Statistica Sinica **31** (2021), 625-646
doi:https://doi.org/10.5705/ss.202018.0322

TESTING HETEROSCEDASTICITY FOR REGRESSION MODELS BASED ON PROJECTIONS

Falong Tan[1], Xuejun Jiang[2], Xu Guo[3] and Lixing Zhu[3,4]

[1]*Hunan University,* [2]*Southern University of Science and Technology*
[3]*Beijing Normal University and* [4]*Hong Kong Baptist University*

Abstract: We propose a new test for heteroscedasticity in parametric and partial linear regression models in multidimensional spaces. When the dimension of the covariates is large, or even moderate, existing tests for heteroscedasticity perform badly, owing to the "curse of dimensionality." To address this problem, we construct a test for heteroscedasticity that uses a projection-based empirical process. Then, we study the asymptotic properties of the test statistic under the null and alternative hypotheses. The results show that the test detects the departure of local alternatives from the null hypothesis at the fastest possible rate during hypothesis testing. Because the limiting null distribution of the test statistic is not asymptotically distribution free, we propose a residual-based bootstrap approach and investigate the validity of its approximations. Simulations verify the finite-sample performance of the test. Two real-data analyses are conducted to demonstrate the proposed test.

Key words and phrases: Heteroscedasticity testing, partial linear models, projection, U-process.

1. Introduction

In many regression models, the error terms are assumed to have a common variance. However, ignoring the presence of heteroscedasticity in a regression model may result in inefficient inferences of the regression coefficients, or even inconsistent estimators of the variance function. Therefore, regression models should be tested for heteroscedasticity whenever the error terms are assumed to have equal variance. Consider the following regression model:

$$Y = m(Z) + \varepsilon, \tag{1.1}$$

where Y is the dependent variable with a p-dimensional covariate Z, $m(\cdot) = E(Y|Z = \cdot)$ is the regression function, and the error term ε satisfies $E(\varepsilon|Z) =$

Corresponding author: Xuejun Jiang, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong Province, China. E-mail: jiangxj@sustech.edu.cn.

0. Thus, the null hypothesis when testing for heteroscedasticity in regression model (1.1) is as follows:

$$H_0 : Var(Y|Z) = E(\varepsilon^2|Z) \equiv C \quad \text{for some constant } C > 0;$$

the alternative hypothesis is that H_0 is totally incorrect:

$$H_1 : Var(Y|Z) = E(\varepsilon^2|Z) \text{ is a nonconstant function of } Z.$$

Many test for heteroscedasticity in regression model (1.1) haven been proposed in the literature. Cook and Weisberg (1983) constructed a score test for heteroscedasticity in parametric regression models with parametric structure variance functions. Simonoff and Tsai (1994) proposed a modified score test for heteroscedasticity in linear models. Muller and Zhao (1995) developed a data-based test for heteroscedasticity in a general semiparametric variance function model with a fixed design. Dette and Munk (1998) proposed a consistent test for heteroscedasticity in a nonparametric regression setting, based on the L^2-distance between the underlying variance function and the constant variance. Zhu, Fujikoshi and Naito (2001) developed a test for heteroscedasticity based on residual marked empirical processes. Extending the work of Zheng (1996) on checking the regression function, Dette (2002); Zheng (2009) proposed residual-based tests for heteroscedasticity under different regression models. Su and Ullah (2013) introduced a nonparametric test for conditional heteroscedasticity in nonlinear regression models. Recently, following Stute, Xu and Zhu (2008); Chown and Müller (2018) proposed a test for heteroscedasticity based on a weighted-residual empirical distribution function. Lin and Qu (2012) developed a test for heteroscedasticity in nonlinear semi-parametric regression models, based on the work of Dette (2002). Furthermore, Dette, Neumeyer and Van Keilegom (2007); Wang and Zhou (2007); Koul and Song (2010); Pardo-Fernández and Jiménez-Gamero (2019) considered a more general problem of checking the parametric form of the conditional variance function in nonparametric regressions.

To motivate the construction of our test statistic, we first give a detailed comment on two representative tests, namely, those of Zhu, Fujikoshi and Naito (2001); Zheng (2009). Let $E(\varepsilon^2) = \sigma^2$ and $\eta = \varepsilon^2 - \sigma^2$. Then, the null hypothesis H_0 is equivalent to $E(\eta|Z) = 0$. Consequently,

$$E[\eta E(\eta|Z)f(Z)] = 0,$$

where $f(\cdot)$ is the density function of Z. Based on a consistent estimator of $E[\eta E(\eta|Z)f(Z)]$, Zheng (2009) proposed the following test statistic:

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^p} K\left(\frac{Z_i - Z_j}{h}\right) \hat{\eta}_i \hat{\eta}_j,$$

where $\hat{\eta}_i = \hat{\varepsilon}_i^2 - \hat{\sigma}^2, \hat{\sigma}^2 = (1/n)\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, $\hat{\varepsilon}_i = Y_i - \hat{m}(Z_i)$ with $\hat{m}(\cdot)$ being an estimator of the regression function, $K(\cdot)$ is a p-dimensional multivariate kernel function, and h is a bandwidth, which converges to zero as n goes to infinity. However, because Zheng (2009) used nonparametric smooth estimators in its construction, the test statistic suffers severely from the "curse of dimensionality." More specifically, the above test can only detect local alternatives that converge to the null at a rate of $O(1/\sqrt{nh^{p/2}})$. When p is large, this rate could be very slow, which would quickly reduce the power of this test.

Zhu, Fujikoshi and Naito (2001) used residual marked empirical processes to construct a test for heteroscedasticity. Note that

$$E(\eta|Z) = 0 \Leftrightarrow E[\eta I(Z \leq t)] = 0 \text{ for all } t \in \mathbb{R}^p.$$

Based on this, Zhu, Fujikoshi and Naito (2001) proposed the following residual marked empirical process:

$$R_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\eta}_i I(Z_i \leq t).$$

Here, $I(Z_i \leq t) = I(Z_{i1} \leq t_1) \cdots I(Z_{ip} \leq t_p)$, and Z_{ij} and t_j are the j-components of Z_i and t, respectively. The test statistic of Zhu, Fujikoshi and Naito (2001) is a functional of $R_n(t)$, such as the Cramér−von Mises or Kolmogorov−Smirnov functional. The authors show that their test can detect local alternatives converging to the null at the parametric rate $1/\sqrt{n}$, which is the fastest documented convergence rate in context of hypothesis testing. However, when the dimension p of the covariates is large, this test also suffers severely from the dimension problem due to the data sparseness in multidimensional spaces.

The purpose of this study is to develop a test for heteroscedasticity in parametric regression models that avoids the drawbacks of those of Zhu, Fujikoshi and Naito (2001); Zheng (2009) and, thus, can be applied when the dimension of the covariates is relatively large. Note that Zhu, Fujikoshi and Naito (2001) test is consistent against local alternatives converging to the null at the parametric rate

$1/\sqrt{n}$, which is not related to the dimension of the covariates. Nevertheless, their test still suffers from the "curse of dimensionality" in practice. This is because their test statistic is based on the indicator function $I(Z_i \leq t)$, which is the product of p indicator functions. Therefore, the vector $(I(Z_1 \leq t), \ldots, I(Z_n \leq t))^\top$ is very sparse for large p, which cause the dimension problem in practice. To overcome this problem, we propose using the projected covariates $\alpha^\top Z_i$, rather than Z_i, to construct a residual marked empirical process, yielding a test statistic that does not involve the product of p indicator functions. Escanciano (2006); Lavergne and Patilea (2008, 2012) adopted this approach to construct goodness-of-fit tests for parametric regression models. Because the test is based on one-dimensional projections, it behaves as if the dimension of the covariates is one. As a result, this method is less sensitive to the dimension p of the regressors than is the method of Zhu, Fujikoshi and Naito (2001). We employed residual marked empirical processes to construct the test statistic. Thus, our test statistic avoids a nonparametric estimation of $E(\eta|Z)$, which was used in Zheng (2009), and can detect local alternatives converging to the null at the parametric rate $1/\sqrt{n}$. Furthermore, the new test is easy to compute, does not involve multidimensional numerical integrations, and exhibits excellent power for large dimension in finite-sample simulations; see Section 4.

We also use this method to check for heteroscedasticity in partial linear regression models. When the dimension of the covariates is large, a nonparametric estimation is less accurate, owing to the "curse of dimensionality." In addition, partial linear regression models provide a more flexible substitution if the researchers know that some of the covariates enter the regression model linearly. As a result, this model is widely used in economics, biology, and other related fields. To construct the test statistic for partial linear regression models, we need to use locally smoothing methods to estimate the nonlinear part of the regression function. Although it involves nonparametric estimators, we show that the limiting distribution has the same form as that in parametric regression models. Furthermore, we show that the proposed detects local alternatives converging to the null at rate $1/\sqrt{n}$ under this semi-parametric setting.

Chown and Müller (2018) employed a similar procedure to develop test for heteroscedasticity that uses a weighted empirical process based on the indicator function $I(\hat{\varepsilon}_j \leq t)$; rather than $I(\alpha^\top Z_j \leq t)$. This procedure was first proposed by Stute, Xu and Zhu (2008) for checking parametric regression models in high-dimensional settings. However, Chown and Müller (2018) test applies only to location-scale models; that is, $Y = m(Z) + \sqrt{Var(Y|Z)}e$, where e is independent

of Z. The independence between e and Z is then employed to construct suitable test statistics. As in Chown and Müller (2018); Pardo-Fernández and Jiménez-Gamero (2019) rely on this same restriction. Moreover, they considered one-dimensional covariate only. Our proposed test statistic does not require this restriction. In fact, we only need $E(\varepsilon|Z) = 0$, and ε may depend on Z in a more general way. Another issue is that the weighted function $\omega(Z)$ of the empirical processes suggested by Chown and Müller (2018) also relies on nonparametric estimations, regardless of the type of regression function. As a result their test still suffers from the "curse of dimensionality" even for parametric regression models.

The rest of the paper is organized as follows. In Section 2, we define the test statistic using a projection-based empirical process. In Section 3, we study the asymptotic properties of the test statistic under the null and the alternative hypotheses in parametric and partial linear regression models. In Section 4, a residual-based bootstrap method is proposed to approximate the null distribution of the test statistic. Here, we also present our simulation results that compare the performance of the proposed test with that of several existing methods. Furthermore, we analyze two real data sets to illustrate the proposed method. Section 5 concludes the paper. All technical proofs are delegated to the online Supplementary Material.

2. Test Construction

Recall that the null hypothesis H_0 is equivalent to $E(\eta|Z) = 0$. According to Lemma 1 of Escanciano (2006) or Lemma 2.1 of Lavergne and Patilea (2008), we have

$$E(\eta|Z) = 0 \Longleftrightarrow E(\eta|\alpha^\top Z) = 0, \quad \forall\ \alpha \in \mathbb{S}^p,$$

where $\mathbb{S}^p = \{\alpha : \alpha \in \mathbb{R}^p \text{ and } \|\alpha\| = 1\}$. Consequently,

$$E(\eta|Z) = 0 \Longleftrightarrow E[\eta I(\alpha^\top Z \le t)] = 0, \quad \forall\ \alpha \in \mathbb{S}^p,\ t \in \mathbb{R}.$$

Therefore, the null hypothesis H_0 is equivalent to

$$\int_{\mathbb{S}^p} \int_{\mathbb{R}} |E[\eta I(\alpha^\top Z \le t)]|^2 F_\alpha(dt) d\alpha = 0, \tag{2.1}$$

where F_α is the cumulative distribution function of $\alpha^\top Z$, and $d\alpha$ is the uniform density on $\mathbb{S}^p$. Then, we propose the following test statistic, which we use to

check heteroscedasticity in model (1.1):

$$HCM_n = \int_{\mathbb{S}^p} \int_{\mathbb{R}} \frac{1}{n} \left| \sum_{j=1}^{n} \hat{\eta}_j I(\alpha^\top Z_j \le t) \right|^2 F_{n,\alpha}(dt) d\alpha, \tag{2.2}$$

where $F_{n,\alpha}$ is the empirical distribution function of the projected covariates $\{\alpha^\top Z_j, 1 \le j \le n\}$.

Note that the test statistic HCM_n involves a multidimensional integral for large p. Indeed, by some elementary calculations,

$$\begin{aligned} HCM_n &= \frac{1}{n} \sum_{i,j=1}^{n} \hat{\eta}_i \hat{\eta}_j \int_{\mathbb{S}^p} \int_{\mathbb{R}} I(\alpha^\top Z_i \le t) I(\alpha^\top Z_j \le t) F_{n,\alpha}(dt) d\alpha \\ &= \frac{1}{n^2} \sum_{i,j,k=1}^{n} \hat{\eta}_i \hat{\eta}_j \int_{\mathbb{S}^p} I(\alpha^\top Z_i \le \alpha^\top Z_k) I(\alpha^\top Z_j \le \alpha^\top Z_k) d\alpha. \end{aligned}$$

It is well known that multidimensional numerical integrations are extremely difficult to handle in practice. However, the following lemma enables us to avoid multidimensional integrations in the numerical calculations and, thus, obtain an analytic expression for the test statistic HCM_n. Its proof can be found in Appendix B of Escanciano (2006).

Lemma 1. *Let $u_1, u_2 \in \mathbb{R}^p$ be two nonzero vectors, and let $\mathbb{S}^p$ be the p-dimensional unit sphere. Then, we have*

$$\int_{\mathbb{S}^p} I(\alpha^\top u_1 \le 0) I(\alpha^\top u_2 \le 0) d\alpha = \frac{\pi - < u_1, u_2 >}{2\pi},$$

where $d\alpha$ is the uniform density on $\mathbb{S}^p$, and $< u_1, u_2 > = \arccos(u_1^\top u_2 / (\|u_1\| \|u_2\|))$ is the angle between u_1 and u_2.

The integral in Lemma 1 can be viewed as a kernel function. Then, our test statistic has similar form to that of Zheng (2009). However, in contrast to the test of Zheng (2009), our test statistic can be viewed as a U-statistic with a fixed, rather than a varying bandwidth. This is important. From the theory on U-statistics, we know that those with a fixed bandwidth have a parametric convergence rate $1/\sqrt{n}$, which is faster than those with a varying bandwidth. This supports our theoretical results derived using empirical processes. For further information on U-statistics with a fixed bandwidth, see Anderson, Hall and Titterington (1994); Fan (1998).

The proposed test works for all regression models, and avoids the following

drawbacks of the tests of Zhu, Fujikoshi and Naito (2001); Zheng (2009): the nonparametric estimation of $E(\eta|Z)$, multidimensional numerical integration, and the low power when the dimension p is large. Note that the test statistic is based on the residuals $\hat{\varepsilon}_j = Y_j - \hat{m}(Z_j)$, that is, it involves the estimator of the regression function $E(Y|Z = \cdot)$. Thus, our test works well if it does not involve multidimensional nonparametric estimations of $E(Y|Z)$. In this study we deal only with parametric and partial linear regression models, because the test statistic only involves parametric estimations for parametric regression models, and only involves one-dimensional kernel estimations for partial linear regression models. It can also be applied to nonparametric regression models. Then, we have to estimate the unknown regression function in a nonparametric way. Owing to the sparsity of the data in multidimensional spaces, the behavior of nonparametric estimations quickly deteriorates when the dimension of the covariates increases. As a result, the test still suffers from the "curse of dimensionality" for nonparametric regression models in practice. This problem is common to all existing tests for heteroscedasticity in nonparametric regression, because they all need to first estimate the unknown regression function. Therefore, dealing with the dimension problem when testing for heteroscedasticity in nonparametric regression models remains a challenging problem.

3. Asymptotic Results

First, we consider the following parametric regression model:

$$Y = m(Z, \beta) + \varepsilon, \quad E(\varepsilon|Z) = 0, \tag{3.1}$$

where $\beta \in \mathbb{R}^d$, and $m(\cdot, \beta) = E(Y|Z = \cdot)$ is the given regression function. Let $\hat{\beta}_n$ be a consistent estimator of β and $\hat{\varepsilon}_i = Y_i - m(Z_i, \hat{\beta}_n)$. Then, $\hat{\eta}_i = \hat{\varepsilon}_i^2 - \hat{\sigma}^2 = \hat{\varepsilon}_i^2 - (1/n)\sum_{i=1}^n \hat{\varepsilon}_i^2$. Define the projected empirical process as follows:

$$V_n(\alpha, t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\eta}_i I(\alpha^\top Z_i \le t).$$

The test statistic becomes

$$HCM_n = \int_{\mathbb{S}^p} \int_{\mathbb{R}} |V_n(\alpha, t)|^2 F_{n,\alpha}(dt) d\alpha.$$

To obtain the asymptotic properties of $V_n(\alpha, t)$ under the null and the alternatives, we impose several regularity conditions:

(A1) $E(\varepsilon^4) < \infty$;

(A2) $\sqrt{n}(\hat{\beta}_n - \beta) = O_p(1)$

(A3) The parametric regression function $m(z, \gamma)$ is twice continuously differentiable at each γ in a neighborhood of β. Set

$$m'(z,\gamma) = \frac{\partial m(z,\gamma)}{\partial \gamma} \quad \text{and} \quad m''(z,\gamma) = \frac{\partial m(z,\gamma)}{\partial \gamma^\top \partial \gamma}.$$

Assume $E\|m'(Z,\beta)\|^2 < \infty$ and $\|m''(z,\gamma)\| \le M(z)$, with $E|M(Z)|^2 < \infty$, for all γ. Here, $\|\cdot\|$ denotes the Frobenius norm.

Conditions (A1) and (A3) are commonly used in the literature on heteroscedasticity testing; see, for example, Zheng (2009). Condition (A2) is satisfied for the ordinary least squares estimator and its robust modifications; see, Chapters 5 and 7 in Koul (2002).

Theorem 1. *Assume that the regularity conditions* A1–A3 *hold. Under* H_0*, we have*

$$V_n(\alpha, t) \longrightarrow V_\infty(\alpha, t) \quad \text{in distribution,}$$

where $V_\infty(\alpha, t)$ *is a zero-mean Gaussian process with covariance function*

$$K\{(\alpha_1,t_1),(\alpha_2,t_2)\} = E\{\eta^2[I(\alpha_1^\top Z \le t_1) - F_{\alpha_1}(t_1)][I(\alpha_2^\top Z \le t_2) - F_{\alpha_2}(t_2)]\}.$$

Furthermore,

$$HCM_n \longrightarrow \int_{\mathbb{S}^p}\int_{\mathbb{R}} V_\infty(\alpha,t)^2 F_\alpha(dt)d\alpha \quad \text{in distribution.}$$

Next we apply this approach to check for heteroscedasticity in partial linear regression models. Consider

$$Y = \beta^\top X + g(T) + \varepsilon, \quad E(\varepsilon|X,T) = 0, \tag{3.2}$$

where $T \in \mathbb{R}$, $\beta \in \mathbb{R}^q$, and $g(\cdot)$ is a smooth function. Because the nonlinear part $g(T)$ in equation (3.2) is unknown, it has to be estimated in a nonparametric way. Thus, in theoretical investigations, the decomposition of the proposed projected empirical process involves a U-process. With the help of the theory on U-process in the literature, e.g., Nolan and Pollard (1987), we obtain the same asymptotical property as that in Theorem 1 for partial linear regression models.

We now use the kernel method to give the estimators of β and $g(T)$. Note that

$$Y - E(Y|T) = \beta^\top [X - E(X|T)] + \varepsilon.$$

Set $\tilde{Y} = Y - E(Y|T)$ and $\tilde{X} = X - E(X|T)$. It is easy to see that

$$\beta = [E\tilde{X}\tilde{X}^\top]^{-1} E(\tilde{X}\tilde{Y}).$$

Let $\{(X_i, T_i, Y_i)\}_{i=1}^n$ be an independent identically distributed (i.i.d) sample from the distribution of (X, T, Y). The estimator of β is given by

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^n [X_i - \hat{E}(X|T_i)][X_i - \hat{E}(X|T_i)]^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n [X_i - \hat{E}(X|T_i)][Y_i - \hat{E}(Y|T_i)] \right), \tag{3.3}$$

where

$$\hat{E}(X|T_i) = \frac{1}{n} \sum_{j=1, j\neq i}^n \frac{X_j K_h(T_i - T_j)}{\hat{f}_i(T_i)},$$
$$\hat{E}(Y|T_i) = \frac{1}{n} \sum_{j=1, j\neq i}^n \frac{Y_j K_h(T_i - T_j)}{\hat{f}_i(T_i)},$$

and $\hat{f}_i(T_i) = (1/n) \sum_{j=1, j\neq i}^n K_h(T_i - T_j)$. Here, $K_h(t) = (1/h)K(t/h)$, and $K(\cdot)$ is a kernel function satisfying the regularity condition (B3), specified below. To obtain the estimator of $g(\cdot)$, note that $g(T) = E(Y - \beta^\top X|T)$. Thus, the kernel estimator of $g(T)$ has the following form:

$$\hat{g}(T_i) = \frac{1}{n} \sum_{j=1, j\neq i}^n \frac{[Y_j - \hat{\beta}_n^\top X_j] K_h(T_i - T_j)}{\hat{f}_i(T_i)}. \tag{3.4}$$

The following regularity conditions are needed in order to derive the asymptotic distribution of HCM_n in partial linear regression models. In the following, C is a constant, although the value may vary depending on the context.

(B1) Let $E'(Y|T = t)$ be the derivative of $E(Y|T = t)$, and let $F(x|t)$ be the conditional distribution function of X, given $T = t$. Suppose there exists an open neighborhood Θ_1 of zero, such that, for all t and x,

$$|E(X|T=t+u)-E(X|T=u)| \le C|u|, \quad \forall\ u \in \Theta_1;$$
$$|E'(X|T=t+u)-E'(X|T=u)| \le C|u|, \quad \forall\ u \in \Theta_1;$$
$$|F(x|t+u)-F(x|t)| \le C|u|, \quad \forall\ u \in \Theta_1.$$

(B2) $E(Y^4) < \infty, E(\|X\|^4) < \infty$, and there exists a constant C, such that $|E(\varepsilon^2|T=t, X=x)| \le C$, for all t and x.

(B3) The kernel function $K(\cdot)$ is bounded, continuous, symmetric about zero and satisfies the following: (a) the support of $K(\cdot)$ is the interval $[-1,1]$; and (b) $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} |u|K(u)du \neq 0$.

(B4) $nh^4 \to 0$ and $nh^2 \to \infty$ as $n \to \infty$.

The Conditions (B1), (B2), and (B3) are commonly used to derive the asymptotic properties of nonparametric estimators; see, for example, Schick (1996); Zhu and Ng (2003). Condition (B4) is necessary to obtain the limiting distribution of the test statistic.

Lemma 2. *Under the regularity conditions* B1–B4*, we have*

$$\sqrt{n}(\hat{\beta}_n - \beta) = [E\tilde{X}\tilde{X}^\top]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_i\varepsilon_i + O_p\left(\frac{1}{\sqrt{n}h} + \sqrt{n}h^2\right)^{1/2}. \tag{3.5}$$

Lemma 2 can be found in Zhu and Ng (2003). It indicates that, under the regularity condition (B4), $\hat{\beta}_n$ is root-n consistent. Now we can obtain the asymptotic properties of HCM_n in partial linear regression models. Set $p = q+1$ and $Z_i = (X_i^\top, T_i)^\top$. The proposed empirical process and the test statistic have the same form as before,

$$V_n(\alpha, t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\eta}_i I(\alpha^\top Z_i \le t),$$
$$HCM_n = \int_{\mathbb{S}^p}\int_{\mathbb{R}} |V_n(\alpha,t)|^2 F_{n,\alpha}(dt)d\alpha.$$

Here $\hat{\eta}_i = \hat{\varepsilon}_i^2 - \hat{\sigma}^2, \hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, and $\hat{\varepsilon}_i = Y_i - \hat{\beta}_n^\top X_i - \hat{g}(T_i)$.

Theorem 2. *Suppose that the regularity conditions* B1–B4 *hold. Then, under partial linear models* (3.2) *and the null hypothesis* H_0*, the results in Theorem* 1 *continue to hold.*

Note that existing tests for heteroscedasticity in partial linear models usually assume that the variance function $Var(Y|X,T)$ depends only on T. This condition is not necessary for our test. Under this condition, we can construct a much simpler test using the covariate T, rather than the projected covariate $\alpha^\top(X^\top,T)^\top$. Because $Var(Y|X,T)$ is a function of T, it follows that $Var(Y|X,T) = E(\varepsilon^2|T)$. Thus, the null hypothesis H_0 is equivalent to $E(\eta|T)=0$. The resulting test statistic is given as follows:

$$CM_n^{(1)} = \int_{\mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\eta}_i I(T_i \le t) \right|^2 dt.$$

More generally, if $T \in \mathbb{R}^d$ is a multiple random variable, we also encounter the dimension problem for large d. Then, we can use the projected covariates $\alpha^\top T$ to construct a test for heteroscedasticity. The test statistic becomes

$$CM_n^{(2)} = \int_{\mathbb{S}^d} \int_{\mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\eta}_i I(\alpha^\top T_i \le t) \right|^2 F_{n,\alpha}(dt)d\alpha,$$

where $F_{n,\alpha}$ is the empirical distribution function of the projected covariates $\{\alpha^\top T_i : i = 1,\dots,n\}$. The limiting distributions of $CM_n^{(1)}$ and $CM_n^{(2)}$ are similar to that of HCM_n, which we derive here.

Now, we investigate the sensitivity of the proposed test to alternative hypotheses. Consider a sequence of local alternatives converging to the null at a convergence rate c_n:

$$H_{1n} : E(\varepsilon^2|Z) = \sigma^2 + c_n s(Z), \tag{3.6}$$

where $s(Z)$ is not a constant function of Z, and $E[s(Z)]=0$ and $E[s^2(Z)]<\infty$. The following theorem shows that the proposed test is consistent against all global alternatives, and can detect local alternatives converging to the null at a parametric convergence rate $1/\sqrt{n}$.

Theorem 3. *Suppose the regularity conditions in Theorem* 1 *or Theorem* 2 *hold. Then,*

(1) under the alternatives H_{1n}, with $\sqrt{n}c_n \to \infty$, we have $HCM_n \to \infty$ in probability;

(2) under the alternatives H_{1n}, with $c_n = 1/\sqrt{n}$, we have

$$HCM_n \longrightarrow \int_{\mathbb{S}^p} \int_{\mathbb{R}} [V_\infty(\alpha,t) + S(\alpha,t)]^2 F_\alpha(dt)d\alpha \quad \text{in distribution,}$$

where $S(\alpha,t)=E\{s(Z)[I(\alpha^\top Z\le t)-F_\alpha(t)]\}$ *is a nonrandom shift term .*

The proofs of Theorems 1–3 are presented in the online Supplementary material. These theorems confirm the claims made in the introduction.

4. Numerical Studies

4.1. Simulation studies

In this subsection, we conduct several simulation studies to investigate the performance of our test. Because the test is not asymptotically distribution free, we suggest a residual-based bootstrap to approximate the distribution of the test statistic. This method was also used by Hsiao and Li (2001); Wang and Zhou (2007); Su and Ullah (2013); Guo et al. (2019). The procedure for the residual-based bootstrap is given as follows:

(1). For a given random sample $\{(Y_i,Z_i):i=1,\dots,n\}$, obtain the residual $\hat\varepsilon_i=Y_i-\hat m(Z_i)$, where $\hat m(\cdot)$ is the estimator of the regression function.

(2). Obtain the bootstrap error ε_i^* by randomly sampling, with replacement, from the center variables $\{\hat\varepsilon_i-\bar{\hat\varepsilon}:i=1,\dots,n\}$, where $\bar{\hat\varepsilon}=(1/n)\sum_{i=1}^n\hat\varepsilon_i$. Then define, $Y_i^*=\hat m(Z_i)+\varepsilon_i^*$.

(3). For the bootstrap sample $\{(Y_i^*,Z_i):i=1,\dots,n\}$, obtain the estimator $\hat m^*(Z_i)$, and then define the bootstrap residual $\hat\varepsilon_i^*=Y_i^*-\hat m^*(Z_i)$. Let $\hat\eta_i^*=\hat\varepsilon_i^{*2}-\hat\sigma_i^{*2}$ and $\hat\sigma_i^{*2}=(1/n)\sum_{i=1}^n\hat\varepsilon_i^{*2}$. Thus, the bootstrap test statistic HCM_n^* is calculated based on $\{(\hat\eta_i^*,Z_i):i=1,\dots,n\}$.

(4). Repeat steps (2) and (3) many times, say, B times. For a given significance level τ, the critical value is determined by the upper $\tau-$ quantile of the bootstrap distribution $\{HCM_{n,j}^*:j=1,\dots,B\}$ of the test statistic.

Note that $\hat m(Z_i)=m(Z_i,\hat\beta_n)$ for the parametric regression model (3.1), and $\hat m(Z_i)=\hat\beta_n^\top X_i+\hat g(T_i)$, with $Z_i=(X_i,T_i)$, for the partial linear regression model (3.2). The bootstrap estimators $\hat m^*(Z_i)$ are defined similarly.

The next theorem establishes the validity of the residual-based bootstrap.

Theorem 4. *Suppose the regularity conditions in Theorem* 1 *or Theorem* 2 *hold. Then,*

(1) under the null H_0 *and the local alternative* H_{1n}*, the distribution of* HCM_n^**, given* $\{(Y_i,Z_i):i=1,\dots,n\}$*, converges to the limiting null distribution of* HCM_n *in Theorem* 1*.*

(2) under the alternative H_1, the distribution of HCM_n^*, given $\{(Y_i, Z_i) : i = 1, \ldots, n\}$, converges to a finite limiting distribution.

Theorem 4 indicates that the bootstrap is asymptotically valid. Under the null hypothesis, the bootstrap distribution gives an asymptotical approximation to the limiting null distribution of HCM_n. Under the local alternatives H_{1n} and the global alternative H_1, the proposed test based on the bootstrap critical values remains consistent.

Next, we report several simulation results that evaluate the finite-sample performance of the proposed test. We also compare the performance of the proposed test with that of the tests of Zhu, Fujikoshi and Naito (2001) T_n^{ZFN}, Zheng (2009) T_n^{ZH}, and Guo et al. (2019) T_n^G under different settings of dimensions. Note that Guo et al. (2019) used the characteristic function to construct a test for heteroscedasticity, based on one-dimensional projections. Thus, their test is also less sensitive to the dimension of the covariates. Specifically, their test statistic is based on the fact that the null hypothesis H_0 is equivalent to $E[\eta \exp(\mathrm{i}t^\top Z)] = 0$, for all $t \in \mathbb{R}^p$. The test statistic of Guo et al. (2019) is given as follows:

$$T_n^G = \int_{\mathbb{R}^p} \left| \frac{1}{n} \sum_{j=1}^{n} \hat{\eta}_j \exp(\mathrm{i}t^\top Z_j) \right|^2 f_{\delta,p}(t)dt,$$

where $f_{\delta,p}(t)$ denotes the density of a spherical stable distribution in $\mathbb{R}^p$, with a characteristic exponent $\delta \in (0, 2]$. Note that

$$\int_{\mathbb{R}^p} \cos(t^\top z) f_{\delta,p}(z)dz = \exp(-\|t\|^\delta).$$

Thus, the test statistic of Guo et al. (2019) has a closed form:

$$T_n^G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{\eta}_i \hat{\eta}_j \exp(-\|Z_i - Z_j\|^\delta).$$

In the following simulations, $a = 0$ corresponds to the null, and $a \neq 0$ corresponds to the alternatives. The sample sizes are 100 and 200. The empirical size and powers are calculated using 1,000 replications at a nominal level 0.05. The bootstrap sample is set to $B = 500$. We choose $\delta = 1.5$ in T_n^G, as suggested by Guo et al. (2019).

Study 1. The data are generated from the following parametric regression models:

$$H_{11} : Y = \beta^\top Z + |a \times \beta^\top Z + 0.5| \times \varepsilon;$$
$$H_{12} : Y = \beta^\top Z + \exp(a \times \beta^\top Z) \times \varepsilon;$$
$$H_{13} : Y = \beta^\top Z + |a \times \sin(\beta^\top Z) + 1| \times \varepsilon;$$
$$H_{14} : Y = \exp(-\beta^\top Z) + |a \times \beta^\top Z + 0.5| \times \varepsilon;$$

where $Z \sim N(0, I_p)$, independent of the standard normal error ε, and $\beta = (1, \ldots, 1)^\top / \sqrt{p}$. To show the effect of the dimension, p is set to 2, 4, and 8 in each model. Note that model H_{13} is a high-frequency model, and the other three are low-frequency models. To determine whether the regression function affects the performance of the tests, we consider a nonlinear regression function in model H_{14}.

The simulation results for models H_{11} and H_{12} are presented in Table 1. The remaining results are relegated to the Supplementary material, for brevity. When $p = 2$, Zheng (2009)'s test T_n^{ZH} and the test of Guo et al. (2019) T_n^G do not maintain the significance level in some cases, although the other two perform better. In terms of the empirical power, the tests all work well. However, the proposed test HCM_n and the test of Zhu, Fujikoshi and Naito (2001) T_n^{ZFN} grow faster than the other two as a increases. When the dimension p becomes large, the tests HCM_n and T_n^{ZFN} still control the empirical size. In contrast, the empirical sizes of T_n^{ZH} and T_n^G are slightly away from the significance level. In terms of empirical power, the tests HCM_n and T_n^G outperform the other two. Here, T_n^{ZFN} performs worst when $p = 8$. These finding validate our theoretical results that the proposed test HCM_n is little affected by the dimension of the covariates, and that the tests T_n^{ZH} and T_n^{ZFN} suffer severely from the dimensionality problem. In the high-frequency model H_{13}, we observe that the locally smoothing test T_n^{ZH} performs much worse than the other tests do. This differs from the case of model checking, where locally smoothing tests usually outperform their globally smoothing counterparts in high-frequency models. Furthermore, we found no significant difference between the empirical size and power of the regression functions in models H_{11} and H_{14}.

In the next simulation study, we investigate the performance of the proposed test in partial linear regression models. We focus on two cases: (1) $Var(\varepsilon|X, T)$ is a function of (X, T), and (2) $Var(\varepsilon|X, T)$ is a function of T.

Study 2. The data are generated from the following models:

$$H_{21} : Y = \beta^\top X + T^2 + |a(\beta^\top X + T) + 0.5| \times \varepsilon;$$

Table 1. Empirical sizes and powers of HCM_n, T_n^G, T_n^{ZH}, and T_n^{ZFN} for H_{11} and H_{12} in Example 1.

	a	HCM_n		T_n^G		T_n^{ZH}		T_n^{ZFN}	
		n=100	n=200	n=100	n=200	n=100	n=200	n=100	n=200
$H_{11}, p=2$	0.0	0.045	0.051	0.058	0.062	0.042	0.033	0.052	0.049
	0.1	0.528	0.895	0.391	0.751	0.123	0.286	0.503	0.889
	0.2	0.966	1.000	0.921	1.000	0.468	0.889	0.961	1.000
	0.3	0.998	1.000	0.990	1.000	0.779	0.990	0.985	1.000
	0.4	0.998	1.000	0.999	1.000	0.885	0.998	0.974	1.000
	0.5	0.994	1.000	0.999	1.000	0.928	1.000	0.965	0.998
$H_{11}, p=4$	0.0	0.055	0.053	0.050	0.057	0.031	0.022	0.063	0.051
	0.1	0.398	0.767	0.233	0.481	0.049	0.095	0.131	0.593
	0.2	0.874	0.997	0.669	0.958	0.145	0.347	0.426	0.956
	0.3	0.963	1.000	0.857	0.999	0.306	0.621	0.541	0.964
	0.4	0.970	0.999	0.943	1.000	0.430	0.821	0.419	0.916
	0.5	0.944	0.998	0.958	1.000	0.492	0.876	0.297	0.809
$H_{11}, p=8$	0.0	0.049	0.049	0.053	0.065	0.045	0.036	0.050	0.049
	0.1	0.289	0.600	0.151	0.257	0.055	0.055	0.004	0.004
	0.2	0.755	0.980	0.352	0.688	0.108	0.132	0.004	0.010
	0.3	0.883	0.997	0.526	0.892	0.138	0.187	0.004	0.010
	0.4	0.874	0.990	0.623	0.946	0.167	0.254	0.009	0.009
	0.5	0.853	0.988	0.647	0.966	0.247	0.324	0.023	0.014
$H_{12}, p=2$	0.0	0.054	0.046	0.043	0.068	0.032	0.056	0.052	0.045
	0.1	0.183	0.347	0.138	0.262	0.059	0.080	0.153	0.327
	0.2	0.564	0.892	0.440	0.753	0.121	0.295	0.502	0.878
	0.3	0.882	0.996	0.747	0.967	0.281	0.692	0.810	0.993
	0.4	0.973	0.999	0.927	0.999	0.514	0.900	0.919	0.997
	0.5	0.987	0.999	0.983	1.000	0.650	0.964	0.944	0.986
$H_{12}, p=4$	0.0	0.050	0.046	0.058	0.048	0.028	0.023	0.057	0.056
	0.1	0.127	0.270	0.103	0.157	0.034	0.038	0.040	0.110
	0.2	0.424	0.789	0.264	0.479	0.048	0.075	0.104	0.529
	0.3	0.702	0.976	0.488	0.856	0.114	0.208	0.210	0.804
	0.4	0.862	0.993	0.727	0.976	0.163	0.436	0.294	0.857
	0.5	0.910	0.993	0.849	0.996	0.272	0.651	0.317	0.802
$H_{12}, p=8$	0.0	0.050	0.046	0.085	0.062	0.039	0.037	0.054	0.047
	0.1	0.112	0.193	0.083	0.111	0.055	0.053	0.014	0.001
	0.2	0.274	0.618	0.156	0.266	0.063	0.057	0.002	0.003
	0.3	0.549	0.919	0.252	0.526	0.089	0.086	0.002	0.000
	0.4	0.757	0.973	0.372	0.727	0.113	0.154	0.001	0.002
	0.5	0.836	0.972	0.494	0.865	0.140	0.207	0.001	0.002

$$H_{22} : Y = \beta^\top X + T^2 + \exp\{a(\beta^\top X + T)\} \times \varepsilon;$$
$$H_{23} : Y = \beta^\top X + T^2 + |a \sin(\beta^\top X + T) + 1| \times \varepsilon;$$

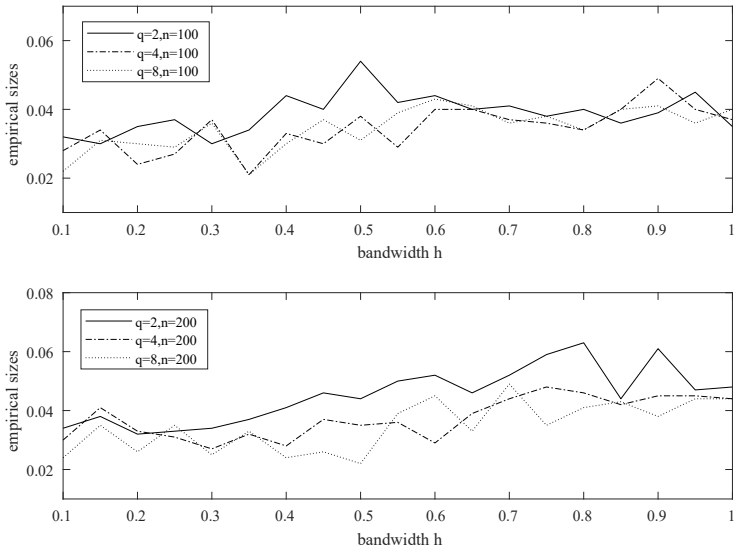

Figure 1. The empirical size curves of HCM_n against the different bandwidths and sample size 100 and 200 with $a = 0$ in Model H_{21}.

$$H_{24} : Y = \beta^\top X + \exp(T) + |a(\beta^\top X + T) + 0.5| \times \varepsilon;$$
$$H_{25} : Y = \beta^\top X + \exp(T) + |a \sin(\beta^\top X + T) + 1| \times \varepsilon;$$
$$H_{26} : Y = \beta^\top X + T^2 + \exp(4aT) \times \varepsilon;$$

where $X \sim N(0, I_q)$, $T \sim U(0, 1)$, $\varepsilon \sim N(0, 1)$, and $\beta = (1, \ldots, 1)^\top/\sqrt{q}$. The error term ε is independent of (X, T). The dimension q of the covariates X is again set to 2, 4, and 8. We use the kernel function $K(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$. A further issue is the selection of the bandwidth h. Several data-driven procedures are available for selecting the bandwidth automatically in estimation problems, (e.g., generalized cross validation; GCV). In hypothesis testing, how best to select a bandwidth remains an open problem. Note that the underlying regression models are different under the null and the alternatives. Eubank and Thomas (1993) stated that the GCV method works well when choosing the bandwidth for a homoscedastic model, but may not be useful for a heteroscedastic model. Thus, it is unknown whether a data-driven procedure exists for selecting the bandwidth in hypothesis testing. On the other hand, Theorems 2 and 3 show that the asymptotic property of the test statistic HCM_n does not rely on the choice of h when the regularity condition (B4) is satisfied. Thus, the proposed

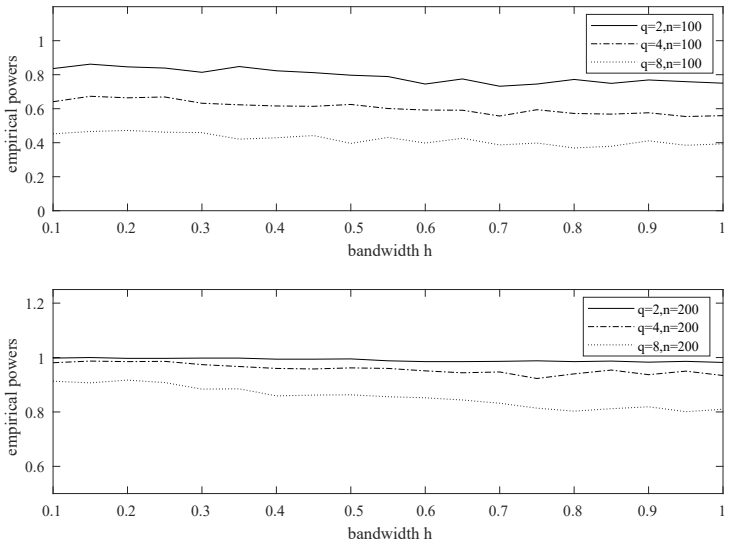

Figure 2. The empirical power curves of HCM_n against the different bandwidths and sample size 100 and 200 with $a = 0.2$ in Model H_{21}.

test is not overly sensitive to the choices of the smoothing parameter h. Thus, we consider a wide range of values of h, and empirically choose one as the bandwidth. This strategy was also adopted by Zhu, Fujikoshi and Naito (2001); Sun and Wang (2009), among many others. Let $h = j/100$, for $j = 10, 15, 20, \ldots, 100$. The empirical size and power of each dimension are presented in Figures 1 and 2.

From these two figures, we can see that when the bandwidth h is too small, HCM_n cannot maintain the significance level. However, when the bandwidth h is greater than 0.5, the test statistic HCM_n seems robust against different bandwidths. Thus, we use the bandwidth $h = 0.65$ in the following simulation studies.

The empirical size and power values are presented in the Supplementary Material. We observe that the results are similar to those of Study 1 for the first five models. The proposed test HCM_n still performs best. It seems the nonlinear part $g(\cdot)$ of a partial linear regression model does not affect the performance of the test. However, this changes in model H_{26}. When the dimension q of the covariate X is relatively large, the tests all perform very poorly, because when q is large, the weight of T that contributes to the test statistics becomes small.

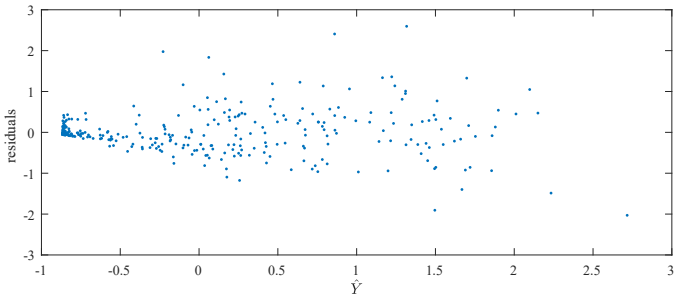

Figure 3. The scatter plot of the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{Y}_i$ for the baseball salary data set.

4.2. Real-data analysis

In this subsection, we analyze two data sets. The first one is a well-known baseball salary data set (available from the website `http://www4.stat.ncsu.edu/~boos/var.select/baseball.html`), with data on the salary Y and 16 performance measures for each of 337 Major League Baseball players for the 1991 and 1992 seasons. Further details about the variables in the data set are available from the above website. Recently, Tan and Zhu (2019) analyzed the data set, and suggested fitting the data set using the following parametric single-index model:

$$Y = a + b(\beta^\top X) + c(\beta^\top X)^2 + \varepsilon.$$

Here, we investigate whether heteroscedasticity exists in this model. We first plot the residuals $\hat{\varepsilon}$ against the fitted values $\hat{Y}$ in Figure 3, where $\hat{\varepsilon} = Y - \hat{a} - \hat{b}(\hat{\beta}_n^\top X) - \hat{c}(\hat{\beta}_n^\top X)^2$ and $\hat{Y} = \hat{a} + \hat{b}(\hat{\beta}_n^\top X) + \hat{c}(\hat{\beta}_n^\top X)^2$. This plot shows that heteroscedasticity may exist. When the proposed test is applied, the p-value is about zero. This indicates the existence of heteroscedasticity. Thus, a parametric single-index model with heteroscedasticity is plausible for the salary data set.

In the next example, we consider the ACTG315 data set, which was used by an AIDS clinical trial group study to identify the relationship between virologic and immunologic responses in AIDS clinical trials. The data set has been studied by Wu and Wu (2001, 2002); Yang, Xue and Cheng (2009). In general, the virologic response RNA (measured by viral load) and immunologic response (measured by CD cell counts) have a negative correlation during clinical trials. Let viral load be the response variable, and let CD4+cell counts and treatment time be the covariates. Liang et al. (2004) find that a linear relationship between viral load and CD4+ cell count, but a nonlinear relationship between viral

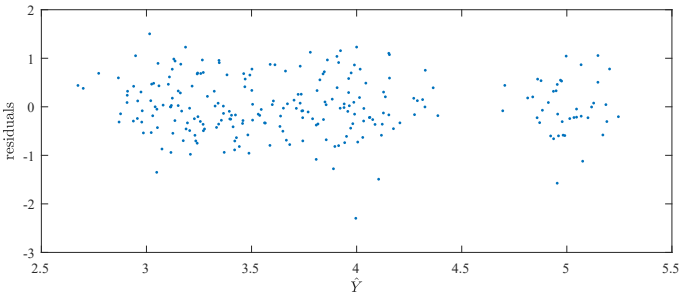

Figure 4. The scatter plot of the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{Y}_i$ for the ACTG 315 data set.

load and treatment time. Base on these findings, Yang, Xue and Cheng (2009) suggested a partial linear regression model to fit the data. Xu and Guo (2013) confirmed this model using a goodness of fit test. The data set contains 317 observations, with 64 CD4+ cell counts missing. To illustrate our test, we clear the observations with missing variables. Let Y be viral load, T be treatment time, and X be CD4+cell count. Yang, Xue and Cheng (2009) use the following model to fit the data:

$$Y = \beta X + g(T) + \varepsilon.$$

We use the proposed test to check for heteroscedasticity in the above models. When the normal kernel and the bandwidth $h = 0.65$ are used, the p-value is about 0.246. Thus, we cannot reject the homoscedasticity assumption in the partial linear regression model. The scatter plot of the residuals $\hat{\varepsilon}$ against the fitted values $\hat{Y}$ is presented in Figure 4, where $\hat{\varepsilon} = Y - \hat{\beta}_n X - \hat{g}(T)$ and $\hat{Y} = \hat{\beta}_n X + \hat{g}(T)$. This plot confirms that a partial linear model with homoscedasticity is appropriate for the data set.

5. Conclusion

We propose a test for heteroscedasticity that uses a projected empirical process. The proposed test can be viewed as a generalization of the test of Zhu, Fujikoshi and Naito (2001). When the dimension of the covariate is one, the proposed test reduces to that of Zhu, Fujikoshi and Naito (2001). Thus, the tests share several common desirable feathers: both are consistent for all global alternatives; the convergence rate does not relate to the dimension of the covariates; and they can detect local alternatives departing from the null at a parametric rate $1/\sqrt{n}$, which is the fastest convergence rate in hypothesis testing. Neverthe-

less, we use the projection of the covariates rather than the covariates themselves to construct the residual marked empirical process. Because the proposed test is based on one-dimensional projections, it performs as if the dimension of the covariates is one. Thus, our test can significantly alleviate the impact of the "curse of dimensionality." The simulation results validate these theoretical results. Furthermore, our method can easily be extended to a more generalized problem of testing the parametric form of a variance function. However, the limiting distributions of the empirical processes may have a more complicated structure, which may lead to the asymptotic test not being available. This is beyond the scope of this study, and is left to further research.

Supplementary Material

The online Supplementary Material contains proofs for the main results and additional simulation results.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (531118010318), National Natural Science Foundation of China (11871263,11601227, 11671042), NSF grant of Guangdong Province of China (No. 2017A030313012), Shenzhen Sci-Tech Fund No. JCYJ20170307110329106, and a grant from the University Grants Council of Hong Kong. All correspondence should be addressed to Xuejun Jiang, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzen, China; E-mail: jiangxj@sustech.edu.cn.

References

Anderson, N. H., Hall, P. and Titterington, D. M. (1994). Two-sample tests for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis* **50**, 41–54.

Chown, J. and Müller, U. U. (2018). Detecting heteroscedasticity in non-parametric regression using weighted empirical processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 951–974.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 693–708.

Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference* **103**, 311–329.

Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 903–917.

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

Eubank, R. L. and Thomas, W. (1993). Detecting heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **55**, 145–155.

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory* **14**, 604–621.

Guo, X., Jiang, X. J., Zhang, S. M. and Zhu, L. X. (2019). Pairwise distance-based heteroscedasticity test for regressions. *Science in China: Mathematics* to appear.

Hsiao, C. and Li, Q. (2001). A consistent test for conditional heteroskedasticity in time-series regression models. *Econometric Theory* **17**, 188–221.

Koul, H. L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*. 2nd Edition. Lecture Notes in Statistics, Springer-Verlag, New York.

Koul, L. H. and Song, W. X. (2010). Conditional variance model checking. *Journal of Statistical Planning and Inference* **140**, 1056–1072.

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

Lavergne, P. and Patilea, V. (2012). One for all and all for one: Regression checks with many regressors. *Journal of Business & Economic Statistics* **30**, 41–52.

Liang, H., Wang, S. J., Robins, J. and Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* **99**, 357–367.

Lin, J. G. and Qu, X. Y. (2012). A consistent test for heteroscedasticity in semi-parametric regression with nonparametric variance function based on the kernel method. *Statistics* **46**, 565–576.

Muller, H. G. and Zhao, P. L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *The Annals of Statistics* **23**, 946–967.

Nolan, D. and Pollard, D. (1987). U-process: Rates of convergence. *The Annals of Statistics* **12**, 780–799.

Pardo-Fernández, J. C. and Jiménez-Gamero, M. D. (2019). A model specification test for the variance function in nonparametric regression. *AStA Advances in Statistical Analysis* **103**, 387–410.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

Schick, A. (1996). Root-n consistent estimation in partly linear regression models. *Statistics & Probability Letters* **28**, 353–358.

Simonoff, J. S. and Tsai, C. L. (1994). Improved tests for nonconstant variance in regression based on the modified profile likelihood. *Journal of Applied Statistics* **43**, 357–370.

Su, L. and Ullah, A. (2013). A nonparametric goodness-of-fit-based test for conditional heteroskedasticity. *Econometric Theory*, **29**, 187–212.

Sun, Z. H. and Wang, Q. H. (2009). Checking the adequacy of a general linearmodel with responses missing at random. *Journal of Statistical Planning and Inference* **139**, 3588–3604.

Stute, W., Xu, W. L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high dimensional spaces. *Biometrika* **95**, 1–17.

Tan, F. L. and Zhu, L. X. (2019). Adaptive-to-model checking for regressions with diverging number of predictors. *The Annals of Statistics* **47**, 1960–1994.

Wang, L. and Zhou, X. H. (2007). Assessing the adequacy of variance function in heteroscedastic regression models. *Biometrics* **63**, 1218–1225.

Wu, H. and Wu, L. (2001). A multiple imputation method for missing covariates in nonlinear mixed-effect models, with application to HIV dynamics. *Statistics in Medicine* **20**, 1755–1769.

Wu, L. and Wu, H. (2002). Nonlinear mixed-effect models with missing time-dependent covariates, with application to HIV viral dynamics. *Journal of the Royal Statistical Society. Series C. (Applied Statistics)* **51**, 297–318.

Xu, W. L. and Guo, X. (2013). Checking the adequacy of partial linear models with missing covariates at random. *Annals of the Institute of Statistical Mathematics* **65**, 473–490

Yang, Y. P., Xue, L. G. and Cheng, W. H. (2009). Empirical likelihood for a partially linear model with covariate data missing at random. *Journal of Statistical Planning and Inference* **139**, 4143–4153.

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

Zheng, J. X. (2009). Testing heteroscedasticity in nonlinear and nonparametric regressions. *Canadian Journal of Statistics* **37**, 282–300.

Zhu, L. X., Fujikoshi, Y. and Naito, K. (2001). Heteroscedasticity test for regression models. *Science China Series A* **44**, 1237–1252.

Zhu, L. X. and Ng, K. W. (2003). Checking the adequacy of a partial linear model. *Statistica Sinica* **13**, 763–781.

Falong Tan

College of Finance and Statistics, Hunan University, Changsha, Hunan, China.

E-mail: falongtan@hnu.edu.cn

Xuejun Jiang

Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong Province, China.

E-mail: jiangxj@sustech.edu.cn

Xu Guo

School of Statistics, Beijing Normal University, Haidian District, Beijing, China.

E-mail: xustat12@bnu.edu.cn

Lixing Zhu

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

E-mail: lzhu@hkbu.edu.hk

(Received August 2018; accepted May 2019)