Lecture 3
Consistency of Extremum Estimators[1]

This lecture shows how one can obtain consistency of extremum estimators. It also shows how one can find the probability limit of extremum estimators in cases where they are not consistent.

The parameter space of interest is $\Theta \subset R^d$. Extremum estimators (EE) $\{\widehat{\theta}_n : n \geq 1\}$ are defined to be random elements of Θ that approximately minimize a stochastic criterion function $\hat{Q}_n(\theta)$. That is, $\widehat{\theta}_n$ is defined to satisfy

Definition EE: $\widehat{\theta}_n \in \Theta$ *and* $\hat{Q}_n(\widehat{\theta}_n) \leq \inf_{\theta\in\Theta} \hat{Q}_n(\theta) + o_p(1)$.

Here are some examples of extremum estimators:

(1) Maximum Likelihood (ML) Estimator: Suppose the data $\{W_i : i \leq n\}$ are iid with density $f(w, \theta)$ (with respect to some measure μ that does not depend on θ). The likelihood function is $\Pi_{i=1}^n f(W_i, \theta)$. The log-likelihood function is $\sum_{i=1}^n \log f(W_i, \theta)$. The ML estimator $\widehat{\theta}_n$ maximizes the likelihood function or the log-likelihood function over Θ. Equivalently, the ML estimator $\widehat{\theta}_n$ minimizes (at least up to $o_p(1)$)

$$\hat{Q}_n(\theta) = -\tfrac{1}{n}\sum_{i=1}^n \log f(W_i, \theta) \tag{1}$$

over $\theta \in \Theta$.

In most cases in econometrics, the vector W_i can be decomposed into two subvectors: $W_i' = (Y_i', X_i')$, where Y_i is a vector of outcome (or response) variables and X_i is a vector of explanatory variables. Suppose the conditional density of Y_i given $X_i = x$ is $f(y|x, \theta)$ and the marginal density of X_i is some density $g(x)$ that does not depend on θ. Then, $f(w, \theta) = f(y|x, \theta)g(x)$ and

$$\begin{aligned} \hat{Q}_n(\theta) &= -\tfrac{1}{n}\sum_{i=1}^n \log f(W_i, \theta) \\ &= -\tfrac{1}{n}\sum_{i=1}^n \log f(Y_i|X_i, \theta) - \tfrac{1}{n}\sum_{i=1}^n \log g(X_i). \end{aligned} \tag{2}$$

Since the second term on the right-hand side does not depend on θ, it can be deleted without affecting the definition of the ML estimator. That is, one can define the ML estimator to minimize minus the conditional log-likelihood

$$\hat{Q}_n(\theta) = -\tfrac{1}{n}\sum_{i=1}^n \log f(Y_i|X_i, \theta) \tag{3}$$

[1]The note for this lecture is largely adapted from the note of Donald Andrews on the same topic. I am grateful for Professor Andrews' generosity and elegant exposition. All errors are mine.

which does not require that one specifies the marginal distribution of the explanatory variables X_i.

In the case of dependent data, we write the log-likelihood function as the product of conditional distributions. Suppose $W_i' = (Y_i', X_i')$, where Y_i and X_i are as above. Let $f(y_i|x_i, x_{i-1}, ..., x_1, y_{i-1}, y_{i-2}, ..., y_1, \theta)$ denote the conditional density of Y_i given $X_i = x_i$, $X_{i-1} = x_{i-1}, ..., X_1 = x_1$ and $Y_{i-1} = y_{i-1}$, $Y_{i-2} = y_{i-2}, ..., Y_1 = y_1$. Let $g(x_i|x_{i-1}, ..., x_1, y_{i-1}, ..., y_1)$ denote the conditional density of X_i given $X_{i-1} = x_{i-1}, ..., X_1 = x_1$ and $Y_{i-1} = y_{i-1}, ..., Y_1 = y_1$. If $g(\cdot)$ does not depend on θ, then the explanatory variables $\{X_1, ..., X_n\}$ are said to be *weakly exogenous.* If $g(\cdot)$ does not depend on θ or $y_{i-1}, ..., y_1$, then the explanatory variables are said to be *strongly exogenous.* Suppose the explanatory variables are *weakly or strongly exogenous.* Then, the likelihood function can be written as

$$\begin{aligned} & g(X_1)f(Y_1|X_1, \theta) \times g(X_2|X_1, Y_1)f(Y_2|X_2, X_1, Y_1, \theta) \times ... \times \\ & g(X_n|X_{n-1}, ..., X_1, Y_{n-1}, ..., Y_1)f(Y_n|X_n, ..., X_1, Y_{n-1}, ..., Y_1, \theta) \\ = & \prod_{i=1}^{n} f(Y_i|X_i, ..., X_1, Y_{i-1}, ..., Y_1, \theta)g(X_i|X_{i-1}, ..., X_1, Y_{i-1}, ..., Y_1). \end{aligned} \tag{4}$$

The log-likelihood function is

$$\sum_{i=1}^{n} \log f(Y_i|X_i, ..., X_1, Y_{i-1}, ..., Y_1, \theta) + \sum_{i=1}^{n} \log g(X_i|X_i, ..., X_1, Y_{i-1}, ..., Y_1). \tag{5}$$

In this case, the ML estimator minimizes

$$\hat{Q}_n(\theta) = -\tfrac{1}{n} \sum_{i=1}^{n} \log f(Y_i|X_i, ..., X_1, Y_{i-1}, ..., Y_1, \theta). \tag{6}$$

If $\{Y_i : i \geq 1\}$ is Markov of order κ, then $f(y_i|x_i, ..., x_1, y_{i-1}, ..., y_1, \theta) = f(y_i|x_i, ..., x_{i-\kappa}, y_{i-1}, ..., y_{i-\kappa}, \theta)$ and we take

$$\hat{Q}_n(\theta) = -\tfrac{1}{n} \sum_{i=1}^{n} \log f(Y_i|X_i, ..., X_{i-\kappa}, Y_{i-1}, ..., Y_{i-\kappa}, \theta). \tag{7}$$

(2) Least Squares (LS) Estimator for Nonlinear Regression: Suppose the data $W_i = (Y_i, X_i')'$ for $i \leq n$ are iid and satisfy the nonlinear regression model

$$Y_i = g(X_i, \theta_0) + \varepsilon_i, \tag{8}$$

where $E(\varepsilon_i|X_i) = 0$ a.s. The LS estimator $\widehat{\theta}_n$ minimizes

$$\hat{Q}_n(\theta) = \tfrac{1}{n}\sum\nolimits_{i=1}^{n}(Y_i - g(X_i,\theta))^2/2 \tag{9}$$

over $\theta \in \Theta$. (The scale factor $1/2$ is used because it is convenient for the asymptotic normality results given below. It has no effect on the consistency results.)

(3) Generalized Method of Moments (GMM) Estimator: Let θ_0 denote the true value. Suppose the data $\{W_i : i \leq n\}$ are iid and the following moment conditions hold

$$Eg(W_i,\theta) \begin{cases} = 0 & \text{if } \theta = \theta_0 \\ \neq 0 & \text{if } \theta \neq \theta_0 \end{cases}, \tag{10}$$

where $g(w,\theta) \in R^k$ for some $k \geq d$. Let A_n be a $k \times k$ random (weight) matrix. Then, the GMM estimator $\widehat{\theta}_n$ minimizes

$$\hat{Q}_n(\theta) = \left|\left|A_n \tfrac{1}{n}\sum\nolimits_{i=1}^{n} g(W_i,\theta)\right|\right|^2/2 \tag{11}$$

over $\theta \in \Theta$, where $||\cdot||$ is the Euclidean norm.

(4) Minimum Distance (MD) Estimator: Let $\widehat{\pi}_n$ be a consistent unrestricted estimator of a k-vector parameter π_0. Suppose π_0 is known to be a function of a d-vector parameter θ_0, where $d \leq k$:

$$\pi_0 = g(\theta_0). \tag{12}$$

Let A_n be a $k \times k$ random weight matrix. Then, the MD estimator $\widehat{\theta}_n$ minimizes

$$\hat{Q}_n(\theta) = ||A_n(\widehat{\pi}_n - g(\theta))||^2/2 \tag{13}$$

over $\theta \in \Theta$. For example, in the correlated random effects panel data model, $\widehat{\pi}_n$ is the OLS (or FGLS) estimator of π_0.[2]

(5) Two-step (TS) Estimator: Suppose the data $\{W_i : i \leq n\}$ are iid, $\widehat{\tau}_n$ is a preliminary consistent estimator of a parameter τ_0, $G_n(\theta,\tau)$ is a random k-vector that should be close to 0 when

[2](Wooldridge (2002), p. 445) A panel data correlated random effect model is of the form:

$$y_{it} = \psi + x_{it}\beta + c_i + v_{it}, \tag{14}$$

where x_{it} is a k-dimensional row vector, $x_i = (x_{i1},...,x_{iT})'$, $c_i = \sum_{t=1}^{T} x_{it}\lambda_t$, $E(v_{it}) = 0$ and $E(x_i'v_{it}) = 0$, $t = 1,...,T$. The structual parameter of the model is $\theta = (\psi, \lambda_1',...,\lambda_T', \beta')'$. To apply the MD approach, one first estimates (by OLS)

$$y_{it} = \pi_{t0} + x_i\pi_t + v_{it},\ t = 1,...,T. \tag{15}$$

Let $\pi = (\pi_{10}, \pi_1',...,\pi_{T0}, \pi_T')'$. Then one forms the MD problem using the restriction

$$\pi = H\theta, \tag{16}$$

$\theta = \theta_0$, $\tau = \tau_0$, and n is large (e.g., in the GMM case, $G_n(\theta,\tau) = \frac{1}{n}\sum_{i=1}^n g(W_i,\theta,\tau)$ and in the MD case $G_n(\theta,\tau) = \widehat{\pi}(\tau) - g(\theta,\tau)$), and A_n is a $k \times k$ random weight matrix. Then, the TS estimator $\widehat{\theta}_n$ minimizes

$$\hat{Q}_n(\theta) = ||A_n G_n(\theta, \widehat{\tau}_n)||^2/2 \tag{17}$$

over $\theta \in \Theta$.

The criterion function $\hat{Q}_n(\theta)$ is assumed to satisfy:

Assumption U–WCON: $\sup_{\theta\in\Theta} |\hat{Q}_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ *for some nonstochastic function* $Q(\theta)$ *on* Θ.

For a given estimator of interest, one has to verify this condition.
What is the function $Q(\theta)$ in our examples? It is given by

(1) ML Estimator: $Q(\theta) = -E\log f(W_i,\theta)$, where E denotes expectation under the true distribution generating the data.

(2) LS Estimator: $Q(\theta) = E(Y_i - g(X_i,\theta))^2/2$.

(3) GMM Estimator: $Q(\theta) = ||AEg(W_i,\theta)||^2/2$, where $A_n \xrightarrow{p} A$.

(4) MD Estimator: $Q(\theta) = ||A(\pi_0 - g(\theta))||^2/2$, where $A_n \xrightarrow{p} A$ and $\widehat{\pi}_n \to_p \pi_0$.

(5) TS Estimator: $Q(\theta) = ||AG(\theta,\tau_0)||^2/2$, where $A_n \xrightarrow{p} A$, $\widehat{\tau}_n \xrightarrow{p} \tau_0$, and $G_n(\theta,\tau) \xrightarrow{p} G(\theta,\tau)$.

The next assumption is that of *identifiable uniqueness* of some non-stochastic element θ_0 of Θ. It is to this parameter value that the estimators $\{\widehat{\theta}_n : n \geq 1\}$ converge as $n \to \infty$. Let $B(\theta,\varepsilon)$ denote an open ball in Θ of radius ε centered at θ.

Assumption ID: There exists $\theta_0 \in \Theta$ such that $\forall \varepsilon > 0$, $\inf_{\theta\notin B(\theta_0,\varepsilon)} Q(\theta) > Q(\theta_0)$.

A necessary condition for Assumption ID is that θ_0 uniquely minimizes $Q(\theta)$ over Θ. What values uniquely minimize $Q(\theta)$ in the examples? The answer is as follows:

---

where $H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & I_{KT} & E_1 \\ 1 & 0 & 0 \\ 0 & I_{KT} & E_2 \\ \cdots & \cdots & \cdots \\ 1 & 0 & 0 \\ 0 & I_{KT} & E_T \end{pmatrix}$, $E_t = e_t \otimes I_K$, and e_t is a T-vector whose tth element is 1 and other elements are 0. The MD objective function is as defined in (13) with $g(\theta) = H\theta$ and some user-chosen weight matrix A_n.

(1) ML Estimator (with iid observations): If the true density $f(w, \theta_0)$ of the data is in the parametric family $\{f(w, \theta) : \theta \in \Theta\}$, then θ_0 minimizes $Q(\theta)$:

$$\begin{aligned} Q(\theta_0) - Q(\theta) &= E_{\theta_0} \log(f(W_i, \theta)/f(W_i, \theta_0)) \\ &\leq \log E_{\theta_0} f(W_i, \theta)/f(W_i, \theta_0) \\ &= \log \int \frac{f(w, \theta)}{f(w, \theta_0)} f(w, \theta_0) d\mu(w) \\ &= 0, \end{aligned}$$

where the inequality holds by Jensen's inequality because $\log(\cdot)$ is a concave function. The inequality is strict, and so, θ_0 uniquely minimizes $Q(\theta)$ over Θ if and only if $P(f(W_i, \theta) \neq f(W_i, \theta_0)) > 0 \ \forall \theta \in \Theta$ with $\theta \neq \theta_0$.

Suppose the true density $f(\cdot)$ of W_i is not in the parametric family $\{f(\cdot, \theta) : \theta \in \Theta\}$. In this case, we proceed as follows. The Kullback–Liebler Information Criterion (KLIC) between $f(\cdot)$ and $f(\cdot, \theta)$ is defined by

$$\text{KLIC}(f, f(\cdot, \theta)) = E_f \log f(W_i) - E_f \log f(W_i, \theta),$$

where E_f denotes expectation when W_i has density f. Note that

$$\text{KLIC}(f, f(\cdot, \theta)) = E_f \log f(W_i) + Q(\theta).$$

In consequence, the ML estimator under misspecification (often called the quasi-ML estimator) converges in probability to the parameter value θ_0 that uniquely minimizes the KLIC between the true density f and the densities in the parametric family $\{f(\cdot, \theta) : \theta \in \Theta\}$ (provided such a unique value exists).

(2) LS Estimator: The true value θ_0 minimizes $Q(\theta)$ if the regression model is correctly specified (i.e., if there is a $\theta_0 \in \Theta$ such that $E(Y_i|X_i) = g(X_i, \theta_0)$ a.s.):

$$\begin{aligned} & Q(\theta) - Q(\theta_0) \\ = \ & E(U_i + g(X_i, \theta_0) - g(X_i, \theta))^2/2 - EU_i^2/2 \\ = \ & E(g(X_i, \theta_0) - g(X_i, \theta))^2/2 + EU_i(g(X_i, \theta_0) - g(X_i, \theta)) \\ = \ & E(g(X_i, \theta_0) - g(X_i, \theta))^2/2 \\ \geq \ & 0. \end{aligned}$$

The inequality is strict, and so, θ_0 uniquely minimizes $Q(\theta)$ over Θ if and only if $P(g(X_i, \theta) \neq g(X_i, \theta_0)) > 0 \ \forall \theta \in \Theta$ with $\theta \neq \theta_0$.

Suppose the nonlinear regression model is not correctly specified. Let $g(x) = E(Y_i|X_i = x)$. Then,

$$\begin{aligned} Q(\theta) &= E(Y_i - g(X_i) + g(X_i) - g(X_i, \theta))^2/2 \\ &= E(Y_i - g(X_i))^2/2 + E(g(X_i) - g(X_i, \theta))^2/2 \end{aligned}$$

and a value θ_0 uniquely minimizes $Q(\theta)$ over Θ if it uniquely minimizes

$$E(g(X_i) - g(X_i, \theta))^2$$

over $\theta \in \Theta$. That is, the LS estimator converges to the point θ_0 that gives the best mean squared approximation in the family $\{g(\cdot, \theta) : \theta \in \Theta\}$ to the conditional mean of Y_i given X_i (provided the best approximation is unique).

(3) GMM Estimator: If A is nonsingular and there exists a unique value $\theta_0 \in \Theta$ such that $Eg(W_i, \theta_0) = 0$, then θ_0 uniquely minimizes $Q(\theta)$ over Θ. Suppose the moment conditions are misspecified and no value $\theta \in \Theta$ is such that $Eg(W_i, \theta) = 0$. Then, θ_0 is the value that uniquely minimizes $Q(\theta) = Eg(W_i, \theta)'A'AEg(W_i, \theta)/2$, if such a value exists.

(4) MD Estimator: If A is nonsingular and there exists a unique value $\theta_0 \in \Theta$ such that $\pi_0 = g(\theta_0)$, then θ_0 uniquely minimizes $Q(\theta)$ over Θ. Suppose the restrictions are misspecified and no value $\theta \in \Theta$ is such that $\pi_0 = g(\theta)$. Then, θ_0 is the value that uniquely minimizes $Q(\theta) = (\pi_0 - g(\theta))'A'A(\pi_0 - g(\theta))/2$, if such a value exists.

(5) TS Estimator: If A is nonsingular and there exists a unique value $\theta_0 \in \Theta$ such that $G(\theta_0, \tau_0) = 0$, then θ_0 uniquely minimizes $Q(\theta)$ over Θ.

The following theorem gives sufficient conditions for $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

Theorem 3.1 (Consistency): EE, ID, & U–WCON $\Rightarrow \widehat{\theta}_n \xrightarrow{p} \theta_0$.

Proof of Theorem 31: Given $\varepsilon > 0$, $\exists \delta > 0$ such that $\theta \notin B(\theta_0, \varepsilon) \Rightarrow Q(\theta) - Q(\theta_0) \geq \delta > 0$. Thus,

$$\begin{aligned} P\left(\widehat{\theta}_n \notin B(\theta_0, \varepsilon)\right) &\leq P(Q(\widehat{\theta}_n) - Q(\theta_0) \geq \delta) \\ &= P(Q(\widehat{\theta}_n) - \hat{Q}_n(\widehat{\theta}_n) + \hat{Q}_n(\widehat{\theta}_n) - Q(\theta_0) \geq \delta) \\ &\leq P(Q(\widehat{\theta}_n) - \hat{Q}_n(\widehat{\theta}_n) + \hat{Q}_n(\theta_0) + o_p(1) - Q(\theta_0) \geq \delta) \\ &\leq P(2 \sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q(\theta)| + o_p(1) \geq \delta) \xrightarrow{p} 0. \end{aligned} \tag{18}$$

Primitive Sufficient Conditions for the Identification Condition

The following assumption is sufficient for Assumption ID.

Assumption ID1: (i) Θ *is compact.*
(ii) $Q(\theta)$ *is continuous on* Θ.
(iii) θ_0 *uniquely minimizes* $Q(\theta)$ *over* $\theta \in \Theta$.

Lemma 3.1: ID1 $\Rightarrow$ ID.

Proof of Lemma 3.1: (Problem Set Question.) Hint: Use a proof by contradiction.

Note that none of the three conditions of Assumption ID1 is redundant. If Θ is non-compact, $Q(\theta)$ is not continuous, or $Q(\theta)$ is not uniquely minimized at θ_0, then Assumption ID fails.
On the other hand, only Assumption ID1(iii) is a necessary condition for Assumption ID.

When does Assumption ID1 hold in the examples? Obviously ID1(i) holds in each example if Θ is compact. We have already given conditions under which ID1(iii) holds in each example. To verify ID1(ii) in the examples, the following lemma is useful.

Lemma 3.2: *Suppose* $m(W_i, \theta)$ *is continuous in* θ *at each* $\theta \in \Theta$ *with probability one and* $E \sup_{\theta \in \Theta} ||m(W_i, \theta)|| < \infty$. *Then,* $Em(W_i, \theta)$ *is continuous in* θ *on* Θ.

For example, when $\hat{Q}_n(\theta)$ is of the form $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^n m(W_i, \theta)$, then $Q(\theta) = Em(W_i, \theta)$. This occurs in the ML and nonlinear regression examples. In the GMM example, $\hat{Q}_n(\theta)$ is of the form $\hat{Q}_n(\theta) = ||A_n n^{-1} \sum_{i=1}^n m(W_i, \theta)||^2/2$ and $Q(\theta) = ||AEm(W_i, \theta)||^2/2$. In this case too, continuity of $Em(W_i, \theta)$ yields continuity of $Q(\theta)$.
Proof of Lemma 3.2: The result holds by the dominated convergence theorem with the dominating function given by $\sup_{\theta \in \Theta} ||m(W_i, \theta)||$, since

$$\lim_{\theta' \to \theta} m(W_i, \theta') = m(W_i, \theta) \text{ a.s. and} \tag{19}$$

$$E \sup_{\theta \in \Theta} ||m(W_i, \theta)|| < \infty. \tag{20}$$

Using Lemma 3.2, we see that ID1(ii) holds for the (1) ML estimator if $f(w, \theta)$ is continuous in θ at each $\theta \in \Theta$ with probability one and $E \sup_{\theta \in \Theta} || \log f(W_i, \theta)|| < \infty$, (2) LS estimator if $g(x, \theta)$ is continuous in θ at each $\theta \in \Theta$ with probability one and $E \sup_{\theta \in \Theta} (Y_i - g(X_i, \theta))^2 < \infty$ (or, equivalently, $E\varepsilon_i^2 < \infty$ and $E \sup_{\theta \in \Theta} (g(X_i, \theta_0) - g(X_i, \theta))^2 < \infty$), and (3) GMM estimator if $g(w, \theta)$ is continuous in θ at each $\theta \in \Theta$ with probability one and $E \sup_{\theta \in \Theta} ||g(W_i, \theta)|| < \infty$.

Assumption ID1(ii) holds for the MD and TS estimators if $g(\theta)$ and $G(\theta, \tau_0)$ are continuous in θ on Θ respectively.

Consistency without Point Identification In recent years, there have been considerable interest in partially identified models, in which the parameters cannot be pinned down to a unique point even if one had infinite amount of data. In the context of extreme estimation, partial identification amounts to the failure of Assumption ID. Typically, Assumption ID is replaced by a set identification condition:

Assumption SetID: There exists $\Theta_0 \subset \Theta$ such that $\sup_{\theta \in \Theta_0} Q(\theta) = \inf_{\theta \in \Theta} Q(\theta)$ and $\forall \varepsilon > 0$, $\inf_{\theta \notin B(\Theta_0, \varepsilon)} Q(\theta) > Q(\theta_0)$

The set Θ_0 is called the identified set.

Assumption EE is replaced with a set estimator $\widehat{\Theta}_n$.

Consistency of the set estimator typically means consistency in the Hausdorff distance, where the Hausdorff distance is a distance measure between two sets:

$$d_H(A, B) = \sup_{a \in A} \inf_{b \in B} \|a - b\| + \sup_{b \in B} \inf_{a \in A} \|a - b\|. \tag{21}$$

A natural set estimator is the argmin set of $\hat{Q}_n(\theta)$. However, under Assumptions U-WCON and SetID, this set estimator is not necessarily consistent in Hausdorff distance. It is not difficult to show (problem set question) that it is half-Hausdorff consistent under these two conditions, i.e.,

$$\sup_{\theta \in \widehat{\Theta}_n} \inf_{\theta' \in \Theta_0} \|\theta - \theta'\| \to_p 0. \tag{22}$$

That is: the argmin set approaches the identified set, but it might not include enough points to approach every point in the identified set. There have been different proposals for solving this problem:

1. Hausdorff consistency of the argmin set can be proved in the special case where $\hat{Q}_n(\theta)$ degenerate in the interior of Θ_0 and the closure of the interior equals Θ_0, i.e.,$\sup_{\theta \in \Theta_0^{-\varepsilon}} \hat{Q}_n(\theta) = \inf_{\theta \in \Theta} \hat{Q}_n(\theta)$ with probability approaching 1 for any $\varepsilon > 0$ and $\Theta_0^{-\varepsilon}$ being the set of points that are at least ε away from the complement of Θ_0, and $\lim_{\varepsilon \downarrow 0} d_H(\Theta_0, \Theta_0^{-\varepsilon}) = 0$. This is a special case of Chernozhukov, Hong and Tamer (2007, Condition C.3)

2. Use a bigger set: let $\widehat{\Theta}_n$ be all points such that $\hat{Q}_n(\theta) \leq \inf_{\theta \in \Theta} \hat{Q}_n(\theta) + \tau_n$ where τ_n is a positive sequence that converges to zero at an appropriate rate. A stronger (convergence rate) assumption is needed to replace U-WCON for such estimators to be consistent. (Chernozhukov, Hong and Tamer (2007, Econometrica),Theorem 3.1)

3. Abandon consistent estimation all together and focus on confidence set. The argument for this proposal is that the identified set is not the true parameter and thus is not the parameter of

interest anyway, while a confidence set for the true parameter can be constructed and interpreted in the conventional sense. (Andrews and Soares (2010, Econometrica), Andrews and Guggenberger (2010, Econometrica), etc.)

4. Make assumptions so that certain features of the identified set (like the upper left corner) is point identified and focus on such features. (Pakes, Porter, Ho and Ishii (2012, Econometrica))

5. Bayesian analysis, where there is no notion of "partial identification". (Liao and Jiang (2010, Annals of Statistics))

Below I describe Chernozhukov, Hong and Tamer (2007)'s results regarding solutions 1 and 2. But before doing so, here are a couple of motivating examples of partially identified models.

Example. Interval Outcome in Regression Model. Suppose that one has a regression model:

$$Y = X^{'}\beta_0 + \varepsilon,\ E(\varepsilon X) = 0, \tag{23}$$

where Y is the outcome variable and X is the regressors. Assume that variables in X are all nonnegative. Suppose that Y is unobserved. Only an interval that includes Y is observed: $[Y_\ell, Y_u]$. No addition information on how $[Y_\ell, Y_u]$ is related to Y is available. Then the model implies (and is implied by) the following moment inequalities:

$$\begin{aligned} &E[XY_u - XX^{'}\beta_0] \geq 0,\ and \\ &E[XX^{'}\beta_0 - XY_\ell] \geq 0\ . \end{aligned} \tag{24}$$

One criterion function used for estimating the identified set of β_0 is:

$$\hat{Q}_n(\beta) = \|A_n[\bar{m}_n(\beta)]_-^{'}\|^2, \tag{25}$$

where A_n is a weighting matrix,$[x]_- = |minx, 0|$, and

$$\bar{m}_n(\beta) = \left(n^{-1}\sum_{i=1}^{n}(X_iY_{u,i} - X_iX_i^{'}\beta)^{'},\ n^{-1}\sum_{i=1}^{n}(X_iX_i^{'}\beta - X_iY_{\ell,i})^{'} \right)^{'}, \tag{26}$$

The function $\hat{Q}_n(\beta)$ converges (under appropriate conditions) uniformly to the function $Q(\theta) = \|A \cdot m(\theta)\|^2$, where

$$m(\theta) = \left(E[XY_u - XX^{'}\beta]^{'},\ E[XX^{'}\beta - XY_\ell]^{'} \right)^{'}. \tag{27}$$

The function $Q(\theta)$ does not have unique minimum.

Proposition 3.1. Under P-WCON, IDSet and the degeneracy condition described in solution 1 above, the argmin set $\widehat{\Theta}_n = \{\theta : \hat{Q}_n(\theta) = \inf_{\theta\in\Theta}\hat{Q}_n(\theta)\}$ is consistent in Hausdorff distance.

Proof. The Hausdorff consistency is implied by Equation (22) and

$$\sup_{\theta\in\Theta_0}\inf_{\theta'\in\widehat{\Theta}_n}\|\theta-\theta'\|\to_p 0. \tag{28}$$

The proof of Equation (22) is left as a problem set question. Here we prove (28). Consider the following derivation: for any $\varepsilon>0$,

$$\begin{aligned}&\Pr(\sup_{\theta\in\Theta_0}\inf_{\theta'\in\widehat{\Theta}_n}\|\theta-\theta'\|>\varepsilon)\\&\leq \Pr\left(\sup_{\theta\in\Theta_0}\inf_{\theta'\in\Theta_0^{-\delta}}\|\theta-\theta'\|>\varepsilon\right)+\Pr(\Theta_0^{-\delta}\not\subseteq\widehat{\Theta}_n).\end{aligned} \tag{29}$$

For any fixed $\delta>0$, the second term of the RHS of the above display converges to zero as $n\to\infty$. The first term can be made arbitrarily small by choosing δ small enough. Therefore the limit of the LHS is zero as $n\to\infty$. ■

The problem with solution 1 is that the degeneracy condition often is not satisfied and certainly is tricky to verify.

The second solution needs more assumptions on the convergence of the sample criterion functions. The additional assumption is given below as Assumption RT, where "RT" stands for "rate (of convergence)". Part (a) of this assumption is without loss of generality because one can always recenter the criterion functions by their infimum and make this assumption hold. In part (b), the rate "n" can be replaced with any other rate, as long as the τ_n choice in the proposition below adjusts with this rate accordingly.

Assumption RT (a)$\inf_{\theta\in\Theta}\hat{Q}_n(\theta)\geq\inf_{\theta\in\Theta}Q(\theta)$ and (b) $n\sup_{\theta\in\Theta_0}|\hat{Q}_n(\theta)-Q(\theta)|=O_p(1)$.

Proposition 3.2. Under Assumptions U-WCON, RT and IDset, if $n\tau_n\to\infty$, then the augmented argmin set defined in solution 2 above is consistent in Hausdorff distance.

Proof. As in the previous propostion, we only prove (28). It suffices to show that $\Theta_0\subseteq\widehat{\Theta}_n$ with probability approaching 1. Consider the following derivation:

$$\begin{aligned}\Pr(\Theta_0\not\subseteq\widehat{\Theta}_n)&\leq\Pr\left(\sup_{\theta\in\Theta_0}\hat{Q}_n(\theta)>\inf_{\theta\in\Theta}\hat{Q}_n(\theta)+\tau_n\right)\\&\leq\Pr\left(\sup_{\theta\in\Theta_0}\hat{Q}_n(\theta)-\inf_{\theta\in\Theta}Q(\theta)>\tau_n\right)\\&\leq\Pr\left(\sup_{\theta\in\Theta_0}|\hat{Q}_n(\theta)-Q(\theta)|>\tau_n\right)\\&\leq\Pr\left(n\sup_{\theta\in\Theta_0}|\hat{Q}_n(\theta)-Q(\theta)|>n\tau_n\right).\end{aligned} \tag{30}$$

The last line of the above display goes to zero as $n \to \infty$ due to Assumption RT(b). ■

The assumptions required for solution 2 are innocuous enough, but it leaves the choice of τ_n open. Other than a loose rate requirement, there is no practical way to choose τ_n. As a result, one man's set estimator can be 10 times bigger than another man's simply due to different choices of τ_n.