Statistica Sinica **32** (2022), 1661-1681
doi:https://doi.org/10.5705/ss.202020.0038

EFFICIENT DIAGNOSTICS FOR PARAMETRIC REGRESSION MODELS WITH DISTORTION MEASUREMENT ERRORS INCORPORATING DIMENSION-REDUCTION

Zhihua Sun, Feifei Chen and Hua Liang

University of Chinese Academy of Sciences, Beijing Normal University and George Washington University

Abstract: In this work, we study the diagnostics of parametric regression models when both the response variable and the covariates are distorted by errors. We employ a projected empirical process to develop Cramér–von Mises and Kolmogorov–Smirnov tests with dimension-reduction effects. We apply random approximation to enable an expedient calculation of the Kolmogorov–Smirnov test for checking the suitability of regression models. The proposed tests are shown to be consistent and can detect an alternative hypothesis close to the null hypothesis at the root-n rate. Simulation studies show that the proposed tests outperform existing methods. A real data set is analyzed for illustration.

Key words and phrases: Cramér–von Mises test, dimension-reduction, empirical process, Kolmogorov–Smirnov test, random approximation.

1. Introduction

Data distortion is a common problem in the biomedical, public health, and economics fields. Kaysen et al. (2002) presented a typical example in which the fibrinogen level and the serum transferrin level are observed with distortion owing to the existence of the body mass index (BMI) as a confounding variable. Şentürk and Müller (2005) showed that distortion fundamentally changes the relationship between the response and the predictor variables, and were the first to introduce a linear covariate-adjustment model. They established an estimation procedure by connecting this model with a varying-coefficient model. Since this pioneering work, a large body of literature has developed attempting to eliminate the adverse effects of distortion measurement errors. However, most studies on the subject have been restricted to the estimation of regression models; see Şentürk and Müller (2006, 2009), Nguyen and Şentürk (2008), Cui et al. (2009),

Corresponding author: Feifei Chen, Center for Statistics and Data Science, Beijing Normal University, Zhuhai, 519087, China. E-mail: chenfeifei12@mails.ucas.ac.cn.

Zhang, Zhu and Liang (2012), Delaigle, Hall and Zhou (2016), and Deng and Zhao (2019), among others.

The correct specification of a regression model suffering from data distortion is undoubtedly important to avoid misleading results in statistical analyses. In this work, we study the diagnostics of parametric models when both the response variable and the covariates are measured with distortion. The models are of the following form:

$$\begin{cases} Y = g(\mathbf{X}, \mathbf{Z}, \beta) + \varepsilon, \\ \widetilde{Y} = \psi(U)Y, \\ \widetilde{\mathbf{X}} = \gamma(U)\mathbf{X}, \end{cases}$$

where Y is the response variable, $\mathbf{X}$ and $\mathbf{Z}$ are p- and q-dimensional covariates, respectively, U is a scalar confounding variable independent of $(Y, \mathbf{X}^\top, \mathbf{Z}^\top)^\top$, β is the unknown parameter vector, and g is a known function. The variables Y and $\mathbf{X}$ are unavailable owing to the measurement error caused by the confounding variable U. Instead of Y and $\mathbf{X}$, the distorted variables $\widetilde{Y}$ and $\widetilde{\mathbf{X}}$ are observed. Here, the function ψ is unknown, and γ is a $p \times p$ diagonal matrix with non-parametric diagonal element functions $\gamma_1, \ldots, \gamma_p$. To ensure identifiability, let $\mathrm{E}\{\psi(U)\} = 1$ and $\mathrm{E}\{\gamma_r(U)\} = 1$ for $r = 1, \ldots, p$.

We write $\varepsilon(Y, \mathbf{X}, \mathbf{Z}) = Y - g(\mathbf{X}, \mathbf{Z}, \beta)$ and aim to test

$$\mathcal{H}_0 : \Pr\left\{\mathrm{E}\{\varepsilon(Y, \mathbf{X}, \mathbf{Z})|\mathbf{X}, \mathbf{Z}\} = 0\right\} = 1, \quad \text{for some } \beta, \tag{1.1}$$

against the alternative hypothesis that $\mathcal{H}_0$ does not hold. Zhang, Li and Feng (2015) proposed a residual-based empirical process test for problem (1.1). The test has the desired merit of dimension reduction, and is made highly suitable for a directional test by choosing the deviation function as the weighting function to maximize the power. However, the test is solely directional and depends on the prespecified weighting function. Recently, Zhao and Xie (2018) developed a local test that is consistent, but suffers from the dimension problem.

We propose omnibus tests rather than directional tests. Our goal is to propose tests that are free of the dimension problem, are easy to calculate, and perform well in terms of test power. We consider empirical process tests with the linear indicator weighting function $\mathbf{1}(\nu^\top\theta \leq t)$, where $\nu = (\mathbf{X}^\top, \mathbf{Z}^\top)^\top$, for any vector $\theta \in \mathcal{R}^{p+q}$ and any real number $t \in \mathcal{R}$.

The empirical-process-based test was first introduced by Stute (1997), and has been studied extensively in the field of regression model checking. In recent years, empirical process tests that consider a linear indicator weighting function and offer the advantage of dimensionality reduction have attracted considerable attention (Escanciano (2006); Conde-Amboage, Sánchez-Sellero and González-Manteiga (2015); Colling and Van Keilegom (2017)). Additional efforts have been made to eliminate the "curse of dimensionality" in the test methods. For example, Ma et al. (2014) proposed a variant of the integrated conditional moment test based on the linear projection approach, where the projection direction was chosen by fitting a single-index model. Furthermore, Guo, Wang and Zhu (2016) and Tan, Zhu and Zhu (2018) developed dimension-reduction model-adaptive approaches to avoid the problems with dimensionality.

The empirical process tests with a linear indicator weighting function involve a nuisance parameter θ, also called a projection direction parameter. To ensure the feasibility of the calculation and the consistency of the tests, the nuisance parameter is assumed to be a random vector following a uniform distribution on the unit sphere. The resultant tests in the literature are of the Cramér–von Mises (CvM) type, which can be transformed into a simple summation by applying a critical transformation formula provided by Escanciano (2006).

One may naturally wonder about the feasibility and effectiveness of different nuisance parameter choices. Furthermore, in addition to the CvM-type tests, is it possible to construct other tests, such as the Kolmogorov–Smirnov (KS) test? We investigate this possibility by applying random approximation to make the estimated empirical process with the linear indicator weighting function computationally convenient, and thus avoid applying the transformation formula in Escanciano (2006). Moreover, even if the nuisance parameter follows distributions other than the uniform distribution on the unit sphere, the tests are realizable.

The remainder of this paper is organized as follows. In Section 2, a CvM test is built based on an empirical process with a linear indicator weighting function. In Section 3, motivated by a random approximation algorithm, a KS test is established. The asymptotic properties of the proposed tests and the determination of the critical values are presented in Section 4. Simulation studies and a real-data analysis are conducted in Section 5. In the Appendices, we provide the conditions needed in the proofs. The proofs of the main results are presented in the online Supplementary Material.

2. CvM Test

2.1. Estimation of the null hypothesis model

Assume that an independent and identically distributed (i.i.d.) sample $\{(\widetilde{Y}_i, \widetilde{\mathbf{X}}_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ is obtained from $(\widetilde{Y}, \widetilde{\mathbf{X}}, \mathbf{Z})$. Because the true variables Y

and $\mathbf{X}$ are unavailable, by calibrating the measurement errors, we obtain their estimators: $\hat{Y}_i = \widetilde{Y}_i \widetilde{Y}_{m,n}/\hat{\psi}_n(U_i)$, $\hat{X}_{ri} = \widetilde{X}_{ri}\widetilde{X}_{m,nr}/\hat{\gamma}_{nr}(U_i)$, for $i = 1,\ldots,n$; $r = 1,\ldots,p$, with $\widetilde{Y}_{m,n}$, $\widetilde{X}_{m,nr}$, $\hat{\psi}_n(u)$ and $\hat{\gamma}_{nr}(u)$ defined in Appendix A. The calibrated method can also refer to that of Zhang, Li and Feng (2015). Here, we apply the local linear method to estimate $\psi(u)$ and $\gamma(u)$.

Based on the calibrated sample $\{(\hat{Y}_i, \hat{\mathbf{X}}_i, \mathbf{Z}_i), i = 1,\ldots,n\}$ with $\hat{\mathbf{X}}_i = (\hat{X}_{1i}, \ldots, \hat{X}_{pi})^\top$, an estimator of β, denoted by $\hat{\beta}_n$, is defined as the minimizer of the least squares objective function:

$$\sum_{i=1}^{n}\left\{\hat{Y}_i - g(\hat{\mathbf{X}}_i, \mathbf{Z}_i, \beta)\right\}^2. \tag{2.1}$$

The asymptotic normality of $\hat{\beta}_n$ is presented in Lemma 3 in the online Supplementary Material. It can be concluded that under the null hypothesis model in (1.1), $\hat{\beta}_n$ is $\sqrt{n}$-consistent.

2.2. CvM test statistic

A direct test of the conditional expectation in (1.1) involves a nonparametric estimation of $\mathrm{E}\{\varepsilon(Y,\mathbf{X},\mathbf{Z})|\mathbf{X},\mathbf{Z}\}$, which would cause the "curse of dimensionality." Therefore, we examine an equivalent form of the null hypothetical condition by transforming it into infinite equations of the unconditional expectations.

Proposition 1. *The following statements are equivalent: (i)* $\mathcal{H}_0$ *in* (1.1) *is true; (ii)* $\mathrm{E}\{\varepsilon(Y,\mathbf{X},\mathbf{Z})\mathbf{1}(\nu^\top\theta \le t)\} = 0$*, for any vector* $\theta \in \mathcal{R}^{p+q}$ *and any real number* $t \in \mathcal{R}$*; (iii)* $\mathrm{E}\{\varepsilon(Y,\mathbf{X},\mathbf{Z})\mathbf{1}(\nu^\top\theta \le t)\} = 0$*, for any vector* $\theta \in \mathcal{R}^{p+q}$ *with* $\|\theta\| = 1$ *and any real number* $t \in \mathcal{R}$*.*

The proof of the equivalence of (i) and (ii) is similar to that of Lemma 2.1 in Lavergne and Patilea (2008). The equivalence of (ii) and (iii) can be obtained by the fact that for any $\theta \ne 0$, the σ-field generated by $\nu^\top\theta$ is the same as the σ-field generated by $\nu^\top\theta/\|\theta\|$. This fact was also mentioned by Lavergne and Patilea (2008).

Denote the estimated model error $\hat{Y}_i - g(\hat{\mathbf{X}}_i, \mathbf{Z}_i, \hat{\beta}_n)$ by $\hat{\varepsilon}_n(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, for $i = 1,\ldots,n$. Based on $\mathrm{E}\{\varepsilon(Y,\mathbf{X},\mathbf{Z})\mathbf{1}(\nu^\top\theta \le t)\}$, we construct an estimated empirical process: $\mathcal{M}_{n,pro}(t) = n^{-1/2}\sum_{i=1}^{n}\hat{\varepsilon}_n(Y_i,\mathbf{X}_i,\mathbf{Z}_i)\mathbf{1}(\mathbf{V}_i^\top\theta \le t)$, where $\mathbf{V}_i = (\hat{\mathbf{X}}_i^\top, \mathbf{Z}_i^\top)^\top$, for $i = 1,\ldots,n$. Then, the CvM test is defined as

$$\mathcal{T}_{n,CvM} = \int\int \{\mathcal{M}_{n,pro}(t)\}^2 f(\theta) F_{n\theta}(dt)d\theta, \tag{2.2}$$

where $f(\theta)$ is the density function of θ, and $F_{n\theta}(t) = n^{-1}\sum_{i=1}^{n}\mathbf{1}(\mathbf{V}_i^\top\theta \le t)$.

Under the null hypothesis in (1.1), the test statistic $\mathcal{T}_{n,CvM}$ tends to zero and becomes larger under alternative hypotheses. Therefore, the null hypothesis is rejected for a sufficiently large value of $\mathcal{T}_{n,CvM}$.

Note that the test statistic $\mathcal{T}_{n,CvM}$ is equal to the following summation:

$$\mathcal{T}_{n,CvM} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\hat{\varepsilon}_n(Y_i,\mathbf{X}_i,\mathbf{Z}_i)\hat{\varepsilon}_n(Y_j,\mathbf{X}_j,\mathbf{Z}_j)A_{ijl}, \tag{2.3}$$

where $A_{ijl} = \int \mathbf{1}(\mathbf{V}_i^\top\theta \le \mathbf{V}_l^\top\theta)\mathbf{1}(\mathbf{V}_j^\top\theta \le \mathbf{V}_l^\top\theta)f(\theta)d\theta$. In general, θ is assumed to follow a uniform distribution on the unit sphere. As shown in Escanciano (2006),

$$A_{ijl} = \frac{\pi^{(p+q)/2-1}}{\Gamma(\frac{p+q}{2}+1)}\left|\pi - \arccos\left\{\frac{(\mathbf{V}_i-\mathbf{V}_l)^\top(\mathbf{V}_j-\mathbf{V}_l)}{\|\mathbf{V}_i-\mathbf{V}_l\|\|\mathbf{V}_j-\mathbf{V}_l\|}\right\}\right|, \tag{2.4}$$

where $\Gamma(\cdot)$ is the gamma function. The proposed test $\mathcal{T}_{n,CvM}$ has the merit of computational expedience because only simple algebraic operations are involved.

2.3. A new random approximation computation procedure

Although the uniform distribution assumption of the projection parameter θ is generally accepted (Escanciano (2006); Conde-Amboage, Sánchez-Sellero and González-Manteiga (2015)), it is interesting to investigate the effect of other distributions. Under these circumstances, A_{ijl} cannot be calculated using formula (2.4). However, it is unclear whether an alternative expression for A_{ijl} is available. Therefore, we develop a new procedure to compute A_{ijl} by employing random approximation.

Note that $A_{ijl} = \int \mathbf{1}(\mathbf{V}_i^\top\theta \le \mathbf{V}_l^\top\theta)\mathbf{1}(\mathbf{V}_j^\top\theta \le \mathbf{V}_l^\top\theta)f(\theta)d\theta = \mathrm{E}_\theta\{\mathbf{1}(\mathbf{V}_i^\top\theta \le \mathbf{V}_l^\top\theta)\mathbf{1}(\mathbf{V}_j^\top\theta \le \mathbf{V}_l^\top\theta)|\mathbf{V}_i,\mathbf{V}_j,\mathbf{V}_l\}$, for $i,j,l = 1,\ldots,n$, which means that A_{ijl} is represented as the conditional expectation of a function of θ. Generate an i.i.d. random sequence $\{\theta_1,\ldots,\theta_m\}$ from the density function $f(\theta)$, and define $\hat{A}_{ijl} = m^{-1}\sum_{k=1}^{m}\mathbf{1}(\mathbf{V}_i^\top\theta_k \le \mathbf{V}_l^\top\theta_k)\mathbf{1}(\mathbf{V}_j^\top\theta_k \le \mathbf{V}_l^\top\theta_k)$. Then, we obtain an approximation of the test statistic $\mathcal{T}_{n,CvM}$ by calculating

$$\hat{\mathcal{T}}_{n,CvM} =: \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\hat{\varepsilon}_n(Y_i,\mathbf{X}_i,\mathbf{Z}_i)\hat{\varepsilon}_n(Y_j,\mathbf{X}_j,\mathbf{Z}_j)\hat{A}_{ijl}. \tag{2.5}$$

Remark 1. The above random approximation method is very similar to the number theoretic method in Zhu, Fang and Bhatti (1997). Comparatively, the random approximation method is easier to implement, and the resulting test maintains good theoretical properties by the law of large numbers.

Remark 2. Formula (2.5) shows that even if θ follows distributions other than the uniform distribution, the tests can be realized based on the random sequence generated from $f(\theta)$. In many cases, however, it is difficult to generate a random sequence from a known density function. This difficulty can be overcome with the aid of uniform random numbers. Note too that $A_{ijl} = \mathrm{E}_\eta\{\mathbf{1}(\mathbf{V}_i^\top \eta \leq \mathbf{V}_l^\top \eta)\mathbf{1}(\mathbf{V}_j^\top \eta \leq \mathbf{V}_l^\top \eta) f(\eta) | \mathbf{V}_i, \mathbf{V}_j, \mathbf{V}_l\} C_{p+q}$, for $i, j, l = 1, \ldots, n$, where η is a uniformly distributed random vector on the unit sphere, and C_{p+q} denotes the volume of the unit sphere in $\mathcal{R}^{p+q}$. Generate an i.i.d. random sequence $\{\eta_1, \ldots, \eta_m\}$ of η, and let $\widetilde{A}_{ijl} = m^{-1} \sum_{k=1}^m \mathbf{1}(\mathbf{V}_i^\top \eta_k \leq \mathbf{V}_l^\top \eta_k)\mathbf{1}(\mathbf{V}_j^\top \eta_k \leq \mathbf{V}_l^\top \eta_k) f(\eta_k) C_{p+q}$. For some large m, $\widetilde{A}_{ijl}$ can approximate A_{ijl} well. Then, we obtain the value of the test statistic $\mathcal{T}_{n,CvM}$ by calculating $\widetilde{\mathcal{T}}_{n,CvM} =: n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \hat{\varepsilon}_n(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ $\hat{\varepsilon}_n(Y_j, \mathbf{X}_j, \mathbf{Z}_j)\widetilde{A}_{ijl}$.

3. KS Test Statistic

The KS test is another popular option for checking the adequacy of regression models. As for problem (1.1), together with the CvM test in (2.2), the KS test statistic should be built as

$$\mathcal{T}_{n,KS} = \sup_t \int |\mathcal{M}_{n,pro}(t)| \, f(\theta) d\theta =: \sup_t B_n(t).$$

Though the linear indicator weighting function is widely used to construct CvM-type tests with dimension-reduction effects (Escanciano (2006); Conde-Amboage, Sánchez-Sellero and González-Manteiga (2015); Colling and Van Keilegom (2017)), the main reason that there is no KS-type test with the linear indicator weighting function is that its calculation is challenging and cannot be achieved analogously to $\mathcal{T}_{n,CvM}$ with the help of (2.4). We fill this gap and propose a strategy for calculating $\mathcal{T}_{n,KS}$ that employs a random approximation to avoid a direct application of (2.4). The strategy is stated as follows.

First, generate an i.i.d. random sequence $\{\theta_1, \ldots, \theta_m\}$ from the density function $f(\theta)$. Then, for given t, define $\hat{B}_n(t) = m^{-1} n^{-1/2} \sum_{k=1}^m \left| \sum_{i=1}^n \hat{\varepsilon}_n(Y_i, \mathbf{X}_i, \mathbf{Z}_i) \mathbf{1}(\mathbf{V}_i^\top \theta_k \leq t) \right|$. By the law of large numbers, it is clear that $\hat{B}_n(t)$ is an appropriate approximation of $B_n(t)$. Thus, $\mathcal{T}_{n,KS}$ can be estimated by

$$\widehat{\mathcal{T}}_{n,KS} =: \sup_t \left\{ \frac{1}{m\sqrt{n}} \sum_{k=1}^m \left| \sum_{i=1}^n \hat{\varepsilon}_n(Y_i, \mathbf{X}_i, \mathbf{Z}_i) \mathbf{1}(\mathbf{V}_i^\top \theta_k \leq t) \right| \right\}. \tag{3.1}$$

Remark 3. Similarly to the discussion in Remark 2, an alternative method to computing $B_n(t)$ is to compute $\widetilde{B}_n(t) =: m^{-1} n^{-1/2} \sum_{k=1}^m \left| \sum_{i=1}^n \hat{\varepsilon}_n(Y_i, \mathbf{X}_i,\right.$

$\mathbf{Z}_i)\mathbf{1}(\mathbf{V}_i^\top \eta_k \le t)|f(\eta_k)C_{p+q}$, where the i.i.d. random sequence $\{\eta_1, \dots, \eta_m\}$ follows a uniform distribution on the unit sphere, and C_{p+q} denotes the volume of the unit sphere in $\mathcal{R}^{p+q}$.

4. Asymptotic Distributions and Determining Critical Values

4.1. Asymptotic distributions under the null hypothesis

In this subsection, we investigate the asymptotic properties of the tests under the null hypothesis in (1.1).

Theorem 1. *Suppose that Conditions* (C1)–(C8) *in Appendix B hold. Under the null hypothesis in* (1.1), $\mathcal{M}_{n,pro}(t)$ *converges in distribution to* $\mathcal{M}_{pro}(t)$, *where* $\mathcal{M}_{pro}(t)$ *is a centered Gaussian process with covariance function* $Cov\{\mathcal{M}_{pro}(t_1), \mathcal{M}_{pro}(t_2)\} = Cov\{\mathcal{IF}_{(t_1,\theta)}(Y, \mathbf{X}, \mathbf{Z}, \nu, U), \mathcal{IF}_{(t_2,\theta)}(Y, \mathbf{X}, \mathbf{Z}, \nu, U)\}$. *Here,* $\mathcal{IF}_{(t,\theta)}(Y, \mathbf{X}, \mathbf{Z}, \nu, U)$ *is defined in Appendix A. Furthermore, we have*

$$\mathcal{T}_{n,CvM} \xrightarrow{L} \int \int \{\mathcal{M}_{pro}(t)\}^2 f(\theta) F_\theta(dt) d\theta,$$
$$\mathcal{T}_{n,KS} \xrightarrow{L} \sup_t \int |\mathcal{M}_{pro}(t)| f(\theta) d\theta,$$

where $F_\theta(t)$ *is the conditional distribution of* $\nu^\top\theta$ *given* θ.

Theorem 1 indicates that the asymptotic distributions of the test statistics $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$ are the distributions of $\int\int\{\mathcal{M}_{pro}(t)\}^2 f(\theta)F_\theta(dt)d\theta$ and $\sup_t \int |\mathcal{M}_{pro}(t)|f(\theta)d\theta$, respectively.

Let $F_m(\theta)$ be the empirical distribution function based on $\{\theta_1, \dots, \theta_m\}$. Then, $\hat{\mathcal{T}}_{n,CvM}$ and $\hat{\mathcal{T}}_{n,KS}$ can be written as $\hat{\mathcal{T}}_{n,CvM} = \int\int\{\mathcal{M}_{n,pro}(t)\}^2 F_{n\theta}(dt) F_m(d\theta)$ and $\hat{\mathcal{T}}_{n,KS} = \sup_t \int |\mathcal{M}_{n,pro}(t)|\, F_m(d\theta)$, respectively. From the results of Theorem 1, the following conclusion holds.

Corollary 1. *Suppose that Conditions* (C1)–(C8) *in Appendix B hold. Under the null hypothesis in* (1.1), *we have* $\hat{\mathcal{T}}_{n,CvM} \xrightarrow{L} \int\int\{\mathcal{M}_{pro}(t)\}^2 f(\theta)\ F_\theta(dt)d\theta$ *and* $\hat{\mathcal{T}}_{n,KS} \xrightarrow{L} \sup_t \int |\mathcal{M}_{pro}(t)|f(\theta)d\theta$.

4.2. Determination of the critical values

The distributions of $\int\int\{\mathcal{M}_{pro}(t)\}^2 f(\theta)F_\theta(dt)d\theta$ and $\sup_t \int |\mathcal{M}_{pro}(t)|f(\theta)d\theta$ are very complex. Thus, their upper quantiles and, in turn, the critical values of the proposed tests cannot be obtained directly. In assessing the adequacy of general parametric models, Stute (1997) approximates the critical values of CvM tests using a principal component decomposition of the covariance operator.

We apply a data-driven bootstrap method to determine the critical values that perform well for both the CvM and the KS tests. The rationale for the bootstrap method can be found in Stute, González Manteiga and Presedo Quindimil (1998). Our implementation is described as follows.

Step 1: Generate an i.i.d. random variable sequence $\{e_1, \ldots, e_n\}$ with mean zero, variance one and a finite third moment. Let $\widetilde{Y}_i^* = g(\hat{\mathbf{X}}_i, \mathbf{Z}_i, \hat{\beta}_n) + \{\hat{Y}_i - g(\hat{\mathbf{X}}_i, \mathbf{Z}_i, \hat{\beta}_n)\}e_i$, for $i = 1, \ldots, n$.

Step 2: Calculate the statistics $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$.

Step 3: Based on the bootstrap sample $\{(\widetilde{Y}_i^*, \hat{\mathbf{X}}_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, calculate the statistics $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$, denoted by $\mathcal{T}^*_{n,CvM}$ and $\mathcal{T}^*_{n,KS}$, respectively.

Step 4: Repeat **Step 3** ρ times and obtain $\{\mathcal{T}^*_{n1,CvM}, \ldots, \mathcal{T}^*_{n\rho,CvM}\}$ and $\{\mathcal{T}^*_{n1,KS}, \ldots, \mathcal{T}^*_{n\rho,KS}\}$. Calculate the $1-\alpha$ empirical quantiles based on $\{\mathcal{T}^*_{n1,CvM}, \ldots, \mathcal{T}^*_{n\rho,CvM}\}$ and $\{\mathcal{T}^*_{n1,KS}, \ldots, \mathcal{T}^*_{n\rho,KS}\}$, which are taken as the α-level critical values.

The above scheme is easy to implement without estimating other quantities, such as the complicated influence function $\mathcal{IF}_{(t,\theta)}(Y, \mathbf{X}, \mathbf{Z}, \nu, U)$ in (A.1). In addition, it is acceptable to take the number of repetitions ρ to be 300, 500, or 1000, in general.

4.3. Asymptotic distributions under alternative hypotheses

In this subsection, the asymptotic distributions of the test statistics $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$ are established under the alternative hypothetical models:

$$\mathcal{H}_{1,local} : Y = g(\mathbf{X}, \mathbf{Z}, \beta) + C_n S(\mathbf{X}, \mathbf{Z}) + \varepsilon, \tag{4.1}$$

where $\mathrm{E}(\varepsilon|\mathbf{X}, \mathbf{Z}) = 0$ and $S(\cdot, \cdot)$ is a measurable function that satisfies $0 < \mathrm{E}\{S^2(\mathbf{X}, \mathbf{Z})\} < \infty$ and cannot take the form of $g(\mathbf{X}, \mathbf{Z}, \beta)$.

Theorem 2. *Suppose that Conditions* (C1)–(C8) *in Appendix B hold.*

*(*1*) Under the local alternative hypothetical models* (4.1) *with $C_n = n^{-1/2}$,*

$$\mathcal{T}_{n,CvM} \xrightarrow{L} \int\int \{\mathcal{M}_{pro}(t) + \mathcal{DR}_t\}^2 f(\theta) F_\theta(dt) d\theta,$$

$$\mathcal{T}_{n,KS} \xrightarrow{L} \sup_t \int |\mathcal{M}_{pro}(t) + \mathcal{DR}_t| f(\theta) d\theta,$$

with $\mathcal{DR}_t$ defined in Appendix A.

(2) Under the local alternative hypothetical models (4.1) *with* $C_n n^{1/2} \to \infty$*, we have* $\mathcal{T}_{n,CvM} \to \infty$ *and* $\mathcal{T}_{n,KS} \to \infty$.

Remark 4. Similarly to the arguments of Corollary 1, we can conclude that $\hat{\mathcal{T}}_{n,CvM}$ ($\hat{\mathcal{T}}_{n,KS}$) has the same asymptotic property as $\mathcal{T}_{n,CvM}$ ($\mathcal{T}_{n,KS}$) under the alternative hypotheses (4.1).

Remark 5. Let $\mathcal{H}_{1n} : Y = g(\mathbf{X}, \mathbf{Z}, \beta) + n^{-1/2} S(\mathbf{X}, \mathbf{Z}) + \varepsilon$, $\mathcal{H}_{2n} : Y = g(\mathbf{X}, \mathbf{Z}, \beta) + S(\mathbf{X}, \mathbf{Z}) + \varepsilon$, and $\mathcal{H}_{3n} : Y = g(\mathbf{X}, \mathbf{Z}, \beta) + C_n S(\mathbf{X}, \mathbf{Z}) + \varepsilon$, with $C_n n^{1/2} \to \infty$. For both $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$, the powers $\Pr\{\text{Reject } \mathcal{H}_0 | \mathcal{H}_{1n}\}$ are larger than the test level α. Therefore, the proposed tests can detect the Pitman alternative hypothesis models converging to the null hypothesis model at a rate of $n^{-1/2}$. Under $\mathcal{H}_{2n}$ and $\mathcal{H}_{3n}$, the tests $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$ converge to infinity, and therefore have asymptotic power one.

Remark 6. Zhang, Li and Feng (2015) also investigated the model checking problem (1.1). For the alternative hypothesis models (4.1) with $C_n = n^{-1/2}$, the asymptotic expansion for the test statistic in Zhang, Li and Feng (2015) also includes a drift function $Cov\{l(\mathbf{X}), S(\mathbf{X}, \mathbf{Z})\} F'_\varepsilon$, where $l(\mathbf{X})$ is a weighting function and F'_ε is the derivative of the distribution of the model error ε. If $l(\mathbf{X})$ is orthogonal to the deviation function $S(\mathbf{X}, \mathbf{Z})$, the test of Zhang, Li and Feng (2015) loses effect. Therefore, the choice of the weighting function is critical. For the proposed tests, the drift function $\mathcal{DR}_t$ is nonzero, and the deficit is effectively avoided.

Remark 7. Assume that the null hypothesis is not true and the data are generated from $\mathcal{H}_{4n} : Y = G(\mathbf{X}, \mathbf{Z}) + \varepsilon$, where the nonzero measurable function $G(\mathbf{X}, \mathbf{Z})$ cannot take the form of $g(\mathbf{X}, \mathbf{Z}, \beta)$. Let $Y = g(\mathbf{X}, \mathbf{Z}, \beta) + \{G(\mathbf{X}, \mathbf{Z}) - g(\mathbf{X}, \mathbf{Z}, \beta)\} + \varepsilon =: g(\mathbf{X}, \mathbf{Z}, \beta) + S^*(\mathbf{X}, \mathbf{Z}) + \varepsilon$. The results that $\mathcal{T}_{n,CvM} \to \infty$ and $\mathcal{T}_{n,KS} \to \infty$ under the alternative hypothesis models in $\mathcal{H}_{4n}$ can be proved from the results of Theorem 2. The tests $\mathcal{T}_{n,CvM}$ and $\mathcal{T}_{n,KS}$ have asymptotic power one for any alternative model in $\mathcal{H}_{4n}$, and are consistent in terms of $\Pr\{\text{Reject } \mathcal{H}_0 | \mathcal{H}_0 \text{ is false}\} \to 1$ as $n \to \infty$.

5. Numerical Studies

5.1. Simulation studies

In this subsection, simulation studies are carried out to evaluate the performance of the proposed tests. The following three settings are considered.

Setting 1. We first consider two-dimensional models of the following forms:

$$Y = \beta_1 X_1 + \beta_2 X_2 + C\exp(0.5X_2) + \varepsilon, \tag{5.1}$$

$$Y = \beta_1 + X_1(1 + X_2)^{\beta_2} + C\exp(0.5X_2) + \varepsilon, \tag{5.2}$$

where $\mathbf{X} \sim \mathcal{U}_2[1,2]$. These models are also considered by Zhang, Li and Feng (2015). Set $(\beta_1, \beta_2) = (2,3), C = 0.0, 0.2, 0.4, 0.6, 0.8$ and $(\beta_1, \beta_2) = (1,2), C = 0.0, 0.1, 0.2, 0.3, 0.4$ for models (5.1) and (5.2), respectively. We further let the distorting functions related to $\mathbf{X}$ be $\gamma_1(U) = 1 + 0.3\cos(2\pi U)$ and $\gamma_2(U) = 1 + 0.2(U^2 - 1/3)$.

Setting 2. Consider the following five-dimensional linear candidate models:

$$Y = \beta^\top \mathbf{X} + 2C\exp(0.5X_2) + \varepsilon, \tag{5.3}$$

where $\mathbf{X} \sim \mathcal{U}_5[1,2]$, $\beta = (1,1,1,1,1)^\top$. The distorting functions related to $\mathbf{X}$ are chosen to be $\gamma_1(U) = 1 + 0.3\cos(2\pi U)$, $\gamma_2(U) = 1 + 0.2(U^2 - 1/3)$, $\gamma_3(U) = U + 1/2$, $\gamma_4(U) = 1 + 0.2(U^2 - 1/3)$, and $\gamma_5(U) = U^2 + 2/3$. The constant C is equal to $0.0, 0.1, 0.2, 0.3, 0.4$.

Setting 3. Consider the following 10-dimensional linear candidate models:

$$Y = \beta_1^\top \mathbf{X} + \beta_2^\top \mathbf{Z} + 0.1C\exp(\beta_3^\top \mathbf{X}) + \varepsilon, \tag{5.4}$$

where $\mathbf{X} \sim \mathcal{U}_6[1,2]$, $\mathbf{Z} \sim \mathcal{U}_4[1,2]$, $\beta_1 = (1,1,1,1,-1,-1)^\top$, $\beta_2 = (-1,-1,-1,-1)^\top$, and $\beta_3 = (1,1,0,0,0,0)^\top$. The distorting functions follow the forms of $\gamma_1(U) = \gamma_2(U) = \gamma_3(U) = 1 + 0.3\cos(2\pi U)$ and $\gamma_4(U) = \gamma_5(U) = \gamma_6(U) = 1 + 0.2(U^2 - 1/3)$. The constant C is set to be $0.0, 0.1, 0.2, 0.3, 0.4$.

In Settings 1–3, the distorting function related to the response variable Y is set to be $\psi(U) = 1 + 0.2\cos(2\pi U)$, with the confounding variable $U \sim \mathcal{U}[0,1]$, and the model error ε is generated from a normal distribution with mean zero and standard deviation 0.15. The null hypothesis holds if and only if $C = 0$. Moreover, $\mathbf{X}$ and ε are independent. To obtain $\hat{\psi}_n(u)$ and $\hat{\gamma}_{nr}(u)$ for $r = 1, \ldots, p$, the Epanechnikov kernel function is employed. A significance level of 0.05 and sample sizes of $n = 100, 200, 300$ are considered. In the bootstrap operation, the number of replications ρ is set to 1,000. The empirical sizes and powers are computed based on 500 repetitions.

The following five test methods are considered: the CvM test $\mathcal{T}_{n,CvM}$ in (2.3) with A_{ijl} computed from (2.4), and the proposed CvM and KS tests with θ following the uniform distribution, denoted by ($\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^U_{n,KS}$), and with θ

following the standard normal distribution, denoted by ($\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^N_{n,KS}$). The approximate formulae (2.5) and (3.1) are employed when calculating the empirical sizes and powers of the last four tests.

Choice of bandwidth: Instead of considering all five tests, we take test $\mathcal{T}_{n,CvM}$ as an example to examine the impact of the bandwidth. Let $\hat{\sigma}_U$ be the sample deviation of the confounding variable U. For the Epanechnikov kernel function, the optimal bandwidth for the local constant kernel estimation of the mean regression function is $2.34\hat{\sigma}_U n^{-1/5}$, according to the rule of thumb (Silverman (1986)). For the considered model checking problem, undersmoothing is necessary, and $2.34\hat{\sigma}_U n^{-1/3}$ may be a reasonable choice.

Based on the above considerations, for the two-dimensional model (5.1), the five-dimensional model (5.3), and the 10-dimensional model (5.4), we calculate the empirical sizes and powers by choosing $h_n = C_h\hat{\sigma}_U n^{-1/3}$ and letting C_h be 11 grid points from 1.34 to 3.34 at equal intervals of 0.2. Figure 1 displays the rejection frequencies of the null hypothesis for the test $\mathcal{T}_{n,CvM}$ with different values of C_h and C. When $C = 0$, these rejection frequencies are empirical sizes that approximate the type-I error of the test. When $C > 0$, these rejection frequencies refer to the empirical power.

Figure 1 shows that with different values of C_h, the empirical type-I error of the test can be controlled well and the empirical power remains almost unchanged for low-dimensional models (5.1) and (5.3). For the 10-dimensional model (5.4), the choice of C_h does affect the empirical size and power, although this effect weakens gradually as C and n increase. The same phenomenon was also reported in Wang et al. (2020).

As shown in Zhu, Guo and Zhu (2017) and Wang et al. (2020), the optimal bandwidth choice in studies on model adequacy tests remains an open problem that requires further research. We employ a bandwidth of $2.34\hat{\sigma}_U n^{-1/3}$ in the following simulation studies for all settings.

Choice of m in random approximation procedures: Random approximation procedures are employed to calculate the empirical sizes and powers of the tests $\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^U_{n,KS}$, and $\mathcal{T}^N_{n,KS}$. We consider the two-dimensional model (5.1) and the 10-dimensional model (5.4) as examples to illustrate the impact of m. Specifically, m is taken as evenly spaced points in the interval $[25, 300]$ with a spacing of 25.

Figures 2 and 3 show the empirical sizes and powers of the tests $\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^U_{n,KS}$, and $\mathcal{T}^N_{n,KS}$ against different values of m and C at the 5% significance level with sample size $n = 100$ and bandwidth $h = 2.34\hat{\sigma}_U n^{-1/3}$ for

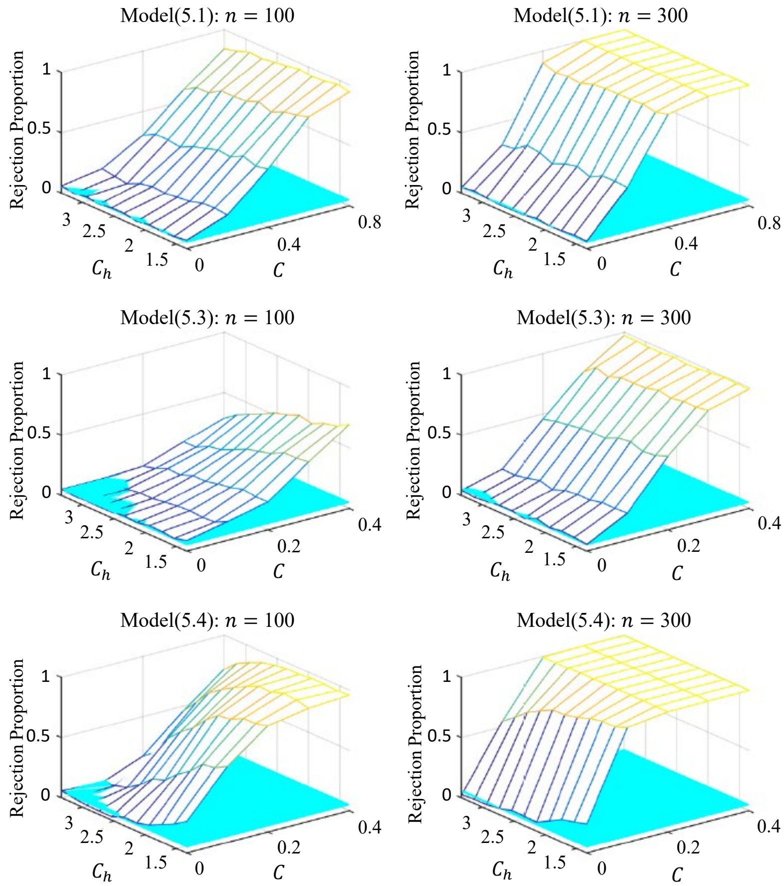

Figure 1. Rejection frequencies of the null hypothesis for the test $\mathcal{T}_{n,CvM}$ against different values of C and C_h at the 5% significance level, with sample sizes 100 (left panel) and 300 (right panel) for models (5.1) (upper row), (5.3) (middle row), and (5.4) (lower row). The horizontal plane corresponds to the 5% significance level.

models (5.1) and (5.4). All four tests are not sensitive to the choice of m. We set $m = 50$, for the sake of simplicity.

We calculate the empirical sizes and powers for models (5.1)–(5.4) and present the results in Tables 1 and 2. For comparison purposes, the tests in Zhang, Li and Feng (2015) and Zhao and Xie (2018) are also considered, which are called $\mathcal{T}_n^{ZLF}$ and $\mathcal{T}_n^{ZX}$, respectively. For the test of Zhang, Li and Feng (2015), the weighting function is set to $l(\mathbf{X}) = \exp(0.5X_2)$. The Epanechnikov kernel function and a bandwidth of $\hat{\sigma}_U n^{-1/3}$ were used. These choices are the same as those in Zhang, Li and Feng (2015). The results are also listed in Tables 1 and 2. The naive

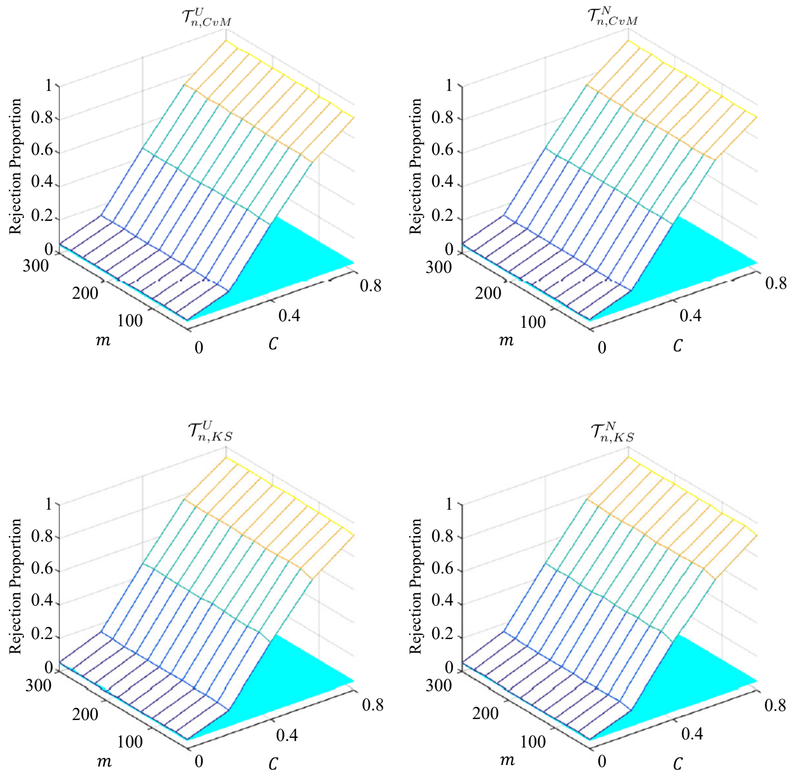

Figure 2. Rejection frequencies of the null hypothesis for the tests $\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^U_{n,KS}$, and $\mathcal{T}^N_{n,KS}$ against different values of m and C at the 5% significance level with sample size $n = 100$ and bandwidth $h = 2.34\hat{\sigma}_U n^{-1/3}$ for model (5.1). The horizontal plane corresponds to the 5% significance level.

method, which ignores the measurement error, is not considered here, because Zhao and Xie (2018) showed that it performs poorly.

From Tables 1 and 2, we observe that the empirical sizes of the five proposed tests are close to the nominal levels in all settings, whereas tests $\mathcal{T}^{ZLF}_n$ and $\mathcal{T}^{ZX}_n$ tend to yield lower empirical sizes, especially for settings 2 and 3, that is, the five-dimensional and 10-dimensional models. Second, with increases in the sample size and the value of C, the empirical power of all seven tests increases, and the five proposed tests perform better than the tests $\mathcal{T}^{ZLF}_n$ and $\mathcal{T}^{ZX}_n$ in terms of empirical power. Moreover, the proposed tests are barely affected by the dimensions of the covariates, whereas the local smoothing test $\mathcal{T}^{ZX}_n$ performs poorly for the five-dimensional and 10-dimensional models. It can be concluded that the proposed methods have advantages in terms of power performance and

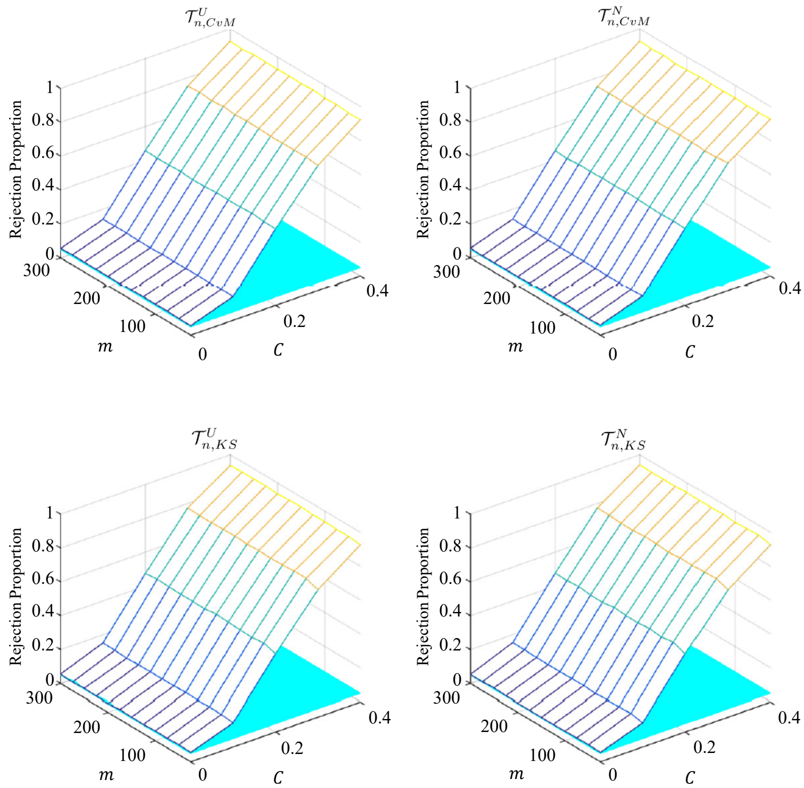

Figure 3. Rejection frequencies of the null hypothesis for the tests $\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^U_{n,KS}$, and $\mathcal{T}^N_{n,KS}$ against different values of m and C at the 5% significance level with sample size $n = 100$ and bandwidth $h = 2.34\hat{\sigma}_U n^{-1/3}$ for model (5.4). The horizontal plane corresponds to the 5% significance level.

dealing with the "curse of dimensionality."

As mentioned above, this study is the first to apply the KS test with the dimension-reduction effect to check the adequacy of regression models. The simulation results show that the proposed KS tests can control the type-I error and yield satisfactory empirical power. As a useful test type, it is worthwhile investigating the performance of KS tests with the dimension-reduction effect when checking other regression models.

Another issue is the effect of the projection parameter selection. The simulation results show that the tests $\mathcal{T}_{n,CvM}$ and $\mathcal{T}^U_{n,CvM}$ yield almost the same results. Note that the tests $\mathcal{T}_{n,CvM}$ and $\mathcal{T}^U_{n,CvM}$ are based on the formula of Escanciano (2006) and on the random approximation to compute A_{ijl} for $i, j, l = 1, \ldots, n$, respectively. Furthermore, the empirical sizes and powers are very similar for the

Table 1. Results for Setting 1. Empirical sizes and powers of $\mathcal{T}_{n,CvM}$, $\mathcal{T}^{U}_{n,CvM}$, $\mathcal{T}^{N}_{n,CvM}$, $\mathcal{T}^{U}_{n,KS}$, $\mathcal{T}^{N}_{n,KS}$, $\mathcal{T}^{ZLF}_{n}$, and $\mathcal{T}^{ZX}_{n}$ at the 5% significance level for the two-dimensional models (5.1) and (5.2).

Model	n	C	$\mathcal{T}_{n,CvM}$	$\mathcal{T}^{U}_{n,CvM}$	$\mathcal{T}^{N}_{n,CvM}$	$\mathcal{T}^{U}_{n,KS}$	$\mathcal{T}^{N}_{n,KS}$	$\mathcal{T}^{ZLF}_{n}$	$\mathcal{T}^{ZX}_{n}$
(5.1)	100	0.0	0.058	0.058	0.058	0.058	0.056	0.028	0.004
		0.2	0.146	0.148	0.146	0.148	0.150	0.064	0.006
		0.4	0.420	0.420	0.418	0.430	0.426	0.214	0.048
		0.6	0.728	0.726	0.728	0.742	0.742	0.476	0.222
		0.8	0.950	0.948	0.946	0.956	0.954	0.732	0.428
	200	0.0	0.046	0.044	0.046	0.058	0.052	0.032	0.004
		0.2	0.272	0.268	0.262	0.288	0.278	0.128	0.040
		0.4	0.756	0.752	0.748	0.768	0.760	0.488	0.220
		0.6	0.970	0.970	0.970	0.976	0.974	0.822	0.682
		0.8	0.996	0.996	0.996	0.996	0.996	0.962	0.952
	300	0.0	0.048	0.050	0.050	0.052	0.058	0.030	0.006
		0.2	0.410	0.404	0.400	0.406	0.416	0.206	0.074
		0.4	0.902	0.902	0.902	0.912	0.916	0.620	0.460
		0.6	0.996	0.996	0.996	0.996	0.996	0.976	0.922
		0.8	1	1	1	1	1	1	1
(5.2)	100	0.0	0.054	0.056	0.058	0.052	0.058	0.040	0.002
		0.1	0.132	0.142	0.136	0.138	0.152	0.102	0.006
		0.2	0.418	0.420	0.426	0.422	0.408	0.224	0.014
		0.3	0.730	0.726	0.722	0.718	0.722	0.422	0.062
		0.4	0.904	0.904	0.904	0.876	0.914	0.630	0.090
	200	0.0	0.054	0.056	0.054	0.044	0.044	0.040	0.002
		0.1	0.308	0.304	0.302	0.268	0.286	0.182	0.012
		0.2	0.716	0.712	0.708	0.712	0.720	0.448	0.078
		0.3	0.964	0.964	0.964	0.956	0.962	0.728	0.226
		0.4	0.994	0.994	0.994	0.996	0.996	0.934	0.572
	300	0.0	0.052	0.052	0.054	0.056	0.048	0.056	0.010
		0.1	0.418	0.414	0.426	0.390	0.402	0.232	0.032
		0.2	0.896	0.904	0.892	0.894	0.888	0.642	0.160
		0.3	0.994	0.994	0.994	0.994	0.994	0.920	0.590
		0.4	1	1	1	1	1	0.998	0.916

CvM test and the KS test, regardless of whether the projection parameter follows the uniform or the normal distribution. Therefore, the random approximation method is a feasible way of eliminating the calculation difficulties caused by the unknown nuisance parameter θ.

Table 2. Results for Settings 2 and 3. Empirical sizes and powers of $\mathcal{T}_{n,CvM}$, $\mathcal{T}_{n,CvM}^{U}$, $\mathcal{T}_{n,CvM}^{N}$, $\mathcal{T}_{n,KS}^{U}$, $\mathcal{T}_{n,KS}^{N}$, $\mathcal{T}_{n}^{ZLF}$, and $\mathcal{T}_{n}^{ZX}$ at the 5% significance level for the five-dimensional model (5.3) and 10-dimensional model (5.4).

Model	n	C	$\mathcal{T}_{n,CvM}$	$\mathcal{T}_{n,CvM}^{U}$	$\mathcal{T}_{n,CvM}^{N}$	$\mathcal{T}_{n,KS}^{U}$	$\mathcal{T}_{n,KS}^{N}$	$\mathcal{T}_{n}^{ZLF}$	$\mathcal{T}_{n}^{ZX}$
(5.3)	100	0.0	0.054	0.058	0.052	0.056	0.060	0.020	0
		0.1	0.062	0.060	0.068	0.088	0.086	0.036	0.002
		0.2	0.202	0.194	0.208	0.188	0.208	0.060	0.004
		0.3	0.302	0.294	0.308	0.312	0.326	0.092	0
		0.4	0.564	0.558	0.570	0.578	0.582	0.158	0
	200	0.0	0.042	0.052	0.042	0.048	0.048	0.022	0.002
		0.1	0.136	0.134	0.136	0.140	0.132	0.046	0
		0.2	0.378	0.376	0.384	0.384	0.394	0.104	0
		0.3	0.698	0.682	0.700	0.726	0.728	0.218	0.002
		0.4	0.888	0.886	0.886	0.892	0.888	0.376	0.002
	300	0.0	0.052	0.048	0.048	0.054	0.044	0.026	0
		0.1	0.212	0.204	0.212	0.206	0.206	0.048	0.004
		0.2	0.548	0.546	0.558	0.566	0.564	0.144	0
		0.3	0.888	0.888	0.892	0.886	0.894	0.350	0
		0.4	0.984	0.980	0.982	0.982	0.984	0.580	0
(5.4)	100	0.0	0.062	0.054	0.052	0.058	0.058	0.026	0
		0.1	0.296	0.286	0.290	0.310	0.320	0.080	0
		0.2	0.658	0.622	0.656	0.668	0.672	0.108	0
		0.3	0.874	0.862	0.840	0.890	0.878	0.164	0
		0.4	0.936	0.930	0.926	0.958	0.958	0.188	0.006
	200	0.0	0.058	0.052	0.056	0.048	0.058	0.024	0
		0.1	0.684	0.654	0.668	0.692	0.694	0.112	0
		0.2	0.984	0.978	0.980	0.986	0.978	0.250	0.004
		0.3	0.996	0.998	0.996	0.998	0.996	0.448	0.020
		0.4	0.998	0.998	0.998	1	1	0.480	0.056
	300	0.0	0.052	0.056	0.052	0.042	0.056	0.028	0
		0.1	0.848	0.818	0.826	0.840	0.834	0.168	0.006
		0.2	1	0.998	1	1	1	0.436	0.014
		0.3	1	1	1	1	1	0.720	0.088
		0.4	1	1	1	1	1	0.768	0.228

5.2. Analyses of diabetes data

In this subsection, we conduct a real-data analysis of a diabetes data set (Schorling et al. (1997); Willems et al. (1997))(`https://hbiostat.org/data`). This data set has also been analyzed by Şentürk and Nguyen (2006) and Delaigle, Hall and Zhou (2016), where covariate-adjusted linear and nonparametric regression models, respectively, were employed. Our aim is to check whether the

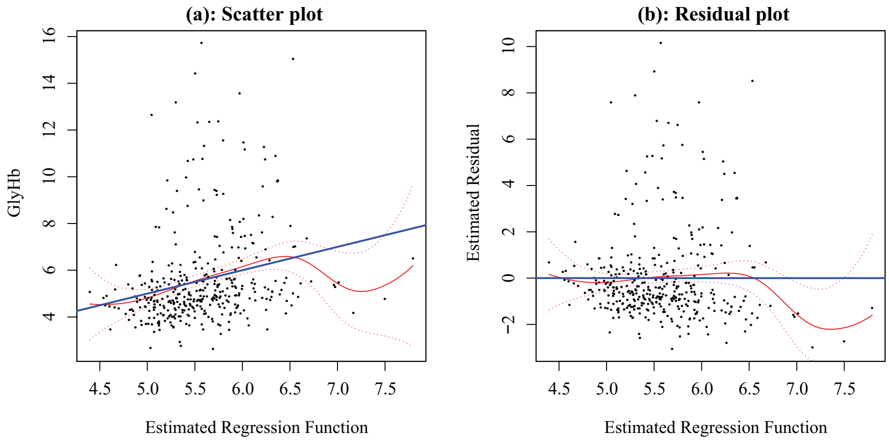

Figure 4. Scatter plots of calibrated GlyHb (a) and the estimated residuals (b) versus the estimated regression function, along with estimated linear (thick lines) and nonparametric (solid lines) regression curves with 95% confidence bands (dotted lines).

following linear model is suitable for these data on 380 individuals:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \varepsilon, \tag{5.5}$$

where Y is the glycosolated haemoglobin level (GlyHb), and X_1, X_2, and Z are the systolic blood pressure (SBP), diastolic blood pressure (DBP), and gender indicator (0, male; 1, female), respectively.

As in Şentürk and Nguyen (2006) and Delaigle, Hall and Zhou (2016), the variables GlyHb, SBP, and DBP are believed to be distorted by the BMI. The settings of the proposed methods are the same as those in the simulation studies. The p-values of the tests $\mathcal{T}_{n,CvM}$, $\mathcal{T}^U_{n,CvM}$, $\mathcal{T}^N_{n,CvM}$, $\mathcal{T}^U_{n,KS}$, and $\mathcal{T}^N_{n,KS}$ are calculated and shown to be 0.005, 0.008, 0.003, 0.007, and 0.002, respectively. The method of Zhang, Li and Feng (2015) was also applied to analyze this data set, yielding p-values of 0.830, 0.265, and 0.599 for different choices of weighting functions $\sin(X)$, $\exp(X)$, and $\cos(X)$. The p-value of the method of Zhao and Xie (2018) was computed to be 0.425. Therefore, the proposed tests suggest rejecting the null hypothesis linear model (5.5), whereas the tests of Zhang, Li and Feng (2015) and Zhao and Xie (2018) cannot reject the null hypothesis. We show scatter plots of the calibrated variable GlyHb and the estimated residuals versus the estimated regression function in Figure 4. The estimated residual curve deviates significantly from a horizontal line, which indicates that the linear model (5.5) is inadequate for this data set.

Supplementary Material

The online Supplementary Material includes the preliminary lemmas, proofs of Theorems 1 and 2, additional simulation studies, and real data analyses.

Acknowledgments

This research was supported by Beijing Natural Science Foundation (Z200001, 1202001), the National Natural Science Foundation of China (12071457, 119710 45, 12101055), NSF grant (DMS-1620898), and Fundamental Research Funds for the Central Universities. Correspondence should be addressed to Feifei Chen.

Appendices

A. Notations

(I) Define $\widetilde{Y}_{m,n} = n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i$, $\widetilde{X}_{m,nr} = n^{-1}\sum_{i=1}^{n}\widetilde{X}_{ri}, r = 1,\ldots,p$, and $S_l(u, h_n) = n^{-1}\sum_{j=1}^{n}(U_j - u)^l K_{h_n}(u - U_j), l = 0,1,2$, where $K(\cdot)$ is a kernel function, h_n is a bandwidth sequence and $K_{h_n}(u) = h_n^{-1}K_h(u/h_n)$.

(II) Denote the derivative of g related to β by $\dot{g}_\beta$. Furthermore, $\ddot{g}_{\beta,x}$, $\ddot{g}_{x,\beta}$ and $g^{(3)}_{\beta,x,\beta}$ can be defined similarly.

(III) Define

$$\hat{\psi}_n(u) = n^{-1}\sum_{j=1}^{n}\frac{\{S_2(u,h_n) - S_1(u,h_n)(U_j - u)\}K_{h_n}(u - U_j)\widetilde{Y}_j}{S_0(u,h_n)S_2(u,h_n) - S_1^2(u,h_n)},$$
$$\hat{\gamma}_n(u) = n^{-1}\sum_{j=1}^{n}\frac{\{S_2(u,h_n) - S_1(u,h_n)(U_j - u)\}K_{h_n}(u - U_j)\widetilde{\mathbf{X}}_j}{S_0(u,h_n)S_2(u,h_n) - S_1^2(u,h_n)},$$

$\Gamma_1(t) = \mathrm{E}\{\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)^\top \mathbf{1}(\nu^\top\theta \le t)\}$, $\Sigma = \mathrm{E}\{\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)^\top\}$, $\mathcal{DR}_t = \mathrm{E}\{S(\mathbf{X},\mathbf{Z})\mathbf{1}(\nu^\top\theta \le t)\} - \Gamma_1(t)\Sigma^{-1}\mathrm{E}\{\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)S(\mathbf{X},\mathbf{Z})\}$, $\Omega = (X_1\dot{g}_{x_1}(\mathbf{X},\mathbf{Z},\beta)/\mathrm{E}(X_1),\ldots,X_p\dot{g}_{x_p}(\mathbf{X},\mathbf{Z},\beta)/\mathrm{E}(X_p))^\top$,
$\Sigma_x = \mathrm{E}\{\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)\Omega^\top\}$.

(IV) Let the symbols $\otimes$ and $\oslash$ indicate multiplying and dividing componentwise, respectively. Denote

$$\begin{aligned}
&\mathcal{IF}_{(t,\theta)}(Y,\mathbf{X},\mathbf{Z},\nu,U)\\
&= \{\mathbf{1}(\nu^\top\theta \le t) - \Gamma_1(t)\Sigma^{-1}\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)\}\varepsilon\\
&\quad + \left\{\mathrm{E}\{Y\mathbf{1}(\nu^\top\theta \le t)|U\} - \Gamma_1(t)\Sigma^{-1}\mathrm{E}\{Y\dot{g}_\beta(\mathbf{X},\mathbf{Z},\beta)\}\right\}\frac{\widetilde{Y} - Y}{\mathrm{E}(Y)}
\end{aligned}$$

$$+ \left\{ \mathrm{E}\left\{ (\mathbf{X} \otimes \dot{g}_x(\mathbf{X}, \mathbf{Z}, \beta) \oslash \mathrm{E}(\mathbf{X}))^\top \mathbf{1}(\nu^\top \theta \le t) | U_j \right\} - \Gamma_1(t)\Sigma^{-1}\Sigma_x \right\} \left(\widetilde{\mathbf{X}} - \mathbf{X} \right). \tag{A.1}$$

(V) Let $\Delta_{ni} = (\Delta_{n1i}, \ldots, \Delta_{npi})^\top$ with $\Delta_{nri} = X_{ri}\{\gamma_r(U_i)\widetilde{X}_{m,nr} - \hat{\gamma}_{nr}(U_i)\}/\hat{\gamma}_{nr}(U_i)$ for $i = 1, \ldots, n$ and $r = 1, \ldots, p$. Define $\tilde{\Delta}_{ij} = (\tilde{\Delta}_{1ij}, \ldots, \tilde{\Delta}_{pij})^\top$ with $\tilde{\Delta}_{rij} = X_{ri}\{\gamma_r(U_j)\mathrm{E}(X_r) - X_{rj}\}/\mathrm{E}(\widetilde{X}_r|U = U_i)$ for $i, j = 1, \ldots, n$ and $r = 1, \ldots, p$.

B. Conditions

(C1) The density function of U, $f_u(u)$, is bounded away from zero and satisfies Lipschitz condition of order 1 on the support of U.

(C2) (i) The functions $\psi(u)$ and $\gamma_r(u)$, $r = 1, \ldots, p$, have bounded and continuous derivatives. (ii) The functions $\psi(u)$ and $\gamma_r(u)$, $r = 1, \ldots, p$, are non-zero on the support set of U.

(C3) $\mathrm{E}(Y)$ and $\mathrm{E}(X_r), r = 1, \ldots, p$, are bounded away from zero. $\mathrm{E}(|Y|^3) < \infty$ and $\mathrm{E}(|X_r|^3) < \infty$, $r = 1, \ldots, p$.

(C4) The matrix $\Sigma = \mathrm{E}\{\dot{g}_\beta(\mathbf{X}, \mathbf{Z}, \beta)\dot{g}_\beta(\mathbf{X}, \mathbf{Z}, \beta)^\top\}$ is positive finite.

(C5) The partial derivatives of $g(\mathbf{X}, \mathbf{Z}, \beta)$ with respect to x and β exist and are continuous; the second-order and third-order partial derivatives of $g(\mathbf{X}, \mathbf{Z}, \beta)$ with respect to x and β exist and are bounded.

(C6) The objective function (2.1) has a unique minimizer.

(C7) (i) The kernel function $K(u)$ is a bounded univariate kernel function of order 2 with a bounded support. (ii) The second derivative of $K(u)$ is bounded and satisfies Lipschitz condition.

(C8) The bandwidth h_n satisfies the following conditions: $h_n \to 0$, $nh_n^4 \to 0$ and $\ln n/(nh_n) \to 0$ as $n \to \infty$.

Remark 8. Conditions (C1)-(C3) are also employed in Şentürk and Müller (2006) and Zhang, Li and Feng (2015) aiming for avoiding the case where the denominator is zero. Conditions (C4)-(C6) are necessary for the asymptotic normality of the nonlinear least squares estimator. Conditions (C7) and (C8) are common for the nonparametric kernel method.

References

Colling, B. and Van Keilegom, I. (2017). Goodness-of-fit tests in semiparametric transformation models using the integrated regression function. *J. Multivariate Anal.* **160**, 10–30.

Conde-Amboage, M., Sánchez-Sellero, C. and González-Manteiga, W. (2015). A lack-of-fit test for quantile regression models with high-dimensional covariates. *Comput. Statist. Data Anal.* **88**, 128–138.

Şentürk, D. and Müller, H.-G. (2005). Covariate-adjusted regression. *Biometrika* **92**, 75–89.

Şentürk, D. and Müller, H.-G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Ann. Statist.* **34**, 654–679.

Şentürk, D. and Müller, H.-G. (2009). Covariate-adjusted generalized linear models. *Biometrika* **96**, 357–370.

Şentürk, D. and Nguyen, D. V. (2006). Estimation in covariate-adjusted regression. *Comput. Statist. Data Anal.* **50**, 3294–3310.

Cui, X., Guo, W., Lin, L. and Zhu, L. (2009). Covariate-adjusted nonlinear regression. *Ann. Statist.* **37**, 1839–1870.

Delaigle, A., Hall, P. and Zhou, W. (2016). Nonparametric covariate-adjusted regression. *Ann. Statist.* **44**, 2190–2220.

Deng, S. and Zhao, X. (2019). Covariate-adjusted regression for distorted longitudinal data with informative observation times. *J. Amer. Statist. Assoc.* **114**, 1241–1250.

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

Guo, X., Wang, T. and Zhu, L. (2016). Model checking for parametric single-index models: A dimension reduction model-adaptive approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78**, 1013–1035.

Kaysen, G. A., Dubin, J. A., Müller, H.-G., Mitch, W. E., Rosales, L. M., Levin, N. W. et al. (2002). Relationships among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney Int.* **61**, 2240–2249.

Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *J. Econometrics* **143**, 103–122.

Ma, S., Zhang, J., Sun, Z. and Liang, H. (2014). Integrated conditional moment test for partially linear single index models incorporating dimension-reduction. *Electron. J. Stat.* **8**, 523–542.

Nguyen, D. V. and Şentürk, D. (2008). Multicovariate-adjusted regression models. *J. Stat. Comput. Simul.* **78**, 813–827.

Schorling, J. B., Roach, J., Siegel, M., Baturka, N., Hunt, D. E., Guterbock, T. M. et al. (1997). A trial of church-based smoking cessation interventions for rural African Americans. *Prev. Med.* **26**, 92–101.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**, 613–641.

Stute, W., González Manteiga, W. and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93**, 141–149.

Tan, F., Zhu, X. and Zhu, L. (2018). A projection-based adaptive-to-model test for regressions. *Statist. Sinica* **28**, 157–188.

Wang, M., Liu, C., Xie, T. and Sun, Z. (2020). Data-driven model checking for errors-in-variables varying-coefficient models with replicate measurements. *Comput. Statist. Data Anal.* **141**, 12–27.

Willems, J. P., Saunders, J. T., Hunt, D. E. and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *South Med. J.* **90**, 814–820.

Zhang, J., Li, G.-R. and Feng, Z.-H. (2015). Checking the adequacy for a distortion errors-in-variables parametric regression model. *Comput. Statist. Data Anal.* **83**, 52–64.

Zhang, J., Zhu, L. and Liang, H. (2012). Nonlinear models with measurement errors subject to single-indexed distortion. *J. Multivariate Anal.* **112**, 1–23.

Zhao, J. and Xie, C. (2018). A nonparametric test for covariate-adjusted models. *Statist. Probab. Lett.* **133**, 65–70.

Zhu, L.-X., Fang, K.-T. and Bhatti, M. I. (1997). On estimated projection pursuit-type Crámer-von Mises statistics. *J. Multivariate Anal.* **63**, 1–14.

Zhu, X., Guo, X. and Zhu, L. (2017). An adaptive-to-model test for partially parametric single-index models. *Stat. Comput.* **27**, 1193–1204.

Zhihua Sun

School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China.

E-mail: sunzh@ucas.ac.cn

Feifei Chen

Center for Statistics and Data Science, Beijing Normal University, Zhuhai, 519087, China.

E-mail: chenfeifei12@mails.ucas.ac.cn

Hua Liang

Department of Statistics, George Washington University, Washington, D.C. 20052, USA.

E-mail: hliang@gwu.edu

(Received January 2020; accepted January 2021)