COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1252

ASYMPTOTICS IN MINIMUM DISTANCE FROM INDEPENDENCE ESTIMATION

Donald J. Brown and Marten H. Wegkamp

April 2000

Asymptotics in Minimum Distance from Independence Estimation

Donald J. Brown and Marten H. Wegkamp*

April 2000

Abstract

In this paper we introduce a family of minimum distance from independence estimators, suggested by Manski's minimum mean square from independence estimator. We establish strong consistency, asymptotic normality and consistency of resampling estimates of the distribution and variance of these estimators. For Manski's estimator we derive both strong consistency and asymptotic normality.

1 Introduction

Manski (1983) introduced minimum distance from independence estimators in his paper on CED estimation. In particular, he considered minimum mean square distance from independence estimation of semi-parametric implicit simultaneous equations models, the central topic of this paper. He proved strong consistency of his estimator, but was unable to derive the first-order asymptotic distribution. At the end of his paper, he conjectured that the theory of empirical processes might prove useful in deriving the asymptotic distributions of CED estimators.

For the case of the minimum mean square distance from independence estimator, we show that his conjecture is correct. That is, using the methods of empirical process theory we give a new proof of the strong consistency of Manski's estimator under less stringent assumptions and demonstrate the normality of the asymptotic distribution.

Recently, Brown and Matzkin (1998) extended Manski's analysis to non-parametric simultaneous equations models where they proposed to minimize the Prohorov distance from independence. They also were unable to derive the limiting distribution of their estimator, but they did derive a necessary and sufficient condition for identification in nonlinear simultaneous equations models, based on earlier work of Brown (1983) and Roehrig (1988). This condition plays a significant role in our applications. Moreover, Brown and Matzkin proposed a random utility model of consumer choice over continuous alternatives. The first-order conditions for the consumer's optimization problem subject to her budget constraint, where prices and incomes are exogenous random variables, constitute the structural equations for this model. If

*The authors thank Don Andrews and David Pollard for their helpful comments and remarks.

the random utility function $V(y,\varepsilon) = U(y) + \varepsilon \cdot y$, then they derived necessary and sufficient conditions on the family of permissible deterministic utility functions $U(y)$ to identify their model assuming ε and $x = (p, I)$ are stochastically independent, where p is the vector of commodity prices and I is the consumer's income.

For $U(\cdot)$ parameterized by a compact subset of Euclidean space, we extend their analysis to pure trade models. Here the structural equations consist of the first-order conditions of each agent in the economy, each of whom has a random utility function of the form specified by Brown and Matzkin, and the market clearing conditions. In this model prices and consumptions are now endogenous variables, i.e., $y = (y_A, y_B, p)$, and the random observable individual endowments and incomes are the exogenous variables, i.e., $x = (\omega_A, \omega_B, I_A, I_B)$. The intended use of this model is in the field of applied general equilibrium theory where the models are estimated rather than calibrated, see Mansur and Whalley (1984) for an extended discussion. This model appears in the next section of the paper.

Manski's estimator, although computationally quite tractable, is difficult to analyze theoretically. Hence prior to our discussion of the asymptotic properties of Manski's estimator, we introduce another family of minimum distance from independence estimators which are easier to analyze and are also computationally attractive. We show strong consistency, asymptotic normality and consistency of the resampling estimates of the sampling distribution and variance of these estimators. For Manski's estimator we derive both strong consistency and asymptotic normality.

The main tools in our analysis are techniques derived from the theory of empirical processes. 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 application of our results is the estimation of a continuous random choice model as in Brown and Matzkin. Two significant differences between our paper and the Pakes and Pollard paper are that our estimator is an extremum estimator, i.e., we minimize a random criterion function, and their estimator is a Z-estimator, i.e., they approximately solve a family of random equations. More importantly, Theorem 7 (Wegkamp (1995, 1999)) employed here subsumes as special cases: M-estimation, Cramer–Von Mises estimation, regression and minimum distance from independence estimation, see 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).

2 Estimating a Simple Pure Trade Model

In applied general equilibrium analysis, there are two methods for determining parameter values: calibration and econometric estimation. The latter method, although theoretically more appealing, suffers from a number of limitations. In particular, the random shocks to tastes and technology enter the model in an ad hoc fashion, i.e.,

in most cases they are simply added to reduced forms of the deterministic structural equations, such as demand or supply functions. In addition, given the nonlinear nature of the structural equations, assumptions of model identification are problematic. In fact, as pointed out by Mansur and Whalley (1984) in their survey article, these issues have not been successfully resolved even for simple textbook models of general equilibrium such as the pure trade model, the Robinson Crusoe model, or the two-sector model. Surprisingly, to our knowledge, this is still the case.

In this section of our paper, we consider a simple pure trade model with two countries, where the tastes of each country is characterized by a random utility function, representing the distribution of tastes within the country. The analysis is partial equilibrium in that the random utility functions $V(y, y_0, \varepsilon) = U(y) + y_0 + \varepsilon \cdot y$ are quasi-linear with a random linear perturbation.

The assumption of quasi-linearity plays a number of roles in our analysis. Most importantly, this specification gives rise to monotone individual demand functions for fixed realizations of ε, see Quah (1999) for discussion. If we posit a distribution economy where the income distribution is fixed, then monotonicity of individual demand implies monotonicity of aggregate demand, a sufficient condition for uniqueness of the equilibrium price vector, see Hildenbrand (1994). This uniqueness of equilibrium price vectors is an essential ingredient in our proof of identification.

Let us denote the two countries as A and B and the aggregate endowment in the world as (ω, ω_0). Then the countrywide endowments are $(\omega_A, \omega_{0A}) = \alpha_A(\omega, \omega_0)$ and $(\omega_B, \omega_{0B}) = \alpha_B(\omega, \omega_0)$, where $\alpha_A, \alpha_B > 0$ and $\alpha_A + \alpha_B = 1$. We now normalize prices (p, p_0) such that $(p, p_0) \cdot (\omega, \omega_0) = 1$.

The observable exogenous random variables are (ω, ω_0). The unobservable exogenous random variables are ε_A and ε_B, the random shocks to tastes. The observable endogenous random variables are the equilibrium price vector (p, p_0) and the consumptions of country A, (y_A, y_{0A}). α_A and α_B are deterministic and fixed.

As noted earlier these assumptions are sufficient for uniqueness of the equilibrium price vector, conditional on the realizations of $\varepsilon = (\varepsilon_A, \varepsilon_B)$ and (ω, ω_0), but they limit our ability to identify each country's characteristics, i.e., $\langle U_A, f_{\varepsilon_A} \rangle$ and $\langle U_B, f_{\varepsilon_B} \rangle$ where f_{ε_A} and f_{ε_B} are the distributions of ε_A and ε_B, respectively, since (ω_A, ω_{0A}) and (ω_B, ω_{0B}) are dependent, i.e., linearly related. Hence we assume that $U_A = U_B$. That is, each country has the same quasi-linear location function, but the distribution of tastes about the location function in each country may differ.

The remaining assumptions follow those of Brown and Matzkin, except we assume that the quasi-linear utility functions under consideration are parameterized by a compact subset of $\mathbb{R}^L$, with nonempty interior, denoted Θ. All distributions have smooth densities and their supports are in the positive orthants of the relevant Euclidean spaces. The final identifying assumption is that $\varepsilon = (\varepsilon_A, \varepsilon_B)$ is stochastically independent of (ω, ω_0).

We now proceed to show that this model is identified, i.e., if $\theta \neq \widetilde{\theta}$ then the resulting distributions of data are unequal. The structural equations can be expressed in terms of each country's F.O.C.'s for utility maximization subject to their budget constraints. We use the market clearing conditions to express the F.O.C.'s for country

B in terms of country A's consumptions.

Structural Equations

$$\varepsilon_A = p/p_0 - DU(y_A) \tag{1}$$

$$\varepsilon_B = p/p_0 - DU(\omega - y_A) \tag{2}$$

$$y_{0A} = (\alpha_A - p \cdot y_A)/p_0 \tag{3}$$

$$y_{0B} = (\alpha_B - p \cdot (\omega - y_A))/p_0 \tag{4}$$

We can solve these equations in two steps, because of the assumption of quasi-linear utility functions. First, we solve (1) and (2) for $q = p/p_0$ and y_A. Then we substitute these values into the budget constraints, (3) and (4), to solve for y_{0A} and p_0.

Hence the relevant structural equations for our estimation procedure are

$$\varepsilon_A = q - DU(y_A) \tag{5}$$

$$\varepsilon_B = q - DU(\omega - y_A) \tag{6}$$

This is a system of $2k$ equations in $2k$ unknowns, q and y_A, with $2k$ unobserved random variables $\varepsilon = (\varepsilon_A, \varepsilon_B)$. We write this system as $\varepsilon = g(q, y_A, \omega, \theta)$ where θ indexes U. The standard assumptions on U that it is smooth, strictly concave and monotone with interior optima on budget sets, together with our earlier assumptions, guarantee the existence of a unique smooth function $h(\varepsilon, \omega, \theta) = (q, y_A)$ such that $\varepsilon \equiv g(h(\varepsilon, \omega, \theta), \omega, \theta)$. Brown and Matzkin (1998) have shown the following necessary and sufficient condition for identification: If $\theta, \widetilde{\theta} \in \Theta$ and $\theta \neq \widetilde{\theta}$ then

$$\frac{\partial h(\varepsilon, \omega, \theta)}{\partial \omega} \neq \frac{h(\widetilde{\varepsilon}, \omega, \widetilde{\theta})}{\partial \omega} \text{ where } \widetilde{\varepsilon} = g(h(\varepsilon, \omega, \theta), \omega, \widetilde{\theta}).$$

Applying the implicit function theorem to the structural equations (1) and (2), we deduce that

$$\frac{\partial h(\varepsilon, \omega, \theta)}{\partial \omega} = \begin{bmatrix} I & -D^2U(y_A) \\ I & D^2U(\omega - y_A) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ D^2U(\omega - y_A) \end{bmatrix}$$

Our assumptions of smoothness and strict concavity of $U(\cdot)$ guarantee that $D^2U(y_A)$ is negative definite, hence invertible. Moreover, these assumptions guarantee that $[D^2U(y_A) + D^2U(\omega - y_A)]^{-1}$ exists.

Let $R = D^2U(y_A)$ and $S = D^2U(\omega - y_A)$, then

$$\begin{bmatrix} I & -R \\ I & S \end{bmatrix}^{-1} = \begin{bmatrix} S(R+S)^{-1} & R(R+S)^{-1} \\ -(R+S)^{-1} & (R+S)^{-1} \end{bmatrix}$$

and

$$\begin{aligned} \begin{bmatrix} I & -R \\ I & S \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ S \end{bmatrix} &= \begin{bmatrix} S(R+S)^{-1} & R(R+S)^{-1} \\ -(R+S)^{-1} & (R+S)^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ S \end{bmatrix} \\ &= \begin{bmatrix} R(R+S)^{-1}S \\ (R+S)^{-1}S \end{bmatrix}. \end{aligned}$$

Therefore $\partial h(\varepsilon,\omega,\theta)/\partial\omega = \partial h(\widetilde{\varepsilon},\omega,\widetilde{\theta})/\partial\omega$ iff

(i) $R(R+S)^{-1}S = \widetilde{R}(\widetilde{R}+\widetilde{S})^{-1}\widetilde{S}$ and

(ii) $(R+S)^{-1}S = (\widetilde{R}+\widetilde{S})^{-1}\widetilde{S}$

(i) and (ii) imply $(R-\widetilde{R})(R+S)^{-1}S = 0$. Since $(R+S)^{-1}S$ is nonsingular we see that $R = \widetilde{R}$. We have proven the following theorem.

Theorem 1 *The structural equations $\varepsilon_A = q - DU(y_A)$ and $\varepsilon_B = q - DU(\omega - y_A)$ are identified iff for all $\theta, \widetilde{\theta} \in \Theta$ if $\theta \neq \widetilde{\theta}$ then $\exists \overline{y}_A$ such that $D^2U(\overline{y}_A) \neq D^2\widetilde{U}(\overline{y}_A)$.*

One obvious example of a family of utility functions with this property are Cobb–Douglas utility functions.

3 Identification

Identification is a necessary condition for the criterion function of extremum estimators to have a unique global optimum. In our case it is also sufficient. Compactness of Θ and the continuity of our criterion functions imply that the optimum is well separated.

Minimum distance from independence estimators, as defined by Manski (1983), are extremum estimators where the criterion function $\gamma(\cdot,\cdot)$ is a metric on the space of joint distributions of (x,ε). Hence $\gamma(f(x,\varepsilon), f(x)f(\varepsilon)) = 0$ iff x and ε are stochastically independent, where $f(x,\varepsilon)$ is the joint distribution of (x,ε) and $f(x)$, $f(\varepsilon)$ are the associated marginal distributions. Our discussion of identification of semiparametric implicit simultaneous equations models follows the expositions of Brown (1983), Roehrig (1988) and Brown–Matzkin (1998).

A structure S is an ordered pair $\langle g(x,y,\theta), f(x,\varepsilon)\rangle$. The structural equations are defined as $\varepsilon = g(x,y,\theta)$. We define a model M as an indexed family of structures $\{S_i\}_{i\in I}$, where $S_i = \langle g(x,y,\theta_i), f_i(x,\varepsilon)\rangle$ and $\Theta = \{\theta_i\}_{i\in I}$. All structures $S = \langle g(x,y,\theta), f(x,\varepsilon)\rangle$ satisfy the following assumptions:

Assumption 1 $\exists!$ *reduced form $y = h(x,\varepsilon,\theta)$ such that*

$$\varepsilon \equiv g(x, h(x,\varepsilon,\theta), \theta).$$

Assumption 2 *$\partial g/\partial y$ has full rank a.e.*

Assumption 3 *$f(x,\varepsilon) = f(x)f(\varepsilon)$, i.e., x and ε are stochastically independent.*

The following definitions will prove useful:

Definition 1 *If $(x,\varepsilon) \sim f(x,\varepsilon)$ then $(x, h(x,\varepsilon,\theta)) \sim f_\theta(x,y)$.*

Definition 2 *If $S_0 = \langle g(x,y,\theta_0), f_0(x,\varepsilon)\rangle$ and $S_1 = \langle g(x,y,\theta_1), f_1(x,\varepsilon)\rangle$ then S_0 and S_1 are observationally equivalent if $f_0(x,y) = f_1(x,y)$ a.e.*

Definition 3 *A structure S_0 is identifiable in M if there is no other structure in M that is observationally equivalent to S_0.*

Definition 4 $\varepsilon_{0,i} \equiv g(x, h(x, \varepsilon, \theta_0), \theta_i)$

Definition 5 $S_{0,i} \equiv \langle g(x, y, \theta_i), f_{0,i}(x, \varepsilon_{0,i}) \rangle$.

Theorem 2 (Brown–Roehrig) *S_0 is observationally equivalent to S_i iff $f_i(x, \varepsilon) = f_{0,i}(x, \varepsilon_{0,i})$* a.e.

Theorem 3 (Brown–Roehrig) $f_i(x, \varepsilon) = f_{0,i}(x, \varepsilon_{0,i})$ a.e. *iff*

$$\frac{\partial g(x, h(x, \varepsilon, \theta_0), \theta_i)}{\partial x} = 0 \quad \text{a.e.}$$

Theorem 4 (Brown–Matzkin) $\partial g(x, h(x, \varepsilon, \theta_0), \theta_i)/\partial x = 0$ a.e. *iff*

$$\frac{\partial h(x, \varepsilon, \theta_0)}{\partial x} = \left. \frac{\partial h(x, \varepsilon, \theta_i)}{\partial x} \right|_{\varepsilon = \varepsilon_{0,i}} \quad \text{a.e.}$$

Theorem 5 (Brown–Matzkin) *If S_0 is identifiable in M, then $f_0(x, \varepsilon)$ is the unique global minimum of $\gamma(f_{0,i}(x, \varepsilon_{0,i}))$, $f_{0,i}(\varepsilon_{0,i}) f_0(x))$ over $i \in I$.*

Proposition 1 *If Θ is compact and $M(\theta_i) = \gamma(f_{0,i}(x, \varepsilon_{0,i}))$, $f_{0,i}(\varepsilon_{0,i}) f_0(x))$ is continuous in θ, then the unique global minimum of $M(\theta_i)$ over $i \in I$ is well separated.*

4 Consistency and Asymptotic Normality of a General Class of Estimators

Under suitable regularity assumptions, we derive a strongly consistent estimator of θ_0 and we obtain its limiting sampling distribution. Before stating our results, we need some notation.

The parameter space Θ is a subset of $\mathbb{R}^k$, and $g : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}^{d'}$. Let F, G_θ and H_θ be the cumulative distribution functions of X, $g(Z, \theta)$ and $(X, g(Z, \theta))$, respectively. Let F_n, $G_{n\theta}$ and $H_{n\theta}$ be their empirical counterparts based on the data $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$. Define

$$D_\theta(s, t) = H_\theta(s, t) - F(s) G_\theta(t), \tag{7}$$

and

$$D_{n\theta}(s, t) = H_{n\theta}(s, t) - F_n(s) G_{n\theta}(t). \tag{8}$$

Motivated by the independence between X and $\varepsilon = g(Z, \theta_0)$, we propose to minimize the empirical criterion

$$M_n(\theta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} D^2_{n\theta}(s, t) \, d\mu(s, t) \tag{9}$$

for some bounded measure μ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}^{d'}$. The resulting minimizer is denoted by $\hat{\theta}$, i.e.,

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

We will assume without loss of generality that a minimizer 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.

The theoretical counterpart of M_n,

$$M(\theta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} D_\theta^2(s,t)\, d\mu(s,t) \tag{10}$$

is minimized for $\theta = \theta_0$, and $M(\theta_0) = 0$ by the aforementioned independence between X and $g(Z, \theta_0)$.

The first result, Theorem 6, states two general conditions for which this procedure yields a consistent estimate. We need that the deterministic function M has a well separated unique minimum at θ_0. Proposition 3.5 shows that this is the case provided the mapping $\theta \mapsto D_\theta(s,t)$ is continuous μ-a.e. and the parameter space Θ is compact. The stochastic assumption needed for the consistency, to wit, the uniform convergence of $D_{n\theta}$ to D_θ is met if the class

$$\mathcal{A} = \{\{z \in \mathcal{Z} : g(z,\theta) \leq t\},\ \theta \in \Theta, t \in \mathcal{Y}\} \tag{11}$$

is a P-Glivenko–Cantelli class, where P is the probability measure of Z. This in turn is satisfied if the collection $\{g(\cdot,\theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space, or if the functions $g(z,\theta)$ are smooth in z. We will specify the type of smoothness later, but first we establish the following lemma.

Lemma 1 *Suppose that the collection $\mathcal{A}$ is a P-Glivenko–Cantelli class. Then*

$$\sup_{\theta \in \Theta} |D_{n\theta}(s,t) - D_\theta(s,t)| \xrightarrow{a.s.} 0,$$

uniformly in $s \in \mathcal{X}$ and $t \in \mathcal{Y}$.

Proof Rewrite the difference

$$\begin{aligned} D_{n\theta}(s,t) - D_\theta(s,t) \;=\; & (H_{n\theta}(s,t) - H_\theta(s,t)) + \\ & + F_n(s)\{G_\theta(t) - G_{n\theta}(t)\} + G_\theta(t)\{F(s) - F_n(s)\}. \end{aligned}$$

The last term tends to zero by the Glivenko–Cantelli theorem, uniformly in s and t, and the fact that $|G_\theta(t)| \leq 1$. For the second term, we observe that

$$G_{n\theta}(t) - G_\theta(t) = (P_n - P)A_{\theta,t},$$

where P_n is the empirical measure putting mass $1/n$ at each observation Z_i, $P = P_Z$ is the probability measure of $Z = (X,Y)$ and the set $A_{\theta,t} \in \mathcal{A}$. The Glivenko–Cantelli

property of $\mathcal{A}$ guarantees the law of large numbers $P_n(A) \to_{\text{a.s.}} P(A)$, uniformly in $A \in \mathcal{A}$, so that indeed

$$\sup_{\theta,t} |G_{n\theta}(t) - G_\theta(t)| = \sup_\theta |(P_n - P)A_{\theta,t}| \to 0.$$

as the VC property yields the uniform law of large numbers. For the first term, we can use the same reasoning as in the above, by simply noting that

$$H_{n\theta}(s,t) - H_\theta(s,t) = (P_n - P)(A_{\theta,t} \cap (B_s \times \mathcal{Y})),$$

where $B_s = \{x \in \mathcal{X} : x \leq s\}$. Because $\{B_s : s > 0\}$ is a P-Donsker class, and the Glivenko–Cantelli property is preserved under pairwise products, it follows that the collection $\{A_{\theta,t} \cap (B_s \times \mathcal{Y}) : \theta \in \Theta, s \in X, t \in \mathcal{Y}\}$ is P-Glivenko–Cantelli as well and consequently

$$\sup_{\theta \in \Theta} |H_{n\theta}(s,t) - H_\theta(s,t)| \xrightarrow{\text{a.s.}} 0.$$

The proof is complete. ■

Corollary 1 *Suppose that $\{g(\cdot,\theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space. Then $\mathcal{A}$ is P-Donsker (and hence P-Glivenko–Cantelli) for all probability measures P.*

Proof Since $t = (t_1, \ldots, t_d)^T$ and $g(z,\theta) = (g_1(z,\theta), \ldots, g_d(z,\theta))^T$, we can write $A_{\theta,t}$ as an intersection

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

where

$$A^{(i)}_{\theta,t} = \{z \in \mathcal{Z} : g_i(z,\theta) \leq t_i\}.$$

It is well known that $\{A^{(i)}_{\theta,t} : \theta \in \Theta\}$ is a VC-class of sets if $\{g_i(z,\theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space, see Pakes and Pollard (1989), Lemma 2.4, p. 1031. Actually, in that case the stronger result that $\{A^{(i)}_{\theta,t} : \theta \in \Theta, t \in \mathcal{Y}\}$ is a VC class, is true. Hence all $\{A^{(i)}_{\theta,t} : \theta \in \Theta, t \in \mathcal{Y}\}$ are VC classes for $i = 1, \ldots, d$. The VC property is closed under intersections, so that $\mathcal{A}$ forms a VC-class of sets, and hence is P-Donsker. ■

The following corollary states that smooth functions typically satisfy the Donsker property as well. Although we only need $\mathcal{A}$ to be Glivenko–Cantelli, rather than Donsker, we need the Donsker property later for the asymptotic normality result. First we need some notation.

For every $k = (k_1, \ldots, k_n) \in \mathbb{N}^n$, define the differential operator D^k by

$$D^k = \frac{\partial^{k_1 + \ldots + k_n}}{\partial x_1^{k_1} \ldots \partial x_n^{k_n}}.$$

Let $\mathcal{C}_M^{p+\alpha}$ be the class of real valued, continuous functions on the the unit cube S^n in $\mathbb{R}^n$ possessing uniformly bounded partial derivatives of order $k \leq p$, i.e., for some constant C_1 independent of f,

$$\max_{k_1+\ldots+k_n \leq p} \max_{x \in S_n} |D^k f(x)| \leq C_1.$$

Moreover the p-th order partial derivatives of each f satisfy a Lipschitz condition of order α $(0 < \alpha \leq 1)$, i.e. there exists a $C_2 > 0$ independent of f such that

$$|D^k f(x) - D^k f(y)| \leq C_2 \|x - y\|^\alpha$$

for all $x, y \in S^n$ and all $k \in \mathbb{N}^n$ with $k_1 + \ldots + k_n = p$. The constants $C_1 + C_2 \leq M$. In our application the dimension n equals $d + d' = \dim(\mathcal{Z})$, and 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 assumptions of the next result.

Corollary 2 *Suppose that $\mathcal{Z} \subset [0,1]^{d+d'}$ with nonempty interior and that each coordinate mapping of $g(z,\theta) \in C_M^\alpha(\mathcal{Z})$. Then the collection $\mathcal{A}$ is P-Donsker for all probability measures P with an uniformly bounded density and $\alpha > d + d'$.*

Proof Corollary 2.7.3 in Van der Vaart and Wellner bounds the entropy of bracketing of the collection of subgraphs of C_1^α. For $\alpha > d + d'$, the bracketing central limit theorem implies that this collection of subgraphs is P-Donsker. Example 2.10.8 in van der Vaart and Wellner states that pairwise products of uniformly bounded Donsker classes is again Donsker, so that the collection $\mathcal{A}$ is P-Donsker. ■

Now we are in the position to state our consistency result.

Theorem 6 *Suppose that $M(\theta)$ has a well separated unique minimum at $\theta = \theta_0$. Furthermore, assume that $D_{n\theta}(s,t) \to D_\theta(s,t)$, almost surely, uniformly in θ. Then $\hat{\theta} = \arg\min_{\theta \in \Theta} M_n(\theta) \to_{a.s.} \theta_0$.*

Proof Using the minimization properties of θ_0 and $\hat{\theta}_n$, we have

$$M_n(\hat{\theta}_n) \leq M_n(\theta_0) + \mathcal{O}_P(n^{-1}) \xrightarrow{P} M(\theta_0) \leq M(\hat{\theta}_n). \tag{12}$$

We will show below that

$$\sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{P} 0, \tag{13}$$

and consequently, $|M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| \to_P 0$. But this in combination with (12) yields that $M(\hat{\theta}) \to_P M(\theta_0)$. Since M has a unique, well separated minimum at θ, we must have that $\hat{\theta} \to_P \theta_0$. This is a fairly standard argument and can be found in Van

der Vaart (1999). We will now establish the uniform convergence (13). But this is immediate from

$$\begin{aligned} |M_n(\theta) - M(\theta)| &= |\mu[D_{n\theta} - D_\theta + D_\theta]^2 - \mu[D_\theta]^2| \\ &\leq \mu[D_{n\theta} - D_\theta]^2 + 2|\mu[D_{n\theta} - D_\theta]D_\theta| \\ &\leq 4\mu|D_{n\theta} - D_\theta|, \end{aligned}$$

since both $|D_\theta| \leq 1$ and $|D_{n\theta}| \leq 1$. By assumption and dominated convergence, the term on the right converges almost surely to zero, uniformly in θ. ■

Before establishing the limit distribution of $\sqrt{n}(\hat{\theta}-\theta_0)$, we repeat some definitions.

Definition 6 (stochastic equicontinuity) *Let (T, d) be a pseudo-metric space and let $\{Z_n(t) : t \in T\}$ be a stochastic process indexed by T. A sequence $\{Z_n\}$ is called stochastically equicontinuous at $t_0 \in T$ iff for every positive η and ε there exists a neighborhood V of t_0 for which*

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{t\in V} |Z_n(t) - Z_n(t_0)| > \eta\right\} \leq \varepsilon.$$

Equivalently, $\{Z_n\}$ is stochastically equicontinuous at $t_0 \in T$ iff for any $\tau_n \to_P t_0$, we have $|Z_n(\tau_n) - Z_n(t_0)| \to_P 0$.

The key to establishing asymptotic normality for the estimators that we consider is the notion of stochastic differentiability.

Definition 7 (stochastic differentiability) *Let $\{Z_n(t) : t \in T\}$ be a stochastic process, indexed by $T \subset \mathbb{R}^k$. A sequence Z_n is called* stochastically differentiable *at $t_0 \in T$ with derivative W_n iff $\forall \eta > 0$ and $\forall \varepsilon > 0$ there exists a neighborhood V of t_0 for which*

$$\limsup_{n\to\infty} \mathbb{P}\left\{\sup_{t\in V} \left|\frac{Z_n(t) - Z_n(t_0) - (t-t_0)'W_n}{|t-t_0|}\right| > \eta\right\} \leq \varepsilon.$$

Equivalently, a sequence Z_n is called stochastically differentiable with derivative W_n at $t_0 \in T$ iff for any $\tau_n \to_P t_0$, we have $|Z_n(\tau_n) - Z_n(t_0) - (\tau_n - t_0)'W_n| = o_P(|\tau_n - t_0|)$.

We will require that in our particular case, $\sqrt{n}(D_{n\theta}-D_\theta)$ is stochastically equicontinuous at θ_0. A sufficient condition is given in the following lemma.

Lemma 2 *Suppose that $\mathcal{A}$ is a P-Donsker class, $\theta \mapsto g(z, \theta)$ is Lipschitz at θ_0, then $\sqrt{n}(D_{n\theta} - D_\theta)(s, t)$ is stochastically equicontinuous at $\theta = \theta_0$.*

Proof Again we write the decomposition

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

This is a sum of three terms, each stochastically equicontinuous at θ_0. We also apply Slutsky's lemma for the $F_n(x)$ in front of the second term, and continuity of $\theta \mapsto G_\theta$ at θ_0 — which follows from the Lipschitz condition on $\theta \mapsto g(z,\theta)$ — in front of the last term. First we observe that the class $\{A(\theta,x,\varepsilon) : \theta \in \Theta\}$ is a P-Donsker class, where

$$A(\theta,x,\varepsilon) = \{z = (z_1,z_2) \in \mathcal{Z} : g(z,\theta) \leq \varepsilon, z_1 \leq x\}.$$

Hence, general empirical process theory dictates that

$$\sqrt{n}(H_{n\theta} - H_\theta)(x,\varepsilon) = \sqrt{n}(P_n - P)(A(\theta,x,\varepsilon)),$$

weakly converges to a tight Gaussian limit, and in particular, the process is stochastically equicontinuous at θ_0. The stochastic equicontinuity is wrt the $L_2(P)$ metric on these sets, *not* the Euclidean distance on Θ. This point is also discussed in Pakes and Pollard (1989) after Lemma 2.16 on p. 1036. That is, we know that for all $\varepsilon, \eta > 0$ there exists a $\delta > 0$ such that

$$\mathbb{P}\left\{ \sup_{P(A_{\theta,x,t} - A_{\theta_0,x,t})^2 \leq \delta} |\sqrt{n}(P_n - P)A_{\theta,x,t} - \sqrt{n}(P_n - P)A_{\theta_0,x,t}| \geq \varepsilon \right\} \leq \eta.$$

By the assumed Lipschitz condition of $\theta \mapsto g(z,\theta)$, we find that

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

So indeed,

$$\sqrt{n}[D_{n\theta_n}(x,t) - D_{\theta_n}(x,t)] - \sqrt{n}D_{n\theta_0}(x,t) \xrightarrow{P} 0,$$

for all $\theta_n \to_P \theta_0$. This concludes the proof. ■

The proof for the asymptotic normality is complicated by the fact that our estimator is not an M-estimator so that we cannot appeal to the standard theory for this large class of estimators (cf. Pollard (1985)). However, Wegkamp (1995, 1999) extended these results to a broader class of estimators which minimize a random criterion. Tailored to our application, it reads as follows.

Theorem 7 *Suppose that*

$\theta_0 = \arg\min_{\theta\in\Theta} M(\theta)$ *lies in the interior of* Θ*;*

M *is twice differentiable at* θ_0 *with a non-singular second derivative* V *at* θ_0*;*

$\hat{\theta} \to_P \theta_0$*;*

$\alpha_n = \sqrt{n}(M_n - M)$ *is stochastically differentiable with derivative* W_n *at* θ_0*, i.e.,*

$$\alpha_n(\theta) = \alpha_n(\theta_0) + (\theta - \theta_0)'W_n + o_P(|\theta - \theta_0|) \text{ for all } \theta \xrightarrow{P} \theta_0,$$

then $\hat{\theta} = \theta_0 - V^{-1}n^{-1/2}W_n + o_P(n^{-1/2})$.

Proof See Andrews(1997, 1999), Pollard(forthcoming), or Wegkamp(1995, 1999).■

We conclude with our main result, the asymptotic normality of the estimator.

Theorem 8 *Suppose that*

*(*A1*)* θ_0 *lies in the interior of* Θ*,*

*(*A2*)* $\hat{\theta} \to_P \theta_0$*,*

*(*A3*) There exists a vector* $\Delta \in L_2(\mu)$ *(coordinatewise) such that*

$$D_\theta(\cdot) = (\theta - \theta_0)^T \Delta(\cdot) + \|\theta - \theta_0\| R_\theta(\cdot),$$

where $\mu R^2(\theta, \cdot) \to 0$ *as* $\theta \to \theta_0$*,*

*(*A4*)* $\sqrt{n}(D_{n\theta} - D_\theta)(s,t)$ *is stochastically equicontinuous at* $\theta = \theta_0$*.*

Then

$$\hat{\theta} = \theta_0 - 2V^{-1} \int \Delta (D_{n,\theta_0} - D_{\theta_0})\, d\mu + o_P(n^{-1/2}), \tag{14}$$

where $V = \mu \Delta \Delta^T$*.*

Before proving this result, we deduce the asymptotic normality from the stochastic expansion (14). For this argument, recall that

$$\begin{aligned} D_{n\theta_0} - D_{\theta_0} \;=\; & D_{n\theta_0} = [H_{n\theta_0}(s,t) - H_{\theta_0}(s,t)] \\ & + F_n(s)[G_{\theta_0}(t) - G_{n\theta_0}(t)] + G_{\theta_0}(t)[F(s) - F_n(s)]. \end{aligned}$$

As a result, the asymptotic normality follows by an application of Donsker's theorem, the continuous mapping theorem (cf. Van der Vaart and Wellner (1996), Theorem 1.3.6, page 20) and Slutsky's lemma, since $\sqrt{n}[H_{n\theta_0}(s,t) - H_{\theta_0}(s,t)]$, $\sqrt{n}[G_{\theta_0}(t) - G_{n\theta_0}(t)]$, and $\sqrt{n}[F(s) - F_n(s)]$ all converge to Gaussian processes. We now prove Theorem 8.

Proof We simply verify the conditions of the preceding Theorem 7. First, by condition A3

$$M(\theta) = (\theta - \theta_0)^T \mu \Delta \Delta^T (\theta - \theta_0) + o(\|\theta - \theta_0\|^2) \text{ for } \theta \to \theta_0.$$

Notice that the matrix $V = \mu \Delta \Delta^T$ is positive definite. It remains to show the stochastic differentiability of $\sqrt{n}(M_n - M)$ at θ_0. Observe that by A4 we have for all $\theta \to_P \theta_0$

$$|D_{n\theta} - D_\theta) - (D_{n\theta_0} - D_{\theta_0})| = o_P(n^{-1/2})$$

so that the continuous mapping theorem implies

$$\mu |(D_{n\theta} - D_\theta) - (D_{n\theta_0} - D_{\theta_0})| = o_P(n^{-1/2}),$$

since μ is a bounded measure. Using A4, Donsker's theorem, the continuous mapping theorem and the Cauchy-Schwarz and Markov inequalities repeatedly, we obtain the following string of order calculations which hold uniformly in a neighborhood of θ_0

$$\begin{aligned}\mu(D_{n\theta} - D_\theta)^2 - \mu D_{n\theta_0}^2 &= \mu\{(D_{n\theta} - D_\theta) - D_{n\theta_0}\}^2 + 2\mu D_{n\theta_0}\{(D_{n\theta} - D_\theta) - D_{n\theta_0}\} \\ &= \mathcal{o}_P(1/n),\end{aligned}$$

and

$$\mu\Delta(D_{n\theta_0} - D_{\theta_0}) = \mathcal{o}_P(n^{-1/2}),$$

$$\begin{aligned}\mu D_\theta[D_{n\theta} - D_\theta] &= \mu(\theta - \theta_0)'\Delta[D_{n\theta} - D_\theta] + \mu||\theta - \theta_0||R_\theta[D_\theta - D_\theta] \\ &= \mu(\theta - \theta_0)'\Delta[D_{n\theta_0} - D_{\theta_0}] + \mathcal{o}_P(n^{-1/2}||\theta - \theta_0||)\end{aligned}$$

since

$$\begin{aligned}|\mu(\theta - \theta_0)'\Delta\{[D_{n\theta} - D_\theta] - D_{n\theta_0}\}| &\leq ||\theta - \theta_0||\sqrt{\mu\Delta\Delta'}\sqrt{\mu\{[D_{n\theta} - D_\theta] - D_{n\theta_0}\}^2} \\ &= \mathcal{o}_P(n^{-1/2}||\theta - \theta_0||)\end{aligned}$$

and

$$\begin{aligned}|\mu\ ||\theta - \theta_0||R_\theta[D_{n\theta} - D_\theta]| &\leq ||\theta - \theta_0||\sqrt{\mu R_\theta^2}\sqrt{\mu[D_{n\theta} - D_\theta]^2} \\ &= \mathcal{o}_P(n^{-1/2}||\theta - \theta_0||)\end{aligned}$$

Combining all the above bounds we obtain

$$\begin{aligned}&M_n(\theta) - M(\theta) \\ &= \mu[D_\theta + D_{n\theta} - D_\theta]^2 - \mu[D_\theta]^2 \\ &= M_n(\theta_0) - M(\theta_0) + 2\mu D_\theta[D_{n\theta} - D_\theta] + \mathcal{o}_P(1/n) \\ &= M_n(\theta_0) - M(\theta_0) + 2(\theta - \theta_0)^T\mu\Delta[D_{n\theta_0} - D_{\theta_0}] + \mathcal{o}_P(1/n) + \mathcal{o}_P(n^{-1/2}\|\theta - \theta_0\|)\end{aligned}$$

This shows that the process $\sqrt{n}(M_n(\theta) - M(\theta))$ is stochastically differentiable at θ_0. The proof follows from Theorem 7 with $W_n = \sqrt{n}\int \Delta D_{n\theta_0}\, d\mu$.

5 Consistency of Resampling Estimators

In this section we show that the bootstrap consistently estimates the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$, and that the delete -d jackknife consistently estimates the variance of linear combinations of $\hat{\theta}$. Let $Z_1^*, \ldots, Z_n^*$ be iid observations from P_n, and let M_n^* be the bootstrap counterpart of M_n based on the bootstrap sample. The resulting estimator minimizing M_n^* over θ is denoted by $\hat{\theta}^*$. It converges in probability to θ_0 provided M has a unique and well-separated minimum at θ_0 and the class $\{A(\theta, x, t) : \theta \in \Theta, x \in \mathcal{X}, t \in \mathcal{Y}\}$ is a Donsker class of sets, see Corollaries 1 and 2 for applications.

Theorem 9 *If M has a unique and well-separated minimum at θ_0 and the class $\{A(\theta, x, t) : \theta \in \Theta, x \in \mathcal{X}, t \in \mathcal{Y}\}$ is a Donsker class of sets, $\theta^* \to_P \theta_0$ in P_n-probability.*

Proof The Donsker property entails that both $\sup_\theta |(M_n - M)(\theta)| \to_{\text{a.s.}} 0$ and $\sup_\theta |(M_n^* - M_n)(\theta)| \to 0$ for almost all samples $z_1, \ldots, z_n$, see e.g. Giné and Zinn (1990). Consequently by the triangle inequality,

$$\sup_\theta |(M_n^* - M)(\theta)| \leq \sup_\theta |(M_n^* - M_n)(\theta)| + \sup_\theta |(M_n - M)(\theta)| \to 0 \text{ in } P_n \text{ probability.}$$

This entails the consistency of the bootstrap estimator by the same reasoning as in the proof of Theorem 6. ■

Theorem 10 *In addition to the conditions (A1),...,(A4) of Theorem 8, suppose that $\theta^* \to \theta_0$ in P_n probability, that $\sqrt{n}(D_{n\theta}^* - D_{n\theta})(s,t)$ is stochastically equicontinuous at θ_0. Then the distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ consistently estimates the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.*

Proof The proof closely follows the arguments of the result for M-estimators obtained by Arcones and Giné (1990). Observe that by similar arguments given in the proof of Theorem 8,

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

where $V = \mu\Delta\Delta'$. Let

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

and let $\theta = \hat{\theta}$ 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 more, we transform $\theta \mapsto V^{-1/2}\theta$ and $\Delta_n^* \mapsto V^{-1/2}\Delta_n^*$ and $\Delta_n \mapsto V^{-1/2}\Delta_n$, but we will suppress this in our notation (equivalently we assume without loss of generality that $V = I$). Hence

$$\begin{aligned}
M_n^*(\theta) - M_n(\eta) \;=\; & (\theta - \eta)^T(\Delta_n^* + \Delta_n) + \tfrac{1}{2}||\theta - \theta_0^2|| - \tfrac{1}{2}||\eta - \theta_0^2|| \\
& + \mathcal{O}_P(||\theta - \theta_0^2|| + ||\eta - \theta_0||^2 + n^{-1/2}||\theta - \theta_0|| + n^{-1/2}||\eta - \theta_0|| + n^{-1})
\end{aligned}$$

So that

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

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 that $\theta^* - \hat{\theta} = \Delta_n^* + o_P(n^{-1/2})$. 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. ■

The required stochastic equicontinuity of $\sqrt{n}(D_{n\theta}^* - D_{n\theta})$ can be verified by the following lemma.

Lemma 3 *Under the same conditions as Lemma 2, $\sqrt{n}(D_{n\theta}^* - D_{n\theta})(s,t)$ converges weakly to a tight Gaussian process.*

Proof This is a direct consequence of the Donsker property of the class $\{A(\theta,x,y)|\,\theta \in \Theta, x \in \mathcal{X}, y \in \mathcal{Y}\}$ and the bootstrap result for empirical processes due to Giné and Zinn (1990), which states that the empirical process can be bootstrapped if the class of functions which index the process is P-Donsker. Actually, they showed that this is a necessary and sufficient condition. ■

Estimation of the Variance

So far, we have not addressed the important issue of estimating the variance of $\hat{\theta}$. We will show that the delete $-d$ jackknife provides a consistent estimate for any linear combination $c'\hat{\theta}$. Before stating the result, we introduce some notation. Let $\hat{\theta}_{d,s}$ be the estimate based on the data set $X_i, i \in s$, 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, denoted by J_{-d}^2, is defined as

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

It is shown in Shao and Tu (1995, Theorem 2.10, page 52), that the stochastic expansion (cf. (12))

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

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

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

Hence it remains to establish uniform integrability of $|\sqrt{n}c'(\hat{\theta} - \theta_0)|^2$.

Lemma 4 *Under the same conditions as Theorem 8,* $|\sqrt{n}c'(\hat{\theta}-\theta_0)|^2$ *is uniformly integrable.*

Proof Let $M_n^c(\theta) = M_n(\theta) - M(\theta)$ be the centered process, and $\kappa > 0$ will denote a generic positive numerical constant. We first bound the tail probabilities as follows:

$$\begin{aligned}
\mathbb{P}\{||\sqrt{n}(\hat{\theta}-\theta_0)|| \geq \delta\} &\leq \mathbb{P}\left\{\sup_{\|\theta-\theta_0\|\geq\delta n^{-1/2}} M_n(\theta_0) - M_n(\theta) \geq 0\right\} \\
&= \mathbb{P}\left\{\sup_{\|\theta-\theta_0\|\geq\delta n^{-1/2}} M_n^c(\theta_0) - M_n^c(\theta) - [M(\theta) - M(\theta_0)] \geq 0\right\} \\
&\leq \sum_{j=\delta}^{\infty} \mathbb{P}\left\{\sup_{jn^{-1/2}\leq\|\theta-\theta_0\|\leq(j+1)n^{-1/2}} M_n^c(\theta_0) - M_n^c(\theta) \geq \frac{\kappa j^2}{n}\right\} := \sum_{j\geq\delta} P_j
\end{aligned}$$

The last inequality follows as [A3] implies the existence of positive, finite constants c_1, c_2 such that

$$c_1\|\theta-\theta_0\|^2 \leq M(\theta) - M(\theta_0) \leq c_2\|\theta-\theta_0\|^2.$$

We continue by examining the difference $M_n^c(\theta_0) - M_n^c(\theta)$. Observe that

$$\begin{aligned}
M_n^c(\theta_0) - M_n^c(\theta) &= \mu D_{n\theta_0}^2 - \mu(D_{n\theta}^2 - D_\theta^2) \\
&= \mu D_{n\theta_0}^2 - \mu(D_{n\theta} - D_\theta)^2 - 2\mu D_\theta(D_{n\theta} - D_\theta) \\
&= -\mu[(D_{n\theta} - D_\theta) - D_{n\theta_0}]^2 - 2\mu D_{n\theta_0}[(D_{n\theta} - D_\theta) - D_{n\theta_0}] \\
&\quad - 2\mu D_\theta(D_{n\theta} - D_\theta)
\end{aligned}$$

Now [A4] and the continuous mapping theorem yield that uniformly in θ over a vicinity of θ_0,

$$|\mu[\sqrt{n}(D_{n\theta} - D_\theta) - \sqrt{n}D_{n\theta_0}]^2| = \mathcal{o}_P(1)$$

and

$$|2\mu(\sqrt{n}D_{n\theta_0}) \cdot (\sqrt{n}[(D_{n\theta} - D_\theta) - D_{n\theta_0}])| = \mathcal{o}_P(1).$$

Also, it is seen that uniformly in a small neighborhood of θ_0

$$\begin{aligned}
|\mu D_\theta(D_{n\theta} - D_\theta)| &= |\mu D_\theta(D_{n\theta_0} + [(D_{n\theta} - D_\theta) - D_{n\theta_0}])| \\
&\leq |\mu D_\theta[(D_{n\theta} - D_\theta) - D_{n\theta_0}]| + |\mu D_\theta D_{n\theta_0}| \\
&\leq \sqrt{\mu D_\theta^2}\sqrt{\mu[(D_{n\theta} - D_\theta) - D_{n\theta_0}]^2} + |\mu(\theta-\theta_0)'\Delta \cdot D_{n\theta_0}| \\
&\quad + |\mu||\theta-\theta_0||R_\theta D_{n\theta_0}| \\
&= \mathcal{o}(||\theta-\theta_0||n^{-1/2}) + |\mu(\theta-\theta_0)'\Delta D_{n\theta_0}| + \mathcal{o}(\|\theta-\theta_0\|n^{-1/2}),
\end{aligned}$$

where we used that $\mu D_\theta^2 = (\theta-\theta_0)'V(\theta-\theta_0) + \mathcal{o}(||\theta-\theta_0||^2)$. Hence each P_j can be written as

$$P_j = \mathbb{P}\left\{\sup_{jn^{-1/2}\leq||\theta-\theta_0||\leq(j+1)n^{-1/2}} M_n^c(\theta_0) - M_n^c(\theta) \geq \frac{\kappa j^2}{n}\right\}$$

$$= \mathbb{P}\left\{\sup_{jn^{-1/2}\le||\theta-\theta_0||\le(j+1)n^{-1/2}} 2\mu(\theta-\theta_0)'\Delta\cdot D_{n\theta_0}+R_n(\theta)\ge\frac{\kappa j^2}{n}\right\}$$
$$\le \mathbb{P}\left\{\sup_{jn^{-1/2}\le||\theta-\theta_0||\le(j+1)n^{-1/2}} 2\mu(\theta-\theta_0)'\Delta\cdot D_{n\theta_0}\ge\frac{\kappa j^2}{2n}\right\}$$
$$+\mathbb{P}\left\{\sup_{jn^{-1/2}\le||\theta-\theta_0||\le(j+1)n^{-1/2}} 2R_n(\theta)\ge\frac{\kappa j^2}{2n}\right\}$$

The second term on the right can be made arbitrarily small (say less than 2^{-j}) as

$$\sup_{\|\theta-\theta_0\|\le jn^{-1/2}} R_n(\theta)=\mathcal{O}_P(1).$$

The dominating term is the first one, and we obtain via Markov's inequality that

$$\mathbb{P}\left\{\sup_{\|\theta-\theta_0\|\le(j+1)n^{-1/2}}\mu(\theta-\theta_0)'\Delta\cdot D_{n\theta_0}\ge\kappa\frac{j^2}{n}\right\}$$
$$\le\frac{\mathbb{E}\sup_{\|\theta-\theta_0\|\le(j+1)n^{-1/2}}|\mu(\theta-\theta_0)'\Delta\sqrt{n}D_{n,\theta_0}|^M}{(\kappa j^2 n^{-1/2})^M}$$
$$\le\frac{\mathbb{E}\sup_{||\theta-\theta_0||\le(j+1)n^{-1/2}}\left|\left(\mu\{(\theta-\theta_0)'\Delta\}^2\right)^{1/2}\left(\mu\{\sqrt{n}D_{n,\theta_0}\}^2\right)^{1/2}\right|^M}{(\kappa j^2 n^{-1/2})^M}$$
$$\le\frac{\mathbb{E}\sup_{\|\theta-\theta_0\|\le(j+1)n^{-1/2}}\left|||\theta-\theta_0|||\left(\mu||\Delta||^2\right)^{1/2}\left(\mu\{\sqrt{n}D_{n,\theta_0}\}^2\right)^{1/2}\right|^M}{(\kappa j^2 n^{-1/2})^M}$$
$$=\mathcal{O}(j^{-M})\text{ for any }M>1.$$

Consequently, $\sum_{j\ge\delta}P_j\le\kappa\delta^{-M'}$ for all $M'>0$. and

$$\mathbb{E}\|\sqrt{n}(\hat\theta-\theta_0)\|^k=\int_0^\infty x^{k-1}\mathbb{P}\{\sqrt{n}\|\hat\theta-\theta_0\|\ge x\}\,dx$$
$$\le C\int_0^\infty x^{k-1}x^{-M'}\,dx<\infty$$

for $k\ge 1, M'>k-1$. In particular, $\mathbb{E}|\sqrt{n}c'(\hat\theta-\theta_0)|^{2+\delta}<\infty$, implying that $\sqrt{n}c'(\hat\theta-\theta_0)$ is uniformly square integrable. ■

A consistent estimate of the covariance matrix of $\hat\theta$ can be derived from the variance estimators of $c'\hat\theta$ for a finite number of c's.

6 Consistency and Asymptotic Normality of Manski's Estimator

Minimizing a slightly different criterion

$$\widetilde{M}_n(\theta)=\int D^2_{n\theta}(x,g(x,y,\theta))dP_n(x,y),$$

over $\theta \in \Theta$ results in the estimate originally proposed by Manski. There is only a slight difference with the preceeding class of estimators, and we need some slight adaptations of our proofs.

Theorem 11 *Suppose that*

$$\widetilde{M}(\theta) = \int D_\theta^2(x, g(x, y, \theta)) dP(x, y)$$

has a well separated minimum at $\theta = \theta_0$. Furthermore, assume that

$$\sup_{\theta, x, t} |D_{n\theta}(x, t) - D_\theta(x, t)| \xrightarrow{\text{a.s.}} 0.$$

Then $\widetilde{\theta} = \arg\min_{\theta \in \Theta} \widetilde{M}_n(\theta) \to_{a.s.} \theta_0$.

Proof It suffices to show that

$$\sup_{\theta \in \Theta} |\widetilde{M}_n(\theta) - \widetilde{M}(\theta)| \xrightarrow{\text{a.s.}} 0.$$

For this matter, write

$$\begin{aligned}
\widetilde{M}_n(\theta) - \widetilde{M}(\theta) &= \int D_{n\theta}^2(x, g(x, y, \theta)) - D_\theta^2(x, g(x, y, \theta))\, dP_n(x, y) \\
&\quad + \int D_\theta^2(x, g(x, y, \theta))\, d(P_n - P)(x, y) \\
&= \int \widetilde{D}_{n\theta}^2(z) - \widetilde{D}_\theta^2(z)\, dP_n(z) + \int \widetilde{D}_\theta^2(z)\, d(P_n - P)(z).
\end{aligned}$$

The first term on the right is bounded by $\sup |D_{n\theta} - D_\theta|$, where the supremum is taken over $x \in \mathcal{X}, \theta \in \Theta$ and $t \in \mathcal{Y}$. Hence it is of order $\mathcal{o}(1)$ with probability one by assumption.

The second term tends to zero almost surely because

$$\{D_\theta^2(x, \varepsilon) : \theta \in \Theta\}$$

is a P-Glivenko–Cantelli class. The latter is a consequence from the fact that H_θ and G_θ are bounded monotone functions, so that $\mathcal{H} = \{H_\theta : \theta \in \Theta\}$ and $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$ are Glivenko–Cantelli classes. Since any cdf. in $\mathbb{R}^k$ can be written as a limit of $\sum_{i=1}^m \alpha_i I_{[0, t_i]}$, for $t_i \in \mathbb{R}^k$ and $\sum_i |\alpha_i| \leq 1$, Theorem 2.6.9, p. 142 in Van der Vaart and Wellner (1996), and the entropy bound for $\{[0, t] : t \in \mathbb{R}^k\}$ yield that indeed $\mathcal{G}$ and $\mathcal{H}$ are Glivenko–Cantelli classes, whence the composition $\{D_\theta^2 : \theta \in \Theta\}$ is Glivenko–Cantelli as well. ■

Set

$$A = A(\theta, x, t) = \{z = (z_1, z_2) \in \mathcal{Z} : g(z, \theta) \leq t, z_1 \leq x\}.$$

The required uniform convergence follows if $\mathcal{A} = \{A(\theta, x, t)\}$ is a Donsker class, as indexed by θ, x, and t. See Corollary 1 and 2 for applications.

Theorem 12 *Suppose that*

*(*C1*)* θ_0 *lies in the interior of* Θ*;*

*(*C2*)* $\widetilde{\theta} \to_P \theta_0$*,*

*(*C3*) There exists a positive definite matrix* $\widetilde{V}$ *such that* $\widetilde{M}(\theta) = (\theta-\theta_0)^T\widetilde{V}(\theta-\theta_0) + \mathcal{O}(\|\theta-\theta_0\|)$ *for* $\theta \to \theta_0$*,*

*(*C4*) There exists a* $\widetilde{\Delta} \in L_2(P)$ *such that* $\widetilde{D}_\theta(\cdot) = (\theta-\theta_0)^T\widetilde{\Delta}(\cdot) + |\theta-\theta_0|\widetilde{R}(\theta,\cdot)$ *where* $P\widetilde{R}^2(\theta,\cdot) \to 0$ *as* $\theta \to \theta_0$*,*

*(*C5*)* $\sqrt{n}(\widetilde{D}_{n\theta} - \widetilde{D}_\theta)(z) - \sqrt{n}(\widetilde{D}_{n\theta_0} - \widetilde{D}_{\theta_0})(z) = \mathcal{O}_P(1)$ *for all* $\theta \to_P \theta_0$*, uniformly in* $z \in \mathcal{Z}$*,*

*(*C6*)* $\{\widetilde{R}_\theta^2 : \theta \in \Theta\}$ *is a P-Glivenko–Cantelli class.*

Then

$$\sqrt{n}(\widetilde{\theta} - \theta_0) = -2\sqrt{n}\widetilde{V}^{-1}\int \widetilde{\Delta}(\widetilde{D}_{n,\theta_0} - \widetilde{D}_{\theta_0})dP + \mathcal{O}_P(1),$$

and in particular, $\widetilde{\theta}$ *is asymptotically normal.*

Proof Notice that

$$\widetilde{M}_n(\theta) = \int \widetilde{D}_{n\theta}^2\, dP_n = \int \widetilde{D}_{n\theta}^2 dP + \int (\widetilde{D}_{n\theta}^2 - \widetilde{D}_\theta^2)d(P_n - P) + \int \widetilde{D}_\theta^2 d(P_n - P).$$

Recalling the proof of the asymptotic normality of the "μ-estimator" $\hat{\theta}$, we need to show that the difference $\widetilde{M}_n(\theta) - \widetilde{M}_n(\theta_0)$ is basically a quadratic function in $\theta - \theta_0$ plus a negligible remainder term of order $\mathcal{O}_P(n^{-1/2}||\theta-\theta_0|| + ||\theta-\theta_0||^2 + n^{-1})$. The difference is equal to

$$\begin{aligned}\widetilde{M}_n(\theta) - \widetilde{M}_n(\theta_0) \;=\; &\int [\widetilde{D}_{n\theta}^2 - \widetilde{D}_{n\theta_0}^2]dP + \int ([\widetilde{D}_{n\theta}^2 - \widetilde{D}_\theta^2)] - [\widetilde{D}_{n\theta_0}^2 - \widetilde{D}_{\theta_0}^2])d(P_n - P) \\ &+ \int [\widetilde{D}_\theta^2 - \widetilde{D}_{\theta_0}^2]d(P_n - P) = I + II + III.\end{aligned}$$

Since I is of the same form as for the preceeding case, it suffices to show that $II + III = \mathcal{O}_P(n^{-1/2}||\theta-\theta_0|| + ||\theta-\theta_0||^2 + n^{-1})$.

First we consider

$$\begin{aligned}III \;&=\; \int \widetilde{D}_\theta^2 d(P_n - P) \\ &=\; \int [(\theta-\theta_0)^T\widetilde{\Delta} + |\theta-\theta_0|\widetilde{R}_\theta]^2\, d(P_n - P) \\ &=\; \mathcal{O}_P(||\theta-\theta_0||^2)\end{aligned}$$

where we invoke that

$$(\theta-\theta_0)^T \int \widetilde{\Delta}\widetilde{\Delta}^T d(P_n - P)(\theta-\theta_0) = \mathcal{O}_P(\|\theta-\theta_0\|^2)$$

as $\widetilde{\Delta} \in L_2(P)$ and

$$\|\theta - \theta_0\|^2 \int \widetilde{R}_\theta^2 d(P_n - P) = \mathcal{O}_P(\|\theta - \theta_0\|^2),$$

by assumption C6. The second term can dealt with as follows

$$\begin{aligned}
&\int [\widetilde{D}_{n\theta}^2 - \widetilde{D}_\theta^2] d(P_n - P) \\
&= \int [(\widetilde{D}_{n\theta} - \widetilde{D}_\theta)^2 + 2\widetilde{D}_\theta(\widetilde{D}_{n\theta} - \widetilde{D}_\theta)] d(P_n - P) \\
&= \int (\widetilde{D}_{n\theta} - \widetilde{D}_\theta)^2 \, d(P_n - P) + 2(\theta - \theta_0)^T \int \widetilde{\Delta}(\widetilde{D}_{n\theta} - \widetilde{D}_\theta) \, d(P_n - P) \\
&\quad + 2||\theta - \theta_0|| \int \widetilde{R}_\theta(\widetilde{D}_{n\theta} - \widetilde{D}_\theta) d(P_n - P) \\
&= \int [\widetilde{D}_{n\theta_0} - \widetilde{D}_{\theta_0}]^2 \, d(P_n - P) + \mathcal{O}_P(n^{-1/2}||\theta - \theta_0|| + n^{-1}),
\end{aligned}$$

using C4, C5 and C6. ■

References

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

[2] Andrews, D. (1999). Estimation When a Parameter Is on the Boundary. Working paper, Yale University.

[3] Andrews, D. (1999). Estimation When a Parameter Is on the Boundary. *Econometrica*, **67**(6), 1341–1384.

[4] Arcones, M. and Giné, E. (1992). On the Bootstrap of M-Estimators and Other Statistical Functionals. In R. LePage and L. Billard (eds.), *Exploring the Limits of the Bootstrap*, Wiley, pp. 13–48.

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

[6] 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.

[7] Dudley, R. (1999). *Uniform Central Limit Theorems*, Cambridge.

[8] Giné, E. & Zinn, J. (1990). Bootstrapping General Empirical Measures. *Annals of Probability*, **18**, 851–869.

[9] Hildenbrand, W. (1994). *Market Demand: Theory and Empirical Evidence*, Princeton University Press, Princeton.

[10] Manski, C.F. (1983). Closest Empirical Distribution Estimation. *Econometrica*, **51**(2), 305–320.

[11] Mansur, A.H. and Whalley, J. (1984). Numerical Specifications of Applied General Equilibrium Models: Estimation, Calibration and Data. In H.E. Scart and J.B. Shoven (eds.), *Applied General Equilibrium Analysis*, Cambridge University Press, Cambridge.

[12] Mas-Colell, A. (1985). *The Theory of General Economic Equilibrium: A Differential Approach*, Cambridge University Press, Cambridge.

[13] Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimaization Estimators. *Econometrica*, **57**(5), 1027–1057.

[14] Pollard, D. (1984). *Convergence of Empirical Processes.* Springer-Verlag, New York.

[15] Pollard, D. (1985). New Ways to Prove Central Limit Theorems. *Econometric Theory*, **1**, 295–314.

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

[17] Quah, J. (1997). The Monotonicity of Individual and Market Demand. *Nuffield WP 127*, January 1997, revised 1998.

[18] Roehrig, C.S. (1988). Conditions for Identification in Nonparametric and Paramteric models. *Econometrica*, **56**, 433–447.

[19] van der Vaart, A & Wellner, J. (1996). *Weak Convergence and Empirical Processes*, Springer.

[20] Wegkamp, M. (1995). Asymptotic Results for Parameter Estimation in General Empirical Processes. *Tech. Report TW9504*, University of Leiden.

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