Moment Inequalities for Empirical Research

Luca Maini*

CONTENTS

*Please contact lmaini@email.unc.edu if you notice any mistakes.

1 AN INTRODUCTION TO PARTIALLY IDENTIFIED MODELS

1.1 Motivating example from theory: multiple equilibria

If you have taken an introductory IO class at the undergraduate level you have already encountered a model with multiple equilibria. In the version I teach, it's called McDonalds and Burger King. It's a simple two-period game where two firms, M and B, have to decide whether to enter in a market or not. The game matrix below describes the game actions and payoffs:

		B	
		Stay out	Enter
M	Stay out	$(0, 0)$	$(0, \alpha_B)$
	Enter	$(\alpha_M, 0)$	$(\alpha_M - \delta, \alpha_B - \delta)$

Depending on the values of α_M, α_B, and δ, any of the four combination of actions can be an equilibrium of the game (and it should be straightforward to figure out the values that lead to each outcome). Furthermore, for a certain set of values, it's possible that both $(\text{Enter}, \text{Stay out})$ and $(\text{Stay out}, \text{Enter})$ could be equilibrium outcomes.[1] This is a standard example in IO: when a market only supports one firm, but there are multiple potential entrants, we can't necessarily predict ex-ante which firm will enter.

Multiple equilibria generate problems in estimation. Suppose that we had data on entry in some markets, and wanted to analyze it to describe firm behavior. We might choose the following simple parametrization for α_M, α_B and δ:

		B	
		Stay out	Enter
M	Stay out	$(0, 0)$	$(0, \beta X_B + \varepsilon_B)$
	Enter	$(\beta X_M + \varepsilon_M, 0)$	$(\beta X_M - \delta + \varepsilon_M, \beta X_B - \delta + \varepsilon_B)$

Here, β and δ are parameters of interest, X_i is an array of observable variables, and ε_i are error terms that are observed by the firm, but not not by the econometrician. We call these *structural* errors; a concept we will go back to repeatedly in these notes. Notice that this is a game of perfect information. I have purposely left out any additional error unobserved by both the firm and the econometrician (that type of error I will call *expectational* error, and those will also reappear many times).

To derive restrictions on the parameters of the model, we'll start by assuming that the distribution of ε_i is known: $\varepsilon_i \sim \Phi_i(\cdot)$. [2] We use the shorthand 0 for Stay Out, and 1 for Enter. Then, we

[1] We are restricting our analysis to equilibria in pure strategies here. One can easily extend to mixed strategy equilibria, but the equilibrium conditions becomes more cumbersome to derive.

[2] One could possibly relax this assumption by allowing the distribution to be known up to some parameter θ (e.g. Φ is normal, with mean θ_1 and variance θ_2. Then θ_1 and θ_2 must be estimated as well. The theoretical derivation is straightforward, but the empirical implications need to be weighted carefully.

can write down the probability that each of the four possible equilibria will be played in the data as a function of our parameters. For the equilibria $(0,0)$ and $(1,1)$, this is straightforward:

$$\begin{aligned}\mathcal{P}(0,0) &= \Phi_M(-\beta X_M) \times \Phi_B(-\beta X_B) \\ \mathcal{P}(1,1) &= (1-\Phi_M(\delta-\beta X_M)) \times (1-\Phi_B(\delta-\beta X_B))\end{aligned}$$

(remember the interpretation of $\Phi_i(x)$ as the probability that ε_i is less than x). The probability of observing a separating equilibrium — either $(1,0)$ or $(0,1)$ — are not as obvious. In fact, you should realize pretty quickly that we have no way of figuring out what those probabilities are unless we have a selection rule that tells us who should enter when the market can support only one firm and both firms would like to enter. At this point, we have three options:

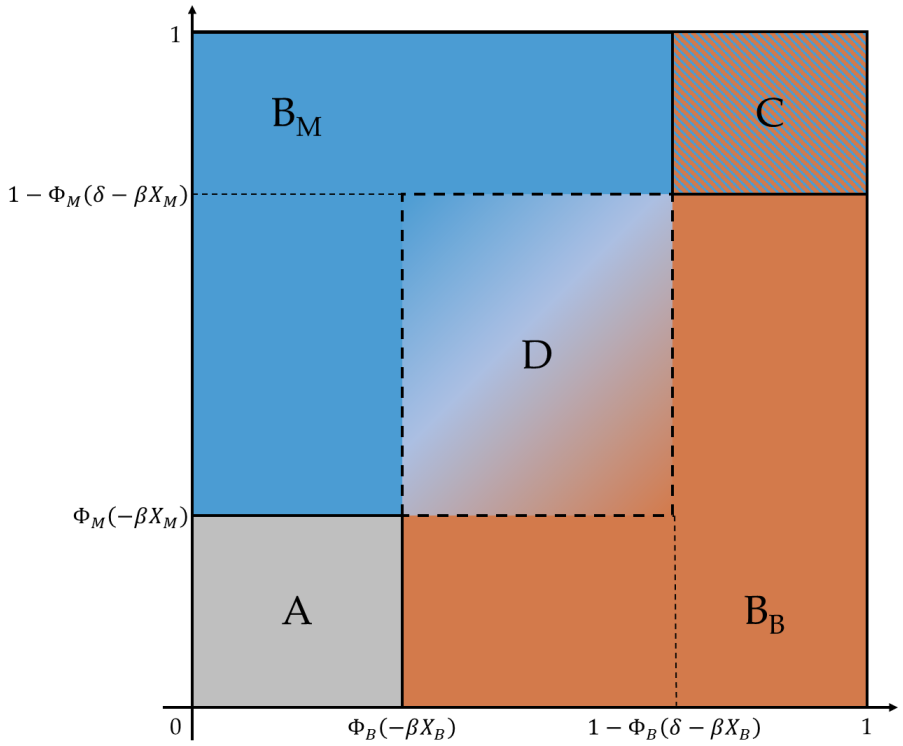
1. We apply an equilibrium selection rule. A simple extension, like sequential entry, will guarantee a unique equilibrium. The downside here is that most rules we can apply are going to be pretty ad-hoc, and likely will not match whatever is going on in the data (possible exceptions may include situations where the problem has a specific architecture, for example, a government procurement rule with clear tie-breaking rules). An example of this approach is Mazzeo (2002), who looks an entry of motels along US interstate highways.

2. We match the probability of observing 1 firm in the market. The downside here is that we might be interested in the identity of the entrant for counterfactual purposes, and furthermore, this reduces the number of moment restrictions we can use, which we may or may not care about depending on the number of parameters we need to identify. Berry (1992) uses this simplification in a model looking at airline entry.

3. We impose inequality conditions. The figure below should give you a pretty clear intuition for how we can do that. Basically, the probability of observing $(1,0)$ will be at least equal to the area B_M, and no more than $B_M + D$, while the probability of observing $(0,1)$ will be at least equal to the area of B_B, and no more than $B_B + M$. You can check that the four inequalities are

$$\begin{aligned}&\mathcal{P}(1,0) \geqslant (1-\Phi_M(-\beta X_M)) \cdot \Phi_B(-\beta X_B) + \Phi_M(\delta-\beta X_M) \cdot (1-\Phi_B(\delta-\beta X_B) - \Phi_B(-\beta X_B)) \\ &\mathcal{P}(1,0) \leqslant (1-\Phi_M(-\beta X_M)) \cdot (1-\Phi_B(\delta-\beta X_B)) \\ &\mathcal{P}(0,1) \geqslant (1-\Phi_B(-\beta X_B)) \cdot \Phi_M(-\beta X_M) + \Phi_B(\delta-\beta X_B) \cdot (1-\Phi_M(\delta-\beta X_M) - \Phi_M(-\beta X_M)) \\ &\mathcal{P}(0,1) \leqslant (1-\Phi_B(-\beta X_B)) \cdot (1-\Phi_M(\delta-\beta X_M))\end{aligned}$$

   The paper that most people reference when discussing this third approach is Ciliberto and Tamer (2009), which also looks at entry in airline markets. Additional papers of interest include Tamer (2003) and Andrews and Barwick (2012).



1.2 Motivating example from empirical work: large choice sets and complex models

This example is originally from an unpublished thesis (Katz, 2007). The discussion in this section is largely based on a treatment of the same problem in Pakes (2010).

Consider the following individual choice problem: a shopper chooses which supermarket to go to, and what to buy once there. Driving to a supermarket further away is costly, but may be worth it if prices are lower, or if the supermarket offers a larger basket of products. In an empirical exercise, we might have scanner data showing purchases, and consumer locations. We want to use this data to estimate the perceived cost of travel for the shopper.

To start tackling this problem, we make some assumptions. First, let expenditure and travel costs be separable in the utility function, so that we can write

$$\pi\left(b_i, s_i; z_i, \theta\right) = U\left(b_i, z_i\right) - e\left(b_i, s_i\right) - \theta_i dt\left(s_i, z_i\right) \tag{1}$$

where

- s_i is the supermarket chosen by the shopper, and b_i is the bundle of goods purchased at that supermarket.
- z_i is a vector of consumer characteristics
- $U(\cdot)$ and $e(\cdot)$ are the utility from the bundle and expenditure functions respectively

- $dt(\cdot)$ is the travel time to supermarket s_i for a consumer with characteristics z_i , and θ_i is that consumer's disutility of travel.

Notice that there should be a coefficient on $e(\cdot)$, but that coefficient has been normalized to 1. That allows us to interpret θ_i as a dollar cost of travel.

If you stop and think about what this problem involves, it'll quickly become clear that there are a myriad of variables and parameters at play. First, there may be a large number of supermarkets within driving distance. Second, each supermarket will have a number of different products (not necessarily the same), and each product will have a price, which will be different across supermarkets. Third, each possible bundle of products is a potential option for the shopper, meaning that the dimensionality of the state space is virtually unbounded.

So how do we tackle this? Absent any concerns about the dimension of the choice set, the most natural thing to do would be to add a logit error to equation (1) and estimate this model as a random utility discrete choice model. Since we have individual level data, we could run something like micro-BLP. One could go down this route (and Katz did, for illustrative purposes) by drastically cutting down on the choice set, and reducing the number of parameters to be estimated. First, limit the set of bundles that can be purchased; second, limit the number of supermarkets in the consideration set to a set included within a specific radius of a shopper's location; third, set $\theta_i = \theta z_i$ (i.e. driving aversion is a function of observables only). A feasible specification might include several expenditure bundles (the goods in each bundle are selected by looking at "typical" purchase patterns conditional on a given level of expenditure), with each bundle having a mean utility parameter, which can vary by supermarket. Even so, there are a number of reasons why the estimates of this model might be biased:

1. Even if we get the expenditure level right, the goods in the expenditure bundle might not be the goods that a specific shopper is going to buy. Shoppers will probably shop at supermarkets where the goods they are interested in are cheaper. Hence, for a given expenditure level, they can get a better-than-expected bundle at the supermarket of their choice, relative to other supermarkets. This means the error term is correlated with the choice of supermarket.

2. The model does not allow for expectational errors (consumers have to know all the prices at all the stores).

3. There is no unobserved heterogeneity in the aversion to drive time. We can expect drive time to be correlated to aversion to driving, so once again the error term is correlated to a choice variable. This is possibly the most problematic assumption because that's the parameter we are most interested in.

So, to recap, a full solution to this model requires (i) unreasonable assumptions on the functional forms of model primitives, (ii) unrealistic information sets for agents, and (iii) arbitrary reductions

to the dimensionality of the choice set. Which hopefully brings you to ask: is there anything else we can do? Fortunately, there is.

Let's go back to the initial model from equation (1) and assume that the agent makes his choice of store by maximizing expected utility. This assumption is ubiquitous in economics, so it's possibly the least demanding assumption you can make. We denote expected utility by the operator $\mathcal{E}$. (We use this operator instead of the usual one $\mathbb{E}[\cdot]$ because we are not requiring this to conform to rational expectations. I will be more clear about what assumptions are needed for $\mathcal{E}(\cdot)$ in a second.) Also denote $\mathcal{J}_i$ as the information set of the agent at time i (i.e. what the agent knows). It bears repeating that the approach above requires all the prices to be in in $\mathcal{J}_i$.

Under the assumption we made, and using a revealed-preference argument, we know that, letting s_i^o be the supermarket chosen by shopper i in the data, for any s_i', we have

$$\mathcal{E}\left[\pi\left(\mathbf{b}, s_i^o; z_i, \theta_i\right) \middle| \mathcal{J}_i\right] \geqslant \mathcal{E}\left[\pi\left(\mathbf{b}, s_i'; z_i, \theta_i\right) \middle| \mathcal{J}_i\right]$$

where the bundle $\mathbf{b}$ is written in boldface to indicate that it is a random variable at the time of the shopper's choice (because prices are unknown, and the exact bundle depends on them). In the data we don't observe the counterfactual bundle b_i' that shopper i would have bought had he chosen supermarket s_i'. However, regardless of what b_i'would have been, we know that

$$\pi\left(b_i', s_i'; z_i, \theta_i\right) \geqslant \pi\left(b_i, s_i'; z_i, \theta_i\right)$$

because that's basically how b_i' is defined—it's the bundle that maximizes the utility of the agent.[3] Therefore, if we compare the utility of choosing (s_i, b_i) to the utility of choosing $\left(s_i', b_i\right)$, we should get that, on average, the former is greater than the latter. In fact, we can be even more precise. Denote, for any function $f(\cdot)$,

$$\Delta f\left(s_i, s_i', \cdot\right) = f\left(s_i, s_i', \cdot\right) - f\left(s_i, s_i', \cdot\right)$$

Then, the shopper's expectation of $\Delta\pi\left(s_i, s_i', \cdot\right)$ is

$$\mathcal{E}\left[\Delta\pi\left(\mathbf{b}, s_i, s_i'; z_i, \theta_i\right) \middle| \mathcal{J}_i\right] = \mathcal{E}\left[-\Delta e\left(\mathbf{b}, s_i, s_i'\right) - \theta_i \Delta dt\left(s_i, s_i', z_i\right) \middle| \mathcal{J}_i\right] \geqslant 0$$

where we have eliminate the term $U\left(b_i, z_i\right)$ because it cancels out in the difference (as it does not depend on s_i).

Denote the difference between the ex-ante expectation and the realization as

$$\nu_{1,i,s,s'} \equiv \Delta\pi\left(b, s_i, s_i'; z_i, \theta_i\right) - \mathcal{E}\left[\Delta\pi\left(\mathbf{b}, s_i, s_i'; z_i, \theta_i\right) \middle| \mathcal{J}_i\right]$$

[3] Technically, b_i' maximizes $U\left(b_i, z_i\right) - e\left(b_i, s_i'\right)$, but we have fixed the supermarket choice, so distance does not matter in this case.

Then, we apply the only restriction we need to $\mathcal{E}[\cdot]$, which is that

$$\frac{1}{N}\sum_i \nu_{1,i,s,s'} \to^p 0 \tag{2}$$

where $\to^p$ denotes convergence in probability. Finally, assume that the shopper knows driving time (i.e. $\Delta dt(\cdot) \in \mathcal{J}_i$. Then, it's straightforward to show that

$$\frac{1}{N}\sum_i \frac{\nu_{1,i,s,s'}}{\Delta dt\left(s_i, s_i', z_i\right)} \to^p 0 \tag{3}$$

At this point we have enough to build our moment inequalities. Let's parametrize $\theta_i = \theta_0 + \nu_{2,i}$, where θ_0 is the average aversion to driving, and $\nu_{2,i}$ is individual-specific aversion that the shopper knows, but the econometrician doesn't. Notice that by construction, $\sum_i \nu_{2,i} \equiv 0$. Finally, to ease notation, let

$$\mathcal{E}\left[\Delta\pi\left(\mathbf{b}, s_i, s_i'; z_i, \theta_i\right) \middle| \mathcal{J}_i\right] = K_{s,s'} \geqslant 0$$

Then, from equation (3):

$$\begin{aligned}
&\frac{1}{N}\sum_i \frac{\Delta\pi\left(b, s_i, s_i'; z_i, \theta_i\right) - K_{s,s'}}{\Delta dt\left(s_i, s_i', z_i\right)} \to^p 0 \\
\implies &\frac{1}{N}\sum_i -\frac{\Delta e\left(b_i, s_i, s_i'\right)}{\Delta dt\left(s_i, s_i', z_i\right)} - \theta_0 - \frac{1}{N}\sum_i \nu_{2,i} \to^p \frac{1}{N}\sum_i \frac{K_{s,s'}}{\Delta dt\left(s_i, s_i', z_i\right)} \\
\implies &-\frac{1}{N}\sum_i \frac{\Delta e\left(b_i, s_i, s_i'\right)}{\Delta dt\left(s_i, s_i', z_i\right)} \to^p \theta_0 + \frac{1}{N}\sum_i \frac{K_{s,s'}}{\Delta dt\left(s_i, s_i', z_i\right)}
\end{aligned}$$

The last line gives us our inequality condition. We can pick supermarkets s' that are *further away* than s_i from shopper i (and where purchasing b_i is feasible). Then, the residual term of the inequality will be negative:

$$\frac{1}{N}\sum_i \frac{K_{s,s'}}{\Delta dt\left(s_i, s_i', z_i\right)} \leqslant 0$$

so those inequalities will yield lower bounds on the parameter θ_0. Conversely, we can pick supermarkets s'' that are closer, and that will yield an upper bound on the parameter θ_0.[4]

To recap, moment inequalities in this setting allow us to (i) drop the functional form assumption on $U(\cdot)$, (ii) flexibly incorporate expectations in the model, and (iii) put no limits on the choice set. The cost is that the model is no longer necessarily point-identified. However, in Katz's estimation, θ_0 in the point-identified model is around \$240 per hour, while in the model using moment inequalities it is between \$16-\$17 (average hourly wage in the area was \$18). So at least in this case it seems the approach is well justified.

[4]One can drop the assumption that $\Delta dt(\cdot) \in \mathcal{J}_i$ and still build moment inequalities to recover θ. However, if we do that, we can no longer allow for the unobserved heterogeneity component $\nu_{2,i}$. See Pakes (2010) for the alternative moment condition.

1.3 Closing remarks

In this section we have seen two versions of partially identified models. In the first version, the model is partially identified because the most natural notion of equilibrium we can use justifies any of a set of possible outcomes. In the second version, the model is partially identified because there are primitives of the models that we don't know, and that we can't identify based only on observable data. These differences are important. In the first case, not only is there no amount of data that could ever tell us which of $(1,0)$ or $(0,1)$ will be played, but in fact, even knowing every single primitive of the model without error would not necessarily help. In the second case, given basic regularity assumptions, the model has a unique equilibrium — trivially, since it's a single-agent model. Therefore, if we knew the primitive $U(\cdot)$ and had perfect information on supermarket offerings and prices (and a powerful machine to boot) we could fully solve the model. Of course no amount of data would probably be enough to correctly estimate $U(\cdot)$, and so we use inequalities.[5]

You should realize that in both cases one can always make additional assumptions to obtain point identification and remove the need for inequalities. However, if there is one important takeaway from IO is that researchers should view structural assumptions as costly sacrifices, not convenient shortcuts. One of the most important contributions of the literature on moment inequalities is that they allow identification with fewer assumptions.

[5]There is a third situation that we haven't considered in which moment inequalities can help, and that is the case of models that would be point-identified given enough time/computing power or the right amount of data, but for which the econometrician is operating under one or more constraints that effectively prevent him/her from deriving the full solution. These types of problems are basically extensions of the second case we considered.

2 REVEALED PREFERENCE INEQUALITIES: THE WORKHORSE MODEL FOR EMPIRICAL APPLICATIONS

The most commonly used moment inequality framework in empirical work is the second of the two examples that we examined in the previous section. This framework is based on the assumption that the choices that agents make reflect a maximization problem, and therefore one can use revealed-preference arguments to build inequality conditions that compare the payoff of the observed action to the payoff of a hypothetical alternative action. In this section, we formally introduce the basic model for revealed preference inequalities.

2.1 A starting point: the second stage estimator in Bajari, Benkard and Levin (2007)

Before laying out the general framework, I am going to briefly discuss the inequalities in Bajari et al. (2007, henceforth BBL). In doing so, I hope to achieve two goals. First, I want to draw a connection with topics you have already covered. Second, I want to use the model in BBL as a jumping off point towards more general and less restrictive forms of moment inequalities that rely on less data and fewer assumptions. To make the connections clearer I am going to make some changes to the notation used in the original paper, so that the variables in the BBL model can map to identical variables in the framework I will present later, which is from Pakes et al. (2015).

Model

The BBL framework is a dynamic model of entry and exit with investment. There are n firms. Firm i maximizes the expected discounted value of the stream of profits:

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\beta^{\tau-t}\pi_i\left(d_{i\tau},\mathbf{d}_{-i\tau},\mathbf{y}_{i\tau};\theta\right)\middle|\mathcal{J}_{it}\right] \tag{4}$$

where:

- $d_{i\tau}$ is the action (decision) of firm i at time τ, and $\mathbf{d}_{-i\tau}$ are the actions of other firms. Boldface is used to denote the fact that this is a random variable from the point of view of firm i (in the original BBL notation, this variable is $\mathbf{a}_\tau$ — though in BBL notation boldface simply denotes vectors).

- $\mathbf{y}_{it}$ denotes any variables that may affect the payoff of firm i, other than the decision variables (in BBL notation this variable represents the state variable $\mathbf{s}_\tau$, as well as the shocks $\nu_{i\tau}$).

- $\pi_i\left(\cdot;\theta\right)$ is the period payoff function, with θ being the parameter we are interested in estimating.

- $\mathcal{J}_{it}$ is the information set of firm i at time t. This is what the firm knows at the time of making the decision (in BBL, this is simply the state variable $\mathbf{s}_t$).

The structural parameters of the model are

- the payoff function $\pi_i(\cdot;\theta)$
- the joint probability distribution of $(\mathcal{J}_{i\tau}, \mathbf{y}_{i\tau})$ as a function of agent's actions — in BBL these are the transition probability matrix $P(s_{t+1}|a_t, s_t)$, and the distribution of the private shocks $\nu_{i\tau}$.

(β is also a parameter, but in virtually all application the value of β is assumed).

BBL focuses on MPEs as equilibrium concept.[6] Denote a Markov strategy profile as a mapping $s : \mathcal{I} \to D$ from information sets to a vector of actions for all firms — $\mathcal{I}$ is the space of information sets, and D is the set of actions agents can take. The definition of MPE implies that if $s = (s_1, s_2, \ldots, s_3)$ is an MPE, then each firm prefers that strategy to any other alternatives, given the strategies of other firms. In other words, if we denote the expression in equation (4) as a value function $V_i(s_i, s_{-i}, \mathcal{J}_{it}; \theta)$, then

$$V_i(s_i, s_{-i}; \theta) \geqslant V_i\left(s_i', s_{-i}, ; \theta\right) \tag{5}$$

holds for all s_i'.

Estimation

The estimation of BBL is done in two stages. The output of the first stage is an estimate of the joint probability distribution of $(\mathcal{J}_{i\tau}, \mathbf{y}_{i\tau})$, which I denote $G(\cdot)$ — in practice, this is done by combining the transition probability matrix, which determines the evolution of the state variable as a function of its past value and the actions of agents, and the policy functions, which determine what agents are going to do given an information set.[7] Given that, BBL provides an efficient simulation method to compute the value function in terms of the structural parameters.

The estimator in the second stage exploits the same intuition previewed in section a revealed preference argument. From equation (5), we can define a set of parameters

$$\Theta_0(s_i, G) \equiv \left\{\theta : \theta, s_i \text{ satisfy } (5) \text{ for all } i, s_i'\right\}$$

Whether $\Theta_0(s_i, G)$ is a singleton depends on rather technical assumptions about the set to which θ is restricted, and the properties of the value function. We are not interested in those here. What we care about is the principle behind the conditions, which is the same revealed preference principle we applied in the supermarket case.

[6] You should be familiar with the definition of properties of MPEs. For a refresher, see the series of papers by Maskin and Tirole: Maskin and Tirole (1987, 1988a,b, 2001).

[7] If you need to refresh your memory on the first stage estimation you can go back to the paper, or to the excellent summary in Ackerberg et al. (2007).

Discussion

BBL describes a very flexible class of dynamic models, and the estimation framework has been sucessfully used to analyze a variety of markets (see e.g. Ryan 2012; Collard-wexler 2013). It can accommodate both discrete and continuous decision variables, multiple firms, unobserved heterogeneity, and a complicated state space. So, does that mean that moment inequalities can be used broadly across those models? Unfortunately, here the answer is no, or at least not in practice. The main difficulty in applying the BBL framework to empirical work are the onerous requirements of the first stage, particularly the ability to recover consistent estimates of the policy functions. Why is this requirement onerous? Because in dynamic games it's not enough to simply observe the actions of players. We must also know how players would react to out-of-equilibrium play. Almost always, this will require knowledge (or assumptions) about firm expectations, as well as the ability to predict behavior in unobserved portions of the state space, and this will turn into a requirement to know the solution to the model (i.e. derive s_i not as a first-stage semiparametric estimate, but rather as a theory-based analytical function). Given that what we prize about moment inequality is the ability to simplify complicated problems, this is a major downside of this particular approach to estimation.

2.2 The reference framework: Pakes, Porter, Ho and Ishii (2015)

To make progress in our understanding of moment inequalities, we now consider what can be done when the econometrician cannot easily recover a first-stage estimate of the value function. This section, which is based on Pakes et al. (2015, henceforth PPHI), will cover a slightly more general model than BBL.[8] As a preview of what we will cover, the final estimator is going to be virtually identical to BBL, but we will spell out the specific conditions under which the inequality constraints generated by the decision problem of the agent can be used for estimation. These conditions are (rather implicitly) satisfied in BBL by having consistent first-stage estimates. However, as we will see, consistent first-stage estimates are sufficient, but not necessary.

Setup

The starting point is quite similar to BBL. There are n decision-making agents, indexed by i. Each agent is endowed with an information set $\mathcal{J}_i$, and plays strategies $s_i : \mathcal{I}_i \to D_i$. We write $s_i(\mathcal{J}_i) = \mathbf{d}_i$ so that, again, boldface indicates that $\mathbf{d}_i$ is a random variable whose realization depends on the realization of $\mathcal{J}_i$. Like in BBL, there are no restrictions on D_i (it can be discrete, or continuous, unbounded, uncountable, etc.). Payoffs are given by a function $\pi(\cdot)$, which depends on actions as well as a set of payoff-relevant random variables $\mathbf{y}_i \in \mathbf{Y}_i$. The structural parameters of the model are the payoff function $\pi_i(\cdot)$ and the joint probability distribution of $(\mathcal{J}_i, \mathbf{y}_i)$. Notice we don't use

[8] I have tried to follow the notation and steps in PPHI as closely as possible. On occasion however, I have changed the definition of certain variables, and the presentation of certain equations in order to make the steps and concepts easier to understand. Therefore, a reader following these notes alongside the original paper should keep in mind that in certain spots my notation will deviate slightly from that of PPHI.

a subscript t, but you can interpret s_i as a series of decision over time, so we don't have any loss of generality: for all intents and purposes this is the same setup as BBL.

We do make one deviation relative to BBL, and that is to substitute the rational expectation operator $\mathbb{E}[\cdot]$, with a generic operator $\mathcal{E}[\cdot]$, which in principle could represent any type of forecasting process on the part of the agent, though we are going to specify some restrictions right now.

Assumptions

The first assumption is the best response condition.

ASSUMPTION 1 — BEST RESPONSE CONDITION: *If s_i is the strategy played by agent i, then*

$$\sup_{d \in \mathcal{D}_i} \mathcal{E}\left[\pi\left(d, \mathbf{d}_{-i}, \mathbf{y}_i\right) \middle| \mathcal{J}_i, \mathbf{d}_i = d\right] \leqslant \mathcal{E}\left[\pi\left(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{y}_i\right) \middle| \mathcal{J}_i, \mathbf{d}_i = s_i\left(\mathcal{J}_i\right)\right]$$

for $i = 1, \ldots, n$.

This assumption comes from maximizing behavior in single-agent problems, and from the conditions for Bayes-Nash equilibria in multiple-agent games. Notice that this is broader than BBL, which is restricted to MPE (a specific type of BNE), and also that there are no constraints on any functional forms.

To complete Assumption 1, we need to also have a model for $\mathbf{d}_{-i}$ and $\mathbf{y}_i$. These two variables are likely to be endogenous — i.e. they will change in response to changes in d_i. For the next assumption, define $\mathbf{z}_i$ as a set of exogenous variables — that is, variables that do not change in response to d_i.

ASSUMPTION 2 — COUNTERFACTUAL CONDITION: $\mathbf{y}_i = y\left(\mathbf{z}_i, \mathbf{d}_i, \mathbf{d}_{-i}\right)$, *and* $\mathbf{d}_{-i} = d^{-i}\left(d, \mathbf{z}_i\right)$, *and the distribution of $\mathbf{z}_i$ conditional on $\mathcal{J}_i$ and $\mathbf{d}_i = d$ does not depend on d.*

In plain terms, Assumption 2 states that if either $\mathbf{y}_i$ or $\mathbf{d}_{-i}$ is endogenous (i.e. is a function of $\mathbf{d}_i$), then we need a model for how they respond to changes in $\mathbf{d}_i$. This assumption can be more or less restrictive depending on the problem you are considering. In single-agent problems we don't need to worry about $\mathbf{d}_{-i}$, and also $\mathbf{y}_i$ will generally consist of variables that a single firm cannot affect (e.g. market conditions). In static simultaneous move games, $\mathbf{d}_{-i}$ will matter, but we can usually assume that deviations are unexpected and so other firms will not deviate from equilibrium play. However, $\mathbf{y}_i$ might contain variables (such as price, or quantity) that are endogenous to $\mathbf{d}_i$, and therefore a model is needed. Finally, in a sequential or dynamic multiple-agent model we need to consider the response of agents that move later on, so a model for both $\mathbf{d}_{-i}$ and $\mathbf{y}_i$ is needed.

Assumption 1 and 2 together yield the following inequality for estimation:

$$\mathcal{E}\left[\Delta\pi\left(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i\right) \middle| \mathcal{J}_i\right] \geqslant 0 \tag{6}$$

where the function $\Delta\pi(\cdot)$ is defined as

$$\Delta\pi\left(d_i, d', d_{-i}, z_i\right) = \pi\left(d_i, d^{-i}\left(d_i, z_i\right), y\left(z_i, d_i, d_{-i}\right)\right) - \pi\left(d', d^{-i}\left(d', z_i\right), y\left(z_i, d', d_{-i}\right)\right)$$

In order to transform equation (6) into something we can use for estimation, we require the following:

- a parametric function $r(\cdot, \theta)$ that approximates $\pi(\cdot)$
- an assumption about the error between $\mathcal{E}[\cdot]$ and the sample moments that are generated in the data.

Notice what we *don't* need here: a specification for $\mathbf{d}_i$ conditional on observables. In other words, we don't need a policy function, or a solution to the model (unlike BBL).

Before specifying our third assumption, let's describe the form of the difference between the agent's expectation of $\pi(\cdot)$, and its sample analog. First, let $z_i^o \subseteq z_i$ be a subset of z_i observable by econometrician. Furthermore, let $r\left(d, d_{-i}, z_i^o; \theta\right)$ be the measurement function for $\pi(\cdot)$, and define $\nu(\cdot)$ to be the difference between the two. Then, the sample analog of equation (6) can be written as

$$\Delta r\left(d, d', d_{-i}, z_i^o; \theta\right) \equiv \Delta\pi\left(d, d', d_{-i}, z_i\right) + \Delta\nu\left(d, d', d_{-i}, z_i^o, z_i, \theta\right) \tag{7}$$

The econometrician observes $\Delta r(\cdot; \theta)$, but the agent makes a decision based on $\mathcal{E}\left[\Delta\pi(\cdot)\middle|\mathcal{J}_i\right]$. The relationship between the two, based on the construction in equation (7), is

$$\Delta r\left(d, d', d_{-i}, z_i^o; \theta\right) \equiv \mathcal{E}\left[\Delta\pi\left(d, d', \mathbf{d}_{-i}, \mathbf{z}_i\right)\middle|\mathcal{J}_i\right] + \nu_{1,i,d,d'} + \nu_{2,i,d,d'} \tag{8}$$

where

$$\nu_{1,i,d,d'} \equiv \underbrace{\left(\Delta\pi\left(d, d', d_{-i}, z_i\right) - \mathcal{E}\left[\Delta\pi\left(d, d', \mathbf{d}_{-i}, \mathbf{z}_i\right)\middle|\mathcal{J}_i\right]\right)}_{\text{expectational error in } \Delta\pi(\cdot)} + \underbrace{\left(\Delta\nu\left(d, d', d_{-i}, z_i^o, z_i, \theta\right) - \mathcal{E}\left[\Delta\nu\left(d, d', \mathbf{d}_{-i}, z_i^o, \mathbf{z}_i, \theta\right)\middle|\mathcal{J}_i\right]\right)}_{\text{expectational error in } \Delta\nu(\cdot)}$$

which we can further simplify to

$$\nu_{1,i,d,d'} \equiv \Delta r\left(d, d', d_{-i}, z_i; \theta\right) - \mathcal{E}\left[\Delta r\left(d, d', \mathbf{d}_{-i}, z_i^o, \theta\right)\middle|\mathcal{J}_i\right]$$

and

$$\nu_{2,i,d,d'} \equiv \mathcal{E}\left[\Delta\nu\left(d, d', \mathbf{d}_{-i}, z_i^o, \mathbf{z}_i, \theta\right)\middle|\mathcal{J}_i\right]$$

(or

$$\nu_{2,i,d,d'} \equiv \mathcal{E}\left[\Delta r\left(d, d', \mathbf{d}_{-i}, z_i^o, \theta\right)\middle|\mathcal{J}_i\right] - \mathcal{E}\left[\Delta\pi\left(d, d', \mathbf{d}_{-i}, \mathbf{z}_i\right)\middle|\mathcal{J}_i\right]$$

in a notation closer to PPHI).

(You can double-check that if you plug the νs back into (8), you get equation (7) back). Let's take a look at the two error terms.

ν_1 combines two terms. The first is the difference between the agent's expectation of payoffs, and the actual realization of the payoffs. This is an *expectational* error, and one that the agent cannot forecast. In the supermarket example, this could be the realization of prices at the supermarket. In BBL, the entry and investment cost shocks of other firms (which affect the realization of $\mathbf{d}_{-i}$). The second term can result from measurement error in variables, or from model specification error that is independent of $\mathcal{J}_i$ (for example, if $r(\cdot)$ is a reduced-form regression of profits onto variables of interest). Both terms are expectational error terms. As such, they do not generate selection issues. In sample analogs, these errors will average out to 0 in expectation.

ν_2 is the more problematic term. It represents the difference between the true profits and the measure $r(\cdot)$ that the agent — but not the econometrician — can observe in advance. Formally, $\nu_{2,i,d} \in \mathcal{J}_i$, and therefore d_i is a function of $\nu_{2,i,d}$ as well. This is the classic selection issue in IO: in entry models, we observe entry only in those markets where the fixed cost of entry is low enough to make entry profitable; in investment models, we see larger investment from firms whose cost is lower; in productivity models we only see data from firms whose productivity is high enough to keep them in the market.

The problem with $\nu_{2,i}$ is that if we simply take an average of $\Delta r(d, d', d_{-i}, z_i; \theta)$ it may not be greater than zero, due to the selection problem. Our third and final assumption provides conditions on the distribution of $\nu_{1,i}$ and $\nu_{2,i}$ that are sufficient to generate non-negative weighted averages of $\Delta r(d, d', d_{-i}, z_i; \theta)$.

ASSUMPTION 3: *Let $h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}) : \mathcal{D}_i \to \mathbb{R}^+$ be a nonnegative function whose value can depend on the alternative choice considered d', on the information set $\mathcal{J}_i$, and on observable components of the other agents' information sets $\mathbf{x}_{-i} \subseteq \times_{j \neq i} \mathcal{J}_j$. Assume that*

$$\text{(a)} \quad \mathcal{E}\left[\sum_{i=1}^{n} \sum_{d' \in \mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right) \cdot \nu_{1,i,\mathbf{d}_i,d'}\right] \geqslant 0$$

and

$$\text{(b)} \quad \mathcal{E}\left[\sum_{i=1}^{n} \sum_{d' \in \mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right) \cdot \nu_{2,i,\mathbf{d}_i,d'}\right] \geqslant 0$$

Let's put aside for a moment what this assumption means. Simply using the two equations in the assumption we can show that using $h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i})$ as weights will yield a sample average

for $\Delta r\left(d, d', d_{-i}, z_i; \theta\right)$ that is nonnegative in expectation. Using the definition from (8):

$$\begin{aligned}
\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right)\Delta r\left(d, d', d_{-i}, z_i; \theta\right)\right] = \\
\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right)\cdot\mathcal{E}\left[\Delta\pi\left(d, d', \mathbf{d}_{-i}, \mathbf{z}_i\right)\middle|\mathcal{J}_i\right]\right] \\
+\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right)\cdot\nu_{1,i,d,d'}\right] \\
+\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right)\cdot\nu_{2,i,d,d'}\right]
\end{aligned}$$

Each term of the sum is nonnegative. The first one, because both $\Delta\pi\left(\cdot\right) \geqslant 0$ by Assumption 1 and 2, while $h^i\left(\cdot\right)$ is non-negative, and the last two by Assumption 3. Hence

$$\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i\left(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}\right)\Delta r\left(d, d', d_{-i}, z_i; \theta\right)\right] \geqslant 0$$

Now let's think about what these assumptions mean. First, how can we think about $h^i\left(\cdot\right)$? This can be any nonnegative function that depends on an agent's choice $\mathbf{d}_i$, any variable they know $\mathcal{J}_i$, and any observable (to the econometrician) variable that the other agents know. Second, what do we need to satisfy the two conditions?

Condition (a) is satisfied in single agent problems by any function $h^i\left(\cdot\right)$. That's because $\nu_{1,i,\mathbf{d}_i,d'}$ is mean independent of $\mathbf{d}_i$ and $\mathcal{J}_i$, and there are no other firms, so$\mathbf{x}_{-i}$ does not matter. Hence, if we can find a function that satisfies (b), it will also satisfy (a). Let's keep that in mind, and put that aside for a second. In multiple agent problems we may run into issues when satisfying condition (a) if $\mathbf{x}_{-i}$ is observable by the econometrician, but not by the agent (this can happen in games of asymmetric information). It's a pretty extreme case, and not one you need to worry too much about (I am unaware of empirical examples where this is a serious issue), but still important to know. In general, condition (a) will be easy to satisfy.

Condition (b) however, will not. Remember that we expect $\nu_{2,i,\mathbf{d}_i,d'}$ to have *negative* expected value in our data because the agent chooses d_i based on ν_2. To see why, notice that by Assumption 1

$$\begin{aligned}
&\mathcal{E}\left[\Delta\pi\left(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{y}_i\right)\middle|\mathcal{J}_i\right] \geqslant 0 \\
\Longrightarrow\ &\mathcal{E}\left[\Delta r\left(\mathbf{d}_i, d', \mathbf{d}_{-i}, z_i^o, \theta\right)\middle|\mathcal{J}_i\right] - \mathcal{E}\left[\nu_{1,i,\mathbf{d}_i,d'}\middle|\mathcal{J}_i\right] - \mathcal{E}\left[\nu_{2,i,\mathbf{d}_i,d'}\middle|\mathcal{J}_i\right] \geqslant 0 \\
\Longrightarrow\ &\mathcal{E}\left[\Delta r\left(\mathbf{d}_i, d', \mathbf{d}_{-i}, z_i^o, \theta\right)\middle|\mathcal{J}_i\right] - \mathcal{E}\left[\nu_{2,i,\mathbf{d}_i,d'}\middle|\mathcal{J}_i\right] \geqslant 0
\end{aligned}$$

Where in the last step we use the fact that $\mathcal{E}\left[\nu_{1,i,\mathbf{d}_i,d'} \middle| \mathcal{J}_i\right] = 0$. If the inequality has to hold, then expected value of $\nu_{2,i,\mathbf{d}_i,d'}$ is more likely to be low (and negative). Finding ways to correct for this selection problem is the chief challenge in papers that use moment inequalities, and indeed, in a number of cases, the answer that researchers used has been to set $\nu_{2,i} = 0$, i.e. assume that there are no payoff-relevant random variables that the agent observes, but the econometrician doesn't. This is *always* an unrealistic assumption. However, in some cases, it may not be overly unrealistic, and sometimes the econometrician can argue that the selection effect is likely to go against finding a result.

2.3 Closing remarks

The moment inequalities we derived in this section have two big advantages over estimation techniques that yield point-estimates. First, you can use moment inequalities to reduce the number of assumptions you are making when estimating a model. This is helpful if you are not sure what variables are part of an agent's information set — as in the supermarket example in the first section, or Dickstein and Morales (2018) — or if you want to assume a more flexible distribution for an error term — as in Ho and Pakes (2014). Second, you can use moment inequalities to obtain restrictions in models that are too complicated to solve using traditional techniques. These types of models are increasingly common in the literature, and include complex network problems (Ishii, 2005; Morales et al., 2019), dynamic models with spillovers and externalities (Holmes, 2011; Maini and Pammolli, 2020) and more.

Keep in mind however, that even if moment inequalities can help relieve the burden of assumptions, you will still have to restrict distributions, functions, etc. The way you do so will depend on the problem at hand, and you're gonna have to decide whether or not those assumptions are realistic in your specific application. In what follows, we examine a number of examples and different approaches.

3 APPLICATIONS AND EXTENSIONS IN THE EMPIRICAL LITERATURE

Note: in presenting the models in these papers I often make simplifications to facilitate exposition. Occasionally, but not always these simplifications are noted in the text. If you are interested in a specific model you should read the paper and not simply rely on these notes.

3.1 A first example: Ishii (2005)

This example is also covered in PPHI. The goal of the paper is to estimate the impact of incompatibility of ATM networks on consumer welfare. This is a version of a question that has been asked many times: given competing networks, what would be the impact of allowing universal compatibility for everyone?[9] One of the parameters that need to be estimated is the cost of installing an additional ATM machine, and we are going to focus on that parameter in what follows.

Setup

The model is a two-stage entry model. Banks first choose the number of ATMs to install, and in the second period interest rates are set conditional on the ATM network, consumers choose banks, make deposits and use ATMs. The second stage of the model yields a revenue function $R\left(d_i, d_{-i}, \mathbf{z}_i^o\right)$ that gives a bank's revenue for a given number of ATMs d_i, conditional on the number of ATMs installed by its competitors d_{-i}, and exogenous variables $\mathbf{z}_i$ (this function is estimated "off-line", and we'll take it as given).

The paper restricts the installation cost to be linear, but allows for an unobserved heterogeneity component, so that the cost of installing d_i machines is $d_i \cdot (\theta_0 + \eta_i)$, where θ_0 is the average cost of installation, and η_i is the unobserved component. This is probably a reasonable assumption, though it's possible that installation costs could be increasing faster than linearly, for example if space for ATMs is scarce. Under this specification, the profit function of a bank is given by

$$\pi\left(d_i, d_{-i}, \mathbf{z}_i\right) = R\left(d_i, d_{-i}, \mathbf{z}_i^o\right) - d_i\left(\theta_0 + \eta_i\right) + \nu_{1,i,d_i}$$

and our inequality condition can be written as

$$\Delta\pi\left(d_i, d', \mathbf{z}_i\right) = \Delta R\left(d_i, d_{-i}, \mathbf{z}_i^o\right) - \left(d_i - d'\right)\left(\theta_0 + \eta_i\right) + \nu_{1,i,d_i,d'}$$

In the notation of section 2.2, the measurement function of the econometrician is

$$\Delta r\left(d_i, d', \mathbf{z}_i^o\right) = \Delta R\left(d_i, d_{-i}, \mathbf{z}_i^o\right) - \left(d_i - d'\right)\theta_0$$

and $\nu_{2,i,d_i,d'} = \left(d_i - d'\right)\eta_i$.

[9] Papers in this area cover all sorts of industries, from hospitals (Ho, 2006), to charging stations for electric cars (Li, 2019), to videogames (Lee, 2013).

Choosing the instrument

One of the tricks to satisfy the requirements for assumption 3 is to come up with moment inequalities that produce a difference in strategies that are linear in terms of the unobserved heterogeneity parameter. In this case, doing so is very easy thanks to the linearity in cost assumption. In fact, for any strategy $d' = d_i + t$, the structural error term will be $t \times \eta_i$. Since $\mathcal{E}[\eta_i] = 0$, it will suffice to compare alternative strategies for fixed t and use a constant term as $h^i(\cdot)$ to obtain a consistent estimator. For example, we could choose $h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}) = \frac{1}{n}$, where n is the number of firms, and compare strategies that include one additional ATM or one less ATM. Doing so we get:

$$\frac{1}{n}\sum_{i=1}^{n}\Delta r(d_i, d_i+1, \mathbf{z}_i^o) \geqslant \underbrace{\frac{1}{n}\sum_{i=1}^{n}\eta_i}_{=0} \geqslant \frac{1}{n}\sum_{i=1}^{n}\Delta r(d_i, d_i-1, \mathbf{z}_i^o)$$

What about assumption 2? Since firms announce n at the same time, it is probably okay to assume that n_{-i} remains constant. The second stage model will account for how endogenous variables (in this case, the interest rate) adjust to deviations.

3.2 Moment inequalities to control for unobserved quality: Ho and Pakes (2014)

This paper addresses the issue of hospital choice, and specifically, how prices and physician incentives affect hospital choice. The specific setting for this paper is choice of hospital for privately insured patients giving birth in California. They choose California because many insurers there have recently switched to capitated payment models, which pay a physician group a fixes sum for the care of one patient. These payment schemes may steer physician incentives towards more cost-effective treatment options. Of course, the danger of these schemes is also that the incentivize underutilization of care, though, so far, all research on the topic shows either a null or slightly positive effect on quality.

The gap in the literature is a modeling assumption that virtually all previous papers on hospital choice made: leaving the price paid by the insurer out of the choice function. This assumption may be okay in some settings, but when it comes to studying the impact of payment models it's clearly inadequate. Introducing price into the choice equation is tricky. There are three issues:

1. The data reports a "list price", but the real negotiated price between an insurer and a hospital is unobserved. They can get average discounted prices, which at least gives them a price measure that has a mean zero error. However, the error term will be correlated with the choice variable, which they need to account for.

2. The price that enters the choice function may not be the real price, but an expected price. This is similar to the issue of supermarket prices. Given that prices can vary by patients (depending on comorbidities, and whatever unexpected issues may come up during the hospital stay), they don't want to assume that everyone knows all the prices. They simply assume that expected prices are right on average.

3. The price will be correlated to quality. This is the classic endogeneity issue you will find in all demand systems.

So they are going to have to deal with all three of these issues.

Setup

This is a static single-agent model. The choice of hospital is based on a *referral* function (basically a joint utility function between the physician and the patient), which they parametrize as

$$W_{i,\pi,h} = \theta_{p,\pi} p\left(c_i, h, \pi\right) + g_\pi\left(q_h\left(s\right), s_i\right) + f_\pi\left(d\left(l_i, l_h\right)\right)$$

where

- i indexes patients, π indexes insurers, and h indexes hospitals
- $p\left(c_i, h, \pi\right)$ is the price that insurer π can expect to pay at hospital h for a patient with condition c_i
- s_i is the severity of the patient condition (they aggregate conditions c_i into severity bins using algorithms from the medical literature), while $q_h\left(s\right)$ is a vector of perceived quality for hospital h for a given severity level s.[10]
- $g_\pi\left(q_h\left(s\right), s_i\right)$ is a plan- and severity-specific function which determines how quality impacts hospital choice
- l_i and l_h are patient and hospital locations respectively, with $d\left(\cdot\right)$ measuring distance and $f_\pi\left(\cdot\right)$ measuring the importance of distance — again, this is an insurance-specific function.

This is a pretty flexible form, and it's important to understand what they achieve with this specification, and also what they leave out.

First, they can capture severity-specific quality with the function $q_h\left(s\right)$. Implicitly, this flexibility acknowledges that quality can have multiple components. For example, a teaching hospital may be used to dealing with more complicated deliveries, but may end up being short-staffed for regular, uncomplicated deliveries, whereas a smaller hospital will struggle to deal with more complicated cases, but will provide a great environment for a standard birth. Note that there could be specific quality variables one could use to explicitly incorporate multi-dimensional quality: past volumes of certain severity types, staffing numbers, etc. That's not what the function incorporates. Instead, this is a more reduced-form way to extract similar information.

Second, they allow different insurers to affect referral functions differently (this is reflected in $\theta_{p,\pi}$, $g_\pi\left(\cdot\right)$, and $f_\pi\left(\cdot\right)$). This is important because variation in physician incentives exists across insurers. In particular, in their empirical setting, they know the fraction of patients that are covered

[10] The interpretation of c_i and s_i are different: the former is supposed to capture the cost of a patient, the latter is supposed to capture patient characteristics that may interact with specific aspects of a hospital's quality.

under capitation incentives for each insurer (the numbers vary from as low as 37% to as high as 97%). Different data may have called for a different specification. For example, if they knew which patients were capitated, they could have used functions that were specific to the payment model instead of the insurer.

Finally, what they leave out: there is no patient-specific error term here, meaning that they assume that patients do not have specific-preferences for certain hospitals that cannot be captured by the variables s_i,$d\left(l_i, l_h\right)$, and c_i.The way to interpret this is as a restriction on the distribution of the $\nu_{2,i,h_i,h'}$ term. It will soon be clear why they make this assumption once we discuss the estimation strategy.

Estimation

The most important thing is the specification of $g_\pi(\cdot)$. If there is an unobserved quality component that we cannot adequately instrument for (and in this case there are no good instruments), then we will run into serious endogeneity issues. A standard logit-type estimation will find that the price coefficient is positive. In theory, the way that this problem is addressed is by utilizing a richer set of control variables. In practice, this quickly makes estimation unfeasible, both because of computational constraints, and because the number of observations in the data is fixed, so adding more and more parameters introduces more measurement error (remember that in the logit model we assume there is no error in prices).

Instead, they use moment inequalities to get rid of $g_\pi(\cdot)$. How? They select couples of individuals i and i' with the same severity s, and the same insurer π, that chose different hospitals h and h', and then sum $\Delta W\left(i, h, h'\right)$ and $\Delta W\left(i, h', h\right)$. To see why this works notice that

$$\Delta W\left(i, h, h'\right) = \theta_{p,\pi} p\left(c_i, h, \pi\right) + g_\pi\left(q_h(s), s\right) + f_\pi\left(d\left(l_i, l_h\right)\right) \\ - \theta_{p,\pi} p\left(c_i, h', \pi\right) - g_\pi\left(q_{h'}(s), s\right) - f_\pi\left(d\left(l_i, l_{h'}\right)\right)$$

and

$$\Delta W\left(i', h', h\right) = \theta_{p,\pi} p\left(c_{i'}, h', \pi\right) + g_\pi\left(q_{h'}(s), s\right) + f_\pi\left(d\left(l_{i'}, l_h\right)\right) \\ - \theta_{p,\pi} p\left(c_{i'}, h, \pi\right) - g_\pi\left(q_h(s), s\right) - f_\pi\left(d\left(l_{i'}, l_h\right)\right)$$

In the sum, the $g_\pi(\cdot)$ terms cancel out and we obtain

$$\Delta W\left(i, h, h'\right) + \Delta W\left(i', h', h\right) \\ = \theta_{p,\pi}\left(\Delta p\left(c_i, h, h', \pi\right) + \Delta p\left(c_{i'}, h', h, \pi\right)\right) + \left(\Delta f_\pi\left(d\left(l_i, l_h, l_{h'}\right)\right) + \Delta f_\pi\left(d\left(l_{i'}, l_{h'}, l_h\right)\right)\right)$$

Revealed preference implies that the expected value of $\Delta W\left(i, h, h'\right) + \Delta W\left(i', h', h\right)$ has positive expected value. Averaging across all pairs i, i' that satisfy the prerequisites for inclusion eliminates all sources of expectational error.

Methodological extension: dealing with $\nu_{2,i}$ when differencing out is not possible

The trick to dealing with the structural error in Ho and Pakes is to make assumptions that enable them to find two observations with the same structural errors, and then combine them in a way that cancels them out. In particular, the key assumption is that ν_2 is not i-specific, but in fact can be specific to groups of individuals that share similar characteristics. In other settings, this may not be possible or palatable, and more recent papers have proposed approaches that relax the assumptions that need to be made on the distribution of $\nu_{2,i}$. The approaches of these papers tend to be computationally more taxing, and also a bit convoluted in their presentation, so I will not delve into details here, though I will try to relay the main intuition behind their results.

The first paper in this list of approaches is Dickstein and Morales (2018). Their motivation to use moment inequalities is to relax and test assumptions on the amount of information firms have access to when making decisions (An aside. The issue of information has come up multiple times in these notes. In the supermarket example, remember how the logit model is biased if shoppers don't know all prices. The same issue arises with hospital in Ho and Pakes. In both situations, moment inequalities can be constructed to remove those concerns. The spirit of Dickstein and Morales (2018) is to go beyond simply alleviating information concerns and actually try to figure out whether they can rule out that firms use certain variables in forming expectations. It's a compelling concept, and I encourage you to read the paper in its entirety if you find the question interesting).

The paper analyzes export decisions of Chilean firms. They have a model to back out firm revenues from sales data. They have panel data, so they see each firm i in a number of markets j, for a number of years t. Their model for revenue allows for country-year specific heterogeneity, but no firm-country-year specific errors that are observed by the firm but not by the econometrician. However, they do allow firm-country-year unobserved heterogeneity in the cost function. This is the structural ν_2 error.[11]

Their solution to solving the selection problem is parametric. They assume ν_2 has a normal distribution. As a result, they can back out an equation, that gives them the likelihood that a firm will export to a given market conditional on their estimated revenue equation. They use this probability estimate to build inequalities that impose as restriction the condition that certain types of firms need to be more likely to export relative to others. They call these odd-based inequalities, and they are slightly different in their form than the traditional moment inequalities we have seen so far. The derivation is not straightforward, and involves an application of Jensen's inequality (you can check the proof in the Online Appendix of the paper), but the intuition is basically the same as other estimator that account for selection in the data, like Heckman (1979). It should also be noted that as these are based on odds, they can only be applied to binary choices, and cannot be extended to multinomial choice. In their paper, this means that they have to assume that the choice to export is independent across countries.

[11] They have an extension that allows for ijt-specific heterogeneity in the revenue as well, though, as far as I can tell, that model is identified only through functional form assumptions.

There is also another paper (Illanes, 2016), which further relaxes the need for a distributional assumption on ν_2 by using a technique called latent integration. His method can be applied to multinomial choice.

3.3 Moment inequalities to infer network gains: Holmes (2011)

The previous two papers offered techniques to get around the selection issue generated by the structural error $\nu_{2,i}$. In both cases, the key was a combination of assumptions on the distribution of $\nu_{2,i}$ together with clever choice of counterfactual choices that allowed the econometrician to remove choice-specific error. I am now going to discuss a paper where finding such counterfactual choices is close to impossible, and where as a result the assumption is that $\nu_{2,i} = 0$.

The paper examines the expansion of Walmart starting in the 1970s. One of the keys to Wal-Mart's success is that its distribution network is vertically integrated, which means that one of its competitive advantages is the ability to rely on a dense distribution network. Indeed, the roll-out of new Walmart stores is quite striking: from its central first location in Arkansas, WalMart always opened new stores close to existing ones (you can download a video of store openings on Holmes' website). The goal of the paper is to quantify the benefit of this dense network for Walmart.[12] The key intuition for how this benefit is identified is that there is a cost in placing stores close together: when market areas overlap, new stores can cannibalize sales from existing stores.

The main difficulty in treating this model is that there are an enormous number of different strategies that Walmart can follow, and no clear way to derive an analytical solution. Hence, in this case, the choice to resort to moment inequalities is motivated by the fact that it's almost impossible to derive estimating conditions in any other way.

Setup

Holmes models the economy as a set of L locations, indexed by ℓ, where $d_{\ell\ell'}$ is distance between two locations ℓ and ℓ'. At any given point in time, a subset $\mathcal{B}_t^{Wal}$ of locations have a Walmart.[13] Sales at a particular store depend on the location, and also on its proximity to other Walmarts (hence, a store draws sales not just from its location but also from neighboring locations): profits at a particular store j at time t is denoted as $\mu R_{jt}\left(\mathcal{B}_t^{Wal}\right)$, where $R_{jt}(\cdot)$ is the revenue function and μ is the gross margin, assumed constant across stores and time.

The model has three sources of cost besides the cost of products sold (which is incorporated by the parameter μ):

1. Distribution costs: supplies for Walmart stores are delivered from distribution centers. The distribution cost is parametrized as $DC_{jt} = \tau d_{jt}$, where d_{jt} is the distance from the closest distribution center (note that distribution centers are not stores themselves). τ is the main parameter of interest here, as stores that are closer to distribution centers will have lower

[12] If you find the question interesting, you might also want to check out Houde and Newberry (2016), which performs a similar exercise using Amazon fulfillment centers.

[13] The paper distinguishes between supercenters and regular stores. I disregard this distinction in this presentation.

costs. Also note that this is modeled as a fixed cost, even though there may be a variable component. The variable component is suppressed for simplicity.

2. Variable store costs: Holmes incorporates the cost of labor, land, plus a residual "other variable" cost term as a percentage of revenue. So variable cost for component k is $v_k R_{jt}\left(\mathcal{B}_t^{Wal}\right)$.

3. Fixed costs: the level of fixed cost will not affect the choice of *where* to locate a store, only the choice to open it or not. Since Holmes is interested only in the location choice, he only cares about capturing differences in fixed costs across stores. He assumes that any differences in fixed costs are related to population density across locations.

The maximization problem that Walmart is solving has 4 parts:

1. How many new stores do we open this year?
2. Where do we locate these stores?
3. How many distribution centers do we open?
4. Where do we locate these distribution centers?

Hopefully it's clear at this point that this is a very complicated problem (if you think back to Ishii (2005), her entire problem is essentially part 1., though in a multi-agent setting). Thankfully, moment inequality allows Holmes to avoid having to directly solve any of the four parts of the maximization because the inequalities he is going to use will remove any unobservables related to parts 1., 3. and 4.. We will see how in a moment.

The maximization problem of Walmart is

$$\max_a \sum_{t=1}^{\infty} \beta^{t-1} \left(\sum_{j \in \mathcal{B}_t^{Wal}} \left(\pi_{jt} - c_{jt} - \tau d_{jt}\right) \right)$$

where π_{jt} is the operating profit (the revenue function minus the variable cost), and c_{jt} is the fixed cost.

Once again, before moving forward, let's think about what this specification leaves out. First, notice that τ does not have a store-specific component (unlike, say, in the supermarket example).

Estimation

Holmes uses store-level sales data to estimate a model of demand for Walmart stores, which he uses to back out an implied *"cannibalization rate"*, i.e. the percentage drop in sales in pre-existing Walmarts that is due to the opening of new Walmarts. Intuitively, this particular moment of the data is important because the higher the degree of cannibalization that Walmart is willing to bear, the higher the returns from store density must be. The demand function is estimated in a way that allows for measurement error, but not for unobserved store characteristics.

The deviations that Holmes considers to build moment inequalities involve swapping the opening period of two stores. For example, if store 1 opened in 1962 and store 2 opened in 1964, then the deviation would open store 2 in 1962 and store 1 in 1964, leaving everything else unchanged.

These deviations are helpful because a lot of parameters cancel out in the difference. Since you are opening the same number of stores, you don't have to worry about fixed cost. Second, they leave profits unchanged both before and after the deviation. In the example I gave, Walmart's profits before 1962 and after 1964 are identical under both strategies. Hence, we don't need to project how a deviation would affect profits in the long run.

A few final thoughts:

1. Deviations that anticipate entry in stores that are further away, and delays entry in stores that are closer will yield a lower bound on τ. Vice versa, deviations that anticipate entry in stores that are closer, and delay entry in stores that are further away yield an upper bound.

2. Since he has measurement error in the demand function he has to aggregate and average multiple moment conditions to smooth this error out. That's because one moment condition may fail simply because of the presence of a large negative measurement error shock. It's not a very taxing requirement, but important to keep in mind nonetheless.

3. The $\nu_{2,i}$ constraint manifests itself in the fact that there is no store-specific cost component. If there was, then stores that are opened earlier would have lower costs in expectation, and the inequalities may not hold (even averaging across multiple moment conditions). In theory, if he had data on store costs, then he may have been able to back those out offline, though I am unaware of any papers that have access to that kind of data.

4. However, I think he would be able to incorporate unobserved store characteristics in the demand system (maybe no given the data he has, though I'm not sure). That's because you can back those parameters out from the demand system and, as long as they are time-invariant, and then account for them in the inequalities.

3.4 Moment inequalities with unobserved strategies: Maini and Pammolli (2020)

Finally, this paper considers a model close to Holmes, but one where firms face unobserved constraints that impose stochastic constraints on their choice set (rather than payoff shocks). It contains an extension of the PPHI framework to account for these different kind of shocks. It turns out, somewhat surprisingly, that the properties of the moment inequalities that can be derived are different than the standard PPHI inequalities. In particular, the bounds will often be one-directional.

The paper examines the impact of external reference pricing policies on the entry strategies of pharmaceutical companies in Europe. External reference pricing is a policy that links drug prices to the prices of the same drug abroad. Hence, launches in countries with low prices can carry a

negative spillover effect on overall revenue if they drag down prices elsewhere. The main goal is to estimate the impact of the policy on launch delays. To do so, the authors need to disentangle strategic delays from delays that arise for idiosyncratic reasons.

Setup

Like the Holmes paper, they use a multi-location entry model. There are fewer locations (25 countries, vs. a partition of the whole US in Holmes), but the revenue implications of the launch sequence are much more complicated to track as we will see in second.

First, we briefly consider the period revenue function. The firm earns additive revenue in each period from each country where their product is available. Demand is estimated as a nested logit. Given that most patients access drugs at deeply discounted prices, they do not estimate a price coefficient (they have a quantity coefficient in the pricing equation, which is somewhat akin to including a single price-elasticity coefficient in the demand system). Price is a weighted average of an internal government benchmark p_{ijt}^{gov}, which varies across drugs, countries and years, and the reference price p_{ijt}^{ref} (which is based on previous period prices, with a functional form that is assumed to be known). The weights are parameters of the model to be estimated.

$$p_{ijt} = \begin{cases} p_{ijt}^{\text{gov}} & \text{if } p_{ijt}^{\text{ref}} \geqslant p_{ijt}^{\text{gov}} \\ (1-\mu_j)\, p_{ijt}^{\text{gov}} + \mu_j p_{ijt}^{\text{ref}} & \text{if } p_{ijt}^{\text{ref}} < p_{ijt}^{\text{gov}} \end{cases} \tag{9}$$

In the dynamic choice problem the firm chooses when to enter in each market. However, rather than entering right away, firms have to apply in each country, and then their application is accepted with probability $(1-\psi_j)$. If an application is not accepted, it is delayed until the next period, when a new shock is drawn. The value function of the problem is

$$V_t\left(S_{t-1}\right) = \max_{\mathcal{A}_t = \{A_\tau(S_{\tau-1})\}_{\tau=t}^{T}} \sum_{S} \left(\sum_{\tau=t}^{T} \beta^{\tau-t} \Pi_\tau\left(S\right) \right) \cdot \mathcal{P}\left(S \middle| S_{t-1}, \mathcal{A}_t\right)$$

where the state space variable $S_t$is the entry sequence as of time t, S is the overall entry sequence, and

$$\Pi_\tau\left(S\right) = \mathbb{E}\left[\sum_j D_{ij\tau} \times p_{ij\tau}\left(S\right) \right]$$

The transition probabilities $\mathcal{P}\left(S \middle| S_{t-1}, \mathcal{A}_t\right)$ depend on the actions of the firm, which are given by the set of state-specific vector $\mathcal{A}_t$, and on the random delay shocks.

Estimation

Due to the complex incentives generated by reference pricing there is no analytical solution to the model. Moreover, with 25 countries, and up to 11 periods of data, the choice set of a firm is too large to compute a solution numerically. This is what motivates the use of moment inequalities.

Building the right inequalities is also complicated. Each launch has dynamic implications through the reference pricing channel. Moreover, the horizon of these implications is not limited to the current, or even the next period. Instead, anticipating or delaying a launch by one period carries implications for the entire pricing sequence. This is because reference prices can take a while to adjust. Many countries use average as their reference function, meaning that the reaction to a small perturbation can take many periods to be fully incorporated as prices converge towards a steady state level. As a result, the small perturbation approach in Holmes cannot be used here to create moment inequalities.

As an additional problem, the econometrician cannot simulate the expected value of the strategy played by the agent, simply because the strategy is not observed: in the data, an application that was delayed is observationally equivalent to an application that was not sent on purpose. So moment inequalities have to be built in a different way.

The solution is to base the restrictions on the revenue implications of the model. In the data, the econometrician sees the revenue earned by each firm. The revealed preference assumption guarantees not only that the strategy played is the preferred one, but that no other strategy earns as much. So for a given guess of the parameter ψ_j we can simulate the profits of a bunch of strategies and see if we can earn higher profits than the firm. If we do, then the parameter guess has to be rejected. There is an additional step that's required here, and that's aggregating across firms. That's because the revenue earned in the data is one draw from the distribution of revenue (which is affected by the random shocks as well as the uncertainty in demand and price). Aggregating across drugs removes the source of error in a large sample.

Some final notes The paper makes several simplifying assumptions to make estimation feasible:

1. Firms are treated as monopolies. While brand drugs do not face competitions from generic products, most face similar (though not identical) competitors. The impact of competitors is incorporated in the demand and price function (through a reduced form parameter), but not when firms make dynamic decisions. Note that this is makes the model not entirely internally consistent (i.e. other firm's actions do affect my profits, but I do not take into account how they would react to my strategies). In practice though, the impact is small, and the assumption is that firms do not think about the reaction of competitors when considering deviations.

2. All payoff-relevant random variables known to the firm are also known to the econometrician. This is the assumption that $\nu_{2,i} = 0$, which is the same as in Holmes. Papers that have looked at dynamic entry models with spillovers have generally made this assumption (for another example, see Morales et al., 2019). In this case, the main concern would be unobservedly high demand in countries where the firm has launched. However, the model does incorporate country-drug fixed effects, so the only possible source of unobserved heterogeneity has to be country-drug-year specific. There could be a concern that firms may delay entry on purpose while waiting for demand to be high enough. However in this case, the

model will have an upward bias in demand prediction, which would lead towards earlier entry, and generally work against finding a result.

3. The price function is essentially a reduced-form control function. This is generally highly problematic, but in this case the concerns are somewhat allayed by the fact that the counterfactual predictions do not rely on price prediction out of equilibrium. The appendix of the paper also shows that the price function can be micro-founded from a Nash Bargaining problem, albeit one that is too simple to really be taken seriously.

4 BIBLIOGRAPHY

Ackerberg, Daniel, C. Lanier Benkard, Steven Berry, and Ariel Pakes, "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 2007, *6* (SUPPL. PART A), 4171–4276.

Andrews, Donald W. K. and Panle Jia Barwick, "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 2012, *80* (6), 2805–2826.

Bajari, Patrick, C. Lanier Benkard, and Jonathan Levin, "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, sep 2007, *75* (5), 1331–1370.

Berry, Steven T., "Estimation of a Model of Entry in the Airline Industry," *Econometrica'*, 1992, *60* (4), 889–917.

Ciliberto, Federico and Elie Tamer, "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 2009, *77* (6), 1791–1828.

Collard-wexler, Allan, "Demand Fluctuations in the Ready-Mix Concrete Industry," *Econometrica*, 2013, *81* (3), 1003–1037.

Dickstein, Michael J. and Eduardo Morales, "What Do Exporters Know?," *Quarterly Journal of Economics*, 2018, *133* (4), 1753–1801.

francois Houde, Jean and Peter Newberry, "Sales Tax, E-commerce, and Amazon's Fulfillment Center Network," 2016, pp. 1–3.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

Ho, Kate and Ariel Pakes, "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians," *American Economic Review*, dec 2014, *104* (12), 3841–3884.

Ho, Katherine, "The welfare effects of restricted hospital choice in the US medical care market," *Journal of Applied Econometrics*, nov 2006, *21* (7), 1039–1079.

Holmes, Thomas J., "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, 2011, *79* (1), 253–302.

Illanes, Gastón, "Switching Costs in Pension Plan Choice," 2016.

Ishii, Joy, "Compatibility , Competition , and Investment in Network Industries : ATM Networks in the Banking Industry," 2005.

Katz, Michael, "Essays in Industrial Organization and Technological Change." PhD dissertation 2007.

Lee, Robin S., "Vertical integration and exclusivity in platform and two-sided markets," *American Economic Review*, 2013, *103* (7), 2960–3000.

Li, Jing, "Compatibility and Investment in the U.S. Electric Vehicle Market," 2019.

Maini, Luca and Fabio Pammolli, "Reference Pricing as a Deterrent to Entry: Evidence from the European Pharmaceutical Market," 2020.

Maskin, Eric and Jean Tirole, "A Theory of Dynamic Oligopoly, III. Cournot Competition," *European Economic Review*, 1987, *31* (4), 947–968.

_ and _, "A Theory of Dynamic Oligopoly, I: Overview and Quantity Competition with Large Fixed Costs," *Econometrica*, 1988, *56* (3), 549–569.

_ and _, "A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles," *Econometrica*, 1988, *56* (3), 571–99.

_ and _, "Markov Perfect Equilibrium," *Journal of Economic Theory*, oct 2001, *100* (2), 191–219.

Mazzeo, Michael J., "Product Choice and Oligopoly Market Structure," *The RAND Journal of Economics*, 2002, *33* (2), 221.

Morales, Eduardo, Gloria Sheu, and Andrés Zahler, "Extended Gravity," *Review of Economic Studies*, 2019, *86* (6), 2668–2712.

Pakes, Ariel, "Alternative Models for Moment Inequalities," *Econometrica*, 2010, *78* (6), 1783–1822.

_ , Jack Porter, Kate Ho, and Joy Ishii, "Moment Inequalities and Their Application," *Econometrica*, 2015, *83* (1), 315–334.

Ryan, Stephen P, "The Costs of Environmental Regulation in a Concentrated Industry," *Econometrica*, 2012, *80* (3), 1019–1061.

Tamer, Elie, "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 2003, *70* (1), 147–165.