Moment Inequalities and Their Application

A. Pakes, J. Porter, Kate Ho, and Joy Ishii*

December 13, 2011

Abstract

This paper provides conditions under which the inequality constraints generated by either single agent optimizing behavior, or by the Nash equilibria of multiple agent problems, can be used as a basis for estimation and inference. We also add to the econometric literature on inference in models defined by inequality constraints by providing a new method of inference for the boundaries of the model's identified set. An application illustrates how the use of inequality constraints can simplify estimation and inference in complex behavioral models, and a Monte Carlo with sample design based on the application considers the performance of alternative inference procedures for boundary points.

1 Introduction

This paper provides conditions under which the inequality constraints generated by single agent optimizing behavior, or by the Nash equilibria of multiple agent games, can be used as a basis for estimation and inference. The conditions do not restrict choice sets (so the controls can be discrete, continuous, or have more complex domains) or require the researcher to specify a parametric form for the disturbance distributions (though some restrictions are imposed on those distributions), and they do allow for endogenous regressors. In addition, the conditions do not require specification of the contents of agents' information sets or an equilibrium selection mechanism (in cases in which there may be multiple equilibria).

The generality provided by these conditions does come with some costs, however. First, perhaps not surprisingly, under our conditions *partial* identification of the parameters of interest is likely. We add to the econometric literature on inference for such models

*The authors are from Harvard University and the National Bureau of Economic Research, the University of Wisconsin, Harvard University, and Harvard University, respectively. Much of this was written while Pakes was at New York University and we thank that institution for their hospitality. We also thank Jeremy Fox, Oliver Hart, and Ali Hortascu, Guido Imbens, and Bill Sandholm for valuable comments. Pakes and Porter thank their respective NSF grants for financial support.

by providing new inferential procedures for boundary points of the model's identified set which are not computationally burdensome. Second, though we provide sufficient conditions to generate a set of moment inequalities for inference, we do not have necessary conditions, and hence do not know the limits of our framework. So there remains the question of precisely which of the models typically used to structure data satisfy our sufficient conditions.

We show that a number of familiar single agent models do and often can be analyzed with our framework using less restrictive assumptions than typically assumed. Moreover the multiple agent analogs of these models, models which typically do not satisfy the assumptions originally used for the single agent problems, satisfy our conditions also. We then use our framework to empirically analyze a multiple agent problem of policy interest. Finally we provide a Monte Carlo analysis based on a sample design taken from the empirical analysis which compares alternative inference procedures for boundary points.

The next section of the paper provides our analytic framework. We begin by assuming that our agents maximize their expected returns. This yields a "revealed preference" inequality; the expected returns from the strategy played should be at least as large as the expected returns from feasible strategies that were not played. Since we do not want to specify how these expectations are formed, we consider only the implications of this assumption on the difference between the realized returns at the agent's observed strategy (or "choice") and the returns the agent would have earned had it played an alternative feasible strategy. What the revealed preference theory tells us is that the expectation of this difference is nonnegative. When there are interacting agents these inequalities are necessary conditions for any (of the possibly many) Nash equilibria.

We assume that the econometrician can construct an approximation to the realized returns from both the actual choice and from at least one feasible alternative, and that these approximations depend on only a finite dimensional parameter vector of interest and observable random variables. We then consider the difference between the returns at the observed choice and at the alternative. This difference has an actual value given by the difference in actual realized returns (whose expectation is positive at the true value of the parameter), and it has an approximated value given by the difference in the approximated returns. The difference between the expectation of the increment in realized values and the increment in the approximated values becomes the disturbance.

This disturbance can be decomposed into two terms. The first is mean independent of the variables known to the agent when the agent makes its decision. It consists of expectational error (due to asymmetric or other forms of incomplete information) and/or measurement and approximation errors that satisfy the mean independence condition. The second, or structural, error is defined as that part of the difference between the approximated and actual difference in returns that *was* known to the agent when it made its decision, but is not observed by the econometrician. Since the agent knew the value of this disturbance when it made its choice, and the disturbance is a component of the expected profit differences from that choice, the choice itself potentially depends on this disturbance's value. So when we observe a decision we know that the value of

the structural disturbance associated with it must have been "selected" from the subset of the possible values that would lead to that decision. As a result, the expectation of the structural disturbance corresponding to the decision can be non-zero, even if the structural disturbance corresponding to any fixed value of the possible choices is mean zero in the relevant population. This non-zero expected disturbance implies that the differences in approximated returns based on observables may not mirror the revealed preference assumption of a non-negative expectation at the true value of the parameter.

We provide a sufficient condition for overcoming this hurdle, and a number of ways of satisfying it. In some models certain linear combinations of differenced returns will not depend on the structural disturbance (they will "difference out" that error). Examples include (generalizations of) the standard assumptions underlying models which use matched observations to control for unobservables, and Industrial Organization (or social interaction) models with market (or network) specific unobservables known to the agents but not to the econometrician. A second possibility results from an ability to choose linear combinations of differenced returns that are additive in the structural errors regardless of the decision made. This case allows us to use standard assumptions on the availability of instruments to construct sample analogues to moments of the structural error that do not condition on decisions, and the expectation of these unconditional moments will be non-negative at the true value of the parameter vector. Examples include multiple agent ordered choice models and contracting models where not all the determinants of the payments specified in the contract are observed by the econometrician. A third possibility occurs when it is reasonable to assume that the distribution of the structural disturbance is symmetric and we can use the symmetry to correct for the selection bias in that disturbance (this extends an idea due to Powell (1986) for use in moment inequalities). Examples include models with either continuous or ordered controls that are bounded on one side (such as Tobit models or auction models with bids that must be above a reservation value).

When there is no structural disturbance our framework is a natural extension of the first order condition estimator for single agent dynamic problems proposed in Hansen and Singleton (1982), and extended to allow for transaction costs, and hence inequalities, by Luttmer (1996). Our extension is to allow for arbitrary (including discrete) choice sets and interacting agents. As in Hansen and Singleton, we do not require either parametric assumptions on the distribution of the disturbance term, or a specification for what each agent knows at the time the decision is made (and we allow for asymmetric and other forms of incomplete information). Since it is rare for the econometrician to know what each agent knows about its competitors' likely actions, the fact that we need not specify information sets is particularly appealing in multiple agents settings. Ciliberto and Tamer (2009) and Andrews, Berry, and Jia (2004) provide alternative methods for estimating models with discrete choice sets and interacting agents. The two approaches are not nested and Pakes (2010) provides a formal comparison of their assumptions and a Monte Carlo analysis of the resultant estimators' robustness to deviations from their modeling assumptions.

Section 3 of this paper adds to the econometric literature on moment inequality inference with methods for providing conservative inference for boundary points of the identified set generated by moment inequalities. In models with three or more dimensions it will be difficult for empirical researchers to provide a full description of the identified set, and they are likely to look for lower dimensional descriptive statistics. Methods of inference for extreme points provide empirical researchers with an analogue to the standard approach in point identified empirical models of reporting dimension-by-dimension parameter estimates with associated standard errors; i.e. we provide dimension-by-dimension extreme points and associated confidence intervals.

Our approach differs from methods that focus on confidence sets for the whole identified set or the true value of the parameter. Such methods could be modified, for instance, through projections onto a given dimension, to provide the same end product, and we compare our procedures to some of the alternative possiblities in a Monte Carlo example in section 4. A difference between our estimators and the estimators designed to form confidence sets for the true point or for the identified set, is that our inference procedure is developed through a local approximation to the limiting distribution of an extreme point estimator. This has computational advantages, as we need not compute the value of the objective function on a grid that covers all possible parameter values. The computational advantages grow with the dimension of the parameter vector being estimated and are likely to be larger in problems where a fixed point must be calculated to evaluate any given value of the parameter vector. Our procedure does, however, assume the identified set has a non-empty interior and that the extreme point in the dimension of interest is not a singleton.

Section 4 provides our empirical example and a Monte Carlo with a sample design based on the empirical problem. While our econometric approach is not computationally demanding, both the precision of inference and the relevance of our behavioral assumptions are still open questions. Our empirical application is both informative and encouraging in this respect. It is a problem which does not satisfy the assumptions which justify the use of more traditional tools; an investment problem with interacting agents and non-convex or "lumpy" investment alternatives (it analyzes banks' choices of the number of their ATM locations). It also illustrates the ease with which the proposed framework can handle environments in which there can be many possible "network" equilibria. Formally it is a multiple agent ordered choice problem; a problem which arises frequently in industrial organization and one which illustrates the intuition underlying the use of our framework quite clearly. The small sample sizes do force us to use parsimonious specifications. However, the results make it clear even with the small sample the new techniques provide useful information on important parameters. The more detailed policy implications of the estimates are discussed in Ishii (2011).

The Monte Carlo is based on a model with only two parameters. This make it easy to compare alternative estimators. The results indicate that different estimators are likely to be preferable in problems with different, observable, properties.

2 A Framework for the Analysis

This section derives the moment inequalities that serve as the basis for econometric inference. We start from a player's (or "agent's") best response condition in a simultaneous move game (single agent problems are treated as a special case), and add two more assumptions; one which allows us to compute counterfactual profits if we had access to the agent's profit function, and one which constrains the relationship between the agent's perceived profits and the profits that we can actually measure. These assumptions place restrictions on the stochastic environment which imply a set of moment inequalities. We discuss the restrictions implied by each assumption immediately after introducing it.

2.1 Agents' Problem

Agents are indexed by $i = 1, \ldots, n$. Let $\mathcal{J}_i$ be a random variable denoting the information set available to agent i when actions are chosen ("decisions" are made), where $\mathcal{J}_i \in \mathcal{I}_i$, the space of such information sets. Let $\mathcal{D}_i$ be the set of actions agent i could take. Then the strategy played by agent i is a mapping $s_i : \mathcal{I}_i \to \mathcal{D}_i$. The strategy and information set for each player generate observed decisions $\mathbf{d}_i = s_i(\mathcal{J}_i)$. For notational convenience we assume these are pure strategies[1], so $\mathcal{D}_i$ is the support for $\mathbf{d}_i$. Note that we distinguish between $\mathbf{d}_i$ and the realization of the decision, say d_i, by using boldface for the former random variable.

When $\mathcal{D}_i \subset \mathcal{R}$ it can be either a finite subset (as in "discrete choice" problems), countable (as in ordered choice problems), uncountable but bounded on one or more sides (as in continuous choice with the choice set confined to the positive orthant), or uncountable and unbounded. If $\mathbf{d}_i$ is vector-valued then $\mathcal{D}_i$ is a subset of the appropriate product space.[2]

The payoff (or profit) to agent i will be determined by agent i's decision, the other agents' decisions, and an additional set of variables $\mathbf{z}_i$ with support $\mathbf{Z}_i$. Profits will be given by the function $\pi : \mathcal{D}_i \times \mathcal{D}_{-i} \times \mathbf{Z}_i \to \mathcal{R}$, where $\mathcal{D}_{-i}$ denotes $\times_{j \neq i} \mathcal{D}_j$. Not all components of $\mathbf{z}_i$ need to be known to the agent at the time it makes its decisions and not all of its components need to be observed by the econometrician. Notice that by indexing profits by $\mathbf{z}_i$ we abrogate the need to keep track of inter-agent differences in profit functions.

[1] We could obtain a moment inequality of exactly the same form as the inequality derived below from a game in which agents used mixed strategies provided each pure strategy with positive probability in the mixed strategy had the same expected return. Notice that this implies that when using our inequalities there is no need for the econometrician to specify whether the underlying strategies are pure or mixed. However if we did know mixed strategies were being played, and we could distinguish the mixed strategies associated with particular information sets, then more information would be available for use in estimation than the information we use; see Beresteanu and Molinari (2008).

[2] For example $\mathcal{D}_i$ might be a vector of contract offers, with each contract consisting of a fixed fee and a price per unit bought (a two-part tariff). If a contract with one buyer precludes a contract with another, as in exclusive deals which ensure a single vendor per market, $\mathcal{D}_i$ becomes a proper subset of the product space of all possible two part tariffs.

The functions π and s_i $(i = 1, \ldots, n)$, and the joint probability distribution for $(\mathcal{J}_i, \mathbf{Z}_i)_{i=1,\ldots,n}$, are basic elements of the game. So the expectation operator introduced below (i.e., $\mathcal{E}(\cdot)$) is with respect to this joint distribution,[3] and the observed decisions $\mathbf{d}_i$ are generated by these strategies and information sets.

We begin with an assumption that characterizes the behavior of agents in the game.

Assumption 1 (Best Response Condition) *If s_i is the strategy played by agent i*

$$sup_{d \in \mathcal{D}_i} \mathcal{E}[\pi(d, \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i, \mathbf{d}_i = d] \leq \mathcal{E}[\pi(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)], \quad (a.s.\ \mathcal{J}_i),$$

for $i = 1, \ldots n$. ♠

In single agent problems, this assumption would simply be derived from optimizing behavior. For instance, with $n = 1$ and $\mathcal{D}_i$ a finite set, Assumption 1 is an implication of a standard discrete choice problem. If $\mathcal{D}_i$ is an interval, then Assumption 1 generates the standard first order (or Kuhn-Tucker complementarity) conditions for optimal choice of a continuous control. When there are multiple interacting agents, Assumption 1 is a necessary condition for any Bayes-Nash equilibrium. It does not rule out multiple equilibria, and it does not assume anything about the selection mechanism used when there are multiple equilibria.

Note that Assumption 1 does not put any restrictions on the information structure of the game, which is an aspect of the problem that the econometrician typically has little information on. In particular we will not have to specify whether (d_{-i}, z_{-i}) is in agent i's information set, $\mathcal{J}_i$, at the time decisions are made. So, $\mathcal{J}_i$, could contain their values, could contain a signal on their likely values, or may not contain any information on their values at all.

Finally note that at the cost of increased notational complexity we could have used a weaker version of Assumption 1. More specifically, Assumption 1 will typically be used to form moment inequalities with particular alternative decisions considered by the econometrician. Formally, we could only require Assumption 1 to hold for this subset of alternative decisions. For example, if we limited the econometrician's choice to $\mathcal{D}(d_i) = \{d : |d - d_i| \geq \alpha d_i, d \in \mathcal{D}_i\} \subset \mathcal{D}_i$ for a known $\alpha > 0$, we would allow agents to take actions which were not optimal but within α percent of the optimal decision. Similarly we could have specified that expected profits from alternative possible decisions were less that $(1 + \delta)$ times expected profits from the decision taken (for a known $\delta > 0$), allowing agents to make decisions which led to expected profits which were close to, but not necessarily equal to, maximal expected profits.

[3] We could have defined the expectation operator that results from the agents' perceptions, and then put constraints on the relationship between the agents' perceptions and the expectation operator emanating from the data generating process. Though correct perceptions are certainly sufficient for Assumption 1 to be true, they are not necessary; see Pakes (forthcoming), and the literature cited there, for further discussion.

Counterfactuals

To use Assumption 1, we want to be able to compare the profits actually earned to those that would have been earned had the agent made a different decision. Under the next assumption, the distribution of i's counterfactual profits can be obtained by simply changing the decision made by agent i and evaluating $\pi(\cdot)$ at the alternative decision.

Assumption 2 (Counterfactual Condition) *The distribution of* $(\mathbf{d}_{-i}, \mathbf{z}_i)$ *conditional on* $\mathcal{J}_i$ *and* $\mathbf{d}_i = d$ *does not depend on* d.

In either single agent problems, or multiple agent problems with simultaneous moves, conditional independence of other agents' decisions (of $\mathbf{d}_{-i}$) from $\mathbf{d}_i$ is an assumption of the model.[4] So then the main restriction in this assumption is that $\mathbf{z}_i$ does not depend on $\mathbf{d}_i$. In many examples the profit function is naturally written in a way that depends on additional variables that, in turn, depend on the decisions of *all* agents. For example, our empirical study analyzes the number of ATMs chosen by banks. The profits a bank earns from its ATM investments depend on the equilibrium interest rates in the periods in which those ATM's will be operative. So the profit function would naturally be written as a function of interest rates, the number of ATM's, and other (exogenous) variables. The interest rates, in turn, depend on the number of ATM's installed by the bank and its competitors. Assumption 2 states that to use a comparison between the profits actually earned to those that would have been earned had the bank chosen a different number of ATMs we will need to compute the interest rate that would result from the counterfactual ATM choice, and then use that calculaton to express the profit difference in terms of (d', d_{-i}) and variables which do not change with a different choice of the number of ATMs.

More formally, suppose a profit function $\tilde{\pi}$ is a function of decisions $(\mathbf{d}_i, \mathbf{d}_{-i})$ and potentially endogenous additional variables $\mathbf{y}_i$ (endogenous in the sense that the realization of $\mathbf{y}_i$ depends upon the realizations of $\mathbf{d}_i$, thus violating Assumption 2). Then, to satisfy Assumption 2, one will typically need a model of how $\mathbf{y}_i$ is related to agents' decisions. That is, we need a function y such that $\mathbf{y}_i = y(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i)$ for a random variable $\mathbf{z}_i$ satisfying Assumption 2. Then the $\pi(d, d_-, z)$ which appears in Assumption 2 is constructed as $\tilde{\pi}(d, d_-, y(d, d_-, z))$. We note that if there is not a one to one map between $\mathbf{y}_i$ and $\mathbf{d}_i$ conditional on $(\mathbf{d}_{-i}, \mathbf{z}_i)$ (or if the researcher is not sure of what that map is), but the researcher can construct a lower bound to the counterfactual profits that the agent could make, the researcher could replace the counterfactual profits in Assumption 2 with that lower bound.

[4] If, in non-simultaneous move games, we were to construct counterfactuals for agents who move early, Assumption 2 would generally be violated because the decisions of those who move later would be components of $\mathbf{d}_{-i}$, and their distribution, conditional on $(\mathcal{J}_i, d_i = d)$, would typically depend on d. To derive counterfactuals for that case we would either need a model for how the agent making the early period decision believes a change in that decision is likely to change the behavior of agents who move later, or we would need to compute the later agent's response that minimizes the profits earned from the counterfactual by the agent who moves early (for examples, see Ho, 2009, Pakes, 2010, and Crawford and Yorukoglu, forthcoming).

Assumption 2 implies that Assumption 1 can be rewritten without conditioning on different d. So if we define

$$\Delta\pi(d, d', d_{-i}, z_i) = \pi(d, d_{-i}, z_i) - \pi(d', d_{-i}, z_i)$$

and recall that $\mathbf{d}_i = s_i(\mathcal{J}_i)$, then Assumptions 1 and 2 imply that for any $d' \in \mathcal{D}_i$

$$\mathcal{E}[\Delta\pi(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i] \geq 0. \tag{1}$$

2.2 Econometrician's Problem

We assume that the econometrician has access to a parametric function r that can be used to approximate the payoff function π. The function $r(\cdot)$ has arguments d_i, d_{-i}, an *observable* vector of the determinants of profits, say z_i^o, and θ. The parameter $\theta \in \Theta$ has an unknown true value, say θ_0, which is the parameter vector of interest. We let $\mathbf{z}_i^o$ be the random variable whose realizations are given by z_i^o. For convenience, we assume $\mathbf{z}_i^o \subset \mathbf{z}_i$ so that Assumption 2 applies to $\mathbf{z}_i^o$ as well. We obtain our approximation to $\Delta\pi(d, d', z)$, which we label $\Delta r(d, d', z^o, \theta)$, by evaluating $r(\cdot)$ at d and d' and taking the difference, so that $\Delta r : \mathcal{D}_i^2 \times \mathcal{D}_{-i} \times Z^o \times \Theta \to \mathcal{R}$. The relationships between $\Delta\pi(\cdot)$ and $\Delta r(\cdot)$ and $\mathbf{z}_i$ and $\mathbf{z}_i^o$ define the following unobservables.

Definitions. For $i = 1, \ldots, n$, and $(d, d') \in \mathcal{D}_i^2$ define

$$\nu_{2,i,d,d'} = \mathcal{E}[\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i] - \mathcal{E}[\Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i], \text{ and} \tag{2}$$

$$\nu_{1,i,d,d'} = \nu_{1,i,d,d'}^{\pi} - \nu_{1,i,d,d'}^{r} \tag{3}$$

where

$$\nu_{1,i,d,d'}^{\pi} = \Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) - \mathcal{E}[\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i], \text{ and}$$

$$\nu_{1,i,d,d'}^{r} = \Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) - \mathcal{E}[\Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i].$$

It follows that

$$\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) = \Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) + \nu_{1,i,d,d'} + \nu_{2,i,d,d'}. \tag{4}$$

The function $\Delta r(\cdot, \theta)$ is the observable measure of the change in profits that would result from a change of $d_i = d$ to $d_i = d'$. The random variables ν_1 and ν_2 are the determinants of the true profit difference that are *not observed* by the econometrician. Different versions of these random variables are defined for every different (d, d') and every agent. We distinguish between two types of unobservables (i.e. ν_1 and ν_2) because the difference in their properties has important implications for alternative estimators.

The unobservables ν_1 and ν_2 differ in what the agent (in contrast to the econometrician) knows about them. While the agent "knows" its ν_2 values *before* it makes its decision ($\nu_{2,i} \in \mathcal{J}_i$), realizations of $\nu_{1,i}$ do not change the agent's expected profits at the time decisions are made. Since the decision depends on the information set, $\mathbf{d}_i = s_i(\mathcal{J}_i)$, $\mathbf{d}_i$ can depend on the values of $\nu_{2,i}$. Consequently the distribution across agents of $\nu_{2,i,d,d'}$ (for a fixed d) and $\nu_{2,i,\mathbf{d}_i,d'}$ (for the $\mathbf{d}_i$ chosen) can differ. In particular even if the former is mean zero, the latter will generally not be. In contrast $\mathcal{E}[\nu_{1,i,d,d'}|\mathcal{J}_i] = 0$ by construction, and since $\mathbf{d}_i$ is a function of the variables in $\mathcal{J}_i$, $\nu_{1,i,\mathbf{d}_i,d'}$ is mean independent of $\mathbf{d}_i$. So the mean of $\nu_{1,i,\mathbf{d}_i,d'}$ across agents is also zero.

The importance of accounting for one or both of (ν_1, ν_2) is likely to be different in different applied problems. The unobservable ν_1 realizations can arise from either expectational errors or measurement errors. There are two sources of expectational errors: (i) incomplete information on the environmental variables (the $\mathbf{z}_i$ or $\mathbf{z}_i^o$); and (ii) asymmetric information resulting from incomplete information on $\mathbf{d}_{-i}$. The measurement error in profits $(\nu_{1,i}^{\pi} - \nu_{1,i}^{r})$ can result from either measurement error in the observables that go into the profit function or from specification error in $r(\cdot)$ per se.

In contrast, ν_2 is a "structural" disturbance, i.e. a source of variance in the difference in profits that the agent conditions its decisions on, but that the econometrician does not observe. Variation in ν_2 will be important when $\Delta r(d, d', \cdot)$ does not account for an important source of variation in $\mathcal{E}[\Delta\pi(d, d', \cdot)|\mathcal{J}_i]$ that the agent accounts for when it makes its decision. The examples below clarify how this can happen.[5]

Selection

The assumptions thus far are not very stringent. In addition to not assuming what each agent knows about either its competitors or $\mathbf{z_i}$, we *have not* specified a particular form for the distribution of either ν_1 or ν_2, and we *have* allowed for discrete choice sets and regressors that are correlated with the unobserved $\nu_{2,i}$. We do, however, require an additional assumption. This assumption is due to the fact that $\mathbf{d}_i$ is both a determinant of profits and is, in part, potentially determined by an unobservable determinant of profits (the $\nu_{2,i}$). This implies that the $\nu_{2,i}$'s that correspond to the observed decisions are a selected subset of the possible values of the $\nu_{2,i}$'s. More formally, equation (4) implies that s_i is a strategy satisfying Assumption 1 only if

$$\mathcal{E}[\Delta r(s_i(\mathcal{J}_i), d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i] \geq -\nu_{2,i,s_i(\mathcal{J}_i),d'}. \tag{5}$$

[5]Notice that equation (4) nests the familiar model in which $\mathbf{z}_i^o = \mathbf{z}_i$, and $\Delta\pi(\cdot)$ consists of a primitive function (our $\Delta r(\cdot)$) and an additively separable disturbance. Then the $\mathbf{z}_i$ are the observed determinants of profits, and the $\nu_{2,i}$ and $\nu_{1,i}$ are, respectively, the unobserved determinants of profits that the agent knows, and does not know, when it makes its decisions. The only change that is required to allow for measurement error in a z_i that enters the model linearly is to increase the sources of ν_1. Note, however, that if $\Delta\pi(\cdot)$ was a non-linear function of an unobservable then this interpretation would fail, and if we wanted to derive $\Delta r(\cdot)$ from $\Delta\pi(\cdot)$ we would require more assumptions.

That is, draws on $\nu_{2,i}$ corresponding to the observed decisions are *selected* from a subset of the support of the ν_2 distribution.

The econometrician only has access to $\Delta r(\cdot, \theta)$ and Assumption 1 is in terms of the conditional expectation of $\Delta\pi(\cdot)$. So, to use Assumption 1 to restrict the observed data we will need to impose restrictions on the distribution of the $\nu_{2,i}$. The next assumption provides such a restriction. It considers certain weighted averages of $\Delta r(\mathbf{d}_i, d', \cdot; \theta)$ across values of d' and agents i. The weights are allowed to be nonnegative functions of observable components of the information sets of the agents. The condition will then imply that the weighted average of $\Delta r(\cdot, \theta_0)$ will have nonnegative mean as shown in section 2.3 below.

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 (on d'), on the information set $\mathcal{J}_i$ (which determines $\mathbf{d_i}$), and (possibly) on some observable component of the other agents' information sets, $\mathbf{x}_{-i} \subset \times_{j \neq i} \mathcal{J}_j$. Assume that*

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

and

$$\mathcal{E}[\sum_{i=1}^{n} \sum_{d' \in \mathcal{D}_i} h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}) \nu^r_{1,i,\mathbf{d}_i,d'}] \geq 0.$$

□

Assumption 3 does not require us to specify particular distributions for ν_1 and/or ν_2, the contents of agents' information sets, or the nature of the agent's choice set. In particular since both the choice set and the distributions of the unobservables are unspecified, Assumption 3 allows us to analyze models with discrete choice sets and regressors that are correlated with $\nu_{2,i}$ without making particular distributional assumptions. On the other hand, particular distributional assumptions can sometimes generate Assumption 3 in models where it would not hold without them. The examples below elaborate on both these points, and in addition, show that our combination of assumptions can help determine difficult to estimate parameters of continuous choice problems.

To see the difficulty in satisfying the condition on ν_2 in Assumption 3 recall that the observed strategy $\mathbf{d}_i = s_i(\mathcal{J}_i)$ must satisfy (5) which favors larger, more positive values of $\nu_{2,i,\mathbf{d}_i,d'}$. Assumption 3 requires that the $\nu_{2,i}$ associated with the observed decision and counterfactual d' is not positively correlated with the $h^i(d', \mathbf{d}_i, \mathcal{J}_i)$ function. We are, however, helped by the fact that both the counterfactual choice (d') and the $h(\cdot)$ function (our "instrument") can vary both with d_i and with individual characteristics.

Notice that the condition on ν^r_1 in Assumption 3 is trivially satisfied in two important cases: 1) the weight function for agent i does not depend on $\mathbf{x}_{-i}$; and 2) $\nu^r_{1,i,\mathbf{d}_i,d'}$ is mean independent of $\cup_j \mathcal{J}_j$. When the weight function for i does not depend on variables outside of $\mathcal{J}_i$, then it can be written as $h^i(d'; \mathbf{d}_i, \mathcal{J}_i)$ and, since we have constructed $\nu^r_{1,i,\mathbf{d}_i,d'}$ so that

$\mathcal{E}[\nu^r_{1,i,\mathbf{d}_i,d'}|\mathcal{J}_i] = 0$, the condition on ν^r_1 in Assumption 3 is satisfied. The assumption that $\nu^r_{1,i,\mathbf{d}_i,d'}$ is mean independent of $\cup_j \mathcal{J}_j$ is natural when the $\nu^r_{1,i,\mathbf{d}_i,d'}$ part of the disturbance represents measurement error as in example 2 below,[6] or when the game is a symmetric information game as then $\mathcal{J}_i = \cup_j \mathcal{J}_j$ (see the discussion in example 4). Typically the ν_2 condition in Assumption 3 is satisfied by finding some combination of: a) an observable instrument in $\mathcal{J}_i$ that is uncorrelated (or negatively correlated) with $\nu_{2,i}$; and/or b) restrictions on $\nu_{2,i,\mathbf{d}_i,d'}$ for carefully chosen values of d'. Examples 1 and 3 below illustrate with two familiar models. Example 4 provides a case where distributional assumptions enable one to satisfy Assumption 3.

The condition on ν^r_1 in Assumption 3 is nontrivial when the weight function depends on observables $\mathbf{x}_{-i}$ from other agents that are not a part of i's information set but might be a determinant of agent i's realized profits. An example where this might occur is in games with asymmetric information where the weight function depends on a competitor's control. We allow the weight function to depend on $\mathbf{x}_{-i}$ because it provides additional flexibility; in particular it enables us to satisfy the ν_2 condition by restricting the relationship between the $\nu_{2,i,\mathbf{d}_i,d'}$ of different agents (see examples 2 and 4 below). This flexibility comes at the cost of the additional requirement on ν^r_1.

2.3 Inequality Conditions

Recall that the data we observe for agent i will be based on his strategy s_i that satisfies Assumption 1. So realized decisions for agent i will be determined by s_i, i.e. $\mathbf{d}_i = s_i(\mathcal{J}_i)$. Accordingly equation (4) implies that

$$\begin{aligned}\Delta r(d,d',\mathbf{d}_{-i},\mathbf{z}^o_i,\theta_0) &= \Delta\pi(d,d',\mathbf{d}_{-i},\mathbf{z}_i) - \nu^\pi_{1,i,d,d'} + \nu^r_{1,i,d,d'} - \nu_{2,i,d,d'} \\ &= \mathcal{E}[\Delta\pi(d,d',\mathbf{d}_{-i},\mathbf{z}_i)|\mathcal{J}_i] + \nu^r_{1,i,d,d'} - \nu_{2,i,d,d'}.\end{aligned}$$

Then,

$$\begin{aligned}&\mathcal{E}\left[\sum_{i=1}^n \sum_{d'\in\mathcal{D}_i} h^i(d';\mathbf{d}_i,\mathcal{J}_i,\mathbf{x}_{-i})\Delta r(\mathbf{d}_i,d',\mathbf{d}_{-i},\mathbf{z}^o_i,\theta_0)\right] \\ &= \mathcal{E}\left[\sum_{i=1}^n \sum_{d'\in\mathcal{D}_i} h^i(d';\mathbf{d}_i,\mathcal{J}_i,\mathbf{x}_{-i})\mathcal{E}[\Delta\pi(\mathbf{d}_i,d',\mathbf{d}_{-i},\mathbf{z}_i)|\mathcal{J}_i]\right] \qquad (6) \\ &\quad + \mathcal{E}\left[\sum_{i=1}^n \sum_{d'\in\mathcal{D}_i} h^i(d';\mathbf{d}_i,\mathcal{J}_i,\mathbf{x}_{-i})\nu^r_{1,i,\mathbf{d}_i,d'}\right] - \mathcal{E}\left[\sum_{i=1}^n \sum_{d'\in\mathcal{D}_i} h^i(d';\mathbf{d}_i,\mathcal{J}_i,\mathbf{x}_{-i})\nu_{2,i,\mathbf{d}_i,d'}\right]\end{aligned}$$

[6] A third case, which can be thought of as a special case of the second, that sometimes occurs is $\nu^r_{1,i,\mathbf{d}_i,d'} \equiv 0$. This happens when the function r depends only on observables in $\mathcal{J}_i$. For instance, in single agent problems, $\mathbf{d}_{-i}$ does not enter profits, and so $\nu^r_{1,i,\mathbf{d}_i,d'} = 0$ if $\mathbf{z}^o_i$ is in agent i's information set (there is no uncertainty). Since $\nu^r_{1,i,\mathbf{d}_i,d'}$ is identically zero, the second part of Assumption 3 is satisfied for any weight function, so that function need only satisfy the first part of Assumption 3.

We consider each of the three terms following the equality in equation (6) in turn. Since $\mathcal{E}[\Delta\pi(s_i(\mathcal{J}_i), d', \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i] \geq 0$ by Assumptions 1 and 2, each term in the first summand is nonnegative by the assumed nonnegativity of the weights $h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i})$.

As noted before, the definition of ν_1^r in equation (3) yields $\mathcal{E}[\nu^r_{1,i,\mathbf{d}_i,d'}|\mathcal{J}_i] = 0$. So, when the weight function does not depend on $\mathbf{x}_{-i}$, the summation over ν_1^r terms in equation (6) is zero. More generally, Assumption 3 states that the last two terms in equation (6) are non-negative. As a result

$$\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i(d'; \mathbf{d}_i, \mathcal{J}_i)\Delta r(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)\right] \geq 0. \tag{7}$$

Equation (11) depends only on observables and θ_0, so we can form its sample analog and look for values of θ that satisfy it.[7]

2.4 Examples

The examples in this section show how conditions in specific applications lead to Assumptions 1, 2, and 3 holding. We especially emphasize conditions for satisfying Assumption 3.

Example 1. Suppose that for each $d_i \in \mathcal{D}_i$ there is some $d'(d_i)$ such that $\Delta\pi(\mathbf{d}_i, d'(\mathbf{d}_i), \mathbf{z}_i)$ is observable up to a parameter vector of interest and an error which is mean zero conditional on the agent's information set (so $\nu_{2,i,\mathbf{d}_i,d'(\mathbf{d}_i)}$ is zero). Then Assumption 3 is satisfied with $h(d'(d_i); d_i, \mathcal{J}_i, \mathbf{x}_{-i}) = h(d'(d_i); d_i, \mathcal{J}_i) > 0$, and $h(\cdot)$ zero elsewhere. This is a familiar special case as it implies that

$$\Delta\pi(d_i, d'(d_i), \mathbf{d}_{-i}, \mathbf{z}_i) = \Delta r(d_i, d'(d_i), \mathbf{d}_{-i}, \mathbf{z}_i, \theta_0) \ + \ \nu_{1,i,d_i,d'},$$

so our assumptions on the disturbance are the analog of those used by Hansen and Singleton (1982) for their first order condition estimator. However our estimator: (i) allows for more general (discrete and/or bounded) choice sets, and (ii) allows explicitly for interacting agents (without having to fully specify the information structure). We now provide a discrete and then a continuous control example that show why these extensions might be useful.

For the discrete control example we show that these assumptions enable us to apply Euler's perturbation method to the analysis of single agent dynamic discrete choice problems. This simplifies the analysis of those problems dramatically. For specificity consider a discrete choice model with switching costs so that the observable model for the agent's profits (or utility) in a given period, say $U(d_{i,t}, d_{i,t-1}, z_{i,t}, \theta)$, depends on the decision in the prior period. For simplicity we drop the i subscript in this discussion and assume there are just two choices, $\mathcal{D} = \{0, 1\}$. The strategy for the agent is a sequence of

[7] In general Assumptions 1, 2, and 3 are sufficient but not necessary for the inequalities in (11), which, in turn, provide the basis for estimation and inference. That is, we expect that there are alternative conditions that will also suffice.

functions, say $\{s_t(\cdot)\}_t$ mapping $\mathcal{J}_t$ into $\{0,1\}$. The observed value generated by the strategy is $r(\cdot) = \sum_t \beta^t U(s_t(\mathcal{J}_t), d_{t-1}, z_t, \theta_0)$. Consider a one-period "perturbation" setting $d'(d_0) = (1-d_0)$. For periods $t > 0$ the strategy function is unchanged so $\forall t > 0,\ d'_t = d_t$, while $U(s_t(\mathcal{J}_t), d_{t-1}, z_t, \theta_0)$ is unchanged $\forall t > 1$. Then the assumptions imply

$$0 \leq \mathcal{E}[\Delta r(\mathbf{d}_0, 1-\mathbf{d}_0, \mathbf{z}_0, \theta_0)|\mathcal{J}_0)] \equiv$$

$$\mathcal{E}[U(\mathbf{d}_0, \mathbf{d}_{-1}, \mathbf{z}_0; \theta_0) - U(1-\mathbf{d}_0, \mathbf{d}_{-1}, \mathbf{z}_0; \theta_0) + \beta(U(s_1(\mathcal{J}_1), \mathbf{d}_0, \mathbf{z}_1; \theta_0) - U(s_1(\mathcal{J}_1), 1-\mathbf{d}_0, \mathbf{z}_1; \theta_0))|\mathcal{J}_0]$$

at $\theta = \theta_0$. Similarly if $x \in \mathcal{J}_0$ and $h(\cdot)$ is positive valued, $\mathcal{E}[\Delta r(\mathbf{d}_0, 1-\mathbf{d}_0, \mathbf{z}_0; \theta_0) h(\mathbf{x}_0)] \geq 0$. Notice that if we base estimation off of such inequalities we obtain our estimates without ever computing the fixed point which defines the value function, thus circumventing the need for a nested fixed point algorithm, and, as a result, significantly increasing our ability to analyze richer models (see, for example, Morales 2011).[8]

For a continuous choice example we consider a uniform-price electricity auction (see Wolak 2001). Each day multiple generators bid in supply functions. An independent system operator sums those functions horizontally, intersects the sum with the hourly demand curve to determine the price, and directs generators to produce the quantity they bid at that price. The econometrican knows the bids each agent makes and the market clearing rule and so can compute the returns an agent would have made had it submitted a different bid function (holding other agents' bids and demand fixed).

The revenues of the generator i equal the equilibrium price, $p(d_i, d_{-i}, D_t) \equiv p(d, D_t)$, where d_i is i's bid function and D_t is the demand function in hour t, times the quantity it produces ($q_{i,t}(d, D_t)$). Its hourly costs are the variable costs of producing that quantity ($vc(q_{i,t}(d, D_t), \theta_1)$) plus the startup cost θ_2 of bringing the generator up when it had been inoperative in the previous hour ($\theta_2\{q_{i,t}(d, D_t) > 0\}\{q_{i,t-1}(d, D_t) = 0\}$). For simplicity we assume there are no shutdown costs and that the generator will go down for maintenance at the end of the day. Then today's bid influences only current profits, and our measure of those profits is

$$r(d_i, d_{-i}, z_i; \theta) =$$

$$\sum_t p(d, D_t) q_{i,t}(d, D_t) - \sum_t vc(q_{i,t}(d, D_t), \theta_1) - \theta_2[\sum_t \{q_{i,t}(d, D_t) > 0\}\{q_{i,t-1}(d, D_t) = 0\}].$$

This is a simultaneous move game. So if $r(\cdot)$ were differentiable in d_i and agents were maximizing expected profits we could differentiate the profit function with respect to a continuous determinant of d_i and average those derivatives to form a moment which, provided the usual regularity conditions hold, has expectation zero at the true θ_0. In fact the derivative of the term $\theta_2[\sum_t \{q_{i,t}(d, D_t) > 0\}\{q_{i,t-1}(d, D_t) = 0\}]$ will be zero almost

[8] We emphasize that this does assume that $\Delta\nu_{2,d,d',t} = 0$. To relax the $\Delta\nu_{2,d,d',t} = 0$ assumption and still derive moment inequalities that circumvent the need for a nested fixed point we would need to restrict the distribution of $\{\nu_{2,\cdot}\}$, though not necessarily to a parametric family (see Example 4 below). On the other hand, in contrast to the literature which uses nested fixed points to analyze this problem, we do allow for measurement and/or specification error (our $\{\nu_{1,\cdot}\}$), and we do not need to specify how the agent forms its perceptions on the evolution of $\mathbf{z_i}$.

everywhere (and undefined elsewhere). So the sample first order conditions would not contain information on θ_2, the parameter we are often most interested in. However if we evaluated counterfactual bid functions which differed from the actual bid by enough to induce changes in the periods in which the generator was called, the difference in returns between the counterfactual and actual bid, together with the Assumptions above, would produce inequalities which could be used for inference on θ_2.[9]

Two final points on the assumptions underlying this example. First its assumptions cover the familiar case where there is a ν_2 but we can find a "control function" that allows us to condition on its value. Second, as can be seen from our examples, they often generate inequalities which are extremely easy to use. So we might want to use moment inequality techniques in conjunction with more computationally intensive procedures either as a check for the robustness of different assumptions or as starting values for the more computationally intensive techniques.

Example 2. This example formalizes conditions that suffice for inequality estimators based on "matched observations," and since each of our "observations" consists of a difference between an observed and a counterfactual choice, these conditions also suffice for inequality estimators based on differences-in-differences. For specificity we describe these conditions in the context of an analysis of the determinants of hospital choice (for more detail on this problem, see Ho and Pakes 2011).

Hospital choices depend on a complex interaction between the patient's health condition and the quality of the hospital's services for that health condition, as well as other hospital and patient specific characteristics. The role of the non-health related patient characteristics, the health plan, and the hospital characteristics in hospital choice have important policy implications (e.g. the interaction between the prices at the hospital a doctor sends its patient to and the incentives the patient's health plan gives the doctor). However, the analysis of those effects must contend with the confounding effects of the patient-health/hospital -quality interaction induced by the endogeneity of the hospital choice decision. The health condition of the patient is recorded in some detail when the patient enters the hospital, but selection and other problems make it difficult to measure hospital quality in treating the patient's condition.

Assume the expected utility which determines hospital choice (d_i) is additively separable in two components: 1) a non-health related function which is known up to a parameter vector of interest; and 2) an unobserved expected health outcome which depends on the patient's initial health condition and hospital characteristics. The non-health related component depends on patient and hospital characteristics to be denoted by z_i (distance from home to hospital, hospital prices, patient's insurance plan,...). The unobserved expected health outcome of patient i at hospital d is denoted by $\nu_{2,i,d}$. If we allow $r(d, z_i^o, \theta)$ to

[9]Reguant (2011) estimates parameters of this type for the Spanish electric utility market by using a computationally intensive simulation to approximate firms' perceptions of the distributions of competitors' bids.

measure the non-hospital component up to a conditional mean zero error[10], we have

$$\pi(d, z_i; \theta_0) = r(d, z_i^o, \theta_0) + \nu_{2,i,d} - \nu_{1,i,d}^r,$$

with decisions made to maximize $\mathcal{E}(\pi(d, \mathbf{z}_i)|\mathcal{J}_i) = \mathcal{E}[r(d, \mathbf{z}_i^o; \theta_0)|\mathcal{J}_i] + \nu_{2,i,d}$. Since $\nu_{1,i,d}^r$ is caused by measurement (or specification) error we assume that $\forall(i,d),\ \mathcal{E}[\nu_{1,i,d}^r|\cup_j \mathcal{J}_j] = 0$.

There are two points about this decision problem that should be noticed. First we assume that hospital choice is a single agent problem in the sense that the decisions of other patients (d_{-i}) does not effect i's choice. Second we have not put any restriction on the interaction between the patient health condition and hospital choice ($\nu_{2,i,d}$) other than its additive separability from $r(\cdot)$. We now show that Assumption 3 can be satisfied by assuming that the $\nu_{2,i,d}$ depend only on the patient's observed initial health condition and the hospital choice. That is, if $\mathbf{q}_i$ gives us the initial health condition of individual i, then we can write $\nu_{2,i,d}$ as $\nu_{2,\mathbf{q}_i,d}$ $(\forall(i,d))$.

We "match" patient i, with a patient, say j, with the same initial condition, *i.e.* $\mathbf{q}_j = \mathbf{q}_i$, but who chose a different hospital, so $d_i \neq d_j$. This match yields two counterfactuals. For patient i consider the counterfactual of choosing hospital d_j, and for patient j consider hospital choice d_i. The first counterfactual yields a difference in expected utility of

$$\mathcal{E}[\Delta\pi(d_i, d_j, \mathbf{z}_i)|\mathcal{J}_i] = \mathcal{E}[\Delta r(d_i, d_j, \mathbf{z}_i^o; \theta_0)|\mathcal{J}_i] + \nu_{2,\mathbf{q}_i,d_i,d_j}$$

where $\nu_{2,\mathbf{q}_i,d_i,d_j} = \nu_{2,\mathbf{q}_i,d_i} - \nu_{2,\mathbf{q}_i,d_j}$. The second counterfactual yields the same difference with the i and j indices reversed. Since $\mathbf{q}_i = \mathbf{q}_j$, we have $\nu_{2,\mathbf{q}_i,d_i,d_j} = -\nu_{2,\mathbf{q}_j,d_j,d_i}$. So, when we sum the difference in utilities from these two individuals and their counterfactuals, the ν_2 component of utility will be eliminated, i.e.

$$\mathcal{E}[\Delta\pi(d_i, d_j, \mathbf{z}_i)|\mathcal{J}_i] + \mathcal{E}[\Delta\pi(d_j, d_i, \mathbf{z}_j)|\mathcal{J}_j] = \mathcal{E}[\Delta r(d_i, d_j, \mathbf{z}_i^o; \theta_0)|\mathcal{J}_i] + \mathcal{E}[\Delta r(d_j, d_i, \mathbf{z}_j^o; \theta_0)|\mathcal{J}_j].$$

The elimination of the ν_2 component in this sum of expected profit increments implies that the first part of Assumption 3 holds with "matched pair" weight functions. More formally we choose the weight functions to pick out all possible matched pairs; or $x_{-i} = \times_{j\neq i}(\mathbf{d}_j, \mathbf{q}_j)$, and $h^i(d', \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}) = \sum_{j\neq i} \mathbf{1}\{\mathbf{q}_i = \mathbf{q}_j\}\mathbf{1}\{\mathbf{d}_i \neq \mathbf{d}_j\}\mathbf{1}\{d' = \mathbf{d}_j\}$. With a similar choice for h^j, all matched pairs with identical health conditions and different choices of hospital will be summed. Since the sum of the ν_2's for each matched pair will be zero

$$\sum_{i=1}^{n} \sum_{d'\in\mathcal{D}_i} h^i(d', \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i})\nu_{2,i,\mathbf{d}_i,d'} = 0,$$

which satisfies the first part of Assumption 3, and since

$$\mathcal{E}[\sum_{i=1}^{n} \sum_{d'\in\mathcal{D}_i} h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i})\nu_{1,i,\mathbf{d}_i,d'}^r] = 0$$

[10]This could be caused either by mis-specification or measurement error, though Ho and Pakes (2011) are particularly worried about measurement errors in their price and distance measures.

by virtue of the assumption that $\forall(i,d),\ \mathcal{E}[\nu^r_{1,i,d}|\cup_j \mathcal{J}_j] = 0$, the second part of Assumption 3 is satisfied also. It follows that for the matched pairs weight function

$$\mathcal{E}\left[\sum_{i=1}^{n}\sum_{d'\in\mathcal{D}_i} h^i(d',\mathbf{d}_i,\mathcal{J}_i,\mathbf{x}_{-i})\Delta r(\mathbf{d}_i,d',\mathbf{z}^o_i,\theta_0)\right] \geq 0.$$

The example above does not have interacting agents, but the matched pair approach can enable one to mitigate estimation problems in multiple agents settings as well.[11] However, as we now show, when one agent's returns depend on another agent's decision to satisfy Assumption 3 we may need to impose constraints on the information structure of the game. For a familiar example consider an Industrial Organization model in which there is a market-specific component of costs or demand that is known to the agents at the time decisions are made, but not to the econometrician; say an entry model with a market-specific sunk cost of entering (for the role of market specific unobservables in the entry literature see Berry and Reiss, 2007). We take matched pairs consisting of potential entrants who did, and did not, enter a specific location, and counterfactuals which are the opposite of the decision the agent made. If we sum the difference in expected profits between the actual and counterfactual behavior of the pair, we would eliminate the market-specific unobservable (our ν_2) and the first part of Assumption 3 would be satisfied. Now however, the other agent's decisions would generally affect each agent's *realized* profits and there is a $\mathbf{x}_{-i}$ component of $h^i(\cdot)$ (the determinants of the entry decisions of the other agent). So if there were asymmetric information and one agent's decision depended on variables not in other agents' information set, then the ν^r_1 error of the second agent could be correlated with the weight assigned to the first agent's counterfactual and this might violate the second part of Assumption 3. Notice that if it was appropriate to assume a "symmetric information" equilibrium, as is often done in applied work, this difficulty would not arise (then $\mathbf{d}_{-i} \in \mathcal{J}_i$ so that the second part of Assumption 3 would follow from the definition of ν^r_1). It could still hold if there were asymmetric information, but additional assumptions would be needed.[12].

Example 3. This example shows that ordered choice models impose a restriction on the structural disturbance (our ν_2) that enables us to satisfy Assumption 3. This implies that we can analyze both the familiar single agent ordered choice model and multiple agent ordered choice models; and we can do so without parametric assumptions on the structural disturbance. We conclude by noting that the restriction used in this example

[11] There are also natural extensions to the more aggregate cases often used in Labor Economics and Public Finance; see for example, Card and Krueger 1994, who use pairs of similar states or regions to eliminate the influence of additive unobservables from before and after a policy intervention.

[12] For instance if there are only two agents, the only uncertainty is with respect to the competitor's decision, and we were willing to assume that an agent's returns to entry fell if the other agent entered, the covariance between both $\nu^r_{1,i,\mathbf{d}_i,d'}$ and $\nu^r_{1,j,\mathbf{d}_j,d'}$ and the weight function can be shown to be positive. In this case our matched pair estimator would satisfy the last part of Assumption 3 with a strict inequality.

is a special case of a more general restriction that is often appropriate. The next section provides an empirical example of a multiple agent ordered choice model and a Monte Carlo study of the performance of alternative estimators based on this empirical example.

Lumpy investment decisions (say in the number of stores or machines) are often treated as ordered choice problems. Our example has interacting firms each deciding on how many units of a machine to purchase and install in a number of independent markets.[13] Let d_i denote the number of units chosen by firm i. Then revenue, $R(d_i, d_{-i}, z_i^o)$ depends on the number of units chosen by i and the other firms in the market, as well as some other observed determinants, z_i^o. The marginal cost to firm i of an additional unit is $\theta_0 + \eta_i$, where θ_0 reflects the average marginal cost (across firms) of the machines, and η_i captures firm-level heterogeneity in costs known to each firm when it makes its decision but not observed by the econometrician. Thus the *unconditional* mean of η is zero, or $\mathcal{E}(\eta_i) = 0$.

Allowing for the uncertainty and measurement errors that generate ν_1, profits to firm $7 * 4.87 = i$ are: $\pi(d_i, \mathbf{d_{-i}}, \mathbf{z}_i) = R(d_i, \mathbf{d}_{-i}, \mathbf{z}_i^o) - d_i(\theta_0 + \eta_i) + \nu_{1,i,d_i}$. So profit differences from a counterfactual d' are

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

where $\Delta\pi$, ΔR, and $\nu_{1,i,d,d'}$ are defined as differences using notation as before. The observable (or econometrician's) approximation to the differenced profits is $\Delta r(d, d', \mathbf{d_{-i}}, \mathbf{z}_i^o; \theta) = \Delta R(d, d', \mathbf{d_{-i}}, \mathbf{z}_i^o) + (d' - d)\theta$. The structural disturbance is, $\nu_{2,i,d,d'} = (d' - d)\eta_i$.

We now show that the assumption that $\nu_{2,i,d,d'} = (d' - d)\eta_i$ enables us to choose weights that satisfy the first half of Assumption 3. Consider a counterfactual choice of $d' = d_i + t$; a fixed number of units (t) away from d_i. Then

$$\nu_{2,i,d_i,d_i+t} = \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+t} = t\eta_i, \quad \text{and,} \quad \mathcal{E}\nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+t} = t\mathcal{E}(\eta_i) = 0.$$

For now assume that the counterfactual $d' = \mathbf{d}_i + t$ is feasible for all i (the case where it may not be is considered in the next example). Taking $h(d'; \mathbf{d}_i, \mathcal{J}_i) = n^{-1}$ if $d' = \mathbf{d}_i + t$, and zero otherwise, we have

$$\mathcal{E}[\sum_{i=1}^{n} \sum_{d' \in \mathcal{D}_i} h(d'; \mathbf{d}_i, \mathcal{J}_i)\nu_{2,i,\mathbf{d}_i,d'}] = \mathcal{E}[n^{-1} \sum_{i=1}^{n} \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+t}] = n^{-1} \sum_{i=1}^{n} t\mathcal{E}(\eta_i) = 0.$$

The form of the structural disturbance and this choice of counterfactual allows the formation of averages that avoid the selection problem by including the structural error *no matter the choice* each agent made. The second half of Assumption 3 is automatically satisfied because the weight function depends only on $\mathbf{d}_i$ and not on variables for the other agents ($\mathbf{x}_{-i}$). So, $\mathcal{E}[h^i(d'; \mathbf{d}_i, \mathcal{J}_i)\nu^r_{1,i,\mathbf{d}_i,d'}] = \mathcal{E}[h^i(d'; \mathbf{d}_i, \mathcal{J}_i)\mathcal{E}(\nu^r_{1,i,\mathbf{d}_i,d'}|\mathcal{J}_i)] = 0$.

In the empirical example in the next section we focus mostly on counterfactuals with t set to 1 or -1. These counterfactuals lead to moment inequalities that form lower and

[13] For a single agent example of the use of moment inequalities for a lumpy investment problem in which there is dependence in the outcomes across markets, see Holmes, 2011.

upper bounds on the parameter θ_0. However additional weight functions can be formed from traditional instrumental variables. That is, if $\mathbf{x}_i \in \mathcal{J}_i$ and $\mathcal{E}\eta_i g(\mathbf{x}_i) = 0$ for some positive function g, then $h(d+t; d, \mathcal{J}_i) = g(\mathbf{x}_i)$ (and zero otherwise). These weights generate additional moment inequalities potentially containing more information on θ_0. The additional moments also enable us to get bounds on each parameter when we consider extensions which allow for richer marginal cost functions indexed by several parameters, as we do in the Monte Carlo example.

The ordered choice model is a particular example of a more general structure which we can analyze. We can form inequalites which satisfy Assumption 3 in any model in which we can always find a counterfactual which generates a difference between the actual and counterfactual choices that is a linear function of the structural error (regardless of the observed choice). The vertical discrete choice model used in Bresnahan (1987) and Katz (2007), and contracting models where the source of the structural error is a component of the transfers among agents that agents condition on but the econometrician does not observe (for details see Pakes, 2010), are two examples of other models which can be shown to satisfy this condition.

Before leaving this example we note an issue that can arise in using an inequality which does not condition on the choice agents make to satisfy Assumption 3. We do so in the context of the ordered choice model discussed above. To construct the (unconditional) moment used to estimate its parameter, it was not necessary to have the counterfactual be $d' = d_i + t$ for the same t for each i, as we could have divided the inequalities formed from the different observations by different t_i and, as long as all the t_i were positive, we would have still obtained a moment whose expectation was positive. However it is essential that there is a feasible counterfactual which has the *same sign* for the t chosen for each observation. If the choice set is bounded this is not always possible. For instance, in choosing how many machines to install, it is sometimes natural to bound the choice set below by zero. The boundary implies we can not find feasible counterfactuals with $d' = d_i - t$ when $d_i = 0$ for any $t > 0$, and if we drop the observations with $d_i = 0$ before we form our averages we generate another selection problem. The next example considers this case, and, in so doing, presents another set of assumptions leading to satisfaction of Assumption 3.

Example 4. The last example illustrated how the selection problem reappears in the inequality approach to the ordered choice problem when there is a boundary to the choice set which is chosen by some agents. A similar problem will occur when controls are continuous but bounded from one side (as in a tobit model, or in a bidding model where there is a cost or benefit which is known to the agent but unobserved to the econometrician and induces some agents to not bid). As we now show if, in these cases, it is reasonable to assume that the distribution of the ν_2 component is symmetric, or at least not skewed in the direction of truncation (see below), we can use the distribution of the ν_2 component in the untruncated tail to ensure we obtain a bound also in the direction of truncation. To do this, we show that ideas similar to those in Powell (1986), can be extended to obtain

moment inequalities that account for truncation.

For specificity, we go back to the example of the last section but assume that choices were bounded from below by zero and that some agents chose $d_i = 0$. As a result we need a truncation correction to obtain an upper bound for θ_0. We now show that an assumption of a symmetric distribution enables us to form a moment which satisfies Assumption 3 and delivers this bound. Let $L = \{i : \mathbf{d}_i > 0\}$, the set of firms that install a positive number of machines and are *not* on the boundary, and let n_L be the number of firms in L. It will be helpful to use standard order statistic notation, i.e. $\eta_{(1)} \leq \eta_{(2)} \leq \cdots \leq \eta_{(n)}$. Let $L_\eta = \{i : \eta_i \leq \eta_{(n_L)}\}$ and $U_\eta = \{i : \eta_i \geq \eta_{n_L+1}\}$. Similarly, let $\Delta R_i^+ = \Delta R(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{z}_i)$ and $\Delta R_{(1)}^+ \leq \Delta R_{(2)}^+ \leq \cdots \leq \Delta R_{(n)}^+$. Let $U_R = \{i : \Delta R_i^+ \geq \Delta R_{(n_L+1)}^+\}$. Sets L and U_R are observable to the econometrician, but sets L_η and U_η are not.

As in section 2.3, we want to obtain a moment inequality based on weighted averages of Δr. We start by looking at a particular weighted average that uses the observations where it is feasible to consider the counterfactual $d' = d_i - 1$ and specific observations with the counterfactual $d' = d_i + 1$. In particular we let

$$h^i(d'; \mathbf{d}_i, \mathcal{J}_i, \mathbf{x}_{-i}) = n^{-1}\Big[\ \mathbf{1}\{d' = \mathbf{d}_i - 1\}\mathbf{1}\{i \in L\} + \ \mathbf{1}\{d' = \mathbf{d}_i + 1\}\mathbf{1}\{i \in U_R\}\Big],$$

then $\sum_i \sum_{d' \in \mathcal{D}_i} h^i(\cdot)\Delta r(\cdot) =$

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in L} \Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) + \frac{1}{n}\sum_{i\in U_R} \Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) \\
&\quad \geq \quad \frac{1}{n}\sum_{i\in L} \Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) + \frac{1}{n}\sum_{i\in U_\eta} \Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) \\
&\quad = \quad \frac{1}{n}\sum_{i\in L} \left\{E[\Delta\pi(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i] - \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i-1} + \nu^r_{1,i,\mathbf{d}_i,\mathbf{d}_i-1}\right\} \\
&\qquad + \frac{1}{n}\sum_{i\in U_\eta} \left\{E[\Delta\pi(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i)|\mathcal{J}_i] - \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+1} + \nu^r_{1,i,\mathbf{d}_i,\mathbf{d}_i+1}\right\} \\
&\quad \geq \quad -\frac{1}{n}\left\{\sum_{i\in L} \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i-1} + \sum_{i\in U_\eta} \nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+1}\right\} \ + \ \frac{1}{n}\left\{\sum_{i\in L} \nu^r_{1,i,\mathbf{d}_i,\mathbf{d}_i-1} + \sum_{i\in U_\eta} \nu^r_{1,i,\mathbf{d}_i,\mathbf{d}_i+1}\right\} \qquad (8)
\end{aligned}$$

The first inequality holds because $\Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) = \Delta R_i^+ + \theta_0$. The second inequality follows by the conditional expectation of the profit increments being nonnegative by Assumptions 1 and 2.

Now consider the asymptotic behavior of the sums in (8). Assume $n_L/n \xrightarrow{p} q$. Then

$$\begin{aligned}
\frac{1}{n}\left\{\sum_{i\in L}\nu_{2,i,\mathbf{d}_i,\mathbf{d}_i-1}+\sum_{i\in U_\eta}\nu_{2,i,\mathbf{d}_i,\mathbf{d}_i+1}\right\} &= \frac{1}{n}\left\{\sum_{i\in L}-\eta_i+\sum_{i\in U_\eta}\eta_i\right\} \\
&\leq \quad \frac{1}{n}\left\{\sum_{i\in L_\eta}-\eta_i+\sum_{i\in U_\eta}\eta_i\right\} = -\frac{1}{n}\sum_{i=1}^{n_L}\eta_{(i)}+\sum_{i=n_L+1}^{n}\eta_{(i)} \\
&\xrightarrow{p} \quad -\mathcal{E}[\eta\mathbf{1}\{\eta\leq F_\eta^{-1}(q)\}]+\mathcal{E}[\eta\mathbf{1}\{\eta\geq F_\eta^{-1}(1-q)\}]=\mathcal{E}(\eta)=0
\end{aligned}$$

where F_η is the c.d.f. of η, and we assume that F_η is continuous and strictly increasing in a neighborhood of q to get the convergence above. The assumed symmetry of the η distribution about zero implies the last two equalities.

This shows that the first half of Assumption 3 holds asymptotically for large n (alternatively we could have averaged over markets and used asymptotics in the number of markets). To show that the second half holds also we have to show that the sums involving ν_1^r in (8) are non-negative asymptotically. The first sum is

$$\frac{1}{n}\sum_{i\in L}\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i-1}^r=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{d}_i>0\}\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i-1}^r\xrightarrow{p}\mathcal{E}[\mathbf{1}\{\mathbf{d}_i>0\}\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i-1}^r]=0$$

since $\mathbf{d}_i\subset\mathcal{J}_i$. The second sum is $\frac{1}{n}\sum_i\mathbf{1}\{i\in U_\eta\}\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i+1}^r$, and since the event, $\{i\in U_\eta\}$, depends on realizations of η_{-i}, the limit of this sum will depend on the information structure of the game. If η_{-i} is in agent i's information set, as would be the case in full information games, then the fact that $\mathcal{E}[\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i+1}^r|\mathcal{J}_i]=0$ would insure that $\frac{1}{n}\sum_i\mathbf{1}\{i\in U_\eta\}\nu_{1,i,\mathbf{d}_i,\mathbf{d}_i+1}^r$converged to zero. If there was asymmetric information then to sign this sum we would require assumptions on the relationship between the unexpected part of agent i's profit measure and η_{-i}.

This example provided a set of assumptions which generates a lower bound for the parameter of interest despite the fact that the choice set is bounded from below. The appendix shows that we can use instruments along with a symmetry assumption to generate more moment inequalities for the lower bound. The example in section 4 provides both empirical and Monte Carlo results which use these estimators.

3 Estimation and Inference for Extreme Points

In section 2.3, we derived the inequality conditions that result from Assumptions 1, 2, and 3. These inequalities fit into the general econometric framework of moment inequalities. A rapidly expanding literature provides a number of possibilities for moment inequality estimation and inference (Andrews and Guggenberger 2009, 2010; Andrews and Jia 2008; Andrews and Shi 2010; Andrews and Soares 2010; Beresteanu and Molinari 2008; Bugni 2010; Canay 2010; Chernozhukov, Hong, and Tamer 2007; Fan and Park 2009;

Guggenberger, Hahn, Kim 2008; Imbens and Manski 2004; Menzel 2008; Romano and Shaikh 2008, 2010; Stoye 2010). Here, we add to this literature with an approach which approximates the distribution of the extreme points of the estimated identified set.

The approach given below is designed to mimic the standard practice, in empirical papers with point identified models, of reporting a table of parameter estimates along with dimension by dimension confidence intervals (typically formed by adding and subtracting twice the estimated standard error). A corresponding notion for set-valued estimates would be to start with dimension-by-dimension extreme point estimates. Then, an analog of "twice the standard error" could be added or subtracted from the extreme point estimates to yield confidence intervals for each dimension of the identified set. We provide such analogs to twice the standard error here.

This approach differs from methods that focus on confidence sets for the whole identified set or coverage of the true parameter alone. Such methods could be modified, for instance, through projections onto a given dimension, to provide the same end product. Here, however, the inference procedure is developed through a local approximation to the limiting distribution of an extreme point estimator. In the next section, a Monte Carlo designed to mimic the empirical example there compares confidence intervals formed in various ways.

Though we will assume that the expectation of the moments used in estimation are continuously differentiable, the fact that the objective function used in estimation in moment inequality problems only penalizes values of the parameter vector that violate the inequalities implies that it can converge to a function which is non-differentiable at its extreme points. As a result the limiting distribution for the extreme point estimator cannot, in general, be obtained from the usual expansion and a normal approximation to the distribution of the objective function. We suggest a simple simulation to conservatively approximate this asymptotic distribution. It is the quantiles of this simulated distribution which we use to mimic the confidence intervals of point identified problems. Andrews and Han (2009) show that this non-differentiability can lead to problems for inference based on simulations of extreme point distributions. In particular, they note that the most straightforward bootstrap of confidence interval endpoints will not generally be uniformly asymptotically valid. The simulation proposed below differs from the Andrews-Han investigation by allowing for some moment selection in the spirit of Andrews-Soares (2010). Additionally, the framework we consider is more restrictive. In particular, our assumptions imply that the class of identified sets under consideration have non-trivial interiors and that the extreme point of interest is a well identified singleton. These assumptions allow for the local approximation that is behind our approach. The pay-off for these more restrictive assumptions is the ability to do direct dimension-by-dimension inference with a computationally simple simulation. In section 4.3 below, Monte Carlo results are shown for the case when our assumptions are "suspect" to alternative degrees.

Our results also allow us to do inference on extreme points in other directions of the parameter space.[14] Further it is straightforward to generalize our results to obtain the

[14]If the identified set is convex, the boundary of that set is defined by the extreme points in all

joint distribution of two or more extreme points and correspondingly alter our inference procedure. For example we could provide the joint distribution to the upper and lower bound of a subvector of the parameter estimates, which, in turn could be used to construct shorter confidence intervals for the actual value of the parameter vector (instead of for the extreme points). Alternatively we could find the joint distribution of the maximum in several directions, thereby approximating the distribution of a section of the boundary of the identified set.

We assume that there is data on J markets indexed by $j = 1, \ldots, J$, and a j superscript will be added to previously defined variables to denote what market they are from. A market is a draw on $\mathbf{w}^j = (\mathbf{d}^j, \mathbf{z}^{oj}, \mathbf{x}^j)$ where $\mathbf{d}^j \equiv \{\mathbf{d}_i^j\}_{i=1}^{n^j}$, $\mathbf{z}^{oj}$, $\mathbf{x}^j$ are defined similarly, and $\mathbf{x}_i^j$ is the observable component of $\mathcal{J}_i^j$ that is used in the weight function h^i in Assumption 3. Assume that the observed markets are independent draws from a population whose distribution is an element of a given class of distributions $\mathcal{F}$ that respects our Assumptions 1, 2, and 3. If F is a distribution function in $\mathcal{F}$, then we let $\mathcal{P}_F$ denote the corresponding measure.

The $M(\equiv \dim(m) \times \dim(h))$ dimensional moment function from equation (11) is:

$$m(\mathbf{w}^j, \theta) = \sum_{i=1}^{n_j} \sum_{d' \in \mathcal{D}_i^j} h^i(d'; \mathbf{d}_i^j, \mathbf{x}_i^j) \Delta r(\mathbf{d}_i^j, d', \mathbf{d}_{-i}^j, \mathbf{z}_i^{oj}, \theta_0)]$$

The inequality in (11) can then be expressed simply as $\mathcal{P}_F m(\mathbf{w}, \theta_0) \geq 0$. Let elements of the moment vector be denoted m_k with covariances $\sigma_{kl,F}(\theta) = \mathcal{P}_F[(m_k(\mathbf{w}, \theta) - \mathcal{P}_F m_k(\mathbf{w}, \theta))$ $(m_l(\mathbf{w}, \theta) - \mathcal{P}_F m_l(\mathbf{w}, \theta))]$. Let $\Omega_F(\theta)$ be the correlation matrix of the moments and $D_F(\theta)$ is the diagonal matrix with diagonal elements $\sigma_{kk,F}(\theta)$. Let $\Gamma_F(\theta) = \frac{\partial}{\partial \theta} \mathcal{P}_F m(\mathbf{w}, \theta)$ be the matrix of partial derivatives of the expected moments at θ.

Let $\Theta \subset \mathbb{R}^K$ denote the parameter space. For a given distribution $F \in \mathcal{F}$, the set of parameters satisfying the moment inequalities is the identified set and denoted by $\Theta_{0,F} = \{\theta \in \Theta : \mathcal{P}_F m(\mathbf{w}, \theta) \geq 0\}$. For notational simplicity, we focus on a particular extreme point of the identified set, the minimizing value of the first dimension of the identified set,

$$\underline{\theta}_F = \{\theta \in \Theta_{0,F} : \theta_1 = \arg \min_{\tilde{\theta} \in \Theta_{0,F}} \tilde{\theta}_1\}$$

where θ_1 denotes the first element of the vector θ, and $\underline{\theta}_F \in \mathbb{R}^K$.[15] In what follows one could equally well consider the minimum or maximum along any other direction in the Θ space. When $\Theta_{0,F}$ is convex, each boundary point can be expressed as the extreme point of some linear combination of dimensions of θ, though convexity of the identified set will not be required for the results to come.

directions. In general, however, by reporting extreme points for each parameter dimension, we are only giving the smallest hypercube containing the set estimate, and this hyper-cube could be a very poor approximation to that set estimate (Stoye 2010)

[15] Assumptions below will ensure that $\underline{\theta}_F$ is well-defined. In general, $\underline{\theta}_F$ could be a set, but the notation and terminology foreshadow our assumption, below, that this set is a singleton.

For specificity we consider estimation of $\underline{\theta}_{1,F}$. Let $\mathbb{P}_J$ denote the empirical distribution so that $\mathbb{P}_J m(\mathbf{w},\theta) = \frac{1}{J}\sum_{j=1}^{J} m(\mathbf{w}^j,\theta)$. Assume that a consistent estimator $\hat{D}_J$ of $D_F(\underline{\theta}_F)$ is available, where $\hat{D}_J$ is itself a diagonal matrix. Then, estimation proceeds as follows. Set

$$\hat{\Theta}_J = \arg\min_{\theta\in\Theta} \| \left(\hat{D}_J^{-1/2} \mathbb{P}_J m(w,\theta) \right)_- \| \tag{9}$$

where $(\cdot)_- = \min\{\cdot, 0\}$, and then set

$$\underline{\hat{\theta}}_{1,J} = \arg\min_{\theta\in\hat{\Theta}_J} \theta_1. \tag{10}$$

Let $\underline{\hat{\theta}}_J$ be any element of $\hat{\Theta}_J$ with first element $\underline{\hat{\theta}}_{1,J}$. Note that the criterion in (9) is in the class of criteria considered in Andrews and Soares (2010), and can also be used for specification testing of the moment inequality model as discussed there.

Consider the implications of two different outcomes of the estimation given in equation (9). If there exists some value of θ such that $\mathbb{P}_J m(w,\theta) \geq 0$, then the solution in (9) is the same as the solution to $\arg\min_{\theta\in\Theta} \|(\mathbb{P}_J m(w,\theta))_-\|$, as long as the diagonal elements of $\hat{D}_J$ are positive. Then $\hat{\Theta}_J = \{\theta \in \Theta : \mathbb{P}_J m(z,\theta) \geq 0\}$, and we can estimate $\hat{\Theta}_J$ without constructing an initial estimator $\hat{D}_J$. The econometric assumptions below will imply that this case occurs with probability approaching one.

The other outcome occurs when there is no $\theta \in \Theta$ such that $\mathbb{P}_J m(w,\theta) \geq 0$. Typically then $\hat{\Theta}_J$ will be a point, so that $\underline{\hat{\theta}}_J = \hat{\Theta}_J$ and the solution to (9) is likely to differ from either the solution to

$$\arg\min_{\theta\in\Theta} \|(\mathbb{P}_J m(w,\theta))_-\|, \quad \text{or to} \quad \arg\min_{\theta\in\Theta} \|(\hat{D}_J(\theta)^{-1/2}\mathbb{P}_J m(w,\theta))_-\|,$$

where $\hat{D}_J(\theta)$ denotes an estimate of $D_F(\theta)$ (so the latter estimator is an analogue of the continuous updating estimator). In this case the estimation in (9) and (10) assumes an estimator $\hat{D}_J$ of $D_F(\underline{\theta}_F)$. A simple way to obtain such an estimator is as follows. Estimate $\Theta_{0,F}$ without the weights, i.e. $\tilde{\Theta}_J = \arg\min_{\theta\in\Theta} \|(\mathbb{P}_J m(w,\theta))_-\|$. Then a consistent estimator for $\underline{\theta}_F$ comes from $\underline{\tilde{\theta}}_{1,J} = \arg\min_{\theta\in\tilde{\Theta}_J} \theta_1$ with $\underline{\tilde{\theta}}_J$ any element of $\tilde{\Theta}_J$ with first element $\underline{\tilde{\theta}}_{1,J}$. Finally the sample variances of the moments evaluated at $\underline{\tilde{\theta}}_J$ can be used as the diagonal elements of the matrix $\hat{D}_J$.

Next we derive the asymptotic distribution of the estimator $\underline{\hat{\theta}}_1$. This result is presented for a fixed choice of F as this clarifies the nature of the estimation problem. However it does not lead directly to inference. For the inference result the uniformity in coverage will require a stronger version of the limiting distribution result under sequences of F's. We will discuss this step following the asymptotic distribution result. The formal assumptions for our results are discussed briefly here. Many of the assumptions require conditions to hold in a uniform sense, which will generally be needed for the inference result but not the limiting distribution result presented first.

The first assumption restricts the identified sets to be closed and not intersecting the boundary of the parameter space. It also imposes what becomes a key condition in

our derivation; the extreme point we focus on is point identified. In principle, one could consider the case when a set of points form the extreme boundary in one direction, but our approach is to localize around the identified point and use the derived behavior to suggest inference. This condition restricts the class of distributions we consider. In particular, we require $\Theta_{0,F}$ to be a set with a nonempty interior (satisfying a degeneracy-type property as in CHT, see Assumption A4) with a singleton extreme point in the direction of interest. Following the inference result, we note how this condition affects its generality.

Assumption A2 gives one side of local identification at every boundary point of the identified set. This could be weakened to focus only on identification of $\underline{\theta}_F$. Assumption A3 ensures that $\underline{\theta}_F$ is a singleton and clearly distinguished from other points in the identified set. Assumption A4 ensures strong local identification of $\underline{\theta}_F$ through conditions on the derivative of the moments evaluated at the extreme point. In particular, there exists a direction toward the interior of the identified set such that, for the nearly binding moments, the derivative in this direction is positive and bounded away from zero. The other part of Assumption A4 ensures that when moving in a direction away from the identified set some binding moment has a strictly negative derivative.

Assumptions A5-A8 impose standard conditions on the moments in a uniform sense. The assumptions ensure that: (A5) the sample average of the moments is uniformly consistent for the mean; (A6) $\hat{D}_J$ is consistent for $D_F(\underline{\theta}_F)$; (A7) the expectation of the moments is continuously differentiable in a neighborhood of $\underline{\theta}_F$; and (A8) the sample moments satisfy a stochastic equicontinuity condition. Assumptions A6 and A7 also bound the variance and derivatives of the moments, which simplifies the uniformity arguments in the inference result and also implies that Lyapunov's central limit theorem can be applied to the sample moments.

Under these assumptions, we can provide a limit distribution for $\sqrt{J}(\hat{\underline{\theta}}_{1,F} - \underline{\theta}_{1,F})$. Let the subvector of moments which are binding at $\underline{\theta}_F$, i.e. for which $\mathcal{P}_F m_k(\mathbf{w}, \underline{\theta}_F) = 0$, be denoted by a 0 superscript, and let any matrix with a 0 superscript refer to the submatrix with rows corresponding to the elements of m^0.

Theorem 1 *Suppose Assumptions A1 -A8 hold and let*

$$\hat{\underline{\tau}}_1 = \min \left\{ \tau_1 : \tau = \arg\min_{\tilde{\tau}} \left\| \left(D_F^0(\underline{\theta}_F)^{-1/2} \Gamma_F^0(\underline{\theta}_F) \tilde{\tau} + \mathcal{Z}^0 \right)_- \right\| \right\},$$

where $\mathcal{Z}^0 \sim N(0, \Omega^0(\underline{\theta}_F))$. *Then,*

$$\sqrt{J}(\hat{\underline{\theta}}_{1,J} - \underline{\theta}_{1,F_J}) \xrightarrow{d} \hat{\underline{\tau}}_1.$$

□

Under our assumptions, there will always exist $\tilde{\tau}$ such that $D_F^0(\underline{\theta}_F)^{-1/2}\Gamma^0(\underline{\theta}_F)\tilde{\tau} + \mathcal{Z}^0 \geq 0$. So, the definition of $\hat{\underline{\tau}}_1$ could be rewritten as

$$\hat{\underline{\tau}}_1 = \min\{\tau_1 : D_F^0(\underline{\theta}_F)^{-1/2}\Gamma^0(\underline{\theta}_F)\tilde{\tau} + \mathcal{Z}^0 \geq 0\}. \tag{11}$$

That is, $\hat{\underline{\tau}}_1$ is simply the solution to a stochastic linear program.

Notice that the asymptotic distribution given by $\hat{\underline{\tau}}_1$ depends only on the characteristics of the subvector of moments, m^0. The remaining moments, m^1, are non-binding when evaluated at $\underline{\theta}_F$, i.e. $\mathcal{P}_F m^1(\mathbf{w}, \underline{\theta}_F) > 0$. Looking at the criterion for estimation in (10), the intuition is that for J sufficiently large the sample average of the non-binding moments will be positive with arbitrarily large probability when evaluated in a neighborhood of $\underline{\theta}_F$. So these moments will not determine the extreme point asymptotically, and, as a result, will have no influence on the asymptotic distribution of the estimator.

Given the binding moments in m^0, the distribution defining $\hat{\underline{\tau}}_1$ could be straightforwardly approximated with estimates of $D_F^0(\underline{\theta}_F)$, $\Gamma_F^0(\underline{\theta}_F)$, and $\Omega_F^0(\underline{\theta}_F)$. Using these estimates, the stochastic linear program above could be simulated to yield the desired approximation. As an inference procedure, this approximation has two related drawbacks. First, *a priori*, which moments fall into the subvector m^0 and which are in m^1 is unknown. Second, the discontinuity in the limit distribution with respect to binding versus non-binding moments corresponds to lack of uniformity in the coverage of this approximation method. As a result, the simulated limit distribution obtained in this way may provide a poor approximation to the finite sample distribution of the estimator (since the simulated data would not reflect the fact that nearly binding moments affect the finite sample distribution). The problem posed by the discontinuity is handled by providing an estimator which accounts for nearly binding moments in addition to the binding moments.

Suppose we have consistent estimators $\hat{\Gamma}_J$ and $\hat{\Omega}_J$ for $\Gamma_F(\underline{\theta}_F)$ and $\Omega_F(\underline{\theta}_F)$ (uniformly in F). Consider the following bootstrap simulation. Take a draw on $Z^* \sim N(0, \hat{\Omega}_J)$ and consider the inequalities $0 \leq \hat{D}_J^{-1/2}\hat{\Gamma}_J\tau + Z^* + r_J \cdot \left(\hat{D}_J^{-1/2}\mathbb{P}_J m(w, \hat{\underline{\theta}}_J)\right)_+$ for a nonnegative sequence $r_J = o(\sqrt{J}/\sqrt{2\ln\ln J})$. If these inequalities have a solution, then we find the solution that minimizes τ_1. If the system of inequalities does not have a solution, then we eliminate inequalities in order of the corresponding value of $\hat{D}_{j,J}^{-1/2}\mathbb{P}_J m_j(w, \hat{\underline{\theta}}_J)$ starting with the largest value. Eliminate inequalities until a solution to the remaining system exists. Let the s subscript denote the indices corresponding to the remaining inequalities. Then,

$$\underline{\tau}_{1,J}^* = \min\left\{\tau_1 : 0 \leq \hat{D}_{s,J}^{-1/2}\hat{\Gamma}_{s,J}\tau + Z_s^* + r_J \cdot \left(\hat{D}_{s,J}^{-1/2}\mathbb{P}_J m_s(w, \hat{\underline{\theta}}_J)\right)_+\right\}. \tag{12}$$

The criterion in (12) mimics the (infeasible) linear program in (11). The influence of non-binding moments is moderated in two ways. First, inequalities corresponding to the largest values of $\hat{D}_{j,J}^{-1/2}\mathbb{P}_J m_j(w, \hat{\underline{\theta}}_J)$ can possibly be eliminated. Second, the shift term $r_J(\hat{D}_{s,J}^{-1/2}\mathbb{P}_J m_s(z, \hat{\underline{\theta}}_J))_+$ is included in the remaining system. Such a term is used in one of the moment selection procedures considered in Andrews and Soares (2010). For the nearly binding moments, this shift term should be close to zero, which will in turn cause these moments to more likely bind in the simulations and have an impact on the simulated distribution. On the other hand for moments which have larger positive values when

evaluated at the extreme point, the shift term will tend to be more strongly positive, so these moments will have little influence on the simulated distribution. The scaling r_J ensures that the simulated distribution is, if anything, a conservative approximation to the distribution obtained from Theorem 1.

Let $\underline{q}_J^*$ denote the α^{th} quantile of the $\underline{\tau}_{1,J}^*$ distribution conditional on the data, so $\Pr^*(\underline{\tau}_{1,J}^* \leq \underline{q}_J^*) = \alpha$. This quantile can be used to provide conservative inference on $\underline{\theta}_{1,F}$ by the following result.

Theorem 2 *Suppose Assumptions A1-A9 hold. Then,*

$$\liminf_{J\longrightarrow\infty} \inf_{F\in\mathcal{F}} \Pr\left(\sqrt{J}(\hat{\underline{\theta}}_{1,J} - \underline{\theta}_{1,F}) \leq q_J^*\right) \geq \alpha. \tag{13}$$

□

The proof of Theorem 2 is obtained by working along a sequence of distributions F_J in $\mathcal{F}$ that attain the limit infimum in (13), and is provided in Appendix 6.2.

As noted above, we focus on distributions where the extreme point of interest is well identified, and, in particular, narrow our focus to $\Theta_{0,F}$ which have non-empty interior. To the extent that this restriction reduces the class of distributions considered relative to moment inequality inference papers like Andrews and Guggenberger (2009, 2010), Andrews and Soares (2010), Romano and Shaikh (2008), the coverage result in Theorem 2 is weaker than in these other papers. That is, our inference method may have coverage uniform over a smaller class of distributions. Computation of confidence intervals is also distinctly different here than in the papers mentioned above. All that's required here are stochastic linear program solutions, which are quite cheap computationally.

The result in Theorem 2 can be used to provide one-sided α-level confidence intervals for $\underline{\theta}_{1,F}$, $[\hat{\underline{\theta}}_{1,J} - q_J^*/\sqrt{J}, \infty)$. By combining this result with the analogous interval for the "upper" extreme point $\bar{\theta}_{1,F} = \arg\max_{\tilde{\theta}\in\Theta_{0,F}} \tilde{\theta}_1$, one can obtain a confidence interval for the first dimension of the identified set $[\underline{\theta}_{1,F}, \bar{\theta}_{1,F}]$. Such a confidence interval can also serve as a confidence interval for the first dimension of the true parameter θ_0, introduced in section 2.2. Imbens and Manski (2004) point out that confidence intervals for the identified set are conservative for interior true points, see also Guggenberger, Hahn, and Kim (2008). Additionally, if we had considered joint estimation of the two endpoints, we could have obtained a possibly shorter confidence interval by accounting for the correlation in estimation. We do not pursue these improvements here.

4 An Empirical Example and Monte Carlo Analogue.

Our empirical example is based on the work of Ishii (2004) who studies the welfare implications of alternative market designs for ATM networks (with particular interest in

designs that do not allow discriminatory surcharges). Her analysis requires estimates of the cost of installing and operating ATMs; estimates we provide here.

The framework for analysis is a two period model with simultaneous moves in each period. In the first period each bank chooses a number of ATMs to maximize its expected profits given its perceptions on the number of ATMs likely to be chosen by its competitors. In the second period interest rates are set conditional on the ATM networks in existence, and consumers choose banks, make deposits, and use ATM's. The second period game is analyzed in Ishii (2004). A careful empirical analysis of the demand system for banking services and an interest rate setting equation enables Ishii (2004) to estimate the parameters needed to compute what the earnings of each bank would have been conditional on alternative choices of ATM networks and to provide an algorithm for doing so. Note that this requires calculating the equilibrium interest rates that would prevail were the alternative possible networks in place.[16]

We now use Ishii's results on the second stage of the game to analyze its first stage; the choice of the number of ATM's. Since the result of these choices is a network, multiple equilibria are likely (for an analysis of the multiple equilibria aspect, see Lee and Pakes 2009). As a result, only the necessary conditions for an equilibrium are available for use in estimation. These conditions form the basis for a multiple agent ordered choice model, and we use the moment inequalities for that model outlined in Examples 3 and 4 of section 2 for inference.

Section 4.1 reviews the model and section 4.2 presents the empirical results. In section 4.3 we build Monte Carlo data sets based on the characteristics of Ishii's data, and use them to compare alternative procedures for inference on the cost parameters.

4.1 The Model and Its Moment Inequalities.

We begin with a brief review of the basic features of the model. Each firm chooses the number of its ATMs, or a $d_i \in \mathcal{D} \subset \mathcal{Z}^+$ (the non-negative integers). The difference in profits from the counterfactual of installing $d'(d_i)$ instead of d_i machines is the sum of a revenue difference and a cost difference. The revenue difference, taken directly from Ishii's (2004) results, is denoted by $\Delta R(d_i, d'(d_i), d_{-i}, z_i^o)$. The cost difference is $(d'(d_i) - d_i)(\theta_0 + \eta_i)$, where $\{\eta_i\}_i$ are firm specific marginal cost differences known to the firms but not to the econometrician, and θ_0 is the mean marginal cost (so $\mathcal{E}(\eta_i) = 0$). Thus the difference in profits from installing and operating d_i rather than $d'(d_i)$ ATMs is

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

[16]The banks' earnings are calculated as the earnings from the credit instruments funded by the deposits minus the costs of the deposits (including interest costs) plus the fees associated with ATM transactions. The ATM fee revenue is generated when non-customers use a bank's ATMs, and revenue is both generated and paid out when customers use a rival's ATMs. Note also that to calculate banks' earnings under alternative ATM networks one must either assume a unique interest rate setting equilibrium, or common knowledge about which equilibrium is selected.

where, as in section 2, $\mathcal{E}[\nu_{1,i,d,d'}|\mathcal{J}_i] = 0$ (the $\nu_{1,i,d,d'}$ result from expectational and/or conditional mean zero specification or measurement error). The assumption that $\nu_{2,i,d,d'} = -\eta_i(d - d')$ mimics the standard ordered choice assumption for the impact of the unobservable known to the agent but not to the econometrician. From Example 3, note that $\Delta r(d_i, d', d_{-i}, z_i^o, \theta) = \Delta R(d_i, d', d_{-i}, z_i^o) - (d_i - d')\theta$.

Two necessary conditions for Assumption 1 are that the expected increment to returns from the last ATM the bank installed ($d'(d_i) = d_i - 1$) were greater than its cost of an ATM, while the expected increment to returns from adding one ATM more than the number actually installed ($d'(d_i) = d_i + 1$) was less than that cost.[17] So,

$$\begin{aligned} 0 \;\leq\; & \mathcal{E}\left[\left(\begin{array}{c} \Delta\pi(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i) \\ \Delta\pi(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i) \end{array}\right)\middle|\mathcal{J}_i\right] = \left(\begin{array}{c} \mathcal{E}[\Delta R(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o)|\mathcal{J}_i] - \theta_0 - \eta_i \\ \mathcal{E}[\Delta R(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o)|\mathcal{J}_i] + \theta_0 + \eta_i \end{array}\right) \\ = \; & \left(\begin{array}{c} \mathcal{E}[\Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i] - \eta_i \\ \mathcal{E}[\Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i] + \eta_i \end{array}\right) \end{aligned}$$

Now, adopt notation for the vector of differences, $\mathbf{\Delta r}_i(\theta)' \equiv [\Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta)$, $\Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta)]$ with $\mathcal{E}(\eta_i) = 0$. More generally, suppose there are "instruments" $\mathbf{x}_i \in \mathcal{J}_i$ with $\mathcal{E}[\eta_i|\mathbf{x}_i] = 0$. If $\mathbf{h}(\cdot)$ is a vector of nonnegative functions while $\otimes$ denotes the Kronecker product, moments formed as

$$m(\mathbf{w}, \theta) = \frac{1}{n}\sum_i \mathbf{\Delta r_i}(\theta) \otimes \mathbf{h}(\mathbf{x}_i)$$

have non-negative expectations, and can be used in estimation and inference as discussed in the last section. Notice that the moment above is for one market. As in section 3, we can add a j superscript to index the markets.

The simplicity of the model makes this a particularly good example for illustrating how inequality analysis works. Consider first using only the moment conditions generated by $h(x_i) \equiv 1$, i.e. $m(\mathbf{w}^j, \theta) = n^{-1}\sum_i \mathbf{\Delta r_i^j}(\theta)$. Then, temporarily assuming all banks have at least on ATM, the moment condition from the profitability difference that arises as a result of decreasing the value of d_i, or the change "to the left", is $(n^j)^{-1}\sum_i[\Delta R(d_i^j, d_i^j - 1, d_{-i}^j, z_i^{oj}) - \theta] \equiv \;\; \Delta R_L^j - \theta$, while the moment condition from the profit change that would result from increasing the value of d_i^j, or the change to the right, is $(n^j)^{-1}\sum_i[\Delta R(d_i^j, d_i^j + 1, d_{-i}^j, z_i^{oj}) + \theta] \equiv \Delta R_R^j + \theta$. Averaging across markets then yields $\Delta\overline{R}_L = \mathbb{P}_J\Delta R_L^j$ and $\Delta\overline{R}_R = \mathbb{P}_J\Delta R_R^j$. Since $(\Delta\overline{R}_L, \Delta\overline{R}_R)$ are the average changes in revenue resulting from first an increase and then a decrease in the number of ATM's, on average we expect $\Delta\overline{R}_L$ to be positive and greater than the average cost of an ATM, while $\Delta\overline{R}_R$ should be negative with absolute value less than that cost. Assuming this to be the case, our estimate of an interval that covers θ_0 is

$$\hat{\Theta}_J = \{\theta : -\Delta\overline{R}_R \leq \theta \leq \Delta\overline{R}_L\}.$$

[17] These conditions will also be sufficient if the expectation of $\pi(\cdot)$ is (the discrete analogue of) concave in d_i for all values of d_{-i}. We can not check this condition without specifying information sets etc., but the realizations of profits evaluated at the estimated value of θ were concave in d_i for almost all banks.

If we add instruments, each new instrument produces a pair of additional inequalities. That is, if k indexes instruments and $\mathcal{E}(\eta_i|\mathbf{x}_{k,i}) = 0$, then

$$0 \leq \mathcal{E}[(\Delta R(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o) - \theta_0)h(\mathbf{x}_{k,i})|\mathcal{J}_i], \;\; 0 \leq \mathcal{E}[(\Delta R(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o) + \theta_0)h(\mathbf{x}_{k,i})|\mathcal{J}_i].$$

Again, including the market index superscript, we compute

$$\begin{aligned}
\Delta\overline{R}_{k,L} &= \frac{\frac{1}{J}\sum_j \frac{1}{n^j}\sum_i (\Delta R(d_i^j, d_i^j - 1, d_{-i}^j, z_i^{oj}))h(x_{k,i}^j)}{\frac{1}{J}\sum_j \frac{1}{n}\sum_i h(x_{k,i}^j)}, \\
\Delta\overline{R}_{k,R} &= \frac{\frac{1}{J}\sum_j \frac{1}{n^j}\sum_i (\Delta R(d_i^j, d_i^j + 1, d_{-i}^j, z_i^{oj}))h(x_{k,i}^j)}{\frac{1}{J}\sum_j \frac{1}{n^j}\sum_i h(x_{k,i}^j)}
\end{aligned}$$

which also estimate bounds for θ_0 from above and below, respectively. The identified set then becomes

$$\hat{\Theta}_J \;=\; [max_k\{-\Delta\overline{R}_{k,R}\},\; min_k\{\Delta\overline{R}_{k,L}\}],$$

when this interval is well defined. So $\hat{\Theta}_J$ becomes shorter (weakly) as the number of instruments increases. Now we expect some of the bounds not to bind, so our estimate of the lower bound is the greatest lower bound while our estimate of the upper bound becomes the least upper bound.

The greatest lower bound is the maximum of a finite number of moments each of which will, in finite samples, distribute approximately normally about a separate mean, say $\underline{\theta_k} \leq \theta_0$. So when we use this max as our estimator in small samples we should expect a positive bias in the greatest lower bound (the expectation of the maximum of normal random variables is greater than the maximum of the expectations), and the bias should be (weakly) increasing in the number of inequalities. Analogously, since the estimate of the upper bound is a minimum, we should expect a small sample negative bias in it. As a result, even if the model is correctly specified, the estimated lower bound can be greater than the estimated upper bound, in which case the estimation criterion from the previous section will choose $\hat{\Theta}_J$ to be a singleton. The other reason why we may obtain a point estimate is that the model is misspecifed; i.e. there is no value of θ which satisfies all the population moment restrictions.[18]

Boundaries. For those agents with $d_i = 0$, $d_i - 1$ is not a feasible choice and $\Delta R(d_i, d_i - 1, \cdot)$ can not be calculated. As noted in Examples 3 if, in cases like this, we base averages only on those observations with $d_i \geq 1$ we introduce a selection problem; those observations with $d_i = 0$ may have disproportionately high costs of aquiring an ATM. In Example 4 of section 2 we showed that we can correct for this selection provided the distribution for the η_i is symmetric. In our example, a small fraction, just under 5.5%, of the observations have $d_i = 0$. We present results both with the symmetry assumption and the associated correction as well as without this correction.

[18]Hirano and Porter (2009) discuss implications of the bias from taking the minimum and maximum as endpoint estimates.

4.2 Empirical Results

The data set consists of a cross-section of all banks and thrifts in Massachusetts metropolitan statistical areas in 2002. A market is defined as a primary metropolitan statistical area, and the sample is small: it contains a total of 291 banks in 10 markets.[19] The number of banks varies quite a bit across markets (from 8 in the smallest market to 148 in Boston), as does the number of distinct ATM locations per bank (which averages 10.1 and has a standard deviation of 40.1). The estimation routine is as described above.

Table 1 contains the inequality estimators of the cost parameter. The first two rows provide the results when only a constant term is used as an instrument; the first row uses only those observations with $d \geq 1$ to calculate the upper bound, while the second row adds in the observations with $d = 0$ and uses the symmetry assumption to correct for the fact that banks which decided not to invest in ATM's may have had higher than average ATM related costs (see Example 4 in section 2). The lower bound of the estimates of the identified set from the two rows have to be identical, but we would expect the upper bound from the estimates that correct for selection to be higher than those that do not. This is what we find, though the difference in upper bounds is relatively small (25,283 vs 26,644), about equal in percentage terms to the fraction of observations with $d_i = 0$. Even after the correction the estimate of the identified interval is quite short ([24,452, 26,444]), but its confidence interval, using the method described in the previous section, is larger ([20,472, 30,402]) .

Rows 3 and 4 repeat the exercise in Rows 1 and 2 after expanding $h(x)$ to include the constant term, the market population, the number of banks in the market, and the number of branches of the bank (its mean is 6 and standard deviation is 15). $\hat{\Theta}_J$ is then a singleton and its value is about equal to the lower bound of the confidence intervals obtained using only the constant term as an instrument. The fact that we obtained a point estimate reflects that there is no parameter value such that all the sample moment inequalities are satisfied. This could be due to the small sample bias discussed above or a misspecification; say a violation of Assumption 3 due to correlation between the instruments and η. When we formally tested for misspecification, we found that we rejected the null.

There is at least one more source of information in the data. Currently, we are only using profit differences based on the counterfactuals $d' = d_i \pm 1$. We also try including profit differences based on counterfactuals $d' = d_i \pm 2$, adding the rows $[\Delta r(\mathbf{d}_i, \mathbf{d}_i - 2, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta), \Delta r(\mathbf{d}_i, \mathbf{d}_i + 2, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta)]'$ to the vector $\mathbf{\Delta r}_i(\theta)$. If the expected profit function is (the discrete analogue of) concave in d, then adding the extra moments should not change the identified interval; but it may change the confidence interval associated with it. We note that the fraction of banks with $d_i < 2$ is much larger than the fraction with $d_i < 1 (25.5\% \; vs \; 5.5\%)$, so the truncation issue corresponding to the counterfactual $d' = d_i - 2$ is potentially worse. However, if the interval itself is set by the $d' = d_i - 1$

[19] The data set is described in Ishii (2004), and is carefully put together from a variety of sources including the Summary of Deposits, the Call and Thrift Financial Reports, the 2000 Census, the Massachusetts Division of Banks, and various industry publications.

Table 1: **Inequality Method, ATM Costs***

	θ_J	95% CI for θ LB	UB
1. $h(x) \equiv 1, d \geq 1$ for u.b. $\hat{\theta}$	[24,452, 25,283]	20,544	29,006
2. $h(x) \equiv 1, d \geq 0$	[24,452, 26,444]	20,472	30,402
$h(x)$ =(1,pop, # Banks in Mkt, # Branches of Bank)			
3. $d \geq 1$ for u.b. $\hat{\theta}$	19,264	16,130	23,283
4. $d \geq 0$	20,273	17,349	24,535
$\{d : \lvert d - d_i \rvert = 1, 2\}$, $h(x) = 1$			
5.$\{d : \lvert d - d_i \rvert = 1, 2\}$; $d \geq 1$ for u.b. $\hat{\theta}$	[24,452, 25,283]	20,691	28,738
6.$\{d : \lvert d - d_i \rvert = 1, 2\}$; $d \geq 0$	[24,452, 26,644]	20,736	29,897
First Order Conditions (analogue of Hansen and Singleton,1982)			
7. h(x)=1	28,528	23,929	33,126
8. h(x)=(1,pop, # Banks in Mkt, # Branches of Bank)	16,039	11,105	20,262

* There are 291 banks in 10 markets. The first order condition estimator requires derivatives with respect to interest rate movements induced by the increment in the number of ATMs. We used two-sided numerical derivatives of the first order conditions for a Nash equilibria for interest rates.

moment, the greater truncation for $d' = d_i - 2$ should not impact our interval estimate. Rows 5 and 6 present the results from interacting these four profit differences with a constant term. The interval estimates are in fact unchanged by this addition. This implies that when we add the moments formed from the counterfactuals $d' = d_i \pm 2$ the confidence intervals can not become larger; in fact they shrink by about a thousand dollars.

Alternative Models and Estimation Methods There are at least two alternative econometric models that have been used in analogous empirical problems; ordered choice models, and models which base estimation on first order conditions. In our notation, the ordered choice model sets $\nu_1 \equiv 0$, assumes a particular distribution for ν_2 conditional on the other determinants of profits, and forms the likelihood of the observed d. For this to be correct in an interacting agent model there must also be a unique equilibrium and no correlation among the η_i of the firms in a market. The first order condition estimator ignores the discrete nature of our control, and then goes to the opposite extreme: it assumes that $\nu_2 \equiv 0$ (there is no structural error). It does not restrict the distribution of the remaining (ν_1) disturbance nor does it require a unique equilibrium.

The ordered choice model can not be estimated with our data. The issue is that for some of our observations the revenue "difference from the left" is less than the revenue "difference to the right". Then there is no value of $\theta + \eta_i$ that rationalizes the observed choice: if it was profitable to purchase the last ATM, the model says that it must have been profitable to purchase the next ATM (the log likelihood is minus infinity for any θ and any distribution for the η_i). If there is either some uncertainty when decisions are made or measurement error in revenues, we should not be surprised to find an agent whose observed revenue "difference from the left" is less than that "difference to the right," even if all agents acted optimally.[20]

Assuming the control to be continuous and that agents act optimally, the first order condition for agents with $d > 0$ must have an expectation of zero conditional on their information sets. Maintaining the continuous control assumption and provided $x_i \in \mathcal{J}_i$, a consistent estimator of θ_0 can be found by minimizing

$$\left\| \frac{1}{n} \sum_i \{d_i > 0\} \left(\left. \frac{\partial R(d, d_{-i}, z_i^o)}{\partial d} \right|_{d=d_i} - \theta \right) \times h(x_i) \right\|$$

with respect to θ. Given the discrete nature of the ATM application, we approximate the first order condition by replacing the derivative in the above equation with one half of the sum of the change in profits from first increasing and then decreasing the observed number of ATM's by one. So the simplest first order condition approximation adds the two moment inequalites used in row 1 together for those observations with $d \geq 1$, divides by two, and imposes that on average the result should converge to zero at $\theta = \theta_0$. The estimates are presented in rows 9 and 10 of the table. Again the estimates that use the added instruments give different results than those that just use the constant term, so we focus on the latter. The first order condition point estimate is outside of the interval estimate obtained from the inequality estimators (by 15 to 20%), but about equal to the upper bound of the confidence interval obtained from the moment inequalities. Interestingly the confidence interval from the first order condition estimator is only a bit shorter than that from the moment inequalities; the latter uses inequalities, the former uses equalities, but the latter uses more inequalities than the former uses equalities.

4.3 Monte Carlo

We focus on the performance of alternative inference methods for the upper and lower bounds of the parameters, corresponding to single dimensional confidence intervals that are likely to be reported in empirical work. By "performance" we mean both distance to the true values of these bounds and coverage.

[20] Of course one could modify the simple ordered choice model and avoid the possibility of events that the model assigns zero probability to. For example, one could specify a particular form for measurement error and then construct a likelihood by numerical integration or simulation. This, however, would require more modelling assumptions and a more complicated estimation algorithm.

We want to use the Monte Carlo to investigate a multi-dimensional parameter model so we modify Ishii's (2004) model to allow marginal cost to be a function of the number of ATM's bought. In the Monte Carlo the cost of the d^{th} ATM for firm i is $\theta_1+\theta_2(d-1)+\eta_i$. The definiton of $\mathbf{\Delta r_i}(\theta)$ is changed accordingly

$$\mathbf{\Delta r_i}(\theta) = \begin{pmatrix} \Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) \\ \Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) \end{pmatrix} = \begin{pmatrix} R(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i^o) - R(\mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o) - \theta_1 - \theta_2(\mathbf{d}_i - 1) \\ R(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i^o) - R(\mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o) + \theta_1 + \theta_2\mathbf{d}_i \end{pmatrix},$$

and Assumptions 1 and 2 yield

$$0 \;\leq\; \mathcal{E}\left[\begin{pmatrix} \Delta\pi(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i) \\ \Delta\pi(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i) \end{pmatrix} \middle| \mathcal{J}_i \right] = \begin{pmatrix} \mathcal{E}[\Delta r(\mathbf{d}_i, \mathbf{d}_i - 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i] - \eta_i \\ \mathcal{E}[\Delta r(\mathbf{d}_i, \mathbf{d}_i + 1, \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|\mathcal{J}_i] + \eta_i \end{pmatrix}.$$

Setting $N = \sum_j n^j$, the equation $N^{-1}\sum_{i,j} \mathbf{\Delta r_i^j}(\theta) = 0$ produces two lines in (θ_1, θ_2) space. One has slope $[N^{-1}\sum_{i,j} d_i^j]^{-1}$ and, using the sample moment inequality, bounds acceptable (θ_1, θ_2) combinations from below. The other has slope $[N^{-1}\sum_{i,j}(d_i^j - 1)]^{-1}$ and bounds them from above. $\hat{\Theta}_J$ is the intersection of these two half-spaces. Each time we add an instrument we add two additional moments to the estimation inequalities. These inequalities generate further half-spaces with boundaries given by the zeroes of the moments. One boundary line has slope $[N^{-1}\sum_{i,j} h(x_i^j)]/[N^{-1}\sum_{i,j} d_i^j h(x_i^j)]$ and the other has slope $[N^{-1}\sum_{i,j} h(x_i^j)]/[N^{-1}\sum_{i,j}(d_i^j - 1)h(x_i^j)]$ The estimated identified set $\hat{\Theta}_J$ then becomes the intersection of a larger set of half-spaces. If that intersection is the null set we obtain a point estimate of Θ_0.

Constructing the Monte Carlo Data Sets. We begin each Monte Carlo data set with a random drawing of firms from Ishii's data set. For each firm, the Monte Carlo data set uses the observables variables (other than d_i) associated with each firm that is drawn.[21] An initial d_i for each firm is also obtained in this way but then is adjusted, as described below, to insure that $\sum_i \eta_i d_i < 0$, as we might expect if d_i were chosen optimally.

We use two different algorithms to adjust the d_i and this produces two different data sets. In one we insure that there are no observations with $d_i = 0$. Then we can compute the change in profits when we decrease or increase the number of ATMs by one for all observations. In the other we allow observations with $d_i = 0$, so there are some observations for which we can not compute the change in profits when we decrease d_i. To correct for the selection problem this induces in moments that involve profit changes from decreasing d_i we use the symmetry of the η_i distribution and the algorithm described in Example 4 of section 2 for constructing our moment inequalities.

In both data sets the "true" values of (θ_1, θ_2) are set equal to $(14,846;\ 1,312.5)$ and four separate unobservables are assigned to each observation. The unobservables are a

[21] This includes Ishii's data on $R(d, d_{-i}, z_i)$ and an x_i vector consisting of the number of branches of bank i, the population of the city it operates in, and the total number of banks in that city.

component of marginal cost (η_i, which determines $\nu_{2,i}$), and three separate "shocks" which combine to form $\nu_{1,i}$. The components of $\nu_{1,i}$ consist of a shock to the increment in returns (to $\Delta R(\cdot)$) which has the same effect on the increment in going from d_i to d_i+1 as it does to the increment in going from d_i-1 to d_i, and separate shocks for each of the increments that are independent of one another and of the common shock.[22]

The procedure for adjusting d_i differed for the truncated and non-truncated case. For both we begin with the draw η_i on the cost shock. If the draw was less than one standard deviation from the mean we left d_i alone. If it was positive (negative) and between one and two standard deviations from the mean we decreased (increased) d_i by one, and if it was positive (negative) and more than two standard deviations from the mean we decreased (increased) d_i by two. In constructing the sample with no observations at $d_i = 0$, after doing the procedure described above we changed all d_i that were less than one to one. For the sample that allowed for observations with $d_i = 0$, at the end we set all d_i that were less than or equal to zero equal to zero.

Inference Methods. We use the bounds estimator introduced in Section 3 along with the simulation-based inference method introduced there. Below we refer to this inference method as the "interval inference method." In addition, we consider three other methods of constructing confidence sets which we will refer to collectively as "grid methods". Each method has a point coverage and a set coverage version. These methods fall in the class of inference procedures developed in Andrews and Soares (2010). Each method compares a sample criterion function to a simulated critical value to determine the confidence set. We do not vary the criterion function, see (9), across methods; only the critical value simulation changes. The first method, introduced in Chernozhukov et. al (2007), simulates each moment after re-centering at zero. The second method selects a subset of "nearly binding" moments to re-center and then simulates. The last method re-centers based on the sample moments as in (12). Details of the confidence interval construction methods follow.

For all but the bounds inference method, we begin by defining a grid of points in (θ_1, θ_2) space; the points that will be considered for inclusion in the confidence sets. Both the boundaries for the grid and the fineness of its partition will typically be determined empirically and require some prior analysis. Since we have the advantage of knowing the true parameter values and will not be concerned with the computational costs of proce-

[22] All of the unobservables are random draws from a normal distribution with mean zero. The η_i shock was assigned a standard error that was 10% of the median of the cross-sectional distribution of $\theta_1 + \theta_2 d_i$. To obtain the standard error of the $\nu_{1,i}$ components that were independent across increments we took one half of the cross-sectional variance of the difference in the two increments, and then took one half of the associated standard error. To obtain the standard error of the $\nu_{1,i}$ component that was common to the two increments we took one half of the cross sectional variance in the sum of the two increments minus the variance in both the cost shock and the idiosyncratic shocks to the increment in returns, and then took one half of the standard error of the resultant variance. So about half of the observed cross-sectional variance in profit increments is being attributed to random variables that were not known when decisions were made.

dures for defining these points, we simply set these details up front. The boundaries were set as zero to two times the parameter value in each dimension, and within the rectangle formed from these boundaries we placed over 20,000 grid points. The distance between the grid points was made smaller for points that were close to, or actually members of, the true identified set.[23] We refer to the separate grid points as $\{\theta^g\}_{g=1}^{n_g}$.

The grid methods treat the criterion function as a test statistic. Then compare the test statistic evaluated at each grid point to a critical value to determine a confidence set. Let l index the Monte Carlo data sets, and consider the criterion function from (9), $Q_l(\theta) = \|\Big((\hat{D}_J^l)^{-1/2}\mathbb{P}_J^l m(w^l,\theta)\Big)_-\|$. To form an α-level pointwise confidence set, critical values, $c_{l,\alpha}(\theta)$, are simulated in three different ways at each grid point, θ^g. Then the pointwise confidence set is defined as those θ^g where $Q_l(\theta^g) \leq c_{l,\alpha}(\theta^g)$.

To compute $c_{l,\alpha}(\theta^g)$, an approximation to the distribution of the criterion function under the null ($\theta^g = \theta_0$) is simulated by three methods. In the first method, the distribution of the sample moments in the expression for $Q_l(\theta^g)$ is approximated by a normal with mean zero and variance given by the sample moment variance estimated using the l^{th} Monte Carlo data set. These Gaussian simulations generate a corresponding approximate distribution for $Q_l(\theta^g)$ (under the null). The $1-\alpha$ quantile of this distribution is defined as the critical value $c_{l,\alpha}(\theta^g)$. Two other methods are used to generate critical values through variants of the first method which take account of the moments that are binding or close to binding. The second method ("moment selection") selects only the nearly binding moments at each θ^g and simulates this subset as in the first method. The k^{th} moment is dropped in the simulation at θ^g if $\sqrt{J}\mathbb{P}_J^l m_k(w^l,\theta^g)/\hat{\sigma}(m_k(w^l,\theta^g)) \geq \sqrt{2\ln(\ln(J))}$. The third method uses "moment shifting" similar to the procedure proposed in section 3. All moments are simulated, as in the first method, but rather than centering the normal distribution at zero, the mean of the normal is set to $\left(\sqrt{J}(\hat{D}_J^l)^{-1/2}\mathbb{P}_J^l m(w^l,\theta^g)/\sqrt{2\ln(\ln(J))}\right)_+$. This shift moderates the influence of highly positive moments at θ^g in the simulated distribution.

To form a setwise confidence set (for coverage of Θ_0), the same criterion function is used, but a critical value, $\bar{c}_{l,\alpha}$ that does not depend on θ^g is obtained through simulation. For a given Monte Carlo data set, we obtain a conservative estimate of the identified set, $\hat{\Theta}^l(\epsilon)$[24]. To obtain $\hat{\Theta}^l(\epsilon)$, we first find the minimum value of the criterion $Q_l(\theta)$ over all θ.

[23] We first divided each dimension of the rectangle into one hundred equally spaced intervals and took the Cartesian product of their endpoints to produce 10,000 points. We then looked over the estimated identified sets from the samples of firms we drew, and added another 10,000 points obtained in the same way but using as endpoints the fifth percentile of the lower bound in a given dimension and the ninety fifth percentile of the upper bound in the same dimension as the endpoints. Finally for samples for which our estimator was a point, we added that point, and for samples where the estimator was a set too small to include any of the points in our grid, we added a midpoint from that set to our set of points. This insured that the identified set contained at least one point for each sample, and therefore insured an ability to compute coverage for each sample.

[24] We used $\epsilon = \sqrt{F^{-1}_{\chi^2_m}(1-\alpha) + (\ln n)}$ as suggested by CHT. The values of ϵ for the four and eight moment models presented below were 9.49 and 15.51, respectively.

Then, $\hat{\Theta}^l(\epsilon)$ consists of the parameter grid points where the criterion function is within ϵ of the minimum, i.e. $Q_l(\theta^g) \leq (\min_\theta Q_l(\theta)) + \epsilon$. This set estimate $\hat{\Theta}^l(\epsilon)$ is used for the setwise versions of all three grid methods. Each method above describes a way to simulate an approximation to the criterion function at a given value of θ. Let $Q^*_{l,s}(\theta)$ denote the s^{th} simulation draw from the simulation distribution for any one of the methods. Fixing s, form $\sup_{\theta \in \hat{\Theta}^l(\epsilon)} Q^*_{l,s}(\theta)$. Then the critical value $\bar{c}_{l,\alpha}$ is the $1-\alpha$ quantile of this simulated distribution. Each method of simulating moments produces a method to simulate $Q^*_{l,s}(\theta)$, which in turn generates a critical value. These critical values determine the confidence set as the θ^g such that $Q_l(\theta^g) \leq \bar{c}_{l,\alpha}$.

For each inference method, we computed: (i) the average of the intervals obtained by projecting the confidence set onto each axis and taking its endpoints, and (ii) empirical coverage rates for; the point θ_0, each $\theta \in \Theta_0$, the set Θ_0, and the interval formed by the extreme points of the identified set $(\underline{\theta}_1, \overline{\theta}_1)$ and $(\underline{\theta}_2, \overline{\theta}_2)$.

The different inference methods differ in their computational burden. The method of Section 3 begins by estimating the extreme points of the identified set. When the inequalities are linear (as in our example), this only requires checking the vertices of a simplex (see Dickstein and Morales, in process, for details). In non-linear problems a search algorithm would be employed to reduce the number of θ values at which the criterion function needs to be evaluated while searching for the extreme points. Once the extreme points are found, the original criterion function is used to obtain the correlation and Jacobian term estimates ($\hat{\Omega}_J$ and $\hat{\Gamma}_J$). In the inference method's simulations, the criterion function is never re-evaluated. For the grid methods that generate confidence sets for the point θ_0 we need to evaluate the criterion function at every grid point. The methods that generate confidence sets for the Θ_0 have a first stage which estimates the identified set, and a second stage which evaluates the criterion at all grid points in the estimate of the identified set. In our simple example it is not difficult to evaluate the criterion at different values of θ. However, in examples that require the computation of a fixed point every time the criterion function is evaluated at a different θ, the computational burden will rise with (i) the number of grid points that need to be evaluated, and (ii) the dimensionality of the fixed point. As either of these grow, and they typically both grow rapidly with the complexity of the problem being analyzed, the grid methods will become increasingly computationally burdensome.

Monte Carlo Results. All Monte Carlo results are based on one thousand samples of eight hundred and seventy three firms each. The simulated critical values for the confidence sets and the simulated distribution for the interval estimator were based on four hundred simulation draws for each sample. We begin with two and then increase the number of instruments. The two instruments are a constant and the number of markets the bank operates in. They generate four moments; just identifying the upper and lower bounds of each parameter. Nominal coverage for all our experiments was 95%.

The interval inference method in section 3 generates a confidence interval for each

dimension of the parameter. The grid methods provide confidence sets which we project onto each dimension. Table 2 provides the average interval endpoints in each dimension for each of the methods for both the non-truncated and the truncated samples. Table 2 also shows the empirical coverage of the intervals given by the upper and lower extreme points of the identified set in each dimension. The coverage numbers show that all the grid coverage methods provide conservative inference for the *intervals*. The grid methods were designed to provide coverage of either the point θ_0 or the set Θ_0, so the finding that they provide conservative coverage for intervals is *not* a sign of poor performance. All methods designed to cover the point θ_0 also generated conservative coverage of that point, but the methods designed for point coverage are designed to provide coverage for any value of $\theta \in \Theta_0$, so it may be more appropriate to judge their coverage by the minimum coverage over all points in the true Θ_0. The minimal coverage of the point coverage methods over all $\theta \in \Theta_0$ was more in line with the nominal coverage, ranging from 93.2 to 97.2%.

The coverage of the interval inference method developed in Section 3 is close to the nominal coverage for $[\underline{\theta}_1, \overline{\theta}_1]$ but it under-covers $[\underline{\theta}_2, \overline{\theta}_2]$. Taking a closer look at the nature of the under-coverage, we found that generally in the Monte Carlo samples where the confidence interval did not cover, the confidence interval endpoints were barely inside the true values of the endpoints. To evaluate the likely impact of the under-coverage we: (i) moved the two endpoints of the actual interval back in equal percentages until the simulated coverage equaled 95% and then reported the required percentage reduction, and (ii) calculated the minimum of the pointwise coverage of for any value of θ_j that is a component of some $\theta \in \Theta_0$ (for $j = 1, 2$). Over the four cases; (i) the maximum perentage reduction of the interval needed was .65%, and (ii) the minimum coverage for any point in the set was 95.3%.

This brings us to the most striking result in the table: the confidence interval endpoints generated by the interval inference method are much closer to the true endpoints than those from the other procedures. In the non-truncated sample the lengths of the average intervals in the θ_1 and θ_2 dimensions generated by the interval estimates were [4,214; 251]. Even if we only consider the confidence intervals designed to cover the point θ_0 the grid methods generated lengths which ranged from [20,184; 1,803] to [10,891; 929]. That is, the average length of the intervals generated by the projections from the grid methods were 2.5 to 5 times (for θ_1) and 3.5 to 7 times (for θ_2) the average length of the interval obtained from the interval method. Interestingly the shortest interval obtained from any of the grid methods is always that produced by the inference method that combines the same shifted mean adjustment the interval method makes with the grid inference procedure. The results from the truncated sample are similar in these respects. Indeed, the only notable difference for the truncated sample is that the truncation together with our correction for it lead to a noticeable increase in the endpoint corresponding to $\overline{\theta}_1$ (but not $\overline{\theta}_2$).

Figure 1, which plots the average value of the sample criterion function evaluated at different θ values (averaged over Monte Carlo data sets), helps to see how these differences can arise. The minimum of the average criterion function is 1.00 and it occurs at $\theta \approx \theta_0$.

Table 2: **Average Intervals and Coverage: Four Moments.**

Inference Method	$[\underline{\theta}_1; \overline{\theta}_1]$	% Cover	$[\underline{\theta}_2; \overline{\theta}_2]$	% Cover
Non-Truncated Sample				
1. True Intervals	[13,876;15,688]		[1,266;1,366]	
2. Interval Inference	[12,754;16,968]	96.1	[1,190;1,441]	93.7
% ↓ in interval for 95% Cov. & Min. Cov. of $\theta \in \Theta_0(*)$		0 & 98.5%		.125 & 96.5%
3. Point Inference	[2,081;22,665]	100	[633;2,436]	100
4. Set Inference	[2,042;22,665]	100	[626;2,441]	100
5. Point with Moment Selection	[2,088;22,666]	100	[633;2,436]	100
6. Set with Moment Selection	[2,042;22,749]	100	[626;2,441]	100
7. Point with Shifted Mean	[8,339;19,230]	99.8	[932;1,860]	100
8. Set with Shifted Mean	[2,349;22,268]	100	[668;2,410]	100
Truncated Sample				
9.True Intervals	[13,538;18,095]		[1,169;1,405]	
10. Interval Inference	[12,409;19,268]	92.9	[1,065;1,500]	89.4
% ↓ in interval for 95% Cov. & Min cov for $\theta \in \Theta_0.(*)$		.5 & 96.5%		.65 & 95.3%
11. Point Inference	[1,623;30,443]	100	[147;2,567]	100
12. Set Inference	[1,552;30,529]	100	[143;2,575]	100
13. Point with Moment Selection	[1,165;31,314]	100	[98;2,623]	100
14. Set with Moment Selection	[719;32,058]	100	[70;2,672]	100
15. Point with Shifted Mean	[7,938;23,802]	100	[633;1,949]	99.9
16. Set with Shifted Mean	[2,186;29,520]	100	[201;2,504]	100

$(*)$Percentage decrease in endpoints of intervals to obtain 95% coverage, and the minimum covarage for any value of the component of the parameter vector that lies inside the identified set.

The criterion function increases rather rapidly as we move one component of θ away from θ_0 holding the other component fixed. However there is a set of (θ_1, θ_2) values that define a diagonal oblong-shaped area at which the criterion function increases only slowly as we move away from θ_0. It is this set of parameter values that are generally below the critical values generated by the grid method procedures and hence form their confidence sets. When these sets are projected onto each dimension, the resulting rectangle covers a much larger area then the sets themselves. Of course in this two-dimensional case, the applied researcher could easily communicate which points are in the confidence set, and the rectangle might never be needed; but, as noted, this gets much harder for higher dimensional problems.

The confidence sets for the grid methods rely on the estimate of a covariance matrix which depends on θ. These methods compare the value of the objective function at different θ, a value which depends on the estimated variances at that θ, to a critical value for that function obtained from simulating data from a mean zero normal distribution with a correlation matrix which depends on θ. To obtain the critical value of the objective function we drew one set of normal draws and used then used the Cholesky transform of the estimated correlation matrix at that θ to find the appropriate critical value. In the interval method only the covariance matrix at the endpoints of the estimated intervals and the derivative matrix at those points are used to determine whether any θ is in the confidence set. So the grid methods will be sensitive to the variance-covariance of the objective function at points outside the estimated identified set, while the interval estimate will be sensitive to the estimate of the derivative matrix at the endpoints of the estimated intervals. In our case the inequalities are derived from discrete analogues to first order conditions and, at least in the non-truncated case, are linear in the parameter vector (so the derivative matrix does not involve estimated parameters).

There is one other aspect of this example we pursued. All the data sets drawn for our Monte Carlo runs yielded set estimates of both parameters. Our experience in using moment inequalities in applied work is that point estimates do occur; even in cases in which one can accept the null that all the moment inequalities are satisfied at some θ. To explore the properties of the various estimators in situations where point estimates are more likely to arise we added moments and then recomputed the estimators. First we added two moments which interact our unit differences in the number of ATM bought with the number of branches the bank has in all markets the bank operates in. This generated only a modest number of samples for which there was a point estimate (1.7%), but it shortened the true intervals dramatically. In particular, the new θ_1 interval was one third of its prior length, and the new θ_2 interval was forty percent of its previous length. Next we added two interactions with the number of branches the bank has in the given market. These additional moments both increased the number of samples that generated point estimates sharply (to 37.1%) and further reduced the true intervals (particularly for the θ_2 parameter). Table 3 presents a summary of the results.[25]

[25]We omit the results for the truncated sample and for the grid methods designed to provide coverage for sets (rather than points) because their relationship to the results Table 3 mimics the relationship

Looking first to the six moment case, all estimators cover the true intervals with more than the nominal coverage rates. The lengths of the confidence intervals estimated by the grid estimators relative to that of the interval estimator have all decreased substantially. Again the grid method which uses a shifted mean does noticeably better than the other grid methods. Now the lengths of the confidence intervals it generates for the θ_1 parameter are comparable to those generated by the interval estimator (the mean length is shorter and the median is longer than those of the interval method), but the interval estimator still generates shorter confidence intervals for the θ_2 parameter.

When we move to the eight moment case the results change rather sharply. This is a case which is quite demanding of the interval estimator; the length of the true θ_2 interval is only about 1% of its value, so we are nearly point identified in that dimension in the population we are drawing from. Accordingly the coverage of the interval estimator falls dramatically, to 64% for the estimator for the θ_1 interval and 70% for θ_2. Looking closer at the data, it is clear that the under-coverage was a result of the fact that many of the Monte Carlo data sets generated point estimators. When we only use the Monte Carlo data sets that generated estimated sets the coverage is above 95%. Just as before, the grid method with the shifted mean does much better than all the other grid methods, and it has more than adequate coverage.

We conclude that, at least in our example, when data sets generated set estimators of the parameters of interest, the interval estimator did quite well, and its advantages over the grid methods grew with the length of the estimated intervals. However, when the sample drawn generated a point estimator, the coverage of the interval estimator went down dramatically. This is not that surprising given that our theoretical results do not necessarily hold when the population generates moments which point identify the true parameter vector. The grid method which uses the same shifted mean adjustment used in the interval estimator does better in all experiments than the rest of the grid methods; and it always has adequate coverage.

In one sense these results are heartening; one estimator of the confidence interval does well when there is an estimated set while the other does well when a point is estimated, so the applied researcher can look at the resulting estimator and choose an appropriate confidence interval. The problem that remains is that there are many cases in which computational burdens make grid point methods infeasible, and some of them will generate point estimators. For example, in Ho and Pakes (2011), there are over one hundred and fifty parameters to estimate, and it takes several hours just to compute the covariance matrix associated with one vector value for those hundred and fifty parameters. Of course when computation of the objective function at a particular θ requires computing a fixed point, as in a nested fixed point algorithm, the grid point methods will be infeasible with a much smaller dimensional parameter space.

between the analogous results in Table 2.

Table 3: **Intervals and Coverage**

Inference Method	$[\underline{\theta}_1; \overline{\theta}_1]$	% Coverage	$[\underline{\theta}_2; \overline{\theta}_2]$	% Coverage
Non-Truncated Sample, Six Moments. Point estimates in 1.7% of Data Sets.				
True Intervals	[14,232;15,417]		[1,293;1,334]	
Interval Inference		97.1		95.6
Mean	[13,081;21,049]		[1,133;1,393]	
Median	[13,104;17,394]		[1,192;1,388]	
Point Inference	[10,325;17,951]	100	[1,100;1,641]	100
Point with Moment Selection	[10,358;17,933]	100	[1,100;1,641]	100
Point with Shifted Mean		99.8		99.9
Mean	[12,145;17,161]		[1,172;1,477]	
median	[12,334;16.960]		[1,210;1,440]	
Non-Truncated Sample, Eight Moments. Point estimates in 37.1% of Data Sets				
True Intervals	[14,376;15,298]		[1,305;1,320]	
Interval Inference		63.8		70.1
Mean	[12,617;19,480]		[1,203;1,353]	
Median	[12,138;17,834]		[1,201;1,351]	
Interval: Estmated Sets Only (*).		95.1		100
Mean	[11,930,21,690]		[1,151;1367]	
Median	[11,863,19,120]		[1,220;1362]	
Point Inference	[11,860;17,243]	100	[1,174;1,498]	100
Point with Moment Selection	[11,961;17,183]	100	[1,221;1,417]	100
Point with Shifted Mean		99.8		99.9
Mean	[12,759;16,276]		[1,221;1,417]	
Median	[12,850;16.632]		[1,242;1,400]	

(*) Drop all Monte Carlo data sets that result in degenerate estimates of the identified sets.

5 Summary

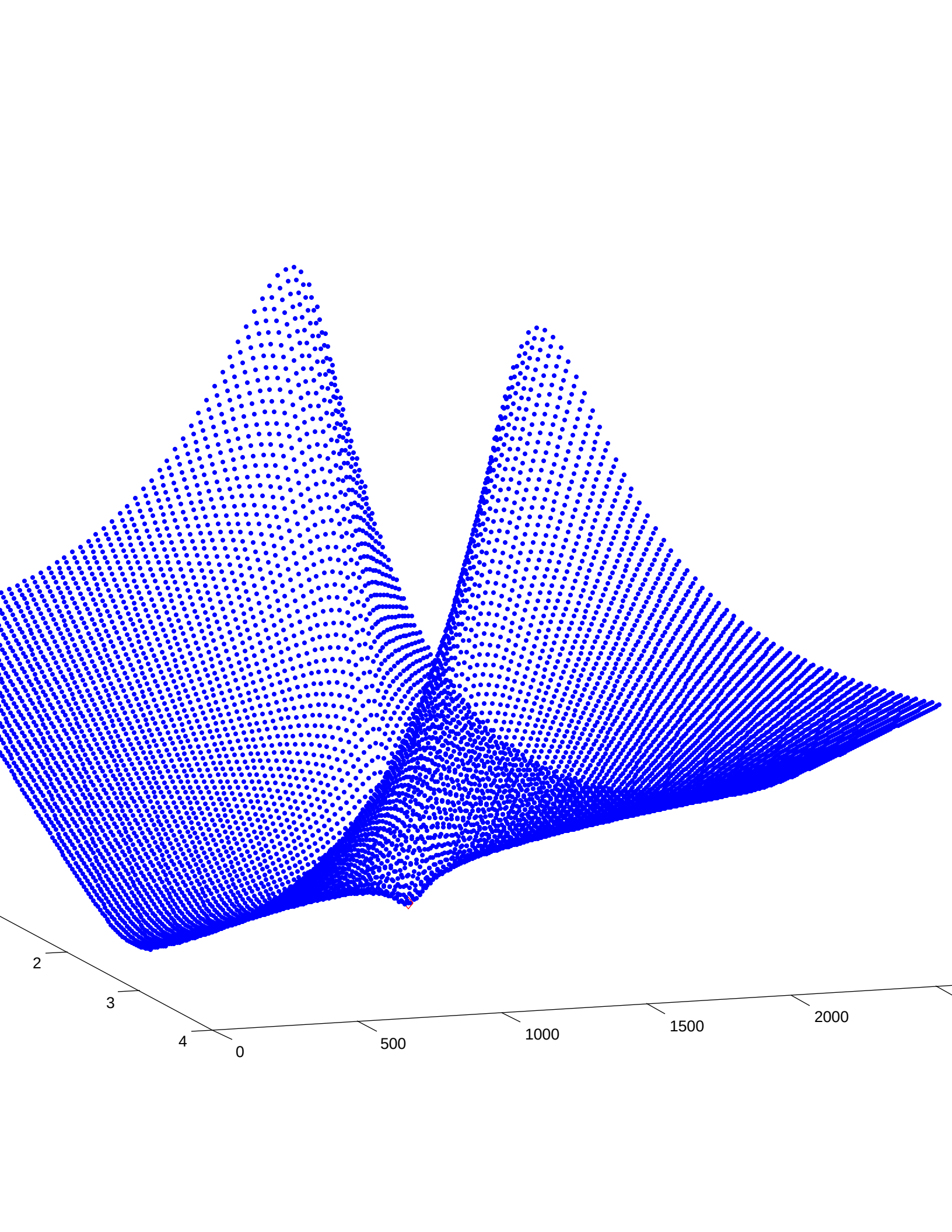
This paper provides conditions which ensure that the inequality constraints generated by either single agent optimizing behavior, or by the best response functions of problems with interacting agents, can be used in estimation. The conditions do not place any restrictions on the choice sets of the agents, or on what the agents know about either the exogenous conditions that will be a determinant of their profits or about their competitors' play.

If agents maximize expected returns conditional on their information sets, then profit realizations will contain a set of disturbances whose expectations, conditional on those information sets, are necessarily zero. These disturbances, together with any conditional mean zero measurement and/or approximation error in the measurement of profits, generate our ν_1. The distribution of ν_1 can be quite complex as it depends on the information sets of agents and, in multiple agent problems, on the details of the equilibria selected by the market participants. However the fact that the realizations of ν_1 have zero conditional expectations allows us to form estimators which account for this complexity without ever specifying these details or computing an equilibria.

The other possible source of error is a difference between the agent's conditional expectation of the profit variable and the conditional expectation that is implicit in the researchers' parametric structural model for realized profits, a difference which we label ν_2. We provide conditions which suffice to obtain inequality constraints for the parameters of interest when both types of disturbances are present. The conditions do impose restrictions on the structure of these errors, but they do not require a parametric specification for the form of the joint distribution of ν_1 and ν_2, and allow for endogenous regressors and discrete choice sets.

We then add to the growing literature on estimation subject to inequality constraints by providing a procedure for generating conservative inference on the boundary points of an identified set; a procedure which can be used to provide dimension by dimension confidence intervals for parameters of interest. An empirical example illustrates the usefulness of our assumptions, and a Monte Carlo example with a sample design based on the empirical example compares the dimension by dimension confidence intervals obtained from direct estimation of the endpoints to those obtained from projecting the confidence sets from other procedures onto alternative axis. It indicates that different methods for constructing confidence intervals are likely to be preferred in different situations.

A number of unanswered questions remain. Among them, we would like to know necessary, as well as sufficient, conditions for our generating the moment inequalities useful for inference. We have not investigated either what can be learned if we replace the parametric structural model of profits with a non-parametric one, or how to combine the many moment inequalities our assumptions generate to achieve more desirable estimators (and the use of moments inequalities, instead of equalities, is likely to accentuate precision problems). Still our example makes it clear that the framework proposed here enables us to obtain information on parameters of interest in environments where estimation has proven difficult in the past, and which are of significant applied interest.


2
3
4
0
500
1000
1500
2000

References

Andrews, D., S. Berry, and P. Jia (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location," manuscript, Yale University.

Andrews, D., and P. Guggenberger (2009): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory,* 25(03), 669-709..

Andrews, D., and P. Guggenberger (2010): "Applications of Subsampling, Hybrid and Size-Correction Methods ," *Journal of Econometrics,* 158(2), 285-305..

Andrews, D., and S. Han (2010): "Invalidity of the Bootstrap and the m out of n Bootstrap for Confidence Interval Endpoints Defined by Moment Inequalities," *The Econometrics Journal,* 12, 172-199..

Andrews, D., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Cowles Foundation Discussion Paper 1676, Yale University..

Andrews, D., and X. Shi (2010): "Inference Based on Conditional Moment Inequalities," Cowles Foundation Discussion Paper 1761, Yale University..

Andrews, D., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica,* 78(1), 119-157.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models" *Econometrica,* 76(4), 763-814..

Steven Berry and Peter Reiss (2007), "Empirical Models of Entry and Market Structure", Chapter 29 in the *Handbook of Industrial Organization*, vol. 3, Mark Armstrong and Robert Porter, eds. North-Holland Press.

Bresnahan T. (1987) "Competition and Collusion in the American Automobile Industry: The 1955 Price War," *Journal of Industrial Economics,* vol. 35, no. 4, pp. 457-482.

Bugni, F. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735-753.

Canay, I. (2010): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408-425.

Card, D., and A. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84(4), 772-793.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243-1284.

Ciliberto, F. and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791-1828.

Crawford, G. and A. Yurukoglu (2010): "The Welfare Effects of Bundling in Multichannel Television Markets," forthcoming, *American Economic Review*.

Dickstein, M. and E. Morales (2011): "A Computational Approach to Moment Inequality Estimation," manuscript, Harvard University.

Fan, Y. and S. Park (2009): "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets," in *Nonparametric Econometric Methods (Advances in Econometrics)*, ed by T.B. Fomby and R.C. Hill, 25, 3-70.

Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification Testing under Moment Inequalities," *Economics Letters,* 99(2), 375-378.

Hansen, L., and K. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50, 1269-86.

Hirano, K., and J. Porter (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77(5), 1683-1701.

Ho, K. (2009): "Insurer-Provider Networks in the Medical Care Market," *American Economic Review,* 99(1): 393-430.

Ho, K., and A. Pakes (2011): "Physician Responses to Financial Incentives: Evidence from Hospital Discharge Data," manuscript, Columbia University.

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

Imbens, G., and C. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-1857.

Ishii, J. (2011): "Compatibility, Competition, and Investment in Network Industries: ATM Networks in the Banking Industry," manuscript, Stanford University.

Katz, M. (2007): "Supermarkets and Zoning Laws," *Unpublished Ph.D. Dissertation,* Harvard University.

Lee, R., and A. Pakes (2009): "Multiple Equilibria and Selection by Learning in an Applied Setting," *Economics Letters,* 104(1), 13-16.

Luttmer, E. (1996): "Asset Pricing in Economies with Frictions," *Econometrica*, 64(6), 1439-1467.

Menzel, K. (2008): "Estimation and Inference with Many Moment Inequalities," manuscript, MIT.

MORALES, E. (2011): "Gravity and Extended Gravity: Estimating a Structural Model of Export Entry," *Unpublished Ph.D Dissertation* Harvard University.

PAKES, A. (2010): "Alternative Models for Moment Inequalities," *Econometrica*, Vol. 78, No. 6, pp 1783-1822.

POWELL, J. (1986): "Symmetrically Trimmed Least Squares Estimation of Tobit Models", *Econometrica*, 54, 1435-1460.

REGUANT, M. (2011): "The Welfare Effects of Complementary Bidding Mechanisms: An Empirical Analysis of the Spanish Wholesale Electricity Market," manuscript, MIT.

ROMANO, J., AND A. SHAIKH (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference - Special Issue in Honor of Ted Anderson*, 138, 2786-2807.

ROMANO, J., AND A. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169-211.

STOYE, J. (2010): "Partial Identification of Spread Parameters," *Quantitative Economics*, 1(2), 232-357.

WOLAK, F (2001); "Identification and estimation of cost functions using observed bid data: an application to electricity markets", NBER Working Paper, 2001 -

6 Appendix

6.1 Example 4 Generalization

This appendix generalizes example 4. Now instead of estimating the unidimensional parameter θ we will be interested in estimating $f(z_i, \theta)$ and instead of using just a constant term as our instrument we will use the positive valued $h(z_i)$. Recall that $L = \{i : d_i > 0\}$, $x_i^l = \mathcal{E}[\Delta r(d_i, d_i - 1, \cdot)|\mathcal{J}_i]$ and $x_i^r = \mathcal{E}[\Delta r(d_i, d_i + 1, \cdot)|\mathcal{J}_i]$, and that our model implies that $\forall i$, $x_i^r - f(z_i, \theta) \leq 0$, while for $i \in L$, $x_i^l - f(z_i, \theta) \geq 0$.

We assume that $h(\cdot)$ is an instrument, or $N^{-1} \sum \nu_{2,i} h(z_i) \to_P 0$, and that the distribution ν_2, $F(\cdot)$, is symmetric so that for any $0 \geq q \geq 1$, $F^{-1}(q) = -F^{-1}(1 - q)$. To proceed as we did in the example for this, more general, case we need to find an upper bound for $N^{-1} \sum_i (x_i^l - f(z_i, \theta)) h(z_i)$. As in the example we have

$$N^{-1} \sum_i (x_i^l - f(z_i, \theta)) h(z_i) \{i \in L\} \geq N^{-1} \sum_i \nu_{2,i} h(z_i) \{i \in L\}.$$

So what is needed is an upper bound for $\sum_i \nu_{2,i} h(z_i) \{i \notin L\}$.

Order the observations so that $i \notin L$ preceed all others and for $i \notin L$ the observations are ordered by their values of $h(z_i)$, and let r permute the i index so that

$r(1) = \{i : i = max_{i=1,...,N}\nu_{2,i}\}$, $r(2)$ is the second largest $\nu_{2,i}$ and so on. Consequently $N^{-1}\sum_i \nu_{2,i}h(z_i)\{i \notin L\} \leq N^{-1}\sum_i \nu_{2,r(i)}h(z_i)\{i \notin L\}$, and from symmetry

$$N^{-1}\sum_i \nu_{2,r(i)}h(z_i)\{i \notin L\} - N^{-1}\sum_i -\nu_{2,(n-r(i)+1)}h(z_i)\{i \notin L\} \to_P 0$$

by summetry of $F(\cdot)$. So for N large enough

$$N^{-1}\sum_i -\nu_{2,(n-r(i)+1)}h(z_i)\{i \notin L\} \geq N^{-1}\sum_i \nu_{2,i}h(z_i)\{i \notin L\}$$

with arbitrarily large probability. Now let $j(i)$ permute the value of the i index such that $j(1) = \{i : i = max_{i=1,...,N}x_i^r\}$, $\jmath(2)$ is the second largest x_i^r and so on. Then from our modeling assumptions

$$N^{-1}\sum_i (f(z_{j(i)},\theta) - x^r_{j(i)})h(z_i)\{i \notin L\} \geq N^{-1}\sum_i -\nu_{2,(n-r(i)+1)}h(z_i)\{i \notin L\}.$$

It follows that for N large enough

$$N^{-1}\sum_i (f(z_{j(i)},\theta) - x^r_{j(i)})h(z_i)\{i \notin L\} \geq N^{-1}\sum_i \nu_{2,i}h(z_i)\{i \notin L\},$$

with arbitrarily large probability. ♠.

6.2 Econometric Assumptions

Assumption A1 *(a) Θ is compact; and for all $F \in \mathcal{F}$ (b) for some $\epsilon > 0$, $\Theta^\epsilon_{0,F} \subset int(\Theta)$, where $\Theta^\epsilon_{0,F} = \{\theta \in \Theta : \inf_{\theta' \in \Theta_{0,F}} \|\theta - \theta'\| \leq \epsilon\}$; (c) $\Theta_{0,F}$ is closed; (d) $\underline{\theta}_F$ is a singleton.*

Assumption A2 *For any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{F\in\mathcal{F}} \inf_{\theta \in (\Theta^\epsilon_{0,F})^c} \| \Big(\mathcal{P}_F m(z,\theta)\Big)_- \| > \delta.$$

Assumption A3 *Define, for each $F \in \mathcal{F}$, $\mathcal{T}_F = \left\{\frac{\theta - \underline{\theta}_F}{\|\theta - \underline{\theta}_F\|} : \theta \in \Theta_{0,F}, \theta \neq \underline{\theta}_F\right\}$. Let $\bar{\delta}$ $= \inf\{\tau_1 : \tau \in \mathcal{T}_F, F \in \mathcal{F}\}$. Assume $\bar{\delta} > 0$.*

Assumption A4 *For some $\underline{\delta} > 0$, there exists $\eta_\Gamma, \varepsilon_\Gamma > 0$ and for each F there is λ_F with $\|\lambda_F\| = 1$ such that (a) $\inf_F \min_{j:D_{j,F}^{-1/2}\mathcal{P}_F m_j(z,\underline{\theta}_F)<\eta_\Gamma} D_{j,F}^{-1/2}\underline{\Gamma}_{j,F}\lambda_F > \varepsilon_\Gamma$; (b) $\sup_F \sup_{\tau:\tau_1\leq\underline{\delta},\|\tau\|=1} \min_{j:\mathcal{P}_F m_j(z,\underline{\theta}_F)=0} D_{j,F}^{-1/2}\underline{\Gamma}_{j,F}\tau < -\varepsilon_\Gamma$, where the index j is running over the elements of the vector m^a.*

Assumption A5 *For any $\delta > 0$,*

$$\sup_F \Pr_F \left(\sup_{\theta\in\Theta} \|\mathbb{P}_{J,F} m(z,\theta) - \mathcal{P}_F m(z,\theta)\| \geq \delta\right) \longrightarrow 0.$$

Assumption A6 *(a) For some* $\underline{d} > 0$ *and* $\delta > 0$, $\underline{d} \leq \inf_F \min_j Var_F(m_j(z, \underline{\theta}_F))$ *and* $\sup_F \max_j \mathcal{P}_F \|m(z, \underline{\theta}_F)\|^{2+\delta} < \infty$*;*
(b) for any $\delta > 0$,

$$\sup_F \Pr_F \left(\left\| \hat{D}_{J,F}^{-1/2} D_F^{1/2} - I \right\| \geq \delta \right) \longrightarrow 0.$$

Assumption A7 *There exists* $\nu > 0$ *such that for all* F, $\mathcal{P}_F m(z, \theta)$ *is continuously differentiable in the neighborhood* $\mathcal{N}_\nu^{\underline{\theta}_F}$ *and there exists* $C < \infty$ *such that* $\| \frac{\partial}{\partial \theta} \mathcal{P}_F m(z, \theta) \| \leq C$ *for* $\theta \in \mathcal{N}_\nu^{\underline{\theta}_F}$*. For any* $\delta \downarrow 0$,

$$\sup_F \sup_{\theta : \|\theta - \underline{\theta}_F\| < \delta} \left\| \frac{\partial}{\partial \theta} \mathcal{P}_F m(z, \theta) - \frac{\partial}{\partial \theta} \mathcal{P}_F m(z, \underline{\theta}_F) \right\| = o(\delta)$$

and there exists $\eta_d > 0$ *such that*

$$\sup_F \sup_{\theta' : \|\theta' - \underline{\theta}_F\| < \eta_d} \sup_{\theta : \|\theta - \theta'\| < \delta} \left\| \left[\mathcal{P}_F m(z, \theta) - \mathcal{P}_F m(z, \theta') - \frac{\partial}{\partial \theta} \mathcal{P}_F m(z, \theta')(\theta - \theta') \right] \right\| = o(\delta).$$

Assumption A8 *For any* $\delta > 0$ *and all sequences* $\eta \downarrow 0$,

$$\sup_F \Pr_F \left(\sup_{\theta : \|\theta - \underline{\theta}_F\| \leq \eta} \left\| \sqrt{J} \left[\mathbb{P}_{J,F} m(z, \theta) - \mathcal{P}_F m(z, \theta)) - (\mathbb{P}_{J,F} m(z, \underline{\theta}_F) - \mathcal{P} m(z, \underline{\theta}_F)) \right] \right\| \geq \delta \right) \longrightarrow 0$$