Presentation of PhD project

Identification in models with discrete variables

Lukáš Lafférs

NHH Norwegian School of Economics

September 30, 2011

Motivation

Model Uncertainty in Economics

Typical approaches:

- Model Selection
- Model Averaging
- Partial Identification

PhD plan

- 1st paper - to be presented today
- 2nd paper - extention of the 1st paper
- 3rd paper - an application of the proposed method

Research questions

done **1st paper**

- Propose a method to determine the identified set in a broad setting
- Replicate existing results in the partial identification literature
- Propose a flexible method to study sensitivity of the identification to different identifying assumptions

planned **2nd paper**

- Find out how inference (confidence regions and specification tests) can be done within this framework
- Study and improve computational aspects of the framework

planned **3rd paper**

- Demonstrate the techniques on a relevant application

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

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 a parameter but there is a set of **observationally equivalent models**.
Meaning that no amount of data would ever help me to distinguish between these models.

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

Examples

Example 1 - Manski (1990) - Missing data
We are interested in $\theta = E(Y)$, it is only observed when $D = 1$.
$\theta = E(Y) = E(Y|D = 1)P(D = 1) + E(Y|D = 0)P(D = 0)$
$\theta = p.\mu_1 + (1 - p).\mu_0$

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 - Games with multiple equilibria
Equilibrium selection mechanism may not be available or plausible.
It is not easy to see what is the identified set even in the very simple setting

▸ Jovanovic (1989) example

Galichon and Henry Framework - literature

Galichon and Henry (2006)
Galichon and Henry (2009, JoE)
Ekeland, Galichon and Henry (2010, Econ. Theory)
Galichon and Henry (forthcoming, REStud)

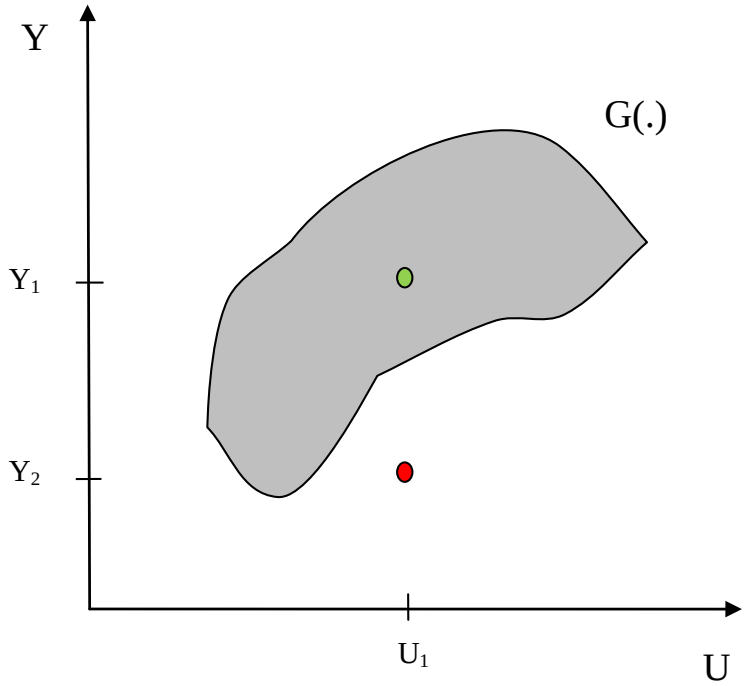
Galichon and Henry Framework (simplified)

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 take 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).

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

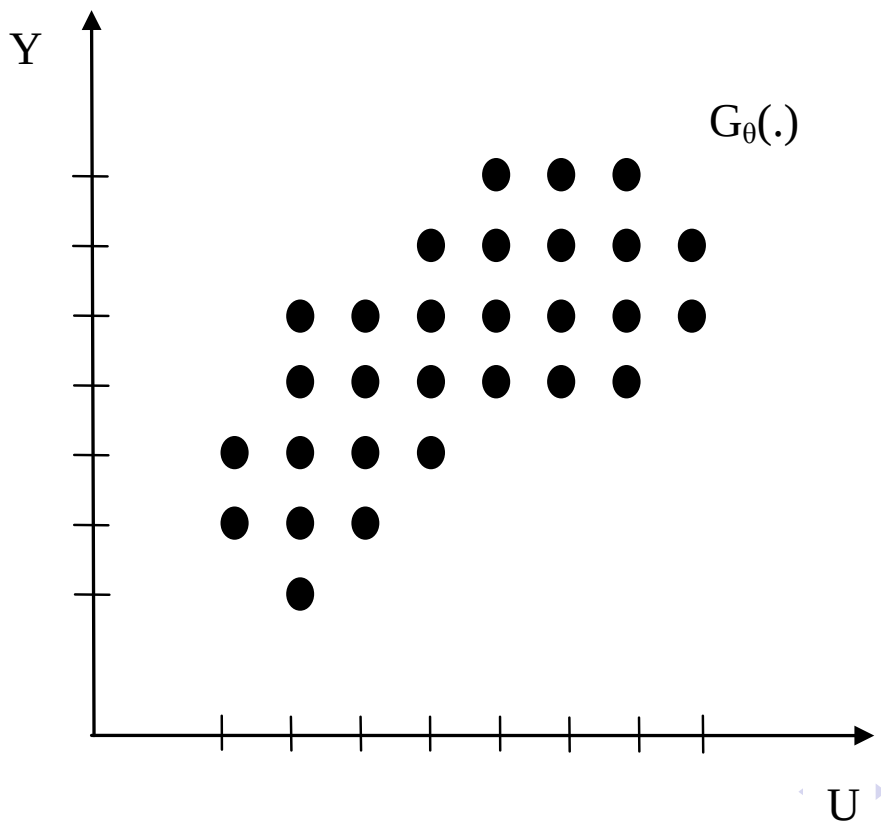
This is a very strong and general result.

Galichon and Henry Framework (simplified) (5)

Consider discrete Y and U.

c_{ij} is a 0-1 penalty on (Y, U) incompatible with G_θ.

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



Galichon and Henry Framework (simplified) (6)

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

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

My Extension of GH Framework

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

My Extension of GH Framework

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

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

My Extension of GH Framework (2)

In general we can add any "distributional" restrictions.

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

My Extension of GH Framework (3)

What can be done using this extension ?

To show that it works I replicate (in a simple fashion) 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

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

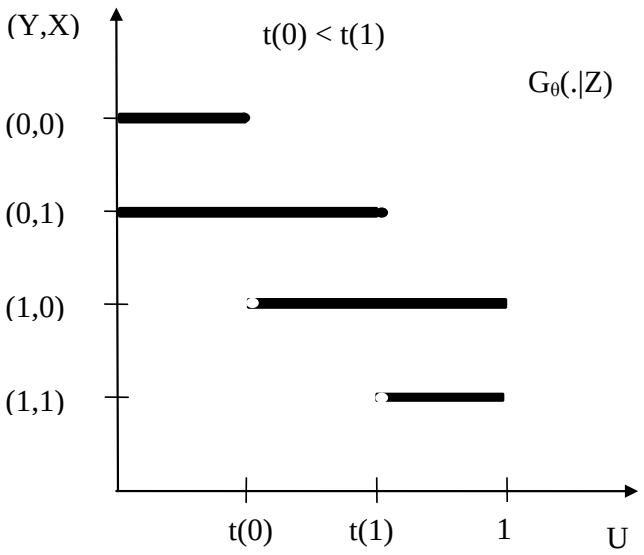
$$Y = h(X, U) \Leftrightarrow (Y, X, Z) \in G_\theta(U)$$

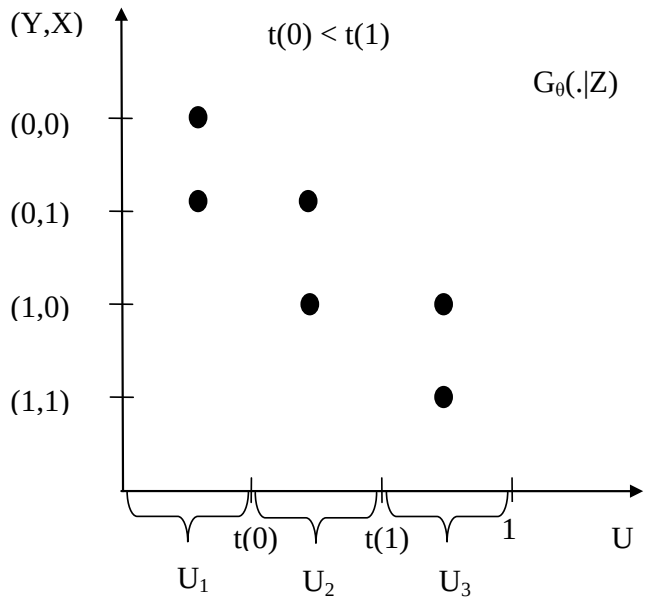
Further assumptions
$U \perp Z$, $t(X) = \Phi(-\theta_0 - \theta_1 X)$.

What can we tell about (θ_0, θ_1) ?

Formulation in the extended GH framework

Support restrictions and Discretization





Formulation in extended GH framework (2)

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

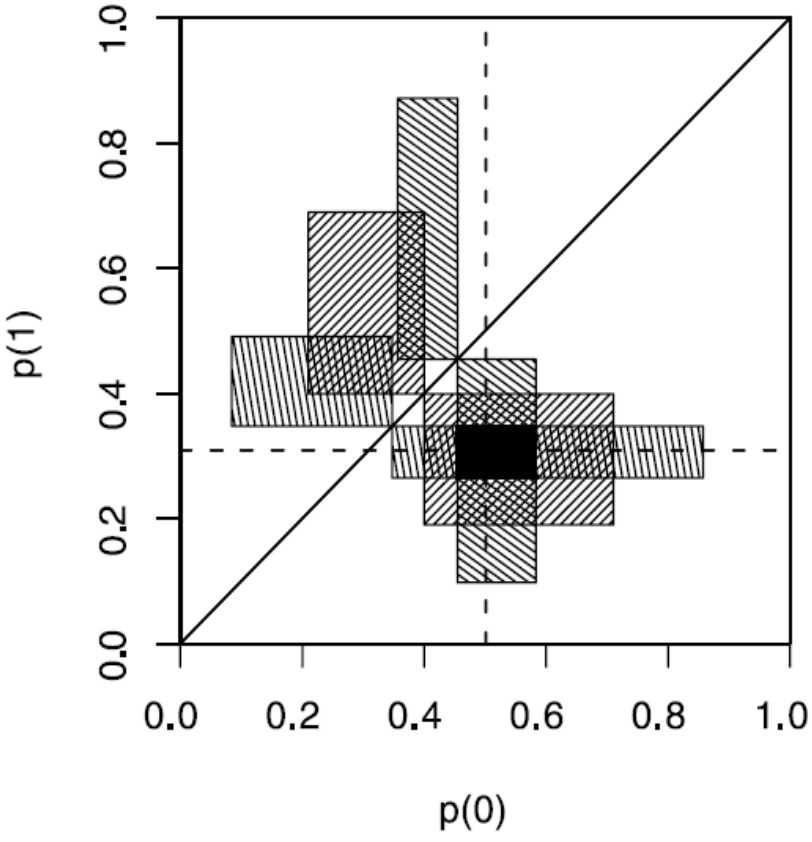
Penalty is given by

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

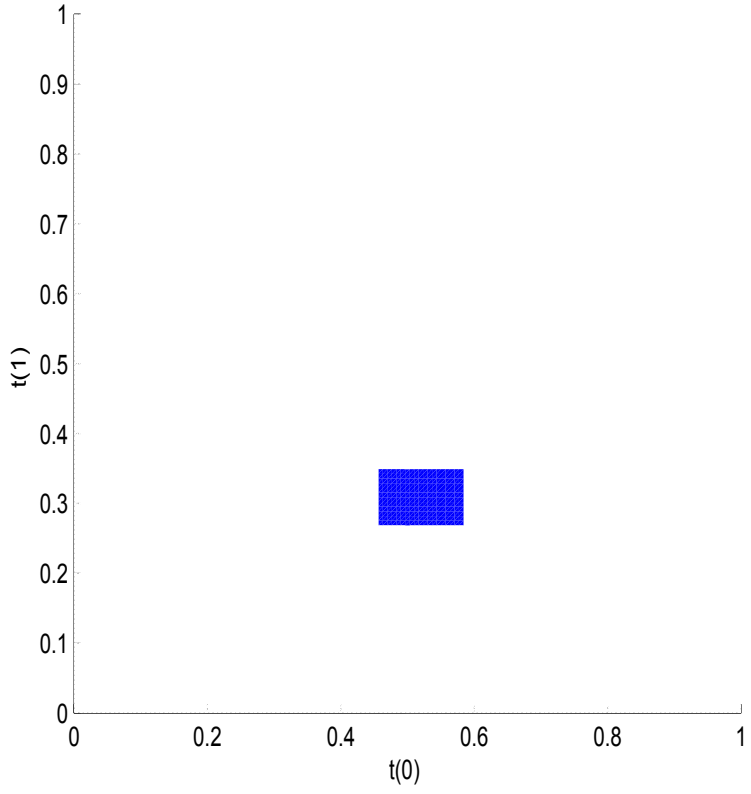
Problem is formulated as

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

Comparison of Results (Identified set)



(Chesher 2010)

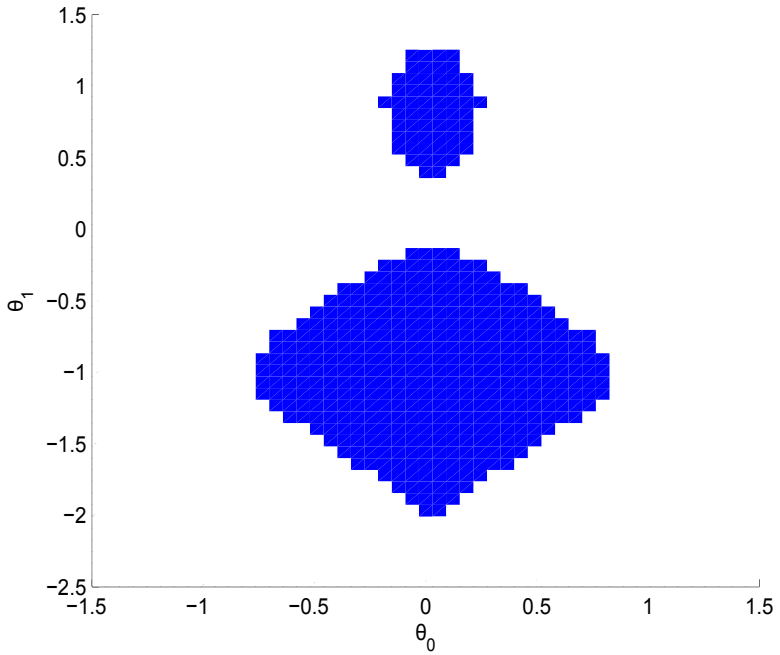


(my approach)

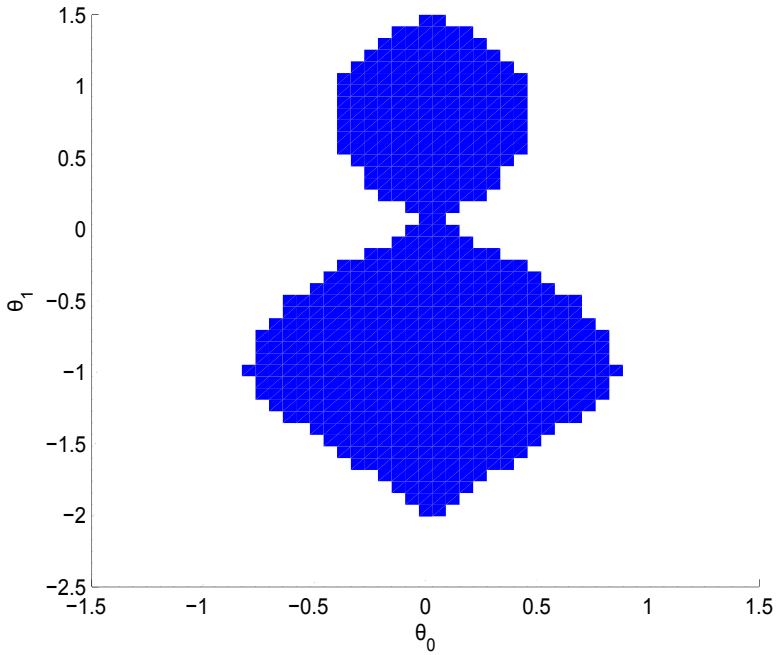
▸ Comparison - continuous X

How independency matters

Example with X continuous which is discretized

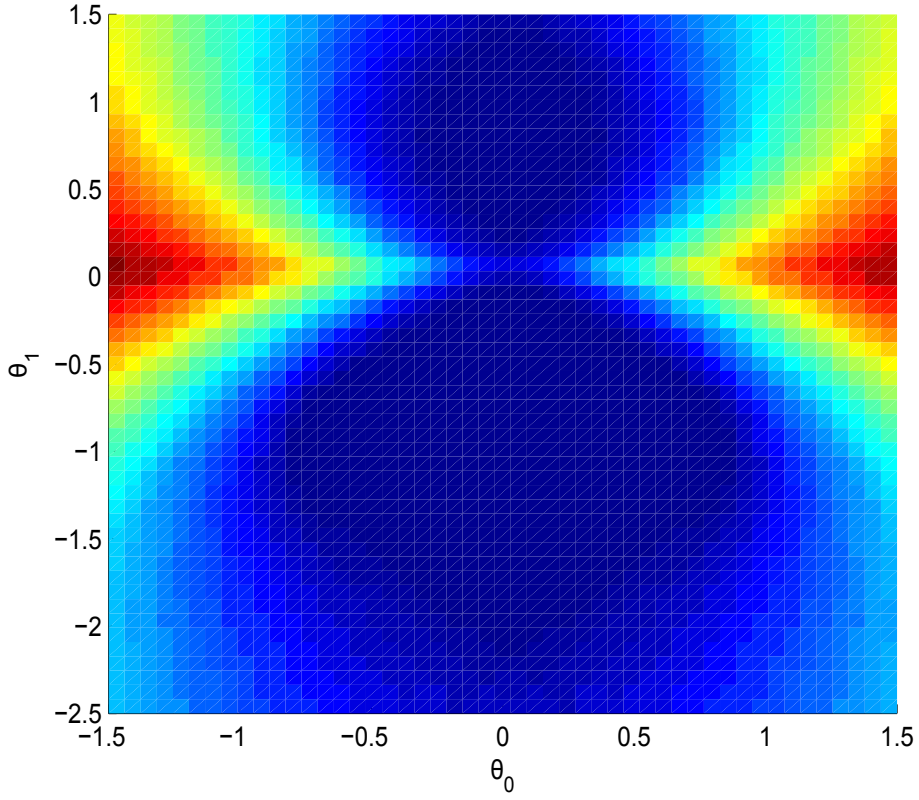


U and Z independent



U and Z not independent

Contour Plot



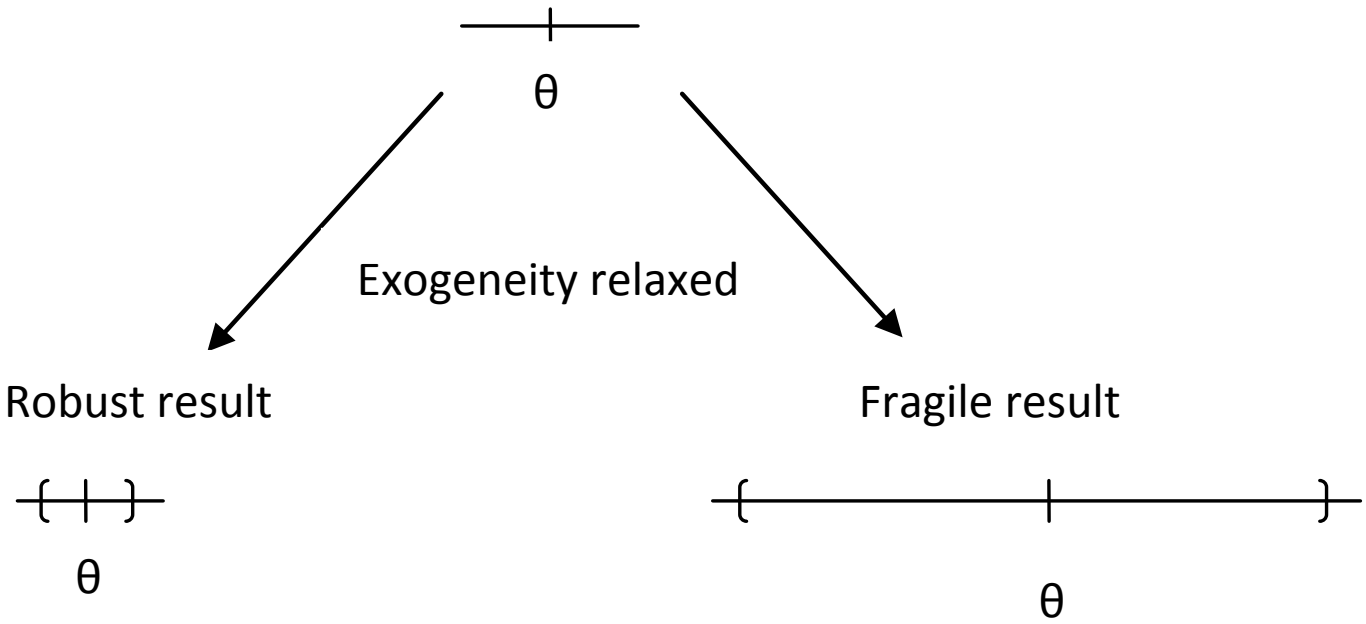
The value of minimized objective function stands for the minimal probability of the event incompatible with economic restrictions

▸ 3D figures of the optimized function

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

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

Exogeneity assumption relaxed (3)

Recall exogenous case

$$\begin{aligned} &\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} && (3) \\ &\text{s.t.} \\ &\textstyle\sum_{l} \pi_{ijkl} = p_{ijk}, && \forall i,j,k \\ &\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_l, && \forall l \\ &\textstyle\sum_{i,j} \pi_{ijkl} = \sum_{i,j} p_{ijk}\nu_l, && \forall k,l \\ &\pi_{ijkl} \geq 0, && \forall i,j,k,l. \end{aligned}$$

$$Pr(Z = z_k, U = u_l) = Pr(Z = z_k)Pr(U = u_l) \quad \forall k, l$$

Exogeneity assumption relaxed (4)

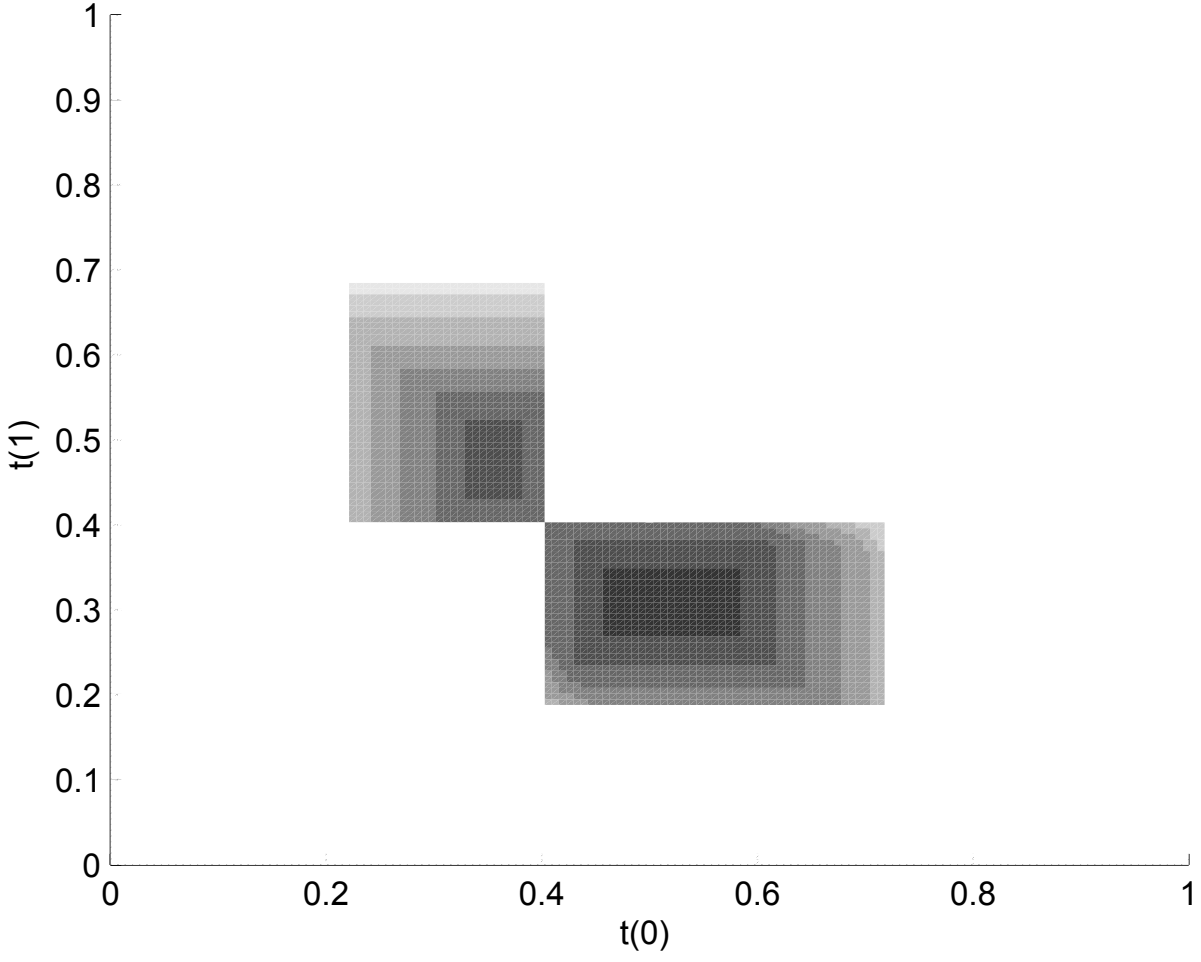
Now the Z and U are only "close" to being independent.

$$
\begin{aligned}
&\min_{(\pi)} \textstyle\sum_{i,j,k,l} \pi_{ijkl} c_{ijkl} && (4)\\
&\text{s.t.} \\
&\textstyle\sum_{l} \pi_{ijkl} = p_{ijk}, && \forall i,j,k \\
&\textstyle\sum_{i,j,k} \pi_{ijkl} = \nu_{l}, && \forall l \\
&\textstyle\sum_{i,j} \pi_{ijkl} - \sum_{i,j} p_{ijk}\nu_{l} \leq \delta, && \forall k,l \\
&-\textstyle\sum_{i,j} \pi_{ijkl} + \sum_{i,j} p_{ijk}\nu_{l} \leq \delta, && \forall k,l \\
&\pi_{ijkl} \geq 0, && \forall i,j,k,l.
\end{aligned}
$$

$$|Pr(Z = z_k, U = u_l) - Pr(Z = z_k)Pr(U = u_l)| \leq \delta \quad \forall k,l$$

Still a linear program - computationally feasible.

Exogeneity assumption relaxed (5)



$$\delta = [0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.075, 1]$$

Conclusion

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

Open Questions

- How to do inference (Dilation Bootstrap, Intersection bounds, Random set theory)
- Continuous variables (or what is the effect of discretization?)
- Computational aspects (size of π max 35000(?))
- Make it more "user friendly"

Thank you for your attention!

Example 2 - Multiple Equilibria

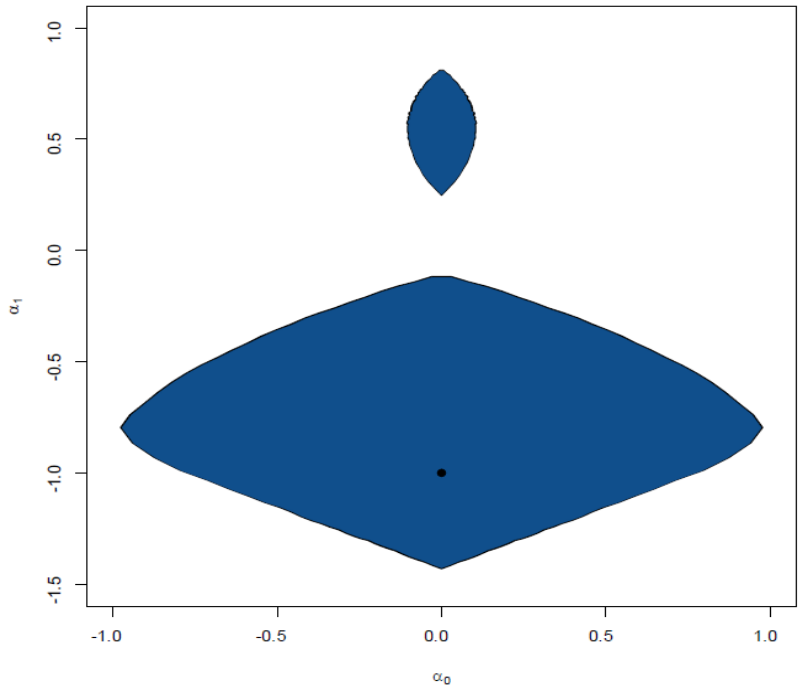
Jovanovic (1989)
Two players' game.
$\Pi_1(Y_1, Y_2, U_1, U_2) = (\theta Y_2 - U_2)1_{(Y_1=1)}$
$\Pi_2(Y_1, Y_2, U_1, U_2) = (\theta Y_1 - U_1)1_{(Y_2=1)}$
$U = (U_1, U_2)$ are exogenous costs, observed to firms, unobserved to econometrician and assumed to be $Unif(0,1)^2$
Two pure strategy Nash equilibria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

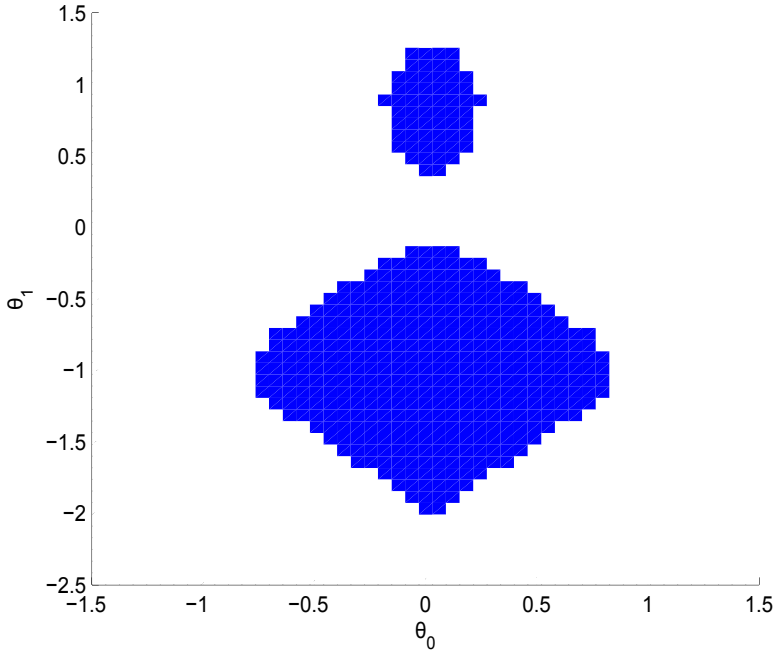
$\theta \in [\sqrt{p}, 1]$

Back

Comparison of Results (with continuous X)

X and U are discretized here

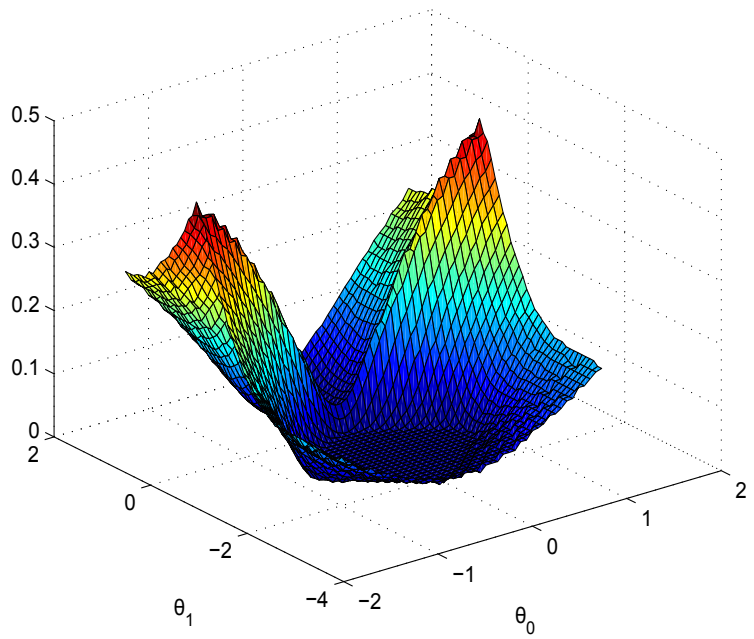


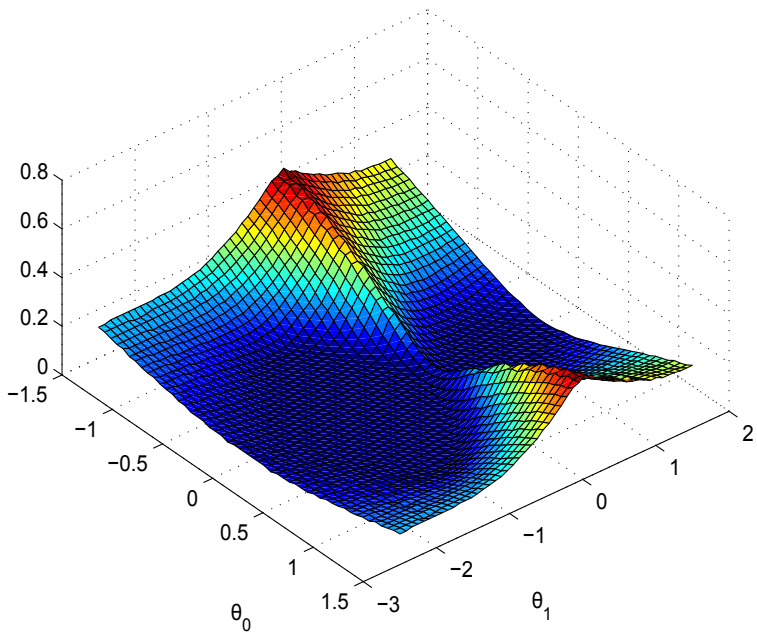


Discretization matters!

▸ Back

3D figures





Back