KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.986

"Rate Optimal Specification Test When the Number of Instruments is Large"

Kohtaro Hitomi, Masamune Iwasawa and Yoshihiko Nishiyama

March 2018

KYOTO UNIVERSITY

KYOTO, JAPAN

Rate Optimal Specification Test When the Number of Instruments is Large*

Kohtaro Hitomi† Masamune Iwasawa‡ Yoshihiko Nishiyama§

March 2, 2018

Abstract

We propose a rate optimal specification test for instrumental variable regression models based on the nearest neighbor observation with respect to instruments. The proposed test has uniform power against a set of non-smooth alternatives. The optimal minimax rate is $n^{-1/4}$ for any dimension of instruments, where n is sample size. This rate coincides with the fastest possible rate achievable by any tests under the local alternative setting when the alternative is constructed by a non-smooth function and/or the dimension of the instrument is large. Since such local alternative belongs to the set of alternatives considered in this study, our test is preferable in a large dimension setting. In the simulation and empirical applications with a large dimension of instruments, we observe that the test works well and the power approaches one reasonably fast as the sample size increase.

Keywords: instrumental variable model; specification test; minimax approach; k-nearest neighbor method

JEL Classification: C12; C14

1 Introduction

The instrumental variable (IV) model is one of the most important tools in a variety of fields of applied economics, and a large number of analytical techniques for it has been developed in econometrics. Another growing field is the literature on specification tests for IV regression models, which is based on and an extension of earlier development of tests for regression models. The earlier development seems to place much emphasis

*Acknowledgments: This work was supported by JSPS KAKENHI Grant Number 16J01227. We are grateful to Hidehiko Ichimura, Arthur Lewbel, Ryo Okui, Naoya Sueishi and the participants of the Asian meeting of the Econometric Society, the Australasia Meeting of the Econometric Society, and Statistische Woche for their useful comments.

†Kyoto Institute of Technology, hitomi@kit.ac.jp

‡Graduate School of Economics, The University of Tokyo and Research Fellow of Japan Society for the Promotion of Science, masamune.iwasawa@gmail.com,

§Institute of Economic Research, Kyoto University, nishiyama@kier.kyoto-u.ac.jp

on the power of tests, because it is an important statistical feature that is a good measure to compare properties of tests. Thus, it is natural that most existing tests are constructed so that they perform well in terms of power, especially, when the dimension of exogenous variables (instruments) is small. In economics, however, the number of instruments could be large because instruments often consist of cross-terms between exogenous variables in the model and instruments from outside of the model, or they include lags of exogenous variables. A well-known example is the return of education investigated in Angrist and Krueger (1991), in which quarter of birth and the cross-terms with all exogenous variables in the model give a total of 240 IVs. In a large dimension setting like this, asymptotic behavior, such as size and power performance of tests, has not been investigated sufficiently. Existing tests might not have good power performance, and the high computational burden can even make these tests inapplicable in praxis.

This study fills this gap in the literature by proposing a specification test that is powerful in the sense that it is rate optimal even when the dimension of instruments is large. In general, the fastest possible rate of a local alternative against which a test can show non-trivial power depends on smoothness of the alternative and dimension of instruments. By contrast, this study shows a set of alternatives against which no tests have non-trivial uniform power when it approaches the null at a rate, say $o(\rho_n)$, that is independent of the dimension of instruments. The rate is $\rho_n = n^{-1/4}$, where n is sample size. This is the fastest rate achievable by any tests under the local alternative setting when the alternative is constructed by a non-smooth function and/or the dimension of the instrument is large. The rate optimality of the proposed test indicates that the test has uniform power against the set of alternatives and is the most powerful under the large dimension setting.

The proposed test is constructed based on the nearest neighbor method. It can be interpreted as a K-nearest neighbor test with uniform K-nearest neighbor weights and fixed $K = 1$. The rate optimality of the simple fixed neighbor test may be surprising, because increasing the number of neighbors is likely to improve the power of the test. Intuitively, our test captures high frequency misspecification well when it appears at the nearest neighbor points. Since the set of alternatives we consider includes non-smooth ones, the simple nearest neighbor test performs quite well.

An important implication of our results is that power performance of K-nearest neighbor tests does not always improve as the number of neighbors K increases. This finding contradicts the results that a K-nearest neighbor test performs better against Pitman local alternatives when K grows with sample size. Indeed, Jun and Pinkse (2009) show that K-nearest neighbor tests detect such alternatives approaching the null at a rate of $(nK)^{-1/4}$, while our minimax result shows the set of functions against which the fastest achievable rate is $n^{-1/4}$. The existing results also indicate that the nearest neighbor tests for IV regression models can detect Pitman local alternatives converging to the null at a rate equal to $n^{-1/4}$. Thus, our results complement the existing result by figuring out the detectable set of alternatives while keeping uniform power.

In the minimax approach, the alternative hypothesis is a set of functions belong-

ing to a smoothness class. The set of alternatives is separated from the null model by L^2-distance but the distance approaches zero at a specific rate. The literature on optimal minimax rates is developed in to test the presence of a signal in the Gaussian white noise model. Ermakov (1991), Y. I. Ingster (1993), and Lepski and Tsybakov (2000) show the optimal minimax rates against alternatives within a Hölder class, while Spokoiny (1996) and Lepski and Spokoiny (1999) do so against alternatives within a Besov class. Y. I. Ingster and Sapatinas (2009) extend these results to test a multivariate non-parametric regression model with Gaussian noise against alternatives within an ellipsoid in the Hilbert space with respect to the tensor product Fourier basis. Another line of literature, such as Abramovich, Feis, Italia, and Theofanis (2009), shows the optimal minimax rate of testing for the additivity assumption of a response function against alternatives within a Besov class.

In econometrics, optimal minimax rates are established in testing for regression function. Guerre and Lavergne (2002) provide the optimal minimax rates of specification testing for a non-linear parametric regression model against alternatives within a Hölder class. For now, let us denote the dimension of regressor by l and smoothness index of a Hölder class by s. Then, the optimal minimax rates of Guerre and Lavergne (2002) are $n^{-2s/(l+4s)}$ if $s > l/4$ and $\tilde{\rho}_n = n^{-1/4}$ if $s \leq l/4$. Using this result, Horowitz and Spokoiny (2001) show that their test for parametric regression models achieves the optimal rate in their adaptive framework when $s \geq \max(2, l/4)$. Another rate optimal test is K-nearest neighbor tests proposed by H. Li, Li, and Liu (2016) when $s > l/4$. Their specification test for regression function without endogeneity is based on the K-nearest neighbor estimator for the residual of the regression function, and is the most related to our test in its construction. Their test is no longer rate optimal when the number of neighbors K is fixed. This study extends the literature by considering the different set of alternatives and expanding the model to linear and non-linear IV regression models.[1] Unlike existing works, this study analyzes optimal minimax rates against alternatives within a cone set of a Hölder class with some normalization. To the best of our knowledge, this is the first study that investigates the optimal minimax rate against the set of alternatives defined in this class.

Technically, deriving optimal minimax rates against alternatives within a cone set of a Hölder class is not obvious. To find the lower bounds for optimal minimax rates, it is necessary to find an example of a function that is difficult to detect (even by optimal Bayesian test) but still belongs to the set of alternatives. Guerre and Lavergne (2002) construct an example of a function within their Hölder class by using an idea from a Fourier series. However, this function does not belong to our alternatives. Since orthogonality seems to play an important role in constructing an example of such functions within our alternatives, we make use of wavelets instead of a Fourier series. Since wavelet methods are known to have good properties and are a useful technical tool in the

[1] Allowing models to be linear and non-linear generates some difficulties by producing asymptotic results compared to non-linear models only. This is because boundedness of derivatives of the parametric function is incompatible with allowing models to be linear. Without boundedness of derivatives, some bothersome steps are required to show the asymptotic features of testing, which requires calculating the upper or lower bounds of some statistics that include derivatives.

general Besov function classes (see, e.g., Daubechies, 1988, Meyer, 1992, and Y. Ingster & Suslina, 2003), rate optimal specification testing against alternatives within a Besov class also use the wavelet techniques (see, Spokoiny, 1996 and Lepski & Spokoiny, 1999).

The rest of the paper proceeds as follows. Section 2 reviews specification tests for IV models. Section 3 describes our testing framework and test statistic. Section 4 derives asymptotic normality of our test statistic under the null hypothesis. The main results, including the optimal minimax rate of the specification test for IV models and rate optimality, are shown in Section 5. Section 6 presents the Monte Carlo experiments. Section 7 applies our test to the return to education of Angrist and Krueger (1991) and Engel curve specifications. Section 8 concludes. The proofs of the primary results are in the Appendix.

2 Literature Review

The review of literature is focused on specification tests for IV models. Such tests were first developed by Donald, Imbens, and Newey (2003) and Tripathi and Kitamura (2003) as general tests of conditional moment restrictions in which conditional mean regressions are a special case. For example, Tripathi and Kitamura (2003) propose a smoothed empirical likelihood ratio-based test that detects Pitman linear local alternatives that converge to the null at rates slower than the parametric rate. Holzmann (2008) extends the specification test for the conditional mean regression proposed by Aït-Sahalia, Bickel, and Stoker (2001) to the IV setting. His test is based on the weighted distance between the smoothed parametric estimator and non-parametric kernel estimator of the IV model. This test can detect the Pitman linear local alternatives approaching the null at rate $n^{-1/2}h^{-l/4}$, where h is bandwidth and l is the dimension of the conditioning variables. Horowitz's (2006) test takes a form resembling the integrated conditional moment test developed by Bierens (1982, 1990), and Bierens and Ploberger (1997) and includes the non-parametric kernel density estimator as weighting. His test detects Pitman linear local alternatives approaching the null at the rate $n^{-1/2}$, making it the best specification test for IV models in terms of power performance. Gørgens and Würtz (2012) propose a test based on a sequence of Lagrange multiplier (LM) statistics. Because the LM statistic requires no estimation under the alternative, their test statistic involves no non-parametric estimation. However, the number of LM statistics to be calculated extends to infinity as the sample size increases. Our proposed test entails no kernel smoothing and no integration while maintaining consistency against all departures from the null hypothesis. Accordingly, it is easily implemented, requires less calculation time, and has no bandwidth choices. The test statistic can be obtained via straightforward calculation that requires only parametric model estimation under the null hypothesis and locating the neighbors nearest to each observation.

3 Framework and Test Statistic

Let $\{Y_i, X_i\}_{i=1}^n$ be a random sample from $(Y, X) \in \mathbb{R} \times \mathbb{R}^{l_x}$. The model is as follows:

$$Y_i = g(X_i, \theta) + u_i, \tag{1}$$

where $g(X_i, \theta)$ is a known function up to parameters $\theta \in \Theta$, Θ is a compact subset of $\mathbb{R}^{l_x}$, and u_i is an error term. Endogeneity of regressors often appears in many economic models, which complicates the estimation of parameters by violating the moment condition, which is essential for basic econometric estimation strategies, such as ordinary least squares methods.

In this study, we denote $E(Y|Z)$ as IV regression, where $Z_i \in \mathbb{R}^l$ is a vector of a random sample that consists of exogenous variables in X and exogenous variables from outside the model, so that conditional moment restriction $E(u_i|Z_i) = 0$ holds. The conditional moment restriction is equivalent to saying that a value of $\theta \in \Theta$ exists such that $E(Y_i|Z_i) = E[g(X_i, \theta)|Z_i]$ for all i. Therefore, we refer to $E[g(X_i, \theta)|Z_i]$ as the IV regression model. The IV regression model or conditional moment restrictions are often derived by economic theory. However, since economic models often aim to capture a simple and specific aspect of complex economic events, conditional moment restrictions derived by economic theory may not be satisfied in reality. Under this circumstance of model misspecification, estimation results obtained by using observed data may induce misleading interpretation.

Thus, we test the specification of the IV regression model. The null and alternative hypotheses are

$$\begin{aligned} H_0 &: E(u_i|Z_i) = 0, \\ H_1 &: E(u_i|Z_i) = h(Z_i) \neq 0, \end{aligned}$$

respectively, where $h(\cdot)$ is an unknown function. Since the null hypothesis implies that the parametric specification of the IV regression is correct, testing the null directly tests the fit of the parametric model to the true IV regression. The testing framework we consider can be interpreted as an extension of the specification test for the regression model by allowing the conditioning variable to contain variables from outside the model, which is not considered in the regression framework. Note that we aim to test the specification of IV regression rather than the functional form of $g(\cdot, \theta)$.

Our test statistic makes use of the feature of the nearest neighbor observations. Let the subscript i^* denote the nearest neighbor observation of i in the sense that observation i^* satisfies

$$||Z_i - Z_{i^*}|| \leq ||Z_i - Z_j|| \text{ for all } j \neq i, \tag{2}$$

where $||\cdot||$ is the Euclidean norm. Then, Y_{i^*} and X_{i^*} are the concomitant statistic to Z_{i^*}, that is, Y_{i^*} and X_{i^*} are the observations of individual i^*, which satisfies (2). We also define $u_{i^*} = Y_{i^*} - g(X_{i^*}, \theta)$. To propose the test statistic, we first assume that a nearest neighbor is uniquely assigned to each observation. However, the unique assignment assumption is not essential to construct the test. When ties exist, we can slightly modify

the test statistic, and the modified test has the same asymptotic properties with the test without ties. How to modify the test in the case of ties is discussed in Remark 1.

Let $\hat{\theta}$ be any $\sqrt{n}$-consistent estimator of θ under the null hypothesis, such as generalized method of moments (GMM) estimator of θ.[2] Let $\hat{u}_i \equiv Y_i - g(X_i, \hat{\theta})$ be the parametric estimator for u_i. Then, our test statistic is

$$T_n = \frac{1}{\hat{\mu}\sqrt{n}} \sum_{i=1}^{n} \hat{u}_i \hat{u}_{i^*}, \tag{3}$$

where $\hat{\mu}^2 \equiv n^{-1} \sum_{i=2}^{n} \sum_{j<i} W_{i,j}^2 \hat{u}_i^2 \hat{u}_j^2$ appears for the standardization. The weighting term $W_{i,j}$ is defined as $W_{i,j} = K_{i,j} + K_{j,i}$, where $K_{i,j} \equiv \mathbb{1}(||Z_i - Z_j|| \leq ||Z_i - Z_{i^*}||)$ for $i \neq j$ and $K_{i,j} = 0$ for $i = j$. A motivation of the test comes from the difference between non-parametric and parametric variance estimators with bias correction. To observe this, we rewrite the test statistic as follows:

$$\frac{\hat{\mu}}{\sqrt{n}} T_n = \hat{\sigma}_p^2 - \hat{\sigma}_d^2 + \frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i + \hat{u}_{i^*})[g(X_i, \hat{\theta}) - g(X_{i^*}, \hat{\theta})] + \frac{1}{2n} \sum_{i=1}^{n} [g(X_i, \hat{\theta}) - g(X_{i^*}, \hat{\theta})]^2,$$

where $\hat{\sigma}_p^2 \equiv \frac{1}{2n} \sum_{i=1}^{n} (\hat{u}_i^2 + \hat{u}_{i^*}^2)$ is the estimator for the variance of error term in the parametric model (1) under the null hypothesis and $\hat{\sigma}_d^2 \equiv \frac{1}{2n} \sum_{i=1}^{n} (Y_i - Y_{i^*})^2$ is the non-parametric difference-based estimator for the variance of $Y - E(Y|Z)$, denoted as σ^2. The final two terms added to the variance estimators are bias-correction terms for the non-parametric difference-based estimator.

$$\begin{aligned}
\hat{\sigma}_d^2 - \sigma^2 &= \frac{1}{2n} \sum_{i=1}^{n} (Y_i - Y_{i^*})^2 - \sigma^2 \\
&= \left[\frac{1}{2n} \sum_{i=1}^{n} (u_i - u_{i^*})^2 - \sigma^2 \right] + \frac{1}{n} \sum_{i=1}^{n} (u_i - u_{i^*})[g(X_i, \theta) - g(X_{i^*}, \theta)] \\
&\quad + \frac{1}{2n} \sum_{i=1}^{n} [g(X_i, \theta) - g(X_{i^*}, \theta)]^2,
\end{aligned}$$

where $\frac{1}{2n} \sum_{i=1}^{n} (u_i - u_{i^*})^2$ can be shown to converge to the true variance σ^2 when the model is true.[3] Therefore, our test measures the difference between the variance estimator under the null hypothesis from its true value that is estimated non-parametrically with bias correction.

To obtain the test statistic, researchers need only estimate the parameters under the null hypothesis and assign nearest neighbor observations for each observation according

[2] A detailed discussion on this point is given in Sections 4 and 5 under the null and alternative hypotheses, respectively.

[3] See, for example, the proof of Theorem 1 of Yatchew (1988). The difference-based estimator is first provided by Von Neumann, Kent, Bellinson, and Hart (1941) and developed by Gasser, Sroka, and Jennen-Steinmetz (1986), P. Hall, Kay, and Titterinton (1990), and Munk, Bissantz, Wagner, and Freitag (2005), among others.

to the assignment rule (2). In contrast to existing specifications testing for IV regression models, such as the integrated conditional moment-type test of Horowitz (2006) and the kernel-based smoothing type test of Holzmann (2008), our test does not include non-parametric estimation of, for example, density and regression functions. Thus, our test is simple, easily implementable, requires less calculation time than alternative tests do, and needs no bandwidth choices, as non-parametric kernel estimation does.

Our test resembles Yatchew's (1988) specification test for regression models without endogeneity. Like us, he shows consistency of the non-parametric difference-based variance estimator and provides a specification test that consists of the difference between the non-parametric (difference-based) and parametric variance estimator. He also shows conditions of ordering rules for choosing neighbors in which ordering rule (2) can be replaced while keeping the asymptotic power results of testing. The primary difference between the test provided in this study and Yatchew's (1988) is that our test can be applied to IV models that allow conditioning variables from outside the model, which standard regression setting does not allow. In addition, there are some notable differences that distinguish our test from his. First, we use nearest observations with respect to instruments, not covariates. Applying an ordering rule to covariates X_i leads the difference-based estimator to be consistent because, the two bias terms vanish as the sample size increases (by assuming $g(x,\cdot)$ is Lipschitz continuous with respect to x). Not surprisingly, however, the rate of convergence depends on the dimension of covariates, which makes $\hat{\sigma}_d^2$ non-$\sqrt{n}$-consistent when the dimension is greater than 3. Since our test does not include the bias terms, no trade-off exists between the rate of convergence and the dimension. Second, we do not split the sample in two. To establish the joint distribution of the non-parametric and parametric variance estimators, it is necessary to estimate their covariance, which may be difficult. To circumvent this problem, Yatchew (1988) divide the sample in two to make the non-parametric and parametric variance estimators independent. However, doing so reduces the power of the test. We do not divide the sample because the standardization term $\hat{\mu}^2$ in our test statistic can be obtained easily.

Remark 1. When ties exist, we can modify the test as follows

$$T_n^{tie} = \frac{1}{\hat{\mu}^{tie}\sqrt{\sum_{i=1}^n \sum_{j\neq i} K_{i,j}}} \sum_{i=1}^n \hat{u}_i \hat{u}_{i^*}, \tag{4}$$

where $(\hat{\mu}^{tie})^2 \equiv (\sum_{i=1}^n \sum_{j\neq i} K_{i,j})^{-1} \sum_{i=2}^n \sum_{j<i} W_{i,j}^2 \hat{u}_i^2 \hat{u}_j^2$. The modified test (4) is proportional to the original test (3), that is, $T_n^{tie} = O(1)T_n$, because $\sum_{i=1}^n \sum_{j\neq i} K_{i,j} = O(n)$. Thus, the asymptotic properties such as size and power of the modified test are the same with those of T_n.

Remark 2. The proposed test can be interpreted as the sample analogue of $E[u_i E(u_i|Z_i)]$ with u_i replaced by $\hat{u}_i$ and conditional expectation replaced by its K-nearest neighbor estimates with uniform K-nearest neighbor weights and fixed $k = 1$. Thus, it is possible to consider a K-nearest neighbor version of testing instead of using just the nearest

neighbor. Indeed, H. Li et al. (2016) provide a K-nearest neighbor specification test for the regression function and show that the proposed test is rate optimal. Interestingly, however, we show later that tests using just a single neighbor, such as ours, are also rate optimal against alternatives within a certain smoothness functional class. In other words, tests that utilize K-nearest neighbors with an appropriate K do not improve power performance against such alternatives.

Remark 3. The proposed test investigates the parametric specification of IV regression model $E[g(X,\theta)|Z]$. In special cases, however, the null hypothesis implies correct specification of $g(X,\theta)$. For example, under the completeness restriction, our test becomes a direct test for the functional form of $g(\cdot)$. The completeness restriction is satisfied, for example, when the conditional distribution of X given Z belongs to exponential families (see Newey & Powell, 2003). For other sufficient conditions, see, for example Hu and Shiu (2017). The correct specification of $g(\cdot)$ may be implied under some shape restrictions on it. For example, under the monotonicity restriction, Theorem 1 in Chetverikov and Wilhelm (2017) implies that testing the null may imply testing the specification of $g(\cdot)$ to some extent. Chetverikov and Wilhelm (2017) argues that monotonicity conditions are plausible in the estimation of Engel curves for normal goods. We test specifications of Engel curves in Section 7.

4 Asymptotic Distribution under the null hypothesis

Let $m(Z_i) \equiv E(Y_i|Z_i)$ denote the true IV regression, and $\omega_i \equiv Y_i - m(Z_i)$ its error. Variance of ω_i conditioned on Z_i is denoted by $\sigma^2(z) \equiv E(\omega_i^2|Z_i = z)$. Under the null hypothesis, conditional variance of parametric error $E(u_i^2|Z_i)$ is equivalent to $\sigma^2(z)$. Now, we list all the regularity conditions that are used in the derivation of the asymptotic distribution of our test statistic under the null hypothesis.

Assumption 1. *$\{Y_i, X_i, Z_i\}_{i=1}^n$ are a random sample on $(Y,X,Z) \in \mathbb{R}\times\mathbb{R}^{l_x}\times\mathbb{R}^l$, where l_x and l are finite. For all i, $M<\infty$ exists such that $E\left(|\omega_i|^p|Z_i\right) < M$, for $p=8$.*

Assumption 2. *For all x, $g(x,\theta)$ is twice continuously differentiable with respect to $\theta\in\Theta$, where Θ is a compact subset of $\mathbb{R}^{l_x}$.*

Assumption 3. $E[\sup_{\theta\in\Theta}\|\frac{\partial}{\partial\theta}g(X_i,\theta)\|^2] < \infty$.

Assumption 4. $E[\sup_{\theta\in\Theta}\|\frac{\partial}{\partial\theta\partial\theta'}g(X_i,\theta)\|^2] < \infty$.

Assumption 5. *For each z and $\theta\in\Theta$, $E\{[\frac{\partial}{\partial\theta}g(X,\theta)]^2|Z\}$ is bounded from above.*

Assumption 6. *Under the null hypothesis, we have an $\sqrt{n}$-consistent estimator $\hat{\theta}_n \equiv \hat{\theta}$ of θ_0, where θ_0 satisfies $E(Y_i|Z_i) = E[g(X_i,\theta_0)|Z_i]$.*

The higher moment restrictions in Assumption 1 are required to guarantee the consistency of $\hat{\mu}$ with respect to the asymptotic variance of the non-standardized test statistic, which is shown in Lemma 3 in the Appendix. This corresponds to, for example, the

finite fourth moments condition for the estimation of asymptotic variance of GMM estimators. In our setting, it is also possible to reduce the finite eighth moment of error term in Assumption 1 to its finite fourth moment. However, to keep the consistency of $\hat{\mu}$, we need a stronger dominance condition, that is, $E[\sup_{\theta\in\Theta} \|\frac{\partial}{\partial\theta} g(X_i,\theta)\|^4] < \infty$ instead of Assumption 3. A detailed discussion is given in the proof of Proposition 1 in the Appendix.

Assumptions 2–5 are all imposed on the family of parametric function $g(x,\theta)$, $\theta \in \Theta$. Dominance conditions 3 and 4 together with Assumption 2 guarantee uniform convergence of $\frac{1}{n}\sum_{i=1}^n \left\|\frac{\partial}{\partial\theta} g(X_i,\theta)\right\|^2$ and $\frac{1}{n}\sum_{i=1}^n \left\|\frac{\partial}{\partial\theta\partial\theta'} g(X_i,\theta)\right\|^2$. The dominance condition for the first derivative is a standard assumption that is also required, for example, for the asymptotic normality of GMM estimators. In addition, dominance conditions 3 and 4 imply boundedness of expectations of the first and second derivatives of $g(\cdot)$. These assumptions differ from Guerre and Lavergne (2002), who assume boundedness of the first and second derivatives of parametric functions and focus on testing the specification of non-linear parametric regression models.

Assumption 6 requires $\sqrt{n}$-consistent estimators. Under the null hypothesis, such estimators can be obtained by, for example, local Cressie–Read minimum distance estimators proposed by Smith (2007), which include smoothed empirical log-likelihood estimators (Kitamura, Tripathi, & Ahn, 2004), local exponential tilting (ET) estimators (see, Kitamura & Stutzer, 1997; Imbens, Spady, & Johnson, 1998 for ET estimators), and local continuous updating estimators. Estimators that use the continuum of moment restrictions, such as estimators proposed by Dominguez and Lobato (2004) and Carrasco and Florens (2000), are also $\sqrt{n}$-consistent.

$\sqrt{n}$-consistent estimators using unconditional moment restrictions, such as GMM, two-stage least squares (2SLS), and IV methods, can be also employed, as long as no identification issues arise. As Dominguez and Lobato (2004) show, estimation procedures based on unconditional moment restrictions may result in inconsistency, especially for non-linear models, when a finite number of unconditional moment restrictions is selected from infinite restrictions implied by conditional moment restrictions. Thus, in this case, we should be aware that parameter identification is guaranteed.

The following proposition shows that our test statistic converges to the standard normal distribution under the null hypothesis.[4]

Proposition 1. *Suppose Assumptions 1, 2, 3, 4, 5, and 6 hold. Then, under the null hypothesis,*

$$T_n \xrightarrow{d} N(0,1).$$

The test is asymptotically one-sided because departures from the null appear with positive values when the distance between Z_i and its nearest neighbor Z_{i^*} approaches 0 as the sample size grows.[5] To observe this, recall that the test detects the misspecification

[4]Not that the asymptotic normality of our test in Proposition 1 does not rely on the continuity of instruments.

[5]Sufficient conditions are given in Assumption 7 about the density of instruments and the subsequent discussion about the assumption.

when $E[h(Z_i)h(Z_{i^*})] \neq 0$. Roughly, $E[h(Z_i)h(Z_{i^*})] \approx E[h(Z_i)^2]$ when Z_i lies close to Z_{i^*} and $h(\cdot)$ is continuous. Thus, the departure from the null is asymptotically positive and the test is asymptotically one-sided, implying that we reject the null when the test statistic lies above the $(1-\alpha)$ quantile of the normal distribution, where α is a significance level.

Remark 4. Proposition 1 is an extension of Theorem 3.1 of H. Li et al. (2016), which shows asymptotic normality of the K-nearest neighbor test for the regression function, to the case of IV regression testing. Since H. Li et al. (2016) consider testing $E(u_i|X_i) = 0$, neighbors in H. Li et al.'s (2016) test are chosen by using covariates X_i. By contrast, neighbors in our test are chosen by using instruments Z_i, which include variables not only from the model but also from the outside of the model, which makes the derivation of asymptotic distribution somewhat involved. For example, to show the asymptotic normality, H. Li et al. (2016) directly employ Lemma B6 of Jun and Pinkse (2012). In our setting, however, the lemma is applicable only partially without imposing additional restrictions on parametric function.[6] Thus, we require some rearrangements to show the asymptotic normality.

Remark 5. Using the second or more distant nearest neighbor observations instead of the first nearest does not change the limiting behavior of the test statistic. The asymptotic distribution of the test statistic under the null hypothesis is derived independently of how far the neighboring observations are located. This can be intuitively understood when we consider the basic idea of the test, namely, $E(u_i u_{i^*}) = E[E(u_i|Z_i)E(u_j|Z_j)] = 0$ under the null hypothesis. Since this equation holds for any observation instead of i^* except $i^* = i$, it can be easily shown that the asymptotic distributions of $\hat{\mu}T_n$ and of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{u}_i\hat{u}_j$ for any $j \neq i$ are identical. A more detailed discussion is given in the Appendix.

5 Asymptotic Property under the Alternative Hypothesis

We examine the asymptotic power property of our test by the minimax approach of Y. I. Ingster (1993), in which the alternative hypothesis is a set of functions belonging to a smoothness class. The set of alternatives is separated from the null model by L^2-distance but the distance approaches 0 at a specific rate. A test is called rate optimal when it has prescribed minimax power uniformly against a set of alternatives that approaches the null hypothesis at a rate faster than any other tests can detect. This rate is then called the optimal minimax rate.[7]

The local power property described in the minimax approach differs from that in more standard approaches, such as Pitman or singular local alternatives. In the Pitman approach, the distance between the true and parametric functions shrinks toward zero

[6] Precisely, we need to assume $E[\|\frac{\partial}{\partial\theta\partial\theta'}g(X_i,\theta)\|^2|Z_i] < \infty$ for all $\theta \in \Theta$, which we do not assume to show Proposition 1.

[7] A formal definition of the optimal minimax rate is given in Definition 1 of Guerre and Lavergne (2002).

at a specific rate, which, in our notation, is $m(Z_i) - E[g(X_i, \theta)|Z_i] = \gamma_n d(Z_i)$, where $d(\cdot)$ is a square-integrable continuous function and a deterministic sequence γ_n approaches zero as the sample size increase. In the singular local alternatives approach, the distance is expressed as a sequence of continuous functions that approach zero. For example, the singular local alternative in our notation is $m(Z_i) - E[g(X_i, \theta)|Z_i] = r_n d_n(Z_i)$, where r_n is a deterministic sequence and $\int d_n(z)^2 dz$ approaches zero as the sample size increases. In contrast to these approaches, the minimax approach captures the distance as a set of functions and evaluates the behavior of a test uniformly against such alternatives. The set of alternatives we consider includes functions that can be treated as Pitman or singular local alternatives approaching the null hypothesis at rate ρ_n, as long as they satisfy our smoothness condition.

Now, we define the smoothness class. Let $\mathcal{M}_{L,s,k}$ be a class of map defined on a compact set Ω such that

$$\mathcal{M}_{L,s,k} = \left\{ m : \sum_{j=0}^{k} \sup_{|\beta|=j} \sup_{x \in \Omega} \|D^\beta m(x)\| + \sup_{|\beta|=j} \sup_{x,y \in \Omega} \frac{\|D^\beta m(x) - D^\beta m(y)\|}{\|x - y\|^s} \leq L \right\},$$

for some smoothness index $s \in [0, 1]$, a non-negative integer k, and a positive constant L. $D^\beta m(x)$ indicates $|\beta|$-times partial derivatives of $m(\cdot)$, and $|\beta|$ is the number of partial derivatives. This class reduces to the Hölder class of functions when smoothness index is restricted within $(0, 1]$. A large body of literature, such as Ermakov (1991), Y. I. Ingster (1993), Lepski and Tsybakov (2000), and Guerre and Lavergne (2002), investigate the optimal minimax rate of testing against alternatives defined in the Hölder class.

The index s and k represent smoothness of functions belonging to the class $\mathcal{M}_{L,s,k}$. To observe the relationship between the smoothness of functions and the index, let us compare the functions in classes $\mathcal{M}_{L,s_1,k}$ and $\mathcal{M}_{L,s_2,k}$, where $s_1 > s_2$. When we focus on the local variation of functions, say $\|x - y\| \leq 1$, it is obvious that $\|x - y\| \leq \|x - y\|^{s_1} < \|x - y\|^{s_2}$. This indicates that functions belonging to $\mathcal{M}_{L,s_2,k}$ are allowed to have locally greater variation than those in $\mathcal{M}_{L,s_1,k}$. Thus, $\mathcal{M}_{L,s_2,k}$ includes functions that are locally less smooth than that in $\mathcal{M}_{L,s_1,k}$, such that $\mathcal{M}_{L,s_1,k} \in \mathcal{M}_{L,s_2,k}$. This result indicates that, given a positive constant L, a class with smaller smoothness index s contains locally less smooth functions.

Guerre and Lavergne (2002) investigate the optimal minimax rate for testing specifications of the regression function against alternatives defined in the Hölder class. Their result is that the optimal minimax rate for non-parametric specification testing in non-linear regression models is $n^{-2s/(l+4s)}$ for $s \geq l/4$, and $n^{-1/4}$ for $s < l/4$, assuming s is known a priori. In the adaptive framework, where s is set to be an unknown nuisance parameter, Horowitz and Spokoiny (2001) propose a uniformly consistent test that uses kernel smoothing to detect the alternative approaching the null hypothesis at the fastest rate $n^{-2s/(l+4s)}(\log \log n)^{s/(l+4s)}$ when $s \geq \max(2, l/4)$.

Our local alternative is defined on a cone set consisting of normalized functions in smoothness class $\mathcal{M}_{L,s,k}$:

$$\mathcal{M} = \left\{ a_n f(\cdot) : f(\cdot) \in \mathcal{M}_{L,s,k}, E[f(\cdot)^2] = 1, a_n \neq 0, a_n < a \right\},$$

where a_n is any scalar valued sequence that is bounded from above by a constant a. Let $\delta_\theta(Z) \equiv m(Z) - E[g(X,\theta)|Z]$ be the difference between the true and parametric models. The set of alternatives is

$$H_{n,1} : \mathcal{M}(\rho_n) = \left\{ \delta_\theta(Z) \in \mathcal{M} : \inf_{\theta \in \Theta} E\left[\delta_\theta(Z)^2\right] \geq \rho_n^2. \right\},$$

The minimax approach finds the fastest rate at which ρ_n approaches 0 while assuring the test uniformly detects alternatives in $\mathcal{M}(\rho_n)$. To the best of our knowledge, this is the first study that investigates the optimal minimax rate against the set of alternatives defined in this class.

The next proposition shows a lower bound $\tilde{\rho}_n$ for the optimal minimax rate against which no specification tests for IV models demonstrate non-trivial power.

Proposition 2 (*Lower Bound*)**.** *Suppose* $\{Y_i, X_i, Z_i\}_{i=1}^n$ *are an independent and identically distributed sample on* $(Y, X, Z) \in \mathbb{R} \times \mathbb{R}^{l_x} \times \mathbb{R}^l$*, where* l_x *and* l *are finite. Let* $\tilde{\rho}_n = n^{-1/4}$*. If each* $\omega_i \equiv Y_i - m(Z_i)$ *is* $N(0,1)$ *conditionally upon* Z_i *and* Z_i *is uniformly distributed, for any test* t_n *with* $\sup_{\delta_\theta(Z) \in H_0} P(t_n > z_\alpha) \leq \alpha + o(1)$*,*

$$\sup_{\delta_\theta(Z) \in \mathcal{M}(\rho_n)} P(t_n \leq z_\alpha) \geq 1 - \alpha + o(1), \qquad \textit{whenever} \quad \rho_n = o(\tilde{\rho}_n),$$

where z_α *indicate the* α *level critical value of test* t_n*.*

Proposition 2 is proved by replacing the minimax problem with a Bayesian problem. This is a standard argument to show the lower bound of the optimal minimax rate (see, e.g., Y. I. Ingster, 1993, Spokoiny, 1996, Lepski & Spokoiny, 1999, Lepski & Tsybakov, 2000, Guerre & Lavergne, 2002, Abramovich et al., 2009, and Y. I. Ingster & Sapatinas, 2009).

Remark 6. Proposition 2 shows the lower bound of specification tests for IV models. No test has non-trivial uniform power against the set of alternatives $\mathcal{M}(\rho_n)$ when it approaches the null at a rate faster than $n^{-1/4}$. This result differs from that of Guerre and Lavergne (2002), in which the lower bounds depend on the smoothness of alternatives and the dimension of conditioning variables. The difference mainly arises because the set of alternatives we consider is constructed based on $\mathcal{M}_{L,s,k}$, which includes the Höder class employed in existing works.[8]

Finding a lower bound requires an example function that is difficult to detect (even by the optimal Bayesian test) but still belongs to the set of alternatives. In the proof of Proposition 2 given in the Appendix, we construct the example function $f_{n,c}(Z_i)$ by

[8]The lower bounds under the IV model setting against the set of alternatives that is the same as that in Guerre and Lavergne (2002) can be proved when following the proof of Theorem 2 in Guerre and Lavergne (2002), and the results are exactly the same as those derived under their regression setting. Thus, the lower bound does not change, even though the model is extended to an IV regression model. By contrast, the set of alternatives in this study consists of a cone of functions belonging to the smoothness class $\mathcal{M}_{L,s,k}$, which includes the smoothness class used in Guerre and Lavergne (2002).

wavelet series, which form like a local alternative function that shrinks toward the null at the rate of ρ_n. The example function has a tuning parameter c, which represents the resolution level of wavelets that determines the frequency of the example function. Roughly, a small c constructs a low frequency function that is detected easily by a test, while a large c makes it difficult for tests to detect the example function. We obtain the following corollary.

Corollary 1. *Let $\tilde{\rho}_n = n^{-2/(4+cl^2)}$ if $l < 2/\sqrt{c}$ and $\tilde{\rho}_n = n^{-1/4}$ if $l \geq 2/\sqrt{c}$. Suppose the setting of Proposition 2 holds. No test has non-trivial power against the local alternative $\rho_n f_{n,c}(Z_i)$ (defined in the proof of Proposition 2), whenever $\rho_n = o(\tilde{\rho}_n)$.*

Remark 7. Since the dimension of the instruments l is restricted by c, a tradeoff exists between the achievable testing power and the dimension of instruments. When an alternative function is smooth so that it can be well approximated by wavelets with low resolution level (c small), tests can exhibit local power at the rate $n^{-2/(4+cl^2)}$ for moderately small l that satisfies $l < 2/\sqrt{c}$. By contrast, when l is large, no tests achieve the rate faster than the minimax rate of $n^{-1/4}$ even against smooth alternatives.

In the following part, we investigate the minimax rate of our test under $H_{n,1}$. The following assumptions are imposed.

Assumption 7. *The density of Z, denoted as $f(\cdot) : \mathbb{R}^l \to \mathbb{R}$, has compact support (without loss of generality $[0,1]^l$), satisfies $0 < \underline{f} \leq f(z) \leq \overline{f} < \infty$ for any $z \in [0,1]^l$, and is continuous on $[0,1]^l$.*

Assumption 8. *For each $\theta \in \Theta$, $E[g(X,\theta)^8|Z] < \infty$.*

Assumption 9. *For each $\theta \in \Theta$, $E[g(X,\theta)|Z] \in \mathcal{M}_{L_g,s,k}$, for some constant $L_g \leq L$.*

Assumption 10. *For each $m(\cdot) \in \mathcal{M}_{(1+a)L,s,k}$, a unique pseudo-true value for θ exists such that*

$$\theta_m^* \equiv \arg\min_{\theta\in\Theta} E\{[m(Z) - g(X,\theta)]Z\}' M E\{[m(Z) - g(X,\theta)]Z\},$$

where M is a symmetric and positive definite $l \times l$ weight matrix. Letting $\theta_m^* = \theta^*$, $\sqrt{n}(\hat{\theta} - \theta^*) = O_p(1)$ uniformly with respect to $m(\cdot) \in \mathcal{M}_{(1+a)L,s,k}$.

As in Assumption 7, we focus on the case in which Z is continuous when investigating the power property under the alternative, although the continuity is not the essential assumption. The source of testing power comes from the distance between the true and the parametric models, that is, $\delta(Z_i)\delta(Z_{i^*}) \approx \delta(Z_i)^2 + (Z_i - Z_{i^*})\partial\delta(Z_i)/\partial Z_i$ under the alternative. Since the sign of the second term is unknown, our test exhibits better power property when the second term is zero. The continuity of Z is one of the sufficient conditions to make Z_{i^*} approach Z_i as sample size increases (see, e.g., Lemma 14.1 of Q. Li & Racine, 2007). Thus, continuity of Z in Assumption 7 is not necessary and can be replaced with other assumptions that ensure Z_{i^*} is close to Z_i. For example, the power property in Proposition 3 below also holds when we assume that $f(\cdot)$ is discrete instead

of continuous as long as the number of realization points of Z with positive probability is not large relative to the sample size. In IV regression models, the compactness of the instruments in Assumption 7 is not restrictive at all because it can be achieved by an appropriate monotone transformation.

Assumptions 8 and 9 restrict the parametric model of interest. Similar assumptions are used in existing works, such as Guerre and Lavergne (2002) and H. Li et al. (2016), to show the minimax power of testing. Assumption 9 is used to replace arguments of the minimax approach with respect to the uniformity in $\delta_\theta(Z)$ with uniformity in $m(Z)$. Under the local alternative that $\delta_\theta(Z) \in \mathcal{M}$, a sequence a_n bounded by a exists and satisfies $m(Z) - E[g(X,\theta)|Z] \in \mathcal{M}_{a_n L,s,k} \subset \mathcal{M}_{aL,s,k}$, implying $m(Z_i) \in \mathcal{M}_{(1+a)L,s,k}$ under Assumption 9. Assumption 8 along with the boundedness of the error term in Assumption 1 guarantees $E(u_i^{*p}|Z_i) < \infty$ for $p = 8$, which is required for the consistency of $\hat{\mu}$ under the alternative hypothesis. Similar to the consistency of $\hat{\mu}$ under the null, it is possible to reduce the finite eighth moment of the error term in Assumption 8 to its finite fourth moment.

Assumption 10 ensures the existence of the pseudo-true values and guarantees its estimator $\hat{\theta}$ to be $\sqrt{n}$-consistent uniformly in $m(Z) \in \mathcal{M}_{(1+a)L,s,k}$. Uniform consistency is essential for developing the minimax approach, because the approach finds the local power of a test while maintaining a test can detect alternatives, in which the IV function $m(\cdot)$ belongs to the smoothness class, uniformly. For further discussion, see Guerre and Lavergne (2002), who give an example of an ordinary least squares estimator for a simple univariate regression model that satisfies uniform consistency.

The weighting matrix M is arbitrary but researchers should choose carefully. As A. R. Hall and Inoue (2003) suggest, the probability limit and the limiting distribution of the GMM estimator depend on the limit of the weighting matrix and the limiting distribution of the elements of the weighting matrix, respectively, when both parameter over-identification and non-local misspecification are present. Although asymptotic behavior of GMM estimators under our alternative hypothesis is not trivial, we leave it for future work, as it exceeds the scope of this study. To circumvent the problem of asymptotic behavior of GMM estimators, we regard M as an identity matrix throughout this study.

Proposition 3 shows that our test has non-trivial uniform power against the alternatives in $H_{n,1}$ that approach the null hypothesis at the rate $\kappa n^{-1/4}$ for a constant κ. Together with the the lower bound, this result indicates that our test is rate optimal and the optimal minimax rate for the specification tests for IV models is $n^{-1/4}$.

Proposition 3. *Suppose Assumptions 1, 2, 3, 4, 5, 7, 8, 9, and 10 hold. Let $\rho_n = n^{-1/4}$. For any prescribed bound $\beta \in (0, 1-\alpha)$, a constant κ exists such that*

$$\sup_{\delta_\theta(Z) \in \mathcal{M}(\kappa\rho_n)} P(T_n \leq z_\alpha) \leq \beta + o(1).$$

Remark 8. The power property against Pitman local alternatives is established by Jun and Pinkse (2009). Their results imply that K-nearest neighbor tests for IV regression models can detect Pitman local alternatives converging to the null at a rate equal to

$n^{-1/4}$ when the number of neighbors is constant. Proposition 3 complements their results by showing a set of alternatives approaching the null at a rate equal to $n^{-1/4}$, against which K-nearest neighbor tests with a fixed number of neighbors have non-trivial uniform power.

Remark 9. Propositions 2 and 3 along with Corollary 1 imply that power performance of K-nearest neighbor tests does not always improve as the number of neighbors K increases. This finding contradicts the results that a K-nearest neighbor test performs better against Pitman local alternatives when K grows with sample size. Jun and Pinkse (2009) show K-nearest neighbor tests detect such alternatives approaching the null at a rate of $(nK)^{-1/4}$. Increasing power property, however, does not apply against a wider class of alternatives such as ours, because Pitman local alternatives focus only on certain departures from the null.[9] Indeed, Proposition 2 and Corollary 1 expose the set of alternatives against which our single nearest neighbor test achieves the fastest possible rate. Since rate optimality of our test holds even for large l, there are no tests that perform better in terms of power when the dimension of instruments is large. Therefore, K-nearest neighbor tests with fixed K neighbors such as ours complement other tests, because they have computational advantage in the large dimension setting.

6 Monte Carlo Experiments

6.1 Size and Power against Non-smooth Alternative

We conduct Monte Carlo studies to investigate the size and power performance of T_n under finite samples. We test the null hypothesis that

$$g(x) = \theta_0 + \theta_1 x + \theta_2 w_1 + \theta_3 w_2 \tag{5}$$

The data-generating processes (DGP) are

$$\begin{aligned} x_i &= \Phi\left(\rho v_{1,i} + (1-\rho^2)^{1/2} v_{2,i}\right), \\ z_i &= \Phi(v_{1,i}), \\ u_i &= 0.2\Phi\left(\eta v_{2,i} + (1-\eta^2)^{1/2} v_{3,i}\right), \\ y_i &= 1 + x_i + w_{1,i} + w_{2,i} + \beta h(w_{1,i}) + u_i, \end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. We have five random variables $v_{1,i}$, $v_{2,i}$, $v_{3,i}$, $w_{1,i}$, and $w_{2,i}$. The first three variables are driven randomly from $N(0,1)$ and the remaining two are random samples from $U[0,\pi]$.

In this experiment, x_i is endogenous, which is instrumented by z_i. The endogeneity accrues via v_2, and correlation between x_i and u_i depends on parameters ρ and η. The exogeneity of the instruments is guaranteed because u_i does not depend on v_1 that generates instruments. Correlation between instruments and the endogenous variable

[9] For features of Pitman local alternatives compared to others, see, Fan and Li (2000)

can be gauged by parameter ρ. The DGPs for x, z, and u are adapted from Horowitz (2006).

The outcome y consists of endogenous variable x, two exogenous variables $w_{1,i}$ and $w_{2,i}$, and a function $h(\cdot)$ that generates misspecification under alternatives (when $\beta \neq 0$). To introduce non-smooth alternatives, $h(\cdot)$ is set to be a Haar wavelet function. In particular, $h(x) = 1$ if $x \in (1, 1.5)$, $h(x) = -1$ if $x \in (1.5, 2)$, and $h(x) = 0$, otherwise as illustrated in Figure 1.

The form of the DGP for the outcome resembles the local alternatives considered in the proof of Proposition 2, where we have fixed parameter β instead of ρ_n, which shrinks as sample size increases. The size of the test can be investigated by testing the null hypothesis (5) under $\beta = 0$, while the power can be examined by letting β be non-zero. According to the results of Propositions 2 and 3, our test has power uniformly as long as $\beta = O(n^{-1/4})$.

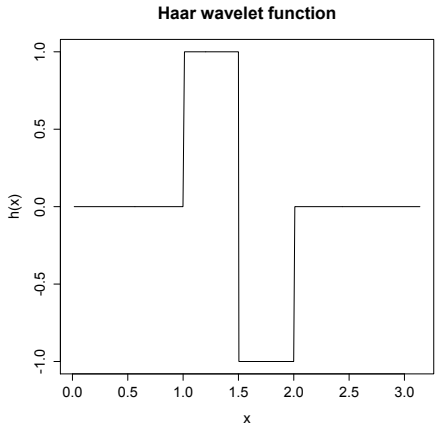

Figure 1: Haar wavelet function.

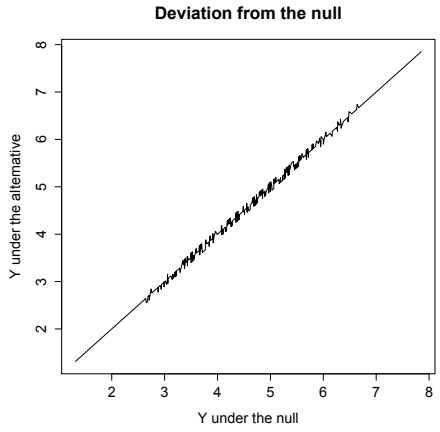

Figure 2: Deviation from the null. **Notes:** Data are generated by DGP1 and $\beta = 0.1$.

We consider three sets of DGPs, called DGP 1, DGP 2, and DGP 3. In each set of DGPs, different parameter values for ρ and η are assigned: DGP 1: $\rho = 0.8$, $\eta = 0.1$; DGP 2: $\rho = 0.8$, $\eta = 0.5$; DGP 3: $\rho = 0.7$, $\eta = 0.1$. Figure 2 illustrates the outcome y under the alternative. The deviations from the 45-degree line show the distance between the null and alternative models.

Table 1 shows the finite sample size performance of our test. Sample sizes are chosen to be $n = \{100, 250, 500, 1000, 5000\}$ and the results are based on $M = 1,000$ simulation runs. The test tends to under-reject the null hypothesis in most cases for all significance levels, indicating that the test is conservative. However, the estimated sizes get closer to their nominal sizes for all DGPs as sample size increases.

Figure 3 displays the power functions of the test statistic depending on β. For each value of $\beta = \{0.01, 0.02, \cdots, 0.20\}$, we conduct $M = 1,000$ simulation runs to obtain fractions for rejecting the null hypothesis. The first row of this figure shows the power

Table 1: Monte Carlo results for size.

	GDP1				GDP2				GDP3			
n	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
100	0.007	0.032	0.070	0.135	0.001	0.027	0.062	0.131	0.007	0.031	0.072	0.140
250	0.009	0.029	0.066	0.142	0.001	0.029	0.068	0.146	0.009	0.028	0.067	0.141
500	0.007	0.042	0.081	0.177	0.008	0.035	0.071	0.159	0.006	0.043	0.083	0.173
1000	0.012	0.046	0.094	0.175	0.008	0.054	0.092	0.175	0.012	0.047	0.094	0.174
5000	0.008	0.045	0.089	0.192	0.013	0.048	0.088	0.188	0.008	0.045	0.090	0.192

Note: Since the test is one-sided, the null is rejected when the test statistic is larger than 1.64.

functions when z_i and two exogenous variables in the model are used as instruments (total of three instruments). Recall that the value for β determines how far an alternative is apart from the null model. Figure 3 clearly illustrates the increasing power performance as β becomes large. In particular, the fractions of rejection become exactly 1 for all DGPs when β is larger than 0.06, 0.08, and 0.13 for $n = 1000$, $n = 500$, and $n = 250$, respectively. Figure 3 shows increasing power performance as sample size grows. These results coincide with those given in Propositions 2 and 3, indicating that our test is powerful when β is not too small relative to the sample size.

6.2 Power under Many IVs

An interesting set-up for the simulation study is an IV model with many instruments. In addition to the DGPs described above, we generate instruments as follows:

$$z_{i,j} = v_{1,i} + v_{4,i,j}, \quad \text{for } j = 1, \ldots, l,$$

where $v_{4,i,j}$ is driven randomly from $N(0, \sigma_j^2)$, and σ_j^2 is driven randomly from $U[0, 0.1]$.

The second row of Figure 3 illustrates the power functions against non-smooth alternatives when $l = 10$, that is, the model has a total of 12 instruments, including 2 exogenous variables in the model. For small sample size with $n = 100$ and $n = 250$, the slopes of power functions are gentler compared to the previous simulation results, with only 3 instruments shown in the first row of Figure 3. In particular, the rejection probabilities are around 0.2 for $\beta = 20$ for all DGPs when $n = 100$, while they are around 0.8 in the previous simulation results. However, the power performance increases as the sample size grows. In particular, the fractions of rejection become exactly 1 for all DGP when β is larger than 0.08 and 0.14 for $n = 1000$ and $n = 500$, respectively. A similar tendency is captured for power functions when the model has a total of 27 instruments, including 2 exogenous variables (the third row of Figure 3). Overall, power declines as the dimension of instruments increases. Rejection probabilities of around 0.10 are achieved for $\beta = 20$ for all DGPs when $n = 100$. The fractions of rejection, however, increase as the sample size grows and the fractions become exactly 1 for all DGPs when β is larger than 0.09 for $n = 1000$. It is rather surprising that the test rejects the null hypotheses of correct model specification with the high probabilities, even the model includes 27 instruments.

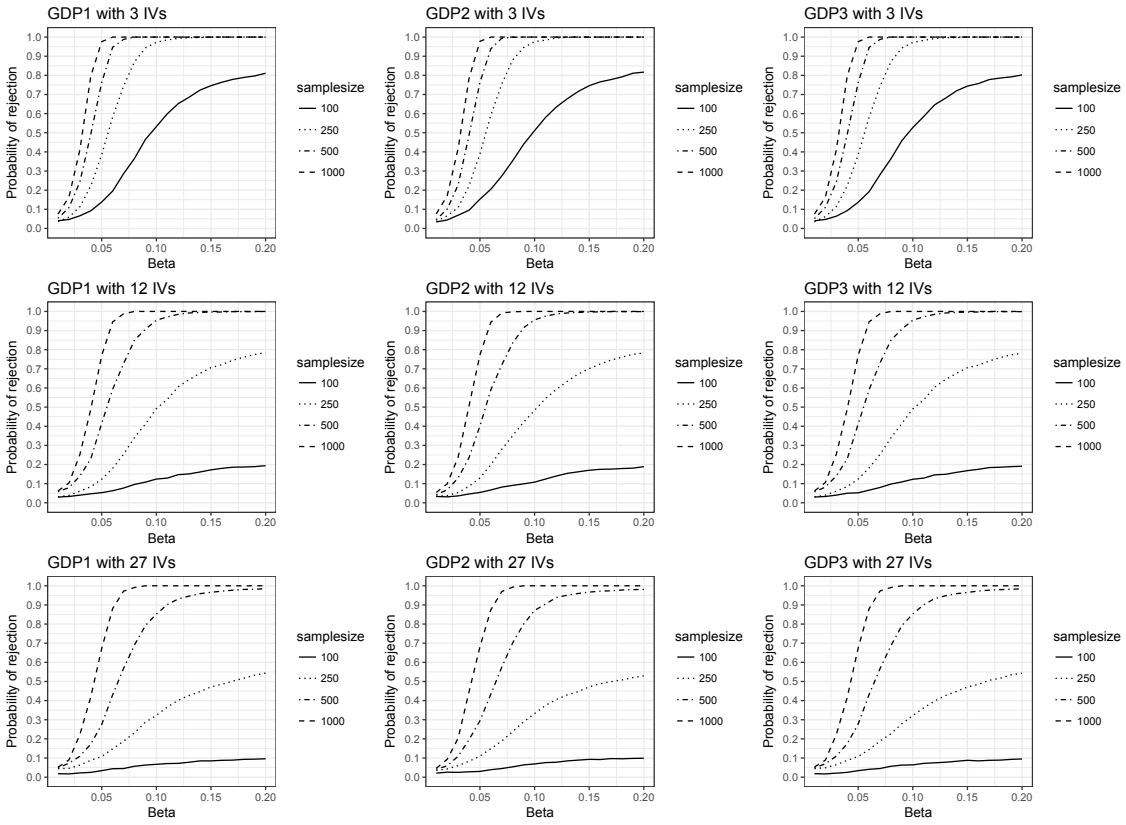

Figure 3: Power functions against non-smooth alternatives. **Notes:** Reported are fractions of rejection of the null hypothesis in 1000 simulation runs.

7 Application

This section applies our test to the models for return to schooling and specification of Engel curves. The empirical application for return to schooling highlights attractive feature of our test when model include many IVs. The appropriate specification of Engel curves depends on the source of endogeneity and assumptions set on it. Since the source is unobserved, researchers need to test the specification of the model. As discussed in Remark 3, monotonicity condition may hold for Engel curves, indicating that the null hypothesis implies correct specification of $g(\cdot)$. In the following application, we use the modified test statistic (4) when ties exist.

7.1 Return to Schooling

A motivating empirical example to apply our test to is return to schooling investigated by Angrist and Krueger (1991). The sample is the 1980 US census and consists of $329,509$ men born between 1930 and 1939. Angrist and Krueger (1991) estimate the following

IV regression model:

$$\text{wage} = \beta\text{education} \ + \alpha' X + \text{error},$$

where wage is the log of weekly wage, education is the year of education, and X is a vector of exogenous variables, such as 9 year-of-birth (YOB) dummies and 50 state-of-birth (SOB) dummies. The vector of instruments Z for endogenous education includes the quarter of birth (QOB) dummies, exogenous variables X, all cross-terms of QOB and X, and a constant. We use a total of 240 instruments, namely, 3 QOB dummies, 9 YOB dummies, 50 SOB dummies, 27 QOB times YOB interactions, 150 QOB times SOB interactions, and a constant. The estimation results are given in the second column of Table VII of Angrist and Krueger (1991). The P-value for Sargan's statistic is 0.800 (the test value is 163 with 179 degrees of freedom), indicating that the test does not reject the null of true model specification. The value of our test is -0.29, thereby not rejecting the null hypothesis. This result amplifies the fact that our test is acceptable in terms of computational cost and well applicable even when the numbers of instruments and sample size are quite large.

7.2 Specification of Engel curves

Next, we apply our test to models for Engel curves (or consumer expansion paths). Engel curves represent the relationship between the shares of expenditure on specific commodities and total expenditure. Since the shape of the curve determines the elasticity of commodities toward total expenditure and displays whether the good is an inferior, normal, or luxury good, empirical analyses of Engel curves are important in understanding, for example, consumer responses to policy reform. However, endogeneity arises from simultaneity and measurement errors in expenditure data, which makes the estimation of Engel curves cumbersome.

The baseline model is the Working–Leser specification of Engel curves, a structural model originating in consumer theory. Expenditure shares of a commodity have linear relationships to the log of total expenditure in the Working–Leser specification, which is shown to have micro foundations (Muellbauer, 1976, Deaton & Muellbauer, 1980, and Jorgenson, Lau, & Stoker, 1982). Let $y_{i,j}$ be the expenditure on good j by individual i, $X_i \equiv \sum_j y_{i,j}$ be total expenditures, and W be a vector of exogenous variables. Then, the Working–Leser specification of Engel curves is

$$\frac{y_{i,j}}{X_i} = \alpha_{0,j} + \alpha_{1,j}' W_{i,j} + \beta_j \log X_i + \epsilon_{i,j}, \tag{6}$$

where ϵ is unobserved regression error and α_0, α_1 and β are unknown parameters to be estimated.

There are at least two plausible ways in which ordinary least squares estimation can fail to obtain consistent parameter estimates. First, total expenditure seems to be simultaneously determined with expenditure for each good. Second, expenditure data may be measured with errors. Since both simultaneity and measurement error can cause correlation between the error term and total expenditure, X is likely to be endogenous in equation (6).

While endogeneity driven by simultaneity can be addressed by employing the conventional IV approach, it fails to obtain consistent parameter estimates when non-linear measurement errors exist. Measurement errors are non-linear in Engel curves, because error-ridden expenditure appears on both sides of equation (6). The only exception in which the IV approach leads to consistent parameter estimates is when errors are multiplicative and identical to all goods. To observe this case, let $\tilde{y}_{i,j}$ be the real expenditure for good j and $\tilde{X}_i \equiv \sum_j \tilde{y}_{i,j}$, and assume multiplicative measurement errors $y_{i,j} = \tilde{y}_{i,j} v_i$, where v_i is a mean zero random variable independent of $\tilde{y}_{i,j}$ and instruments. By summing up over all goods and taking the logarithm, we obtain $\log X_i = \log \tilde{X}_i + \log v_i$. Moreover, errors on the left-hand side of equation (6) disappear when errors do not depend on the type of good so that $y_{i,j}/X_{i,j} = \tilde{y}_{i,j}/\tilde{X}_i$. However, homogeneous measurement errors are not realistic in many cases, and thus, it is more plausible to allow non-linear measurement errors to appear in Engel curves.

We focus on three estimation strategies for Engel curves. They are IV, Lewbel's (1996), and Battistin and De Nadai's (2015) approaches. Lewbel (1996) proposes an estimation procedure when measurement error is the only source of endogeneity. Recently, Battistin and De Nadai (2015) show an identification strategy when both simultaneity and measurement error are present. Although all of the IV, Lewbel (1996), and Battistin and De Nadai (2015) approaches aim to estimate the same parameter β_j, different assumptions about the source of endogeneity lead to some differences in the specification of Engel curves.

Let Z_i be a vector of IVs that includes exogenous variables $W_{i,j}$. In the IV approach, it is assumed that

$$E\left(\frac{y_{i,j}}{X_i}\middle| Z_i\right) = \alpha_{0,j} + \alpha'_{1,j} W_{i,j} + \beta_j E(\log X_i | Z_i).$$

The parameters are identified through a 2SLS regression of y/X on constant W and $\log X$, where $\log X$ is instrumented by Z.

When additive measurement errors are assumed such that $y_{i,j} = \tilde{y}_{i,j} + \tilde{X}_i v_j$, equation (6) turns out to be

$$\frac{y_{i,j}}{X_i} = \frac{\tilde{y}_{i,j}/\tilde{X}_i + v_j}{V} = \frac{\alpha_{0,j} + \alpha'_{1,j} W_{i,j} + \beta_j \log \tilde{X}_i + \epsilon_{i,j} + v_j}{V}, \tag{7}$$

where $V \equiv 1 + \sum_j v_j$.

Under the assumption that v_i is a mean zero random variable independent of $\tilde{X}_i$, W_i, $\epsilon_{i,j}$, and instruments Z_i, $E(X_i|Z_i) \neq 0$, and $E(\epsilon_{i,j}|Z_i) = 0$ and the fact that $X_i = \tilde{X}_i V$, it holds that $E(X_i|Z_i) = E(\tilde{X}_i V|Z_i) = E(\tilde{X}_i|Z_i)E(V) = E(\tilde{X}_i|Z_i)$, $E(W_{i,j}X_i|Z_i) = E(W_{i,j}\tilde{X}_i|Z_i)$, and $E(X_i \log X_i|Z_i) = E[\tilde{X}_i V(\log \tilde{X}_i + \log V)|Z_i] = E[\tilde{X}_i \log \tilde{X}_i|Z_i] + E(\tilde{X}_i|Z_i)E(V \log V|Z_i)$. Thus, multiplying either side of equation (7) by X_i and taking conditional expectations with respect to Z_i yields

$$E\left(y_{i,j}|\, Z_i\right) = \tilde{\alpha}_{0,j} E(X_i|Z_i) + \alpha'_{1,j} E(W_{i,j}X_i|Z_i) + \beta_j E(X_i \log X_i|Z_i) + E(\tilde{X}_i \epsilon_{i,j}|Z_i), \tag{8}$$

where $\tilde{\alpha}_{0,j} \equiv \alpha_{0,j} - \beta_j E(V \log V | Z_i)$. In Lewbel's (1996) approach, it is assumed that $E(\epsilon_{i,j}|\tilde{X}_i) = 0$, so that the fourth term of equation (8) is zero. The parameters are identified through a 2SLS regression of y on X, WX, and $X \log X$ without a constant and Z as instruments.

To address the violation of $E(\tilde{X}_i \epsilon_{i,j}|Z_i)$ being zero assumed in Lewbel (1996), Battistin and De Nadai (2015) use a control function approach. Let η_i be the residual term from the regression of $\log X_i$ on the set of instruments Z_i and $\tilde{\eta}_i$ be the residual using $\log \tilde{X}_i$ instead of $\log X_i$. The authors set a parametric assumption that $E(\epsilon_{i,j}|Z_i, \tilde{\eta}_i) = \rho_i \tilde{\eta}_i$, which yields $E(\tilde{X}_i \epsilon_{i,j}|Z_i) = E[\tilde{X}_i E(\epsilon_{i,j}|Z_i, \tilde{\eta}_i)|Z_i] = \rho_i E[\tilde{X}_i \tilde{\eta}_i|Z_i]$.

Since $E[X_i \eta_i|Z_i] = E[\tilde{X}_i \tilde{\eta}_i|Z_i] + \mathrm{cov}(V, \log V)$ by using $\eta_i = \tilde{\eta}_i + \log V - E(\log V)$, we obtain

$$E\left(y_{i,j}|\, Z_i\right) = \alpha_{0,j} E(X_i|Z_i) + \alpha'_{1,j} E(W_{i,j} X_i|Z_i) + \beta_j E(X_i \log X_i|Z_i) + \rho_i E(X_i \eta_{i,j}|Z_i),$$

where $\alpha_{0,j} \equiv \alpha_{0,j} - \beta_j E(V \log V|Z_i) - \rho_i \mathrm{cov}(V, \log V)$. By replacing η with its fitted values $\hat{\eta}$, parameters, including ρ_j, are identified through a 2SLS regression of y on X, WX, $X \log X$, and $X\hat{\eta}$ without a constant and Z as instruments.

In summary, each of IV, Lewbel's (1996), and Battistin and De Nadai's (2015) approaches has its own econometric specification for Engel curves. They are represented in the following moment restrictions:

$$E\left(\epsilon^{\mathrm{IV}}_{i,j}\middle|\, Z_i\right) = 0, \quad E\left(\epsilon^{\mathrm{L}}_{i,j}\middle|\, Z_i\right) = 0, \text{ and } E\left(\epsilon^{\mathrm{BN}}_{i,j}\middle|\, Z_i\right) = 0, \tag{9}$$

where

$$\begin{aligned}
\epsilon^{\mathrm{IV}}_{i,j} &\equiv y_{i,j}/X_i - \alpha_{0,j} - \alpha'_{1,j} W_{i,j} - \beta_j \log X_i \\
\epsilon^{\mathrm{L}}_{i,j} &\equiv y_{i,j} - \tilde{\alpha}_{0,j} X_i - \alpha'_{1,j} W_{i,j} X_i - \beta_j X_i \log X_i \\
\epsilon^{\mathrm{BN}}_{i,j} &\equiv y_{i,j} - \alpha_{0,j} X_i - \alpha'_{1,j} W_{i,j} X_i - \beta_j X_i \log X_i - \rho_i X_i \hat{\eta}_{i,j}.
\end{aligned}$$

When $\epsilon_{i,j}$ is exogenous to both X_i and Z_i, all moment restrictions in (9) hold. Consider that the source of endogeneity is only the simultaneous determination (or omitted variables). Then, moment restriction of the IV approach holds, while that of Battistin and De Nadai's (2015) approach holds only if the parametric specification for $E(\epsilon_{i,j}|Z_i, \tilde{\eta}_i)$ is correct. The moment restriction of Lewbel's (1996) approach may not hold, since $E(\epsilon_{i,j}|\tilde{X}_i) = E(\epsilon_{i,j}|X_i) \neq 0$. By contrast, when endogeneity arises only from the measurement error of the form discussed above, moment restrictions of Lewbel's (1996) and Battistin and De Nadai's (2015) approaches hold, while those of the IV approach fail. When both simultaneity and measurement errors are present, only the moment restrictions of Battistin and De Nadai's (2015) approach hold under the correct parametric assumption for $E(\epsilon_{i,j}|Z_i, \tilde{\eta}_i)$.

However, in the empirical analysis, we are unaware of the source of endogeneity, indicating that testing the model specifications (9) enables us to investigate the fitness of each model. We apply these models to two data sets, and test the model specifications by employing our test.

7.2.1 Example 1: Italian Household Survey

This section adapts the application given in Battistin and De Nadai (2015) to test the Engel curve specifications of IV, Lewbel's (1996), and Battistin and De Nadai's (2015) approaches. The data are the 2010 wave of the Bank of Italy's *Survey on Households' Income and Wealth* (SHIW).

Battistin and De Nadai (2015) focuses on Engel curves for food, the linearity of which is supported by substantial empirical evidence, and runs separate regressions depending on the number of children in the household (couples without children, couples with one child, and couples with more than one child). Exogenous variables are the household regional variation represented by macro area dummies (North, Center, and South) and instruments for the total expenditure are the average of male logged wages across areas. The detailed explanation for data sets and estimation results is given in Battistin and De Nadai (2015).

Table 2: Test for Engel curve specification. SHIW 2010 data

	IV		Lewbel (1996)		BN (2015)	
No children	0.414	(0.339)	5.161	(0.000)	2.230	(0.013)
One child	2.765	(0.003)	3.801	(0.000)	4.929	(0.000)
More than one child	3.193	(0.001)	5.612	(0.000)	5.595	(0.000)

Note: Presented are the test statistics T_n in equation (3). Sample size for groups "No children," "One child," and "More than one child" are 345, 709, and 1257, respectively. P-values are given in parentheses. BN (2015) denotes Battistin and De Nadai (2015).

Table 2 presents the test statistics for each model specification and household group. The null hypothesis is that the specification of the model is true. We test each model specification (IV, Lewbel's 1996, and Battistin and De Nadai's 2015 model) in three populations, that is, households without children, those with one child, and those with more than one child. Since the test is one-sided, the null hypothesis is rejected when the P-value (in parentheses) is smaller than the significance level. For households without children (the first row), the test rejects only Lewbel's (1996) specification at the 1% significance level. Recall that Lewbel's (1996) approach is appropriate when the measurement error is the only source of endogeneity. This result coincides with the suggestion of Battistin and De Nadai (2015) that total expenditure endogeneity caused by simultaneity might be a more serious problem than measurement error, at least in these data. For households with one child and more than one child (the second and third rows, respectively), all model specifications are rejected even at the 1% significance level. Overall, our results suggest that a model should be chosen carefully, because the source of endogeneity and thus, model specifications may depend on the population of interest.

7.2.2 Example 2: Japanese Household Survey

We use data from the 2004 wave of the *National Survey of Family Income and Expenditure* (NSFIE). Every 5 years since 1959, the NSFIE gathers detailed information on

household consumption and income in Japan. We focus on households formed by couples with and without children, where the household head was 25 to 60 years old in 2004. Furthermore, we select a subsample of households whose disposable income lies between the 25th and 75th percentiles of income distribution of all households, resulting in a sample of $10,037$ households.

Table 3: Descriptive statistics.

	No children		One child		More than one child	
Variable	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Log expenditure	12.50	0.43	12.50	0.40	12.54	0.35
Log income	12.81	0.17	12.79	0.17	12.81	0.16
Food share	0.22	0.09	0.23	0.08	0.26	0.08
Lisure share	0.09	0.07	0.09	0.07	0.10	0.06
Sample size	1,936		2,716		5,385	

Note: The data source is the NSFIE.

We focus on Engel curves for food and leisure, and run separate regressions depending on the number of children in the household (couples without children, couples with one child, and couples with more than one child). Total expenditure represents average monthly living expenditure, which does not include non-living expenditure, such as taxes and social insurance. Exogenous variables are an urban–rural indicator, which is 1 for households in one of three metropolitan areas (Tokyo, Chukyo, or Kinki) in Japan, and gender rate, which is defined as the number of male household members divided by the number of all household members. Instruments for the total expenditure are disposable income, which is salary or wage from which non-living expenditure, such as taxes and social insurance premiums, is deducted. Table 3 documents the means and standard deviations of key variables for couples with no children, one child, and more than one child, where the sample sizes are $1,936$, $2,716$, and $5,385$, respectively.

Let X, W, and Z be total expenditure, a vector of exogenous variables (urban–rural indicator and gender rate), and disposable income, respectively. In the IV approach, the parameters in Engel curves are identified through a 2SLS regression of outcomes (food share and leisure share) on W, $\log X$ with a constant, where $\log X$ is instrumented by Z. Lewbel's (1996) approach identifies parameters through a 2SLS regression of outcomes (log food and leisure expenditure) on X, WX, and $X \log X$ without a constant, where all covariates are instrumented by Z, $\log Z$, $Z \log Z$, and their interactions with W without a constant in the first stage (for a total of 9 instruments). The first stage in Battistin and De Nadai's (2015) approach obtains $\hat{\eta}$, which is the residual term of regression $\log X$ on Z, $\log Z$, $Z \log Z$, and their interactions with W without a constant. Parameters are identified through a 2SLS regression of outcomes (log food and leisure expenditure) on X, WX, $X \log X$, and $X\hat{\eta}$ without a constant, where all covariates are instrumented by Z, $\log Z$, $Z \log Z$, and their interactions with W without a constant in the first stage

(for a total of 9 instruments).

Table 4: Estimation results for Engel curve. NSFIE 2004 data

	IV		Lewbel (1996)		BN (2015)	
Results for food share						
No children	-0.084^{***}	(0.013)	-0.083^{***}	(0.013)	-0.075^{***}	(0.028)
One child	-0.020^{*}	(0.012)	-0.030^{**}	(0.014)	-0.032^{***}	(0.013)
More children	-0.039^{***}	(0.008)	-0.052^{***}	(0.008)	-0.050^{***}	(0.008)
Results for leisure share						
No children	0.027^{**}	(0.013)	0.008	(0.019)	0.003	(0.020)
One child	0.037^{***}	(0.010)	0.020	(0.014)	0.008	(0.021)
More children	0.032^{***}	(0.007)	0.017^{**}	(0.008)	0.021	(0.013)

Note: Presented are the estimation results for β in the Engel curve given in equation (6). Standard errors are given in parentheses. The level of significance are shown by ***, **, and * for 1%, 5%, and 10%, respectively. BN (2015) denotes Battistin and De Nadai (2015).

Table 4 displays estimates for β in equation (6), which is identified by IV, Lewbel's (1996), and Battistin and De Nadai's (2015) approaches. The first and second panels show results for food and leisure shares in total expenditure as outcomes, respectively, where separate regressions are run for couples without children, couples with one child, and couples with more than one child. The results suggest that estimates for β depend on model identification.

Table 5: Test for Engel curve specification. NSFIE 2004 data

	IV		Lewbel (1996)		BN (2015)	
Results for food share						
No children	0.503	(0.307)	0.584	(0.280)	0.960	(0.169)
One child	1.208	(0.114)	-0.173	(0.569)	-0.259	(0.602)
More than one child	0.072	(0.471)	0.885	(0.188)	2.199	(0.014)
Results for leisure share						
No children	0.354	(0.362)	0.040	(0.484)	-0.280	(0.610)
One child	-0.381	(0.649)	-0.176	(0.570)	3.593	(0.000)
More than one child	0.068	(0.473)	0.810	(0.188)	-0.209	(0.583)

Note: Presented are the test statistics T_n in equation (3). P-values are given in parentheses. BN (2015) denotes Battistin and De Nadai (2015).

Table 5 reports the test results. The first panel focuses on the Engel curve of food share. For households without children and with one child, all models are not statistically different from the true model even at the 10% significance level. For households with more than one child, the null is rejected for Battistin and De Nadai's (2015) model

at the 5% significance level, implying that additional assumptions in their approach, such as parametric assumptions for $E(\epsilon_{i,j}|Z_i, \tilde{\eta}_i)$, may be inappropriate in this case. For the Engel curve of leisure share given in the second panel, the null is not rejected for all models even at the 10% significance level except Battistin and De Nadai's (2015) model for households with one child. Overall, similar to the results from SHIW in Example 1, these results indicate that the model should be chosen carefully, because appropriate model specifications might depend on the population of interest. For NSFIE data, both measurement error and simultaneity might be a serious problem that cause total expenditure endogeneity, and the test results are likely to be sensitive to the additional parametric restriction in Battistin and De Nadai's (2015) model.

8 Conclusion

This study proposes a rate optimal specification test for IV regression models. Our test is rate optimal against a set of alternatives $\mathcal{M}(\rho_n)$ with $\rho_n = n^{-1/4}$, where the alternative consists of functions belonging to the cone set of the Hölder class with some normalization. This rate coincides with the fastest possible rate achievable by any tests under the local alternative setting when the alternative is constructed by a non-smooth function and/or the dimension of instrument is large. Since the non-smooth function belongs to $\mathcal{M}(\rho_n)$, our test is preferable in a large dimension setting. Indeed, our simulation results show that the rejection probability of our test against a misspecified model with many instruments ($l = 27$) approaches one reasonably fast with the sample size. Moreover, we observe that the test works well in the empirical application to the return of education investigated in Angrist and Krueger (1991), even though their model includes a total of 240 IVs.

Although the literature for estimation and inference of parameters in linear IV regression models with many or many weak instruments is recently growing (see, e.g., Andrews & Stock, 2007, Newey & Windmeijer, 2009, Anatolyev & Gospodinov, 2011, Lee & Okui, 2012, Chao, Hausman, Newey, Swanson, & Woutersen, 2014, and references therein), specification testing and its rate optimality are not sufficiently investigated. A possible extension of this study, on which we are currently working, is rate optimal specification testing under many weak instruments.

References

Abramovich, F., Feis, D., Italia, S., & Theofanis. (2009). Optimal testing for additivity in multiple nonparametric regression. *Annals of the Institute of Statistical Mathematics*, *61(3)*, 691–714.

Aït-Sahalia, Y., Bickel, P. J., & Stoker, T. M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, *105(2)*, 363–412.

Anatolyev, S., & Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, *27*, 427–441.

Andrews, D. W., & Stock, J. H. (2007). Testing with many weak instruments. *Journal of Econometrics*, *138*, 24–46.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106(4)*, 979–1014.

Battistin, E., & De Nadai, M. (2015). Identification and estimation of engel curves with endogenous and unobserved expenditures. *Journal of Applied Econometrics*, *30(3)*, 487–508.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20(1)*, 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica*, *58(6)*, 1443–1458.

Bierens, H. J., & Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica*, *65(5)*, 1129–1151.

Billingsley, P. (2012). *Probability and measure, anniversary edition*. Wiley.

Carrasco, M., & Florens, J.-P. (2000). Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, *16(6)*, 797–834.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R., & Woutersen, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics*, *178*, 15–21.

Chetverikov, D., & Wilhelm, D. (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, *85(4)*, 1303–1320.

Cohen, A., Daubechies, I., & Vial, P. (1993). Wavelets and fast wavelet transform on the interval. *Applied and Computational Harmonic Analysis*, *1*, 54–81.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, *41(7)*, 909–996.

Deaton, A., & Muellbauer, J. (1980). An almost ideal demand system. *The American economic review*, *70(3)*, 312–326.

Dominguez, M. A., & Lobato, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, *72(5)*, 1601–1615.

Donald, S. G., Imbens, G. W., & Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, *117(1)*, 55–93.

Ermakov, M. S. (1991). Minimax detection of a signal in a Gaussian white noise. *Theory of Probability & Its Applications*, *35(4)*, 667–679.

Fan, Y., & Li, Q. (2000). Consistent model specification tests. *Econometric Theory*, *16(06)*, 1016–1041.

Gasser, T., Sroka, L., & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, *73(3)*, 625–633.

Gørgens, T., & Würtz, A. (2012). Testing a parametric function against a non-parametric alternative in IV and GMM settings. *The Econometrics Journal*, *15(3)*, 462–489.

Guerre, E., & Lavergne, P. (2002). Optimal minimax rates for nonparametric specification testing in regression models. *Econometric Theory*, *18(5)*, 1139–1171.

Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method

of moments estimator in misspecified models. *Journal of Econometrics*, *114(2)*, 361–394.

Hall, P., & Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press.

Hall, P., Kay, J. W., & Titterinton, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, *77(3)*, 521–528.

Holzmann, H. (2008). Testing parametric models in the presence of instrumental variables. *Statistics & Probability Letters*, *78(6)*, 629–636.

Horowitz, J. L. (2006). Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica*, *74(2)*, 521–538.

Horowitz, J. L., & Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, *69(3)*, 599–631.

Hu, Y., & Shiu, J.-L. (2017). Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Econometric Theory*, *69*, 1–35.

Imbens, G., Spady, R. H., & Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, *66(2)*, 333–357.

Ingster, Y., & Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics 169. Springer-Verlag New York.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Mathematical Methods of Statistics*, *2(2)*, 85–114.

Ingster, Y. I., & Sapatinas, T. (2009). Minimax goodness-of-fit testing in multivariate nonparametric regression. *Mathematical Methods of Statistics*, *18(3)*, 241–269.

Jorgenson, D. W., Lau, L. J., & Stoker, T. M. (1982). The transcendental logarithmic model of aggregate consumer behavior. In R. Basmann & G. Rhodes (Eds.), *Advances in econometrics vol. 1* (pp. 97–238). Greenwich: JAI Press.

Jun, S. J., & Pinkse, J. (2009). Semiparametric tests of conditional moment restrictions under weak or partial identification. *Journal of Econometrics*, *152(1)*, 3–18.

Jun, S. J., & Pinkse, J. (2012). Testing under weak identification with conditional moment restrictions. *Econometric Theory*, *28(6)*, 1229–1282.

Kitamura, Y., & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, *65(4)*, 861–874.

Kitamura, Y., Tripathi, G., & Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, *72(6)*, 1667–1714.

Lee, Y., & Okui, R. (2012). Hahn–hausman test as a specification test. *Journal of Econometrics*, *167*, 133–139.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. New York, USA: Springer.

Lepski, O. V., & Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, *5(2)*, 333–358.

Lepski, O. V., & Tsybakov, A. (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*,

117(1), 17–48.

Lewbel, A. (1996). Demand estimation with expenditure measurement errors on the left and right hand side. *The Review of Economics and Statistics*, *78(4)*, 718–725.

Li, H., Li, Q., & Liu, R. (2016). Consistent model specification tests based on k-nearest-neighbor estimation method. *Journal of Econometrics*, *194(1)*, 187–202.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice.* Princeton University Press.

Mack, Y.-P., & Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, *9(1)*, 1–15.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge university press.

Muellbauer, J. (1976). Community preferences and the representative consumer. *Econometrica*, *44(5)*, 979–999.

Munk, A., Bissantz, N., Wagner, T., & Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67(1)*, 19–41.

Newey, W. K., & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, *71(5)*, 1565–1578.

Newey, W. K., & Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, *77*, 687–719.

Smith, R. J. (2007). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*, *138(2)*, 430–460.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, *24(6)*, 2477–2498.

Tripathi, G., & Kitamura, Y. (2003). Testing conditional moment restrictions. *The Annals of Statistics*, *31(6)*, 2059–2095.

Von Neumann, J., Kent, R. H., Bellinson, H. R., & Hart, B. I. (1941). The mean square successive difference. *The Annals of Mathematical Statistics*, *12(2)*, 153–162.

Yatchew, A. J. (1988). Dynamic econometric modeling, proceedings of the third international symposium in economic theory and econometrics. In W. A. Barnett, E. R. Berndt, & H. White (Eds.), (pp. 121–135).

APPENDIX A

Proof of Proposition 1

Proof of Proposition 1. We define a indicator function $K_{i,j}$ that takes 1 if the observation j is the nearest neighbor of observation i. Formally, $K_{i,j} = \mathbb{1}(||Z_i - Z_j|| \leq ||Z_i - Z_{i^*}||)$ for $i \neq j$ and $K_{i,j} = 0$ for $i = j$. Without loss of generality, we assume that nearest neighbors are uniquely determined.[10] The frequencies that an observation is assigned to be nearest neighbors are finite because of the boundedness of the kissing number. The boundedness also holds when we use k-nearest neighbor for any fixed $k > 1$ instead of the first nearest neighbor. These features of the boundedness are crucial for the derivation of asymptotic properties of our test statistic.

Under the null hypothesis, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}_i\hat{u}_{i^*} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g(X_i,\hat{\theta})][Y_{i^*} - g(X_{i^*},\hat{\theta})] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(X_i,\theta_0) - g(X_i,\hat{\theta}) + u_i][g(X_{i^*},\theta_0) - g(X_{i^*},\hat{\theta}) + u_{i^*}] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(X_i,\theta_0) - g(X_i,\hat{\theta})][g(X_{i^*},\theta_0) - g(X_{i^*},\hat{\theta})] \quad : A_1 \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(X_i,\theta_0) - g(X_i,\hat{\theta})]u_{i^*} \quad : A_2 \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(X_{i^*},\theta_0) - g(X_{i^*},\hat{\theta})]u_i \quad : A_3 \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_iu_{i^*}. \quad : A_4
\end{aligned}$$

It can be straightforwardly shown that A_1, A_2, and A_3 are $o_p(1)$. Under Assumptions 1, 2, 3, and 6, we have $A_1 = o_p(1)$. Intuitively, it comes from the $\sqrt{n}$-consistency of $\hat{\theta}$, smoothness and finite moment assumption imposed on $g(\cdot)$, and the boundedness of the number of nearest neighbors, $\sum_{i\neq j}^{n} K_{i,j} \leq \infty$. We can apply Lemma B6 of Jun and Pinkse (2012) to show the convergences of A_2 and A_3 in probability under Assumptions 1, 2, 4, 5, and 6. The results for A_3 can be shown by replacing Assumption 5 with Assumption 3 by dividing them into martingale difference sequences with vanishing variances. We show it in the followings.

Lemma 1. *Under Assumptions 1, 2, 3, 4, and 6, we have $A_3 = o_p(1)$.*

[10] Nearest neighbors are uniquely determined when the density of Z is assumed to be continuous. When Z is discrete, some observations may take exactly the same value and both of them could be assigned to be the nearest neighbors of a observation. In this case, one can just arbitrary choose one of them to determine unique neighbors.

Proof. From the mean value theorem, we obtain

$$A_3 = \sqrt{n}(\hat{\theta} - \theta_0)\frac{1}{n}\sum_{i=1}^{n}\underline{\mu}_i + \sqrt{n}(\hat{\theta} - \theta_0)\frac{1}{n}\sum_{i=1}^{n}\overline{\mu}_i + \sqrt{n}(\hat{\theta} - \theta_0)\mu_n\sqrt{n}(\tilde{\theta} - \theta_0),$$

where $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$, $\underline{\mu}_i \equiv \sum_{j<i} K_{i,j}\frac{\partial}{\partial\theta}g(X_j, \theta_0)u_i$, $\overline{\mu}_i \equiv \sum_{j>i} K_{i,j}\frac{\partial}{\partial\theta}g(X_j, \theta_0)u_i$, and $\mu_n \equiv \frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_{i,j}\frac{\partial}{\partial\theta\partial\theta}g(X_j, \theta)|_{\theta=\tilde{\tilde{\theta}}}u_i$ for a interior point $\tilde{\tilde{\theta}}$ between $\hat{\theta}$ and θ_0. Note that $\underline{\mu}_i$ and $\overline{\mu}_i$ are martingale difference sequences with respect to σ-fields generated by $\{X_1, X_2, \ldots, X_i, Z_1, Z_2, \ldots, Z_n\}$, and $\{X_i, X_{i+1}, \ldots, X_n, Z_1, Z_2, \ldots, Z_n\}$, respectively. The variances of $\frac{1}{n}\sum_{i=1}^{n}\underline{\mu}_i$ and $\frac{1}{n}\sum_{j=1}^{n}\overline{\mu}_i$ can be straightforwardly shown to be $O(1/n)$. Thus, $\frac{1}{n}\sum_{j=1}^{n}\underline{\mu}_i \xrightarrow{p} 0$ and $\frac{1}{n}\sum_{j=1}^{n}\overline{\mu}_i \xrightarrow{p} 0$ from the Chebyshev's inequality. We can also show $\mu_n = o_p(1)$ by using the bounded second moments for u_j and $\frac{\partial}{\partial\theta\partial\theta}g(X_j, \theta_0)$. □

In Lemma 2, we apply martingale central limit theorem (C.L.T.) to A_4.

Lemma 2. *Under Assumption 1, we have* $A_4 \xrightarrow{d} N(0, \mu^2)$*, where* μ^2 *is asymptotic variance of* A_4*.*

Proof. A_4 is represented as follows:

$$A_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i u_{i^*} = \frac{1}{\sqrt{n}}\sum_{i=2}^{n}\sum_{j=1}^{i-1} W_{i,j}u_i u_j = \sum_{i=2}^{n}\epsilon_{n,i},$$

where $W_{i,j} \equiv K_{i,j} + K_{j,i}$ and $\epsilon_{n,i} \equiv \frac{1}{\sqrt{n}}\sum_{j=1}^{i-1} W_{i,j}u_i u_j$. Let $\mathcal{F}_{n,i}$ be a σ-field generated by $\{Y_1, Y_2, \ldots, Y_i, X_1, X_2, \ldots, X_i, Z_1, Z_2, \ldots, Z_n\}$. It is obvious that $\mathcal{F}_{n,i}$ form a filtration, that is, $\mathcal{F}_{n,k} \subset \mathcal{F}_{n,k+1}$ holds, and $\epsilon_{n,i}$ is a martingale difference with respect to $\mathcal{F}_{n,i}$. We employ C.L.T. to prove the asymptotic normality of $\sum_{i=2}^{n}\epsilon_{n,i}$. According to Theorem 35.12 of Billingsley (2012),

$$\sum_{i=2}^{n}\epsilon_{n,i} \xrightarrow{d} N(0, \mu^2),$$

if

$$\mu^2 \equiv \lim_{n\to\infty}\sum_{i=2}^{n} E[\epsilon_{n,i}^2|\mathcal{F}_{n,i-1}] < \infty, \tag{A.1}$$

and

$$\lim_{n\to\infty}\sum_{i=1}^{n} E[\epsilon_{n,i}^2 \mathbb{1}\{|\epsilon_{n,i}| \geq \epsilon\}] = 0 \text{ for each } \epsilon. \tag{A.2}$$

Note that $W_{i,j_1}W_{i,j_2} = 0$ if $j_1 \neq j_2$. Then, (A.1) can be shown as follows:

$$\mu^2 = \lim_{n\to\infty}\sum_{i=2}^{n} E(\epsilon_{n,i}^2|\mathcal{F}_{n,i-1}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} W_{i,j}^2 u_j^2\sigma^2(Z_i) < \infty, \tag{A.3}$$

because $E(u_i^2|\mathcal{F}_{n,i-1}) = \sigma^2(Z_i) < \infty$ by Assumption 1 and $\sum_{i=1}^n W_{i,j}^2 \leq \infty$. We rewrite the condition (A.2) as

$$\sum_{i=1}^{n} E[\epsilon_{n,i}^2 \mathbb{1}\{|\epsilon_{n,i}| \geq \epsilon\}] \leq n \sup_i E\left[\frac{|\epsilon_{n,i}|^3}{|\epsilon_{n,i}|}\mathbb{1}\{|\epsilon_{n,i}| \geq \epsilon\}\right] \leq \frac{n}{\epsilon} E\left[|\epsilon_{n,i}|^3\right].$$

By using the Hölder's inequality. the third absolute moment of $\epsilon_{n,i}$ is

$$E[|\epsilon_{n,i}|^3] \leq \frac{M}{(\sqrt{n})^3} E\left[\left|\sum_{j=1}^{i-1} W_{i,j} u_j\right|^3\right] \leq \frac{M}{(\sqrt{n})^3}\left(E\left\{E\left[\left(\sum_{j=1}^{i-1} W_{i,j} u_j\right)^4 \middle| Z\right]\right\}\right)^{\frac{3}{4}},$$

for some constant M. The boundedness of $E(u^4|Z)$, $\sum_{j=1}^{i-1} W_{i,j}^4$, and $\sum_{j=1}^{i-1}\sum_{l\neq j}^{i-1} W_{i,j}^2 W_{i,l}^2$ implies $E[|\epsilon_{n,i}|^3] = O(n^{-3/2})$, and we obtain $\sum_{i=1}^n E[\epsilon_{n,i}^2 \mathbb{1}\{|\epsilon_{n,i}| \geq \epsilon\}] = O(n^{-1/2})$. Therefore, equations (A.1) and (A.2) hold and we yield

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i u_{i^*} \xrightarrow{d} N(0, \mu^2).$$

□

Lastly, we show $\hat{\mu}^2 \xrightarrow{p} \mu^2$, where $\hat{\mu}^2 = n^{-1}\sum_{i=1}^n (\hat{u}_i \hat{u}_{i^*})^2$, by decomposing $\hat{\mu}$ into vanishing terms and sum of martingale sequences in the following lemma.

Lemma 3. *Under 1, 2, 3, 4, 5, and 6, we have $\hat{\mu}^2 \xrightarrow{p} \mu^2$ under the null hypothesis.*

Proof. We show that $\hat{\mu}^2$ converges to μ^2 defined in equation (A.1) almost surely and μ^2 is equivalent to $\lim_{n\to\infty} \frac{1}{n}\sum_{i=2}^n \sum_{j=1}^{i-1} W_{i,j}^2 \sigma^2(Z_j)\sigma^2(Z_i)$. $\hat{\mu}^2$ is represented as follows:

$$\hat{\mu}^2 = \frac{2}{n}\sum_{i=1}^{n} [g(X_{i^*}, \theta_0) - g(X_{i^*}, \hat{\theta})] u_i^2 u_{i^*} + \frac{1}{n}\sum_{i=1}^{n} u_i^2 u_{i^*}^2 + B, \tag{A.4}$$

where B includes terms that converges to zero in probability. The convergence can be shown straightforwardly by using the $\sqrt{n}$-consistency of parameter estimates in Assumption 6, uniform convergence of the first and the second derivative of $g(x, \theta)$ with respect to $\theta \in \Theta$ under Assumptions 2, 3, and 4, and Assumptions 1 and 5 about the boundedness.

The absolute value of the first term of equation (A.4) is

$$\begin{aligned}
\frac{2}{n}\left|\sum_{i=1}^{n} [g(X_{i^*}, \theta_0) - g(X_{i^*}, \hat{\theta})] u_i^2 u_{i^*}\right| &\leq \frac{2}{n}\left|\sum_{i=1}^{n} (\hat{\theta} - \theta_0)' \frac{\partial}{\partial \theta} g(X_{i^*}, \theta_0) u_{i^*} u_i^2\right| \\
&\quad + \frac{2}{n}\left|\sum_{i=1}^{n} (\hat{\theta} - \theta_0)' \left.\frac{\partial g(X_{i^*}, \theta)}{\partial \theta \partial \theta'}\right|_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0) u_{i^*} u_i^2\right| \\
&\equiv B_1 + B_2,
\end{aligned}$$

where $\tilde{\theta} \in \Theta$ is an interior point between θ_0 and $\hat{\theta}$. We show that B_1 and B_2 are $o_p(1)$. First, B_1 is represented as follows:

$$B_1 \leq 2\|\sqrt{n}(\hat{\theta} - \theta_0)\| \left[\frac{1}{n\sqrt{n}} \sum_{i=1}^{n} \left\| \frac{\partial}{\partial \theta} g(X_{i^*}, \theta_0) \right\|^2 u_i^4 \right]^{1/2} \left(\frac{1}{n\sqrt{n}} \sum_{i=1}^{n} u_{i^*}^2 \right)^{1/2},$$

where

$$\frac{1}{n\sqrt{n}} \sum_{i=1}^{n} E\left[\left\| \frac{\partial g(X_{i^*}, \theta_0)}{\partial \theta} \right\|^2 u_i^4 \right] = \frac{1}{n\sqrt{n}} \sum_{i=1}^{n} \sum_{j \neq i} E\left[K_{i,j} \left\| \frac{\partial g(X_j, \theta_0)}{\partial \theta} \right\|^2 E(u_i^4 | Z_i) \right] \quad = o(1),$$

because $E(u_i^4|Z_i)$, $\sum_{i \neq j} K_{i,j}$, and $E[\|\frac{\partial}{\partial \theta} g(X_j, \theta_0)\|^2]$ are bounded by Assumption 1, the boundedness of the kissing number, and Assumption 3, respectively. Furthermore,

$$\frac{1}{n\sqrt{n}} \sum_{i=1}^{n} u_{i^*}^2 = \frac{1}{n\sqrt{n}} \sum_{i=1}^{n} \sum_{j \neq i} K_{i,j} u_j^2 = O(n^{-1/2}) E\left[\sigma^2(Z_j)\right] + o_p(1) = o_p(1).$$

Thus, we yield $B_1 = o_p(1)$. Second, B_2 is represented as follows:

$$B_2 \leq O_p(1) \left[\frac{1}{n^2} \sum_{i=1}^{n} \left\| \left. \frac{\partial g(X_{i^*}, \theta)}{\partial \theta \partial \theta'} \right|_{\theta = \tilde{\theta}} \right\|^2 \right]^{1/2} \left(\frac{1}{n^2} \sum_{i=1}^{n} u_i^4 u_{i^*}^2 \right)^{1/2}$$

Note that $\frac{1}{n^2} \sum_{i=1}^{n} \| \frac{\partial g(X_{i^*}, \theta)}{\partial \theta \partial \theta'} |_{\theta = \tilde{\theta}} \|^2 = o_p(1)$ by the uniform convergence under Assumptions 1, 2, 4, and 6. Furthermore,

$$\frac{1}{n^2} \sum_{i=1}^{n} u_i^4 u_{i^*}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} K_{i,j} u_i^4 u_j^2 \leq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} u_i^4 u_j^2 = C + o_p(1),$$

for some constant C because $E(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} u_i^4 u_j^2) = E(u_i^4) E(u_j^2) + o(1)$ is bounded and $var(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} u_i^4 u_j^2) = o(1)$, where we need Assumption 1 with $p = 8$ to show the vanishing variance.[11] Therefore, we yield $B_2 = o_p(1)$.

We apply Theorem 2.17 of P. Hall and Heyde (1980) to show the probability limit of the second term of equation (A.4). We have

$$\frac{1}{n} \sum_{i=1}^{n} u_i^2 u_{i^*}^2 = \frac{1}{n} \sum_{i=2}^{n} \sum_{j=1}^{i-1} W_{i,j}^2 u_i^2 u_j^2 + o_p(1) = \sum_{i=2}^{n} \nu_{n,i} + o_p(1),$$

where $\nu_{n,i} \equiv \frac{1}{n} \sum_{j=1}^{i-1} W_{i,j}^2 u_i^2 u_j^2$ is a martingale with respect to $\mathcal{F}_{n,i}$. According to Theorem 2.17 of P. Hall and Heyde (1980), $\sum_{i=2}^{n} \nu_{n,i}$ converges to $\lim_{n \to \infty} \sum_{i=2}^{n} E(\nu_{n,i} | \mathcal{F}_{n,i-1})$

[11] We can also show $B_2 = o_p(1)$ without assuming the eighth moment of the error term in Assumption 1. This can be done by showing the first and the second moment of the second term of equation (A.4) approaches to zero instead of showing the absolute value of the second term of equation (A.4) approaches to zero. For the proof, however, we need stronger dominance condition, that is, $E[\sup_{\theta \in \Theta} \| \frac{\partial}{\partial \theta} g(X_i, \theta) \|^4] < \infty$ instead of Assumption 3.

almost surely if $\lim_{n\to\infty}\sum_{i=2}^{n} E(|\nu_{n,i}||\mathcal{F}_{n,i-1}) < \infty$. Note that the condition holds because

$$\lim_{n\to\infty}\sum_{i=2}^{n} E(|\nu_{n,i}||\mathcal{F}_{n,i-1}) \leq \lim_{n\to\infty}\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(W_{i,j}^2 u_i^2 u_j^2 \middle| \mathcal{F}_{n,i-1}\right) < \infty,$$

by (A.3). Thus, $\sum_{i=2}^{n}\nu_{n,i}$ converges to $\lim_{n\to\infty}\sum_{i=2}^{n} E(\nu_{n,i}|\mathcal{F}_{n,i-1})$ almost surely, which is equivalent to μ^2 in (A.3) because

$$\sum_{i=2}^{n} E(\nu_{n,i}|\mathcal{F}_{n,i-1}) = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} W_{i,j}^2\sigma^2(Z_i)u_j^2 = \frac{1}{n}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} W_{i,j}^2\sigma^2(Z_i)u_j^2 = \sum_{j=1}^{n-1} v_{n,j}$$

where $v_{n,j} \equiv \frac{1}{n}\sum_{i=j+1}^{n} W_{i,j}^2\sigma^2(Z_i)u_j^2$.

Let $\overline{\mathcal{F}}_{n,j}$ be a σ-field generated by $\{Y_j, Y_{j+1}, \ldots, Y_n, X_j, X_{j+1}, \ldots, X_n, Z_1, Z_2, \ldots, Z_n\}$. Then, $v_{n,j}$ is a reversed martingale with respect to $\overline{\mathcal{F}}_{n,j}$. According to Theorem 2.17 of P. Hall and Heyde (1980), $\sum_{j=1}^{n-1} v_{n,j}$ converges to $\lim_{n\to\infty}\sum_{i=2}^{n} E(v_{n,j}|\overline{\mathcal{F}}_{n,j+1})$ almost surely if $\lim_{n\to\infty}\sum_{i=2}^{n} E(|v_{n,j}||\overline{\mathcal{F}}_{n,j+1}) < \infty$, which can be shown as folllows:

$$\sum_{i=2}^{n} E(|v_{n,j}||\overline{\mathcal{F}}_{n,j+1}) \leq \frac{1}{n}\sum_{i=2}^{n}\sum_{i=j+1}^{n} W_{i,j}^2\sigma^2(Z_i)\sigma^2(Z_j) < \infty.$$

Therefore, we obtain

$$\sum_{j=1}^{n-1} v_{n,j} \xrightarrow{a.s.} \mu^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n} W_{i,j}^2\sigma^2(Z_i)\sigma^2(Z_j).$$

□

Remark 10. Lemma 3 shows that $\hat{\mu}^2$ converges to $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sigma^2(Z_i)\sigma^2(Z_{i^*})$ in probability. The probability limit of $\hat{\mu}^2$ does not change when we use the second or more distant nearest neighbor observation instead of the first nearest. To avoid complexity, we assume for now that the density of Z has compact support and is continuous and $\sigma^2(z)$ is continuously differentiable with respect to Z. Then, there is an interior point z between Z_i and Z_{i^*} such that

$$\frac{1}{n}\sum_{i=1}^{n}\sigma^2(Z_i)\sigma^2(Z_{i^*}) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2(Z_i)\left[\sigma^2(Z_i) + (Z_{i^*} - Z_i)'\frac{\partial\sigma^2(z)}{\partial z}\right].$$

Since $E[\|Z_i - Z_{i_k}\|^s] = O((k/n)^{s/l})$ by Mack and Rosenblatt (1979), the second term converges to zero in probability. Thus, $\hat{\mu}^2$ is the consistent estimator of $E\{[\sigma^2(Z_i)]^2\}$, and it does not depend on the choice of the nearest observation. That is, using the $k-$nearest for any fixed k instead of the first nearest observation does not alter the limiting behavior of $\hat{\mu}^2$.

□

Proofs of Proposition 2 and Corollary 1

Proof of Proposition 2.
Let $\psi_{j,\kappa}(z) = 2^{lj/2}\Psi(2^j z - \kappa) = 2^{lj/2}\psi(2^j z_1 - \kappa_1)\cdots\psi(2^j z_l - \kappa_l)$ for some $j \in \mathbb{Z}$ and $\kappa \equiv (\kappa_1, \ldots, \kappa_l)'$ be a k-times continuously differentiable orthonormal wavelet function defined on $[0, 2p-1]^l$ for some integer p that satisfies $|\psi_{j,\kappa}(z)| \leq 2^{-jl/2}C$ for some constant C. The orthonormality implies that $E[\psi_{j,\kappa}(Z)\psi_{j',\kappa'}(Z)] = \mathbb{1}\{j = j'\}\mathbb{1}\{\kappa = \kappa'\}$, when random variable Z is assumed to be uniformly distributed, where $\mathbb{1}\{j = j'\}$ is an indicator function taking 1 if $j = j'$ and zero otherwise.

The wavelet series $\psi_{j,\kappa}(z)$ defined above can be constructed by using, for example, Daubechies's (1988) orthonormal wavelets.[12] Let $\psi_D(\cdot)$ be Daubechies's orthonormal wavelet with support on $[-p+1, p]$ for some integer $p \geq 1$. The wavelet function becomes $\psi_{j,\kappa}(z) = 2^{lj/2}\psi_D(2^j z_1 - \kappa_1)\cdots\psi_D(2^j z_l - \kappa_l)$. By defining an appropriate collection of κ for each j, the support of $\psi_{j,\kappa}(z)$ become $[0, 2p-1]^l$. Let $\mathcal{K}_j$ denote the collection of all possible distinct values for κ such that $\mathcal{K}_j = \{\kappa \in \mathbb{Z}^l : \kappa_\iota = (p-1) + c(2p-1), c = 0, 1, \ldots, 2^j - 1, \iota = 1, 2, \ldots, l\}$. Then, $\mathcal{K}_j$ includes 2^{jl} elements for each j. Then dyadic cubes, $I_{j,\kappa} \equiv \prod_{\iota=1}^{l}((-p+1+\kappa_\iota)2^{-j}, (\kappa_\iota + p)2^{-j}]$, satisfy $\cup_{\kappa\in\mathcal{K}_j} I_{j,\kappa} \subset [0, 2p-1]^l$. Since $\psi_D(2^j z_\iota - \kappa_\iota)$ is zero if z_ι lies outside of $((-p+1+\kappa_\iota)2^{-j}, (\kappa_\iota + p)2^{-j}]$, $\psi_{j,\kappa}(z)$ is zero if $z \notin I_{j,\kappa}$. Thus, the support of $\psi_{j,\kappa}(z)$ with $\kappa \in \mathcal{K}_j$ is $[0, 2p-1]^l$. Any intersection of two different cubes are always empty, i.e. $I_{j,\kappa} \cap I_{j,\kappa'} = \varnothing$ for any $\kappa, \kappa' \in \mathcal{K}_j$ $(\kappa \neq \kappa')$, which implies $\psi_{j,\kappa}(z)\psi_{j,\kappa'}(z) = 0$. Our wavelet function is orthonormal, because it is the tensor product of Daubechies's wavelets. Furthermore, Daubechies's wavelets is known to be νp times continuously differentiable, where $\nu \approx 0.2$. Thus, $\psi_{j,\kappa}(\cdot)$ can be constructed to be k-times continuously differentiable by taking p large enough and satisfy $|\psi_{j,\kappa}(z)| \leq 2^{jl/2}C$ for some constant C. This implies that we are able to construct $\psi_{j,\kappa}(z)$ so that $\Psi(z)$ belongs to $M_{L_\psi,s,k}$ for some constant L_ψ and any $s \in [0, 1]$ and integer $k \geq 0$ by taking p large enough.

Let B_κ be a sequence with $|B_\kappa| = 1$. We define for a positive constant λ that

$$\delta_{n,\theta_0}(\cdot) \equiv m_n(\cdot) - E[g(X,\theta_0)|\cdot], \qquad m_n(\cdot) = E[g(X,\theta_0)|\cdot] + \lambda\rho_n 2^{-jl/2}\sum_{\kappa\in\mathcal{K}_j} B_\kappa\psi_{j,\kappa}(\cdot).$$

Let the resolution revel of wavelets j depend on sample size so that it increase as sample size grows. Specifically, $j = \infty$ for $s + k = 0$ and $j = \lfloor -\log(\rho_n^{1/(s+k)})/\log(2)\rfloor$ otherwise, where $\lfloor z\rfloor$ is the floor functions such that $\lfloor z\rfloor = \max\{z \in \mathbb{Z} | z \leq x\}$, implying $2^{-j} = O(\rho_n^{1/(s+k)})$.

Lemma 4. *$\delta_{n,\theta_0}(Z_i)$ belongs to the class of alternative $\mathcal{M}(\rho_n)$ when $\lambda \geq 1$.*

Proof. It is enough to show (i) $\delta_{n,\theta_0}(Z_i) \in \mathcal{M}$ and (ii) $E[\delta_{n,\theta_0}(Z_i)^2] \geq \rho_n^2$.

[12] Construction of a wavelet function with support $[0, 1]$ is also possible by using, for example, the method proposed by Cohen, Daubechies, and Vial (1993).

(i) Since $\psi_{j,\kappa}(z)$ is orthonormal and $\mathcal{I}_j$ includes 2^{jl} location shifts, we obtain

$$E[\delta_{n,\theta_0}(Z_i)^2] = \lambda^2\rho_n^2 2^{-jl} E\left[\sum_{\kappa\in\mathcal{K}_j} B_\kappa^2 \psi_{j,\kappa}(Z_i)^2\right] = \lambda^2\rho_n^2 2^{-jl} \sum_{\kappa\in\mathcal{K}_j} E\left[\psi_{j,\kappa}(Z_i)^2\right]$$
$$= \rho_n^2\lambda^2. \tag{A.5}$$

Then, it is enough to show that $f_n(z) \equiv 2^{-jl/2}\sum_{\kappa\in\mathcal{K}_j} B_\kappa \psi_{j,\kappa}(z)$ belongs to $\mathcal{M}_{L,s,k}$ for some $L < \infty$, $s \in [0,1]$, and $k \geq 0$ uniformly in n.

For any $z \in I_{j,\kappa'}$, we obtain

$$\left|D^k f_n(z)\right| = \left|2^{-jl/2}\sum_{\kappa\in\mathcal{K}_j} B_\kappa 2^{jl/2} D^k\Psi(2^j z - \kappa)\right| = \left|B_{\kappa'} 2^{jk}\Psi^{(k)}(2^j z - \kappa')\right|,$$

where $\Psi^{(k)}(\cdot)$ indicates k-times partial derivative of $\Psi(\cdot)$. By using this and fact that $\Psi(z) \in M_{L_\psi,s,k}$ for some constant L_ψ and any $s \in [0,1]$ and $k \geq 0$, we obtain for any $z, y \in I_{j,\kappa'}$,

$$\begin{aligned}\left|D^k f_n(z) - D^k f_n(y)\right| &= \left|2^{jk} B_{\kappa'}[\Psi^{(k)}(2^j z - \kappa') - \Psi^{(k)}(2^j y - \kappa')]\right| \\ &\leq 2^{j(s+k)} \frac{\left|\Psi^{(k)}(2^j z - \kappa') - \Psi^{(k)}(2^j y - \kappa')\right|}{2^{js}\|z-y\|^s}\|z-y\|^s \\ &\leq 2^{j(s+k)} L_\psi \|z-y\|^s. \end{aligned} \tag{A.6}$$

When $z \in I_{j,\kappa_z}$ and $y \in I_{j,\kappa_y}$ for $\kappa_z \neq \kappa_y$,

$$\begin{aligned}&\left|D^k f_n(z) - D^k f_n(y)\right| = \left|2^{jk}[B_{\kappa_z}\Psi^{(k)}(2^j z - \kappa_z) - B_{\kappa_y}\Psi^{(k)}(2^j y - \kappa_y)]\right| \\ &\leq 2^{jk}\left|\Psi^{(k)}(2^j z - \kappa_z) - \Psi^{(k)}(2^j y - \kappa_z)\right| + 2^{jk}\left|\Psi^{(k)}(2^j z - \kappa_y) - \Psi^{(k)}(2^j z - \kappa_y)\right| \\ &\leq 2 2^{j(s+k)} L_\psi \|z-y\|^s. \end{aligned} \tag{A.7}$$

Equations (A.6), and (A.7) imply $f_n(z) \in \mathcal{M}_{2^{j(s+k)}3L_\psi,s,k}$. Focusing on a wide smoothness class with $s + k = 0$ eliminates the dependency of the class with sample size through j, so that $f_n(z) \in \mathcal{M}_{L,s,k}$ for any L larger than $3L_\psi$. This concludes the proof of $\delta_{n,\theta_0}(Z_i) \in \mathcal{M}$.

(ii) It immediately holds from equation (A.5) that $E[\delta_{n,\theta_0}(Z_i)^2] - \rho_n^2 = \rho_n^2[\lambda^2 - 1]$. Thus, $E[\delta_{n,\theta_0}(Z_i)^2] \geq \rho_n^2$ when $\lambda \geq 1$.

□

Remark 11. The fact that $f_n(z) \in \mathcal{M}_{L,0,0}$ uniformly in n implies that $\delta_{n,\theta_0}(\cdot)$ is one of the least smooth function that belongs to the set of alternative. Thus, the constructed function $\delta_{n,\theta_0}(\cdot)$ represents an element in the alternative that is difficult to detect.

In what follows we construct a Bayesian a priori measure by using the result of Lemma 4 and show even the optimal Bayesian test that has the smallest errors of testing does not have non-trivial power. Replacing the minimax problem by a Bayesian problem is standard arguments to show the lower bound of testing power (see, for example, Y. I. Ingster, 1993; Spokoiny, 1996; Lepski & Spokoiny, 1999; Lepski & Tsybakov, 2000; Guerre & Lavergne, 2002; Abramovich et al., 2009; Y. I. Ingster & Sapatinas, 2009). To prove Proposition 2, it suffices to show that

$$\sup_{\delta_\theta(Z)\in\mathcal{M}(\tilde{\rho}_n)} P(t_n \le z_\alpha) + \sup_{\delta_\theta(Z)\in H_0} P(t_n > z_\alpha) \ge 1 + o(1). \tag{A.8}$$

To give a lower bound of the left hand side of (A.8), we consider a Bayesian a priori measure over H_0 and $H_{n,1}$ by considering $\delta_\theta(\cdot)$ as a random variable defined on $H_0 \cup H_{n,1}$.

First, let Π_0 be the priori distribution defined on H_0 that has Dirac mass:

$$\Pi_0[\delta_{\theta_0}(\cdot) = 0] = \Pi_0\{m(\cdot) = E[g(X,\theta_0)|\cdot]\} = 1.$$

Second, let B_κ be an i.i.d. Rademacher random variable independent of the observations with $P(B_\kappa = 1) = P(B_\kappa = -1) = 1/2$. Let $\Pi_{n,1}$ be the priori distribution defined on $H_{n,1}$:

$$\Pi_{n,1}\left[\delta_{\theta_0}(\cdot) = \lambda\rho_n 2^{-jl/2}\sum_{\kappa\in\mathcal{K}_j} b_\kappa\psi_{j,\kappa}(\cdot)\right] = \prod_{\kappa\in\mathcal{K}_j} P\left(B_\kappa = b_\kappa\right), \quad b_\kappa \in \{-1,1\},$$

where Lemma 4 guarantees $\Pi_{n,1}$ to be an a priori measure over $H_{n,1}$. Then, $\Pi_n = \Pi_0 + \Pi_{n,1}$ is an a priori Bayesian measure over $H_0 \cup H_{n,1}$. This gives the lower bound

$$\sup_{\delta_\theta(Z)\in\mathcal{M}(\tilde{\rho}_n)} P(t_n \le z_\alpha) + \sup_{\delta_\theta(Z)\in H_0} P(t_n > z_\alpha) \ge \int P(t_n \le z_\alpha)d\Pi_{n,1} + \int P(t_n > z_\alpha)d\Pi_0.$$

The right hand side of the above equation is the Bayes error of the test t_n that is the sum of type I and type II errors of testing. It is known that the optimal Bayesian test based on the likelihood ratio has the smallest error, which we now introduce.

Let $\mathcal{Y}$ and $\mathcal{Z}$ be the set of observations Y and Z, respectively, where the joint distribution of Y and Z (specifically, the conditional mean of Y given Z) is described by $m(\cdot)$, which suggests that the relation between Y and Z depends on $\delta_\theta(\cdot)$. Then, we denote by $p_\delta(\mathcal{Y},\mathcal{Z})$ the joint density of $\mathcal{Y}$ and $\mathcal{Z}$. Average densities under the null and alternative hypotheses are $p_0(\mathcal{Y},\mathcal{Z}) \equiv \int p_\delta(\mathcal{Y},\mathcal{Z})d\Pi_0$ and $p_{n,1}(\mathcal{Y},\mathcal{Z}) \equiv \int p_\delta(\mathcal{Y},\mathcal{Z})d\Pi_{n,1}$, respectively. Let L_n denotes the likelihood ratio of the optimal Bayesian test, which is

$$L_n = \frac{p_{n,1}(\mathcal{Y},\mathcal{Z})}{p_0(\mathcal{Y},\mathcal{Z})} = \frac{\int p_\delta(\mathcal{Y}|\mathcal{Z})d\Pi_{n,1}}{\int p_\delta(\mathcal{Y}|\mathcal{Z})d\Pi_0} \equiv \frac{p_{n,1}(\mathcal{Y}|\mathcal{Z})}{p_0(\mathcal{Y}|\mathcal{Z})}.$$

By using the The Bayesian error of the optimal Bayes test (see, Theorem 13.3.1 of Lehmann & Romano, 2005, p.528), Guerre and Lavergne (2002) show that (A.8) holds if

$$\int L_n^2 p_0(\mathcal{Y}|\mathcal{Z})d\mathcal{Y} \equiv E_0(L_n^2|\mathcal{Z}) \xrightarrow{p} 1, \tag{A.9}$$

where E_0 is the expectation under p_0.

By assumption, each ω_i is standard normal conditionally upon Z_i, where $\omega_i = Y_i - m(Z_i)$. We define $\omega_{i,0} = Y_i - E[g(X_i, \theta_0)|Z_i]$. The conditional density of $\mathcal{Y}$ given $\mathcal{Z}$ under Π_0 is normal with mean $E[g(X_i, \theta_0)|Z_i]$. Since we have n observations, $\omega_i = \omega_{i,0}$ almost surely under Π_0, and $\omega_{i,0}$ is measurable with respect to Π_0 given Z_i,

$$p_0(\mathcal{Y}|\mathcal{Z}) = (2\pi)^{-n/2} \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \omega_{i,0}^2\right) d\Pi_0 = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \omega_{i,0}^2\right).$$

Since $\omega_i = Y_i - m(Z_i) = Y_i - m_n(Z_i)$ and $\delta_{\theta_0}(Z_i) = \delta_{n,\theta_0}(Z_i)$ almost surely under $\Pi_{1,n}$, we yield

$$\begin{aligned} p_{n,1}(\mathcal{Y}|\mathcal{Z}) &= (2\pi)^{-n/2} \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n} [Y_i - m_n(Z_i)]^2\right) d\Pi_{n,1} \\ &= p_0(\mathcal{Y}|\mathcal{Z}) \int \exp\left(\sum_{i=1}^{n} \omega_{i,0}\delta_{n,\theta_0}(Z_i) - \frac{1}{2}\sum_{i=1}^{n} \delta_{n,\theta_0}(Z_i)^2\right) d\Pi_{n,1}, \end{aligned}$$

where $\sum_{i=1}^{n} \omega_{i,0}\delta_{n,\theta_0}(Z_i) = \lambda\rho_n 2^{-jl/2} \sum_{i=1}^{n} \omega_{i,0} \sum_{\kappa\in\mathcal{K}_j} B_\kappa \psi_{j,\kappa}(Z_i)$ and $\sum_{i=1}^{n} \delta_{n,\theta_0}(Z_i)^2 = \lambda^2\rho_n^2 2^{-jl} \sum_{i=1}^{n} \left[\sum_{\kappa\in\mathcal{K}_j} B_\kappa \psi_{j,\kappa}(Z_i)\right]^2 = \lambda^2\rho_n^2 2^{-jl} \sum_{i=1}^{n} \sum_{\kappa\in\mathcal{K}_j} \psi_{j,\kappa}(Z_i)^2$. Thus,

$$\begin{aligned} L_n = \frac{p_{n,1}(\mathcal{Y},\mathcal{Z})}{p_0(\mathcal{Y},\mathcal{Z})} &= \int \exp\left(\sum_{i=1}^{n} \omega_i\delta_{n,\theta_0}(Z_i)\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \delta_{n,\theta_0}(Z_i)^2\right) d\Pi_{n,1} \\ &= \exp\left(-\frac{1}{2}\lambda^2\rho_n^2 2^{-jl} \sum_{i=1}^{n} \sum_{\kappa\in\mathcal{K}_j} \psi_{j,\kappa}(Z_i)^2\right) \\ &\quad \prod_{\kappa\in\mathcal{K}_j} \frac{1}{2}\left[\exp\left(\lambda\rho_n 2^{-jl/2} \sum_{i=1}^{n} \omega_{i,0}\psi_{j,\kappa}(Z_i)\right) + \exp\left(-\lambda\rho_n 2^{-jl/2} \sum_{i=1}^{n} \omega_{i,0}\psi_{j,\kappa}(Z_i)\right)\right]. \end{aligned}$$

Thus,

$$\begin{aligned} L_n^2 &= \exp\left(-\lambda^2\rho_n^2 2^{-jl} \sum_{i=1}^{n} \sum_{\kappa\in\mathcal{K}_j} \psi_{j,\kappa}(Z_i)^2\right) \\ &\quad \prod_{\kappa\in\mathcal{K}_j} \frac{1}{4}\left[\exp\left(2\lambda\rho_n 2^{-jl/2} \sum_{i=1}^{n} \omega_{i,0}\psi_{j,\kappa}(Z_i)\right) + 2 + \exp\left(-2\lambda\rho_n 2^{-jl/2} \sum_{i=1}^{n} \omega_{i,0}\psi_{j,\kappa}(Z_i)\right)\right]. \end{aligned}$$

Conditionally on $\mathcal{Z}$ and under p_0, $\{2\lambda\rho_n 2^{-jl/2}\omega_{i,0}\psi_{j,\kappa}(Z_i)\}_{i=1}^{n}$ is independent centered Gaussian with conditional variance given by $4\lambda^2\rho_n^2 2^{-jl}\psi_{j,\kappa}(Z_i)^2$. Since $E[\exp(u)] = \exp(\sigma^2/2)$ for any random variable u that follows centered gaussian with variance σ^2, we get

$$E_0(L_n^2|\mathcal{Z}) = \prod_{\kappa\in\mathcal{K}_j} \exp\left(-\lambda^2\rho_n^2 2^{-jl} \sum_{i=1}^{n} \psi_{j,\kappa}(Z_i)^2\right)$$

$$\times \frac{1}{4}\left[\exp\left(2\lambda^2\rho_n^2 2^{-jl}\sum_{i=1}^n \psi_{j,\kappa}(Z_i)^2\right) + 2 + \exp\left(2\lambda^2\rho_n^2 2^{-jl}\sum_{i=1}^n \psi_{j,\kappa}(Z_i)^2\right)\right]$$
$$= \prod_{\kappa\in\mathcal{K}_j} \cosh\left(\lambda^2\rho_n^2 2^{-jl}\sum_{i=1}^n \psi_{j,\kappa}(Z_i)^2\right),$$

where $\cosh(\cdot)$ is the hyperbolic cosine function. By using $1 \leq \cosh(z) \leq \exp(z^2)$, we obtain,[13]

$$1 \leq E_0(L_n^2|\mathcal{Z}) \leq \exp\left(\sum_{\kappa\in\mathcal{K}_j}\left[\lambda^2\rho_n^2 2^{-jl}\sum_{i=1}^n \psi_{j,\kappa}(Z_i)^2\right]^2\right)$$

Then, (A.9) holds if

$$\sum_{\kappa\in\mathcal{K}_j}\left[\lambda^2\rho_n^2 2^{-jl}\sum_{i=1}^n \psi_{j,\kappa}(Z_i)^2\right]^2 \xrightarrow{p} 0. \tag{A.10}$$

Then, to see (A.10), we consider the expectation of the right hand side positive random variable:

$$\begin{aligned}
&\sum_{\kappa\in\mathcal{K}_j} E\left[\lambda^4\rho_n^4 2^{-2jl}\sum_{i_1=1}^n\sum_{i_2=1}^n \psi_{j,\kappa}(Z_{i_1})^2\psi_{j,\kappa}(Z_{i_2})^2\right] \\
&\quad= \lambda^4\rho_n^4 2^{-2jl}\sum_{i=1}^n\sum_{\kappa\in\mathcal{K}_j} E\left[\psi_{j,\kappa}(Z_i)^4\right] + \lambda^4\rho_n^4 2^{-2jl}\sum_{i_1=1}^n\sum_{i_2\neq i_1}\sum_{\kappa\in\mathcal{K}_j} E\left[\psi_{j,\kappa}(Z_{i_1})^2\right]E\left[\psi_{j,\kappa}(Z_{i_2})^2\right] \\
&\quad\leq \lambda^4\rho_n^4 2^{-jl}\sum_{i=1}^n\sum_{\kappa\in\mathcal{K}_j} E\left[\psi_{j,\kappa}(Z_i)^2\right] + \lambda^4\rho_n^4 n(n-1)2^{-2jl}2^{jl} \\
&\quad= \lambda^4\rho_n^4 n 2^{-jl}2^{jl} + \lambda^4\rho_n^4 n(n-1)2^{-2jl}2^{jl} \\
&\quad= \lambda^4\rho_n^4 n + \lambda^4\rho_n^4 n(n-1)2^{-jl},
\end{aligned} \tag{A.11}$$

where the last term is equal to $O(\rho_n^4 n)$ because $\delta_{n,\theta_0}(\cdot)$ belongs to the set of alternative uniformly in n when $s+k=0$, which implies $j=\infty$. Thus, the last term approaches to zero because $\rho_n = o(\tilde{\rho}_n) = o(n^{-1/4})$. □

Proof of Corollary 1. Let us consider a situation in which the resolution revel of wavelets is set to grow with sample size and depend on the dimension of instruments. Especially, we set $j = (\lfloor -cl\log(\rho_n)/\log(2)\rfloor \wedge 1)$ for some positive constant c, where $\lfloor z\rfloor$ is the floor functions such that $\lfloor z\rfloor = \max\{x\in\mathbb{Z}|x\leq z\}$. Then, we have $2^{-j} = O(\rho_n^{cl})$ for sufficiently large n. Lemma 4 implies that $\delta_{n,\theta_0}(Z_i)$ belongs to $\mathcal{M}(\rho_n)$ for any constant c. Thus, for the uniform power in the optimal minimax approach, the lower bound

[13] $\cosh(x) = 2^{-1}[\exp(x)+\exp(-x)]$. On the one hand Maclaurin expansion yields $\cosh(x) = 1+x^2/2!+x^4/4!+\ldots$. On the other hand, Maclaurin expansion of $\exp(x^2/2)$ yields $\exp(x^2/2) = 1+x^2/2!+2x^4/4!+\ldots$. Therefore, we yield $\cosh(x)\leq\exp(x^2/2)\leq\exp(x^2)$.

should be determined independently with c. In contrast, when we get rid of uniform power and regard $\lambda\rho_n f_n(z)$ as a sequence of a local alternative, the lower bound of testing power can depend on it. In the local alternative setting, (A.11) turns out to be $O(\rho_n^4 n) + O(\rho_n^{4+cl^2} n^2)$. Then, we consider the following two cases:

(i) $l < 2/\sqrt{c}$: if $\rho_n = o(\tilde{\rho}_n) = o(n^{-2/(4+cl^2)})$, we have $\rho_n^4 n = o(n^{(-4+cl^2)/(4+cl^2)}) = o(1)$ and $\rho_n^{4+cl^2} n^2 = o(1)$.

(ii) $l \geq 2/\sqrt{c}$: if $\rho_n = o(\tilde{\rho}_n) = o(n^{-1/4})$, we have $\rho_n^4 n = o(1)$ and $\rho_n^{4+cl^2} n^2 = o(n^{(4-cl^2)/4}) = o(1)$.

□

Proof of Proposition 3

Proof of Proposition 3. We first consider the asymptotic behavior of $\hat{\mu}$ under $H_{n,1}$.

Lemma 5. *Let Assumptions 1, 2, 3, 4, 5, 8, 9, and 10 hold. Let*

$$\bar{\mu} \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} \sum_{j=1}^{i-1} W_{i,j}^2 E\left(u_i^{*2} \middle| Z_i\right) E\left(u_j^{*2} \middle| Z_j\right)$$

where $u_i^* \equiv Y_i - g(X_i, \theta^*)$. *Then, under* $H_{n,1}$, $\hat{\mu}^2 = \bar{\mu} + o_p(1)$ *and* $\bar{\mu}$ *is bounded from above uniformly in* $\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})$.

Proof. Under the local alternative that $\delta_\theta(Z_i) \in \mathcal{M}$, there exists a sequence $a_n < a$ that satisfies $m(Z_i) - E[g(X_i, \theta)|Z_i] \in \mathcal{M}_{a_n L,s,k} \subset \mathcal{M}_{aL,s,k}$. Together with Assumption 9, this implies $m(Z_i) \in \mathcal{M}_{(1+a)L,s,k}$. Thus, we have the following relationship: $\{\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})\} \subset \{\delta_\theta(Z) \in \mathcal{M}\} = \{m(Z) - E[g(X,\theta)|Z] \in \mathcal{M}\} \subset \{m(Z) - E[g(X,\theta)|Z] \in \mathcal{M}_{aL,s,k}\} \subset \{m(Z) \in \mathcal{M}_{(1+a)L,s,k}\}$. This implies that it suffices to show the boundedness of $\bar{\mu}$ uniformly in $m(Z) \in \mathcal{M}_{(1+a)L,s,k}$. Under the alternative, we can show

$$E(u_i^{*p}|Z_i) < \infty, \tag{A.12}$$

for $p \leq 8$ uniformly in $m(Z) \in \mathcal{M}_{(1+a)L,s,k}$. Indeed, $E(u_i^{*4}|Z_i)$ can be decomposed into $m(Z_i)^p$, $E(\omega_i^p|Z_i)$, $E[g(X,\theta)^p|z]$, and cross products of them, where $m(Z_i)^p$ is bounded uniformly under the alternative, $E(\omega_i^p|Z_i)$ is bounded by Assumption 1, and $E[g(X,\theta)^p|z]$ is bounded by Assumption 8. Thus, $\bar{\mu}$ is bounded from above uniformly in $m(Z) \in \mathcal{M}_{(1+a)L,s,k}$.

Next, we show the limiting behavior of $\hat{\mu}^2$, which is represented as follows:

$$\hat{\mu}^2 = \frac{2}{n} \sum_{i=1}^{n} [g(X_{i^*}, \theta^*) - g(X_{i^*}, \hat{\theta})] u_i^{*2} u_{i^*}^* + \frac{1}{n} \sum_{i=1}^{n} u_i^{*2} u_{i^*}^{*2} + D, \tag{A.13}$$

where D includes terms that converges to zero in probability. Equation (A.13) is a version of equation (A.4) in Lemma **??** under the alternative, and differs from (A.4) in

points that it has θ^* instead of θ_0 and error term under the pseudo true value u_i^* instead of u_i. Thus, the proof for Lemma 5 goes along with that for Lemma **??** except points which asymptotic behavior of parameter estimates under the alternative affects.

The convergence of D can be shown straightforwardly by using the $\sqrt{n}$-consistency of parameter estimates in Assumption 10, uniform convergence of the first and the second derivative of $g(x,\theta)$ with respect to $\theta \in \Theta$ under Assumptions 1, 2, 3, 4, and 10, and the boundedness in Assumptions 3 and 5. The boundedness for the conditional expectation of error terms is now guaranteed by equation (A.12) under Assumptions 8 and 9.

The absolute value of the first term of equation (A.13) is

$$\begin{aligned}\frac{2}{n}\left|\sum_{i=1}^{n}[g(X_{i^*},\theta^*)-g(X_{i^*},\hat{\theta})]u_i^{*2}u_{i^*}^*\right| &\leq \frac{2}{n}\left|\sum_{i=1}^{n}(\hat{\theta}-\theta^*)'\frac{\partial}{\partial\theta}g(X_{i^*},\theta^*)u_{i^*}^*u_i^{*2}\right| \\ &\quad + \frac{2}{n}\left|\sum_{i=1}^{n}(\hat{\theta}-\theta^*)'\left.\frac{\partial g(X_{i^*},\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}^*}(\hat{\theta}-\theta^*)u_{i^*}^*u_i^{*2}\right| \\ &\equiv D_1+D_2,\end{aligned}$$

where $\tilde{\theta}^*$ is an interior point between $\hat{\theta}$ and θ^*. We show that D_1 and D_2 are $o_p(1)$. First, D_1 is represented as follows:

$$D_1 \leq 2\|\sqrt{n}(\hat{\theta}-\theta^*)\|\left[\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\left\|\frac{\partial}{\partial\theta}g(X_{i^*},\theta^*)\right\|^2 u_i^{*4}\right]^{1/2}\left(\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}u_{i^*}^{*2}\right)^{1/2},$$

where

$$\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}E\left[\left\|\frac{\partial g(X_{i^*},\theta^*)}{\partial\theta}\right\|^2 u_i^{*4}\right] = \frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j\neq i}E\left[K_{i,j}\left\|\frac{\partial g(X_j,\theta^*)}{\partial\theta}\right\|^2 E(u_i^{*4}|Z_i)\right] \quad = o(1),$$

because $E(u_i^{*4}|Z_i)$, $\sum_{i\neq j}K_{i,j}$, and $E[\|\frac{\partial}{\partial\theta}g(X_j,\theta^*)\|^2]$ are bounded by (A.12), the boundedness of the kissing number, and Assumption 3. Furthermore,

$$\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}u_{i^*}^{*2} = \frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\sum_{j\neq i}K_{i,j}u_j^{*2} = O(n^{-1/2})E\left(u_j^{*2}\right)+o_p(1) = o_p(1),$$

Thus, we yield $D_1 = o_p(1)$. Second, D_2 is represented as follows:

$$D_2 \leq O_p(1)\left[\frac{1}{n^2}\sum_{i=1}^{n}\left\|\left.\frac{\partial g(X_{i^*},\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}^*}\right\|^2\right]^{1/2}\left(\frac{1}{n^2}\sum_{i=1}^{n}u_i^{*4}u_{i^*}^{*2}\right)^{1/2}$$

Note that $\frac{1}{n^2}\sum_{i=1}^{n}\|\frac{\partial g(X_{i^*},\theta)}{\partial\theta\partial\theta'}|_{\theta=\tilde{\theta}^*}\|^2 = o_p(1)$ by the uniform convergence under Assumptions 1, 2, 4, and 10. Furthermore,

$$\frac{1}{n^2}\sum_{i=1}^{n}u_i^{*4}u_{i^*}^{*2} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}K_{i,j}u_i^{*4}u_j^{*2} \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i^{*4}u_j^{*2} = C+o_p(1),$$

for some constant C because $E(n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i^{*4}u_j^{*2}) = E(u_i^{*4})E(u_j^{*2})+o(1)$ is bounded and $var(n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i^{*4}u_j^{*2}) = o(1)$. Therefore, we yield $D_7 = o_p(1)$.

We apply Theorem 2.17 of P. Hall and Heyde (1980) to show the probability limit of the second term of equation (A.13) . We have

$$\frac{1}{n}\sum_{i=1}^{n}u_i^{*2}u_{i^*}^{*2} = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}W_{i,j}^2u_i^{*2}u_j^{*2} + o_p(1) = \sum_{i=2}^{n}\bar{\nu}_{n,i} + o_p(1),$$

where $\bar{\nu}_{n,i} \equiv \frac{1}{n}\sum_{j=1}^{i-1}W_{i,j}^2u_i^{*2}u_j^{*2}$ is a martingale with respect to $\mathcal{F}_{n,i}$. According to Theorem 2.17 of P. Hall and Heyde (1980), $\sum_{i=2}^{n}\bar{\nu}_{n,i}$ converges to $\lim_{n\to\infty}\sum_{i=2}^{n}E(\bar{\nu}_{n,i}|\mathcal{F}_{n,i-1})$ almost surely if $\lim_{n\to\infty}\sum_{i=2}^{n}E(|\bar{\nu}_{n,i}||\mathcal{F}_{n,i-1}) < \infty$. Note that the condition holds because

$$\sum_{i=2}^{n}E(|\bar{\nu}_{n,i}||\mathcal{F}_{n,i-1}) = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}W_{i,j}^2E\left(u_i^{*2}\middle|Z_i\right)u_j^{*2} = O(1)E(u_j^{*2}) + o_p(1).$$

Thus, $\sum_{i=2}^{n}\bar{\nu}_{n,i}$ converges to $\lim_{n\to\infty}\sum_{i=2}^{n}E(\bar{\nu}_{n,i}|\mathcal{F}_{n,i-1})$ almost surely. Now we consider the limit of $\sum_{i=2}^{n}E(\bar{\nu}_{n,i}|\mathcal{F}_{n,i-1})$, which is represented as $\sum_{j=1}^{n-1}\bar{v}_{n,j}$, where $\bar{v}_{n,j} \equiv \frac{1}{n}\sum_{i=j+1}^{n}W_{i,j}^2E\left(u_i^{*2}\middle|Z_i\right)u_j^{*2}$. Note that $\bar{v}_{n,j}$ is a reversed martingale with respect to $\overline{\mathcal{F}}_{n,j}$. According to Theorem 2.17 of P. Hall and Heyde (1980), $\sum_{j=1}^{n-1}\bar{v}_{n,j}$ converges to $\lim_{n\to\infty}\sum_{i=2}^{n}E(\bar{v}_{n,j}|\overline{\mathcal{F}}_{n,j+1})$ almost surely if $\lim_{n\to\infty}\sum_{i=2}^{n}E(|\bar{v}_{n,j}||\overline{\mathcal{F}}_{n,j+1}) < \infty$, which can be shown as folllows:

$$\sum_{i=2}^{n}E(|\bar{v}_{n,j}||\overline{\mathcal{F}}_{n,j+1}) = \frac{1}{n}\sum_{i=2}^{n}\sum_{i=j+1}^{n}W_{i,j}^2E\left(u_i^{*2}\middle|Z_i\right)E\left(u_j^{*2}\middle|Z_j\right) < \infty.$$

Therefore, we obtain

$$\sum_{j=1}^{n-1}\bar{v}_{n,j} \xrightarrow{a.s.} \lim_{n\to\infty}\sum_{j=1}^{n-1}E(\bar{v}_{n,j}|\overline{\mathcal{F}}_{n,j+1}) = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}W_{i,j}^2E\left(u_i^{*2}\middle|Z_i\right)E\left(u_j^{*2}\middle|Z_j\right),$$

and we yield $\hat{\mu}^2 = \bar{\mu} + o_p(1)$, where $\bar{\mu}$ is bounded from above uniformly in $\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})$. □

Next, we consider the asymptotic behavior of the test statistics under the alternative. $\hat{\mu}T_n$ can be decomposed as follows:

$$\hat{\mu}T_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}_i\hat{u}_{i^*} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g(X_i,\hat{\theta})][Y_{i^*} - g(X_{i^*},\hat{\theta})] = T_n^* - C_1 + C_2,$$

where $T_n^* \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g(X_i,\theta^*)][Y_{i^*} - g(X_{i^*},\theta^*)]$, $C_1 \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_{\theta^*}(Z_i)[g(X_{i^*},\hat{\theta}) - g(X_{i^*},\theta^*)]$, and C_2 includes terms that consists of vanishing term $g(X_{i^*},\hat{\theta}) - g(X_{i^*},\theta^*)$ times a random variable whose expectation conditioned on instruments is zero. It is straightforward to show that $C_2 = o_p(1)$ uniformly in $\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})$. We show that $C_1 = O_p(1)$

Lemma 6. *Under Assumptions 1, 2, 4, 5, and 10, we have* $\sup_{\delta_\theta(Z)\in\mathcal{M}(\kappa n^{-1/4})} C_1 = O_p(1)$.

Proof. Since $\sqrt{n}(\hat{\theta}-\theta^*) = O_p(1)$ uniformly in $m(Z) \in \mathcal{M}_{(1+a)L,s,k}$ from Assumption 10, we have $C_1 = O_p(1)(C_1^{'} + C_1^{''})$, where $C_1^{'} \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\theta^*}(Z_i)\frac{\partial}{\partial\theta} g(X_{i^*},\theta^*)$ and $C_1^{''} = \frac{1}{n\sqrt{n}}\sum_{i=1}^n \delta_{\theta^*}(Z_i)\frac{\partial}{\partial\theta\partial\theta} g(X_{i^*},\theta)|_{\theta=\tilde{\theta}^*}$. It is obvious to show that $C_1^{''} = o_p(1)$ by using the Schwarz inequality, boundedness of $E\left[|\delta_{\theta^*}(Z_i)|^2\right]$ under $H_{n,1}$ and the uniform convergence of the second moment of the second derivative under Assumptions 1, 2, 4, and 10. Second, there is a constant $c > 0$ such that

$$E(\|C_1^{'}\|) \le \sum_{j\neq i} E\left\{K_{i,j}\,|\delta_{\theta^*}(Z_i)|\,E\left[\left\|\frac{\partial}{\partial\theta} g(X_j,\theta^*)\right\|\middle| Z\right]\right\} \le cE\left(|\delta_{\theta^*}(Z_i)|\right) < \infty.$$

From the Markov's inequality, $P(\sup\|C_1^{'}\| > c) < E(\sup|\delta_\theta(Z_i)|) < \infty$, which indicates $C_1^{'}$ is stochastically bounded. Therefore, we yield $\sup_{\delta_\theta(Z)\in\mathcal{M}(\kappa n^{-1/4})} C_1^{'} = O_p(1)$. □

There is a constant $C > 0$ such that $P(T_n \le z_\alpha) \le P\left(T_n^* \le z_\alpha^{'} + C\right) + o(1)$, where $z_\alpha^{'} \equiv \bar{\mu} z_\alpha$ is bounded uniformly by Lemma 5. Further,

$$P\left(T_n^* \le z_\alpha^{'} + C\right) = P\left(-[T_n^* - E(T_n^*)] \ge E(T_n^*) - z_\alpha^{'} - C\right) \quad \le \frac{var(T_n^*)}{\left\{E(T_n^*) - z_\alpha^{'} - C\right\}^2},$$

if $E(T_n^*) - z_\alpha^{'} - C > 0$. It is then sufficient to show that κ can be chosen so that

$$E(T_n^*) - z_\alpha^{'} - C > 0, \tag{A.14}$$

$$\frac{var(T_n^{*'})}{\left\{E(T_n^*) - z_\alpha^{'} - C\right\}^2} \le \beta, \tag{A.15}$$

uniformly in $\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})$. We cam decompose T_n^* as follows:

$$\begin{aligned} T_n^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [m(Z_i) - g(X_i,\theta^*) + \omega_i][m(Z_{i^*}) - g(X_{i^*},\theta^*) + \omega_{i^*}] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [m(Z_i) - g(X_i,\theta^*)][m(Z_{i^*}) - g(X_{i^*},\theta^*)] + \bar{T}_n^*, \end{aligned}$$

where $\bar{T}_n^*$ consists of terms that satisfies $E(\bar{T}_n^*) = 0$. Then, we have

$$E(T_n^*) = E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{j\neq i}^n K_{i,j}\delta_{\theta^*}(Z_i)\delta_{\theta^*}(Z_j)\right] = \sqrt{n}E\left[\delta_{\theta^*}(Z_i)\delta_{\theta^*}(Z_{i^*})\right].$$

Since $\delta_{\theta^*}(Z) \in \mathcal{M}$ under $H_{n,1}$, there exist a positive constant L such that $|\delta_{\theta^*}(Z_i) - \delta_{\theta^*}(Z_{i^*})| \le a_n L\|Z_i - Z_{i^*}\|^s$. Thus, we obtain

$$\left|\sqrt{n}E\left[\delta_{\theta^*}(Z_i)\delta_{\theta^*}(Z_{i^*})\right]\right| \ge \sqrt{n}E\left[\delta_{\theta^*}(Z_i)^2\right] - \sqrt{n}E[|\delta_{\theta^*}(Z_i) - \delta(Z_{i^*})|\,|\delta_{\theta^*}(Z_i)|]$$

$$\geq \sqrt{n}E\left[\delta_{\theta^*}(Z_i)^2\right]\left\{1-\frac{E[|\delta_{\theta^*}(Z_i)-\delta_{\theta^*}(Z_{i^*})|^2]^{1/2}}{E[|\delta_{\theta^*}(Z_i)|^2]^{1/2}}\right\}$$
$$\geq \sqrt{n}\kappa^2\rho_n^2\left[1-O(n^{-s/l})\right],$$

where the last equality holds because $E[\|Z_i - Z_{i^*}\|^s] = O(n^{-s/l})$ under Assumption 7 (see, for example, Lemma 14.1 of Q. Li & Racine, 2007). For large enough n that satisfies $\left[1-O(n^{-s/l})\right] > 0$, $E(T^*_{n,1})$ is always positive by taking large κ. Then, we obtain

$$\frac{E\left(T_n^*\right)-z'_\alpha-C}{|\sqrt{n}E\left[\delta_{\theta^*}(Z_i)\delta_{\theta^*}(Z_{i^*})\right]|} \geq 1-\frac{|z'_\alpha|+C}{\sqrt{n}\kappa^2\rho_n^2\left\{1-O(n^{-s/l})\right\}}.$$

Since $\rho_n^2 = n^{-1/2}$, a large value of κ makes the last term in the above equation arbitrary close to one. Therefore, (A.14) holds by taking κ large enough.

To prove (A.15), we represent T_n^* as follows:

$$T_n^* = \frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{j\neq i}K_{i,j}[Y_i-g(X_i,\theta^*)][Y_j-g(X_j,\theta^*)] \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n\eta_i,$$

where $\eta_i \equiv \sum_{j\neq i}K_{i,j}[Y_i-g(X_i,\theta^*)][Y_j-g(X_j,\theta^*)]$. Let $Z=\{Z_1,Z_2,\ldots,Z_n\}$. From the law of total variance, we obtain

$$var\left(T_n^*\right) = \frac{1}{n}var\left(\sum_{i=1}^n\eta_i\right) = \frac{1}{n}\sum_{i=1}^n E\left[var\left(\eta_i|\,Z\right)\right] + \frac{1}{n}var\left[\sum_{i=1}^n E\left(\eta_i|\,Z\right)\right],$$

where the last equality holds because η_i's are uncorrelated given Z, that is, η_i's are i.i.d conditionally upon Z. Let $\bar{\eta}_i = E[g(X_i,\theta^*)|Z_i] - g(X_i,\theta^*) + \omega_i$. Then, it is obvious that $E(\bar{\eta}_i|Z_i) = 0$ and $E(\bar{\eta}_i^2|Z_i) \leq E[g(X_i,\theta^*)|Z_i]^2 + E[g(X_i,\theta^*)^2|Z_i] + \sigma^2(Z_i)$ is bounded by Assumptions 1 and 8. By using these and boundedness of $\delta_{\theta^*}(Z_j)$, we can show that $E(\eta_i^2|Z) = \sum_{j\neq i}K_{i,j}E\left\{\left[\delta_{\theta^*}(Z_i)+\bar{\eta}_i\right]^2\middle|\,Z\right\}E\left\{\left[\delta_{\theta^*}(Z_j)+\bar{\eta}_j\right]^2\middle|\,Z\right\}$ is bounded from above by a constant Λ. Similarly, there is a constant $\bar{\Lambda}$ such that $\frac{1}{n}var[\sum_{i=1}^n E(\eta_i|Z)] \leq \bar{\Lambda}^2$. Thus, we yield $var(T_n^*) \leq \Lambda + \bar{\Lambda}^2$. For large enough n that satisfies $\left[1-O(n^{-s/l})\right] > 0$, we obtain

$$\frac{var(T_n^*)}{\left|\sqrt{n}E\left[\delta_{\theta^*}(Z_i)\delta_{\theta^*}(Z_{i^*})\right]\right|^2} \leq \frac{\Lambda+\bar{\Lambda}^2}{\left|\kappa^2\left\{1-O(n^{-s/l})\right\}\right|^2}.$$

Since this upper bound is bounded and decreasing in κ, (A.15) holds uniformly in $\delta_\theta(Z) \in \mathcal{M}(\kappa n^{-1/4})$.

□

APPENDIX B

Table B.1 shows Monte Carlo results for the power of the test. The results corresponding to those illustrated in Figure 3.

Table B.1: Monte Carlo results of power.

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
GDP1 with 3 IVs																				
$n = 100$	0.04	0.05	0.07	0.09	0.14	0.20	0.29	0.37	0.46	0.53	0.60	0.65	0.69	0.72	0.75	0.76	0.78	0.79	0.80	0.81
$n = 250$	0.04	0.06	0.11	0.22	0.39	0.58	0.74	0.87	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 500$	0.05	0.11	0.24	0.49	0.76	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.08	0.16	0.41	0.80	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP2 with 3 IVs																				
$n = 100$	0.03	0.04	0.07	0.10	0.15	0.20	0.27	0.35	0.44	0.51	0.58	0.64	0.68	0.72	0.75	0.77	0.78	0.79	0.81	0.82
$n = 250$	0.04	0.07	0.11	0.22	0.39	0.59	0.76	0.88	0.94	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 500$	0.05	0.10	0.23	0.49	0.76	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.08	0.17	0.42	0.78	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP3 with 3 IVs																				
$n = 100$	0.04	0.05	0.06	0.09	0.14	0.19	0.28	0.36	0.46	0.53	0.58	0.65	0.68	0.72	0.74	0.76	0.78	0.79	0.79	0.80
$n = 250$	0.04	0.06	0.11	0.22	0.38	0.58	0.74	0.87	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 500$	0.05	0.11	0.24	0.49	0.76	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.08	0.16	0.41	0.80	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP1 with 12 IVs																				
$n = 100$	0.03	0.03	0.04	0.05	0.05	0.06	0.08	0.10	0.11	0.12	0.13	0.15	0.15	0.16	0.17	0.18	0.19	0.19	0.19	0.19
$n = 250$	0.03	0.04	0.06	0.08	0.12	0.18	0.25	0.34	0.41	0.49	0.54	0.61	0.64	0.68	0.71	0.72	0.74	0.76	0.78	0.78
$n = 500$	0.06	0.08	0.14	0.23	0.42	0.59	0.73	0.85	0.91	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.06	0.10	0.25	0.49	0.77	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP2 with 12 IVs																				
$n = 100$	0.03	0.03	0.04	0.05	0.05	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.16	0.17	0.18	0.18	0.18	0.18	0.19
$n = 250$	0.04	0.04	0.05	0.09	0.13	0.20	0.28	0.35	0.42	0.48	0.55	0.59	0.64	0.67	0.70	0.72	0.75	0.76	0.78	0.78
$n = 500$	0.04	0.07	0.13	0.24	0.40	0.59	0.72	0.84	0.92	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.05	0.10	0.24	0.51	0.77	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP3 with 12 IVs																				
$n = 100$	0.03	0.03	0.04	0.05	0.05	0.07	0.08	0.10	0.11	0.12	0.13	0.15	0.15	0.16	0.17	0.18	0.18	0.19	0.19	0.19
$n = 250$	0.03	0.04	0.06	0.08	0.12	0.18	0.25	0.34	0.41	0.49	0.54	0.60	0.64	0.68	0.71	0.72	0.74	0.76	0.77	0.78
$n = 500$	0.06	0.08	0.14	0.23	0.41	0.59	0.73	0.85	0.91	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$n = 1000$	0.06	0.11	0.25	0.49	0.77	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP1 with 27 IVs																				
$n = 100$	0.02	0.02	0.02	0.02	0.03	0.04	0.04	0.06	0.06	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.10
$n = 250$	0.04	0.05	0.06	0.09	0.11	0.15	0.19	0.23	0.29	0.32	0.37	0.40	0.43	0.45	0.47	0.48	0.50	0.52	0.53	0.54
$n = 500$	0.05	0.07	0.11	0.17	0.28	0.43	0.57	0.69	0.79	0.85	0.90	0.93	0.95	0.96	0.97	0.97	0.98	0.98	0.98	0.98
$n = 1000$	0.05	0.09	0.22	0.42	0.67	0.88	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP2 with 27 IVs																				
$n = 100$	0.02	0.03	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.07	0.08	0.08	0.08	0.09	0.09	0.09	0.10	0.10	0.10	0.10
$n = 250$	0.04	0.04	0.06	0.08	0.11	0.15	0.19	0.24	0.29	0.33	0.38	0.41	0.43	0.45	0.47	0.49	0.50	0.51	0.52	0.53
$n = 500$	0.04	0.06	0.11	0.20	0.30	0.43	0.57	0.70	0.80	0.87	0.90	0.94	0.95	0.96	0.97	0.97	0.97	0.98	0.98	0.98
$n = 1000$	0.05	0.09	0.20	0.45	0.68	0.88	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GDP3 with 27 IVs																				
$n = 100$	0.02	0.02	0.02	0.02	0.03	0.04	0.04	0.06	0.06	0.06	0.07	0.08	0.08	0.08	0.09	0.08	0.09	0.09	0.09	0.10
$n = 250$	0.04	0.05	0.06	0.09	0.11	0.14	0.19	0.23	0.28	0.32	0.36	0.40	0.43	0.45	0.47	0.48	0.51	0.52	0.54	0.54
$n = 500$	0.05	0.07	0.11	0.18	0.28	0.43	0.57	0.69	0.79	0.85	0.90	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.98	0.98
$n = 1000$	0.05	0.09	0.22	0.42	0.68	0.88	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00