

Moment Inequalities and Their Application

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

November, 2006
(First Version, August 2004)

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 specification test and methods of inference for the boundaries of the model's identified set. Two applications illustrate how the use of inequality constraints can simplify the problem of obtaining estimators from complex behavioral models of substantial applied interest.

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

*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.

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 (and we allow for incomplete information) 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 by providing a new specification test and inferential procedures for boundary points of the model's identified set. Both the test and the confidence intervals we introduce are easy to construct. Second, though we provide sufficient conditions for the inferential procedures, we do not have necessary conditions, and hence do not know the limits of our framework. Moreover, there remains the question of precisely which of the models typically used to structure data satisfy our sufficient conditions. We provide two empirical applications which are helpful in this respect as they allow us to examine in detail how our general framework can be specialized to two cases used extensively in applied work. The examples also illustrate how the use of inequality constraints can simplify the problem of obtaining estimators of the parameters of complex behavioral models. Finally, both examples generate empirical results of considerable substantive interest.

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. Note that when there are interacting agents these inequalities are necessary conditions for a Nash equilibria, but there may be many vectors of decisions that satisfy them.

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. When there are multiple interacting agents the approximations would typically de-

pend upon the other agents' choices. Note that to compute the approximation we need a procedure which accounts for any change in other determinants of returns that would result from the agent making the alternative choice.

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 actual returns and has an approximated value given by the approximating return function. The difference between the actual and approximated values is then defined as the disturbance. That is, the disturbance is given by the difference between the actual differenced returns of the agent and the difference in the econometrician's approximation of returns when evaluated at the true value of the parameter vector.

We decompose this disturbance term into two parts. The first part arises because agents need not know exactly what their returns will be when they make their decisions (due to asymmetric or other forms of incomplete information), or it could arise simply due to measurement error in the approximation of returns. The defining characteristic of the first part of the disturbance is that it is mean independent of the variables known to the agent when its decision is made. The second part of the disturbance is a structural error that *is* known to the agent when it makes its decision, but is not observed by the econometrician. Since the agent knows the value of the structural disturbance when it makes its choices, the choice itself may depend on this part of the disturbance. Thus when we observe a decision we know that the value of the structural disturbance must have been "selected" from the subset of its possible values that would lead to that decision. As a result, the expectation of the structural disturbance conditional on the observed decision can be nonzero and without further conditions the nonnegativity of the expected difference in actual returns may *not* extend to the expected difference in the econometrician's parametric approximation of returns. That is, the theory only implies that the sum of the expected difference between the parametric difference in returns and the structural error is nonnegative, but we can only construct an approximation to the first term in that sum since the structural disturbance is unobservable.

We provide a sufficient condition for overcoming this hurdle. There are at least two ways of fulfilling this condition. First, in some models, certain linear combinations of differenced returns will not depend on the structural disturbance (it is "differenced out"). Our examples of this case all involve multiple decisions determined by the same structural error. They include limited dependent variable panel data models with choice specific fixed ef-

fects, and IO or social network models with market or network specific effects. The second possibility results from an ability to choose a linear combination of differenced returns that is additive in the structural error 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 have an expectation which is nonpositive at the true value of the parameter vector. We show that we can do this in both ordered choice models and in contracting problems when not all the determinants of the payments specified in the contract are observed by the econometrician.

When there is no structural disturbance our framework is a natural extension of the estimation framework proposed in Hansen and Singleton (1982). Hansen and Singleton consider first order conditions for a continuous control. We replace this condition by a “revealed preference” difference inequality, which enables us to consider unrestricted choice sets. Hansen and Singleton consider a single agent expected utility maximization problem. We replace this by the Nash equilibrium condition of a game. 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.

A second strand in the literature that is related to our work is the work on entry games in Ciliberto and Tamer (2003) and Andrews, Berry, and Jia (2004). As in our work, when there is the possibility of multiple equilibria these papers do not take any stance on equilibrium selection (for an earlier paper based on similar ideas see Tamer (2003)). Moreover, they also allow for the possibility that the multiple equilibria may lead to only partial identification of the parameters of interest. However, there is a sense in which this work makes exactly the opposite assumptions to those made in Hansen and Singleton (1982). Following the previous entry game literature,¹ Ciliberto and Tamer and Andrews, Berry, and Jia assume that there is no expectational or measurement error; i.e. that *all* of the unaccounted variance

¹See for example Brenahan and Resiss (1990) and Berry (1992).

in realized returns is due to factors that the agent knew when it made its decision but the econometrician does not observe. These papers then make a parametric assumption on the distribution of this structural disturbance and specify precisely what each agent knows about the determinants of the returns from its actions (including what the agent knows about the distribution of its competitor's actions). This enables them to develop estimators that apply to entry games with endogenous regressors and allow for the possibility of multiple equilibria. The estimators require the econometrician to check the necessary equilibrium conditions for the observed choices each time a parameter vector is evaluated in the search routine and as a result often have a much larger computational burden than the estimators proposed here.²

The prediction of our framework that we take to data is an expectational inequality based on observable variables and a known parametric returns function. That is, our framework generates moment *inequalities*. Moment inequality models and, more generally, partially identified models have received considerable recent interest in the econometrics literature. Manski and co-authors have advanced the literature on partial identification through a series of papers (and a book), including Manski (1990, 2003), Horowitz and Manski (1998, 2006), Imbens and Manski (2004), Manski and Pepper (2000), and Manski and Tamer (2002). Hansen, Heaton, and Luttmer (1995) develop moment inequality tests in the context of asset pricing models with market frictions. Moon and Schorfheide (2004) consider the case of point identified models of moment equalities and inequalities. Both Chernozhukov, Hong, and Tamer (2003) and Andrews, Berry, and Jia (2004) establish general identification, estimation, and inference results for moment inequality models. The methods in these papers are directly applicable to the framework developed formally below. Other recent and directly applicable inferential methods (to our framework) are developed in Shaikh (2005), Rosen (2005), Guggenberger, Hahn, and Kim (2006), and Soares (2006).

We add to the econometric literature on moment inequality inference with an emphasis on computationally simple methods for practitioners. In

²Pakes (2005) provides a more detailed comparison of the discrete games applications of the framework provided in this paper and the framework that is implicit in the articles cited above (including a Monte Carlo comparison of performance and computational burdens). We should note that there has also been some work on discrete games which use more detailed specifications and typically result in full, rather than partial, identification of parameters; see Bajari, Hong, Ryan (2004) and Bajari, Hong, Krainer, and Nekipelov (2006).

this literature the set of parameters that satisfy the inequality constraints is called the “identified set.” We focus on estimation of a boundary point of the identified set. By taking this focus, we are able to derive an explicit form of the asymptotic distribution of the most natural estimator. While this limit distribution is not available directly for inference, its form suggests two straightforward simulation methods, requiring only linear programming, for obtaining confidence regions. One simulation method provides conservative inference, while the other provides an upper bound on one source of conservatism in the first method. Moreover, under certain regularity conditions, we expect the confidence interval produced by the second method to have better coverage properties (and be shorter) for large enough samples.

We also consider a specification test of the moment inequalities. Specification testing is likely to be important in our context. We expect practitioners to want to be able to test for the importance of allowing for structural disturbances, and inequality tests are likely to be more robust to small deviations in modelling assumptions than tests of a point null hypothesis. The test statistic is obtained by modifying the logic of traditional tests of overidentifying restrictions in method of moment models for the presence of inequalities. We develop a simple method for obtaining conservative critical values. Our procedures can be simplified further when the moment conditions are linear, and we conclude the econometric section with a brief overview of this case.

While the econometric approach described is not computationally demanding, both the precision of inference and the relevance of our behavioral assumptions are still open questions. Our two empirical applications are both informative and encouraging in this respect. They are both problems: (i) which could not have been analyzed with more traditional tools, and (ii) with sample sizes that are quite small. The small sample sizes do force us to use parsimonious specifications. However, the results make it quite clear that the new techniques provide useful information on important parameters; information which could not have been unraveled using more traditional estimation methods.

The first example shows how our setup can be used to analyze investment problems with 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 multiple agent environments in which there can be many possible “network” equilibria. Formally this example is a multiple agent ordered choice problem; a problem which arises in many different contexts in industrial organization. This example

also illustrates the intuition underlying the estimators' properties particularly clearly, and provides rather robust estimates of a parameter of considerable policy importance.

The second example illustrates how the proposed approach can be used to analyze the nature of contracts emanating from a market with a small number of *both* buyers and sellers. Though markets with a small number of buyers and sellers appear frequently in industrial organization, econometric analysis of their equilibrium outcomes had not been possible prior to this work without restrictive simplifying assumptions. Our particular example analyzes the nature of the contracts between health insurance plans and hospitals, a largely unstudied problem of particular importance in determining the investment incentives facing hospitals.

In both examples, the results we obtain are compared to alternative estimators that come to mind for the respective problems. In one example the alternative procedure ignores endogenous regressors. In the other, one of the two alternatives assumes away the non-structural error in the profit measures and the other alternative assumes away the discreteness in the choice set. The empirical results make it clear that accounting for both endogenous regressors and non-structural errors in discrete choice problems can be extremely important. The more detailed substantive implications of the parameter estimates are discussed in Ho (2004) and Ishii (2004).

The next section begins with the assumptions underlying our framework. It then shows how these assumptions lead to moments that have nonnegative expectation when evaluated at the true value of the parameter vector and concludes by reviewing some familiar examples which satisfy our assumptions. Section 3 takes our inequality condition as a starting point and provides methods of inference for the parameter vector. Section 4 applies these techniques to two empirical problems that are of substantive interest and could not have been analyzed using more traditional techniques (at least not without further assumptions).

2 A Framework for the Analysis

This section derives the moment inequalities that serve as the basis for econometric inference. The setting for this derivation is a Nash equilibrium to a simultaneous move game in pure strategies. Within this framework, our assumptions do not restrict choice sets nor require a unique equilibrium, and

we allow for both incomplete and asymmetric information. After stating the assumptions, we illustrate their use with some familiar examples, and then show how they generate the moment inequalities for use in the econometric analysis. We conclude with a number of generalizations which show how to allow for: mixed strategies, various forms of non-Nash behavior, and non-simultaneous moves.

2.1 Agents' Problem

Suppose players (or “agents”) are indexed by $i = 1, \dots, n$. Let \mathcal{J}_i denote the information set available to agent i before any decisions are made, where $\mathcal{J}_i \in \mathcal{I}_i$, the space of such information sets. \mathcal{D}_i will be the set of possible decisions by agent i , and $s_i : \mathcal{I}_i \rightarrow \mathcal{D}_i$ will denote the strategy actually played by agent i . The observed decision \mathbf{d}_i is generated by i 's strategy and its information set, that is, $\mathbf{d}_i = s_i(\mathcal{J}_i)$ when i plays strategy s_i . For now we assume these are pure strategies, so \mathcal{D}_i is the set of possible values or decisions (the support) for \mathbf{d}_i .³ 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.⁴

Let payoffs (or profits) to agent i be given by the function $\pi : \mathcal{D}_i \times \mathcal{D}_{-i} \times \mathbf{Y} \rightarrow \mathcal{R}$, where \mathcal{D}_{-i} denotes $\times_{j \neq i} \mathcal{D}_j$. In particular, returns to i are determined by agent i 's decision, d_i , other agents' decisions, d_{-i} , and an additional set of variables $y_i \in \mathbf{Y}$. Not all components of y_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.

Let \mathcal{E} be the expectation operator and \mathbf{y}_i be the random variable whose

³Some of our examples require us to distinguish agents of different types (e.g. buyers and sellers in buyer-seller networks), but we refrain from introducing a type index until we need it.

⁴For 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 “exclusives” which ensure a single vendor per market, \mathcal{D}_i becomes a proper subset of the product space of all possible two part tariffs.

realizations are given by y_i .⁵ The following assumption characterizes the behavior of agents in the game.

Assumption 1 (Nash Condition.) *If s_i is the strategy played by agent i , then*

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

for $i = 1, \dots, n$.⁶ ♠

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. In section 2.4, we discuss the relaxation of Assumption 1 to cover certain kinds of sub-optimal behavior.

Counterfactuals

In many examples the realization of \mathbf{y} will depend on (d_i, d_{-i}) . For example 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 and these interest rates, in turn, depend on the number of ATM's installed by the bank and its competitors. Since we want to compare the profits actually earned to those that would have been earned had the agent made a different decision,

⁵Formally this is the expectation corresponding to the joint distribution of the random variables we define and our assumptions place restrictions on that distribution. We could have defined the expectation operator corresponding to each agent's perceptions, and then assumed that these perceptions are in fact correct for the play that generated our data. Though this is certainly sufficient for Assumption 1 to be true, our assumptions do not require agents to have correct perceptions everywhere.

⁶Formally, we will only make use of a weaker version of Assumption 1. In particular, it will be sufficient that expected returns at the strategy played are higher than expected returns at the particular alternative/counterfactual decisions considered by the econometrician.

we will need an assumption which allows us to determine what y would have been had the agent's decision been different.

This will require additional notation. Let $\mathbf{y} : \mathcal{D}_i \times \mathcal{D}_{-i} \times \mathcal{Z} \rightarrow \mathcal{R}^{\#y}$, so that $\mathbf{y}_i = \mathbf{y}(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i)$ for a random variable \mathbf{z} , with realizations that will be denoted by z . Then we make the following assumption.

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 .*

Conditional independence of other agents' decisions (of \mathbf{d}_{-i}) from \mathbf{d}_i is an implication of play in simultaneous move games. The assumption also requires that, if there is a \mathbf{y} which is endogenous in the sense that its realization depends on d_i , then we have a model of that dependence. This enables us to form an estimate of $\pi(d', d_{-i}, y(d, d_{-i}, z_i))$ for $d' \neq d_i$ that has an expectation which conforms to Assumption 1.

More precisely if we define

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

then Assumptions 1 and 2 imply that for any $d' \in \mathcal{D}_i$

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

Section 2.4 provides a generalization which allows us to analyze some cases where Assumption 2 is not satisfied; both non-simultaneous move games and cases where $y(\cdot)$ does depend on d_i but we do not have a model for that dependence. The generalization is, however, both more computationally burdensome and likely to provide inequalities which are less informative than those that are valid when Assumption 2 is satisfied.

2.2 Econometrician's Problem

The econometrician may not be able to measure profits exactly but can calculate an approximation to $\pi(\cdot)$, say $r(\cdot; \theta)$, which is known up to the parameter vector θ . The function $r(\cdot)$ has arguments d_i, d_{-i} , an *observable* vector of the determinants of profits, say \mathbf{z}_i^o , and θ . The parameter $\theta \in \Theta$ and its true value will be denoted by θ_0 . We obtain our approximation to the difference in profits that would have been earned had the agent chosen d' instead of d , say $\Delta r(d, d', \cdot)$, by evaluating $r(\cdot)$ at d and d' and taking the difference.

More formally $\Delta r(\cdot) : \mathcal{D}_i^2 \times \mathcal{D}_{-i} \times Z^o \times \Theta \rightarrow \mathcal{R}$ is a *known* function of; (d, d') , other agents' decisions, or d_{-i} , our observable determinants of profits, $z_i^o \in Z$, and a parameter $\theta \in \Theta$. Let $\mathbf{z}_i^o \subset \mathbf{z}_i$ be the random variable whose realizations are given by z_i^o . Then the relationships between $\Delta\pi(\cdot)$ and $\Delta r(\cdot)$ and z_i and z_i^o define the following two unobservables.

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

$$\nu_{2,i,d,d'} = \mathcal{E}[\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i] - \mathcal{E}[\Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) | \mathcal{J}_i], \text{ and } (1)$$

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

It follows that

$$\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) = \Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) + \nu_{1,i,d,d'} + \nu_{2,i,d,d'}. \quad (3)$$

The function $\Delta r(\cdot, \theta)$ is our observable measure of the change in profits that would result from a change of $d_i = d$ to $d_i = d'$. ν_1 and ν_2 are the determinants of the true profit difference that are *not observed* by the econometrician. They have different values for every different (d, d') and every agent. We distinguish between two types of unobservables 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. The agent knows its ν_2 values *before* it makes its decision, i.e. $\nu_{2,i} \in \mathcal{J}_i$. Since the realized decision depends on the information set, $\mathbf{d}_i = s_i(\mathcal{J}_i)$, we expect d_i to depend on the values of $\nu_{2,i,d_i,d'}$. Consequently even if the unconditional mean of $\nu_{2,i,d_i,d'}$ is zero, an observation of $d = d_i$ tells us something about the realization of $\nu_{2,i,d_i,d'}$. As a result the conditional mean of $\nu_{2,i,d_i,d'}$ given $\mathbf{d}_i = s_i(\mathcal{J}_i)$ (where s_i satisfies Assumption 1) will not generally be zero, i.e. $\mathcal{E}[\nu_{2,i,d_i,d'} | \mathbf{d}_i = s_i(\mathcal{J}_i)] \neq 0$.

In contrast, the agent's decision does not depend on $\nu_{1,i}$. These random variables are all mean zero conditional on the information set and so do not affect the conditional expectation of profits that determine decisions (via Assumption 1). Moreover since $\mathcal{E}[\nu_{1,i,d_i,d'} | \mathcal{J}_i] = 0$ by construction and \mathbf{d}_i is a function of the variables in \mathcal{J}_i for the realized decisions, the conditional mean given the information set and decision is zero, $\mathcal{E}[\nu_{1,i,d_i,d'} | \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)] = 0$.

The importance of accounting for one or both of (ν_1, ν_2) is likely to be different in different applied problems. Differences between ν_1 and zero do not change the agent’s expected profits at the time decisions are made. So ν_1 realizations can be caused by either expectational or measurement errors. There are two sources of expectational errors: (i) incomplete information on the environmental variables that will determine the profits that result from the agent’s decision, and (ii) asymmetric information possibly resulting from incomplete information on either the z_{-i} ’s or the $\nu_{2,-i}$ ’s that determine the decisions of the agent’s competitors. Measurement error in profits can result from measurement error in the components of either revenues or costs.

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 $\Delta \pi(d, d', \cdot)$ that the agent accounts for when it makes its decision (we are more explicit about how this can happen in discussing our examples).

A few other points about these definitions are worth noting. Typically it is difficult for researchers to know what agents know about each other. In this context we note that we have not had to specify whether $(z_{-i}, \nu_{2,-i})$ is in agent i ’s information set at the time decisions are made. The information set, \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.⁷ Relatedly we need not make a particular assumption on the relationship of the $\{\nu_{2,i}\}$ draws of the different agents in a given market.⁸

We have assumed that the \mathbf{z}_i^o which determines $\Delta r(\cdot)$ is the observable part of \mathbf{z}_i which determines the $\Delta \pi(\cdot)$. Alternatively, the relationship between \mathbf{z}_i^o and \mathbf{z}_i could have been left unrestricted but then we would also need to extend Assumption 2 to hold for \mathbf{z}_i^o as well. There is a sense then that including \mathbf{z}_i^o in \mathbf{z}_i is just a notational convenience (it allows us to state

⁷The fact that we need not be explicit about the contents of information sets differentiates our setup and from the setups used in most prior applied work in Industrial Organization. For example, Bresnahan and Reiss (1991) and Berry (1992) assume full information; Seim (2002) and Pakes, Ostrovsky, and Berry (2003) assume no knowledge of $\nu_{2,-i}$; and Fershtman and Pakes (2004) allow for signals. Of course if we knew (or were willing to assume) more on the properties of the $\nu_{2,i}$ we might well be able to provide more precise estimators of θ (see, for example, Bajari, Hong, and Ryan 2004).

⁸Note however that the combination of assuming that $\nu_{2,-i}$ is unobservable while z_i^o is observable, and of not making $\Delta r(\cdot)$ a function of $\nu_{2,-i}$, implies that the ν_2 ’s of the firm’s competitors only affects its profits indirectly, through $\nu_{2,-i}$ ’s effects on (z_i^o, d_i, d_{-i}) .

Assumption 2 just in terms of \mathbf{z}_i). Note, however, that if $\mathbf{z}_i = (\mathbf{z}_i^o, \nu_{1,i}, \nu_{2,i})$, then the \mathbf{z}_i^o are the observable determinants of the profit differences, the $\nu_{2,i}$ are the unobservable determinants of profit differences that the agent knows when it makes its decisions, and the $\nu_{1,i}$ are the unobservable determinants that the agent did not know when it made its decision. In this case the unobservable determinants of profitability that the agent knew when decisions were made enter profit differences in an *additively separable* way. If this is not the case and we wanted to derive $\nu_{2,i}$ from more detailed assumptions on primitives, then we would need to explicitly consider the expectation in equation (1).

Selection

Our assumptions thus far are not very stringent. In addition to not assuming what each agent knows about its competitors, 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 endogenous regressors. 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, 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, s_i is a strategy satisfying Assumption 1 only if $\nu_{2,i,s_i(\mathcal{J}_i),d'} \geq -\mathcal{E}[\Delta r(s_i(\mathcal{J}_i), d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) | \mathcal{J}_i]$, so draws on $\nu_{2,i}$ corresponding to the observed decisions are *selected* from a subset of the support of the ν_2 distribution.

The next assumption offers a route for overcoming this selection problem. The observables enable us to form sample means of nonnegative linear combinations (over alternative decisions) of our observed proxies for the profit differences (of $\Delta r(\mathbf{d}_i, d', \cdot; \theta)$ given $\mathbf{d}_i = s_i(\mathcal{J}_i)$), and consider values of θ which make the sample averages positive. Assumption 1 ensures that the analogous linear combinations of the $\Delta \pi(\mathbf{d}_i, d', \cdot)$ have a positive conditional expectation. Equation (1) then implies that the conditional expectation of the observable linear combination of $\Delta r(\mathbf{d}_i, d', \cdot; \theta)$ values will be positive at $\theta = \theta_0$ provided the conditional expectation of $\nu_{1,i,\mathbf{d}_i,d'}$ and $\nu_{2,i,\mathbf{d}_i,d'}$ are not positive. If the weights in the linear combination are functions of the agents' information sets, the definition of $\nu_{1,i,\mathbf{d}_i,d'}$ ensures that the relevant linear combinations of the $\nu_{1,i,\mathbf{d}_i,d'}$ will have zero conditional expectation. Assumption 3 provides conditions which suffice to ensure that the conditional

expectation of the same linear combination of the $\nu_{2,i,\mathbf{d}_i,d'}$ is not positive.

Assumption 3 constrains the relationship between the ν_2 and the $\mathcal{E}[\Delta r(\cdot)|\mathcal{J}]$ in equation (1). Special cases of this assumption occur when we can find a linear combination of the $\Delta r(\cdot)$ that either does not involve ν_2 , or generates the same ν_2 value no matter the realization of \mathbf{d}_i (for this to occur the $\nu_{2,i,d,d'}$ values must typically be constrained in some fashion). In the latter case we employ instruments to account for possible correlations between ν_2 and the other observable determinants of profits (e.g. $(\mathbf{d}_{-i}, \mathbf{z}_i^o)$). A third case occurs when the linear combination of ν_{2_i} 's have a negative correlation with a $\mathbf{x}_i \in \mathcal{J}_i$ conditional on $\mathbf{d}_i = s_i(\mathcal{J}_i)$. After presenting Assumption 3 we consider four familiar examples which clarify how it can be used.

Assumption 3 *Let h be a function which maps x_i into a nonnegative Euclidean orthant. Assume that for an \mathbf{x}_i that is both in \mathcal{J}_i and is observed by the econometrician, and a nonnegative weight function $\chi_{\mathbf{d}_i, \mathcal{J}_i}^i : \mathcal{D}_i \rightarrow \mathbb{R}^+$ whose value can depend on the realization of \mathbf{d}_i (and the information set \mathcal{J}_i)*

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \nu_{2,i,\mathbf{d}_i,d'} h(\mathbf{x}_i) | \mathbf{d}_i = s_i(\mathcal{J}_i)\right] \leq 0,^9$$

where $\nu_{2,i,\mathbf{d}_i,d'} = \sum_{d \in \mathcal{D}_i} \mathbf{1}\{\mathbf{d}_i = d\} \nu_{2,i,d,d'}$.

This assumption 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.¹⁰ In particular since both the choice set and the distributions of the unobservables are unspecified, Assumption 3 allows us to analyze some models with discrete choice sets and endogenous regressors without making particular distributional assumptions (examples are given below). Recall that strategy s_i is defined as the strategy that will generate the decisions for agent i actually observed by the econometrician (i.e. in the data). So, by conditioning on the event $\mathbf{d}_i = s_i(\mathcal{J}_i)$ in the expectation, we are simply focusing on the expectation corresponding to the distribution of

⁹An inequality applied to a vector means the inequality holds for every element of the vector.

¹⁰More stringent assumptions about information sets and distributions of disturbances can lead to alternative estimators than those considered here; see Pakes (2005) for a discussion. The alternatives are quite a bit more computationally burdensome than the estimators discussed below.

decisions \mathbf{d}_i actually observed by the econometrician. Note also that fixing the set of realizations of variables observed by the econometrician does not fix the agents' information sets. So agents with the same set of observables can make different decisions. Finally though there is a sense in which the \mathbf{x} play the role of an "instrument" in prior work, our "instrument" need not generate a traditional zero correlation moment *equality*; we require only an inequality. In particular, it is sufficient for \mathbf{x} (actually $h(\mathbf{x})$) to be negatively correlated with the unobservable known to the agent when its decision is made (ν_2).

In the examples below, it will be convenient to take \mathbf{d}_i to always denote the realized decision $\mathbf{d}_i = s_i(\mathcal{J}_i)$, so that it is not used to denote counterfactual decisions. By adopting this convention, we avoid the need to write the expectations below as conditional expectations given $\mathbf{d}_i = s_i(\mathcal{J}_i)$, as in Assumption 3. So, to be clear, in the examples, any expectations involving \mathbf{d}_i are taken with respect to the realized decision distribution, $s_i(\mathcal{J}_i)$ (and so need not include the conditioning event in Assumption 3).

Example 1. Suppose $\pi(\cdot)$ is observable up to a parameter vector of interest and an error which is mean zero conditional on the agent's information set. Formally this is the special case where $\nu_{2,i,d,d'}$ is identically zero for all d, d' , so that Assumption 3 is satisfied with $h(\cdot) = 1$ and any χ^i which weights a $d' \in \mathcal{D}_i$. For example pick any d' and set $\chi^i(d') = 1$ and zero elsewhere. Then

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

and our assumptions are satisfied.

We note that our functional form and stochastic assumptions are then those of Hansen and Singleton (1982), but our estimator: (i) allows for more general (discrete and/or bounded) choice sets; (ii) allows explicitly for interacting agents (without having to fully specify the information structure); and (iii) as we discuss in the generalizations below, allows for agents whose choices are not always exactly optimal conditional on a prespecified information set. We are able to do this because we assume an ability to compute the profits that would have been earned if the alternative actions had been made up to the parameter of interest and a mean zero disturbance (Hansen and Singleton, 1982, assume an ability to calculate the first derivative of expected returns).

Note that these assumptions allow us to apply Euler's perturbation method

to the analysis of single agent dynamic discrete choice problems, and this, in turn, simplifies the econometric analysis of those problems markedly. That is, consider a perturbation to the chosen dynamic program which changes the decisions made in consecutive periods but ensures that in subsequent years the return function is the same (with probability one). Under our assumptions the conditional expectation of the difference between the original and the perturbed program's discounted change in utility over the two periods must be positive at the true θ_0 . This moment inequality enables estimation without ever needing to compute the value function¹¹.

More generally these assumptions are relevant for any problem for which we can measure profits up to a mean zero error, so they constitute a special case we might often want to test for.

Example 2. (*Choice Specific Fixed Effects*) This example assumes the existence of a set of weights such that $\sum_{i=1}^n \sum_{d'} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'} = 0$. Then there are certain (vectors of) decisions (d_i), feasible alternatives (d'_i), and weights $\chi_{d_i}^i(d'_i)$, with the property that $\sum_i \chi_{d_i}^i(d'_i) \Delta\pi(d_i, d'_i, \cdot)$ is positive (in expectation) and does not depend on any of the ν_2 disturbances.

Choice specific fixed effects are probably the most familiar case in point.¹² They occur in both the market level analysis that is familiar from industrial organization, and in the analysis of single agent problems. In the market level analysis the fixed effects represent a determinant of a decision which is unobserved to the econometrician and common to the agents in a given market, but which varies across markets. Our second empirical example contains an illustration of this case.

The individual level analysis either concerns multiple decisions by the same individual in a given period, or panel data problems where the unobserved determinant of a repeated decision varies across agents but is the same for a given agent over time. For simplicity suppose we observe two decisions

¹¹We emphasize that this does assume that $\nu_{2,i,d,t} = 0$, where t indexes time, whereas most of the dynamic discrete choice literature assumes a set of $\{\nu_{2,i,d,t}\}$ which are i.i.d. both across different choices and over time. One alternative is to assume that $\nu_{2,i,d,t} = \nu_{2,i,d}$ for all t and use the choice specific fixed effects estimator described in the next example. Another is to use one of these easy to compute estimators as starting values in a nested fixed point algorithm that allows for a richer joint distribution of the $\{\nu_{2,i,d,t}\}$. Note that, in contrast to the nested fixed point estimators, neither of the two inequality estimators require a particular specification for agents' information sets.

¹²Fixed effects which do not interact with agents' decisions are differenced out in the estimation algorithm, and do not effect the properties of the estimator.

so $d = (d_a, d_b)$ with $d_w \in \{0, 1\}$ for $w = \{a, b\}$, and assume that the profit from decision w is given by

$$\Delta\pi_w(d_{i,w}, d'_{i,w}, \cdot) = \Delta r_w(d_{i,w}, d'_{i,w}, \cdot) + (d_{i,w} - d'_{i,w})\nu_{2,i} + \nu_{1,i,d_{i,w},d'_{i,w}}^w,$$

for $w = \{a, b\}$. In the cross sectional variant w would index different agents in a market and d_b would typically be a determinant of $\pi_a(d_a, d'_a, \cdot)$. In the panel data context the w would index different decisions made by the same agent and one might want to allow the part of the profit function given by $r(\cdot)$ to be non-additive across the agent's choices.¹³ To reflect the choice specific fixed effects case, we have made an assumption that restricts the form of the ν_2 's. In particular, the equation above follows from our more general formulation by assuming there is a $\nu_{2,i}$ such that $\nu_{2,i,d_w,d_w'}^w = (d_w - d_w')\nu_{2,i}$. In this case, we will be able to find weights such that Assumption 3 is satisfied.

Consider setting $\chi_d^i(d')$ to one whenever $d = (1, 0)$ and $d' = (0, 1)$ (or vice versa), and zero otherwise. Then if $\mathbf{1}\{\cdot\}$ is notation for the indicator function

$$\begin{aligned} \sum_{d'} \chi_{d_i}^i(d') \Delta\pi(d_i, d', \cdot) &= \sum_{d'} \chi_{d_i}^i(d') [\Delta\pi_a(d_{i,a}, d'_a, \cdot) + \Delta\pi_b(d_{i,b}, d'_b, \cdot)] \\ &= \mathbf{1}\{d_i = (1, 0)\} \left[\Delta r_a(d_{i,a} = 1, d'_a = 0, \cdot) + \Delta r_b(d_{i,b} = 0, d'_b = 1, \cdot) \right] \\ &\quad + \mathbf{1}\{d_i = (0, 1)\} \left[\Delta r_a(d_{i,a} = 0, d'_a = 1, \cdot) + \Delta r_b(d_{i,b} = 1, d'_b = 0, \cdot) \right] \\ &\quad + \sum_{d'} \chi_{d_i}^i(d') \nu_{1,i,d_i,d'}, \end{aligned}$$

and the last term is mean independent of any $x \in \mathcal{J}_i$. Note that the moments used in estimation for models with choice specific fixed effects are the average of differences, across choices, of the difference in returns between the optimal and an alternative feasible choice; i.e. they are *difference in difference inequalities*.

The crucial assumptions of this example are that there are repeated observations determined by the same (possibly vector of) $\nu_{2,i}$, and that choices depend only on the observed regression function and these $\nu_{2,i}$; i.e. the idiosyncratic disturbance is a result of expectational or measurement errors. With this understanding the example covers many familiar cases including: (i) discrete choice panel data models with agent specific unobservable (fixed)

¹³In the nonadditive case we would write $\Delta\pi(d, d', \cdot) = \Delta r(d, d', \cdot) + [(d_a - d'_a) + (d_b - d'_b)]\nu_{2,i} + \nu_{1,i,d,d'}$.

effects and endogenous regressors (without requiring a parametric distribution for either the $\nu_{2,i}$ or the idiosyncratic disturbance), (ii) models in which there are either returns to, or information flows from, the size of a network and a common unobservable determinant of the choices of agents of whether to join the network, and (iii) models in which there is a distribution of response parameters among agents who make the same choice repeatedly.

Example 3. (*Ordered choice*). This example assumes weights that yield a mean zero unconditional expectation for the ν_2 's, or

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d'} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \nu_{2,i,\mathbf{d}_i,d'}\right] = 0.$$

Ordered choice, or any discrete choice with an order to the choice set and a determinant of the agent's ordering that is not observed by the econometrician (which becomes ν_2), could potentially generate this condition. Lumpy investment decisions (say in the number of stores or machines) are often treated as ordered choice problems, and our first empirical example is a case in point.¹⁴ It has markets consisting of sets of interacting firms each of whom decides how many units of a machine to purchase and install. The parameter of interest (θ) determines the average (across firms) of the cost of installing and operating machines, and the model allows costs to differ across firms in a way which is known to the firm when they make their decisions but not observed by the econometrician (the ν_2).

With θ denoting the average marginal cost across firms and $\theta + \nu_{2,i}$ the constant marginal costs for some firm i , the difference in profits from installing d versus d' machines includes a cost difference equal to $(d - d')(\theta + \nu_{2,i})$ for firm i . So if $r(\cdot)$ provides the revenues, the incremental profits from the next machine bought are

$$\Delta\pi(d_i, d_i + 1, \cdot) = \Delta r(d_i, d_i + 1, \cdot) + (\theta_0 + \nu_{2,i}) + \nu_{1,i,d,d+1}.$$

¹⁴Another case is the vertical discrete choice model introduced by Mussa and Rosen (1978) and used in Bresnahan (1987). This is a consumer discrete choice problem where consumers all agree on the quality of the products available to choose from, prices are increasing in that quality, and there is a consumer specific marginal utility of income which is not observed by the econometrician but does determine the consumers' disutility from price. Differences in consumer utility between the good chosen and a good of higher quality divided by the negative of the price difference between the two goods then become an inequality with properties similar to the profit difference inequality developed below.

Since θ_0 is the average marginal cost across firms, $\mathcal{E}\nu_{2,i} = 0$, and Assumption 3 is satisfied with $h = 1$ and $\chi_{d_i}^i(d') = 1$ only if $d' = d_i + 1$. Consequently

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'}\right] = \sum_{i=1}^n \mathcal{E}[\nu_{2,i,d_i,d_i+1}] = \sum_{i=1}^n \mathcal{E}[\nu_{2,i}] = 0.$$

In this example $\nu_{2,i,d,d+1} = \nu_{2,i}$ regardless of d . Hence our choice of weight function generates an unconditional average of the ν_2 's (it includes a value of ν_2 regardless of the choice), and this ensures that there is no selection problem. Assumptions 1 and 2 then give the needed moment inequality. We provide a fuller discussion of this case, including a discussion of identification, below.¹⁵

Example 4. Suppose the ν_2 's (or a weighted sum of ν_2 's) are mean independent of a *subset* of the variables that the agents know when they make their decisions, a subset which will become our “instruments,” \mathbf{x} ,

$$\mathcal{E}\left[\sum_i \sum_{d'} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \nu_{2,i,\mathbf{d}_i,d'} | \mathbf{x}_i\right] = 0.$$

Example 3 could be extended to include this case by assuming an $\mathbf{x}_i \in \mathcal{J}_i$ that satisfies $\mathcal{E}[\nu_{2,i} | \mathbf{x}_i] = 0$. Our second empirical example is another case, and since it is of some applied interest, it is dealt with explicitly here.

Buyer-seller networks with unobserved transfers. Here we need to introduce a set of types, say $\mathcal{T} = \{\mathbf{b}, \mathbf{s}\}$ for buyers and sellers respectively, and let the choice of inequalities and the form of the primitives (the choice set, profit function, etc.) differ by type. Type \mathbf{b} 's incremental cost is the cost of purchase, and its incremental expected returns are the expected profit from resale. Type \mathbf{s} 's incremental returns are \mathbf{b} 's purchase cost and its incremental costs are the costs of production. Assume that sellers make take it or leave it offers to buyers. Note that since buyers know the sellers' offers before they determine whether to accept, this is our first example which is not a simultaneous move game, and so we will have to adjust our framework (we deal with non-simultaneous move games in more generality in section 2.4). The offers themselves are not public information (they are proprietary), and

¹⁵The discussion above has implicitly assumed that there are no corners to the choice set (there are feasible choices that are higher than every possible observed choice.). The discussion below considers corners in some detail.

it is their properties that we want to investigate empirically. We assume that the offers are a parametric function of observables (e.g. an unknown markup per unit purchased) and an error (ν_2).

For now assume there is only one seller and one buyer in each market studied. \mathcal{D}_s is the set of contracts which can be offered. We assume it includes a null contract, say $d_s = \phi$, that is never accepted. A contract which is not accepted does not generate any profit for the seller. $\mathcal{D}_b = \{0, 1\}$ with $d_b = 1$ indicating the contract was accepted. Note that any transfer cost to the buyer is a revenue for the seller, so there is only one value of ν_2 per market and it enters the profits of the buyer and the seller with opposite signs.¹⁶ Assume that the profits of the buyers and sellers are both known up to measurement error and the value of ν_2 .

Assumption 1 implies that (i) the expected profits to the seller should the contract be accepted are larger than those from the null contract, and (ii) if the buyer rejects the offer it is because profits without the contract are expected to be higher than profits with the contract. If there is a contract the seller earns ν_2 , while if there is no contract the buyer saves ν_2 by rejecting the contract. Thus by taking the change in profits of the seller from offering the contract offered rather than the null contract when there is a contract, and the change in profits to the buyer from rejecting the contract instead of accepting when there is no contract, we obtain a difference which contains ν_2 *no matter* the outcome (whether or not there was a contract).

To be more formal we amend our notation for this example only to distinguish between the buyer's information set before and after the contract offer has been made, and allow the buyer's strategy to depend explicitly on the seller's offer, i.e. now $\mathbf{d}_b = \mathbf{s}_b(\mathcal{J}_b, \mathbf{d}_s)$. Also, the difference in seller profits between when d_s is offered and when ϕ is offered is now

$$\Delta\pi^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot) = \pi^s(d_s, \mathbf{s}_b(\mathcal{J}_b, d_s), \cdot) - \pi^s(\phi, \mathbf{d}_b = 0, \cdot).$$

Define $\Delta r^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot)$ analogously. For the buyer

$$\Delta\pi^b(d_b, d'_b, d_s, \cdot) = \pi^b(d_b, d_s, \cdot) - \pi^b(d'_b, d_s, \cdot),$$

with $\Delta r^b(d_b, d'_b, d_s, \cdot; \theta)$ defined analogously.

Note that $\mathbf{s}_b(\mathcal{J}_b, d_s)$ is an indicator function which takes the value of one when the contract is accepted and zero elsewhere. Then the behavioral

¹⁶Without loss of generality, we could allow ν_2 to vary by the terms of the seller's contract offer d_s , but we opt for the simpler notation here.

condition for the seller is that

$$\begin{aligned} & \mathcal{E} \left[\mathbf{s}_b(\mathcal{J}_b, d_s) \Delta \pi^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot) | \mathcal{J}_s \right] = \\ & \mathcal{E} \left[\mathbf{s}_b(\mathcal{J}_b, d_s) \left[\Delta r^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot; \theta) + \nu_2 \right] | \mathcal{J}_s \right] \geq 0, \end{aligned}$$

while that condition for the buyer if the buyer rejects the contract offer is

$$\begin{aligned} & \mathcal{E} \left[\left[1 - \mathbf{s}_b(\mathcal{J}_b, d_s) \right] \Delta \pi^b(0, 1, d_s, \cdot) | \mathcal{J}_b \right] = \\ & \mathcal{E} \left[\left[1 - \mathbf{s}_b(\mathcal{J}_b, d_s) \right] \left[\Delta r^b(0, 1, d_s, \cdot; \theta) + \nu_2 \right] | \mathcal{J}_b \right] \geq 0. \end{aligned}$$

Let $\mathbf{x} \in \mathcal{J}_s \cap \mathcal{J}_b$ and assume \mathbf{x} is an instrument in the sense that $\mathcal{E}[\nu_2 | \mathbf{x}] = 0$. Then since $\mathcal{E} \left[((1 - \mathbf{s}_b(\mathcal{J}_b, d_s)) + \mathbf{s}_b(\mathcal{J}_b, d_s)) \nu_2 | x \right] = \mathcal{E}[\nu_2 | x]$ our assumptions imply

$$\begin{aligned} 0 & \leq \mathcal{E} \left[\mathbf{s}_b(\mathcal{J}_b, d_s) \Delta \pi^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot) + [1 - \mathbf{s}_b(\mathcal{J}_b, d_s)] \Delta \pi^b(0, 1, d_s, \cdot) | x \right] \\ & = \mathcal{E} \left[\mathbf{s}_b(\mathcal{J}_b, d_s) \Delta r^s(d_s, \phi, \mathbf{s}_b(\mathcal{J}_b, d_s), 0, \cdot; \theta) + [1 - \mathbf{s}_b(\mathcal{J}_b, d_s)] \Delta r^b(0, 1, d_s, \cdot; \theta) | x \right]. \end{aligned}$$

Our second empirical example is a generalization of this one.

Note. All of the examples generated ν_2 averages with zero, in contrast to negative, expectations. However strict inequalities are often found. For one example add a non-negative cost of switching decisions to the panel discrete choice problem in example 2. Alternatively assume we start with a set of existing relationships in a buyer-seller network (example 4), and investigate whether a new set of contract offers makes it worthwhile for the HMO to make a change when any change involves a cost. The discussion of boundaries in ordered choice models (see our first empirical example) is yet another case.

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. $\mathbf{d}_i = s_i(\mathcal{J}_i)$. It follows that the averages over realizations

of the random variable \mathbf{d}_i will actually be approximating expectations conditional on $\mathbf{d}_i = s_i(\mathcal{J}_i)$. Hence, we will show that our assumptions lead to a corresponding moment inequality in such a conditional expectation.

Equation (3) implies that

$$\begin{aligned}
& \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \Delta r(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)\right] \quad (4) \\
&= \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \Delta \pi(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)\right] \\
&\quad - \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \nu_{1,i,\mathbf{d}_i,d'} h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)\right] \\
&\quad - \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \nu_{2,i,\mathbf{d}_i,d'} h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)\right]
\end{aligned}$$

We consider each of the three terms following the equality in equation (4) in turn. Each summand in the first term can be written as

$$\begin{aligned}
& \mathcal{E}[\chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \Delta \pi(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)] \\
&= \mathcal{E}[\chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \mathcal{E}[\Delta \pi(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) \mid \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)] h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)] \\
&\geq 0
\end{aligned}$$

Note that $\mathcal{E}[\Delta \pi(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) \mid \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)] = \mathcal{E}[\Delta \pi(s_i(\mathcal{J}_i), d', \mathbf{d}_{-i}, \mathbf{z}_i) \mid \mathcal{J}_i]$ by Assumption 2, and this last term is nonnegative by Assumptions 1 and 2 as discussed above. The inequality above then follows by the fact that both $\chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d')$ and $h(\mathbf{x}_i)$ are non-negative.

As discussed above, the definition of ν_1 in equation (2) and Assumption 2 yield $\mathcal{E}[\nu_{1,i,\mathbf{d}_i,d'} \mid \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)] = 0$. So, the ν_1 term above is zero. Assumption 3 states that the last term in equation (4) is non-negative, so we have

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{\mathbf{d}_i, \mathcal{J}_i}^i(d') \Delta r(\mathbf{d}_i, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) h(\mathbf{x}_i) \mid \mathbf{d}_i = s_i(\mathcal{J}_i)\right] \geq 0. \quad (5)$$

Equation (5) depends only on observables and θ_0 , so we can form its sample

analog and look for values of θ that satisfy it.¹⁷ Note that, alternatively, we could have plugged in $\mathbf{d}_i = s_i(\mathcal{J}_i)$ and taken expectations without conditioning. That is, our assumptions also lead to

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

2.4 Generalizations

For expositional ease, the assumptions used in sections 2.1 and 2.2 were not as general as they could have been. Here we discuss a number of generalizations and show how they generate moment inequalities that are analogous to those in equation (5).

Generalization 1. (*Non-optimal decision-making.*) It is possible to weaken Assumption 1 considerably. Consider the generalization

$$\sup_{d \in \mathcal{D}_i(s_i(\mathcal{J}_i))} \mathcal{E}[\pi(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{y}_i) | \mathcal{J}_i, \mathbf{d}_i = d] \leq (1 + \delta) \mathcal{E}[\pi(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{y}_i) | \mathcal{J}_i, \mathbf{d}_i = s_i(\mathcal{J}_i)]$$

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

When $\mathcal{D}(d_i) = \mathcal{D}_i$ and $\delta = 0$, we are back to Assumption 1. However this version of Assumption 1 allows the agent to make decisions which are only within a multiplicative factor $1 + \delta$ of the decision that maximizes the expected value of the outcome and allows the decision space of the alternative, $\mathcal{D}_i(d_i)$, to be a subset of \mathcal{D}_i . If, for example, $\delta = .5$, then non-optimal choices would be allowed provided they did not, on average, reduce expected profits more than 50% from the expected profits that would be earned from optimal strategies. Complementary reasoning applies to the actions per se when $\delta = 0$ but $\mathcal{D}_i(d_i) \neq \mathcal{D}_i$. For example, if there was a continuous control, and we specified that $\mathcal{D}_i(d_i) = \{d : |d - d_i| \geq \alpha d_i, d \in \mathcal{D}\}$ for some small $\alpha > 0$, then we would be specifying that though small deviations about optimal behavior can occur (deviations that leave the choice within 100 α % of the optimal decision), at least on average large deviations do not occur.¹⁸

¹⁷In general Assumptions 1, 2, and 3 are sufficient but not necessary for the inequalities in (5), which, in turn, provide the basis for estimation and inference. That is, we expect that there are alternative conditions that will also suffice.

¹⁸One would typically assume δ and $\mathcal{D}_i(\cdot)$ are set exogenously though we could begin the

The inequalities carry through with this assumption provided we alter the definitions of $\Delta\pi$ and Δr to account for the $1 + \delta$ factor.

Generalization 2. (*Non-simultaneous Move Games and Related Violations of Assumption 2.*) We begin with a generalization to example 4 which illustrates the problems that can arise in non-simultaneous move games. In particular we consider a “vertical” market with multiple sellers and multiple buyers. Sellers make simultaneous take it or leave it offers to buyers. Buyers respond simultaneously at some later date. The buyers can still accept or reject any given contract without changing any other contract outcome. However, relative to example 4, this multiple buyer-seller setting requires additional consideration for counterfactual responses to alternative seller contract offers. In particular, if seller s^* changes his contract offer to buyer b^* , then b^* ’s optimal response may well include a change in response to the offers from *other* sellers.

The contract offer of seller s to buyer b will be denoted by $d_s^b \in \mathcal{D}_s$, where \mathcal{D}_s is the space of possible contracts (e.g. all two part tariffs), $d_s = (d_s^b, d_s^{-b}) \in \times_b \mathcal{D}_s$, and $d_S = (d_{s=1}, \dots, d_{s=S}) \in (\times_b \mathcal{D}_s)^S$ where S is the number of sellers. The take-it-or-leave-it decisions of buyer b are collected in the vector $d_b = (d_b^s, d_b^{-s}) \in [0, 1]^S$, and $d_B = (d_{b=1}, \dots, d_{b=B}) \in [0, 1]^{B \times S}$, where B is the number of buyers. The argument above implies that the distribution of \mathbf{d}_b^{-s} conditional on the seller’s information set is not independent of the seller’s offer, i.e. of $d_s^b \in \mathcal{D}_s$. Assumption 1 implies

$$\begin{aligned} & \mathcal{E}[\pi^s(\mathbf{d}_s, \mathbf{d}_B, y(\mathbf{d}_S, \mathbf{d}_B, \mathbf{z})) | \mathcal{J}_s, \mathbf{d}_s = (d_s^b, d_s^{-b})] \\ & - \mathcal{E}[\pi^s(\mathbf{d}_s, \mathbf{d}_B, y(\mathbf{d}_S, \mathbf{d}_B, \mathbf{z})) | \mathcal{J}_s, \mathbf{d}_s = (\phi, d_s^{-b})] \geq 0. \end{aligned}$$

Since the buyers move simultaneously, if we could observe or construct random draws from the distribution of $(\mathbf{d}_b^s, \mathbf{d}_b^{-s})$ conditional on $\mathbf{d}_s = (\phi, d_s^{-b})$ and \mathcal{J}_s , we could construct random draws from the right hand side of this

analysis assuming $\delta = 0$ and $\mathcal{D}(d_i) = \mathcal{D}_i$, and then test whether the data is consistent with those assumptions. If it is not, find a relaxation of those assumptions that *is* consistent with the data; for example find a value for δ that satisfies the inequalities (up to sampling error) and the implied estimator of the parameter vector. Note that this procedure maintains our assumptions on functional forms and asks only whether, given those functional forms, the relaxation of optimizing behavior needed to rationalize the data is too large to be *a priori* reasonable. Of course, inference based on such a procedure would need to account for this sample-based method of suggesting a δ , which would require an extension of the econometric results in section 3.

inequality and proceed as we did in the simultaneous move games analyzed above.

The problem is that we do not know how to construct a random draw from the distribution of $(\mathbf{d}_b^s, \mathbf{d}_b^{-s})$ conditional on $\mathbf{d}_s = (\phi, d_s^{-b})$ and \mathcal{J}_s . In particular without further assumptions we do not know how the buyer would change its responses to other sellers were it faced with a null contract from a given seller. One way around this problem is to compute the minimum of $\pi^s(\cdot)$ over all possible choices buyer b could make given the observed realization of (d_{-s}, d_{-b}, z) and then use the average of the difference between the realized profit and this minimized counterfactual profit as the theoretical inequality on which estimation is based.¹⁹

That is, since

$$\begin{aligned} & \min_{d \in [0,1]^{s-1}} \pi^s \left((\phi, d_s^{-b}), [d_b^s = 0, d], d_{-b}, y((\phi, d_s^{-b}), d_{-s}, [d_b^s = 0, d], d_{-b}, z) \right) \\ & \leq \pi^s \left((\phi, d_s^{-b}), [d_b^s = 0, d_b^{-s}], d_{-b}, y((\phi, d_s^{-b}), d_{-s}, [d_b^s = 0, d], d_{-b}, z) \right) \end{aligned}$$

for every realization of $(\mathbf{d}_{-s}, \mathbf{d}_{-b}, \mathbf{z})$,

$$\mathcal{E}[\pi^s(\mathbf{d}_s, \mathbf{d}_B, y(\mathbf{d}_S, \mathbf{d}_B, \mathbf{z})) | \mathcal{J}_s, \mathbf{d}_s = (d_s^b, d_s^{-b})] \quad (6)$$

$$- \mathcal{E} \left[\min_{d \in [0,1]^{s-1}} \pi^s \left(\mathbf{d}_s, [d_b^s = 0, d], \mathbf{d}_{-b}, y(\mathbf{d}_S, [d_b^s = 0, d], \mathbf{d}_{-b}, \mathbf{z}) \right) | \mathcal{J}_s, \mathbf{d}_s = (\phi, d_s^{-b}) \right] \geq 0.$$

To use (6) to generate moment inequalities we need to actually compute a minimum over alternative choices, and this increases the computational burden of the estimator. Moreover the minimum may not be terribly informative about the parameters of interest. Still the inequality in equation (6) can be used when games are not simultaneous.

The non-simultaneous move game is a special instance of a more general problem. The problem occurs when Assumption 2 is not satisfied because; (i) there is component of \mathbf{d}_{-i} whose distribution, conditional on $(\mathcal{J}_i, d_i = d)$,

¹⁹We are implicitly assuming here both that; (i) the offers themselves are not public information, i.e. the offers to a particular buyer are known only to that buyer and not to the other buyers, and (ii) passive expectations, i.e. that the fact that a buyer gets an alternative offer from a given seller will not change this buyer's perceptions on the offers the particular seller was likely to have made to the buyer's competitors. If the offers were public information then the minimum in equation (6) below must be taken over d_{-b} as well as d_b^{-s} for the inequality in that equation to follow from our assumptions.

depends on d , and (ii) we do not have a model for what that component's value would be were we to change d_{-i} . If, for those components, we can compute the minimum profits over the values that d_{-i} could take, then we can use the minimal profit value as the “counterfactual” in our inequalities.

Generalization 3. There are a number of generalizations that can be incorporated into our framework without making any change in the inequalities taken to data.

Individual Effects. Note that unobserved factors whose effect on the agent's profits do not depend on d are differenced out of (5). As a result, additively separable individual effects that do not interact with the alternative chosen can be added to the returns functions π and r without affecting the inequalities in (5). This automatic differencing out of individual effects implies that the unobservable ν_2 need only capture the effects of omitted variables that impact on the change in profits in response to a change in d .²⁰

Mixed Strategies. If agent i plays a mixed strategy then Assumption 1 would require slight modification to condition on the whole strategy played. It would imply that each pure strategy with positive probability in the mixed strategy must have the same expected return. Minor notational differences aside, the other two assumptions would still apply and yield a moment inequality of the same form. So, there is no need for the econometrician to specify whether the underlying strategies are pure or mixed. If we did know mixed strategies were being played, and we could distinguish the mixed strategies associated with particular information sets, then there would be more information available than the information being used in our current inequalities.

Conditioning Sets and Heterogenous Primitives. The notion that \mathcal{J}_i denotes agent i 's information set at the time decisions are made is only used as motivation for Assumption 1. If Assumptions 1, 2, and 3 were known to hold for a set of conditioning variables which were not the actual information set, then the required moment conditions could still be formed,²¹ despite the fact that some of the natural interpretations of the unobservables would no longer necessarily hold. Also we note that the $\pi(\cdot)$, and $r(\cdot)$ functions could be indexed by i as could the instruments (i.e. $\mathbf{x}_{i,d,d'}$).

²⁰This discussion of individual effects being automatically differenced out assumes $\delta = 0$ in our first generalization).

²¹Of course insuring that Assumptions 2 and 3 are satisfied will put conditions on \mathcal{J}_i .

3 Estimation and Inference.

In section 2.3, we derived the inequality conditions that result from Assumptions 1, 2, and 3. These inequalities fit naturally into the general econometric framework of moment inequalities. This section addresses certain estimation and inference issues in a general moment inequality setting. The goal is to enable us to use the conditions from section 2 to analyze data.

A key feature of the moment inequality setting is the possibility that the parameters of interest are only partially identified (Manski 2003). The set of parameters satisfying the moment inequalities is called the identified set. A number of papers have focused on methods of constructing confidence regions for this set (or for the true parameter value which is contained in the set; see Imbens and Manski 2003, Andrews, Berry, and Jia 2004, Chernozhukov, Hong, and Tamer 2003, Rosen 2005, Shaikh 2005, and Soares 2006).

Our focus is on estimation of an extreme (or boundary) point of the identified set, and we list a set of assumptions under which we can provide a complete characterization of the asymptotic distribution of the extreme point estimator. Empirical research typically provides a table of estimates with dimension by dimension standard errors or confidence intervals. One corresponding notion for set-valued estimates would be dimension by dimension extreme point estimates along with confidence intervals, either for the extreme points themselves, or for the parameter of interest. Our results allow us to do inference on extreme points of other directions of the parameter space as well.²²

The limiting distribution we obtain is, in general, non-normal and we do not always have a way of precisely approximating it under the the general assumptions listed in our theorem. Instead we consider two distributions both of which are easy to simulate. One of these stochastically dominates the limiting distribution of the extreme point estimator asymptotically, while the second is stochastically dominated by the limiting distribution asymptotically. These simulated distributions allow us to compute “outer” and an

²²If the identified set is convex the boundary of that set is defined by the extreme points in all 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 2005). We note that with additional regularity conditions it is possible to generalize to extreme points of a function of the parameter vector, or of expectations of functions of the data and the parameter vector; topics we do not pursue here.

“inner” confidence intervals for the extreme point. Asymptotically, the outer confidence interval will contain the corresponding infeasible confidence interval generated by the limit distribution, and in this sense the outer confidence interval is conservative. The inner confidence interval will, asymptotically, lie within the infeasible limit distribution confidence interval. The inner confidence interval will be used to provide a bound on the conservatism of the outer confidence interval. Moreover, for an important special case, the inner confidence interval will provide improved coverage.

Our formal results on extreme point estimation are contained in section 3.1. Subsection 3.1.1 deals with consistency, 3.1.2 with the asymptotic distribution, and 3.1.3 with simulated approximations to that distribution. A heuristic explanation of the arguments leading to both the asymptotic distribution and to the simulated approximations to those distributions precedes the presentation of formal results in sections 3.1.2 and 3.1.3. The reader who is interested in the linear moments case may want to read the heuristic arguments and then move directly to section 3.3 which provides slightly different simulation procedures which are applicable to that case. Both of our empirical examples are linear, so section 3.3 should enable the reader to understand how we obtained the results in the rest of the paper.

For clarity we focus the discussion on one extreme point, but it is straightforward to generalize and obtain the joint distribution of two or more extreme points. 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). Section 3.1.3 explicitly considers the implications of our results on the construction of confidence intervals for the parameter vector.

In section 3.2, we present a specification test of the moment inequalities. This test is a natural extension of the usual GMM specification test to the case with inequalities, but the test statistic does not have a pivotal distribution. We provide a computationally simple method for obtaining conservative critical values.

3.1 Estimation and Inference for Extreme Points

Assume that there is data on J markets indexed by $j = 1, \dots, J$. A market is a draw on $z^j = (\mathbf{y}^j, x^j, d^j)$ where $\mathbf{y}^j \equiv \{\mathbf{y}_i^j\}_{i=1}^{n^j}$, and d^j , x^j , and z^j are defined similarly. We will assume that the observed markets are independent

draws from a population of such vectors with a distribution, say \mathcal{P} , that respects our Assumptions 1, 2, and 3.

The $M(\equiv m \times h)$ dimensional moment function from equation (5) is:

$$m(z^j, \theta) = \sum_{i=1}^{n_j} \sum_{d' \in \mathcal{D}_i^j} \chi_{d_i^j, \mathcal{J}_i^j}^i(d') \Delta r(d_i^j, d', \mathbf{d}_{-i}^j, \mathbf{z}_i^j, \theta) h(x_i^j).$$

The inequality in (5) can then be expressed simply as $\mathcal{P}m(z, \theta) \geq 0$.

Let $\Theta \subset \mathbb{R}^K$ denote the parameter space. The set of parameters satisfying the moment inequalities will be referred to as the identified set and denoted by $\Theta_0 = \{\theta \in \Theta : \mathcal{P}m(z, \theta) \geq 0\}$. To estimate Θ_0 we find the values of θ that satisfy the sample analog of the moment inequalities, or if no such values exist, we take the value(s) that are “closest” to satisfying the inequalities. Specifically, letting P_J denote the empirical distribution so that $P_J m(z, \theta) = \frac{1}{J} \sum_{j=1}^J m(z^j, \theta)$, our estimate of the identified set is

$$\Theta_J = \arg \min_{\theta \in \Theta} \left\| \left(P_J m(z, \theta) \right)_- \right\|$$

where $(\cdot)_- = \min\{\cdot, 0\}$.²³

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,

$$\underline{\theta} = \{\theta \in \Theta_0 : \theta_1 = \arg \min_{\tilde{\theta} \in \Theta_0} \tilde{\theta}_1\}$$

where θ_1 denotes the first element of the vector θ , and $\underline{\theta} \in \mathbb{R}^K$.²⁴ In what follows, one could equally well consider the minimum or maximum of other dimensions of $\theta \in \Theta_0$, or more generally, extremes of linear combinations of various dimensions of $\theta \in \Theta_0$. Corresponding asymptotic results for extremes of linear combinations of the dimensions of θ are immediate from the results given below. When Θ_0 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.

²³This choice of criterion corresponds to an identity weight matrix in GMM. We do not explore other weight matrix choices here, but note that, relative to GMM, the weight matrix choice here is restricted to maintain the inequalities.

²⁴In general, $\underline{\theta}$ could be a set, but the notation and terminology foreshadow our assumption, below, that this set is a singleton.

Given Θ_J , a natural estimator for $\underline{\theta}$ is²⁵

$$\hat{\underline{\theta}} = \{\theta \in \Theta_J : \theta_1 = \arg \min_{\tilde{\theta} \in \Theta_J} \tilde{\theta}_1\}. \quad (7)$$

For completeness, we briefly provide a consistency result for this estimator. Then we derive the asymptotic distribution and discuss simulation methods for inference. The reader who is only interested in the asymptotic distribution should be able to go directly to section 3.1.2.

3.1.1 Consistency

The conditions for consistency follow.

Assumption A1 (a) Θ is compact; (b) $\Theta_0 \subset \text{int}(\Theta)$; (c) Θ_0 is closed; (d) $\underline{\theta}$ is a singleton.

The key part of Assumption A1 is that $\underline{\theta}$ is a singleton. This condition could likely be relaxed if one were to focus only on the limiting distribution of the first component of $\hat{\underline{\theta}}$, but here it greatly simplifies the asymptotic distribution expression and is used in the proofs. Closure of the identified set assures that $\underline{\theta} \in \Theta_0$. The other parts of Assumption A1 are standard.

The next assumption formally characterizes the definition of our estimator. The definition in (7) certainly suffices, but is more restrictive than necessary.

Assumption A2 *The estimator satisfies*

$$\hat{\underline{\theta}}_1 = \inf_{\theta \in \Theta_J} \theta_1 + o_p(1/\sqrt{J})$$

and $\hat{\underline{\theta}} \in \Theta_J$.

The $o_p(1/\sqrt{J})$ term could be relaxed to $o_p(1)$ for the consistency result, but the faster rate will be used for the asymptotic distribution result.

The next two assumptions ensure “local” identification. Assumption A3 ensures that at points which are at least ϵ away from the identified set the inequalities cannot be arbitrarily close to holding. Assumption A4 ensures that the inequalities are satisfied strictly for some point in each neighborhood

²⁵ $\hat{\underline{\theta}}$ could be a set. The asymptotic results will refer to any sequence of points taken from the $\hat{\underline{\theta}}$ for each sample size.

of the extreme point. Note that when this assumption is combined with continuity of the population moments (as will be required for the asymptotic distribution), it will ensure that the boundary point is not a single point isolated from the remainder of the identified set.

Assumption A3 *For any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{\theta \in (\Theta_0^\epsilon)^c} \left\| \left(\mathcal{P}m(z, \theta) \right)_- \right\| > \delta$$

where

$$\Theta_0^\epsilon = \{ \theta \in \Theta : \inf_{\theta' \in \Theta_0} \|\theta - \theta'\| \leq \epsilon \}.$$
²⁶

Assumption A3 gives one side of the local identification condition at every boundary point of the identified set. By definition, at any point outside the identified set, the inequalities cannot all hold.

Assumption A4 *Every neighborhood of $\underline{\theta}$ contains a point θ such that $Pm(Z, \theta) > 0$.*

Finally we need a standard uniform consistency condition for the sample moments.

Assumption A5

$$\sup_{\theta \in \Theta} \|P_J m(Z, \theta) - \mathcal{P}m(Z, \theta)\| \xrightarrow{p} 0.$$

Then we have the following result. Formal proofs of each theorem are provided in a technical appendix.

Theorem 1 *Under Assumptions A1-A5, $\hat{\underline{\theta}} \xrightarrow{p} \underline{\theta}$.*

PROOF SKETCH: Fix $\epsilon > 0$. Assumptions A3 and A5 imply that with probability approaching one (w.p.a. 1) $\Theta_J \subset \Theta_0^\epsilon$, i.e. there is not a $\theta \in \Theta_J$ that is more than a distance of ϵ from some point in Θ_0 . So $\hat{\underline{\theta}} \in \Theta_0^\epsilon$ as defined in A3. Consequently, the first component of $\hat{\underline{\theta}}$, i.e. $\hat{\underline{\theta}}_1$, cannot be more than ϵ away from the first component of some $\theta \in \Theta_0$ w.p.a. 1, and so $\hat{\underline{\theta}}_1 \geq \underline{\theta}_1 - \epsilon$ (w.p.a. 1). By Assumptions A4 and A5, there exists a point θ' in an ϵ -neighborhood of $\underline{\theta}$ such that $P_J m(Z, \theta') > 0$ w.p.a. 1. Hence $\theta' \in \Theta_0$, and

²⁶Note that Θ_0^ϵ is the largest set such that $d_H(\Theta_0^\epsilon, \Theta_0) \leq \epsilon$, where d_H is the Hausdorff metric.

the first dimension of θ' cannot be too much larger than the first dimension of $\underline{\theta}$, $\theta_1' \leq \underline{\theta}_1 + \varepsilon$. Then, by the definition of $\hat{\underline{\theta}}$ in Assumption A2, $\hat{\underline{\theta}}_1 \leq \underline{\theta}_1 + \varepsilon$ w.p.a. 1. Together these arguments show that $|\hat{\underline{\theta}}_1 - \underline{\theta}_1| < \varepsilon$ w.p.a. 1. Lemma 2 in the Appendix then shows that by Assumption A1 consistency of the first component, $\hat{\underline{\theta}}_1$, implies consistency of $\hat{\underline{\theta}}$.

3.1.2 Asymptotic Distribution

The estimator, $\hat{\underline{\theta}}$, is the minimizer of the first dimension of Θ_J (up to the $o_p(1/\sqrt{J})$ error in Assumption A2). With probability approaching one, Θ_J is defined as the set of θ 's satisfying $P_J m(z, \theta) \geq 0$. Given consistency, we can focus on the local behavior of this sample moment inequality (around $\underline{\theta}$) to provide intuition for the form of the limit distribution for $\hat{\underline{\theta}}$. Multiplying the sample moments by \sqrt{J} , we have

$$\begin{aligned} 0 &\leq \sqrt{J} P_J m(z, \hat{\underline{\theta}}) \\ &= \sqrt{J} \left(\mathcal{P}m(z, \hat{\underline{\theta}}) - \mathcal{P}m(z, \underline{\theta}) \right) + \sqrt{J} \left(P_J m(z, \hat{\underline{\theta}}) - \mathcal{P}m(z, \hat{\underline{\theta}}) \right) + \sqrt{J} \mathcal{P}m(z, \underline{\theta}) \\ &= \sqrt{J} \left(\mathcal{P}m(z, \hat{\underline{\theta}}) - \mathcal{P}m(z, \underline{\theta}) \right) + \sqrt{J} \left(P_J m(z, \underline{\theta}) - \mathcal{P}m(z, \underline{\theta}) \right) + \sqrt{J} \mathcal{P}m(z, \underline{\theta}) + o_p(1) \end{aligned}$$

where the second equality assumes stochastic equicontinuity of the moments in a neighborhood of $\underline{\theta}$ and consistency of $\hat{\underline{\theta}}$.²⁷

Now partition the moments into two sets: those that bind at the extreme point (i.e. those with a value of zero at θ) and those that do not. The binding constraints will be denoted m_0 (so $\mathcal{P}m_0(z, \underline{\theta}) = 0$), and the non-binding constraints will be denoted m_1 (so $\mathcal{P}m_1(z, \underline{\theta}) > 0$), and $m = (m_0', m_1')'$. The non-binding constraints will not play a role in the limiting distribution of the extreme point estimator. To see this, note that the term $\sqrt{J} \mathcal{P}m_1(z, \underline{\theta})$ will dominate the expression in the inequality above for the non-binding moments, so that these terms will be strictly greater than zero for all θ local to $\underline{\theta}$ with probability approaching one. For the binding moments, i.e. for the moments which do determine the form of the limit distribution, $\sqrt{J} \mathcal{P}m_0(z, \underline{\theta})$ is zero, so the other terms in the inequality determine the local behavior.

If we assume that the binding population moments are continuously differentiable in a neighborhood of $\underline{\theta}$, the consistency of $\hat{\underline{\theta}}$ implies that we can

²⁷Actually, we will only require stochastic equicontinuity for a subset of the moments in the assumptions below.

rewrite the inequality above for the binding moments as

$$0 \leq \underline{\Gamma}_0 \sqrt{J}(\hat{\underline{\theta}} - \underline{\theta}) + \sqrt{J}P_J m_0(z, \underline{\theta}) + o_p(1) \quad (8)$$

where $\underline{\Gamma}_0 = \frac{\partial}{\partial \underline{\theta}} \mathcal{P}m_0(z, \underline{\theta})$.

Since $\mathcal{P}m_0(z, \underline{\theta}) = 0$, $\sqrt{J}P_J m_0(z, \underline{\theta})$ will typically obey a central limit theorem, $\sqrt{J}P_J m_0(z, \underline{\theta}) \xrightarrow{d} N(0, \underline{\Sigma}_0)$, where $\underline{\Sigma}_0 = \text{Var}(m(z, \underline{\theta}))$. Substituting this normal approximation for the last term of equation (8) yields the inequality which determines the limiting distribution of $\sqrt{J}(\hat{\underline{\theta}} - \underline{\theta})$. That is if we knew $\underline{\Gamma}_0$ and $\underline{\Sigma}_0$, the limiting distribution could be simulated by substituting draws from $N(0, \underline{\Sigma}_0)$ for the $\sqrt{J}P_J m_0(z, \underline{\theta}) + o_p(1)$ term in equation (8), and finding the values of $\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1)$ that minimize the negative part of the resulting equations.

We now provide the assumptions we use in formalizing the result just described. The first three assumptions are familiar from traditional method of moments estimators. Note that the conditions on the non-binding moments, m_1 , are less stringent than those on the binding moment functions, m_0 . We begin with an assumption on the smoothness of the *population* moments.

Assumption A6 (a) $\mathcal{P}m_0(z, \theta)$ is continuously differentiable in a neighborhood of $\underline{\theta}$; (b) $\mathcal{P}m_1(Z, \theta)$ is continuous at $\underline{\theta}$.

The next two assumptions place restrictions on the asymptotic behavior of the sample (binding) moments.

Assumption A7 $\sqrt{J}P_J m_0(z, \underline{\theta}) \xrightarrow{d} N(0, \underline{\Sigma}_0)$.

A law of large numbers condition is also needed for the non-binding moments, but Assumption A5 suffices.

Assumption A8 For all sequences $\delta_n \downarrow 0$,

$$\sup_{\|\underline{\theta} - \underline{\theta}\| \leq \delta_n} \|\sqrt{J}(P_J m_0(z, \theta) - \mathcal{P}m_0(z, \theta)) - \sqrt{J}(P_J m_0(z, \underline{\theta}) - \mathcal{P}m_0(z, \underline{\theta}))\| = o_p(1)$$

Assumption A9 is less familiar. It places conditions on the linear program that is derived from linearizing the moments, cf. equation (8).

Assumption A9 For each Γ in some open neighborhood of $\underline{\Gamma}_0$, the unique solution to $\min_{\tau: \Gamma \tau \geq 0} \tau_1$ is zero. There exists some λ such that $\underline{\Gamma}_0 \lambda > 0$.

The first part of assumption A9 ensures uniqueness of the linear program solution. This assumption, for instance, implies Assumption A1(d); i.e. that $\underline{\theta}$ is a singleton. It also implies that $\underline{\Gamma}_0$ (and each Γ in the neighborhood) is full column rank. The second part of the assumption assures that the local identification given in Assumption A4 is occurring along some particular direction. This assumption implies Assumption A4.

We can now provide a limit distribution for $\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1)$.

Theorem 2 *Suppose Assumptions A1-A9 hold. Let $\hat{\tau}_1 = \min\{\tau_1 : 0 \leq \Gamma_0\tau + \mathcal{Z}\}$ where $\mathcal{Z} \sim N(0, \underline{\Sigma}_0)$. Then,*

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

PROOF SKETCH: First show that $\hat{\underline{\theta}}$ is \sqrt{J} -consistent. To do this, we begin by showing $P_J m(Z, \underline{\theta} + c\lambda/\sqrt{J}) \geq 0$ with probability approaching one for c large enough (this follows from Assumptions A6, A7, and A8). This implies that $\underline{\theta} + c\lambda/\sqrt{J} \in \Theta_J$ w.p.a. 1, so $\hat{\underline{\theta}}_1 \leq \underline{\theta}_1 + c\lambda_1/\sqrt{J} + o_p(1/\sqrt{J})$. Rearranging yields $\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1) \leq c\lambda_1 + o_p(1)$. Then we show similarly that $P_J m(Z, \hat{\underline{\theta}} + c\lambda/\sqrt{J}) \geq 0$ w.p.a. 1. This implies that $\hat{\underline{\theta}} + c\lambda/\sqrt{J} \in \Theta_0$ w.p.a. 1. Consequently $\hat{\underline{\theta}}_1 + c\lambda_1/\sqrt{J} \geq \underline{\theta}_1 + o_p(1/\sqrt{J})$, or $\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1) \geq -c\lambda_1 + o_p(1)$ w.p.a. 1. So $\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1) = O_p(1)$, which, in turn, yields $\sqrt{J}(\hat{\underline{\theta}} - \underline{\theta}) = O_p(1)$ by Assumption A9 using Lemma 4 in the Appendix.

Second we consider the estimator, θ^* defined as follows. Let $L_J(\theta) = \underline{\Gamma}_0(\theta - \underline{\theta}) + P_J m_0(z, \underline{\theta})$, and set $\theta^* = \arg \inf\{\theta_1 : 0 \leq L_J(\theta)\}$. Since $\underline{\Gamma}_0$ and $\underline{\theta}$ are unknown, θ^* is infeasible, but we can show that it is well-defined and \sqrt{J} -consistent. Moreover from Assumption A7, θ_1^* has the limit distribution given in the theorem. To complete the proof, we show that $\sqrt{J}(\hat{\underline{\theta}}_1 - \theta_1^*) = o_p(1)$ so the limit distributions of $\hat{\underline{\theta}}_1$ and θ_1^* are the same. This final step follows by showing that there exist deterministic sequences h_J and r_J , both $o(1)$, such that $L_J(\hat{\underline{\theta}} + h_J\lambda/\sqrt{J}) \geq P_J m_0(z, \hat{\underline{\theta}})$ and $P_J m_0(z, \theta^* + r_J\lambda/\sqrt{J}) \geq L_J(\theta^*)$ w.p.a. 1. □

Note that when the number of binding moments equals the dimension of the parameter vector (K), we can explicitly solve for θ_1^* as a linear combination of normals and this implies that $\hat{\underline{\theta}}_1$ has a normal limit distribution.

However when there are more binding moments than the dimension of the parameter vector, the distribution is a mixture of truncated normals with endogenous truncation points, and hence is not normal.

Except in particular cases where the economic model provides reasons to believe that there are more than K inequalities binding at a specific boundary point, one might think that the special case in which there are exactly K binding population moments at $\underline{\theta}$ is likely to be appropriate. Though this may well be true we have found that the limiting normal distribution generated by that special case provides a poor approximation to the finite sample distribution of our estimators for samples of the sizes we use in our examples. This deficiency in the approximation is a result of the fact that the limiting distribution ignores the influence of all the non-binding moments. If some of those non-binding population moments are “close” to binding, then the corresponding sample moments will actually bind with positive probability in finite samples. Having these additional moments bind in finite samples creates variation in the finite sample distribution for the extreme point estimator that is not captured by the limiting distribution emanating from the assumption that only K moments bind.

3.1.3 Approximating the Limit Distribution

Approximation of the limiting distribution provided in Theorem 2 is hampered by the fact that it depends on the identity of the binding moments, and the researcher will generally not know which moments bind *a priori*. Our goal is to provide inferential procedures which do not depend on prior knowledge of which moments bind and are easy for the applied person to use. In particular, our methods are computationally simple, requiring only simulation from a normal distribution and solving linear programs.

To this end, we introduce two simulated distributions for each of the lower and the upper bound estimators; i.e. two for each of $\hat{\theta}_1$ and $\hat{\bar{\theta}}_1$. Asymptotically, one of each couple stochastically dominates the true asymptotic distribution and the other is stochastically dominated by the true asymptotic distribution. These simulation distributions can be used to produce “outer” and “inner” confidence intervals for various parameters of interest: the interval defined by the upper and lower bounds, the true parameter value, and the upper and lower bounds themselves.

One of the two simulation methods yields an outer confidence interval for the interval $[\underline{\theta}_1, \bar{\theta}_1]$. This outer confidence interval is asymptotically conservative, and can also be used as a conservative confidence interval for the true parameter value, $\theta_{0,1}$. The other simulation method leads to an inner confidence interval for $[\underline{\theta}_1, \bar{\theta}_1]$. This confidence interval is informative about the conservatism of the outer confidence interval, as described below. Perhaps more importantly, when there are exactly K binding population moments at $\underline{\theta}_1$ and at $\bar{\theta}_1$, the simulation distributions used for the inner confidence interval converge to the true limiting distributions of the boundary estimators and so are providing the desired inference for the boundary points. The limiting distributions of the boundary estimators, for this case, are normals based on only the K binding population moments. Like the actual finite sample distributions of the estimators, the inner simulation distributions will not generally be normal, in this case, and will reflect the fact that nonbinding population moments may bind in finite samples and affect the simulated distributions. Note also that if there are more than K population moments binding at either boundary, then our inner confidence interval may have inadequate coverage (even in the limit).²⁸

The outer confidence interval for the true value of the parameter, $\theta_{0,1}$, has three sources of conservatism, and two of them are shared by the inner interval. To see the two common sources of conservatism, let $[\hat{a}, \hat{b}]$ be a random interval and note that since $\theta_{0,1} \in [\underline{\theta}_1, \bar{\theta}_1]$,

$$\Pr \left\{ \theta_{0,1} \in [\hat{a}, \hat{b}] \right\} \geq \Pr \left\{ [\underline{\theta}_1, \bar{\theta}_1] \subset [\hat{a}, \hat{b}] \right\} \geq 1 - \Pr \left\{ \hat{a} > \underline{\theta}_1 \right\} - \Pr \left\{ \hat{b} < \bar{\theta}_1 \right\}.$$

A choice of \hat{a} and \hat{b} that sets the far right expression to $1 - \alpha$ is clearly conservative for $1 - \alpha$ level coverage of $\theta_{0,1}$. The first inequality above comes from using inference on $[\underline{\theta}_1, \bar{\theta}_1]$ to generate inference on the point $\theta_{0,1}$. This source of conservatism has been considered in detail by Imbens and Manski (2003), see also Guggenberger, Hahn, and Kim (2006). Note that it does not contribute to conservatism for coverage of the interval, $[\underline{\theta}_1, \bar{\theta}_1]$, itself. The second source of conservatism stems from the fact that we use approximations to the marginal distributions of each bound estimator and do not adjust for the correlation between these distributions. This source of conservatism is easily corrected by approximating the joint distribution directly. For expositional ease, we do not pursue this correction here.

²⁸Below we suggest a way of evaluating the possible extent of the coverage problem with the inner confidence interval when more than K moments are binding.

If we were able to use the true limiting distributions of the boundary estimators to obtain \hat{a} and \hat{b} above, then we could have $\Pr(\hat{a} > \underline{\theta}_1) = \alpha/2$ and $\Pr(\hat{b} < \bar{\theta}_1) = \alpha/2$. The true limiting distributions are not generally available, so the outer confidence interval is constructed from simulation distributions that ensure (asymptotically) $\Pr(\hat{a} > \underline{\theta}_1) \leq \alpha/2$, and $\Pr(\hat{b} < \bar{\theta}_1) \leq \alpha/2$. When these probability inequalities are strict, a third source of conservatism appears. Because these inequalities are reversed for the simulated distributions used for the *inner* confidence interval, the inner confidence interval can provide an upper bound on this source of conservatism in the outer confidence interval. Note also that in the leading case of exactly K binding population moments, the inner confidence interval yields the desired quantiles so that $\Pr(\hat{a} > \underline{\theta}_1) \rightarrow \alpha/2$ and $\Pr(\hat{b} < \bar{\theta}_1) \rightarrow \alpha/2$. In this case, the inner confidence interval is asymptotically conservative due to only the first two sources.

The simulated distributions can also be used to provide confidence intervals for the upper and lower bounds themselves. Here our approximations provide an asymptotically conservative confidence interval for the lower bound, $\underline{\theta}_1$, and can produce another confidence interval which will give a bound on the level of conservatism of the first confidence interval. When the binding moments for the lower bound just identify that bound, one of the simulated distributions converges to the true limit distributions and we can obtain a shorter confidence interval for $\underline{\theta}_1$. This is the approximation we expect to be correct coverage for large enough samples. The procedures for inference on the upper bound, $\bar{\theta}_1$, are symmetric.

The Approximating Distributions: Heuristics We begin with a heuristic argument for the approximating distributions for $\hat{\theta}_1$, starting with the distribution which is (asymptotically) stochastically dominated by the limiting distribution for that estimator. The argument for the approximating distributions for $\hat{\bar{\theta}}_1$ is analogous. The additional regularity conditions required to make the argument precise, and the formal statement of our results, are presented immediately after the heuristic argument.

Let $\underline{\Gamma} = \frac{\partial}{\partial \theta} \mathcal{P}m(z, \underline{\theta})$ and $\underline{\Sigma} = \text{Var}(m(z, \underline{\theta}))$, with estimators $\hat{\underline{\Gamma}}$ and $\hat{\underline{\Sigma}}$ computed from the sample analogues of these moments evaluated at $\hat{\underline{\theta}}$. Then the distribution which, in the limit, is stochastically dominated by the true limiting distribution for $\hat{\underline{\theta}}_1$ is simulated by taking random draws from $Z^* \sim$

$N(0, \hat{\Sigma})$ and computing

$$\underline{\tau}_1^*(Z^*) = \min\{\tau_1 : 0 \leq \hat{\Gamma}\tau + Z^* + \sqrt{J}P_Jm(z, \hat{\theta})\}. \quad (9)$$

This procedure simulates a stochastic linear program based on *all* the moments. The last term in the inequality, $\sqrt{J}P_Jm(z, \hat{\theta})$, plays a crucial role. The components of this vector that correspond to the non-binding moments approach infinity as the sample size grows, $\sqrt{J}P_Jm_1(z, \hat{\theta}) \rightarrow +\infty$. So asymptotically the non-binding moments do not contribute to this distribution. That is, asymptotically the solution to the above program, $\underline{\tau}_1^*$, will be found as the lowest value of τ_1 for which the vector of inequalities corresponding to the *binding* moments are non-negative.

Now consider the binding moments. The term $\sqrt{J}P_Jm_0(z, \hat{\theta})$ will generally be stochastically bounded, and the procedure which constructs $\hat{\theta}$ will insure this term is non-negative with probability approaching one (see below). Let $O_p^+(1)$ be notation for a non-negative stochastically bounded random variable. Then, $\sqrt{J}P_Jm_0(z, \hat{\theta}) = O_p^+(1)$. Further $(\hat{\theta}, \hat{\Gamma}, \hat{\Sigma})$ converges to $(\underline{\theta}, \underline{\Gamma}, \underline{\Sigma})$. So, asymptotically, we expect the distribution obtained from simulating Z^* and calculating $\underline{\tau}_1^*(Z^*)$, as defined in equation (9), to approach the distribution obtained by simulating $\tilde{Z} \sim N(0, \underline{\Sigma})$ and calculating

$$\tilde{\tau}_1(\tilde{Z}_0) = \min\{\tau_1 : 0 \leq \underline{\Gamma}_0\tau + \tilde{Z}_0 + O_p^+(1)\}, \quad (10)$$

where \tilde{Z}_0 contains the sub-vector of elements of \tilde{Z} corresponding to the binding moments.

Now compare the distribution of $\tilde{\tau}_1(\tilde{Z}_0)$ to the distribution of $\hat{\tau}_1(\tilde{Z}_0)$ defined in Theorem 2 as $\hat{\tau}_1(\tilde{Z}_0) = \min\{\tau_1 : 0 \leq \underline{\Gamma}_0\tau + \tilde{Z}_0\}$. Fix \tilde{Z}_0 . Then any τ which satisfies the inequalities defining $\hat{\tau}_1$ will automatically satisfy the inequalities in (10). Consequently $\tilde{\tau}_1(\tilde{Z}_0) \leq \hat{\tau}_1(\tilde{Z}_0)$, which implies that $\hat{\tau}_1(\tilde{Z}_0)$ stochastically dominates $\tilde{\tau}_1(\tilde{Z}_0)$. Since we have argued that the limit distributions of $\underline{\tau}_1^*(\cdot)$ and $\tilde{\tau}_1(\cdot)$ are the same, we should then expect $\hat{\tau}_1(\cdot)$ to stochastically dominate $\underline{\tau}_1^*(\cdot)$, i.e. for any x , $Pr\{\underline{\tau}_1^* \leq x\} \geq Pr\{\hat{\tau}_1 \leq x\}$ (asymptotically).

To understand why $P_Jm_0(z, \hat{\theta})$ will be non-negative (element by element) with probability approaching one, consider the definition of $\hat{\theta}$. If there exists any θ with $P_Jm(z, \theta) \geq 0$, then $\hat{\theta}$ is taken as the θ satisfying this inequality with the lowest value of $\hat{\theta}_1$ (up to $o_p(1/\sqrt{J})$). Since our model implies that a solution θ to $P_Jm(z, \theta) \geq 0$ will exist with probability approaching one,

$P_J m_0(z, \hat{\theta}) \geq 0$ with probability approaching one. If this is the case and there are M_0 binding moments where $M_0 > K$, then $M_0 - K$ of those moments will typically be positive when evaluated at $\hat{\theta}$. These positive sample moments generate the $O_p^+(1)$ term that differentiates our simulated distribution from the limit distribution in Theorem 2. Note, however, that if there are exactly K binding moments (so $M_0 = K$), then the $O_p^+(1)$ term will actually be $o_p(1)$. In this case the limit distribution of $\underline{\tau}_1^*$ is the same as the distribution of $\hat{\tau}_1$.

We now obtain a distribution which stochastically dominates the distribution of $\hat{\tau}_1$. Let m_a denote a sub-vector of the moments that contains the binding moments with probability one. The simplest choice for m_a would be all the moments. Gather the rows of $\hat{\Gamma}$ and Z^* (from (9)) corresponding to the sub-vector of moments in m_a into the matrix $\hat{\Gamma}_a$ and the vector Z_a^* . To obtain the second approximating distribution take draws from Z^* but this time for each draw solve the linear program

$$\underline{\tau}_1^{**}(Z_a^*) = \min\{\tau_1 : 0 \leq \hat{\Gamma}_a \tau + Z_a^*\}. \quad (11)$$

By analogous reasoning to that given above for $\underline{\tau}_1^*(Z^*)$, we expect the distribution of $\underline{\tau}_1^{**}(Z_a^*)$ from (11) with $Z^* \sim N(0, \hat{\Sigma})$ to behave asymptotically as does the distribution of

$$\tilde{\tau}_1(\tilde{Z}_a) = \min\{\tau_1 : 0 \leq \Gamma_a \tau + \tilde{Z}_a\}, \quad (12)$$

where $\tilde{Z} \sim N(0, \hat{\Sigma})$ and \tilde{Z}_a is formed from \tilde{Z} as Z_a^* is formed from Z^* .

Take a fixed \tilde{Z} . Then any τ which satisfies the inequality in (12) will also necessarily satisfy the inequality defining $\hat{\tau}_1$ in Theorem 2. Consequently $\hat{\tau}_1(\tilde{Z}_0) \leq \tilde{\tau}_1(\tilde{Z}_a)$, and since the limit distributions of $\tilde{\tau}_1(\cdot)$ and $\underline{\tau}_1^{**}(\cdot)$ are the same, this implies that asymptotically $\underline{\tau}_1^{**}$ stochastically dominates $\hat{\tau}_1$, i.e. $Pr\{\hat{\tau}_1 \leq x\} \geq Pr\{\underline{\tau}_1^{**} \leq x\}$ for any x , asymptotically.

Given these simulated distributions for approximating the limiting distribution from Theorem 2, we can form the confidence intervals discussed earlier. Let q_α^* denote the α^{th} quantile of the $\underline{\tau}_1^*$ distribution, so $Pr^*(\underline{\tau}_1^* \leq q_\alpha^*) = \alpha$. Define q_α^{**} as the α^{th} quantile for $\underline{\tau}_1^{**}$ similarly. Also, let $\hat{\tau}_1$ denote the limiting distribution for $\hat{\theta}$, and define $\bar{\tau}_1^*$, $\bar{\tau}_1^{**29}$ and their α^{th} quantiles

²⁹The set of moments (m_a) used to construct $\bar{\tau}_1^{**}$ will generally differ from the set of moments used to construct $\underline{\tau}_1^{**}$.

\bar{q}_α^* and \bar{q}_α^{**} in exact analogy with the correspondingly denoted lower bound random variables.

Note that the stochastic dominance relations between $\hat{\tau}_1$ and both $\bar{\tau}_1^*$ and $\bar{\tau}_1^{**}$ will be reversed relative to the corresponding lower bound relationships. As a result the $1 - \alpha$ “outer” and “inner” confidence intervals for $[\underline{\theta}_1, \bar{\theta}_1]$ (and hence $\theta_{0,1}$) are given, respectively, by

$$(\hat{\theta}_1 - \underline{q}_{1-\alpha/2}^{**}/\sqrt{J}, \hat{\theta}_1 - \bar{q}_{\alpha/2}^{**}/\sqrt{J}), \text{ and } (\hat{\theta}_1 - \underline{q}_{1-\alpha/2}^*/\sqrt{J}, \hat{\theta}_1 - \bar{q}_{\alpha/2}^*/\sqrt{J}).$$

As noted the outer confidence interval is asymptotically conservative and the inner confidence interval will be also if there are just K binding moments at each bound.³⁰

The outer and inner $1 - \alpha$ confidence intervals for the lower bound per se (for $\underline{\theta}_1$) are, respectively,

$$(\hat{\theta}_1 - \underline{q}_{1-\alpha/2}^{**}/\sqrt{J}, \hat{\theta}_1 - \underline{q}_{\alpha/2}^*/\sqrt{J}), \text{ and } (\hat{\theta}_1 - \underline{q}_{1-\alpha/2}^*/\sqrt{J}, \hat{\theta}_1 - \underline{q}_{\alpha/2}^{**}/\sqrt{J}),$$

while the respective bounds for $\bar{\theta}_1$ are

$$(\hat{\theta}_1 - \bar{q}_{1-\alpha/2}^*/\sqrt{J}, \hat{\theta}_1 - \bar{q}_{\alpha/2}^{**}/\sqrt{J}), \text{ and } (\hat{\theta}_1 - \bar{q}_{1-\alpha/2}^{**}/\sqrt{J}, \hat{\theta}_1 - \bar{q}_{\alpha/2}^*/\sqrt{J}).$$

Again if the binding moments just identify the lower bound, then the inner confidence interval would produce asymptotically correct coverage.

A simple modification of the confidence interval methods using the second simulated distribution is to suppose m_a contains the binding moments with probability $1 - \alpha_1$. Then the $(1 - \alpha_0)^{th}$ quantile of the τ_1^{**} simulation distribution is larger than the $1 - \alpha_0 - \alpha_1$ quantile of the asymptotic distribution of $\sqrt{J}(\hat{\theta}_1 - \underline{\theta}_1)$. So one could use a multi-step procedure to construct the confidence intervals whose endpoints use quantiles of the distribution of

³⁰One way of providing guidance on the extent of any possible coverage problem with the inner confidence interval is through the following Monte Carlo procedure. Readjust the sample means of the moments so that more moments are binding at $\hat{\theta}$ than the researcher thinks could be binding at $\underline{\theta}$. Simulate samples from that distribution and construct the estimators from equation (9) generated by the simulated samples with these moments. Calculate the fraction of those estimators that fall within the inner confidence interval. This would be a consistent estimate of the coverage of the inner confidence interval if $\hat{\theta} = \underline{\theta}$, $\hat{\Sigma}(\hat{\theta}) = \Sigma(\underline{\theta})$, and the specified moments were in fact binding.

$\underline{\tau}_1^{**}$ or $\bar{\tau}_1^{**}$. First the researcher would do one (or if levels are adjusted more than one) “pre-test” of whether a particular subset of the moments are less than or equal to zero at $\underline{\theta}$ (here all the moments *not* contained in m_a). If a test with size α_1 is rejected, those moments would be dropped from the set of moments used to construct $\underline{\tau}_1^{**}$ (or similarly $\bar{\tau}_1^{**}$). Second, one would adjust the level of the quantiles of $\underline{\tau}_1^{**}$ used for the confidence interval construction and the coverage of the resulting interval accordingly³¹.

As noted there is another intuitive way of simulating the two approximating distributions when the moment inequalities are linear. That method is presented in section 3.3 and the reader who is not interested in the formalities may want to go directly to it.

Approximating Distributions: Formalities The simulation estimator $\underline{\tau}_1^*$ treats the binding and the non-binding moments symmetrically. So, to develop its asymptotic properties we will need to extend our assumptions on the binding moments to also hold for the non-binding moments. In particular, we make use of differentiability, asymptotic normality, and stochastic equicontinuity for the whole moment function. Formally, we require the extensions of Assumptions A6, A7, and A8 to hold for all the moments in m , not just the binding moments.

Assumption A6' $\mathcal{P}m(z, \theta)$ is continuously differentiable in a neighborhood of $\underline{\theta}$.

Assumption A7' $\sqrt{J}P_Jm(z, \underline{\theta}) \xrightarrow{d} N(0, \underline{\Sigma})$.

Assumption A8' For all sequences $\delta_n \downarrow 0$,

$$\sup_{\|\theta - \underline{\theta}\| \leq \delta_n} \|\sqrt{J}(P_Jm(z, \theta) - \mathcal{P}m(z, \theta)) - \sqrt{J}(P_Jm(z, \underline{\theta}) - \mathcal{P}m(z, \underline{\theta}))\| = o_p(1).$$

³¹If we hold the size of the pre-test constant, then for a large enough sample we will be able to reject that *all* non-binding moments are less than or equal to zero with arbitrarily large probability. However the fact that there was a pre-test requires us to adjust the levels of the quantiles of $\underline{\tau}_1^{**}$ or $\bar{\tau}_1^{**}$ used to construct boundaries for confidence intervals, so without further refinements even the limiting confidence interval from this procedure will be conservative relative to a confidence interval based on quantiles of the limiting distributions $\hat{\tau}_1$ and $\hat{\tau}_1$. Alternatively, we could consider a sequence of pre-tests with size declining as the sample size grows, but we do not pursue this here.

We also give formal definitions of the simulation distributions. Let $\mathcal{T}_J^* = \arg \min_{\tau} \left\| \left(\hat{\Gamma} \tau + Z^* + \sqrt{J} P_J m(z, \hat{\theta}) \right)_- \right\|$, where $Z^* \sim N(0, \underline{\Sigma})$, and $\mathcal{T}_J^{**} = \arg \min_{\tau} \left\| \left(\hat{\Gamma}_a \tau + Z_a^* \right)_- \right\|$. Then

$$\underline{\tau}_1^* = \min\{\tau_1 : \tau \in \mathcal{T}_J^*\}$$

and

$$\underline{\tau}_1^{**} = \min\{\tau_1 : \tau \in \mathcal{T}_J^{**}\}.$$

With probability approaching one, these definition coincide with the definitions given in (9) and (11) of the heuristics. The definitions here remain valid when the inequalities in (9) and (11) have no solutions. The quantiles of the simulated distributions are as defined before, though in the following theorem they are denoted with a J subscript to emphasize their dependence on sample size.

Theorem 3 *Suppose Assumptions A1-A9 hold and $(\hat{\Gamma}, \hat{\Sigma}) \xrightarrow{p} (\underline{\Gamma}, \underline{\Sigma})$. Take any $0 < \alpha < 1$.*

(a) *If m_a contains the binding moments m_0 , then*

$$\liminf_{J \rightarrow \infty} \Pr \left(\sqrt{J}(\hat{\theta}_1 - \underline{\theta}_1) \leq q_{\alpha, J}^{**} \right) \geq \alpha. \quad (13)$$

(b) *If Assumptions A6'-A8' also hold, then*

$$\limsup_{J \rightarrow \infty} \Pr \left(\sqrt{J}(\hat{\theta}_1 - \underline{\theta}_1) \leq q_{\alpha, J}^* \right) \leq \alpha. \quad (14)$$

PROOF:³² See the appendix. ♠

Note that for sufficiently large J , (13) implies that $\alpha \leq \Pr(\underline{\theta}_1 \geq \hat{\theta}_1 - q_{\alpha, J}^{**}/\sqrt{J})$, while equation (14) yields $1 - \alpha \leq \Pr(\underline{\theta}_1 < \hat{\theta}_1 - q_{\alpha, J}^*/\sqrt{J})$. These findings lead to the conservative confidence interval endpoints given above for the lower bound. Also, to the extent that Theorem 3 is used to form confidence intervals, this result only shows pointwise coverage of those intervals (see, e.g., Shaikh 2005 for more discussion of pointwise and uniform coverage).

³²It is worth noting that in the proof of Theorem 3 we show that if the first part of Assumption A9 is extended to uniqueness of the solution to $\min\{\tau_1 : \Gamma_a \tau \geq 0\}$ for Γ_a in a neighborhood of $\underline{\Gamma}_a$, then the inequality (13) is nontrivial. That is, the limit on the left of (13) is strictly less than one.

3.2 Specification Analysis and Testing

There are a number of reasons why specification testing is likely to be particularly important in our context. First, as noted above, the actual estimator the researcher uses will depend on the importance of unobservables that are known to the agent when decisions are made but not to the econometrician (ν_2). For every model that does allow for such a disturbance, there is a restricted version which does not and should provide for more efficient estimators provided the restriction is true. So often it will make sense to start out by testing whether it is necessary to allow for the ν_2 .

Second, the use of inequalities enables us to apply tools developed for the specification analysis of models for continuous unbounded outcomes to models with more complex choice sets. For example when we use inequality estimators on models with discrete outcomes the likely impact of a left out variable can be analyzed by projecting those variables down onto the included variables and analyzing the sign of the resulting projection coefficients (an analysis that is independent of the particular distributional assumptions made on the disturbances). The fact that the inequality estimators are easy to compute makes this type of specification analysis particularly useful (see the empirical examples below).

Finally, the use of inequalities adds another dimension to specification analysis; it provides some ability to investigate whether deviations from the null are likely to be due to the behavioral assumption (Assumption 1). Typically specification analyses focuses on the model's functional form or stochastic assumptions (Assumptions 2 and 3). Generalization 1 provides two sets of alternative behavioral assumptions to investigate. One alternative allows choices with expected returns that are a fraction δ less than the optimal returns, and the other alternative decreases the number of choices that can be used for comparison. Of course a direct implementation of the generalization would implicitly condition on the functional form and stochastic assumptions. We have not investigated the extent to which it is possible to distinguish between the two types of specification errors.

A Specification Test

If there is a value of $\theta \in \Theta_J$ for which $P_J m(z, \theta) \geq 0$, any reasonable specification test will yield acceptance of the hypothesis $\mathcal{P}m(z, \theta_0) \geq 0$. However, as noted above, there are frequently good reasons to expect $\min_{\theta} \| (P_J m(z, \theta))_- \|$

to be different from zero even if the underlying model is correct.

The typical GMM specification test is based on the minimized criterion function value, so it measures the distance between the sample moments and zero. With moment inequalities, a natural specification test of $H_0 : \mathcal{P}m(z, \theta_0) \geq 0$ vs. $H_1 : \mathcal{P}m(z, \theta_0) \not\geq 0$ would be based on the extent to which the inequalities are violated, or on $T_J \equiv \min_{\theta} \left\| \left(\sqrt{J} P_J m(z, \theta) \right)_- \right\|$.

In general T_J does not have a standardized limit distribution (it is not asymptotically pivotal), so to use this type of test one needs a method for obtaining appropriate critical values. First, note that under the null

$$\min_{\theta} \left\| \left(\sqrt{J} P_J m(z, \theta) \right)_- \right\| \leq \left\| \left(\sqrt{J} P_J m(z, \theta_0) \right)_- \right\| \leq \left\| \left(\sqrt{J} [P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)] \right)_- \right\|.$$

So for any ϵ ,

$$\Pr(T_J \geq \epsilon) \leq \Pr\left(\left\| \left(\sqrt{J} [P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)] \right)_- \right\| \geq \epsilon\right).$$

If θ_0 were known, the asymptotic distribution of this latter term could be approximated by simulation. Draw from a normal distribution with mean zero and variance taken as the sample variance of $m(z, \theta_0)$. Then the distribution of the norm of the negative part of this draw could be used to approximate any desired quantile.

Since θ_0 is unknown, we consider an $1 - \alpha/2$ level confidence interval for it, denoted $CI_{1-\alpha/2}$. Suppose we construct a family of random variables indexed by θ (a stochastic process in θ), say $\{Z_J(\theta)\}$, with approximately the same distribution at each θ (the marginal distributions) as $\{\sqrt{J}[P_J m(z, \theta) - \mathcal{P}m(z, \theta)]\}$. Let $\bar{z}_{\alpha, J}(\theta)$ be the $1 - \alpha/2$ quantile of $Z_J(\theta)$ and $\bar{z}_{\alpha, J}$ be the supremum of these quantiles over the values of θ in a $1 - \alpha/2$ confidence interval, i.e.

$$\Pr\left\{\|(Z_J(\theta))_-\| \geq \bar{z}_{\alpha, J}(\theta)\right\} = \alpha/2, \quad \text{and} \quad \bar{z}_{\alpha, J} \equiv \sup_{\theta \in CI_{1-\alpha/2}} \bar{z}_{\alpha, J}(\theta).$$

Then,

$$\Pr\{T_J \geq \bar{z}_{\alpha, J}\} \leq \Pr\{\theta_0 \notin CI_{1-\alpha/2}\} + \Pr\{T_J \geq \bar{z}_{\alpha, J} \mid \theta_0 \in CI_{1-\alpha/2}\} \leq \alpha,$$

so $\bar{z}_{\alpha, J}$ is a size α critical value for T_J . A more formal statement of the result follows.

Let $\Sigma(\theta)$ denote $\text{Var}(m(Z, \theta))$, and suppose $\hat{\Sigma}(\theta)$ is an estimator for $\Sigma(\theta)$. Given the data, let $Z_J^*(\theta)$ be a stochastic process such that at each θ , $Z_J^*(\theta) \sim$

$\mathcal{N}(0, \hat{\Sigma}(\theta))$. Now define $\bar{z}_{\alpha, J} = \sup_{\theta \in CI_{1-\alpha/2, J}} \bar{z}_{\alpha, J}(\theta)$, where $\Pr^*(\|Z_J^*(\theta)_-\| \geq \bar{z}_{\alpha, J}(\theta)) \leq \alpha/2$ and \Pr^* denotes probabilities taken with respect to the distribution of the $Z_J^*(\theta)$ conditional on the data. Also, let F denote the c.d.f. of the limiting distribution for $\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\|$.

Theorem 4 Suppose (a) $\sqrt{J}P_J m(z, \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0))$.; (b) $\hat{\Sigma}(\theta_0) \xrightarrow{as} \Sigma(\theta_0)$; and (c) $CI_{1-\alpha/2, J}$ is such that $\liminf_{J \rightarrow \infty} \Pr(\theta_0 \in CI_{1-\alpha/2, J}) \geq 1 - \alpha/2$. Take α such that $1 - \alpha/2 > F(0)$. Then under $H_0 : \mathcal{P}m(z, \theta_0) \geq 0$,

$$\limsup_{J \rightarrow \infty} \Pr(\min_{\theta} \|(\sqrt{J}P_J m(z, \theta))_-\| \geq \bar{z}_{\alpha, J}) \leq \alpha.$$

Proof Sketch:

Define $c_{\alpha/2}$ by $\Pr^*(\|Z_J^*(\theta_0)_-\| \geq c_{\alpha/2}) = \alpha/2$. Now note that

$$\begin{aligned} & \Pr(\inf_{\theta} \|(\sqrt{J}P_J m(z, \theta))_-\| \geq \bar{z}_{\alpha, J}) \\ & \leq \Pr(\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\| \geq \bar{z}_{\alpha, J}) \\ & = \Pr(\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\| \geq \bar{z}_{\alpha, J} \cap \{\bar{z}_{\alpha, J} \geq c_{\alpha/2}\}) \\ & \quad + \Pr(\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\| \geq \bar{z}_{\alpha, J} \cap \{\bar{z}_{\alpha, J} < c_{\alpha/2}\}) \\ & \leq \Pr(\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\| \geq c_{\alpha/2}) + \Pr(\bar{z}_{\alpha, J} < c_{\alpha/2}) \\ & \leq \Pr(\|(\sqrt{J}[P_J m(z, \theta_0) - \mathcal{P}m(z, \theta_0)]_-)\| \geq c_{\alpha/2}) + \Pr(\theta_0 \notin CI_{1-\alpha/2, J}) \end{aligned}$$

The result follows by taking limits. \square

It still remains to construct $\{Z_J^*(\theta)\}$ and compute $\bar{z}_{\alpha, J}$. Perhaps the computationally simplest method for constructing $\{Z_J^*(\theta)\}$ and finding the associated $\bar{z}_{\alpha, J}$ is as follows. Take repeated draws on $\varepsilon^* \sim N(0, I)$. For each draw set $Z_J^*(\theta) = \Sigma(P_J, \theta)^{1/2} \varepsilon^*$. Now find the largest value of $\bar{z}_{\alpha, J}$ that is less than a fraction $\alpha/2$ of the values of $\sup_{\theta \in CI_{1-\alpha/2}} \|Z_J^*(\theta)_-\|$.³³ This

³³Note that Theorem 4 does not actually require weak convergence of the process $\sqrt{J}[P_J m(z, \theta) - \mathcal{P}m(z, \theta)]$ to a Gaussian process (it only requires asymptotic normality at θ_0). We impose no conditions on the covariances of $\{Z_J^*(\theta)\}$ at different θ 's, i.e. $Cov(Z_J^*(\theta), Z_J^*(\theta'))$ is unrestricted. Any covariance process for components of $\{Z_J^*(\theta)\}$ will be sufficient as long as it doesn't violate existence of the process and satisfies the variance requirement given above. Consequently a natural alternative to the construction above would be to take $\{Z_J^*(\theta)\}$ as the Gaussian process with mean zero and covariance process given by the sample covariances evaluated at different θ .

test becomes particularly simple when the underlying moments are linear. There are, however, other ways of computing test statistics for this problem, and, generally, one would like a method that obtains a critical value as close as possible to $c_{\alpha/2}$ (as defined in the proof of Theorem 4) with minimal computational burden.³⁴

3.3 The Linear Case

In section 3.1.3 two simulation distributions were defined for the general nonlinear case. These distributions, specialized to the linear case, provide valid inference for that case. However there is another intuitive way of simulating the two approximating distributions for the linear case. Moreover though the limit distributions from these alternative simulators are the same as the limit distributions from the respective simulators provided in the earlier subsection, some preliminary Monte Carlo analysis has suggested that the approximation methods given in this subsection have better small sample performance.

Suppose that all our inequality restrictions are linear so that the j^{th} observation on the data can be partitioned as $(w_{1,j}, w_{2,j})$ and

$$m(z_j, \theta) = w_{1,j}\theta - w_{2,j}.$$

Recall that there are M moments and θ is K -dimensional, so $w_{1,j}$ is $M \times K$ and $w_{2,j}$ is a vector with M elements.

Now the population and sample moments are, respectively,

$$\mathcal{P}m(z, \theta) = (\mathcal{P}w_1)\theta - \mathcal{P}w_2 \quad \text{and} \quad P_J m(z, \theta) = (P_J w_1)\theta - P_J w_2,$$

where \mathcal{P} is a probability distribution satisfying our assumptions in section 2. Similarly the identified set and its estimator are

$$\Theta_0 = \{\theta \in \Theta : (\mathcal{P}w_1)\theta \geq \mathcal{P}w_2\}, \quad \text{and} \quad \Theta_J \equiv \arg \inf_{\theta} \left\| \left((P_J w_1)\theta - P_J w_2 \right)_- \right\|,$$

where $(f, 0)_- = \min(f, 0)$. Finally, the minimum value of the first element of the parameter vector in the identified set and its estimator are

$$\underline{\theta}_1 = \arg \inf \{\theta_1 : \theta \in \Theta_0\} \quad \text{and} \quad \hat{\underline{\theta}}_1 = \arg \inf \{\theta_1 : \theta \in \Theta_J\}.$$

³⁴There is a question of whether one could base a more powerful test on the $\bar{z}_{\alpha,J}(\theta)$. Clearly if one knew θ_0 a test which rejected if $T_J \geq \bar{z}_{2\alpha,J}(\theta_0)$ would be more powerful. One possibility is to present $\bar{z}_{2\alpha,J}(\hat{\theta}_J)$, a statistic which should approximate the more powerful test statistic but whose size will generally be greater than α .

When the moment inequalities are linear the $\underline{\Gamma}$ of section 3.1 is $\mathcal{P}w_1$, and the $\underline{\Sigma}$ of that section is $\text{Var}(w_{1,j}\underline{\theta} - w_{2,j})$.

We want to approximate the distribution of the estimator for $\underline{\theta}_1$ and use it for inference on $\underline{\theta}_1$ (the lower bound for the first dimension of θ_0). A formal expression for the asymptotic distribution is given in Theorem 2, which applies to the special case of linear moments as well. Note that the zero subscript in that theorem denotes the rows corresponding to the binding moments, i.e. the rows of w_1 and w_2 such that the population inequality holds with equality.

Before describing the two simulation methods for approximating the distribution of the estimator, we introduce some notation. Let $w_j = \left(\text{vec}(w_{1,j})', w_{2,j}' \right)'$. Define $\Sigma_w = \text{Var}(w_j)$, and let the sample covariance of $\{w_j\}_{j=1}^J$ be $\hat{\Sigma}_w$.

To obtain the first linear case approximation, we take ns independent draws from a normal centered at $P_J w$ with variance-covariance $\hat{\Sigma}_w/J$. Then, treat each draw as a random data sample, and compute the distribution of the inequality estimators formed from these ns samples.

More formally let $y_s^* \equiv \left(\text{vec}(y_{1,s}^*)', (y_{2,s}^*)' \right)$, where $y_{1,s}^*$ and $y_{2,s}^*$ have the same dimensions as $w_{1,j}$ and $w_{2,j}$, be independent draws from a normal centered at zero with covariance matrix equal to $\hat{\Sigma}_w$. For each draw find

$$\Theta_s^I = \arg \inf_{\theta} \left\| \left((P_J w_1 + y_{1,s}^*/\sqrt{J})\theta - P_J w_2 - y_{2,s}^*/\sqrt{J} \right) \right\|_-$$

which may be a set, and

$$\hat{\underline{\theta}}_{1,s}^I = \inf \{ \theta_1 : \theta \in \Theta_s^I \}.$$

As J and ns grow large, the distribution of $\{(\hat{\underline{\theta}}_{1,s}^I - \hat{\underline{\theta}}_{1,J})\}_{s=1}^{ns}$ will be stochastically dominated by the asymptotic distribution of $(\hat{\underline{\theta}}_{1,J} - \underline{\theta}_1)$. It is used for the lower end point for the ‘‘inner’’ confidence interval for both $\theta_{0,1}$, and for $\underline{\theta}_1$.

To see the relationship of this distribution to that used for our general case, let $\tau_{1,s}^I = \sqrt{J}(\hat{\underline{\theta}}_{1,s}^I - \hat{\underline{\theta}}_1)$, and note that the $\{\tau_{1,s}\}_s$ generated in this way can also be obtained by finding the minimum value of τ_1 in the set

$$\tau_s^I = \arg \inf_{\tau} \left\| \left((P_J w_1 + y_{1,s}^*/\sqrt{J})\tau + (y_{1,s}^*\hat{\underline{\theta}} - y_{2,s}^*) + \sqrt{J}(P_J w_1 \hat{\underline{\theta}} - P_J w_2) \right) \right\|_- \quad (15)$$

To compare the set in equation (15) to the linear version of the analogous set in section 3.1.3 (i.e. to the set defined in (9)), note that in the linear case $y_{1,s}^* \hat{\underline{\theta}} - y_{2,s}^*$ is a random draw on the Z^* needed for (9). Thus the only difference between the definition of the set in equation (15) and the linear case of the set in (9) is the $y_{1,s}^*/\sqrt{J}$ in (15), a term which goes to zero with sample size.

Just as in section 3.1.3, for the second linear case approximation, we consider only those moments which could possibly be binding.³⁵ Denote the corresponding rows of w_1 and w_2 with an ‘‘a’’ subscript, $w_{1,a}$, $w_{2,a}$. Let $w_{a,j} = (\text{vec}(w_{1,a,j})', w_{2,a,j}')'$. Finally let $\Sigma_w^a = \text{Var}(w_{a,j})$ with estimator $\hat{\Sigma}_w^a$, based on the sample covariance of $\{w_{a,j}\}_{j=1}^J$.

To obtain the second approximation, we take ns independent draws from a normal centered at $P_J(\text{vec}(w_{1,a})', 0)'$ with variance-covariance $\hat{\Sigma}_w^a/J$. Next, form the moments from each of these pseudo random samples, subtract $P_J w_1^a \hat{\underline{\theta}}_1$ from those moments, and compute the distribution of the resultant inequality estimators. Centering the draws on the last $\dim(w_2^a)$ components of the normal at zero (rather than at $P_J w_2^a$) and subtracting $P_J w_1^a \hat{\underline{\theta}}_1$ from the moments, is a way of recentering the moments at zero (just as is done in section 3.1.3).

More formally let $y_{a,s}^* \equiv (\text{vec}(y_{1,a,s}^*)', (y_{2,a,s}^*)')$, where $y_{1,a,s}^*$ and $y_{2,a,s}^*$ have the same dimensions as $w_{1,a,j}$ and $w_{2,a,j}$, and their elements are drawn from a normal centered at zero with covariance matrix equal to $\hat{\Sigma}_w^a$. For each draw find

$$\Theta_s^O = \arg \inf_{\theta} \left\| \left((P_J w_{1,a} + y_{1,a,s}^*/\sqrt{J})(\theta - \hat{\underline{\theta}}) + (y_{1,a,s}^* \hat{\underline{\theta}} - y_{2,a,s}^*)/\sqrt{J} \right) \right\|$$

and

$$\hat{\underline{\theta}}_{1,s}^O = \arg \inf \{ \theta_1 : \theta \in \Theta_s^O \}.$$

As J and ns grow large, the distribution of $\{\sqrt{J}(\hat{\underline{\theta}}_{1,s}^O - \hat{\underline{\theta}}_1)\}_{s=1}^{ns}$ will stochastically dominate the asymptotic distribution of $\sqrt{J}(\hat{\underline{\theta}}_{1,J} - \underline{\theta}_1)$. It is used for

³⁵If this is determined by a ‘‘pretest’’ which throws out moments which are positive and highly significant, the $1 - \alpha$ level of the confidence interval given below must be adjusted appropriately.

the lower end point for the “outer” confidence interval for both $\theta_{0,1}$, and for $\underline{\theta}_1$.

Similar to the inner approximation, this method produces a simulation distribution for $\tau_{1,s}^O = \sqrt{J}(\hat{\underline{\theta}}_{1,s}^O - \underline{\theta}_1)$ which we could have obtained directly from

$$\tau_s^O = \arg \inf_{\tau} \left\| \left((P_J w^{1,a} + y_{1,a,s}^*/\sqrt{J})\tau + (y_{1,a,s}^*\hat{\underline{\theta}} - y_{2,a,s}^*) \right) \right\|. \quad (16)$$

The distribution obtained in this fashion is asymptotically equivalent to the distribution obtained for the linear case of the general procedure which uses the set defined in equation (11). The only difference in the equation defining the set estimators is the $y_{1,a,s}^*/\sqrt{J}$ term, a term which converges to zero with the sample size.

Since the simulation methods provided in this section are distinct from the earlier ones, we provide a formal statement of their properties.

Theorem 5 *Suppose Assumptions A1-A9 hold and $\hat{\Sigma}_w \xrightarrow{p} \Sigma_w$. Take $0 < \alpha < 1$.*

(a) *If the binding moments are contained in the subset of rows of (w_1, w_2) denoted by $(w_{1,a}, w_{2,a})$, then*

$$\liminf_{J \rightarrow \infty} \Pr(\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1) \leq q_{\alpha,J}^O) \geq \alpha. \quad (17)$$

(b) *If Assumptions A6'-A8' also hold, then*

$$\limsup_{J \rightarrow \infty} \Pr(\sqrt{J}(\hat{\underline{\theta}}_1 - \underline{\theta}_1) \leq q_{\alpha,J}^I) \leq \alpha. \quad (18)$$

where $q_{\alpha,J}^I$ and $q_{\alpha,J}^O$ are, respectively, the α^{th} quantiles of the distributions $\hat{\tau}_1^I$ and $\hat{\tau}_1^O$ conditional on the data.

PROOF SKETCH: The linear case result follows the proof of Theorem 3 after modification to account for simulation affecting the “slope” term in the stochastic linear programs. \square

Specification testing for the linear case is the same as for the general nonlinear case in section 3.2.

4 Empirical Examples

We now introduce our two empirical examples. One is an ordered choice problem while the other is a bargaining problem. In each case we begin by outlining the substantive problem. Next we describe how the application fits into the framework of section 2. Finally, we provide our method of moments inequality estimators, discuss their properties and compare them to familiar alternatives. We conclude with a brief discussion of the empirical results.

4.1 Ordered Choice

This section is based on Ishii (2004). She analyzes how ATM networks affect market outcomes in the banking industry. The part of her study considered here is the choice of the number of ATMs. More generally this example shows how the techniques proposed in this paper can be used to empirically analyze multiple agent “lumpy” investment problems, or investment problems which are not convex for some other reason.³⁶

Ishii uses 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. Note that there are likely to be many possible Nash equilibria to this game.

Ishii (2004) estimates a demand system for banking services and an interest rate setting equation. Both are estimated conditional on the number of ATMs of the bank and its competitors, i.e. on (d_i, d_{-i}) . The demand system has consumers choosing among a finite set of banks with consumer and bank specific unobservables (as in Berry, Levinsohn, and Pakes 1995). The indirect utility of the consumer depends on the distance between the consumer’s home and the nearest bank branches, the consumer’s income, interest rates on deposits, bank level of service proxies, the size of the ATM network, and the distribution of ATM surcharges (surcharges are fees that ATM users pay to an ATM owner when that owner is not the user’s bank). Interest rates are set in a simultaneous move Nash game. This setup provides Ishii (2004)

³⁶Actually Ishii’s problem has two sources of non-convexities. One stems from the discrete nature of the number of ATMs chosen, the other from the fact that network effects can generate increasing returns as we increase the number of ATMs.

with the parameters needed to compute the banks' earnings conditional on its own and its competitors ATM networks.³⁷

To complete her analysis of ATM networks Ishii requires estimates of the cost of setting up and running ATMs. These costs are central to the public debate on alternative “market designs” for the ATM network (of particular interest is the analysis of systems that do not allow surcharges). This paper provides initial estimates of those costs, while Ishii (2004) provides robustness tests and considers the implications of the results.

4.1.1 The ATM Choice Model: Theory and Econometric Issues

To obtain the cost estimates we model the choice of the size of a network, that is the choice of $d_i \in \mathcal{D} \subset \mathcal{Z}^+$, the non-negative integers. Here we only attempt to estimate an average (across banks) of the marginal cost of buying and installing an ATM. Suppose bank profits take the following form

$$\pi(d, d_{-i}, y_i, \nu_{1,i,d}, \nu_{2,i}) = R(d, d_{-i}, y_i) - (\nu_{2,i} + \theta_0)d + \nu_{1,i,d}, \quad (19)$$

where $R(d, d_{-i}, y_i)$ is the profits, aside from ATM cost, that would be earned in the second stage if the firm chose d and its competitors chose d_{-i} in the first stage, θ_0 is the average (across banks) of the marginal cost of purchasing and installing ATM's.

The function $R(\cdot)$ in equation (19) is obtained from the first stage of Ishii's analysis. Note that to find the returns that would be earned were $d \neq d_i$ (the firm's actual choice), we have to solve out for the equilibrium interest rates that would prevail were the alternative network chosen. The unobservables $\nu_{1,i,d}$ and $\nu_{2,i}$ are directly interpretable in terms of the unobservables defined in section 2.2. Specifically, $\nu_{1,i,d,d'} = \nu_{1,i,d} - \nu_{1,i,d'}$ results from expectational or measurement error so $\mathcal{E}[\nu_{1,i,d,d'} | \mathcal{J}_i] = 0$. Additionally, $\nu_{2,i}$ is part of bank i 's information set when it makes its decision so it represents a bank specific component of marginal cost that the agent knew when it made its decision but the econometrician does not observe. We assume $\mathcal{E}(\nu_{2,i}) = 0$, which is equivalent to defining θ_0 to be the average of the marginal costs.³⁸

³⁷These 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.

³⁸Note that in terms of our prior notation this implies that $\nu_{2,i,d,d'} = -\nu_{2,i}(d-d')$, which

Clearly two necessary conditions for Assumption 1 are that the expected increment to returns from the last ATM the bank installed were greater than its cost of an ATM, while the expected increment to returns from adding one ATM more than the number actually installed was less than that cost. We use these two differences as our $\Delta\pi(\cdot)$.³⁹ So $m = 2$, and our moment condition is based on the vector of profit differences

$$0 \leq \mathcal{E}\Delta\pi(\cdot) = \mathcal{E} \begin{pmatrix} \mathcal{E}[R(d_i, d_{-i}, y_i) - R(d_i - 1, d_{-i}, y_i)|\mathcal{J}_i] - \theta_0 - \nu_{2,i} \\ \mathcal{E}[R(d_i, d_{-i}, y_i) - R(d_i + 1, d_{-i}, y_i)|\mathcal{J}_i] + \theta_0 + \nu_{2,i} \end{pmatrix}.$$

The simplicity of the model makes this a particularly good example for illustrating how inequality analysis works. Set

$$\Delta\mathbf{r}(\cdot, \theta) = \begin{pmatrix} R(d_i, d_{-i}, y_i) - R(d_i - 1, d_{-i}, y_i) - \theta \\ R(d_i, d_{-i}, y_i) - R(d_i + 1, d_{-i}, y_i) + \theta \end{pmatrix}.$$

Recall that we form moment conditions by interacting $\Delta\mathbf{r}(\cdot)$ with $h(x)$. Consider first using only the moment conditions generated by $h(x_i) \equiv 1$, i.e. by $\Delta\mathbf{r}(\cdot) \otimes 1$. Then 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

$$\begin{aligned} P_J m_L(z, \theta) &= \frac{1}{J} \sum_j \frac{1}{n_j} \sum_i [R(d_i^j, d_{-i}^j, y_i^j) - R(d_i^j - 1, d_{-i}^j, y_i^j) - \theta] \quad (20) \\ &= \frac{1}{J} \sum_j [\Delta\bar{R}_L^j(\cdot) - \theta] = \Delta\bar{R}_L - \theta, \end{aligned}$$

is the standard restriction on the form of the structural disturbances in ordered choice models. To see that that this specification satisfies Assumption 3 for some $(\chi^i(\cdot), h(\cdot))$, take $h(\cdot) = 1$ and $d' = d + \kappa$, where κ is a fixed integer. Then $\mathcal{E}(\nu_{2,i,d,d'}) = \kappa\mathcal{E}(\nu_{2,i}) = 0$ which implies Assumption 3 with $\chi_{d_i}^i(d_i + \kappa) = 1$ and 0 otherwise.

³⁹These 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. Note also that in forming profits at alternative d' we are assuming either that the solution to the second stage problem for interest rates is unique, or, if not unique, that our method for computing interest rates picks out the equilibria that would have been played.

where

$$\Delta \bar{R}_L^j(\cdot) \equiv \frac{1}{n_j} \sum_i [R(d_i^j, d_{-i}^j, y_i^j) - R(d_i^j - 1, d_{-i}^j, y_i^j)], \text{ and } \Delta \bar{R}_L \equiv \frac{1}{J} \sum_j \Delta \bar{R}_L^j(\cdot).$$

Analogously 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

$$P_{Jm_R}(z, \theta) \equiv \Delta \bar{R}_R + \theta. \quad (21)$$

The set of θ 's that satisfy the sample analogues of our theoretical restrictions are the values of θ that make both equations (20) and (21) nonnegative. Since $(\Delta \bar{R}_L, \Delta \bar{R}_R)$ are the changes in revenue resulting from first an increase and then a decrease in the number of ATM's, we expect $\Delta \bar{R}_L$ to be positive while $\Delta \bar{R}_R$ should be negative. If $-\Delta \bar{R}_R < \Delta \bar{R}_L$ then

$$\Theta_J = \{\theta : -\Delta \bar{R}_R \leq \theta \leq \Delta \bar{R}_L\}$$

while if $-\Delta \bar{R}_R \geq \Delta \bar{R}_L$ there is a no θ which satisfies our restrictions in sample and Θ_J is the singleton $.5[-\Delta \bar{R}_R + \Delta \bar{R}_L]$.⁴⁰

Increasing The Number of Instruments. If we add instruments each new instrument produces a pair of additional inequalities (one for the change from the left and one for the change from the right). For implementation, one needs variable(s) x_i and a nonnegative function h satisfying Assumption 3. For instance, if k indexes instruments and $\mathcal{E}(\nu_{2,i}|x_i) = 0$, then a corresponding moment inequality is

$$0 \leq \mathcal{E} \left[\left(R(d_i, d_{-i}, y_i) - R(d_i - 1, d_{-i}, y_i) - \theta_0 \right) h(x_{k,i}) \right].$$

which yields an upper bound on θ_0 ,

$$\theta_0 \leq \frac{\mathcal{E} \left[\left(R(d_i, d_{-i}, y_i) - R(d_i - 1, d_{-i}, y_i) \right) h(x_{k,i}) \right]}{\mathcal{E}(h(x_{k,i}))}. \quad (22)$$

⁴⁰In the simple case where $h(x) \equiv 1$, if $r(\cdot)$ is concave in d_i then, at least in expectation, $\Delta \bar{R}_L \geq -\Delta \bar{R}_R$, so we do not expect the set of minimizers to be a singleton. Once we add instruments, however, concavity no longer ensures that the inequalities will be satisfied by a set of θ values.

Let

$$\Delta\bar{R}_{k,L} = \frac{\frac{1}{J} \sum_j \frac{1}{n_j} \sum_i (R(d_i^j, d_{-i}^j, y_i^j) - R(d_i^j - 1, d_{-i}^j, y_i^j)) h(x_{k,i}^j)}{\frac{1}{J} \sum_j \frac{1}{n_j} \sum_i h(x_{k,i}^j)}$$

(and similarly for $\Delta\bar{R}_{k,R}$). Then moment inequality estimation leads to

$$\Theta_J = [\max_k \{-\Delta\bar{R}_{k,R}\}, \min_k \{\Delta\bar{R}_{k,L}\}].$$

So Θ_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 $\theta_k < \theta_0$. By using this max as our estimator 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, so in expectation the greatest lower bound estimator will be larger than $\max_k \theta_k$). The extent of the bias should increase with the number of inequalities. So when there are a large number of inequalities and some give lower bounds close to θ_0 , we should not be surprised if the estimated lower bound is greater than θ_0 . Analogously, since the estimate of the upper bound is a minimum, it should not be surprising if the upper bound estimate is less than θ_0 . Of course, if the lower bound is greater than θ_0 *and* the upper bound is less than θ_0 , then the estimate Θ_J is just a singleton (even if the true Θ_0 is an interval). This accentuates the need for a test of the moment inequalities with good small sample properties.⁴¹

Tests for the Presence of ν_2 . Assume the x used as instruments are contained in the appropriate information sets and are orthogonal to any unobserved cost differences known to the agents. Then one key difference between models with and without ν_2 is that in the model without ν_2 we can use the actual decision, or d , as an instrument, and in the model with a ν_2 use of d as an instrument would violate Assumption 3. Accordingly one

⁴¹Similar issues arise in obtaining the upper and lower bounds to the distribution of values in independent private value auctions; see Haile and Tamer (2003). It suggests a role for a small sample correction, but that is a topic beyond the scope of this paper.

way of determining the importance of ν_2 is to compare the specification test statistics for the moments excluding d as an instrument with those including d as an instrument.

Increasing the Number of Parameters. Change the specification so that the cost of setting up and operating an ATM equals $\theta_0 + \theta_1 w$ where w can be either bank or market specific. Again beginning with the case that $h(x) \equiv 1$, we have two sample moments

$$P_J m_L(z, \theta) = \Delta \bar{R}_L - \theta_0 - \theta_1 \bar{w},$$

where $\bar{w} = J^{-1} \sum_j n_j^{-1} \sum_i w_i^j$, and

$$P_J m_R(z, \theta) = \Delta \bar{R}_R + \theta_0 + \theta_1 \bar{w}.$$

If we plot the two inequalities $P_J m_L(z, \theta) \geq 0$ and $P_J m_R(z, \theta) \geq 0$ on a graph, their boundaries will be given by two parallel lines. If $\Delta \bar{R}_L > -\Delta \bar{R}_R$, then Θ_J , the estimate of Θ_0 , will be the area between the two parallel lines. If we add the product of the two profit differences with another instrument, say the number of branches, then provided Θ_J is not a singleton, it will be the intersection of the area between two sets of parallel lines with different slopes, or a parallelogram. If further moments are added, we obtain the intersection of the areas between a larger number of parallel lines. With three parameters we would look for the intersection between planes, and so on.

Boundaries. If the choice set has a boundary that is chosen by some agents, then there may be moments which cannot be constructed for those agents. In our example the choice set is bounded below by zero, and there are markets in which a number of banks chose not to purchase ATM's. In these cases, we cannot compute the change from the left, which corresponds to comparing profits at the observed choice to profits at one less than the observed choice. The sample mean of the structural error for those who are not at the boundary then converges to the expected value of the structural error conditional on not being at the boundary, and we have to check that the sign of that expectation is negative as required by Assumption 3. But,

$$\mathcal{E}[-\nu_{2,i} | d_i \geq 1] \geq \mathcal{E}[-\nu_{2,i}] = 0.$$

To show this inequality note that $\mathcal{E}(-\nu_{2,i}) = 0$, and $\mathcal{E}(-\nu_{2,i})$ is a weighted average of $\mathcal{E}[-\nu_{2,i}|d_i \geq 1]$ and $\mathcal{E}[-\nu_{2,i}|d_i = 0]$. So it is enough to show that

$$\mathcal{E}[-\nu_{2,i}|d_i = 0] \leq \mathcal{E} \left[-\nu_{2,i} - \nu_{2,i} \leq -\left(R(1, d_{-i}, y_i) - R(0, d_{-i}, y_i) - \theta_0 \right) \right] \geq \mathcal{E}[-\nu_{2,i}].$$

Hence the fact that there is a boundary on the left can indeed cause a violation of Assumption 3. Note that if the structural error represented a component of returns (instead of costs), or if the boundary was from above rather than below,⁴² then the conditional expectation of ν_2 for those observations that were not bounded would have had the opposite sign. In these cases the boundaries do not violate our Assumption 3 and all we have to do to deal with the boundary is to drop those observations which are constrained by it.

In cases where a boundary causes a violation of Assumption 3, one way to circumvent the boundary problem is by substituting a random variable which is known to have the appropriate inequality relationship to the $\nu_{2,i}$ for the missing observations, and averaging across the full sample. For the ATM example we could substitute a number which is larger than any reasonable ATM cost for the missing change from the left for banks with no ATM's.⁴³

Alternative Estimators: Ordered Choice Models. Ordered choice is one of two models traditionally used for such problems. In our notation the typical ordered choice model sets $\nu_1 \equiv 0$ in equation (19), assumes a particular distribution for ν_2 conditional on the other determinants of profits, and forms the likelihood of the observed d . This model, then, does *not* allow for expectational errors (which, in multiple agent problems, means that it does not allow for asymmetric information), or measurement error. Moreover it can only allow for a non-zero correlation of the $\nu_{2,i}$ of the different agents if none of the agents' decisions affect any other agent's profits.

Regardless of the distribution chosen, the ordered log-likelihood of any θ in our data is *minus infinity*, and so can not be estimated. This occurs

⁴²Note, however, that if the boundary was from above, then it would create a similar problem for the other inequality (using differences "from the right").

⁴³To get some indication of whether a potentially problematic boundary is of empirical importance, we can use a function of an instrument to select a subsample of firms which are unlikely to be at the boundary (in our case large firms), and redo the estimation procedure. A large difference in the estimates from the selected sample could indicate a need to worry about the boundary problem.

because if our “difference from the left” is less than our “difference from the right” for one or more observations there will be no value of $\theta + \nu_2$ that rationalizes the observed choices (in this case if it was profitable to purchase the last ATM, the model says that it must have been profitable to purchase the next ATM). Note that as long as there is some uncertainty when decisions are made we should expect some agent’s difference from the left to be less than its difference from the right even if all agents are behaving optimally.⁴⁴

Alternative Estimators: First Order Conditions (FOC). If we were willing to ignore the discrete nature of our control we could apply Hansen and Singleton’s (1982) FOC estimator to the ordered choice problem. The FOC estimator assumes that there is no structural error ($\nu_2 \equiv 0$), and attributes all differences in outcomes not explained by observables to ν_1 .

Given these assumptions and some mild regularity conditions, if the agents are maximizing with respect to continuous controls, the first order condition for agents with a $d > 0$ (i.e. away from the boundary of the choice set) must have an expectation of zero conditional on their information sets. As a result, provided $x \in \mathcal{J}$, a consistent estimator of θ can be found by minimizing

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

There are two differences between these moment conditions and those that define the inequality estimator. First the inequality estimator uses a discrete analogue of the derivative; i.e. the derivative is replaced with inequalities from two discrete changes (one from the left and one from the right).⁴⁵ Second, as originally formulated the first order condition estimator

⁴⁴One might be able to modify the simple ordered model to allow for some of the phenomena it rules out and, by doing so, avoid the possibility of events occurring 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. Alternatively, one might be able to allow for asymmetric information of a particular form, select among the multiple equilibria for that form, construct optimal strategies for that selection, and then construct the associated likelihood. However, this both needlessly complicates the estimation problem and makes the estimates dependent on much more detailed assumptions than those maintained above.

⁴⁵This assumes an inequality model with maximizing behavior and that the inequality estimator only uses the inequalities generated from the two adjacent possible choices.

does not allow for ν_2 and as a result (i) could use d as an instrument, and (ii) does not face any selection issues due to boundaries. We note, however, that we could reformulate the Hansen-Singleton model to allow for an additive ν_2 error, in which case consistency would require us to choose instruments and treat boundaries precisely as we do for our inequality estimator.

4.1.2 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.⁴⁶ The moment inequalities are derived as described above. The instruments used when we refer to the full set of instruments (and for the first order condition estimator) include a 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).

Table 1 contains the inequality estimators of the cost parameter (they represent costs over a six month period).⁴⁷ All runs were done twice, once using a Euclidean norm as the distance metric and once using an absolute value norm. We only present the parameter estimates and confidence intervals obtained using a Euclidean norm, as those using the absolute value norm hardly differed (see Ishii 2005). Since the test statistics using the different

The FOC model is not sufficiently flexible to estimate subject to the weaker behavioral assumptions we considered in our generalizations, and does not enable the researcher to add other inequalities to improve the efficiency of the estimator.

⁴⁶The 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. 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). Since the number of banks per market varies so widely, we weighted our market averages with the square root of the number of banks in each market before averaging across markets (this generates a small improvement in confidence intervals).

⁴⁷All estimators for both empirical problems analyzed in this paper were obtained using the “fmincon” algorithm in Matlab. In the linear case, “fmincon” finds the *argmin* of $F(\theta)$ subject to the linear constraints $A\theta \leq B$. By setting $F(\theta) = \theta_k$ and then $F(\theta) = -\theta_k$ for the different components of θ we obtain the vertices of Θ_J . For details on the search method used in fmincon see <http://design1.mae.ufledu/enkin/egm6365/AlgorithmConstrained.pdf>.

norms did differ somewhat, we present two sets of test results.

The first row provides the results when only a constant term is used as an instrument (the $h(x) = 1$ case). Then the estimate is an interval, $\Theta_J = [32,006, 32,492]$, but the interval is quite short. With the same number of parameter estimates as inequalities, there is an unambiguous confidence interval and it places the true θ_0 between \$23,416 to \$41,082 dollars with 95% probability. Not surprisingly then when the rest of our instruments are added, the interval collapses to a point \$32,338. The “inner” simulated confidence interval now shortens to [\$31,114, \$36,185], but the outer confidence interval actually increases slightly, indicating just how large a difference there can be between the two confidence intervals for samples of this size. Accordingly we stick with [\$23,400, \$41,000] for our conservative confidence interval.

Since the estimates in row 2 are point estimates we want to test whether the data is consistent with the inequalities holding, that is we want to use the test for $H_0 : \mathcal{P}m(z, \theta_0) \geq 0$ provided in Theorem 4. The simulated distribution of the test statistic from two thousand simulation draws is described in Figure 1 (which partitions the draws on the test statistic into twenty-five bins) and Figure 2 (into fifty bins). With 25 bins it is hard to tell the difference between that distribution and a half normal; recall that the test takes its values from the negative parts of mean zero moments. In the 50 bin figure we see the differences between the simulated and half normal distributions generated by the fact that different moments will bind in different simulation draws.

The bottom rows provide the ratio of the value of our objective function to the simulated critical value of the test statistic when $\alpha = .05$ for both the Euclidean and absolute value norms. When the actual decision was not included in the instrument set, the ratio was .42 using the Euclidean norm and .96 using the absolute value norm. The critical value is one, so we accept the null in both cases.⁴⁸ Next we added the actual number of ATMs chosen to the instrument set. The test statistic jumped to .6 using the Euclidean norm, and to 1.36 using the absolute value norm. A test result of 1.36 indicates rejection at any reasonable significance level. We conclude that the data indicate a need to allow for unobserved cost components, but provide no reason to worry about the specification once we do.

⁴⁸Though close to one when we used the absolute value norm, we could not reject the null in this case even if we artificially assumed away the variance in the test statistic caused by the fact that we did not know θ_0 exactly. That is, when we assumed $\hat{\theta}$ exactly equalled the true θ , the test statistic was .97, still indicating acceptance of the null.

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

	θ_J	95% CI for θ	
		LB	UB
1. $h(x) \equiv 1$	[32,006, 32,492]	23,416	41,082
2. $h(x) = \text{full } d > 0$	32,338	31,114	36,185
Same; Conservative CI		23,064	43,206
3. $h(x) = \text{full}, d \geq 0$	32,477	31,245	36,153
<i>Different Choices of $D(d_i)$ ($h(x) = \text{full}$)</i>			
4. $\{d : d - d_i = 2\}$	32,432	31,289	37,170
5. $\{d : d - d_i = 1, 2\}$	32,412	31,513	35,310
<i>Extending the Model ($h(x) = \text{full}$)</i>			
6. θ_b (in branch ATM)	34,990	34,161	38,129
7. θ_r (remote ATM)	35,358	31,022	43,825
<i>Test Statistics</i>		$d \notin IV$	$d \in IV$
T(observed)/T(critical at 5%), Euclidean		.42	.60
T(observed)/T(critical at 5%), Abs. Value		.96	1.36

Table 2: **First Order Conditions, ATM Costs***

	Coeff.	Std. Error
θ_{01} (constant)	38,491	7,754
θ_0 (in-branch constant)	50,132	11,102
θ_1 (remote constant)	55,065	12,010

* There are 291 banks in 10 markets. The FOC 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.

Five percent of the observations have $d = 0$. In the first two columns we keep the inequality from the right for these observations, and simply drop those observations in constructing the inequalities from the left. The banks that did not have ATM's were the smallest banks, and our estimates indicate that the return to the first ATM is increasing in bank size. The estimates in the “full, $d \geq 0$ ” row, then, come from substituting the average return of the first ATM among the banks that had ATMs for the unobserved returns for banks that did not have ATMs. As expected this increases the estimates, but by under 1%, indicating that boundary problems are likely to have only a minor impact on the empirical results.

Next we return to the model in Assumption 1 and assume that $D(d_i) = \{d : |d - d_i| = 2\}$ and $\delta = 0$. This allows agents to make ATM choices that are one ATM more or less than the optimal, but not more than one. The results using the weaker restriction on choices yield an estimate which is very close to the original estimate. When we consider alternatives that are one or two ATMs from the observed number, $D(d_i) = \{d : |d - d_i| = j \text{ for } j = 1, 2\}$, the estimate is unchanged but the simulated confidence interval is now only \$31,513 to \$35,310.

Finally we consider if there is a difference in cost for “in-branch” and “remote” ATM locations. To do so the model is extended to allow for a choice of in-branch ATMs, say d_b , and remote ATMs, say d_r . The amended model has $\pi_i(d, \cdot) = R_i(d, \cdot) - d_b\theta_b - d_r\theta_r - \nu_{2,i}(d_b + d_r) + \nu_{1,i,d}$. We get point estimates for each cost, but θ_r is only about 1% higher than θ_b ,⁴⁹ and both are within the confidence interval for the model with one θ parameter.

The fact that the results when we set $D(d_i) = \{d : |d - d_i| = 2\}$ are similar to those when we set $D(d_i) = \{d : |d - d_i| = 1\}$, indicates that there is no reason to doubt that firms, at least on average, act optimally. This bodes well for the first order condition estimator that we now turn to, as that estimator can not be modified to allow for non-optimal choices. Table 2 shows the FOC cost estimate to be \$38,491, with a standard error of \$7,754 (about equal to the difference between the FOC point estimate and the inequality estimate). When we allow for a separate θ_r and θ_b , the differences between the FOC

⁴⁹We initially expected a cost advantage to in-branch locations. However on going back to the data we found that 16 banks own remote ATM sites while having branches that lack an ATM; a fact which indicates either lower costs or greater benefits to remote ATM's for at least some banks. Also banks may find it optimal to install more, and/or more expensive, machines in their branches thus offsetting other branch cost advantages. Unfortunately we do not have the data nor the model needed to investigate these possibilities further.

estimator and the inequality estimator become more noticeable, as now the FOC estimator is outside any of the confidence intervals for the inequality estimators, and the standard errors of the FOC estimator grow.

Perhaps the most notable characteristic of the inequality estimators is how stable they were across alternative perturbations (we changed instruments, the d' , and the model specification to allow for different types of ATM's). Ishii (2004) notes that the cost estimates presented here are higher than what had previously been thought appropriate and uses them to examine: the effect of surcharges on concentration in banking, the welfare impacts of alternative market designs conditional on the given ATM network, and the optimality of the size of that network.

4.2 Discrete Choice and Buyer/Seller Networks.

The section complements Ho (forthcoming and 2004b). Her goal is to analyze the interactions between privately insured consumers, Health Maintenance Organizations (HMOs), and hospitals in the U.S. health care market. Ho structures the analysis as a three stage game. In the last stage consumers choose an HMO given the networks of hospitals the HMOs have contracted with and the premiums set by the HMOs. In the second stage HMOs set their premiums in a Nash equilibrium which conditions on both consumer preferences and the hospital networks of the HMOs. The first stage sets contracts which determine which hospitals each HMO has access to and the transfers the hospitals receive for the services supplied to the HMOs they contract with.

This section provides an analysis of the first stage, that is of the HMO-hospital contracting game. We develop here a framework capable of empirically analyzing the nature of contracts that arise in a market in which there are a small number of both buyers and sellers all of whom have some “market power”. Similar issues arise in the analysis of many markets where vertical relationships are important.

There are a number of ways to model buyer-seller interactions in these markets and then use Assumption 1 to provide inequalities which can be matched to data. We assume that sellers (or hospitals) simultaneously make take it or leave it offers to buyers (or HMOs) of contracts. The HMOs respond simultaneously by either accepting or rejecting each offer. The offers themselves are assumed to be proprietary; when the HMO makes its decision

it does not know the offers made to its competitors.⁵⁰

We then analyze the “reduced form” relationship between contract parameters and buyer, seller, and market characteristics. Note that we do not attempt to uncover the structural model which determines why the contracts offered were in fact offered. Moreover there are likely to be many configurations of hospital networks that satisfy the inequalities we take to data, and we do not investigate how this multiplicity of possible equilibria gets resolved. The hope is that the characterization of contracts obtained here will help determine the relevance of alternative more detailed models, and, to the extent policy and environmental changes do not effect the reduced form relationship per se, provide some idea of how changes in the environment are likely to impact HMO/hospital transfers.

4.2.1 The Model

We begin with a brief overview of how the consumer demand for HMO’s was obtained (for more detail see Ho, forthcoming). A consumer’s utility from a hospital network conditional on the consumer having a given diagnosis is estimated from consumer level data on hospital choices and a discrete choice demand system. The consumer’s expected utility from a given network is then constructed as the sum of (demographic group specific) probabilities of various medical conditions times the utility the consumer gets from the network should it have a medical condition. The individual chooses its HMO as a function of this expected utility, the premiums the HMOs charge, and other observed and unobserved plan characteristics. The function determining the utility from different HMO’s is estimated from market level data on consumers’ HMO choices (as in Berry, Levinsohn, and Pakes 1995).

We assume that conditional on any set of HMO-hospital contracts, the premiums the HMOs charges their members are set in a Nash premium setting equilibrium (and if that equilibrium is not unique, we know the selection mechanism). Given premiums, we construct each HMO’s profits as premiums from the consumers who chose that HMO minus the costs of hospital care for those consumers.

Hospitals are indexed by h , HMO’s by m , the hospital network of HMO m by H_m (this is just a vector of indicator functions which tell us whether there are contracts between the HMO and each of the hospitals), and, analogously,

⁵⁰For a discussion of some of the implications of these assumptions, and a partial comparison to alternatives, see Pakes,2006.

the HMO network of hospital h by M_h . We let $R_m(H_m, H_{-m}, z)$ be the revenues of the HMO (premiums times the number of consumers who choose to join the HMO), and $T_{m,h}$ be the transfers it sends to hospital h , so the HMO's profits are

$$\pi_m^M(H_m, H_{-m}, z) = R_m(H_m, H_{-m}, z) - \sum_{h \in H_m} T_{m,h}. \quad (23)$$

Analogously if c_h is the per patient costs of hospital h and $q_{m,h}(M_h, M_{-h}, z)$ is the number of patients HMO m sends to hospital h , the hospital's profits are

$$\pi_h^H(M_h, M_{-h}, z) = \sum_{m \in M_h} T_{m,h} - c_h \sum_{m \in M_h} q_{m,h}(M_h, M_{-h}, z). \quad (24)$$

We do not observe the transfers and so will have to model them. The actual rules determining the transfers are set in contracts which are largely proprietary (and all indications are that even if these contracts were accessible they would be too complicated to summarize in a small number of variables). This is not unusual for markets of this sort and as a result we focus our discussion of structural errors on the errors appearing in our model for the transfers. That is if $T_{m,h}(H_m, H_{-m}, z; \theta)$ is the econometrician's parametric approximation to the transfers between HMO m and hospital h , then we let

$$\mathcal{E}(T_{m,h} | \mathcal{J}_m) = \mathcal{E}(T_{m,h}(H_m, H_{-m}, z; \theta) | \mathcal{J}_m) + \nu_{2,m,h}, \quad (25)$$

so our parametric model only captures the transfers implicit in the contracting process up to $\nu_{2,m,h}$. Moreover we assume that this structural error is known to the hospital before it makes its decision, so that

$$\mathcal{E}(T_{m,h} | \mathcal{J}_h) = \mathcal{E}(T_{m,h}(H_m, H_{-m}, z; \theta) | \mathcal{J}_h) + \nu_{2,m,h}. \quad (26)$$

As suggested by the notation we are treating the structural error as a fixed fee specified by the contract. Finally note that though the parametric approximation to transfers is identical for hospitals and HMOs (it is set by the contract), the information sets of the two agents may differ leading to differences in expected transfers.

Let $R_m(H_m, H_{-m}, z; \theta)$ denote the econometrician's parametric approximation to the revenues of HMO m . Set

$$r_m^M(H_m, H_{-m}, z; \theta) \equiv R_m(H_m, H_{-m}, z; \theta) - \sum_{h \in H_m} T_{m,h}(H_m, H_{-m}, z; \theta).$$

Then, profits of the HMO are

$$\pi_m^M(H_m, H_{-m}, z) = r_m^M(H_m, H_{-m}, z, \theta) - \sum_{h \in H_m} \nu_{2,m,h} + \nu_{1,m,H_m} \quad (27)$$

Similarly the hospitals profits are

$$\pi_h^H(M_h, M_{-h}, z) = r_h^H(M_h, M_{-h}, z; \theta) + \sum_{m \in M_h} \nu_{2,m,h} + \nu_{1,h,M_h} \quad (28)$$

where

$$r_h^H(M_h, M_{-h}, z, \theta) \equiv \sum_{m \in M_h} \left[T_{m,h}(H_m, H_{-m}, z; \theta) - c_h q_{m,h}(M_h, M_{-h}, \theta) \right].$$

Note again that because $\nu_{2,m,h}$ is a disturbance in the transfers between HMO's and hospitals, under the assumptions discussed above, each $\nu_{2,m,h}$ that appears in (28) for a given hospital appears with an opposite sign in (27) for the respective HMO.⁵¹

The purpose of the empirical exercise is to determine what the contracts must have been related to for the equilibrium we observe in the data to have satisfied Assumption 1. With this in mind we assume that the $T_{m,h}(\cdot, \theta)$ function specifies a per-patient cost and investigate how that cost varies across hospital/HMO pairs.⁵²

Moment Inequalities When $\nu_2 \equiv 0$. Recall that the model has hospitals making simultaneous take it or leave it offers to HMOs. Then Assumption 1 implies that HMO's accept or reject contract offers according as the offers increase or decrease their expected profits. With $\nu_2 = 0$, we can use the difference between our estimate of the HMO's profits from the observed HMO

⁵¹Again we note that this assumes that the only structural error is in the transfers, and that the structural error for a given hospital-HMO contract does not vary with the network that is established. We could weaken these assumptions provided we insured that the resulting disturbance had a negative expectation conditional on the instruments, and therefore satisfied our Assumption 3.

⁵²Thus one could think of the contracts themselves as two-part tariffs with $\nu_{2,m,h}$ as the fixed fee (or reimbursement as the case may be). We note that there are two previous papers which analyze the marginal value of the hospital to hospital profits, but they do not condition on the networks, and do not attempt to analyze the determinants of the payments from the hospitals to the HMO; see Capps, Dranove and Satterthwaite (2003), and Town and Vistnes (2001).

network and our estimate of what those profits would have been from the alternative network obtained by reversing the plan's contract choice with each of the hospitals in the market as the $\Delta\pi^M(\cdot)$, which should have positive expectation. That is, if the plan did contract with the hospital we compare to a situation in which it did not contract with the hospital, and vice versa.⁵³

The game is sequential, so without further assumptions the implications of Assumption 2 on the inequalities we can derive for the hospitals require the second generalization in section 2.4. That generalization notes that the change in hospital profits between a contract that was accepted and the minimum profits the hospital could earn were the HMO to not accept the hospital's contract should be positive in expectation (the minimum being over the HMO's possible decisions with other hospitals). On the other hand were we to assume the existence of a contract that would induce an HMO which accepted a hospital's offer to reverse its decision with that hospital without changing its decisions with other hospitals, then Assumption 1 would imply a stronger inequality; that the difference in hospital's profits from a network that includes an HMO which accepted its offer and one that does not is expected to be positive. We start with this assumption, and then see whether relaxing it affects the results.

Also, though we allowed the plan membership and the patient flows to adjust for the alternative hospital networks formed by changing the contract status of one hospital (HMO), we did not allow the premiums to adjust (since this requires computing a new equilibrium for premiums and the premium adjustment for a change in one component of the network is likely to be small). Under these assumptions it took about ten minutes for a run of two hundred simulation draws on a pentium three processor with a 1.33 GHZ harddrive and 512 MB of RAM, so it was easy to try numerous specifications. We then examined the robustness of the results to this simplification.

Alternative Estimators: A Logit Model We compare the results to those from a logit model. The logit model assumes that the plan chooses the network that maximizes its profits knowing the value of all the unobserved disturbances in the profit equation. Profits for the different networks are

⁵³More precisely we used the change in profits from reversing the decision of each HMO with each of the six largest hospitals separately, and then formed one more HMO inequality by summing over the $\Delta\pi^M(\cdot)$ of the remaining hospitals. So $\Delta\pi^M(\cdot)$ had seven elements. The six largest hospitals by capacity cover an average of 57% of the admissions to hospitals.

calculated as in equation (27) and disturbances are assumed to distribute i.i.d. extreme value. Estimation is by maximum likelihood.

As is well known from the entry literature in Industrial Organization, the assumptions on the disturbance term in the logit model are problematic. First they imply that there is no expectational or measurement error in the profit measure (i.e. $\nu_1 \equiv 0$). This only leaves the structural disturbance, but that disturbance can not be both independent of the observed determinants of profits and a determinant of the firm's decision (since those decisions determine profits). Accordingly we expect maximum likelihood to lead to inconsistent estimates.⁵⁴

4.2.2 The Data and Empirical Results with $\nu_2 = 0$.

The primary data set contains every HMO in 42 major US markets, and considers the network of hospitals these HMOs offered to enrollees in March/April 2003. It has 441 plans and 633 hospitals, and contains a number of plan, market, and hospital characteristics put together from different data sources (for more detail, see Ho 2004b). As in the ATM example, market size varies quite a bit, and we found that we obtain somewhat shorter confidence intervals when the market averages are weighted by the square root of the number of plans in the market before averaging across markets to form the moments used in estimation.

This is not a large data set and we had no guidance from prior results, so we experimented with parsimonious specifications. It soon became evident that we needed to allow per patient costs to depend on a constant term, a measure of the extent particular hospitals are likely to be capacity-constrained (obtained by calculating the number of patients treated at each hospital under the thought experiment that every plan contracts with every hospital in the market), and a measure of hospital costs per admission. Hospital costs and patient flows were not used as instruments as we worried about measurement error in the cost variable and prediction error in

⁵⁴Note also that the errors for the different possible networks consist of the sum of the errors for the hospitals in those networks. Since the same hospitals are in different networks it is unlikely that the composite error in the network choice equation is either independent across choices for a given HMO, or independent across different HMO's. As a result we should expect the disturbance in the profit equation to be correlated with the choices of the firm's competitors as well as with its own choice, and the choices of the firm's competitors are also determinants of the firm's profits.

our model for the patient flow variable (though including the patients flow variable in the instrument set had little effect on the results). The additional instruments used included other market and plan characteristics known to the agents at the time of the contracting decision.⁵⁵

Estimates (with $\nu_2 = 0$). Table 3 provides the base results and Table 4 presents a selection of (various) robustness checks. We subtracted costs per patient from the revenues in all specifications, so the coefficients appearing on the table are the coefficients of the markup implicit in the per patient payment.

The estimate of Θ_0 from every specification was a singleton, so there was no parameter vector that satisfied all the inequality constraints. All specifications have eighty-eight or more inequality constraints so this should not be surprising. None of the test statistics for these specifications were close to their critical values, however the variance in the moment conditions in this example was quite large, so the test results may not be too powerful. Eleven of the inequalities were negative at the estimated parameter value, though only one was significant at the 5% level.

The point estimates in Table 3 all have the expected sign, and are significantly different from zero. Interestingly the point estimates imply an equilibrium configuration where almost all the cost savings from low cost hospitals are captured by the HMO's who do business with those hospitals, and that markups increase sharply when a hospital is capacity constrained. Thus, though the estimates do have reasonably large confidence intervals, especially when we use the conservative confidence intervals, if these results were interpreted as causal they would imply significantly lower incentives for hospitals to invest in either cost savings or in capacity expansion than would occur in a price taking equilibrium.

Using the estimates, average population figures for probability of admis-

⁵⁵The hospitals that are "capacity constrained" are hospitals for which the predicted number of patients exceeds the number of beds \times 365 / average length of stay in the hospital. The additional instruments included indicator variables for a high proportion of population aged 55-64, a high number of beds per population, a high proportion of hospitals in the market being integrated into systems of hospitals (see below), whether the plan is local, whether the plan has good breast cancer screening services, whether the plan has poor mental health services, and interactions between some of these characteristics. Here low proportion means less than the mean percentile, except for beds per population and breast cancer screening rates where quartiles of the distribution were used.

sion to a hospital, and our data on costs per admission and premiums, we worked out profits per patient for the hospitals and net premiums (over hospital payments) per enrollee for the HMO's. The cost figures we use are for "total facility expenses" and hence should include a (sometimes imputed) cost for buildings and equipment.

The estimates imply that capacity constrained hospitals earn profits per HMO admission of 20 to 25% of their costs for those admissions.⁵⁶ The non-capacity constrained hospitals do much worse. The point estimates imply they generate a loss per admission of about 10% of their costs per admission. HMO's, on the other hand, have 28%-47% of their premiums left over after hospital costs to devote to non-hospital related enrollee and administrative costs, and to profits. On the whole these implications seem plausible, though one might think that the hospital profit figures are a little low (a point we return to below).

The logit estimates, on the other hand, make little sense. The negative constant term implies that hospitals which are not capacity constrained incur per patient losses which are *larger* than their per patient costs (though only slightly larger), while even capacity constrained hospitals incur per patient losses of about half of their costs. That is, the logit estimates indicate a system which could not survive without a constant massive infusion of funds to cover hospital losses.

Table 4 presents results from a selection of robustness checks. The first two specifications remove the simplifying assumptions used to decrease the computational burden in obtaining our estimates. First we remove the assumption that the hospital could offer an alternative contract to an HMO which accepted its contract which would have induced that HMO to reject the hospital's offer but not change the HMO's responses to the offers of other hospitals. In particular we allow the HMO to either not change its other responses or add any hospital which it had previously rejected, and then take the minimum of the resulting change in hospital profits for the counterfactual profits in the hospital inequality (see the second generalization in section 2.4). The point estimates do not change a lot, but there is an increase in the length of confidence intervals. Next we allow premiums to adjust when evaluating the counterfactual. Then both the constant and the cost coefficient increased in absolute value. The third set of robustness

⁵⁶Perhaps not surprisingly, capacity constrained hospitals also tend to be the lower cost hospitals in their markets.

results accounts for the non-hospital related costs associated with plan enrollees. These costs clearly exist but the number of enrollees do not change very much when we change networks by one hospital (if they did not change at all they would act as the fixed effect in generalization three and could be omitted). Consequently the enrollee variable gets a positive, but imprecisely estimated coefficient, while the rest of the coefficients are similar to those in the base specification.

The results in Table 4 are representative of the differences we saw across other specifications not reported here. It is clear that capacity-constrained hospitals get higher markups and that hospitals with lower costs have to share a significant portion of their cost savings with HMO's. However the relative contributions of the capacity constraints and costs to profits do vary across specifications, and as we shall see, sometimes the profit implications themselves do also.

4.2.3 Allowing For Structural Errors.

We begin by allowing for a ν_2 which can differ freely across contracts, and then show how the special case in which ν_2 takes the form of HMO effects delivers additional structure which can be used in estimation.

Some additional notation will prove useful. Let $H_m \setminus h$ be the set of all hospitals in the network of HMO m but hospital h , $H_m \cup h$ be the network obtained when we add hospital h to H_m , and $M_h \setminus m$ be the the set of HMO's obtained when we take HMO m out of the network of hospital h . Now define

$$\begin{aligned}\Delta\pi_m^M(+h) &= \pi_m^M(H_m, H_{-m}, z) - \pi_m^M(H_m \setminus h, H_{-m}, z) \quad \forall h \in H_m, \\ \Delta\pi_m^M(-h) &= \pi_m^M(H_m, H_{-m}, z) - \pi_m^M(H_m \cup h, H_{-m}, z) \quad \forall h \notin H_m, \\ \Delta\pi_h^H(+m) &= \pi_h^H(M_h, M_{-h}, z) - \pi_h^H(M_h \setminus m, M_{-h}, z) \quad \forall m \in M_h.\end{aligned}$$

So, $\Delta\pi_m^M(+h)$ is the HMO's incremental profit from accepting the contract with a hospital (h) that it did contract with, $\Delta\pi_m^M(-h)$ is the HMO's incremental profit from rejecting the contract of a hospital that it did not contract with, and $\Delta\pi_h^H(+m)$ is the increment in hospital's h 's profit from the contract it offered to an HMO which it contracted with. Define

$$\Delta r_m^M(-h, \theta), \quad \Delta r_m^M(+h, \theta), \quad \text{and} \quad \Delta r_h^H(+m, \theta)$$

Table 3: **Determinants of Hospital/HMO Contracts (ν_1 only).**

Characteristics of Hospitals	θ	Simulated 95% CI		Cons. 95% CI		Logits θ SE	
Per Patient Markups (Units = \$/thousand, per patient).							
const	9.45	3.9	16.2	.5	23.4	-6.95	2.94
capcon	3.52	1.0	8.5	1.2	11.6	6.69	1.48
c_h	-.95	-1.6	-.5	-2.6	-.5	-.44	.19

Table 4: **Robustness Analysis (ν_1 only).**

Model	Min.Hos. π			Prem. Adj.			Enrollees		
Variable	θ	Sim CI		θ	Sim CI		θ	Sim CI	
Enrole	-	-	-	-	-	-	.01	-.035	.03
(cons.)	-	-	-	-	-	-	-	(-.1	.1)
Per Patient Markups (Units=\$/thousand, per patient).									
const	9.3	2.7	17.7	12.7	6.7	18.4	9.0	4.7	13.9
(cons.)	-	(-5.2	23.1)	-	(6.3	27.1)	-	(-.5	21.3)
capcon	4.0	1.5	10.3	1.7	-2.4	4.7	2.8	.2	6.8
(cons.)	-	(2.1	16.5)	-	(-1.7	7.6)	-	(.1	13.4)
c_h	-.96	-1.9	-.4	- 1.25	-1.8	-.7	-.90	-1.4	-.6
(cons.)	-	(-3.0	-.4)	-	(-2.8	-1.0)	-	(-2.5	-.5)

Notes: Enrole refers to the total number of enrollees of the plan, “capcon” to the capacity constraint, and c_h to the cost per hospital admission.

analogously. Finally let $\chi(\mathbf{m}, h)$ be the indicator function which takes the value of one if HMO \mathbf{m} and hospital h contract, and zero elsewhere.

Assumption 1 implies that hospitals expect to increase their profit from signed contracts and that HMOs only reject offers when their expected profits are higher without them. Thus the expectation of

$$U^\pi(\mathbf{m}, h) \equiv \chi(\mathbf{m}, h)\Delta\pi_h^H(+\mathbf{m}) + (1 - \chi(\mathbf{m}, h))\Delta\pi_{\mathbf{m}}^M(-h),$$

conditional on any $x \in \mathcal{J}_{\mathbf{m}} \cap \mathcal{J}_h$ is nonnegative.

Note that when a contract is accepted the hospital gains ν_2 whereas when the contract is rejected the HMO saves ν_2 . Thus if we define $U^r(\mathbf{m}, h; \theta)$ as the parametric analog to $U^\pi(\mathbf{m}, h)$, and substitute the profit functions for the HMOs and the hospitals (equations (27,28)) into the equation above

$$\begin{aligned} U^\pi(\mathbf{m}, h; \theta) &= \chi(\mathbf{m}, h)[\Delta r_h^H(+\mathbf{m}, \theta) + \nu_{1,h,M_h,M_h \setminus \mathbf{m}} + \nu_{2,\mathbf{m},h}] \\ &\quad + (1 - \chi(\mathbf{m}, h))[\Delta r_{\mathbf{m}}^M(-h, \theta) + \nu_{1,\mathbf{m},H_{\mathbf{m}},H_{\mathbf{m}} \setminus h} + \nu_{2,\mathbf{m},h}] \\ &= U^r(\mathbf{m}, h, ; \theta) + \chi(\mathbf{m}, h)\nu_{1,h,M_h,M_h \setminus \mathbf{m}} + \nu_{2,\mathbf{m},h} \\ &\quad + (1 - \chi(\mathbf{m}, h))\nu_{1,\mathbf{m},H_{\mathbf{m}},H_{\mathbf{m}} \setminus h}. \end{aligned}$$

This equation contains the same linear function of ν_2 regardless of whether a contract was established or not, so the ν_2 appearing in it *are not selected* on the outcome being modeled. As a result if we take an $x \in \mathcal{J}_{\mathbf{m}} \cap \mathcal{J}_h$ that is also an instrument (i.e. $\mathcal{E}[\nu_{2,\mathbf{m},h}|x] = 0$), then since $\mathcal{E}[U^\pi(\cdot)|x] \geq 0$

$$\mathcal{E}[U^r(\mathbf{m}, h, ; \theta)|x] \geq 0, \quad (29)$$

and $U^r(\cdot)$ can be one component of the $\Delta \mathbf{r}(\cdot)$ used to construct the moment inequalities in our objective function.

A second component can be obtained as the sum of HMO and hospital profits if they contract and zero otherwise. Assumption 1 ensures that each component of the sum has positive expectation and since the sum does not depend on the transfers between these two agents it does not depend on $\nu_{2,\mathbf{m},h}$ (though it does depend on transfer between both of them and other agents, and hence on θ). Let

$$S^r(\mathbf{m}, h; \theta) = \chi(\mathbf{m}, h)[\Delta r_h^H(+\mathbf{m}, \theta) + \Delta r_{\mathbf{m}}^M(+h, \theta)], \quad (30)$$

then $\mathcal{E}[S^r(\mathbf{m}, h; \theta)|x] \geq 0$ provided $x \in \mathcal{J}_{\mathbf{m}} \cap \mathcal{J}_h$.

“Effects” Models. We now assume $\nu_{2,m,h} = \nu_{2,m}, \forall h$, that is that there are HMO effects.⁵⁷ This assumption generates two inequalities with positive expectation.

If an HMO accepts at least one hospital’s contract and rejects the contract of another, then the sum of the increment in profits from accepting the contract accepted and rejecting the contract rejected both differences out the HMO effect and has a positive expectation. More formally for every $\tilde{h} \notin H_m$ and $h \in H_m$ we have

$$\begin{aligned} & \Delta\pi_m^M(H_m, H_m \cup \tilde{h}, \cdot) + \Delta\pi_m^M(H_m, H_m \setminus h, \cdot) = \\ & \Delta r_m^M(H_m, H_m \cup \tilde{h}, \cdot) + \Delta r_m^M(H_m, H_m \setminus h, \cdot) + \nu_{1,m,H_m,H_m \cup \tilde{h}} + \nu_{1,m,H_m,H_m \setminus h}. \end{aligned}$$

Since $\mathcal{E}[\nu_1 | \mathcal{J}_m] = 0$ by construction, Assumption 1 ensures

$$\mathcal{E} \left[\Delta r_m^M(H_m, H_m \cup \tilde{h}, \cdot) + \Delta r_m^M(H_m, H_m \setminus h, \cdot) | \mathcal{J}_m \right] \geq 0. \quad (31)$$

If $\#$ denotes the cardinality of a set, equation (31) gives us $\sum_m \#H_m(\#H - \#H_m)$ *difference in difference inequalities*.

Also note that if $\nu_{2,m,h} = \nu_{2,m}$ then

$$U^\pi(\mathbf{m}, h) = U^r(\mathbf{m}, h; \theta) + [\chi(\mathbf{m}, h)\nu_{1,h,M_h,M_h \setminus \mathbf{m}} + (1 - \chi(\mathbf{m}, h))\nu_{1,m,H_m,H_m \setminus h}] + \nu_{2,m}.$$

So if we define

$$\bar{U}^r(\mathbf{m}) = \frac{1}{\#H} \sum_h U^r(\mathbf{m}, h),$$

then

$$\mathcal{E}[\bar{U}^r(\mathbf{m}) | \mathcal{J}] \geq -\nu_{2,m},$$

where \mathcal{J} represents the intersection of the information sets of the hospitals in H and HMO m . Consequently for $h \in H_m$

$$\begin{aligned} 0 & \leq \mathcal{E} [\Delta\pi_m^M(H_m, H_m \setminus h, \cdot) | \mathcal{J}] = \mathcal{E} [\Delta r_m^M(H_m, H_m \setminus h, \cdot) | \mathcal{J}] - \nu_{2,m} \quad (32) \\ & \leq \mathcal{E} [\Delta r_m^M(H_m, H_m \setminus h, \cdot) | \mathcal{J}] + \mathcal{E}[\bar{U}^r(\mathbf{m}) | \mathcal{J}], \end{aligned}$$

⁵⁷A more complete analysis of effects models in buyer seller networks would allow for both buyer and seller effects. This is a straightforward, though somewhat tedious, extension of the results below. We examine the HMO effects case in detail because all the contract correlates we use in our analysis are hospital specific, and we wanted to make sure that the absence of HMO characteristics did not bias the analysis of the impacts of these hospital specific variables.

providing us with another $\sum_h \#H_m$ inequalities that difference out the structural disturbances.

Empirical Results (allowing for ν_2 disturbances). Column I of Table 5 provides the estimates when we allow both ν_1 and ν_2 to be free and use the inequalities in equations (29) and (30), column II provides the estimates when we assume $\nu_{2,m,h} = \nu_{2,m}$ and use the inequalities from (30), (31), and (32), while column III uses all four sets of inequalities.

Though none of the test statistics are significant, there is reason to question the appropriateness of the effects models. Only 6 of the 88 inequalities in model I were negative at the estimated parameter value, and *none* were significantly so. In contrast about a third of the inequalities in models II and III were negative, and a third of these were significantly negative. We note that this explains why the “conservative” confidence intervals for the effects models sometimes have shorter lengths than the confidence intervals which are not conservative. The conservative confidence intervals can only be shorter when there are moments whose values are negative at the estimated parameter values.

The difference between the estimates with free ν_2 (I of Table 5) and the ν_1 only model (Table 3) is that the cost coefficient is lower and the capacity constraint higher when we allow for ν_2 . In particular the estimates in column I imply a higher markup to capacity constrained hospitals and that 60% of low cost hospitals’ cost savings are being transferred to HMOs (not the 94% in Table 3). As a result these estimates imply higher hospital profits; even non-capacity constrained hospitals earn profits of about 14% of the revenues (16% of costs) from their HMO patients, while the average capacity constrained hospital earns profits of over 58% of revenue. Of course these profits have to account for any deficits from other parts of the hospitals’ activities (emergency room walk-ins, medicare, and medicaid). The flip side of this comparison is that plan profits decline, and they now depend a great deal on the nature of costs and capacity in the markets they are operating in.

We have made several auxiliary calculations which support the general nature of the results in Table 5; in particular, these calculations reinforce the finding that low cost hospitals only get to keep a modest fraction of their cost savings and capacity constrained hospitals earn higher markups. Ho (2004b) reports the estimates that we obtained from a ν_1 only model that allowed,

Table 5: **Allowing for ν_1 & ν_2 Disturbances.**

Model	I			II			III		
Description	Ineq. (29) & (30). Free ν_2 (IV).			Ineq. (30),(31) & (32). $\nu_{2,m,h} = \nu_2$.			All Four Ineq. $\nu_{2,m,h} = \nu_2$ & IV.		
Variable	θ	Sim CI		θ	Sim CI		θ	Sim CI	
Per Patient Markups (Units=\$/thousand, per patient).									
const	8.2	2.9	13.7	11.1	5.2	15.3	11.2	5.4	15.2
(cons.)	-	(-11.9	47.4)	-	(4.5	15.5)	-	(8.0	16.3)
capcon	13.5	1.9	16.8	6.6	2.1	9.4	6.4	1.8	9.2
(cons.)	-	(3.9	73)	-	(4.6	12.9)	-	(4.2	10.7)
c_h	-.58	-1.0	-.2	-.95	-1.8	-.4	-.94	-1.3	-.4
(cons.)	-	(- 3.1	1.5)	-	(-1.4	-.6)	-	(-1.4	-.6)

See the notes to Table 4.

in addition to the variables considered above, for a lump sum transfer to hospitals which negotiate as a system, and an increment to this lump sum when the HMO accepts a contract from some members of a system but rejects it from others. She finds significant effects for the system variables, and results for the rest of the variables that are similar to those in model I in table 5, though with a bit smaller coefficient on capacity (implying somewhat smaller profits for capacity constrained hospitals and larger profits for plans). One summary then is that if we allow for the system variables we do not need to allow for the ν_2 , while if we do not allow for the system variables we need the ν_2 .⁵⁸ Also Pakes (2006) presents results from a numerical analysis of equilibria in markets with characteristics which are similar to those in these markets and obtains results which reinforce those above, though the magnitudes of the capacity constrained coefficients does depend on what other variables are allowed in the markup equation.

⁵⁸We tried to estimate the model with free ν_2 and ν_1 allowing for the system variables. We got similar point estimates but standard errors that were extremely large.

5 Conclusion

This paper provides conditions which ensure that the inequality constraints generated by either single agent optimizing behavior, or by the Nash equilibria of interacting agents, can be used in estimation. We assume an ability to construct an approximation to profits from a counterfactual; a structural model that enables us to obtain an approximation to the returns that would have been earned under an alternative feasible action. We do not, however, place any restrictions on the choice sets of the agents, or on what the agents know about either their competitors' play or about the exogenous conditions that will rule when the profits from their actions materialize.

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 measurement error in the profit proxy generate our ν_1 . The distribution of ν_1 can be quite complex as it depends on the information sets of different agents and 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 either specifying these details, or computing an equilibria.

The only 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 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 contribute to the growing literature on estimation subject to inequality constraints by providing a new test of the null that there is a value for the parameter vector at which all the inequalities are satisfied, and by providing a limit theorem for points on the boundary of the identified set. However, this limit distribution depends on the number of binding moments at the boundary point; a parameter which is generally not known. So we provide two approximating distributions, which allow us to form an outer and an inner approximation to the confidence interval generated by the true asymptotic distribution. Moreover, in a leading case, the inner confidence

interval should be generated by a simulated distribution that converges to the true limiting distribution for the boundary estimator. These tools enable conservative inference on the actual parameter values and the interval defined by the boundaries for the parameters, as well as for the boundary points per se.

A number of important unanswered questions remain. We would like to know necessary, as well as sufficient, conditions for combining inequality constraints with counterfactuals as is done here, as this would enable us to clarify when we need more detailed assumptions. We have not investigated what can be learned if we replace the parametric structural model of profits with a non-parametric one. Moreover we have not investigated the efficiency properties of the estimators we propose, and the use of method of moments inequalities, instead of equalities, is likely to accentuate precision problems.

Despite these issues our empirical examples show that the framework proposed here can enable 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. Moreover the estimators themselves were fairly easy to construct, and were obtained from data sets that were not large by modern standards.

References.

- American Hospital Association Annual Survey Database: Fiscal Year 2001.
- Andrews, D., Berry, S., 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.
- Bajari, P., Benkard, L., and J. Levin (2004), "Estimating Dynamic Models of Imperfect Competition", manuscript, Stanford University.
- Bajari, P., Hong, H., and P. Ryan (2004), "Identification and Estimation of Discrete Games of Complete Information", manuscript, Duke University.
- Berry, S. (1992); "Estimation of a Model of Entry in the Airline Industry", *Econometrica*, vol. 60, no. 4, pp. 889-917.
- Steve Berry, Jim Levinsohn, and Ariel Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, vol. 63, no. 4, pp. 841-890.
- Bickel, P. and D. Freedman (1981), "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9, 1196-1217.

- Bresnahan, Timothy and Peter Reiss (1991): "Entry and Competition in Concentrated Markets", *Journal of Political Economy*, vol. 99, no. 5. pp. 977-1009.
- Chernozhukov, V., Hong, H., and E. Tamer (2003), "Parameter Set Inference in a Class of Econometric Models," manuscript.
- Cliberto, and E. Tamer (2004), "Market Structure and Multiple Equilibria in the Airline Markets," manuscript.
- Dove Consulting, 2002 ATM Deployer Study, Executive Summary, February 2002.
- Fershtman, C. and A. Pakes, 2004, "Dynamic Games with Asymmetric Information; A Computational Framework", *mimeo*, Harvard University.
- Guggenberger, P., J. Hahn, and K. Kim, 2006, "Specification Testing under Moment Inequalities," *mimeo*, UCLA.
- Haile, P. and E. Tamer, 2003, "Inference with an Incomplete Model of English Auctions", *The Journal of Political Economy*, 111, 1-51.
- Hansen, Lars, 1982, "Large Sample Properties of Method of Moments Estimators", *Econometrica*, 50, 1029-1054.
- Hansen, Lars Peter, and Kenneth J. Singleton, 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, vol. 50, no. 5, pp. 1269-86.
- Ho, K., (forthcoming), "The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market", *The Journal of Applied Econometrics*.
- Ho, K., (2004b), "Insurer-Provider Networks in the Medical Care Market", *mimeo*, Columbia University.
- Horowitz, J. and C. Manski (1998), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281-302.
- Imbens, G. and C. Manski (2003), "Confidence Intervals for Partially Identified Parameters," manuscript.
- Ishii, Joy, 2004, "Compatibility, Competition, and Investment in Network Industries: ATM Networks in the Banking Industry", *mimeo* Stanford University.
- Kaiser Family Foundation report, "Trends and Indicators in the Changing Health Care Marketplace", 2004.

Manski, C. (2003), *Partial Identification of Probability Distributions*, Springer: New York.

Pakes, A. and D. Pollard, 1989; "Simulation and the Asymptotics of Optimization Estimators" *Econometrica* vol 57, pp. 1027-57.

Pakes, A., M. Ostrovsky, and S. Berry, 2003; "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry-Exit Examples)" *National Bureau of Economic Research* WP 10506.

Pakes, A., 2005; "Theory and Econometrics in Empirical I.O.", The Fischer-Schultz lecture, World Congress of the Econometric Society, *mimeo* Harvard University.

Rosen, A., 2005; "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *mimeo*, Northwestern University.

Seim, K.(2002);"Geographic Differentiation and Firms' Entry Decisions: The Video Retail Industry", *mimeo*, GSB, Stanford.

Shaikh, A. (2005): "Inference for Partially Identified Econometric Models," *mimeo*, Stanford University.

Soares, G. (2006): "Inference for Partially Identified Models with Inequality Moment Constraints," *mimeo*, Yale University.

Stoye, J. (2005): "Partial Identification of Spread Parameters," *mimeo*, New York University.