Identification in Models with Discrete Variables

Lukáš Lafférs

Norwegian School of Economics

May 19, 2011

Outline

- Short Introduction to Partial Identification
- Galichon and Henry setup
- My Extension of Galichon and Henry Setup
- Demonstration that it works
- What can I do using it: Imperfect Instruments example

Introduction to Partial Identification

Econometricians typically work with point-identified models, e.g.
$Y_i = X_i'\beta + U_i \quad E(U_i|X_i) = 0$, elements of X_i not perfectly correlated
there exists **only one** β that satisfies these assumptions and is compatible with the distribution of (Y_i, X_i) which is revealed by the data.

in certain situations our assumptions are not strong enough to determine a unique value of parameter but there is a set of **observationally equivalent models**.

surveys Manski(1995,2003), Tamer(2010)

Throughout this presentation I will discuss **Identification** not Inference.
It is assumed that we know the true data generating process of observable variables.

Example 1 - Missing Data

Manski (1990)
Suppose that we observe $Z_i = (D_i, D_i.Y_i)$,
where $D_i = 1$ when Y_i is observed
and $D_i = 0$ if we do not observe Y_i.

We would like to say something about the quantity of interest $\theta = E(Y_i)$.

We can learn something about $p = E(D_i)$ and $\mu_1 = E(Y_i|D_i = 1)$ but the data tell us **nothing** about $\mu_0 = E(Y_i|D_i = 0)$.

$\theta = p.\mu_1 + (1 - p).\mu_0$

Additionial assumptions needed, if e.g. $Y_i \in \{0, 1\}$ then

$\theta \in [\theta_{low}, \theta_{high}] = [p.\mu_1, p.\mu_1 + (1 - p)]$.

Example 1 - Missing Data

Manski (1990)
Suppose that we observe $Z_i = (D_i, D_i.Y_i)$,
where $D_i = 1$ when Y_i is observed
and $D_i = 0$ if we do not observe Y_i.

We would like to say something about the quantity of interest $\theta = E(Y_i)$.

We can learn something about $p = E(D_i)$ and $\mu_1 = E(Y_i|D_i = 1)$ but the data tell us **nothing** about $\mu_0 = E(Y_i|D_i = 0)$.

$\theta = p.\mu_1 + (1 - p).\mu_0$

Additionial assumptions needed, if e.g. $Y_i \in \{0, 1\}$ then

$\theta \in [\theta_{low}, \theta_{high}] = [p.\mu_1, p.\mu_1 + (1 - p)]$.

Example 2 - Multiple Equilibria

Jovanovic (1989)

Two players' game.

$\Pi_1(Y_1, Y_2, U_1, U_2) = (\theta Y_2 - U_2)1_{(Y_1=1)}$

$\Pi_2(Y_1, Y_2, U_1, U_2) = (\theta Y_1 - U_1)1_{(Y_2=1)}$

$U = (U_1, U_2)$ are exogenous costs, observed to firms, unobserved to econometrician and assumed to be $Unif(0,1)^2$

Two pure strategy Nash equilibria $\{(0,0),(1,1)\}$.

p - probability of observing $(1,1)$

θ is the unknown parameter of interest, what can we say about it?

$\theta \in [0, \sqrt{p}]$

Example 2 - Multiple Equilibria

Jovanovic (1989)
Two players' game.
$\Pi_1(Y_1, Y_2, U_1, U_2) = (\theta Y_2 - U_2)1_{(Y_1=1)}$
$\Pi_2(Y_1, Y_2, U_1, U_2) = (\theta Y_1 - U_1)1_{(Y_2=1)}$
$U = (U_1, U_2)$ are exogenous costs, observed to firms, unobserved to econometrician and assumed to be $Unif(0,1)^2$
Two pure strategy Nash equilibria $\{(0,0),(1,1)\}$.
p - probability of observing $(1,1)$
θ is the unknown parameter of interest, what can we say about it?
$\theta \in [0, \sqrt{p}]$

Galichon and Henry Framework - literature

Galichon and Henry (2006)
Galichon and Henry (2008, JoE)
Ekeland, Galichon and Henry (2010, Econ. Theory)
Galichon and Henry (forthcoming, REStud)

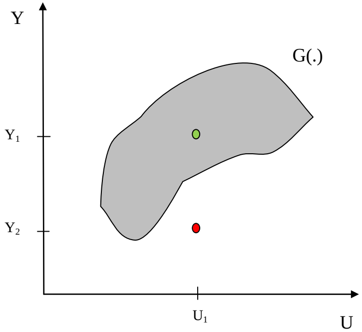
Galichon and Henry Framework (simplified)

Two types of variables:
Y - **Observable** variables ($Y \in \mathcal{Y}$ with density p)
U - **Unobservable** variables ($U \in \mathcal{U}$ with density ν)

Economic restrictions takes the form of
G - many-to-many mapping ($G : \mathcal{U} \mapsto \mathcal{Y}$)

Galichon and Henry Framework (simplified) (2)

Not all pairs (Y, U) are compatible with economic restrictions



(Y_1, U_1) is compatible $(Y_1 \in G(U_1))$
(Y_2, U_1) is not $(Y_2 \notin G(U_1))$

Galichon and Henry Framework (simplified) (3)

- *Structure* S is defined as a triplet $S = (G, \nu, p)$ (Jovanovich 1989, Koopmans and Reiersol 1950)
- Structure S is said to be *internally consistent* if and only if there exists a joint distribution π of (Y, U) on $\mathcal{Y} \times \mathcal{U}$ with marginals p and ν such that $\pi(\{Y \in G(U)\}) = 1$
- It means that S is compatible with data at hand and satisfies economic restrictions almost surely

Galichon and Henry Framework (simplified) (4)

- Now parametrize ν and G by $\theta \in \Theta$ (possibly $\theta = (\theta_\nu, \theta_G)$).
- $S_\theta = (G_\theta, \nu_\theta, p)$
- *Identified set* is
  $\Theta_I = \{\theta \in \Theta : S_\theta \text{ is internally consistent}\}$

Necessary and Sufficient condition for inclusion of θ into the identified set (Galichon and Henry 2009 - Theorem 1).

$$0 = \max_{A \in \mathcal{Y}} (Pr(A) - \nu_\theta(G_\theta^{-1}(A))),$$

This is a very strong and general result and its proof relies on the optimal transportation theory.

Galichon and Henry Framework (simplified) (5)

Consider discrete Y and U

c_{ij} is a 0-1 penalty on those pairs (Y, U) that are incompatible with the economic model.

$$c_{ij} = \begin{cases} 0, y_i \in G_\theta(u_j), \\ 1, \text{ otherwise} \end{cases}$$

Now the existence of a joint density which assures internal consistency of S can be formulated as a following linear program:

$$\begin{array}{cc} \min_{(\pi)} \sum_{i,j} \pi_{ij} c_{ij} & \\ \text{s.t.} & \\ \sum_j \pi_{ij} = p_i, & \forall i \\ \sum_i \pi_{ij} = \nu_j, & \forall j \\ \pi_{ij} \geq 0, & \forall i,j. \end{array}$$

My Extension of GH Framework

What if we had extra information:
$E(\phi_\beta(Y,U)) = 0$ and $|cov(Y,U)| \leq 0.1$?

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\
&\text{s.t.} \\
&\textstyle\sum_j \pi_{ij} = p_i, && \forall i \\
&\textstyle\sum_i \pi_{ij} = \nu_j, && \forall j \\
&\textstyle\sum_{i,j} \pi_{ij} \phi_\beta(y_i, u_j) = 0, \\
&\textstyle\sum_{i,j} \pi_{ij}(y_i - \sum_k y_k)(u_j - \sum_l u_l) \leq 0.1, \\
&-\textstyle\sum_{i,j} \pi_{ij}(y_i - \sum_k y_k)(u_j - \sum_l u_l) \leq 0.1, \\
&\pi_{ij} \geq 0, && \forall i,j.
\end{aligned}
$$

My Extension of GH Framework (2)

In general we can add any distributional restrictions.

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j} \pi_{ij} c_{ij} \\
&\text{s.t.} \\
&\textstyle\sum_j \pi_{ij} = p_i, && \forall i \\
&\textstyle\sum_i \pi_{ij} = \nu_j, && \forall j \\
&\psi_{1,\theta}(\pi, \mathbf{y}, \mathbf{u}) = \mathbf{0}_{\mathbf{k_1}}, \\
&\psi_{2,\theta}(\pi, \mathbf{y}, \mathbf{u}) \leq \mathbf{0}_{\mathbf{k_2}}, \\
&\pi_{ij} \geq 0, && \forall i, j.
\end{aligned}
$$

My Extension of GH Framework (3)

What can be done using this extension ?

To show that it works I replicate few results from partial identification literature that were obtained by distinct approaches.

In addition: I show how to see the strength of the assumption of a strict exogeneity of instruments in a nonlinear model with discrete variables.

Single Equation Endogenous Binary Response Model

Model studied in Chesher (2010, ECTA).

- (Y, X, Z) - **Observable** variables
  $((Y, X, Z) \in \{0,1\} \times \{0,1\} \times \{z_1, z_2, ..., z_k\}$ with density $p_{ijk})$
- U - **Unobservable** variables $(U \sim Unif(0,1))$

The economic restrictions are

$$Y = h(X, U) \quad = \quad \begin{cases} 0, & \text{if } U \leq t(X), \\ 1, & \text{if } U > t(X) \end{cases} \tag{1}$$

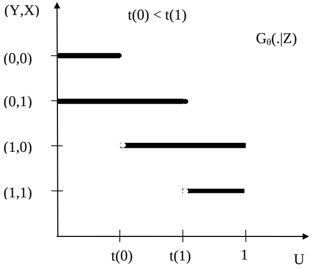
$$Y = h(X, U) \Leftrightarrow (Y, X, Z) \in G_\theta(U)$$

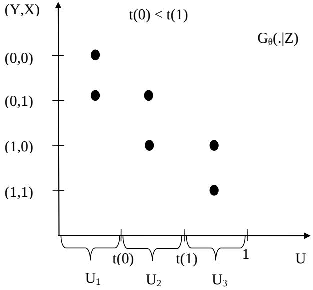
Further assumptions
$U \perp Z$, $t(X) = \Phi(-\theta_0 - \theta_1 X)$.

What can we tell about (θ_0, θ_1) ?

Formulation in the extended GH framework

Discretization





Formulation in extended GH framework (2)

$$\pi_{ijkl} = Pr(Y = y_i, X = x_j, Z = z_k, U = u_l)$$

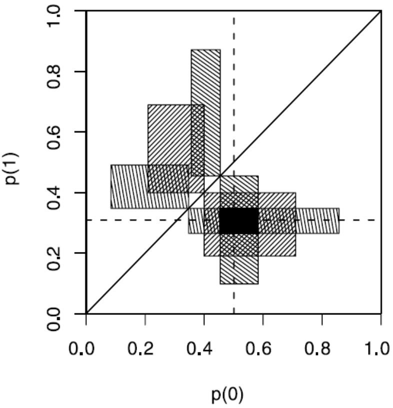
Penalty is given by

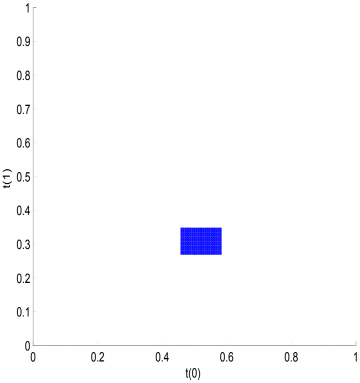
$$c_{ijkl} = \begin{cases} 0, y_i = h(x_j, u_l)), \\ 1, \text{ otherwise.} \end{cases}$$

Problem is formulated as

$$\begin{array}{rl} \min_{(\pi)} \sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} & \\ \text{s.t.} & \\ \sum_l \pi_{ijkl} = p_{ijk}, & \forall i, j, k \\ \sum_{i,j,k} \pi_{ijkl} = \nu_l, & \forall l \\ \sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk} \nu_l, & \forall k, l \\ \pi_{ijkl} \geq 0, & \forall i, j, k, l. \end{array}$$

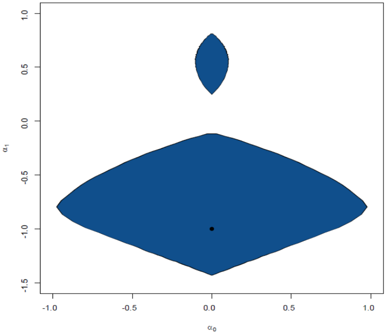
Comparison of Results

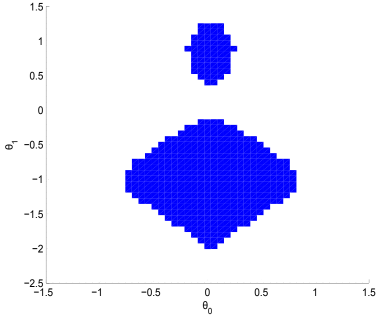




Comparison of Results (with continuous X)

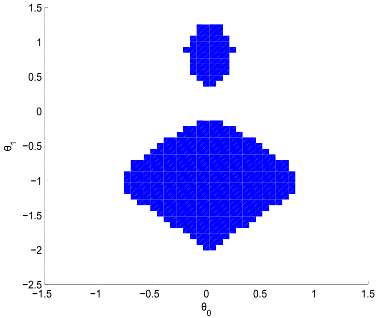
X and U are discretized here



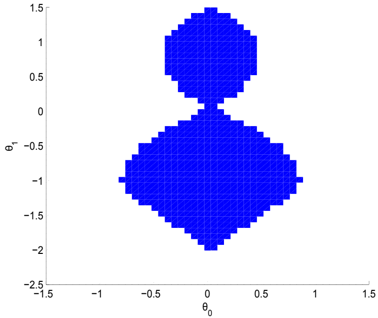


Discretization matters!

How independency matters

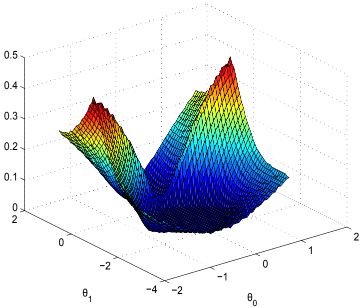


U and Z independent

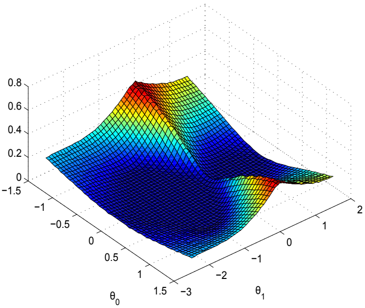


U and Z not independent

3D figures

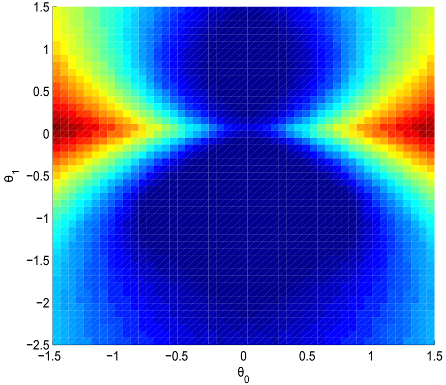


0.5
0.4
0.3
0.2
0.1
0
2
0
−2
−4
θ1
−2
−1
0
1
2
θ0



0.8
0.6
0.4
0.2
0
−1.5
−1
−0.5
0
0.5
1
1.5
θ0
−3
−2
−1
0
1
2
θ1

Contour Plot



The value of minimized objective function stands for the minimal probability of the event incompatible with economic restrictions

Exogeneity assumption relaxed

Why is it interesting?

- To see the strength of the assumption that cannot be tested
- Sensitivity analysis

Exogeneity assumption relaxed (2)

Recall exogenous case

$$
\begin{aligned}
\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} & \\
\text{s.t.} & \\
\textstyle\sum_{l} \pi_{ijkl} = p_{ijk}, & \quad \forall i,j,k \\
\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, & \quad \forall l \\
\textstyle\sum_{i,j} \pi_{ijkl} = \textstyle\sum_{i,j} p_{ijk} \nu_l, & \quad \forall k,l \\
\pi_{ijkl} \geq 0, & \quad \forall i,j,k,l.
\end{aligned}
\tag{2}
$$

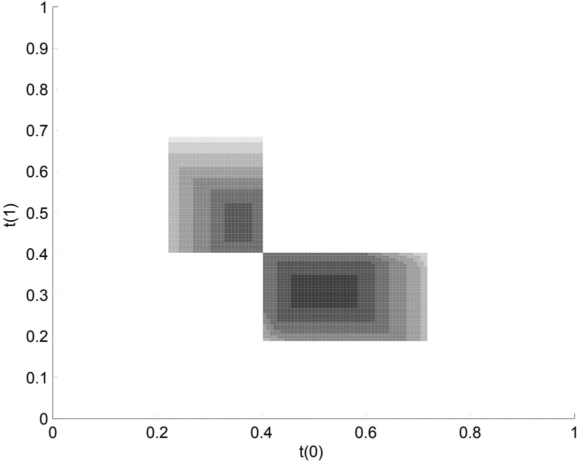
Exogeneity assumption relaxed (2)

Now the Z and U are only "close" to be independent.

$$\begin{array}{rl}
\min_{(\pi)} \sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} & \\
\text{s.t.} & \\
\sum_l \pi_{ijkl} = p_{ijk}, & \forall i,j,k \\
\sum_{i,j,k} \pi_{ijkl} = \nu_l, & \forall l \\
\sum_{i,j} \pi_{ijkl} - \sum_{i,j} p_{ijk}\nu_l \leq \delta, & \forall k,l \\
-\sum_{i,j} \pi_{ijkl} + \sum_{i,j} p_{ijk}\nu_l \leq \delta, & \forall k,l \\
\pi_{ijkl} \geq 0, & \forall i,j,k,l.
\end{array} \tag{3}$$

Still a linear program - computationally feasible.

Exogeneity assumption relaxed (3)



$\delta = [00.010.020.030.040.050.060.0751]$

Open Questions

- How to do inference
- Continuous variables (or what is the effect of discretization?)
- Computational aspects (size of π max 35000(?))
- Make it more "user friendly"

Conclusion

- Extension of an existing framework for incompletely specified models with discrete variables
- Can replicate some existing results from partial identification literature
- It is possible to see the identification "strength" of the exogeneity of instruments in non-linear models with discrete variables

Thank you for your attention!