COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1140

CONDITIONAL INDEPENDENCE RESTRICTIONS: TESTING AND ESTIMATION

Oliver Linton and Pedro Gozalo

November 1996

CONDITIONAL INDEPENDENCE RESTRICTIONS: TESTING AND ESTIMATION[1]

By Oliver Linton and Pedro Gozalo[2]

Yale University and Brown University

Abstract

We propose a nonparametric test of an 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 non-model 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. Finally, we invert the test statistic to provide a method for estimating the parameters identified through the conditional independence restriction. They are asymptotically normal at rate root-n.

Key words. Conditional Independence; Empirical distribution; Independence; Nonparametric; Smooth Bootstrap; Test.

JEL classification: C12, C14, C15, C52.

[1] **March 25, 1997**.

[2] Address correspondence to: Oliver Linton, Cowles Foundation for Research in Economics, Yale University, Yale Station Box 208281, New Haven, CT 06520-8281, USA. Phone: (203) 432-3699. Fax: (203) 432-6167. E-mail address: linton@econ.yale.edu; http:\\www.econ.yale.edu\~linton.

1 Introduction

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

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

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 independent of X), in that it imposes an infinite number of restrictions on the joint distribution.[3] 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 $E(Y|X) = 0$, or that Y is mean independent of X given Z, i.e. that $E(Y|X, Z) = E(Y|Z)$. We now give two concrete reasons for interest in the conditional independence hypothesis.

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*

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

If it exceeds the appropriate measure of cost, the program should be maintained, see for example Heckman, Ichimura, Smith and Todd (1995). The main problem in the estimation of $E(Y_1 - Y_0|D = 1, X)$ is that the second term, $E(Y_0|D = 1, X)$, cannot be observed. Replacing it with the observable average outcomes of nonparticipants $E(Y_0|D = 0, X)$ leads to the presence of the self-selection bias

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

term $B(X) = E(Y_0|D = 1, X) - E(Y_0|D = 0, X)$. One can try to characterize $B(X)$ by using a control group (people that applied to participate in the program but were randomly denied access to program) to estimate $E(Y_0|D = 1, X)$ and a comparison group (eligible non-participants) to estimate $E(Y_0|D = 0, X)$. 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, $\Pr(D = 1|X) = P(X)$, 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

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

and $0 < \Pr(D = 1|X) < 1$ for all X, then, conditioning on X is equivalent to conditioning on the univariate index $P(X)$. In particular, $E(Y_0|D, X) = E(Y_0|X) = E(Y_0|P(X))$, 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. The weaker mean independence restriction is not sufficient here.

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

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

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

$$Y \perp\!\!\!\perp X|\, \beta^T X, \tag{4}$$

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), $Y = D \cdot F(\beta^T X, \varepsilon)$, where D is monotonic and non-degenerate, 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 $Y_t = F(\alpha + \beta^T X_t + \varepsilon_t)$, $t = 1, \ldots, T$, where the composite random term $\alpha + \varepsilon_t$ is independent of $\beta^T X_1, \ldots, \beta^T X_T$, then

$$Y_1, \ldots, Y_T \perp\!\!\!\perp X_1, \ldots, X_T|\, \beta^T X_1, \ldots, \beta^T X_T.$$

In this case, it has only been possible to consistently estimate β under the independence assumption, see Powell (1994, p2513).

We now come to the contribution of this paper. We first provide a nonparametric test of (1) based on an empirical distribution function. A large literature now exists on testing parametric regression models against general alternatives, see for example Bierens and Ploberger (1996), 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 (1995) 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, Kiefer and Rosenblatt (1961) and Skaug and Tjøstheim (1993) and Delgado (1996), and those based on smoothing methods like Robinson (1991) and Zheng (1994). However, there do not appear to be any fully nonparametric tests for conditional independence.[4] We propose a nonparametric test of (1) based on an empirical measure using the fact that under (1),

$$\Pr(C)\Pr(A\cap B\cap C)=\Pr(A\cap C)\Pr(B\cap C) \tag{5}$$

for any events $A\in\sigma(Y), B\in\sigma(X)$, and $C\in\sigma(Z)$. Here, $\sigma(X)$, for example, denotes the sigma algebra of events generated by X. If Z were a null random variable, so that $\sigma(Z)\supseteq\sigma(Y)\bigcup\sigma(X)$, then we are considering the (unconditional) independence of Y and X. See Chow and Teicher (1988, pp 221-223) for further discussion.[5] A key question addressed in the sequel is how to choose the events A, B, and C so as to apply this principle to mixed continuous and discrete data. We extend

[4] When X, Y, Z are jointly normal with mean μ and covariance matrix $\Sigma=(\sigma_{ij})$, $Y \perp\!\!\!\perp X\ |Z$ is equivalent to $\sigma_{YX}=0$, while $Y \perp\!\!\!\perp X$ 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, p228).

[5] The condition can be more generally stated in terms of arbitrary sigma algebras $\mathcal{F}_1,\mathcal{F}_2$, and $\mathcal{F}_3$, which allows for intermediate cases between independence and conditional independence of random variables.

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 concerned observable quantities, while in the second case there were unknown parameters involved; our test is therefore devised to allow for estimated parameters.[6] Our test statistic is easy to compute and to analyze; its asymptotic distribution is a functional of a Gaussian process whose quantiles can be found by the bootstrap. Empirical measure based tests like ours have a major theoretical advantage over smoothing based tests in that they have power against all alternatives at distance $n^{-1/2}$; this holds, in theory, for any dimension.

Our second contribution is to propose a new estimation method for parameters that are defined as the unique values that make an hypothesis of (1) between certain residual quantities be satisfied. The estimator minimizes a criterion function which is based on one of the test statistics. It can be interpreted as a version of Manski's (1983) Closest Empirical Distribution (CED) estimator. We show, using an extension of the asymptotic analysis of Pakes and Pollard (1989), that the estimator is asymptotically normal at rate root-n. Our estimator can be viewed as a substitute for the average derivative estimator of Stoker (1986,1992) or the nonlinear least squares estimator of Ichimura (1995) in estimating semiparametric index models. It has the benefit of not requiring smoothing methods for its construction. In some cases, this at the cost of the stronger conditional independence assumption; however, with a binary Y, conditional independence is equivalent to conditional mean independence and we are not requiring any stronger condition.

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.

We treat first the testing problem: in section 2 we introduce our test statistic, 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 gives a bootstrap method for obtaining critical values. Section 5 treats with the estimation problem. We provide a small simulation experiment in section 6.

Notation. We use $\mathbf{1}(\cdot)$ for the indicator function, i.e. $\mathbf{1}(A) = 1$ if event A occurs and $\mathbf{1}(A) = 0$

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

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

Our 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 (5) for some (separating) class of subsets and replace the population probability measure by an empirical measure. Most previous work has been based on quadrants, 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 $\Pr(Y = 1, X = 0) = \Pr(Y = 0, X = 1) = 1/2$, 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

[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 $A = \left\{V \in \mathfrak{B}(y) \times (-\infty, \infty)^{m+k}\right\} \in \sigma(Y)$, where

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

be the joint rectangular distribution function of V, and denote the corresponding probability functions of (Y, Z), (X, Z) and Z by $G(y, z|\theta)$, $H(x, z|\theta)$, and $L(z|\theta)$, respectively; also, let $F(v) = F(v|\theta^0)$, $G(y, z) = G(y, z|\theta^0)$, $H(x, z) = H(x, z|\theta^0)$, and $L(z) = L(z|\theta^0)$. 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 6 below.

Letting $A(v|\theta) = L(z|\theta)F(v|\theta) - G(y, z|\theta)H(x, z|\theta)$, the null hypothesis is equivalent to[10]

$$\mathbf{H_0}\text{:} \quad A(v|\theta^0) = 0, \quad \text{for all } v \in \mathbb{R}^d, \quad \text{some } \theta^0 \in \Theta \subset \mathbb{R}^p.$$

The alternative hypothesis $\mathbf{H_A}$ is the negation of this. A number of functionals of A can be used to test $\mathbf{H_0}$; specifically, the Kolmogorov-Smirnov $KS = \sup_v |A(v|\theta^0)|$ and the Cramér von-Mises $CM = \int A^2(v|\theta^0)d\mu(v)$ 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. Note that $\mathbf{H_0}$ is true if and

$\sigma(Y)$ is the sigma algebra generated by the random variable Y, $B = \left\{V \in \mathfrak{B}(x) \times (-\infty, \infty)^{l+k}\right\} \in \sigma(X)$, and $C = \left\{V \in \mathfrak{B}(z) \times (-\infty, \infty)^{l+m}\right\} \in \sigma(Z)$. Then, for example $A \cap B \cap C = \{V \in \mathfrak{B}(v)\}$. Note that, for example, $\sigma(A(y)\text{: } y \in \mathbb{R}^l) = \sigma(Y)$.

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

[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.

only if $I = 0$, for $I = CM, KS$. The quantities CM and KS provide a general measure of the amount of conditional independence there is.

We suppose that there exists estimates $\widehat{\theta}$ of θ^0 that are root-n consistent under the null hypothesis. In some cases there are many candidate estimates; we provide a general method in the second part of the paper that provides root-n consistent estimates under the assumption of conditional independence. To implement the test we replace $A(v|\theta^0)$ by the empirical analogue $A_n = L_n F_n - G_n H_n$, suppressing dependence on $\widehat{\theta}$, where

$$\begin{aligned} L_n(z) &= n^{-1}\sum_{j=1}^{n} 1\left\{\widehat{Z}_j \in \mathfrak{B}(z)\right\} \quad ; \quad G_n(y,z) = n^{-1}\sum_{j=1}^{n} 1\left\{(\widehat{Y}_j, \widehat{Z}_j) \in \mathfrak{B}(y,z)\right\} \\ F_n(v) &= n^{-1}\sum_{j=1}^{n} 1\left\{\widehat{V}_j \in \mathfrak{B}(v)\right\} \quad ; \quad H_n(x,z) = n^{-1}\sum_{j=1}^{n} 1\left\{(\widehat{X}_j, \widehat{Z}_j) \in \mathfrak{B}(x,z)\right\} \end{aligned}$$

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

$$CM_n = n^{-1}\sum_{i=1}^{n} A_n^2(\widehat{V}_i) \quad ; \quad KS_n = \max_{1\le i\le n} \left|A_n(\widehat{V}_i)\right|. \tag{6}$$

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 (1995) in another context. It has the advantage of requiring only $O(n^2)$ computations. Computation of both tests can be completely vectorized.[11] Given critical values $\widehat{c}_\alpha$, our level-α test based on either test statistic $I_n = CM_n, KS_n$ is then

[11] 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,

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

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 6 we used a random grid of points based on the observations.

$$\text{reject if}: \qquad I_n > \widehat{c}_\alpha. \tag{7}$$

In section 4 below we discuss how to compute critical values $\widehat{c}_\alpha$ with the property that $\Pr[I_n > \widehat{c}_\alpha | \ \mathbf{H}_0] \rightarrow \alpha$ and $\Pr[I_n > \widehat{c}_\alpha | \ \mathbf{H}_A] \rightarrow 1$.

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 apparently 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 recent book by 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 (1995), and Koul (1996).

First of all we introduce some notation. Define: $\delta_1(\cdot, v|\theta) = 1\left\{V(\cdot,\theta) \in \mathfrak{B}(v)\right\} - F(v)$, $\delta_2(\cdot, v|\theta) = \mathbf{1}\left\{Z(\cdot,\theta) \in \mathfrak{B}(z)\right\} - L(z)$, $\delta_3(\cdot, v|\theta) = \mathbf{1}\left\{(X(\cdot,\theta), Z(\cdot,\theta)) \in \mathfrak{B}(x,z)\right\} - H(x,z)$, $\delta_4(\cdot, v|\theta) = \mathbf{1}\left\{(Y(\cdot,\theta), Z(\cdot,\theta)) \in \mathfrak{B}(y,z)\right\} - G(y,z)$, and

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

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

$$A_n(v|\theta) - A(v|\theta) = \Delta_n(v|\theta) + O_p(n^{-1}), \tag{8}$$

where the error is uniform in both v and θ. If θ^0 were known, then the asymptotic distribution of the empirical process $\Delta_n(v|\theta^0)$ 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 $\Delta_n(v|\theta)$ changes with movements in θ. Letting

$$\Delta_j(u|\theta) = E_{\theta^0}\left\{\delta_j(U, V(u,\theta)|\theta)\right\}$$

for $j = 0, \ldots, 4$, we have

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

We make the following assumptions:

Assumption 1. *Under* $\mathbf{H}_0$,

$$\sqrt{n}(\widehat{\theta} - \theta^0) = n^{-1/2}\sum_{i=1}^{n}\psi(U_i|\theta^0) + o_p(1),$$

where $E\left\{\psi(U_i|\theta^0)\right\} = 0$ *and* $E\left\{\psi(U_i|\theta^0)\psi(U_i|\theta^0)^T\right\} < \infty$.

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

Assumption 3. *The functions* $\Delta_j(\cdot|\theta)$, $j = 0, \ldots, 4$ *are continuously differentiable in* θ *on* Θ_0, *and the derivative vector* $\Gamma(\cdot|\theta) = \partial \Delta_0(\cdot|\theta)\Big/\partial \theta^T$ *satisfies*

$$\int \Gamma(U|\theta^0)\Gamma(U|\theta^0)^T dP(U) < \infty, \tag{9}$$

where P *is the distribution of* U.

Assumption 2 could, perhaps, be weakened to once differentiability at the cost of a longer proof. In general though these assumptions are fairly standard and can be verified for the linear model as we now discuss. Suppose that

$$y = \beta_0^T X + \varepsilon, \tag{10}$$

where $\beta_0 = (1,1)^T$ and $X = (X_1, X_2)^T$, with $(\varepsilon, X^T)^T \sim N(0, I_3)$. Consider testing the hypothesis that

$$y \perp\!\!\!\perp \alpha_0^T X | \beta_0^T X, \tag{11}$$

where $\alpha_0 = (1,-1)^T$.[12] For any α and β we have

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

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 $\beta_0^T \alpha_0 = 0$. The derivatives of Δ_j are fairly easy to compute in this case. For example,

$$\begin{aligned} \frac{\partial}{\partial \beta} E\left[\mathbf{1}\left\{ \beta^T X \le \beta^T x \right\} \right] &= \frac{\partial}{\partial \beta} \Phi \left\{ \frac{\beta^T x}{(\beta^T \beta)^{1/2}} \right\} \\ &= \left[I - (\beta^T \beta)^{-1} \beta \beta^T \right] \frac{x}{(\beta^T \beta)^{1/2}} \phi \left\{ \frac{\beta^T x}{(\beta^T \beta)^{1/2}} \right\}. \end{aligned}$$

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 $\beta^T x$ replaced by 0.

The large sample properties of our test statistics are given in the following theorem which is proved in the appendix:

[12] Note that the index model defined by (4) can be equivalently stated as (11).

THEOREM 1. *Suppose that assumptions A1-A3 hold. (i) Under* H_0,

$$nCM_n \Rightarrow \sum_{\ell=1}^{\infty} \lambda_\ell \chi_{1\ell}^2, \tag{12}$$

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*

$$Tq(\cdot) = \int h(\cdot, y) q(y) dP(y)$$

in which

$$h(u_1, u_2) = \int \zeta(u_1, V(U, \theta^0)|\theta^0)\zeta(u_2, V(U, \theta^0)|\theta^0) dP(U)$$

with $\zeta(u, v|\theta) = \delta(u, v|\theta) + \Gamma(v|\theta)\psi(u|\theta)$; *and (ii) Under* H_0,

$$n^{1/2}KS_n \Rightarrow \sup_{t \in \mathbb{R}^q} |W(t)|, \tag{13}$$

in which W *is a Gaussian process with mean zero and covariance function*

$$\omega(u, u') = \int \zeta(U, V(u, \theta^0)|\theta^0)\zeta(U, V(u', \theta^0)|\theta^0) dP(U).$$

The limiting distributions are non-Gaussian. Also, there is a "correction factor" [the term $\Gamma(v|\theta^0)\psi(u|\theta^0)$ inside $\zeta(u, v|\theta^0)$] 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 (1996). 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, p194). 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.

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

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

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

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

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

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

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

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

[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).

4 Bootstrap Critical Values

We use the bootstrap because it performs well in many other related situations and because the alternative methods tried in Bierens and Ploberger (1996), for example, are quite complicated to implement.

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

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

where $\hat{P}_n^{X|Z}$ and $\hat{P}_n^{Y|Z}$ denote the bootstrap conditional distributions of X and Y, respectively. We just explain the procedure for computing $\hat{P}_n^{X|Z}$, since $\hat{P}_n^{Y|Z}$ 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 $\hat{P}_n^{X|Z}$. 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

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

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

$$\hat{P}_n^{X|Z}(A \mid Z_i) = \frac{n^{-1}\sum_{j=1}^{n} K_h(\|Z_j - Z_i\|)\mathbf{1}(X_j \in A)}{n^{-1}\sum_{j=1}^{n} K_h(\|Z_j - Z_i\|)}, \tag{14}$$

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 (14); this involves choosing

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

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 k'th nearest 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 $\widehat{c}_\alpha$ are computed as an approximate solution to

$$\Pr{}^*[CM_n^* > \widehat{c}_\alpha] = \alpha, \tag{15}$$

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

The consistency of this procedure:

$$\sup_{c \in \mathbb{R}} |\Pr{}^*(I_n^* \leq c) - \Pr(I_n \leq c)| \to 0 \quad a.s., \quad n \to \infty,$$

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

[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.

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 $\hat{P}_n^{X|Z}$ and $\hat{P}_n^{Y|Z}$ are almost surely 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 $\widehat{\theta}$. 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 $\widehat{\theta}$ estimate $\hat{P}_n^{X|Z}$ and $\hat{P}_n^{Y|Z}$ but also $\hat{P}_n^{Z}$ [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 $\hat{P}_n^{X|Z} \cdot \hat{P}_n^{Y|Z} \cdot \hat{P}_n^{Z}$.

2. Compute $U_i^* = r(Y_i^*, X_i^*, Z_i^*)$ and reestimate $\widehat{\theta}^*$ using the bootstrap sample $\{U_i^*\}_{i=1}^n$.

3. Compute $\widehat{Y}_i^* = Y(U_i^*, \widehat{\theta}^*)$, $\widehat{X}_i^* = X(U_i^*, \widehat{\theta}^*)$, and $\widehat{Z}_i^* = Z(U_i^*, \widehat{\theta}^*)$, $i = 1, \ldots, n$.

4. Compute CM_n^* and KS_n^* using the bootstrap sample $\left\{\widehat{Y}_i^*, \widehat{X}_i^*, \widehat{Z}_i^*\right\}_{i=1}^n$.

Repeat the above B times and compute $\widehat{c}_\alpha$ as in (15).

5 Estimation

5.1 Method

There has been much work in the statistics literature on models defined by conditional independence restrictions, see for example Wermuth and Lauritzen (1990). Suppose our "model" is that there is a unique θ^0 such that $Y(U;\theta^0)$ and $X(U;\theta^0)$ are independent conditional on $Z(U;\theta^0)$. For example, one might specify the binary choice model (2) in this way; this would be a slight generalization allowing the distribution of ε to depend on the index. We now examine an estimation method for θ^0 based on using only the assumption of conditional independence; specifically, we invert our Cramér von Mises test statistic.[16] We work with the Cramér von Mises criterion because of its analytical tractability. Alternative procedures based on the supremum and other norms can also be implemented; however, their asymptotic distribution is not guaranteed to be normal, see for example Rao, Schuster, and Littell (1975).

The conditional independence assumption is equivalent to saying that θ^0 globally minimizes (setting to zero) the criterion $Q(\theta) = \int A^2(v|\theta)d\mu(v)$, where $\mu(\cdot)$ is absolutely continuous with respect to Lebesgue measure. We work with a fixed measure $\mu(\cdot)$ for convenience here. Now let $\widetilde{\theta}$ be an approximate minimizer of

$$Q_n(\theta) = \int A_n^2(v|\theta)d\mu(v) = \|A_n(\theta)\|^2 . \tag{16}$$

The criterion function is discontinuous, so that $\widetilde{\theta}$ is not unique; some arbitrary tie breaking rule is needed to obtain a unique estimate. The Nelder-Mead simplex method is recommended for computing the estimates when the dimensions of θ exceed one [perhaps taking as starting values estimates computed as minimizers of a smoothed version of Q_n].

We now give some examples. (a) Suppose that we wish to estimate the parameters β of the

[16] See Pollard (1980) for a similar strategy [except that in his case the parameter to be estimated lies inside a smooth distribution function].

There is also the question about how to best estimate the distribution function F, see Brown and Newey (1996).

linear regression (10), where $\varepsilon \perp\!\!\!\perp X$ is the identifying assumption made concerning the continuously distributed random variables (ε, X). In this case, one version of the criterion function is

$$Q_n(\beta) = \int \left\{ F_n^{e,X}(e,x) - F_n^e(e) F_n^X(x) \right\}^2 d\mu(e,x) \tag{17}$$

for some positive weighting function $\mu(e,x)$, where $F_n^{e,X}(e,x)$, $F_n^e(e)$, and $F_n^X(x)$ are the corresponding empirical distribution functions, for example, $F_n^{e,X}(e,x) = n^{-1}\sum_{j=1}^n \mathbf{1}(e_j \leq e, X_j \leq x)$. This is example (iii) given in Manski (1983) who uses empirical weighting, i.e. $\mu(e,x) = F_n^{e,X}(e,x)$. (b) Suppose that instead we only assume that $Y \perp\!\!\!\perp X \,\middle|\, \beta^T X$, which allows for some heteroskedasticity albeit depending only on the index, then we might take

$$Q_n(\beta) = \int \left\{ F_n^{\beta^T X}(z) F_n^{Y,X,\beta^T X}(y,x,z) - F_n^{Y,\beta^T X}(y,z) F_n^{X,\beta^T X}(x,z) \right\}^2 d\mu(y,x,z), \tag{18}$$

as criterion function.

A competitor to our estimation method would be to use the semiparametric profile likelihood method, where available: for example, in (b) choose β to maximize $\sum_{i=1}^n \ln \widehat{f}(Y_i|\beta^T X_i)$, where $\widehat{f}$ is a kernel estimate of the conditional density $Y|\beta^T X$ evaluated at the observation points. This method has the disadvantage of requiring smoothing methods which can contribute a very large second order effect [see Linton (1995)]. It is also rather hard to define in more complicated situations like when all variables are subject to a parametric transformation.

5.2 Asymptotic Properties

The consistency argument is standard: given the identification assumption it is sufficient to establish to show that

$$\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \to_p 0,$$

which follows by $\sup_{v\in\mathbb{R}^d}\sup_{\theta\in\Theta}|A_n(v|\theta) - A(v|\theta)| \to_p 0.$[17]

[17] This result is established in the appendix for shrinking Θ_n.

Asymptotic distribution theory for estimators derived from this type of criterion function have been given in Pakes and Pollard (1989) for the case that the norm is finite dimensional, and more recently in Van der Vaart (1995) for the case that both the norm and Θ are infinite dimensional. In our case, the parameter space is finite dimensional but the norm is infinite dimensional, i.e. there are an infinite number of restrictions. We have the following theorem.

THEOREM 2. *Let $\widetilde{\theta}$ be a consistent estimator of θ^0, the unique point of Θ for which $A(u|\theta^0) = 0$ for all u. Suppose also that:*

(a) $\left\|A_n(\widetilde{\theta})\right\| \leq o_p(n^{-1/2}) + \inf_{\theta\in\Theta} \|A_n(\theta)\|$ *;*

(b) *For all s,*

$$\Delta(s|\theta) = E_{\theta^0}\left\{\delta_j(U, s|\theta)\right\}$$

is differentiable in θ at θ^0, with a vector of continuous derivatives $\Gamma(s|\theta) = \partial\Delta(s|\theta)\Big/\partial\theta^T$. The matrix $\mathcal{H} = \int \Gamma\Gamma^T(s|\theta^0)d\mu(s)$ is of full rank.

(c) *For every sequence $\{\epsilon_n\}$ of positive numbers that converges to zero,*

$$\sup_{|\theta-\theta^0|\leq\epsilon_n} n^{1/2}\left\|A_n(\theta) - A(\theta) - A_n(\theta^0)\right\| = o_p(1)$$

(d) $n^{1/2}\Delta_n(\cdot|\theta^0) \Rightarrow B(\cdot)$, *a mean zero Gaussian process with positive definite covariance function $\omega(s,t)$*

(e) θ^0 *is an interior point of Θ.*

Then

$$n^{1/2}(\widetilde{\theta} - \theta^0) \Rightarrow N(0, \Omega),$$

with $\Omega = \mathcal{H}^{-1}\mathcal{J}\mathcal{H}^{-1}$, *where* $\mathcal{J} = \int \Gamma(s|\theta^0)\Gamma^T(t|\theta^0)\omega(s,t)d\mu(s)d\mu(t)$.

Remark 1. The assumptions have been written down in such a way as to emphasize the connection with Pakes and Pollard (1989). The stochastic equicontinuity condition might also be expressed in terms of the more fundamental empirical processes as defined in Theorem 1. The derivative condition (b) was discussed before Theorem 1; the weak convergence condition (d) was verified in the proof of Theorem 1.

Remark 2. The estimator $\widetilde{\theta}$ is not generally efficient as is evident from the sandwich form of the asymptotic variance. If we could find a measure $\mu_{opt}(\cdot|\theta^0)$ for which

$$\int \Gamma(s|\theta^0)\Gamma^T(t|\theta^0)\omega(s,t)d\mu_{opt}(s|\theta^0)d\mu_{opt}(t|\theta^0) = \int \Gamma\Gamma^T(s|\theta^0)d\mu_{opt}(s|\theta^0), \qquad (19)$$

then the asymptotic variance of the estimator computed using $\mu_{opt}(\cdot|\theta^0)$ in (16) is $\Omega_{opt} = \left[\int \Gamma\Gamma^T(s|\theta^0)d\mu_{opt}(s|\theta^0)\right]^{-1}$ which is minimal among all procedures of the form (16).[18] In a simpler example, Boos (1981) gives the optimal weighting function - it depended on derivatives of an unknown density function. In our situation too, estimating the optimal weighting function will likely require smoothing methods and seems practically not very desirable.

5.3 Examples

We now discuss the linear regression examples. Suppose that we are in case (a) where we identify the parameters β by the (unconditional) independence assumption $\varepsilon \perp\!\!\!\perp X$. Furthermore, assume that

[18] It also ought to be semiparametrically efficient, see Newey (1990)

the variables are continuously distributed and we use the standard empirical distribution functions. We have

$$\begin{aligned}
\Delta_n &= F_n^{e,X} - F^{e,X} - F^e(F_n^X - F^X) - F^X(F_n^e - F^e) + O_p(n^{-1}) \\
&= \frac{1}{n}\sum_{j=1}^{n} \{\mathbf{1}(e_i \leq e) - F^e(e)\} \left\{\mathbf{1}(X_i \leq x) - F^X(x)\right\} + O_p(n^{-1})
\end{aligned}$$

and the asymptotic covariance function of $\Delta_n(e,x)$ [at the true β] is

$$\omega((e_1,x_1),(e_2,x_2)) = \{F^e(e_1 \wedge e_2) - F^e(e_1)F^e(e_2)\} \left\{F^X(x_1 \wedge x_2) - F^X(x_1)F^X(x_2)\right\}, \tag{20}$$

where $e_1 \wedge e_2$ denotes coordinate-wise minimum. Furthermore, $\Gamma(e,x|\beta^0) =$

$$\frac{\partial E\left[\Delta_n(\beta^0)\right]}{\partial\beta} = \frac{\partial E\left[F_n^{e,X}(\beta^0)\right]}{\partial\beta} - F^X \frac{\partial E\left[F_n^e(\beta^0)\right]}{\partial\beta} = f^e(e)\int_{-\infty}^{x} \{X - \mu_X\} f^X(X)dX, \tag{21}$$

where $\mu_X = E(X)$, while f^e and f^X denote the densities of ε and X respectively, since

$$\begin{aligned}
\frac{\partial E\left[F_n^{e,X}(\beta)\right]}{\partial\beta} &= \frac{\partial}{\partial\beta}E\left[\mathbf{1}\left\{\varepsilon - (\beta-\beta^0)^T X \leq e\right\}\mathbf{1}\{X \leq x\}\right] \quad \text{by identical distribution} \\
&= \frac{\partial}{\partial\beta}E_X\left[F^e(e + (\beta-\beta^0)^T X)\mathbf{1}\{X \leq x\}\right] \quad \text{byiteratedexpectation} \\
&= E_X\frac{\partial}{\partial\beta}\left[F^e(e + (\beta-\beta^0)^T X)\mathbf{1}\{X \leq x\}\right] \quad \text{by reverse order} \\
&= E_X\left[Xf^e(e + (\beta-\beta^0)^T X)\mathbf{1}\{X \leq x\}\right] \\
&= f^e(e)\int_{-\infty}^{x} Xf^X(X)dX \quad \text{when } \beta = \beta^0;
\end{aligned}$$

and by similar reasoning, $\partial E\left[F_n^e(\beta)\right]/\partial\beta = f^e(e)\int_{-\infty}^{\infty} Xf^X(X)dX$ when $\beta = \beta^0$.

Similar calculations can be produced for the case (b) where identification is achieved through the assumption that $Y \perp\!\!\!\perp X|\beta^T X$, although the formulae in this case are quite lengthy and are not repeated here.

5.4 Standard Errors

Unlike the test statistic, the estimator $\widetilde{\theta}$ is asymptotically normal, and one can construct confidence intervals using variations on the usual methods. In the example discussed in the previous section, there are some obvious estimates of the quantities (20) and (21) that appear in Ω. For example, $\Gamma(e,x|\theta^0)$ can be estimated by $\widetilde{\Gamma}(e,x|\widetilde{\theta}) = \widehat{f}^e(e)n^{-1}\sum_{i=1}^n (X_i - \overline{X})\mathbf{1}\left\{X_i \leq x\right\}$, where $\widehat{f}^e(e) = (nh)^{-1}\sum_{i=1}^n K\left\{(e-e_i)/h\right\}$ is a kernel density estimate based on the residuals $e_i = Y_i - \widetilde{\beta}^T X_i$ in which K is a scalar probability density function symmetric about zero with $h = h(n) \downarrow 0$ a scalar bandwidth sequence. The covariance function $\omega((e_1,x_1),(e_2,x_2))$ can be estimated by

$$\widetilde{\omega}((e_1,x_1),(e_2,x_2)) = \left\{F_n^e(e_1 \wedge e_2) - F_n^e(e_1)F_n^e(e_2)\right\}\left\{F_n^X(x_1 \wedge x_2) - F_n^X(x_1)F_n^X(x_2)\right\},$$

where F_n^e is computed using the residuals e_i. Then compute $\widehat{\Omega} = \widehat{\mathcal{H}}^{-1}\widehat{\mathcal{J}}\widehat{\mathcal{H}}^{-1}$, where

$$\widehat{\mathcal{H}} = \int \widetilde{\Gamma}\widetilde{\Gamma}(s|\widetilde{\theta})^T d\mu(s) \quad ; \quad \widehat{\mathcal{J}} = \int \widetilde{\Gamma}(s|\widetilde{\theta})\widetilde{\Gamma}(t|\widetilde{\theta})^T \widetilde{\omega}(s,t)\tau_n(|s-t|)d\mu(t)d\mu(s), \tag{22}$$

where the truncator function $\tau_n(r)$ eliminates contributions too far way from the diagonal, as in for example Newey and West (1987). The integrals can be computed by numerical methods.

It is convenient to give a general method for computing standard errors when Γ is hard to calculate. In view of the approximate linearity of $\Delta_0(v|\theta)$ in θ near θ^0, we propose using the local linear regression smoother [see Fan (1992) and Härdle and Linton (1994)] to estimate $\Gamma(v|\theta^0)$ which seems preferable to the numerical derivative approach suggested in Sherman (1993) and Pakes and Pollard (1989). Let $\{\theta_1, \ldots, \theta_m\}$ be a grid of parameter values chosen in a neighbourhood of $\widetilde{\theta}$ and let $\widehat{\alpha}(v,\theta)$ and $\widehat{\beta}(v,\theta)$ minimize

$$\sum_{j=1}^{m}\left\{A_n(v|\theta_j)-\alpha-\beta^T(\theta_j-\theta)\right\}^2 K\left(\frac{\theta-\theta_j}{h}\right), \tag{23}$$

where $K(\cdot)$ is a kernel function and $h=h(n)$ is a scalar bandwidth; then estimate $\Gamma(v|\theta^0)$ by

$$\widetilde{\Gamma}(v|\widetilde{\theta})=\widehat{\beta}(v,\widetilde{\theta}).$$

Note that the asymptotic variance of $\widetilde{\Gamma}(v|\widetilde{\theta})$ is $O((nh^{p+2})^{-1/2})$, while its bias is determined by the smoothness of Γ: if this function is twice continuously differentiable, then the bias of $\widetilde{\Gamma}(v|\widetilde{\theta})$ is $O(h^2)$.[19] For pointwise consistency of $\widetilde{\Gamma}(v|\widetilde{\theta})$, we therefore require that $h\to 0$ and $nh^{p+2}\to\infty$ as $n\to\infty$.[20] These results can be extended to hold uniformly in $v\in\mathbb{R}^d$ and $\theta\in\Theta_n$, see Andrews (1995) and Masry (1996). In this general case, $\omega(s,t)$ is the covariance of an empirical process

$$[L(\cdot),F(\cdot),-G(\cdot),-H(\cdot)]\,n^{1/2}\begin{bmatrix}F_n(\cdot)-F(\cdot)\\ L_n(\cdot)-L(\cdot)\\ H_n(\cdot)-H(\cdot)\\ G_n(\cdot)-G(\cdot)\end{bmatrix},$$

with the probability functions being defined on a general class of rectangles. The process $n^{1/2}\left\{F_n(\cdot)-F(\cdot)\right\}$ has covariance function $\omega_F(s,t)=F(s\sqcap t)-F(s)F(t)$, where $F(s\sqcap t)=\Pr\left\{V(U;\theta^0)\in\mathfrak{B}(s)\cap\mathfrak{B}(t)\right\}$; this quantity can be estimated consistently by $\widehat{\omega}_F(s,t)=F_n(s\sqcap t)-F_n(s)F_n(t)$, where $F_n(s\sqcap t)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{V(U_i;\widetilde{\theta})\in\mathfrak{B}(s)\cap\mathfrak{B}(t)\right\}$. Likewise for the marginal processes $n^{1/2}\left\{L_n(\cdot)-L(\cdot)\right\}$, $n^{1/2}\left\{G_n(\cdot)-G(\cdot)\right\}$, and $n^{1/2}\left\{H_n(\cdot)-H(\cdot)\right\}$. The scaling factor $[L(\cdot),F(\cdot),-G(\cdot),-H(\cdot)]$ can be estimated by $[L_n(\cdot),F_n(\cdot),-G_n(\cdot),-H_n(\cdot)]$.

An alternative method for getting standard errors is to use the bootstrap as in Wellner and Zhan (1996).

[19] If additional smoothness is present in Γ, then higher polynomials can be used in (23) and one can expect smaller asymptotic bias.

[20] Note there is no restriction on m apart from the requirement that $m\geq p+1$.

6 Simulations

6.1 Testing

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

$$\begin{aligned} D1 \quad Y &= \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 > \varepsilon), \\ D2 \quad Y &= \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 + 10n^{-1/2}(X_1^2 + X_2^2) > \varepsilon), \\ D3 \quad Y &= \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 + (X_1^2 + X_2^2) > \varepsilon), \\ D4 \quad Y &= \mathbf{1}(\beta_1 X_1 + \beta_2 X_2 > \varepsilon(X_1^2 + X_2^2)^{1/2}), \end{aligned}$$

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.[21] 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 $Y \perp\!\!\!\perp (X_1, X_2) |\ \beta_1 X_1 + \beta_2 X_2$ reduces to testing $\mathbf{H}_0 : Y \perp\!\!\!\perp X_1 |\ \beta_1 X_1 + \beta_2 X_2$.

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 $\widehat{\beta}_P$ and the Klein-Spady semiparametrically efficient estimate $\widehat{\beta}_{KS}$. 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 $\widehat{\beta}_P$. This, however, leaves one open to the criticism that an estimate too efficient under the null

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

is being used relative to $\widehat{\beta}_{KS}$, the central semiparametric estimator here. To address all these issues, we used the following first order approximation to $\widehat{\beta}_{KS}$. Consider the local alternative in which $\Pr(Y=1|X) = \Phi\left\{\beta^{0T}X + n^{-1/2}g(X)\right\}$ for some function $g(\cdot)$. Let

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

with $\Phi_i = \Phi(\beta^{0T}X_i)$, $\phi_i = \phi(\beta^{0T}X_i)$, $\overline{X}_i = E(X_i|\beta^{0T}X_i)$, $g_i = g(X_i)$ and $\overline{g}_i = E(g_i|\beta^{0T}X_i)$.[22] Then, under both $D1$ and $D2$,

$$n^{1/2}(\widetilde{\beta}_{KS} - \widehat{\beta}_{KS}) = o_p(1).$$

In fact this result holds under the global alternatives $D3$ and $D4$, except that in those misspecified cases $\widetilde{\beta}_{KS}$ nor $\widehat{\beta}_{KS}$ 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 $\mathfrak{B}(z) = [z - \widehat{\sigma}_{50}/4, z - \widehat{\sigma}_{50}/4]$ for the index Z, where $\widehat{\sigma}_{50}$ 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.

For each estimated index value $\widetilde{Z}_i = \widetilde{\beta}_1 X_{1,i} + \widetilde{\beta}_2 X_{2,i}$, 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.[23]

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

[22] 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 $E(\sum_{i=1}^{p} X_i^2|\sum_{i=1}^{p} X_i) = (p-1) + p^{-1}(\sum_{i=1}^{p} X_i)^2$.

[23] 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 $(\widetilde{Z}_i)$. This cut m to a more manageable 3510 points on average.

replication to calculate the critical values at significance levels $\alpha = 1\%,\ 5\%,\ 10\%$, and 20%. To compute the bootstrap test, we used (14) 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 k'th nearest neighbor. The values of k chosen were $k = 5$, 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 $\widetilde{\beta}^*$ with which we form a new set of index values $\widetilde{Z}_i^*$. 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.

*** Table 1,2 here ***

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 $D4$.

6.2 Estimation

We close with a small simulation study designed to illustrate the estimation procedure. We generated data from

$$Y_i = F(X_i \sin\theta + Z_i \cos\theta) + \sigma_i \varepsilon_i,$$

where $(\varepsilon_i, X_i, Z_i)$ were trivariate standard normal. The first two designs were homoskedastic, in fact $\sigma_i = 0.1$, and: in design (E1) we took F to be the identity, while in design (E2) F is the normal c.d.f. The third design (E3) has F the identity and $\sigma_i = 0.1(X_i \sin\theta + Z_i \cos\theta)^2$. We investigated three estimation procedures: nonlinear least squares (*nlls*); the "independence" estimator (*ind*) that minimized the criterion (17) with weight function the empirical measure, i.e. $\mu = F_n^{e,X}$; finally, the "conditional independence" estimator (*cind*) that minimized (18) again with empirical weights. The Nelder-Mead algorithm was used to find the minimum of each criterion function with respect to θ which was throughout set equal to zero. The root mean squared errors for the estimates of $\sin\theta$ are shown below based on 200 replications

*** Table 3 here***

In the homoskedastic designs, (E1) and (E2), the independence and conditional independence estimators perform quite similarly, both exceeding the efficient estimator (in this case, *nlls*) by somewhat less than 50% when $n = 500$. In the heteroskedastic design (E3), *ind* appears to be inconsistent, while *cind* is actually more efficient than *nlls.* We next show the density functions of the *ind* and *cind* estimators in the design E1 in comparison with a normal density.[24]

Figures Here

There is some evidence of non-normality for the smaller sample sizes but by $n = 200$ both estimates are pretty close to the normal shape.

It should be noted that the conditional independence criterion function is the same for all d.g.p.'s, while the other two estimators take account of the special known structure of the mean through the parametrically defined residuals. These latter estimators are therefore sensitive to the misspecification of this parametric function, while the conditional independence estimator is not.

[24] The densities are computed using a Kernel density routine with Silverman's rule of thumb bandwidth h. The variables are standardised and the comparable normal density is $N(0, 1 + h^2)$ reflecting the small sample smoothing bias.

7 Concluding Remarks

Our set up throughout has been with i.i.d. data suitable for microeconometric work and adopted by the many papers we cited in the introduction. However, there is one important application of our test statistic in time series: that of nonlinear Granger causality in which the null hypothesis of non-causality from the time series X_t to the series Y_t might be expressed as

$$Y_t \perp\!\!\!\perp (X_{t-1}, \ldots, X_{t-q}) | \, (Y_{t-1}, \ldots, Y_{t-p})$$

for some fixed lags p and q, see Florens and Fougere (1996).[25] A version of Theorem 1 can surely be proved here with some considerable additional work, and the technology for calculating critical values by bootstrap is in the process of being developed, see for example Hall and Horowitz (1996).

A Appendix

Let $\Theta_n(c) = \{\theta\colon \sqrt{n}\,|\theta - \theta^0| \leq c\}$. Since $\Pr(\Theta_n^c)$ can be made arbitrarily small, we can essentially confine our attention to this neighborhood.

We begin by providing some background results concerning the process

$$\nu_n(\theta, v) = n^{-1/2} \sum_{i=1}^{n} [\mathbf{1}\{V(U_i, \theta) \in \mathfrak{B}(v)\} - E\{\mathbf{1}(V(U_i, \theta) \in \mathfrak{B}(v)\}], \quad \theta \in \Theta_n, v \in \mathbb{R}^d$$

and the related process

$$\nu_n'(\theta, u) = n^{-1/2} \sum_{i=1}^{n} [\mathbf{1}\{V(U_i, \theta) \in \mathfrak{B}(V(u, \theta))\} - E\{\mathbf{1}(V(U_i, \theta) \in \mathfrak{B}(V(u, \theta))\}], \quad \theta \in \Theta_n, u \in \mathbb{R}^q.$$

The same results hold for the corresponding empirical processes involving subvectors of $V(U_i, \theta)$, but for convenience we just state results for $\nu_n(\theta, v)$ and $\nu_n'(\theta, v)$. Define the pseudo-metric

[25] See Granger and Thomson (1987) for an alternative definition and Hiemstra and Jones (1994) for an application to financial data.

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

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 $\Gamma \times \mathbb{R}^d$ and $\Gamma \times \mathbb{R}^q$ are totally bounded. In the sequel we shall just use the generic notation $\rho(\cdot,\cdot)$ for a metric.

By writing $\theta = \theta^0 + \gamma n^{-1/2}$, we shall make a reparameterization to $\nu_n(\gamma,v)$ and $\nu_n'(\gamma,u)$, where $\gamma \in \Gamma(c) \subset \mathbb{R}^p$. We establish the following

$$\sup_{\gamma\in\Gamma, v\in\mathbb{R}^d} |\nu_n(\gamma,v) - \nu_n(0,v)| = o_p(1) \tag{24}$$

and

$$\sup_{\gamma\in\Gamma, v\in\mathbb{R}^d} |\nu_n'(\gamma,u) - \nu_n'(0,u)| = o_p(1). \tag{25}$$

To prove (24) and (25) it is sufficient to show finite dimensional (fidi) convergence and stochastic equicontinuity. The fidi result is immediate. To complete the proof of (24) and (25) we shall use the following lemma, proved below, which states that these processes are stochastically equicontinuous in θ and v and θ and u respectively. Recall definition (2.3) of Andrews (1994).

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

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

We have the following result.

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

The proof of Lemma SE for $\nu_n'(\gamma, u)$ uses the following result

Lemma C. *If a stochastic process $\nu_n(t)$ is stochastically equicontinuous in t, and if $t = g(\cdot)$, where g is a uniformly continuous function, then the reparameterized process $\nu_n'(s) = \nu_n(g(s))$ is stochastically equicontinuous in s.*

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

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

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

$$\overline{\lim_{n\to\infty}} \Pr\left[\sup_{\rho(s_1,s_2)<\zeta} |\nu_n'(s_1) - \nu_n'(s_2)| > \eta\right] < \epsilon$$

which satisfies the definition of stochastic equicontinuity for the process $\nu_n'(\cdot)$. ■

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

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

for some intermediate point $\overline{\theta}$. Define the processes:

$$\nu_{n1}(\gamma, v) = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{1}\left\{T(U_i; \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\} - E\left\{\mathbf{1}(T(U_i; \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\}\right]$$

$$\nu_{n2}(\gamma, v) = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{1}\left\{V(U_i; \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\} - \mathbf{1}\left\{T(U_i; \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\}\right],$$

where $T = (T_1, \ldots, T_d)^T$ with $T_\ell(U_i; \theta) = V_\ell(U_i, \theta^0) + \sum_{k=1}^{p} \frac{\partial V_\ell}{\partial \theta_k}(U_i, \theta^0)(\theta_k - \theta_k^0)$, $\ell = 1, \ldots, d$, and the deterministic centering term

$$m_{n3}(\gamma, v) = n^{1/2} E\left[\mathbf{1}\left\{V(U_i, \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\}\right] - E\left[\mathbf{1}\left\{T(U_i, \theta^0 + \gamma n^{-1/2}) \in \mathfrak{B}(v)\right\}\right].$$

Then,

$$\nu_n(\gamma, v) = \nu_{n1}(\gamma, v) + \nu_{n2}(\gamma, v) - m_{n3}(\gamma, v).$$

Using the triangle inequality it suffices to establish that

(a) $\nu_{n1}(\gamma, v)$ is stochastically equicontinuous

(b) $\nu_{n2}(\gamma, v)$ is stochastically equicontinuous

(c) $m_{n3}(\gamma, v)$ is equicontinuous.

Proof of (a) Our argument is very similar to that contained in Sherman (1993). We show that the following class $\mathcal{F}$ is Euclidean for the envelope 1,

$$\mathcal{F} = \left\{f(\cdot, \tau), \quad \tau \in \Gamma \times \mathbb{R}^d\right\},$$

where for each U and τ,

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

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

$$\text{subgraph}[f(\cdot,\tau)] = \{(U,r) : 0 < r < f(U,\tau)\}$$

This can be written as the set of all (U,r) for which the following product is 1,

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

which can be written as the set of all (U,r) for which the following is 1,

$$\prod_{\ell=1}^{2d} \mathbf{1}\left\{g_\ell \geq 0\right\} \mathbf{1}\left\{g_{d+1} \geq 1\right\}^c \mathbf{1}\left\{g_{d+2} > 0\right\}^c$$

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 $\{g \geq a\}$ 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) For any $\delta > 0$, we can find an ε such that $\delta \geq E\left[\left\{\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta)\right|\right\}^2\right]/\varepsilon^2$. Then, by the Bonferroni and Chebychev inequalities,

$$\begin{aligned}
\Pr\left[n^{-1/2}\max_{1\leq i\leq n}\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta)\right| > \varepsilon\right] &\leq n\Pr\left[n^{-1/2}\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta)\right| > \varepsilon\right] \\
&\leq \frac{E\left[\left\{\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 V_\ell}{\partial\theta_k\partial\theta_r}(U_i,\theta)\right|\right\}^2\right]}{\varepsilon^2} \\
&\leq \delta
\end{aligned}$$

for $\ell = 1, \ldots, d$ and $k, r = 1, \ldots, p$. Thus, we can restrict our attention to the process

$$\nu_{n3}(\pi, v) = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{1}(T(U_i; \theta) + n^{-1/2}\pi \in \mathfrak{B}(v)) - E\left\{ \mathbf{1}(T(U_i; \theta) + n^{-1/2}\pi \in \mathfrak{B}(v)) \right\} \right],$$

where $\pi \in \Pi$ a compact set, and the deterministic centering term

$$m_{n4}(\pi, v) = n^{1/2} \left[E\left\{ \mathbf{1}(T(U_i; \theta) + n^{-1/2}\pi \in \mathfrak{B}(v)) \right\} - E\left\{ \mathbf{1}(T(U_i; \theta) \in \mathfrak{B}(v)) \right\} \right].$$

The process $\nu_{n3}(\pi, v)$ is stochastically equicontinuous by a modification of the argument given in (a). The centering term is handled by Taylor expansion:

$$\begin{aligned} |m_{n4}(\pi_1, v_1) - m_{n4}(\pi_2, v_2)| &\leq \sup_{\pi, v} \left| \frac{\partial m_{n4}}{\partial \pi}(\pi, v) \right| |\pi_1 - \pi_2| \\ &\quad + \sup_{\pi, v} \left| \frac{\partial m_{n4}}{\partial v}(\pi, v) \right| |v_1 - v_2| \\ &\rightarrow 0 \end{aligned}$$

as $|\pi_1 - \pi_2| + |v_1 - v_2| \rightarrow 0$.

Proof of (c). The same Taylor expansion method.

We now argue that the stochastic equicontinuity of $\nu'_n(\gamma, u)$ is a consequence of the result for $\nu_n(\gamma, v)$ combined with the uniform continuity of $V(u, \theta)$. Define the new process

$$\nu''_n(\theta, \pi, u) = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{1}\left\{ V(U_i, \theta) \in \mathfrak{B}(V(u, \pi)) \right\} - E\left\{ \mathbf{1}(V(U_i, \theta) \in \mathfrak{B}(V(u, \pi)) \right\} \right],$$

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

■

Proof of Theorem 1.

(i) Write

$$CM_n = \frac{1}{n}\sum_{i=1}^{n}\left\{A_n(\widehat{V}_i|\widehat{\theta})\right\}^2 = \int\left\{A_n(V(U,\widehat{\theta})|\widehat{\theta})\right\}^2 dP_n(U),$$

where $P_n(\cdot) = n^{-1}\sum 1\left(U_i \leq \cdot\right)$ is the empirical measure of $\{U_i\}_{i=1}^n$, and let

$$\begin{aligned} CM_n^* &= \int A_n^2(V(U,\widehat{\theta})|\widehat{\theta})dP(U) \quad ; \quad CM_n^{**} = \int \Delta_n^2(V(U,\widehat{\theta})|\widehat{\theta})dP(U) \quad ; \\ CM_n^{***} &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(U_i,U_j). \end{aligned}$$

We have to establish the following results:

$$n(CM_n - CM_n^*) \xrightarrow{p} 0 \tag{26}$$

$$n(CM_n^* - CM_n^{**}) \xrightarrow{p} 0 \tag{27}$$

$$n(CM_n^{**} - CM_n^{***}) \xrightarrow{p} 0. \tag{28}$$

The result (i) then follows since

$$nCM_n^{***} \Rightarrow \sum_{\ell=1}^{\infty}\lambda_\ell\chi_{1,\ell}, \tag{29}$$

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

(27) We have for any θ,

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

since $P \leq 1$. We show that

$$n^{1/2} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} |A_n(v|\theta) - \Delta_n(v|\theta)| \to_p 0 \tag{30}$$

$$n^{1/2} \sup_{\theta \in \Theta_n} \sup_{v \in \mathbb{R}^d} |\Delta_n(v|\theta)| = O_p(1). \tag{31}$$

Since $\Pr(\Theta_n^c)$ can be made arbitrarily small, (27) will then follow. Firstly,

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

so that (30) will follow if

$$\begin{aligned} n^{1/4} \sup |G_n(y,z|\theta) - E[G_n(y,z|\theta)]| &\to {}_p 0 \quad ; \quad n^{1/4} \sup |E[G_n(y,z|\theta)] - G(y,z)| \to_p 0 \\ n^{1/4} \sup |H_n(x,z|\theta) - E[H_n(x,z|\theta)]| &\to {}_p 0 \quad ; \quad n^{1/4} \sup |E[H_n(x,z|\theta)] - H(x,z)| \to_p 0 \\ n^{1/4} \sup |F_n(v|\theta) - E[F_n(v|\theta)]| &\to {}_p 0 \quad ; \quad n^{1/4} \sup |E[F_n(v|\theta)] - F(v)| \to_p 0 \\ n^{1/4} \sup |L_n(z|\theta) - E[L_n(z|\theta)]| &\to {}_p 0 \quad ; \quad n^{1/4} \sup |E[L_n(z|\theta)] - L(z)| \to_p 0. \end{aligned}$$

We show just the third line, since the argument is the same for the others. The first part is a consequence of the stochastic equicontinuity result Lemma SE established above. Also, by the mean value theorem, we have for some intermediate vector $\overline{\theta}$,

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

by assumption (A4).

Secondly, (31) is a consequence of fidi and Lemma SE [see the un-named proposition given in Andrews (1994, p2251)].

(28) This follows from the stochastic equicontinuity of the empirical process $\nu_n(\gamma, v)$ and Taylor expansion of the mean, see Andrews (1994). Write

$$\begin{aligned} 0 &= n^{1/2}\Delta_0(u,\theta^0) \\ &= n^{1/2}\Delta_0(u,\widehat{\theta}) + \frac{\partial \Delta_0(u,\theta^*)}{\partial\theta} n^{1/2}\left(\widehat{\theta}-\theta^0\right) \end{aligned}$$

by the mean value theorem, where θ^* are intermediate between $\widehat{\theta}$ 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. Then, writing

$$\Delta_0(u,\theta) = \Delta_n(u,\theta) - \{\Delta_n(u,\theta) - \Delta_0(u,\theta)\}$$

we obtain

$$\begin{aligned} n^{1/2}\Delta_n(u,\widehat{\theta}) &= n^{1/2}\Delta_n(u,\theta^0) + \frac{\partial \Delta_0(u,\theta^0)}{\partial\theta} n^{1/2}\left(\widehat{\theta}-\theta^0\right) \\ &\quad + n^{1/2}\left\{\Delta_n(u,\widehat{\theta}) - \Delta_0(u,\widehat{\theta})\right\} - n^{1/2}\left\{\Delta_n(u,\theta^0) - \Delta_0(u,\theta^0)\right\}. \end{aligned}$$

We now invoke (25) and the triangle inequality to argue that the second line is $o_p(1)$ uniformly in u. 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.

(26) Follows from the fact that

$$n^{1/2}\sup_{u\in\mathbb{R}^q} |P_n(u) - P(u)| = O_p(1)$$

and (27) and (28).

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 (28).

■

Proof of Theorem 2: This uses essentially the same arguments as in Pakes and Pollard (1989, p1041); we just give an outline. Define the linear function

$$L_n(s|\theta) = \Delta_n(s|\theta^0) + \Gamma(s|\theta^0)(\theta - \theta^0).$$

Asymptotically, $A_n(s|\theta)$ is approximately linear in θ, in the sense that minimizing $\|A_n(\theta)\|$ is equivalent to minimizing $\|L_n(\theta)\|$, up to order $n^{-1/2}$. This latter minimization problem can be solved explicitly to give

$$n^{1/2}(\theta^* - \theta^0) = \left[\int \Gamma\Gamma^T(s|\theta^0)d\mu(s)\right]^{-1} \int \Gamma(s|\theta^0)n^{1/2}\Delta_n(s|\theta^0)d\mu(s).$$

Since $n^{1/2}\Delta_n(\cdot|\theta^0)$ obeys a functional CLT, the right hand side is asymptotically normal with the stated mean and covariance matrix. ■

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.* John Wiley, New York.

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

Andrews, D.W.K (1995): “A conditional Kolmogorov test.” Cowles Foundation Discussion Paper, no 1111R.

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

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

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

Bierens, H.J., and W. Ploberger (1996): “Asymptotic theory of integrated conditional moment tests,” Southern Methodist University working paper. Forthcoming in *Econometrica.*

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

Boos, D.D. (1981): ”Minimum distance estimators for location and goodness of fit,” *Journal of the American Statistical Association* **76**, 663-670.

Brown, B.W. and W.K. Newey (1996): ”Bootstrapping for GMM,” Preprint, Rice University.

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

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

Csörgó, S., and J.J. Faraway (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, J. (1992): ”Design-adaptive nonparametric regression,” *Journal of the American Statistical Association* **87**, 998-1003.

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

Florens, J.P. and D. Fougere (1996): ”Noncausality in continuous time,” *Econometrica* **64**, 1195-1212.

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

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

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

HECKMAN, J.J., H. ICHIMURA, J. SMITH, AND P. TODD (1996): "Characterizing selection bias using experimental data," Preprint, Chicago University.

HÄRDLE, W., P. JANSSEN AND R. SERFLING (1988): "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *Annals of Statistics*, **16**, 1428-1449.

HÄRDLE, W., AND O.B. LINTON (1994): "Applied nonparametric methods," *The Handbook of Econometrics*, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland.

HIEMSTRA, C., AND J. D. JONES (1994): "Testing for linear and nonlinear Granger Causality in the Stock Price-Volume Relation," *The Journal of Finance* **44**, 1639-1664.

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

HONG, Y., AND H. WHITE (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 *Advances in Economics and Econometrics: 7th World Congress*, D. Kreps and K.W. Wallis, eds., Cambridge: Cambridge University Press, forthcoming.

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

JOAG-DEV, KUMAR (1984): "Measures of dependence" *Handbook of Statistics*, Vol. 4. Eds P.R. Krishnaiah and P.K. Sen.

KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *The Annals of Statistics* **18**, 191-219.

KLEIN, R.W., AND R.H. SPADY (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., AND P.L. GOZALO (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.F., (1983): "Closest Empirical Distribution Estimation," *Econometrica* **51**, 305-319.

MANSKI, C. (1994): "Analog estimation of econometric models," in *The Handbook of Econometrics, vol. IV.* eds R.F. Engle III and D.F. McFadden. North Holland.

MASRY, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

NEWY, W.K. AND K.D. WEST (1987): "A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* **55**, 703-708.

NEWEY, W. K. (1990), "Semiparametric efficiency bounds," *Journal of Applied Econometrics,* **5**, 99-135.

NIKITIN, Y. (1995): *Asymptotic efficiency of nonparametric tests.* Cambridge University Press: Cambridge.

PAKES, A., AND D. POLLARD (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., AND K.J. KOPECKY (1979): "Testing goodness of fit for the distribution of errors in regression model," *Biometrika* **66**, 1-5.

POLLARD, D. (1980): "The minimum distance method of testing," *Metrika* **23**, 43-70.

POWELL, J.L. (1994): "Estimation in semiparametric models." in *The Handbook of Econometrics, vol. IV.* eds R.F. Engle III and D.F. McFadden. North Holland.

RAO, P.V., E.F. SCHUSTER, AND R.C. LITTELL (1975): "Estimation of shift and center of symmetry based on Kolmogorov-Smirnov statistics," *The Annals of Statistics* **3**, 862-873.

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

ROSENBAUM, P.R., AND D.B. RUBIN (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. AND J.A. WELLNER (1986): *Empirical processes with applications to statistics* John Wiley, New York.

SKAUG, H.J. AND D. TJØSTHEIM, (1993): "A nonparametric test of serial independence based on the empirical distribution function." *Biometrika.* **80**, 591-602.

Stoker, T.M. (1986): "Consistent estimation of scaled coefficients," *Econometrica* **54**, 1461-1481.

Stoker, T.M. (1992): *Lectures on Semiparametric Econometrics.* Core Lecture Series: Louvain-la-Neuve

Van der Vaart, A.W. (1995): "Efficiency of infinite dimensional M-estimators," *Statistica Neerlandica* **49**, 9-30.

Van der Vaart, A.W. and J.A. Wellner (1996): *Weak convergence and empirical processes.* Springer-Verlag: Berlin.

Wellner, J.A., and Y. Zhan (1996): "Bootstrapping Z-estimators" Manuscript, Washington University.

Wermuth, N and S.L. Lauritzen (1990): "On substantive research hypotheses, conditional independence graphs and graphical chain models," (with discussion) *Journal of the Royal Statistical Society, Series B* **52**, 21-50.

Zheng, J.Z. (1994): “A consistent specification test of independence.” Manuscript, University of Texas, Austin.

Percentage Rejections of Null Hypothesis ($Z = \widetilde{\beta}_{KS}^{T} X$)

		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

TABLE 1. Size and Power with Estimated Conditioning Variable Z.

Percentage Rejections of Null Hypothesis ($Z = \beta^T X$)

		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

TABLE 2. Size and Power with Known Conditioning Variable Z.

Table 3

		n=20	*n=50*	*n=100*	*n=200*	*n=300*	*n=500*
E1	*ind*	0.4328	0.2415	0.1608	0.1048	0.0798	0.0588
	cind	0.3744	0.2388	0.1632	0.1127	0.0827	0.0598
	nlls	0.2326	0.1316	0.0955	0.0747	0.0572	0.0412
E2	*ind*	***[26]	0.5739	0.4630	0.3065	0.2574	0.2073
	cind	***	0.5368	0.4763	0.3202	0.2433	0.1795
	nlls	***	0.4526	0.3587	0.2326	0.1869	0.1439
E3	*ind*	***	0.5989	0.5706	0.5425	0.5645	0.5895
	cind	***	0.2556	0.1548	0.0898	0.0678	0.0525
	nlls	***	0.2966	0.2415	0.1307	0.1097	0.0799

[26] Could not compute