TESTS OF INDEPENDENCE IN SEPARABLE ECONOMETRIC MODELS: THEORY AND APPLICATION

DONALD J. BROWN*, RAHUL DEB†, AND MARTEN H. WEGKAMP‡

ABSTRACT. A common stochastic restriction in econometric models separable in the latent variables is the assumption of stochastic independence between the unobserved and observed exogenous variables. Both simple and composite tests of this assumption are derived from properties of independence empirical processes and the consistency of these tests is established. As an application, we simulate estimation of a random quasilinear utility function, where we apply our tests of independence.

1. INTRODUCTION

Recently, Brown and Wegkamp (2002) proposed a family of extremum estimators for semiparametric econometric models separable in the latent variables W, where $W = \rho(X, Y, \theta)$, X a random vector of observed exogenous variables, Y a random vector of observed endogenous variables, W is drawn from a fixed but unknown distribution and θ is a vector of unknown parameters. An important special case is the implicit nonlinear simultaneous equations model, where a reduced form function $Y = \rho^{-1}(X, W, \theta)$ exists. Of course, in general $Y = \rho^{-1}(X, W, \theta)$ is non-additive in W, e.g., consider the random quasilinear utility model $V(Y, W, \theta)$ proposed by Brown and Calsamiglia (2007), where $V(Y, W, \theta) = U(Y, \theta) + W \cdot Y + Y_0$. In this case the structural equations defined by $W = \rho(X, Y, \theta)$ are equivalent to the first order conditions of maximizing $V(Y, W, \theta)$ subject to the budget constraint $P \cdot Y + Y_0 = I$ (P and I stand for prices and income, respectively and Y_0 is the numeraire good). The details can be found in Section 2 below.

The principal maintained assumption in Brown and Wegkamp (2002) is the stochastic independence between W and X. In this paper we propose tests of this assumption using the elements of empirical independence processes. We present both simple tests, i.e., the null hypothesis states that for a given θ_0, $\rho(X, Y, \theta_0)$ and X are independent, as well as composite tests where the null hypothesis is that there exists some $\theta_0 \in \Theta$, the set of possible parameter values, such that X and $\rho(X, Y, \theta_0)$ are independent.

*DEPARTMENT OF ECONOMICS, YALE UNIVERSITY, BOX 208268, NEW HAVEN, CT 06520-8268
†DEPARTMENT OF ECONOMICS, YALE UNIVERSITY, BOX 208268, NEW HAVEN, CT 06520-8268
‡DEPARTMENT OF STATISTICS, FLORIDA STATE UNIVERSITY, TALLAHASSEE, FL 32306-4330
E-mail addresses: donald.brown@yale.edu, rahul.deb@yale.edu, wegkamp@stat.fsu.edu.
Date: July 20, 2008.
Key words and phrases. Cramér–von Mises distance, empirical independence processes, random utility models, semiparametric econometric models, specification test of independence.

Here we extend the analysis of Brown and Wegkamp (2002) beyond the characterization of the independence of random vectors in terms of their distribution functions. In particular, we define a family of weighted minimum mean-square distance from independence estimators in terms of characteristic or moment generating functions. The latter characterization is well suited for estimating separable econometric models with non-negative endogenous and exogenous variables. These estimates are computationally more tractable than the ones considered by Brown and Wegkamp (2002). We show asymptotic normality, and consistency of the bootstrap for our estimates and consistency of the tests for independence.

The paper is organized as follows. In Section 2 of this paper we present both the general econometric model and the example which motivated this research. Properties of empirical independence processes are reviewed in Section 3. Asymptotic properties of our estimators are derived in Section 4, and Section 5 discusses tests of independence between the observed and unobserved exogenous variables. Simulations results are in the Appendix.

2. THE ECONOMETRIC MODEL

In this paper we consider semiparametric econometric models, which are separable in the latent variables. In these models we have a triple $(X, Y, W) \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \times \mathbb{R}^{k_2}$ of random vectors, where X and W are stochastically independent. The exogenous variable $W = \rho(X, Y) \in \mathbb{R}^{k_2}$ is unobserved and drawn from a fixed but unknown distribution. In this paper we consider structural equations ρ of the parametric form $\rho(x, y) = \rho(x, y, \theta)$ for some $\theta \in \Theta \subseteq \mathbb{R}^p$.

In general, two random vectors $X \in \mathbb{R}^{k_1}$ and $W \in \mathbb{R}^{k_2}$ are independent if and only if

$$\mathbb{E}f(X)g(W) = \mathbb{E}f(X)\mathbb{E}g(W) \text{ for all } f \in \mathcal{F}_1,\ g \in \mathcal{F}_2, \tag{2.1}$$

where $\mathcal{F}_\ell$ ($\ell = 1, 2$) are

$$\mathcal{F}_\ell = \left\{ 1_{(-\infty,t]}(\cdot),\ t \in \mathbb{R}^{k_\ell} \right\}. \tag{2.2}$$

Note that each $\mathcal{F}_\ell$ in (2.2) is a universal Donsker class, indexed by a set of finite dimensional parameters $(s, t) \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$ only. This situation has been considered in Brown and Wegkamp (2002). Indeed, there are other classes $\mathcal{F}_\ell$, for which (2.1) holds as well. For example, the classes

$$\mathcal{F}_\ell = \left\{ \exp(< t, \cdot >),\ t \in \mathbb{R}^{k_\ell} \right\}, \tag{2.3}$$

or the classes

$$\mathcal{F}_\ell = \left\{ \exp(i\ < t, \cdot >),\ t \in \mathbb{R}^{k_\ell} \right\} \text{ where } i = \sqrt{-1}, \tag{2.4}$$

or the classes of all $\mathcal{C}^\infty$ functions on $\mathbb{R}^{k_\ell}$. The first two sets of classes are Donsker, provided t ranges in a bounded subset. In (2.3) we compare the joint moment generating functions (m.g.f.'s)

with the product of its marginal m.g.f.'s, and in (2.4) the comparison is based on characteristic functions. The class of all $\mathcal{C}^\infty$ functions is not finite dimensional, and therefore is uninteresting from a computational perspective. We note in passing that this formulation using expected values does not allow for comparison between the joint density of X and $\rho(X,Y,\theta)$, and the product of its marginal densities. In fact, our estimators can be viewed as moment estimators as (2.1) is a family, albeit infinite, of moment conditions.

Let $(X_1,Y_1),\cdots,(X_n,Y_n)$ be independent copies of the pair (X,Y). Motivated by the equivalence (2.1), we compare the empirical version

$$\frac{1}{n}\sum_{i=1}^n f(X_i)g(\rho(X_i,Y_i,\theta)) = \frac{1}{n}\sum_{i=1}^n f(X_i) \ \cdot \ \frac{1}{n}\sum_{i=1}^n g(\rho(X_i,Y_i,\theta)),$$

for all $f\in\mathcal{F}_1$ and $g\in\mathcal{F}_2$. Letting $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i,Y_i}$ be the empirical measure based on the sample $(X_1,Y_1),\cdots,(X_n,Y_n)$, we can write the preceding display more compactly as

$$\mathbb{P}_n f(x)g(\rho(x,y,\theta)) = \mathbb{P}_n f(x)\mathbb{P}_n g(\rho(x,y,\theta)) \text{ for all } f\in\mathcal{F}_1,\ g\in\mathcal{F}_2.$$

Observe that this amounts to comparing the joint cumulative distribution functions (c.d.f.'s) with the product of the marginal c.d.f.'s.

In order to obtain a tractable large sample theory, we consider the statistics

$$\mathbb{M}_n(\theta;\mathbb{P}_n;\mu) \equiv \int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \left\{\mathbb{P}_n f_s(x)g_t(\rho(x,y,\theta)) - \mathbb{P}_n f_s(x)\mathbb{P}_n g_t(\rho(x,y,\theta))\right\}^2 \, d\mu(s,t),$$

where μ is a c.d.f. acting as a weight function. We require that μ has a strictly positive density. In this way, we guarantee that all values s and t, that is, all functions $f_\ell\in\mathcal{F}_\ell$, are taken into account. The heuristic idea is that the unique minimizer (if it exists, conditions are necessary to guarantee this) of $\mathbb{M}_n(\theta;\mathbb{P}_n;\mu)$ should be close to the unique minimizer of

$$M(\theta;P;\mu) \equiv \int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \left\{Pf_s(x)g_t(\rho(x,y,\theta)) - Pf_s(x)Pg_t(\rho(x,y,\theta))\right\}^2 \, d\mu(s,t),$$

where P is the probability measure of the pair (X,Y). The unique minimizer of this criterion is denoted by $\theta_P = \theta(P;\mu)$. Observe that $M(\theta;P;\mu)$ is finite for all θ since μ is a distribution function, and that $M(\theta_P;P;\mu)=0$ if and only if $\rho(X,Y,\theta_P)$ and X are independent. In this case $\theta(P;\mu)$ does not depend on μ and we say that the model is identified. We can interpret $M(\theta)$ as the Cramér-von Mises distance between the actual distribution of the pair $(X,\rho(X,Y,\theta))$ and the (product) distribution of (X,W_θ), where the marginals X and W_θ are independent and W_θ has the same distribution as $\rho(X,Y,\theta)$. Observe that

$$M(\widehat{\theta}_n) \ = \ M(\theta_P) + \frac{1}{2}(\widehat{\theta}_n-\theta_P)'M''(\theta_n)(\widehat{\theta}_n-\theta_P),$$

provided $M \in \mathcal{C}^2(\Theta)$, for some θ_n between θ_P and $\widehat{\theta}_n$, where $\widehat{\theta}_n$ is the estimator of θ_P as defined in section 4. We can view the first term on the right as the approximation error due to the finite dimensional model, and the last term can be thought of as the estimation error, which has an asymptotic χ^2_p distribution (cf. Theorem 4.1 below) under some regularity assumptions. For instance, suppose that $\rho(X, Y)$ and X are independent for some ρ which we approximate by some finite series

$$\rho(x, y) \cong \rho(x, y, \theta) \equiv \sum_{i=1}^{p} \theta_i \psi_i(x, y)$$

based on some finite dimensional basis $\psi_1, \cdots, \psi_p$.

We end this section with an example of an implicit nonlinear simultaneous equations model separable in the latent variables, which motivated our research. In this example, we show that the econometric model is identified for the class of extremum estimators proposed in this paper and hence can be estimated by these methods.

Example. (A Random Quasilinear Utility Model of Consumer Demand)

We consider a consumer with a random demand function $Y(P, I, W, \theta_0)$ derived from maximizing a random utility function $V(Y, W, \theta_0)$ subject to her budget constraint $P \cdot Y + Y_0 = I$. First, the consumer draws W from a fixed and known distribution. Then nature draws $X = (P, I)$, from a fixed but unknown distribution. The main model assumption is that W and X are stochastically independent. The consumer solves the following optimization problem:

$$\text{maximize } V(y, w, \theta_0) \text{ over } y \text{ such that } p \cdot y + y_0 = I. \tag{2.5}$$

The econometrician knows $V(y, w, \theta)$ and Θ, the set of all possible values for the parameter θ, but does not know θ_0, the true value of θ. Nor does the econometrician observe W or know the distribution of W. The econometrician does observe $X = (P, I)$. The econometrician's problem is to estimate θ_0 and the distribution of W from a sequence of observations $Z_i = (X_i, Y_i)$ for $i = 1, 2, ..., n$. The structural equations for this model are simply the first-order conditions of the consumer's optimization problem. These conditions define an implicit nonlinear simultaneous equations model of the form $W = \rho(X, Y, \theta)$, where the reduced form function is the consumer's random demand function $Y(P, I, W, \theta_0)$ for the specification of $V(y, w, \theta)$ proposed by Brown and Calsamiglia (2007), i.e., $V(y, w, \theta) = U(y, \theta) + w \cdot y + y_0$. They assume that for all $\theta \in \Theta$, $U(y, \theta)$ is a smooth monotone strictly concave utility function on the positive orthant of $\mathbb{R}^k$, i.e., $DU(y, \theta) > 0$ and $D^2U(y, \theta)$ is negative definite for all y in the positive orthant of $\mathbb{R}^k$, and $W \geq 0$.

Our examples are suggested by their model, where first we consider:

$$V(y, w, \theta) = y_0 + \sum_{k=1}^{K} \theta_k g_k(y_k) + \sum_{k=1}^{K} w_k y_k, \tag{2.6}$$

where $\theta_k \in (0,1)$ and y_0 is the numeraire good. Then the first-order conditions for this optimization problem can be written as $W = \rho(X, Y, \theta)$, where $X = (P_1, P_2, \ldots, P_K)$, $Y = (Y_1, Y_2, \ldots, Y_K)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ and each g_k is smooth, strictly concave and increasing. Note that, because our utility is linear in the numeraire and we assume an interior solution, our variable X does not include the income I and Y does not include Y_0. Our first order conditions are thus

$$w_k = p_k - \theta_k \frac{\partial g_k(y_k)}{\partial y_k} \tag{2.7}$$

Because of our assumptions on g_k the above system can be solved uniquely for the random demand functions $Y_k(X, W, \theta)$. This verifies that there exists a unique reduced form $Y = \gamma(X, W, \theta)$ such that $W = \rho(X, \gamma(X, W, \theta), \theta)$. Clearly an important special case of the above form is the Cobb-Douglas function where we can take each $g_k(y_k) = \ln y_k$.

We now need to show that the above system is identified. We will use the necessary and sufficient condition for observational equivalence in an econometric model, where W and X are independent, from Matzkin (2005)[1]. If we can find an x and y such that Matzkin's identity is not satisfied for different θ, $\tilde{\theta}$ then our model is identified.

Matzkin's Identity is given by

$$\begin{aligned} &\frac{\partial \log(f_W(\rho(y,x,\theta)))}{\partial w}\left[\frac{\partial \rho(y,x,\theta)}{\partial x} - \frac{\partial \rho(y,x,\theta)}{\partial y}\left(\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1}\frac{\partial \rho(y,x,\tilde{\theta})}{\partial x}\right] \\ &\qquad + \left[\frac{\partial}{\partial x}\log\left(\left|\frac{\partial \rho(y,x,\theta)}{\partial y}\right|\right) - \frac{\partial}{\partial x}\log\left(\left|\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right|\right)\right] \\ &- \left[\left(\frac{\partial}{\partial y}\log\left(\left|\frac{\partial \rho(y,x,\theta)}{\partial y}\right|\right) - \frac{\partial}{\partial y}\log\left(\left|\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right|\right)\right)\left(\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1}\frac{\partial \rho(y,x,\tilde{\theta})}{\partial x}\right] \equiv 0 \end{aligned} \tag{2.8}$$

where $f_W(w)$ is the fixed but unknown distribution of our parameter W and the function ρ is the system of first order conditions.

The identity

$$\frac{\partial}{\partial y}\log\left(\left|\frac{\partial \rho(y,x,\theta)}{\partial y}\right|\right) = \frac{\partial}{\partial y}\log\left(\left|\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right|\right) = \begin{pmatrix} \frac{g_1'''(y_1)}{g_1''(y_1)} & \frac{g_2'''(y_2)}{g_2''(y_2)} & \cdots \end{pmatrix} \tag{2.9}$$

follows from (2.7). Hence the third term is zero. The second term is also zero, since

$$\frac{\partial}{\partial x}\log\left(\left|\frac{\partial \rho(y,x,\theta)}{\partial y}\right|\right) = \frac{\partial}{\partial x}\log\left(\left|\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right|\right) = 0$$

[1]Lenkard and Berry (2006) show that the necessary and sufficient conditions for identification, proposed by Brown (1983) and Roehrig (1988), which are widely cited in the literature and used in Brown and Wegkamp (2002), are incorrect. This paper corrects the error in Brown and Wegkamp (2002).

Simplifying the remaining term we get the following equation:

$$\frac{\partial \log(f_W(\rho(y,x,\theta)))}{\partial w}\left[\mathbf{I}-\begin{pmatrix}\frac{\theta_1}{\tilde{\theta}_1} & 0 & 0 & \dots\\ 0 & \frac{\theta_2}{\tilde{\theta}_2} & 0 & \dots\\ \vdots & \vdots & \ddots & \end{pmatrix}\right]=0 \tag{2.10}$$

Since θ and $\tilde{\theta}$ are different, there exists a k such that $\theta_k \neq \tilde{\theta}_k$. Assuming that the k^{th} component of the derivative of f_W is not identically zero implies there exist x and y such that Matzkin's identity is not satisfied. That is, the model is identified.

This is the example we compute in the Appendix. We can show that general utility functions of the form

$$V(y,w,\theta)=y_0+U(y_1,\dots,y_K,\theta_1,\dots,\theta_K)+\sum_{k=1}^{K}w_k y_k, \tag{2.11}$$

(where U is some concave, monotone function of y which does not contain linear terms in y, as they can be absorbed in the error term) are identified under the following restrictions:

Theorem 2.1. *The system $w_k = p_k - \frac{\partial U(y_1,y_2,\dots,y_K,\theta)}{\partial y_k}$ is identified if*

1. *$w=\rho(y,x,\theta)$ or equivalently $w_k = p_k - \frac{\partial U(y_1,y_2,\dots,y_K,\theta)}{\partial y_k}$ is an invertible function in y and w.*
2. *For any fixed $\theta \neq \theta'$ and $\forall c$, $\exists \overline{y}$ such that $U(\overline{y}_1,\overline{y}_2,\dots,\overline{y}_K,\theta)-U(\overline{y}_1,\overline{y}_2,\dots,\overline{y}_K,\theta')\neq c$.*
3. *$\forall y,x \qquad \frac{\partial}{\partial x}\frac{\partial \log(f_W(\rho(y,x,\theta)))}{\partial w}$ is invertible.*

Proof. The proof follows from the following observation - Matzkin's identity holds only if all derivatives of the identity w.r.t. x and y are also zero. We differentiate Matzkin's identity with respect to x and examine the individual terms.

For all x, y

$$\frac{\partial}{\partial x}\left[\left(\frac{\partial}{\partial y}\log\left(\left|\frac{\partial\rho(y,x,\theta)}{\partial y}\right|\right)-\frac{\partial}{\partial y}\log\left(\left|\frac{\partial\rho(y,x,\tilde{\theta})}{\partial y}\right|\right)\right)\left(\frac{\partial\rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1}\frac{\partial\rho(y,x,\tilde{\theta})}{\partial x}\right]=0 \tag{2.12}$$

because the term inside the square brackets does not depend on x.

Similarly the second term is independent of x, hence

$$\left[\frac{\partial}{\partial x}\log\left(\left|\frac{\partial\rho(y,x,\theta)}{\partial y}\right|\right)-\frac{\partial}{\partial x}\log\left(\left|\frac{\partial\rho(y,x,\tilde{\theta})}{\partial y}\right|\right)\right]=0 \tag{2.13}$$

This leaves the first term. Once again the term inside the square brackets does not depend on x, thus our sufficient condition for identification is that for some y, x

$$\frac{\partial}{\partial x}\left[\frac{\partial\log(f_W(\rho(y,x,\theta)))}{\partial w}\right]\left[\frac{\partial\rho(y,x,\theta)}{\partial x}-\frac{\partial\rho(y,x,\theta)}{\partial y}\left(\frac{\partial\rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1}\frac{\partial\rho(y,x,\tilde{\theta})}{\partial x}\right]\neq 0 \tag{2.14}$$

But using assumption (3) we need only consider the claim:

$$\frac{\partial \rho(y,x,\theta)}{\partial x} - \frac{\partial \rho(y,x,\theta)}{\partial y}\left(\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1}\frac{\partial \rho(y,x,\tilde{\theta})}{\partial x} \neq 0 \tag{2.15}$$

However,

$$\frac{\partial \rho(y,x,\theta)}{\partial x} = \frac{\partial \rho(y,x,\tilde{\theta})}{\partial x} = \mathbf{I}$$

Hence

$$\frac{\partial \rho(y,x,\theta)}{\partial y}\left(\frac{\partial \rho(y,x,\tilde{\theta})}{\partial y}\right)^{-1} \neq \mathbf{I} \tag{2.16}$$

for some y and x suffices for identification. Suppose not, then $(\forall y)$, $\quad U(y_1, y_2, \ldots, y_K, \theta) - U(y_1, y_2, \ldots, y_K, \tilde{\theta}) =$ a constant, as U does not contain linear terms in y; which contradicts assumption (2). □

As an example assume that the w's are independent and half normally distributed with parameter θ or

$$f_W(w) = \prod_{k=1}^{K} \frac{2\theta_k}{\pi} e^{-w_k^2\theta_k^2/\pi} \tag{2.17}$$

Then

$$\frac{\partial \log(f_W(w))}{\partial w} = \left[-\frac{2\theta_1^2 w_1}{\pi}, \ldots, -\frac{2\theta_K^2 w_K}{\pi}\right]$$

and therefore

$$\frac{\partial}{\partial x}\frac{\partial \log(f_W(\rho(y,x,\theta))}{\partial w} = \begin{pmatrix} -\frac{2\theta_1^2}{\pi} & 0 & \cdots \\ 0 & -\frac{2\theta_2^2}{\pi} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{2.18}$$

which is invertible for all y, x.

Table 1 summarizes the distributions for which our assumption (3) holds. Note that for the checked distributions it holds for all values of parameters. For the other distributions our assumption holds only if we consider a restricted subset of parameters.

TABLE 1. List of Distributions

Weibull (Exponential)	X
Gamma (Chi Square)	√
Half Normal	√
Log Normal	X
Pareto	√
Rayleigh	√
Type 2 Gumbell	X
Wald	X
Levy	X

3. INDEPENDENCE EMPIRICAL PROCESSES

Given the classes $\mathcal{F}_1$ and $\mathcal{F}_2$, we define $\mathcal{F} \equiv \mathcal{F}_1$ and

$$\mathcal{G} \equiv \{f(\rho(\cdot,\cdot,\theta)) : \ f \in \mathcal{F}_2, \, \theta \in \Theta\} = \{g_t(\rho(\cdot,\cdot,\theta)) : \ t \in \mathbb{R}^{k_2}, \, \theta \in \Theta\}.$$

As before, we denote the joint probability measure of the pair (X, Y) by P, and the empirical measure based on the sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ by $\mathbb{P}_n$. For any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, set

$$\mathbb{D}_n(f, g) \equiv \mathbb{P}_n fg - \mathbb{P}_n f \mathbb{P}_n g$$

and

$$D(f, g) \equiv Pfg - PfPg,$$

so that

$$\mathbb{M}_n(\theta) = \int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} \mathbb{D}_n^2(f_s, g_{t,\theta}) \, d\mu(s, t)$$

in the new notation. Finally, we define the independence empirical process $\mathbb{Z}_n$ indexed by $\mathcal{F} \times \mathcal{G}$ by

$$\mathbb{Z}_n(f, g) \equiv \sqrt{n}(\mathbb{D}_n - D)(f, g).$$

Observe that [cf. Van der Vaart and Wellner (1996, page 367)]

$$\begin{aligned} \mathbb{Z}_n(f,g) &= \sqrt{n}\left\{(\mathbb{P}_n - P)(fg) - (\mathbb{P}_n g)(\mathbb{P}_n - P)(f) - (Pf)(\mathbb{P}_n - P)(g)\right\} \\ &= \sqrt{n}(\mathbb{P}_n - P)((f - Pf)(g - Pg)) - \sqrt{n}(\mathbb{P}_n - P)(f)(\mathbb{P}_n - P)(g) \end{aligned} \tag{3.1}$$

The minor difference with the original formulation of independence empirical processes in Van der Vaart and Wellner (1996, Chapter 3.8) is that we consider the marginal distributions of $(X, \rho(X, Y, \theta))$ rather than (X, Y). The next result states sufficient conditions for weak convergence of the independence empirical process $\mathbb{Z}_n$ in $\ell^\infty(\mathcal{F} \times \mathcal{G})$. Let $\|P\|_{\mathcal{F}}$ be the sup-norm on $\ell^\infty(\mathcal{F})$ for any class $\mathcal{F}$, i.e. $\|P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} P|f|$.

Theorem 3.1. *Let $\mathcal{F}, \mathcal{G}$ and $\mathcal{F} \times \mathcal{G}$ be P-Donsker classes, and assume that $\|P\|_{\mathcal{F}} < \infty$ and $\|P\|_{\mathcal{G}} < \infty$. Then $\mathbb{Z}_n$ converges weakly to a tight Gaussian process Z_P in $\ell^\infty(\mathcal{F} \times \mathcal{G})$ as $n \to \infty$.*

Proof. The first term on the right in (3.1) converges weakly as $\mathcal{F} \times \mathcal{G}$ is P-Donsker. The second term in this expression is asymptotically negligible, since $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker. We invoke Slutsky's lemma to conclude the proof. □

We can also bootstrap the limiting distribution of $\mathbb{Z}_n$. Let $(X_1^*, Y_1^*), \cdots, (X_n^*, Y_n^*)$ be an i.i.d. sample from $\mathbb{P}_n$, and let $\mathbb{P}_n^*$ be the corresponding bootstrap empirical measure. Then we define the bootstrap counterpart of $\mathbb{Z}_n$ by

$$\mathbb{Z}_n^*(f, g) = \sqrt{n}(\mathbb{D}_n^* - \mathbb{D}_n)(f, g),$$

where $\mathbb{D}_n^*(f,g) = \sqrt{n}(\mathbb{P}_n^* fg - \mathbb{P}_n^* f \mathbb{P}_n^* g)$.

Theorem 3.2. *Let $\mathcal{F}, \mathcal{G}$ and $\mathcal{F} \times \mathcal{G}$ be P-Donsker classes, and assume that $\|P\|_{\mathcal{F}} < \infty$ and $\|P\|_{\mathcal{G}} < \infty$. Then $\mathbb{Z}_n^*$ converges weakly to a tight Gaussian process Z_P in $\ell^\infty(\mathcal{F} \times \mathcal{G})$, given P^∞-almost every sequence $(X_1, Y_1), (X_2, Y_2), \cdots$, as $n \to \infty$.*

Proof. We first note that

$$\mathbb{Z}_n^*(f,g) = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)((f - \mathbb{P}_n f)(g - \mathbb{P}_n g)) - \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)(f)(\mathbb{P}_n^* - \mathbb{P}_n)(g)$$

and recall that $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ converges weakly [cf. Theorem 3.9.11 in Van der Vaart and Wellner (1996)]. □

4. Estimation of θ_P

4.1. **A general result.** Given P-Donsker classes $\mathcal{F} = \{f_s : s \in \mathbb{R}^{k_1}\}$ and $\mathcal{G} = \{g_{t,\theta} : t \in \mathbb{R}^{k_2}, \theta \in \Theta\}$ and a c.d.f. μ, we can define

$$\mathbb{M}_n(\theta) = \int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} \mathbb{D}_n^2(f_s, g_{t,\theta})\, d\mu(s,t)$$

and

$$M(\theta) = \int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} D^2(f_s, g_{t,\theta})\, d\mu(s,t).$$

We propose to estimate $\theta_P = \theta(P;\mu)$ by $\widehat{\theta}_n = \theta(\mathbb{P}_n;\mu)$ which minimizes the random criterion function $\mathbb{M}_n$ over Θ. Then, provided M has a unique, well-separated minimum at an interior point θ_P of Θ, it follows immediately by the weak convergence of $\mathbb{Z}_n$ (cf. Theorem 3.1) and Theorem 5.9 in Van der Vaart (1998, page 46) that

$$\widehat{\theta}_n \in \arg\min \mathbb{M}_n(\theta) \to \arg\min M(\theta) = \theta_P,$$

in probability. We will now show the asymptotic normality of the standardized distribution $\sqrt{n}(\widehat{\theta}_n - \theta_P)$.

We impose the following set of assumptions:

(A1) M has a unique global, well-separated minimum at θ_P in the interior of Θ and $M(\theta;P) \in \mathcal{C}^2(\Theta)$ and $M''(\theta_P;P)$ is non-degenerate.

(A2) $D(f_s, g_{t,\theta})$ is differentiable with respect to θ for all s,t, and its derivative satisfies

$$\left|\dot{D}(s,t,\theta) - \dot{D}(s,t,\theta_P)\right| \le |\theta - \theta_P| \Delta(s,t)$$

for some $\Delta \in \mathcal{L}^2(\mu)$.

(A3) $\sup_{s,t} P|f_s g_{t,\theta} - f_s g_{t,\theta_P}|^2 \to 0$ as $\theta \to \theta_P$.

(A4) The map $\rho(\cdot,\cdot,\theta)$ is continuously differentiable in θ.

(A5) The classes $\mathcal{F}, \mathcal{G}$ and $\mathcal{F}\times\mathcal{G}$ are P-Donsker.

We have the following result:

Theorem 4.1. *Assume (A1) – (A5). Then, $\sqrt{n}(\widehat{\theta}_n - \theta_P)$ has a non-degenerate Gaussian limiting distribution, as $n\to\infty$.*

Proof. The result follows from Theorem 3.2 in Wegkamp (1999, page 48). We need to verify the following three conditions:

(i) $\widehat{\theta}_n \to \theta_P$ in probability.

(ii) M has a non-singular second derivative at θ_P.

(iii) $\sqrt{n}(\mathbb{M}_n - M)(\theta)$ is stochastically differentiable at θ_P.

As noted above, (i) follows from general theory. Condition (ii) is subsumed in (A1). It remains to establish (iii). Let the symbol $\rightsquigarrow$ denote weak convergence in general metric spaces. (A3) implies that, for all s,t,

$$\mathbb{Z}_n(f_s, g_{t,\theta}) - \mathbb{Z}_n(f_s, g_{t,\theta_P}) \rightsquigarrow 0 \text{ as } \theta \rightsquigarrow \theta_P,\ n\to\infty.$$

Consequently, by the continuous mapping theorem

$$\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \left[\mathbb{Z}_n^2(f_s, g_{t,\theta}) - \mathbb{Z}_n^2(f_s, g_{t,\theta_P})\right]\, d\mu(s,t) \rightsquigarrow 0$$

as $\theta \rightsquigarrow \theta_P,\ n\to\infty$. (A2), (A3) and the continuous mapping theorem yield also that

$$\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} [D(f_s, g_{t,\theta})\mathbb{Z}_n(f_s, g_{t,\theta}) - D(f_s, g_{t,\theta_P})\mathbb{Z}_n(f_s, g_{t,\theta_P}) - (\theta-\theta_P)'\dot{D}(s,t,\theta_P)\mathbb{Z}_n(f_s, g_{t,\theta_P})\Big]\, d\mu(s,t) \rightsquigarrow 0$$

as $\theta \rightsquigarrow \theta_P,\ n\to\infty$. Conclude that

$$\begin{aligned}
&\sqrt{n}(\mathbb{M}_n - M)(\theta)\\
&= \int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \mathbb{Z}_n^2(f_s, g_{t,\theta})\, d\mu(s,t) + 2\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} D(f_s, g_{t,\theta})\mathbb{Z}_n(f_s, g_{t,\theta})\, d\mu(s,t)\\
&= \int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \mathbb{Z}_n^2(f_s, g_{t,\theta_P})\, d\mu(s,t) + 2\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} D(f_s, g_{t,\theta_P})\mathbb{Z}_n(f_s, g_{t,\theta_P})\, d\mu(s,t) +\\
&\quad +2(\theta-\theta_P)'\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \dot{D}(s,t,\theta_P)\mathbb{Z}_n(f_s, g_{t,\theta_P})\, d\mu(s,t) + o_p(1+\|\theta-\theta_P\|)\\
&= \sqrt{n}(\mathbb{M}_n - M)(\theta_P) + 2(\theta-\theta_P)'\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \dot{D}(s,t,\theta_P)\mathbb{Z}_n(f_s, g_{t,\theta_P})\, d\mu(s,t)\\
&\qquad +o_p(1+\|\theta-\theta_P\|),
\end{aligned}$$

which establishes (iii). □

In fact, the asymptotic linear expansion

$$\begin{aligned}(4.1)\qquad &\sqrt{n}(\widehat{\theta}_n - \theta_P) \\ &= -2\left[M''(\theta_P)\right]^{-1} \int\!\!\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \dot{D}(s,t,\theta_P)\mathbb{Z}_n(f_s, g_{t,\theta_P})\, d\mu(s,t) + o_p(1)\end{aligned}$$

holds. This expression coincides with the one derived in Brown and Wegkamp (2002, page 2045).

In addition, the conditional distribution of the bootstrap estimators $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$ has the same limit in probability. Here $\widehat{\theta}_n^*$ is based on i.i.d. sampling from $\mathbb{P}_n$, see Section 3. The proof of this assertion follows from similar arguments as Theorem 4.1, see Brown and Wegkamp (2002, pages 2046 - 2048) and is for this reason omitted.

Theorem 4.2. *Assume (A1) – (A5). Then,*

$$\sup_{h\in BL_1}\left|\mathbb{E}^* h\left(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)\right) - \mathbb{E}h\left(\sqrt{n}(\widehat{\theta} - \theta_P)\right)\right| \to 0$$

in probability, as $n \to \infty$. Here $\mathbb{E}^*$ means the conditional expectation given the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and BL_1 is the space of all functions $h : \mathbb{R} \to [-1, 1]$ with $|h(x) - h(y)| \le |x - y|$ for all $x, y \in \mathbb{R}$.

We apply the developed theory to the special cases where $\mathcal{F}$ and $\mathcal{G}$ are indicator functions of half-spaces $(-\infty, \cdot]$ or exponential functions $\exp(t'x)$.

4.2. **Estimators based on the distribution functions.** For every $s \in \mathbb{R}^{k_1}, t \in \mathbb{R}^{k_2}$ and $\theta \in \Theta$, define the empirical distribution functions

$$\mathbb{F}_n(s) = \frac{1}{n}\sum_{i=1}^n \{X_i \le s\},\ \mathbb{G}_{n\theta}(t) = \frac{1}{n}\sum_{i=1}^n \{\rho(X_i, Y_i, \theta) \le t\} \text{ and}$$

$$\mathbb{H}_{n\theta}(s,t) = \frac{1}{n}\sum_{i=1}^n \{X_i \le s,\ \rho(X_i, Y_i, \theta) \le t\}.$$

The criterion function $\mathbb{M}_n$ becomes in this case

$$\mathbb{M}_n(\theta) \equiv M(\theta; \mathbb{P}_n; \mu) = \int\!\!\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} \left\{\mathbb{F}_n(s)\mathbb{G}_{n\theta}(t) - \mathbb{H}_{n\theta}(s,t)\right\}^2 d\mu(s,t).$$

This is essentially the empirical criterion proposed by Brown and Wegkamp (2002). We obtain its theoretical counterpart $M(\theta) = M(\theta; P; \mu)$ by replacing the empirical distributions $\mathbb{F}_n, \mathbb{G}_{n\theta}$ and $\mathbb{H}_{n\theta}$ by the population distributions.

Assumption (A3) is verified if $\rho(x, y, \theta)$ is Lipschitz in θ, see Brown and Wegkamp (2002, page 2043, proof of Lemma 3). Assumptions (A2), and (A4) follow from smoothness assumptions on

$\rho(\cdot,\cdot,\theta)$ and P. For (A1), we refer to Brown and Wegkamp (2002, Theorem 3, page 2038). We now show how to verify (A5).

We define the sets

$$A_{\theta,t} = \left\{(x,y) \in \mathbb{R}^{k_1+k_2} : \rho(x,y,\theta) \leq t\right\},\ t \in \mathbb{R}^{k_2},\ \theta \in \Theta,$$

and the associated collection

$$\mathcal{A} = \left\{A_{\theta,t} : \theta \in \Theta,\ t \in \mathbb{R}^{k_2}\right\}.$$

Note that $\mathcal{G}$ corresponds to the indicators I_A of sets $A \in \mathcal{A}$, and $\mathcal{F}$ corresponds to corresponds to the indicators I_B of sets $B \in \mathcal{B} \equiv \left\{\{x \in \mathbb{R}^{k_1} : x \leq t\},\ t \in \mathbb{R}^{k_1}\right\}$, which is universally Donsker. Condition (A5) becomes in this specific setting

(A5′) The classes of sets $\mathcal{A}$, $\mathcal{A} \times \mathcal{B}$ are P-Donsker.

Sufficient conditions for $\mathcal{A}$ to be P-Donsker are either smoothness of $\rho(x,y,\theta)$ (with respect to x and y, not θ) or that ρ ranges over a finite dimensional vector space. See Brown and Wegkamp (2002) for a discussion.

Example 4.3. Let $\{\rho(\cdot,\cdot,\theta),\ \theta \in \Theta\}$ be a subset of a finite dimensional vector space. Then both $\mathcal{A}$ and $\mathcal{B}$ are VC-classes, see Van der Vaart and Wellner (1996, section 2.1.1) for a description and many examples of VC classes. From Van der Vaart and Wellner (1996, page 147) it follows that and $\mathcal{A} \times \mathcal{B}$, the product of two VC-classes, is again VC. Hence $\mathcal{A}, \mathcal{B}$ and $\mathcal{A} \times \mathcal{B}$ are universally Donsker.

Example 4.4. Let the support of (X,Y) be a bounded, convex subset of $\mathbb{R}^{k_1+k_2}$ with non-empty interior, and, for each θ, $\rho(x,y,\theta)$ have uniformly bounded (by K) partial derivatives through order $\beta = \lfloor\alpha\rfloor$, and the derivatives of order β satisfy a uniform Hölder condition of order $\alpha - \beta$, and with Lipschitz constant bounded by K. For a complete description of the space $C_K^{\alpha}[X \times \mathcal{Y}]$, we refer to Van der Vaart and Wellner (1996), page 154. If $\alpha > d$ and P has a bounded density, then $\mathcal{A}$ and $\mathcal{A} \times \mathcal{B}$ are P-Donsker. To see why, we first notice that $\mathcal{A} \times \mathcal{B}$ has constant envelope 1, and that

$$Q|fg - \tilde{f}\tilde{g}|^2 \leq 2Q|f - \tilde{f}|^2 + 2Q|g - \tilde{g}|^2,$$

and that $f_L \leq f \leq f_U$ and $g_L \leq g \leq g_U$ implies $f_L g_L \leq fg \leq f_U g_U$. Hence

$$\mathcal{N}_B(2\varepsilon, L^2(Q), \mathcal{F} \times \mathcal{G}) \leq \mathcal{N}_B(\varepsilon, L^2(Q), \mathcal{F})\mathcal{N}_B(\varepsilon, L^2(Q), \mathcal{G}),$$

where $\mathcal{N}_B(\varepsilon, L^2(Q), \mathcal{F})$ is the ε-bracketing number of the set $\mathcal{F}$ with respect to the $L^2(Q)$ norm. Since $\log \mathcal{N}_B(\varepsilon, L^2(Q), \mathcal{B}) \lesssim \log(1/\varepsilon)$, the bound on the bracketing numbers in Corollary 2.7.3 in Van der Vaart and Wellner (1996) on $\mathcal{A}$ implies that $\mathcal{A} \times \mathcal{B}$ is P-Donsker.

4.3. **Estimators based on the moment generating functions.** Assume that X and $\rho(X, Y, \theta)$ are bounded, so that in particular their m.g.f.'s exist. For every $s \in \mathbb{R}^{k_1}, t \in \mathbb{R}^{k_2}$ and $\theta \in \Theta$, define the empirical m.g.f.'s

$$\phi_n(s) = \frac{1}{n}\sum_{i=1}^{n} \exp(< s, X_i >),\ \psi_{n\theta}(t) = \frac{1}{n}\sum_{i=1}^{n} \exp\{< t, \rho(X_i, Y_i, \theta) >\}$$

and $$\zeta_{n\theta}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \exp\{< s, X_i > + < t, \rho(X_i, Y_i, \theta) >\}.$$

Let $k = k_1 + k_2$, and $C_\varepsilon > 0$ be such that $\mu[-C_\varepsilon, C_\varepsilon]^k = 1 - \varepsilon$. In this case we take the random criterion function $\mathbb{M}_n$

$$\mathbb{M}_n(\theta) \equiv M(\theta; \mathbb{P}_n; \mu) = \int\int_{[-C_\varepsilon, +C_\varepsilon]^k} \{\phi_n(s)\psi_{n\theta}(t) - \zeta_{n\theta}(s,t)\}^2\, d\mu(s,t).$$

This setting corresponds to

$$\mathcal{F}_\varepsilon = \{\exp(< t, x >),\ t \in [-C_\varepsilon, +C_\varepsilon]^{k_1}\}$$

and

$$\mathcal{G}_\varepsilon = \{\exp(< t, \rho(x, y, \theta) >),\ t \in [-C_\varepsilon, +C_\varepsilon]^{k_2},\ \theta \in \Theta\}.$$

Van de Geer (2000, Lemma 2.5) shows that the box $[-C_\varepsilon, C_\varepsilon]^{k_1}$ can be covered by $\left(4C_\varepsilon\delta^{-1} + 1\right)^{k_1}$ many δ-balls in $\mathbb{R}^{k_1}$. Since

$$\mathbb{P}_n\left|\exp(< s, X >) - \exp(< t, X >)\right|^2 \lesssim \mathbb{P}_n\|X\|^2\|s - t\|^2,$$

it follows from the above covering number calculation that the uniform entropy condition (cf. Van der Vaart and Wellner (1996, page 127) is met, and consequently the class $\mathcal{F}_\varepsilon$ is P-Donsker. Restricting the integration over $[-C_\varepsilon, C_\varepsilon]^k$, which has μ-probability equal to $1-\varepsilon$, forces the function M to be within ε of the original criterion function, since

$$\left|\int\int_{\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}} D^2(s,t)\, d\mu(s,t) - \int\int_{[-C_\varepsilon, C_\varepsilon]^k} D^2(s,t)\, d\mu(s,t)\right| \le \mu\left(\mathbb{R}^k \setminus [-C_\varepsilon, C_\varepsilon]^k\right) \le \varepsilon.$$

Assumption (A1) will force the corresponding unique minimizers to be close as well. Notice that $\mathcal{F}_\varepsilon$ is not a Donsker class if we take $C_\varepsilon = +\infty$. $\mathcal{G}_\varepsilon$ will be a P-Donsker class if $\{\rho(\cdot,\cdot,\theta) :\ \theta \in \Theta\}$ has this property. This is a consequence of the fact that the Donsker property of a class is preserved under Lipschitz transformations, see Theorem 2.10.6 in Van der Vaart and Wellner (1996, page 192).

Assumptions (A2) and (A3) follow from (A4), smoothness of $\rho(\cdot,\cdot,\theta)$, and the smoothness of the exponential function. Again, for (A1) we refer to Brown and Wegkamp (2002, Theorem 3, page 2038).

5. TESTS OF INDEPENDENCE

Our null hypothesis is that $\rho(X,Y)$ and X are independent for some specified structural equation $\rho(x,y) = \rho(x,y,\theta_0)$. Following the discussion in Van der Vaart and Wellner (Chapter 3.8, 1996), a reasonable test is based on the Kolmogorov-Smirnov type statistic

$$\mathbb{K}_n \equiv \sup_{s,t} \sqrt{n} \left| \mathbb{P}_n f_s(x) g_t(\rho(x,y)) - \mathbb{P}_n f_s(x) \mathbb{P}_n g_t(\rho(x,y)) \right|.$$

Provided $\mathcal{F} \times \mathcal{G}$, $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, the limiting distribution of $\mathbb{K}_n$ under the null is known and can be bootstrapped (see Van der Vaart and Wellner, 1996, pages 367 -369).

Alternatively, we propose tests based on the criteria $\mathbb{M}_n$ defined above. Given observations (X_i, Y_i) we can compute $(X_i, W_i) \equiv (X_i, \rho(X_i, Y_i))$. Next, we note that

$$\begin{aligned}
\mathbb{Z}_n(f,g) &\equiv \sqrt{n}(\mathbb{D}_n - D)(f, g \circ \rho) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ f(X_i) g(W_i) - \mathbb{E} f(X_i) \mathbb{E} g(W_i) \} \\
&\qquad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ f(X_i) - \mathbb{E} f(X_i) \} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ g(W_i) - \mathbb{E} g(W_i) \}
\end{aligned}$$

is the same independence empirical process discussed in Van der Vaart and Wellner (1996, Section 3.8). Theorem 3.8.1 in Van der Vaart and Wellner (1996, page 368) states that $\mathbb{Z}_n(f,g)$ converges weakly to a tight Gaussian process Z_H in $\mathcal{F} \times \mathcal{G}$. Consequently, under the null hypothesis

$$n\mathbb{M}_n = \int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} \left\{ \mathbb{Z}_n(f_s, g_t) + \sqrt{n} D(f_s, g_t) \right\}^2 d\mu(s,t) \tag{5.1}$$

converges weakly to

$$\int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} Z_H^2(f_s, g_t) \, d\mu(s,t) \tag{5.2}$$

by the continuous mapping theorem. However,

$$n\mathbb{M}_n \to +\infty \text{ (in probability)}$$

under any alternative $P_{X,W}$ with

$$\int D^2(f_s, g_t) \, d\mu(s,t) > 0,$$

which, provided $\mathcal{F}$ and $\mathcal{G}$ are generating classes as in (2.2), (2.3) or (2.4), is equivalent with X and $W = \rho(X,Y)$ are dependent. This implies that the power of the test converges to one under each alternative, that is, the test is consistent.

In lieu of the normal limiting distribution (5.2), we can also rely on the following bootstrap approximation for the distribution of the test statistic under the null. Let P^X and P^W be the probability measures of X and W, respectively, with empirical counterparts denoted by $\mathbb{P}_n^X$ and $\mathbb{P}_n^W$, respectively. Under the null hypothesis, the joint distribution of (X, W) is the product measure $P^X \times P^W$, and a natural estimate for the joint distribution of (X, W) is $\mathbb{P}_n^X \times \mathbb{P}_n^W$. In order to imitate the independence structure under the null hypothesis, we sample from the product measure $\mathbb{P}_n^X \times \mathbb{P}_n^W$. Let $(X_1^*, W_1^*), \cdots, (X_n^*, W_n^*)$ be the resulting i.i.d. sample from $\mathbb{P}_n^X \times \mathbb{P}_n^W$, and define

$$\frac{1}{\sqrt{n}}\mathbb{Z}_n^*(f,g) = \frac{1}{n}\sum_{i=1}^n f(X_i^*)g(W_i^*) - \frac{1}{n}\sum_{j=1}^n f(X_j^*)\frac{1}{n}\sum_{k=1}^n g(W_k^*).$$

Since the bootstrap sample is taken from $\mathbb{P}_n^X \times \mathbb{P}_n^W$ and not the ordinary empirical measure $\mathbb{P}_n$, the variables $\mathbb{Z}_n^*(f,g)$ have conditional mean zero. Van der Vaart and Wellner (1996, Theorem 3.8.3) obtain sufficient conditions ($\mathcal{F} \times \mathcal{G}$ satisfies the uniform entropy condition (cf. Van der Vaart and Wellner (1996, page 171) for envelope functions F, G and $\mathcal{F} \times G$ in $L_2(P)$) that $\mathbb{Z}_n^*(f,g)$ converges weakly to $Z_{P^X \times P^W}$ almost surely. Since this limit coincides with the limiting distribution of $\mathbb{Z}_n$ under the null hypothesis, $n\mathbb{M}_n^* = \int\{\mathbb{Z}_n^*(f_s, g_t)\}^2 \, d\mu(s,t)$ can be used to approximate the finite sample distribution of $n\mathbb{M}_n$ in a consistent manner (under the null hypothesis). Note that this procedure is model based as the resampling is done from the estimated model under the null hypothesis.

In addition, we present a specification test where the composite null hypothesis is the existence of a $\theta_0 \in \Theta$ such that X and $\rho(X, Y, \theta_0)$ are independent. We base the test on the statistic $T_n \equiv n\mathbb{M}_n(\widehat{\theta})$, and we show that T_n equals in distribution approximately $n\mathbb{M}_n(\theta_0)$ plus some drift due to $\widehat{\theta}_n$. In general the limiting distribution depends on θ_0, but it can be bootstrapped.

Theorem 5.1. *Assume (A1) – (A5) and $M(\theta_0) = 0$. Then*

$$n\mathbb{M}_n(\widehat{\theta}_n) - \int \left[\mathbb{Z}_n(f_s, g_{t,\theta_0}) + \sqrt{n}(\hat{\theta}_n - \theta_0)'\dot{D}(s,t,\theta_0)\right]^2 d\mu(s,t) \rightsquigarrow 0, \tag{5.3}$$

and

$$\int \left[\mathbb{Z}_n(f_s, g_{t,\theta_0}) + \sqrt{n}(\hat{\theta}_n - \theta_0)'\dot{D}(s,t,\theta_0)\right]^2 d\mu(s,t) \tag{5.4}$$

is asymptotically tight.

Proof. First, we note that $\mathbb{Z}_n(f_s, g_{t,\theta})$ is stochastically differentiable in θ for all s, t by Condition (A3). An application of the functional continuous mapping theorem yields that

$$\int\int_{\mathbb{R}^{k_1} \times \mathbb{R}^{k_2}} \left[\mathbb{Z}_n(f_s, g_{t,\theta'}) - \mathbb{Z}_n(f_s, g_{t,\theta})\right]^2 d\mu(s,t) \rightsquigarrow 0, \text{ for } \theta' \rightsquigarrow \theta.$$

The stochastic equicontinuity, weak convergence of $\widehat{\theta}_n$ and (A2) yield the following expansion of $\mathbb{M}_n(\hat{\theta}_n)$:

$$\begin{aligned} n\mathbb{M}_n(\widehat{\theta}_n) &= \int n\mathbb{D}_n^2(f_s, g_{t,\widehat{\theta}_n})\, d\mu(s,t) \\ &= \int \left[\left\{\mathbb{Z}_n(f_s, g_{t,\widehat{\theta}_n}) - \mathbb{Z}_n(f_s, g_{t,\theta_0})\right\} + \mathbb{Z}_n(f_s, g_{t,\theta_0}) + \sqrt{n}D(f_s, g_{t,\widehat{\theta}_n})\right]^2 d\mu(s,t) \\ &= \int \left[\mathbb{Z}_n(f_s, g_{t,\theta_0}) + \sqrt{n}(\widehat{\theta}_n - \theta_0)'\dot{D}(s,t,\theta_0)\right]^2 d\mu(s,t) + o_p(1). \end{aligned}$$

Since $\widehat{\theta}_n$ is asymptotically linear [cf. (4.1)], the vector

$$\left(\mathbb{Z}_n(f_s, g_{t,\theta_0})\ ,\ \sqrt{n}(\widehat{\theta}_n - \theta_0)\right)$$

converges weakly to a tight limit. Claim (5.3) and (5.4) follow from the continuous mapping theorem. □

Notice that we may write under the null hypothesis

$$\mathbb{T}_n = n\mathbb{M}_n(\widehat{\theta}) = \int \left[\sqrt{n}\{\mathbb{D}_n(f_s, g_{t,\widehat{\theta}}) - D(f_s, g_{t,\theta_0})\}\right]^2 d\mu(s,t).$$

Motivated by this expression, we propose the following bootstrap procedure. Let $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ be an i.i.d. bootstrap sample from $\mathbb{P}_n$. The distribution of $\mathbb{T}_n$ can be approximated by

$$\mathbb{T}_n^* = \int \left[\sqrt{n}\{\mathbb{D}_n^*(f_s, g_{t,\theta^*}) - \mathbb{D}_n(f_s, g_{t,\widehat{\theta}})\}\right]^2 d\mu(s,t).$$

Theorem 5.2. *Assume (A1) – (A5) and $M(\theta_0) = 0$. Then*

$$\sup_{h \in BL_1} \left|\mathbb{E}^* h\left(\mathbb{T}_n^*\right) - \mathbb{E}h\left(\mathbb{T}_n\right)\right| \to 0 \tag{5.5}$$

in probability, as $n \to \infty$.

Proof. By Theorems 3.1, 3.2 and 4.2, the functional continuous mapping theorem and condition (A2), we find

$$\begin{aligned} \mathbb{T}_n^* &= \int \left[\mathbb{Z}_n^*(f_s, g_{t,\theta^*}) + \mathbb{Z}_n(f_s, g_{t,\theta^*}) - \mathbb{Z}_n(f_s, g_{t,\widehat{\theta}}) + \sqrt{n}\{D(f_s, g_{t,\theta^*}) - D(f_s, g_{t,\widehat{\theta}})\}\right]^2 d\mu(s,t) \\ &= \int \left[\mathbb{Z}_n^*(f_s, g_{t,\theta^*}) + \sqrt{n}(\theta^* - \widehat{\theta})\dot{D}(s,t,\theta_0)\right]^2 d\mu(s,t) + o_p(1) \\ &= \int \left[\mathbb{Z}_n^*(f_s, g_{t,\theta_0}) + \sqrt{n}(\theta^* - \widehat{\theta})\dot{D}(s,t,\theta_0)\right]^2 d\mu(s,t) + o_p(1) \end{aligned}$$

Finally invoke Theorems 3.1, 3.2 and 4.2 and the conclusion follows easily. □

This result says that the distribution of $\mathbb{T}_n^*$ can be used to approximate the finite sample distribution of our test statistics $\mathbb{T}_n$. Again, we note that the power of the test converges to one, as $n\mathbb{M}_n(\widehat{\theta}_n) \to +\infty$ under any alternative $P_{X,W}$ with $\int D^2(f_s, g_t)\, d\mu(s,t) > 0$, that is, $P^X P^W \neq P^{X,W}$.

Remark: A model based bootstrap as described in the introduction would resample $X_1^*, \ldots, X_n^*$ from $X_1, \ldots, X_n$ and $W_1^*, \ldots, W_n^*$ from $\widehat{W}_1 = \rho(X_1, Y_1, \widehat{\theta}), \ldots, \widehat{W}_n = \rho(X_n, Y_n, \widehat{\theta})$. Let $\widehat{\mathbb{D}}_n^*$ be the bootstrap equivalent of $\mathbb{D}_n$ based on this bootstrap sample. The bootstrap equivalent of $\mathbb{T}_n$, namely $n \int [\widehat{\mathbb{D}}_n^*(f_s, g_t)]^2$ has the same limiting distribution as $\mathbb{T}_n$ following section 2.8.3 in Van der Vaart and Wellner (1996, pp 173-174).

APPENDIX A. SIMULATION RESULTS : ESTIMATING A ONE PARAMETER MODEL

We simulate a data set for the simple one-dimensional parameter model

$$U(y_0, y_1, y_2) = \theta \log y_1 + (1 - \theta) \log y_2 + W_1 y_1 + W_2 y_2 + y_0$$

subject to

$$p_1 y_1 + p_2 y_2 + y_0 = I$$

where $0 \leq \theta \leq 1$ is the parameter. We set the true parameter $\theta_0 = .4$. The first order conditions are

$$w_1 = p_1 - \frac{\theta}{y_1}$$
$$w_2 = p_2 - \frac{1 - \theta}{y_2}$$

We use the estimator based on the moment generating functions (section 4.3) to compute our estimate $\hat{\theta}$. Because of the exponential form of the mgf's, the random criterion function has a simple exponential form which is computationally inexpensive to evaluate, since the integral is not explicitly computed. To minimize the random criterion function, we use a simple grid search. Below we give the expression for $\mathbb{M}_n(\theta)$ for the k dimensional version of the above 2 dimensional model. In this section we set $k = 2$ whereas in Section B of the appendix we set $k = 4$.

$$\mathbb{M}_n(\theta) = \int_{[-2,+2]^k} \int \left\{ \phi_n(s) \psi_{n\theta}(t) - \zeta_{n\theta}(s,t) \right\}^2 \, d\mu(s,t).$$

Plugging in values for ϕ , ψ and ζ, and setting $W_i(\theta) = \rho(X_i, Y_i, \theta)$, we get

$$\begin{aligned}
\mathbb{M}_n(\theta) &= \int\int_{[-2,+2]^k} \left\{ \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \prod_{l=1}^{k} e^{s^l X_{i_1}^l} e^{t^l W_{i_2}^l(\theta)} - \frac{1}{n} \sum_{j_1=1}^{n} \prod_{l=1}^{k} e^{s^l X_{j_1}^l} e^{t^l W_{j_1}^l(\theta)} \right\}^2 d\mu(s,t) \\
&= \int\int_{[-2,+2]^k} \left\{ \frac{1}{n^4} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} \prod_{l=1}^{k} e^{s^l X_{i_1}^l} e^{t^l W_{i_2}^l(\theta)} e^{s^l X_{i_3}^l} e^{t^l W_{i_4}^l(\theta)} \right. \\
&\quad + \frac{1}{n^2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \prod_{l=1}^{k} e^{s^l X_{j_1}^l} e^{t^l W_{j_1}^l(\theta)} e^{s^l X_{j_2}^l} e^{t^l W_{j_2}^l(\theta)} \\
&\quad \left. - \frac{2}{n^3} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j_1=1}^{n} \prod_{l=1}^{k} e^{s^l X_{i_1}^l} e^{t^l W_{i_2}^l(\theta)} e^{s^l X_{j_1}^l} e^{t^l W_{j_1}^l(\theta)} \right\} d\mu(s,t) \\
&= \frac{1}{n^4} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} \int\int_{[-2,+2]^k} \left\{ \prod_{l=1}^{k} e^{s^l (X_{i_1}^l + X_{i_3}^l)} e^{t^l (W_{i_2}^l(\theta) + W_{i_4}^l(\theta))} \right\} d\mu(s,t) \\
&\quad + \frac{1}{n^2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \int\int_{[-2,+2]^k} \left\{ \prod_{l=1}^{k} e^{s^l (X_{j_1}^l + X_{j_2}^l)} e^{t^l (W_{j_1}^l(\theta) + W_{j_2}^l(\theta))} \right\} d\mu(s,t) \\
&\quad - \frac{2}{n^3} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j_1=1}^{n} \int\int_{[-2,+2]^k} \left\{ \prod_{l=1}^{k} e^{s^l (X_{i_1}^l + X_{j_1}^l)} e^{t^l (W_{i_2}^l(\theta) + W_{j_1}^l(\theta))} \right\} d\mu(s,t)
\end{aligned}$$

We take for μ the uniform distribution on $[-2, +2]^k$. This makes computing the integral computationally inexpensive since it is simply the integral of exponentials.

In this simulation we draw parameter W from a uniform distribution. This seems to contradict Theorem 2.1 which requires the distribution for W to be smooth, but we can approximate a uniform distribution arbitrarily closely by a smooth distribution.

We run two simulations:

(1) The first simulation corresponds to Theorem 4.1. It demonstrates that our estimates are normally distributed around the true value of the parameter we are estimating.
(2) The second simulation corresponds to Theorem 4.2. It demonstrates that the bootstrap estimates are normally distributed around the estimated value of the parameter.

In the first simulation, we randomly sample p_1, p_2 from $\mathbb{U}[1, 2]$ and we sample w_1, w_2 from $\mathbb{U}[0, 1]$. The true value θ_0 allows us to calculate the corresponding consumer demands. We pick a 100 price and corresponding consumer demands. The supports of the uniform distributions are chosen to ensure an interior solution. We resample values of p and w a 1000 times and calculate

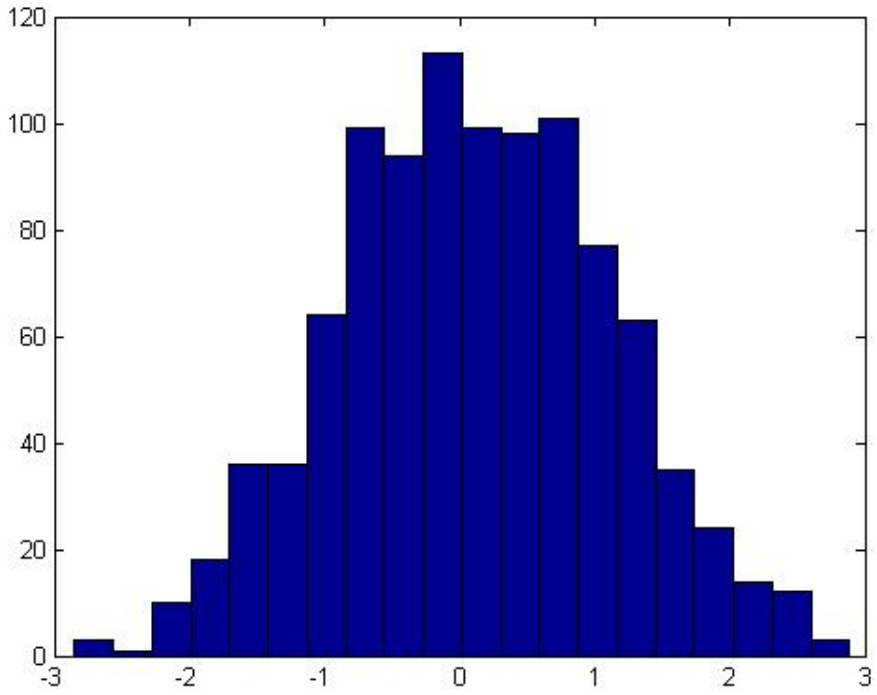

FIGURE 1. Resampled histogram

the estimated $\hat{\theta}$ each time. Recall that $\theta_0 = .4$. We obtain the following results

$$\overline{\theta} = \frac{1}{1000}\sum_{n=1}^{1000}\hat{\theta}_n = 0.406116 \qquad \text{std} = \sqrt{\frac{1}{999}\sum_{n=1}^{1000}(\hat{\theta}_n - \overline{\theta})^2} = 0.051247$$

$$\text{mse} = \frac{1}{1000}\sum_{n=1}^{1000}(\hat{\theta}_n - \theta)^2 = 0.00266103$$

and plot the standardized histogram centered around θ_0 (Figure 1)

In our second simulation, we sample p, W (100 points) only once from the same uniform distributions. We estimate our parameter θ, bootstrap the sample a thousand times and obtain the

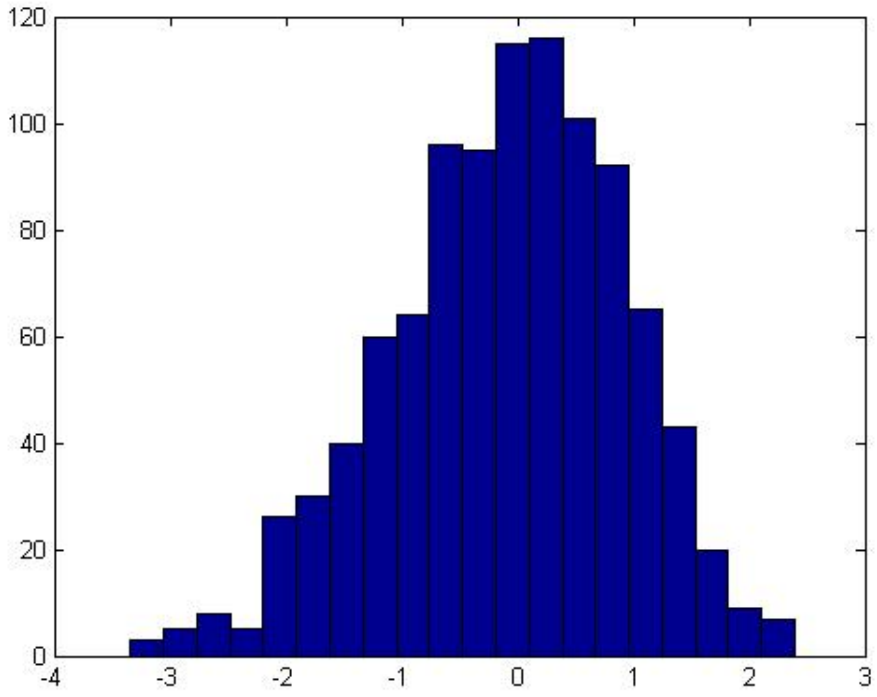

FIGURE 2. Bootstrapped histogram

following results

$$\hat{\theta} = \text{estimated value without bootstrapping} = 0.405$$

$$\overline{\theta} = \frac{1}{1000}\sum_{b=1}^{1000}\hat{\theta}_b = \text{mean of bootstrapped estimates} = 0.400463$$

$$\text{std} = \sqrt{\frac{1}{999}\sum_{b=1}^{1000}(\hat{\theta}_b - \overline{\theta})^2} = 0.051756$$

$$\text{mse} = \frac{1}{1000}\sum_{b=1}^{1000}(\hat{\theta}_b - \hat{\theta})^2 = 0.00269659$$

And finally we plot the standardized histogram of the bootstrapped estimates centered around our estimate $\hat{\theta}$ to get Figure 2.

APPENDIX B. SIMULATION RESULTS : ESTIMATING A THREE PARAMETER MODEL

The purpose of this section is to show that the estimator works well even with as little as 100 data points. The estimator naturally works better with larger data sets. Due to computational complexity we estimate only up to 3 parameters. We also estimate up to second decimal place

accuracy for the same reason. It is possible to make the computations relatively inexpensive using the fast Laplace transform and more sophisticated optimization techniques.

We simulate a data set as in Appendix A with multiple parameters. Hence, our multidimensional model is

$$U(y_0, y_1, y_2, y_3, y_4) = \sum_{i=1}^{4} \theta_i \log y_i + \sum_{i=1}^{4} W_i y_i + y_0$$

subject to

$$p_1 y_1 + p_2 y_2 + p_3 y_3 + p_4 y_4 + y_0 = I$$

where $0 \leq \theta_i \leq 1$, $i \in \{1, 2, 3\}$ are the parameters we must estimate, since $\theta_4 = 1 - (\theta_1 + \theta_2 + \theta_3)$. The true parameters values are: $\theta_1 = .2$, $\theta_2 = .3$ and $\theta_3 = .4$.

First, we sample p_1, p_2, p_3, p_4 from $\mathbb{U}[1, 2]$ and w_1, w_2, w_3, w_4 from $\mathbb{U}[0, 1]$. Then we choose 100 price vectors and the corresponding consumer demands. The supports of the uniform distributions are chosen to ensure an interior solution. We repeat this process a 100 times and obtain the following mean estimates for our parameters:

$$\widehat{\theta}_1 = .228, \qquad \widehat{\theta}_2 = .262, \qquad \widehat{\theta}_3 = .410$$

APPENDIX C. SIMULATION RESULTS (TESTS FOR INDEPENDENCE)

We will simulate both the simple and composite null hypotheses outlined in Section 5. Below is an outline of the simulations

(1) The first set of simulations are for the null hypothesis that $\rho(X, Y)$ and X are independent for some specified structural equation $\rho(x, y) = \rho(x, y, \theta_0)$. We test both when the null hypothesis is true and also some local alternatives when the null is false.
(2) The second set of simulations are for the composite null hypothesis i.e. the existence of a $\theta_0 \in \Theta$ such that X and $\rho(X, Y, \theta_0)$ are independent. We test separately the independent case as well as cases where there is perfect and slight correlations.

In the first simulations we test the two parameter model used in appendix A. In particular we fix the true value $\theta = .4$. We generate X and W independently and then back out the Y's using the true value of θ, and then test for independence of $\rho(X, Y)$ and X for specified structural equations $\rho(x, y) = \rho(x, y, \theta_0)$, where we allow θ_0 to take the true value $.4$ as well as local alternatives .3 and .5. We approximate the distribution of the test statistic by bootstrapping the sample and ordering the values of the test statistic from the bootstrapped distribution in ascending order, where we take the 95^{th} percentile value as the critical value. The null is rejected if the value of the test statistic

from the original sample is greater than this critical value and repeat this for various sample sizes. Our results are summarized in the following tables

TABLE 2. Independence Test Results $\theta_0 = .4$ (True $\theta = .4$)

Sample Size	No of Simulations	No of Accepts
500	1000	892
1000	1000	989

TABLE 3. Independence Test Results $\theta_0 = .3$ (True $\theta = .4$)

Sample Size	No of Simulations	No of Rejects
500	1000	924
1000	1000	984

TABLE 4. Independence Test Results $\theta_0 = .5$ (True $\theta = .4$)

Sample Size	No of Simulations	No of Rejects
500	1000	887
1000	1000	971

The second simulations requires the generation of independent random vectors X and W and testing their independence. We generate independent data, dependent data as well as correlated data and test for various sample sizes. The details of the tests are summarized in the tables below.

We first test for dependence

TABLE 5. Independence Test Results ($w_k = p_k - 1$)

Sample Size	No of Simulations	No of Rejects
500	1000	997
1000	1000	1000

For correlated data we generate multivariate (X_1, X_2, W_1, W_2) with mean $(0, 0, 0, 0)$ and covariance matrix

$$\begin{array}{c} \\ X_1 \\ X_2 \\ W_1 \\ W_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & W_1 & W_2 \end{array} \\ \begin{bmatrix} 1 & 0 & \sigma & 0 \\ 0 & 1 & 0 & \sigma \\ \sigma & 0 & 1 & 0 \\ 0 & \sigma & 0 & 1 \end{bmatrix} \end{array}$$

We test for different values of the correlation σ and report the results below

TABLE 6. Independence Test Results ($\sigma = .5$)

Sample Size	No of Simulations	No of Rejects
500	1000	986
1000	1000	998

TABLE 7. Independence Test Results ($\sigma = .1$)

Sample Size	No of Simulations	No of Rejects
500	1000	541
1000	1000	712

TABLE 8. Independence Test Results ($\sigma = 0$)

Sample Size	No of Simulations	No of Accepts
500	1000	935
1000	1000	972

Remark: The parameter estimation procedure works well for small samples sizes of $n = 100, 200$ but the tests for independence are not effective for these values of n.

REFERENCES

[1] Benkard, C. and Berry, S. (2006). On The Nonparametric Identification Of Nonlinear Simultaneous Equations Models: Comment On Brown (1983) And Roehrig (1988). *Econometrica*, 74(5), 1429-1440.

[2] Brown, B. W. (1983). The Identification Problem in Systems Nonlinear in the Variables. *Econometrica*, 51, 175-196.

[3] Brown, D. and Calsamiglia, C. (2007). The Nonparametric Approach to Applied Welfare Analysis. *Economic Theory*, 31, 183-188.

[4] Brown, D. and Wegkamp, M. (2002). Weighted Minimum Mean-Square Distance from Independence Estimation. *Econometrica*, 70(5), 2035-2051.

[5] Matzkin, R.L. (2005). Identification in Nonparametric Simultaneous Equations. *Mimeo*. Northwestern University.

[6] Roehrig, C. S. (1988). Conditions for Identification in Nonparametric and Parametric Models. *Econometrica*, 56, 433-447.

[7] Vaart, A. van der (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[8] Vaart, A. van der and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

[9] Wegkamp, M. (1999) *Entropy Methods in Statistical Estimation.* CWI-tract 125, Amsterdam.