC($1/\sqrt{3}$)	0.117	0.121	0.158	0.244	0.371	0.509
	mHSIC	0.111	0.134	0.213	0.367	0.555	0.729
	KS	0.110	0.132	0.214	0.336	0.500	0.650
	CvM	0.121	0.148	0.246	0.386	0.563	0.731
	AN	0.121	0.116	0.133	0.156	0.189	0.246

Model 3. In our last example, we consider the univariate regression model

$$Y = 5X_1 + aX_1^2 + \eta, \tag{4.3}$$

where X_1 follows a uniform distribution on the unit interval, and η is drawn from a normal distribution with mean zero and variance one. This example is adapted from Model 1 of Stute, Manteiga and Quindimil (1998). The sample size is 50, and the null hypothesis is true if and only if $a = 0$.

Table 4 presents the empirical size and power associated with the VT, UT, KS, CvM, AN, and HSIC tests for different a and α. From Table 4, in this example with one-dimensional data, the proposed tests behave comparably well

to existing methods, such as the well-known KS and CvM tests. A comparison with the mHSIC and HSIC(1) tests shows that, for this particular example, the HSIC(1) test is slightly more effective than the mHSIC test. Slightly surprisingly, in contrast to Models 1 and 2, the HSIC($1/\sqrt{2}$) and HSIC($1/\sqrt{3}$) tests perform satisfactorily in this example. These results suggest that the HSIC-based goodness-of-fit test depends on the data-generating process. How to choose the optimal tuning parameters associated with the HSIC-based method seems to be a difficult practical issue, and deserves a deeper study.

5. Discussion

Based on the popular dCov (Székely, Rizzo and Bakirov (2007)), we propose two tests that determine the goodness-of-fit of linear models. Our tests can be viewed as an extension of the independent test of no-effect model of Székely, Rizzo and Bakirov (2007) to a test of the lack-of-fit of a regression model. Our tests successfully break the curse of dimensionality found in some nonparametric tests such as the KS and CvM tests (Stute (1997)), when the dimension of the regressors is larger than one. Compared with the HSIC-based test (Sen and Sen (2014)), our methods successfully avoid having to make subjective choices of parameters, such as bandwidths and kernels. Our simulations results demonstrate the good behavior of the tests in small samples, as compared with that of other well-known tests, such as the AN test (Fan and Huang (2001)).

Finally, we conclude by noting three potential topics for future research. First, an interesting extension of our methodology would be to include a general semi-parametric class. Second, it is possible to extend our methodology to include missing, censored, or dependent data. Third, when the number of useless predictors included in the working model increases, it is useful to mitigate the impact of the large dimensionality and, thus, obtain power enhancement tests using techniques, such as projection pursuit and sufficient dimension reduction. We are currently investigating these issues.

Supplementary Material

The online Supplementary Material contains proofs of Theorems 2–3 and the second assertion of Theorem 1, as well as additional numerical results on some aspects of limiting distributions and a real data set.

Acknowledgments

We appreciate the constructive suggestions from the referees and the editors. This research was supported by grant no. 11901006 from the National Natural Science Foundation of China, grant no. 1908085QA06 from the Natural Science Foundation of Anhui Province, and grant no. 751811 from the Talent Foundation of Anhui Normal University.

Appendix

A. Some Technical Proofs

Lemmas 1 and 2 will be used repeatedly in the proof of main results. Lemma 1 is employed to deduce the Euclidean (Pakes and Pollard (1989, Def. 2.7)) property of a class of functions and directly extracted from Lemma 2.13 of Pakes and Pollard (1989). Lemma 2 states a uniform convergence result for U-statistics indexed by parameters and is a direct consequence of Corollary 8 of Sherman (1994).

Lemma 1. *Let $\mathcal{F} = \{\mathfrak{F}(\cdot, \boldsymbol{t}) : \boldsymbol{t} \in \boldsymbol{T}\}$ be a class of functions on $\mathfrak{X}$ indexed by a bounded subset $\boldsymbol{T}$ of $\mathbb{R}^d$. If there exists a $C > 0$ and a nonnegative function $\phi(\cdot)$ such that $|\mathfrak{F}(\boldsymbol{x}, \boldsymbol{t}) - \mathfrak{F}(\boldsymbol{x}, \boldsymbol{t}')| \leq \phi(\boldsymbol{x})\|\boldsymbol{t} - \boldsymbol{t}'\|^C$ for $\boldsymbol{x} \in \mathfrak{X}$ and $t, t' \in \boldsymbol{T}$, then $\mathcal{F}$ is Euclidean for the envelope $|\mathfrak{F}(\cdot, \boldsymbol{t}_0)| + M\phi(\cdot)$, where $\boldsymbol{t}_0$ is an arbitrary point of $\boldsymbol{T}$, $M = \{2d^{1/2} \sup_{\boldsymbol{T}} \|\boldsymbol{t} - \boldsymbol{t}_0\|\}^C$ and we say that $\mathcal{F}$ has an envelope $L(\cdot)$ if $\sup_{\mathfrak{F} \in \mathcal{F}} |\mathfrak{F}(\cdot, \boldsymbol{t})| \leq L(\cdot)$.*

According to the notations of Sherman (1994), given a random sample $\{\mathbf{z}_i\}_{i=1}^n$ with distribution on a set $\mathcal{S}$, the k-order U-statistic indexed by $\boldsymbol{\theta}$ is defined as $U_n^k H(\cdot, \boldsymbol{\theta}) = \{n(n-1)\cdots(n-k+1)\}^{-1} \sum_{i_1 \neq i_2 \neq \cdots \neq i_k} H(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_k}, \boldsymbol{\theta})$. The kernel function $H(\mathbf{z}_1, \ldots, \mathbf{z}_k, \boldsymbol{\theta})$ can be asymmetrical in $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$.

Lemma 2. *Let $\mathcal{H}$ a class of real-valued functions on $\mathcal{S}^k = \mathcal{S} \otimes \cdots \otimes \mathcal{S}$. Supposing (i) $\mathcal{H}$ be a class of degenerate functions on $\mathcal{S}^k$, (ii) $H(\cdot, \boldsymbol{\theta}_0) \equiv 0$, (iii) $\mathcal{H}$ is Euclidean for an envelope L satisfying $E(L^2) < \infty$ and (iiii) $E|H(\cdot, \boldsymbol{\theta})| \to 0$, as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$, then $U_n^k H(\cdot, \boldsymbol{\theta}) = o_p(n^{-k/2})$ uniformly over $o_p(1)$ neighborhoods of $\boldsymbol{\theta}_0$.*

Proof of Theorem 1. Let us start by focusing on the statistic U_n. Let $I(\cdot)$ be the indicator function. By the identity (Knight (1998))

$$|x - y| - |x| = -y\{I(x > 0) - I(x < 0)\} + 2\int_0^y \{I(x \leq z) - I(x \leq 0)\}dz,$$

which is valid for $x \neq 0$, and the fact $\eta_{in} - \eta_i = -\mathbf{g}(\mathbf{x}_i)^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)$, it follows

$$|\eta_{in} - \eta_{jn}| = |\eta_i - \eta_j| - \{\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_j)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\{I(\eta_i > \eta_j) - I(\eta_i < \eta_j)\} + 2\int_0^{\{\mathbf{g}(\mathbf{x}_i)-\mathbf{g}(\mathbf{x}_j)\}^{\mathrm{T}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)} \{I(\eta_i - \eta_j \leq z) - I(\eta_i \leq \eta_j)\}dz. \tag{A.1}$$

Define $c(n,m) = n_m/m_m$ where $n_m = n(n-1)\cdots(n-m+1)$, and write $\mathbf{x}_{(i)}^{(j)} = \mathbf{x}_i - \mathbf{x}_j$ and $\eta_{(i)}^{(j)} = \eta_i - \eta_j$. From Lemma 1 in Yao, Zhang and Shao (2018), it is easy to show $U_n = \{c(n,4)\}^{-1}\sum_{i<j<k<l} h_0(\mathbf{z}_{in}, \mathbf{z}_{jn}, \mathbf{z}_{kn}, \mathbf{z}_{ln})$, where $\mathbf{z}_{in} = (\eta_{in}, \mathbf{x}_i)$ and

$$h_0(\mathbf{z}_{in}, \mathbf{z}_{jn}, \mathbf{z}_{kn}, \mathbf{z}_{ln}) = 6^{-1}\sum_{s<t,u<v}^{(i,j,k,l)} |\eta_{(sn)}^{(tn)}|\Big(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\|\Big) - 12^{-1}\sum_{(s,t,u)}^{(i,j,k,l)} |\eta_{(sn)}^{(tn)}|\|\mathbf{x}_{(s)}^{(u)}\| \in \mathbb{R}. \tag{A.2}$$

By definition, combining (A.1) and (A.2), we decompose U_n into three parts

$$U_n = U_{0n} + (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}U_{1n} + U_{2n}, \tag{A.3}$$

where $U_{0n} = \{c(n,4)\}^{-1}\sum_{i<j<k<l} h_0(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ with

$$h_0(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = 6^{-1}\sum_{s<t,u<v}^{(i,j,k,l)} |\eta_{(s)}^{(t)}|\Big(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\|\Big) - 12^{-1}\sum_{(s,t,u)}^{(i,j,k,l)} |\eta_{(s)}^{(t)}|\|\mathbf{x}_{(s)}^{(u)}\| \in \mathbb{R}, \tag{A.4}$$

and $\mathbf{z}_i = (\eta_i, \mathbf{x}_i)$, $U_{1n} = \{c(n,4)\}^{-1}\sum_{i<j<k<l} h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ with

$$h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = 6^{-1}\sum_{s<t,u<v}^{(i,j,k,l)} \delta_{1st}\Big(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\|\Big) - 12^{-1}\sum_{(s,t,u)}^{(i,j,k,l)} \delta_{1st}\|\mathbf{x}_{(s)}^{(u)}\| \in \mathbb{R}^d, \tag{A.5}$$

and $\delta_{1st} = -\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}\{I(\eta_s > \eta_t) - I(\eta_s < \eta_t)\}$, $U_{2n} = \{c(n,4)\}^{-1}\sum_{i<j<k<l}$

$h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ with

$$h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = 6^{-1} \sum_{s<t,u<v}^{(i,j,k,l)} \delta_{2st} \Big(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\| \Big) - 12^{-1} \sum_{(s,t,u)}^{(i,j,k,l)} \delta_{2st} \|\mathbf{x}_{(s)}^{(u)}\| \in \mathbb{R},$$

and

$$\begin{aligned} \delta_{2st} = & \int_0^{\{\mathbf{g}(\mathbf{x}_s)-\mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)} \{I(\eta_s - \eta_t \le z) - I(\eta_s \le \eta_t)\} dz \\ & + \int_0^{\{\mathbf{g}(\mathbf{x}_t)-\mathbf{g}(\mathbf{x}_s)\}^{\mathrm{T}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)} \{I(\eta_t - \eta_s \le z) - I(\eta_t \le \eta_s)\} dz. \end{aligned}$$

Therefore, to obtain the asymptotic distribution of U_n, it suffices to investigate the asymptotic expansion of U_{0n}, U_{1n}, U_{2n} and $\boldsymbol{\beta}_n - \boldsymbol{\beta}_0$.

In what follows, we first focus on the term U_{2n} with the help of Lemmas 1 and 2 and the derivations on U_{0n}, U_{1n}, and $\boldsymbol{\beta}_n - \boldsymbol{\beta}_0$ are reported later. Clearly, U_{2n} does not belong to a class of degenerate functions on $\mathcal{S}^4$. Therefore, we further decompose U_{2n} into two parts

$$U_{2n} = U_{21n} + U_{22n}, \tag{A.6}$$

where $U_{21n} = \{c(n,4)\}^{-1} \sum_{i<j<k<l} h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ with $h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = 24^{-1} \sum_{(s,t,u,v)}^{(i,j,k,l)} \{\delta_{2st} - E(\delta_{2st}|\mathbf{x}_s, \mathbf{x}_t)\}(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\| - 2\|\mathbf{x}_{(s)}^{(u)}\|)$, and $U_{22n} = \{c(n,4)\}^{-1} \sum_{i<j<k<l} h_{22}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ with $h_{22}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = 24^{-1} \sum_{(s,t,u,v)}^{(i,j,k,l)} \{E(\delta_{2st} \mid \mathbf{x}_s, \mathbf{x}_t)\}$ $(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\| - 2\|\mathbf{x}_{(s)}^{(u)}\|)$. We shall show that the kernel $h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l)$ of U_{21n} satisfies conditions (i)–(iiii) in Lemma 2. According to the smoothing property of conditional expectation and the independence between η and $\mathbf{x}$, we have $E[\{\delta_{2st} - E(\delta_{2st}|\mathbf{x}_s, \mathbf{x}_t)\}|\mathbf{z}_s] = 0$ and therefore $E[\{\delta_{2st} - E(\delta_{2st}|\mathbf{x}_s, \mathbf{x}_t)\}(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\| - 2\|\mathbf{x}_{(s)}^{(u)}\|)|\mathbf{z}_i] = 0$ for $i = s, t, u, v$. That is, condition (i) in Lemma 2 holds. The fact that $\delta_{2st} = 0$ when $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0$ implies condition (ii) in Lemma 2 holds. For ease of our derivations, write $h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l; \boldsymbol{\beta}_n)$. We shall use C to denote a positive constant that do not depend on n and whose value may change from place to place. By definition, Minkowski and Cauchy-Schwarz inequalities, $|h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l; \boldsymbol{\beta}_1) - h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l; \boldsymbol{\beta}_2)| \le C| \int_0^{\{\mathbf{g}(\mathbf{x}_i)-\mathbf{g}(\mathbf{x}_j)\}^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)} \{I(\eta_i - \eta_j \le z) - I(\eta_i \le \eta_j) - EI(\eta_i - \eta_j \le z) + I(\eta_i \le \eta_j)\} dz - \int_0^{\{\mathbf{g}(\mathbf{x}_i)-\mathbf{g}(\mathbf{x}_j)\}^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)} \{I(\eta_i - \eta_j \le z) - I(\eta_i \le \eta_j) - EI(\eta_i - \eta_j \le z) + EI(\eta_i \le \eta_j)\} dz| (\|\mathbf{x}_i\| + \|\mathbf{x}_j\| + \|\mathbf{x}_k\| + \|\mathbf{x}_l\|) \le 4C|\{\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_j)\}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)|(\|\mathbf{x}_i\| + \|\mathbf{x}_j\| + \|\mathbf{x}_k\| + \|\mathbf{x}_l\|) \le$

$C\|\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_j)\|(\|\mathbf{x}_i\| + \|\mathbf{x}_j\| + \|\mathbf{x}_k\| + \|\mathbf{x}_l\|)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$. This, together with Lemma 1 and Condition C2, implies condition (iii) in Lemma 2. Similarly, taking $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_0$ leads to $E|h_{21}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l; \boldsymbol{\beta}_1)| \le C\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\| E\{\|\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_j)\|(\|\mathbf{x}_i\| + \|\mathbf{x}_j\| + \|\mathbf{x}_k\| + \|\mathbf{x}_l\|)\}$, which, together with Condition C2, indicates condition (iiii) in Lemma 2. Consequently,

$$nU_{21n} = o_p(1), \tag{A.7}$$

uniformly over $o_p(1)$ neighborhoods of $\boldsymbol{\beta}_0$. The condition $E\{\|\mathbf{g}(\mathbf{x})\|^{2+\gamma}\} < \infty$ implies $\max_{1\le i\le n} \|\mathbf{g}(\mathbf{x}_i)\| = o_p(n^{1/2})$, which yields $\max_{1\le s,t\le n} |\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)| \le 2\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| \max_{1\le i\le n} \|\mathbf{g}(\mathbf{x}_i)\| = o_p(1) n^{1/2}\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\|$, which implies

$$\max_{1\le s,t\le n} |\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)| \le 2\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| \max_{1\le i\le n} \|\mathbf{g}(\mathbf{x}_i)\| = o_p(1), \tag{A.8}$$

due to $n^{1/2}\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| = O_p(1)$ and Slutsky's theorem. Let $F_{\eta_{(1)}^{(2)}}$ and $f_{\eta_{(1)}^{(2)}}$ be the cdf and pdf of $\eta_{(1)}^{(2)} = \eta_1 - \eta_2$. By Taylor's expansion and Condition A, we have uniformly over $1 \le s, t \le n$,

$$\begin{aligned} E(\delta_{2st}|\mathbf{x}_s, \mathbf{x}_t) &= \int_0^{\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)} \Big\{F_{\eta_{(1)}^{(2)}}(z) - F_{\eta_{(1)}^{(2)}}(0)\Big\} dz \\ &\quad + \int_0^{\{\mathbf{g}(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_s)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)} \Big\{F_{\eta_{(1)}^{(2)}}(z) - F_{\eta_{(1)}^{(2)}}(0)\Big\} dz, \\ &= \{1 + o_p(1)\} f_{\eta_{(1)}^{(2)}}(0)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}. \\ &\quad \{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \end{aligned} \tag{A.9}$$

From (A.6), (A.7) and (A.9), it is easy to show

$$nU_{2n} = f_{\eta_{(1)}^{(2)}}(0)\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}^{\mathrm{T}} U_{2n}^{\natural}\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\} + o_p(1), \tag{A.10}$$

where $U_{2n}^{\natural} = \{c(n,4)\}^{-1} \sum_{i<j<k<l} h_2^{\natural}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) \in \mathbb{R}^{d\times d}$ with

$$\begin{aligned} h_2^{\natural}(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) &= 6^{-1} \sum_{s<t,u<v}^{(i,j,k,l)} \{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}\Big(\|\mathbf{x}_{(s)}^{(t)}\| + \|\mathbf{x}_{(u)}^{(v)}\|\Big) \\ &\quad -12^{-1} \sum_{(s,t,u)}^{(i,j,k,l)} \{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}\{\mathbf{g}(\mathbf{x}_s) - \mathbf{g}(\mathbf{x}_t)\}^{\mathrm{T}}\|\mathbf{x}_{(s)}^{(u)}\|. \end{aligned}$$

Together with (A.3), we get

$$nU_n = nU_{0n} + \{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}^{\mathrm{T}}\{n^{1/2}U_{1n}\} + f_{\eta^{(2)}_{(1)}}(0)\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}^{\mathrm{T}}U^{\natural}_{2n}\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\} + o_p(1). \quad \text{(A.11)}$$

It is observed that U_{0n} is degenerate and U_{1n} and $U^{\natural}_{2n}$ are non-degenerate. Define $h^{(2)}_0(\mathbf{z}_i, \mathbf{z}_j) = 6E\{h_0(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) \mid \mathbf{z}_1, \mathbf{z}_2\}$ and $h^{(1)}_1(\mathbf{z}_1) = 4E\{h_1(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) \mid \mathbf{z}_1\}$. According to Hoeffding decomposition in technical appendix of Yao, Zhang and Shao (2018),

$$h^{(2)}_0(\mathbf{z}_i, \mathbf{z}_j) = C_\eta(\eta_i, \eta_j)C_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_j). \quad \text{(A.12)}$$

Similarly, it is straightforward to verify that $h^{(1)}_1(\mathbf{z}_1) = -2E[\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}\{2F_\eta(\eta_1) - 1\}C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)|\mathbf{x}_1] - 2E[\{\mathbf{g}(\mathbf{x}_2) - \mathbf{g}(\mathbf{x}_1)\}\{1 - 2F_\eta(\eta_1))\}C_{\mathbf{x}}(\mathbf{x}_2, \mathbf{x}_1)|\mathbf{x}_1]$, which further yields

$$h^{(1)}_1(\mathbf{z}_1) = 4\{1 - 2F_\eta(\eta_1))\}E[\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}C_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}_2)|\mathbf{x}_1], \quad \text{(A.13)}$$

where F_η is the cdf of η. From the standard theory of U-statistics, invoking Conditions C1-C2 entails

$$nU_{0n} - n^{-1}\sum_{i\neq j} h^{(2)}_0(\mathbf{z}_i, \mathbf{z}_j) = o_p(1), \text{ and}$$
$$n^{1/2}U_{1n} - n^{-1/2}\sum_{i=1}^{n} h^{(1)}_1(\mathbf{z}_i) = o_p(1). \quad \text{(A.14)}$$

Condition C2 also implies $E[(\|\mathbf{x}_i\| + \|\mathbf{x}_j\|)\|\mathbf{g}(\mathbf{x}_k)\|\|\mathbf{g}(\mathbf{x}_l)|\}] < \infty, 1 \le i, j, k, l \le 4$. By careful calculations, $EU^{\natural}_{2n} = E[\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}\{\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\}^{\mathrm{T}}\{\|\mathbf{x}^{(2)}_{(1)}\| - E(\|\mathbf{x}^{(2)}_{(1)}\|\|\mathbf{x}_2) - E(\|\mathbf{x}^{(2)}_{(1)}\|\|\mathbf{x}_1) + E(\|\mathbf{x}^{(2)}_{(1)}\|)\}] = 2\boldsymbol{\Lambda}$, where

$$\boldsymbol{\Lambda} = -E\left[\{\mathbf{g}(\mathbf{x}_1) - E\mathbf{g}(\mathbf{x}_1)\}\{\mathbf{g}(\mathbf{x}_2) - E\mathbf{g}(\mathbf{x}_2)\}^{\mathrm{T}}\|\mathbf{x}_1 - \mathbf{x}_2\|\right]. \quad \text{(A.15)}$$

By the law of large numbers for U-statistics, we have $U^{\natural}_{2n} \to 2\boldsymbol{\Lambda}$ in probability. From Slutsky's theorem and Conditions C2-C3, we have $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{-1/2}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i)\eta_i + o_p(1)$. Combination of these and Slutsky's theorem entails

$$\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}^{\mathrm{T}}\{n^{1/2}U_{1n}\}$$

$$= \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} h_1^{(1)}(\mathbf{z}_i) \right\} + o_p(1), \text{ and}$$

$$\{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}^{\mathrm{T}} U_{2n}^{\natural} \{n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)\}$$

$$= 2 \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\} + o_p(1). \tag{A.16}$$

To obtain the desired result, we now need to decompose the term $n^{-1}\sum_{i<j} h_0^{(2)}(\mathbf{z}_i, \mathbf{z}_j)$. By (A.12) and Lemma 1.1 in technical appendix of Yao, Zhang and Shao (2018), it is easy to check $E\{h_0^{(2)}(\mathbf{z}_i, \mathbf{z}_j)^2\} < CE\{\|\mathbf{x}\|^2 + \eta^2\} < \infty$. By Mercer's theorem, we have $h_0^{(2)}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{z}_1)\phi_i(\mathbf{z}_2)$, where $\{\phi_i(\cdot)\}_{i=1}^{\infty}$ are orthonormal eigenfunctions, i.e., $E\{\phi_i(\mathbf{z})\phi_j(\mathbf{z})\} = 1$ if $i = j$ and zero otherwise, corresponding to the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ which are defined in connection with $h_0^{(2)}(\cdot,\cdot)$. Since $h_0^{(2)}(\cdot,\cdot)$ is degenerate, $E\{\varphi_i(\mathbf{z})\} = 0$ for any $i \geq 1$. Due to the orthogonality, $E\{h_0^{(2)}(\mathbf{z}_1, \mathbf{z}_1)\} = \sum_{i=1}^{\infty} \lambda_i$. By Slutsky's theorem and the law of large numbers, nU_{0n} can be expressed as $nU_{0n} = (n-1)^{-1}\sum_{i<j} h_0^{(2)}(\mathbf{z}_i, \mathbf{z}_j) + o_p(1) = (n-1)^{-1}\sum_{i,j=1}^{n} K_0(\mathbf{z}_i, \mathbf{z}_j) - (n-1)^{-1}\sum_{i=1}^{n} K_0(\mathbf{z}_i, \mathbf{z}_i) = \sum_{i=1}^{\infty} \lambda_i [\{n^{-1/2}\sum_{j=1}^{n} \phi_i(\mathbf{z}_j)\}^2 - 1] + o_p(1)$. Therefore, further combination of (A.11) and (A.16) yields $nU_n = nU_n^{\ddagger} + o_p(1)$, where

$$\begin{aligned} nU_n^{\ddagger} = & \sum_{i=1}^{\infty} \lambda_i \left[\left\{ n^{-1/2} \sum_{j=1}^{n} \phi_i(\mathbf{z}_j) \right\}^2 - 1 \right] \\ & + \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} h_1^{(1)}(\mathbf{z}_i) \right\} \\ & + 2 f_{\eta_{(1)}^{(2)}}(0) \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i \right\}. \end{aligned} \tag{A.17}$$

As the first term in (A.17) contains infinitely many λ_i and ϕ_i, we here follow Theorem 5.5.2 of Serfling (1980) to finish the proof of the first assertion on U_n. We aim to show that $E\exp(ixnU_n^{\ddagger}) \to E\exp(ixU^{\ddagger})$, $n \to \infty$, for each x, where $U^{\ddagger}$ is defined as in (3.6). Write $nU_{n,K}^{\ddagger} = nU_n^{\ddagger} - \sum_{i=K}^{\infty} \lambda_i [\{n^{-1/2}\sum_{j=1}^{n} \phi_i(\mathbf{z}_j)\}^2 - 1]$ and $U_K^{\ddagger} = U^{\ddagger} - \sum_{i=K}^{\infty} \lambda_i(\mathcal{Z}_i^2 - 1)$. Due to the spectral decomposition of $h_0^{(2)}(\mathbf{z}_1, \mathbf{z}_2)$, we have $E(nU_n^{\ddagger} - nU_{n,K}^{\ddagger})^2 \leq 2\sum_{i=K}^{\infty} \lambda_i^2 \to 0$, as $K \to \infty$. For any $\varepsilon > 0$, by the inequality $| E\exp(ixnU_n^{\ddagger}) - E\exp(ixnU_{n,K}^{\ddagger}) | \leq | x | \{E(nU_n^{\ddagger} - nU_{n,K}^{\ddagger})^2\}^{1/2}$ and

by choosing and fixing K large enough,

$$| E\exp(ixnU_n^{\ddagger}) - E\exp(ixnU_{n,K}^{\ddagger}) | \leq \varepsilon, \tag{A.18}$$

for all n sufficiently large. Similarly, using the inequality

$$| E\exp(ixU^{\ddagger}) - E\exp(ixU_K^{\ddagger}) | \leq | x | \{E(\mathcal{Z}_1^2 - 1)^2\}^{1/2} \left(\sum_{i=K}^{\infty} \lambda_i^2 \right)^{1/2},$$

we obtain, by choosing and fixing K large enough,

$$| E\exp(ixU^{\ddagger}) - E\exp(ixU_K^{\ddagger}) | \leq \varepsilon, \tag{A.19}$$

for all n sufficiently large. For any fixed K, and by multivariate central limit theorem,

$$| E\exp(ixnU_{n,K}^{\ddagger}) - E\exp(ixU_{n,K}^{\ddagger}) | \leq \varepsilon, \tag{A.20}$$

for all n sufficiently large. Combining (A.18), (A.19) and (A.20), we have for any x and any $\varepsilon > 0$

$$| E\exp(ixnU_n^{\ddagger}) - E\exp(ixU^{\ddagger}) | \leq 3\varepsilon, \tag{A.21}$$

for all n sufficiently large. That is, $nU_n^{\ddagger}$ converges in distribution to $U^{\ddagger}$.

On the other hand, similar to appendix of Székely, Rizzo and Bakirov (2007), we can also show

$$V_n = n^{-2} \sum_{i,j=1}^{n} \widehat{A}_{ij} \widehat{B}_{ij} = n^{-4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} h_0(\mathbf{z}_{in}, \mathbf{z}_{jn}, \mathbf{z}_{kn}, \mathbf{z}_{ln}), \tag{A.22}$$

where $h_0(\mathbf{z}_{in}, \mathbf{z}_{jn}, \mathbf{z}_{kn}, \mathbf{z}_{ln})$ is defined as in (A.2). Combining the standard V- and U-statistic theories, and the equation (6) in Sherman (1994) , we can obtain $nV_n = \sum_{i=1}^{\infty} \lambda_i \{n^{-1/2} \sum_{j=1}^{n} \phi_i(\mathbf{z}_j)\}^2 + \{n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i\}^{\tau} \boldsymbol{\Sigma}^{-1} \{n^{-1/2} \sum_{i=1}^{n} h_1^{(1)}\} + 2 f_{\eta_{(1)}^{(2)}}(0) \{n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i\}^{\tau} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \{n^{-1/2} \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i)\eta_i\} + o_p(1)$. Apply arguments exactly similar to those for dealing with U_n to finish the proof of the first assertion. The remaining technical proofs are available in Supplementary Material.

References

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics* **20**, 105-134.

Christensen, R. and Sun, S. K. (2010). Alternative goodness-of-fit tests for linear models. *Journal of the American Statistical Association* **105**, 291-301.

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

Fan, J. and Huang, L. (2001). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association* **96**, 640-652.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *Test* **22**, 361-411.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. and Smola, A. (2008). A Kernel Statistical Test of Independence. In *International Conference on Neural Information Processing Systems 20*, 585–592. MIT Press, Cambridge, Massachusetts.

Guerre, E. and Lavergne, P. (2005). Rate-optimal data-driven specification testing for regression models. *The Annals of Statistics* **33**, 840-870.

Khmaladze, E. V. and Koul, H. L. (2004). Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics* **32**, 995-1034.

Knight, K. (1998). Limiting distribution for L_1 regression estimators under general conditions. *The Annals of Statistics* **26**, 755-770.

Koul, H. L. and Ni, P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference* **119**, 109-144.

Lee, C. E. and Shao, X. F. (2018). Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *Journal of the American Statistical Association* **113**, 216-229.

Leucht, A. and Neumann, M. H. (2009). Consistency of general bootstrap methods for degenerate U-type and V-type statistics. *Journal of Multivariate Analysis* **100**, 1622-1633.

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

Matteson, D. and Tsay, R. (2017). Independent component analysis via distance covariance. *Journal of the American Statistical Association* **112**, 623-637.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027-1057.

Sen, A. and Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika* **101**, 927-942.

Serfling, R. L. (1980). *Approximation Theorems in Mathematical Statistics.* Wiley, New York.

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

Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics* **22**, 439-459.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**, 613-41.

Stute, W., Manteiga, W. G. and Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **93**, 141-149.

Stute, W., Xu, W. L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika* **95**, 451-467.

Székely, G. and Rizzo, M. (2009). Brownian distance covariance. *Annals of Applied Statistics* **3**, 1236-1265.

Székely, G. and Rizzo, M. (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters* **82**, 2278-2282.

Székely, G. and Rizzo, M. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143**, 1249-1272.

Székely, G. and Rizzo, M. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* **42**, 2382-2412.

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

Yao, S., Zhang, X. and Shao, X. (2018). Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 455-480.

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

Kai Xu

School of Mathematics and Statistics, Anhui Normal University, Wuhu 241002, China.

E-mail: tjxxukai@163.com

Daojiang He

School of Mathematics and Statistics, Anhui Normal University, Wuhu 241002, China.

E-mail: djheahnu@yahoo.com

(Received July 2018; accepted August 2019)