Econometric Reviews, 0(0):1–30, 2013

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2013.825135

Taylor & Francis
Taylor & Francis Group

TESTING CONDITIONAL INDEPENDENCE RESTRICTIONS

Oliver Linton[1] **and Pedro Gozalo**[2]

[1]*Faculty of Economics, Cambridge University, Cambridge, UK*
[2]*Department of Community Health, Brown University, Providence, Rhode Island, USA*

□ *We propose a nonparametric test of the hypothesis of conditional independence between variables of interest based on a generalization of the empirical distribution function. This hypothesis is of interest both for model specification purposes, parametric and semiparametric, and for nonmodel-based testing of economic hypotheses. We allow for both discrete variables and estimated parameters. The asymptotic null distribution of the test statistic is a functional of a Gaussian process. A bootstrap procedure is proposed for calculating the critical values. Our test has power against alternatives at distance $n^{-1/2}$ from the null; this result holding independently of dimension. Monte Carlo simulations provide evidence on size and power.*

Keywords Conditional independence; Empirical distribution; Independence; Nonparametric; Smooth bootstrap; Test.

JEL Classification C12; C14; C15; C52.

1. INTRODUCTION

Les Godfrey has made some important contributions to hypothesis testing in econometrics. He wrote a series of papers tackling central issues in misspecification testing, culminating in his Econometric Society monograph, Godfrey (1988). This paper is about testing for conditional independence and related hypotheses.

We investigate the application of the hypothesis of conditional independence in econometrics. Let Y, X, and Z be random variables; following Dawid (1979), we write

$$Y \perp\!\!\!\perp X \mid Z \tag{1}$$

to denote that Y is independent of X given Z. This assumption is related to the more commonly treated hypothesis that $Y \perp\!\!\!\perp X$ (Y is

Address correspondence to Oliver Linton, Faculty of Economics, Cambridge University, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, UK; E-mail: obl20@cam.ac.uk

independent of X), in that it imposes an infinite number of restrictions on the joint distribution.[1] These assumptions are stronger than the mean independence conditions usually employed in regression analysis: for example, that Y is mean independent of X, i.e., that $E(Y \mid X) = 0$, or that Y is mean independent of X given Z, i.e., that $E(Y \mid X, Z) = E(Y \mid Z)$. We now give two concrete reasons for interest in the conditional independence hypothesis.

A large literature now exists on testing parametric regression models against general alternatives, see for example Bierens and Ploberger (1997), Hong and White (1995), and Fan and Li (1996). These amount to testing a null hypothesis of mean independence of the regressors from some parametrically defined residual. Andrews (1997) extends this to testing the null hypothesis of a parametric conditional distribution against a general nonparametric alternative. There are many nonparametric tests of independence for continuous data, starting with Hoeffding (1948), including those based on empirical distribution functions such as Blum et al. (1961) and Skaug and Tjøstheim (1993) and Delgado (1996), and those based on smoothing methods like Robinson (1991) and Zheng (1997). However, there do not appear to be any fully nonparametric tests for conditional independence.[2]

Our first application concerns the evaluation of the impact of a social program such as a job training program. Let D denote the dummy variable such that $D = 1$ when the person receives treatment (participates), and $D = 0$ if not treated. Let Y_1 and Y_0 be the outcomes associated with the participation values $D = 1$ and $D = 0$, respectively, and let X denote individual observed characteristics. A common measure of the impact of partial coverage programs, such as job training programs, is the *average treatment effects on the treated*

$$E(Y_1 - Y_0 \mid D = 1, X) = E(Y_1 \mid D = 1, X) - E(Y_0 \mid D = 1, X).$$

If it exceeds the appropriate measure of cost, the program should be maintained, see for example Heckman et al. (1998). The main problem in the estimation of $E(Y_1 - Y_0 \mid D = 1, X)$ is that the second term, $E(Y_0 \mid D = 1, X)$, cannot be observed. Replacing it with the observable average outcomes of nonparticipants $E(Y_0 \mid D = 0, X)$ leads to the presence of the self-selection bias term $B(X) = E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X)$. One can try to characterize $B(X)$ by using a control group (people

[1] See Phillips (1988) for a discussion of the difference between independence and conditional independence.

[2] When X, Y, Z are jointly normal with mean μ and covariance matrix $\Sigma = (\sigma_{ij})$, $Y \perp\!\!\!\perp X$ is equivalent to $\sigma_{YX} = 0$, while $Y \perp\!\!\!\perp X \mid Z$ is equivalent to $\sigma^{YX} = 0$, where the concentration matrix $\Sigma^{-1} = (\sigma^{ij})$. In this case, there are simple parametric tests of both independence and conditional independence. For categorical data, there are also numerous tests of independence and conditional independence, see (Agresti, 1990, p. 228).

that applied to participate in the program but were randomly denied access to program) to estimate $E(Y_0 \mid D = 1, X)$ and a comparison group (eligible non-participants) to estimate $E(Y_0 \mid D = 0, X)$. The potentially high dimension of X, however, makes direct nonparametric estimation of $B(X)$ problematic. Instead, the most common approach in this literature is to use the probability of program participation given observed characteristics, $\Pr(D = 1 \mid X) = P(X)$, also referred to as the *propensity score*, to characterize the bias. The important role of the propensity score is often motivated by the results of Rosenbaum and Rubin (1983). They show that if there exists an X such that

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid X,$$

and $0 < \Pr(D = 1 \mid X) < 1$ for all X, then, conditioning on X is equivalent to conditioning on the univariate index $P(X)$: specifically, $E(Y_0 \mid D, X) = E(Y_0 \mid X) = E(Y_0 \mid P(X))$, so that

$$B(X) = B\{P(X)\} = 0, \quad \text{for all } X.$$

This index sufficiency restriction is essentially the conditional independence restriction that treatment D is ignorable given the observables X.

Our second application concerns semiparametric model specification. Consider the semiparametric binary choice model

$$Y = \begin{cases} 1 & \text{if } \beta' X \geq \varepsilon \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where β is a vector of unknown parameters and ε is an unobservable stochastic error term. The semiparametric literature divides into two broad categories according to whether ε is assumed to be independent of X or only mean (actually median) independent (see the recent review papers by Manski, 1994, and Powell, 1994, for discussion). In the latter case, Manski (1975) developed the maximum score procedure for estimating β which was subsequently shown by Kim and Pollard (1990) to converge, on centering, at rate $n^{1/3}$ to a non-normal limit. Horowitz (1992) suggested a smoothed version of the maximum score procedure obtaining, under smoothness conditions, asymptotic normality at a rate faster than $n^{1/3}$, but still less than $n^{1/2}$. In fact, Chamberlain (1987) showed that the semiparametric information in this model is zero: i.e., that one cannot estimate β in this model at the usual $n^{1/2}$ rate. By contrast, in the case that ε is assumed to be independent of X, it is possible to estimate β with the $n^{1/2}$ rate of convergence; the Ichimura (1993) and Klein and Spady (1993) procedures both achieve this. If

$$\varepsilon \perp\!\!\!\perp X, \tag{3}$$

then

$$Y \perp\!\!\!\perp X \mid \beta^{\intercal} X, \tag{4}$$

so that the independence condition on the unobservable random variable ε implies the conditional independence of the observable quantities. The conditional independence restriction (4) is weaker than (3), which suggests that an alternative way of specifying (2) would be to assume the weaker condition (4). Note also that when Y is binary, for example, independence is equivalent to mean independence.

The above discussion holds much more generally in the class of transformation models considered in Han (1987), $Y = D \circ F(\beta^{\intercal} X, \varepsilon)$, where D is a monotonic function, while F is monotonic in each of its arguments. This includes transformation models, binary choice, duration models and censored regression. We can also extend the discussion to include panel data models where the independence assumption is even more crucial. Suppose that $Y_t = F(\alpha + \beta^{\intercal} X_t + \varepsilon_t)$, $t = 1, \ldots, \intercal$, where the composite random term $\alpha + \varepsilon_t$ is independent of $\beta^{\intercal} X_1, \ldots, \beta^{\intercal} X_{\intercal}$. Then

$$Y_1, \ldots, Y_{\intercal} \perp\!\!\!\perp X_1, \ldots, X_{\intercal} \mid \beta^{\intercal} X_1, \ldots, \beta^{\intercal} X_{\intercal}.$$

In this case, it has only been possible to consistently estimate β under the independence assumption, see (Powell, 1994, p. 2513).

Let $\sigma(Y), \sigma(X)$, and $\sigma(Z)$ be, respectively, the sigma algebras generated by the random variables Y, X, and Z defined on the same probability space $(\Omega, \mathcal{F}, P)$. We say that $\sigma(Y)$ is independent of $\sigma(X)$ given $\sigma(Z)$, if for all $A \in \sigma(Y)$ and $B \in \sigma(X)$, we have

$$P(A \cap B \mid \sigma(Z)) = P(A \mid \sigma(Z))\, P(B \mid \sigma(Z)),$$

where $P(A \mid \sigma(Z))$ is the conditional probability of the event A given $\sigma(Z)$.

When Y, X, Z are all continuous random variables with joint density function $f(y, x, z)$ with respect to Lebesgue measure on $\mathbb{R}^d$, then conditional independence of Y, X given $Z = z$ becomes

$$\frac{f(y, x, z)}{f(z)} = \frac{f(y, z)}{f(z)} \frac{f(x, z)}{f(z)},$$

for all y, x, z in the support of Y, X, Z or, equivalently,

$$f(z) f(y, x, z) = f(y, z) f(x, z).^{3} \tag{5}$$

[3]Similarly, when Y, X, Z are all discrete random variables with joint probability mass function $p(y, x, z)$, then conditional independence implies that $p(z)p(y, x, z) = p(y, z)p(x, z)$ for all y, x, z. In the sequel we will use $f()$ to denote both density and probability mass functions to simplify notation and let the continuous or discrete nature of the random variables determine its interpretation.

A consistent nonparametric test of conditional independence based on the relationship between joint and marginal density functions given by (5) would appear to require (except in the special case where all variables are discrete) nonparametric smoothing estimation of potentially high dimensional density functions $f(y, x, z)$. The main disadvantage of this approach is the slower rate of growth of the noncentrality of the test statistic which results in lower power against local alternatives. In particular, a smoothing based test would not be able to detect alternatives at distance $n^{-1/2}$ from the null detected by parametric and some nonparametric test statistics.

Our strategy to avoid this dimensionality problem is to verify the relationship (5) over subsets (with positive Lebesgue measure) in the support of Y, X, Z rather than at individual values, thereby replacing densities with distribution functions. Expression (5) is extended to

$$P(C)P(A \cap B \cap C) = P(A \cap C)P(B \cap C), \tag{6}$$

for all subsets A, B in the support of Y, X, and C a subset in the open support of Z.[4]

We now come to the contribution of this paper. We first provide a nonparametric test of (1) based on the empirical distribution functions representation of condition (6). A key question addressed in the sequel is how to choose the subsets A, B, and C so as to apply this principle to mixed continuous and discrete data. We extend the usual treatment based on quadrants to a general class of rectangles suitable for these types of data. In our first example the central hypothesis is simple (although it is about a function), while the second case was composite and there were unknown nuisance parameters involved; our test is therefore devised to allow for estimated parameters.[5] A key technical issue we face in this paper is the verification of the stochastic equicontinuity property for processes involving both discontinuous indicator functions of rectangles and nonlinear functions of the underlying parameters. Our test statistic is easy to compute and to analyze. Its asymptotic distribution is a functional of a Gaussian process whose quantiles are found by the bootstrap.

The use of (6) over (5) has the advantage of increased power over local alternatives for any dimension, but it introduces a difficulty. As the conditioning set C in (6) deviates from the singleton set $C = \{z\}$, we will see in the next section that the exact equivalence between conditional

[4]We need to exclude C from being the entire support since (6) would reduce to $P(A \cap B) = P(A)P(B)$ which implies $Y \perp\!\!\!\perp X$ but it is well known that conditional independence does not not imply independence (Chow and Teicher, 1998, p. 221).

[5]Note that (3) itself is not directly testable because one cannot estimate ε consistently.

independence (1) and expression (5) is only partially preserved by (6). In particular we will see that while (6) implies (1), the reverse is not always true.[6] In other words, the set of distributions P for which (6) holds is a subset of the set of distributions for which conditional independence (1) holds. The practical implication of this is that a test of conditional independence based on (6) may have a noncentrality different from zero not only in cases of conditional dependence but also in some cases where conditional independence holds. However, our proposed bootstrap is able to correct this by automatically adjusting for the noncentrality present under the null of conditional independence. Our simulations illustrate the performance of our test in such a case where (6) does not hold exactly under the null.

We remark that since this paper was first written (Linton and Gozalo, 1995), a number of authors have proposed tests of conditional independence, see for example Fernandes and Flores (2001), Su and White (2008), and Song (2009). These authors all use smoothing techniques (i.e., kernels) and obtain tests that are consistent of the full null hypothesis against all alternatives in a large class.

The rest of the paper is organized as follows. In Section 2 we introduce our test statistics, in Section 3 we make our assumptions and present the limiting distributions (we use an i.i.d. setup suitable for cross-sectional data); Section 4 introduces a bootstrap method for obtaining critical values. We provide a small simulation experiment in Section 5.

Notation. We use $\mathbf{1}(\cdot)$ for the indicator function, i.e., $\mathbf{1}(A) = 1$ if event A occurs and $\mathbf{1}(A) = 0$ otherwise. Let $\Rightarrow$ and $\to_p$ denote weak convergence of probability measures and convergence in probability, respectively; all limits are taken as sample size $n \to \infty$.

2. TEST STATISTICS

The underlying population is the random vector $U \in \mathbb{R}^q$ from which we observe an independent and identically distributed (i.i.d.) sample $\{U_i\}_{i=1}^n$. Of interest are certain residual [or index] functions computed from U, that is $V(U;\theta) = (Y(U;\theta), X(U;\theta), Z(U;\theta)) \in \mathbb{R}^d$, where the parameter $\theta \in \Theta \subset \mathbb{R}^p$ and $d = l + m + k$. The null hypothesis to be tested is that $Y(U;\theta^0)$ and $X(U;\theta^0)$ are independent conditional on $Z(U;\theta^0)$ for some particular θ^0 whose value is not known.

We shall base our test on the equality (6) for some (separating) class of subsets and replace the population probability measure by an empirical measure. Most previous work has been based on quadrants,

[6]This was pointed out to us by Jon Wellner to whom we are grateful.

i.e., the empirical distribution function.[7] These sets apparently work well for continuous data but, as currently applied, are unsuited for discrete data as the following example illustrates. Suppose that (Y, X) are binary with $\Pr(Y = 1, X = 0) = \Pr(Y = 0, X = 1) = 1/2$, then Y and X are perfectly dependent with the same marginals $\Pr(Y = 1) = \Pr(X = 1) = 1/2$. Unfortunately, quadrants located at the observations will not uncover this dependence; in fact, $\Pr(Y \leq 1, X \leq 0) = \Pr(Y \leq 1)\Pr(X \leq 0)$, $\Pr(Y \leq 1, X \leq 1) = \Pr(Y \leq 1)\Pr(X \leq 1)$, etc. In view of this, we consider the more general class of all rectangular subsets of $\mathbb{R}^d$ of a certain (possibly zero) width. Let a_α, b_α, $\alpha = 1, \ldots, d$ be given nonnegative numbers, possibly infinite, and let

$$\mathfrak{B}_\alpha(v_\alpha) = [v_\alpha - a_\alpha, v_\alpha + b_\alpha]$$

be a rectangle in the component α. Let also $\mathfrak{B}(v) = \times_{\alpha=1}^{d} \mathfrak{B}_\alpha(v_\alpha)$, and $\mathfrak{B}(y)$, $\mathfrak{B}(x)$, and $\mathfrak{B}(z)$ be the rectangles obtained by intersecting the corresponding intervals.[8] Then let

$$F(v \mid \theta) = P\left(V(U; \theta) \in \mathfrak{B}(v)\right)$$

be the joint rectangular distribution function of V, and denote the corresponding probability functions of (Y, Z), (X, Z), and Z by $G(y, z \mid \theta)$, $H(x, z \mid \theta)$, and $L(z \mid \theta)$, respectively; also, let $F(v) = F(v \mid \theta^0)$, $G(y, z) = G(y, z \mid \theta^0)$, $H(x, z) = H(x, z \mid \theta^0)$, and $L(z) = L(z \mid \theta^0)$. When $b_\alpha = 0$ and $a_\alpha = \infty$, these functions correspond to the usual distribution functions. For discrete variables, events of the form $\{V \leq v\}$ are not a wise choice, as discussed above, and would give zero power against some alternatives, see Joag-Dev (1984). For these variables we shall take $a_\alpha = b_\alpha = 0$.[9] For continuously distributed data we take $a_\alpha > 0$ and $b_\alpha \geq 0$, except we also rule out the case $a_\alpha = b_\alpha = \infty$. Note that the choice of rectangles can vary with location, so that a data series with both continuous and discrete components can be accommodated by choosing atomic rectangles at points of discreteness but intervals elsewhere. There is, therefore, wide latitude in choosing which rectangles to use for a given application. We discuss this further in section 5 below.

[7] There has also been some work using multivariate half spaces, i.e., hyperplanes, see Beran and Millar (1986).

[8] Formally speaking, the sets we examine are of the form $A = \{V \in \mathfrak{B}(y) \times (-\infty, \infty)^{m+k}\} \in \mathbb{R}^d$, $B = \{V \in \mathfrak{B}(x) \times (-\infty, \infty)^{l+k}\} \in \mathbb{R}^d$, and $C = \{V \in \mathfrak{B}(z) \times (-\infty, \infty)^{l+m}\} \in \mathbb{R}^d$. Then, for example $A \cap B \cap C = \{V \in \mathfrak{B}(v)\}$.

[9] In our discrete example, the dependence is uncovered by this choice of events, since clearly $\Pr(Y = 1, X = 0) \neq \Pr(Y = 1)\Pr(X = 0)$.

Letting $A(v \mid \theta, P) = L(z \mid \theta)F(v \mid \theta) - G(y, z \mid \theta)H(x, z \mid \theta)$, the Eq. (6) is equivalent to.[10]

$$A(v \mid \theta^0, P) = 0, \quad \text{for all } v \in \mathbb{R}^d, \quad \text{some } \theta^0 \in \Theta \subset \mathbb{R}^p. \tag{7}$$

A number of functionals of A can be used to test conditional independence; specifically, the Kolmogorov-Smirnov $KS = \sup_v |A(v \mid \theta^0, P)|$ and the Cramér von-Mises $CM = \int A^2(v \mid \theta^0, P)d\mu(v)$ for some measure $\mu(\cdot)$ (for example $\mu = F$). Shorack and Wellner (1986) discuss a number of alternative test functionals in a variety of contexts. The quantities KS and CM provide a general measure of the amount of conditional independence there is.

Let now P^{XYZ} be the joint distribution of (X, Y, Z), and write $P^{XYZ} = P^{XY \mid Z} \cdot P^Z$, where $P^{XY \mid Z}$ and P^Z are the distribution of (X, Y) conditional on Z, and the marginal distribution of Z, respectively. Similarly, let $P^{X \mid Z}$ and $P^{Y \mid Z}$ denote the distributions of X and Y conditional on Z, respectively. Then we define P_{CI}^{XYZ} as the joint distribution of (X, Y, Z) generated by P^{XYZ} subject to the null hypothesis $P^{XY \mid Z} = P^{X \mid Z} \cdot P^{Y \mid Z}$ of independence between X and Y conditional on Z. That is,

$$P_{CI}^{XYZ} = P^{X \mid Z} \cdot P^{Y \mid Z} \cdot P^Z.$$

This is just a functional of the original probability measure P^{XYZ}. We can then express the null hypothesis of conditional independence as

$$\mathbf{H_0} : P^{XYZ}(v \mid \theta^0) = P_{CI}^{XYZ}(v \mid \theta^0), \quad \text{for all } v \in \mathbb{R}^d : \lambda(\mathfrak{B}(z)) = 0,$$
$$\text{some } \theta^0 \in \Theta \subset \mathbb{R}^p,$$

where $\lambda(\cdot)$ denotes Lebesgue measure. Thus, for some parameter value θ^0, $P^{XYZ}(v \mid \theta^0)$ and $P_{CI}^{XYZ}(v \mid \theta^0)$ agree under $\mathbf{H_0}$ over all rectangles $\mathfrak{B}(v)$ of zero width in the conditional variable Z. The alternative hypothesis $\mathbf{H_A}$ is the negation of this for each $\theta \in \Theta \subset \mathbb{R}^p$.

Consider now how (7), which is evaluated over all rectangles $\mathfrak{B}(v)$, relates to $\mathbf{H_0}$, which is defined over all rectangles $\mathfrak{B}(v)$ with single-valued conditioning sets $\mathfrak{B}(z) = \{z\}$. Suppose that (7) holds. If (X, Y, Z) has a discrete distribution, (7) holding over all rectangles $\mathfrak{B}(v)$ implies that it also holds for single-valued sets $\mathfrak{B}(v) = \{v\}$. Hence (5) holds and this implies $\mathbf{H_0}$. Similarly, when (X, Y, Z) has a continuous distribution, a limiting argument for $\mathfrak{B}(v)$ shrinking towards a single-valued set $\{v\}$ shows that (5) holds and this again implies $\mathbf{H_0}$. Therefore, (7) implies $\mathbf{H_0}$.

[10]This is because the class of rectangles of a given width separates probability measures. That is, if two probability measures P_1 and P_2 agree on the class of all rectangles of given width, then they agree on all Borel sets.

Suppose now that $\mathbf{H_0}$ holds. The first element of $A(v \mid \theta, P)$ (omitting the dependence on θ) is

$$\begin{aligned} L(z)F(v) &= \int_{\mathfrak{B}(z)} f(z)\,dz \int_{\mathfrak{B}(z)} \int_{\mathfrak{B}(y)} \int_{\mathfrak{B}(x)} f(y, x, z)\,dy dx dz \\ &= \int_{\mathfrak{B}(z)} f(z)\,dz \int_{\mathfrak{B}(z)} \left[\int_{\mathfrak{B}(y)} \int_{\mathfrak{B}(x)} f(y, x \mid z)\,dy dx \right] f(z)\,dz \\ &= \int_{\mathfrak{B}(z)} f(z)\,dz \int_{\mathfrak{B}(z)} \left[\int_{\mathfrak{B}(y)} f(y \mid z)\,dy \right] \left[\int_{\mathfrak{B}(x)} f(x \mid z)\,dx \right] f(z)\,dz, \end{aligned}$$

where the last equality follows by $\mathbf{H_0}$. But, in general,

$$\begin{aligned} &\int_{\mathfrak{B}(z)} f(z)\,dz \int_{\mathfrak{B}(z)} \left[\int_{\mathfrak{B}(y)} f(y \mid z)\,dy \right] \left[\int_{\mathfrak{B}(x)} f(x \mid z)\,dx \right] f(z)\,dz \\ &\quad \neq \int_{\mathfrak{B}(z)} \left[\int_{\mathfrak{B}(y)} f(y \mid z)\,dy \right] f(z)\,dz \int_{\mathfrak{B}(z)} \left[\int_{\mathfrak{B}(x)} f(x \mid z)\,dy \right] f(z)\,dz \\ &\quad = G(y, z)H(x, z), \end{aligned}$$

whenever $\mathfrak{B}(z)$ is not a single-valued set. Therefore, in general, $\mathbf{H_0}$ does not imply $A(v \mid \theta^0, P) = 0$.[11] In other words, the set of distributions P for which (7) holds is a subset of the set of distributions for which conditional independence $\mathbf{H_0}$ holds.

Our choice of functional based on multivalued conditioning sets is therefore nonstandard in the sense that it is zero only when additional restrictions are true, and therefore can have noncentrality different from zero for some values of the null hypothesis. However, we show that this noncentrality can be estimated and that our test statistics are consistent for the null hypothesis of interest. Specifically, we will see that $A_n(v \mid \theta)$ can be decomposed as

$$A_n(v \mid \theta) = A(v \mid \theta, P) + \Delta_n(v \mid \theta) + O_p(n^{-1}),$$

where $\Delta_n(v \mid \theta)$ is a zero-mean stochastic term that determines the limiting distribution of our test. Let $A(v \mid \theta, P_{CI})$ denote $A(v \mid \theta, P)$ when P is replaced by the distribution it generates under $\mathbf{H_0}$, P_{CI}, and let $\mu(v \mid \theta, P) = A(v \mid \theta, P) - A(v \mid \theta, P_{CI})$. Then, adding and subtracting $A(v \mid \theta, P_{CI})$, we can express $A_n(v \mid \theta)$ as

$$\begin{aligned} A_n(v \mid \theta) &= \mu(v \mid \theta, P) + A(v \mid \theta, P_{CI}) + \Delta_n(v \mid \theta) + O_p(n^{-1}) \\ &= \mu(v \mid \theta, P) + D_n(v \mid \theta) + O_p(n^{-1}), \end{aligned}$$

[11] A case where $\mathbf{H_0}$ would imply $A(v \mid \theta^0, P) = 0$ is when $f(z)$ is constant for all z in $\mathfrak{B}(z)$ and $P(\mathfrak{B}(y)|z)$ or $P(\mathfrak{B}(x)|z)$ are constant for all z in $\mathfrak{B}(z)$.

say. Note that $\mathbf{H_0}$ is true if and only if the noncentrality $\mu(v \mid \theta^0, P) = 0$ for all v, while $\mu(v \mid \theta, P) > 0$ under $\mathbf{H_A}$ for each θ and some v (which may depend on θ). Typical test statistics based on empirical distributions require an approximation to the case-dependent distribution of the zero-mean term $\Delta_n(v \mid \theta)$. In Section 4 below, we propose a bootstrap that approximates the distribution of the (possibly nonzero mean) term $D_n(v \mid \theta)$.

To implement the test suppose that there exists estimates $\hat{\theta}$ of θ^0 that are root-n consistent under $\mathbf{H_0}$, and replace $A(v \mid \theta^0, P)$ by its empirical analogue $A(v|\hat{\theta}, P_n)$ or $A_n = L_n F_n - G_n H_n$, suppressing dependence on $\hat{\theta}$ and P_n, where

$$L_n(z) = n^{-1} \sum_{j=1}^{n} 1\{\widehat{Z}_j \in \mathfrak{B}(z)\}; \quad G_n(y,z) = n^{-1} \sum_{j=1}^{n} 1\{(\widehat{Y}_j, \widehat{Z}_j) \in \mathfrak{B}(y,z)\}$$

$$F_n(v) = n^{-1} \sum_{j=1}^{n} 1\{\widehat{V}_j \in \mathfrak{B}(v)\}; \quad H_n(x,z) = n^{-1} \sum_{j=1}^{n} 1\{(\widehat{X}_j, \widehat{Z}_j) \in \mathfrak{B}(x,z)\}$$

with $\widehat{Z}_j = Z_j(U_j; \hat{\theta})$, $\widehat{X}_j = X_j(U_j; \hat{\theta})$, and $\widehat{Y}_j = Y_j(U_j; \hat{\theta})$. We then estimate *CM* and *KS* by

$$CM_n = n^{-1} \sum_{i=1}^{n} A_n^2(\widehat{V}_i); \quad KS_n = \max_{1 \le i \le n} \left| A_n(\widehat{V}_i) \right|. \tag{8}$$

Note that a maximum is used in KS_n instead of the usual supremum. This particular version of the Kolmogorov–Smirnov statistic has recently been suggested by Andrews (1997) in another context. It has the advantage of requiring only $O(n^2)$ computations. Computation of both tests can be completely vectorized.[12] Given critical values $\hat{c}_\alpha$, our level-α test based on either test statistic $I_n = CM_n, KS_n$ is then

$$\text{Reject if}: \quad I_n > \hat{c}_\alpha. \tag{9}$$

[12]Although CM_n and KS_n are desirable from a computational point of view, they can have poor (small sample) performance for large d, because the evaluation points $\widehat{V}_i$ are not representative enough. In practice, the following statistics may work better with large d and small n,

$$CM_n^f = m^{-1} \sum_{i=1}^{m} A_n^2(t_i); \quad KS_n^f = \max_{1 \le i \le m} |A_n(t_i)|,$$

where $\{t_i; i = 1, \ldots, m\}$ is a fixed or random grid of points. The number of evaluation points, m, is under the control of the practitioner, but should increase with sample size, see Beran and Millar (1986) for justification of this device. In the simulations presented in Section 5, we used a random grid of points based on the observations.

The bootstrap critical values $\hat{c}_\alpha$ have the property that $\Pr[I_n > \hat{c}_\alpha \mid \mathbf{H}_0] \to \alpha$ and $\Pr[I_n > \hat{c}_\alpha \mid \mathbf{H}_A] \to 1$.

3. ASYMPTOTIC PROPERTIES OF THE TEST

We now establish the asymptotic properties of CM_n and KS_n. The main technical difficulty here is that V is a nonlinear function of both the data and the parameters and occurs inside an indicator which is itself a non-smooth function. Empirical processes with estimated parameters were first studied by Durbin (1973). There followed a number of papers that extended his results to a variety of situations, including nonlinear and dependent data. See the books Van der Vaart and Wellner (1996) and Shorack and Wellner (1986) for many references. Some recent works of special interest to econometricians include Bai (1994), Andrews (1997), and Koul (1996).

First of all we introduce some notation. Define: $\delta_1(\cdot, v \mid \theta) = \mathbf{1}\{V(\cdot, \theta) \in \mathfrak{B}(v)\} - F(v)$, $\delta_2(\cdot, v \mid \theta) = \mathbf{1}\{Z(\cdot, \theta) \in \mathfrak{B}(z)\} - L(z)$, $\delta_3(\cdot, v \mid \theta) = \mathbf{1}\{(X(\cdot, \theta), Z(\cdot, \theta)) \in \mathfrak{B}(x, z)\} - H(x, z)$,

$\delta_4(\cdot, v \mid \theta) = \mathbf{1}\{(Y(\cdot, \theta), Z(\cdot, \theta)) \in \mathfrak{B}(y, z)\} - G(y, z)$, and

$$\delta_0(\cdot, v \mid \theta) = L(z)\delta_1(\cdot, v \mid \theta) + F(v)\delta_2(\cdot, v \mid \theta) - G(y, z)\delta_3(\cdot, v \mid \theta) - H(x, z)\delta_4(\cdot, v \mid \theta).$$

The process $\Delta_n(v \mid \theta) = n^{-1} \sum_{i=1}^n \delta_0(U_i, v \mid \theta)$ is an approximation to $A_n(v \mid \theta)$ in the sense that

$$A_n(v \mid \theta) - A(v \mid \theta) = \Delta_n(v \mid \theta) + O_p(n^{-1}), \tag{10}$$

where the error is uniform in both v and θ. If θ^0 were known, then the asymptotic distribution of the empirical process $\Delta_n(v \mid \theta^0)$ determines the limiting distribution of our test. When θ^0 is replaced by an estimate we must also take account of its variation. For this we must calculate how $\Delta_n(v \mid \theta)$ changes with movements in θ. Letting $\Delta_j(u \mid \theta) = E_{\theta^0}[\delta_j(U, V(u, \theta) \mid \theta)]$ for $j = 0, \ldots, 4$, we have

$$\Delta_n(V(u, \theta) \mid \theta) = \Delta_n(V(u, \theta^0) \mid \theta^0) + \frac{\partial \Delta_0(u \mid \theta^0)}{\partial \theta^\intercal}(\theta - \theta^0) + O_p(|\theta - \theta^0|^2).$$

We make the following assumptions.

Assumption A1. *Under* $\mathbf{H}_0$,

$$\sqrt{n}(\hat{\theta} - \theta^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(U_i \mid \theta^0) + o_p(1),$$

where $E[\psi(U_i \mid \theta^0)] = 0$ *and* $E[\psi(U_i \mid \theta^0)\psi(U_i \mid \theta^0)^\intercal] < \infty$.

Assumption A2. *The function $V(u;\theta)$ is uniformly continuous in u and twice continuously differentiable in θ on $\Theta_0 = \{\theta: \left|\theta - \theta^0\right| \le c\}$ for some $c > 0$, with $E[|\frac{\partial V_\ell}{\partial \theta_k}(U_i, \theta^0)|^2] < \infty$ and for some $\kappa \ge 2$, $E[\sup_{\theta\in\Theta_0} |\frac{\partial^2 V_\ell}{\partial\theta_k \partial\theta_r}(U_i,\theta)|^\kappa] < \infty$ for $\ell = 1,\dots,d$ and $k, r = 1,\dots,p$.*

Assumption A3. *The functions $\Delta_j(\cdot\,|\,\theta)$, $j = 0,\dots,4$ are continuously differentiable in θ on Θ_0, and the derivative vector $\Gamma(\cdot\,|\,\theta) = \partial\Delta_0(\cdot\,|\,\theta)\big/\partial\theta^\intercal$ satisfies*

$$\int \Gamma(U\,|\,\theta^0)\Gamma(U\,|\,\theta^0)^\intercal dP(U) < \infty, \tag{11}$$

where P is the distribution of U.

Remarks. Assumption 2 could, perhaps, be weakened to once differentiability at the cost of a longer proof. In general these assumptions are fairly standard and can be verified directly. For illustrative purposes, we just consider the linear model

$$y = \beta_0^\intercal X + \varepsilon, \tag{12}$$

where $\beta_0 = (1,1)^\intercal$ and $X = (X_1, X_2)^\intercal$, with $(\varepsilon, X^\intercal)^\intercal \sim N(0, I_3)$. Consider testing the hypothesis that $y \perp\!\!\!\perp \alpha_0^\intercal X \,|\, \beta_0^\intercal X$, where $\alpha_0 = (1,-1)^\intercal$. For any α and β, we have

$$\begin{pmatrix} y \\ \alpha^\intercal X \\ \beta^\intercal X \end{pmatrix} \sim N\left[0, \begin{pmatrix} 1+\beta_0^\intercal\beta_0 & \beta_0^\intercal\alpha & \beta_0^\intercal\beta \\ \beta_0^\intercal\alpha & \alpha^\intercal\alpha & \alpha^\intercal\beta \\ \beta_0^\intercal\beta & \alpha^\intercal\beta & \beta^\intercal\beta \end{pmatrix}\right].$$

In order for Assumption 3 to be satisfied, this covariance matrix should be nonsingular at $\alpha = \alpha_0$ and $\beta = \beta_0$; this certainly holds, since by construction $\beta_0^\intercal\alpha_0 = 0$. The derivatives of Δ_j are fairly easy to compute in this case. For example,

$$\frac{\partial}{\partial\beta} E\left[\mathbf{1}\left\{\beta^\intercal X \le \beta^\intercal x\right\}\right] = \left[I - (\beta^\intercal\beta)^{-1}\beta\beta^\intercal\right]\frac{x}{(\beta^\intercal\beta)^{1/2}}\phi(\beta^\intercal x/(\beta^\intercal\beta)^{1/2}).$$

This quantity is mean zero and has finite variance with respect to the distribution of x. Kim and Pollard (1990) carry out similar calculations for general distributions with instead $\beta^\intercal x$ replaced by 0.

The large sample properties of our test statistics are given in the following theorem which is proved in the Appendix. Let

$$CM_n^c = n^{-1}\sum_{i=1}^n \left(A_n(\widehat{V}_i) - \mu(V_i)\right)^2; \quad KS_n^c = \max_{1\le i\le n}\left|A_n(\widehat{V}_i) - \mu(V_i)\right|.$$

Theorem 1. *Suppose that Assumptions A1–A3 hold. (i) Under* H_0,

$$nCM_n^c \Rightarrow \sum_{\ell=1}^{\infty} \lambda_\ell \chi_{1\ell}^2, \tag{13}$$

where $\chi_{1\ell}^2$, $\ell = 1, 2, \ldots$ *are independent chi-squared [with one degree of freedom] random variables, while* $\{\lambda_\ell\}_{\ell=1}^{\infty}$ *are the eigenvalues of the operator* T, *where* $\mathsf{T}q(\cdot) = \int h(\cdot, y) q(y) dP(y)$ *in which* $h(u_1, u_2) = \int \zeta(u_1, V(U, \theta^0) \mid \theta^0) \zeta(u_2, V(U, \theta^0) \mid \theta^0) dP(U)$ *with* $\zeta(u, v \mid \theta) = \delta(u, v \mid \theta) + \Gamma(v \mid \theta)\psi(u \mid \theta)$; *and (ii) under* H_0,

$$n^{1/2} KS_n^c \Rightarrow \sup_{t \in \mathbb{R}^q} |W(t)|, \tag{14}$$

in which W *is a Gaussian process with mean zero and covariance function* $\omega(u, u') = \int \zeta(U, V(u, \theta^0) \mid \theta^0) \zeta(U, V(u', \theta^0) \mid \theta^0) dP(U)$.

The limiting distributions are non-Gaussian. Also, there is a "correction factor" [the term $\Gamma(v \mid \theta^0)\psi(u \mid \theta^0)$ inside $\zeta(u, v \mid \theta^0)$] in the limiting distribution of both test statistics due to the estimation of θ^0. When the parameters are known, this term disappears and $\zeta = \delta$. Similarly, when the parameters enter in a linear fashion, the correction term can be zero, see for example Pierce and Kopecky (1979). Even in this case, the null distributions of our tests are complicated functionals of a Gaussian process and depend on the underlying distribution, i.e., neither test is distribution free. This is why we use the bootstrap, see below, to construct critical values.

Consider next the power of our tests against local alternatives. Our tests should have power against all root-n alternatives, just like the Bierens and Ploberger test (1997). This is true for discrete variables by virtue of the events $\mathfrak{B}$ we have taken; this is also true regardless of the dimensionality of V.[13] We suppose, for simplicity, that the parameters are known. The choice of how to specify alternatives even for the (unconditional) independence test is not universally agreed on, see (Nikitin, 1995, p. 194). For simplicity we shall assume that the data are generated by a sequence of distribution functions $\overline{F}$ shrinking towards a distribution function F that does satisfy the null hypothesis, i.e.,

$$\mathbf{H_n} : \overline{F}(v) = F(v) + \frac{a_V(v)}{n^{1/2}}$$

[13]Also note that the rate of convergence to the limiting distributions in Theorem 1 is $n^{1/2}$ independently of dimensions which implies that the size distortion is of order $n^{-1/2}$ independently of dimensions. See Csörgó and Faraway (1996).

for some function $a(\cdot)$ which is not identically zero (and which makes $\overline{F}(v)$ a probability for all v for n larger than some n_0). This implies that

$$\overline{L}(z) = L(z) + \frac{a_Z(z)}{n^{1/2}}; \quad \overline{G}(y,z) = G(y,z) + \frac{a_{YZ}(y,z)}{n^{1/2}};$$
$$\overline{H}(x,z) = H(x,z) + \frac{a_{XZ}(x,z)}{n^{1/2}}$$

for functions $a_Z(z) = a_V(\infty,\infty,z)$, $a_{YZ}(y,z) = a_V(y,\infty,z)$, and $a_{XZ}(x,z) = a_V(\infty,x,z)$. For $\overline{F}(v)$ to be a proper alternative hypothesis, we require that

$$\tau(v) = L(z)a_V(v) + F(v)a_Z(z) - G(y,z)a_{XZ}(x,z) - H(x,z)a_{YZ}(y,z)$$

is not identically zero. In any case, under the sequence of hypotheses $\mathbf{H}_n$, we have

$$n^{1/2}KS_n^c \Rightarrow \sup_{t\in\mathbb{R}^q} |W(t) + \tau(t)|.$$

This guarantees nontrivial power against such alternatives. A similar result holds for the Cramér-von Mises test.

4. BOOTSTRAP CRITICAL VALUES

We use the bootstrap because it performs well in many other related situations (and can lead to some improvements, see for example Canepa and Godfrey, 2007) and because the alternative bounding methods suggested in Bierens and Ploberger (1997), for example, are quite complicated to implement and unintuitive.

We first discuss the method for the case that the parameters θ^0 are known. The basic problem for the bootstrap is how to impose the null hypothesis in the resampling scheme. Simple resampling from the empirical joint distribution of V_i will not impose the null restriction. In the independence case, one can resample from the marginal empiricals thereby imposing independence, see for example Skaug and Tjøstheim (1993). We essentially do the same here except that our marginals are conditional on Z. Let P_n^{XYZ} be the joint empirical distribution, then write $P_n^{XYZ} = P_n^{XY|Z} \cdot P_n^Z$, where $P_n^{XY|Z}$ and P_n^Z are the empirical distribution of (X,Y) conditional on Z, and the empirical marginal distribution of Z, respectively. The conditioning variable Z is an ancillary statistic, so that we can conduct inference conditional on the sample $\{Z_i\}_{i=1}^n$ without any loss of information, i.e., we can work with $P_n^{XY|Z}$. Our proposal consists of drawing resamples $\{X_i^*, Y_i^*, Z_i^*\}_{i=1}^n$, where $Z_i^* = Z_i$, from a conditional distribution

$\widehat{P}_n^{XY\,|\,Z}$ in which we impose the null hypothesis of independence between X and Y conditional on Z. That is,

$$\widehat{P}_n^{XY\,|\,Z} = \widehat{P}_n^{X\,|\,Z} \cdot \widehat{P}_n^{Y\,|\,Z},$$

where $\widehat{P}_n^{X\,|\,Z}$ and $\widehat{P}_n^{Y\,|\,Z}$ denote the bootstrap conditional distributions of X and Y, respectively. We just explain the procedure for computing $\widehat{P}_n^{X\,|\,Z}$, since $\widehat{P}_n^{Y\,|\,Z}$ is constructed in the same manner. Unlike with the joint distribution P^{XZ} of X and Z where the (naive) bootstrap distribution can be chosen to be the empirical distribution $P_n^{XZ} = n^{-1}\sum_{i=1}^n \mathbf{1}(X = X_i)\mathbf{1}(Z = Z_i)$, the analogy does not carry over to $\widehat{P}_n^{X\,|\,Z}$. The reason for this is simple: unless one has repeated observations for Z among the observed values $\{Z_i\}_{i=1}^n$, only one value of X, namely X_i, will be associated with each Z_i, $i = 1, \ldots, n$, so that

$$\widehat{P}_n^{X\,|\,Z}\ (X_i^* \mid Z_i) = \frac{n^{-1}\sum_{j=1}^n \mathbf{1}(Z_j = Z_i)\mathbf{1}(X_j = X_i^*)}{n^{-1}\sum_{j=1}^n \mathbf{1}(Z_j = Z_i)},$$

or $X_i^* = X_i$ with probability 1, $i = 1, \ldots, n$. Even in the rare event that each value of Z in our sample is associated with two or three distinct values of X, it will still be inadequate to produce a good approximation of $P^{X\,|\,Z}$ through the empirical distribution $P_n^{X\,|\,Z}$.

One way to solve this problem is to smooth $P_n^{X\,|\,Z}$. We choose the following smoothing procedure in our simulations and application below. For any set A, including singletons, let

$$\widehat{P}_n^{X\,|\,Z}\ (A \mid Z_i) = \frac{n^{-1}\sum_{j=1}^n K_h(\|Z_j - Z_i\|)\mathbf{1}(X_j \in A)}{n^{-1}\sum_{j=1}^n K_h(\|Z_j - Z_i\|)}, \tag{15}$$

where $K_h(u) = h^{-1}K(u/h)$, and the univariate kernel K is a symmetric, nonnegative function that integrates to one, and is absolutely integrable. In practice, a weighted distance is chosen to reflect the different scales of the vector components.[14] We resample from (15); this involves choosing

$$X_i^* = X_j \quad \text{with probability } \frac{K_h(\|Z_j - Z_i\|)}{\sum_{j=1}^n K_h(\|Z_j - Z_i\|)}, \quad j = 1, \ldots, n.$$

In practice, it will be advisable in small samples to choose the bandwidth parameter h to be, for example, the distance from Z_i to its kth nearest

[14]See Härdle and Linton (1994) for discussion of smoothing methods and Horowitz (1995) for background on the bootstrap.

neighbor (k-NN).[15] This guarantees that each X_i^* is drawn from at least k observations of X whose associated Z are the k closest to Z_i. We generate B bootstrap samples and with each sample compute CM_n^* and KS_n^* in analogous fashion to CM_n and KS_n. The level α critical values $\hat{c}_\alpha$ are computed as an approximate solution to

$$\Pr^*[CM_n^* > \hat{c}_\alpha] = \alpha, \tag{16}$$

where $\Pr^*$ denotes probability conditional on the sample.

The consistency of this procedure,

$$\sup_{c \in \mathbb{R}} |\Pr^*(I_n^* \le c) - \Pr(I_n \le c)| = o_p(1), \quad n \to \infty,$$

where I_n denotes either CM_n or KS_n and I_n^* denotes either CM_n^* or KS_n^*, should follow from the following argument. Firstly, suppose that instead of the fixed distribution P of U, there was a deterministic sequence P_n of probability measures which for each n satisfies the null hypothesis. Theorem 1 can be extended to include this triangular array and, provided $P_n \to P$, the limiting distribution is the same as given in Theorem 1. Secondly, one can show that $\widehat{P}_n^{X|Z}$ and $\widehat{P}_n^{Y|Z}$ are uniformly consistent, under regularity conditions such as can be found in Härdle et al. (1988). Thus, the bootstrap test statistic has the same asymptotic distribution.

Suppose now that θ^0 is replaced by the estimated value $\hat{\theta}$. In some special cases the correction term due to parameter estimation is zero, i.e., $E(\Gamma) = 0$. This occurs in the linear index model when X is mean zero. In this case, one can use the above algorithm without re-estimating θ each time. In general, however, $E(\Gamma) \neq 0$. We suppose that there is a well defined inversion mapping $r(Y, X, Z) = U$ [which is certainly the case in linear index models]. In this case, we recommend the following procedure:

1. With the original data and $\hat{\theta}$ estimate $\widehat{P}_n^{X|Z}$ and $\widehat{P}_n^{Y|Z}$ but also $\widehat{P}_n^{Z}$ (for the latter just take the unsmoothed empirical distribution function). Now draw a random sample $\{Y_i^*, X_i^*, Z_i^*\}_{i=1}^n$ from the joint distribution $\widehat{P}_n^{X|Z} \cdot \widehat{P}_n^{Y|Z} \cdot \widehat{P}_n^{Z}$.
2. Compute $U_i^* = r(Y_i^*, X_i^*, Z_i^*)$ and re-estimate $\hat{\theta}^*$ using the bootstrap sample $\{U_i^*\}_{i=1}^n$.
3. Compute $\widehat{Y}_i^* = Y(U_i^*, \hat{\theta}^*)$, $\widehat{X}_i^* = X(U_i^*, \hat{\theta}^*)$, and $\widehat{Z}_i^* = Z(U_i^*, \hat{\theta}^*)$, $i = 1, \ldots, n$.
4. Compute CM_n^* and KS_n^* using the bootstrap sample $\{\widehat{Y}_i^*, \widehat{X}_i^*, \widehat{Z}_i^*\}_{i=1}^n$.

[15] For such h, and K the uniform distribution, we get a k-nearest neighbor smooth distribution with $X_i^* = X_j$ with probability $1/k$ for all X_j, $j = 1, \ldots, n$, such that $Z_j \in \mathcal{N}_k(Z_i)$, where $\mathcal{N}_k(Z_i)$ denotes a k-neighborhood of Z_i.

Repeat the above B times and compute $\hat{c}_\alpha$ as in (16).

In the cases where there is no an inversion mapping r, one could instead use the approach described in Horowitz (1995), where the recentred test statistic $CM_n^* - CM_n$ is used and the resample does not need to impose the null hypothesis.

5. SIMULATIONS

We evaluate the performance of our test in a binary choice model to test the conditional independence restriction (4). Specifically, we take the following designs:

$$D1. \quad Y = \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 > \varepsilon),$$

$$D2. \quad Y = \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 + 10 n^{-1/2}(X_1^2 + X_2^2) > \varepsilon),$$

$$D3. \quad Y = \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 + (X_1^2 + X_2^2) > \varepsilon),$$

$$D4. \quad Y = \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 > \varepsilon(X_1^2 + X_2^2)^{1/2}),$$

where in all cases $\beta_1 = \beta_2 = 1$, $X = (X_1, X_2)$ is bivariate standard normal, and ε is standard normal independent of X.[16] The first design satisfies the null hypothesis of conditional independence of Y from X given the index $Z = \beta_1 X_1 + \beta_2 X_2$. The second is an order $n^{-1/2}$ local alternative to this hypothesis. The third and fourth designs represent global alternatives arising from location and scale shifts, respectively. Note that conditioning on the index $Z = \beta_1 X_1 + \beta_2 X_2$ implies that Y is independent of (X_1, X_2) given $\beta_1 X_1 + \beta_2 X_2$ reduces to testing H_0**:** Y is independent of X_1 given $\beta_1 X_1 + \beta_2 X_2$.

In order to implement the tests of index sufficiency we first need to estimate the index. There are two estimators to consider; the probit estimator $\hat{\beta}_P$ and the Klein–Spady semiparametrically efficient estimate $\hat{\beta}_{KS}$. The implementation of the Klein–Spady estimator is perhaps problematic in that one has to select bandwidth and trimming parameters for which there is very little theoretical guidance as yet. Furthermore, this procedure is also quite time consuming for our purposes, since we have to re-compute this quantity for each bootstrap sample. We could have, therefore, based our simulations on $\hat{\beta}_P$. This, however, leaves one open to the criticism that an estimate too efficient under the null is being used relative to $\hat{\beta}_{KS}$, the central semiparametric estimator here. To address all these issues, we used the following first order approximation to $\hat{\beta}_{KS}$. Consider the

[16]Additional simulations with (Y, X, Z) trivariate normal are reported in Linton and Gozalo (1995).

local alternative in which $\Pr(Y=1|X)=\Phi\left\{\beta^{0\intercal}X+n^{-1/2}g(X)\right\}$ for some function $g(\cdot)$. Let

$$\tilde{\beta}_{KS}=\beta^0-\left\{n^{-1}\sum_{i=1}^{n}\frac{\phi_i^2\left(X_i-\overline{X}_i\right)\left(X_i-\overline{X}_i\right)^{\intercal}}{\Phi_i(1-\Phi_i)}\right\}^{-1}\times\left\{n^{-1}\sum_{i=1}^{n}\frac{Y_i-\Phi_i-\phi_i\overline{g}_i\,/\,n^{1/2}}{\Phi_i(1-\Phi_i)}\phi_i\left(X_i-\overline{X}_i\right)\right\},$$

with $\Phi_i=\Phi(\beta^{0\intercal}X_i)$, $\phi_i=\phi(\beta^{0\intercal}X_i)$, $\overline{X}_i=E(X_i\,|\,\beta^{0\intercal}X_i)$, $g_i=g(X_i)$, and $\overline{g}_i=E(g_i\,|\,\beta^{0\intercal}X_i)$.[17] Then, under both $D1$ and $D2$, $n^{1/2}(\tilde{\beta}_{KS}-\hat{\beta}_{KS})=o_p(1)$. In fact this result holds under the global alternatives $D3$ and $D4$, except that in those misspecified cases $\tilde{\beta}_{KS}$ nor $\hat{\beta}_{KS}$ will converge to β^0.

In the construction of the test we used zero width rectangles for the discrete variable Y, $\mathfrak{B}(x)=(-\infty,x]$ for the continuous variable X, and $\mathfrak{B}(z)=[z-\hat{\sigma}_{50}/4,z-\hat{\sigma}_{50}/4]$ for the index Z, where $\hat{\sigma}_{50}$ denotes the estimated standard deviation of Z from one fix sample of size $n=50$. Depending on the location of the evaluation point z, the interval $\mathfrak{B}(z)$ will contain different number of observations. The interval width of $\mathfrak{B}(z)$ was kept fix and independent of n throughout the simulations. We did not experiment greatly wiht the choice of rectangles and presumably one could achieve superior performance via tuning of this choice.

For each estimated index value $\widetilde{Z}_i=\tilde{\beta}_1X_{1,i}+\tilde{\beta}_2X_{2,i}$, the dependence score $A_n=L_nF_n-G_nH_n$ was evaluated at the m_i sample points $V_j=(Y_j,X_j,\widetilde{Z}_i)$ for observations j whose $\widetilde{Z}_j$ is in the interval $\mathfrak{B}(\widetilde{Z}_i)$. This results in $m=\sum_{i=1}^{n}m_i$ evaluation points. For the sample sizes considered of $n=50$, $n=100$, and $n=500$, the average value of m_i was approximately 7.5, 14.7, and 70.2, respectively, resulting on $m=375$, 1470, and 35100 evaluation points, respectively.[18]

We conducted 500 replications of the Cramér-von Mises and Kolmogorov–Smirnov type tests under the null, and 100 under each alternative design. We used 100 bootstrap samples in each replication to calculate the critical values at significance levels $\alpha=1\%$, 5%, 10%, and 20%. To compute the bootstrap test, we used (15) to obtain the bootstrap observations X_i^* with Z_i and Z_j replaced by the estimated indexes $\widetilde{Z}_i$ and $\widetilde{Z}_j$, Epanechnikov kernel $K(\cdot)$, and bandwidth h equal to the distance from $\widetilde{Z}_i$ to its kth nearest neighbor. The values of k chosen were $k=5$,

[17]Note that $E(X_j\,|\,\sum_{i=1}^{p}X_i)=p^{-1}\sum_{i=1}^{p}X_i$, for $j=1,\ldots,p$, and $E(\sum_{i=1}^{p}X_i^2\,|\,\sum_{i=1}^{p}X_i)=(p-1)+p^{-1}(\sum_{i=1}^{p}X_i)^2$.

[18]Given the large value of m using this procedure for $n=500$, we decided to evaluated the test at only 10% of the points in each interval $\mathfrak{B}(\widetilde{Z}_i)$. This cut m to a more manageable 3510 points on average.

TABLE 1 Size and power with estimated conditioning variable Z.

		Percentage rejections of null hypothesis ($Z = \tilde{\beta}_{KS}^{\intercal} X$)					
		Cramér-von mises			Kolmogorov–Smirnov		
Design	α (%)	$n = 50$	$n = 100$	$n = 500$	$n = 50$	$n = 100$	$n = 500$
D1	20	15.4	20.6	20.4	24.0	25.6	20.0
	10	5.8	8.6	10.2	12.4	12.4	10.6
	5	2.8	3.4	5.4	5.0	6.2	5.8
	1	1.2	1.0	1.2	1.2	1.4	2.8
D2	20	19.0	37.0	86.0	32.0	38.0	60.0
	10	11.0	15.0	60.0	20.0	21.0	36.0
	5	2.0	3.0	38.0	9.0	9.0	19.0
	1	1.0	0.0	12.0	3.0	1.0	6.0
D3	20	15.0	37.0	97.0	24.0	38.0	90.0
	10	6.0	15.0	92.0	15.0	21.0	66.0
	5	3.0	3.0	74.0	6.0	9.0	44.0
	1	1.0	0.0	45.0	1.0	1.0	23.0
D4	20	11.0	22.0	38.0	14.0	19.0	32.0
	10	6.0	11.0	24.0	7.0	9.0	18.0
	5	2.0	4.0	18.0	1.0	3.0	13.0
	1	0.0	1.0	4.0	0.0	1.0	2.0

10, and 10 for $n = 50$, 100, and 500, respectively. Similarly for Y_i^*. The bootstrap sample (Y_i^*, X_i^*), $i = 1, \ldots, n$, was then used to obtain new parameter estimates β^* with which we form a new set of index values $\tilde{Z}_i^*$.

TABLE 2 Size and power with known conditioning variable Z

		Percentage rejections of null hypothesis ($Z = \tilde{\beta}_{KS}^{\intercal} X$)					
		Cramér-von mises			Kolmogorov–Smirnov		
Design	α (%)	$n = 50$	$n = 100$	$n = 500$	$n = 50$	$n = 100$	$n = 500$
D1	20	18.4	19.6	20.2	23.2	21.6	20.4
	10	7.8	11.0	10.2	11.6	11.0	10.2
	5	3.8	5.4	5.2	5.8	6.2	5.4
	1	0.4	1.0	0.8	1.4	1.4	1.8
D2	20	20.0	51.0	94.0	26.0	40.0	61.0
	10	8.0	22.0	80.0	13.0	20.0	41.0
	5	3.0	9.0	52.0	4.0	11.0	23.0
	1	1.0	2.0	18.0	2.0	3.0	8.0
D3	20	20.0	51.0	100.0	28.0	40.0	92.0
	10	9.0	22.0	100.0	18.0	20.0	76.0
	5	3.0	9.0	97.0	8.0	11.0	63.0
	1	0.0	2.0	82.0	2.0	3.0	31.0
D4	20	14.0	33.0	59.0	20.0	30.0	42.0
	10	8.0	21.0	37.0	8.0	11.0	26.0
	5	3.0	8.0	28.0	2.0	4.0	17.0
	1	0.0	2.0	9.0	1.0	2.0	4.0

Finally, $(Y_i^*, X_i^*, \widetilde{Z}_i^*)$, $i = 1, \ldots, n$, is used to compute the bootstrap tests CM_n^* and KS_n^*,.

Our results using the Klein–Spady index estimate are given in Table 1.

The Cramér-von Mises test appears to have good size, even in relatively small samples, while the Kolmogorov–Smirnov test requires a larger sample size to achieve values close to the nominal values. Both tests have power against the root-n local alternative design $D2$, and have power against the global alternatives $D3$ and $D4$ of shifts in the location and the scale of the distribution of the error term (particularly against $D3$). The Cramér-von Mises test has higher power against all alternatives except for $n = 50$.

To evaluate the size/power loss due to having to estimate the index, we computed the two tests with $Z = \beta_1 X_1 + \beta_2 X_2$ assumed known. The results are given in Table 2. There is not much difference in size performance between the two tests. The power has increased for all designs and sample sizes, as expected, but particularly for the scale-shift design $D4$.

6. CONCLUSIONS

We have proposed a test of conditional independence or rather a stronger condition than conditional independence. The test has the advantage that it does not involve an explicit bandwidth parameter and has good statistical properties. One key question is how to choose the rectangles used in the construction of our test. The default rectangles are the half spaces used in the usual construction of c.d.f.'s and these may work well for most applications, but may be improved in certain cases. We leave this for future work.

A. APPENDIX

A.1. Preliminaries

Let $\Theta_n(c) = \{\theta\colon\ \sqrt{n}\left|\theta - \theta^0\right| \leq c\}$. Since $\sqrt{n}(\hat{\theta} - \theta^0) = O_p(1)$, for all $\epsilon > 0$ there exists a c_ϵ such that $\Pr[\hat{\theta} \in \Theta_n(c_\epsilon)] \geq 1 - \epsilon$. Let $\mathscr{A}_\theta$ be any event depending on θ, and let $\mathscr{B}$ be the event that $\hat{\theta} \in \Theta_n(c_\epsilon)$. Then $\Pr[\mathscr{A}_{\hat{\theta}}] = \Pr[\mathscr{A}_{\hat{\theta}} \cap \mathscr{B}] + \Pr[\mathscr{A}_{\hat{\theta}} \cap \mathscr{B}^c] \leq \Pr[\mathscr{A}_{\hat{\theta}} \cap \mathscr{B}] + \Pr[\mathscr{B}^c] = \Pr[\mathscr{A}_{\hat{\theta}} \cap \mathscr{B}] + o(1)$. Finally, we can replace $\hat{\theta}$ by the fixed value in $\Theta_n(c)$ that makes $\Pr[\mathscr{A}_\theta \cap \Theta_n(c)]$ the largest.

We begin by providing some background results concerning the processes:

$$v_n(\theta, v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{1}\left\{V(U_i, \theta) \in \mathfrak{V}(v)\right\} - E\left\{\mathbf{1}(V(U_i, \theta) \in \mathfrak{V}(v)\right\}\right],$$

$$\theta \in \Theta_n, v \in \mathbb{R}^d$$

$$v'_n(\theta, u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{1}\left\{ V(U_i, \theta) \in \mathfrak{B}(V(u, \theta)) \right\} - E\left\{ \mathbf{1}(V(U_i, \theta) \in \mathfrak{B}(V(u, \theta)) \right\} \right],$$
$$\theta \in \Theta_n, u \in \mathbb{R}^q.$$

The same results hold for the corresponding empirical processes involving subvectors of $V(U_i, \theta)$, but for convenience we just state results for $v_n(\theta, v)$ and $v'_n(\theta, v)$. Note that for given θ, these are standard empirical processes with the slight difference driven by the shape of the sets $\mathfrak{B}(v)$. Define the pseudo-metric

$$\rho((\theta, v), (\theta', v')) = E\left(\left[\mathbf{1}\left\{ V(U_i, \theta) \in \mathfrak{B}(v) \right\} - \mathbf{1}\left\{ V(U_i, \theta') \in \mathfrak{B}(v') \right\} \right]^2 \right),$$

on $\Theta_0 \times \mathbb{R}^d$. Likewise, define the pseudo-metric $\rho'((\theta, u), (\theta', u'))$ on $\Theta_0 \times \mathbb{R}^q$. Under these metrics, the parameter spaces $\Theta_0 \times \mathbb{R}^d$ and $\Theta_0 \times \mathbb{R}^q$ are totally bounded. In the sequel we shall just use the generic notation $\rho(\cdot, \cdot)$ for a metric.

We are only interested in the behavior of these processes as θ varies in the small set Θ_n. By writing $\theta = \theta^0 + \gamma n^{-1/2}$, we shall make a reparameterization to $v_n(\gamma, v)$ and $v'_n(\gamma, u)$, where $\gamma \in \Gamma(c) \subset \mathbb{R}^p$. We establish the following results:

$$\sup_{\gamma \in \Gamma, v \in \mathbb{R}^d} \left| v_n(\gamma, v) - v_n(0, v) \right| = o_p(1) \tag{17}$$

$$\sup_{\gamma \in \Gamma, v \in \mathbb{R}^d} \left| v'_n(\gamma, u) - v'_n(0, u) \right| = o_p(1). \tag{18}$$

To prove (17) and (18) it is sufficient to show a pointwise law of large numbers, e.g., $v_n(\gamma, v) - v_n(0, v) = o_p(1)$ for any $\gamma \in \Gamma, v \in \mathbb{R}^d$, and stochastic equicontinuity of the processes v_n and v'_n. The pointwise result is immediate because the random variables are sums of bounded i.i.d. random variables with zero mean; the limits of $v_n(\gamma, v)$ and $v_n(0, v)$ are the same by smoothness of the expected value in γ. To complete the proof of (17) and (18) we shall use the following lemma, proved below, which states that v_n and v'_n are stochastically equicontinuous in θ and v and θ and u respectively. First, recall definition (2.3) of Andrews (1994).

Definition. A process $v_n(\cdot)$ is stochastically equicontinuous if for all $\epsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$\limsup_{n \to \infty} \Pr\left[\sup_{\rho(t_1, t_2) < \delta} \left| v_n(t_1) - v_n(t_2) \right| > \eta \right] < \epsilon.$$

Lemma SE. *Under the above assumptions, the processes $v_n(\gamma, v)$ and $v'_n(\gamma, u)$ are stochastically equicontinuous.*

Before we give the proof of Lemma SE, we state and prove a lemma which is useful in the sequel.

Lemma C. *If a stochastic process $v_n(t)$ is stochastically equicontinuous in t, and if $t = g(s)$, where g is a uniformly continuous function on its domain, then the induced process $v_n^*(s) = v_n(g(s))$ is stochastically equicontinuous in s.*

Proof. Let $\varepsilon, \eta > 0$ be given. Take $\delta > 0$ for which

$$\limsup_{n\to\infty} \Pr\left[\sup_{\rho(t_1,t_2)<\delta} \left|v_n(t_1) - v_n(t_2)\right| > \eta\right] < \epsilon.$$

By the definition of continuity: there exists $\zeta > 0$ such that $\rho(s_1, s_2) < \zeta \Rightarrow \rho(g(s_1), g(s_2)) < \delta$. But this implies that

$$\limsup_{n\to\infty} \Pr\left[\sup_{\rho(s_1,s_2)<\zeta} \left|v_n^*(s_1) - v_n^*(s_2)\right| > \eta\right] < \epsilon,$$

which satisfies the definition of stochastic equicontinuity for the process $v_n^*(\cdot)$. □

We now give the proof of Lemma SE.

Proof of Lemma SE. We first prove the stochastic equicontinuity of $v_n(\gamma, v)$. Make a Taylor series expansion of $V(U_i, \theta)$ about $V(U_i, \theta^0)$

$$V_\ell(U_i, \theta^0 + \gamma n^{-1/2}) = V_\ell(U_i, \theta^0) + \frac{1}{\sqrt{n}} \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U_i, \theta^0)\gamma_k + \frac{1}{n} \sum_{k=1}^{p}\sum_{r=1}^{p} \frac{\partial^2 V_\ell}{\partial \theta_k \partial \theta_r}(U_i; \overline{\theta}_\ell)\gamma_k\gamma_r \tag{19}$$

for some intermediate points $\overline{\theta}_\ell$. Define the processes:

$$v_{n1}(\gamma, v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{1}\left\{ \tau\left(U_i; \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} - E\left\{ \mathbf{1}\left(\tau\left(U_i; \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} \right] \right.$$

$$v_{n2}(\gamma, v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{1}\left\{ V\left(U_i; \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} \right.$$
$$\left. - \mathbf{1}\left\{ \tau\left(U_i; \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} \right],$$

where $\tau = (\tau_1, \ldots, \tau_d)^\intercal$ with $\tau_\ell(U_i; \theta) = V_\ell(U_i, \theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U_i, \theta^0)(\theta_k - \theta_k^0)$, $\ell = 1, \ldots, d$, being the linear approximation to $V_\ell(U_i; \theta)$. Define also the deterministic centering term

$$m_{n3}(\gamma, v) = \sqrt{n} E\left[\mathbf{1}\left\{ V\left(U_i, \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} \right]$$
$$- E\left[\mathbf{1}\left\{ \tau\left(U_i, \theta^0 + \frac{\gamma}{\sqrt{n}} \right) \in \mathfrak{B}(v) \right\} \right].$$

Then, since $v_n(\gamma, v) = v_{n1}(\gamma, v) + v_{n2}(\gamma, v) + m_{n3}(\gamma, v)$, for stochastic equicontinuity of $v_n(\gamma, v)$ it suffices to establish as follows:

(a) The process $v_{n1}(\gamma, v)$ is stochastically equicontinuous in γ, v;
(b) The process $v_{n2}(\gamma, v)$ is stochastically equicontinuous in γ, v;
(c) The process $m_{n3}(\gamma, v)$ is equicontinuous in γ, v.

Proof of (a). Our argument is very similar to that contained in Sherman (1993) because we basically have a linear index structure in this part. We show that the following class $\mathcal{F}$ is Euclidean for the envelope 1, $\mathcal{F} = \{f(\cdot, \tau),\ \tau \in \Gamma \times \mathbb{R}^d\}$, where for each U and τ,

$$f(U, \tau) = \prod_{\ell=1}^{d} \mathbf{1}\left\{ V_\ell(U, \theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U, \theta^0)\gamma_k \le v_\ell + b_\ell \right\}$$
$$\times \mathbf{1}\left\{ V_\ell(U, \theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U, \theta^0)\gamma_k \ge v_\ell - a_\ell \right\}.$$

For each U, define

$$g(U, v, r, \varkappa_1, \varkappa_2, \varkappa_3, \varkappa_4) = \varkappa_1 r + \sum_{\ell=1}^{d} \varkappa_{2\ell} v_\ell + \sum_{\ell=1}^{d} \varkappa_{3\ell} V_\ell(U, \theta^0)$$
$$+ \sum_{\ell=1}^{d} \sum_{k=1}^{p} \varkappa_{2\ell k} \frac{\partial V_\ell}{\partial \theta_k}(U, \theta^0)$$

and

$$\mathcal{G} = \left\{ g(\cdot,\cdot,\cdot,\varkappa_1,\varkappa_2,\varkappa_3,\varkappa_4) : \varkappa_1 \in \mathbb{R}, \varkappa_2 \in \mathbb{R}^d, \varkappa_3 \in \mathbb{R}^d, \varkappa_4 \in \mathbb{R}^{dp} \right\}.$$

The vector space of real-valued functions $\mathcal{G}$ is of dimension $dp + 2d + 1$. For each τ, we have subgraph$[f(\cdot,\tau)] = \{(U,r) : 0 < r < f(U,\tau)\}$. This can be written as the set of all (U,r) for which the following product is equal to one

$$\prod_{\ell=1}^{d} \mathbf{1}\left\{ V_\ell(U,\theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U,\theta^0)\gamma_k \le v_\ell + b_\ell \right\}$$
$$\times \prod_{\ell=1}^{d} \mathbf{1}\left\{ V_\ell(U,\theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U,\theta^0)\gamma_k \ge v_\ell - a_\ell \right\} \mathbf{1}\{r \ge 1\}^c \, \mathbf{1}\{r > 0\}^c$$

which can be written as the set of all (U,r) for which the following quantity $\prod_{\ell=1}^{2d} \mathbf{1}\{g_\ell \ge 0\}\,\mathbf{1}\{g_{d+1} \ge 1\}^c\,\mathbf{1}\{g_{d+2} > 0\}^c$ is equal to one for some choice of $g_1,\ldots,g_{2d+2} \in \mathcal{G}$. Thus the subgraph of $f(\cdot,\tau)$ is the intersection of $2d+2$ sets each of which belongs to a polynomial class (by Lemma 2.4 in Pakes and Pollard, 1989, the class of sets of the form $\{g \ge a\}$ or $\{g > a\}$ with $g \in \mathcal{G}$ and $a \in \mathbb{R}$ is a VC class). Therefore, $\{\text{subgraph}(f),\quad f \in \mathcal{F}\}$ forms a VC class of sets. Finally, one can apply Lemma 2.12 in Pakes and Pollard (1989).

Proof of (b). By virtue of assumption A2, for any $\delta > 0$, we can find an $\varepsilon > 0$ such that $\max_{\ell=1,\ldots,d;k,r=1,\ldots,p} E[\sup_{\theta\in\Theta_n} |\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta)|^2]/\varepsilon^2 \le \delta/dp^2$. Then, by the Bonferroni and Chebychev inequalities,

$$\Pr\left[\frac{1}{\sqrt{n}} \max_{\substack{1\le i\le n\\ \ell=1,\ldots,d\\ k,r=1,\ldots,p}} \sup_{\theta\in\Theta_n} \left| \frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta) \right| > \varepsilon \right]$$
$$\le n \sum_{\substack{\ell=1,\ldots,d\\ k,r=1,\ldots,p}} \Pr\left[\frac{1}{\sqrt{n}} \sup_{\theta\in\Theta_n} \left| \frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta) \right| > \varepsilon \right]$$
$$\le \frac{\sum_{\substack{\ell=1,\ldots,d\\ k,r=1,\ldots,p}} E\left[\sup_{\theta\in\Theta_n} \left| \frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta) \right|^2 \right]}{\varepsilon^2} \le \delta.$$

Therefore, with probability tending to one

$$\sup_{\gamma\in\Gamma(c)}\max_{1\le i\le n}\left|\frac{1}{n}\sum_{k=1}^{p}\sum_{r=1}^{p}\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i;\overline{\theta}_\ell)\gamma_k\gamma_r\right|\le\frac{\overline{\pi}}{\sqrt{n}}$$

for some $\overline{\pi}<\infty$. Therefore, it suffices to show that the process

$$v_{n3}(\gamma,\pi,v)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbf{1}\left(\tau\left(U_i;\theta^0+\frac{\gamma}{\sqrt{n}}\right)+\frac{\pi}{\sqrt{n}}\in\mathfrak{B}(v)\right)\right.$$
$$\left.-E\left\{1\left(\tau\left(U_i;\theta^0+\frac{\gamma}{\sqrt{n}}\right)+\frac{\pi}{\sqrt{n}}\in\mathfrak{B}(v)\right)\right\}\right],$$

where $\pi\in\Pi$ a compact set, is stochastically equicontinuous in γ,π,v, and the deterministic centering term

$$m_{n4}(\gamma,\pi,v)=\sqrt{n}\left[E\left\{\mathbf{1}\left(\tau\left(U_i;\theta^0+\frac{\gamma}{\sqrt{n}}\right)+\frac{\pi}{\sqrt{n}}\in\mathfrak{B}(v)\right)\right\}\right.$$
$$\left.-E\left\{\mathbf{1}\left(\tau\left(U_i;\theta^0+\frac{\gamma}{\sqrt{n}}\right)\in\mathfrak{B}(v)\right)\right\}\right]$$

is also equicontinuous. The process $v_{n3}(\gamma,\pi,v)$ is stochastically equicontinuous by an obvious modification of the argument given in part (a) because the parameter π enters in a linear fashion. The centering term is handled by Taylor expansion:

$$\begin{aligned}&\left|m_{n4}(\gamma_1,\pi_1,v_1)-m_{n4}(\gamma_2,\pi_2,v_2)\right|\\&\quad\le\sup_{\gamma,\pi,v}\left|\frac{\partial m_{n4}}{\partial\gamma}(\gamma,\pi,v)\right|\left|\gamma_1-\gamma_2\right|\\&\qquad+\sup_{\gamma,\pi,v}\left|\frac{\partial m_{n4}}{\partial\pi}(\gamma,\pi,v)\right|\left|\pi_1-\pi_2\right|+\sup_{\gamma,\pi,v}\left|\frac{\partial m_{n4}}{\partial v}(\gamma,\pi,v)\right|\left|v_1-v_2\right|\\&\quad\to 0\end{aligned}$$

as $\left|\gamma_1-\gamma_2\right|+\left|\pi_1-\pi_2\right|+\left|v_1-v_2\right|\to 0$, since $\sup_{\gamma,\pi,v}\left|\partial m_{n4}(\gamma,\pi,v)/\partial\phi\right|<\infty$ with $\phi=(\gamma,\pi,v)$ by assumption A.

Proof of (c). It is straightforward to see that $|m_{n3}(\gamma,v)|\le c(\gamma,v)/n^{1/2}$ for some bounded continuous function c, so that (c) follows.

We now argue that the stochastic equicontinuity of $v_n'(\gamma, u)$ is a consequence of the result for $v_n(\gamma, v)$ combined with the uniform continuity of the function $(u, \vartheta) \longmapsto v = V(u, \vartheta)$. Define the new process

$$v_n''(\theta, \vartheta, u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[\mathbf{1}\{V(U_i, \theta) \in \mathfrak{B}(V(u, \vartheta))\} - E\{\mathbf{1}(V(U_i, \theta) \in \mathfrak{B}(V(u, \vartheta))\} \Big],$$

where $\theta, \vartheta \in \Theta_n, u \in \mathbb{R}^q$. Lemma C implies that $v_n''(\theta, \vartheta, u)$ is stochastically equicontinuous. Therefore, so is $v_n'(\theta, u)$.

A.2 Proof of Theorem 1

(i) Write $CM_n = n^{-1} \sum_{i=1}^n \{A_n(\widehat{V}_i \mid \hat{\theta})\}^2 = \int \{A_n(V(U, \hat{\theta}) | \hat{\theta})\}^2 \, dP_n(U)$, where $P_n(\cdot) = n^{-1} \sum 1(U_i \le \cdot)$ is the empirical measure of $\{U_i\}_{i=1}^n$, and let $CM_n^* = \int A_n^2(V(U, \hat{\theta}) \mid \hat{\theta}) dP(U)$, $CM_n^{**} = \int \Delta_n^2(V(U, \hat{\theta}) \mid \hat{\theta}) dP(U)$, and $CM_n^{***} = n^{-2} \sum_{i=1}^n \sum_{j=1}^n h(U_i, U_j)$, where the function h was defined in the theorem. By the triangle inequality,

$$n \left| CM_n - CM_n^{***} \right| \le n \left| CM_n - CM_n^* \right| + n \left| CM_n^{**} - CM_n^* \right| + n \left| CM_n^{**} - CM_n^{***} \right|.$$

We establish the following results:

$$n(CM_n - CM_n^*) \xrightarrow{p} 0 \tag{20}$$

$$n(CM_n^* - CM_n^{**}) \xrightarrow{p} 0 \tag{21}$$

$$n(CM_n^{**} - CM_n^{***}) \xrightarrow{p} 0. \tag{22}$$

Then, the result (i) follows because

$$nCM_n^{***} \Rightarrow \sum_{\ell=1}^{\infty} \lambda_\ell \chi_{1,\ell},$$

by standard U-statistic theory, see Skaug and Tjøstheim (1993).

Proof of (21). We have for any θ,

$$\begin{aligned} n(CM_n^* - CM_n^{**}) &= \int \{A_n^2(V(U, \theta) \mid \theta) - \Delta_n^2(V(U, \theta) \mid \theta)\} dP(U) \\ &\le 2 \sup_{v \in \mathbb{R}^d} \left| \Delta_n(v \mid \theta) \right| \sup_{v \in \mathbb{R}^d} \left| A_n(v \mid \theta) - \Delta_n(v \mid \theta) \right| \\ &\quad + \sup_{v \in \mathbb{R}^d} \left| A_n(v \mid \theta) - \Delta_n(v \mid \theta) \right|^2, \end{aligned}$$

since the probability measure $P \leq 1$. We show that

$$\sqrt{n} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} \left| A_n(v \mid \theta) - \Delta_n(v \mid \theta) \right| \to_p 0 \tag{23}$$

$$\sqrt{n} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} \left| \Delta_n(v \mid \theta) \right| = O_p(1). \tag{24}$$

Since $\Pr(\hat{\theta} \in \Theta_n^c)$ can be made arbitrarily small, (21) will then follow. Firstly,

$$\begin{aligned} A_n(v \mid \theta) - \Delta_n(v \mid \theta) = &\{F_n(v \mid \theta) - F(v)\}\{L_n(z \mid \theta) - L(z)\} \\ &- \{G_n(y, z \mid \theta) - G(y, z)\}\{H_n(x, z \mid \theta) - H(x, z)\}, \end{aligned}$$

so that (23) will follow if $n^{1/4} \sup |G_n(y, z \mid \theta) - E[G_n(y, z \mid \theta)]| \to_p 0$; $n^{1/4} \sup |E[G_n(y, z \mid \theta)] - G(y, z)| \to_p 0$; $n^{1/4} \sup |H_n(x, z \mid \theta) - E[H_n(x, z \mid \theta)] | \to_p 0$; $n^{1/4} \sup |E[H_n(x, z \mid \theta)] - H(x, z) | \to_p 0$; $n^{1/4} \sup |F_n(v \mid \theta) - E[F_n(v \mid \theta)]| \to_p 0$; $n^{1/4} \sup |E[F_n(v \mid \theta)] - F(v)| \to_p 0$; $n^{1/4} \sup |L_n(z \mid \theta) - E[L_n(z \mid \theta)] | \to_p 0$; and $n^{1/4} \sup |E[L_n (z \mid \theta)] - L(z)| \to_p 0$, where the suprema are taken over $\theta \in \Theta_n$ and $v \in \mathbb{R}^d$. We just show

$$n^{1/4} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} \left| F_n(v \mid \theta) - E\left[F_n(v \mid \theta)\right] \right| \to {}_p 0 \tag{25}$$

$$n^{1/4} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} \left| E\left[F_n(v \mid \theta)\right] - F(v) \right| \to {}_p 0, \tag{26}$$

since the argument is the same for the other terms.

The proof of (25) is a consequence of the facts that: $n^{1/4}|F_n(v \mid \theta) - E[F_n(v \mid \theta)]| \to_p 0$ for any $\theta \in \Theta_n$ and $v \in \mathbb{R}^d$ by a simple law of large numbers, along with the stochastic equicontinuity result Lemma SE established above. By the mean value theorem, we have for some intermediate vector $\overline{\theta}$,

$$\begin{aligned} n^{1/4} \left| E[F_n(v \mid \theta)] - F(v) \right| &= n^{1/4} \left| \frac{\partial F}{\partial \theta^{\intercal}}(v \mid \overline{\theta})(\theta - \theta^0) \right| \\ &\leq n^{-1/4} \sup_{\theta \in \Theta_n} \left| \frac{\partial F}{\partial \theta^{\intercal}}(v \mid \theta) \right| \sup_{\theta \in \Theta_n} |n^{1/2}(\theta - \theta^0)| \\ &\to 0, \end{aligned}$$

by assumption (A3).

The result (24) is also a consequence of Lemma SE (see the un-named proposition given in Andrews, 1994, p. 2251).

Proof of (22). This follows from the stochastic equicontinuity of the empirical process $\nu_n(\gamma, v)$ and Taylor expansion of the mean, see Andrews

(1994). Write $0 = n^{1/2}\Delta_0(u,\theta^0) = n^{1/2}\Delta_0(u,\hat{\theta}) + [\partial\Delta_0(u,\theta^*)/\partial\theta]n^{1/2}(\hat{\theta}-\theta^0)$ by the mean value theorem, where θ^* are intermediate between $\hat{\theta}$ and θ^0. By the uniform continuity of $\partial\Delta_0(u,\theta)/\partial\theta$ near θ^0, we can replace θ^* by θ^0. Specifically, writing $\Delta_0(u,\theta) = \Delta_n(u,\theta) - \{\Delta_n(u,\theta) - \Delta_0(u,\theta)\}$, we obtain

$$\sqrt{n}\Delta_n(u,\theta) = \sqrt{n}\Delta_n(u,\theta^0) + \left\{\frac{\partial\Delta_0(u,\theta^0)}{\partial\theta} + o(1)\right\}\sqrt{n}(\theta-\theta^0) + \sqrt{n}\{\Delta_n(u,\theta) - \Delta_0(u,\theta)\} - \sqrt{n}\{\Delta_n(u,\theta^0) - \Delta_0(u,\theta^0)\}$$

for any $\theta \in \Theta_n$. We now invoke (18) and the triangle inequality to argue that the second line is $o_p(1)$ uniformly in $u \in \mathbb{R}^q$ and $\theta \in \Theta_n$. Finally, substituting in the expansion for $n^{1/2}(\theta-\theta^0)$ we are done. Therefore,

$$CM_n^{**} = \int \left\{\frac{1}{n}\sum_{i=1}^n \zeta(U_i, V(U,\theta^0)\,|\,\theta^0)\right\}^2 dP(U) + o_p(n^{-1}).$$

The result follows by interchanging summation and integration.

Proof of (20). This follows from the fact that $\sqrt{n}\sup_{u\in\mathbb{R}^q}|P_n(u) - P(u)| = O_p(1)$ and (21) and (22).

Proof of (ii). We have already shown that $A_n(v\,|\,\theta)$ can be approximated by $\Delta_n(v\,|\,\theta)$ with error of order smaller than $n^{-1/2}$. Also use the argument given in (22). □

ACKNOWLEDGEMENTS

We both thank seminar participants for their helpful comments. The first author would like to thank Tilburg University for its hospitality and Pierre Chaussé for research assistance. Financial support from the National Science Foundation and the North Atlantic Treaty Organization is gratefully acknowledged.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

Andrews, D. W. K. (1994). Empirical process methods in econometrics. In: *Handbook of Econometrics*. Vol. IV. Engle, R. F., McFadden, D. L., eds. Elsevier Science B.V. Q1

Andrews, D. W. K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11:560–596. Q2

Andrews, D. W. K. (1997). A conditional Kolmogorov test. *Econometrica* 65:1097–1128.

Bai, J. (1994). Weak convergence of the sequential residual empirical process in ARMA models. *The Annals of Statistics* 22:2051–2061.

Beran, R., Millar, P. W. (1986). Confidence sets for a multivariate distribution. *The Annals of Statistics* 14:431–443.

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

Blum, J. R., Kiefer, J., Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics* 32:485–498.

Canepa, A., Godfrey, L. (2007). Improvement of the quasi-likelihood ratio test in ARMA models: some results for bootstrap methods. *Journal of Times Series Analysis* 28:434–453.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34:305–334.

Chow, Y. S., Teicher, H. (1988). *Probability Theory*. 2nd ed. Berlin: Springer Texts in Statistics.

Csörgó, S., Faraway, J. J. (1996). The exact and asymptotic distribution of Cramér von Mises statistics. *Journal of the Royal Statistical Society Series B* 58:221–234.

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

Delgado, M. (1996). Testing serial independence using the sample distribution function. *Journal of Time Series Analysis* 17:271–287.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics* 1:279–290.

Fan, Y., Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* 64:865–890.

Fernandes, M., Flores, R. G. (2001). Tests for Conditional Independence, Markovian Dynamics, and Noncausality. http://www.vwl.uni-mannheim.de/brownbag/flores.pdf **Q3**

Hall, P., Horowitz, J. (1996). Bootstrap critical values for tests based on Generalized Methods of Moments estimation. *Econometrica*. **Q2,Q4**

Godfrey, L. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. ix + pp. 252. Econometric Society Monograph No. 16. Cambridge: Cambridge University Press, 1988, reprinted 1989.

Linton, O., Gozalo, P. (1997). *Conditional Independence Restrictions: Testing and Estimation*. Discussion Paper 1140, Cowles Foundation for Research in Economics, Yale Univ.

Granger, C. W. J., Thomson, P. J. (1987). Predictive consequences of using conditioning or causal variables. *Econometric Theory* 3:150–152. **Q2**

Härdle, W., Janssen, P., Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Annals of Statistics* 16:1428–1449.

Härdle, W., Linton, O. B. (1994). Applied nonparametric methods. *The Handbook of Econometrics*. Vol. IV. McFadden, D. F., Engle III, R. F., eds. North Holland. **Q5**

Han, A. K. (1987). A non-parametric analysis of transformations. *Journal of Econometrics* 35:191–209.

Heckman, J. J., Ichimura, H., Smith, J., Todd, P. (1998). *Characterizing Selection Bias Using Experimental Data. Econometrica*. Vol. 66, No. 5 (Sep., 1998), pp. 1017–1098.

Hiemstra, C., Jones, J. D. (1994). Testing for linear and nonlinear Granger Causality in the Stock Price-Volume Relation. *The Journal of Finance* 44:1639–1664. **Q2**

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics* 58:546–557.

Hong, Y., White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica* 63:1133–1159.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* 60:505–531.

Horowitz, J. (1995). Bootstrap methods in econometrics: Theory and numerical performance. In: Kreps, D., Wallis, K. W., eds. *Advances in Economics and Econometrics: 7th World Congress*, Cambridge: Cambridge University Press, forthcoming. **Q6**

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58:71–120.

Joag-Dev, K. (1984). Measures of dependence. In: Krishnaiah, P. R., Sen, P. K., eds. *Handbook of Statistics*, Vol. 4. **Q5**

Kim, J., Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* 18:191–219.

Klein, R. W., Spady, R. H. (1993). An efficient semiparametric estimator for discrete choice models. *Econometrica* 61:387–421.

Koul, H. L. (1996). Asymptotics of some estimators and sequential residual empiricals in nonlinear time series models. *The Annals of Statistics* 24:380–404.

Linton, O. B., Gozalo, P. L. (1995). A nonparametric test of conditional independence. *Cowles Foundation Discussion Paper* no 1106.

Manski, C. F. (1975). The maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3:205–228.

Manski, C. (1994). Analog estimation of econometric models. In: Engle III, R. F., McFadden, D. F., eds. *The Handbook of Econometrics, vol. IV*. North Holland. **Q5**

Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge: Cambridge University Press.

Pakes, A., Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* 57:1027–1057.

Phillips, P. C. B. (1988). Conditional and unconditional statistical independence. *Journal of Econometrics* 38:341–348.

Pierce, D. A., Kopecky, K. J. (1979). Testing goodness of fit for the distribution of errors in regression model. *Biometrika* 66:1–5.

Powell, J. L. (1994). Estimation in semiparametric models. In: Engle III, R. F., McFadden, D. F., eds. *The Handbook of Econometrics*. Vol. IV. North Holland. **Q5**

Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Review of Economic Studies* 58:437–453.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61:123–138.

Shorack, G. R., Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: John Wiley.

Skaug, H. J., Tjøstheim, D. (1993). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika* 80:591–602.

Song, K. (2009). Testing conditional independence via rosenblatt transforms. *The Annals of Statistics* 37(6B):4011–4045.

Su, L., White, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* 24:829–864.

Van der Vaart, A. W., Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag: Berlin.

Zheng, J. Z. (1997). A consistent specification test of independence. *Journal of Nonparametric Statistics* 7(4). **Q7**