UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays On Testing Conditional Independence

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Meng Huang

Committee in charge:

Professor Halbert White, Chair
Professor Graham Elliott
Professor Patrick J. Fitzsimmons
Professor Dimitris N. Politis
Professor Yixiao Sun

2009

The dissertation of Meng Huang is approved, and it is acceptable in quality and form for publication on microfilm and eletronically:

Chair

University of California, San Diego

2009

DEDICATION

To my parents and my husband.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am deeply indebted to my advisor, Halbert White, for his constant inspiration, effective guidance, generous advice and steady support. The dissertation would not exist without his help. I am amazingly lucky to have him as my advisor. I am especially grateful to Yixiao Sun, for his keen insights, thoughtful suggestions and detailed instructions. I have learned so much from him. I would sincerely thank Graham Elliott, for his helpful criticism and constructive comments. I have also been especially fortunate to have Dimitris N. Politis and Patrick J. Fitzsimmons on my committee, their intellectual comments and great patience are truly appreciated. I also benefit from discussions with Liangjun Su, Jin Seo Cho, Xun Lu, James D. Hamilton, Andres Santos, Brendan Beare and my UCSD classmates, and I would like to thank them as well. Helpful comments from participants at the Econometrics seminars at UC-San Diego, Hong Kong University of Science, Beijing University and Bates White are also gratefully acknowledged. All remaining errors are mine.

VITA

1999	B.A., International Finance Nankai University, Tianjin, China
2002	M.A., Finance Nankai University, Tianjin, China
2003	M.A., Economics The Ohio State University, Columbus
2009	Ph.D., Economics University of California, San Diego

ABSTRACT OF THE DISSERTATION

Essays On Testing Conditional Independence

by

Meng Huang

Doctor of Philosophy in Economics
University of California, San Diego, 2009

Professor Halbert White, Chair

Conditional independence is of interest for testing unconfoundedness assumptions in causal inference, for selecting among semiparametric models, and for testing Granger noncausality, etc. This dissertation propose flexible tests for conditional independence, which are simple to implement yet powerful in the sense that they are consistent and achieve $\sqrt{n}$ local power.

In the literature, there are many tests available for the case in which the variables are categorical. But there are only a few nonparametric tests for the continuous case. On the other hand, in economics applications, it is common to condition on continuous variables. Chapter 1 provides a nonparametric test for continuous variables. The test statistic is a Wald type test based on an estimator of the topological "distance" between the 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 indexed by a nuisance parameter vector.

Although the test in chapter 1 is easy to calculate and has a tractable limiting null distribution, its consistency relies on the randomization of the choice of the nuisance parameters. In chapter 2, I obtain a Bierens type Integrated Conditional Moment test by integrating out the nuisance parameters. The test still achieves $\sqrt{n}$ local power and its consistency does not rely on the randomization any more. Its limiting null distribution is a functional of a mean zero Gaussian process. I simulate the critical values by a conditional simulation approach. As an example of application, I test the key assumption of unconfoundedness in the context of estimating the returns to schooling.

In applied microeconomics, many variables are categorical or binary. For example, in the returns-to-schooling example, the conditioning variables usually include a number of discrete variables such as sex, race, union or industry. However, in previous chapters I assume the conditioning variables to be continuous. In chapter 3, I extend the conditional independence tests to incorporate the case of mixed conditioning random variables, using the frequency approach and the smoothing approach.

1

A Chi-square Test for Conditional Independence

1.1 Introduction

In this chapter, I propose a 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, which can be denoted as

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

Intuitively, this means that given the information in Z, X cannot provide additional information to predict Y. The notation and definition of conditional independence is as given by Dawid (1979).

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.

In economics, conditional independence is often used as a part of the identifying restrictions, as in the identification for nonseparable models. It is also of interest for selecting among semiparametric models and for testing Granger noncausality, as mentioned in Su and White (2008). Another potential application of conditional independence testing is to test a key assumption identifying causal effects. For example, if we are interested in estimating the effect of additional year of schooling on income, but we are worried that unobserved ability will affect both years of schooling and income, then OLS will generally fail to give us a consistent estimator. Nevertheless, if we can find a set of covariates, say AFQT scores, such that ability will be independent of schooling given the covariates, then we can estimate the returns to schooling by various methods provided in the causal inference literature. The conditional independence assumption is a key assumption identifying the causal effect. Since ability is unobservable, we cannot directly test this assumption using the proposed test. But if there is another set of observable covariates satisfying certain conditions (see White and Chalak (2006)), we may test an implied conditional independence assumption: the new set of observable covariates is independent of schooling given the AFQT scores.

In the literature, there are many tests for conditional independence for the case in which the variables are categorical. But there are only a few nonparametric tests for the continuous case. On the other hand, in economics applications, it is common to condition on continuous variables. Previous work on testing conditional independence for continuous random variables includes that of Linton and Gozalo (1997), Fernandez and Flores (2002), and Delgado and Gonzalez-Manteiga (2001). More recently, Su and White have several papers (Su and White (2003), Su and White (2007) and Su and White (2008), “SW”) attacking 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 $1/\sqrt{n}$ and hence suffer from the "curse of dimensionality".

The test proposed here achieves $\sqrt{n}$ local power. The philosophy of the test follows a series of papers of consistent specification tests in Bierens (1982), Bierens (1990), Bierens and Ploberger (1997) and Stinchcombe and White (1998), among others. Whereas Bierens (1982), Bierens (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 Stinchcombe and White (1998), basing the test on estimators of the topological "distance" between unrestricted and restricted *probability measures*, corresponding to conditional independence or its absence.

This distance between the two probability measures is measured indirectly by a family of moments, which use *Generically Comprehensively Revealing* (GCR) functions such as the logistic cumulative distribution function (CDF) or the normal probability density function (PDF) as the test functions, indexed by a nuisance parameter vector $\boldsymbol{\gamma}$. Under the null, all such moments are zeroes. under the alternatives, the use of GCR functions makes the test consistent, because of its property that any deviation from the null will result in nonzero moments for essentially all choices of $\boldsymbol{\gamma}$.

I estimate these moments by their sample analogs, using kernel smoothing. A Wald type test statistic based on these estimators is obtained and its limiting distribution is a Chi-square distribution.

The plan of the remaining paper is as follows. In section 2, I explain the hypothesis and the basic idea of my test for conditional independence, and derive an equivalent null hypothesis which is based on the topological distance between the restricted and unrestricted models. I then suggest an estimator for that distance and a test statistic based on that. In section 3, I establish asymptotic normality of the properly centered and scaled moment estimators using extended U-statistic theory, and prove the test is consistent. I use a simple plug-in bandwidth following Powell and Stoker (1996). In section 4, I perform simulations to examine the finite sample properties of the test. Section 5 concludes.

1.2 The Hypothesis and the Test Statistic

1.2.A The Hypothesis

Let X, Y, and Z be three random vectors, with dimensions d_X, d_Y, and d_Z, respectively. For convenience, I assume throughout that the sample observations $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ are independent and identically distributed (IID). Analogous results hold under weaker conditions, but I leave consideration of these aside here. Formally, I make the following assumption:

Assumption 1.1 $\{W_i \equiv (X_i', Y_i', Z_i')'\}$ *is an IID sequence of random variables on the complete probability space* $(\Omega_W, \mathcal{F}_W, P_W)$. $X_i, Y_i,$ *and* Z_i *take values in* $\mathbb{R}^{d_X}$, $\mathbb{R}^{d_Y}$*,and* $\mathbb{R}^{d_Z}$, *respectively, and* $d_W \equiv d_X + d_Y + d_Z$.

We exploit the identical distribution assumption to drop the i subscript when convenient, and we write $W \equiv (X, Y, Z)$. The null hypothesis to be tested is that Y is independent of X given Z, i.e.

$$H_0 : Y \perp X \mid Z, \tag{1.2.1}$$

whereas the alternative is that Y and X are dependent conditioning on Z, i.e.

$$H_a : Y \not\perp X \mid Z. \tag{1.2.2}$$

The definition of conditional independence is as given by Dawid (1979). We say that Y is independent of X given Z, if for all x

$$f_{Y|XZ}(y \mid x, z) = f_{Y|Z}(y \mid z), \tag{1.2.3}$$

where $f_{Y|XZ}(y \mid x, z)$and $f_{Y|Z}(y \mid z)$ denote the conditional density of Y given $(X, Z) = (x, z)$ and the conditional density of Y given $X = x$, respectively. The intuition of the meaning of conditional independence is that given Z, X cannot provide further information to predict Y. Note that (1.2.3) and the following three expressions are equivalent to one another:

$$f_{X|YZ}(x \mid y, z) = f_{X|Y}(x \mid y), \tag{1.2.4}$$

$$f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z)\ f_{Y|Z|}(y \mid z), \tag{1.2.5}$$

and

$$f_{XYZ}(x, y, z)\ f_Z(z) = f_{XZ}(x, z)\ f_{YZ}(y, z). \tag{1.2.6}$$

From (1.2.5) and (1.2.6) we see clearly that the definition is symmetric for X and Y. So we use "X and Y are independent given Z", "X is independent of Y given Z" and "Y is independent of X given Z" interchangeably, and we use "$Y \perp X \mid Z$" and "$X \perp Y \mid Z$" interchangeably.

As mentioned in the introduction, there are many conditional independence tests in the literature designed for categorical random variables. So in this paper, I am particularly interested in the case for continuous random variables.

1.2.B An Equivalent Hypothesis in the Form of Moment Conditions

For testing the null hypothesis (1.2.1), what I will do is to test an equivalent hypothesis which is a family of moment conditions. Since we can easily estimate the moment conditions by their sample analogs, we can construct a test statistic based on these estimators. I will first establish that equivalent hypothesis.

The idea is inspired by a series of papers of consistent specification tests: Bierens (1982), Bierens (1990), Bierens and Ploberger (1997), and Stinchcombe and White (1998), among others. Those tests 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 $g(x) \equiv E(Y \mid X = x)$. Bierens tests the hypothesis that the parametric functional form, $f(x, \theta)$, 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 estimators 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 zeroes for all $\boldsymbol{\gamma}$. lemma 1 in Bierens (1990) shows that the converse essentially holds, due to the properties of the exponential function, making possible a test able to detect all deviations from the null.

Stinchcombe and White (1998) find that there is a broader class of functions having this property. They extend Bierens's result by replacing the exponential function in the moment conditions by any *Generically Comprehensively Revealing* (GCR) function, e.g. nonpolynomial analytic functions, and by extending the probability measures considered in the Bierens's approach (Bierens (1990)) to signed measures. Further, they point out that such specification tests are based on estimates of topological distance between a restricted and an unrestricted model. Following this idea, we can construct a test for conditional independence based on estimates of the topological distance between unrestricted and restricted probability measures corresponding to conditional independence or its absence.

We first restate the definition of GCR function from Stinchcombe and White (1998). We let $C(B)$ denote the set of continuous functions on a compact set $B \subset \mathbb{R}^{d_W}$, and we write $sp\ H_\varphi(\Gamma)$ to denote the span of a collection of functions $H_\varphi(\Gamma)$.

Definition 1 *(Definition 3.6 in Stinchcombe and White (1998)) We say that* $H_\varphi = \left\{h : \mathbb{R}^{d_W} \to \mathbb{R} \mid h(w) = \varphi\left(\tilde{w}'\boldsymbol{\gamma}\right),\ \boldsymbol{\gamma} \in \Gamma \subset \mathbb{R}^{1+d_W}, \tilde{w} := (1, w')'\right\}$ *is generically comprehensively revealing if for all* Γ *with non-empty interior, the uniform closure of* $sp\ H_\varphi(\Gamma)$ *contains* $C(B)$ *for every compact set* $B \subset \mathbb{R}^{d_W}$.

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 a GCR if the generated H_φ is a GCR. As stated in Stinchcombe and White (1998), GCR functions include real analytic functions except for the polynomials, e.g. exp, logistic CDF, sine, cosine, and also some nonanalytic functions like the normal CDF or its density.

I now establish an equivalent hypothesis in the form of a family of moment conditions following the argument of Stinchcombe and White (1998). Let P be the joint distribution[1] of the random vector $W = (X, Y, Z)$, 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,

$$P(A) = \int 1[(x, y, z) \in A] dF_{XYZ}(x, y, z) \tag{1.2.7}$$

and

$$Q(A) \equiv \int 1[(x, y, z) \in A] dF_{Y|Z}(y|z) dF_{XZ}(x, z), \tag{1.2.8}$$

where $1[\cdot]$ is an indicator function, and $dF_{Y|Z}(y \mid z)$, $dF_{XZ}(x, z)$, $dF_{XYZ}(x, y, z)$ denote the conditional or joint densities indicated by their subscripts and arguments.

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

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

To test the null hypothesis is thus equivalent to test whether there is any deviation of P from Q.

Let E_P and E_Q denote expectations with respect to P and Q, respectively, and define

$$\Delta_{\varphi}(\boldsymbol{\gamma}) \equiv E_P(\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)) - E_Q(\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)),$$

where $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1', \gamma_2', \gamma_3')' \in \mathbb{R}^{1+d_W}$ is a vector of nuisance parameters. Under the null hypothesis, $\Delta_{\varphi}(\boldsymbol{\gamma})$ is obviously zero for any choice of function φ including

[1] The distribution of X is the measure $F := P \circ X^{-1}$ on $(R, \beta(R))$ defined by $(A \in \beta(R))$, $F(A) = P \circ X^{-1} = P[X \in A]$.

GCR functions and for any choice of $\boldsymbol{\gamma}$. To build up a test, we want $\Delta_{\varphi}(\boldsymbol{\gamma})$ to be nonzero under the alternatives. 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 of $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. An arbitrary function may fail to detect some alternatives. Nevertheless, according to Stinchcombe and White (1998), if W is a bounded random vector, the properties of GCR functions imply that GCR functions can detect all alternatives for essentially any choice of $\boldsymbol{\gamma} \in \Gamma \subset \mathbb{R}^{1+d_W}$ with Γ having non-empty interior. "Essentially any" $\boldsymbol{\gamma} \in \Gamma$ means that the set of "bad" $\boldsymbol{\gamma}$'s, $\{\boldsymbol{\gamma} \in \Gamma{:}\Delta_{\varphi}(\boldsymbol{\gamma}) = 0 \text{ and } Y \not\perp X \mid Z\}$ i.e. the $\boldsymbol{\gamma}$'s which cannot detect the alternative with the choice of test function φ, has Lebesgue measure zero and is not dense in Γ.

This remarkable result implies that any deviation of P from Q can be detected by essentially any choice of $\boldsymbol{\gamma} \in \Gamma$, provided we choose φ to be a GCR function. In other words, to test $H_0 : Y \perp X \mid Z$ is equivalent to test that

$$H_0 : \Delta_{\varphi}(\boldsymbol{\gamma}) = 0 \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma \tag{1.2.9}$$

against

$$H_A : H_0 \text{ is false,}$$

where φ is GCR and Γ has non-empty interior.

Note that to get the result, we require W to be bounded. But as in Bierens (1990) and Stinchcombe and White (1998), boundedness can be ensured by replacing W by $\Phi(W)$, where Φ is a bounded one-to-one mapping such as $\Phi(W) = \Phi(W^{(1)}, W^{(2)}, ..., W^{(d_W)}) = [\tan^{-1}(W^{(1)}), \tan^{-1}(W^{(2)}), ..., \tan^{-1}(W^{(d_W)})]$. This replacement will not affect the conditional independence (or its absence) since the sigma fields will not be affected by this transformation. We leave this transformation implicit in this paper to make the notation simpler.

A straightforward test would be to estimate $\Delta_{\varphi}(\boldsymbol{\gamma})$ and to see how far it is from zero. But if we proceed in that way, we will encounter a nonparametric estimator $\hat{f}_Z$ of the density f_Z in the denominator of the estimator, making the asymptotic normality of the estimator hard to prove.

For this technical reason, I replace φ by $\varphi^* \equiv \varphi f_Z$ to avoid $\hat{f}_Z$ appearing in the denominator. To ensure the continued validity of the test, I need to show that φ^* is still a GCR function under certain assumptions. For that purpose, I extend Stinchcombe and White (1998)'s definition of generically comprehensively revealing functions to permit multiplication by a function.

Definition 2 *Let $d_W \in \mathbb{N}$, and let $\varphi : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R}^{d_W} \to \mathbb{R}$ be measurable functions. For $\Gamma \subset \mathbb{R}^{1+d_W}$, let $H_{\varphi f}(\Gamma) \equiv \{h : \mathbb{R}^{d_W} \to \mathbb{R} \mid h(w) = \varphi(\tilde{w}'\boldsymbol{\gamma})f(w), \boldsymbol{\gamma} \in \Gamma\}$, where $w \in \mathbb{R}^{d_W}$ and $\tilde{w} \equiv (1, w')'$. We say that $H_{\varphi f}$ is generically comprehensively revealing (GCR) if for all Γ with non-empty interior, the uniform closure of sp $H_{\varphi f}(\Gamma)$ contains $C(B)$ for every compact set $B \subset \mathbb{R}^{d_W}$.*

When $H_{\varphi f}$ is a GCR, we also say that φf is a GCR. The following proposition shows that φf_Z is still a GCR if φ is a GCR and the density f_Z satisfies some conditions.

Proposition 1 *Let Z be a $d_Z \times 1$ sub-vector of the random $d_W \times 1$ vector W, and suppose that f_Z, the density of Z, is continuous, positive, and bounded on supp(Z). Let φ be a GCR. Then $\varphi^* \equiv \varphi f_Z$ is a GCR.*

I impose these conditions in the following assumption.

Assumption 1.2 f_Z, the density of Z, is continuous, positive, and bounded on *supp*(Z).

Then I conclude that a null hypothesis equivalent to conditional independence is

$$H_0 : \Delta\left(\boldsymbol{\gamma}\right) = 0 \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma, \tag{1.2.10}$$

where

$$\begin{aligned} & \Delta\left(\boldsymbol{\gamma}\right) \\ \equiv\; & E_P\left[\varphi^{*}(W;\boldsymbol{\gamma})\right] - E_Q\left[\varphi^{*}(W;\boldsymbol{\gamma})\right] \qquad (1.2.11) \\ =\; & E_P\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] \\ & -E_Q\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right], \end{aligned}$$

φ is a GCR, and $\Gamma \subset \mathbb{R}^{1+d_W}$ with Γ having non-empty interior. By "essentially all" $\boldsymbol{\gamma} \in \Gamma$, I mean that the set $\{\boldsymbol{\gamma} \in \Gamma{:}\Delta\left(\boldsymbol{\gamma}\right) = 0 \text{ and } Y \not\perp X \mid Z\}$ has Lebesgue measure zero and is not dense in Γ. $\Delta\left(\boldsymbol{\gamma}\right) = 0$ is a moment condition, as it is the difference of two expectations. The null hypothesis of conditional independence is thus equivalent to a family of moment conditions indexed by $\boldsymbol{\gamma}$.

1.2.C Estimators of the Moments and a Wald Type Test Statistic

Since we are testing whether $\Delta\left(\boldsymbol{\gamma}\right)$ is zero or not, it is natural to construct a test statistic based on an estimator of $\Delta\left(\boldsymbol{\gamma}\right)$. In the following, I will use a sample analog $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ to estimate $\Delta\left(\boldsymbol{\gamma}\right)$, and then propose to use a finite collection of $\boldsymbol{\gamma}$'s to construct a Chi-square test statistic.

First, Ie construct a sample analog to estimate $\Delta\left(\boldsymbol{\gamma}\right)$. Note that $\Delta\left(\boldsymbol{\gamma}\right)$ is the difference between two expectations:

$$\begin{aligned}
& \Delta\left(\boldsymbol{\gamma}\right) \\
\equiv\ & E_P\left[\varphi^*(W;\boldsymbol{\gamma})\right] - E_Q\left[\varphi^*(W;\boldsymbol{\gamma})\right] \\
=\ & E_P\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] \\
& -E_Q\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] \\
=\ & \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)dF_{XYZ}(x,y,z) \qquad \text{(1.2.12)} \\
& -\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)f_{Y|Z}(y|z)dy\ dF_{XZ}(x,z) \\
\equiv\ & E_{X,Y,Z}\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] - \int g_{XZ}(x,z;\boldsymbol{\gamma})dF_{XZ}(x,z) \\
=\ & E_{X,Y,Z}\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] - E_{X,Z}\left[g_{XZ}(X,Z;\boldsymbol{\gamma})\right]
\end{aligned}$$

where

$$\begin{aligned}
g_{XZ}(x,z;\boldsymbol{\gamma}) &\equiv \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)f_{Y|Z}(y,z)dy \\
&= E\left[\varphi(\gamma_0 + x'\gamma_1 + Y'\gamma_2 + z'\gamma_3)f_Z(z)|Z=z\right].
\end{aligned}$$

The first term of $\Delta\left(\boldsymbol{\gamma}\right)$ is a mean of φf_Z, where φ is chosen by us and f_Z can be estimated by a kernel smoothing method. The second term is a mean of g_{XZ}, where the function $g_{XZ}(x,z;\boldsymbol{\gamma})$ is a conditional expectation that can be estimated by a Nadaraya Watson estimator. Thus I propose to use the following estimator to estimate $\Delta\left(\boldsymbol{\gamma}\right)$:

$$\begin{aligned}
\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\hat{f}_Z(Z_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{XZ}(X_i,Z_i;\boldsymbol{\gamma}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}K_h(Z_i - Z_j)\right] \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right] \\
&= \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) \qquad \text{(1.2.13)} \\
&\quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3]K_h(Z_i - Z_j)\},
\end{aligned}$$

where $K_h(\cdot)$ is defined as

$$K_h(u) \equiv \frac{1}{h^{d_u}} K(\frac{u}{h}),$$

with $K(\cdot)$ a symmetric product kernel density function, d_u the dimension of u, and the bandwidth $h \equiv h_n$ depending on n.

Intuitively, $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ should be close to zero under the null and nonzero under the alternatives for essentially all $\boldsymbol{\gamma} \in \Gamma$. By choosing multiple $\boldsymbol{\gamma}$'s we can get a vector of such estimators. For convenience, I denote the vector $[\bar{\Delta}_{n,h_1}(\boldsymbol{\gamma}_1)$, $\bar{\Delta}_{n,h_2}(\boldsymbol{\gamma}_2)$, $...\bar{\Delta}_{n,h_s}(\boldsymbol{\gamma}_s)]'$ by $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$, where $\Gamma_s \equiv \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_s\}$ is the set of chosen $\boldsymbol{\gamma}$'s, and $\mathbf{h} = (h_1, h_2, ..., h_s)'$ is the corresponding set of chosen bandwidths. The choice of smoothing kernel K could in principle depend on $\boldsymbol{\gamma}$ too, but here I use the same kernel K as this choice will not affect the results significantly and I want to keep the notation simple. Similarly, I define $\Delta(\Gamma_s) \equiv [\Delta(\boldsymbol{\gamma}_1)$,$\Delta(\boldsymbol{\gamma}_2)$, ..., $\Delta(\boldsymbol{\gamma}_s)]'$, etc. In the next section, I will show that after proper centering and scaling, $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ converges to an s-dimension normal distribution. Given $\hat{\Omega}$, a consistent estimator of the variance-covariance matrix, I can construct a Wald type test statistic

$$S_n(\Gamma_s) \equiv n \left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]' \hat{\Omega}^{-1} \left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]. \tag{1.2.14}$$

And the test is consistent if we choose Γ_s randomly from a distribution which is absolutely continuous with respect to Lebesgue measure on Γ.

1.3 Asymptotic Behavior of the Test Statistic

To show that the proposed test statistic $S_n(\Gamma_s)$ follows a Chi-square asymptotic distribution under the null, the key is to show $\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is asymptotically normal. Note that $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is not an average of independent random variables because of the double summation. Indeed, it is a U-statistic of degree two. I apply asymptotic theory from the literature on U-statistics (see Hoeffding (1948), Lee (1990), Powell et al. (1989), among others) to show its asymptotic normality. In particular, I apply the extended U-statistic theory developed in Powell et al. (1989). The idea is to use the H-decomposition (Hoeffding (1948)) to decompose $\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ into three parts: its mean vector, $\Delta_{\mathbf{h}}(\Gamma_s)$; an average of independent zero mean random vectors, $2H_{n,\mathbf{h},1}(\Gamma_s)$; and a residual, $R_{n,\mathbf{h},1}(\Gamma_s)$. The first two parts constitute the "projection" of $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$. I first show that the residual is small enough so that $\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ has the same limiting distribution as its projection, and I then derive the asymptotic normality of the projection.

Before I proceed, I first list the assumptions I will use immediately:

Assumption 2 (Kernel function) Let $q \geq 2$ be an even integer. The kernel K is a product of a symmetric qth order univariate kernel $k : \mathbb{R} \to \mathbb{R}$ s.t. $\int k(v)dv = 1$, $\int v^j k(v)dv = 0$ for $j = 1, 2, ...q - 1$, and $0 < \int v^q k(v)dv < \infty$.

Assumption 3 Z_i takes values in the interior of the support of Z, $i = 1, 2, ...$.

Assumption 4 (Smoothness of the densities) The density of Z, f_Z, is continuously differentiable of order $q + 1$; and all partial derivatives of $f_{YZ}(y, z)$, $f_{XZ}(y, z)$, $f_{XYZ}(x, y, z)$ with respect to z of order q exist.

Assumption 5 $\varphi(\cdot)$ is a bounded GCR function.

Assumption 6 (Bandwidth) The bandwidth $h_l \equiv h_{l,n}$, for $l = 1, 2, ..., s$, satisfies

(A 6.1) $nh_l^{dz} \to \infty$ as $n \to \infty$, and

(A 6.2) $\sqrt{n}h_l^q = o(1)$, i.e. $h_l = o(n^{-1/(2q)})$ as $n \to \infty$.

Remark 1 *I impose Assumption 3 to avoid the boundary bias problem. I could relax this assumption and use boundary kernels. For now I assume interior points to simplify the problem. To ensure this condition, we may need to trim the data when* Z *is boundedly supported. In that case, I use* $K_h(Z_i - Z_j) \cdot 1[Z_i \in Z_\varepsilon] \cdot 1[Z_j \in Z_\varepsilon]$ *with* $Z_\varepsilon \subset Supp(Z)$ *defined by* $P[Z_i \in Z_\varepsilon] = 1 - \varepsilon$, $\varepsilon > 0$ *small, instead of* $K_h(Z_i - Z_j)$*, and modify all proofs accordingly.*

Remark 2 *In Assumption 4, we do not assume that the marginal distributions of* X *and* Y *are smooth. In fact,* X *and* Y *could be discrete or continuous.*

Remark 3 *Assumption 5 is stronger than necessary. We only need certain moments of* $\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)$ *bounded. For example, if* φ *is continuous, we don't need it to be bounded since* W *is bounded. But Assumption 5 is a convenient one to ensure those conditions, and we can easily choose a bounded GCR function e.g. normal PDF.,* sin *or* cos.

Remark 4 *In Assumption 6, (A 6.1) ensures that the residual* $R_{n,\mathbf{h},1}(\Gamma_s)$ *is small, while (A 6.2) is used to kill the bias of* $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ *as* $n \to \infty$. *Together these conditions imply that* $2q > d_Z$.

1.3.A H-decomposition

We observe that $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) = [\bar{\Delta}_{n,h_1}(\boldsymbol{\gamma}_1), \bar{\Delta}_{n,h_2}(\boldsymbol{\gamma}_2), ...\bar{\Delta}_{n,h_s}(\boldsymbol{\gamma}_s)]'$ is a vector U-statistic of degree 2, where each element of $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ has the form

$$\begin{aligned}\bar{\Delta}_{n,h}(\boldsymbol{\gamma}_l) &= \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_{l0}+X_i'\gamma_{l1}+Y_i'\gamma_{l2}+Z_i'\gamma_{l3}) \\ &\quad -\varphi(\gamma_{l0}+X_i'\gamma_{l1}+Y_j'\gamma_{l2}+Z_i'\gamma_{l3}]K_h(Z_i-Z_j)\} \qquad (1.3.1)\\ &= \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1}\sum_{(n,2)}\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l),\end{aligned}$$

where $\kappa_h(W_i, W_j; \boldsymbol{\gamma}_l)$ is a symmetric kernel in W_i and W_j defined as

$$\begin{aligned} & \kappa_h(W_i,W_j;\boldsymbol{\gamma}_l) \\ \equiv\ & \frac{1}{2}[\varphi(\gamma_{l0}+X_i'\gamma_{l1}+Y_i'\gamma_{l2}+Z_i'\gamma_{l3}) \\ & -\varphi(\gamma_{l0}+X_i'\gamma_{l1}+Y_j'\gamma_{l2}+Z_i'\gamma_{l3})]K_h(Z_i-Z_j) \\ & +\frac{1}{2}[\varphi(\gamma_{l0}+X_j'\gamma_{l1}+Y_j'\gamma_{l2}+Z_j'\gamma_{l3}) \\ & -\varphi(\gamma_{l0}+X_j'\gamma_{l1}+Y_i'\gamma_{l2}+Z_j'\gamma_{l3})]K_h(Z_j-Z_i) \\ =\ & \kappa_h(W_j,W_i;\boldsymbol{\gamma}_l),\end{aligned}$$

and $l = 1, 2, ...s$, and $\sum_{(n,2)}$ means the summation is over different pairs of i and j.

The idea is to use H-decomposition, named after its inventor Hoeffding, to decompose $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ into three parts:

$$\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) = \Delta_{\mathbf{h}}(\Gamma_s) + 2H_{n,\mathbf{h},1}(\Gamma_s) + R_{n,\mathbf{h},1}(\Gamma_s), \qquad (1.3.2)$$

where

$$\Delta_{\mathbf{h}}(\Gamma_s) \equiv E\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right] \qquad (1.3.3)$$

$$\begin{aligned} H_{n,\mathbf{h},1}(\Gamma_s) &\equiv \frac{1}{n}\sum_{i=1}^{n}\tilde{\kappa}_{\mathbf{h},1}(W_i;\Gamma_s),\ \text{with}\ \{\tilde{\kappa}_{\mathbf{h},1}(W_i;\Gamma_s)\}\ \text{IID}, \\ & \tilde{\kappa}_{\mathbf{h},1}(W_i;\Gamma_s) \equiv \kappa_{\mathbf{h},1}(W_i;\Gamma_s) - \Delta_{\mathbf{h}}\left(\Gamma_s\right) \qquad (1.3.4) \\ & \kappa_{\mathbf{h},1}(W_i;\Gamma_s) \equiv E\left[\kappa_{\mathbf{h}}(W_i,W_j;\Gamma_s)|W_i\right]\ i\neq j,\end{aligned}$$

and

$$R_{n,\mathbf{h},1}\left(\Gamma_s\right) \equiv \bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta_{\mathbf{h}}\left(\Gamma_s\right) - 2H_{n,\mathbf{h},1}\left(\Gamma_s\right). \tag{1.3.5}$$

The subscript 1 in the above notations denotes that the item is a projection on the first argument of $\kappa_{\mathbf{h}}$, W_i. The first part $\Delta_{\mathbf{h}}(\Gamma_s)$ is the mean part, which depends on $\mathbf{h}$. When $\mathbf{h}$ is small, It turns out that $\Delta_{\mathbf{h}}(\Gamma_s)$ is equal to $\Delta(\Gamma_s)$ plus some small terms. $2H_{n,\mathbf{h},1}(\Gamma_s)$ is a leading term, which is an average of IID random variables whose asymptotic behavior is straightforward to derive. And $R_{n,\mathbf{h},1}\left(\Gamma_s\right)$ is the remainder. Both $H_{n,\mathbf{h},1}\left(\Gamma_s\right)$ and $R_{n,\mathbf{h},1}\left(\Gamma_s\right)$ have zero means and are uncorrelated to each other. The projection $\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is defined as

$$\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) \equiv \Delta_{\mathbf{h}}(\Gamma_s) + 2H_{n,\mathbf{h},1}(\Gamma_s). \tag{1.3.6}$$

The idea is to first show the remainder is small so that the projection is the leading term, and then to derive the asymptotics of the projection.

Remark 5 *According to the U-statistic theory (e.g. Lee (1990)),*

$$\begin{aligned} R_{n,\mathbf{h},1}\left(\boldsymbol{\gamma}_l\right) &\equiv \bar{\Delta}_{n,\mathbf{h}}(\boldsymbol{\gamma}_l) - \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}_l\right) - 2H_{n,\mathbf{h},1}\left(\boldsymbol{\gamma}_l\right) \\ &= \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum_{(n,2)} \left[\tilde{\kappa}_{h,2}(W_i, W_j; \boldsymbol{\gamma}_l)\right] \end{aligned} \tag{1.3.7}$$

is another U-statistic of degree 2 *with the kernel*

$$\tilde{\kappa}_{h,2}(W_i, W_j; \boldsymbol{\gamma}_l) \equiv \left[\kappa_h(W_i, W_j; \boldsymbol{\gamma}_l) - \tilde{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}_l) - \tilde{\kappa}_{h,1}(W_j; \boldsymbol{\gamma}_l) - \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}_l\right)\right].$$

And

$$E\left[R_{n,\mathbf{h},1}\left(\boldsymbol{\gamma}_l\right)\right] = E\left[\tilde{\kappa}_{h,2}(W_i, W_j; \boldsymbol{\gamma}_l)\right] = 0,$$

$$\begin{aligned} Var\left[R_{n,\mathbf{h},1}\left(\boldsymbol{\gamma}_l\right)\right] &= \frac{2}{n(n-1)}\left[2\left(n-2\right)\xi_1^2 + \xi_2^2\right] \\ &= \frac{2\xi_2^2}{n(n-1)} \end{aligned}$$

where

$$\begin{aligned}
\xi_1^2 &\equiv VAR\{E\left[\tilde{\kappa}_{h,2}(W_i, W_j; \boldsymbol{\gamma}_l)|W_j\right]\} \\
&= VAR\{E\left[\kappa_h(W_i, W_j; \boldsymbol{\gamma}_l) - \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}_l\right)|W_j\right] - E\left[\tilde{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}_l)\right] - \tilde{\kappa}_{h,1}(W_j; \boldsymbol{\gamma}_l)\} \\
&= VAR\{\tilde{\kappa}_{h,1}(W_j; \boldsymbol{\gamma}_l) - 0 - \tilde{\kappa}_{h,1}(W_j; \boldsymbol{\gamma}_l)\} \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\xi_2^2 &\equiv VAR\left[\tilde{\kappa}_{h,2}(W_i, W_j; \boldsymbol{\gamma}_l)\right] \\
&= E\left[\kappa_h(W_i, W_j; \boldsymbol{\gamma}_l) - \tilde{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}_l) - \tilde{\kappa}_{h,1}(W_j; \boldsymbol{\gamma}_l) - \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}_l\right)\right]^2.
\end{aligned}$$

So for a fixed h, the remainder $R_{n,\mathbf{h},1}\left(\boldsymbol{\gamma}_l\right)$ has a smaller order variance than the leading term $2H_{n,\mathbf{h},1}(\Gamma_s)$.

1.3.B $\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ Has the Same Limiting Distribution As Its Projection $\sqrt{n}\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$

We have seen that if $\mathbf{h}$ were fixed, $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ would be a conventional U-statistic and $R_{n,\mathbf{h},1}\left(\Gamma_s\right) = \bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ would be $o_p\left(n^{-1/2}\right)$. But now the bandwidth vector $\mathbf{h} \equiv \mathbf{h}_n$ is going to 0 as $n \to \infty$, so we need the theory for the extended U-statistics. I apply Lemma 3.1 in Powell et al. (1989) to show that $R_{n,\mathbf{h}_n,1}\left(\Gamma_s\right)$ is still $o_p\left(n^{-1/2}\right)$ if we properly control $\mathbf{h}_n$ so that it does not shrink too fast. I summarize the result precisely in the next lemma.

Lemma 2 *Under Assumptions 1-5, if each h_l $(l = 1, 2, ..., s)$ satisfies assumption 6.1, i.e. $nh_l^{dz} \to \infty$, and $h_l \to 0$ as $n \to \infty$, then $\sqrt{n}\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right] = o_p(1)$.*

1.3.C Asymptotic Distribution of $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$

Given the asymptotic equivalence of $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ and $\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$, the remaining task is to show that $\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ has a $\sqrt{n}$ limiting normal distribution with mean $\Delta(\Gamma_s)$. Note that both $\Delta_{\mathbf{h}}(\Gamma_s)$ and $H_{n,\mathbf{h},1}(\Gamma_s)$ depend on $\mathbf{h}$. When $\mathbf{h}$ is fixed, $\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is obviously normal, but with a bias depending on $\mathbf{h}$. Using Taylor expansion, I can separate the parts independent of $\mathbf{h}$ and the parts associated with $\mathbf{h}$ in $\Delta_{\mathbf{h}}(\Gamma_s)$ and $H_{n,\mathbf{h},1}(\Gamma_s)$. I then use a higher order kernel K and control the rate of $\mathbf{h}$ to shrink fast enough so that the parts associated with $\mathbf{h}$ will vanish asymptotically, as is the case in Powell et al. (1989).

I now discuss the results in detail. In the following lemma, I first show that $\Delta_{\mathbf{h}}(\Gamma_s) = \Delta(\Gamma_s) + O(\mathbf{h}^q)$, where q is the order of the kernel K and $\mathbf{h}^q = (h_1^q, h_2^q, ..., h_s^q)'$. Then I show that $H_{n,\mathbf{h},1}(\Gamma_s) = n^{-1}\sum_{i=1}^{n}\{\kappa_1(W_i;\Gamma_s) - E[\kappa_1(W_i;\Gamma_s)]\} + O_p(\mathbf{h}^q)$, where

$$\begin{aligned}\kappa_1(W_i;\boldsymbol{\gamma}) \equiv\ & \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i) \\ & -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{YZ}(y, Z_i)dy \\ & +\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{XYZ}(x, y, Z_i)dxdy \\ & -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x, Z_i)dx.\end{aligned} \tag{1.3.8}$$

Under Assumption 6.2, $\sqrt{n}h_l^q \to 0$ for $l = 1, 2, ..., s$, which makes both the second term of $\Delta_{\mathbf{h}}(\Gamma_s)$ and the second term of $H_{n,\mathbf{h},1}(\Gamma_s)$ vanish asymptotically. The leading term of $H_{n,\mathbf{h},1}(\Gamma_s)$ is an average of IID random variables independent of h, which obeys the Lindeberg-Levy Central Limit Theorem. The following lemma summarizes the facts that the projection $\sqrt{n}\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is asymptotically normal with mean zero under the null and diverges under the alternatives.

Lemma 3 *Under Assumptions 1-5 and if each h_l $(l = 1, 2, ..., s)$ satisfies assump-*

tion 6.2, i.e. $\sqrt{n}h_l^q \to 0$, *then* $\Delta_{\mathbf{h}}(\Gamma_s) \equiv E\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right] = \Delta(\Gamma_s) + o\left(n^{-1/2}\right)$ *and* $H_{n,\mathbf{h},1}(\Gamma_s) = n^{-1}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + o_p\left(n^{-1/2}\right)$, *hence*

$$\sqrt{n}\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) = \sqrt{n}\Delta(\Gamma_s) + 2n^{-1/2}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + o_p(1).$$

So

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

where

$$\Omega(l,k) \equiv \sigma_\Delta(\boldsymbol{\gamma}_l,\boldsymbol{\gamma}_k) = 4cov\left[\kappa_1(W_i;\boldsymbol{\gamma}_l),\kappa_1(W_i;\boldsymbol{\gamma}_k)\right]. \quad (1.3.9)$$

If in addition H_0 *holds, then* $\Delta(\Gamma_s) = 0$, $\sqrt{n}\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) \xrightarrow{d} N(0,\Omega)$, *and*

$$cov\left[\kappa_1(W_i;\boldsymbol{\gamma}_l),\kappa_1(W_i;\boldsymbol{\gamma}_k)\right] = E\left[\Lambda(W_i;\boldsymbol{\gamma}_l)\Lambda(W_i;\boldsymbol{\gamma}_k)\right],$$

where

$$\begin{aligned}\Lambda(W_i;\boldsymbol{\gamma}) \;=\; & \frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i,Y_i,Z_i\right] \quad (1.3.10)\\ & -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i,Z_i\right]\\ & +\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Z_i\right]\\ & -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Y_i,Z_i\right].\end{aligned}$$

From the H-decomposition and the results in lemma 2 and 3, we can conclude immediately that $\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is also asymptotically normal with mean zero under the null and diverges under the alternatives. I state these results in the following theorem.

Theorem 4 *Under Assumptions 1-6,*

$$\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) = \sqrt{n}\Delta(\Gamma_s) + 2n^{-1/2}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + o_p(1).$$

So

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

where Ω *is defined as in (1.3.9). If in addition* H_0 *holds, then* $\Delta(\Gamma_s) = 0$ *and*

$$cov\left[\kappa_1(W_i; \boldsymbol{\gamma}_l), \kappa_1(W_i; \boldsymbol{\gamma}_k)\right] = E\left[\Lambda(W_i; \boldsymbol{\gamma}_l)\Lambda(W_i; \boldsymbol{\gamma}_k)\right];$$

hence

$$\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) \xrightarrow{d} N(0, 4E\left[\Lambda(W_i; \Gamma_s)\Lambda(W_i; \Gamma_s)'\right]).$$

The consistency of $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ as an estimator of $\Delta(\Gamma_s)$ is a by-product of the previous theorem, which is stated as a corollary in the following.

Corollary 5 *Under Assumptions 1-6,* $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) \xrightarrow{P} \Delta(\Gamma_s)$.

1.3.D A Consistent Variance-Covariance Matrix Estimator

To construct a Chi-square test statistic based on theorem 4, we need a consistent estimator for the variance-covariance matrix Ω. Similar to the case in Powell et al. (1989), the U-statistic theory we used before also suggests a natural estimator for the asymptotic variance. Note that

$$\begin{aligned} E[\kappa_1(W_i; \boldsymbol{\gamma})] &= \frac{1}{2}E_{XYZ}[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)] \qquad (1.3.11) \\ &\quad -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{YZ}(y,z)f_{XZ}(x,z)dydxdz \\ &\quad +\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XYZ}(x,y,z)f_Z(z)dxdydz \\ &\quad -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XZ}(x,Z_i)f_{YZ}(y,z)dxdydz \\ &= E_{XYZ}[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)] - E_{X,Z}\left[g(X,Z;\gamma)\right] \\ &= \Delta(\boldsymbol{\gamma}), \end{aligned}$$

so

$$\begin{aligned}\Omega &= 4cov\left[\kappa_1(W_i;\Gamma_s)\kappa_1(W_i;\Gamma_s)'\right] \\ &= 4E\left[\kappa_1(W_i;\Gamma_s)\kappa_1(W_i;\Gamma_s)'\right] - 4\Delta(\Gamma_s)\Delta(\Gamma_s)'.\end{aligned}$$

I can use

$$\hat{\kappa}_{\mathbf{h},1}(W_i;\Gamma_s) \equiv (n-1)^{-1}\sum_{j=1,j\neq i}^{n}\kappa_{\mathbf{h}}(W_i,W_j;\Gamma_s)$$

to replace the unknown $\kappa_1(W_i;\Gamma_s)$, then the proposed estimator of Ω is

$$\begin{aligned}\hat{\Omega} &= 4n^{-1}\sum_{i=1}^{n}\left[\hat{\kappa}_{h,1}(W_i;\Gamma_s)\hat{\kappa}_{h,1}(W_i;\Gamma_s)'\right] - 4\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)' \\ &= 4n^{-1}\sum_{i=1}^{n}\left\{\left[(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\kappa_{\mathbf{h}}(W_i,W_j;\Gamma_s)\right]\right. \qquad (1.3.12) \\ &\quad \times\left.\left[(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\kappa_{\mathbf{h}}(W_i,W_j;\Gamma_s)\right]'\right\} - 4\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)'.\end{aligned}$$

The following theorem establishes the consistency of $\hat{\Omega}$.

Theorem 6 *Under Assumptions 1-6,* $\hat{\Omega} \xrightarrow{P} \Omega$.

As $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) \xrightarrow{P} 0$ under the null hypothesis, I can immediately get another estimator of Ω which is consistent only under the null:

Corollary 7 *Under the Assumptions 1-6 and* H_0*,*

$$\begin{aligned}\tilde{\Omega} &\equiv 4n^{-1}\sum_{i=1}^{n}\left[\hat{\kappa}_{h,1}(W_i;\Gamma_s)\hat{\kappa}_{h,1}(W_i;\Gamma_s)'\right] \\ &\xrightarrow{P} \Omega.\end{aligned}$$

Remark 6 *Since* $E\left[\sqrt{n}\left(\bar{\Delta}_n(\boldsymbol{\gamma}) - \Delta\right)\right] = E\left[B_5(W_i;\boldsymbol{\gamma})\right]\sqrt{n}h^q + o(\sqrt{n}h^q)$ $= O\left(\sqrt{n}h^q\right)$ *, the order of the kernel* $K\left(\cdot\right)$ *will affect the bias of the mean. And* *since we require $nh^{d_Z} \to \infty$, $\sqrt{n}h$ through $\sqrt{n}h^{\frac{d_Z}{2}}$ will explode. Hence we require the order q is bigger than $\frac{d_Z}{2}$ to make the bias term vanish asymptotically. On the other hand, the choice of $K(\cdot)$ does not affect Ω, the leading term of the variance. So it looks reasonable to use higher order kernel to reduce the bias. We may even consider an infinity order kernel like $k_\infty(u) = \frac{1+\cos u - 2\cos^2 u}{\pi u^2}$. Then we will need to modify our theorems and proofs. And of course we need to assume accordingly the higher smoothness of the densities.*

Remark 7 *The asymptotic variance-covariance matrix of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_1(W_i;\Gamma_s)$ would be cov $[\kappa_1(W_i;\Gamma_s)\kappa_1(W_i;\Gamma_s)']$. Now that*

$$\begin{aligned}
\sqrt{n}\bar{\Delta}_{n,h}(\Gamma_s) &= \sqrt{n}\begin{pmatrix} n \\ 2 \end{pmatrix}^{-1}\sum_{(n,2)}\kappa_h(W_i,W_j;\Gamma_s) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\kappa_h(W_i,W_j;\Gamma_s)\right\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\kappa}_{h,1}(W_i;\Gamma_s).
\end{aligned}$$

Due to the dependence between $\{\hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma})\}_{i=1}^{n}$, the asymptotic variance-covariance matrix of $\sqrt{n}\bar{\Delta}_{n,h}(\Gamma_s)$ turns out to be four times bigger than that of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_1(W_i;\Gamma_s)$.

1.3.E The Chi-square Test Is Consistent

$\sqrt{n}\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ is asymptotically normal with mean zero under the null and diverges under the alternatives, so I can use these facts to construct a Chi-square test statistic. As proposed in the previous subsections, the test statistic is

$$S_n(\Gamma_s) \equiv n\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]'\hat{\Omega}^{-1}\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]. \tag{1.3.13}$$

It has a limiting Chi-square distribution of degree s if Ω is not degenerate. Let's impose the following assumption:

Assumption 7 Ω is positive definite.

This and the consistency of $\hat{\Omega}$ ensures that $\hat{\Omega}$ is positive definite with probability approaching one.

The following result defines a Wald type test statistic.

Corollary 8 *(Chi-square Test) Under Assumptions 1-7,*

$$n\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta(\Gamma_s)\right]' \hat{\Omega}^{-1} \left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta(\Gamma_s)\right] \xrightarrow{d} \chi^2\left(s\right)$$

If in addition H_0 holds, then

$$S_n(\Gamma_s) \equiv n\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]' \hat{\Omega}^{-1} \left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right] \xrightarrow{d} \chi^2\left(s\right)$$

Thus $S_n(\Gamma_s)$ is a Wald type test statistic whose asymptotic null distribution is a Chi-square distribution with degree s.

As the sample size goes to infinity, $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$ goes to $\Delta(\Gamma_s) \equiv [\Delta(\boldsymbol{\gamma}_1)$, $\Delta(\boldsymbol{\gamma}_2)$, ... $\Delta(\boldsymbol{\gamma}_s)]'$. Under the null, $\Delta(\Gamma_s)$ is zero. under the alternatives, each element of $\Delta(\Gamma_s)$, $\Delta(\boldsymbol{\gamma}_l)$, is nonzero for essentially all $\boldsymbol{\gamma}_l \in \Gamma$, since the GCR function can detect any difference between the restricted and unrestricted probability measures for essentially all $\boldsymbol{\gamma}_l \in \Gamma$. That means $S_n(\Gamma_s)$ goes to infinity for essentially all $\Gamma_s \in \Gamma$ under the alternatives. The test is thus consistent against all alternatives with probability one if we choose Γ_s randomly from a distribution which is absolutely continuous with respect to Lebesgue measure on Γ.

Proposition 9 *Under assumptions 1-7 and assume φ is a GCR function, then for essentially all $\Gamma_s \in \Gamma$,*

$$\lim_{n\to\infty} \Pr(S_n(\Gamma_s) \in R) = \lim_{n\to\infty} \Pr(|S_n(\Gamma_s)| > C_a) = 1 \tag{1.3.14}$$

where R represents the rejection region of the test, and C_a represents the critical value of a test of size a.

1.3.F The Symmetry Problem

We observe that $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ (hence $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)$) is not symmetric in X and Y, whereas the hypothesis $Y \perp X \mid Z$ is. From the previous subsection, we see that $\sqrt{n}\left(\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta_h(\boldsymbol{\gamma})\right) = \sqrt{n}\left(\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta(\boldsymbol{\gamma})\right) + o\left(1\right) \overset{d}{\to} N(0, \sigma^2_\Delta\left(\boldsymbol{\gamma}\right))$ where $\sigma^2_\Delta\left(\boldsymbol{\gamma}\right) = 4VAR\left[\kappa_1\left(W_i;\boldsymbol{\gamma}\right)\right]$. Note that Y and X are symmetric in both

$$\begin{aligned}
\kappa_1\left(W_i;\boldsymbol{\gamma}\right) \;\equiv\;& \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_Z(Z_i) \\
& -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3) f_{YZ}(y, Z_i) dy \\
& +\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3) f_{XYZ}(x, y, Z_i) dxdy \\
& -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_{XZ}(x, Z_i) dx
\end{aligned} \tag{1.3.15}$$

and in

$$\begin{aligned}
\Delta(\boldsymbol{\gamma}) \;&\equiv\; E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_Z(Z_i)\right] - E\left[g_{XZ}(X_i, Z_i; \gamma)\right] \\
&= E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_Z(Z_i)\right] \\
&\quad - \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_{YZ}(y, z) f_{XZ}\left(x, z\right) dxdydz \\
&= E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_Z(Z_i)\right] - E\left[g_{YZ}(Y_i, Z_i; \boldsymbol{\gamma})\right].
\end{aligned}$$

In fact, if we construct another estimator $\tilde{\Delta}_n(\boldsymbol{\gamma})$ by switching the roles of X and Y, we will find that this is asymptotically equivalent to $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$. Define

$$\begin{aligned}\tilde{\Delta}_n(\boldsymbol{\gamma}) \quad &\equiv \quad \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3) \\ &\qquad -\varphi(\gamma_0+X_j'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3]K_h(Z_i-Z_j)\}. \qquad (1.3.16)\end{aligned}$$

Inspecting the proofs in the previous subsections, we can see that

$$\begin{aligned}\sqrt{n}\tilde{\Delta}_n(\boldsymbol{\gamma}) \quad &= \quad \sqrt{n}\Delta(\boldsymbol{\gamma})+\sqrt{n}2H_{n,h,1}(\boldsymbol{\gamma})+o_p\left(1\right) \\ &= \quad \sqrt{n}\Delta(\boldsymbol{\gamma})+n^{-1/2}\sum_{i=1}^{n}\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\}+o_p\left(1\right).\end{aligned}$$

The leading term, $\sqrt{n}\Delta(\boldsymbol{\gamma})+n^{-1/2}\sum_{i=1}^{n}\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\}$, is the same as in $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$, which follows from the symmetry of $\Delta(\boldsymbol{\gamma})$ and $\kappa_1(W_i;\boldsymbol{\gamma})$ in X and Y. As a result, $\sqrt{n}\tilde{\Delta}_n(\boldsymbol{\gamma})$ has the same asymptotic distribution as $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ under both the null and the alternatives. Further, the asymptotic correlation of $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $\sqrt{n}\tilde{\Delta}_n(\boldsymbol{\gamma})$ is 1. So, if we consider a weighted average of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $\tilde{\Delta}_n(\boldsymbol{\gamma})$, the resulting new test statistic would have the same asymptotic distribution as $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ under both the null and the alternatives for any choice of the weights. For a symmetry of the test statistic, we may take the average of $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $\sqrt{n}\tilde{\Delta}_n(\boldsymbol{\gamma})$. But to make the notation simpler, we don't do this here.

1.3.G The Bandwidth Selection

Although theoretically speaking the specific choice of bandwidth **h** will not affect the first order results as long as it satisfies the order restrictions in assumption 6, in practice we need some guidance on how to select **h**. Ideally we should select **h** that would give us the greatest power given particular size, but that procedure would be complicated enough for another study and only make difference in the higher order results. Therefore, for the purpose of this study, I provide a simple "plug-in" estimator of the MSE-minimizing bandwidth proposed by Powell and Stoker (1996).

Since the test statistic is based on $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$, which is an estimator of $\Delta(\boldsymbol{\gamma})$, it is reasonable to choose h which minimizes the mean squared error (MSE) of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$.

After some tedious yet straightforward calculation, I get

$$\begin{aligned}
MSE\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] &= \left(\Delta_h(\boldsymbol{\gamma}) - \Delta(\boldsymbol{\gamma})\right)^2 + VAR\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \\
&= \left\{E\left[B_5(W_i;\boldsymbol{\gamma})\right]h^q + o(h^q)\right\}^2 + VAR\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \\
&= \left\{E\left[B_5(W_i;\boldsymbol{\gamma})\right]\right\}^2 h^{2q} + o(h^{2q}) + VAR\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \\
&= \left\{E\left[B_5(W_i;\boldsymbol{\gamma})\right]\right\}^2 h^{2q} + o(h^{2q}) \qquad (1.3.17) \\
&\quad +4n^{-1}VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + 4n^{-1}C_0(\boldsymbol{\gamma})h^q + o(n^{-1}h^q) \\
&\quad -4n^{-2}VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + 2n^{-2}E\left[\delta(W_{i;\boldsymbol{\gamma}})\right]h^{-dz} \\
&\quad +o\left(n^{-2}h^{-dz}\right) - 2n^{-2}\Delta(\boldsymbol{\gamma})^2 + o(n^{-2}),
\end{aligned}$$

where B_5 is defined in (1.6.6), and $\delta(W_i)$ is defined by

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

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

$$L\_MSE\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \equiv \left\{E\left[B_5(W_i;\boldsymbol{\gamma})\right]\right\}^2 h^{2q} + 2n^{-2}E\left[\delta(W_i;\boldsymbol{\gamma})\right]h^{-dz}. \qquad (1.3.18)$$

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

$$h^{+}=\left[\frac{d_Z\cdot E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right]}{q\cdot\left\{E\left[B_5\left(W_i;\boldsymbol{\gamma}\right)\right]\right\}^2}\right]^{1/(2q+d_Z)}\cdot\left[\frac{1}{n}\right]^{2/(2q+d_Z)}+o\left(\left[\frac{1}{n}\right]^{2/(2q+d_Z)}\right). \quad (1.3.19)$$

Under the condition $2q>d_Z$, which is implied by Assumption 6, the right-hand remainder term is $o\left(n^{-1}\right)$, but necessarily greater than $O\left(n^{-2}\right)$. h^{+} could be approximated by its leading term

$$h^{++}=\left[\frac{d_Z\cdot E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right]}{q\cdot\left\{E\left[B_5\left(W_i;\boldsymbol{\gamma}\right)\right]\right\}^2}\right]^{1/(2q+d_Z)}\cdot\left[\frac{1}{n}\right]^{2/(2q+d_Z)}=O\left(n^{-2/(2q+d_Z)}\right). \quad (1.3.20)$$

Now (**A 6.1)** is satisfied:

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

And so is (**A 6.2):**

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

But $E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right]$ and $E\left[B_5\left(W_i;\boldsymbol{\gamma}\right)\right]$ are unknown since the densities are unknown. So I use a simple plug-in estimator of h^{+} proposed by Powell and Stoker (1996). Let h_0 be an initial bandwidth. Suppose $E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma})^4\right]=O\left(h_0^{-\eta-2d_Z}\right)$ for some $\eta>0$, and let $\rho=\max\left\{\eta+2d_Z,2q+d_Z\right\}$. If $h_0\to 0$ and $nh_0^{\rho}\to\infty$, then by Proposition 4.2 of Powell and Stoker (1996),

$$\hat{\delta}\equiv\hat{\delta}\left(h_0\right)=\frac{1}{\begin{pmatrix}n\\2\end{pmatrix}}\sum_{(n,2)}h_0^{d_Z}\cdot\kappa_{h_0}(W_i,W_j;\boldsymbol{\gamma})^2\xrightarrow{P}E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right] \quad (1.3.21)$$

and

$$\begin{aligned}\hat{B}_5 &\equiv \frac{\bar{\Delta}_{n,\tau h_0}(\boldsymbol{\gamma})-\bar{\Delta}_{n,h_0}(\boldsymbol{\gamma})}{(\tau h_0)^q-h_0^q}\ for\ some\ \text{positive}\ \tau\neq 1 \quad (1.3.22)\\ &\xrightarrow{P} E\left[B_5\left(W_i;\boldsymbol{\gamma}\right)\right].\end{aligned}$$

The estimator $\hat{B}_5$ suggested in (1.3.22) is a “slope” of two point $(h_0^q,\bar{\Delta}_{n,h_0}(\boldsymbol{\gamma}))$ and $(\tau h_0^q,\bar{\Delta}_{n,\tau h_0}(\boldsymbol{\gamma}))$. To get a more stable estimator $\hat{B}_5$, we could use a regression of $\bar{\Delta}_{n,h_0}(\boldsymbol{\gamma})$ on h_0^q for various values of h_0.

Thus the proposed bandwidth selection is

$$\hat{h} = \left[\frac{d_Z \cdot \hat{\delta}}{q \cdot \hat{B}_5^2}\right]^{1/(2q+d_Z)} \cdot \left[\frac{1}{n}\right]^{2/(2q+d_Z)}. \tag{1.3.23}$$

Remark 8 *In practice we can choose q large enough such that $\rho = \max\{\eta + 2d_Z, 2q + d_Z\} = 2q + d_Z$, then we can choose the initial bandwidth $h_0 = o\left(n^{-1/(2q+d_Z)}\right)$.*

Remark 9 *Powell and Stoker (1996) mentioned one technical proviso: $\bar{\Delta}_n(\gamma; \hat{h})$ is not guaranteed to be asymptotically equivalent to $\bar{\Delta}_n(\gamma; h^{++})$ since the MSE calculations used to derive the form of h^+ held h fixed in calculating the moments of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$. The suggested straightforward solution is to discretize the set of possible scaling constants, replacing $\hat{h}$ with the closest value, $\hat{h}'$, in some finite set.*

1.4 Monte Carlo Experiments

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

For all the simulations, I generate $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ IID. The bandwidth I use is a value close to $\hat{h}$, and I use $h_0 = n^{-1/[3(2q+d_Z)]}$ and $\tau = 0.5$ when calculating $\hat{h}$ by (1.3.23). I choose $\varphi(\cdot)$ to be the standard normal PDF, and $k(u)$ the sixth order Gaussian kernel. In the following experiments, I only choose one $\boldsymbol{\gamma}$, i.e. I let $\Gamma_s = \{\boldsymbol{\gamma}\}$.

1.4.A The Distribution of Test Statistic

We consider the following DGP to study the distributions of the test statistic under the null and under the alternatives.

DGP 1:

$$\begin{aligned} Y &= \beta X + Z + \epsilon_Y \\ X &= Z + Z^2 + \epsilon_X \end{aligned}$$

$$\begin{pmatrix} \epsilon_X \\ \epsilon_Y \end{pmatrix} \tilde{} N \left(0, \begin{pmatrix} \sigma_X^2 \, 0 \\ 0 \, \sigma_Y^2 \end{pmatrix} \right) = N \left(0, \begin{pmatrix} 4 \, 0 \\ 0 \, 1 \end{pmatrix} \right)$$

$$Z \tilde{} N(0, \sigma_Z^2) = N(0, 3)$$

We are testing

$$H_0 : Y \perp X | Z.$$

Note that when $\beta = 0$, the null is true; otherwise the alternative is true.

Under the Null

Let $\beta = 0$ in DGP 1. Hence the null hypothesis holds.

Although selecting $\boldsymbol{\gamma}$ at random from a smooth density will deliver a consistent test with probability 1 (Bierens (1990)), in practice we should avoid to choose $\boldsymbol{\gamma}$ which makes $|\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3|$ too large or too small. The reason is that the value of $\varphi(u)$ will be very close to zero if $|u|$ too large and $\varphi(u)$ will be close to linear if $|u|$ too small, in which cases the test will not have good power. In our simulation, we choose $\boldsymbol{\gamma}$ which make $|\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3|$

around one. To be specific, we choose

$$\begin{aligned}
\gamma_0 &\approx 1 - \left[\frac{\bar{X}}{std(\{X_i\}_{i=1}^n)} + \frac{\bar{Y}}{std(\{Y_i\}_{i=1}^n)} + \frac{\bar{Z}}{std(\{Z_i\}_{i=1}^n)}\right] \\
\gamma_1 &\approx \left(\frac{1}{std(\{X_i\}_{i=1}^n)}\right) \\
\gamma_2 &\approx \left(\frac{1}{std(\{Y_i\}_{i=1}^n)}\right) \\
\gamma_3 &\approx \left(\frac{1}{std(\{Z_i\}_{i=1}^n)}\right).
\end{aligned}$$

The QQ plot (figure 1.1) shows that the distribution of $\tilde{T}_n(\boldsymbol{\gamma}) \equiv \frac{\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_{\Delta}(\boldsymbol{\gamma})}$ is approximately standard normal. Other choices of φ give similar results. We also examined other DGP's under the null, but results are not reported here because the resulting figures are similar.

Under the Alternative

Now let $\beta = 0.2887$ such that

$$\rho_{X,Y|Z} = \frac{cov\,(X, Y|Z)}{\sigma_{X|Z}\sigma_{Y|Z}} = \frac{4\beta}{2\sqrt{4\beta^2+1}} = 0.5.$$

In this case the alternative is true.

The QQ plots (figure 1.2)shows that the distribution of $\tilde{T}_n(\boldsymbol{\gamma})$ is still close to normal but the mean of our proposed statistic is not zero. In fact, the mean increases with the sample size n.

1.4.B Size and Power Study

DGP 1

In this section, we use DGP 1 to study the finite sample size/power of the test against conditional mean dependence. We use

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

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

We want to plot the power of test against ρ from -0.9 to 0.9. To do that, we choose

$$\beta = \frac{\rho_{X,Y|Z}}{2\sqrt{\left(1-\rho_{X,Y|Z}^2\right)}} \text{ such that } \rho_{X,Y|Z} = -0.9, -0.8, ..., 0.9$$

Figure 1.3 shows the power of the test when choosing $\varphi(\cdot)$ as standard normal PDF and $k(u)$ the sixth order Gaussian kernel. The size and power looks not bad when the sample size is as big as 500.

DGP 2

DGP 2 is a modification of the first one by choosing ϵ_X and ϵ_Y to be student t distribution of degree 3:

$$\epsilon_X \sim 2t_3,\ \epsilon_Y \sim t_3,\ \epsilon_X \perp \epsilon_Y.$$

Then the results showed in figure 1.4 are a little worse than the previous normal distribution case.

DGP 3

DGP 3 is again a modification of the first one by choosing ϵ_X and ϵ_Y to be centered Chi square distribution:

$$\epsilon_X \sim 2\left(\chi_1^2 - 1\right),\ \epsilon_Y \sim \left(\chi_1^2 - 1\right),\ \epsilon_X \perp \epsilon_Y.$$

Then the results showed in figure 1.5 are a little better than previous two cases.

1.5 Concluding Remarks

In this chapter, I proposed a nonparametric GCR test for conditional independence. The basic idea is to test if the topological distance between a restricted and an unrestricted probability measures corresponding to conditional independence or its absence is zero. The test statistic has a simple closed form and hence easy to compute. The limiting null distribution of the test statistic is a Chi-square distribution.

I use a collection of $\boldsymbol{\gamma}$'s for constructing the test statistic. $\boldsymbol{\gamma}$ is a nuisance parameter "present only under the alternative"; it is also described as "identified only under the alternative". Although Bierens (1990) points out that selecting $\boldsymbol{\gamma}$ at random from a smooth density will deliver a consistent test with probability 1, this method will introduce a degree of arbitrariness into both the size and the power of the test. In chapter 2, I will study how to "integrate out" $\boldsymbol{\gamma}$.

In the assumptions, I assume the conditioning variables to be continuous. But in applied microeconomics, many variables are discrete. I will discuss how to deal with mixed data in chapter 3. Another limitation is that I assume IID data. But the IID assumption is not suitable for time series. For example, IID.

assumption fails when we want to test nonlinear Granger causality. We may extend this approach to a time series framework in the future.

1.6 Appendix: Proofs

Proof of Proposition 1: It suffices to verify that φ^* satisfies our extension of the GCR definition. By assumption, H_φ is GCR, so that for every T with non-empty interior, sp $H_\varphi(T)$ is uniformly dense in $C(B)$ for any compact B. Let B be a given compact set belonging to *supp*(Z). Then for every such T, all $g \in C(B)$, and all $\epsilon > 0$, there exists $h_{g,\epsilon,B,T} \in$ sp $H_\varphi(T)$ such that

$$\sup_{w\in B} \left|g(w) - h_{g,\epsilon,B,T}(w)\right| < \epsilon.$$

By assumption, there are constants $C_{L,B}$ and C_U such that $0 < C_{L,B} < f_Z(z) < C_U < \infty$ for all $z \in \mathit{supp}(Z)$. It follows that with $f(w) \equiv f_Z(z)$, we have $0 < C_{L,B} < f(w) < C_U < \infty$ for all $w \in B$. Because $0 < C_{L,B} < f(w)$, it follows that if $g \in C(B)$, then also $g/f \in C(B)$. Let $\varphi^* \equiv f\varphi$. Then there exists $h^*_{g/f,\epsilon,B,T} = fh_{g/f,\epsilon,B,T} \in$ sp $H_{\varphi^*}(T)$ such that

$$\begin{aligned} \sup_{w\in B} |g(w) - h^*_{g/f,\epsilon,B,T}(w)| &= \sup_{w\in B} |g(w) - f(w)h_{g/f,\epsilon,B,T}(w)| \\ &= \sup_{w\in B} |f(w)[g(w)/f(w) - h_{g/f,\epsilon,B,T}(w)]| \\ &\leq C_U \sup_{w\in B} |[g(w)/f(w) - h_{g/f,\epsilon,B,T}(w)]| \\ &\leq C_U\ \epsilon. \end{aligned}$$

As B, T, g, and ϵ are arbitrary, the definition is verified, and the result follows. ■

Proof of Lemma 2: Lemma 3.1 in Powell et al. (1989) shows that if $E\left\|\kappa_{\mathbf{h}}(W_i, W_j; \Gamma_s)\right\|^2 = o(n)$,then $\sqrt{n}\left(\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \hat{\Delta}_n(\boldsymbol{\gamma})\right) = o_p(1)$. Now we need to find the condition on $\mathbf{h}$ such that $E\left\|\kappa_{\mathbf{h}}(W_i, W_j; \Gamma_s)\right\|^2 = o(n)$ satisfied.

We first show that $E|\kappa_{h_l}(W_i, W_j; \boldsymbol{\gamma}_l)|^2 = o(n)$ for $l = 1, 2, ..., s$, if $nh_l^{dz} \to \infty$ and $h_l \to 0$ as $n \to \infty$. We observe

$$
\begin{aligned}
& E\left|\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right|^2 \\
= & \int \varphi^2(\gamma_0 + x\gamma_1 + y\gamma_2 + z_1\gamma_3)\frac{1}{h^{2dz}}\left|K(\frac{z_1 - z_2}{h})\right|^2 \\
& \times f_{XYZ}(x, y, z_1)f_Z(z_2)dxdydz_1dz_2 \\
= & \int \frac{1}{h^{dz}}\varphi^2(\gamma_0 + x\gamma_1 + y\gamma_2 + z_1\gamma_3)\left|K(u)\right|^2 \\
& \times f_{XYZ}(x, y, z_1)f_Z(z_1 + uh)dxdz_1dydu \\
= & O(\frac{1}{h^{dz}})
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left|\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right|^2 \\
= & E\left\{\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right\}^2 \\
= & \int \varphi^2(\gamma_0 + x\gamma_1 + y\gamma_2 + z_1\gamma_3)\frac{1}{h^{2dz}}\left|K(\frac{z_1 - z_2}{h})\right|^2 \\
& \times f_{XZ}(x, z_1)f_{YZ}(y, z_2)dxdz_1dydz_2 \\
= & \int \frac{1}{h^{dz}}\varphi^2(\gamma_0 + x\gamma_1 + y\gamma_2 + z_1\gamma_3)\left|K(u)\right|^2 \\
& \times f_{XZ}(x, z_1)f_{YZ}(y, z_1 + uh)dxdz_1dydu \\
= & O(\frac{1}{h^{dz}}),
\end{aligned}
$$

where in the second last step we use $h \to 0$.

Since

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

we have

$$
\begin{aligned}
& E|\kappa_h(W_i,W_j;\boldsymbol{\gamma})|^2 \\
\leq\ & 2E\left|\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)K_h(Z_i-Z_j)\right|^2 \\
& +2E\left|\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3)K_h(Z_i-Z_j)\right|^2 \\
& +2E\left|\varphi(\gamma_0+X_j'\gamma_1+Y_j'\gamma_2+Z_j'\gamma_3)K_h(Z_j-Z_i)\right|^2 \\
& +E\left|\varphi(\gamma_0+X_j'\gamma_1+Y_i'\gamma_2+Z_j'\gamma_3)K_h(Z_j-Z_i)\right|^2 \\
=\ & O\left(\frac{1}{h^{d_Z}}\right)=O(n\frac{1}{nh^{d_Z}}) \\
\big(if\ nh^{d_Z}\to\infty\big) =\ & o(n)
\end{aligned}
$$

So $E\left\|\kappa_h(W_i,W_j;\Gamma_s)\right\|^2 \leq \sum_{l=1}^{s} E\left|\kappa_h(W_i,W_j;\gamma_l)\right|^2 = O(n\frac{1}{nh^{d_Z}})$. And if $nh_l^{d_Z}\to\infty$ for $l=1,2,...,s=o(n)$, $E\left\|\kappa_h(W_i,W_j;\Gamma_s)\right\|^2=o(n)$. ■

Proof of Lemma 3: (a) Show that $\Delta_{\mathbf{h}}(\Gamma_s)\equiv E\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]=\Delta(\Gamma_s)+O(\mathbf{h}^q)$.

Using Taylor expansions, we get

$$
\begin{aligned}
& E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)K_h(Z_i-Z_j)\right] \\
=\ & E\left\{E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)K_h(Z_i-Z_j)|W_i\right]\right\} \\
=\ & E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\int K(u)f_Z(Z_i+uh)du\right] \\
=\ & E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_Z(Z_i)\right]+h^qC_1(\gamma)+o(h^2)
\end{aligned}
$$

with

$$
C_1(\boldsymbol{\gamma})\equiv E\left\{\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\sum_{s=1}^{d_Z}\left[\frac{\partial^q f_Z\left(Z_i\right)/\partial\left(Z_{is}\right)^q}{q!}\right]\right\}\int u^qk(u)du,
$$

and similarly

$$\begin{aligned} & E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j)\right] \\ = \ & E\left[g_{XZ}(X_i, Z_i; \gamma)\right] + C_2(\boldsymbol{\gamma})h^q + o(h^q) \end{aligned}$$

with

$$\begin{aligned} C_2(\boldsymbol{\gamma}) &= E\left\{\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)\sum_{s=1}^{d_Z}\left[\frac{\partial^q f_Z\left(Z_i\right)/\partial\left(Z_{is}\right)^q}{q!}\right]dy\right\} \\ & \times \int u^q k(u)du. \end{aligned}$$

So

$$\begin{aligned} \Delta_{\mathbf{h}}(\boldsymbol{\gamma}) &\equiv E\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] \\ &= E\{[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)K_h(Z_i - Z_j) \\ & \quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)]K_h(Z_i - Z_j)\} \\ &= E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)\right] + C_1(\boldsymbol{\gamma})h^q + o(h^q) \\ & \quad - \{E\left[g_{XZ}(X, Z; \boldsymbol{\gamma})\right] + C_2(\boldsymbol{\gamma})h^q + o(h^q)\} \\ &= \Delta\left(\boldsymbol{\gamma}\right) + C_3(\boldsymbol{\gamma})h^q + o(h^q) \end{aligned}$$

with

$$C_3(\boldsymbol{\gamma}) \equiv C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma}).$$

It follows that

$$E\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right] = \Delta\left(\Gamma_s\right) + C_3(\Gamma_s)\mathbf{h}^q + o(\mathbf{h}^q)$$

hence the result.

(b) Show that $H_{n,\mathbf{h},1}(\Gamma_s) = \frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\}$ $+O_p(\mathbf{h}^q)$.

$$
\begin{aligned}
\kappa_{h,1}(W_i;\boldsymbol{\gamma}) &\equiv E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma})|W_i\right] \\
&= \frac{1}{2}\{\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_Z(Z_i) \\
&\quad +h^qB_1(X_i,Y_i,Z_i;\boldsymbol{\gamma})+o(h^q)\} \\
&\quad -\frac{1}{2}\{\int\varphi(\gamma_0+X_i'\gamma_1+y'\gamma_2+Z_i'\gamma_3)f_{YZ}(y,Z_i)dy \\
&\quad +h^qB_2(X_i,Z_i;\boldsymbol{\gamma})+o(h^q)\} \\
&\quad +\frac{1}{2}\{\int\varphi(\gamma_0+x\gamma_1+y\gamma_2+Z_i'\gamma_3)f_{XYZ}(x,y,Z_i)dxdy \\
&\quad +h^qB_3(Z_i;\boldsymbol{\gamma})+o(h^q)\} \\
&\quad -\frac{1}{2}\{\int\varphi(\gamma_0+x'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_{XZ}(x,Z_i)dx \\
&\quad +h^qB_4(Y_i,Z_i;\boldsymbol{\gamma})+o(h^q)\} \\
&\equiv \kappa_1(W_i;\boldsymbol{\gamma})+h^qB_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})+o(h^q)
\end{aligned}
$$

where

$$
\begin{aligned}
\kappa_1(W_i;\boldsymbol{\gamma}) &\equiv \frac{1}{2}\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_Z(Z_i) \\
&\quad -\frac{1}{2}\int\varphi(\gamma_0+X_i'\gamma_1+y'\gamma_2+Z_i'\gamma_3)f_{YZ}(y,Z_i)dy \\
&\quad +\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+Z_i'\gamma_3)f_{XYZ}(x,y,Z_i)dxdy \\
&\quad -\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_{XZ}(x,Z_i)dx,
\end{aligned} \tag{1.6.1}
$$

$$
\begin{aligned}
B_1(X_i,Y_i,Z_i;\boldsymbol{\gamma}) &\equiv \varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\sum_{s=1}^{d_Z}\left[\frac{\partial^q f_Z\left(Z_i\right)/\partial\left(Z_{is}\right)^q}{q!}\right] \\
&\quad \times\int u^qk(u)du,
\end{aligned} \tag{1.6.2}
$$

$$
\begin{aligned}
B_2(X_i,Z_i;\boldsymbol{\gamma}) &= \int\varphi(\gamma_0+X_i'\gamma_1+y'\gamma_2+Z_i'\gamma_3)\sum_{s=1}^{d_Z}\frac{1}{q!}\frac{\partial^q f_{YZ}(y,Z_i)}{\partial\left(Z_{is}\right)^q}dy \\
&\quad \times\int u^qk\left(u\right)du,
\end{aligned} \tag{1.6.3}
$$

$$\begin{aligned} B_3(Z_i;\boldsymbol{\gamma}) &= \int \sum_{s=1}^{d_Z} \frac{1}{q!} \frac{\partial^q \left[\varphi(\gamma_0 + x\gamma_1 + y\gamma_2 + Z_i\gamma_3) f_{XYZ}(x,y,Z_i)\right]}{\partial \left(Z_{is}\right)^q} dxdy \\ &\quad \times \int u^q k\left(u\right) du, \end{aligned} \tag{1.6.4}$$

$$\begin{aligned} B_4(Y_i,Z_i;\boldsymbol{\gamma}) &\equiv \int \sum_{s=1}^{d_Z} \frac{1}{q!} \frac{\partial^q \left[\varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_{XZ}(x,Z_i)\right]}{\partial \left(Z_{is}\right)^q} \\ &\quad \times \int u^q k\left(u\right) dudx, \end{aligned} \tag{1.6.5}$$

and

$$\begin{aligned} B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma}) &\equiv \frac{1}{2}\left[B_1(X_i,Y_i,Z_i;\boldsymbol{\gamma}) - B_2(X_i,Z_i;\boldsymbol{\gamma})\right. \\ &\quad \left. + B_3(Z_i;\boldsymbol{\gamma}) - B_4(Y_i,Z_i;\boldsymbol{\gamma})\right]. \end{aligned} \tag{1.6.6}$$

Define

$$\begin{aligned} t_h(W_i;\boldsymbol{\gamma}) &\equiv \kappa_{h,1}(W_i;\boldsymbol{\gamma}) - \kappa_1(W_i;\boldsymbol{\gamma}) \qquad (1.6.7) \\ &= h^q B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma}) + o(h^q), \end{aligned}$$

then

$$\begin{aligned} E\left[t_h(W_i;\boldsymbol{\gamma})\right] &= E\left[\kappa_{h,1}(W_i;\boldsymbol{\gamma})\right] - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] \\ &= \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}\right) - \Delta\left(\boldsymbol{\gamma}\right). \end{aligned}$$

So

$$\begin{aligned} H_{n,\mathbf{h},1}(\boldsymbol{\gamma}) &\equiv \frac{1}{n}\sum_{i=1}^{n} \tilde{\kappa}_{\mathbf{h},1}(W_i;\boldsymbol{\gamma}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{\kappa_{\mathbf{h},1}(W_i;\boldsymbol{\gamma}) - \Delta_{\mathbf{h}}\left(\boldsymbol{\gamma}\right)\right\} \qquad (1.6.8) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{\kappa_1(W_i;\boldsymbol{\gamma}) - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \left\{t_h(W_i;\boldsymbol{\gamma}) - E\left[t_h(W_i;\boldsymbol{\gamma})\right]\right\}, \end{aligned}$$

where

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left\{t_h(W_i;\boldsymbol{\gamma}) - E\left[t_h(W_i;\boldsymbol{\gamma})\right]\right\} \\
= \quad & \frac{1}{n}\sum_{i=1}^{n}\left\{h^q B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma}) - E\left[B_5(X_i,Y_i,Z_i;\boldsymbol{\gamma})\right]h^q\right\} + o_p(h^q) \\
= \quad & O_p(h^q) + o_p(h^q).
\end{aligned}$$

So

$$H_{n,\mathbf{h},1}(\Gamma_s) = \frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + O_p(\mathbf{h}^q).$$

(c) Show that under assumption (6.2) $\sqrt{n}H_{n,\mathbf{h},1}(\Gamma_s) \xrightarrow{d} N\left(0,\Omega_H\right)$, hence $\sqrt{n}\left(\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta_{\mathbf{h}}(\Gamma_s)\right) = \sqrt{n}\left(\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta(\Gamma_s)\right) + o\left(1\right) \xrightarrow{d} N\left(0,\Omega\right)$.

Under assumption (6.2), $\sqrt{n}(\mathbf{h}^q) = o\left(1\right)$, so $\Delta_{\mathbf{h}}\left(\Gamma_s\right) = \Delta\left(\Gamma_s\right) + o(\frac{1}{\sqrt{n}})$ and $H_{n,\mathbf{h},1}(\Gamma_s) \ = \frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + o_p(\frac{1}{\sqrt{n}})$. Then

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

The leading term of $H_{n,\mathbf{h},1}(\Gamma_s)$ converges to a multivariate normal distribution by applying the Lindeberg-Levy Central Limit Theorem. Hence

$$\begin{aligned}
& \sqrt{n}\left(\hat{\Delta}_{n,\mathbf{h}}(\Gamma_s) - \Delta(\Gamma_s)\right) \\
= \quad & \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\Gamma_s) - E\left[\kappa_1(W_i;\Gamma_s)\right]\right\} + o_p(1) \\
& \xrightarrow{d} N\left(0,\Omega\right)
\end{aligned}$$

where

$$\Omega(l,k) \equiv \sigma_\Delta\left(\boldsymbol{\gamma}_l,\boldsymbol{\gamma}_k\right) = 4cov\left[\kappa_1(W_i;\boldsymbol{\gamma}_l),\kappa_1(W_i;\boldsymbol{\gamma}_k)\right]. \qquad (1.6.9)$$

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

$$\begin{aligned}
\kappa_1(W_i;\gamma) &= \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i) \\
&\quad -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{YZ}(y, Z_i)dy \qquad (1.6.10) \\
&\quad +\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{XYZ}(x, y, Z_i)dxdy \\
&\quad -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x, Z_i)dx \\
(under\ H_0) &= \frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i, Y_i, Z_i\right] \\
&\quad -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i, Z_i\right] \\
&\quad +\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Z_i\right] \\
&\quad -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Y_i, Z_i\right] \\
&\equiv \Lambda(W_i;\boldsymbol{\gamma}).
\end{aligned}$$

So

$$\begin{aligned}
& cov\left[\kappa_1(W_i;\boldsymbol{\gamma}_l), \kappa_1(W_i;\boldsymbol{\gamma}_k)\right] \\
& (under\ H_0) = E\left[\Lambda(W_i;\boldsymbol{\gamma}_l)\Lambda(W_i;\boldsymbol{\gamma}_k)\right].
\end{aligned}$$

■

Proof of Theorem 4: It follows directly from the H-decomposition and lemma 2 and 3. ■

Proof of Corollary 5: It follows from theorem 4. ■

Proof of Theorem 6: The proof is similar to the proof of theorem 3.4 in Powell et al. (1989). I have shown that $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s) \overset{P}{\to} \Delta(\Gamma_s)$ in previous corollary, so I only need to show the consistency of $\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\kappa}_{h,1}(W_i;\Gamma_s)\hat{\kappa}_{h,1}(W_i;\Gamma_s)'\right]$ for

$E\left[\kappa_1(W_i;\Gamma_s)\kappa_1(W_i;\Gamma_s)'\right]$. First, I show that $E\left[\|\hat{\kappa}_{h,1}(W_i;\Gamma_s)-\kappa_{h,1}(W_i;\Gamma_s)\|^2\right]=o\left(1\right)$.

$$\begin{aligned}
& E\left[\|\hat{\kappa}_{h,1}(W_i;\Gamma_s)-\kappa_{h,1}(W_i;\Gamma_s)\|^2\right] \\
= \;& E\left\{\sum_{l=1}^{s}\left[\hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma}_l)-\kappa_{h,1}(W_i;\boldsymbol{\gamma}_l)\right]^2\right\} \\
= \;& \sum_{l=1}^{s}E\left(\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)-E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right]^2\right) \\
= \;& \sum_{l=1}^{s}E\left(\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\left\{\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)-E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\}\right)^2 \\
= \;& \sum_{l=1}^{s}E\left[\frac{1}{\left(n-1\right)^2}\sum_{j=1,j\neq i}^{n}\left\{\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)-E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\}^2\right] \\
& \left(\begin{array}{c} \textit{cross product} \\ \textit{terms are zeros}\end{array}\right) \\
= \;& \frac{1}{n-1}\sum_{l=1}^{s}E\left[\left\{\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)-E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\}^2\right] \\
= \;& \frac{1}{n-1}\sum_{l=1}^{s}E\left\{E\left[\left\{\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)-E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\}^2|W_i\right]\right\} \\
= \;& \frac{1}{n-1}\sum_{l=1}^{s}E\left\{VAR\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\},\ j\neq i \\
= \;& \frac{1}{n-1}\sum_{l=1}^{s}\left(E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)\right]^2-E\left\{E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]\right\}^2\right) \\
\leq \;& \frac{1}{n-1}\sum_{l=1}^{s}E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)\right]^2,\ \textit{equality holds iff } E\left[\kappa_h(W_i,W_j;\boldsymbol{\gamma}_l)|W_i\right]=0 \\
= \;& \frac{1}{n-1}E\left\|\kappa_h(W_i,W_j;\Gamma_s)\right\|^2 \\
\leq \;& \frac{1}{n-1}O\left(\frac{1}{h^{dz}}\right)=O\left(\frac{1}{nh^{dz}}\right)=o\left(1\right),
\end{aligned}$$

where in the last step I use the assumption (A 6.1).

Secondly,

$$\begin{aligned}
& E\left\|\kappa_{h,1}(W_i;\Gamma_s)-\kappa_1\left(W_i;\Gamma_s\right)\right\|^2 \\
= & \left\|t_h(W_i;\Gamma_s)\right\|^2 \\
= & \sum_{l=1}^{s}\left[t_h(W_i;\boldsymbol{\gamma}_l)\right]^2 \\
= & \sum_{l=1}^{s}\left[O_p\left(h_l^q\right)\right]^2=o_p\left(\frac{1}{n}\right)
\end{aligned}$$

where in the last step I use the assumption (A 6.2) i.e. $\sqrt{n}h_l^q\to 0$.

So.

$$\begin{aligned}
& E\left\|\hat{\kappa}_{h,1}(W_i;\Gamma_s)-\kappa_1\left(W_i;\Gamma_s\right)\right\|^2 \\
= & E\left\|\left[\hat{\kappa}_{h,1}(W_i;\Gamma_s)-\kappa_{h,1}(W_i;\Gamma_s)\right]+\left[\kappa_{h,1}(W_i;\Gamma_s)-\kappa_1\left(W_i;\Gamma_s\right)\right]\right\|^2 \\
\leq & E\left\|\hat{\kappa}_{h,1}(W_i;\Gamma_s)-\kappa_{h,1}(W_i;\Gamma_s)\right\|^2+E\left\|\kappa_{h,1}(W_i;\Gamma_s)-\kappa_1\left(W_i;\Gamma_s\right)\right\|^2 \\
= & o_p\left(1\right)
\end{aligned}$$

which implies

$$E\left[\left\|\hat{\kappa}_{h,1}(W_i;\Gamma_s)\hat{\kappa}_{h,1}(W_i;\Gamma_s)'-\kappa_1(W_i;\Gamma_s)\kappa_1(W_i;\Gamma_s)'\right\|\right]=o\left(1\right).$$

where for a matrix A, $\|A\|\equiv\left[trace\left(A'A\right)\right]^{1/2}$. Using Markov's inequality and the SLLN, I obtain

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

hence

$$\hat{\Omega}\xrightarrow{P}\Omega$$

■

Proof of Corollary 7: $\tilde{\Omega} - \hat{\Omega} = 4\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)' \xrightarrow{P} 4\Delta(\Gamma_s)\Delta(\Gamma_s)'$, by Corollary 5. And under the null, $\Delta(\Gamma_s) = 0$. ■

Proof of Corollary 8: The results follow from theorem 4 and 6. ■

Proof of Proposition 9: The result follows from corollary 8 and that under the alternatives $\Delta(\Gamma_s) \neq 0$ for essentially all $\Gamma_s \subset \Gamma$. ■

1.7 Figures

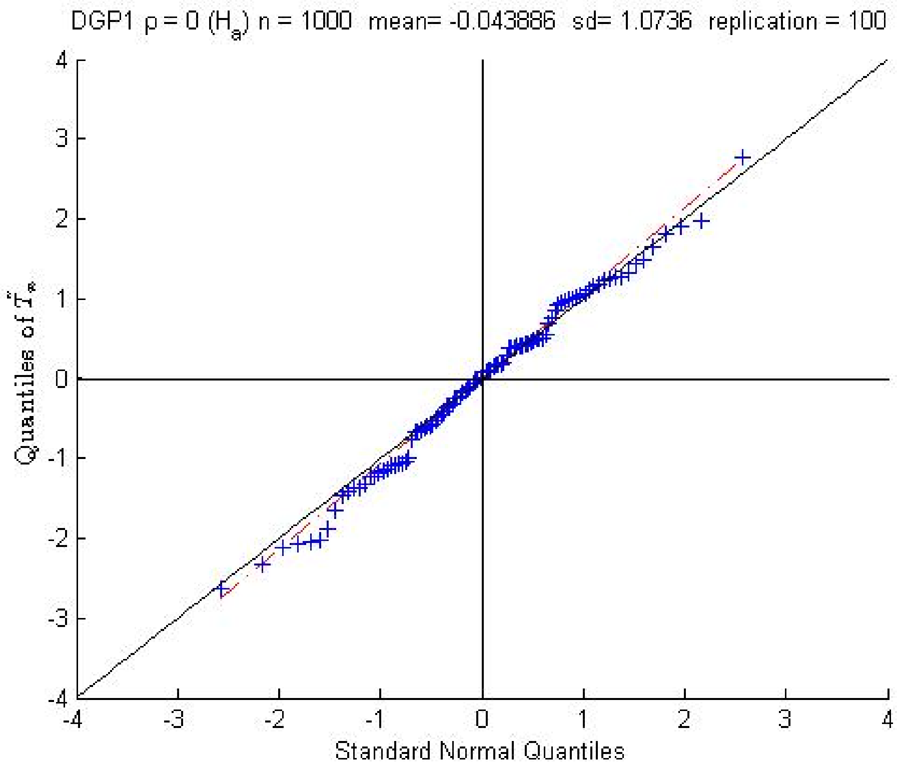

Figure 1.1: QQ plot of $\tilde{T}_n$ under the null vs. standard normal

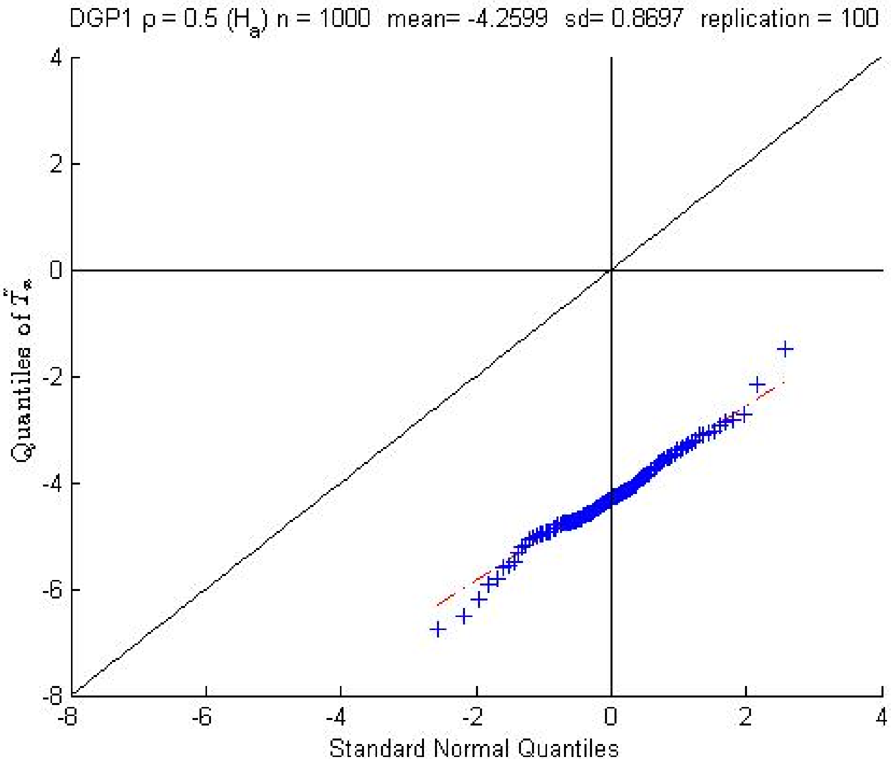

Figure 1.2: QQ plot of $\tilde{T}_n$ under the alternative vs. standard normal

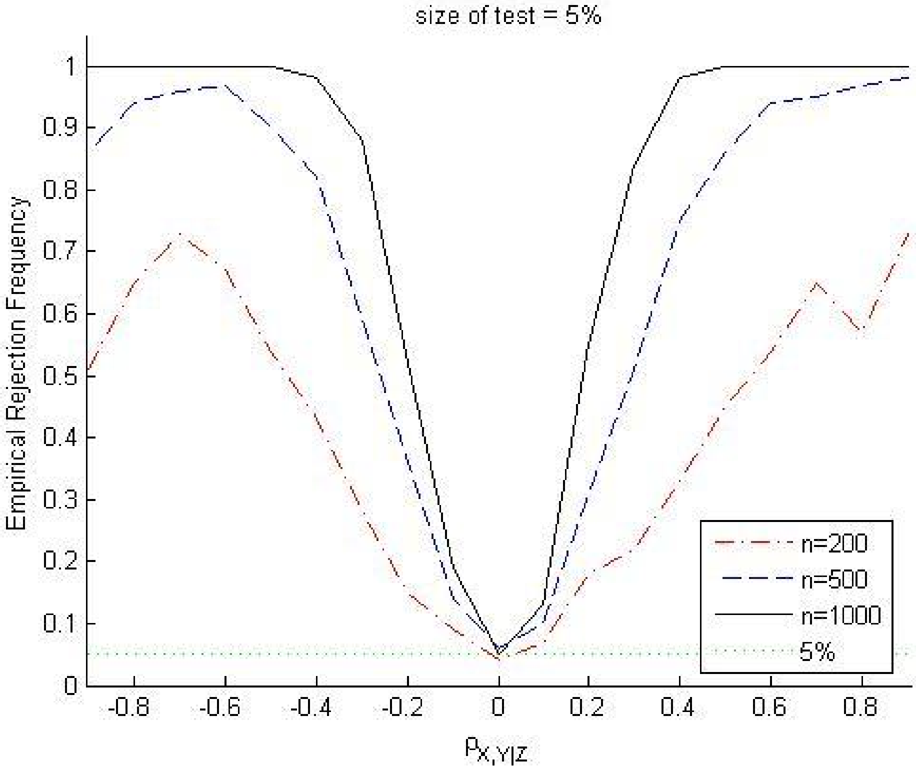

Figure 1.3: Power functions of Chi-square test for DGP 1

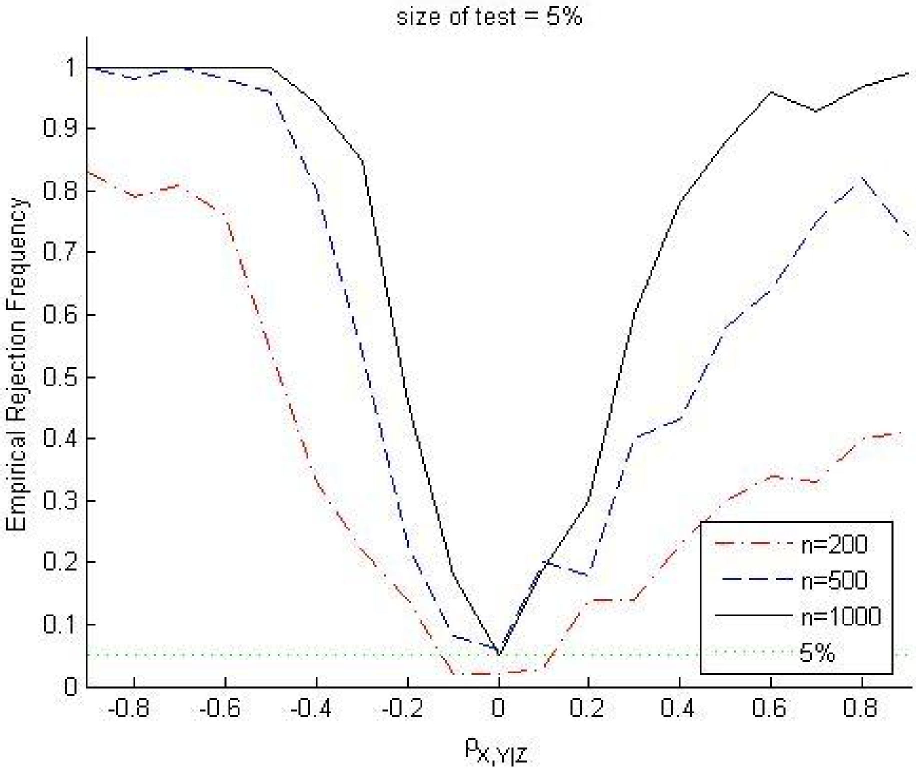

Figure 1.4: Power functions of Chi-square test for DGP 2

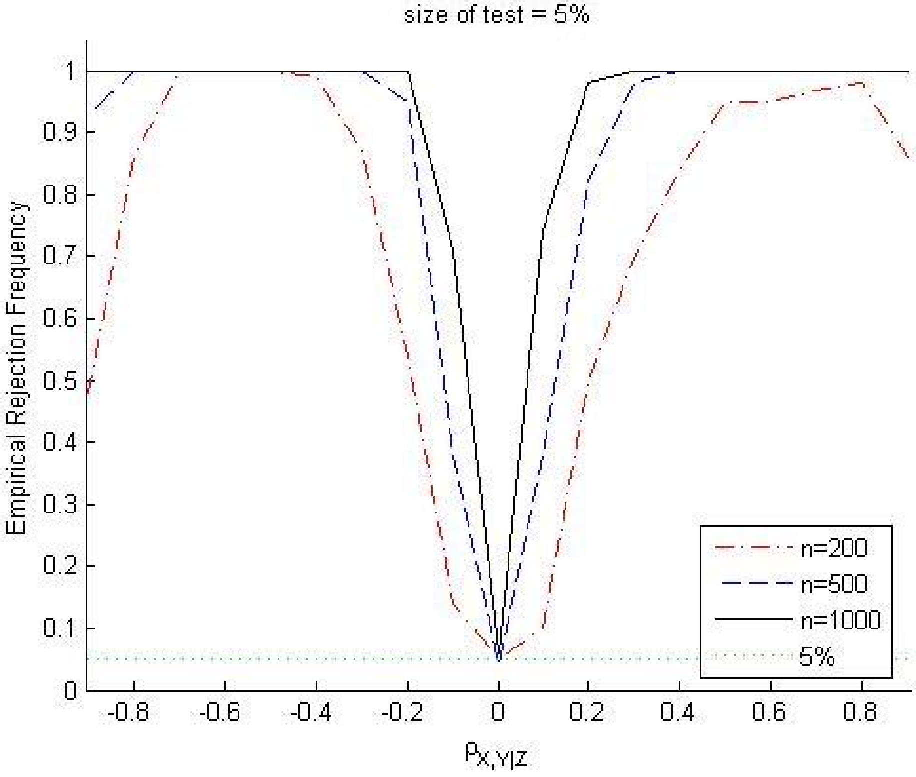

Figure 1.5: Power functions of Chi-square test for DGP 3

2

An Integrated Conditional Moment Test for Conditional Independence

2.1 Introduction

In chapter 1, I proposed a nonparametric test for conditional independence. Suppose we have the data for three random vectors X, Y and Z. The null hypothesis we are testing is that Y is independent of X given Z. The idea of the test in chapter 1 follows a series of papers of consistent specification tests in Bierens (1982), Bierens (1990), Bierens and Ploberger (1997) and Stinchcombe and White (1998), among others. 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 indexed by a nuisance parameter vector. The use of GCR functions makes the test consistent. Under the null, the limiting distribution of the test statistic is a Chi-square distribution.

Although the test statistic in chapter 1 is easy to calculate and has a tractable limiting null distribution, its consistency relies on the randomization of the choice of the test parameters. In this chapter I obtain a Bierens type Integrated Conditional Moment (ICM) test by integrating out the nuisance parameters. The test still achieves $\sqrt{n}$ local power and its consistency does not rely on the randomization any more. Its limiting null distribution is a functional of a mean zero Gaussian process. I simulate critical values using a conditional simulation approach suggested by Hansen (1996). I also examine the use of convenient case-independent upper bounds suggested by Bierens and Ploberger (1997).

One potential 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 X and Y are related by the structural equation

$$Y = \beta_0 + \beta_1 X + U,$$

where U (e.g. ability) is an unobserved cause of Y (income) and β_0 and β_1 are unknown coefficients with β_1 representing the effect of X on Y. Since X is typically not randomly assigned and is correlated with U (e.g. unobserved ability will affect both the schooling and the income), OLS will generally fail to consistently estimate β_1. Nevertheless, if we can find a set of covariates Z (e.g. AFQT scores) such that given Z, U and X are independent, i.e.

$$U \perp X \mid Z, \tag{2.1.1}$$

we can estimate β_1 consistently by various methods: covariance adjustment, matching, methods using the propensity score such as weighting and blocking, or combinations of these approaches.

Assumption (2.1.1) is the key assumption for identifying β_1; this is called a conditional exogeneity assumption by White and Chalak (2006). This enforces the "ignorability" or "unconfoundedness" condition, also known as "selection on observables" (Barnow et al. (1981)).

Note that assumption (2.1.1) cannot be directly tested since U is unobservable. But if there are other observable covariates V satisfying certain conditions (see White and Chalak (2006)), we have

$$U \perp X \mid Z \Rightarrow V \perp X \mid Z,$$

so we can test (2.1.1) by testing its implication,

$$V \perp X \mid Z. \tag{2.1.2}$$

Section 5 of this paper applies this test in the context of the study of the returns to schooling.

The plan of this chapter is as follows. In section 2, I state the hypotheses to be tested and explain the basic idea of the test. In section 3, I develop an ICM test and discuss its local and global properties. I derive the asymptotic distribution of the test statistic and simulate the critical values. In section 4, I report some Monte Carlo results which show the test performs well for finite samples. In section 5, I apply the ICM test in the context of the study of the returns to schooling, testing the key assumption of unconfoundedness. Section 6 concludes and discusses directions for further research.

2.2 The Hypotheses and the Idea of the Test

2.2.A The Hypotheses

As in chapter 1, let X, Y, and Z be three random vectors, with dimensions d_X, d_Y, and d_Z, respectively. I still assume the sample observations $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ are independent and identically distributed (IID) and drop the i subscript when convenient. Formally, I keep Assumption 1.1 in chapter 1:

Assumption 1 $\{W_i \equiv (X_i', Y_i', Z_i')'\}$ *is an IID sequence of random variables on the complete probability space* $(\Omega_W, \mathcal{F}_W, P_W)$. X_i, Y_i, *and* Z_i *take values in* $\mathbb{R}^{d_X}$, $\mathbb{R}^{d_Y}$*,and* $\mathbb{R}^{d_Z}$, *respectively, and* $d_W \equiv d_X + d_Y + d_Z$.

The null hypothesis is that X and Y are independent given Z, whereas the alternative is its negation. Using the same notation introduced in chapter 1, we are testing

$$H_0 : Y \perp X \mid Z \text{ vs. } H_a : Y \not\perp X \mid Z. \tag{2.2.1}$$

Conditional independence can be defined via conditional densities or densities, as given by Dawid (1979). The following four equations are equivalent to each other and each defines that X and Y are independent conditioning on Z:

$$f_{Y|X,Z}(y \mid x, z) = f_{Y|Z}(y \mid z), \tag{2.2.2}$$

$$f_{X|YZ}(x \mid y, z) = f_{X|Y}(x \mid y), \tag{2.2.3}$$

$$f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z)\ f_{Y|Z|}(y \mid z), \tag{2.2.4}$$

and

$$f_{XYZ}(x,y,z)\ f_Z(z) = f_{XZ}(x,z)\ f_{YZ}(y,z) \tag{2.2.5}$$

where $f_{\cdot|\cdot}$ denotes the conditional densities and $f_{\cdot}$ denotes the densities.

2.2.B The Idea of the Test

One way to test conditional independence is to compare the densities in its definition to see if the equality condition holds. For example, Su and White's (2008) test essentially compares $f_{XYZ}(x,y,z)f_Z(z)$ to $f_{XZ}(x,z)f_{YZ}(y,z)$. But to do that, they estimate $f_{XYZ}(x,y,z)$, $f_Z(z)$, $f_{XZ}(x,z)$, and $f_{YZ}(y,z)$ nonparametrically, so their test has a power against local alternatives only at a rate of $n^{-1/2}h^{-d_W/4}$, the slowest rate of the four nonparametric density estimators i.e. the rate of $\hat{f}_{XYZ}(x,y,z)$. This is a rate slower than $1/\sqrt{n}$ and hence suffer from the "curse of dimensionality". The dimension here is $d_W = d_X + d_Y + d_Z$, which is at least three and potentially high.

To achieve a rate of $1/\sqrt{n}$, I do not compare the density functions directly. Instead, I use a family of "average" indexed by a nuisance parameter vector $\boldsymbol{\gamma}$ to indirectly measure the distance between $f_{XYZ}(x,y,z)f_Z(z)$ and $f_{XZ}(x,z)f_{YZ}(y,z)$, so that for each given $\boldsymbol{\gamma}$, the test statistic is based on an estimator of an average which could achieve $1/\sqrt{n}$ rate just like what a semiparametric estimator would do. This idea is analogous to Bierens (1982), Bierens (1990) and Stinchcombe and White (1998)'s specification tests, among others.

In chapter 1, I have established a pair of hypotheses equivalent to (2.2.1), i.e.

$$\begin{aligned} H_0' : \Delta(\boldsymbol{\gamma}) &\equiv E_P(\varphi^*(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)) \\ &\quad - E_Q(\varphi^*(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)) \qquad (2.2.6) \\ &= 0, \forall \boldsymbol{\gamma} \in \Gamma \end{aligned}$$

versus

$$H_a' : \Delta(\boldsymbol{\gamma}) \neq 0, \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma.$$

P denotes the (unrestricted) joint distribution of the random vector $W = (X, Y, Z)$, Q denotes the (restricted) joint distribution of W with $Y \perp X \mid Z$. E_P and E_Q denote the expectations with respect to P and Q , respectively. The function we choose to measure the distance of P from Q is

$$\varphi^* \equiv \varphi f_Z,$$

where φ is a univariate *Generically Comprehensively Revealing* (GCR) function. And the index parameter vector is

$$\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1, \gamma_2, \gamma_3)' \in \Gamma \subset \mathbb{R}^{1+d_W}$$

with Γ having non-empty interior. “Essentially any” $\boldsymbol{\gamma} \in \Gamma$ means that the set of “bad” $\boldsymbol{\gamma}$’s, $\{\boldsymbol{\gamma} \in \Gamma$:$\Delta_\varphi(\boldsymbol{\gamma}) = 0$ and $Y \not\perp X \mid Z\}$, has Lebesgue measure zero and is not dense in Γ.

To understand the moment conditions given in the equivalent null hypothesis, we notice

$$\begin{aligned}
& \Delta(\boldsymbol{\gamma}) \\
= \;& E_P\left[\varphi^*(W;\boldsymbol{\gamma})\right] - E_Q\left[\varphi^*(W;\boldsymbol{\gamma})\right] \qquad (2.2.7) \\
= \;& E_P\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] \\
& -E_Q\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] \\
= \;& \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)dF_{XYZ}(x,y,z) \\
& -\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)f_{Y|Z}(y|z)dy\; dF_{XZ}(x,z) \\
= \;& \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_Z(z)f_{XYZ}(x,y,z)dxdydz \\
& -\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{YZ}(y,z)f_{XZ}(x,z)dxdydz.
\end{aligned}$$

Instead of comparing $f_{XYZ}(x,y,z)f_Z(z)$ to $f_{YZ}(y,z)f_{XZ}(x,z)$, we are now comparing their transformation $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XYZ}(x,y,z)f_Z\left(z\right)dydxdz$ and $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XZ}(x,z)f_{YZ}\left(y,z\right)dydxdz$. Before the transformation, the density functions $f_{XYZ}(x,y,z)f_Z(z)$ and $f_{YZ}(y,z)f_{XZ}(x,z)$ are functions in (x,y,z), the data points, so that those functions can only be estimated at a nonparametric rate which is slower than $n^{-1/2}$. After the transformation, $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XYZ}(x,y,z)f_Z\left(z\right)dydxdz$ and $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XZ}(x,z)f_{YZ}\left(y,z\right)dydxdz$ are now functions in $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$. For each $\boldsymbol{\gamma}$, the transformation is an average of the data, so that semiparametric techniques could be used here to get a $n^{-1/2}$ rate. Essentially, we are comparing two functions by comparing an infinite number of their weighted averages. And the two comparison are equivalent because of the properties of the test functions we choose. Intuitively, a family of GCR functions indexed by $\boldsymbol{\gamma}$ is a class of functions with a span that comes arbitrarily close to any function. If we choose GCR functions as our test functions that run though an index space Γ and couldn't detect any difference between P and Q, then P and Q should agree for any function. Moreover, the index space Γ could be "small" as long as it has non-empty interior.

2.3 Integrated Conditional Moment Type Test

In chapter 1, I estimate $\Delta(\boldsymbol{\gamma})$ by its sample analog $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and choose a finite collection of $\boldsymbol{\gamma}$'s to construct a Chi-square test statistic. The consistency of the Chi-square test is due to randomization of the choice of $\boldsymbol{\gamma}$'s, which introduces a degree of arbitrariness into both the size and the power of the test. In this section, I will integrate out $\boldsymbol{\gamma}$ to get an Integrated Conditional Moment (ICM) type test statistic, following Bierens (1990), Bierens and Ploberger (1997) and Stinchcombe and White (1998).

2.3.A The Test Statistic

As I have showed, testing $H_0 : Y \perp X \mid Z$ vs. $H_a : Y \not\perp X \mid Z$ is equivalent to testing

$$H_0' : \Delta(\boldsymbol{\gamma}) = 0 \text{ for } \boldsymbol{\gamma} \in \Gamma \text{ where } \Gamma \text{ has a non-empty interior.}$$

If we in addition choose Γ to be compact, it turns out that $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ converges to a Gaussian process with a mean function $\sqrt{n}\Delta(\boldsymbol{\gamma})$. Under the null, that mean function is a zero function. In other words, if we view $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$, the estimator of $\Delta(\boldsymbol{\gamma})$, as a random function in $\boldsymbol{\gamma}$, we are testing if its mean function $\Delta(\boldsymbol{\gamma})$ is zero on Γ.

Based on $\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$, I can construct an integrated conditional moment test statistic

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

where μ is a probability measure on Γ which is chosen absolutely continuous with respect to the Lebesgue measure on Γ. As introduced in chapter 1, I use the sample analog $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ to estimate $\Delta\left(\boldsymbol{\gamma}\right)$:

$$\begin{aligned}\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3) \\ &\quad -\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3]K_h(Z_i-Z_j)\},\end{aligned}$$

where $K_h(\cdot)$ is defined as

$$K_h(u)\equiv\frac{1}{h^{d_u}}K(\frac{u}{h}),$$

with $K(\cdot)$ a symmetric product kernel density function, d_u the dimension of u, and bandwidth $h\equiv h_n$ depending on n.

The test statistic M_n uses an L^2 norm to integrate out $\boldsymbol{\gamma}$. We could also use other norms to do this. For example, we can use a uniform norm to get another test statistic $\sup_{\Gamma}\left|\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right|$. Intuitively, which norm is better will depend on the underlying data generating process which is unknown.

2.3.B Asymptotic Distribution of the Test Statistic

The proposed ICM type test statistic M_n is a functional of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$. To derive its asymptotic distribution, the key is to show that $\sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta\left(\boldsymbol{\gamma}\right)\right]$ converges to a Gaussian process. In chapter 1 I have showed that under certain assumptions, for a finite collection of $\boldsymbol{\gamma}$'s, $\Gamma_s\equiv\{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,...,\boldsymbol{\gamma}_s\}\subset\Gamma$, The vector $\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)=[\bar{\Delta}_{n,h_1}(\boldsymbol{\gamma}_1),\bar{\Delta}_{n,h_2}(\boldsymbol{\gamma}_2),...\bar{\Delta}_{n,h_s}(\boldsymbol{\gamma}_s)]'$ is asymptotically normal after proper centering and scaling, i.e. $\sqrt{n}\left(\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)-\Delta(\Gamma_s)\right)\xrightarrow{d}N\left(0,\Omega\right)$. If we in addition choose Γ to be compact, we can further show a stronger result, i.e. the weak convergence of $\sqrt{n}\left[\bar{\Delta}_{n,h}(\cdot)-\Delta\left(\cdot\right)\right]$ to a Gaussian process on Γ with a mean function zero. We keep assumptions 2-6 in chapter 1 and restate those assumptions here

Assumption 2 (Kernel function) Let $q\geq 2$ be an even integer. The kernel K is a product of a symmetric qth order univariate kernel $k : \mathbb{R} \to \mathbb{R}$ s.t. $\int k(v)dv = 1$, $\int v^j k(v)dv = 0$ for $j = 1, 2, ...q - 1$, and $0 < \int v^q k(v)dv < \infty$.

Assumption 3 Z_i takes values in the interior of the support of Z, $i = 1, 2, ...$.

Assumption 4 (Smoothness of the densities) The density of Z, f_Z, is continuously differentiable of order q; and all partial derivatives of $f_{YZ}(y, z)$, $f_{XZ}(y, z)$, $f_{XYZ}(x, y, z)$ with respect to z of order q exist.

Assumption 5 $\varphi(\cdot)$ is a bounded GCR function.

Assumption 6 (Bandwidth) The bandwidth $h \equiv h_n$, satisfies

(**A 6.1)** $nh^{d_Z} \to \infty$ as $n \to \infty$, and

(**A 6.2)** $\sqrt{n}h^q = o(1)$, i.e. $h = o(n^{-1/(2q)})$ as $n \to \infty$.

In addition, we need the compactness of Γ which is stated in the following assumption:

Assumption 7 The index parameter space Γ is compact with non-empty interior, which is a subset of $\mathbb{R}^{1+d_W}$.

To show the weak convergence result, we first show the leading term of $\sqrt{n}\left[\bar{\Delta}_{n,h}(\cdot) - \Delta(\cdot)\right]$ converges to a zero mean Gaussian process and then show that the remainder term is negligible. I still use U-statistic theory and Taylor expansion to get the leading term as what I did in chapter 1, and the summarized result is as follows.

First, $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ is a U-statistic of degree 2:

$$\begin{aligned} \bar{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \{[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) \\ & \quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3] K_h(Z_i - Z_j)\} \qquad (2.3.1) \\ &= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{(n,2)} \kappa_h(W_i, W_j; \boldsymbol{\gamma}), \end{aligned}$$

where $\kappa_h(W_i, W_j; \boldsymbol{\gamma})$ is a symmetric kernel

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

As in chapter 1, we use H-decomposition to decompose $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ into three parts such that

$$\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) = \Delta_h(\boldsymbol{\gamma}) + 2H_{n,h,1}(\boldsymbol{\gamma}) + R_{n,h,1}(\boldsymbol{\gamma}) \qquad (2.3.2)$$

where

$$\Delta_h(\boldsymbol{\gamma}) \equiv E\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right] = E\left[\kappa_h(W_i, W_j)\right]$$

$$\begin{aligned} H_{n,h,1}(\boldsymbol{\gamma}) &\equiv \frac{1}{n} \sum_{i=1}^{n} \{\kappa_{h,1}(W_i; \boldsymbol{\gamma}) - \Delta_h(\boldsymbol{\gamma})\} \qquad (2.3.3) \\ & \kappa_{h,1}(W_i; \boldsymbol{\gamma}) \equiv E\left[\kappa_h(W_i, W_j; \boldsymbol{\gamma}) | W_i\right], \; i \neq j \qquad (2.3.4) \end{aligned}$$

and

$$\begin{aligned} R_{n,h,1}(\boldsymbol{\gamma}) &\equiv \bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta_h(\boldsymbol{\gamma}) - 2H_{n,h,1}(\boldsymbol{\gamma}) \qquad (2.3.5) \\ & \quad \text{and uncorrelated with } H_{n,h,1}(\boldsymbol{\gamma}). \end{aligned}$$

The first two terms $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and $H_{n,h,1}(\boldsymbol{\gamma})$ constitute the projection of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ and the remainder $R_{n,h,1}(\boldsymbol{\gamma})$ is a smaller term if assumption 6.1 holds, which is shown in detail in the proof of lemma 2 in chapter 1. Using Taylor expansion, I have shown in the proof of lemma 3 in chapter 1 that $\Delta_h(\boldsymbol{\gamma}) = \Delta(\boldsymbol{\gamma}) + O(h^q)$ and $H_{n,h,1}(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^n \{\kappa_1(W_i;\boldsymbol{\gamma}) - E[\kappa_1(W_i;\boldsymbol{\gamma})]\} + O_p(h^q)$, where

$$\begin{aligned}
\kappa_1(W_i;\boldsymbol{\gamma}) \equiv\; & \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i) \\
& -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{YZ}(y, Z_i)dy \qquad (2.3.6)\\
& +\frac{1}{2}\int \varphi(\gamma_0 + x\gamma_1 + y\gamma_2 + Z_i\gamma_3)f_{XYZ}(x, y, Z_i)dxdy \\
& -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x, Z_i)dx.
\end{aligned}$$

$O(h^q)$ and $O_p(h^q)$ should be smaller terms if assumption 6.2 holds.

To summarize, I have shown that under assumptions 1-6,

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

Define

$$\zeta_n(\boldsymbol{\gamma}) \equiv \frac{2}{\sqrt{n}}\sum_{i=1}^n \{\kappa_1(W_i;\boldsymbol{\gamma}) - E[\kappa_1(W_i;\boldsymbol{\gamma})]\}, \qquad (2.3.7)$$

which is the leading term of $\sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta(\boldsymbol{\gamma})\right]$. I will first show that $\zeta_n(\boldsymbol{\gamma})$ converges to a zero mean Gaussian process and then show that $\sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta(\boldsymbol{\gamma})\right]$ converges to the same zero mean Gaussian process. The results are given precisely in the following theorem.

Theorem 1 *Under assumptions 1-7,*

$$(a)\ \zeta_n(\boldsymbol{\gamma}) \Longrightarrow \mathcal{Z}(\boldsymbol{\gamma})$$

$$(b)\ \sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta\left(\boldsymbol{\gamma}\right)\right]\Longrightarrow\mathcal{Z}\left(\boldsymbol{\gamma}\right),$$

where $\boldsymbol{\gamma}\in\Gamma$, *and* $\mathcal{Z}$ *is a Gaussian process on* Γ *with a mean function zero and a covariance function*

$$\begin{aligned} cov\left(\mathcal{Z}(\boldsymbol{\gamma}_1),\mathcal{Z}(\boldsymbol{\gamma}_2)\right) &= 4cov\left[\ \kappa_1\left(W_i;\boldsymbol{\gamma}_1\right),\ \kappa_1\left(W_i;\boldsymbol{\gamma}_2\right)\right] && (2.3.8)\\ &\equiv \sigma_\Delta\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right). \end{aligned}$$

$\zeta_n(\boldsymbol{\gamma})$ *is as defined in (2.3.7),* $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ *is as defined in (2.3.1), and* $\kappa_1\left(W_i;\boldsymbol{\gamma}\right)$ *is as defined in (2.3.6). If in addition that* H_0 *holds, then*

$$T_n(\boldsymbol{\gamma})\equiv\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\Longrightarrow\mathcal{Z}\left(\boldsymbol{\gamma}\right).$$

Remark 1 *The theorem looks fine for* $\{X_i,Y_i,Z_i\}$ *a strictly stationary and absolutely regular with mixing coefficients* ψ_j *s.t.* $\sum\limits_{j=1}^{\infty}j^{1/(r-1)}\psi_j<\infty$.

Moreover, by applying the continuous mapping theorem (Billingsley (1999), p.20), I get the following corollary:

Corollary 2 *Under assumptions 1-7 and assuming* H_0 *holds, let* $m:C\left(\Gamma\right)\to\mathbb{R}^+$ *be* $\left\|\cdot\right\|_\infty$ *continuous and* $m\left(x\right)=0$ *if and only if* $x=0$. *Then*

$$m\left[T_n(\boldsymbol{\gamma})\right]\Longrightarrow m\left[\mathcal{Z}(\boldsymbol{\gamma})\right].$$

where $\boldsymbol{\gamma}\in\Gamma$ *and* $\mathcal{Z}\left(\boldsymbol{\gamma}\right)$ *is the zero mean Gaussian process with covariance function defined by (2.3.8).*

For example, we could choose m to be the L^2 norm to get an ICM test statistic

$$\begin{aligned} M_n\equiv m\left[T_n(\boldsymbol{\gamma})\right] &= \int\limits_\Gamma\left[T_n(\boldsymbol{\gamma})\right]^2d\mu\left(\boldsymbol{\gamma}\right) && (2.3.9)\\ &= n\int\limits_\Gamma\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2d\mu\left(\boldsymbol{\gamma}\right)\Rightarrow\int\limits_\Gamma\left[\mathcal{Z}(\boldsymbol{\gamma})\right]^2d\mu\left(\boldsymbol{\gamma}\right), \end{aligned}$$

where $\mu(\boldsymbol{\gamma})$ is a probability measure on Γ which is absolutely continuous with respect to the Lebesgue measure on Γ.

2.3.C Calculate the Critical Values

Under the null, the ICM type test statistic M_n has limiting distribution as a functional of a zero mean Gaussian process, whose covariance function depends on the data generating process. Hence the asymptotic critical values will depend on the data generating process and cannot be tabulated. In this section, I will use the conditional Monte Carlo approach suggested by Hansen (1996) to simulate the asymptotic null distribution. Additionally, I will give convenient case-independent upper bounds as suggested by Bierens and Ploberger (1997).

Simulate the Asymptotic Null Distribution

In this subsection, I applied the Monte Carlo approach provided by Hansen (1996) to simulate the asymptotic null distribution of the test statistic. The idea is to construct $T_n(\boldsymbol{\gamma})^*$ which follows a zero mean Gaussian process conditional on W_i. The conditional covariance function is

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

I have shown in corollary 7 of chapter 1 that under the Assumptions 1-6 and the null hypothesis,

$$\tilde{\sigma}_{\Delta}\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right) \overset{P}{\to} \sigma_{\Delta}\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right).$$

So $T_n(\boldsymbol{\gamma})^*$ should have the same unconditional limiting distribution as $T_n(\boldsymbol{\gamma})$ under the null, and we can simulate a large enough sample of $T_n(\boldsymbol{\gamma})^*$'s to approximate the limiting null distribution of $T_n(\boldsymbol{\gamma})$.

A candidate $T_n(\boldsymbol{\gamma})^*$ can be generated by generating $\{v_i\}_{i=1}^n$ to be IID standard normal random variables and setting

$$T_n(\boldsymbol{\gamma})^* = \frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}) v_i. \tag{2.3.10}$$

The following proposition states that $T_n(\boldsymbol{\gamma})^*$ has the same limiting distribution as $T_n(\boldsymbol{\gamma})$ under the null.

Proposition 3 *Under assumption 1-7 and H_0, $T_n(\boldsymbol{\gamma})^* \Longrightarrow \mathcal{Z}(\boldsymbol{\gamma})$ where $\boldsymbol{\gamma} \in \Gamma$ and $\mathcal{Z}(\boldsymbol{\gamma})$ is the zero mean Gaussian process with covariance function defined by (2.3.8). Hence $M_n^* \equiv m\left[T_n(\boldsymbol{\gamma})^*\right] = \int\limits_\Gamma \left[T_n(\boldsymbol{\gamma})^*\right]^2 d\mu(\boldsymbol{\gamma}) \Rightarrow \int\limits_\Gamma \left[\mathcal{Z}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma})$.*

Figure 2.1 shows the empirical PDF of M_n and M_n^* are pretty close.

Thus theorem 2 of Hansen (1996) is applicable to our case and we can simulate a large enough sample of M_n^*'s to approximate the distribution of $M_n \equiv m\left[T_n(\boldsymbol{\gamma})\right] = \int\limits_\Gamma \left[T_n(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma})$. To be specific, I execute the following procedure J times for $j = 1, ..., J$ to get$\{M_n^{j*}\}_{j=1}^J$:

- generate $\{v_{ij}\}_{i=1}^n$ IID $N(0,1)$ random variables
- set

$$\begin{aligned} T_n^j(\boldsymbol{\gamma})^* &\equiv \frac{2}{\sqrt{n}} \sum_{i=1}^n \hat{\kappa}_{h,1}(W_i; \boldsymbol{\gamma}) v_{ij} \\ &= \frac{2}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[\frac{1}{n-1} \sum_{j=1, j \neq i}^n \kappa_h(W_i, W_j; \boldsymbol{\gamma}) \right] v_{ij} \right\} \end{aligned}$$

- set $M_n^{j*} \equiv m\left[T_n^j(\boldsymbol{\gamma})^*\right] = \int\limits_{\Gamma} \left[T_n^j(\boldsymbol{\gamma})^*\right]^2 d\mu\left(\boldsymbol{\gamma}\right)$

This gives a simulated sample $\left(M_n^{1*}, ..., M_n^{J*}\right)$, 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 percentage of $\{M_n^{j*}\}_{j=1}^J$ which exceed M_n to get the simulated asymptotic p value. We reject the null hypothesis if the size of test is greater than the simulated p value. As argued in Hansen (1996), J is under the control of the econometrician and can be chosen to be large enough in order to get a good enough approximation.

Upper Bounds of the Critical Values

Although the conditional Monte Carlo approach is straightforward, it needs simulation. Bierens and Ploberger (1997) suggested case-independent upper bounds demanding lighter computation. Their results can be applied here. I restate theorem 7 in Bierens and Ploberger (1997) in terms of our test:

Theorem 4 *(Theorem 7 in Bierens and Ploberger (1997)) Let ε_j be IID $N(0,1)$ and let*

$$\bar{W} = \sup_{m \geq 1} \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j^2.$$

For $\eta > 0$, under the H_0 and assumptions 1-7,

$$\lim_{n \to \infty} \Pr\left[T_n > \eta \int \hat{\sigma}_{\Delta}\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}\right) d\mu\left(\boldsymbol{\gamma}\right)\right] \leq P\left[\bar{W} > \eta\right],$$

where

$$\begin{aligned}\hat{\sigma}_{\Delta}(\boldsymbol{\gamma},\boldsymbol{\gamma}) &= \hat{\sigma}^2_{\Delta}(\boldsymbol{\gamma}) \\ &= 4\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\kappa}_{h,1}(W_i;\boldsymbol{\gamma})\right]^2 - 4\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2 \\ &= 4\frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\kappa_h(W_i,W_j;\boldsymbol{\gamma})\right]^2\right\} - 4\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})\right]^2.\end{aligned}$$

Bierens and Ploberger (1997) simulated $\bar{W}$ using 10,000 replications, and they derived the 10%, 5%, and 1% quantiles:

$$\Pr\left(\bar{W} > 3.23\right) = 0.10;\ \Pr\left(\bar{W} > 4.26\right) = 0.05;\ \Pr\left(\bar{W} > 6.81\right) = 0.01.$$

Using the upper bounds, we would reject the null hypothesis at 5% significance level if

$$T_n > 4.26\int \hat{\sigma}_{\Delta}(\boldsymbol{\gamma},\boldsymbol{\gamma})\, d\mu(\boldsymbol{\gamma}).$$

2.3.D Global and Local Alternatives

The global alternatives which our conditional independence test is against could be defined as

$$H_a^G : f_Z(z)f_{XYZ}(x,y,z) = f_{YZ}(y,z)f_{XZ}(x,z) + \alpha(y,x,z) \tag{2.3.11}$$

where $\alpha(y,x,z)$ is a nontrivial nonzero function. Then under H_a^G, we have

$$\begin{aligned} & \Delta(\boldsymbol{\gamma}) \\ = & \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_Z(z) f_{XYZ}(x,y,z)dxdydz \\ & - \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_{YZ}(y,z) f_{XZ}(x,z)dxdydz \\ = & \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)\alpha(y,x,z)\, dxdydz, \end{aligned} \tag{2.3.12}$$

which should be nonzero for essentially all $\boldsymbol{\gamma} \in \Gamma$ because of the property of the GCR function φ. Form the proof of the theorem 1, we can see

$$\lim_{n\to\infty} \Pr(M_n \in \text{Rejection Region}) = 1. \tag{2.3.13}$$

That is the test is consistent. As the sample size increases, the test will eventually detect the alternative H_a^G from the null hypothesis.

The local alternative can be defined as the following:

$$H_a^L : f_Z(z)f_{XYZ}(x,y,z) = f_{YZ}(y,z)f_{XZ}(x,z) + \alpha\left(y,x,z\right)/\sqrt{n}, \tag{2.3.14}$$

where $\alpha\left(y,x,z\right)$ is a nontrivial nonzero function. As argued in section 2, H_a^L could be equivalently defined as

$$H_a^L : \Delta\left(\boldsymbol{\gamma}\right) = c_\varphi\left(\boldsymbol{\gamma}\right)/\sqrt{n}, \tag{2.3.15}$$

where

$$c_\varphi\left(\boldsymbol{\gamma}\right) = \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)\alpha\left(y,x,z\right)dxdydz.$$

The properties of GCR functions make $c_\varphi\left(\boldsymbol{\gamma}\right)$ to be nonzero for essentially any choice of $\boldsymbol{\gamma}$.

The following result summarizes the asymptotic properties of our test statistic under the local alternatives (2.3.15).

Proposition 5 *Under assumptions 1-7 and under the local alternative H_a^L,*

$$T_n(\boldsymbol{\gamma}) \equiv \sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) \Longrightarrow \mathcal{Z}_c\left(\boldsymbol{\gamma}\right),$$

where $\boldsymbol{\gamma} \in\Gamma$ and $\mathcal{Z}_c$ is a Gaussian process on Γ with a mean function $c_\varphi\left(\boldsymbol{\gamma}\right)$ and a covariance function

$$\begin{aligned} cov\left(\mathcal{Z}_c(\boldsymbol{\gamma}_1), \mathcal{Z}_c(\boldsymbol{\gamma}_2)\right) &= 4cov\left[\ \kappa_1\left(W_i;\boldsymbol{\gamma}_1\right),\ \kappa_1\left(W_i;\boldsymbol{\gamma}_2\right)\right] \qquad (2.3.16) \\ &= \sigma_\Delta\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right). \end{aligned}$$

Moreover,

$$M_n \equiv m\left[T_n(\boldsymbol{\gamma})\right] = \int_{\Gamma} \left[T_n(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right) \Rightarrow \int_{\Gamma} \left[\mathcal{Z}_c(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right). \tag{2.3.17}$$

2.3.E The Relationship Between the Chi-square and ICM Tests

The proposed Chi-square test in chapter 1 and the ICM test in this chapter are related. The Chi-square test statistic $S_n(\Gamma_s) \equiv n\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]' \hat{\Omega}^{-1}\left[\bar{\Delta}_{n,\mathbf{h}}(\Gamma_s)\right]$ is a quadratic form based on a finite number of $\bar{\Delta}_n(\boldsymbol{\gamma}_l)$'s with $\hat{\Omega}^{-1}$ as the weighting matrix. M_n is an average over a continuum number of $\left[T_n(\boldsymbol{\gamma})\right]^2$'s. In practice when we cannot get a closed form integral, we use Monte Carlo integration method for this high dimensional integral so that the approximation errors shrink at a faster rate. In that case, the integral is approximated by an average. If we choose μ to be a uniform distribution over Γ, M_n becomes an average of many $\left[T_n(\boldsymbol{\gamma})\right]^2$'s for different $\boldsymbol{\gamma}$'s. Then this practical M_n is a quadratic form with a large s and with the identity matrix as the weighting matrix. In other words, M_n exploits more $\boldsymbol{\gamma}$'s than S_n but ignores the heteroskedasticity and dependence among $\bar{\Delta}_n(\boldsymbol{\gamma}_l)$'s.

A variant ICM type statistic which takes care of the diagonal heteroskedasticity of the variance covariance matrix of Ω would be based on

$$\tilde{T}_n(\boldsymbol{\gamma}) \equiv \frac{\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_{\Delta}\left(\boldsymbol{\gamma}\right)}.$$

Hence we could use

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

as an alternative ICM test statistic and simulate its critical values by the same way as for M_n. The results for the variant statistic are summarized in the following corollary:

Corollary 6 *(a) Under assumptions 1-7 and assume* $\sigma_n(\boldsymbol{\gamma}) > 0$*,*

$$\tilde{T}_n(\boldsymbol{\gamma}) - \sqrt{n}\frac{\Delta(\boldsymbol{\gamma})}{\sigma_\Delta(\boldsymbol{\gamma})} = \sqrt{n}\left[\frac{\bar{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_\Delta(\boldsymbol{\gamma})} - \frac{\Delta(\boldsymbol{\gamma})}{\sigma_\Delta(\boldsymbol{\gamma})}\right] \Longrightarrow \tilde{\mathcal{Z}}(\boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} \in \Gamma$*, and* $\tilde{\mathcal{Z}}$ *is a Gaussian process on* Γ *with a mean function zero and a covariance function*

$$\begin{aligned} cov\left(\tilde{\mathcal{Z}}(\boldsymbol{\gamma}_1), \tilde{\mathcal{Z}}(\boldsymbol{\gamma}_2)\right) &= \frac{\sigma_\Delta(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)}{\sigma_\Delta(\boldsymbol{\gamma}_1)\,\sigma_\Delta(\boldsymbol{\gamma}_2)} \\ &\equiv \rho_\Delta(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2). \end{aligned} \tag{2.3.18}$$

If in addition that H_0 *holds, then*

$$\tilde{T}_n(\boldsymbol{\gamma}) \equiv \frac{\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_\Delta(\boldsymbol{\gamma})} \Longrightarrow \tilde{\mathcal{Z}}(\boldsymbol{\gamma}).$$

(b) Under assumptions 1-7, assume $\hat{\sigma}_n(\boldsymbol{\gamma}) > 0$*, and assume* H_0 *holds, let* $m : C(\Gamma) \to \mathbb{R}^+$ *be* $\|\cdot\|_\infty$ *continuous and* $m(x) = 0$ *if and only if* $x = 0$*. Then*

$$m\left[\tilde{T}_n(\boldsymbol{\gamma})\right] \Longrightarrow m\left[\tilde{\mathcal{Z}}(\boldsymbol{\gamma})\right].$$

(c) Under assumption 1-7 and assume $\hat{\sigma}_n(\boldsymbol{\gamma}) > 0$*,*

$$\tilde{T}_n(\boldsymbol{\gamma})^* \Longrightarrow \tilde{\mathcal{Z}}(\boldsymbol{\gamma})$$

where

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

with $\{v_{ij}\}_{i=1}^n$ *IID* $N(0,1)$ *random variables. Hence*

$$\tilde{M}_n^* \equiv m\left[\tilde{T}_n(\boldsymbol{\gamma})^*\right] = \int\limits_\Gamma \left[\tilde{T}_n(\boldsymbol{\gamma})^*\right]^2 d\mu(\boldsymbol{\gamma}) \Rightarrow \int\limits_\Gamma \left[\tilde{\mathcal{Z}}(\boldsymbol{\gamma})\right]^2 d\mu(\boldsymbol{\gamma}).$$

Figure 2.2 shows the empirical PDF of $\tilde{M}_n$ and $\tilde{M}_n^*$ are pretty close.

M_n, which is based on $T_n(\boldsymbol{\gamma}) = \sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$, is easier to calculate since it needs not to calculate $\hat{\sigma}_\Delta(\boldsymbol{\gamma})$, and it could get a sharper bound for the critical values since $\hat{\sigma}_\Delta(\boldsymbol{\gamma})$ is typically very small. But the Monte Carlo results in the next section suggest that $\tilde{M}_n$, which is based on $\tilde{T}_n(\boldsymbol{\gamma}) = \frac{\sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma})}{\hat{\sigma}_\Delta(\boldsymbol{\gamma})}$, gets a little better power for most experiments.

2.4 Monte Carlo Experiments

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

For all the simulations, I generate $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ IID. The bandwidth I use is a value close to $\hat{h}$ given in chapter 1 (1.3.23), and I use $h_0 = n^{-1/[3*(2q+d_Z)]}$ and $\tau = 0.5$ when calculating $\hat{h}$. As in chapter 1, I choose $\varphi(\cdot)$ to be the standard normal PDF, and $k(u)$ the sixth order Gaussian kernel. The number of replication is 100, and the number of simulated M_n^* or $\tilde{M}_n^*$ is 100.

2.4.A Size and Power Studies

DGP 1

I first generate the sample of $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ using the DGP 1 in chapter 1. DGP 1 is the following data generating process

$$
\begin{aligned}
Y &= \beta X + Z + \epsilon_Y \\
X &= Z + Z^2 + \epsilon_X
\end{aligned}
$$

where

$$
\begin{pmatrix} \epsilon_X \\ \epsilon_Y \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_X^2 \; 0 \\ 0 \; \sigma_Y^2 \end{pmatrix}\right) = N\left(0, \begin{pmatrix} 4 \; 0 \\ 0 \; 1 \end{pmatrix}\right)
$$

and

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

The null hypothesis we are testing is

$$H_0 : Y \perp X|Z,$$

which is true only when $\beta = 0$. As in chapter 1, we use

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

to indicate the strength of the dependence for X and Y, conditional on Z, which is suitable since $X|Z$ and $Y|Z$ are jointly normal. The power functions are plotted against ρ from -0.9 to 0.9.

Although selecting $\boldsymbol{\gamma}$ from any Γ, which is a compact set having a non-empty interior, should deliver a consistent test, in practice we should avoid choosing Γ which would make $|\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3|$ too large or too small. This is because the value of $\varphi(u)$ will be very close to zero if $|u|$ is too large and $\varphi(u)$ will be close to linear if $|u|$ is in a too small range. In both cases the test will not have a good power. In the simulation, I choose Γ which makes $|\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3|$ around one. To be specific, I choose $\boldsymbol{\gamma}$ $\sim unif(\Gamma)$ where $\Gamma = [\boldsymbol{\gamma}_{center} - 0.5, \boldsymbol{\gamma}_{center} + 0.5]$ with

$$\begin{aligned}
\gamma_{center,0} &\approx 1 - \left[\frac{\bar{X}}{std(\{X_{1i}\}_{i=1}^n)} + \frac{\bar{Y}}{std(\{Y_i\}_{i=1}^n)} + \frac{\bar{Z}}{std(\{Z_i\}_{i=1}^n)}\right] \\
\gamma_{center,1} &\approx \left(\frac{1}{std(\{X_{1i}\}_{i=1}^n)}, \frac{1}{std(\{X_{2i}\}_{i=1}^n)}\right) \\
\gamma_{center,2} &\approx \left(\frac{1}{std(\{Y_i\}_{i=1}^n)}\right) \\
\gamma_{center,3} &\approx \left(\frac{1}{std(\{Z_i\}_{i=1}^n)}\right).
\end{aligned}$$

The size and power does not look bad when the sample size is as small as 100, and it looks pretty good when the sample size reaches 200. The "non-standardized" results in figure 2.3 correspond to M_n and the "standardized" results in figure 2.4 correspond to $\tilde{M}_n$. The power functions show that $\tilde{M}_n$ performs better than M_n in this experiment. The reason might be that the standardized ICM test $\tilde{M}_n$ has more reasonable weights than M_n when summing up T_n's, just like GLS is more efficient than OLS.

DGP 2

The DGP 2 I simulate and test is the DGP 2 in chapter 1, a modification of the DGP 1. This time I am focusing on the outcome of the fat tailed distributions. So I choose ϵ_X and ϵ_Y to be student t distribution of degree 3:

$$\epsilon_X \sim 2t_3, \ \epsilon_Y \sim t_3, \ \epsilon_X \perp \epsilon_Y.$$

The power functions of M_n is plotted in figure 2.5 and the power functions of $\tilde{M}_n$ is plotted in figure 2.6. We can see the power is a little worse than the previous one with normal distributions.

DGP 3

The following DGP is the DGP 3 in chapter 1, which is again a modification of DGP 1. This time I choose both ϵ_X and ϵ_Y to be centered chi-square distributions:

$$\epsilon_X \sim 2\left(\chi_1^2 - 1\right), \ \epsilon_Y \sim \left(\chi_1^2 - 1\right), \ \epsilon_X \perp \epsilon_Y.$$

The power functions of M_n is plotted in figure2.7 and the power functions of $\tilde{M}_n$ is plotted in figure 2.8. The power looks a little better than that of DGP 1 which uses normal distributions:

2.4.B Comparison to Other Tests

In this section I will compare the standardized ICM test to other conditional independence tests by some simulations. Su and White's (2008) test essentially compares $f_{XYZ}(x,y,z)f_Z(z)$ to $f_{XZ}(x,z)f_{YZ}(y,z)$ and can detect local alternatives at a rate of $n^{-\frac{1}{2}}h^{-\frac{d_X+d_Y+d_Z}{4}}$. Su and White's (2007) test essentially compares $f_{Y|X,Z}(y|x,z)$ to $f_{Y|Z}(y|z)$ and can detect local alternatives at a rate of $n^{-\frac{1}{2}}h^{-\frac{d_X+d_Z}{4}}$. My test compares $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XYZ}(x,y,z)f_Z\,(z)$ $dydxdz$ to $\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{XZ}(x,z)f_{YZ}\,(y,z)\,dydxdz$ and can detect local alternatives at a rate of $n^{-1/2}$. We first compare all three tests using DGP1. Figure 2.9 shows the power functions when the sample size is 100. SW 2007 performs better on the negative correlation part while my test is better on the positive correlation part, and both tests performs better than the SW 2008. Figure 2.10 and figure 2.11 show the power functions when the sample size is increased to 200 and to 500, respectively. We can see the power of my test improves faster than the power of SW 2007, which again improves faster than the power of SW 2008. This result is consistent with the rates of local alternatives these tests could detect.

I also compares the power function of my test to the power functions of other tests, and the results are shown in figure 2.12, where the "t-test" represents the t-test for $\beta = 0$, "LG 1997 CM" represents the Cramer-von Mises type test statistic of Linton and Gozalo (1997), "LG 1997 KS" represents the Kolmogorov-Smirnov type test statistic of Linton and Gozalo (1997), "DG 2001 CM" and "DG 2001 KS" represent the Cramer-von Mises type test statistic and the Kolmogorov-Smirnov type test statistic of Delgado and Gonzalez-Manteiga (2001), respectively. Although our test loses power compared to the t-test, which is reasonable since t-test uses more information, it performs better than other nonparametric tests when the sample size is 500 for DGP1.

2.5 Application to Returns-To-Schooling Example

As stated in the introduction, one potential application of the test is to test a key assumption identifying causal effects. In this section, I will provide an example to demonstrate this.

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

$$\ln Y_i = \beta_0 + \beta_1 S_i + \beta_2 EXP_i + \beta_3 EXP_i^2 + U_i, \tag{2.5.1}$$

where the subscript i stands for 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 is the residual term with a mean of zero. β_1 is the effect of additional year of schooling on wage. Our goal is to consistently estimate β_1.

However, direct estimation of Mincer functions suffers from the well known ability bias problem, which is caused by the dependence of schooling on unobserved ability. To make this explicit, let U_i be defined as $U_i = A_i + \varepsilon_i$ and rewrite the Mincer equation (2.5.1) as

$$\ln Y_i = \beta_0 + \beta_1 S_i + \beta_2 EXP_i + \beta_3 EXP_i^2 + A_i + \varepsilon_i, \tag{2.5.2}$$

where A_i stands for unobserved "ability".

Over the years, one method empirical researchers have adopted to tackle the ability bias issue is to find proxies of ability, for example IQ or AFQT scores, and include those as regressors (e.g. Griliches and M. (1972), Griliches (1977), and

Blackburn and Neumark (1993)). In much empirical work, other terms affecting earnings are also included in the regression. Hence, the following empirical wage equation is often estimated

$$\ln Y_i = \beta_0 + \beta_1 S_i + TS_i \gamma + X_i \theta + V_i \tag{2.5.3}$$

where TS_i stands for ability proxies e.g. IQ or AFQT scores, and X_i includes EXP_i, EXP_i^2 and other regressors like tenure, region, sex, race, union, etc.

If the conditional independence assumption

$$A \perp S \mid (TS, X) \tag{2.5.4}$$

holds, regression on the empirical wage equation (2.5.3) will then deliver a consistent estimator of β_1, the effect of schooling on wage in the Mincer function (2.5.2). In fact, assumption (2.5.4) is the key assumption to identify β_1. It is called a "conditional exogeneity assumption" by White and Chalak (2006). That enforced the "ignorability" or "unconfoundedness" condition, also known as "selection on observables" in the literature. If assumption (2.5.4) holds, even if the separability of Mincer function (2.5.2) does not hold, we can still identify β_1 and consistently estimate it by various methods.

We cannot test the conditional independence assumption (2.5.4) directly since A is unobservable. However, following White H. and Chalak K. (2005), if we could find TS_2 s.t.

$$\begin{aligned} TS_2 &= f\left(A, TS, X, \eta\right) \\ \eta &\perp S \mid (A, TS, X), \end{aligned} \tag{2.5.5}$$

where f denotes some function form, then

$$A \perp S \mid (TS, X) \Rightarrow TS_2 \perp S \mid (TS, X).$$

Thus we can test the implied conditional independence condition

$$H_0 : TS_2 \perp S \mid (TS, X). \tag{2.5.6}$$

From (2.5.5), we get some guidance about how to choose TS_2. A candidate for TS_2 could be a vector of variables that are driven by A, TS, X and some error term η. Intuitively, TS and TS_2 could be error-ridden proxies for ability.

Now I would like to test (2.5.6) using some data set. The data I use are from the National Longitudinal Survey of Youth 1979 (NLSY 79). In particular, I use the data from the survey year 2000 and restrict the sample to white males[1]. I use the age-adjusted standardized AFQT in year 1980 as TS. TS_2 includes math and verbal scores for preliminary scholastic aptitude tests from 1981 high school transcripts. To satisfy (2.5.5), I use years of schooling beyond high school as S, so that TS_2 should not be affected by S. X consists of actual experience in survey year 2000 and total tenure with employer in survey year 2000. To implement the test, I choose $\varphi(\cdot)$ to be the standard normal p.d.f., and $k(\cdot)$ the sixth order Gaussian kernel. For the same reason stated in the Monte Carlo section, $\boldsymbol{\gamma}$ is chosen from a uniform distribution so that $|\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3|$ won't be too big or too small. At a size of 5%, I cannot reject the null hypothesis (2.5.6). That provides some evidence supporting the empirical wage equation (2.5.3) used by empirical researchers.

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

2.6 Concluding Remarks

In this chapter, I develop a flexible test for conditional independence, which is simple to implement yet powerful. It is consistent against any deviation from the null and achieves $\sqrt{n}$ local power.

Throughout the chapter, I assume the used data are IID. But the IID assumption is not essential for the results. We may extend the approach to a time series framework so that we could test, say, nonlinear Granger causality. Another extension could be to alter the test so that it could be used on mixed variables of Z. This need arises because in applied microeconomics, many variables are categorical or binary while for the current version of conditional independence test, Z is assumed continuous. A third extension could be further studies about the bandwidth selection problem. Currently I choose the bandwidth to minimize the mean square error of $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$. But ideally, one should choose optimal bandwidths considering the size and power tradeoff. That could be another topic for further research.

2.7 Appendix: Proofs

Proof of Theorem 1: (a) Obviously for a finite number of $\boldsymbol{\gamma}$'s, $\{\zeta_n(\boldsymbol{\gamma}_1)$, $\zeta_n(\boldsymbol{\gamma}_2)$, ... ,$\zeta_n(\boldsymbol{\gamma}_s)\}$ is asymptotically normal. Also, we assume $\boldsymbol{\gamma} \in \Gamma \subset \mathbb{R}^{1+d_W}$ with Γ a compact (hence totally bounded) set. To complete the proof, I need to show that $\zeta_n(\boldsymbol{\gamma})$ is stochastic equicontinuous (Andrews (1994), Billingsley (1999)).

To prove that, I use theorem 4-6 in Andrews (1994). Note that

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

by theorem 6 in Andrews (1994), I only need to verify that each of the four terms satisfies Ossiander's L^2 entropy condition.

For the first term

$$\begin{aligned}\varphi^*(W_i;\boldsymbol{\gamma}) &\equiv \varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)f_Z(Z_i)\\ &= \varphi\left(W_i;\boldsymbol{\gamma}\right)f_Z(Z_i),\end{aligned}$$

then it is a type IV class if I can verify that

$$E\left\{[f_Z(Z_i)]^2\sup_{\boldsymbol{\gamma}_1:\|\boldsymbol{\gamma}_1-\boldsymbol{\gamma}\|<\delta}\left|\varphi\left(W_i;\boldsymbol{\gamma}_1\right)-\varphi\left(W_i;\boldsymbol{\gamma}\right)\right|^2\right\}\leq C\delta^{\psi} \qquad (2.7.1)$$

for any $\boldsymbol{\gamma}\in\Gamma$, for any $\delta>0$ in a neighborhood of 0, and for some finite constants $C>0$ and $\psi>0$, where $\varphi\left(W_i;\boldsymbol{\gamma}\right)\equiv\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)$. Under assumption 5, $\varphi\left(W_i;\boldsymbol{\gamma}\right)$ 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, I can show that (2.7.1) holds by mean value theorem and Cauchy-Schwarz inequality.

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

(b)second, I show that the difference between these two terms is $o_p\left(1\right)$, uniformly in $\boldsymbol{\gamma}$, so that they have the same limiting distribution. I want to verify that

$$sup_{\boldsymbol{\gamma}\in\Gamma}|\sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma})-\Delta\left(\boldsymbol{\gamma}\right)\right]-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma})-E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}|=o_p\left(1\right).$$

Note that

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

The first term is

$$\sqrt{n}\left[\Delta_h\left(\boldsymbol{\gamma}\right)-\Delta\left(\boldsymbol{\gamma}\right)\right]=\sqrt{n}[C_3(\boldsymbol{\gamma})h^q+o(h^q)]$$

with

$$\begin{aligned}C_3(\boldsymbol{\gamma})\;=\;&E\left\{\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\int\frac{1}{q!}\left[\left(\frac{\partial^q}{\partial Z}\right)f_Z\left(Z_i\right)\left(u\right)^q\right]K(u)du\right\}\\ &-E\left[\int\varphi(\gamma_0+X_i'\gamma_1+y'\gamma_2+Z_i'\gamma_3)\frac{1}{q!}\left(\frac{\partial^q}{\partial Z}\right)f_{YZ}(y,Z_i)u^qK(u)dydu\right],\end{aligned}$$

so that

$$\begin{aligned}&\sup_{\gamma\in\Gamma}\left|\sqrt{n}\left[\Delta_h\left(\boldsymbol{\gamma}\right)-\Delta\left(\boldsymbol{\gamma}\right)\right]\right|\\ =\;&\sup_{\gamma\in\Gamma}\left|\sqrt{n}C_3(\boldsymbol{\gamma})h^q+\sqrt{n}o(h^q)\right|\\ =\;&o_p\left(1\right)\end{aligned}$$

since $C_3(\boldsymbol{\gamma})$ is continuous in $\boldsymbol{\gamma}$ and Γ is bounded, and $\sqrt{n}h^q=o_p\left(1\right)$ under assumption 6.2. Similarly, we can see that the other two terms are also $o_p\left(1\right)$, uniformly in $\boldsymbol{\gamma}$. ■

Proof of Corollary 2: The result follows from theorem 10 and the continuous mapping theorem. ■

Proof of Proposition 3: The proof is similar to that of theorem 2 in Hansen (1996). ■

Proof of Theorem 4: This is the theorem 7 in Bierens and Ploberger (1997).

■

Proof of Proposition 5: I have proven that under assumptions 1-7,

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

and

$$(b)\ \sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta\left(\boldsymbol{\gamma}\right)\right] \Longrightarrow \mathcal{Z}\left(\boldsymbol{\gamma}\right).$$

Inspecting the proof, we can see immediately that $T_n(\boldsymbol{\gamma}) \equiv \sqrt{n}\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) \Longrightarrow \mathcal{Z}_c\left(\boldsymbol{\gamma}\right)$ under the local alternative, with a mean function $c_{\varphi}\left(\boldsymbol{\gamma}\right)$ and a covariance function the same as $\mathcal{Z}\left(\boldsymbol{\gamma}\right)$'s. Hence the result follows. ■

Proof of Corollary 6: (a) Since

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

It's sufficient to show that $\sup_{\gamma}\left|\hat{\sigma}_n\left(\boldsymbol{\gamma}\right) - \sigma_{\Delta}\left(\boldsymbol{\gamma}\right)\right| = o_p(1)$ and $\hat{\sigma}_n\left(\boldsymbol{\gamma}\right) > 0$, where the former is implied in the proof of consistency of $\hat{\Omega}$ (theorem 6 of chapter 1) and the latter is assumed.

(b) Similar to corollary 2.

(c) Similar to proposition 3. ■

2.8 Figures

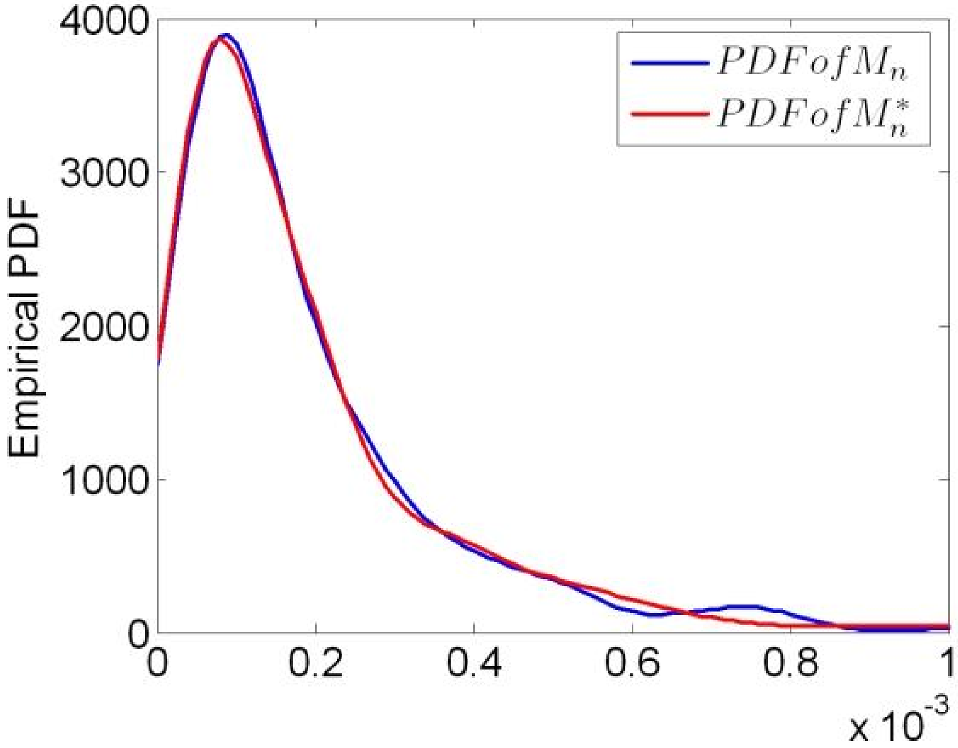

Figure 2.1: Conditional simulation for the asymptotic null distribution of M_n

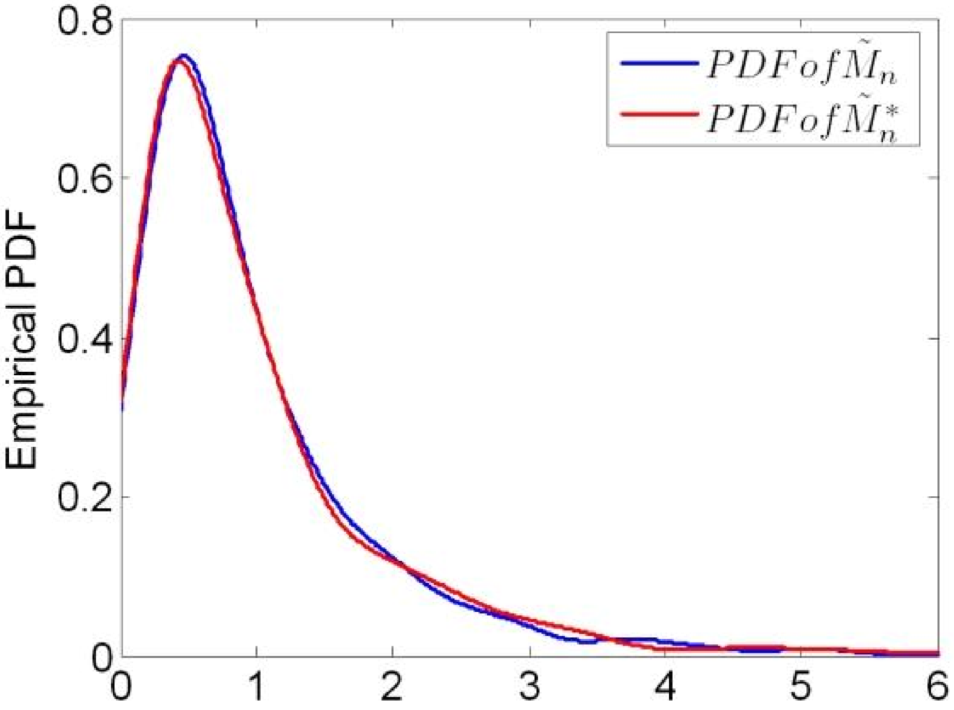

Figure 2.2: Conditional simulation for the asymptotic null distribution of $\tilde{M}_n$

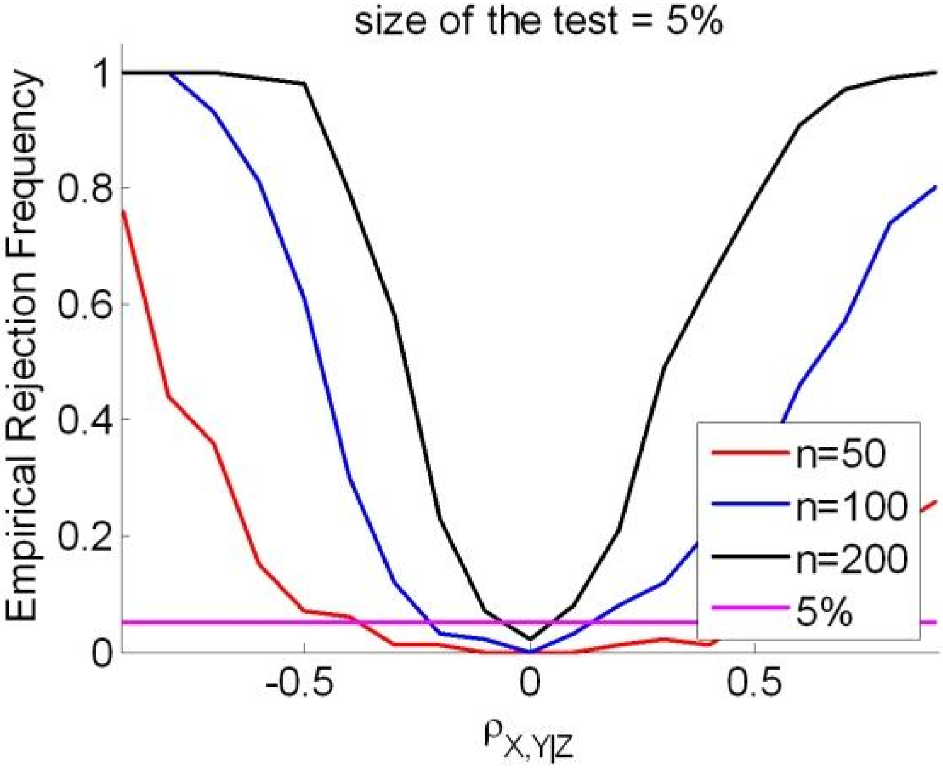

Figure 2.3: Power functions of non-standardized ICM test (M_n) for DGP 1

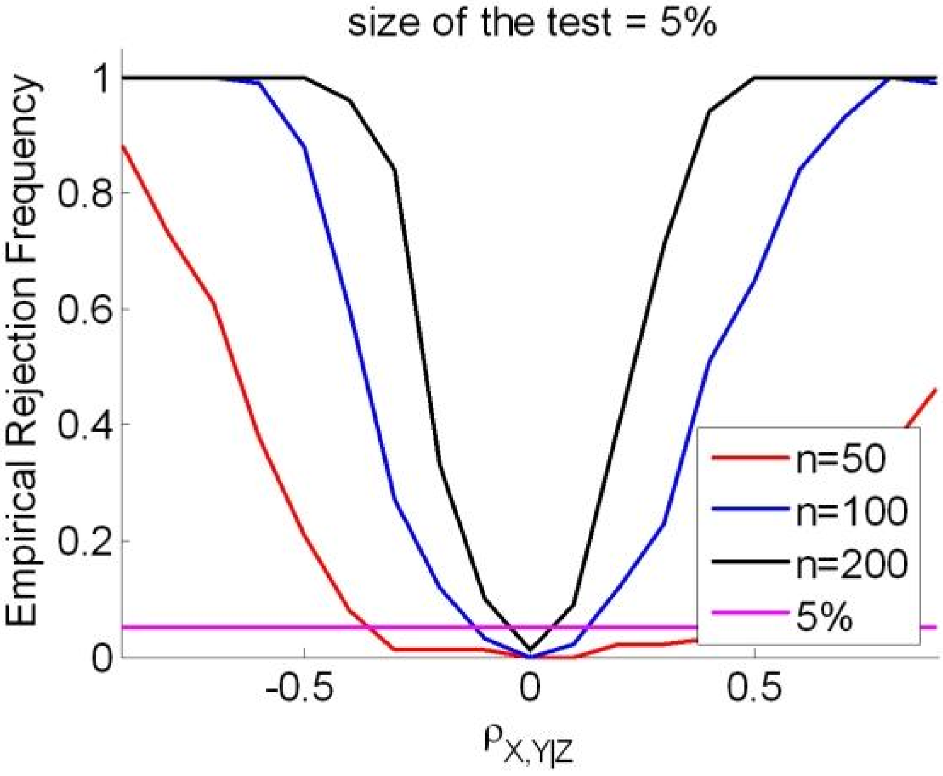

Figure 2.4: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 1

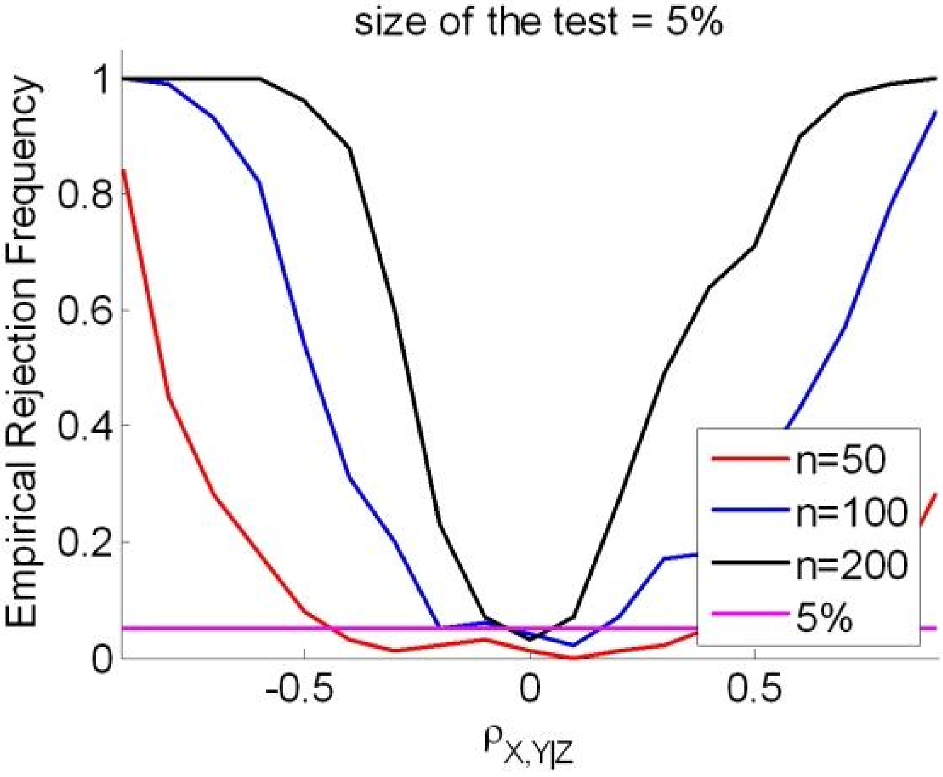

Figure 2.5: Power functions of non-standardized ICM test (M_n) for DGP 2

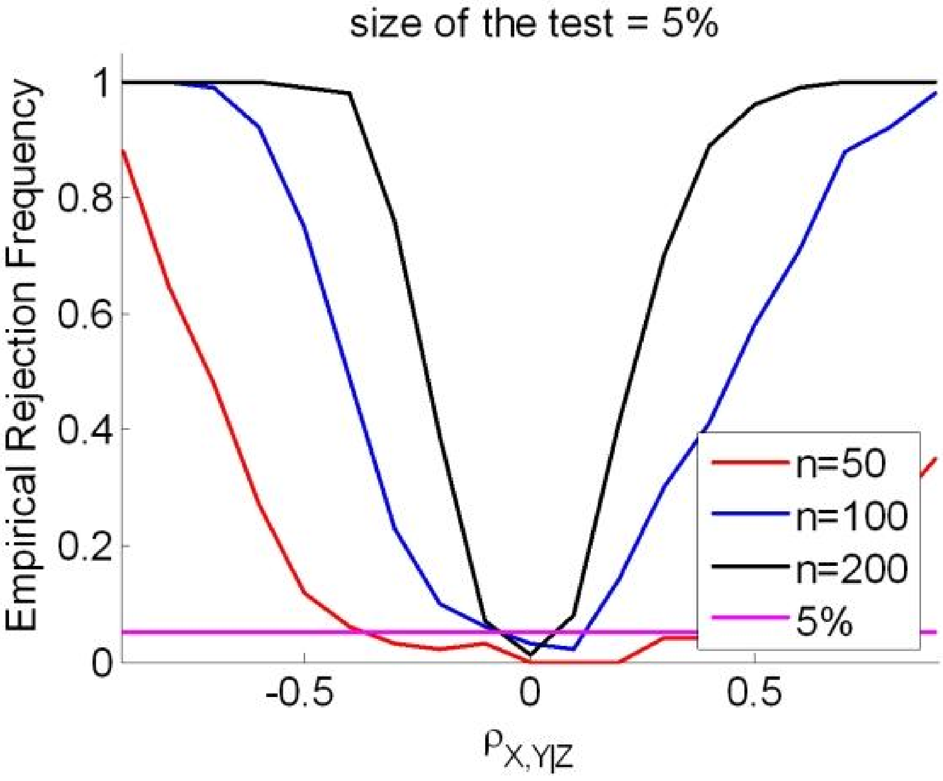

Figure 2.6: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 2

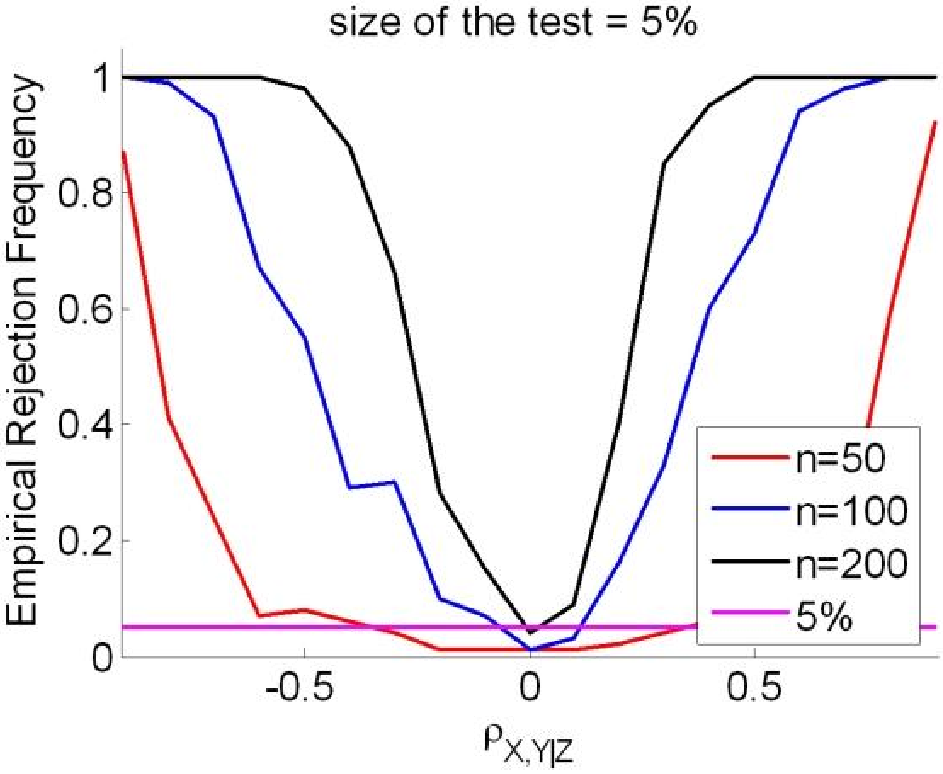

Figure 2.7: Power functions of non-standardized ICM test (M_n) for DGP 3

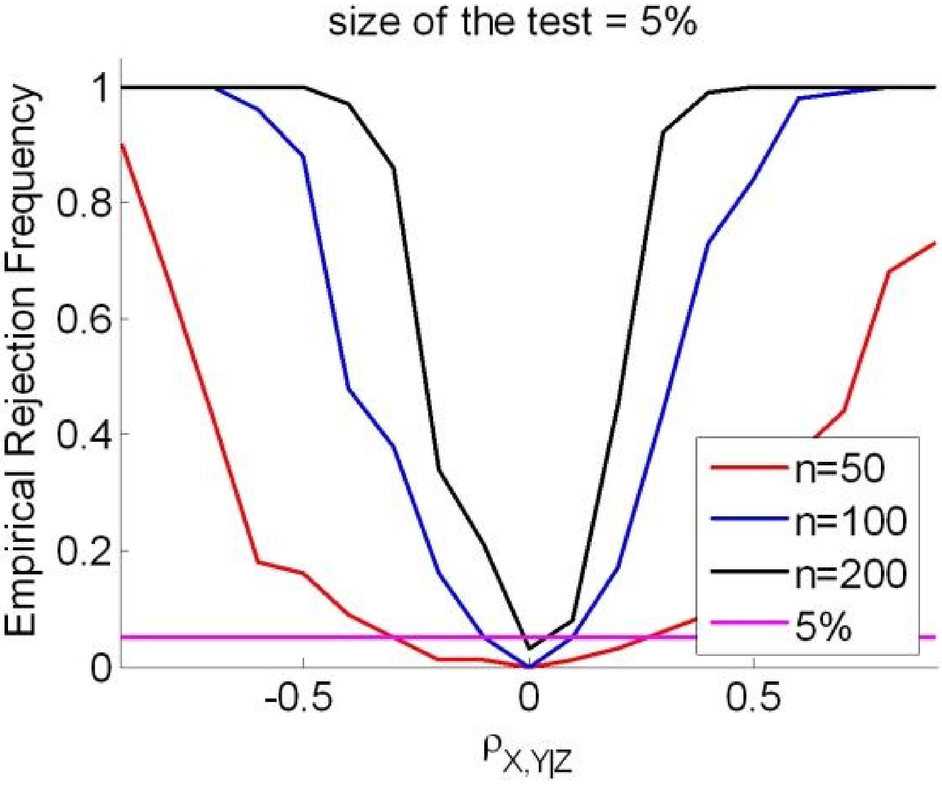

Figure 2.8: Power functions of standardized ICM test ($\tilde{M}_n$) for DGP 3

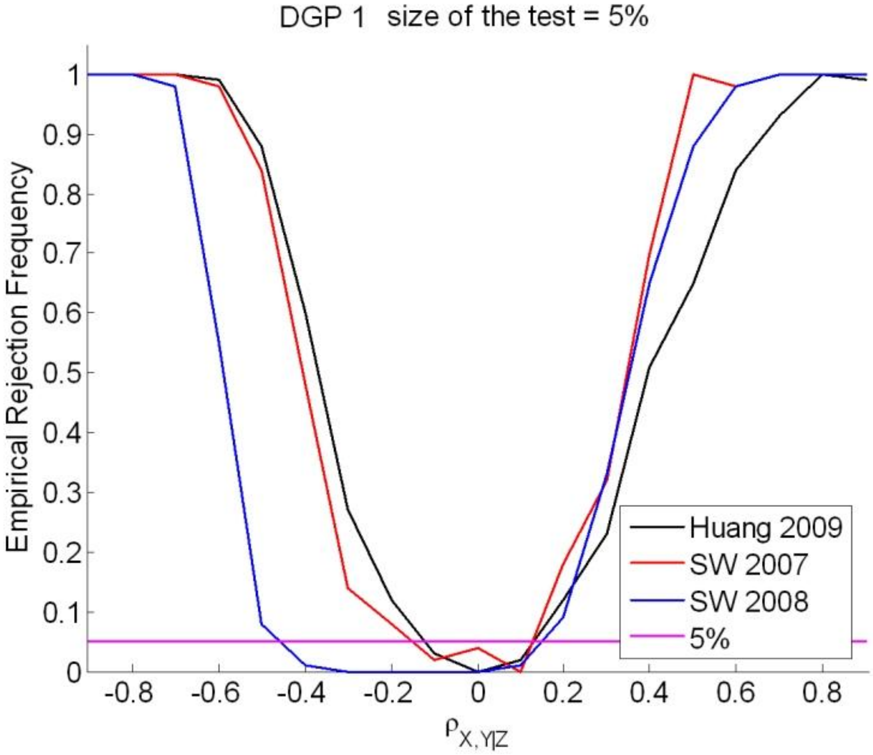

Figure 2.9: Comparison to SW 2007 and SW 2008, sample size 100

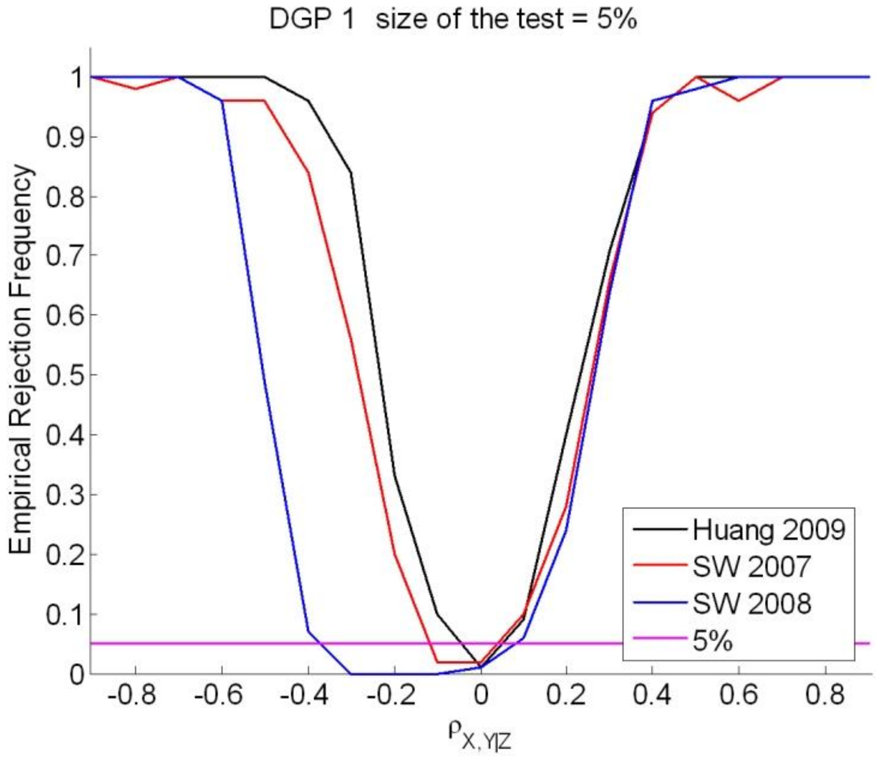

Figure 2.10: Comparison to SW 2007 and SW 2008, sample size 200

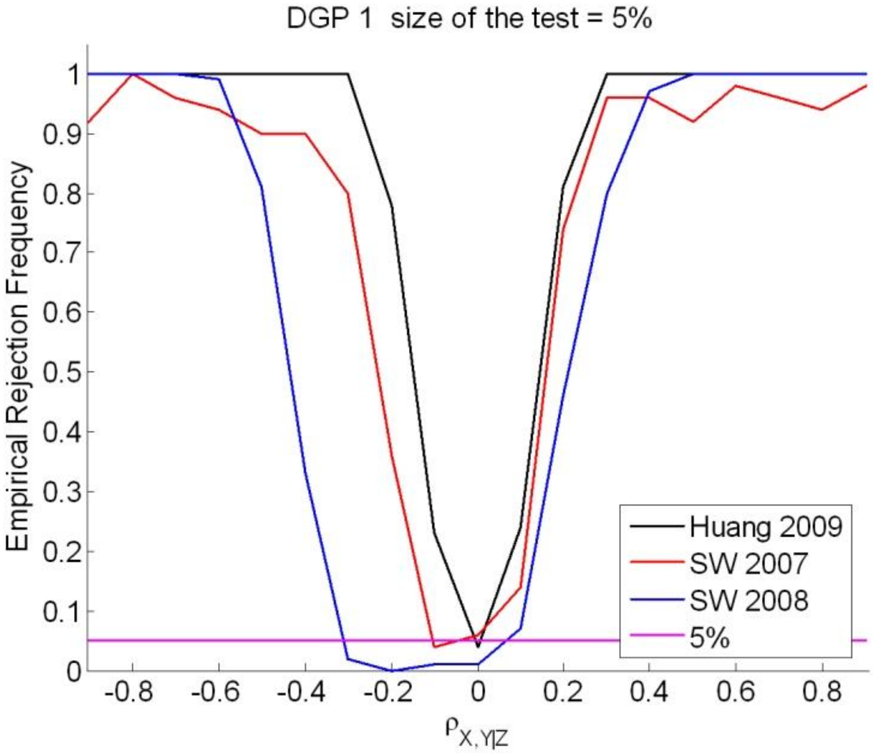

Figure 2.11: Comparison to SW 2007 and SW 2008, sample size 500

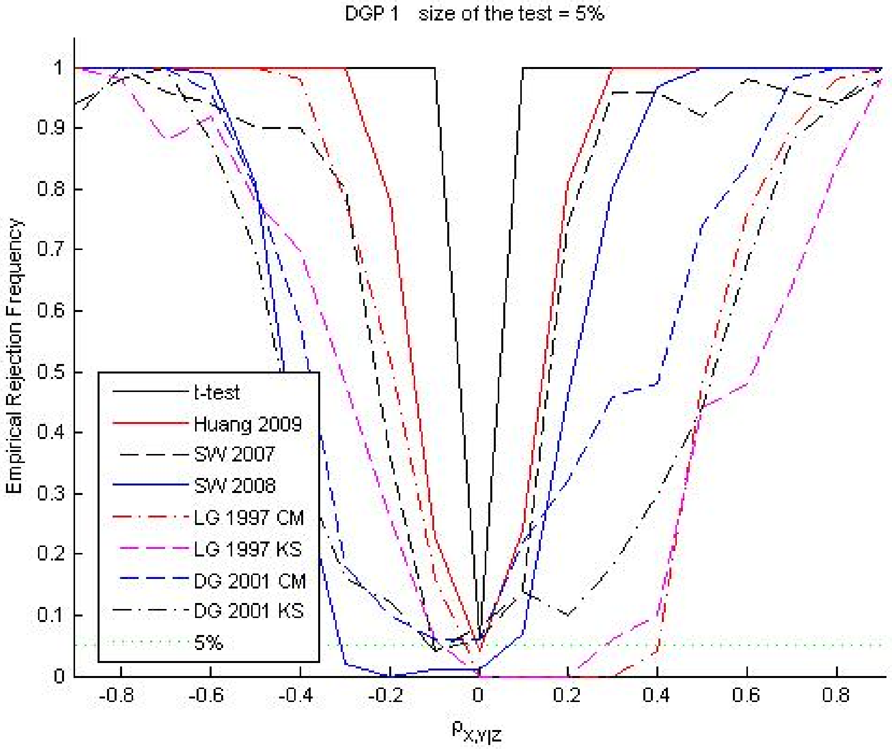

Figure 2.12: Comparison to other tests, sample size 500

3

Conditional Independence Tests for Mixed Discrete and Continuous Conditioning Random Variables

3.1 Introduction

In applied microeconomics, many variables are categorical or binary. For a conditional independence test, the conditioning variables are usually a mix of continuous and discrete variables. For example, in the returns-to-schooling example we discussed in the first two chapters, the conditioning variables usually include a number of discrete variables such as sex, race, union, or industry etc.

However, in previous chapters I assume the conditioning variables to be continuous, which makes the test difficult to directly apply. In this chapter, I extend the conditional independence test to incorporate the case of mixed conditioning random variables.

A straightforward way to do this would be using frequency estimators to handle the discrete variables. This means we split the sample into a collection of subsets ("cells") according to the value of discrete random variables and calculate the test statistics for each cell. For example, if the conditioning variables include one discrete variable, say sex, we can divide the data into two cells: the data for the male and the data for the female. If the conditioning variables include sex and race, we can divide the data into the subsamples for white males, black males, white females, black females, etc. Then we can apply the conditional independence tests introduced in previous chapters to each cell and get a test statistic based on that.

Li and Racine (2003, 2004 etc.) instead advocated the method of smoothing the discrete variables, which originates in the work of Aitchison and Aitken (1976). The idea is to use smooth estimators over different cells. The frequency approach could be viewed as a special case of the smoothing approach, where the frequency approach chooses zeroes as the smoothing parameters for discrete random variables. Li and Racine argued that although smoothing may introduce some estimation bias, it may also reduce the finite sample variance hence reduce the finite-sample mean squared error. Li and Racine (2003) suggested a cross-validation (CV) smoothing method to estimate an unknown distributions of categorical and continuous data. Racine and Li (2004) extended that method to estimate regression functions. Hsiao et al. (2007) applied that method to model specification test with mixed discrete and continuous data. In this chapter, I apply their idea to my conditional independence test.

The plan of this chapter is as follows. In section 2, I state the hypotheses I want to test and summarize the testing idea from the previous chapters. In section 3, I modify the test statistics introduced in previous chapters by the frequency approach, so that we can apply the test for the case where the conditioning random variables are mixed discrete and continuous. In section 4, I use a smoothing approach to deal with the mixed data. In section 5, I report some Monte Carlo results. Section 6 concludes.

3.2 The Hypotheses and the Idea of the Test

As in previous two chapters, we let X, Y, and Z be three random vectors, with dimensions d_X, d_Y, and d_Z, respectively. For convenience, I still assume that the sample observations $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ are independent and identically distributed (IID). As introduced in chapter 1 and 2 , the null hypothesis is that X and Y are independent given Z, and the alternative is that they are dependent given Z, i.e.

$$H_0 : Y \perp X \mid Z \text{ vs. } H_a : Y \not\perp X \mid Z. \tag{3.2.1}$$

In chapter 1, I established equivalent hypotheses in the form of moment conditions, i.e.

$$\begin{aligned} H_0' : \Delta(\boldsymbol{\gamma}) &\equiv E_P(\varphi^*(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)) \\ &\quad - E_Q(\varphi^*(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)) \\ &= 0, \forall \boldsymbol{\gamma} \in \Gamma \end{aligned} \tag{3.2.2}$$

versus

$$H_a' : \Delta(\boldsymbol{\gamma}) \neq 0, \text{ for essentially all } \boldsymbol{\gamma} \in \Gamma.$$

We restate the notations here: P denotes the unrestricted joint distribution of the random vector W, whereas Q denotes the (restricted) joint distribution of W with the null holding. E_P and E_Q represent the expectations with respect to P and Q , respectively. The distance between P and Q is measured using a family of tests functions $\varphi^* \equiv \varphi f_Z$ indexed by $\boldsymbol{\gamma} \equiv (\gamma_0, \gamma_1, \gamma_2, \gamma_3)' \in \Gamma \subset \mathbb{R}^{1+d_W}$, where φ is a univariate *Generically Comprehensively Revealing* (GCR) function and Γ has non-empty interior. "Essentially any" $\boldsymbol{\gamma} \in \Gamma$ means that the set of "bad" $\boldsymbol{\gamma}$'s, $\{\boldsymbol{\gamma} \in \Gamma{:}\Delta_{\varphi}(\boldsymbol{\gamma}) = 0 \text{ and } Y \not\perp X \mid Z\}$, has Lebesgue measure zero and is not dense in Γ. Note that $\Delta(\boldsymbol{\gamma})$ is an estimatable moment:

$$\begin{aligned}
& \Delta(\boldsymbol{\gamma}) \\
\equiv\ & E_P\left[\varphi^*(W;\boldsymbol{\gamma})\right] - E_Q\left[\varphi^*(W;\boldsymbol{\gamma})\right] \\
=\ & E_P\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3) f_Z(Z)\right] \\
& -E_Q\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3) f_Z(Z)\right] \\
=\ & \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_Z(z) dF_{XYZ}(x,y,z) \qquad (3.2.3) \\
& -\int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_Z(z) f_{Y|Z}(y|z) dy\ dF_{XZ}(x,z) \\
\equiv\ & E_{X,Y,Z}\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3) f_Z(Z)\right] - \int g_{XZ}(x,z;\boldsymbol{\gamma}) dF_{XZ}(x,z) \\
=\ & E_{X,Y,Z}\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3) f_Z(Z)\right] - E_{X,Z}\left[g_{XZ}(X,Z;\boldsymbol{\gamma})\right]
\end{aligned}$$

where

$$\begin{aligned}
g_{XZ}(x,z;\boldsymbol{\gamma}) &\equiv \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3) f_Z(z) f_{Y|Z}(y,z) dy \qquad (3.2.4) \\
&= E\left[\varphi(\gamma_0 + x'\gamma_1 + Y'\gamma_2 + z'\gamma_3) f_Z(z) | Z = z\right].
\end{aligned}$$

I use a sample analog $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ to estimate $\Delta(\boldsymbol{\gamma})$, and construct a Chi-square test in chapter 1, an Integrated Conditional Moment (ICM) test in chapter 2. In this chapter, I will modify $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ so that it is suitable for the mixed continuous and discrete conditioning variables, and then we can construct a Chi-square test or an ICM test in the same way as before.

3.3 Mixed Data and the Frequency Approach

3.3.A Mixed Discrete and Continuous Conditioning Random Variables

In this chapter, we allow the conditioning random variables to be mixed continuous and discrete. Let the conditioning random variables Z be denoted by

$$Z = \left(Z^{c\prime}, Z^{d\prime}\right)',$$

where Z^d is a vector of d_d categorical random variables and Z^c is a vector of d_c continuous random variables. Let z_s^d be the s-th component of z^d $(s = 1, 2, ..., d_d)$, which can assume c_s different values, where $c_s \geqslant 2$ is a positive integer. We assume that Z^d has finite support, so $z_s^d \in D_{Z_s^d} \equiv \left\{z_{s,1}^d, z_{s,2}^d, ..., z_{s,c_s}^d\right\}$. We denote the joint density of Z by $f_Z(z) = f_Z\left(z^c, z^d\right)$.

3.3.B Summarized Assumptions

I keep assumptions 5 and 7 in chapter 2 and modify other assumptions to incorporate the discrete conditioning random variables.

Assumption 1 $\{W_i \equiv (X_i', Y_i', Z_i')'\}$ *is an IID sequence of random variables on the complete probability space* $(\Omega_W, \mathcal{F}_W, P_W)$. The random vector $Z = \left(Z^{c'}, Z^{d\prime}\right)'$, where Z^d is a vector of d_d categorical random variables and Z^c is a vector of d_c

continuous random variables. $d_Z = d_c + d_{d.}$ *$X_i, Y_i,$ and Z_i take values in* $\mathbb{R}^{d_X}$, $\mathbb{R}^{d_Y}$*,and* $\mathbb{R}^{d_c} \times \Pi_{s=1}^{d_d} D_{Z_s^d}$*, respectively,* $d_W \equiv d_X + d_Y + d_Z$.

Assumption 2 (Kernel functions for continuous conditioning variables) Let $q \geq 2$ be an even integer. The kernel K^c is a product of a symmetric qth order univariate kernel $k : \mathbb{R} \to \mathbb{R}$ s.t. $\int k(v)dv = 1$, $\int v^j k(v)dv = 0$ for $j = 1, 2, ...q-1$, and $0 < \int v^q k(v)dv < \infty$.

Assumption 3 Z_i^c takes values in the interior of the support of Z^c, $i = 1, 2, ...$.

Assumption 4 (Smoothness of the densities) The density of Z^c, f_{Z^c}, is continuously differentiable of order q; and all partial derivatives of $f_Z(z)$, $f_{YZ}(y, z)$, $f_{XZ}(y, z)$, $f_{XYZ}(x, y, z)$ with respect to z^c of order q exist.

Assumption 5 $\varphi(\cdot)$ is a bounded GCR function.

Assumption 6 (Bandwidths for continuous conditioning variables) The bandwidth for the continuous kernel, $h_s \equiv h_{s,n}$ for $s = 1, 2, ..., d_c$, satisfies

(A 6.1) $n\Pi_{s=1}^{d_c} h_s \to \infty$ as $n \to \infty$, and

(A 6.2) $\sqrt{n}h_s^q = o_p(1)$, i.e. $h_s = o_p(n^{-1/(2q)})$ as $n \to \infty$.

Assumption 7 The index parameter space Γ is compact with non-empty interior.

3.3.C The Frequency Approach

Our test is based on the distance of the restricted and unrestricted joint probabilities. The distance is indirectly measured by

$$\Delta\left(\boldsymbol{\gamma}\right) = E_{X,Y,Z}\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)\right] - E_{X,Z}\left[g_{XZ}(X, Z; \boldsymbol{\gamma})\right] \quad (3.3.1)$$

where

$$g_{XZ}(x, z; \boldsymbol{\gamma}) = E\left[\varphi(\gamma_0 + x'\gamma_1 + Y'\gamma_2 + z'\gamma_3)f_Z(z)|Z = z\right].$$

We thus need to estimate $\Delta\left(\boldsymbol{\gamma}\right)$ to construct the test statistic. To do that, we can estimate $f_Z(z)$ and $g_{XZ}(x, z; \boldsymbol{\gamma})$, then use a sample average to estimate the expectations.

Discrete Conditioning Random Variables

A special case would be that the conditioning random vector is discrete. That is, $d_c = 0$ and $Z = Z^d$. Then f_Z would be the probability density function of Z and can be estimated by the leave-one-out estimator

$$\hat{f}_Z(Z_i) = \hat{f}_{Z^d}(Z_i^d) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} 1\left(Z_j = Z_i\right).$$

The conditional mean g_{XZ} can be estimated by

$$\hat{g}_{XZ}(X_i, Z_i; \boldsymbol{\gamma}) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)1\left(Z_j = Z_i\right).$$

Hence the estimator of $\Delta\left(\boldsymbol{\gamma}\right)$ would be

$$\begin{aligned}
\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\hat{f}_Z(Z_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{XZ}(X_i, Z_i; \boldsymbol{\gamma}) \\
&= \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\{[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) \\
&\quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3]1\left(Z_j = Z_i\right)\}.
\end{aligned}$$

Remark 1 *If Z is discrete, there is no need to use a density weighted expectation (see the reasoning in chapter 1). So we can instead based our test statistic on*

$$\begin{aligned}&\Delta_{\varphi}\left(\boldsymbol{\gamma}\right)\\ =\;& E_P\left[\varphi(W;\boldsymbol{\gamma})\right]-E_Q\left[\varphi(W;\boldsymbol{\gamma})\right]\\ =\;& E_{X,Y,Z}\left[\varphi(\gamma_0+X'\gamma_1+Y'\gamma_2+Z'\gamma_3)\right]-E_{X,Z}\left[g_{\varphi;XZ}(X,Z;\boldsymbol{\gamma})\right],\end{aligned}$$

where

$$\begin{aligned}g_{\varphi;XZ}(x,z;\boldsymbol{\gamma}) &\equiv \int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_{Y|Z}(y,z)dy\\ &= E\left[\varphi(\gamma_0+x'\gamma_1+Y'\gamma_2+z'\gamma_3)|Z=z\right].\end{aligned}$$

And we can then estimate $\Delta_{\varphi}\left(\boldsymbol{\gamma}\right)$ by

$$\begin{aligned}\bar{\Delta}_{\varphi;n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\right]-\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{\varphi;XZ}(X_i,Z_i;\boldsymbol{\gamma})\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\right]\\ &\quad -\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\left[\frac{1}{\sum_{j=1,j\neq i}^{n}1\left(Z_j=Z_i\right)/\left(n-1\right)}\right.\\ &\quad\left.\times\sum_{j=1,j\neq i}^{n}\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3)1\left(Z_j=Z_i\right)\right].\end{aligned}$$

Mixed Conditioning Random Variables

Now suppose we have mixed discrete and continuous Z. The leave-one-out version of the nonparametric kernel estimator of $f_Z\left(Z_i\right)$ would be

$$\hat{f}_Z(Z_i)=\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}K_h(Z_i^c-Z_j)1\left(Z_j^d=Z_i^d\right),$$

and the estimator for the conditional mean $g_{XZ}(X_i,Z_i;\boldsymbol{\gamma})$ would be

$$\hat{g}_{XZ}(X_i,Z_i;\boldsymbol{\gamma})=\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3)K_h^c(Z_i^c-Z_j)1\left(Z_i^d=Z_i^d\right).$$

Hence the estimator of $\Delta\left(\boldsymbol{\gamma}\right)$ would be

$$\begin{aligned}
\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)\hat{f}_Z(Z_i)\right]-\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{XZ}(X_i,Z_i;\boldsymbol{\gamma}) \\
&= \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3) \qquad (3.3.2) \\
&\quad -\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3]K_h^c(Z_i^c-Z_j)1\left(Z_j^d=Z_i^d\right)\},
\end{aligned}$$

where $K_h^c(\cdot)$ is defined as

$$K_h^c(u)\equiv\frac{1}{h^{d_u}}K^c(\frac{u}{h}),$$

with $K^c(\cdot)$ a symmetric product kernel density function, d_u the dimension of u, and the bandwidth $h\equiv h_n$ depending on n.

The frequency estimator $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ is a linear combination of the estimators defined in chapter 2 for each cell. If we split the sample into a collection of cells according to the value of Z^d and denote the number of observations in each cell as $n_{z^d}\equiv\sum_{i=1}^{n}1\left(Z_i^d=z^d\right)$, then the estimator for each cell is

$$\begin{aligned}
\bar{\Delta}_{n,h,z^d}(\boldsymbol{\gamma}) &= \frac{1}{n_{z^d}\left(n_{z^d}-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3) \\
&\quad -\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3]K_h(Z_i-Z_j)1\left(Z_j^d=Z_i^d=z^d\right)\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) \\
&= \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3) \qquad (3.3.3) \\
&\quad -\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3]K_h^c(Z_i^c-Z_j)1\left(Z_j^d=Z_i^d\right)\} \\
&= \sum_{z^d}\bar{\Delta}_{n,h,z^d}(\boldsymbol{\gamma})\cdot\tilde{f}_{Z^d}\left(z^d\right)\cdot\hat{f}_{Z^d}\left(z^d\right)
\end{aligned}$$

where $\sum_{z^d}$ denotes the summation over the set $\Pi_{s=1}^{d_d}D_{Z_s^d}$,

$$\tilde{f}_{Z^d}\left(z^d\right)\equiv\frac{1}{n}\sum_{i=1}^{n}1\left(Z_i^d=z^d\right),$$

and

$$\hat{f}_{Z^d}\left(z^d\right) = \frac{1}{n-1}\left[\sum_{j=1}^{n} 1\left(Z_i^d = z^d\right) - 1\right].$$

Both $\tilde{f}_{Z^d}\left(z^d\right)$ and $\hat{f}_{Z^d}\left(z^d\right)$ are estimators for $f_{Z^d}\left(z^d\right)$, so that $\bar{\Delta}_{n,h}(\boldsymbol{\gamma})$ can be approximately viewed as a linear combination of the estimators for each subsample with the weight $\left[f_{Z^d}\left(z^d\right)\right]^2$.

3.4 Discrete Kernels and The Smoothing Approach

Li and Racine (2003, 2004, 2007 etc.) recommend using a smoothing method to deal with the mixed data instead of splitting the sample into a number of cells. I apply their idea to our test in this section.

3.4.A Discrete Kernels

Now we modify the product kernel K_h to be

$$K_{\lambda,h} = K_\lambda^d \cdot K_h^c$$

where K_λ^d is a product kernel with a smoothing parameter vector λ and K_h^c is a product kernel with a smoothing parameter vector h. K_λ^d is introduced for discrete variables Z^d while K_h^c is a high-order-bias-reduction kernel for continuous variables Z^c as before.

We use the kernel functions suggested by Racine and Li (2004) which have a simple form and treat ordered and unordered discrete random variables separately. Let $Z^d = (\bar{Z}^{d\prime}, \tilde{Z}^{d\prime})'$. $\bar{Z}^d$ denote categorical random variables which have no natural ordering while $\tilde{Z}^d$ denotes discrete random variables which have a natural ordering. Correspondingly, the kernel for $\bar{Z}^d$ is defined as $\bar{k}^d_\lambda\left(\bar{Z}^d_{is}, \bar{z}^d_s\right) = 1$ if $\bar{Z}^d_{is} = \bar{z}^d_s$ and $\bar{k}^d_\lambda\left(\bar{Z}^d_{is}, \bar{z}^d_s\right) = \lambda_s$ if $\bar{Z}^d_{is} \neq \bar{z}^d_s$, and the kernel for $\tilde{Z}^d$ is defined as $\tilde{k}^d_\lambda\left(\tilde{Z}^d_{is}, \tilde{z}^d_s\right) = 1$ if $\tilde{Z}^d_{is} = \tilde{z}^d_s$ and $\tilde{k}^d_\lambda\left(\tilde{Z}^d_{is}, \tilde{z}^d_s\right) = \lambda_s^{\left|\tilde{Z}^d_{is} - \tilde{z}^d_s\right|}$ if $\tilde{Z}^d_{is} \neq \tilde{z}^d_s$, $\lambda_s \in [0,1]$. Then the product kernel K^d_λ is

$$\begin{aligned} K^d_\lambda\left(Z^d_i, z^d\right) &= \left[\prod_{s=1}^{\bar{d}_d} \bar{k}^d_\lambda\left(\bar{Z}^d_{is}, \bar{z}^d_s\right)\right]\left[\prod_{s=\bar{d}_d+1}^{d_d} \tilde{k}^d_\lambda\left(\tilde{Z}^d_{is}, \tilde{z}^d_s\right)\right] \\ &= \left[\prod_{s=1}^{\bar{d}_d} \lambda_s^{1\left(\bar{Z}^d_{is} \neq \bar{z}^d_s\right)}\right]\left[\prod_{s=\bar{d}_d+1}^{d_d} \lambda_s^{\left|\tilde{Z}^d_{is} - \tilde{z}^d_s\right|}\right]. \end{aligned}$$

Note that here we allow the smoothing parameters λ's to be different for different discrete random variables. If we also allow the smoothing parameters h's to be different for different continuous random variables, we get the kernel

$$\begin{aligned} & K_{\lambda,h}\left(Z_i, z\right) \\ =\; & K^d_\lambda \cdot K^c_h \\ =\; & K^d_\lambda\left(Z^d_i, z^d\right) \cdot K^c_h\left(Z^c_i - z^c\right) \qquad (3.4.1) \\ =\; & \left[\prod_{s=1}^{\bar{d}_d} \bar{k}^d_\lambda\left(\bar{Z}^d_{is}, \bar{z}^d_s\right)\right]\left[\prod_{s=\bar{d}_d+1}^{d_d} \tilde{k}^d_\lambda\left(\tilde{Z}^d_{is}, \tilde{z}^d_s\right)\right] \cdot \prod_{s=1}^{d_c} \frac{1}{h_s} k\left(\frac{Z^c_{is} - z^c_s}{h_s}\right) \\ =\; & \left[\prod_{s=1}^{\bar{d}_d} \lambda_s^{1\left[\bar{Z}^d_{is} \neq \bar{z}^d_s\right]}\right]\left[\prod_{s=\bar{d}_d+1}^{d_d} \lambda_s^{\left|\tilde{Z}^d_{is} - \tilde{z}^d_s\right|}\right] \cdot \prod_{s=1}^{d_c} \frac{1}{h_s} k\left(\frac{Z^c_{is} - z^c_s}{h_s}\right). \end{aligned}$$

3.4.B The Test Statistic

Replacing the kernel K_h defined in previous chapters by the new one, we can see that the estimator for $\Delta\left(\boldsymbol{\gamma}\right)$ becomes

$$\begin{aligned}
\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\hat{f}_Z(Z_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{XZ}(X_i, Z_i; \boldsymbol{\gamma}) \\
&= \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) \\
&\quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3]K_{\lambda,h}(Z_i, Z_j)\}.
\end{aligned} \tag{3.4.2}$$

When λ takes the value zero, $K_{\lambda,h}(Z_i, Z_j)$ becomes $K_h^c(Z_i^c - Z_j)1\left(Z_j^d = Z_i^d\right)$ and $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ becomes the frequency estimator.

Based on $\sqrt{n}\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$, I can construct an integrated conditional moment test statistic in the same way as in chapter 2:

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

where μ is a probability measure on Γ which is chosen to be absolutely continuous with respect to Lebesgue measure on Γ. Under the null, we expect M_n is close to zero.

A standardized version of the ICM type statistic introduced in chapter 2 is

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

where

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

with a consistent estimator of variance

$$\begin{aligned}
\hat{\sigma}_{\Delta}\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}\right) &= \hat{\sigma}_{\Delta}^2\left(\boldsymbol{\gamma}\right) \\
&= 4\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\kappa}_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right]^2 - 4\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right]^2 \\
&= 4\frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right]^2\right\} - 4\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right]^2.
\end{aligned}$$

3.4.C U-statistics Theory and the H-decomposition

Since $K_{\lambda,h} = K_{\lambda}^{d} \cdot K_{h}^{c}$ is still symmetric, $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ is still a U-statistic of degree 2:

$$\begin{aligned}
\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) &= \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} [\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) \\
&\quad -\varphi(\gamma_0 + X_i'\gamma_1 + Y_j'\gamma_2 + Z_i'\gamma_3)\big]\, K_{\lambda,h}(Z_i, Z_j) \\
&= \frac{1}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{(i,j)} \kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma}),
\end{aligned}$$

with

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

We still use H-decomposition to decompose $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ into three parts:

$$\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) = \Delta_{\lambda,h}(\boldsymbol{\gamma}) + 2H_{n,\lambda,h,1}(\boldsymbol{\gamma}) + R_{n,\lambda,h,1}(\boldsymbol{\gamma}), \tag{3.4.3}$$

where

$$\Delta_{\lambda,h}(\boldsymbol{\gamma}) \equiv E\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] \tag{3.4.4}$$

$$\begin{aligned}
H_{n,\lambda,h,1}(\boldsymbol{\gamma}) &\equiv \frac{1}{n} \sum_{i=1}^{n} \tilde{\kappa}_{\lambda,h,1}(W_i; \boldsymbol{\gamma}), \text{ with } \{\tilde{\kappa}_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\} \text{ IID} \\
&\quad \tilde{\kappa}_{\lambda,h,1}(W_i; \boldsymbol{\gamma}) \equiv \kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma}) - \Delta_{\lambda,h}(\boldsymbol{\gamma}) \qquad (3.4.5) \\
&\quad \kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma}) \equiv E\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma}) | W_i\right] \; i \neq j,
\end{aligned}$$

and

$$R_{n,\lambda,h,1}(\boldsymbol{\gamma}) \equiv \bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) - \Delta_{\lambda,h}(\boldsymbol{\gamma}) - 2H_{n,\lambda,h,1}(\boldsymbol{\gamma}). \tag{3.4.6}$$

The subscript 1 in the above notations denotes that the item is a projection on the first argument of $\kappa_{\lambda,h}$, W_i. $H_{n,\lambda,h,1}(\boldsymbol{\gamma})$ and $R_{n,\lambda,h,1}(\boldsymbol{\gamma})$ have zero means and are uncorrelated. I define the projection $\hat{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ as:

$$\hat{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) \equiv \Delta_{\lambda,h}(\boldsymbol{\gamma}) + 2H_{n,\lambda,h,1}(\boldsymbol{\gamma}). \tag{3.4.7}$$

The mean $\Delta_{\lambda,h}(\boldsymbol{\gamma})$ is not random (although it depends on (λ, h)), and $H_{n,\lambda,h,1}(\boldsymbol{\gamma})$ is just an average of IID random variables whose asymptotic behavior is straightforward to derive.

3.4.D The Bias Term

We first show that under the null $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ is close to zero if we choose the smoothing parameters h and λ small enough. To do that, we add a suitable assumption for λ.

Assumption 8 (Bandwidths for discrete conditioning variables) The bandwidth for the discrete kernel, $\lambda_s \equiv \lambda_{s,n}$, satisfies $\lambda_s = o_p(n^{-1/2})$ for $s = 1, 2, ..., d_d$, as $n \to \infty$.

Because we will use $\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})$ later to calculate the variance term, we first find the leading term of $\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})$:

$$\begin{aligned} & \kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma}) \\ = \; & E\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})|W_i\right] \\ \equiv \; & \kappa_1(W_i; \boldsymbol{\gamma}) + \sum_{s=1}^{d_c} \left[B_{5,s}^{c}(W_i; \boldsymbol{\gamma}) h_s^q\right] + \sum_{s=1}^{d_d} \left[B_{5.s}^{d}(W_i; \boldsymbol{\gamma}) \lambda_s\right] + s.o., \end{aligned}$$

where

$$\begin{aligned}
& \kappa_1(W_i;\boldsymbol{\gamma}) \\
\equiv\ & \frac{1}{2}\varphi^*(W_i;\boldsymbol{\gamma}) - \frac{1}{2}g_{XZ}(X_i, Z_i;\boldsymbol{\gamma}) + \frac{1}{2}g_Z(Z_i;\boldsymbol{\gamma}) - \frac{1}{2}g_{YZ}(Y_i, Z_i;\boldsymbol{\gamma}) \\
=\ & \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i) \\
& -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{YZ}(y, Z_i)dy \qquad (3.4.8) \\
& +\frac{1}{2}\int \varphi(\gamma_0 + x\gamma_1 + y\gamma_2 + Z_i\gamma_3)f_{XYZ}(x, y, Z_i)dxdy \\
& -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x, Z_i)dx \\
(under\ H_0) =\ & \frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|W_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i, Z_i\right]\ (under\ H_0) \\
& +\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Z_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Y_i, Z_i\right]\ (under\ H_0).
\end{aligned}$$

Note that

$$E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] = \Delta\left(\boldsymbol{\gamma}\right),$$

So the bias term would be $\sum_{s=1}^{d_c}\left[B_{5,s}^c(W_i;\boldsymbol{\gamma})h_s^q\right] + \sum_{s=1}^{d_d}\left[B_{5.s}^d(W_i;\boldsymbol{\gamma})\lambda_s\right] + s.o.$. The following proposition summarized the result.

Proposition 1 *Under Assumptions 1-5 and* $h_s \to 0$ *and* $\lambda_s \to 0$, $\Delta_{\lambda,h}(\boldsymbol{\gamma}) \equiv \Delta\left(\boldsymbol{\gamma}\right) + \sum_{s=1}^{d_c} B_{5,s}^c(W_i;\boldsymbol{\gamma})h_s^q + \sum_{s=1}^{d_d} B_{5.s}^d(W_i;\boldsymbol{\gamma})\lambda_s + s.o.$ *If in addition* H_0 *holds, then* $\Delta\left(\boldsymbol{\gamma}\right) = 0$. *If we furthermore assume* h_s *satisfies (A 6.2) and* λ *satisfies assumption 8, then* $\Delta_{\lambda,h}(\boldsymbol{\gamma}) = o\left(\frac{1}{\sqrt{n}}\right)$. $\Delta_{\lambda,h}(\boldsymbol{\gamma})$, $\Delta\left(\boldsymbol{\gamma}\right)$, $B_{5,s}^c(W_i;\boldsymbol{\gamma})$, *and* $B_{5.s}^d(W_i;\boldsymbol{\gamma})$ *are defined as in (3.4.4), (3.3.1), (3.7.2) and (3.7.3), respectively.*

Remark 2 *The leading term of the bias for the discrete variables is of order* $\sum_{s=1}^{d_d}\lambda_s$, *which is determined by those data which are differ from* z^d *only for one* *element AND by distance 1. When we smooth around z^d(say it's a vector with dimension d_d, for example those are discrete characteristics of the subjects), those close to z^d in the sense that only differ from z^d in one value of its variable (say only the sex is different) should carry more information about z^d. If they differ for, say, two variables, the term is of order $\lambda_s\lambda_t$ where $Z^d_{is} \neq z^d_s$ and $Z^d_{it} \neq z^d_t$, then it is of a smaller order. And if they differ for, say, one variable but by by distance 2, the term is of order λ_s^2 which is of a smaller order.*

3.4.E The Variance Term

According to Lee (1990) pp 12 Theorem 3, the variance of the U-statistic $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ is

$$\begin{aligned} VAR\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] &= \left(\begin{array}{c} n \\ 2 \end{array}\right)^{-1} \left\{2\left(n-2\right)VAR\left[\kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma})\right]\right. \\ &\quad \left. +VAR\left[\kappa_{\lambda,h}(W_i,W_j;\boldsymbol{\gamma})\right]\right\}. \end{aligned}$$

After some calculation, we get

$$\begin{aligned} & VAR\left[\kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma})\right] \\ = \;& VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + \sum_{s=1}^{d_c} 2COV\left[\kappa_1(W_i;\boldsymbol{\gamma}), B^c_{5,s}(W_i;\boldsymbol{\gamma})\right] h^q_s \\ & + \sum_{s=1}^{d_d} 2COV[\kappa_1(W_i;\boldsymbol{\gamma}), B^d_{5.s}(W_i;\boldsymbol{\gamma})]\lambda_s + s.o. \\ \equiv \;& VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + \sum_{s=1}^{d_c} C^c_{0,s}\left(\boldsymbol{\gamma}\right)h^q_s + \sum_{s=1}^{d_d} C^d_{0,s}\left(\boldsymbol{\gamma}\right)\lambda_s + s.o. \end{aligned}$$

and

$$VAR\left[\kappa_{\lambda,h}(W_i,W_j;\boldsymbol{\gamma})\right] = E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right]\prod_{s=1}^{d_c}\frac{1}{h_s} + s.o..$$

So the variance of $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ would be

$$\begin{aligned} & VAR\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] \\ = \quad & 4n^{-1}VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + 4n^{-1}\sum_{s=1}^{d_c} C_{0,s}^{c}\left(\boldsymbol{\gamma}\right)h_s^q + 4n^{-1}\sum_{s=1}^{d_d} C_{0,s}^{d}\left(\boldsymbol{\gamma}\right)\lambda_s \\ & +2n^{-2}E\left[\delta\left(W_i;\boldsymbol{\gamma}\right)\right]\prod_{s=1}^{d_c}\frac{1}{h_s} + s.o.. \qquad (3.4.9) \end{aligned}$$

Details of the calculation are in the appendix. As long as assumptions 6 and 8 are satisfied, the leading term of the variance would be

$$VAR\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] = 4n^{-1}VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] + s.o.(\text{Assumption 6 and 8}). \qquad (3.4.10)$$

3.4.F Asymptotic Distribution of the Test Statistic

$\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ is different from a conventional U-statistic since its kernel $\kappa_{\lambda,h}(W_i, W_j;\boldsymbol{\gamma})$ depends on the smoothing parameters λ and h, which are shrinking to zero as $n \to \infty$. As shown in previous chapters, we need the theory for extended U-statistics. Lemma 3.1 in Powell et al. (1989) shows that if $E\left\|\kappa_{\lambda,h}(W_i, W_j;\boldsymbol{\gamma})\right\|^2 = o(n)$,then $\sqrt{n}\left(\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \hat{\Delta}_n(\boldsymbol{\gamma})\right) = o_p(1)$. When calculating the variance term, we have already gotten that $E\left[\kappa_{\lambda,h}^2(W_i, W_j;\boldsymbol{\gamma})|W_i\right] \equiv O_p\left(\Pi_{s=1}^{d_c}h_s^{-1}\right)$. With assumption 6.1, that will make $E\left\|\kappa_{\lambda,h}(W_i, W_j;\boldsymbol{\gamma})\right\|^2$ to be $o(n)$. Intuitively, if the bandwidth for the continuous variables does not shrink too fast, the remainder term will be of smaller order. I summarize the result precisely in the following lemma:

Lemma 2 *Under Assumptions 1-5, and assumption 6.1 i.e.* $n\Pi_{s=1}^{d_c}h_s \to \infty$ *as* $n \to \infty$*, and* $h_s \to 0$ *as* $n \to \infty$ *for* $s = 1, 2, ..., d_c$*, and assumption 8 i.e.* $\lambda_s = o_p(n^{-1/2})$ *for* $s = 1, 2, ..., d_d$*, then* $\sqrt{n}\left[\bar{\Delta}_{n,\mathbf{h}}(\boldsymbol{\gamma}) - \hat{\Delta}_{n,\mathbf{h}}(\boldsymbol{\gamma})\right] = o_p(1)$.

To summarize, I have shown that under assumptions 1-6, and assumption 8 that

$$\begin{aligned}\sqrt{n}\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) &= \sqrt{n}\left[\Delta_{\lambda,h}(\boldsymbol{\gamma}) + 2H_{n,\lambda,h,1}(\boldsymbol{\gamma}) + R_{n,\lambda,h,1}(\boldsymbol{\gamma})\right] \\ &= \sqrt{n}\Delta\left(\boldsymbol{\gamma}\right) + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma}) - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\} \\ &\quad + small\ terms.\end{aligned}$$

If we further assume assumption 8 holds, the results will hold uniformly in $\boldsymbol{\gamma}$. Define the leading term to be

$$\zeta_n(\boldsymbol{\gamma}) \equiv \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left\{\kappa_1(W_i;\boldsymbol{\gamma}) - E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]\right\}, \tag{3.4.11}$$

we can show that $\zeta_n(\boldsymbol{\gamma})$ converges to a zero mean Gaussian process and thus $\sqrt{n}\left[\bar{\Delta}_{n,h}(\boldsymbol{\gamma}) - \Delta\left(\boldsymbol{\gamma}\right)\right]$ converges to the same zero mean Gaussian process. The results are given precisely in the following theorem.

Theorem 3 *Under assumptions 1-8,*

$$(a)\ \zeta_n(\boldsymbol{\gamma}) \Longrightarrow \mathcal{Z}\left(\boldsymbol{\gamma}\right)$$

$$(b)\ \sqrt{n}\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) - \Delta\left(\boldsymbol{\gamma}\right)\right] \Longrightarrow \mathcal{Z}\left(\boldsymbol{\gamma}\right),$$

where $\boldsymbol{\gamma} \in \Gamma$ *with* Γ *a compact set having a non-empty interior, and* $\mathcal{Z}$ *is a Gaussian process on* Γ *with a mean function zero and a covariance function*

$$\begin{aligned}cov\left(\mathcal{Z}(\boldsymbol{\gamma}_1), \mathcal{Z}(\boldsymbol{\gamma}_2)\right) &= 4cov\left[\ \kappa_1\left(W_i;\boldsymbol{\gamma}_1\right),\ \kappa_1\left(W_i;\boldsymbol{\gamma}_2\right)\right] \\ &= \sigma_\Delta\left(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\right),\end{aligned} \tag{3.4.12}$$

where $\zeta_n(\boldsymbol{\gamma})$ *is as defined in (3.4.11),* $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ *is as defined in (3.4.2), and* $\kappa_1\left(W_i;\boldsymbol{\gamma}\right)$ *is as defined in (3.4.8). If in addition that* H_0 *holds,*

$$T_n(\boldsymbol{\gamma}) \equiv \sqrt{n}\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma}) \Longrightarrow \mathcal{Z}\left(\boldsymbol{\gamma}\right).$$

According to corollary 2 in chapter 2, we will get

$$M_n = n \int\limits_{\Gamma} \left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right) \underset{H_0}{\Rightarrow} \int\limits_{\Gamma} \left[\mathcal{Z}(\boldsymbol{\gamma})\right]^2 d\mu\left(\boldsymbol{\gamma}\right),$$

so we can still use the conditional Monte Carlo approach as in chapter 2 to simulate the asymptotic null distribution of the modified ICM test.

As discussed in chapter 1 and 2, h, the smoothing parameter for the continuous kernel, needs to shrink to zero fast enough as the sample size increases so that the bias term associated with h will vanish asymptotically. On the other hand, h cannot shrink too fast so that the variance term won't blow up.

The rate requirement for the smoothing parameter of the discrete kernel is only that $\lambda_s = o_p(n^{-1/2})$. So λ has to shrink fast enough to kill the introduced bias asymptotically. Unlike h, the shrinking rate of λ does not have a upper bound since since it will not blow up the variance term even if we let it be zero. Note that when $\lambda = 0$, the discrete kernel reduces to an indicator function so that the frequency approach could be viewed as a special case of the smoothing approach.

Li and Racine (2003, 2004) argued that although smoothing may introduce some estimation bias, it may also reduce the finite sample variance to reduce the *finite*-sample mean squared error. They suggested a cross-validation (CV) smoothing method to select the smoothing parameters. But here we cannot apply the cross-validation method directly since we are not estimating a density or regression function.

3.5 Monte Carlo Experiments

In this section, I perform some simple Monte Carlo simulation experiments to examine the finite sample performance of the conditional independence tests for mixed discrete and continuous conditioning variables using frequency and smoothing approaches.

For all the simulations, I generate $\{(X_i, Y_i, Z_i)_{i=1}^n\}$ IID. The bandwidth for the continuous kernel I use is a value close to $\hat{h}$ as given in chapter 1 (1.3.23). If not indicated otherwise, the bandwidth for the discrete kernel, λ, is a value close to $\hat{h}^q$ in DGP2 since I want the bias associated with λ to be of the same order as the bias associated with h; and I set λ in DGP 1 the same as in DGP 2. I choose $\varphi(\cdot)$ to be the standard normal PDF, and $k^c(u)$, the kernel for the continuous variables, to be the sixth order Gaussian kernel. The number of replications is 100.

3.5.A DGP 1: A binary conditioning random variable

I first generate DGP 1 where the conditioning random variable is binary, as in the following data generating process

$$\begin{aligned} Y &= \beta X + Z^d + \epsilon_Y \\ X &= Z^d + \epsilon_X \end{aligned}$$

where

$$\begin{pmatrix} \epsilon_X \\ \epsilon_Y \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_X^2 \; 0 \\ 0 \; \sigma_Y^2 \end{pmatrix}\right) = N\left(0, \begin{pmatrix} 4 \; 0 \\ 0 \; 1 \end{pmatrix}\right),$$

and

$$Z^d = 0, \text{ with} \Pr\left(Z^d = 1\right) = 0.3,$$

$$Z^d = 1, \text{ with} \Pr\left(Z^d = 1\right) = 0.7.$$

As for the DGP 1 in chapter 1 and 2, I use

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

to indicate the strength of the dependence for X and Y, conditional on Z. Because $X|Z$ and $Y|Z$ are jointly normal, the conditional correlation represents the dependence between X and Y given Z.

I use the frequency approach to test the conditional independence and plot the power of test against ρ from -0.9 to 0.9. The size and power does not look bad when the sample size is as small as 100, and it looks pretty good when the sample size reaches 200. The "standardized" results in figure 3.1 correspond to $\tilde{M}_n$ and the "non-standardized" results in figure 3.2 correspond to M_n. Again the simulation results show that $\tilde{M}_n$ performed better than M_n in this experiment.

I also use the smoothing approach with a positive smoothing parameter for the discrete kernel. The results for $\tilde{M}_n$ and M_n are reported in figure 3.3 and figure 3.4, respectively. The size and power looks similar to the results from the frequency approach. We can only notice a very tiny improvement of the test power over some area. Figure 3.5 shows how the power function of $\tilde{M}_n$ will change when we change the choice of λ, the bandwidth for discrete variable, where the sample size is 200. Note that when $\lambda = 0$, the smoothing method becomes the frequency method. The results show that smoothing may improve the power slightly for this DGP.

3.5.B DGP 2: Mixed discrete and continuous conditioning random variables

DGP 2 generates two conditioning variables, where one is binary as in DGP 1 and the other is continuous:

$$\begin{aligned} Y &= \beta X + Z^d + Z^c + \epsilon_Y \\ X &= Z^d + \left(2 - Z^d\right) Z^c + \left(Z^c\right)^2 + \epsilon_X \end{aligned}$$

where

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

and

$$Z^d = 0, \text{ with} \Pr\left(Z^d = 1\right) = 0.3,$$

$$Z^d = 1, \text{ with} \Pr\left(Z^d = 1\right) = 0.7,$$

$$\text{Given } Z^d = 0,\ Z^{c} \tilde{} N(-2, 4),$$

$$\text{Given } Z^d = 1,\ Z^{c} \tilde{} N(2, 4).$$

This DGP has more cells than the previous one.

I first use the frequency approach to test the conditional independence and plot the power function for ρ from -0.9 to 0.9. The "standardized" results in figure 3.6 correspond to $\tilde{M}_n$ and the "non-standardized" results in figure 3.7 correspond to M_n. The power becomes good when the sample size increased to 200 and the dependence is moderate.

I then use the smoothing approach with a positive smoothing parameter for the discrete kernel. The results for $\tilde{M}_n$ and M_n are reported in figure 3.8 and figure 3.9, respectively. The size and power looks similar to the results from the frequency approach while there is very small improvement of the test power over some area. Figure 3.10 shows how the power function of $\tilde{M}_n$ will change when we change the choice of λ, where the sample size is 200. This results show that when we select the correct bandwidth for the discrete variable (in this case 0.04), smoothing may significantly improve the power. But when we choose a too big discrete bandwidth, the size of the test could be way off.

3.6 Concluding Remarks

In this chapter I extend the nonparametric test for conditional independence to incorporate the case where the conditioning random variables are mixed discrete and continuous. The frequency approach divides the sample into a collection of cells according to the value of the discrete conditioning variables, and constructs the test statistic based on a linear combination of estimators for all cells. The frequency approach is easy to calculate and performs well in the simulations. Another method would be to use kernel smoothing for both continuous and discrete variables. This can incorporate the frequency method as a special case. The two tests have the same asymptotic distributions as long as the bandwidth of the discrete kernel shrinks to zero faster enough. The smoothing approach may increase the power for finite samples especially when the number of cells is relatively large. Nevertheless the choice of the bandwidth of the discrete kernel is challenging.

3.7 Appendix: Proofs

Proof of Proposition 1:

$$
\begin{aligned}
& \Delta_{\lambda,h}(\boldsymbol{\gamma}) \\
= \; & E\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right] \\
= \; & E\left\{E\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})|W_i\right]\right\} \\
\equiv \; & E\left\{\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right\} \\
= \; & E\left\{\kappa_1(W_i; \boldsymbol{\gamma}) + \sum_{s=1}^{d_c} B_{5,s}^c(W_i; \boldsymbol{\gamma}) h_s^q + \sum_{s=1}^{d_d} B_{5.s}^d(W_i; \boldsymbol{\gamma}) \lambda_s + s.o\right\} \\
= \; & \Delta\left(\boldsymbol{\gamma}\right) + \sum_{s=1}^{d_c} E\left[B_{5,s}^c(W_i; \boldsymbol{\gamma})\right] h_s^q + \sum_{s=1}^{d_d} E\left[B_{5.s}^d(W_i; \boldsymbol{\gamma})\right] \lambda_s + s.o
\end{aligned}
$$

To calculate the last second step in above derivation, we notice that

$$
\begin{aligned}
& \kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma}) \\
= \; & E\left[\kappa_{\lambda,h}(W_i,W_j;\boldsymbol{\gamma})|W_i\right] \\
= \; & \frac{1}{2}E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_i'\gamma_2+Z_i'\gamma_3)K_{\lambda,h}(Z_i,Z_j)|W_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0+X_i'\gamma_1+Y_j'\gamma_2+Z_i'\gamma_3)K_{\lambda,h}(Z_i,Z_j)|W_i\right] \\
& +\frac{1}{2}E\left[\varphi(\gamma_0+X_j'\gamma_1+Y_j'\gamma_2+Z_j'\gamma_3)K_{\lambda,h}(Z_j,Z_i)|W_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0+X_j'\gamma_1+Y_i'\gamma_2+Z_j'\gamma_3)K_{\lambda,h}(Z_j,Z_i)|W_i\right] \\
= \; & \frac{1}{2}\left\{\varphi^*(W_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{1,s}^c(W_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{1.s}^d(W_i;\boldsymbol{\gamma})\lambda_s\right]\right\} \\
& -\frac{1}{2}\left\{g_{XZ}(X_i,Z_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{2,s}^c(X_i,Z_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{2.s}^d(X_i,Z_i;\boldsymbol{\gamma})\lambda_s\right]\right\} \\
& +\frac{1}{2}\left\{g_Z(Z_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{3,s}^c(Z_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{3.s}^d(Z_i;\boldsymbol{\gamma})\lambda_s\right]\right\} \\
& -\frac{1}{2}\left\{g_{YZ}(Y_i,Z_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{4,s}^c(Y_i,Z_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{4.s}^d(Y_i,Z_i;\boldsymbol{\gamma})\lambda_s\right]\right\}+s.o \\
\equiv \; & \kappa_1(W_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{5,s}^c(W_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{5.s}^d(W_i;\boldsymbol{\gamma})\lambda_s\right]+s.o
\end{aligned}
$$

$$
\begin{aligned}
& \kappa_1(W_i;\boldsymbol{\gamma}) \\
\equiv\ & \frac{1}{2}\varphi^*(W_i;\boldsymbol{\gamma}) - \frac{1}{2}g_{XZ}(X_i,Z_i;\boldsymbol{\gamma}) + \frac{1}{2}g_Z(Z_i;\boldsymbol{\gamma}) - \frac{1}{2}g_{YZ}(Y_i,Z_i;\boldsymbol{\gamma}) \\
=\ & \frac{1}{2}\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i) \\
& -\frac{1}{2}\int \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{YZ}(y,Z_i)dy \qquad (3.7.1) \\
& +\frac{1}{2}\int \varphi(\gamma_0 + x\gamma_1 + y\gamma_2 + Z_i\gamma_3)f_{XYZ}(x,y,Z_i)dxdy \\
& -\frac{1}{2}\int \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x,Z_i)dx \\
(under\ H_0)\quad =\ & \frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|W_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|X_i,Z_i\right]\ (under\ H_0) \\
& +\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Z_i\right] \\
& -\frac{1}{2}E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Y_i,Z_i\right]\ (under\ H_0)
\end{aligned}
$$

where

$$
\begin{aligned}
g_{XZ}(x,z;\boldsymbol{\gamma}) \ &\equiv\ \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + z'\gamma_3)f_{YZ}(y,z)dy \\
&=\ E\left[\varphi(\gamma_0 + x'\gamma_1 + Y'\gamma_2 + z'\gamma_3)f_Z(z)|Z=z\right] \\
(under\ H_0)\ &=\ E\left[\varphi(\gamma_0 + X'\gamma_1 + Y'\gamma_2 + Z'\gamma_3)f_Z(Z)|X=x,Z=z\right],
\end{aligned}
$$

$$
\begin{aligned}
g_Z(Z_i;\boldsymbol{\gamma}) \ &\equiv\ \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)f_{XYZ}(x,y,Z_i)dxdy \\
&=\ E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Z_i\right],
\end{aligned}
$$

$$
\begin{aligned}
g_{YZ}(Y_i,Z_i;\boldsymbol{\gamma}) \ &\equiv\ \int_{x_L}^{x_U} \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_{XZ}(x,Z_i)dx \\
(under\ H_0)\ &=\ E\left[\varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)f_Z(Z_i)|Y_i,Z_i\right],
\end{aligned}
$$

$$
B_{1,s}^c(W_i;\boldsymbol{\gamma}) \equiv \varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)\left[\frac{\partial^q f_Z\left(Z_i^c, Z_i^d\right)/\partial\left(Z_{is}^c\right)^q}{q!}\right]\int u^q k(u)du,
$$

$$
\begin{aligned}
B_{1,s}^d(W_i;\boldsymbol{\gamma}) \quad &\equiv \quad \varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)(\sum_{z^d}\left[1_s\left(Z_i^d, z^d\right) f_Z(Z_i^c, z^d)\right]) \\
&\quad \text{without natural ordering} \\
or \quad &\quad \varphi(\gamma_0 + X_i'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3)(\sum_{z^d}\left[1_s\left(\left|Z_i^d - z^d\right| = 1\right) f_Z(Z_i^c, z^d)\right]) \\
&\quad \text{with natural ordering,}
\end{aligned}
$$

with

$$
\begin{aligned}
1_s\left(Z_i^d, z^d\right) \quad &\equiv \quad 1\left(Z_{is}^d \neq z_s^d\right)\prod_{s'\neq s} 1\left(Z_{is}^d = z_s^d\right) \\
&\quad \text{i.e. } Z_i^d \text{ and } z^d \text{ only differ for the } s\text{th element,}
\end{aligned}
$$

$$
\begin{aligned}
1_s\left(\left|Z_i^d - z^d\right| = 1\right) \quad &\equiv \quad 1\left(\left|Z_i^d - z^d\right| = 1\right) 1\left(Z_{is}^d \neq z_s^d\right)\prod_{s'\neq s} 1\left(Z_{is}^d = z_s^d\right) \\
&\quad \text{only the } s\text{th variable is different by distance 1,}
\end{aligned}
$$

$$
\begin{aligned}
B_{2,s}^c(X_i, Z_i;\boldsymbol{\gamma}) \quad &\equiv \quad \int u^q k\left(u\right) du \\
&\quad \times \int_{y_L}^{y_U} \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3)\frac{\partial^q f_{YZ}(y, Z_i)/\partial\left(Z_{is}^c\right)^q}{q!}dy,
\end{aligned}
$$

$$
\begin{aligned}
B_{2,s}^d(X_i, Z_i;\boldsymbol{\gamma}) \quad &\equiv \quad \sum_{z^d}\left[1_s\left(Z_i^d, z^d\right)\right. \\
&\quad \left.\times \int_{y_L}^{y_U} \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3) f_{YZ}(y, Z_i^c, z^d) dy\right] \\
&\quad \text{without natural ordering} \\
or \quad &= \quad \sum_{z^d}\left[1_s\left(\left|Z_i^d - z^d\right| = 1\right)\right. \\
&\quad \left.\times \int_{y_L}^{y_U} \varphi(\gamma_0 + X_i'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3) f_{YZ}(y, Z_i^c, z^d) dy\right] \\
&\quad \text{with natural ordering,}
\end{aligned}
$$

$$
\begin{aligned}
&B_{3,s}^c(Z_i;\boldsymbol{\gamma}) \\
\equiv \quad &\int u^q k\left(u\right) du \\
&\times \int_{y_L}^{y_U}\int_{x_L}^{x_U}\left[\frac{\partial^q \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i'\gamma_3) f_{XYZ}\left(x, y, Z_i\right)/\partial\left(Z_{is}^c\right)^q}{q!}\right] dxdy,
\end{aligned}
$$

$$\begin{aligned}
B^d_{3,s}(Z_i;\boldsymbol{\gamma}) \quad &\equiv \quad \sum_{z^d}[1_s\left(Z_i^d, z^d\right) \\
&\quad \times \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i^{c\prime}\gamma_3^c + z^{d\prime}\gamma_3^d) f_{XYZ}(x, y, Z_i^c, z^d)dxdy] \\
&\quad \text{without natural ordering} \\
or \quad &= \quad \sum_{z^d}[1_s\left(\left|Z_i^d - z^d\right| = 1\right) \\
&\quad \times \int \varphi(\gamma_0 + x'\gamma_1 + y'\gamma_2 + Z_i^{c\prime}\gamma_3^c + z^{d\prime}\gamma_3^d) f_{XYZ}(x, y, Z_i^c, z^d)dxdy] \\
&\quad \text{with natural ordering,}
\end{aligned}$$

$$\begin{aligned}
B^c_{4,s}(Y_i, Z_i;\boldsymbol{\gamma}) \quad &\equiv \quad \int u^q k\left(u\right) du \\
&\quad \times \int_{x_L}^{x_U} \left[\frac{\partial^q \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i'\gamma_3) f_{XZ}(x, Z_i)/\partial\left(Z_{is}^c\right)^q}{q!}\right] dx,
\end{aligned}$$

$$\begin{aligned}
B^d_{4,s}(Y_i, Z_i;\boldsymbol{\gamma}) \quad &\equiv \quad \sum_{z^d}[1_s\left(Z_i^d, z^d\right) \\
&\quad \times \int_{x_L}^{x_U} \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i^{c\prime}\gamma_3^c + z^{d\prime}\gamma_3^d) f_{XZ}(x, Z_i^c, z^d)dx] \\
&\quad \text{without natural ordering} \\
or \quad &= \quad \sum_{z^d}[1_s\left(\left|Z_i^d - z^d\right| = 1\right) \\
&\quad \times \int_{x_L}^{x_U} \varphi(\gamma_0 + x'\gamma_1 + Y_i'\gamma_2 + Z_i^{c\prime}\gamma_3^c + z^{d\prime}\gamma_3^d) f_{XZ}(x, Z_i^c, z^d)dx] \\
&\quad \text{with natural ordering,}
\end{aligned}$$

$$B^c_{5,s}(W_i;\boldsymbol{\gamma}) \equiv \frac{1}{2}\left[B^c_{1,s}(W_i;\boldsymbol{\gamma}) - B^c_{2,s}(X_i, Z_i;\boldsymbol{\gamma}) + B^c_{3,s}(Z_i;\boldsymbol{\gamma}) - B^c_{4,s}(Y_i, Z_i;\boldsymbol{\gamma})\right], \tag{3.7.2}$$

and

$$B^d_{5.s}(W_i;\boldsymbol{\gamma}) \equiv \frac{1}{2}\left[B^d_{1,s}(W_i;\boldsymbol{\gamma}) - B^d_{2,s}(X_i, Z_i;\boldsymbol{\gamma}) + B^d_{3,s}(Z_i;\boldsymbol{\gamma}) - B^d_{4,s}(Y_i, Z_i;\boldsymbol{\gamma})\right]. \tag{3.7.3}$$

Note that

$$
\begin{aligned}
& E\left[\kappa_1(W_i;\boldsymbol{\gamma})\right] \\
\equiv\ & \frac{1}{2}E\left[\varphi^*(W_i;\boldsymbol{\gamma})\right]-\frac{1}{2}E\left[g_{XZ}(X_i,Z_i;\boldsymbol{\gamma})\right]+\frac{1}{2}E\left[g_Z(Z_i;\boldsymbol{\gamma})\right]-\frac{1}{2}E\left[g_{YZ}(Y_i,Z_i;\boldsymbol{\gamma})\right] \\
=\ & \frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_Z(z)f_{XYZ}(x,y,z)dxdydz \\
& -\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_{YZ}(y,z)f_{XZ}(x,z)dydxdz \qquad (3.7.4) \\
& +\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_{XYZ}(x,y,z)f_Z(z)dxdydz \\
& -\frac{1}{2}\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_{XZ}(x,z)f_{YZ}(y,z)dxdydz \\
=\ & \int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_Z(z)f_{XYZ}(x,y,z)dxdydz \\
& -\int\varphi(\gamma_0+x'\gamma_1+y'\gamma_2+z'\gamma_3)f_{XZ}(x,z)f_{YZ}(y,z)dxdydz \\
=\ & E_P\left[\varphi^*(\gamma_0+X'\gamma_1+Y'\gamma_2+Z'\gamma_3)\right]-E_Q\left[\varphi^*(\gamma_0+X'\gamma_1+Y'\gamma_2+Z'\gamma_3)\right] \\
=\ & \Delta\left(\boldsymbol{\gamma}\right),
\end{aligned}
$$

where we use

$$E\left[\varphi^*(W_i;\boldsymbol{\gamma})\right]=E\left[g_Z(Z_i;\boldsymbol{\gamma})\right]$$

and

$$E\left[g_{XZ}(X_i,Z_i;\boldsymbol{\gamma})\right]=E\left[g_{YZ}(Y_i,Z_i;\boldsymbol{\gamma})\right].$$

■

Proof of (3.4.9): According to Lee (1990) pp 12 Theorem 3, the variance of the U-statistic $\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})$ is

$$
\begin{aligned}
VAR\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] &= \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1}\{2\left(n-2\right)VAR\left[\kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma})\right] \\
&\quad +VAR\left[\kappa_{\lambda,h}(W_i,W_j;\boldsymbol{\gamma})\right]\},
\end{aligned}
$$

where

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

and

$$\begin{aligned}&\kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma})\\ =\ & E\left[\kappa_{\lambda,h}(W_i,W_j;\boldsymbol{\gamma})|W_i\right]\\ =\ & \kappa_1(W_i;\boldsymbol{\gamma})+\sum_{s=1}^{d_c}\left[B_{5,s}^c(W_i;\boldsymbol{\gamma})h_s^q\right]+\sum_{s=1}^{d_d}\left[B_{5.s}^d(W_i;\boldsymbol{\gamma})\lambda_s\right]+s.o..\end{aligned}$$

$$\begin{aligned}&VAR\left[\kappa_{\lambda,h,1}(W_i;\boldsymbol{\gamma})\right]\\ =\ & VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]+\sum_{s=1}^{d_c}2COV\left[\kappa_1(W_i;\boldsymbol{\gamma}),B_{5,s}^c(W_i;\boldsymbol{\gamma})\right]h_s^q\\ &+\sum_{s=1}^{d_d}2COV[\kappa_1(W_i;\boldsymbol{\gamma}),B_{5.s}^d(W_i;\boldsymbol{\gamma})]\lambda_s+s.o.\\ \equiv\ & VAR\left[\kappa_1(W_i;\boldsymbol{\gamma})\right]+\sum_{s=1}^{d_c}C_{0,s}^c\left(\boldsymbol{\gamma}\right)h_s^q+\sum_{s=1}^{d_d}C_{0,s}^d\left(\boldsymbol{\gamma}\right)\lambda_s+s.o.,\end{aligned}$$

where

$$C_{0,s}^c\left(\boldsymbol{\gamma}\right)\equiv 2COV\left[\kappa_1(W_i;\boldsymbol{\gamma}),B_{5,s}^c(W_i;\boldsymbol{\gamma})\right],$$

and

$$C_{0,s}^d\left(\boldsymbol{\gamma}\right)\equiv 2COV[\kappa_1(W_i;\boldsymbol{\gamma}),B_{5.s}^d(W_i;\boldsymbol{\gamma})].$$

$$E\left[\kappa_{\lambda,h}^2(W_i,W_j;\boldsymbol{\gamma})|W_i\right]\equiv\delta\left(W_i;\boldsymbol{\gamma}\right)\prod_{s=1}^{d_c}\frac{1}{h_s}+s.o.$$

$$
\begin{aligned}
& VAR\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right] \\
= \; & E\left[\kappa^2_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right] - \Delta_{\lambda,h}(\boldsymbol{\gamma})^2 \\
= \; & E\left[\delta\left(W_i; \boldsymbol{\gamma}\right)\right] \prod_{s=1}^{d_c} \frac{1}{h_s} - \Delta_{\lambda,h}(\boldsymbol{\gamma})^2 + s.o. \\
= \; & E\left[\delta\left(W_i; \boldsymbol{\gamma}\right)\right] \prod_{s=1}^{d_c} \frac{1}{h_s} + \Delta^2(\boldsymbol{\gamma}) + s.o. \\
= \; & E\left[\delta\left(W_i; \boldsymbol{\gamma}\right)\right] \prod_{s=1}^{d_c} \frac{1}{h_s} + s.o. \text{ if } h_s = o\left(1\right).
\end{aligned}
$$

$$
\begin{aligned}
VAR\left[\bar{\Delta}_{n,\lambda,h}(\boldsymbol{\gamma})\right] = \; & \frac{2}{n(n-1)} \left\{2\left(n-2\right) VAR\left[\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right]\right. \\
& \left. + VAR\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right]\right\} \\
= \; & 4n^{-1} VAR\left[\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right] - \frac{4}{n\left(n-1\right)} VAR\left[\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right] \\
& + 2n^{-2} VAR\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right] + s.o. \\
= \; & 4n^{-1} VAR\left[\kappa_{\lambda,h,1}(W_i; \boldsymbol{\gamma})\right] \\
& + 2n^{-2} VAR\left[\kappa_{\lambda,h}(W_i, W_j; \boldsymbol{\gamma})\right] + s.o. \\
= \; & 4n^{-1} VAR\left[\kappa_1(W_i; \boldsymbol{\gamma})\right] + 4n^{-1} \sum_{s=1}^{d_c} C^c_{0,s}\left(\boldsymbol{\gamma}\right) h^q_s \\
& + 4n^{-1} \sum_{s=1}^{d_d} C^d_{0,s}\left(\boldsymbol{\gamma}\right) \lambda_s + s.o. \\
& + 2n^{-2} E\left[\delta\left(W_i; \boldsymbol{\gamma}\right)\right] \prod_{s=1}^{d_c} \frac{1}{h_s} + s.o.
\end{aligned}
$$

■

Proof of Theorem 3: Replacing the kernel K_h in chapter 2 by $K_{\lambda,h}$ defined in (3.4.1), and using the results in the proof of proposition 1 and the proof of (3.4.9), we find that the proof of theorem 1 in chapter 2 still goes through. ■

3.8 Figures

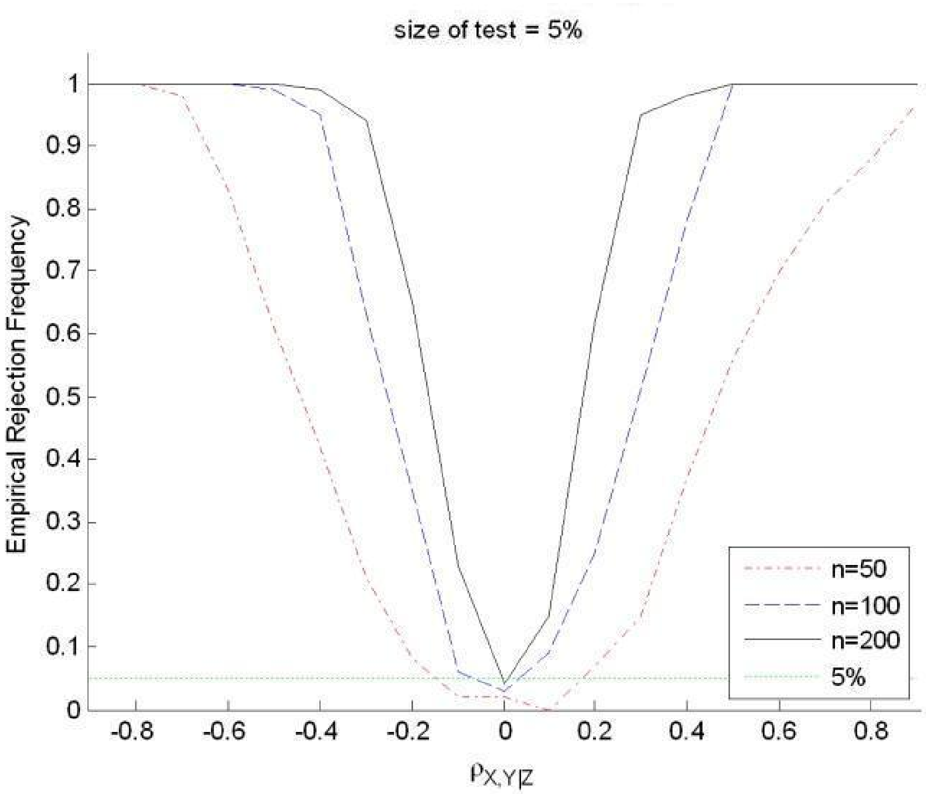

Figure 3.1: Power functions of $\tilde{M}_n$ for DGP 1, frequency approach

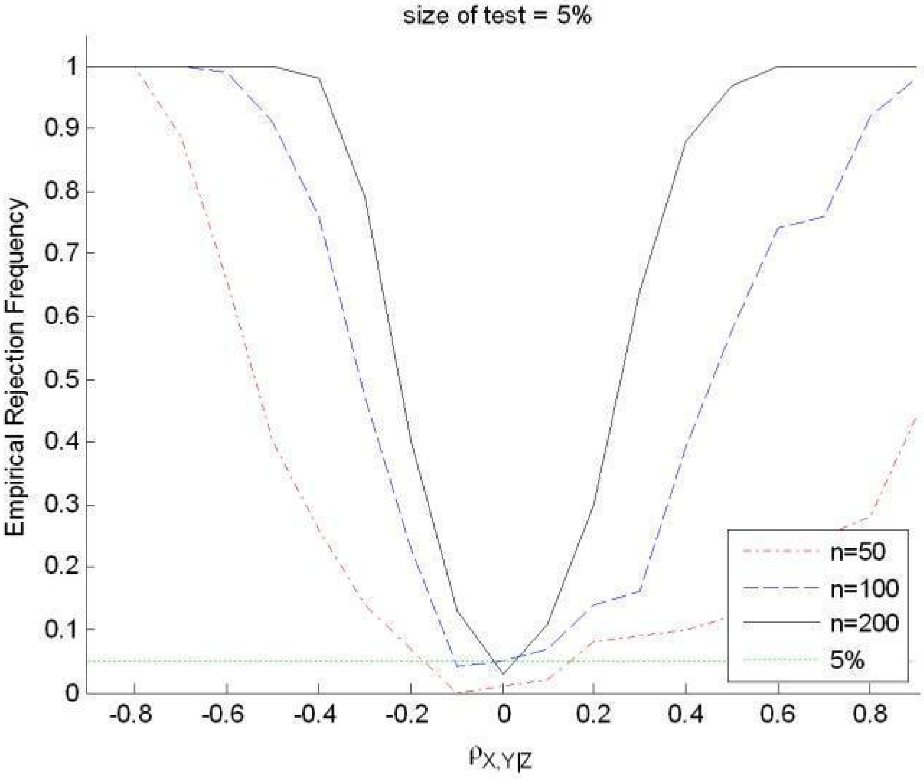

Figure 3.2: Power functions of M_n for DGP 1, frequency approach

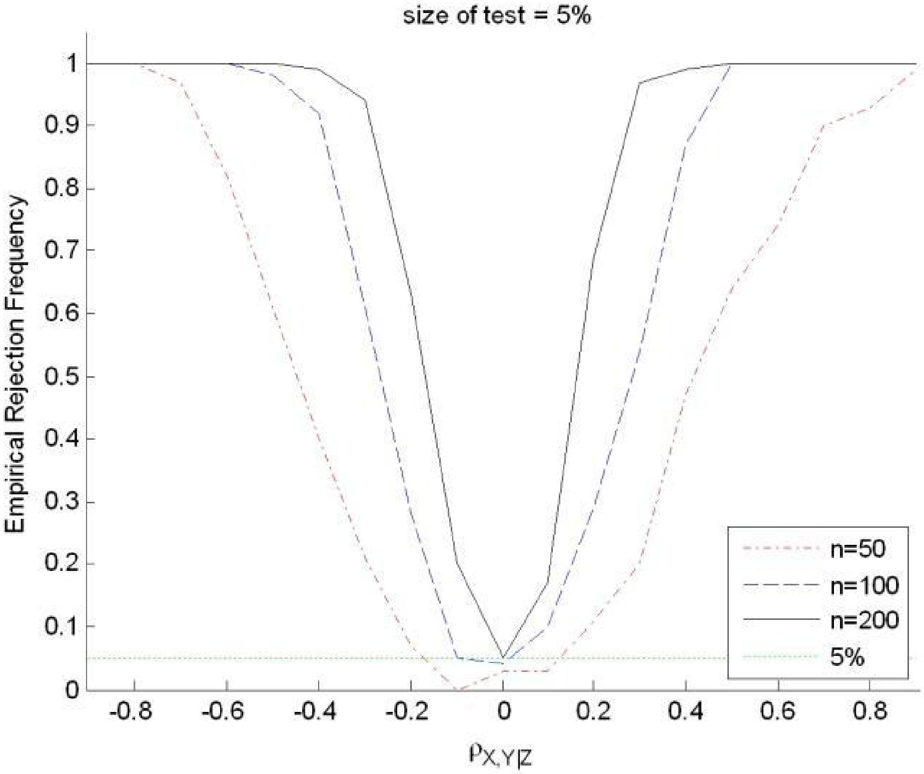

Figure 3.3: Power functions of $\tilde{M}_n$ for DGP 1, smoothing approach

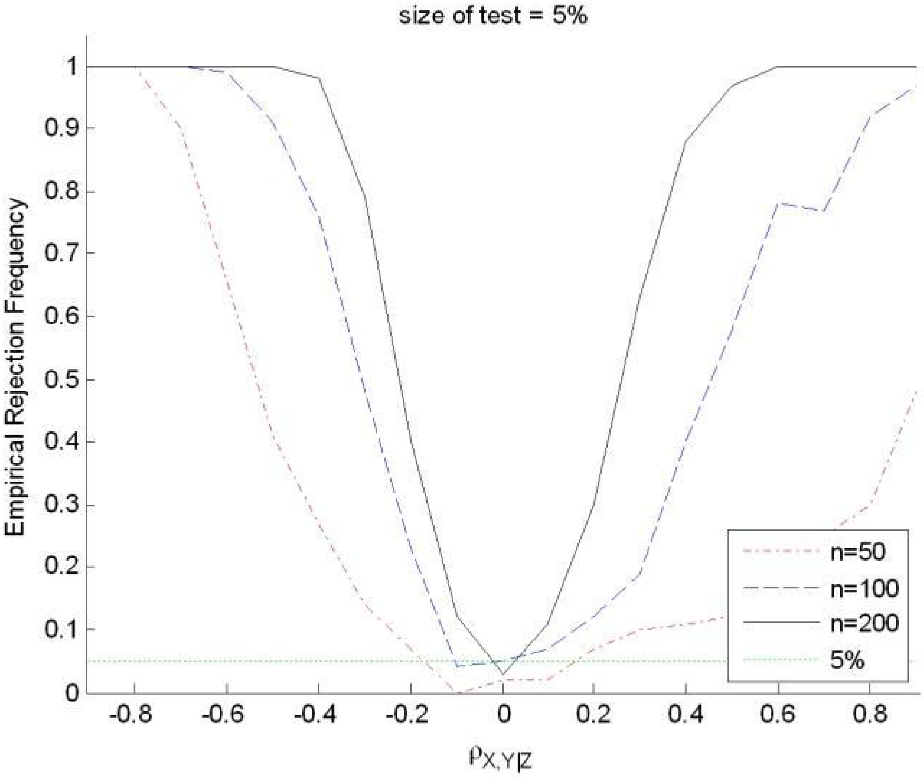

Figure 3.4: Power functions of M_n for DGP 1, smoothing approach

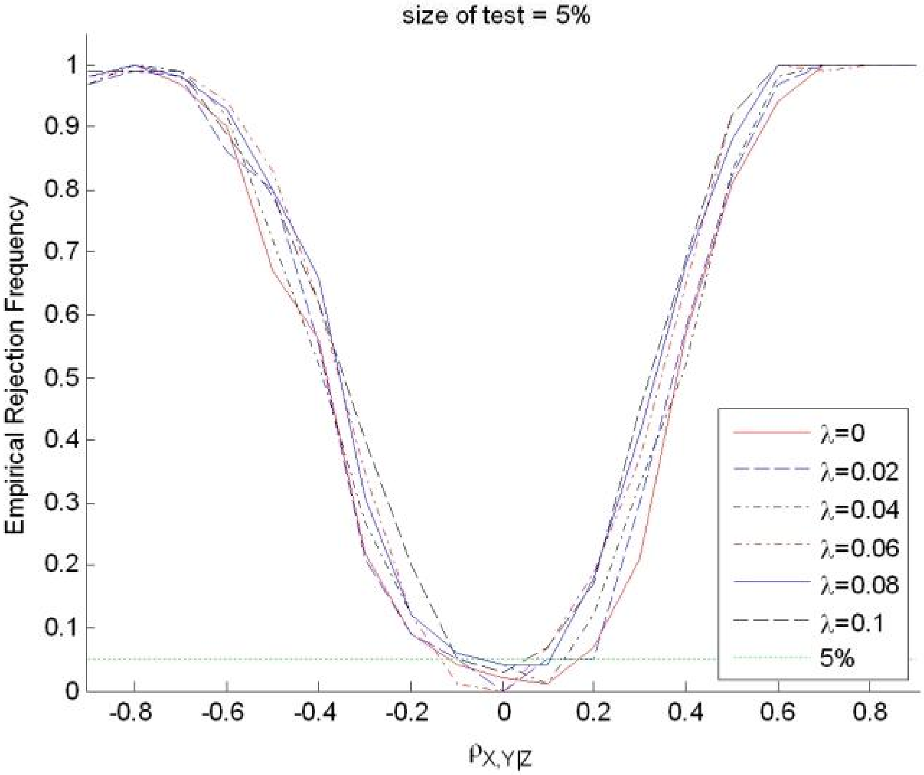

Figure 3.5: Power functions of $\tilde{M}_n$ for DGP 1, different bandwidths for the discrete variable

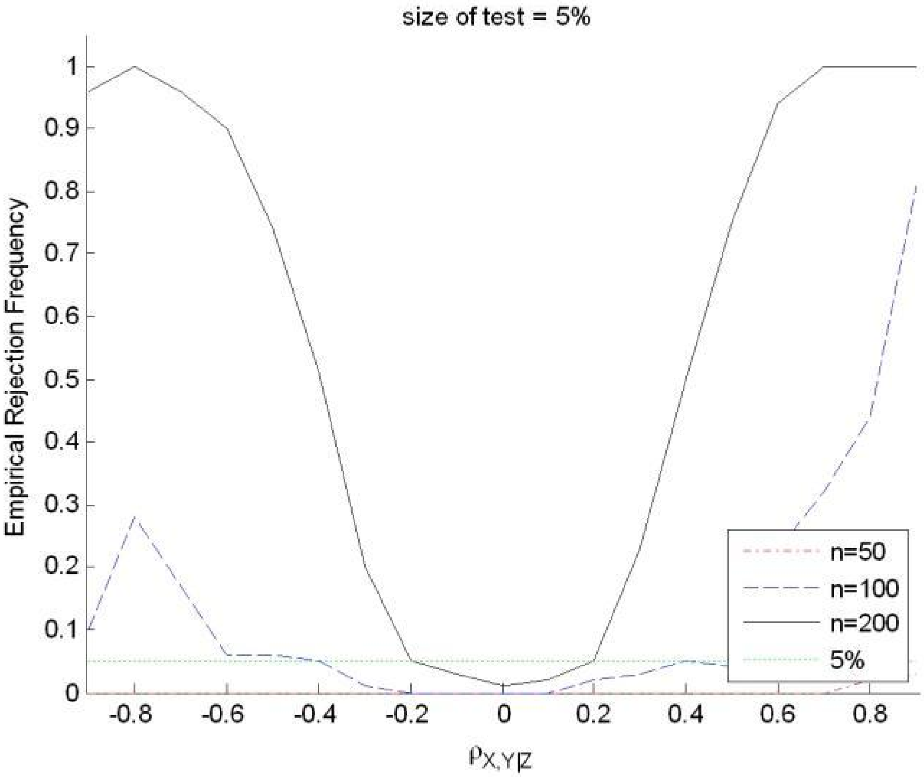

Figure 3.6: Power functions of $\tilde{M}_n$ for DGP 2, frequency approach

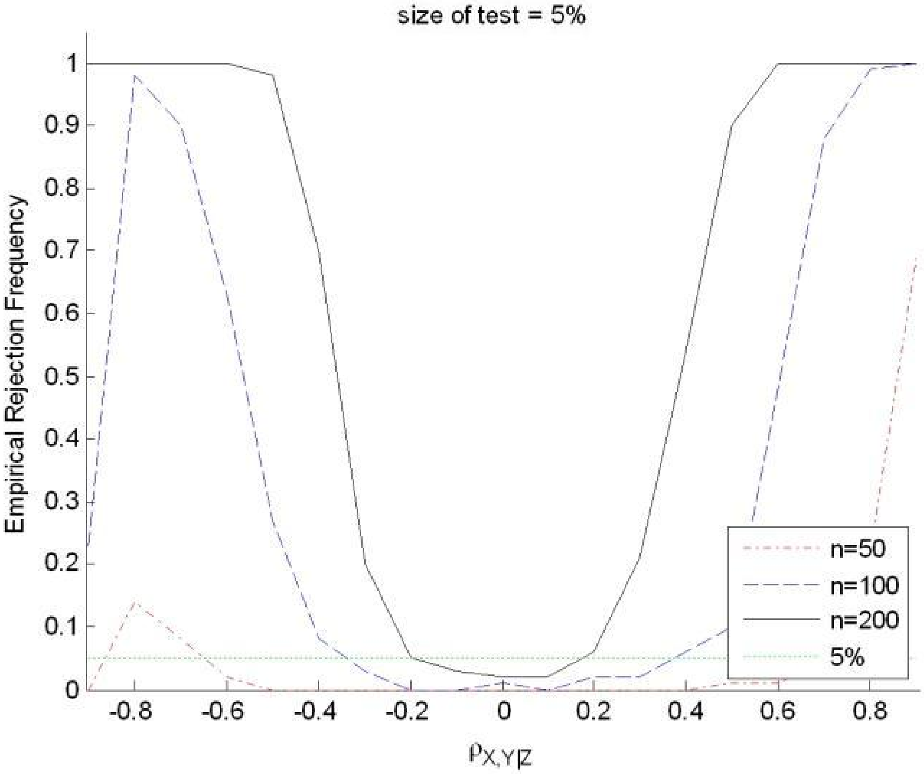

Figure 3.7: Power functions of M_n for DGP 2, frequency approach

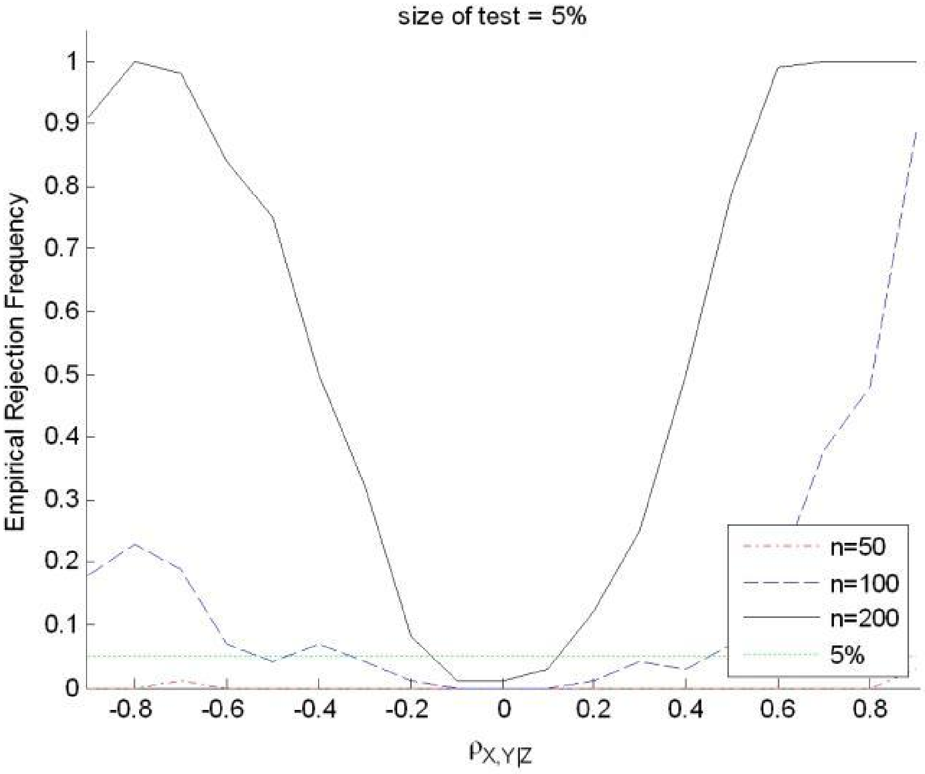

Figure 3.8: Power functions of $\tilde{M}_n$ for DGP 2, smoothing approach

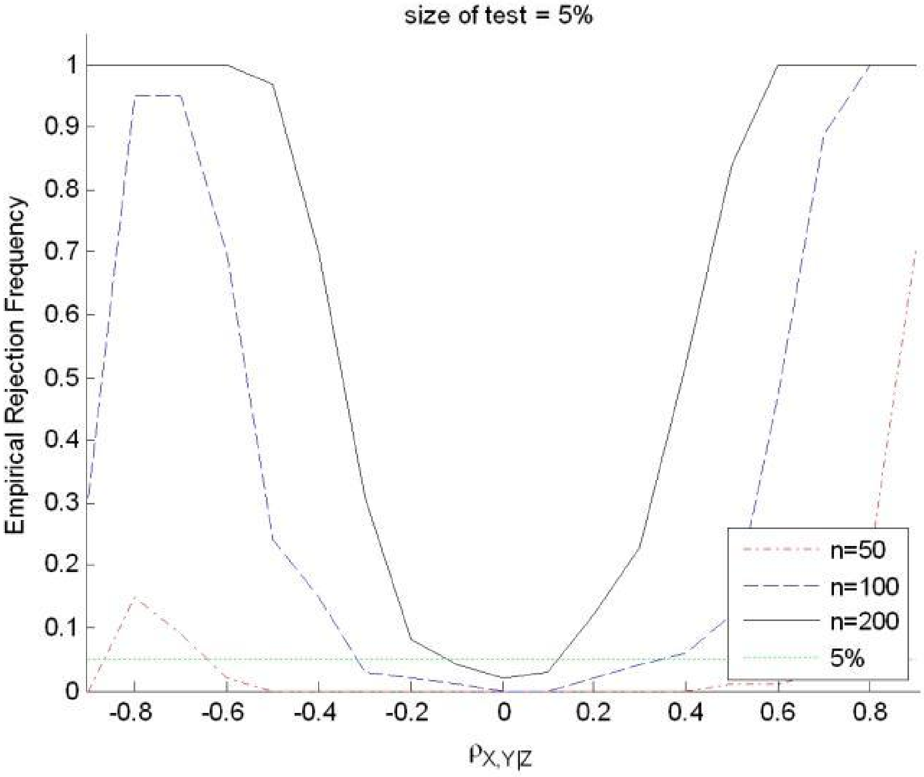

Figure 3.9: Power functions of M_n for DGP 2, smoothing approach

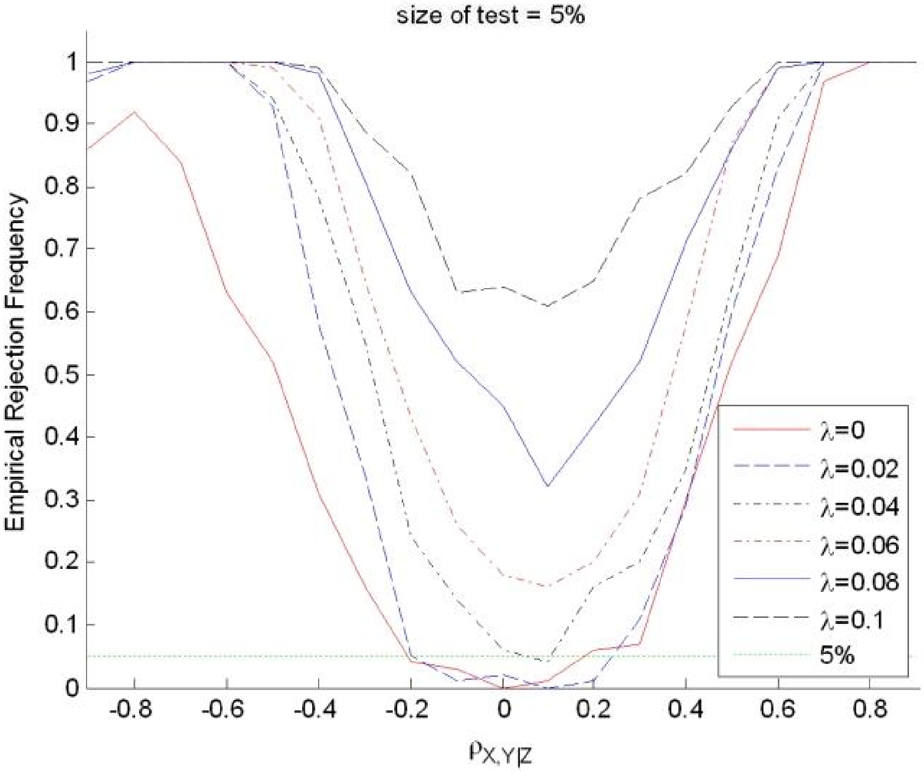

Figure 3.10: Power functions of $\tilde{M}_n$ for DGP 2, different bandwidths for the discrete variable

References

Aitchison, J., and Aitken, C. G. G., 1976: Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Andrews, D. W. K., 1994: Empirical process methods in econometrics. In *Handbook of Econometrics*, editors R. F. Engle, and D. L. McFadden, 2248–2296. Elsevier Science.

Barnow, B. S., Cain, G. G., and Goldberger, A. S., 1981: Selection on observables. *Evaluation Studies Review Annual*, **5**, 43–59.

Bierens, H. J., 1982: Consistent Model Specification Tests. *Journal of Econometrics*, **20**, 105–134.

Bierens, H. J., 1990: A Consistent Conditional Moment Test of Functional Form. *Econometrica*, **58**, 1443–1458.

Bierens, H. J., and Ploberger, W., 1997: Asymptotic Theory of Integrated Conditional Moment Tests. *Econometrica*, **65**(5), 1129–1151.

Billingsley, P., 1999: *Convergence of Probability Measures, second edition.* John Wiley and Sons, Inc.

Blackburn, M. L., and Neumark, D., 1993: Omitted-Ability Bias and the Increase in the. Return to Schooling. *Journal of Labor Economics*, 521–544.

Chung, K. L., 1974: *A course in probability theory.* Academic Press, New York.

Davidson, J., 1994: *Stochastic Limit Theory.* Oxford University Press.

Dawid, A. P., 1979: Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, **41**(1), 1–31.

de Jong, R. M., and Bierens, H. J., 1994: On the Limit Behavior of a Chi-Square Type Test If the Number of Conditional Moments Tested Approaches Infinity. *Econometric Theory*, **10**(1), 70–90.

Delgado, M., and Gonzalez-Manteiga, W., 2001: Significance Testing in Nonparametric Regression Based on the Bootstrap. *Annals of Statistics*, **29**, 1469–1507.

Dunford, N., and Schwartz, J. T., 1958: *Linear Operators*. Wiley, New York.

Fernandez, M., and Flores, R., 2002: Tests for Conditional Independence, Markovian Dynamics and Noncausality. European University Institute Discussion Paper.

Griliches, Z., 1977: Estimating the returns to schooling: some econometric problems. *Econometrica*, **45**(1), 1–22.

Griliches, Z., and M., M. W., 1972: Education, Income, and Ability. *Journal of Political Economy*, **80**(3), S74–S103.

Hansen, B. E., 1996: Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, **64**(2), 413–430.

Hansen, B. E., 2006: Interval forecasts and parameter uncertainty. *Journal of Econometrics*, **135**, 377–398.

Hoeffding, W., 1948: A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, **19**, 293–325.

Hsiao, C., Li, Q., and Racine, J., 2007: A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, **140**, 802–826.

Lee, A. J., 1990: *U-statistics: theory and practice*. CRC Press.

Li, Q., and Racine, J., 2003: Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, **86**, 266–292.

Li, T., Perrigne, I., and Vuong, Q., 2000: Conditionally independent private information in OCS wildcat auctions. *Journal of Econometrics*, **98**, 129–161.

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

Mincer, J., 1974: *Schooling, Experience, and Earnings*. Columbia University Press.

Powell, J. L., Stock, J. H., and Stoker, T. M., 1989: Semiparametric Estimation of Index Coefficients. *Econometrica*, **57**(6), 1403–1430.

Powell, J. L., and Stoker, T. M., 1996: Optimal Bandwidth Choice for Density-weighted Averages. *Journal of Econometrics*, **75**, 291–316.

Racine, J., and Li, Q., 2004: Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, **119**, 99–130.

Ruppert, D., and Wand, M. P., 1994: Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**(3), 1346–1370.

Stinchcombe, M., and White, H., 1998: Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative. *Econometric Theory*, **14**, 295–324.

Su, L. J., and White, H., 2003: Testing Conditional Independence Via Empirical Likelihood. UCSD Department of Economics Discussion Paper.

Su, L. J., and White, H., 2007: A Consistent Characteristic Function-Based Test for Conditional Independence. *Journal of Econometrics*, **141**, 807–834.

Su, L. J., and White, H., 2008: A Nonparametric Hellinger Metric Test for Conditional Independence. *Econometric Theory*, **24**, 829–864.

White, H., and Chalak, K., 2006: Parametric and Nonparametric Estimation of Covariate-Conditioned Average Causal Effects. Working paper.