A Flexible Nonparametric Test for Conditional Independence[*]

Meng Huang[†]
Bates White, LLC

Yixiao Sun[‡]
Department of Economics
UC San Diego

Halbert White[§]
Department of Economics
UC San Diego

May 30, 2013

Abstract

This paper proposes a nonparametric test for conditional independence that is easy to implement, yet powerful in the sense that it is consistent and achieves $n^{-1/2}$ local power. The test statistic is based on an estimator of the topological "distance" between restricted and unrestricted probability measures corresponding to conditional independence or its absence. The distance is evaluated using a family of *Generically Comprehensively Revealing* (GCR) functions, such as the exponential or logistic functions, which are indexed by nuisance parameters. The use of GCR functions makes the test able to detect any deviation from the null. We use a kernel smoothing method when estimating the distance. An integrated conditional moment (ICM) test statistic based on these estimates is obtained by integrating out the nuisance parameters. We simulate the critical values using a conditional simulation approach. Monte Carlo experiments show that the test performs well in finite samples. As an application, we test the key assumption of unconfoundedness in the context of estimating the returns to schooling.

1 Introduction

In this paper, we propose a flexible nonparametric test for conditional independence. Let X, Y, and Z be three random vectors. The null hypothesis we want to test is that Y is independent of X given Z, denoted

$$Y \perp X \mid Z.$$

[*] We thank Graham Elliott, Dimitris Politis, Patrick Fitzsimmons, Jin Seo Cho, Liangjun Su, James Hamilton, Andres Santos, Brendan Beare, and seminar participants at the University of California San Diego, Hong Kong University of Science and Technology, Peking University Guanghua school of management, and Bates White, LLC for helpful comments and suggestions. Special thanks to Liangjun Su for sharing computer programs.

[†] Bates White, LLC, 3580 Carmel Mountain Rd, Suite 420, San Diego, CA 92130 (email: nkhuangmeng@gmail.com)

[‡] Department of Economics 0508, University of California, San Diego, La Jolla, CA 92093 (email: yisun@ucsd.edu)

[§] Department of Economics 0508, University of California, San Diego, La Jolla, CA 92093 (email: hwhite@ucsd.edu)

Intuitively, this means that given the information in Z, X cannot provide additional information useful in predicting Y. Dawid (1979) showed that some simple heuristic properties of conditional independence can form a conceptual framework for many important topics in statistical inference: sufficiency and ancillarity, parameter identification, causal inference, prediction sufficiency, data selection mechanisms, invariant statistical models, and a subjectivist approach to model-building.

An important application of conditional independence testing in economics is to test a key assumption identifying causal effects. Suppose we are interested in estimating the effect of X (e.g., schooling) on Y (e.g., income), and that X and Y are related by the equation

$$Y = \theta_0 + \theta_1 X + U,$$

where U (e.g., ability) is an unobserved cause of Y (income) and θ_0 and θ_1 are unknown coefficients, with θ_1 representing the effect of X on Y. (We write a linear structural equation here merely for concreteness.) Since X is typically not randomly assigned and is correlated with U (e.g., unobserved ability will affect both schooling and income), OLS will generally fail to consistently estimate θ_1. Nevertheless, if, as in Griliches and Mason (1972) and Griliches (1977), we can find a set of covariates Z (e.g., proxies for ability, such as AFQT scores) such that

$$U \perp X \mid Z, \tag{1}$$

we can estimate θ_1 consistently by various methods: covariate adjustment, matching, methods using the propensity score such as weighting and blocking, or combinations of these approaches.

Assumption (1) is a key assumption for identifying θ_1. It is called a conditional exogeneity assumption by White and Chalak (2008). It enforces the "ignorability" or "unconfoundedness" condition, also known as "selection on observables" (Barnow, Cain, and Goldberger, 1981).

Note that assumption (1) cannot be directly tested since U is unobservable. But if there are other observable covariates V satisfying certain conditions (see White and Chalak, 2010), we have

$$U \perp X \mid Z \qquad \text{implies} \qquad V \perp X \mid Z,$$

so we can test (1) by testing its implication, $V \perp X \mid Z$. Section 6 of this paper applies this test in the context of a nonparametric study of returns to schooling.

In the literature, there are many tests for conditional independence when the variables are categorical. But in economic applications it is common to condition on continuous variables, and there are only a few nonparametric tests for the continuous case. Previous work on testing conditional independence for continuous random variables includes Linton and Gozalo (1997, "LG"), Fernandes and Flores (1999, "FF"), and Delgado and Gonzalez-Manteiga (2001, "DG"). Su and White have several papers (2003, 2007, 2008, 2010, "SW") addressing this question. Although SW's tests are consistent against any deviation from the null, they are only able to detect local alternatives converging to the null at a rate slower than $n^{-1/2}$ and hence suffer from the "curse of dimensionality."

Recently, Song (2009) has proposed a distribution-free conditional independence test of two continuous random variables given a parametric single index that achieves the local $n^{-1/2}$ rate. Specifically, Song (2009) tests the hypothesis

$$Y \perp X \mid \lambda_\theta (Z),$$

where $\lambda_\theta(\cdot)$ is a scalar-valued function known up to a finite-dimensional parameter θ, which must be estimated.

A main contribution here is that our proposed test also achieves $n^{-1/2}$ local power, despite its fully nonparametric nature. In contrast to Song (2009), the conditioning variables can be multi-dimensional; and there are no parameters to estimate. The test is motivated by a series of papers on consistent specification testing by Bierens (1982, 1990), Bierens and Ploberger (1997), and Stinchcombe and White (1998, "StW"), among others. Whereas Bierens (1982, 1990) and Bierens and Ploberger (1997) construct tests essentially by comparing a restricted parametric and an unrestricted *regression* model, the test in this paper follows a suggestion of StW, basing the test on estimates of the topological distance between unrestricted and restricted *probability measures*, corresponding to conditional independence or its absence.

This distance is measured indirectly by a family of moments, which are the differences of the expectations under the null and under the alternative for a set of test functions. The chosen test functions make use of *Generically Comprehensively Revealing* (GCR) functions, such as the logistic or normal cumulative distribution functions (CDFs), and are indexed by a continuous nuisance parameter vector $\boldsymbol{\gamma}$. Under the null, all moments are zero. Under the alternative, the moments are nonzero for essentially all choices of $\boldsymbol{\gamma}$. This is in contrast with DG (2001), which employs an indicator testing function that is not generally and comprehensively revealing. By construction, the indicator function takes only the values one and zero, whereas the GCR function is more flexible and hence may better present the information.

We estimate these moments by their sample analogs, using kernel smoothing. An integrated conditional moment (ICM) test statistic based on these is obtained by integrating out the nuisance parameters. Its limiting null distribution is a functional of a mean zero Gaussian process. We simulate critical values using a conditional simulation approach suggested by Hansen (1996) in a different setting.

The plan of the paper is as follows. In Section 2, we explain the basic idea of the test and specify a family of moment conditions and their empirical counterparts. This family of moment conditions is (essentially) equivalent to the null hypothesis of conditional independence and forms a basis for the test. In Section 3, we establish stochastic approximations of the empirical moment conditions uniformly over the nuisance parameters. We derive the finite-dimensional weak convergence of the empirical moment process. We also provide bandwidth choices for practical use: a simple "plug-in" estimator of the MSE-optimal bandwidth. In Section 4, we formally introduce and analyze our ICM test statistic. In particular, we establish its asymptotic properties under the null and alternatives and provide a conditional simulation approach to simulate the critical values. In Section 5, we report some Monte Carlo results examining the size and power properties of our test and comparing its performance with that of a variety of other tests in the literature. In Section 6, we study the returns to schooling, using the proposed statistic to test the key assumption of unconfoundedness. The last section concludes and discusses directions for further research.

2 The Null Hypothesis and Testing Approach

2.1 The Null Hypothesis

Let X, Y, and Z be three random vectors, with dimensions d_X, d_Y, and d_Z, respectively. Denote $W = (X', Y', Z') \in \mathbb{R}^d$ with $d = d_X + d_Y + d_Z$. Given an IID sample $\{X_i, Y_i, Z_i\}_{i=1}^n$, we want to test the null that Y is independent of X conditional on Z, i.e.,

$$H_0 : Y \perp X \mid Z, \tag{2}$$

against the alternative that Y and X are dependent conditional on Z, i.e.,

$$H_a : Y \not\perp X \mid Z.$$

Let $F_{Y|XZ}(y \mid x, z)$ be the conditional distribution function of Y given $(X, Z) = (x, z)$ and $F_{Y|Z}(y \mid z)$ be the conditional distribution function of Y given $Z = z$. Then we can express the null as

$$F_{Y|XZ}(y \mid x, z) = F_{Y|Z}(y \mid z). \tag{3}$$

The following three expressions are equivalent to one another and to (3):

$$F_{X|YZ}(x \mid y, z) = F_{X|Z}(x \mid z), \tag{4}$$

$$F_{XY|Z}(x, y \mid z) = F_{X|Z}(x \mid z)\ F_{Y|Z}(y \mid z), \tag{5}$$

$$F_{XYZ}(x, y, z)\ F_Z(z) = F_{XZ}(x, z)\ F_{YZ}(y, z), \tag{6}$$

where we have used the standard notations for distribution functions.

Let $\Psi : \mathbb{R} \to [0, 1]$ be a one-to-one mapping with Boreal measurable inverse. Define $\Psi_Y(Y) = (\Psi(Y_1), \ldots, \Psi(Y_{d_Y}))$ and define $\Psi_X(X)$ and $\Psi_Z(Z)$ similarly. Then $Y \perp X \mid Z$ is equivalent to $\Psi_Y(Y) \perp \Psi_X(X) \mid \Psi_Z(Z)$. The equivalence holds because the sigma fields are not affected by the transformation. An example of such a transformation is the normal CDF. In practice, we may also use a linear map such as $Y_i \to [Y_i - \min(Y_i)] / [\max(Y_i) - \min(Y_i)]$ to map the data into a bounded set. So without loss of generality, we assume that $P(W \in [0, 1]^d) = 1$ throughout the rest of the paper.

2.2 An Equivalent Null Hypothesis in Moment Conditions

The approach adopted in this paper is inspired by a series of papers on consistent specification testing: Bierens (1982, 1990), Bierens and Ploberger (1997), and StW, among others. The tests in those papers are based on an infinite number of moment conditions indexed by nuisance parameters. Bierens (1990) provides a consistent test of specification of nonlinear regression models. Consider the regression function $g(x) = E(Y \mid X = x)$. Bierens tests the hypothesis that the parametric functional form, $f(x, \lambda)$, is correctly specified in the sense that $g(x) = f(x, \theta_0)$ for some $\theta_0 \in \Theta$. The test statistic is based on an estimator of a family of moments $E\left[(Y - f(X, \theta_0))e^{\boldsymbol{\gamma}' X}\right]$ indexed by a nuisance parameter vector $\boldsymbol{\gamma}$. Under the null hypothesis of correct specification, these moments are zero for all $\boldsymbol{\gamma}$. Bierens's (1990) Lemma 1 shows that the converse essentially holds, due to the properties of the exponential function, making the test capable of detecting all deviations from the null.

StW find that a broader class of functions has this property. They extend Bierens's result by replacing the exponential function in the moment conditions with any GCR function, and by extending the probability measures considered in the Bierens (1990) approach to signed measures. As stated in StW, GCR functions include non-polynomial real analytic functions, e.g., exp, logistic CDF, sine, cosine, and also some nonanalytic functions like the normal CDF or its density. Further, they point out that such specification tests are based on estimates of topological distances between a restricted model and an unrestricted model. Following this idea, we can construct a test for conditional independence based on estimates of a topological distance between unrestricted and restricted probability measures corresponding to conditional independence or its absence.

To define the GCR property formally, let $\mathcal{C}(F)$ be the set of continuous functions on a compact set $F \subset \mathbb{R}^d$, and $sp\left[H_\varphi(\Gamma)\right]$ be the span of a collection of functions $H_\varphi(\Gamma)$. We write $\tilde{w} := (1, w')'$. The definition below is the same as Definition 3.6 in StW.

Definition 1 (StW, Definition 3.6) *We say that* $H_\varphi = \{H : \mathbb{R}^d \to \mathbb{R} \mid H(w) = \varphi\left(\tilde{w}'\boldsymbol{\gamma}\right),\ \gamma \in \Gamma \subset \mathbb{R}^{1+d}\}$ *is generically comprehensively revealing if for all* Γ *with non-empty interior, the uniform closure of* $sp[H_\varphi(\Gamma)]$ *contains* $\mathcal{C}(F)$ *for every compact set* $F \subset \mathbb{R}^d$*.*

Intuitively, GCR functions are a class of functions indexed by $\boldsymbol{\gamma} \in \Gamma$ whose span comes arbitrarily close to any continuous function, regardless of the choice of Γ, as long as it has non-empty interior. When there is no confusion, we simply call φ GCR if the generated H_φ is GCR.

We now establish an equivalent hypothesis in the form of a family of moment conditions following StW. Let P be the joint distribution of the random vector W, and let Q be the joint distribution of W with $Y \perp X \mid Z$. Thus, P is an unrestricted probability measure, whereas Q is restricted. To be specific, P and Q are defined such that for any event A,

$$P(A) \equiv \int 1[(x,y,z) \in A] dF_{XYZ}(x,y,z) = \int 1[(x,y,z) \in A] dF_{XY|Z}(x,y|z) dF_Z(z) \quad (7)$$

and

$$Q(A) \equiv \int 1[(x,y,z) \in A] dF_{X|Z}(x|z) dF_{Y|Z}(y|z) dF_Z(z), \quad (8)$$

where $1[\cdot]$ is an indicator function. Since $W \in [0,1]^d$ with probability 1, the domain of the integration in the above integrals is a cube in $\mathbb{R}^d$, and is omitted for notational simplicity. We will follow the same practice hereafter.

Note that the measure P will be the same as the measure Q if and only if the null is true:

$$\begin{aligned} P(A) &= \int 1[(x,y,z) \in A] dF_{XY|Z}(x,y|z) dF_Z(z) \\ \overset{H_0}{=} \int 1[(x,y,z) &\in A] dF_{X|Z}(x|z) dF_{Y|Z}(y|z) dF_Z(z) = Q(A). \end{aligned}$$

To test the null hypothesis is thus equivalent to test whether there is any deviation of P from Q. It should be pointed out that the marginal distribution of Z is the same under P and Q regardless of whether the null is true or not.

Let E_P and E_Q be the expectation operators with respect to the measure P and the measure Q. Define

$$\Delta_\varphi(\boldsymbol{\gamma}) \equiv E_P\left[\varphi(\tilde{W}'\boldsymbol{\gamma})\right] - E_Q\left[\varphi(\tilde{W}'\boldsymbol{\gamma})\right],$$

where $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1', \gamma_2', \gamma_3')' \in \mathbb{R}^{1+d}$ is a vector of nuisance parameters, $\tilde{W} = (1, W')'$, and φ is such that the indicated expectations exist for all $\boldsymbol{\gamma}$. Under the null hypothesis, $\Delta_\varphi(\boldsymbol{\gamma})$ is obviously zero for any choice of $\boldsymbol{\gamma}$ and any choice of φ, including GCR functions. To construct a powerful test, we want $\Delta_\varphi(\boldsymbol{\gamma})$ to be nonzero under the alternative. If $\Delta_{\varphi_0}(\boldsymbol{\gamma}_0)$ is not zero under some alternative, we say that φ_0 can detect that particular alternative for the choice $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. An arbitrary function φ_0 may fail to detect some alternatives for some choices of $\boldsymbol{\gamma}$. Nevertheless, according to StW, given the boundedness of W, the properties of GCR functions imply that they can detect all possible alternatives for essentially all $\boldsymbol{\gamma} \in \Gamma \subset \mathbb{R}^{1+d}$ with Γ having non-empty interior. "Essentially all" $\boldsymbol{\gamma} \in \Gamma$ means that the set of "bad" $\boldsymbol{\gamma}$'s, i.e., the set $\{\boldsymbol{\gamma} \in \Gamma\colon \Delta_\varphi(\boldsymbol{\gamma}) = 0 \text{ and } Y \not\perp X \mid Z\}$, has Lebesgue measure zero and is not dense in Γ.

Given that any deviation of P from Q can be detected by essentially any choice of $\boldsymbol{\gamma} \in \Gamma$, testing $H_0 : Y \perp X \mid Z$ is equivalent to testing

$$H_0 : \Delta_\varphi(\boldsymbol{\gamma}) = 0 \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma \tag{9}$$

for a GCR function φ and a set Γ with non-empty interior. The alternative is $H_a : H_0$ is false.

A straightforward testing approach would be to estimate $\Delta_\varphi(\boldsymbol{\gamma})$ and to see how far the estimate is from zero. But if we proceed in that way, we encounter a nonparametric estimator $\hat{f}_Z$ of the density f_Z in the denominator of the test statistic, making the analysis of limiting distributions awkward. To avoid this technical issue, we compute the expectations of φf_Z rather than those of φ, leading to a new "distance" metric between P and Q:

$$\Delta_{\varphi f}(\boldsymbol{\gamma}) = E_P\left[\varphi(\tilde{W}'\boldsymbol{\gamma}) f_Z(Z)\right] - E_Q\left[\varphi(\tilde{W}'\boldsymbol{\gamma}) f_Z(Z)\right].$$

Using the change-of-measure technique, we have

$$\Delta_{\varphi f}(\boldsymbol{\gamma}) = C\left\{E_{P^*}\left[\varphi(\tilde{W}'\boldsymbol{\gamma})\right] - E_{Q^*}\left[\varphi(\tilde{W}'\boldsymbol{\gamma})\right]\right\},$$

where P^* and Q^* are probability measures defined according to

$$\begin{aligned} P^*(A) &= \int 1[(x,y,z) \in A] f_Z(z) dF_{XY|Z}(x,y|z) dF_Z(z)/C \\ Q^*(A) &= \int 1[(x,y,z) \in A] f_Z(z) dF_{X|Z}(x|z) dF_{Y|Z}(y|z) dF_Z(z)/C \end{aligned} \tag{10}$$

with $C = \int f_Z^2(z)\, dz$ being the normalizing constant. Under the null of $H_0 : Y \perp X \mid Z$, P^* and Q^* are the same measure, and so $\Delta_{\varphi f}(\boldsymbol{\gamma}) = 0$ for all $\boldsymbol{\gamma} \in \Gamma$. Under the alternative of $H_a : Y \not\perp X \mid Z$, P^* and Q^* are different measures. By definition, if φ is GCR, then its revealing property holds for any probability measure (see Definition 3.2 of StW). So under the alternative, we have $\Delta_{\varphi f}(\boldsymbol{\gamma}) \neq 0$ for essentially all $\boldsymbol{\gamma} \in \Gamma$. The behaviors of $\Delta_{\varphi f}(\boldsymbol{\gamma})$ under the H_0 and H_a imply that we can employ $\Delta_{\varphi f}(\boldsymbol{\gamma})$ in place of $\Delta_\varphi(\boldsymbol{\gamma})$ to perform our test.

To sum up, when φ is a GCR function, Γ has non-empty interior, and $\int f_Z^2(z)\,dz < \infty$, a null hypothesis equivalent to conditional independence is

$$H_0 : \Delta_{\varphi f}(\boldsymbol{\gamma}) = 0 \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma.$$

That is, the null hypothesis of conditional independence is equivalent to a family of moment conditions indexed by $\boldsymbol{\gamma}$. For notational simplicity, we drop the subscript and write $\Delta(\boldsymbol{\gamma}) := \Delta_{\varphi f}(\boldsymbol{\gamma})$ hereafter.

2.3 Heuristics for Rates

When the probability density functions exist, the conditional independence is equivalent to any of the following:

$$\begin{aligned} f_{Y|XZ}(y \mid x, z) &= f_{Y|Z}(y \mid z), \\ f_{X|YZ}(x \mid y, z) &= f_{X|Z}(x \mid z), \\ f_{XY|Z}(x, y \mid z) &= f_{X|Z}(x \mid z)\ f_{Y|Z}(y \mid z), \\ f_{XYZ}(x, y, z)\ f_Z(z) &= f_{XZ}(x, z)\ f_{YZ}(y, z), \end{aligned} \tag{11}$$

where the notation for density functions is self explanatory. One way to test conditional independence is to compare the densities in a given equation to see if the equality holds. For example, Su and White's (2008) test essentially compares $f_{XYZ}f_Z$ with $f_{XZ}f_{YZ}$. To do that, they estimate f_{XYZ}, f_Z, f_{XZ}, and f_{YZ} nonparametrically, so their test has power against local alternatives at a rate of only $n^{-1/2}h^{-d/4}$, the slowest rate of the four nonparametric density estimators, i.e., the rate for $\hat{f}_{XYZ}$. This rate is slower than $n^{-1/2}$ and hence reflects the "curse of dimensionality." The dimension here is $d = d_X + d_Y + d_Z$, which is at least three and could potentially be larger.

To achieve the rate $n^{-1/2}$, we do not compare the density functions directly. Instead, our family of moment conditions indirectly measures the distance between $f_{XYZ}f_Z$ and $f_{XZ}f_{YZ}$, so that for each given $\boldsymbol{\gamma}$, the test statistic is based on an estimator of an average that can achieve an $n^{-1/2}$ rate, just as a semiparametric estimator would.

To better understand the moment conditions of the equivalent null, we write

$$\Delta(\boldsymbol{\gamma}) = \int \varphi(\tilde{w}'\boldsymbol{\gamma})\ f_Z(z)\ f_{XYZ}(x, y, z)\ dxdydz - \int \varphi(\tilde{w}'\boldsymbol{\gamma}) f_{YZ}(y, z)\ f_{XZ}(x, z)\ dxdydz.$$

Instead of comparing $f_{XYZ}f_Z$ with $f_{YZ}f_{XZ}$, we now compare their integral transforms. Before the transformation, $f_{XYZ}f_Z$ and $f_{YZ}f_{XZ}$ are functions of (x, y, z), the data points, and those functions can only be estimated at a nonparametric rate slower than $n^{-1/2}$. But their integral transforms are now functions of $\boldsymbol{\gamma}$. For each $\boldsymbol{\gamma}$, the transformation is an average of the data so that semiparametric techniques could be used here to get an $n^{-1/2}$ rate. Essentially, we compare two functions by comparing their weighted averages. The two comparisons are equivalent because of the properties of the chosen test functions. That is, if we choose GCR functions for our test functions, defined on a compact index space Γ with non-empty interior, and we do not detect any difference between P^* and Q^* transforms at an arbitrary point $\boldsymbol{\gamma}$, then P^* and Q^* must agree, and as a consequence P and Q must agree. We gain robustness by integrating over many points $\boldsymbol{\gamma}$.

2.4 Empirical Moment Conditions

With some abuse of notation, we write $\varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) \equiv \varphi(x, y, z; \boldsymbol{\gamma})$. Define

$$g_{XZ}(x, z; \boldsymbol{\gamma}) = E\left[\varphi(x, Y, z; \boldsymbol{\gamma})|Z = z\right] = \int \varphi(x, Y, z; \boldsymbol{\gamma}) f_{Y|Z}(y|z)dy. \tag{12}$$

Then the moment conditions can be rewritten as

$$\Delta(\boldsymbol{\gamma}) = E\left[\varphi(X, Y, Z; \boldsymbol{\gamma}) f_Z(Z)\right] - E\left[g_{XZ}(X, Z; \boldsymbol{\gamma}) f_Z(Z)\right].$$

The first term of $\Delta(\boldsymbol{\gamma})$ is a mean of φf_Z, where φ is known and f_Z can be estimated by a kernel smoothing method. The second term is a mean of $g_{XZ} f_Z(Z)$, where the function $g_{XZ}(x, z; \boldsymbol{\gamma})$ is a conditional expectation that can be estimated by a Nadaraya-Watson estimator. Thus we can estimate $\Delta(\boldsymbol{\gamma})$ by

$$\begin{aligned}
\hat{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})\hat{f}_Z(Z_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{XZ}(X_i, Z_i; \boldsymbol{\gamma}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} K_h(Z_i - Z_j)\right] \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j=1, j\neq i}^{n}\varphi(\tilde{W}_{i,j}'\boldsymbol{\gamma})K_h(Z_i - Z_j)\right] \qquad (13) \\
&= \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\{[\varphi(\tilde{W}_i'\boldsymbol{\gamma}) - \varphi(\tilde{W}_{i,j}'\boldsymbol{\gamma})]K_h(Z_i - Z_j)\},
\end{aligned}$$

where $\tilde{W}_{i,j}'\boldsymbol{\gamma} = \gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3$ and $K_h(u)$ is a multivariate kernel function. In this paper, we follow the standard practice and use a product kernel of the form:

$$K_h(u) = \frac{1}{h^{d_u}} K\left(\frac{u_1}{h}, \ldots, \frac{u_{d_u}}{h}\right) \text{ with } K\left(u_1, \ldots, u_{d_u}\right) = \prod_{\ell=1}^{d_u} k(u_\ell),$$

where d_u is the dimension of u and $h \equiv h_n$ is the bandwidth that depends on n.

$\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ is an empirical version of $\Delta(\boldsymbol{\gamma})$. For each $\boldsymbol{\gamma} \in \Gamma$, $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ is a second order U-statistic. When $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ is regarded as a process indexed by $\boldsymbol{\gamma} \in \Gamma$, $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ is a U-process. Note that $[\varphi(\tilde{W}_i'\boldsymbol{\gamma}) - \varphi(\tilde{W}_{i,j}'\boldsymbol{\gamma})]K_h(Z_i - Z_j)$ is not symmetric in i and j. To achieve the symmetry so that the theory of U-statistics and U-processes can be applied, we rewrite $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ as

$$\hat{\Delta}_{n,h}(\boldsymbol{\gamma}) = \binom{n}{2}^{-1}\sum_{i<j}\kappa_{h,2}(W_i, W_j; \boldsymbol{\gamma}), \tag{14}$$

where

$$\begin{aligned}
\kappa_{h,2}(W_i, W_j; \boldsymbol{\gamma}) &= \frac{1}{2}\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma}) - \varphi(\tilde{W}_{i,j}'\boldsymbol{\gamma})\right]K_h(Z_i - Z_j) \\
&\quad +\frac{1}{2}\left[\varphi(\tilde{W}_j'\boldsymbol{\gamma}) - \varphi(\tilde{W}_{j,i}'\boldsymbol{\gamma})\right]K_h(Z_j - Z_i) = \kappa_{h,2}(W_j, W_i; \boldsymbol{\gamma}).
\end{aligned}$$

3 Stochastic Approximations and Finite Dimensional Convergence

3.1 Assumptions

In this subsection, we state the assumptions that are required to establish the asymptotic properties of $\hat{\Delta}_{n,h}(\gamma)$. We start with a definition, which uses the following multi-index notation: for $j = (j_1, \ldots, j_m)$ with j_ℓ being nonnegative integers, we denote $|j| = j_1 + j_2 + \cdots + j_m$, $j! = j_1! \cdots j_m!$, $u^j = u_1^j \cdots u_m^{j_m}$, and $D^j g(u) = \partial^{|j|} g(u)/\partial u_1^{j_1} \cdots \partial u_m^{j_m}$.

Definition 2 *$\mathcal{G}_\beta(\mathcal{A}, \epsilon, \rho, m)$, $\beta > 1$, is a class of functions $g_\alpha(\cdot) : \mathbb{R}^m \to \mathbb{R}$ indexed by $\alpha \in \mathcal{A}$ satisfying the following two conditions:*

(a) for each α, $g_\alpha(\cdot)$ is b times continuously differentiable, where b is the greatest integer that is smaller than β;

(b) let $Q_\alpha(u, v)$ be the Taylor series expansion of $g_\alpha(u)$ around v of order b :

$$Q_\alpha(u,v) = \sum_{j:|j|\leq b} \frac{D^j g_\alpha(v)}{j!}(u-v)^j$$

then

$$\sup_{\alpha\in\mathcal{A}} \sup_{\|u-v\|\leq\epsilon} \frac{\|g_\alpha(u) - g_\alpha(v) - Q_\alpha(u,v)\|}{\|u-v\|^\beta} \leq \rho$$

for some constants $\epsilon > 0$ and $\rho > 0$.

In the absence of the index set $\mathcal{A}$, we use $\mathcal{G}_\beta(\epsilon, \rho, m)$ to denote the class of functions. In this case, our definition is similar to Definition 2 in Robinson (1988) and Definition 2 in DG (2001). A sufficient condition for condition (b) is that the partial derivative of the b-th order is uniformly Hölder continuous:

$$\sup_{\alpha\in\mathcal{A}} \sup_{\|v-u\|\leq\epsilon} \left|D^j g_\alpha(u) - D^j g_\alpha(v)\right| \leq \|v-u\|^{\beta-b}$$

for all j such that $|j| = b$.

We are ready to present our assumptions.

Assumption 1 (IID) *(a) $\{W_i \in [0,1]^d\}_{i=1}^n$ is an IID sequence of random variables on the complete probability space $(\Omega, \mathcal{F}, P)$; (b) each element Z_ℓ of Z is supported on $[0,1]$; (c) the distribution of Z admits a density function $f_Z(z)$ with respect to the Lebesgue measure.*

Assumption 2 (Smoothness of the Densities) *(a) $f_Z(\cdot) \in \mathcal{G}_{q+1}(\epsilon, \rho, d_Z)$ for some integer $q > 0$ and some constants $\epsilon > 0$ and $\rho > 0$; (b) $D^j f_Z(\breve{z}) = 0$ for all $0 \leq |j| \leq q$ and all $\breve{z}$ on the boundary of $[0,1]^{d_Z}$; (c) the conditional distribution functions $F_{Y|Z}$, $F_{X|Z}$, and $F_{XY|Z}$ admit the respective densities $f_{Y|Z}(y|z)$, $f_{X|Z}(x|z)$, and $f_{XY|Z}(x,y|z)$ with respect to a finite counting measure, or the Lebesgue measure or their product measure; (d) as functions of z indexed by x, y, or $(x,y) \in \mathcal{A}$, $f_{X|Z}(x|z)$, $f_{Y|Z}(y|z)$ and $f_{XY|Z}(x|z)$ belong to $\mathcal{G}_{q+1}(\mathcal{A}, \epsilon, \rho, d_Z)$ with $\mathcal{A} = [0,1]^{d_X}$, $[0,1]^{d_Y}$ or $[0,1]^{d_X+d_Y}$.*

Assumption 3 (GCR) *(a) Γ is compact with non-empty interior; (b) $\varphi \in \mathcal{G}_\beta(\epsilon, \rho, 1)$.*

Assumption 4 (Kernel Function) *The univariate kernel $k(\cdot)$ is the qth order symmetric and bounded kernel $k : \mathbb{R} \to \mathbb{R}$ such that*

(a) $\int k(v)dv = 1$, $\int v^j k(v)dv = 0$ *for* $j = 1, 2, \ldots, q-1$;

(b) $k(v) = O((1 + |v|^{\xi})^{-1})$ *for some* $\xi > \left(q^2 + 2q + 2\right)$.

Assumption 5 (Bandwidth) *The bandwidth* $h = h_n$ *satisfies*

(a) $nh^{d_Z} \to \infty$ *as* $n \to \infty$;

(b) $\sqrt{n}h^q = o(1)$, *i.e.,* $h = o(n^{-1/(2q)})$ *as* $n \to \infty$.

Some discussions on the assumptions are in order. The IID condition in Assumption 1 is maintained for convenience. Analogous results hold under weaker conditions, but we leave explicit consideration of these aside. If we know the support of Z_ℓ, then a linear map, if necessary, can be used to ensure that Z_ℓ is supported on $[0, 1]$. In this case, the support condition in Assumption 1(b) is innocuous. When the support of Z_ℓ is not known, we can estimate the endpoints of the support by $\min_{i=1,\ldots,n}(Z_{\ell i})$ and $\max_{i=1,\ldots,n}(Z_{\ell i})$. Under some conditions, these estimators converge to the true endpoints at the rate of $1/n$. As a result, the estimation uncertainty has no effect on our asymptotic results.

Assumptions 2(a) and (d) are needed to control the smoothing bias. Under Assumptions 1(b) and 2(a), we have $\int f_Z^2(z)\,dz < \infty$. So it is not necessary to state the square integrability of $f_Z(z)$ as a separate assumption. In assumption 2(d), the smoothness condition is with respect to the conditioning variable Z. It does not require the marginal distributions of X and Y to be smooth. In fact, X and Y could be either discrete or continuous. In addition, from a technical point of view, we only need to assume that there exists a version of the conditional density functions satisfying Assumption 2(d).

Assumption 2(b) is a technical condition, which helps avoid the boundary bias problem, a well-known problem for density estimation at the boundary. The GCR approach of StW requires the boundedness of the random vectors, and so we have to deal with the boundary bias problem. If Assumption 2(b) does not hold, we can transform Z into $\tilde{Z} = (\Theta^{-1}(Z_1), \Theta^{-1}(Z_2), \ldots, \Theta^{-1}(Z_{d_Z}))'$, where $\Theta : [0, 1] \to [0, 1]$ is strictly increasing and $q + 1$ times continuously differentiable with inverse Θ^{-1}. Now

$$
\begin{aligned}
P\left\{\tilde{Z} < z\right\} &= P\left\{Z_1 < \Theta(z_1), \ldots, Z_{d_Z} < \Theta(z_{d_Z})\right\} \\
&= F_Z\left(\Theta(z_1), \ldots, \Theta(z_{d_Z})\right),
\end{aligned}
$$

and the density of $\tilde{Z}$ is $f_{\tilde{Z}}(z) = f_Z(\Theta(z))\,\Theta'(z_1)\ldots\Theta'(z_{d_Z})$. So if $\Theta^{(i)}(0) = \Theta^{(i)}(1) = 0$ for $i = 0, \ldots, q$, then Assumption 2(b) is satisfied for the transformed random vector $\tilde{Z}$ and we can work with $\tilde{Z}$ rather than Z. We can do so because $Y \perp X \mid Z$ if and only if $Y \perp X \mid \tilde{Z}$. An example of Θ is the CDF of a beta distribution:

$$
\Theta(v) = \frac{1}{B(q+1, q+1)} \int_0^v x^q (1-x)^q\,dx := \frac{\mathbb{B}(v, q+1, q+1)}{\mathbb{B}(1, q+1, q+1)}
$$

where $\mathbb{B}(v, q+1, q+1) = \int_0^v x^q (1-x)^q\,dx$ is the incomplete beta function.

If a kernel with compact support is used, we can remove the dominating boundary bias by normalization. See, for example, Li and Racine (2007, pp. 31). In this case, we do not need to assume $f_Z(\cdot)$ to be zero on the boundary.

From a theoretical point of view, it is necessary to reduce the boundary bias to a certain order so that $\hat{\Delta}_{n,h}(\gamma)$ is asymptotically centered at $\Delta(\gamma)$. However, if Z_i takes values in a closed subset of its support with probability close to one, the boundary effect will be small. In this case, we may skip the transformation and ignore the boundary bias in practice.

Assumption 3(a) is needed only when we attempt to establish the uniformity of some asymptotic properties over Γ. Like Assumption 2, Assumption 3(b) helps control the smoothing bias. It is satisfied by many GCR functions such as $\exp(\cdot)$, normal PDF, $\sin(\cdot)$, and $\cos(\cdot)$.

The conditions on the high order kernel in Assumption 4 are fairly standard. For example, both Robinson (1988) and DG (2001) make a similar assumption. The only difference is that Robinson (1988) and DG (2001) require that $\xi > q+1$, while we require a stronger condition that $\xi > \left(q^2+2q+2\right)$ in Assumption 4(b). The stronger condition is needed to control the boundary bias, which is absent in Robinson (1988) and DG (2001), as they assume that Z has an unbounded support. Assumption 4(b) is not restrictive. It is satisfied by typical kernels used in practice, as they are either supported on $[0,1]$ or have exponentially decaying tails.

Assumption 5(a) ensures that the degenerate U-statistic in the Hoeffding decomposition of $\hat{\Delta}_{n,h}(\gamma)$ is asymptotically negligible. Assumption 5(b) removes the dominating bias of $\hat{\Delta}_{n,h}(\gamma)$. See Lemmas 1 and 2 below. A necessary condition for Assumption 5 to hold is that $2q > d_Z$.

3.2 Stochastic Approximations

To establish the asymptotic properties of $\hat{\Delta}_{n,h}(\gamma)$, we develop some stochastic approximations, using the theory of U-statistics and U-processes pioneered by Hoeffding (1948).

Let $\kappa_{h,1}(w;\gamma) = E\kappa_{h,2}(w, W_j;\gamma)$. Using Hoeffding's H-decomposition, we can decompose $\hat{\Delta}_{n,h}(\gamma)$ as

$$\hat{\Delta}_{n,h}(\gamma) = \Delta_h(\gamma) + H_{n,h}(\gamma) + R_{n,h}(\gamma),$$

where

$$\Delta_h(\gamma) = E\kappa_{h,2}(W_j, W_i;\gamma) = E\kappa_{h,1}(W_i;\gamma) \tag{15}$$

$$H_{n,h}(\gamma) = \frac{2}{n}\sum_{i=1}^{n}\tilde{\kappa}_{h,1}(W_i;\gamma) \tag{16}$$

$$R_{n,h}(\gamma) = \binom{n}{2}^{-1}\sum_{i<j}\tilde{\kappa}_{h,2}(W_i, W_j, \gamma) \tag{17}$$

and

$$\begin{aligned}\tilde{\kappa}_{h,1}(W_i;\gamma) &= \kappa_{h,1}(W_i;\gamma) - \Delta_h(\gamma)\\ \tilde{\kappa}_{h,2}(W_i, W_j, \gamma) &= \kappa_{h,2}(W_i, W_j;\gamma) - \kappa_{h,1}(W_i;\gamma) - \kappa_{h,1}(W_j;\gamma) + \Delta_h(\gamma).\end{aligned}$$

The sum of the first two terms in the H-decomposition is known as the Hájek projection. For easy reference, we denote it as

$$\tilde{\Delta}_{n,h}(\gamma) = \Delta_h(\gamma) + H_{n,h}(\gamma). \tag{18}$$

By construction, $H_{n,h}(\boldsymbol{\gamma})$ and $R_{n,h}\left(\boldsymbol{\gamma}\right)$ are uncorrelated zero mean random variables. We show that the projection remainder $R_{n,h}(\boldsymbol{\gamma})$ is asymptotically negligible, and as a result $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ and its Hájek projection $\tilde{\Delta}_{n,h}(\boldsymbol{\gamma})$ have the same limiting distribution.

For each given $\boldsymbol{\gamma}$ and h, $R_{n,h}\left(\boldsymbol{\gamma}\right)$ is a degenerate second order U-statistic with kernel $\tilde{\kappa}_{h,2}\left(\cdot,\cdot;\boldsymbol{\gamma}\right)$. According to the theory of U-statistics (e.g., Lee, 1990), we have

$$var\left[R_{n,h}\left(\boldsymbol{\gamma}\right)\right]=\frac{2}{n(n-1)}var\left[\tilde{\kappa}_{h,2}(W_i,W_j,\boldsymbol{\gamma})\right].$$

This can also be proved directly by observing that $\tilde{\kappa}_{h,2}(W_i,W_j,\boldsymbol{\gamma})$ is uncorrelated with $\tilde{\kappa}_{h,2}(W_\ell,W_m,\boldsymbol{\gamma})$ if $(i,j)\neq(\ell,m)$.

If h were fixed, then it follows from the basic U-statistic theory that $R_{n,h}(\boldsymbol{\gamma})=o_p\left(1/\sqrt{n}\right)$ for each $\boldsymbol{\gamma}\in\Gamma$. However, in the present setting, $h\to 0$ as $n\to\infty$, so the basic U-statistic theory does not directly apply. Nevertheless, we can still show that $R_{n,h}\left(\boldsymbol{\gamma}\right)$ is still $o_p\left(n^{-1/2}\right)$ under Assumption 5(a). In fact, we can prove a stronger result, as Lemma 1 shows.

Lemma 1 *Under Assumptions 1–5(a), if* $h\to 0$ *as* $n\to\infty$*, then* $\sup_{\boldsymbol{\gamma}\in\Gamma}\sqrt{n}R_{n,h}\left(\boldsymbol{\gamma}\right)=o_p(1)$*.*

We proceed to establish a stochastic approximation of the Hájek projection $\tilde{\Delta}_{n,h}(\boldsymbol{\gamma})$. Note that both $\Delta_h(\boldsymbol{\gamma})$ and $H_{n,h}(\boldsymbol{\gamma})$ depend on h. Using a Taylor expansion, we can separate terms independent of h from those associated with h in $\Delta_h(\boldsymbol{\gamma})$ and $H_{n,h}(\boldsymbol{\gamma})$. By using a higher order kernel K and controlling the rate of h so that it shrinks fast enough, we can ensure that the terms associated with h vanish asymptotically, as in Powell, Stock, and Stoker (1989).

More specifically, we first show that $\Delta_h(\boldsymbol{\gamma})=\Delta(\boldsymbol{\gamma})+O(h^q)$, where q is the order of the kernel k. Then we show that $H_{n,h}(\boldsymbol{\gamma})=2n^{-1}\sum_{i=1}^n\left\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}+O_p(h^q)$, where

$$\begin{aligned}\kappa_1(W_i;\boldsymbol{\gamma})\ \equiv\ &\frac{1}{2}\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_Z(Z_i)\\ &-\frac{1}{2}\int\varphi(\gamma_0+X_i'\gamma_1+y'\gamma_2+Z_i'\gamma_3)\ f_{YZ}(y,Z_i)dy\\ &+\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+Z_i'\gamma_3)\ f_{XYZ}(x,y,Z_i)dxdy\\ &-\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\ f_{XZ}(x,Z_i)dx.\end{aligned}$$

Under Assumption 5(b), $\sqrt{n}h^q\to 0$, which makes both the second term of $\Delta_h(\boldsymbol{\gamma})$ and the second term of $H_{n,h}(\boldsymbol{\gamma})$ vanish asymptotically. The following lemma presents these results formally.

Lemma 2 *Let Assumptions 1–4 and 5*(b) *hold. Then*

(a) $\sqrt{n}\left[\Delta_h\left(\boldsymbol{\gamma}\right)-\Delta\left(\boldsymbol{\gamma}\right)\right]=o\left(1\right)$ *uniformly over* $\boldsymbol{\gamma}\in\Gamma$;

(b) $\sqrt{n}H_{n,h}(\boldsymbol{\gamma})=2/\sqrt{n}\sum_{i=1}^n\left\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}+o_p\left(1\right)$ *uniformly over* $\boldsymbol{\gamma}\in\Gamma$.

It follows from Lemmas 1 and 2 that

$$\begin{aligned}
& \sqrt{n}\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta(\boldsymbol{\gamma})\right] \\
= \ & \sqrt{n}H_{n,h}(\boldsymbol{\gamma})+\sqrt{n}R_{n,h}(\boldsymbol{\gamma})+\sqrt{n}\left[\Delta_h(\boldsymbol{\gamma})-\Delta(\boldsymbol{\gamma})\right] \\
= \ & \sqrt{n}H_{n,h}(\boldsymbol{\gamma})+o_p(1)=\frac{2}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}+o_p(1)
\end{aligned}$$

uniformly over $\boldsymbol{\gamma}\in\Gamma$. So $\sqrt{n}\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta(\boldsymbol{\gamma})\right]$ and $2/\sqrt{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}$ have the same limiting distribution for each $\boldsymbol{\gamma}\in\Gamma$.

3.3 Finite Dimensional Convergence

In this subsection, we view $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ as a U-process indexed by $\boldsymbol{\gamma}$ and consider its finite-dimensional convergence.

Let $\Gamma_s=\{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,...,\boldsymbol{\gamma}_s\}$ for some $s<\infty$ and $\boldsymbol{\gamma}_\ell\in\Gamma$, and define

$$\hat{\Delta}_{n,h}(\Gamma_s):=[\hat{\Delta}_{n,h}(\boldsymbol{\gamma}_1),\hat{\Delta}_{n,h}(\boldsymbol{\gamma}_2),...,\hat{\Delta}_{n,h}(\boldsymbol{\gamma}_s)]'.$$

Similarly, we define $\Delta(\Gamma_s):=[\Delta(\boldsymbol{\gamma}_1),\Delta(\boldsymbol{\gamma}_2),...,\Delta(\boldsymbol{\gamma}_s)]'$. Theorem 3 below establishes the asymptotic normality of $\sqrt{n}\left[\hat{\Delta}_{n,h}(\Gamma_s)-\Delta(\Gamma_s)\right]$.

Theorem 3 *Let Assumptions 1–5 hold. Then*

$$\sqrt{n}\left[\hat{\Delta}_{n,h}(\Gamma_s)-\Delta(\Gamma_s)\right]\xrightarrow{d}N(0,\Omega),$$

where the (ℓ,m) *element of* Ω *is*

$$\Omega(\ell,m):=\sigma_\Delta(\boldsymbol{\gamma}_\ell,\boldsymbol{\gamma}_m)=4cov\left[\kappa_1(W_i;\boldsymbol{\gamma}_\ell),\kappa_1(W_i;\boldsymbol{\gamma}_m)\right]. \tag{19}$$

If, in addition, H_0 *holds, then* $\Delta(\boldsymbol{\gamma})=0$, *and*

$$\sigma_\Delta(\boldsymbol{\gamma}_\ell,\boldsymbol{\gamma}_m)=4E\left[\Lambda(W_i;\boldsymbol{\gamma}_\ell)\Lambda(W_i;\boldsymbol{\gamma}_m)\right],$$

where

$$\begin{aligned}
\Lambda(W_i;\boldsymbol{\gamma}) \ = \ & \frac{1}{2}E\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})f_Z(Z_i)|X_i,Y_i,Z_i\right]-\frac{1}{2}E\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})f_Z(Z_i)|X_i,Z_i\right] \\
& -\frac{1}{2}E\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})f_Z(Z_i)|Y_i,Z_i\right]+\frac{1}{2}E\left[\varphi(\tilde{W}_i'\boldsymbol{\gamma})f_Z(Z_i)|Z_i\right].
\end{aligned} \tag{20}$$

Theorem 3 is of interest in its own right. For example, we can use it to construct a Wald test. There may be some power loss if s is small. When s is large enough such that Γ_s approximates Γ very well, then the power loss will be small. The idea can be motivated from the method of sieves. We do not pursue this here but refer to Huang (2009) for more discussions. Instead, we consider the ICM tests in the next section. Theorem 3 is an important first step in obtaining the asymptotic distributions of the ICM statistics.

Observe that $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ (hence $\tilde{\Delta}_{n,h}(\boldsymbol{\gamma})$) is not symmetric in X and Y, whereas the hypothesis $Y\perp X\mid Z$ is. However $\sqrt{n}[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta_h(\boldsymbol{\gamma})]$ is asymptotically equivalent to

$2/\sqrt{n}\sum_{i=1}^{n}\left[\kappa_1\left(W_i;\boldsymbol{\gamma}\right)-E\kappa_1\left(W_i;\boldsymbol{\gamma}\right)\right]$. It can be readily checked that $\kappa_1\left(W;\boldsymbol{\gamma}\right)$ is symmetric in Y and X. Alternatively, we can follow the definition of g_{XZ} in (12) and define $g_{YZ}(y,z)$, $g_Z\left(z\right)$, and $g_{XYZ}\left(x,y,z;\boldsymbol{\gamma}\right)$ as

$$\begin{aligned}
g_{YZ}(y,z;\boldsymbol{\gamma}) &= E\left[\varphi\left(X,y,z;\boldsymbol{\gamma}\right)|Z=z\right]\\
g_Z(z;\boldsymbol{\gamma}) &= E\left[\varphi\left(X,Y,z;\boldsymbol{\gamma}\right)|Z=z\right]\\
g_{XYZ}\left(x,y,z;\boldsymbol{\gamma}\right) &= E\left[\varphi\left(x,y,z;\boldsymbol{\gamma}\right)|Z=z\right]=\varphi\left(x,y,z;\boldsymbol{\gamma}\right)
\end{aligned}$$

where the last equality is tautological. Then

$$\kappa_1(W;\boldsymbol{\gamma})=\frac{1}{2}\left[g_{XYZ}\left(X,Y,Z;\boldsymbol{\gamma}\right)-g_{XZ}\left(X,Z;\boldsymbol{\gamma}\right)-g_{YZ}\left(Y,Z;\boldsymbol{\gamma}\right)+g_Z\left(Z;\boldsymbol{\gamma}\right)\right]f_Z\left(Z\right),$$

which is clearly symmetric in Y and X. If we construct another estimator, say $\breve{\Delta}_{n,h}(\boldsymbol{\gamma})$, by switching the roles of X and Y, we can show that $\breve{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ are asymptotically equivalent in the sense that $\sqrt{n}[\breve{\Delta}_{n,h}(\boldsymbol{\gamma})-\hat{\Delta}_{n,h}(\boldsymbol{\gamma})]=o_p\left(1\right)$ uniformly over $\boldsymbol{\gamma}\in\Gamma$. So there is no asymptotic gain in taking an average of $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $\breve{\Delta}_{n,h}(\boldsymbol{\gamma})$. This point is further supported by the symmetry of $\Lambda(W;\boldsymbol{\gamma})$ in X and Y.

3.4 Bandwidth Selection

Although any choice of bandwidth h satisfying Assumption 5 will deliver the asymptotic distribution in Theorem 3, in practice we need some guidance on how to select h. Ideally we should select an h that would give us the greatest power for a given size of test, but deriving that procedure would be complicated enough to justify another study. Moreover, it would only make a difference for higher order results. Thus, for the present purposes, we just provide a simple “plug-in” estimator of the MSE-minimizing bandwidth proposed by Powell and Stoker (1996).

Since the test statistic is based on $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$, which estimates $\Delta\left(\boldsymbol{\gamma}\right)$, it is appealing to choose an h that minimizes the mean squared error (MSE) of $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$. After some tedious but straightforward calculations, we get

$$\begin{aligned}
MSE\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right] &= \left(\Delta_h\left(\boldsymbol{\gamma}\right)-\Delta\left(\boldsymbol{\gamma}\right)\right)^2+var\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]\\
&= \left\{E\left[B_5(W;\boldsymbol{\gamma})\right]h^q+o(h^q)\right\}^2+var\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]\\
&= \left\{E\left[B_5(W;\boldsymbol{\gamma})\right]\right\}^2h^{2q}+o(h^{2q})+var\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]\\
&= \left\{E\left[B_5(W;\boldsymbol{\gamma})\right]\right\}^2h^{2q}+o(h^{2q})\\
&\quad +4n^{-1}var\left[\kappa_1\left(W;\boldsymbol{\gamma}\right)\right]+4n^{-1}C_0\left(\boldsymbol{\gamma}\right)h^q+o(n^{-1}h^q)\\
&\quad -4n^{-2}var\left[\kappa_1\left(W;\boldsymbol{\gamma}\right)\right]+2n^{-2}E\left[\delta\left(W;\boldsymbol{\gamma}\right)\right]h^{-d_Z}\\
&\quad +o\left(n^{-2}h^{-d_Z}\right)-2n^{-2}\Delta\left(\boldsymbol{\gamma}\right)^2+o(n^{-2}),
\end{aligned}$$

where B_5 is defined in (43) in the appendix, and $\delta\left(W;\boldsymbol{\gamma}\right)$ is defined by

$$\begin{aligned}
E\left[\left\|\kappa_{h,2}\left(W_i,W_j,\boldsymbol{\gamma}\right)\right\|^2|W_i\right] &= \delta\left(W_i;\boldsymbol{\gamma}\right)h^{-d_Z}+\delta^*\left(W_i,h;\boldsymbol{\gamma}\right),\ \text{where}\\
E\left(\left\|\delta^*\left(W_i,h;\boldsymbol{\gamma}\right)\right\|\right) &= o\left(h^{-d_Z}\right).
\end{aligned}$$

The term $4n^{-1}var\left[\kappa_1(W;\boldsymbol{\gamma})\right]-4n^{-2}var\left[\kappa_1\left(W;\boldsymbol{\gamma}\right)\right]$ does not depend on h. The term $2n^{-2}\Delta\left(\boldsymbol{\gamma}\right)^2$ must be of smaller order than $4n^{-1}C_0h^q$, and $4n^{-1}C_0h^q$ must be of smaller order than $\left\{E\left[B_5\left(W;\boldsymbol{\gamma}\right)\right]\right\}^2h^{2q}$; otherwise there would be a contradiction to Assumption 5(b). So the leading term of $MSE[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})]$ that involves h is

$$MSE_1\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]=\left\{E\left[B_5\left(W;\boldsymbol{\gamma}\right)\right]\right\}^2h^{2q}+2n^{-2}E\left[\delta\left(W;\boldsymbol{\gamma}\right)\right]h^{-d_Z}. \tag{21}$$

By minimizing $MSE_1\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]$, we obtain the optimal bandwidth

$$h^*=\left[\frac{d_Z\cdot E\left[\delta\left(W;\boldsymbol{\gamma}\right)\right]}{q\cdot\left\{E\left[B_5\left(W;\boldsymbol{\gamma}\right)\right]\right\}^2}\right]^{1/(2q+d_Z)}\cdot\left[\frac{1}{n}\right]^{2/(2q+d_Z)}. \tag{22}$$

Now Assumption 5(a) is satisfied:

$$n\left(h^*\right)^{d_Z}\asymp n^{1-2d_Z/(2q+d_Z)}\asymp n^{(2q-d_Z)/(2q+d_Z)}\to\infty,\text{ given }2q>d_Z.$$

And so is Assumption 5(b):

$$\sqrt{n}\left(h^*\right)^q\asymp n^{1/2-2q/(2q+d_Z)}\asymp n^{-(2q-d_Z)/2(2q+d_Z)}=o(1),\text{ given }2q>d_Z.$$

The optimal bandwidth depends on the unknown quantities $E\left[\delta\left(W;\boldsymbol{\gamma}\right)\right]$ and $E\left[B_5\left(W;\boldsymbol{\gamma}\right)\right]$. Here we follow the standard practice (e.g., Powell and Stoker (1996)) and use a simple plug-in estimator of h^*. Let h_0 be an initial bandwidth. Suppose $E\left[\kappa_{h,2}(W_i,W_j;\boldsymbol{\gamma})^4\right]=O(h_0^{-\eta-2d_Z})$ for some $\eta>0$, and let $\varrho=\max\left\{\eta+2d_Z,2q+d_Z\right\}$. If $h_0\to0$ and $nh_0^{\varrho}\to\infty$, then by Proposition 4.2 of Powell and Stoker (1996),

$$\hat{\delta}\equiv\hat{\delta}\left(h_0\right)=\binom{n}{2}^{-1}\sum_{i<j}h_0^{d_Z}\cdot\left[\kappa_{h_0,2}(W_i,W_j,\boldsymbol{\gamma})\right]^2\xrightarrow{p}E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right], \tag{23}$$

and

$$\begin{aligned}\hat{B}_5 &\equiv \frac{\hat{\Delta}_{n,\tau h_0}(\boldsymbol{\gamma})-\hat{\Delta}_{n,h_0}(\boldsymbol{\gamma})}{(\tau h_0)^q-h_0^q}\text{ for some }0<\tau\neq1 \\ &\xrightarrow{p} E\left[B_5\left(W;\boldsymbol{\gamma}\right)\right].\end{aligned} \tag{24}$$

The estimator $\hat{B}_5$ given above is a "slope" between two points $(h_0^q,\hat{\Delta}_{n,h_0}(\boldsymbol{\gamma}))$ and $(\tau h_0^q,\hat{\Delta}_{n,\tau h_0}(\boldsymbol{\gamma}))$. To get a more stable estimator, we could use a regression of $\hat{\Delta}_{n,h_0}(\boldsymbol{\gamma})$ on h_0^q for various values of h_0. Given $\hat{\delta}$ and $\hat{B}_5$, the plug-in estimator of h^* is

$$\hat{h}=\left[\frac{d_Z\cdot\hat{\delta}}{q\cdot\hat{B}_5^2}\right]^{1/(2q+d_Z)}\cdot\left[\frac{1}{n}\right]^{2/(2q+d_Z)}. \tag{25}$$

In practice we can choose q large enough so that $\varrho=\max\{\eta+2d_Z,\,2q+d_Z\}=2q+d_Z$; then we can choose the initial bandwidth to be $h_0=o\left(n^{-1/(2q+d_Z)}\right)$. The data driven $\hat{h}$ depends on $\boldsymbol{\gamma}$. We may choose different bandwidths for different $\boldsymbol{\gamma}$'s. This is what we follow in our Monte Carlo experiments.

Powell and Stoker (1996) mention one technical proviso: $\hat{\Delta}_n(\boldsymbol{\gamma};\hat{h})$ is not guaranteed to be asymptotically equivalent to $\hat{\Delta}_n(\boldsymbol{\gamma};h^*)$ since the MSE calculations are based on the assumption that h is deterministic. The suggested solution is to discretize the set of possible scaling constants, replacing $\hat{h}$ with the closest value, $\hat{h}^\dagger$, in some finite set. The estimation uncertainty in $\hat{h}^\dagger$ is small enough that it will not affect the asymptotic MSE.

4 An Integrated Conditional Moment Test

In this section, we "integrate out" $\boldsymbol{\gamma}$ to get an integrated conditional moment (ICM) type test statistic, following Bierens (1990) and StW (1998).

4.1 The Test Statistic

If φ is GCR, testing $H_0 : Y \perp X \mid Z$ is equivalent to testing $H_0 : \Delta(\boldsymbol{\gamma}) = 0$ for essentially all $\boldsymbol{\gamma} \in \Gamma$. In other words, if we view $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ as a random function in $\boldsymbol{\gamma}$, we are testing whether its mean function $\Delta(\boldsymbol{\gamma})$ is zero on Γ. If Γ is compact, we can show that $\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ converges to a zero mean Gaussian process under the null. Based on $\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$, we construct the ICM test statistic

$$M_n = n \int_{\boldsymbol{\Gamma}} \left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}),$$

where μ is a probability measure on Γ that is absolutely continuous with respect to the Lebesgue measure on Γ. Here we integrate $[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})]^2$, which gives a Cramer-von Mises (CM) type test. Alternatively, we could integrate $|\hat{\Delta}_{n,h}(\boldsymbol{\gamma})|^p, 1 \leq p \leq \infty$. The choice $p = \infty$ (which gives the maximum over Γ) yields a Kolmogorov-Smirnov (KS) type test. We work with $p = 2$ for concreteness and because CM-type tests often outperform KS-type tests. As Boning and Sowell (1999) show, choosing μ to be the uniform density has a certain optimality property in a closely related context.

4.2 Asymptotic Distribution of the Test Statistic

To establish the weak convergence of M_n, we first show that $\sqrt{n}\left[\hat{\Delta}_{n,h}(\cdot) - \Delta(\cdot)\right]$ converges to a Gaussian process. Define

$$\zeta_n(\boldsymbol{\gamma}) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \left\{\kappa_1(W_i; \boldsymbol{\gamma}) - E\left[\kappa_1(W_i; \boldsymbol{\gamma})\right]\right\}.$$

Then Lemmas 1 and 2 imply that

$$\sup_{\boldsymbol{\gamma} \in \Gamma} \left| \sqrt{n}\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta(\boldsymbol{\gamma})\right] - \zeta_n(\boldsymbol{\gamma}) \right| = o_p(1).$$

Theorem 4 shows that $\zeta_n(\cdot)$ converges to a zero mean Gaussian process and so does $\sqrt{n}\left[\hat{\Delta}_{n,h}(\cdot) - \Delta(\cdot)\right]$.

Theorem 4 *Let Assumptions 1–5 hold. Then*

(a) $\zeta_n(\cdot) \xrightarrow{d} \mathcal{Z}(\cdot)$;

(b) $\sqrt{n}\left[\hat{\Delta}_{n,h}(\cdot) - \Delta(\cdot)\right] \xrightarrow{d} \mathcal{Z}(\cdot)$, *where* $\mathcal{Z}$ *is a zero mean Gaussian process on* Γ *with covariance function*

$$cov\left(\mathcal{Z}(\boldsymbol{\gamma}_1), \mathcal{Z}(\boldsymbol{\gamma}_2)\right) = 4 cov\left[\ \kappa_1(W; \boldsymbol{\gamma}_1),\ \kappa_1(W; \boldsymbol{\gamma}_2)\right] \equiv \sigma_\Delta(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2). \quad (26)$$

If H_0 *also holds, then*

$$T_n(\cdot) \equiv \sqrt{n}\hat{\Delta}_{n,h}(\cdot) \xrightarrow{d} \mathcal{Z}(\cdot).$$

Let $M : \mathcal{C}(\Gamma) \to \mathbb{R}^+$ be $\|\cdot\|_\infty$ continuous. Then applying the continuous mapping theorem (Billingsley 1999 p. 20), we get

$$M[T_n(\cdot)] \xrightarrow{d} M[\mathcal{Z}(\cdot)]$$

under the null hypothesis. For example, with $M[T_n(\cdot)] = \int_{\mathbf{\Gamma}} [T_n(\boldsymbol{\gamma})]^2 d\mu(\boldsymbol{\gamma})$, we have

$$M_n \equiv M[T_n(\cdot)] = \int_{\mathbf{\Gamma}} [T_n(\boldsymbol{\gamma})]^2 d\mu(\boldsymbol{\gamma}) = n \int_{\mathbf{\Gamma}} \left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}) \xrightarrow{d} \int_{\mathbf{\Gamma}} [\mathcal{Z}(\boldsymbol{\gamma})]^2 d\mu(\boldsymbol{\gamma})$$

under H_0.

4.3 Global and Local Alternatives

The global alternatives for our conditional independence test can always be written as

$$H_a: \; f_Z(z) f_{XYZ}(x,y,z) - f_{YZ}(y,z) f_{XZ}(x,z) = \alpha(x,y,z), \tag{27}$$

for some nontrivial and nonzero function $\alpha(x,y,z)$. Then under H_a, we have

$$\Delta(\boldsymbol{\gamma}) = \int \varphi(\tilde{w}'\boldsymbol{\gamma}) \alpha(x,y,z) dx dy dz.$$

This will be nonzero for essentially all $\boldsymbol{\gamma} \in \Gamma$ provided that φ is GCR. It follows from Theorem 4 that

$$\lim_{n\to\infty} \Pr(M_n > c_n) = 1$$

for any critical value $c_n = o(n)$. That is, the test is consistent: as the sample size increases, the test will eventually detect the alternative H_a.

To construct a local alternative, we consider a mixture distribution of the form

$$H_{a,n}: f_{XYZ}(x,y,z) = \left[\left(1 - \frac{c}{\sqrt{n}}\right) f_{Y|Z}(y|z) + \frac{c}{\sqrt{n}} \tilde{\alpha}(y|x,z)\right] f_{XZ}(x,z), \tag{28}$$

where c is a constant and $\tilde{\alpha}(y|x,z)$ is a conditional density function of $\tilde{Y}$ given $(\tilde{X}, \tilde{Z})$ such that $\tilde{Y} \not\perp \tilde{X} \mid \tilde{Z}$. By construction, $\tilde{\alpha}(y|x,z)$ is a nontrivial function of x and z. That is, the distribution of W is a mixture of two distributions: one satisfies the null of conditional independence and the other does not. The mixing proportion is local to unity. Equivalently, we can rewrite the local alternative as

$$H_{a,n}: f_{XYZ}(x,y,z) = f_{Y|Z}(y,z) \; f_{XZ}(x,z) + \frac{\alpha(x,y,z)}{\sqrt{n}}$$

for $\alpha(x,y,z) = c\left[\tilde{\alpha}(y|x,z) - f_{Y|Z}(y|z)\right] f_{XZ}(x,z)$. Since $\tilde{\alpha}(y|x,z)$ depends on x, $\tilde{\alpha}(y|x,z) - f_{Y|Z}(y|z)$ cannot be a zero function. Hence when φ is GCR and $c > 0$,

$$\pi_\varphi(\boldsymbol{\gamma}) := \int \varphi(\tilde{w}'\boldsymbol{\gamma}) \alpha(x,y,z) \, dx dy dz \neq 0$$

for essentially all $\boldsymbol{\gamma} \in \Gamma$.

Under Assumptions 1–5 and the local alternative $H_{a,n}$, we can use the same arguments as in the proof of Theorem 4 to show that

$$M_n = \int_{\mathbf{\Gamma}} [T_n(\boldsymbol{\gamma})]^2 d\mu(\boldsymbol{\gamma}) \xrightarrow{d} \int_{\mathbf{\Gamma}} [\mathcal{Z}(\boldsymbol{\gamma}) + \pi_\varphi(\boldsymbol{\gamma})]^2 d\mu(\boldsymbol{\gamma}).$$

The essentially nonzero mean is the source of the power of the ICM test against the local alternative.

4.4 Calculating the Asymptotic Critical Values

Under the null, M_n has a limiting distribution given by a functional of a zero mean Gaussian process whose covariance function depends on the DGP. The asymptotic critical values thus depend on the DGP and cannot be tabulated. One could follow Bierens and Ploberger (1997) and obtain upper bounds for the asymptotic critical values. Here, we use the conditional Monte Carlo approach suggested by Hansen (1996) to simulate the asymptotic null distribution.

To apply this approach, we construct a process $T_n^*(\cdot)$, which follows the desired zero mean Gaussian process conditional on $\{W_i\}$. The desired conditional covariance function for T_n^* is

$$cov\left[T_n^*(\boldsymbol{\gamma}_1), T_n^*(\boldsymbol{\gamma}_2) \middle| \{W_i\}_{i=1}^n\right] = \frac{4}{n}\sum_{i=1}^n \hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma}_1)\ \hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma}_2) \equiv \hat{\sigma}_\Delta\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right),$$

where

$$\hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma}) = (n-1)^{-1}\sum_{j=1,j\neq i}^n \kappa_{h,2}(W_i,W_j;\boldsymbol{\gamma}).$$

It is straightforward to show that under Assumptions 1-5 and the null hypothesis,

$$\hat{\sigma}_\Delta\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right) \xrightarrow{p} \sigma_\Delta\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right).$$

A typical $T_n^*(\cdot)$ is constructed by generating $\{V_i\}_{i=1}^n$ as IID standard normal random variables independent of $\{W_i\}$ and setting

$$T_n^*(\boldsymbol{\gamma}) = \frac{2}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma})V_i. \tag{29}$$

Following the arguments similar to the proof of Theorem 2 in Hansen (1996), we can show that under the null hypothesis,

$$M_n^* = \int_{\boldsymbol{\Gamma}} \left[T_n^*(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right) \xrightarrow{d} \int_{\boldsymbol{\Gamma}} \left[\mathcal{Z}(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right),$$

provided that Assumptions 1-5 hold.

Simulation results show that the empirical PDFs of M_n and M_n^* are fairly close. To save space, we do not report the results here, but they are available in Huang (2009).

To approximate the distribution of M_n, we follow the steps below:

- generate $\{V_{ib}\}_{i=1}^n$ IID $N(0,1)$ random variables;
- set
$$T_{n,b}^*(\boldsymbol{\gamma}) \equiv \frac{2}{\sqrt{n}}\sum_{i=1}^n \hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma})V_{ib};$$
- set $M_{n,b}^* \equiv M\left[T_{n,b}^*(\cdot)\right] = \int_{\boldsymbol{\Gamma}} \left[T_{n,b}^*(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right).$

This gives a simulated sample $(M_{n,1}^*, ..., M_{n,B}^*)$, whose empirical distribution should be close to the true distribution of the actual test statistic M_n under the null. Then we can compute the proportion of simulated values that exceed M_n to get the simulated asymptotic p value. We reject the null hypothesis if the simulated p value lies below the specified level for the test. As Hansen (1996) points out, B is under the control of the econometrician and can be chosen sufficiently large to obtain a good approximation.

4.5 A Rescaled ICM Test

The variance of $\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ depends on $\boldsymbol{\gamma}$. It is plausible that by rescaling $\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$ by its standard deviation, one might obtain a somewhat better test. Thus, consider

$$\tilde{T}_n(\boldsymbol{\gamma}) \equiv \frac{\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_\Delta(\boldsymbol{\gamma})} \text{ and}$$

$$\tilde{M}_n \equiv M\left[\tilde{T}_n(\cdot)\right] = \int_{\boldsymbol{\Gamma}} \left[\tilde{T}_n(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}),$$

where

$$\hat{\sigma}_\Delta^2(\boldsymbol{\gamma}) = \hat{\sigma}_\Delta(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = 4n^{-1}\sum_{i=1}^n \left[\hat{\kappa}_{h,1}(W_i; \boldsymbol{\gamma})\right]^2 - 4\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2.$$

Proposition 5 *Suppose Assumptions 1-5 hold and that* $\inf_{\boldsymbol{\gamma}\in\Gamma}\sigma_\Delta(\boldsymbol{\gamma}) > 0$. *Then under the null hypothesis,*

$$\tilde{T}_n(\cdot) = \frac{\sqrt{n}\hat{\Delta}_{n,h}(\cdot)}{\hat{\sigma}_\Delta(\cdot)} \xrightarrow{d} \tilde{\mathcal{Z}}(\cdot),$$

where $\tilde{\mathcal{Z}}$ *is a zero mean Gaussian process on* Γ *with covariance function*

$$cov\left(\tilde{\mathcal{Z}}(\boldsymbol{\gamma}_1), \tilde{\mathcal{Z}}(\boldsymbol{\gamma}_2)\right) = \frac{\sigma_\Delta(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)}{\sigma_\Delta(\boldsymbol{\gamma}_1)\sigma_\Delta(\boldsymbol{\gamma}_2)} \equiv \rho_\Delta(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2).$$

By the continuous mapping theorem, we have

$$\tilde{M}_n \xrightarrow{d} \int_{\boldsymbol{\Gamma}} \left[\tilde{\mathcal{Z}}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}).$$

Define

$$\tilde{T}_n^*(\boldsymbol{\gamma}) \equiv \frac{2}{\hat{\sigma}_n(\boldsymbol{\gamma})} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}) V_i$$

with $\{V_i\}_{i=1}^n$ IID $N(0,1)$, independent of $\{W_i\}$. Then we can follow the proof of Theorem 2 in Hansen (1996) to show that

$$\tilde{M}_n^* \equiv \int_{\boldsymbol{\Gamma}} \left[\tilde{T}_n^*(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}) \xrightarrow{d} \int_{\boldsymbol{\Gamma}} \left[\tilde{\mathcal{Z}}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}).$$

As a result, the critical value of $\tilde{M}_n$ can be obtained by simulating $\tilde{M}_n^*$. Simulation results not reported here show that the empirical PDFs of $\tilde{M}_n$ and $\tilde{M}_n^*$ are fairly close.

Although we do not give formal statements, results analogous to those for M_n hold under the local and global alternatives. Simulation results in the next section suggest that the rescaled ICM test has somewhat better power for most experiments.

5 Monte Carlo Experiments

In this section, we perform some Monte Carlo simulation experiments to examine the finite sample performance of our conditional independence test.

For all simulations, we generate IID $\{(X_i, Y_i, Z_i)\}$. We choose $\varphi(\cdot)$ to be the standard normal PDF, and $k(u)$ be the sixth-order Gaussian kernel ($q = 6$). The number of replications for each experiment is 1000, and the number of replications for simulating M_n^* or $\tilde{M}_n^*$ is 999.

5.1 Level and Power Studies

5.1.1 DGP 1

We first generate a sample $\{(X_i, Y_i, Z_i)\}$ using the DGP

$$\begin{aligned} Y &= \theta X + Z + \varepsilon_Y \\ X &= Z + Z^2 + \varepsilon_X, \end{aligned}$$

where

$$\begin{pmatrix} \varepsilon_X \\ \varepsilon_Y \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}\right) = N\left(0, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

and

$$Z \sim N(0, \sigma_Z^2) = N(0, 3).$$

When $\theta = 0$, the null is true; otherwise the alternative holds.

We normalize each variable so that its support is comparable to that of the GCR function $\varphi(\cdot)$. For the standard normal PDF, the support is the real line but the function is effectively zero out of the interval $[-4, 4]$. We normalize each variable to be supported on this interval. This can be achieved by taking $\tilde{X}_i = 8\left[X_i - \min(X_i)\right] / [\max(X_i) - \min(X_i)] - 4$. We normalize Y_i and Z_i analogously. The conditional independence test is then applied to $\tilde{X}_i, \tilde{Y}_i$, and $\tilde{Z}_i$. Although any compact Γ with a non-empty interior can be used, we take $\Gamma = [-1, 1]^4$. This choice ensures that $\{\tilde{W}_i' \boldsymbol{\gamma}, \boldsymbol{\gamma} \in \Gamma\}$ can take any value in the effective support of $\varphi(\cdot)$.

To compute the ICM statistic M_n, we need to compute the integral $\int_{\boldsymbol{\Gamma}} [T_n(\boldsymbol{\gamma})]^2 \, d\mu(\boldsymbol{\gamma})$. In the absence of a closed-form expression, we recommend using the Monte Carlo integration method. For each simulation replication, we choose 100 $\boldsymbol{\gamma}_s$'s randomly from the uniform distribution on $[-1, 1]^4$ and approximate the integral by the average $\sum_{s=1}^{100} T_n^2(\boldsymbol{\gamma}_s)/100$. We have also tried using 50 random draws, but the results are effectively the same. Note that $T_n^2(\boldsymbol{\gamma}_s)$ depends on the bandwidth parameter h. In our simulation experiments, we employ the data-driven bandwidth $\hat{h}(\boldsymbol{\gamma}_s)$ in (25) with $h_0 = n^{-1/[3(2q+d_Z)]}$ and $\tau = 0.5$. We use different bandwidths for different $\boldsymbol{\gamma}$'s. Given the bandwidth $\hat{h}(\boldsymbol{\gamma}_s)$, we compute the statistic $T_n^2(\boldsymbol{\gamma}_s)$ as $T_n^2(\boldsymbol{\gamma}_s) = n\hat{\Delta}^2_{n,\hat{h}(\boldsymbol{\gamma}_s)}(\boldsymbol{\gamma}_s)$. The average of $T_n^2(\boldsymbol{\gamma}_s)$ gives us the ICM statistic M_n. The rescaled ICM statistic $\tilde{M}_n$ is computed similarly.

We use DGP 1 to study the finite sample size and power of the test against conditional mean dependence. We use

$$\rho_{X,Y|Z} = \frac{cov\,(X, Y|Z)}{\sigma_{X|Z}\sigma_{Y|Z}} = \frac{\theta\sigma_X^2}{\sigma_X\sqrt{\theta^2\sigma_X^2 + \sigma_Y^2}} = \frac{4\theta}{2\sqrt{4\theta^2 + 1}}$$

to indicate the strength of the dependence between X and Y, conditional on Z. Since both $X|Z$ and $Y|Z$ are normal, $\rho_{X,Y|Z}$ fully captures the dependence between X and Y, conditional on Z.

We plot the power of the tests for ρ ranging from -0.9 to 0.9. For this, we choose

$$\theta = \frac{\rho_{X,Y|Z}}{2\sqrt{\left(1-\rho^2_{X,Y|Z}\right)}} \quad \text{for} \quad \rho_{X,Y|Z} = -0.9, -0.8, ..., 0.9.$$

The size and power look fairly good for sample sizes as small as 100, and they look very good when the sample size reaches 200. The "non-standardized" results in Figure 1 correspond to M_n, and the "standardized" results in Figure 2 correspond to $\tilde{M}_n$. When the sample size is small, the levels of the tests approach their nominal value from below, delivering conservative tests. When the sample size increases to 200, our tests become fairly accurate in size. The power functions show that $\tilde{M}_n$ performs better than M_n in this experiment. This may be due to some efficiency improvements associated with the partial GLS correction embodied in $\tilde{M}_n$.

5.1.2 DGP 2

DGP 2 is a modification of DGP 1 that focuses on the consequences of fat-tailed distributions. Here, ε_X and ε_Y are proportional to the Student t with 3 degrees of freedom:

$$\varepsilon_X \sim 2t_3,\ \varepsilon_Y \sim t_3,\ \varepsilon_X \perp \varepsilon_Y.$$

The power functions for M_n are plotted in Figure 3, and those for $\tilde{M}_n$ are plotted in Figure 4. We see that the power is a little but not a lot worse than for the normal distributions of DGP 1.

5.1.3 DGP 3

DGP 3 is another modification of DGP 1. This time we allow skewness, choosing both ε_X and ε_Y to be centered chi-square distributions:

$$\varepsilon_X \sim 2\left(\chi_1^2 - 1\right),\ \varepsilon_Y \sim \left(\chi_1^2 - 1\right),\ \varepsilon_X \perp \varepsilon_Y.$$

The power functions of M_n are plotted in Figure 5 and those for $\tilde{M}_n$ are plotted in Figure 6. Here, the power is slightly better than that for DGP 1. Overall, the size and power properties of our tests are robust to the data distribution.

5.2 Comparison with Other Tests

In this section we compare the standardized ICM test $\tilde{M}_n$ with other conditional independence tests. Su and White's (2008) test essentially compares $f_{XYZ}\ f_Z$ with $f_{XZ}\ f_{YZ}$ and can detect local alternatives at the rate $n^{-1/2}h^{-d/4}$. Su and White's (2007) test essentially compares $f_{Y|X,Z}$ with $f_{Y|Z}$ and can detect local alternatives at the rate $n^{-1/2}h^{-(d_X+d_Z)/4}$. Our test compares integral transforms and can detect local alternatives at the rate $n^{-1/2}$. We first compare all three tests using DGP1. Figure 7 shows the power functions when the sample size is 100. The GCR test in the figure is the test we propose. It is clear that our test outperforms the SW 2007 test, which in turn outperforms the SW 2008 test. More specifically, while our GCR test has almost the same empirical size as the SW 2007 test, it is more powerful than the SW 2007 test. The SW 2008 test is very conservative and has almost no power when ρ is small in absolute value. That is, when the departure from the null is small, the SW 2008 test is less able to detect it, compared with our GCR test and the SW 2007 test.

Figure 8 shows the power functions when the sample size is increased to 200. We see that the power of our GCR test improves faster than the power of SW 2007, which again improves faster than the power of SW 2008. These results are consistent with the local alternative rate results.

Finally, we compare the power function of our $\tilde{M}_n$ test with the tests proposed by LG (1997) and DG (2001). Figure 9 reports the results for DGP 1 with $n = 200$. We report only the results for the Cramer-von Mises type test for each method, as the results for the Kolmogorov-Smirnov type test are qualitatively similar. In the figure, “LG” and “DG” represent the Cramer-von Mises type tests of LG (1997) and DG (2001), respectively. The figure demonstrates the clear advantage of our GCR test. It is as accurate in size as the LG test but more powerful than the latter test. The GCR test has better finite sample performances than the DG test in terms of both size and power properties.

In all the figures, we also report the “gold standard” t-test. This is as good a test as one could want, in the sense that it is the parametric maximum likelihood test for $\theta = 0$ in a correctly specified linear model. Although our test is not as powerful as the t-test, which is reasonable since our test is fully nonparametric, our GCR test does outperform all other nonparametric tests. On the other hand, the t-test measures only linear dependence. In the presence of nonlinear dependence, the t-test may be less powerful than the nonparametric tests. This is supported by simulation results not reported here.

6 Application to Returns to Schooling

As stated in the introduction, one important application of tests for conditional independence is to test a key assumption identifying causal effects. In this section, we provide an example.

In the literature on returns to schooling, the most widely investigated structural equation is a Mincer (1974) type semi-logarithmic human capital earnings function:

$$\ln Y_i = \theta_0 + \theta_1 S_i + \theta_2 EXP_i + \theta_3 EXP_i^2 + U_i, \tag{30}$$

where the subscript i indexes individuals, $\ln Y_i$ is log hourly wage, S_i is years of completed schooling, EXP_i is years of work experience, EXP_i^2 is work experience squared, and U_i represents unobserved drivers of $\ln Y_i$, centered at zero. The effect of interest is θ_1, the effect of an additional year of schooling on wage. In what follows, we drop the i subscript.

Least squares estimates of the Mincer equation suffer from the well-known ability bias problem, which is caused by the dependence of schooling on unobserved ability. To make this explicit, let $U = A + \varepsilon$, where A represents unobserved ability, and rewrite the Mincer equation as

$$\ln Y = \theta_0 + \theta_1 S + \theta_2 EXP + \theta_3 EXP^2 + A + \varepsilon. \tag{31}$$

One method empirical researchers have adopted to address the ability bias issue is to find proxies Z for ability, for example IQ or AFQT scores, and include these as regressors (e.g., Griliches and Mason, 1972; Griliches, 1977; and Blackburn and Neumark, 1993). Now consider the regression of $\ln Y$ on S, EXP, and Z :

$$\begin{aligned} \mu(S, EXP, Z) &= E(\ln Y \mid S, EXP, Z) \\ &= E(\theta_0 + \theta_1 S + \theta_2 EXP + \theta_3 EXP^2 + A + \varepsilon \mid S, EXP, Z) \\ &= \theta_0 + \theta_1 S + \theta_2 EXP + \theta_3 EXP^2 + E(A + \varepsilon \mid S, EXP, Z) \\ &= \theta_0 + \theta_1 S + \theta_2 EXP + \theta_3 EXP^2 + E(A + \varepsilon \mid EXP, Z). \end{aligned}$$

The last equality is justified by a conditional mean independence assumption,

$$E(A + \varepsilon \mid S, EXP, Z) = E(A + \varepsilon \mid EXP, Z).$$

If this holds, then we have

$$(\partial/\partial s)\mu(S, EXP, Z) = \theta_1,$$

so that the effect of interest, θ_1, is identified and can be consistently estimated.

There is no reason *a priori* that the wage equation must have the specific Mincer form, however. More generally, one can consider a nonparametric specification

$$\ln Y = r(S, X, U),$$

where r is an unknown function; X contains observable factors determining wages, including EXP, as well as other factors like job tenure, region, sex, race, etc.; and $U = (A, \varepsilon)$.

An important effect of interest here is

$$\phi_1(S, X, U) = (\partial/\partial s)r(S, X, U),$$

the marginal effect of schooling on wage. This effect depends on all drivers of wage, including unobservables, U, so $\phi_1(S, X, U)$ is not identifiable without further potentially strong restrictions. Nevertheless, just as in the linear case, it is possible to identify and estimate certain expectations of $\phi_1(S, X, U)$ given suitable ability proxies Z, as

$$\begin{aligned} (\partial/\partial s)\mu(s, x, z) &= (\partial/\partial s)E(\ln Y \mid S = s, X = x, Z = z) \\ &= E((\partial/\partial s)r(S, X, U) \mid S = s, X = x, Z = z) \\ &= E(\phi_1(s, X, U) \mid X = x, Z = z) \equiv \bar{\phi}_1(s, x, z). \end{aligned}$$

The crucial condition justifying the third equality is conditional independence:

$$(A, \varepsilon) \perp S \mid (X, Z) \tag{32}$$

This is called a "conditional exogeneity" assumption by White and Chalak (2008). It implies the "ignorability" or "unconfoundedness" condition, also known as "selection on observables" in the literature, ensuring identification of causal effects.

Thus, if (32) holds, and even if the specific Mincer function (31) does not, we can still identify the average marginal effect of schooling $\bar{\phi}_1(s, x, z)$ and consistently estimate this by various methods. If (32) fails, then the marginal effect of interest is no longer identified (see, e.g., White and Chalak, 2008, theorem 4.1).

We cannot test (32) directly, as A and ε are unobservable. However, following White and Chalak (2010), if we can observe V such that

$$\begin{aligned} V &= f(A, \varepsilon, X, Z, \eta) \\ \eta &\perp S \mid (A, X, Z), \end{aligned} \tag{33}$$

where f denotes some unknown function and η is unobserved, then

$$(A, \varepsilon) \perp S \mid (X, Z) \text{ implies } V \perp S \mid (X, Z).$$

Thus, we can test unconfoundedness by testing the implied condition

$$H_0 : V \perp S \mid (X, Z). \tag{34}$$

Equation (33) provides some guidance about how to choose V. The conditional independence requirement on η is particularly plausible when η is a measurement error, so that both Z and V could be error-laden proxies for ability. Here, we test (34) using data from the National Longitudinal Survey of Youth 1979 (NLSY 79). In particular, we use the data from survey year 2000 and restrict the sample to white males.[1] We use the age-adjusted standardized AFQT in year 1980 as Z. V includes math and verbal scores for preliminary scholastic aptitude tests from 1981 high school transcripts. To satisfy (33), we use years of schooling beyond high school as S, so that V is not affected by S. X includes actual work experience in survey year 2000 and total tenure with employer in survey year 2000.

To implement the test, we choose $\varphi(\cdot)$ to be the standard normal PDF, and let $k(\cdot)$ be the sixth-order Gaussian kernel. We choose $\boldsymbol{\gamma}$ and other metaparameters as described in the Monte Carlo section. Applying our $\tilde{M}_n$ test, we find that we do not reject the null hypothesis (34) at the 5% level. Thus, we do not find evidence refuting the approach commonly used by empirical researchers, providing some support for parametric or nonparametric estimation of effects of interest.

7 Concluding Remarks

In this paper, we develop a flexible nonparametric test for conditional independence that is simple to implement, yet powerful. It is consistent against any deviation from the null and achieves local power at the parametric $n^{-1/2}$ rate, despite its nonparametric character. It is also very flexible as it allows for a rich class of GCR functions.

There are several useful directions for future research. First, we have assumed that the data are IID. But this is not essential for the results. We may straightforwardly extend the approach to a time-series framework, so that we could test, for example, nonlinear Granger causality. Another extension could be to modify the test so that it can be used when Z contains both discrete and continuous variables. This is often relevant in applied microeconomics. This extension has been considered in Chapter 3 of Huang (2009). A third direction is to further study the bandwidth selection problem. Here, we choose the bandwidth to minimize the mean squared error of $\hat{\Delta}_{n,h}(\boldsymbol{\gamma})$. Ideally, however, one should choose the bandwidth that optimizes the trade-off between size and power.

[1] To restrict the sample so that it is suitable for estimating a wage equation for survey year 2000, we drop those who were enrolled in high school or college in survey year 2000, and we exclude those who were in active armed forces, self-employed, or working in a family business in survey year 2000. We also drop those whose hourly wage was not in the range $(\$1, \$1000]$.

8 Appendix of Proofs

Throughout the proofs, we use C to denote a constant that may be different across different equations or lines.

Proof of Lemma 1: For the pointwise result, we use Assumption 1 and the theory of U-statistics to obtain

$$\begin{aligned} var\left[\sqrt{n}R_{n,h}\left(\boldsymbol{\gamma}\right)\right] &= \frac{2}{(n-1)} var\left[\tilde{\kappa}_{h,2}(W_i, W_j, \boldsymbol{\gamma})\right] \\ &\leq \frac{2}{(n-1)} var\left[\kappa_{h,2}(W_i, W_j, \boldsymbol{\gamma})\right] \leq \frac{2}{(n-1)} E\left[\kappa_{h,2}^2(W_i, W_j, \boldsymbol{\gamma})\right]. \end{aligned}$$

So it suffices to show that $E\left[\kappa_{h,2}^2(W_i, W_j, \boldsymbol{\gamma})\right] = o(n)$. But

$$\begin{aligned} & \kappa_{h,2}(W_i, W_j, \boldsymbol{\gamma}) \\ = & \frac{1}{2}\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\right. \\ & \left. -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)\right] K_h(Z_i - Z_j) \\ & +\frac{1}{2}\left[\varphi(\gamma_0 + X_j'\gamma_1 + Y_j'\gamma_2 + Z_j'\gamma_3)\right. \\ & \left. -\varphi(\gamma_0 + X_j'\gamma_1 + Y_i'\gamma_2 + Z_j'\gamma_3)\right] K_h(Z_j - Z_i), \end{aligned}$$

and so

$$\begin{aligned} & E\left[\kappa_{h,2}^2(W_i, W_j, \boldsymbol{\gamma})\right] \\ \leq & E\left|\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right|^2 \\ & +E\left|\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right|^2 \\ & +E\left|\varphi(\gamma_0 + X_j'\gamma_1 + Y_j'\gamma_2 + Z_j'\gamma_3)K_h(Z_j - Z_i)\right|^2 \\ & +E\left|\varphi(\gamma_0 + X_j'\gamma_1 + Y_i'\gamma_2 + Z_j'\gamma_3)K_h(Z_j - Z_i)\right|^2 \\ \leq & 4\varphi_{\max}^2 E K_h^2(Z_i - Z_j), \end{aligned} \tag{35}$$

where $\varphi_{\max} = \sup_{\gamma\in\Gamma}\sup_{W\in[0,1]^d}\varphi(\tilde{W}'\gamma)$, which is finite under Assumption 3. Using Assumption 2, we have

$$\begin{aligned} & EK_h^2(Z_i - Z_j) \\ = & \int \frac{1}{h^{2d_Z}}\left|K(\frac{z_1 - z_2}{h})\right|^2 f_Z\left(z_1\right) f_Z(z_2)dz_1dz_2 \\ \leq & \int_{[0,1]^d} \frac{1}{h^{d_Z}}\left(\int_{-\infty}^{\infty} K^2(u) f_Z(z_2 + uh)du\right) f_Z\left(z_2\right) dz_2 \\ \leq & \frac{1}{h^{d_Z}}\left(\int_{[0,1]^d} f_Z^2\left(z\right) dz\right)\left(\int K^2(u)du\right) + \frac{\rho}{h^{d_Z - 1}}\left(\int f_Z\left(z\right) dz\right)\int K^2(u)\left\|u\right\| du \\ = & \frac{1}{h^{d_Z}} E\left[f_Z\left(Z\right)\right]\left(\int \left|K(u)\right|^2 du\right) + \frac{1}{h^{d_Z - 1}}\int K^2(u)\left\|u\right\| du. \end{aligned} \tag{36}$$

It follows from Assumption 4 that $\int K^2(u)du = \left(\int k^2(v)dv\right)^{d_Z} < \infty$ and

$$\begin{aligned}
& \int K^2(u)\,\|u\|\,du \\
= & \int k^2(u_1)\cdots k^2(u_{d_Z})\sqrt{u_1^2+\cdots+u_{d_Z}^2}du_1\cdots du_{d_Z} \\
\leq & \sqrt{d_Z}\int k^2(u_1)\cdots k^2(u_{d_Z})\max_{i=1,\ldots,d_Z}|u_i|\,du_1\cdots du_{d_Z} \\
= & \sqrt{d_Z}\int k^2(u_1)\cdots k^2(u_{d_Z})\left(|u_1|+\cdots+|u_{d_Z}|\right)du_1\cdots du_{d_Z} \\
= & d_Z\sqrt{d_Z}\left(\int k^2(v)\,|v|\,dv\right)\left(\int k^2(v)dv\right)^{d_Z-1} < \infty.
\end{aligned}$$

Therefore

$$EK_h^2(Z_i - Z_j) = O\left(\frac{1}{h^{d_Z}}\right).$$

Combining this with (35), we have, using Assumption 5(a):

$$E|\kappa_{h,i,j}(\boldsymbol{\gamma})|^2 = O\left(\frac{1}{h^{d_Z}}\right) = O(n\times\frac{1}{nh^{d_Z}}) = o(n).$$

This implies that $R_{n,h}(\boldsymbol{\gamma}) = o_p(1/\sqrt{n})$ pointwise for each $\boldsymbol{\gamma}\in\Gamma$.

To show the uniformity result that $\sup_{\boldsymbol{\gamma}\in\Gamma} R_{n,h}(\boldsymbol{\gamma}) = o_p(1/\sqrt{n})$, we employ the theory of U-processes. In particular, we apply Proposition 4 in DG (2001) with their $k = 2$. The class of functions under consideration is $\mathcal{K} = \{\kappa_{h,2}(W_i, W_j, \boldsymbol{\gamma}) : \boldsymbol{\gamma}\in\Gamma\}$. Since $|\kappa_{h,2}(W_i,W_j,\boldsymbol{\gamma})| \leq 2\varphi_{\max}|K_h(Z_i - Z_j)|$, we can use $\mathbb{K}(W_i,W_j) = 2\varphi_{\max}|K_h(Z_i-Z_j)|$ as the envelope function. As sets of linear functions whose subgraphs are half planes, both $\{\tilde{W}_i\boldsymbol{\gamma} : \boldsymbol{\gamma}\in\Gamma\}$ and $\{\tilde{W}_{ij}\boldsymbol{\gamma} : \boldsymbol{\gamma}\in\Gamma\}$ are VC-type. Under Assumption 3(b), it is clear that $\{\varphi(\tilde{W}_i\boldsymbol{\gamma}) : \boldsymbol{\gamma}\in\Gamma\}$ and $\{\varphi(\tilde{W}_{ij}\boldsymbol{\gamma}) : \boldsymbol{\gamma}\in\Gamma\}$ also are VC-type. Multiplying by a fixed function $K_h(\cdot)$ will not change their VC property and the associated VC characteristics. Therefore $\{\kappa_{h,2}(W_i,W_j,\boldsymbol{\gamma}) : \boldsymbol{\gamma}\in\Gamma\}$ is VC type with VC characteristics independent of h. Applying Proposition 4 in DG (2001), we have

$$E\sup_{\boldsymbol{\gamma}\in\Gamma}\left|\frac{n(n-1)}{n}R_{n,h}(\boldsymbol{\gamma})\right|^2 \leq CE\mathbb{K}^2(W_i,W_j)$$

for some constant C that does not depend on h. But $E\mathbb{K}^2(W_i,W_j) = O\left(1/h^{d_Z}\right)$, and so $E\sup_{\boldsymbol{\gamma}\in\Gamma}|\sqrt{n}R_{n,h}(\boldsymbol{\gamma})|^2 = O(1/\left(nh^{d_Z}\right)) = o(1)$. As a result, $\sup_{\boldsymbol{\gamma}\in\Gamma}\sqrt{n}R_{n,h}(\boldsymbol{\gamma}) = o_p(1)$.■

Proof of Lemma 2: Part (a). We first establish an expansion of $\int_{[0,1]^{d_Z}} K_h(u -$

$z)f_Z(u)\,du$, starting with

$$\begin{aligned}
&\int_{[0,1]^{d_Z}} K_h(u-z)f_Z(u)\,du \\
=\ & \int_{[0,1]^{d_Z}} \frac{1}{h^{d_Z}}k(\frac{u_1-z_1}{h})\cdots k(\frac{u_{d_Z}-z_{d_Z}}{h})f_Z(u)\,du_1\cdots du_{d_Z} \\
=\ & \int_{-z_1/h}^{(1-z_1)/h}\cdots\int_{-z_{d_Z}/h}^{(1-z_{d_Z})/h} k(v_1)\cdots k(v_{d_Z})f_Z(z_1+v_1h,\ldots,z_{d_Z}+v_{d_Z}h)\,dv_1\cdots dv_{d_Z} \\
=\ & \prod_{\ell=1}^{d_Z}\left[\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)\,dv_\ell\right] f_Z(z) + \sum_{0<|j|\le q-1} h^{|j|}\frac{D^j f_Z(z)}{j!}\prod_{\ell=1}^{d_Z}\left[\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)v_\ell^{j_\ell}dv_\ell\right] \\
& + \sum_{|j|=q} h^{|j|}\frac{D^j f_Z(z)}{j!}\prod_{\ell=1}^{d_Z}\left[\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)v_\ell^{j_\ell}dv_\ell\right] + C_K h^{q+1},
\end{aligned}$$

where

$$\begin{aligned}
|C_K| &= \left|\sum_{|j|=q+1}\prod_{\ell=1}^{d_Z}\int_{-z_\ell/h}^{(1-z_\ell)/h}\frac{D^j f_Z(z+\tilde{v}h)}{j!}k(v_\ell)v_\ell^{j_\ell}dv_\ell\right| \\
&\le \left[\max_{j:|j|=q+1}\max_{z\in[0,1]^{d_Z}} D^j f_Z(z)\right]\sum_{|j|=q+1}\frac{1}{j!}\prod_{\ell=1}^{d_Z}\int_{-\infty}^{\infty}\left|k(v_\ell)v_\ell^{j_\ell}\right|dv_\ell \le C.
\end{aligned}$$

Here we have used Assumptions 2(a) and 4(b).

When $z_\ell \in [h^\alpha, 1-h^\alpha]$ for some $\alpha\in(0,1)$, we have

$$\begin{aligned}
&\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)\,dv_\ell\right) f_Z(z) \\
=\ & f_Z(z) - \left(\int_{z_\ell/h}^{\infty} k(v_\ell)\,dv_\ell\right) f_Z(z) - \left(\int_{(1-z_\ell)/h}^{\infty} k(v_\ell)\,dv_\ell\right) f_Z(z).
\end{aligned}$$

But under Assumption 4(b), we have, for some constants $\tilde{C}$ and C,

$$\left|\int_{z_\ell/h}^{\infty} k(v_\ell)\,dv_\ell\right| \le \int_{h^{\alpha-1}}^{\infty}|k(v_\ell)|\,dv_\ell \le \tilde{C}\int_{h^{\alpha-1}}^{\infty}\frac{1}{1+|v|^\xi}dv_\ell \le Ch^{(1-\alpha)(q^2+q+1)},$$

and similarly $\left|\int_{(1-z_\ell)/h}^{\infty} k(v_\ell)\,dv_\ell\right| \le Ch^{(1-\alpha)(q^2+q+1)}$. Hence

$$\left|\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)\,dv_\ell\right) f_Z(z) - f_Z(z)\right| \le Ch^{(1-\alpha)(q^2+q+1)}.$$

When $z_\ell\in[0,h^\alpha)$, we have, for some $z_\ell^*\in(0,z_\ell)$,

$$\begin{aligned}
&\left|\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)\,dv_\ell\right) f_Z(z)\right| = \left|\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)\,dv_\ell\right) f_Z(z_1,\ldots,z_\ell,\ldots z_{d_Z})\right| \\
\le\ & \tilde{C}\left|\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} |k(v_\ell)|\,dv_\ell\right)\frac{(z_\ell^*)^{q+1}}{(q+1)!}\right| \le Ch^{\alpha(q+1)},
\end{aligned}$$

where we have used Assumption 2(b). Similarly when $z_\ell \in (1-h^\alpha, 1]$,

$$\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k\left(v_\ell\right) dv_\ell\right) f_Z\left(z\right) \leq Ch^{\alpha(q+1)}.$$

If we choose $\alpha \in (\frac{q}{q+1}, 1-\frac{q}{q^2+q+1})$, which is feasible, then

$$\sup_{z\in[0,1]^{d_Z}}\left|\left(\int_{-z_\ell/h}^{(1-z_\ell)/h} k\left(v_\ell\right) dv_\ell\right) f_Z\left(z\right) - f_Z\left(z\right)\right| \leq Ch^{q+e}$$

for some $e>0$. Repeating the above arguments for other elements of z, we obtain

$$\sup_{z\in[0,1]^{d_Z}}\left|\prod_{\ell=1}^{d_Z}\left[\int_{-z_\ell/h}^{(1-z_\ell)/h} k\left(v_\ell\right) dv_\ell\right] f_Z\left(z\right) - f_Z\left(z\right)\right| \leq Ch^{q+e}.$$

By the same argument, we can show that under Assumption 4 and 2(a)(b):

$$\sup_{z\in[0,1]^{d_Z}}\left|\sum_{j:0<|j|\leq q-1} h^{|j|}\frac{D^j f_Z(z)}{j!}\prod_{\ell=1}^{d_Z}\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)v_\ell^{j_\ell}dv_\ell\right| \leq Ch^{q+e}$$

and

$$\sup_{z\in[0,1]^{d_Z}}\left|\sum_{|j|=q} h\left|^{j}\right|\frac{D^j f_Z(z)}{j!}\prod_{\ell=1}^{d_Z}\left[\int_{-z_\ell/h}^{(1-z_\ell)/h} k(v_\ell)v_\ell^{j_\ell}dv_\ell\right] - \frac{\mu_q}{q!}\left[\sum_{\ell=1}^{d_Z}\frac{\partial^q f_Z\left(z\right)}{\partial z_\ell^q}\right]h^q\right| \leq Ch^{q+e},$$

where

$$\mu_q = \int v^q k(v)dv.$$

We have therefore proved that

$$\sup_{z\in[0,1]^{d_Z}}\left|\int_{[0,1]^{d_Z}} K_h(u-z)f_Z\left(u\right)du - \left\{f_Z\left(z\right) + \frac{\mu_q}{q!}\left[\sum_{\ell=1}^{d_Z}\frac{\partial^q f_Z\left(z\right)}{\partial z_\ell^q}\right]h^q\right\}\right| \leq Ch^{q+e}. \quad (37)$$

Using the above result, we have

$$\begin{aligned}
& E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right] \\
= \ & E\left\{E\left[\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma})K_h(Z_j - Z_i)|W_i\right]\right\} \\
= \ & E\left\{\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma})\left[\int_{[0,1]^{d_Z}} K_h(u-Z_i)f_Z\left(u\right)du\right]\right\} \\
= \ & E\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma})\left\{f_Z\left(Z_i\right) + \frac{\mu_q}{q!}\left[\sum_{\ell=1}^{d_Z}\frac{\partial^q f_Z\left(Z_i\right)}{\partial Z_{i\ell}^q}\right]h\right\} + o(h^q) \\
= \ & E\left[\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma})f_Z(Z_i)\right] + h^q C_1(\gamma) + o(h^q),
\end{aligned}$$

where

$$C_1(\boldsymbol{\gamma}) \equiv \frac{\mu_q}{q!} E\left\{\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma})\left[\sum_{\ell=1}^{d_Z} \frac{\partial^q f_Z\left(Z_i\right)}{\partial Z_{i\ell}^q}\right]\right\}$$

and the $o(h^q)$ term holds uniformly over $\boldsymbol{\gamma} \in \Gamma$.

Next, let

$$\psi\left(z; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) = \int_{[0,1]^{d_Y}} \varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma}) f_{YZ}(y, z) dy$$

be a function of z indexed by $(\tilde{x}, \tilde{z}, \boldsymbol{\gamma})$. Since $\varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma})$ and $f_{YZ}(y, z)$ are bounded, we can exchange differentiation with integration to obtain

$$D_z^j\left[\psi\left(z; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right)\right] = \int_{[0,1]^{d_Y}} \varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma}) D_z^j\left[f_{YZ}(y, z)\right] dy,$$

where $D_z^j[\cdot]$ is the partial differentiation operator with respect to z. So according to Assumption 2(a) and (d), $\psi\left(z; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right)$ is $q+1$ times continuously differentiable with respect to z. Furthermore, under Assumption 3 and for j with $|j| = q$, we have,

$$\begin{aligned}
& \sup_{x_1, z_1, \boldsymbol{\gamma}} \sup_{\left\|z^{(1)} - z^{(2)}\right\| \leq \epsilon} \left|D_{z^{(1)}}^j\left[\psi\left(z^{(1)}; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right)\right] - D_{z^{(2)}}^j\left[\psi\left(z^{(2)}; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right)\right]\right| \\
= \; & \sup_{x_1, z_1, \boldsymbol{\gamma}} \sup_{\left\|z^{(1)} - z^{(2)}\right\| \leq \epsilon} \int_{[0,1]^{d_Y}} \varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma}) \cdot \left|D_z^j\left[f_{YZ}(y, z^{(1)})\right] - D_z^j\left[f_{YZ}(y, z^{(2)})\right]\right| dy \\
\leq \; & \varphi_{\max} \sup_{\left\|z^{(1)} - z^{(2)}\right\| \leq \epsilon} \int_{[0,1]^{d_Y}} \left|D_z^j\left[f_{YZ}(y, z^{(1)})\right] - D_z^j\left[f_{YZ}(y, z^{(2)})\right]\right| dy \\
= \; & \varphi_{\max} \int_{[0,1]^{d_Y}} \tilde{\rho}\left(\left\|z^{(1)} - z^{(2)}\right\|\right) dy \\
\leq \; & \tilde{\rho} \varphi_{\max} \times \left\|z^{(1)} - z^{(2)}\right\|
\end{aligned}$$

for some constant $\tilde{\rho} > 0$. Therefore $\psi\left(z; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) \in \mathcal{G}_{q+1}\left([0,1]^{d_X + d_Z} \times \Gamma, \epsilon, \tilde{\rho}\varphi_{\max}\right)$. In addition, note that

$$\psi\left(z; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) = \left[\int_{[0,1]^{d_Y}} \varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma}) f_{Y|Z}(y|z) dy\right] f_Z\left(z\right),$$

which, combined with Assumption 2(b), implies that $D_z^j \psi\left(\breve{z}; x_1, z_1, \boldsymbol{\gamma}\right) = 0$ for all $\breve{z}$ on the boundary on $[0,1]^{d_Z}$. Given these two properties, we can follow the same steps in showing (37) to obtain

$$\begin{aligned}
& \int_{[0,1]^{d_Z}} K_h(u - \tilde{z}) \psi\left(u; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) du \\
= \; & \psi\left(\tilde{z}; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) + \frac{\mu_q}{q!}\left[\sum_{\ell=1}^{d_Z} \left.\frac{\partial^q \psi\left(u; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right)}{\partial u_\ell^q}\right|_{u = \tilde{z}}\right] h^q + o\left(h^q\right) \\
= \; & \psi\left(\tilde{z}; \tilde{x}, \tilde{z}, \boldsymbol{\gamma}\right) + \frac{\mu_q}{q!}\left[\sum_{\ell=1}^{d_Z} \int \varphi(\tilde{x}, y, \tilde{z}; \boldsymbol{\gamma}) \frac{\partial^q f_{YZ}(y, \tilde{z})}{\partial \tilde{z}_\ell^q} dy\right] h^q + o\left(h^q\right)
\end{aligned}$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$ and $(\tilde{x}, \tilde{z}) \in [0,1]^{d_X + d_Z}$. Using this result, we have

$$\begin{aligned}
& E\left[\varphi(X_i, Y_j, Z_i; \boldsymbol{\gamma}) K_h(Z_i - Z_j)\right] \\
= \;& E\left[\varphi(X_i, Y_j, Z_i; \boldsymbol{\gamma}) K_h(Z_j - Z_i)\right] \\
= \;& E\left\{\int K_h(u - Z_i)\left[\int \varphi(X_i, y, Z_i; \boldsymbol{\gamma}) f_{YZ}(y, u)\, dy\right] du\right\} \\
= \;& E\left\{\int K_h(u - Z_i)\psi(u; X_i, Z_i, \boldsymbol{\gamma})\, dz\right\} \\
= \;& E\psi(Z_i; X_i, Z_i, \boldsymbol{\gamma}) + C_2(\boldsymbol{\gamma}) h^q + o(h^q)
\end{aligned}$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$ where

$$C_2(\boldsymbol{\gamma}) = \frac{\mu_q}{q!} E\left\{\sum_{\ell=1}^{d_Z} \int \varphi(X_i, y, Z_i) \frac{\partial^q f_{YZ}(y, Z_i)}{\partial Z_{i\ell}^q} dy\right\}.$$

By definition, $\psi(Z_i; X_i, Z_i, \boldsymbol{\gamma}) = g_{XZ}(X_i, Z_i; \boldsymbol{\gamma})$. So

$$E\left[\varphi(X_i, Y_j, Z_i; \boldsymbol{\gamma}) K_h(Z_i - Z_j)\right] = E g_{XZ}(X_i, Z_i; \boldsymbol{\gamma}) + C_2(\boldsymbol{\gamma}) h^q + o(h^q)$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$.

Let $C_3(\boldsymbol{\gamma}) \equiv C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma})$, then

$$\begin{aligned}
\Delta_h(\boldsymbol{\gamma}) \;\equiv\;& E\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \\
=\;& E\{[\varphi(X_i, Y_i, Z_i; \boldsymbol{\gamma}) K_h(Z_i - Z_j) - \varphi(X_i, Y_j, Z_i; \boldsymbol{\gamma}) K_h(Z_i - Z_j)\} \\
=\;& E\left[\varphi(X_i, Y_i, Z_i) f_Z(Z_i)\right] + C_1(\boldsymbol{\gamma}) h^q + o(h^q) \\
& - \{E\left[g_{XZ}(X_i, Z_i; \boldsymbol{\gamma})\right] + C_2(\boldsymbol{\gamma}) h^q + o(h^q)\} \\
=\;& \Delta(\boldsymbol{\gamma}) + C_3(\boldsymbol{\gamma}) h^q + o(h^q)
\end{aligned}$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$. It then follows that under Assumption 5(b)

$$E\left[\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right] = \Delta(\boldsymbol{\gamma}) + o\left(n^{-1/2}\right)$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$.

Part (b). By definition

$$H_{n,h}(\boldsymbol{\gamma}) = \frac{2}{n} \sum_{i=1}^{n} \tilde{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}) = \frac{2}{n} \sum_{i=1}^{n} \{\kappa_{h,1}(W_i; \boldsymbol{\gamma}) - \Delta_h(\boldsymbol{\gamma})\},$$

where $\kappa_{h,1}(W_i; \boldsymbol{\gamma}) = E\left[\kappa_h(W_i, W_j; \boldsymbol{\gamma}) | W_i\right]$ for $j \neq i$. Using the same arguments in proving part (a), we have

$$\sup_{\boldsymbol{\gamma} \in \Gamma} \sup_{W_i \in [0,1]^d} \left| \kappa_{h,1}(W_i; \boldsymbol{\gamma}) - \left[\kappa_1(W_i; \boldsymbol{\gamma}) + \frac{1}{2} B_5(X_i, Y_i, Z_i; \boldsymbol{\gamma}) h^q\right]\right| \leq C h^{q+e},$$

where

$$
\begin{aligned}
& \kappa_1(W_i;\boldsymbol{\gamma}) \\
= \; & \frac{1}{2}\varphi(X_i,Y_i,Z_i;\boldsymbol{\gamma})f_Z(Z_i) - \frac{1}{2}\int \varphi(X_i,y,Z_i;\boldsymbol{\gamma})f_{YZ}(y,Z_i)dy \\
& + \frac{1}{2}\int \varphi(x,y,Z_i;\boldsymbol{\gamma})f_{XYZ}(x,y,Z_i)dxdy - \frac{1}{2}\int \varphi(x,Y_i,Z_i;\boldsymbol{\gamma})f_{XZ}(x,Z_i)dx, \quad (38)
\end{aligned}
$$

$$B_1(X_i,Y_i,Z_i;\boldsymbol{\gamma}) \equiv \frac{\mu_q}{q!}\varphi(X_i,Y_i,Z_i;\boldsymbol{\gamma})\sum_{\ell=1}^{d_Z}\frac{\partial^q f_Z\left(Z_i\right)}{\partial Z_{i\ell}^q}, \quad (39)$$

$$B_2(X_i,Z_i;\boldsymbol{\gamma}) \equiv \frac{\mu_q}{q!}\sum_{\ell=1}^{d_Z}\int \varphi(X_i,y,Z_i;\boldsymbol{\gamma})\frac{\partial^q f_{YZ}(y,Z_i)}{\partial Z_{i\ell}^q}dy, \quad (40)$$

$$B_3(Z_i;\boldsymbol{\gamma}) \equiv \frac{\mu_q}{q!}\sum_{\ell=1}^{d_Z}\int \frac{\partial^q\left[\varphi(x,y,Z_i;\boldsymbol{\gamma})f_{XYZ}(x,y,Z_i)\right]}{\partial Z_{i\ell}^q}dxdy, \quad (41)$$

$$B_4(Y_i,Z_i;\boldsymbol{\gamma}) \equiv \frac{\mu_q}{q!}\sum_{\ell=1}^{d_Z}\int \frac{\partial^q\left[\varphi(x,Y_i,Z_i;\boldsymbol{\gamma})f_{XZ}(x,Z_i)\right]}{\partial Z_{i\ell}^q}dx, \quad (42)$$

and

$$B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma}) = B_1(X_i,Y_i,Z_i;\boldsymbol{\gamma}) - B_2(X_i,Z_i;\boldsymbol{\gamma}) - B_4(Y_i,Z_i;\boldsymbol{\gamma}) + B_3(Z_i;\boldsymbol{\gamma}). \quad (43)$$

It is easy to see that $E\kappa_1(W_i;\boldsymbol{\gamma}) = \Delta\left(\boldsymbol{\gamma}\right)$. So

$$
\begin{aligned}
H_{n,h}(\boldsymbol{\gamma}) &= \frac{2}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma}) + \frac{1}{2}B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})h^q\right\} - \Delta_h\left(\boldsymbol{\gamma}\right) \\
&= \frac{2}{n}\sum_{i=1}^{n}\left[\kappa_1(W_i;\boldsymbol{\gamma}) - E\kappa_1(W_i;\boldsymbol{\gamma})\right] + \frac{1}{2n}\sum_{i=1}^{n}B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})h^q \\
&\quad - \left(\Delta_h\left(\boldsymbol{\gamma}\right) - \Delta\left(\boldsymbol{\gamma}\right)\right) + o(h^q)
\end{aligned}
$$

where the $o(h^q)$ term holds uniformly over $\boldsymbol{\gamma} \in \Gamma$.

Since $B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})$ is continuous in γ, $E\sup_{\boldsymbol{\gamma}\in\Gamma}\left|B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})\right| < \infty$, (X_i,Y_i,Z_i) is IID, and Γ is compact, we can use a standard textbook argument to show that a ULLN applies to $n^{-1}\sum_{i=1}^{n}B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})$. That is, $\sup_{\boldsymbol{\gamma}\in\Gamma}\left|n^{-1}\sum_{i=1}^{n}B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})\right| = O(1)$. Combining this with part (a), we have

$$
\begin{aligned}
H_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma}) - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\} + O_p(h^q) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma}) - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\} + o_p(\frac{1}{\sqrt{n}})
\end{aligned}
$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$.

■

Proof of Theorem 3. As a direction implication of Lemmas 1 and 2, we have

$$\begin{aligned}&\sqrt{n}\left[\hat{\Delta}_{n,h}(\Gamma_s)-\Delta(\Gamma_s)\right]\\ =\ &\frac{2}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s)-E\left[\kappa_1(W_i;\Gamma_s)\right]\right\}+o_p\left(1\right)\end{aligned}$$

uniformly over $\gamma\in\Gamma$. The asymptotic normality now follows by applying the Lindeberg-Levy CLT.

If in addition H_0 holds, then $\Delta(\Gamma_s)=0$ and

$$\begin{aligned}\kappa_1(W_i;\gamma)&=\frac{1}{2}\varphi(X_i,Y_i,Z_i;\boldsymbol{\gamma})f_Z(Z_i)-\frac{1}{2}\int\varphi(X_i,y,Z_i;\boldsymbol{\gamma})f_{YZ}(y,Z_i)dy\\&\quad+\frac{1}{2}\int\varphi(x,y,Z_i;\boldsymbol{\gamma})f_{XYZ}(x,y,Z_i)dxdy-\frac{1}{2}\int\varphi(x,Y_i,Z_i;\boldsymbol{\gamma})f_{XZ}(x,Z_i)dx,\\(under\ H_0)&=\frac{1}{2}E\left[\varphi\left(X_i,Y_i,Z_i;\boldsymbol{\gamma}\right)f_Z(Z_i)|X_i,Y_i,Z_i\right]-\frac{1}{2}E\left[\varphi\left(X_i,Y_i,Z_i;\boldsymbol{\gamma}\right)f_Z(Z_i)|X_i,Z_i\right]\\&\quad+\frac{1}{2}E\left[\varphi\left(X_i,Y_i,Z_i;\boldsymbol{\gamma}\right)f_Z(Z_i)|Z_i\right]-\frac{1}{2}E\left[\varphi\left(X_i,Y_i,Z_i;\boldsymbol{\gamma}\right)f_Z(Z_i)|Y_i,Z_i\right]\\&=\Lambda(W_i;\boldsymbol{\gamma}).\end{aligned}$$

Thus, given H_0 we have

$$\Omega\left(\ell,m\right)=4E\left[\Lambda(W_i;\boldsymbol{\gamma}_\ell)\Lambda(W_i;\boldsymbol{\gamma}_m)\right].$$

■

Proof of Theorem 4: Given Lemmas 1 and 2, it suffices to prove part (a). Theorem 3 shows that for a finite number of $\boldsymbol{\gamma}$'s, $\{\zeta_n(\boldsymbol{\gamma}_1),\zeta_n(\boldsymbol{\gamma}_2),\ldots,\zeta_n(\boldsymbol{\gamma}_s)\}$ is asymptotically normal. Also, $\boldsymbol{\gamma}\in\Gamma\subset\mathbb{R}^{1+d}$ with Γ a compact (hence totally bounded) set. To complete the proof, we need to show that $\zeta_n(\boldsymbol{\gamma})$ is stochastically equicontinuous (e.g., see Andrews, 1994). For this, we use Theorems 4–6 in Andrews (1994). In view of the definition of $\kappa_1(W_i;\boldsymbol{\gamma})$ in (38) and Theorem 6 in Andrews (1994), we only need to verify that each of the four terms satisfies Ossiander's L^2 entropy condition.

For the first term in (38), $\varphi\left(W_i;\boldsymbol{\gamma}\right)f_Z(Z_i)$ belongs to the type IV class if we can verify that

$$E\left\{\left[f_Z(Z_i)\right]^2\sup_{\boldsymbol{\gamma}_1:\|\boldsymbol{\gamma}_1-\boldsymbol{\gamma}\|<\nu}\left|\varphi\left(W_i;\boldsymbol{\gamma}_1\right)-\varphi\left(W_i;\boldsymbol{\gamma}\right)\right|^2\right\}\leq C\nu^{\psi}\tag{44}$$

for any $\boldsymbol{\gamma}\in\Gamma$, for any $\nu>0$ in a neighborhood of 0, and for some finite constants $C>0$ and $\psi>0$. Under Assumption 3, $\varphi\left(W_i;\boldsymbol{\gamma}\right)$ is differentiable in $\boldsymbol{\gamma}$. Given that

$$E\left\|f_Z(Z_i)\sup_{\boldsymbol{\gamma}\in\Gamma}\partial\left[\varphi\left(W_i;\boldsymbol{\gamma}\right)/\partial\boldsymbol{\gamma}\right]\right\|^2<\infty$$

and Γ is bounded, we can show that (44) holds by the mean value theorem and Cauchy-Schwarz inequality.

Similarly, we can show that the other three terms in $\kappa_1(W_i;\boldsymbol{\gamma})$ also belong to the type IV class. Hence $\zeta_n(\cdot) \overset{d}{\to} \mathcal{Z}(\cdot)$.

$\blacksquare$

Proof of Proposition 5: Since

$$\begin{aligned}
&\sup_{\gamma}\left|\frac{\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_n(\boldsymbol{\gamma})}-\frac{\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})}{\sigma_{\Delta}(\boldsymbol{\gamma})}\right| \\
\leq\ &\sup_{\gamma}\left[\sqrt{n}\hat{\Delta}_{n,h}(\boldsymbol{\gamma})\right]\sup_{\gamma}\left|\frac{\hat{\sigma}_n(\boldsymbol{\gamma})-\sigma_{\Delta}(\boldsymbol{\gamma})}{\hat{\sigma}_n(\boldsymbol{\gamma})}\right|\sup_{\gamma}\frac{1}{\sigma_{\Delta}(\boldsymbol{\gamma})},
\end{aligned}$$

it suffices to show that $\sup_{\gamma}|1-\sigma_{\Delta}(\boldsymbol{\gamma})/\hat{\sigma}_n(\boldsymbol{\gamma})| = o_p(1)$. Under the given conditions, this follows from the proof of uniform consistency of $\hat{\Omega}$ given in Huang (2009, Ch. 1, Theorem 6). $\blacksquare$

References

[1] Andrews, D.W.K. (1994), "Empirical Process Methods in Econometrics," in Engle, R.F. and McFadden, D.L. (eds.), Handbook of Econometrics, vol. IV. Amsterdam: Elsevier, pp. 2248–2296.

[2] Barnow, B.S., Cain, G.G., Goldberger, A.S. (1981), "Issues in the Analysis of Selectivity Bias," in Stromsdorfer, W.E. and Farkas, G. (eds.), Evaluation studies review annual, Vol. 5. Beverly Hills, CA: Sage, pp. 43–59.

[3] Bierens, H.J. (1982), "Consistent Model Specification Tests," Journal of Econometrics, 20,105-134.

[4] Bierens, H.J. (1990), "A Consistent Conditional Moment Test of Functional Form," Econometrica, 58, 1443–1458.

[5] Bierens, H.J., Ploberger, W. (1997), "Asymptotic Theory of Integrated Conditional Moment Tests," Econometrica, 65, 1129–1151.

[6] Billingsley, P. (1999). Convergence of Probability Measures. New York, NY: John Wiley & Sons, Inc.

[7] Blackburn, M., Neumark, D. (1993), "Omitted-Ability Bias and the Increase in the Return to Schooling," Journal of Labor Economics, vol. 11, 521–544.

[8] Boning, W.B., Sowell, F. (1999), "Optimality for the integrated conditional moment test," Econometric Theory, Vol. 15, 710–718.

[9] Dawid, A.P. (1979), "Conditional Independence in Statistical Theory," Journal of the Royal Statistical Society, Series B 41, 1-31.

[10] Delgado, M., Gonzalez-Manteiga, W. (2001), "Significance Testing in Nonparametric Regression Based on the Bootstrap," Annals of Statistics, 29, 1469–1507.

[11] Fernandes, M., Flores, R. (2002), "Tests for Conditional Independence, Markovian Dynamics and Noncausality," European University Institute Discussion Paper.

[12] Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems," Econometrica, 45, l–22.

[13] Griliches, Z., Mason, W.M., (1972), "Education, Income, and Ability," The Journal of Political Economy, Vol. 80, No. 3, pp. S74–S103.

[14] Hansen, B. E. (1996), "Inference when a Nuisance Parameter is Not Identified under the Null Hypothesis," Econometrica, 64, 413–430.

[15] Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," Annals of Mathematical Statistics, 19, 293–325.

[16] Huang, M. (2009), "Essays On Testing Conditional Independence," Ph.D. dissertation, University of California, San Diego. Available at http://escholarship.org/uc/item/15t6n3h6.

[17] Lee, A.J. (1990), U-statistics: Theory and Practice. New York: CRC Press.

[18] Li, Q. and Racine, J. S. (2007), Nonparametric Econometrics, Princeton University Press.

[19] Linton, O., and Gozalo, P. (1997), "Conditional Independence Restrictions: Testing and Estimation," Yale University Cowles Foundation for Research in Economics Discussion Paper.

[20] Mincer, J. (1974), Schooling, Experience, and Earnings. New York: Columbia University Press.

[21] Powell, J.L., Stock, J.H., Stoker, T.M. (1989), "Semiparametric Estimation of Index Coefficients," Econometrica, 57, 1403–1430.

[22] Powell, J.L., Stoker, T.M. (1996), "Optimal Bandwidth Choice for Density-weighted Averages", Journal of Econometrics, 75, 291–316.

[23] Robinson, P. M. (1988), "Root-N-consistent Semiparametric Regression," Econometrica, 56, 931–954.

[24] Song, K. (2009), "Testing Conditional Independence Via Rosenblatt Transforms," Annals of Statistics, 37, 4011–4045.

[25] Stinchcombe, M., White, H. (1998), "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative," Econometric Theory, 14, 295–324.

[26] Su, L., White, H. (2003), "Testing Conditional Independence Via Empirical Likelihood," UCSD Department of Economics Discussion Paper.

[27] Su, L., White, H. (2007), "A Consistent Characteristic Function-Based Test for Conditional Independence," Journal of Econometrics, 141, 807–834.

[28] Su, L., White, H. (2008), "A Nonparametric Hellinger Metric Test for Conditional Independence," Econometric Theory, 24, 829-864.

[29] Su, L., White, H. (2010), "Testing Structural Change in Partially Linear Models," Econometric Theory, 26, 1761–1806.

[30] White, H., Chalak, K. (2008), "Identifying Structural Effects in Nonseparable Systems Using Covariates," UCSD Department of Economics Discussion Paper.

[31] White, H., Chalak, K. (2009), "Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning," Journal of Machine Learning Research, 10, 1759–1799.

[32] White, H., Chalak, K. (2010), "Testing a Conditional Form of Exogeneity," Economics Letters, 109, 88–90.

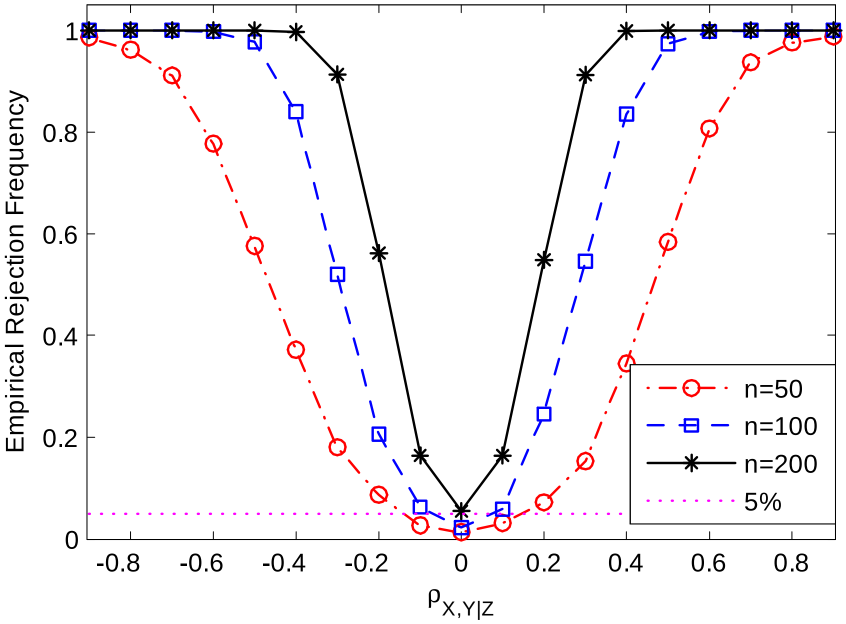

Figure 1: Power functions of non-standardized ICM test (M_n) for DGP 1 with nominal size 5%

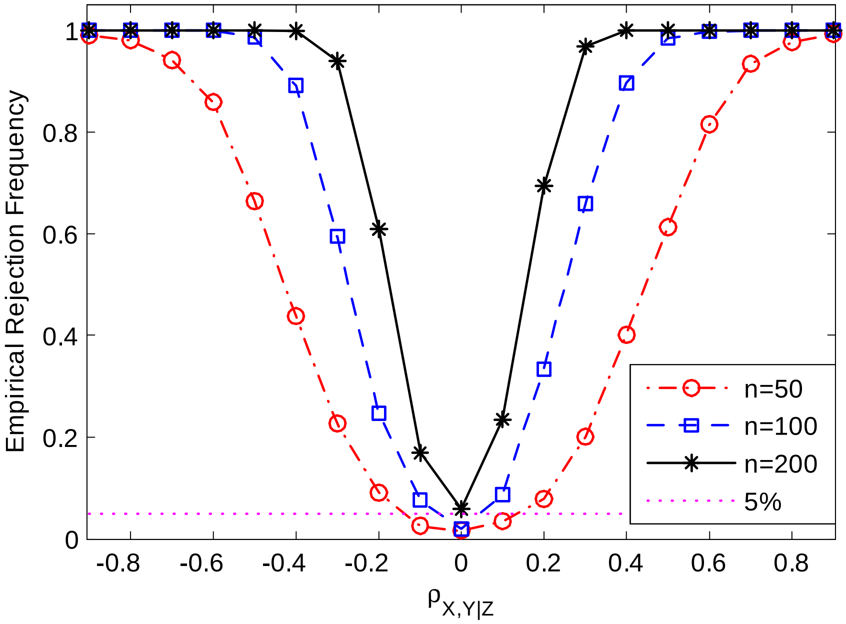

Figure 2: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 1 with nominal size 5%

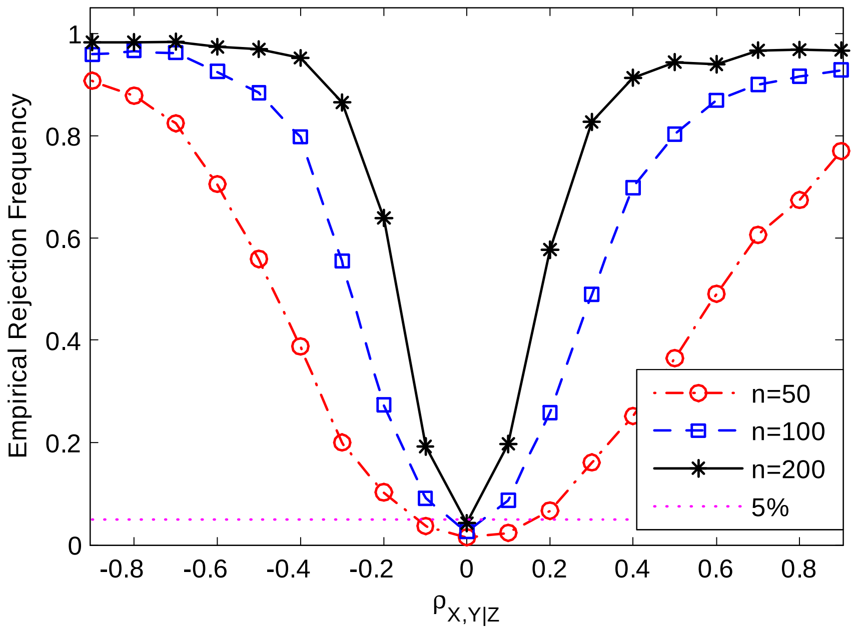

Figure 3: Power functions of non-standardized ICM test (M_n) for DGP 2 with nominal size 5%

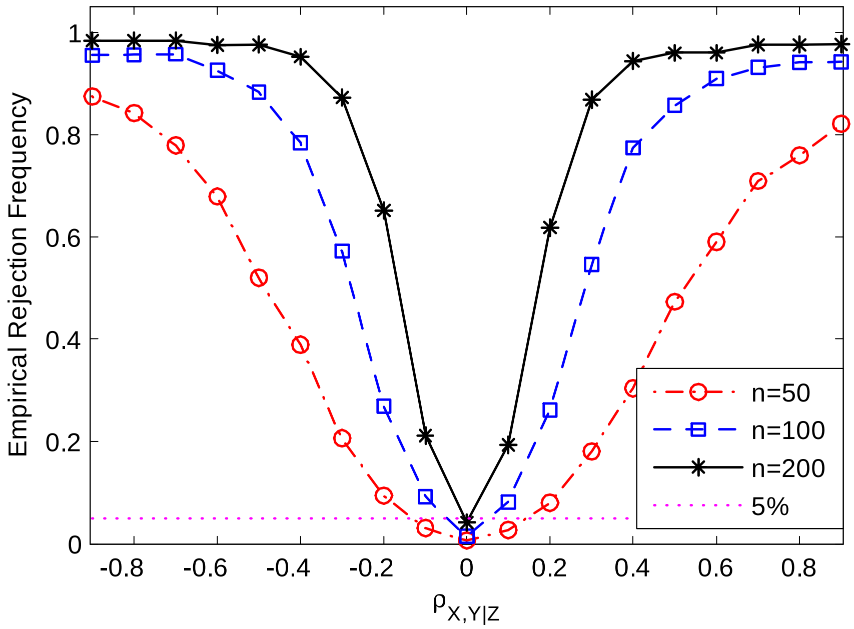

Figure 4: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 2 with nominal size 5%

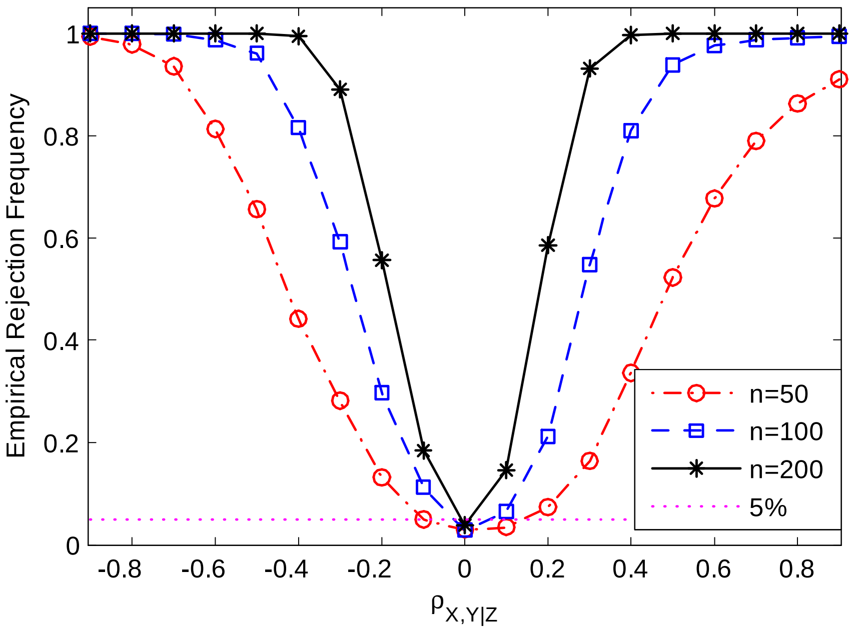

Figure 5: Power functions of non-standardized ICM test (M_n) for DGP 3 with nominal size 5%

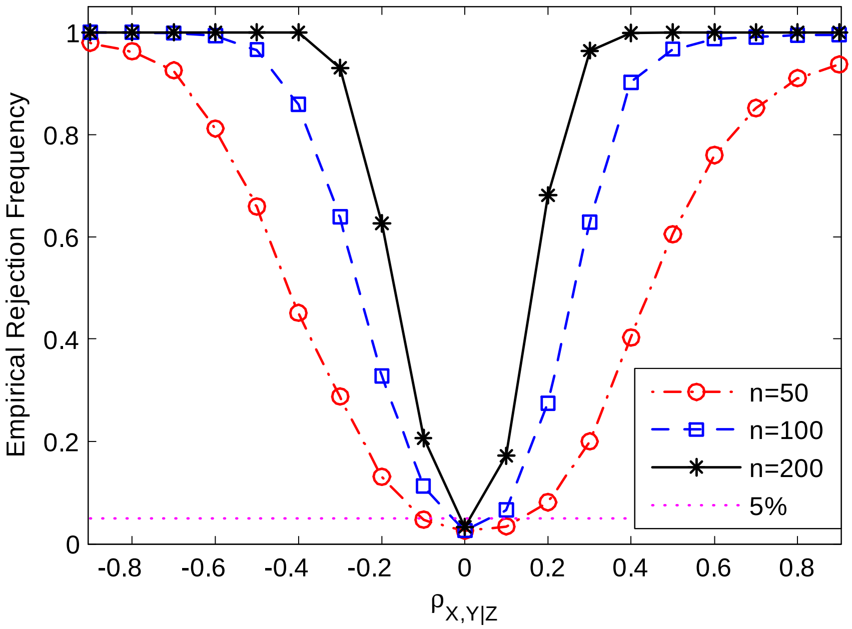

Figure 6: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 3 with nominal size 5%

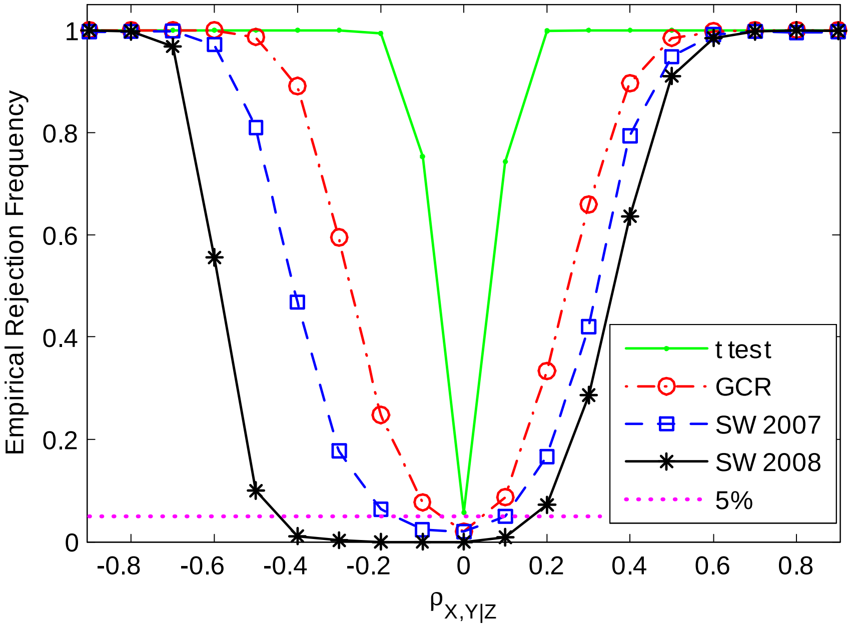

Figure 7: Power functions of the 5% standardized GCR test, SW2007 test, and SW2008 test under DGP1 with sample size 100

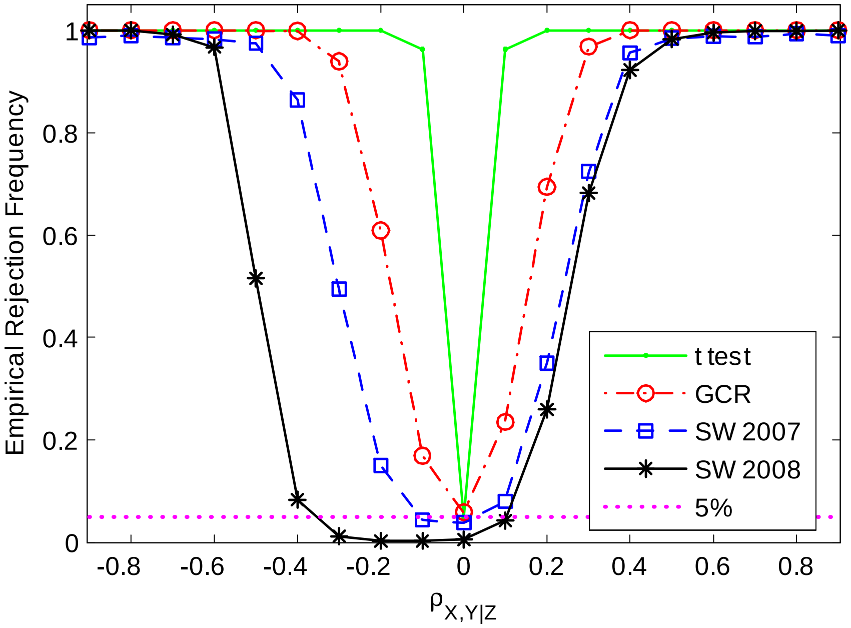

Figure 8: Power functions of the 5% standardized GCR test, SW2007 test, and SW2008 test under DGP1 with sample size 200

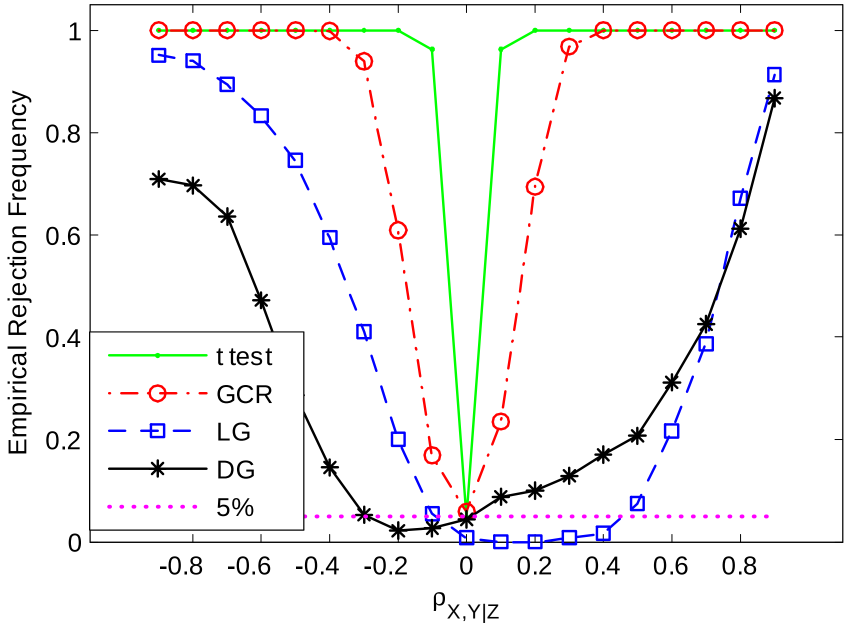

Figure 9: Power functions of the 5% standardized GCR test, LG 1997 test, and DG 2001 test under DGP1 with sample size 200