Weighted Minimum Mean-Square Distance from Independence Estimation

By

Donald J. Brown and Marten H. Wegkamp

January 2001

COWLES FOUNDATION DISCUSSION PAPER NO. 1288

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

WEIGHTED MINIMUM MEAN-SQUARE DISTANCE FROM INDEPENDENCE ESTIMATION

DONALD J. BROWN AND MARTEN H. WEGKAMP

Abstract. In this paper we introduce a family of semi-parametric estimators, suggested by Manski's minimum mean-square distance from independence estimator. We establish the strong consistency, asymptotic normality and consistency of bootstrap estimates of the sampling distribution and the asymptotic variance of these estimators.

1. Introduction

Manski (1983) introduced minimum mean–square distance from independence estimation of semi-parametric econometric models separable in the unobserved exogenous variables ε, i.e., $\varepsilon = \rho(X, Y, \theta)$ where X is a random vector of observed exogenous variables, Y is a random vector of observed endogenous variables, θ is a vector of unknown parameters, ε is drawn from a fixed but unknown distribution, and ε is stochastically independent of X. An important special case is the implicit nonlinear simultaneous equations model, where a reduced form function $Y = \rho^{-1}(X, \varepsilon, \theta)$ exists. This model is a central topic of this paper. Manski proved strong consistency of his estimator, but was unable to derive the first-order asymptotic distribution. The criterion function for Manski's estimator is the mean-square distance between the joint empirical cumulative distribution function of ε and X and the product of its marginal cumulative distribution functions. The criterion function for our estimator is the mean-square distance between the joint empirical cumulative distribution of ε and X and the product of its marginal cumulative distribution functions, weighted by a probability measure on the product space of ε and x.

These weighted minimum mean-square distance from independence estimators offer tractable procedures for those applications where the econometrician assumes that ε is stochastically

Date: December 2000.

The authors thank Don Andrews, Steve Berry and David Pollard for their helpful comments and remarks.

independent of X and ε is drawn from a fixed but unknown distribution. Such is the case for the continuous random utility model proposed by Brown and Matzkin (1998).

Comparing minimum mean-square distance from independence estimation to maximum likelihood estimation and GMM, Manski observes that the joint maximization of the likelihood over the product of the parameter space Θ and the family of possible distributions of ε is often computationally intractable. Also GMM by imposing only a finite number of moment restrictions cannot use all of the information contained in the independence assumption. In contrast, weighted minimum mean-square distance from independence estimation only requires that we minimize over the finite dimensional parameter space Θ. If $\hat{\theta}$ is a consistent estimate of θ_0, the true parameter value, then we give sufficient conditions on the mapping $\theta \mapsto \rho(x, y, \theta)$ for consistent estimation of the true distribution of ε.

Since weighted minimum mean-square distance from independence is an extremum estimator, identification in the econometrics literature means that the asymptotic criterion function has a unique minimum at the truth — see Newey and McFadden (1994). Of course, identification in the usual statistical sense means that the distribution of data at θ_0 differs from that at any other value of θ. As Newey and McFadden note, statistical identification is necessary but not in general sufficient for an extremum estimator to have a unique minimum (maximum) at the truth. An important exception is maximum likelihood estimation. In particular, if $\rho(X, Y, \theta)$ is nonlinear in θ then as they point out "primitive conditions for identification (existence of a unique global minimum) become quite difficult." In practice, global GMM identification for nonlinear simultaneous equations models is simply assumed by the econometrician.

A striking and important feature of weighted minimum mean-square distance from independence estimation is that the standard statistical notion of identification is sufficient for uniqueness of the global minimum. For implicit nonlinear simultaneous equations models, Brown (1983) and more generally Roehrig (1989) have given sufficient conditions on the primitive $\rho(X, Y, \theta)$ for statistical identification, if ε is assumed to be stochastically independent of X.

Unfortunately, Manski's regularity conditions for strong consistency of minimum mean-square distance from independence estimation — see the Corollary on page 314 of Manski

(1983) — are unattractive in at least two respects. First, he simply assumes the existence of a unique minimum, a high-level assumption for which he provides no sufficient conditions on the model's primitives. Second and more importantly — given our prior discussion of the Brown and Roehrig results — is his assumption that the sets $S(\nu,\eta,\theta) = \{(x,y) : (x,y) \in S, x < \nu, \rho(x,y,\theta) < \eta\}$, where S is a compact convex subset of Euclidean space and $\rho(x,y,\theta)$ is continuous on $S \times \Theta$, are convex with boundaries having measure zero with respect to the true fixed but unknown distribution of ε. A technical assumption difficult to verify in practice. The latter assumption is crucial for his consistency argument, since it allows him to invoke a uniform law of large numbers due to Rao (1962).

In this paper, we introduce the family of weighted minimum mean-distance from independence estimators which are computationally tractable and identified. Moreover our regularity conditions for consistency and asymptotic normality are satisfied in many applications. That is, we show if $\rho(x,y,\theta)$ is sufficiently smooth in (x,y,θ) and the possible distributions of ε have sufficiently smooth densities then our estimators are strongly consistent and asymptotically normal. Also, we prove under these assumptions that bootstrap estimates of the sampling distribution and the asymptotic variance are also consistent.

As conjectured by Manski, the main tools of our analysis are techniques derived from the theory of empirical processes, necessitated by our non-smooth criterion function. For instance, see Pakes and Pollard (1989) for a lucid discussion and econometric application of empirical process theory. Their paper and this paper are related both in method and economic motivation. An application of their results is the estimation of a discrete random choice model and an intended application of our results is the estimation of the continuous random choice model of Brown and Matzkin.

Two significant differences between our paper and the paper of Pakes and Pollard are that our estimator is an extremum estimator, i.e., we minimize a non-smooth random criterion function and their estimator is a Z-estimator, i.e., they approximately solve a family of possibly non-smooth random equations. More importantly, Theorem 3.2 in Wegkamp (1999, page 40) employed here subsumes as special cases: M-estimation, Cramer–Von Mises estimation, regression and weighted minimum mean-square distance from independence estimation. See

Wegkamp (1995), Andrews (1997, 1999) and Pollard (forthcoming) for similar results. Additional references on empirical process theory and their statistical applications can be found in Dudley (1999), Pollard (1984, 1985) and Van der Vaart and Wellner (1996). Econometric applications can also be found in Andrews (1994).

This paper is organized as follows. We discuss in turn identification, consistency, asymptotic normality and resampling. In the final section of the paper we present simulation results on estimating the random utility model proposed by Brown and Matzkin. An appendix contains a sufficient condition for identifying nonlinear simultaneous equations models with multiple equilibria.

2. Identification of Minimum Distance from Independence Estimators

Statistical identification is a necessary condition for the asymptotic criterion function of extremum estimators to have a unique global minimum (maximum). In our case it is also sufficient. Compactness of the parameter space Θ and the continuity of the asymptotic criterion function imply that the optimum is well-separated.

We assume that Θ is a compact subset of an Euclidean space. Moreover, let $\mathcal{X}$ be a subset of $\mathbb{R}^L$, $\mathcal{Y}$ be a subset of $\mathbb{R}^K$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a subset of R^{L+K}. Then $Z = (X, Y)$ is a random vector taking values in $\mathcal{Z}$. For all $\theta \in \Theta$, $\rho(\cdot, \theta)$ is a mapping from $\mathcal{Z}$ into $\mathbb{R}^K$.

We define minimum distance from independence estimators, as extremum estimators where the asymptotic criterion function $d(\cdot, \cdot)$ is a metric on the space of joint cumulative distribution functions (c.d.f.'s) of (X, ε), where ε takes values in $\mathbb{R}^K$. If $H(x, \varepsilon)$ is the joint c.d.f. of (X, ε) and $F(x)$, $G(\varepsilon)$ the associated marginal c.d.f.'s then $d(H(x, \varepsilon), F(x)G(\varepsilon)) = 0$ iff X and ε are stochastically independent. In Manski (1983) d is the mean-square distance, in Brown and Matzkin (1998) d is the metric on the space of c.d.f.'s induced by the Prohorov metric on the space of measures and in this paper d is the weighted mean-square distance. Our discussion of implicit nonlinear simultaneous equations models follows the expositions of Brown (1983), Roehrig (1988) and Brown–Matzkin (1998).

A structure S is an ordered pair $\langle \rho(X, Y, \theta), H(x, \varepsilon) \rangle$. The structural equations are defined as $\varepsilon = \rho(X, Y, \theta)$. Our model consists of all structures S that satisfy the following assumptions:

(I.1) $\exists!$ *reduced form* $Y = \gamma(X, \varepsilon, \theta)$ *such that*

$$\varepsilon \equiv \rho(X, \gamma(X, \varepsilon, \theta), \theta).$$

(I.2) *The matrix* $\partial\rho/\partial y$ *has full rank* a.e.

(I.3) $H(x, \varepsilon) = F(x)G(\varepsilon)$ *for all* $(x, \varepsilon) \in \mathcal{X} \times \mathcal{Y}$, *i.e.,* X *and* ε *are stochastically independent.*

(I.4) $H(x, \varepsilon)$ *is absolutely continuous, with respect to Lebesgue measure, with positive density.*

The following notation will prove useful:

(i) $H_\theta(s, t) \equiv P\{X \leq s,\ \rho(X, \gamma(X, \varepsilon, \theta_0), \theta) \leq t\}$.

(ii) $F(x)$ and $G_\theta(\varepsilon)$ are the associated marginal c.d.f.'s of $H_\theta(x, \varepsilon)$.

(iii) $M(\theta) \equiv d(H_\theta(x, \varepsilon)\,,\, F(x)G_\theta(\varepsilon))$.

Given assumption I.1 each structure generates a joint c.d.f. of (X, Y). Our maintained assumption is that the observed c.d.f. of $Z = (X, Y)$ is generated by some structure $S_0 = \langle \rho(X, Y, \theta_0), H_{\theta_0}(x, \varepsilon) \rangle$ in our model.

Brown (1983, pp. 180, 181) in his seminal paper on identification proved the fundamental result, Theorem 2.1, for the special case of semiparametric implicit simultaneous equations models that are only nonlinear in the variables. Subsequently, Roehrig (1988) extended Brown's analysis to nonparametric and semiparametric implicit nonlinear simultaneous equations models. Recently Brown and Matzkin (1998) derived a consequence of Theorem 2.1, Theorem 2.2, which we use in the identification of a random utility model in the final section of the paper. The structural equations in this example are nonlinear in both the variables and the parameters.

Theorem 2.1 (Roehrig (1988, Lemma 3.3, p. 437))**.**

$$H_\theta(x, \varepsilon) = F(x)G_\theta(\varepsilon) \ \ a.e. \quad \textit{iff} \quad \frac{\partial \rho(x, \gamma(x, \varepsilon, \theta_0), \theta)}{\partial x} = 0 \ \ a.e.$$

Theorem 2.2 (Brown–Matzkin (1998, Theorem 1′, p. 6))**.** *If* $\partial\rho(x, \gamma(x, \varepsilon, \theta_0), \theta)/\partial x = 0$ *a.e., then*

$$\frac{\partial \gamma(x, \varepsilon, \theta_0)}{\partial x} = \frac{\partial \gamma(x, \varepsilon, \theta)}{\partial x} \ \ a.e.$$

The identification condition for minimum distance from independence estimators is an immediate consequence of Theorem 2.1.

Theorem 2.3. [1] *θ_0 is the unique global minimum of $M(\theta)$ iff $\forall \theta \neq \theta_0$, $\exists(\overline{x},\overline{\varepsilon})$ such that*

$$\left.\frac{\partial \rho(x,\gamma(x,\varepsilon,\theta_0),\theta)}{\partial x}\right|_{(\overline{x},\overline{\varepsilon})} \neq 0.$$

The following result is well known, but necessary for our proof of consistency. First we recall the definition of a well-separated minimum.

Definition. θ_0 is a well-separated minimum of $M(\theta)$ if $\inf_{\{\theta\in\Theta: m(\theta,\theta_0)\geq\varepsilon\}} M(\theta) > M(\theta_0)$, where m is a metric on Θ.

Theorem 2.4 (Newey–McFadden (1994, Theorem 2.1, p. 2121))**.** *If Θ is compact, $M(\theta)$ is continuous on Θ and θ_0 is the unique global minimum of $M(\theta)$, then θ_0 is a well-separated minimum of M.*

3. Consistency and Asymptotic Normality of Weighted Minimum Mean-Square Distance from Independence Estimators

Our main model assumption in this and the next section is that $\rho(X,Y,\theta)$ is independent of X if and only if $\theta = \theta_0$. Based on independent observations $Z_1,\ldots,Z_n$, we will now construct an estimate of θ_0, and establish its limiting sampling distribution under a set of regularity conditions. The independence assumption between X and $\rho(X,Y,\theta_0)$ is equivalent with

$$H_\theta(x,\varepsilon) = F(x)G_\theta(\varepsilon)\ \forall(x,\varepsilon)\in\mathcal{X}\times\mathcal{Y} \iff \theta=\theta_0.$$

As a consequence, for any bounded measure μ on $\mathcal{Z}$, the criterion function

$$M(\theta) = \int \left[H_\theta(x,\varepsilon) - F(x)G_\theta(\varepsilon)\right]^2\, d\mu(x,\varepsilon)$$

[1] In the appendix, this sufficient condition for identification is extended to nonlinear simultaneous equations models with multiple equilibria.

is minimized at $\theta = \theta_0$. Motivated by this observation, we propose to minimize the empirical counterpart of $M(\theta)$,

$$M_n(\theta) = \int \left[H_{n\theta}(x,\varepsilon) - F_n(x)G_{n\theta}(\varepsilon)\right]^2 \, d\mu(x,\varepsilon)$$

over $\theta \in \Theta$. Here F_n, $G_{n\theta}$ and $H_{n\theta}$ are the empirical c.d.f.'s associated with F, G_θ and H_θ, respectively, based on the observed data $Z_1, \ldots, Z_n$. For instance,

$$G_{n\theta}(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} I\{\rho(Z_i,\theta) \le \varepsilon\}.$$

The resulting minimum is denoted by $\hat{\theta}$ [2], which satisfies

$$M_n(\hat{\theta}) \le M_n(\theta) \text{ for all } \theta \in \Theta.$$

Next we describe the set of regularity assumptions, followed by a brief discussion.

(A.1) The parameter space Θ is compact, and θ_0 is an interior point.

(A.2) The model is identified.

(A.3) The collection of functions $\{\rho(\cdot,\cdot,\theta) : \theta \in \Theta\}$ is either

- a subset of a finite dimensional space *or*
- each coordinate of the mapping $(x,y) \mapsto \rho(x,y,\theta)$ is an element of $C_K^\alpha[\mathcal{X} \times \mathcal{Y}]$ [3], $K > 0$ and $\mathcal{X}$ and $\mathcal{Y}$ compact, for all $\theta \in \Theta$. In this case we require that $H(x,\varepsilon)$ has a bounded density.

(A.4) The random vector ε has a continuous c.d.f. G.

(A.5) The mapping $\theta \mapsto \rho(x,y,\theta)$ is Lipschitz at θ_0 uniformly in $x \in \mathcal{X}, y \in \mathcal{Y}$.

(A.6) The mapping $\theta \mapsto D_\theta(x,\varepsilon) \equiv H_\theta(x,\varepsilon) - F(x)G_\theta(\varepsilon)$ is differentiable at θ_0 in $L_2(\mu)$, that is, there exists $\Delta \in L_2(\mu)$ such that

$$\lim_{\|\theta-\theta_0\|\to 0} \int \left(\frac{D_\theta(x,\varepsilon) - (\theta-\theta_0)'\Delta(x,\varepsilon)}{\|\theta-\theta_0\|}\right)^2 d\mu(x,\varepsilon) = 0.$$

(A.7) The mapping $\theta \mapsto M(\theta)$ has a positive definite second derivative matrix V at θ_0.

[2] We will assume without loss of generality that a minimum exists, since otherwise we can always take any $\theta \in \Theta$ which minimizes M_n within a constant $1/n^2$ without affecting the results.

[3] Each coordinate mapping must have uniformly bounded (by K) partial derivates through order $\beta = \lfloor \alpha \rfloor$, and the derivatives of order β will 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 Dudley (1999), page 252, or Van der Vaart & Wellner (1996), page 154.

The first assumption A.1 is a standard condition in the literature. The second assumption A.2 was the main issue in the previous section, where we derived a necessary and sufficient condition under the assumptions I.1 – I.4, i.e., Theorem 2.3. Concerning the third assumption A.3, we observe that the standard compactness conditions for spaces of smooth functions used in economic theory, e.g. see Mas-Colell (1985), Section K in Chapter 1, are sufficient to guarantee the third assumption. We note in passing that only certain metric entropy properties of $\{\rho(\cdot,\theta) : \theta \in \Theta\}$ are needed to conduct our proof; conditions on these spaces other than A.3 may also give the desired metric entropy property.

Assumptions A.5 and A.6 are implied by pointwise smoothness of the mapping $\theta \mapsto \rho(\cdot,\theta)$. It should be noted that A.6 is weaker than pointwise differentiability (cf. Van der Vaart (1998), Lemma 7.6, page 95, and Chapter 4 in Pollard (forthcoming)).

We are now in the position to state our main results.

Theorem 3.1 (consistency)**.** *Under assumptions A1, A2, A3 and continuity of M at θ_0 (which is implied by A.7), $\hat{\theta}$ is strongly consistent, i.e., $\hat{\theta} \xrightarrow{a.s.} \theta_0$.*

Corollary 3.2. *Under the assumptions of Theorem 3.1 and A.5, $H_{n\hat{\theta}}(x,\varepsilon) \xrightarrow{a.s.} H(x,\varepsilon)$ for all $(x,\varepsilon) \in \mathcal{X} \times \mathcal{Y}$.*

Theorem 3.3 (asymptotic normality)**.** *Under the regularity assumptions A.1 – A.3 described above, $\sqrt{n}(\hat{\theta}-\theta_0)$ converges to a mean zero, non-degenerate multivariate normal distribution. The limiting covariance matrix Σ is $4V^{-1}WV^{-1}$, where V is defined in A.7 and*

$$\begin{aligned} W \;=\; & \int\int \Delta(x,\varepsilon)\Delta'(\overline{x},\overline{\varepsilon})\,[F(x)F(\overline{x})G(\min(\varepsilon,\overline{\varepsilon})) + F(\min(x,\overline{x}))G(\varepsilon)G(\overline{\varepsilon}) + \\ & + H(\min(x,\overline{x}),\min(\varepsilon,\overline{\varepsilon})) - 3H(x,\varepsilon)H(\overline{x},\overline{\varepsilon})]\; d\mu(x,\varepsilon)\, d\mu(\overline{x},\overline{\varepsilon}). \end{aligned}$$

The minimum between two vectors x and $\overline{x}$ should be understood coordinatewise.

Proofs. The aspects of empirical process theory employed in our proofs are generalizations of two fundamental theorems in probability theory: The Glivenko-Cantelli theorem and Donsker's theorem. These results are the paradymatic examples of uniform laws of large numbers and uniform central limit theorems. To illustrate our methodlogy, we present a brief discussion of uniform laws of large numbers. See Van der Vaart (1998), Chapter 19 for a

lucid introduction to this field. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common probability measure P. From the first n observations we construct the empirical measure P_n. This measure puts mass $1/n$ at each observation X_i, $i = 1, \ldots, n$. Given a measurable function f we write $P_n f$ for the expectation of f under the empirical measure, i.e. $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and Pf for the expectation of f under P, i.e. $Pf = \int f\, dP$. By the strong law of large numbers, $P_n f$ converges almost surely to Pf for every f for which Pf is defined. The classical Glivenko-Cantelli theorem states that this convergence is uniform over the class of indicator functions $1_{(-\infty,r]}$, $r \in \mathbb{R}$. A class $\mathcal{F}$ of measurable functions is called P-Glivenko-Cantelli if $\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |P_n f - Pf| \xrightarrow{\text{a.s.}} 0$. In a similar vein, a class of measurable functions is called P-Donsker if the empirical process $\sqrt{n}(P_n - P)(f)$ indexed by $f \in \mathcal{F}$ convergences weakly to a tight Gaussian process in $l^\infty(\mathcal{F})$. By Slutsky's lemma, a P-Donsker class is P-Glivenko-Cantelli; the converse is not necessarily true.

Important classes of functions which are P-Donsker are the so called VC classes (named after Vapnik and Chervonenkis), discussed at length in Pakes and Pollard (1989). VC-classes are determined by purely combinatorial arguments, and they are small in terms of metric entropy. That the classical Glivenko-Cantelli theorem is a special case of the abstract uniform law of large numbers follows from the observation that the class of half sets $(-\infty, r]$, $r \in \mathbb{R}$ constitutes a VC-class.

Many proofs of central limit theorems reduce to a careful application of Taylor's theorem. In these applications the empirical criterion function is written as a second order Taylor's expansion around θ_0, the true value of the parameter θ. However, since our criterion function M_n is not pointwise differentiable, we need a weaker notion of differentiability, to wit, stochastic differentiability of $\sqrt{n}(M_n - M)$, a notion introduced by Pollard (1985).

We now turn to the proofs of Theorem 3.1, Corollary 3.2 and Theorem 3.3. First we need some additional notation and results. Define the sets

$$A_{\theta,y} = \{z \in \mathcal{Z} : \rho(z,\theta) \le y\} \text{ and } B_x = \{t \in \mathcal{X} : t \le x\},$$

and the associated collections

$$\mathcal{A} = \{A_{\theta,y} : \theta \in \Theta, y \in \mathcal{Y}\},\ \mathcal{B} = \{B_x : x \in \mathcal{X}\} \text{ and } \mathcal{C} = \{A \cap (B \times \mathcal{Y}) : A \in \mathcal{A}, B \in \mathcal{B}\}.$$

Let P be the probability measure of $Z = (X, Y)$, and let P_n be its empirical measure based on $Z_1, \ldots, Z_n$, which puts mass $1/n$ at each observation. Recall the definition of

$$D_\theta(x, \varepsilon) = H_\theta(x, \varepsilon) - F(x)G_\theta(\varepsilon),$$

and define further

$$D_{n\theta}(x, \varepsilon) = H_{n\theta}(x, \varepsilon) - F_n(x)G_{n\theta}(\varepsilon).$$

Observe that

$$\begin{aligned} &D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon) \\ &= \{H_{n\theta}(x, \varepsilon) - H_\theta(x, \varepsilon)\} + F_n(x)\{G_\theta(\varepsilon) - G_{n\theta}(\varepsilon)\} + G_\theta(\varepsilon)\{F(x) - F_n(x)\} \end{aligned}$$

and consequently,

$$\begin{aligned} &\sup_{\theta\in\Theta, x\in\mathcal{X}, \varepsilon\in\mathcal{Y}} |D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon)| \\ &\leq \sup_{A\in\mathcal{A}} |(P_n - P)(A)| + \sup_{C\in\mathcal{C}} |(P_n - P)(C)| + \sup_{B\in\mathcal{B}} |(P_n - P)(B)|. \end{aligned}$$

For any measure Q on $\mathcal{Z}$, any class of functions $\mathcal{F} \subset L_2(Q)$ and any positive number δ, let $N(\delta, \mathcal{F}, L_2(Q))$ be the δ-covering number (possibly infinite) of the class $\mathcal{F}$ with respect to the $L_2(Q)$ metric, that is, the number of closed balls with radius δ in $L_2(Q)$ needed to cover $\mathcal{F}$. The δ-bracketing number is denoted by $N_B(\delta, \mathcal{F}, L_2(Q))$, i.e. the number of δ-brackets needed to cover $\mathcal{F}$. A δ-bracket of a function $f \in \mathcal{F}$ is the pair (f_L, f_U) such that $f_L \leq f \leq f_U$ and $\int |f_u - f_L|^2 \, dQ \leq \delta$.

Lemma 3.4. *Suppose that $\{\rho(\cdot, \theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space. Then*

$$\sup_{Q \text{ discrete}} N(\delta, \mathcal{F}, L_2(Q)) \leq C\delta^{-V}$$

for $\mathcal{F} = \mathcal{A}, \mathcal{B}, \mathcal{C}$ and $V > 1$.

Proof. The statement is well known for the class of sets $\mathcal{B}$. For the class $\mathcal{A}$ we argue as follows. Let d be the dimension of $\mathcal{Y}$. Since $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)'$ and $\rho(z, \theta) = \left(\rho^{(1)}(z, \theta), \ldots, \rho^{(d)}(z, \theta)\right)'$,

we can write $A_{\theta,\varepsilon}$ as an intersection, v.i.z.,

$$A_{\theta,\varepsilon} = A^{(1)}_{\theta,\varepsilon} \cap \cdots \cap A^{(d)}_{\theta,\varepsilon},$$

where

$$A^{(i)}_{\theta,\varepsilon} = \left\{ z \in \mathcal{Z} : \rho^{(i)}(z,\theta) \leq \varepsilon_i \right\}.$$

It is well known that $\{A^{(i)}_{\theta,\varepsilon} : \theta \in \Theta, \varepsilon \in \mathcal{Y}\}$ is a VC-class of sets if $\{\rho^{(i)}(z,\theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space, see for instance Van der Vaart & Wellner (1996), Lemma 2.6.15, page 146. Hence all $\left\{ A^{(i)}_{\theta,\varepsilon} : \theta \in \Theta, \varepsilon \in \mathcal{Y} \right\}$ are VC classes for $i = 1, \ldots, d$. The VC property is closed under taking intersections (cf. Van der Vaart & Wellner (1996), Lemma 2.6.17, page 147), so that $\mathcal{A}$ forms a VC-class of sets, and hence the claim for $\mathcal{A}$ follows by Theorem 2.6.4 in Van der Vaart & Wellner (1996), page 136.

The claim for $\mathcal{C}$ follows since $\mathcal{A}$ and $\mathcal{B} \times \mathcal{Y}$ are VC, and hence $\mathcal{C} = \{A \cap (B \times \mathcal{Y}),\ A \in \mathcal{A}, B \in \mathcal{B}\}$ is VC as shown in Van der Vaart & Wellner (1996), Lemma 2.6.17, page 147. □

Lemma 3.5. *Suppose that $\mathcal{Z}$ is compact with nonempty interior and that each coordinate mapping of $\rho(z,\theta) \in C^\alpha_K(\mathcal{Z})$ for all $\theta \in \Theta$. Then the collections $\mathcal{F} = \mathcal{A}, \mathcal{B}, \mathcal{C}$ all satisfy*

$$\log N_B(\delta, \mathcal{F}, L_2(P)) \leq C\delta^{-V}$$

for some $V = 2D/\alpha < 2$ and for all probability measures P with an uniformly bounded density and $\alpha > D$.

Proof. Corollary 2.7.3, page 157 in Van der Vaart and Wellner (1996) bounds the entropy of bracketing of the collection of subgraphs of $C^\alpha_K(\mathcal{Z})$, and the result for $\mathcal{A}$ is immediate. The condition $\alpha > D$ is needed to ensure that $V < 2$. It is easy to show that the δ bracketing number of $\mathcal{C}$ is the product of the δ bracketing numbers of $\mathcal{A}$ and $\mathcal{B}$ as they are bounded classes. Taking the logarithm entails the desired result. □

In particular, the entropy bounds above show that the classes $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are P-Donsker classes under assumption A.3.

Now we are in the position to prove the consistency result, Theorem 3.1.

Proof of Theorem 3.1. First, observe that

$$\begin{aligned}
&\sup_{\theta\in\Theta} |(M_n - M)(\theta)| \\
&\leq 4 \sup_{\theta\in\Theta} \int |D_{n\theta} - D_\theta| \, d\mu \\
&\leq 4 \sup_{\theta\in\Theta, x\in\mathcal{X}, \varepsilon\in\mathcal{Y}} |(D_{n\theta} - D_\theta)(x,\varepsilon)| \\
&\leq 4 \sup_{A\in\mathcal{A}} |(P_n - P)A| + 4 \sup_{B\in\mathcal{B}} |(P_n - P)B| + 4 \sup_{C\in\mathcal{C}} |(P_n - P)C| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

since $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are Glivenko-Cantelli classes. Hence with probability one,

$$M_n(\hat{\theta}) \leq M_n(\theta_0) + o(1/n) = M(\theta_0) + o(1/n) \leq M(\hat{\theta}) + o(1/n).$$

The compactness assumption on Θ and the identifiability assumption yield that $M(\theta)$ has a unique, well-separated minimum (at θ_0) (cf. Theorem 2.4). □

Observe that sufficient (high level) conditions are

(i) $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are P-Glivenko-Cantelli classes.

(ii) $M(\theta)$ has a unique, well-separated minimum.

Lemma 3.6. *Under A.3, A.4, A.5, the process $\sqrt{n}(D_{n\theta} - D_\theta)(x,\varepsilon)$ is stochastically equicontinuous at θ_0 with respect to the Euclidean metric on Θ, for all $x \in \mathcal{X}$ and $\varepsilon \in \mathcal{Y}$, i.e.*

$$\sqrt{n}(D_{n\theta} - D_\theta)(x,\varepsilon) - \sqrt{n}(D_{n\theta_0} - D_{\theta_0})(x,\varepsilon) \xrightarrow{P} 0 \text{ as } \theta \xrightarrow{P} \theta_0.$$

Proof. The decomposition

$$\begin{aligned}
&\sqrt{n}(D_{n\theta} - D_\theta)(x,\varepsilon) \\
&= \sqrt{n}(H_{n\theta} - H_\theta)(x,\varepsilon) - F_n(x)\sqrt{n}(G_{n\theta} - G_\theta)(\varepsilon) - G_\theta(\varepsilon)\sqrt{n}(F_n - F)(x)
\end{aligned}$$

forms a sum of three terms, each stochastically equicontinuous at θ_0. This is a consequence of the already mentioned Donsker property of $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ and the fact that the mapping $\theta \mapsto G_\theta$ is continuous at θ_0, since the Lipschitz condition on $\theta \mapsto \rho(\cdot,\theta)$ yields both $\mathbb{P}\{\rho(Z,\theta) \leq \lambda\} \geq G(\lambda - C\|\theta - \theta_0\|)$ and $\mathbb{P}\{\rho(Z,\theta) \leq \lambda\} \leq G(\lambda + C\|\theta - \theta_0\|)$ and the continuity follows.

It should be stressed that the Donsker property implies that the process is stochastically equicontinuous with respect to the $L_2(P)$ metric on the sets $A \in \mathcal{A}$ and $C \in \mathcal{C}$, *not* necessarily the Euclidean distance on Θ. [4] However, the Lipschitz condition on $\theta \mapsto \rho(\cdot, \theta)$ yields that

$$\begin{aligned}
&\mathbb{P}\{\rho(Z,\theta) \le \lambda, \rho(Z,\theta_0) \le \lambda\} \\
&\ge \mathbb{P}\{\rho(Z,\theta_0) \le \lambda - C\|\theta - \theta_0\|\} \\
&= G_{\theta_0}(\lambda - C\|\theta - \theta_0\|) \\
&\to G_{\theta_0}(\lambda) \text{ for } \theta \to \theta_0
\end{aligned}$$

as $G_{\theta_0} \equiv G$ is continuous. On the other hand,

$$\begin{aligned}
&\mathbb{P}\{\rho(Z,\theta) \le \lambda, \rho(Z,\theta_0) \le \lambda\} \\
&\le \mathbb{P}\{\rho(Z,\theta_0) \le \lambda\} \\
&= G_{\theta_0}(\lambda).
\end{aligned}$$

We have shown that $\mathbb{P}\{A_{\theta,\lambda} \cap A_{\theta_0,\lambda}\} \to \mathbb{P}\{A_{\theta_0,\lambda}\}$ as $\theta \to \theta_0$. By a similar argument we see that $\mathbb{P}\{A_{\theta,\lambda}\} \to \mathbb{P}\{A_{\theta_0,\lambda}\}$ as $\theta \to \theta_0$, so that

$$P(A_{\theta,\lambda} - A_{\theta_0,\lambda})^2 = P\{A_{\theta,\lambda}\} + P\{A_{\theta_0,\lambda}\} - 2P\{A_{\theta,\lambda} \cap A_{\theta_0,\lambda}\} \to 0,$$

as $\theta \to \theta_0$. In other words, the parametrization $\theta \mapsto I_{A_{\theta,y}}$ is continous at θ_0 in $L_2(P)$ sense. Hence in view of the stochastic equicontinuity with respect to the $L_2(P)$ distance

$$G_{n\theta}(\varepsilon) - G_\theta(\varepsilon) - G_{n,\theta_0}(\varepsilon) \xrightarrow{P} 0,$$

for all $\theta \xrightarrow{P} \theta_0$ and all $\varepsilon \in \mathcal{Y}$. A similar argument applies to the first term and the claim follows. □

Sufficient high level conditions for the previous theorem are

(i) $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are P-Donsker classes,

[4] Recall that the empirical process $\sqrt{n}(P_n - P)$, indexed by (indicator functions of) sets $A \in \mathcal{A}$, is stochastically equicontinuous at I_{A_0}, iff for all $\varepsilon, \eta > 0$ there exists a $\delta > 0$ such that $\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{P|I_A - I_{A_0}|\le\delta^2} |\sqrt{n}(P_n - P)(A)| > \varepsilon\right\} < \eta$.

(ii) the parametrization $\theta \mapsto I_{A_{\theta,y}}$ (the indicator function of the set $A_{\theta,y}$) is continous at θ_0 in the $L_2(P)$ sense. [5]

Proof of Corollary 3.2. The calculation in the proof of Lemma 3.6 using A.5 shows that $\theta \mapsto H_\theta$ is continuous. The consistency of $\hat{\theta}$ above implies that $H_{\hat{\theta}}(x,\varepsilon) \to H_{\theta_0}(x,\varepsilon) \equiv H(x,\varepsilon)$ a.s. by the continuous mapping theorem. The proof of Theorem 3.1 implies further that $\sup_{\theta\in\Theta} |H_{n\theta}(x,\varepsilon) - H_\theta(x,\varepsilon)| \xrightarrow{\text{a.s.}} 0$. In particular, $|H_{n\hat{\theta}}(x,\varepsilon) - H_{\hat{\theta}}(x,\varepsilon)| \xrightarrow{\text{a.s.}} 0$, and Corollary 3.2 follows after an application of the triangle inequality. □

Finally, we prove Theorem 3.3.

Proof of Theorem 3.3. The result follows after an application of Theorem 3.2, page 48 in Wegkamp (1999). We need to check the following conditions:

(i) $\hat{\theta} \xrightarrow{P} \theta_0$

(ii) $M(\theta)$ has a non-singular second derivative V at θ_0

(iii) $\sqrt{n}(M_n - M)(\theta)$ is stochastically differentiable at θ_0, that is, there exists W_n which converges weakly to a tight Gaussian distribution such that

$$\sqrt{n}(M_n - M)(\theta) = \sqrt{n}(M_n - M)(\theta_0) + (\theta - \theta_0)'W_n + o_P\left(1 + \sqrt{n}\|\theta - \theta_0\|\right)$$

for $\theta \to \theta_0$.

We have already established consistency, and we assumed the second requirement (ii). It remains to verify the stochastic differentiability requirement (iii). Recall that

$$D_\theta(x,\varepsilon) = (\theta - \theta_0)'\Delta(x,\varepsilon) + r_\theta(x,\varepsilon),$$

[5] This point is also discussed in Pakes and Pollard (1989) after Lemma 2.16 on pages 1036 – 1037. They require that

(i) each element in the interior of $A_{\theta_0,y}$ belongs to $A_{\theta,y}$ for θ close to θ_0

(ii) each element in the interior of the complement of $A_{\theta_0,y}$ belongs to the complement of $A_{\theta,y}$ for θ close to θ_0

(iii) the boundary of $A_{\theta_0,y}$ has zero P measure

where $\int \Delta^2 \, d\mu < \infty$, and $\int r_\theta^2(x,\varepsilon) \, d\mu = o(\|\theta - \theta_0\|^2)$ for $\theta \to \theta_0$. Next, observe that for $\theta \xrightarrow{P} \theta_0$

$$\begin{aligned}
&M_n(\theta) - M(\theta) \\
&= \int (D_{n\theta} - D_\theta + D_\theta)^2 \, d\mu - \int D_\theta^2 \, d\mu \\
&= \int (D_{n\theta} - D_\theta)^2 \, d\mu + 2 \int D_\theta (D_{n\theta} - D_\theta) \, d\mu \\
&= \int D_{n\theta_0}^2 \, d\mu + o_P(1/n) + 2(\theta - \theta_0)' \int \Delta (D_{n\theta} - D_\theta) \, d\mu + o_P(n^{-1/2} \|\theta - \theta_0\|) \\
&= M_n(\theta_0) + 2(\theta - \theta_0)' \int \Delta D_{n\theta_0} \, d\mu + o_P(n^{-1/2} \|\theta - \theta_0\| + 1/n).
\end{aligned}$$

In the above calculations we used that

$$\begin{aligned}
&\int (D_{n\theta} - D_\theta)^2 \, d\mu \\
&= \int D_{n\theta_0}^2 \, d\mu + \int (D_{n\theta} - D_\theta - D_{n\theta_0})^2 \, d\mu + 2 \int D_{n\theta_0} (D_{n\theta} - D_\theta - D_{n\theta_0}) \, d\mu \\
&\equiv I + II + III,
\end{aligned}$$

where $I = M_n(\theta_0)$ by definition, $II = o_P(1/n)$ for $\theta \xrightarrow{P} \theta_0$ by Lemma 3.6 above and the continuous mapping theorem (cf. Van der Vaart and Wellner (1996), Theorem 1.3.6, page 20), and finally

$$\begin{aligned}
III &\leq 2 \left(\int D_{n\theta_0}^2 \, d\mu \right)^{1/2} \left(\int (D_{n\theta} - D_\theta - D_{n\theta_0})^2 \, d\mu \right)^{1/2} \\
&= 2 \mathcal{O}_P(n^{-1/2}) \cdot o_P(n^{-1/2}) = o_P(1/n).
\end{aligned}$$

Also,

$$\begin{aligned}
&\int D_\theta (D_{n\theta} - D_\theta) \, d\mu \\
&= (\theta - \theta_0)' \int \Delta (D_{n\theta} - D_\theta) \, d\mu + \int r_\theta (D_{n\theta} - D_\theta) \, d\mu \\
&= (\theta - \theta_0)' \int \Delta (D_{n\theta} - D_\theta) \, d\mu + o_P(\|\theta - \theta_0\|)
\end{aligned}$$

as for all $\theta \xrightarrow{P} \theta_0$,

$$\begin{aligned}\left|\int r_\theta (D_{n\theta} - D_\theta)\, d\mu\right| &\leq \left(\int r_\theta^2\, d\mu\right)^{1/2} \cdot \left(\int (D_{n\theta} - D_\theta)^2\, d\mu\right)^{1/2} \\ &= \mathcal{O}_P(n^{-1/2}\|\theta - \theta_0\|).\end{aligned}$$

The other calculations are quite similar and have been omitted for this reason. Thus the conditions of Theorem 3.2 in Wegkamp (1999) are met, and consequently

$$\begin{aligned}\hat{\theta} &= \theta_0 - 2V^{-1} \cdot \left(\int \Delta(z) D_{n,\theta_0}(z)\, d\mu(z)\right) + \mathcal{O}_P(n^{-1/2}) \\ &\equiv \theta_0 - 2V^{-1}\Gamma_n + \mathcal{O}_P(n^{-1/2})\end{aligned}$$

holds true. The independence between $\varepsilon = \rho(Z, \theta_0)$ and X and Fubini's theorem imply that $\mathbb{E}\Gamma_n = 0$. Writing $H \equiv H_{\theta_0}$, $G \equiv G_{\theta_0}$, $H_n \equiv H_{n\theta_0}$ and $G_n \equiv G_{n\theta_0}$, the covariance term of the leading linear term equals

$$\begin{aligned}D(\Gamma_n) &= \mathbb{E}\Gamma_n\Gamma_n' = \mathbb{E}\left(\int \Delta(z) D_{n\theta_0}(z)\right)\left(\int \Delta(\overline{z}) D_{n\theta_0}(\overline{z})\right)' \\ &= \int\int \Delta(z)\Delta'(\overline{z})\, \mathbb{E}D_n(z)D_n(\overline{z})\, d\mu(z)\, d\mu(\overline{z}) \\ &= \int\int \Delta(z)\Delta'(\overline{z}) \cdot \mathbb{E}\left[(H_n - H)(z)(H_n - H)(\overline{z}) + (H - F_nG_n)(z)(H - F_nG_n)(\overline{z}) + \right. \\ &\qquad \left. (H_n - H)(z)(H - F_nG_n)(\overline{z}) + (H_n - H)(\overline{z})(H - F_nG_n)(z)\right] d\mu(z)\, d\mu(\overline{z})\end{aligned}$$

A tedious, but straightforward calculation further reveals that

$$\begin{aligned}D(\Gamma_n) &= \frac{1}{n}\int\int \Delta(x,\varepsilon)\Delta'(\overline{x},\overline{\varepsilon})\left[F(x)F(\overline{x})G(\min(\varepsilon,\overline{\varepsilon})) + F(\min(x,\overline{x}))G(\varepsilon)G(\overline{\varepsilon}) + \right. \\ &\qquad \left. + H(\min(x,\overline{x}),\min(\varepsilon,\overline{\varepsilon})) - 3H(x,\varepsilon)H(\overline{x},\overline{\varepsilon})\right] d\mu(x,\varepsilon)\, d\mu(\overline{x},\overline{\varepsilon}) + \mathcal{O}(\frac{1}{n}) \\ &\equiv \frac{1}{n}W + \mathcal{O}(\frac{1}{n}).\end{aligned}$$

In view of the preceeding stochastic expansion of $\hat{\theta}$ and since $\sqrt{n}(H_n - H)(x,\varepsilon)$, $\sqrt{n}(G_n - G)(\varepsilon)$, and $\sqrt{n}(F_n - F)(x)$ all converge to Gaussian processes, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to $\mathcal{N}(0, 4V^{-1}WV^{-1})$, by an application of Donsker's theorem, the continuous mapping theorem and Slutsky's lemma. □

4. Resampling estimates of the sampling distribution and asymptotic variance

In this section we provide an alternative to the normal approximation of the sampling distribution of $\hat{\theta}$ by means of resampling. We show that the ordinary nonparametric bootstrap is consistent. To formulate our result, let the pairs $Z_1^*, \ldots, Z_n^*$ be the (bootstrap) sample drawn from the data $Z_1, \ldots, Z_n$ with replacement. We denote the bootstrap counterpart of M_n based on the bootstrap sample by M_n^*, and let $\hat{\theta}^*$ be its minimum over Θ.

Theorem 4.1. *Under the regularity assumptions A.1 – A.7 described above, the conditional distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ consistently estimates the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ (in probability).*

Proof. Before proving the theorem, we first establish some auxiliary results, to wit, the bootstrap counterparts of Theorem 3.1 and Lemma 3.6.

Lemma 4.2. *Under the same assumptions as in Theorem 3.1, $\theta^* \xrightarrow{a.s.} \theta_0$ for almost all samples $Z_1, \ldots, Z_n$.*

Proof. By the triangle inequality, for almost all samples $Z_1, \ldots, Z_n$, we have

$$\begin{aligned}
&\sup_\theta |(M_n^* - M)(\theta)| \\
&\le \sup_\theta |(M_n^* - M_n)(\theta)| + \sup_\theta |(M_n - M)(\theta)| \\
&\le 4 \sup_{\theta, x, \varepsilon} |(D_{n,\theta}^* - D_{n,\theta})(x, \varepsilon)| + 4 \sup_{\theta, x, \varepsilon} |(D_{n,\theta} - D_\theta)(x, \varepsilon)| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

since $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are P-Donsker classes. The remainder of the proof goes as the one for Lemma 3.1 and has therefore been omitted. □

Lemma 4.3. *Assume A.3, A.4 and A.5. Then the process $\sqrt{n}(D_{n\theta}^* - D_{n\theta})(x, \varepsilon)$, is stochastically equicontinuous at θ_0 with respect to the Euclidean distance on Θ for all $x \in \mathcal{X}$ and $\varepsilon \in \mathcal{Y}$, conditionally given $Z_1, \ldots, Z_n$.*

Proof. Giné and Zinn (1990) proved that the empirical process $\sqrt{n}(P_n - P)$ can be bootstrapped if and only if the class of functions which index the process is P-Donsker. Therefore, as a consequence of the Donsker property of $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$,

$$\sqrt{n}(D_{n\theta}^* - D_{n\theta})(x, \varepsilon) - \sqrt{n}(D_{n\theta} - D_\theta)(x, \varepsilon) \xrightarrow{P} 0,$$

and the desired result follows from Lemma 3.6. □

Now we are in the position to prove Theorem 4.1. The proof closely follows the arguments for M-estimators obtained by Arcones and Giné (1990). Observe that by similar arguments given in the proof of Theorem 3.3, for all $\theta \xrightarrow{P} \theta_0$

$$
\begin{aligned}
&M_n^*(\theta) - M_n(\theta) \\
&= \int (D_{n\theta}^* - D_{n\theta})^2 \, d\mu + 2 \int D_{n\theta}(D_{n\theta}^* - D_{n\theta}) \, d\mu \\
&= \int (D_{n\theta_0}^* - D_{n\theta_0})^2 \, d\mu + \mathcal{o}_P(1/n) + 2 \int D_\theta(D_{n\theta}^* - D_{n\theta}) \, d\mu + 2 \int (D_{n\theta} - D_\theta)(D_{n\theta}^* - D_{n\theta}) \, d\mu \\
&= \int (D_{n\theta_0}^* - D_{n\theta_0})^2 \, d\mu + 2 \int D_\theta(D_{n\theta}^* - D_{n\theta}) \, d\mu + 2 \int D_{n\theta_0}(D_{n\theta_0}^* - D_{n\theta_0}) \, d\mu + \mathcal{o}_P(1/n) \\
&= \int (D_{n\theta_0}^* - D_{n\theta_0})^2 \, d\mu + 2 \int D_{n\theta_0}(D_{n\theta_0}^* - D_{n\theta_0}) \, d\mu + 2(\theta - \theta_0)' \int \Delta(D_{n\theta}^* - D_{n\theta}) \, d\mu + \\
&\quad + \mathcal{o}_P(n^{-1/2}\|\theta - \theta_0\| + n^{-1}).
\end{aligned}
$$

Consequently, for $\theta \xrightarrow{P} \theta_0$ and $\eta \xrightarrow{P} \theta_0$,

$$
\begin{aligned}
&M_n^*(\theta) - M_n^*(\eta) \\
&= [(M_n^* - M_n)(\theta) - (M_n^* - M_n)(\eta)] + [(M_n - M)(\theta) - (M_n - M)(\eta)] \\
&\quad + [M(\theta) - M(\eta)] \\
&= 2(\theta - \eta)^T \int \Delta \left[(D_{n\theta_0}^* - D_{n\theta_0}) + (D_{n\theta_0} - D_{\theta_0}) \right] \, d\mu + \\
&\quad + \frac{1}{2}(\theta - \theta_0)^T V(\theta - \theta_0) - \frac{1}{2}(\eta - \theta_0)^T V(\eta - \theta_0) + \\
&\quad + \mathcal{o}_P(\|\theta - \theta_0\|^2 + \|\eta - \theta_0\|^2 + n^{-1/2}\|\theta - \theta_0\| + n^{-1/2}\|\eta - \theta_0\| + n^{-1}).
\end{aligned}
$$

We define

$$
\Delta_n = 2 \int \Delta(D_{n\theta_0} - D_{\theta_0}) \, d\mu \text{ and } \Delta_n^* = 2 \int \Delta(D_{n\theta_0}^* - D_{n\theta_0}) \, d\mu
$$

and we take $\theta = \hat{\theta}$ and $\eta = \theta_0 - (\Delta_n + \Delta_n^*)$. Observe that $\eta \in \Theta$ for n sufficiently large, as θ_0 is an interior point of Θ. To simplify matters, we assume without loss of generality that

$V = I$. Hence

$$\begin{aligned}
&M_n^*(\theta) - M_n^*(\eta) \\
&= (\theta - \eta)^T(\Delta_n^* + \Delta_n) + \frac{1}{2}\|\theta - \theta_0\|^2 - \frac{1}{2}\|\eta - \theta_0\|^2 \\
&\quad + o_P(\|\theta - \theta_0\|^2 + \|\eta - \theta_0\|^2 + n^{-1/2}\|\theta - \theta_0\| + n^{-1/2}\|\eta - \theta_0\| + n^{-1})
\end{aligned}$$

and

$$\begin{aligned}
0 &\geq M_n^*(\theta^*) - M_n^*(\theta_0 - (\Delta_n + \Delta_n^*)) \\
&= -(\theta^* - \theta_0)^T(\Delta_n^* + \Delta_n) + \frac{1}{2}\|\Delta_n + \Delta_n^*\|^2 + \frac{1}{2}\|\theta^* - \theta_0\|^2 - \frac{1}{2}\|\Delta_n^* + \Delta_n\|^2 + \\
&\quad + o_P(\|\theta^* - \theta_0\|^2 + \|\Delta_n + \Delta_n^*\|^2 + n^{-1/2}\|\theta^* - \theta_0\| + n^{-1/2}\|\Delta_n + \Delta_n^*\| + n^{-1}) \\
&= \frac{1}{2}\|\theta^* - \theta_0 - (\Delta_n^* + \Delta_n)\|^2 + o_P(\|\theta^* - \theta_0\|^2 + n^{-1/2}\|\theta^* - \theta_0\| + n^{-1}).
\end{aligned}$$

whence

$$n\|\theta^* - \theta_0 - (\Delta_n^* + \Delta_n)\|^2 \to 0$$

in P_n- probability. By the preceding theorem

$$\hat{\theta} - \theta_0 = \Delta_n + o_P(n^{-1/2}),$$

so that combination yields $\theta^* - \hat{\theta} = \Delta_n^* + o_P(n^{-1/2})$. The term Δ_n^* has the same limiting distribution as Δ_n by the bootstrap theorem for the mean in $\mathbb{R}^d$. This concludes the proof. □

We end this section with a discussion of the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$. In principle, under sufficient smoothness assumptions, we could plug in $\hat{\theta}$ and P_n in the covariance matrix $\Sigma = \Sigma(\theta_0, P)$. Here P is the probability measure of Z, and P_n is the empirical measure, putting mass $1/n$ at each observation Z_i. However, $\Sigma(\theta_0, P)$ has a complicated structure, and the bootstrap estimator of the variance provides an attractive alternative. Second, we show that the delete $-d$ jackknife estimator [6] of the variance of

[6] Let $\hat{\theta}_{d,s}$ be the estimate based on the data set $Z_i, i \in s$, where s is a subset of $\{1, 2, ..., n\}$ with size $n - d$. Let $\mathcal{S}$ be the collection of all possible subsets of $\{1, 2, ..., n\}$ of size $n - d$, and let $N = \binom{n}{d}$ be its cardinality. The delete $-d$ jackknife of $c'\hat{\theta}$ is defined as

$$\frac{n-d}{dN}\sum_{s\in\mathcal{S}}\left(c'\hat{\theta}_{d,s} - \frac{1}{N}\sum_s c'\hat{\theta}_{d,s}\right)^2.$$

linear combinations $c'\hat{\theta}$ is consistent for d satisfying

$$d/n \geq \varepsilon \text{ for some } \varepsilon > 0 \text{ and } n - d \to \infty. \tag{4.1}$$

We were not able to show that the ordinary jackknife ($d = 1$) works due to the lack of smoothness of the map $\theta \mapsto M_n(\theta)$. For the same reason, the jackknife estimator of the variance of the sample median is inconsistent (cf. Shao and Tu (1989)).

Theorem 4.4. *Under the regularity conditions A.1 – A.7, the nonparametric bootstrap and delete $-d$ jackknife estimators of the variance of $c'\sqrt{n}(\hat{\theta} - \theta_0)$, where d satisfies (4.1) are consistent, for all $c \in \mathbb{R}^{dim(\Theta)}$.*

Proof. Again we set out with the technical lemma's first, concerning uniform integrability of $\|\sqrt{n}(\hat{\theta} - \theta_0)\|^2$.

Lemma 4.5. *If A.3 holds, we have for all $k > 0$*

$$\mathbb{E}\left(\sup_{x,\varepsilon,\theta}\left|\sqrt{n}(D_{n\theta}(x,\varepsilon) - D_\theta(x,\varepsilon)\right|\right)^k < \infty,$$

and

$$\mathbb{E}^*\left(\sup_{x,\varepsilon,\theta}\left|\sqrt{n}(D^*_{n\theta}(x,\varepsilon) - D_{n\theta}(x,\varepsilon)\right|\right)^k < \infty \text{ a.s. }.$$

Proof. First notice that

$$\begin{aligned}
&\mathbb{E}\left(\sup_{x,\varepsilon,\theta}\left|\sqrt{n}(D_{n\theta}(x,\varepsilon) - D_\theta(x,\varepsilon)\right|\right)^k \\
&\leq C_k\left\{\mathbb{E}\sup_{A\in\mathcal{A}}|\sqrt{n}(P_n - P)A|^k + \mathbb{E}\sup_{B\in\mathcal{B}}|\sqrt{n}(P_n - P)B|^k + \mathbb{E}\sup_{C\in\mathcal{C}}|\sqrt{n}(P_n - P)C|^k\right\}
\end{aligned}$$

If $\{\rho(\cdot,\theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space, an application of Theorem 2.14.1, page 237 in Van der Vaart & Wellner (1996) yields

$$\left(\mathbb{E}\sup_{A\in\mathcal{A}}|\sqrt{n}(P_n - P)A|^k\right)^{1/k} \leq C \sup_{Q \text{ discrete}}\int_0^1 \sqrt{1 + \log N(\delta, \mathcal{A}, L_2(Q))}\, d\delta,$$

The right hand side is finite by the bounds obtained in Lemma 3.4. The same applies for $\mathcal{B}$ and $\mathcal{C}$, and combination of the previous two displays establishes the first part for finite dimensional spaces. The bootstrap counterpart follows by the same argument. For the case of smooth functions we do not have an uniform bound for the covering numbers, but a bound on the bracketing numbers instead. Another difference is that we needed to assume the existence of a bounded probability density for $H(x,\varepsilon)$. For this case, Theorem 2.14.5, page 244 and Theorem 2.12.2, page 240 in Van der Vaart & Wellner (1996) yield respectively

$$\begin{aligned}
&\left\{ \mathbb{E}\left(\sup_{x,\varepsilon,\theta} \left|\sqrt{n}(D_{n\theta}(x,\varepsilon) - D_\theta(x,\varepsilon)\right| \right)^k \right\}^{1/k} \\
&\leq C\mathbb{E} \sup_{A\in\mathcal{A}} |\sqrt{n}(P_n - P)A| + Cn^{-\frac{1}{2}+\frac{1}{k}} \text{ for } k \geq 2 \\
&\leq C\int_0^1 \sqrt{1+\log N_B(\delta,\mathcal{A},L_2(P))}\, d\delta + Cn^{-\frac{1}{2}+\frac{1}{k}}
\end{aligned}$$

The bound on the bracketing numbers in Lemma 3.5 shows that the right hand side is finite. The same is true of course for the classes $\mathcal{B}$ and $\mathcal{C}$, and the first claim follows for the case of smooth functions. Also, by the same reasoning, for $k \geq 2$,

$$\begin{aligned}
&\left\{ \mathbb{E}\left(\sup_{x,\varepsilon,\theta} \left|\sqrt{n}(D^*_{n\theta}(x,\varepsilon) - D_{n\theta}(x,\varepsilon)\right| \right)^k \right\}^{1/k} \\
&\leq C\int_0^1 \sqrt{1+\log N_B(\delta,\mathcal{A},L_2(P_n))}\, d\delta + Cn^{-\frac{1}{2}+\frac{1}{k}} \\
&\leq C\int_0^1 \sqrt{1+\log N_B(\delta/2,\mathcal{A},L_2(P))}\, d\delta + Cn^{-\frac{1}{2}+\frac{1}{k}} \text{ a.s.},
\end{aligned}$$

where we used the uniform law of large numbers in the last line. This completes our proof. □

Lemma 4.6. *Under A.1, A.2, A.3, A.6 and A.7, $\|\sqrt{n}(\hat{\theta}-\theta)\|$ is uniformly square integrable.*

Proof. It suffices to show that $\mathbb{E}\|\sqrt{n}(\hat{\theta}-\theta_0)\|^3 < \infty$. First, observe that by A.7, there exist $\delta > 0$ and $c > 0$ such that for all $\|\theta - \theta_0\| < \delta$,

$$c\|\theta-\theta_0\|^2 \leq M(\theta) - M(\theta_0) = M(\theta).$$

Second, for any fixed $\delta > 0$ (not depending on n), there exists an $\eta > 0$ such that $\|\theta-\theta_0\| \geq \delta$ implies that $M(\theta) - M(\theta_0) > \eta$. This is a consequence of θ_0 being a well separated minimum

of M, which follows from A1, A2 and A7 (cf. the proof of Theorem 3.1). Combining these two observations, we find

$$\begin{aligned}\mathbb{E}\|\hat{\theta}-\theta_0\|^3 &= \mathbb{E}\|\hat{\theta}-\theta_0\|^3\{\|\hat{\theta}-\theta_0\|<\delta\}+\mathbb{E}\|\hat{\theta}-\theta_0\|^3\{\|\hat{\theta}-\theta_0\|\geq\delta\}\\ &\leq C\left(\mathbb{E}M^{3/2}(\hat{\theta})+\mathbb{P}\{\|\hat{\theta}-\theta_0\|\geq\delta\}\right)\end{aligned}$$

The constant $C>0$ is a generic constant independent of n. In the last line we invoked A.1 as well. We will bound the two terms on the right hand side separately. Notice that

$$\mathbb{P}\{\|\hat{\theta}-\theta_0\|\geq\delta\}\leq\mathbb{P}\{M(\hat{\theta})-M(\theta_0)\geq\eta\}=\mathbb{P}\{M(\hat{\theta})-M_n(\hat{\theta})+M_n(\hat{\theta})-M(\theta_0)\geq\eta\}$$
$$\leq\mathbb{P}\left\{2\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\geq\eta\right\}\leq(\eta/2)^{-\alpha}\mathbb{E}\sup_{x,\varepsilon,\theta}|D_\theta(x,\varepsilon)-D_{n\theta}(x,\varepsilon)|^\alpha=\mathcal{O}(n^{-\alpha/2})$$

for any $\alpha>2$ by Lemma 4.5. The other term can be handled as follows:

$$\mathbb{E}M^{3/2}(\hat{\theta})=\mathbb{E}\left((M-M_n)(\hat{\theta})+M_n(\hat{\theta})\right)^{3/2}\leq\mathbb{E}\left(|(M_n-M)(\hat{\theta})|+|M_n(\theta_0)|\right)^{3/2}$$
$$\leq\mathbb{E}\left(\int(D_{n\hat{\theta}}-D_{\hat{\theta}})^2\,d\mu+2\int|D_{\hat{\theta}}||D_{n\hat{\theta}}-D_{\hat{\theta}}|\,d\mu+M_n(\theta_0)\right)^{3/2}$$
$$\leq\mathbb{E}\left(2\sup_{x,\varepsilon,\theta}(D_{n\theta}-D_\theta)^2(x,\varepsilon)+C\|\hat{\theta}-\theta_0\|\sup_{x,\varepsilon,\theta}|(D_{n\theta}-D_\theta)(x,\varepsilon)|\right)^{3/2}$$
$$\leq C\mathbb{E}\sup_{x,\varepsilon,\theta}|(D_{n\theta}-D_\theta(x,\varepsilon)|^3+C\left(\mathbb{E}\|\hat{\theta}-\theta_0\|^3\right)^{1/2}\left(\mathbb{E}\sup_{x,\varepsilon,\theta}|D_{n\theta}(x,\varepsilon)-D_\theta(x,\varepsilon)|^3\right)^{1/2}$$
$$=\mathcal{O}(n^{-3/2})+\mathcal{O}(n^{-3/4})\left(\mathbb{E}\|\hat{\theta}-\theta_0\|^3\right)^{1/2}.$$

Combining all three preceding displays, we obtain that

$$\mathbb{E}\|\hat{\theta}-\theta_0\|^3\leq\mathcal{O}(n^{-3/2})+\mathcal{O}(n^{-3/4})\left(\mathbb{E}\|\hat{\theta}-\theta_0\|^3\right)^{1/2}.$$

Since $\mathbb{E}\|\hat{\theta}-\theta_0\|^3<\infty$ by A.1, the conclusion follows. □

It is shown in Shao and Tu (1995, Theorem 2.10, page 52), that the stochastic expansion

$$c'\hat{\theta}=c'\theta_0-2V^{-1}\int c'\Delta\,D_{n\theta_0}\,d\mu+\mathcal{o}_P(1/\sqrt{n})$$

and the uniform integrability of $\|\sqrt{n}c'(\hat{\theta}-\theta_0)\|^2$ imply the consistency of J^2_{-d}, provided the tuning parameter d satisfies

$$d/n\geq\varepsilon\text{ for some }\varepsilon>0\text{ and }n-d\to\infty.$$

The weak convergence result Theorem 4.1 and uniform square integrability of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ imply that the bootstrap provides a consistent alternative for estimating the variance of a linear combination of θ_0. It remains to prove the uniform square integrability, which is immediate from Lemma 4.7 below.

Lemma 4.7. *Under A.1, A.2, A.3, A.6 and A.7,*

$$\mathbb{E}^* \|\sqrt{n}(\hat{\theta}^* - \hat{\theta})\|^3 < \infty \text{ a.s..}$$

Proof. As in the proof of Lemma 4.6, there exists $c = c(\theta_0, \delta) > 0$ such that

$$\begin{aligned}
&c\|\hat{\theta}^* - \theta_0\|^2 \{\|\hat{\theta}^* - \theta_0\| \le \delta\} \\
&\le M(\hat{\theta}^*) - M_n(\hat{\theta}^*) + M_n(\hat{\theta}^*) - M_n^*(\hat{\theta}^*) + M_n^*(\hat{\theta}^*) \\
&\le \{M(\hat{\theta}^*) - M_n(\hat{\theta}^*)\} + \{M_n(\hat{\theta}^*) - M_n^*(\hat{\theta}^*)\} + \{M_n^*(\theta_0) - M_n(\theta_0)\} + M_n(\theta_0).
\end{aligned}$$

Most terms can be handled as before in the proof of Lemma 4.6, and for the additional terms we invoke Lemma 4.5. □

The proof of Theorem 4.4 is complete. □

5. Estimation of a Simulated Model of Consumer Demand

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

$$\text{maximize } V(y, \varepsilon, \theta_0) \text{ over } y \text{ such that } p \cdot y = I.$$

The econometrician knows $V(y, \varepsilon, \theta)$ 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 the ε or know the distribution of ε. The econometrician does observe $X = (P, I)$. The econometrician's problem is to estimate θ_0 and the distribution of ε 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. In general, these conditions define an implicit nonlinear simultaneous equation model of the form $\varepsilon = \rho(X, Y, \theta)$, where the reduced form function is the consumer's random demand function $Y(P, I, \varepsilon, \theta_0)$. The specification of $V(y, \varepsilon, \theta)$ proposed by Brown–Matzkin (1998) where $V(y, \varepsilon, \theta) = U(y, \theta) + \varepsilon \cdot y$ generates an econometric model of this type. 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 $\mathbb{R}^k$.

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

$$V(y, \varepsilon, \theta) = y_1^{\theta_1} y_2^{\theta_2} + \ln y_0 + \varepsilon_1 y_1 + \varepsilon_2 y_2, \text{ where } \theta_1, \theta_2 \in (0, 1).$$

Then the first-order conditions for this optimization problem can be written as $\varepsilon = \rho(X, Y, \theta)$, where $X = (P_1, P_2, I)$, $Y = (Y_0, Y_1, Y_2)$ and $\theta = (\theta_1, \theta_2)$

$$\text{(i)} \quad \varepsilon_1 = \frac{P_1}{(I - P_1 Y_1 - P_2 Y_2)} - \theta_1 Y_1^{(\theta_1 - 1)} Y_2^{\theta_2}$$

$$\text{(ii)} \quad \varepsilon_2 = \frac{P_2}{(I - P_1 Y_1 - P_2 Y_2)} - \theta_2 Y_1^{\theta_1} Y_2^{(\theta_2 - 1)}$$

Equations (i) and (ii) can (in principle) be solved uniquely for the random demand functions $Y_1(X, \varepsilon, \theta)$ and $Y_2(X, \varepsilon, \theta)$, if $\theta_1 + \theta_2 < 1$. This verifies assumption (I.1). To verify assumption (I.2) we consider the matrix $(\partial\rho/\partial y)$,

$$\left(\frac{\partial \rho}{\partial y}\right) = \begin{bmatrix} \dfrac{\partial \rho_1}{\partial y_1} & \dfrac{\partial \rho_1}{\partial y_2} \\ \dfrac{\partial \rho_2}{\partial y_1} & \dfrac{\partial \rho_2}{\partial y_2} \end{bmatrix}$$

where

$$\begin{aligned} \frac{\partial \rho_1}{\partial y_1} &= \frac{p_1^2}{(I - p_1 y_1 - p_2 y_2)^2} + \theta_1(1 - \theta_1) y_1^{(\theta_1 - 2)} y_2^{\theta_2} \\ \frac{\partial \rho_1}{\partial y_a} &= \frac{\partial \rho_2}{\partial y_1} = \frac{p_1 p_2}{(I - p_1 y_1 - p_2 y_2)^2} + \theta_1 \theta_2 y_1^{(\theta_1 - 1)} y_2^{(\theta_2 - 1)} \\ \frac{\partial \rho_2}{\partial y_2} &= \frac{p_2^2}{(I - p_1 y_1 - p_2 y_2)^2} + \theta_2(1 - \theta_2) y_1^{\theta_1} y_2^{(\theta_2 - 2)} \end{aligned}$$

$\det(\partial\rho/\partial y) > 0$, if $\theta_1 + \theta_2 < 1$. Hence $\partial\rho/\partial y$ has rank 2 and (I.2) is verified. It follows from (I.2) and the implicit function theorem that $\partial y(x, \varepsilon, \theta)/\partial x$ can be computed from the

structural equations $\varepsilon = \rho(x, y, \theta)$. In fact,

$$\begin{aligned}\left(\frac{\partial y}{\partial x}\right) &= -\left[\frac{\partial \rho}{\partial y}\right]^{-1}\left[\frac{\partial \rho}{\partial x}\right] \\ \left[\frac{\partial \rho}{\partial x}\right] &= \begin{bmatrix} \frac{\partial \rho_1}{\partial p_1} & \frac{\partial \rho_1}{\partial p_2} & \frac{\partial \rho_1}{\partial I} \\ \frac{\partial \rho_2}{\partial p_1} & \frac{\partial \rho_2}{\partial p_2} & \frac{\partial \rho_2}{\partial I} \end{bmatrix}\end{aligned}$$

where

$$\begin{aligned}\frac{\partial \rho_1}{\partial p_1} &= \frac{1}{(I - p_1 y_1 - p_2 y_2)} + \frac{p_1 y_1}{(I - p_1 y_1 + p_2 y_2)^2} \\ \frac{\partial \rho_1}{\partial p_2} &= \frac{p_2 y_1}{(I - p_1 y_1 + p_2 y_2)^2} \\ \frac{\partial \rho_1}{\partial I} &= \frac{-p_1}{(I - p_1 y_1 + p_2 y_2)^2} \\ \frac{\partial \rho_2}{\partial p_1} &= \frac{1}{(I - p_1 y_1 + p_2 y_2)} + \frac{p_2 y_2}{(I - p_1 y_1 + p_2 y_2)^2} \\ \frac{\partial \rho_2}{\partial p_2} &= \frac{p_1 y_1}{(I - p_1 y_1 + p_2 y_2)^2} \\ \frac{\partial \rho_2}{\partial I} &= \frac{-p_2}{(I - p_1 y_1 + p_2 y_2)^2}\end{aligned}$$

We see that $\partial \rho / \partial x$ has rank 2 and is independent of θ and ε.

Therefore,

$$\begin{aligned}\frac{\partial y(x, \varepsilon, \theta_0)}{\partial x} &= \frac{\partial y(x, \varepsilon, \theta)}{\partial x} \text{ a.e. iff} \\ \frac{\partial \rho(x, y, \theta_0)}{\partial y} &= \frac{\partial \rho(x, y, \theta)}{\partial y} \text{ a.e.}\end{aligned}$$

But

$$\begin{aligned}\text{(iii)}\ \frac{\partial \rho(x, y, \theta_0)}{\partial y} &= \frac{\partial \rho(x, y, \theta)}{\partial y} \text{ a.e. iff} \\ \theta &= \theta_0\end{aligned}$$

If (iii) holds, then $\theta_{0,1}\theta_{0,2} y_1^{(\theta_{0,1}-1)} y_2^{(\theta_{0,2}-1)} \equiv \theta_1 \theta_2 y_1^{(\theta_1 - 1)} y_2^{(\theta_2 - 1)}$. This implies

$$\begin{aligned}&\ln \theta_{0,1} + \ln \theta_{0,2} + (\theta_{0,1} - 1)\ln y_1 + (\theta_{0,2} - 1)\ln y_2 \\ \equiv\ &\ln \theta_1 + \ln \theta_2 + (\theta_1 - 1)\ln y_1 + (\theta_2 - 1)\ln y_2.\end{aligned}$$

Taking partial derivatives with respect to y_1 and y_2, we see that $\theta_{0,1} = \theta_1$ and $\theta_{0,2} = \theta_2$.

If $\theta \neq \theta_0$ then $\partial y(x,\varepsilon,\theta_0)/\partial x \neq \partial y(x,\varepsilon,\theta)/\partial x$ and by Theorem 2 $\partial \rho(x, y(x,\varepsilon,\theta_0),\theta)/\partial x \neq 0$ for some $(\overline{x},\overline{\varepsilon})$. It follows from Theorem 2.3 that the model is identified if $\theta_1 + \theta_2 < 1$, i.e., Assumption (A.2) holds. It is important to notice that the structural equations for this model are nonlinear in *both* the parameters and the variables.

If we assume that p_1 and p_2 are uniformly bounded away from 0 and I is bounded above, then this model satisfies all our verifiable regularity conditions, i.e., Assumptions (A.1), (A.2), (A.3), (A.5) and (A.6).

A more tractable model for simulation, where we can derive explicit expressions for the random demand functions, is the following consumer optimization problem:

$$\begin{aligned} &\max_{y_1,y_2} \theta \ln y_1 + (1-\theta)\ln y_2 + \varepsilon y_1, \text{ where } \theta \in (0,1) \\ &\text{s.t. } py_1 + y_2 = I \\ &\quad y_1, y_2 \geq 0 \end{aligned}$$

An equivalent optimization problem for the consumer is:

$$\begin{aligned} &\max_{y_1} \theta \ln y_1 + (1-\theta)\ln(I - py_1) + \varepsilon y_1 \\ &\text{s.t. } 0 \leq y_1 \leq I/p \end{aligned}$$

The first-order condition for this problem is:

$$\varepsilon = \frac{(1-\theta)p}{(I-py_1)} - \frac{\theta}{y_1}$$

This equation can be solved explicitly for the random demand function $Y_1(P,I,\varepsilon,\theta)$. Then $Y_2(P,I,\varepsilon,\theta) = I - PY_1(P,I,\varepsilon,\theta)$, where

$$Y_1(P,I,\varepsilon,\theta) = \frac{(\varepsilon I - P) + \sqrt{(\varepsilon I - P)^2 + 4\theta I P_\varepsilon}}{2P_\varepsilon}$$

In our simulations we choose $\theta_0 = 0.3579$, then randomly draw I_i , p_i and ε_i, $i = 1, \ldots, n = 100$ from uniform distributions on (0,1) and compute $y_1(p,I,\varepsilon,\theta)$ and $y_2(p,I,\varepsilon,\theta)$. We chose the uniform measure on $[0,1]^2$ for the measure μ. Figure 1 below is a histogram based on 1000 simulations of the model described above. Let $\hat{\theta}_b$, $b = 1, \ldots, 1000$ be the weighted minimum mean-square distance from independence estimate for θ_0 in the b-th simulation. We found in

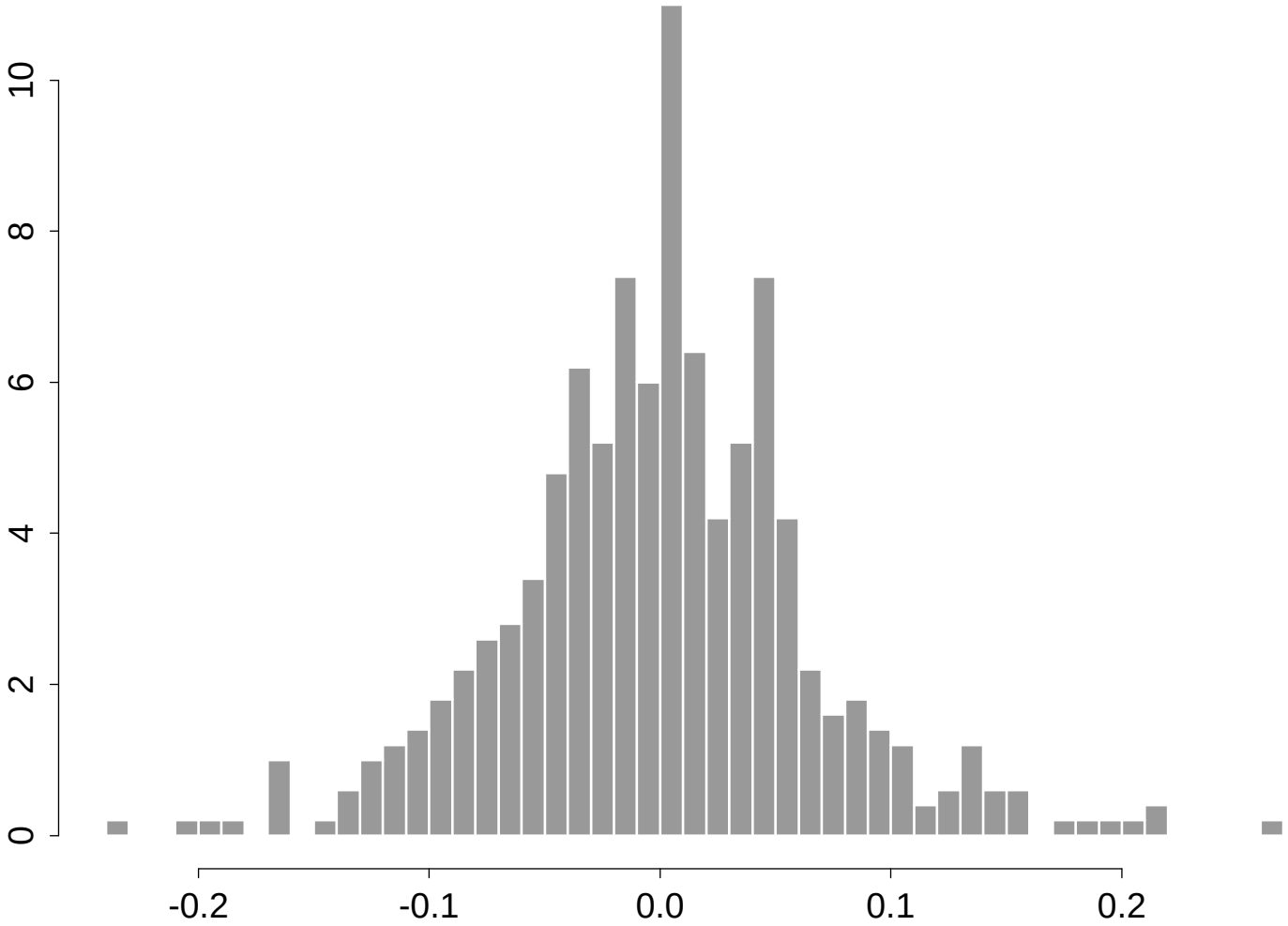
our simulation that

$$\overline{\theta} = \frac{1}{1000}\sum_{b=1}^{1000}\hat{\theta}_b = 0.3576, \text{ bias} = \frac{1}{1000}\sum_{b=1}^{1000}\hat{\theta}_b - \theta_0 = -3.0750 \times 10^{-4},$$

$$\text{std} = \sqrt{\frac{1}{999}\sum_{b=1}^{1000}(\hat{\theta}_b - \overline{\theta})^2} = 0.0071, \text{ and MSE} = \frac{1}{1000}\sum_{b=1}^{1000}(\hat{\theta}_b - \theta_0)^2 = 5.033 \times 10^{-5}.$$

The histogram confirms that our estimators are consistent and asymptotically normal.

Figure 1: Histogram based on 1000 simulations.



6. Appendix: Identification in Nonlinear Simultaneous Equations Models with Multple Equilibria

Nonlinear simultaneous equations models with multiple equilibria are commonplace in applications of game theory or general equilibrium theory in industrial organization and macroeconomics. Amemiya (1985) in his discussion of estimation in nonlinear simultaneous equations models observed that nonlinear two-stage and nonlinear three-stage estimators of θ_0 are consistent even in the presence of multiple equilibria. Unfortunately, this consistency is predicated on his (unstated) assumption that the model is identified – see section 18.7 in Davidson and Mackinnon (1993) for discussion.

Weighted minimum mean-square distance from independence estimators are also consistent in the presence of multiple equilibria, if the model is identified. In section 2 of this paper we presented a necessary and sufficient condition for identifying nonlinear simultaneous equations models with unique equilibria, i.e. Theorem 2.3. Surprisingly, this condition is also *sufficient* for identifying nonlinear simultaneous equations models with multiple equilibria. In addition to assumptions I.2, I.3 and I.4, we assume only that there are a finite number of equilibria. If $\mathcal{Y}$ is compact then this condition folows from I.2 and the implicit function theorem.

Theorem 6.1. *Let $\varepsilon = \rho(x, y, \theta)$ and suppose*

(1) $(\forall \theta \in \Theta)(\forall (x, \varepsilon))\exists N = N(x, \varepsilon, \theta)$ *and* $\{y_j\}_{j=1}^N$ *such that* $\varepsilon = \rho(x, y_j, \theta)$.

(2) *The matrix $\partial\rho/\partial y$ has full rank.*

(3) *X and ε are stochastically indpendent.*

(4) *The joint c.d.f. of (X, ε) is absolutely continuous with respect to Lebesgue measure with positive density.*

(5) $(\forall \theta \neq \theta_0)(\exists(\overline{x}, \overline{\varepsilon}))$ *such that* $\left.\frac{\partial \rho(x, y_j(x,\varepsilon,\theta_0),\theta)}{\partial x}\right|_{(\overline{x},\overline{\varepsilon})} \neq 0$ *for some j.*

Then θ_0 is the unique minimum of $M(\theta)$.

Proof. Without loss of generality, we may assume that (5) implies $\exists(i, k)$ such that

$$\left.\frac{\partial \rho_i(x, y_j(x, \varepsilon, \theta_0), \theta)}{\partial x_k}\right|_{(\overline{x},\overline{\varepsilon})} > 0$$

for all $\theta \neq \theta_0$. Given $(\overline{x}, \overline{\varepsilon})$ and assumptions (1) and (2), we obtain by the implicit function theorem the existence of neighborhoods $U_{\overline{x}}$ of $\overline{x}$ and $U_{\overline{\varepsilon}}$ of $\overline{\varepsilon}$ and functions $\{y_l\}_{l=1}^N$ such that

(i) $y_l : U_{\overline{x}} \times U_{\overline{\varepsilon}} \to \mathbb{R}^K$.

(ii) y_l are smooth.

(iii) y_l have disjoint ranges.

Given $(\overline{x}, \overline{\varepsilon})$ and assumption (5), there exist neighborhoods $V_{\overline{x}}$ of $\overline{x}$ and $V_{\overline{\varepsilon}}$ of $\overline{\varepsilon}$ such that $\forall (x, \varepsilon) \in V_{\overline{x}} \times V_{\overline{\varepsilon}}$

$$\left. \frac{\partial \rho_i(x, y_j(x, \varepsilon, \theta_0), \theta)}{\partial x_k} \right|_{(\overline{x}, \overline{\varepsilon})} > 0.$$

In particular, $\exists \delta > 0$ such that $\rho_i(\overline{x}, Y_j(\overline{x}, \varepsilon, \theta_0), \theta) < \rho_i(\overline{x} + \delta e_k, Y_j(\overline{x} + \delta e_k, \varepsilon, \theta_0), \theta)$ for all $\varepsilon \in V_{\overline{\varepsilon}}$. Let $W_{\overline{x}} = U_{\overline{x}} \cap V_{\overline{x}}$ and $W_{\overline{\varepsilon}} = U_{\overline{\varepsilon}} \cap V_{\overline{\varepsilon}}$ then $\exists 0 < \eta \leq \delta$ such that $\overline{x} + \eta e_k \in W_{\overline{x}}$ and $\rho_i(\overline{x}, y_j(\overline{x}, \varepsilon, \theta_0), \theta) < \rho_i(\overline{x} + \eta e_k, y_j(\overline{x} + \eta e_k, \varepsilon, \theta_0), \theta)$ for all $\varepsilon \in W_{\overline{\varepsilon}}$. Moreover, the restrictions of the y_l to $W_{\overline{x}} \times W_{\overline{\varepsilon}}$ have disjoint ranges. Let $D_j = \{(\overline{x}, y_j(\overline{x}, \varepsilon, \theta_0)) \mid \varepsilon \in W_{\overline{\varepsilon}}\} \cup \{(\overline{x} + \eta e_k, y_j(\overline{x} + \eta e_k, \varepsilon, \theta_0)) \mid \varepsilon \in W_{\overline{\varepsilon}}\}$ and $g_\theta(x, y) = \rho_i(x, y, \theta)$ if $(x, y) \in D_j$ and $g_\theta(x, y) = 0$ otherwise. The change of variables formula for densities, see equation (8.10.2) in Hoffmann-Jørgensen (1994), yields $\mathbb{E}(g_\theta(X, Y) \mid X = \overline{x}) < \mathbb{E}(g_\theta(X, Y) \mid X = \overline{x} + \eta e_k)$. Hence $\varepsilon = \rho(x, y, \theta)$ depends on x for all $\theta \neq \theta_0$, and θ_0 is the unique minimum of $M(\theta)$. The above argument is derived from Brown (1983, pp. 180, 181). □

References

[1] Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press, Cambridge, Mass.

[2] Andrews, D. (1994). Empirical Process Methods in Econometrics. In *Handbook of Econometrics*, Volume 4 (R.F. Engle and D.L. McFadden eds), 2247–2294.

[3] Andrews, D. (1999). Estimation when a parameter is on the boundary. *Working paper, Yale University.*

[4] Andrews, D. (1999). Estimation when a parameter is on the boundary. *Econometrica*, **67**(6), 1341 – 1384.

[5] Arcones, M. and Giné, E. (1992). On the bootstrap of M-estimators and other statistical functionals. In *Exploring the limits of the bootstrap* (R. LePage and L. Billard eds.), pp. 13-48, Wiley.

[6] Brown, B.W. (1983). The identification problem in systems nonlinear in the variables. *Econometrica*, **51**, 175–196.

[7] Brown, D.J. and Matzkin, R. (1998). Estimation of Nonparametric functions in simultaneous equations models, with an application to consumer demand. *Working paper, Yale University*

[8] Davidson, R. and Mackinnon, J. (1993). *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

[9] Dudley, R. (1999). *Uniform Central limit theorems*, Cambridge.

[10] Giné, E. & Zinn, J. (1990). Bootstrapping general empirical measures. *Annals of probability*, **18**, 851 – 869.

[11] Hoffmann-Jørgensen, J.(1994). *Probability with a view toward statistics: Volume II*, Chapman & Hall, New York.

[12] Manski, C.F. (1983). Closest empirical distribution estimation. *Econometrica*, **51**(2), 305 – 320.

[13] Mas-Colell, A. (1985). *The theory of general economic equilibrium: a differential approach*, Cambridge University Press, Cambridge.

[14] Newey, W. and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *In Handbook of Econometrics, Volume 4, Engle and McFadden Eds.*, Elsevier, Amsterdam.

[15] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimaization estimators. *Econometrica*, **57**(5), 1027 –1057.

[16] Pollard, D. (1984). *Convergence of empirical processes.* Springer Verlag. New-York.

[17] Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory*, **1**, 295 – 314.

[18] Pollard, D. (forthcoming). *Asymptopia*

[19] Roehrig, C.S. (1988). Conditions for identification in nonparametric and paramteric models. *Econometrica*, **56**, 433 – 447.

[20] Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*, Springer.

[21] van der Vaart, A. (1999) *Asymptotic Statistics*, Cambridge.

[22] van der Vaart, A. & Wellner, J. (1996). *Weak convergence and empirical processes*, Springer.

[23] Wegkamp, M. (1995). Asymptotic results for parameter estimation in general empirical processes. *Tech. Report TW9504*, University of Leiden.

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