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 of China*

[3] *Beijing Normal University,* [4] *Hong Kong Baptist University*

Abstract: In this paper we propose a new test of heteroscedasticity for parametric regression and partial linear regression models in multidimensional spaces. When the dimension of covariates is large, or even moderate, existing tests of heteroscedasticity perform badly due to the "curse of dimensionality". To attack this problem, we construct a test of heteroscedasticity by using a projection-based empirical process. We study the asymptotic properties of the test statistic under the null hypothesis and alternative hypotheses. It is shown that the test can detect local alternatives departure from the null hypothesis at the fastest possible rate in hypothesis testing. As the limiting null distribution of the test statistic is not asymptotically distribution free, we propose a residual-based bootstrap. The validity of the bootstrap approximations is investigated. We present some simulation results to show the finite sample performances of the test. Two real data analyses are conducted for illustration.

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 common variance. Ignoring the presence of heteroscedasticity in regression models may result in inefficient inferences of the regression coefficients, or even inconsistent estimators of the variance function. Therefore, testing heteroscedasticity in regression models should be conducted when 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 $E(\varepsilon|Z) = 0$. Thus the null hypothesis in testing heteroscedasticity for the regression model (1.1) is that

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

while the alternative hypothesis is that H_0 is totally incorrect:

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

Testing heteroscedasticity for the regression model (1.1) has been studied by many authors 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) further proposed a modified score test of heteroscedasticity for linear models.

Muller and Zhao (1995) developed a data-based test for heteroscedasticity in a general semiparametric variance function model with fixed design. Dette and Munk (1998) proposed a consistent test for heteroscedasticity in a nonparametric regression setting based on a L^2-distance between the underlying variance function and the constant variance. Zhu, Fujikoshi and Naito (2001) developed a test of heteroscedasticity based on residual marked empirical processes. Built on the work of Zheng (1996) for checking the regression function, Dette (2002) and Zheng (2009) respectively proposed two 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 the idea of Stute, Xu and Zhu (2008), Chown and Müller (2018) proposed a test of heteroscedasticity based on a weighted residual empirical distribution function. Lin and Qu (2012) developed a test of heteroscedasticity for 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), and Pardo-Fernández and Jiménez-Gamero (2018) 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 in this paper, we first give a detailed comment on two representative tests: Zhu, Fujikoshi and Naito (2001)'s test and Zheng (2009)'s test. Let $E(\varepsilon^2) = \sigma^2$ and $\eta = \varepsilon^2 - \sigma^2$.

Then the null hypothesis H_0 is tantamount 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 a test statistic as follows:

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

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 would converge to 0 as n goes to infinity. As Zheng (2009) used nonparametric smooth estimators to construct the test statistic, it suffers severely from the "curse of dimensionality". More specifically, Zheng (2009)'s test can only detect the 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 and the power of Zheng (2009)'s test drops quickly.

Zhu, Fujikoshi and Naito (2001) used residual marked empirical processes to construct a test of 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 a residual marked empirical process as follows:

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

4

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. It is shown that the test of Zhu, Fujikoshi and Naito (2001) can detect the local alternatives converging to the null at the parametric rate $1/\sqrt{n}$ which is the fastest convergence rate in hypothesis testing. But 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 paper is to develop a test of heteroscedasticity in parametric regression models that would avoid the drawbacks of Zhu, Fujikoshi and Naito (2001) and Zheng (2009) and be suitable for the case in which the dimension of covariates is relatively large. We notice that Zhu, Fujikoshi and Naito's (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 covariates. Nevertheless, their test still suffers from the "curse of dimensionality" in practice. Note that their test statistic is based on the indictor function $I(Z_i \leq t)$ which is the product of p indictor functions. This means that the vector $(I(Z_1 \leq t), \cdots, I(Z_n \leq t))^\top$ would be very sparse for large p. Thus Zhu, Fujikoshi and Naito's (2001) test cannot avoid the dimension problem in practice. To overcome this problem, we suggest to use the projected covariates $\alpha^\top Z_i$, rather than Z_i, to construct a residual marked empirical process and then the resulting test statistic would

not involve the product of p indictor functions. Escanciano (2006) and Lavergne and Patilea (2008, 2012) also adopted this approach to construct goodness-of-fit tests for parametric regression models. As the test is based on one-dimensional projections, it behaves as if the dimension of covariates was one. Thus this method is less sensitive to the dimension p of the regressors than that in Zhu, Fujikoshi and Naito (2001). On the other hand, we also use residual marked empirical processes to construct the test statistic. Thus our test statistic avoids nonparametric estimation of $E(\eta|Z)$ as used in Zheng (2009) and can detect local alternatives converging to the null at the parametric rate $1/\sqrt{n}$. Besides, the new test is easy to compute, does not involve multidimensional numerical integrations, and presents an excellent power performance for large dimension in finite sample simulations, see Section 4.

We also use this method to check heteroscedasticity in partial linear regression models. When the dimension of covariate is large, nonparametric estimation is less accurate due to the "curse of dimensionality", and partial linear regression models provide a more flexible substitution if the researchers already know some of the covariates enter the regression model linearly. Thus 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 will show that the limiting distribution has the same for-

m as that in parametric regression models and the proposed test can also detect local alternatives converging to the null at a rate $1/\sqrt{n}$ under this semi-parametric setting.

A procedure similar to ours is that of Chown and Müller (2018), where they introduced a test of heteroscedasticity by using a weighted empirical process based on the indictor function $I(\hat{\varepsilon}_j \leq t)$ rather than $I(\alpha^\top Z_j \leq t)$. This procedure is first proposed by Stute, Xu and Zhu (2008) for checking parametric regression models in high dimension settings. However, Chown and Müller (2018)'s test is constructed only for location-scale models. That is, $Y = m(Z) + \sqrt{Var(Y|Z)}e$ with e being independent with Z. The independence between e and Z is then employed to construct suitable test statistics. Same to Chown and Mller (2018), Pardo-Fernandez and Jimenez-Gamero (2018) also relies on this restriction. Moreover, they only considered one dimensional covariate. While 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 Mller (2018) also relies on nonparametric estimations, regardless of the type of the regression functions. Then 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 by using a projection-based empirical process. In section 3 we study the asymptotic properties of the test statistic under the null and al-

ternative hypotheses in parametric regression and partial linear regression models, respectively. In section 4, a residual-based bootstrap method is proposed to approximate the null distribution of the test statistic, simulation results comparing the proposed test with some existing competitors in the literature are presented, and two real data sets are analyzed to illustrate the proposed method. Section 5 contains a discussion. Technical proofs are postponed in the Appendix.

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,

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

Therefore, the null hypothesis H_0 is tantamount to

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

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 a test statistic for checking heteroscedasticity of model (1.1) as

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

8

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

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

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

It is well known that multidimensional numerical integrations are extremely difficult to handle in practice. However, the following Lemma enables us to avoid the multidimensional integrations in numerical calculations and obtain an analytic expression of 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 non-zero vectors and $\mathbb{S}^p$ be the p-dimensional unit sphere. Then we have*

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

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

As noted by an anonymous referee, the integral in Lemma 1 can be considered as a kernel function. Then our test statistic has similar form as Zheng (2009)'s test. However, we should note that different from Zheng (2009), our test statistic can be viewed as an U-statistic with a fixed bandwidth instead of varying bandwidth. This makes a big difference. From the

9

theories of U-statistics, we know that U-statistics with a fixed bandwidth have a parametric convergence rate $1/\sqrt{n}$ which is faster than that with a varying bandwidth. This coincides with the theoretical results we derive here by using empirical processes. For the references of the U-statistic with a fixed bandwidth, see Anderson et al. (1994) and Fan (1998).

The proposed test works for all regression models. It avoids some deficiencies of Zhu, Fujikoshi and Naito (2001) and Zheng (2009), namely, the nonparametric estimation of $E(\eta|Z)$, multidimensional numerical integration and the low power performance when the dimension p is large. Note that the test statistic is based on the residuals $\hat{\varepsilon}_j = Y_j - \hat{m}(Z_j)$, i.e., 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 paper we only deal with parametric regression and partial linear regression models, as the test statistic only involves parametric estimations for parametric regression models and 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. Due to the sparsity of data in multidimensional spaces, the behavior of nonparametric estimations quickly deteriorates when the dimension of covariates increases. Then the resulting test still suffers from the "curse of dimensionality" for nonparametric regression models in practice. However, we should note that this is a common problem for all existing tests of heteroscedasticity for nonparametric

regression model since all tests require firstly to obtain an estimator for the unknown regression function. Therefore, how to deal with the dimension problem in testing heteroscedasticity for nonparametric regression models is still a challenging problem.

3. Asymptotic results

First we consider a 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,

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

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 some 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 neighbourhood of β. Set

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

11

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

Conditions (A1) and (A3) are commonly used in the heteroscedasticity testing literature, see, e.g., Zheng (2009). Condition (A2) is satisfied, e.g., for the ordinary least square estimator and its robust modifications see, e.g., 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 a covariance function

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

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 heteroscedasticity for 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 an smooth function. As 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 an U-process. With the help of the

12

theory of U-process in the literature, see, e.g. Nolan and Pollard (1987), we will 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 i.i.d. sample from the distribution of (X, T, Y). The resulting 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

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

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

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 conditions (B3) that will be specified below. To obtain the estimator of $g(\cdot)$, we notice that $g(T) = E(Y - \beta^\top X|T)$. Thus the kernel estimator of $g(T)$ have the following form:

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

13

The following regularity conditions are needed for deriving the asymptotic distribution of HCM_n in partial linear regression models. In the following, C always stands for a constant that may be different from place to place.

(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 that there exists an open neighborhood Θ_1 of 0 such that for all t and x,

$$|E(X|T = t + u) - E(X|T = u)| \leq C|u|, \quad \forall\, u \in \Theta_1;$$

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

$$|F(x|t + u) - F(x|t)| \leq C|u|, \quad \forall\, u \in \Theta_1.$$

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

(B3) The kernel function $K(\cdot)$ is bounded, continuous, symmetric about 0 and satisfies: (a) the support of $K(\cdot)$ is the interval $[-1, 1]$; (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 in deriving the asymptotic properties of the nonparametric estimators, see, e.g., Schick (1996) and 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*

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

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,

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

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

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.*

It is worth mentioning that existing tests of heteroscedasticity for partial linear models in the literature usually assumed that the variance function $Var(Y|X,T)$ only depends 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 $\alpha^\top(X^\top, T)^\top$. As $Var(Y|X,T)$ is a function of T, it follows that $Var(Y|X,T) = E(\varepsilon^2|T)$. Thus the null hypothesis H_0 is tantamount to $E(\eta|T) = 0$. The resulting

15

test statistic is given as follows,

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

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 of heteroscedasticity. The test statistic becomes

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

where $F_{n,\alpha}$ is the empirical distribution function of projected covariates $\{\alpha^\top T_i : i = 1, \cdots, n\}$. The limiting distributions of $CM_n^{(1)}$ and $CM_n^{(2)}$ are similar as that of HCM_n 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 with $E[s(Z)] = 0$ and $E[s^2(Z)] < \infty$. The following Theorem shows that the proposed test is consistent against all global alternatives and it can detect the local alternatives converging to the null at a parametric convergence rate $1/\sqrt{n}$.

Theorem 3. *Suppose that 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

16

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 $S(\alpha, t) = E\{s(Z)[I(\alpha^\top Z \leq t) - F_\alpha(t)]\}$ is a non-random shift term.

The proofs of Theorems 1-3 are presented in Appendix. These theorems confirm the claims that we 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. As the test is not asymptotically distribution free, we suggest a residual-based bootstrap to approximate the distribution of the test statistic. This method has been previously adopted by Hsiao and Li (2001), Wang and Zhou (2007), Su and Ullah (2013), and Guo et al. (2019). The procedure of the residual-based bootstrap is given as follows:

(1). For a given random sample $\{(Y_i, Z_i) : i = 1, \cdots, n\}$, obtain the residual $\hat{\varepsilon}_i = Y_i - \hat{m}(Z_i)$ with $\hat{m}(\cdot)$ being an estimator of the regression function.

(2). Obtain the bootstrap error ε_i^* by randomly sampling with replacement from the center variables $\{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}} : i = 1, \cdots, n\}$ 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 $\{(Y_i^*, Z_i) : i = 1, \cdots, n\}$, obtain the estimator $\hat{m}^*(Z_i)$ and then define the bootstrap residual $\hat{\varepsilon}_i^* = Y_i^* - \hat{m}^*(Z_i)$.

17

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 $\{(\hat{\eta}_i^*, Z_i) : i = 1, \cdots, n\}$.

(4). Repeat step (2) and (3) a large number of times, say, B times. For a given significant level τ, the critical value is determined by the upper τ quantile of the bootstrap distribution $\{HCM_{n,j}^* : j = 1, \cdots, B\}$ of the test statistic.

Note that $\hat{m}(Z_i) = m(Z_i, \hat{\beta}_n)$ for a 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 a partial linear regression model (3.2). The bootstrap estimators $\hat{m}^*(Z_i)$ are defined similarly.

The next theorem establishes the validity of the proceeding 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 $\{(Y_i, Z_i) : i = 1, \cdots, n\}$ converges to the limiting null distribution of HCM_n in Theorem 1.

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

Theorem 4 indicates that the proceeding bootstrap is asymptotically valid. Under the null hypothesis, the bootstrap distribution gives an asymptotically 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 is still consistent.

Next we report some simulation results to show the finite sample performances of the proposed test. We also make a comparison with Zhu, Fujikoshi and Naito (2001)'s test T_n^{ZFN}, Zheng (2009)'s test T_n^{ZH} and Guo et al. (2019)'s test T_n^G under different settings of dimensions. Note that Guo et al. (2019) used the characteristic function to construct a test of heteroscedasticity, which is also based on one-dimensional projections. Thus their test is also less sensitive to the dimension of covariates. More concretely, their test statistic is based on the fact that the null hypothesis H_0 is tantamount 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,

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

where $f_{\delta,p}(t)$ denotes the density of a spherical stable distribution in $\mathbb{R}^p$ with 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 Guo et al. (2019)'s test statistic has a closed form as follows,

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

In the following studies, $a = 0$ corresponds to the null and $a \neq 0$ to the alternatives. The sample sizes are 100 and 200. The empirical sizes and powers are calculated through 1000 replications at a nominal level 0.05.

The number of the bootstrap sample is set to be $B = 500$. 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:

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

where $Z \sim N(0, I_p)$, independent of the standard normal error ε and $\beta = (1, \cdots, 1)^\top / \sqrt{p}$. To show the impact of the dimension, p is set to be 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 see whether the regression function can affect the performance of the tests, we consider a nonlinear regression function in model H_{14}.

The simulation results for the models H_{11} and H_{12} are presented in Table 1. The rest simulation results are put in the Supplement for saving space. It can be observed that when $p = 2$, Zheng (2009)'s test T_n^{ZH} and Guo et al. (2019)'s test T_n^G cannot maintain the significance level for some cases, while the other two perform better. For the empirical power, all these tests work well. But the proposed test HCM_n and Zhu, Fujikoshi and Naito (2001)'s test 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} can still

control the empirical sizes. In contrast, the empirical sizes of T_n^{ZH} and T_n^G are slightly away from the significant level. For the empirical power, the tests HCM_n and T_n^G work much better than the other two and T_n^{ZFN} becomes the worst one as it almost has no empirical powers when $p = 8$. These phenomena validate the theoretical results that the proposed test HCM_n is little affected by the dimension of covariates and the tests T_n^{ZH} and T_n^{ZFN} suffer severely from the dimensionality. In the high frequency model H_{13}, we can observe that the locally smoothing test T_n^{ZH} performs much worse than the other tests. This is different from the case in model checking where locally smoothing tests usually perform better than globally smoothing tests in high frequency models. Further, we found no significant difference in empirical sizes and powers from different regression functions in models H_{11} and H_{14}.

<p align="center">Tables 1 are about here</p>

In the next simulation study we further investigate the performance of the proposed test in partial linear regression models. We focus on two different 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:

$$
\begin{aligned}
H_{21} : Y &= \beta^\top X + T^2 + |a(\beta^\top X + T) + 0.5| \times \varepsilon; \\
H_{22} : Y &= \beta^\top X + T^2 + \exp\{a(\beta^\top X + T)\} \times \varepsilon; \\
H_{23} : Y &= \beta^\top X + T^2 + |a\sin(\beta^\top X + T) + 1| \times \varepsilon; \\
H_{24} : Y &= \beta^\top X + \exp(T) + |a(\beta^\top X + T) + 0.5| \times \varepsilon; \\
H_{25} : Y &= \beta^\top X + \exp(T) + |a\sin(\beta^\top X + T) + 1| \times \varepsilon; \\
H_{26} : Y &= \beta^\top X + T^2 + \exp(4aT) \times \varepsilon;
\end{aligned}
$$

where $X \sim N(0, I_q)$, $T \sim U(0,1)$, $\varepsilon \sim N(0,1)$ and $\beta = (1, \cdots, 1)^\top / \sqrt{q}$. The error term ε is independent of (X, T). The dimension q of covariates X is also set to be 2, 4 and 8.

We use the kernel function $K(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$. Another issue is the selection of the bandwidth h. There are a number of data-driven procedures available to select the bandwidth automatically in estimation problems, such as the generalized cross validation (GCV). In hypothesis testing, how to select a bandwidth is still an open problem. Note that the underlying regression models are different under the null and alternatives. Eubank and Hart (1993) stated that the GCV method works well for choosing the bandwidth in homoscedastic models while it may not be useful with heteroscedastic models. Thus it is unknown whether there exists a data-driven procedure 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 it may be said that the proposed test is not very sensitive to the choices of the smoothing parameter h. Thus we consider several values of h in a considerable wide range and empirically choose one as the bandwidth. This strategy is also adopted by many authors, such as Zhu, Fujikoshi and Naito (2001) and Sun and Wang (2009). Let $h = j/100$ for $j = 10, 15, 20, \cdots, 100$. The empirical sizes and powers for different dimensions are presented in Figure 1 and 2.

<div align="center">Figures $1 - 2$ is about here</div>

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

The empirical sizes and powers are presented in the Supplement. We can observe that the results are similar to the case in Study 1 for the first five models. The proposed test HCM_n still performs the best. It seems the nonlinear part $g(\cdot)$ in partial linear regression models does not impact the performance of the test. The situation becomes different in model H_{26}. When the dimension q of the covariate X is relatively large, all tests perform very bad. This can be explained that when q is large, the weight of T contributed to the test statistics becomes small.

4.2. Real data analysis

In this subsection we analyze two data sets for illustrations. The first one is the well-known baseball salary data set that can be obtain through the website `http://www4.stat.ncsu.edu/~boos/var.select/baseball.html`. It contains 337 Major League Baseball players on the salary Y and 16 performance measures during both the 1991 and 1992 seasons. More descriptions of the variables in the salary data set can be found in the above website. Recently, Tan and Zhu (2018) analysed the data set and suggest to fit the data set by a parametric single-index model as following:

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

Here we further investigate whether there exists a heteroscedasticity structure in the present 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 the heteroscedasticity structure may exist. When the proposed test is applied, we found the p-value is about 0. This indicates the existence of heteroscedasticity. Thus a parametric single index model with heteroscedasticity is plausible for the salary data set.

<div align="center">Figures 3 is about here</div>

In the next example we consider the ACTG315 data set which is obtain from an AIDS clinical trial group study. This study tries to find the relationship between virologic and immunologic responses in AIDS clinical trials. The data set has been studied by Wu and Wu (2001, 2002) and

Yang, Xue and Cheng (2009). Generally speaking, the virologic response RNA (measured by viral load) and immunologic response (measured by CD cell counts) have a negative correlation during the clinical trials. Let viral load be the response variable and CD4+cell counts and treatment time be the covariates variables. Liang et al. (2004) find that there is a linear relationship between viral load and CD4+ cell count, but a nonlinear relationship between viral load and treatment time. Base on this, Yang, Xue and Cheng (2009) suggested a partial linear regression model to fit the data. Xu and Guo (2013) further confirmed this by using a goodness of fit test. There are totally 317 observations available in the data set 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 counts. Yang, Xue and Cheng (2009) uses the following model for data fitting:

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

We further use the proposed test to check the existence of 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 also shows that a partial linear model with homoscedasticity is appropriate for

the data set.

<div align="center">Figures 4 is about here</div>

5. Conclusions and discusses

In this paper we propose a test of heteroscedasticity by using a projected empirical process. The proposed test can be viewed as a generalization of Zhu, Fujikoshi and Naito (2001)'s test. When the dimension of covariate is one, the proposed test reduces to Zhu, Fujikoshi and Naito (2001)'s test. Thus they share some common desirable feathers: both of them are consistent for all global alternatives; the convergence rate does not relate to the dimension of covariates; they can detect local alternatives departing from the null at a parametric rate $1/\sqrt{n}$, that is the fastest convergence rate in hypothesis testing. Nevertheless, we use the projection of covariates rather than covariates themselves to construct the residual marked empirical process. As the proposed test is based on one-dimensional projections, it performs as if the dimension of covariates was one. Thus our test can significantly alleviate the impact of the "curse of dimensionality". The simulation results validate these theoretical results. Further, our method can be easily extended to a more generalized problem of testing the parametric form of the variance function. But the limiting distributions of the empirical processes may have a more complicated structure which may lead the asymptotic test not available. This is beyond the scope of this paper and deserves a further study.

Supplementary Material

The online Supplementary Material contains two parts with the proofs of the main results and the additional simulation results.

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**, 4154.

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, 80*, 951-974.

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

Dette, H., Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal of*

the Royal Statistical Society: Series B., **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., 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.* **69 (5)**, 903-917.

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

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory*, **14**, 604621.

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 ed. Lecture Notes in Statistics, vol. 166. Springer-Ver

Koul, L. H. and Song, W. X. (2010). Conditional variance model checking. *Journal of Statistical Planning and Inference*, **140(4)**, 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.

REFERENCES

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. (2018). A model specification test for the variance function in nonparametric regression, *AStA Advances in Statistical Analysis*, To appear.

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 andInference*, **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.

REFERENCES

Tan, F. L. and Zhu, L. X. (2018). Adaptive-to-model checking for regressions with diverging number of predictors. *The Annals of Statistics*, accpeted.

Tsai, C. L. (1986). Score test for the first-order autoregressive model with heteroscedasticity. *Biometrika*, **73**, 455-460.

Wang, L. and Zhou, X. H. (2007). Assessing the adequacy of variance function in heteroscedastic regression models. *Biometrics*, **63(4)**, 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 C*, **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., 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., Naito, K. (2001). Heteroscedasticity test for regression models. *Science*

REFERENCES

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.

College of Finance and Statistics, Hunan University, China

E-mail: falongtan@hnu.edu.cn

Department of Mathematics, Southern University of Science and Technology of China, China

jiangxj@sustech.edu.cn

School of Statistics, Beijing Normal University, China

E-mail: xustat12@bnu.edu.cn

Department of Mathematics, Hong Kong Baptist University, Hong Kong
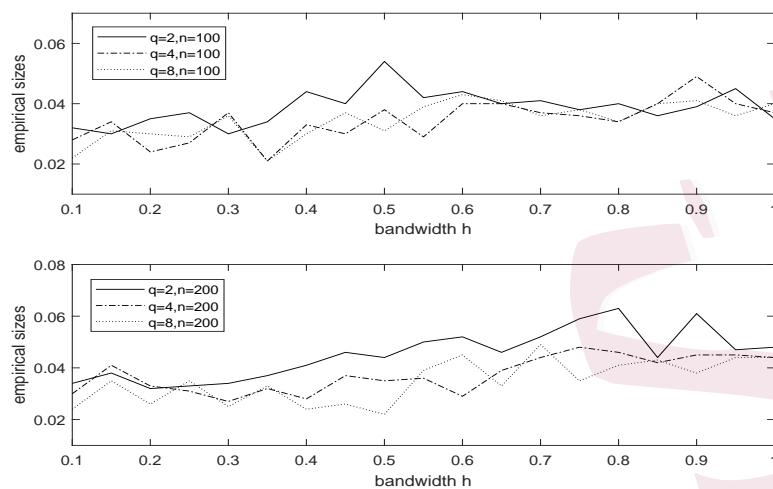
E-mail: lzhu@hkbu.edu.hk

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}.
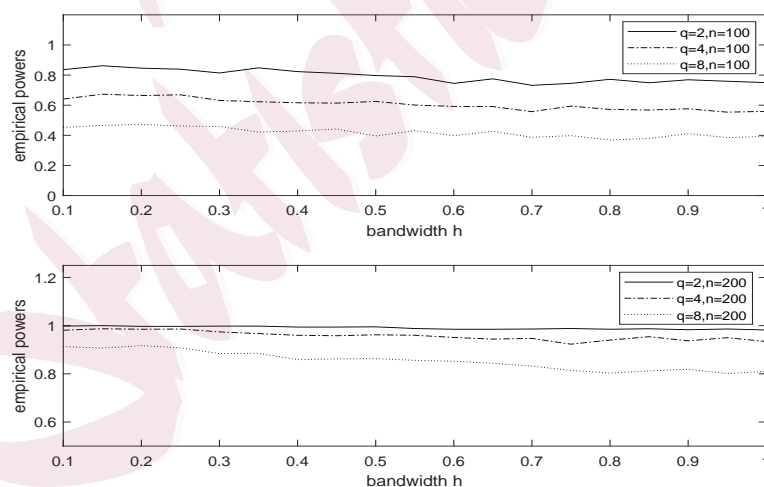

Figure 2: The empirical power curves of HCM_n against the different bandwidths and sample size 100 and 200 with $a = 0.2$ in Model H_{21}.

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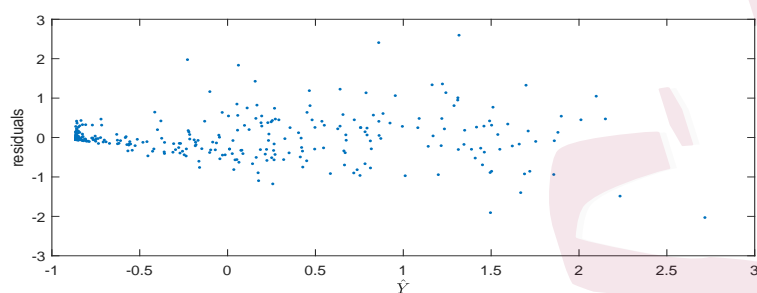
Figure 3: The scatter plot of the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{Y}_i$ for the baseball salary data set.
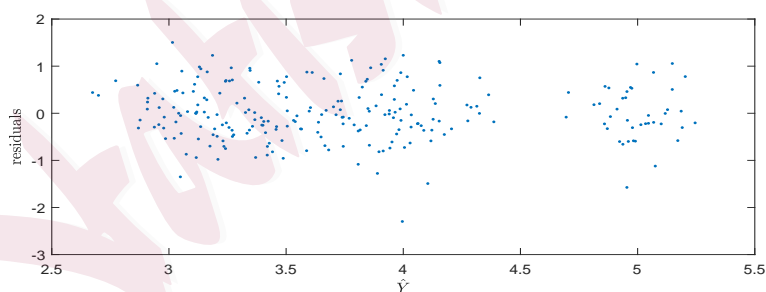

Figure 4: The scatter plot of the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{Y}_i$ for the ACTG 315 data set.