Recent Developments in Semi and Nonparametric Econometrics

Part 3: partial identification

Xavier D'Haultfœuille
CREST

Outline

Outline

Introduction

- Sample data are not sufficient to draw conclusions on the whole population. This issue is similar to the general problem of induction in science.
- Statistical theory has defined assumptions which allows to draw such conclusions. But it has mainly focused on the issue of the accuracy of point estimators. Point identification is usually assumed from the beginning.
- Before recently, identification was indeed considered as a binary event, the parameter of interest being either point identified or not identified at all.
- But, sometimes, overly strong assumptions must be imposed to point identify parameters of interest. Instead, we may get *partial identification* under more credible conditions.

Introduction

- We first present important examples where partial identification arises:
  - missing data problems: nonresponse, interval data, treatment effects...
  - incomplete models: models which do not predict a unique outcome for given "inputs".
- We then consider some issues in inference on partially identified parameters/models.

A formal definition

- Let us consider a statistical model $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, M a set of probability measures such that $P_\theta \in M$ and define the function

$$\begin{aligned} \varphi : \Theta &\rightarrow M \\ \theta &\mapsto P_\theta \end{aligned}$$

- The minimal *identification region* of θ_0 is $R(\theta_0) = \varphi^{-1}(P_{\theta_0})$. It represents the set of all parameters θ that can be rationalized by the data. An identification region is any set R such that $R(\theta_0) \subset R$.
- If $R(\theta) = \{\theta\}$ for all θ, the model is said to be identifiable.
- The minimal identification region of $g(\theta_0)$ is $R(g(\theta_0)) = \{g(\theta), \theta \in R(\theta_0)\}$. $g(\theta_0)$ is point identified if $R(g(\theta_0)) = \{g(\theta_0)\}$.
- Suppose that $g(\theta_0) \in \mathbb{R}$ and we have an identification interval $R = [\underline{g}, \overline{g}]$. The upper bound $\overline{g}$ is *sharp* if $\overline{g} = \sup R(g(\theta_0))$.

A formal definition

- To establish point identification in a model where $X \sim P_\theta$, we often prove that there exists q such that $\theta = q(P^X)$ (constructive proof).
- Consider for instance the model $(P_\theta)_{\theta\in\Theta} = (\mathcal{N}(\mu,\sigma^2))_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+}$. Then μ and σ can be identified by $\mu = E(X)$ and $\sigma^2 = V(X)$.
- Now consider the model $(\mathcal{N}(\mu_1+\mu_2,1))_{(\mu_1,\mu_2)\in\mathbb{R}^2}$. This model is not point identified. The identification region for (μ_1,μ_2) is

$$R(\mu_1,\mu_2) = \{(\mu_1',\mu_2') \in \mathbb{R}^2 / \mu_1' + \mu_2' = \mu_1 + \mu_2\}.$$

- On the other hand, $g(\mu_1,\mu_2) = \mu_1 + \mu_2$ is identified since for all $(\mu_1',\mu_2') \in R(\mu_1,\mu_2)$, we have $\mu_1' + \mu_2' = \mu_1 + \mu_2$.

Outline

Missing data: identification with the data alone

- We are interested in the distribution of a variable Y, but we face a missing data problem. Letting D be the binary indicator of response, we only observe D and Y when $D = 1$.
- A standard assumption is to ignore nonresponse by supposing that $Y \perp\!\!\!\perp D$. In this case the distribution of Y, P^Y, is identified since $P^Y = P^{Y|D=1}$.
- However, this assumption is restrictive. In a labor force survey, for instance, it is likely that unemployment status is related to nonresponse.
- Suppose that $Y \in \mathcal{Y}$ and let M be the set of probability measures on $\mathcal{Y}$. Without any assumption, we can still infer that P^Y belongs to

$$R(P^Y) = \left\{ P^{Y|D=1} p + \mu(1-p), \mu \in M \right\},$$

with $p = P(D = 1)$.

Missing data: identification with the data alone

- If we focus on a parameter $g(P^Y)$, the corresponding identification region is

$$R(g(P^Y)) = \{g(P),\ P \in R(P^Y)\}.$$

- This set is abstract and can be rather difficult to characterize. Consider for instance $g(P^Y) = V(Y)$, with $Y \in [0,1]$ and let $\mathcal{D}$ denote the set of cdf on $[0,1]$. Then the upper bound of $g(P^Y)$ solves the functional problem:

$$\max_{F \in \mathcal{D}} \int \left[y - \int y dF(y) \right]^2 dF(y) \quad \text{s.t. } (F - pF_{Y|D=1})/(1-p) \in \mathcal{D}.$$

Missing data: identification with the data alone

- However, solutions are simple when $Y \in \mathbb{R}$ and for $D-$ parameters, i.e. parameters that respect (first order) stochastic dominance:

$$F_1(y) \leq F_2(y)\ \forall y \Rightarrow g(F_1) \geq g(F_2).$$

- This includes $E(h(Y))$ for all h increasing, quantiles...
- For these parameters, we have, letting $\underline{y} = \inf \mathcal{Y}$ and $\overline{y} = \sup \mathcal{Y}$,

$$R(g(P^Y)) = \left[g\left(pP^{Y|D=1} + (1-p)\delta_{\underline{y}}\right), g\left(pP^{Y|D=1} + (1-p)\delta_{\overline{y}}\right)\right].$$

Missing data: identification with the data alone

- A simple example: suppose that we observe 10% of unemployed people in a labor force survey with 15% of nonresponse. Then without any assumption, the identification region on unemployment rate is between $8,5\%$ (all nonrespondents work) and $23,5\%$ (all nonrespondents are unemployed).
- Another example: we observe the following distribution of wages on a survey with 20% of nonresponse:

Mean	D1	D2	Q1	D3	D4	D5
1,800	800	1,000	1,200	1,300	1,400	1,600

- What is the minimal identification region of the fourth decile of the wages? Of the mean?

Incomplete data: identification with the data alone

- In some cases, rather than observing completely Y or not at all, we have partial observation of it. A usual example is *interval data*, for example on wages: we only observe $Y \in [\underline{Y}, \overline{Y}]$.
- Missing data is a particular case, with $D = 1$ when $\underline{Y} = \overline{Y}$, and $D = 0$ when $\underline{Y} = \underline{y}$ and $\overline{Y} = \overline{y}$.
- Inference on $D-$parameters is easy in this case, since we have

$$F_{\overline{Y}} \leq F_Y \leq F_{\underline{Y}}.$$

In this case,

$$R(g(P^Y)) = \left[g\left(P^{\underline{Y}}\right), g\left(P^{\overline{Y}}\right) \right].$$

Missing data: identification with IV

- Usually, using the data alone does not provide much information on Y. We can shrink bounds on parameters of interest using various restrictions. Let Z be a variable which is always observed. Then suppose

$$Y \perp\!\!\!\perp Z.$$

- Such an assumption is similar to that used in sample selection models, where we use instruments which affect D but not Y.
- Under this assumption, we have, for all z,

$$\begin{aligned} P^{Y}(A) &= P^{Y|Z=z}(A) \\ &= P^{Y|Z=z,D=1}(A)P(D=1|Z=z) + P^{Y|Z=z,D=0}(A)P(D=0|Z=z). \end{aligned}$$

Thus, letting $\mathcal{Z}$ denote the support of Z,

$$R(P^{Y}) = \cap_{z\in\mathcal{Z}} \left\{ P^{Y|Z=z,D=1}P(D=1|Z=z) + \mu P(D=0|Z=z), \mu \in M \right\}.$$

Missing data: identification with IV

- If $P(D=1|Z=z)=1$ for some z, then P^Y is point identified. More generally, the identification region shrinks as the effect of Z on D increases.
- The set may also be empty. In this case, we reject the assumption $Y \perp\!\!\!\perp Z$. This assumption is thus refutable.
- Example (continued): suppose that nonresponse is related to the kind of interviewer (experimented, $Z=1$, or not). For experimented interviewer, we have 13% of nonresponse and an unemployment rate of 8%, while for unexperimented the nonresponse rate is 17% and the unemployement rate is 14%. Finally, $P(Z=1)=0.5$.
- Then the identification interval for the unemployment rate is $[7\%; 20\%] \cap [11.6\%; 28.6\%] = [11.6\%; 20\%]$. Without any assumption it would be equal to $[9.3\%, 24.3\%]$.

Missing data: identification with monotonicity

- Instead of using $Y \perp\!\!\!\perp Z$, we may suppose that Y is *missing at random*, i.e., for covariates X,

$$Y \perp\!\!\!\perp D|X.$$

- This assumption points identify the distribution of Y conditional on X, since $P^{Y|X} = P^{Y|D=1,X}$. The marginal distribution P^Y is also point identified.
- However, this assumption is often considered too stringent. We may instead suppose a *mean missing monotonically assumption*, which asserts that for all increasing h,

$$E(h(Y)|D = 0, X) \leq E(h(Y)|X) \leq E(h(Y)|D = 1, X).$$

Missing data: identification with monotonicity

- In this case, the minimal identification region of $g(P^Y) = E(h(Y))$ is

  $$R(g(P^Y)) = \left[pE(h(Y)|D=1) + (1-p)h(\underline{y}), E(E(h(Y)|D=1,X))\right].$$

- Remark 1: the lower bound is $-\infty$ when $h(\underline{y}) = -\infty$, whereas the upper bound is always finite.
- Remark 2: contrary to the independence assumption $Y \perp\!\!\!\perp Z$, the *mean missing monotonically* assumption is non-refutable.

Treatment effects: introduction

- An important example of missing data arises in analyzing treatment effects. Let $D \in \mathcal{D}$ denote a treatment and $Y(d)$ denote the potential outcome corresponding to $D = d$. We do not observe, for each individual, $(Y(d))_{d \in \mathcal{D}}$ but only $Y \equiv Y(D)$.
- The aim is to recover the distribution of $Y(d)$, or treatment effects such as $E(Y(d) - Y(d')), d \neq d'$.
- Using the empirical evidence alone, we get, as previously:

$$R(P^{Y(d)}) = \left\{ P^{Y|D=d} P(D = d) + \mu(1 - P(D = d)), \mu \in M \right\}.$$

Treatment effects: first assumptions

- We can rely on similar assumptions as previously. Suppose for instance that treatment is exogenous:

$$(Y(d))_{d\in\mathcal{D}} \perp\!\!\!\perp D|X.$$

Then $P^{Y(d)|X}$ (and $P^{Y(d)}$) is point identified by

$$P^{Y(d)|X} = P^{Y|X,D=d}.$$

- Suppose that we have an instrument affecting D and such that

$$(Y(d))_{d\in\mathcal{D}} \perp\!\!\!\perp Z. \quad (1)$$

Then we have an identification region defined by

$$R(P^{Y(d)}) = \cap_{z\in\mathcal{Z}} \left\{P^{Y|Z=z,D=d}P(D=d|Z=z) + \mu P(D\neq d|Z=z), \mu\in M\right\}.$$

- N.B.: randomized experiments are a special case of (1), where $Z=D$. In this case $P(D=d|Z=z) = \mathbb{1}\{z=d\}$ and $R(P^{Y(d)})$ is reduced to the singleton $\{P^{Y|D=d}\}$.

Treatment effects: monotonicity conditions

- Another interesting assumption is *monotone treatment response* (MTR). Suppose that $\mathcal{D}$ is ordered, and assume that:

$$d \geq d' \Rightarrow Y(d) \geq Y(d').$$

- Such an assumption is relevant in production analysis, demand estimation, returns to schooling...
- Let

$$\begin{aligned} \underline{Y}(d) &= Y\mathbb{1}\{D \leq d\} + \underline{y}\mathbb{1}\{D > d\} \\ \overline{Y}(d) &= Y\mathbb{1}\{D \geq d\} + \overline{y}\mathbb{1}\{D < d\}. \end{aligned}$$

- Then the minimal identification region of any $D-$ parameter $g(P^{Y(d)})$ is

$$R(g(P^{Y(d)})) = [g(P^{\underline{Y}(d)}), g(P^{\overline{Y}(d)})].$$

Treatment effects: monotonicity conditions

- Let for instance $g(P^{Y(d)}) = E[h(Y(d))]$, with h increasing. In this case,

$$\begin{aligned} R(g(P^{Y(d)})) &= \big[h(\underline{y})P(D > d) + E[h(Y)|D \leq d]P(D \leq d), \\ &\quad h(\overline{y})P(D < d) + E[h(Y)|D \geq d]P(D \geq d)\big]. \end{aligned}$$

- We can also characterize the minimal identification region of $\Delta_1 = g(P^{Y(d)}) - g(P^{Y(d')})$, whenever $g(P^{Y(d)})$ is a $D-$ parameter and $d > d'$:

$$R(\Delta_1) = [0, g(P^{\overline{Y}(d)}) - g(P^{\underline{Y}(d')})].$$

- Similarly, we have the following minimal identification region for $\Delta_2 = g(P^{Y(d)-Y(d')})$:

$$R(\Delta_2) = [g(\delta_0), g(P^{\overline{Y}(d)-\underline{Y}(d')})].$$

Treatment effects: monotonicity conditions

- Monotonicity allows in general to shrink the identification region.
- However, when $Y(d)$ is not bounded a priori, $\underline{y} = -\infty$ and $\overline{y} = \infty$, the identification region on $E[Y(d)]$ (or $E[Y(d)] - E[Y(d')]$) is $[-\infty, +\infty]$.
- Now suppose that we combine MTR with a *monotone treatment selection* assumption (MTS):

$$d_2 \geq d_1 \Rightarrow E[Y(d)|D = d_2] \geq E[Y(d)|D = d_1].$$

- In the returns to schooling example, this assumption posits a positive ability bias: individuals who make more schooling are those with the higher "intrinsic productivity".

Treatment effects: monotonicity conditions

- Under MTR-MTS, the minimal identification interval of $m(d) = E[Y(d)]$ is:

$$R(m(d)) = \left[\sum_{d'<d} E(Y|D=d')P(D=d') + E(Y|D=d)P(D \geq d), \right. \\ \left. \sum_{d'>d} E(Y|D=d')P(D=d') + E(Y|D=d)P(D \leq d)\right].$$

- Both bounds are finite here.

Treatment effects: monotonicity conditions

- Similarly, we have a finite (not minimal) identification region for treatment effect parameters $\Delta(d_1, d_2) = E[Y(d_2) - Y(d_1)]$, $d_2 > d_1$:

$$\begin{aligned} R &= \Bigg[0, \sum_{d>d_2} E[Y|D=d)P(D=d) + E(Y|D=d_2)P(D \leq d_2) \\ &\quad - \sum_{d'<d_1} E(Y|D=d')P(D=d') - E(Y|D=d_1)P(D \geq d_1)\Bigg]. \end{aligned}$$

- The upper bound is sharp, but not the lower bound.

Treatment effects: example of the returns to schooling

- Manski and Pepper (2000) investigate returns to schooling using MTR-MTS instead of standard instrumental variables satisfying $Z \perp\!\!\!\perp (Y(d))_{d \in \mathcal{D}}$.
- They use data from the NLSY 1979, and restrict themselves to the 1257 white males who reported in 1994 that they were full-time year-round workers but not self-employed, and who reported their wages:

Variable	12	13	14	15	16	17	18	19
$E(Y\|D=d)$	2.50	2.66	2.64	2.69	2.87	2.78	3.01	3.01
$P(D=d)$	0.413	0.074	0.083	0.035	0.189	0.038	0.051	0.020

Empirical mean hourly log(wage) and distribution of years of schooling (taken from Manski & Pepper, 2000)

Treatment effects: example of the returns to schooling

- The previous result allows to compute the upper bound of $\Delta(d_1, d_2) = E(Y(d_2) - Y(d_1))$:

Upper bound	12-13	13-14	14-15	15-16	16-17	17-18	18-19	12-16
Estimate	0.25	0.16	0.20	0.30	0.16	0.39	0.37	0.40
0.95 quantile	0.31	0.23	0.29	0.37	0.26	0.48	0.54	0.45

- The upper bounds on $\Delta(t-1, t)$ are above the usual estimates in the literature. Thus MTS-MTR does not, in this application, have sufficient identifying power to change the consensus on the returns to schooling.
- On the other hand, the upper bound on $\Delta(12, 16)$ implies that the average value of the four year-by-year treatment effects are at most 0.10, which is below the point estimates of Card (1993) and Ashenfelter and Krueger (1994).

Outline

Introduction

First examples: missing data

Second example: incomplete models

Inference for partially identified models

Introduction

- In many economic settings such as entry games (see, e.g., Tamer, 2003), bargaining models (see Engers and Stern, 2002) or discrete choice models with social interactions (see, e.g., Krauth, 2006), we are confronted with multiple equilibria.
- In this case, we only know that the decision $Y \in \mathcal{Y}$ satisfies $Y \in G(\varepsilon|X,\theta)$, where ε (resp. X) denotes the unobserved component (resp. covariates), θ is the parameter of interest and $G(.|X,\theta)$ is a correspondence, i.e., a set valued function.
- Such models are called *incomplete*, because they do not specify a unique outcome for given (X,ε). Because of this, θ will generally be set identified only.

Example: a simple entry game

- Consider the two player entry game of the introduction:

	2 enters	2 does not enter
1 enters	$(\alpha_1 + \beta + \varepsilon_1, \alpha_2 + \beta + \varepsilon_2)$	$(\alpha_1 + \varepsilon_1, 0)$
1 does not enter	$(0, \alpha_2 + \varepsilon_2)$	$(0, 0)$

- We also assume that $\beta \leq 0$ and suppose for simplicity that $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}(0, I_2)$.
- When $(\varepsilon_1, \varepsilon_2) \in [-\alpha_1, -\alpha_1 - \beta] \times [-\alpha_2, -\alpha_2 - \beta]$, there are two equilibria, $(0, 1)$ and $(1, 0)$ (otherwise there is just one equilibrium).

The problem

- Even if the model is fully parametric, the problem is that we may not be able to write down the likelihood of $Y = y$, because this event may not correspond to a well specified region of ε.
- Instead, we have inequalities which come from the definition of (for instance) Nash equilibria. Let $\mathcal{U}_i(y_i, y_{-i}, X_i, \varepsilon_i, \theta)$ denote the utility for i when playing y_i, given that others play y_{-i}. We have:

$$\mathcal{U}_i(Y_i, Y_{-i}, X_i, \varepsilon_i, \theta) \geq \mathcal{U}_i(y, Y_{-i}, X_i, \varepsilon_i, \theta) \quad \forall y. \tag{2}$$

- This provides *necessary* conditions for observing $Y = (Y_i, Y_{-i})$.
- In the example above, for instance, we have

$$(Y_1 = 0, Y_2 = 1) \Rightarrow (\alpha_1 + \beta + \varepsilon_1 < 0, \alpha_2 + \varepsilon_2 \geq 0).$$

- This implies the inequality:

$$P(Y_1 = 0, Y_2 = 1|\theta) \leq \Phi(-\alpha_1 - \beta)\Phi(\alpha_2).$$

Results for parametric models

- More generally, for any $A \subseteq \mathcal{Y}$, let

$$G^{-1}(A|x,\theta) = \{u : G(u|x,\theta) \cap A \neq \emptyset\}.$$

- We have, for the true parameter θ_0 and for all $A \subseteq \mathcal{Y}$,

$$\begin{aligned} Y \in A &\iff Y \in G(\varepsilon|X,\theta_0) \cap A \\ &\implies G(\varepsilon|X,\theta_0) \cap A \neq \emptyset \\ &\iff \varepsilon \in G^{-1}(A|X,\theta_0). \end{aligned}$$

As a result,

$$P(Y \in A|X) \leq P(\varepsilon \in G^{-1}(A|X,\theta_0)).$$

Results for parametric models

- In a parametric model, the distribution of ε is fully specified and depends only on θ. Thus, we can compute, for a given θ, the function

$$\mathcal{L}(A|x,\theta) \equiv P(\varepsilon \in G^{-1}(A|x,\theta)|\theta).$$

and the true parameter satisfies, for all $A \subseteq \mathcal{Y}$,

$$P(Y \in A|X = x) \leq P(\varepsilon \in G^{-1}(A|x,\theta_0)|\theta_0).$$

- This implies

$$R(\theta_0) \subset \{\theta \in \Theta : \forall A \subseteq \mathcal{Y}, \forall x, P(Y \in A|X = x) \leq \mathcal{L}(A|x,\theta)\}.$$

- Actually, Galichon and Henry (2011) prove the stronger result:

$$\begin{aligned} R(\theta_0) &= \{\theta \in \Theta : \forall A \subseteq \mathcal{Y}, \forall x, P(Y \in A|X = x) \leq \mathcal{L}(A|x,\theta)\} \\ &= \{\theta \in \Theta : \min_x \min_{A \subseteq \mathcal{Y}} \mathcal{L}(A|x,\theta) - P(Y \in A|X = x) \geq 0\}. \end{aligned}$$

Results for parametric models

- The latter form is convenient as it shows that finding $R(\theta)$ amounts to minimize several times the *submodular* function $A \mapsto \mathcal{L}(A|X,\theta) - P(Y \in A|X)$. A real function f on sets is submodular if, for all (A, B), it satisfies the inequality

$$f(A \cap B) + f(A \cup B) \leq f(A) + f(B).$$

- Submodularity for functions on sets is equivalent to convexity for functions on $\mathbb{R}^k$, and there are efficient algorithms for minimizing them.
- Another way to compute $R(\theta_0)$ is to reduce the number of inequalities to check for each (x, θ) (namely, $2^{|\mathcal{Y}|} - 2$ if $\mathcal{Y}$ is discrete). This is possible if outcomes satisfy a monotonicity condition (see Galichon and Henry, 2011).

Application to the entry game

- At a first glance, we have $2^4 - 2 = 14$ inequalities to check. However, this number can be much reduced.
- Let $p_{ij} = P(Y_1 = i, Y_2 = j)$, we have indeed

$$\begin{aligned} p_{00} &\leq \Phi(-\alpha_1)\Phi(-\alpha_2) \\ p_{11} &\leq \Phi(\alpha_1 + \beta)\Phi(\alpha_2 + \beta) \\ 1 - p_{00} - p_{11} &\leq 1 - \Phi(-\alpha_1)\Phi(-\alpha_2) - \Phi(\alpha_1 + \beta)\Phi(\alpha_2 + \beta) \end{aligned}$$

- Thus, the three inequalities are actually equalities. Then one can show that the identification region is defined by

$$\begin{aligned} p_{00} &= \Phi(-\alpha_1)\Phi(-\alpha_2) \\ p_{11} &= \Phi(\alpha_1 + \beta)\Phi(\alpha_2 + \beta) \\ p_{10} &\leq \Phi(\alpha_1)\Phi(-\alpha_2 - \beta) \\ p_{01} &\leq \Phi(-\alpha_1 - \beta)\Phi(\alpha_2). \end{aligned}$$

Application to the entry game

- Note that the identification region has at most dimension 1 since α_2 and β are functions of α_1.
- Example: suppose that the true values are $\alpha_1 = 1$, $\alpha_2 = 0.5$ and $\beta = -1$. When there are two equilibria, each one is drawn with probability 0.5. In this case,

$$\begin{aligned} R(\alpha_1) &= [0.86, 1.10] \\ R(\alpha_2) &= [0.35, 0.67] \\ R(\beta) &= [-1.03, -0.95]. \end{aligned}$$

  Such a model is very informative on β, less so for α_1 and α_2.
- Same model with $\alpha_1 = \alpha_2 = 1$. In this case we get almost point identification of β: $R(\beta) = [-1, -0.99]$.
- Same model with a true probability of selection of equilibria at 0, we get $R(\alpha_1) = [1, 1.20]$, $R(\alpha_2) = [0.19, 0.50]$, $R(\beta_0) = [-1, -0.87]$. Intuition?

Results for semiparametric models

- Consider an additively separable utility function:

$$\mathcal{U}_i(y, Y_{-i}, X_i, \varepsilon_i, \theta) = g(y, Y_{-i}, X_i, \theta) + \varepsilon_{iy},$$

where ε_{iy} is a shock specific to decision y. Suppose that $E(\varepsilon_{iy}|X)$ is independent of y. Then (2) implies:

$$E\left[g(Y_i, Y_{-i}, X_i, \theta_0) - g(y, Y_{-i}, X_i, \theta_0)|X\right] \geq 0 \quad \forall y.$$

- We can use these *moment inequalities* to characterize $R(\theta_0)$.
- Note that moment inequalities conditions also arise in linear models with interval-valued outcome:

$$Y^* = X'\theta_0 + \varepsilon, \; E(\varepsilon|X) = 0, \; Y^* \in [\underline{Y}, \overline{Y}].$$

Outline

Introduction

- Inference for identified sets has been developed very recently, since the middle of the 00's only (see, e.g., Imbens and Manski, 2004, Chernozhukov et al, 2007, Beresteanu and Molinari, 2008, Rosen, 2008, Bontemps et al., 2012, and several papers by Andrews, Shaikh...).
- There still lacks a unifying theory for the moment.
- I will only present inference when either the identification region is an interval or is defined by moment inequalities.

Inference with interval region

- Suppose that $R(\theta_0) = [\underline{\theta}, \overline{\theta}]$. We can estimate it by $\widehat{R(\theta)} = [\widehat{\underline{\theta}}, \widehat{\overline{\theta}}]$.
- The choice of a confidence interval is more delicate. We can impose one of the two conditions:

$$\lim_{n\to\infty} P(\text{CI}_{1-\alpha} \supset R(\theta_0)) \geq 1-\alpha \qquad (3)$$

$$\inf_{\theta\in R(\theta_0)} \lim_{n\to\infty} P(\text{CI}_{1-\alpha} \ni \theta) \geq 1-\alpha \qquad (4)$$

- The first condition is more restrictive and thus leads to larger confidence intervals.

Inference with interval region

- Suppose that
$$\sqrt{n}\left(\widehat{\underline{\theta}}-\underline{\theta},\widehat{\overline{\theta}}-\overline{\theta}\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\Sigma\right).$$
- Then (3) is satisfied if we take
$$\mathrm{CI}^1_{1-\alpha}=\left[\widehat{\underline{\theta}}-q_{1-\alpha/2}\sqrt{\Sigma_{11}},\widehat{\overline{\theta}}+q_{1-\alpha/2}\sqrt{\Sigma_{22}}\right],$$
where q_β is the β-th quantile of a $\mathcal{N}(0,1)$.
- On the other hand, if $\underline{\theta}<\overline{\theta}$, (4) is satisfied with the smaller interval
$$\mathrm{CI}^2_{1-\alpha}=\left[\widehat{\underline{\theta}}-q_{1-\alpha}\sqrt{\Sigma_{11}},\widehat{\overline{\theta}}+q_{1-\alpha}\sqrt{\Sigma_{22}}\right].$$
- A problem is that the latter is valid only if $\underline{\theta}<\overline{\theta}$. There is a discontinuity at $\underline{\theta}=\overline{\theta}$ since the asymptotic coverage of the interval is only $1-2\alpha$ in this case. See Imbens and Manski (2004) for a modification of $\mathrm{CI}^2_{1-\alpha}$ that overcomes this issue.

Inference with moment inequalities

- Consider the case where $R(\theta_0)$ corresponds to all θ satisfying

$$E\left[m(Y,\theta)\right] \geq 0,$$

where $m(.,.) \in \mathbb{R}^k$ and inequalities hold componentwise. This is the case in some missing data problems, in games with multiple equilibria or in regression with interval data outcomes:

$$E(\overline{Y} - X'\theta | X) \geq 0, \; E(X'\theta - \underline{Y} | X) \geq 0.$$

These conditional inequalities imply indeed that for all positive $g(.)$,

$$E\left(g(X)\left(\overline{Y} - X'\theta\right)\right] \geq 0, \; E\left(g(X)\left(X'\theta - \underline{Y}\right)\right] \geq 0.$$

Inference with moment inequalities: estimation

- Let $E_n(.)$ denote the empirical mean. Following Rosen (2008), consider the objective function

$$Q_n(\theta) = \min_{t \geq 0} \left[E_n(m(Y,\theta)) - t\right]' \widehat{V}_\theta^{-1} \left[E_n(m(Y,\theta)) - t\right],$$

  where $\widehat{V}_\theta$ is the estimator of the variance matrix of $m(Y,\theta)$.

- Note that $Q_n(\theta)$ can be computed more easily by applying Kuhn-Tucker conditions (which are necessary and sufficient here).
- Under standard assumptions,

$$Q_n(\theta) \xrightarrow{P} Q(\theta) = \min_{t \geq 0} \left[E(m(Y,\theta)) - t\right]' V_\theta^{-1} \left[E(m(Y,\theta)) - t\right].$$

- Moreover, $Q(\theta) = 0$ if and only if $\theta \in R(\theta_0)$. Thus, to estimate $R(\theta_0)$, the idea is to consider the set of θ such that $Q_n(\theta)$ is small.

Inference with moment inequalities: estimation

- More formally, let ε_n be such that $\varepsilon_n \to \infty$ and $\varepsilon_n / n \to 0$, and define $\widehat{R(\theta_0)}$ by

$$\widehat{R(\theta_0)} = \{\theta : nQ_n(\theta) \leq \varepsilon_n\}.$$

Theorem

(Chernozhukov et al, 2007) Suppose that $\sup_\theta |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0$. *Then, letting* $d_H(.,.)$ *denote the Hausdorff distance between sets (i.e.,* $d_H(A, B) = \max(\sup_{a \in A} \inf_{b \in B} ||a - b||, \sup_{b \in B} \inf_{a \in A} ||a - b||)$*) :*

$$d_H(\widehat{R(\theta_0)}, R(\theta_0)) \xrightarrow{P} 0.$$

Inference with moment inequalities: estimation

- Intuition behind: we have $Q_n(\theta) \to 0$ for all $\theta \in R(\theta_0)$ and $Q_n(\theta) \to c > 0$ otherwise. Thus, choosing the θ such that $Q_n(\theta) \leq \varepsilon_n / n \to 0$ will asymptotically pick only elements of $R(\theta_0)$.
- Note however that such a set may be large for finite n. Under some additional restrictions, we can show that taking the smaller set $\widetilde{R(\theta_0)}$ defined by

$$\widetilde{R(\theta_0)} = \arg\min_{\theta} nQ_n(\theta)$$

is also consistent (see Chernozhukov et al, 2007).

Inference with moment inequalities: confidence interval

- Confidence intervals can be built using the following distributional result.

Theorem

(Rosen, 2008) Let $b(\theta)$ denote the number of components $m_j(Y,\theta)$ such that $E(m_j(Y,\theta)) = 0$, $V^*(\theta)$ denotes the variance matrix of these components and let $Z \sim \mathcal{N}(0, V^*(\theta))$. Then

$$
\begin{aligned}
nQ_n(\theta) &\xrightarrow{\mathcal{L}} \min_{s\geq 0}(Z-s)'V^*(\theta)^{-1}(Z-s) \quad \text{if } \theta \in R(\theta_0), b(\theta) > 0, \\
nQ_n(\theta) &\xrightarrow{\mathcal{L}} 0 \quad \text{if } \theta \in R(\theta_0), b(\theta) = 0, \\
nQ_n(\theta) &\xrightarrow{P} \infty \quad \text{if } \theta \notin R(\theta_0).
\end{aligned} \tag{5}
$$

- To construct a confidence region on $R(\theta_0)$, we will therefore consider $CI_{1-\alpha} = \{\theta : nQ_n(\theta) \leq c_{1-\alpha}\}$, for an appropriate $c_{1-\alpha}$.

Inference with moment inequalities: confidence interval

- The theorem shows that strict moment inequalities are not binding asymptotically and we may only focus on moment equalities (as when we test $\theta \leq 0$ versus $\theta > 0$).
- If $b(\theta)$ and $V^*(\theta)$ were known, the distribution of $T = \min_{s \geq 0}(Z - s)' V^*(\theta)^{-1}(Z - s)$ could be tabulated (by simulations).
- However, $b(\theta)$ and $V^*(\theta)$ are unknown.

Inference with moment inequalities: confidence interval

- To overcome this issue, one can use the following inequality (see Rosen, 2008, Corollary 1):

$$\inf_{\theta \in R(\theta_0)} \lim_{n \to \infty} P(nQ_n(\theta) \leq c) \geq \frac{1}{2} P(\chi^2_b \leq c) + \frac{1}{2} P(\chi^2_{b-1} \leq c).$$

  for any $b \geq \sup_{\theta \in R(\theta_0)} b(\theta)$. One can choose for instance $b =$dim$(m(Y, \theta))$.

- Thus, we can define $c_{1-\alpha}$ to be such that

$$\frac{1}{2} P(\chi^2_b \leq c_{1-\alpha}) + \frac{1}{2} P(\chi^2_{b-1} \leq c_{1-\alpha}) = 1 - \alpha.$$

- Such a $c_{1-\alpha}$ yields an asymptotically conservative confidence interval. It is possible to build exact confidence intervals when $V^*(\theta)$ is diagonal (see Rosen, 2008).
- Note that the above procedure can also be used to test whether $\theta_1 \in R(\theta_0)$. Such a test is consistent by (5).