Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location[*]

Donald W. K. Andrews
Dept of Economics and Cowles Foundation
Yale University

Steven Berry
Dept of Economics and Cowles Foundation
Yale University

Panle Jia
Dept of Economics Yale University

May 3, 2004

1 Introduction

In this paper, we consider the empirical problem of estimating the underlying profit function of firms from "revealed choice" data on entry in a cross-section of oligopoly markets. One problem faced in the earlier literature is the problem of multiple equilibrium, which makes traditional maximum likelihood estimation (MLE) techniques difficult to implement. The problem of multiple equilibria is serious enough that Sutton (2000), for example, suggests giving up on parameter estimation in the case of realistic entry models. However, Manski and co-authors (e.g. Manski and Tamer (2002)) encourage work on a related class of "incomplete" parametric models. In this paper, we combine the two ideas and consider the estimation of complicated entry models that exhibit multiple equilibria, without attempting to resolve the equilibrium selection problem.

In the context of entry models, the necessary conditions for a pure-strategy Nash equilibrium can be easily shown to place a set of inequality restrictions on the parameters of the

[*] This is still VERY PRELIMINARY AND INCOMPLETE. A much earlier version of this paper was given at the SITE conference in Palo Alto in the summer of 2002 and at Northwestern University in March 2003

underlying profit functions. These restrictions may or may not provide point identification of the parameters. We consider the problems of estimation and inference when the model's asymptotic (population) inequalities may define a region of parameters rather than a single point in the parameter space.

Our idea of using the inequality restrictions implied by necessary conditions is nearly the same as in the independently developed work-in-progress of Ciliberto and Tamer (2003), which uses the general method for inference on incomplete models described in Chernozhukov, Hong, and Tamer (2003). However, we take a different approach to both consistency and inference. While those papers focus on the distribution of a minimized criterion function, we focus directly on the distribution of the inequality constraints and how the variance in those constraints affects inference about the identified region of parameters. Our method may have computational advantages. We also develop the case where a simple analog estimator is consistent, allowing us to avoid the problem of using a potentially arbitrary "tolerance" parameter. At the least, the methods of Chernozhukov, Hong, and Tamer (2003) and the present paper together provide alternative starting points for new set estimation strategies in oligopoly entry games (and in a large set of related models).

1.1 Idea for Estimation

Most earlier research on oligopoly entry takes a traditional ML approach to estimation. One assumes a parametric form for the profit function—typically with additive errors. The errors are often assumed to be observed by the firms but not by the econometrician and are assumed to be drawn from a parametric family of distributions.

If there is a unique equilibrium to the game, then one can solve for the set of errors that is consistent with each equilibrium and construct an ML estimator using the implied probabilities of each possible equilibrium outcome. For early examples, see Bresnahan and Reiss (1988), Berry (1989), Bresnahan and Reiss (1991), and Berry (1992). In these models, identification is usually assumed rather than proved (Tamer (2003) is an exception), but the restrictions of the parametric model are so strong that parametric identification seems very likely to hold.

Realistic entry games, however, are likely to admit multiple equilibria for at least some combinations of parameters and unobservables. This invalidates the traditional MLE approach, which requires a one-to-one match between regions of the unobservables and observed equilibrium outcomes. Many papers work hard to add additional assumptions that guarantee uniqueness of at least some observed outcomes (for example, see Berry (1992) and Mazzeo (2002)). As games become more complicated, these additional assumptions become either implausible or impossible.[1]

In many cases, however, the necessary conditions for equilibrium still place restrictions on the parameters. In any pure strategy equilibrium, each firm's action maximizes profits conditional on the actions of the other firms. There is some subset of the unobservables that

[1] Seim (2001) suggests another approach which is to look at games with private information which in some cases may reduce the problem of multiple equilibria.

is consistent with the necessary conditions for each equilibrium and so we can think of the probability that a given set of necessary conditions holds.

By definition, when there are multiple equilibria there are regions of unobservables that are consistent with the necessary conditions for more than one equilibrium. Thus, there is no one-to-one match between regions of unobservables and outcomes. However, the probability of the necessary condition for a given event is *greater than or equal to* the probability of the event (because "necessary" is broader than "necessary and sufficient"). Thus, the necessary conditions for Nash equilibrium provide a set of inequality constraints on the parameters.

As a simple estimation method, one could use the sample analog of the population necessary conditions. There is some set of parameters (perhaps the null set) that is consistent with the model necessary conditions being at least as great as the sample probabilities of equilibrium events.

As in Chernozhukov, Hong, and Tamer (2003), we consider the estimation and inference problem that is implied by this simple idea for estimation. Traditional proofs of consistency and simple notions of confidence regions must be extended to consider the non-standard nature of the estimation problem. This is the topic of the rest of this paper.

1.2 Outline of the Paper

The next section of the paper provides a brief and broad overview of a class of relevant models and of the ideas we use for estimation and inference. The following sections then give a formal treatment of the econometric issues. Two final sections describe a set of Monte Carlo results and a preliminary set of estimates for an empirical application to chain store location.

2 The Basic Model, Notation, and an Outline of the Results

In this section, we provide a description of the broad class of entry models that we consider. In fact, the techniques of the paper apply to a much broader class of "incomplete" models. We lay out a simple class of entry models and the necessary conditions for a pure strategy Nash equilibrium in that class of models. We sketch the basic ideas of identification and estimation in this section, before turning to a more formal discussion.

2.1 Necessary Conditions in an Entry Model

Consider a profit function for firm j in market i given by

$$\pi_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \varepsilon_{i,j}, \theta_0), \tag{1}$$

where $Y_{i,j}$ is firm j's strategy, $Y_{i,-j}$ is the vector of firm j's opponents' strategies, $X_{i,j}$ is a vector of profit shifters (some of which are specific to the firm and some of which may be common across firms within the market), $\theta_0 \in \Theta \subset R^p$ is a vector of parameters to be estimated, and $\varepsilon_{i,j}$ is an unobserved (to the econometrician) profit shifter. The number of firms is J and the number of markets is n. For market i, the observed strategy vector is $Y_i = (Y_{i,j}, ..., Y_{i,J})'$ $(\in R^{d_Y})$, the observed vector of profit shifters X_i $(\in R^{d_X})$ is comprised of the non-redundant elements of $X_{i,j}$ for $j = 1, ..., J$, and the unobserved vector of profit shifters is $\varepsilon_i = (\varepsilon_{i,1}, ..., \varepsilon_{i,J})'$ $(\in R^{d_\varepsilon})$.

For example, a common form for the profit function is

$$\pi_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \varepsilon_{i,j}, \theta_0) = \bar{\pi}_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \theta_0) + Y_{i,j}\varepsilon_{i,j}. \tag{2}$$

Often $\bar{\pi}_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \theta_0)$ is taken to be linear in $Y_{i,-j}$ and $X_{i,j}$.

Prior to the game, X_i and ε_i are observed by all participants in market i and we assume that Y_i is generated by a pure-strategy Nash equilibrium. The data available to the researcher is the set of observable profit shifters, X_i, and equilibrium outcomes, Y_i, for all markets. The researcher does not observe the ε_i's, but does know the parametric form of $\pi_{i,j}$ and does know that ε_i is iid across i and independent of the X_i's. Further, we assume that the distribution of ε_i is known up to a set of unknown parameters (that are included in the parameter vector θ_0).

In different models, a strategy $Y_{i,j}$ might be a continuous variable (a level of investment), an indicator function (an "entry" variable), an integer-valued variable (the number of store locations), or a vector of multiple strategies. In the case of a continuous $Y_{i,j}$, first-order conditions often allow for estimation even in the case of multiple equilibria. The discrete case, however, is harder. We focus on this case here. Henceforth, we assume that Y_i has a discrete distribution.

In any pure strategy equilibrium in market i, it must be the case that the action $Y_{i,j}$ taken by each firm j is at least as good as any other possible action Y^*, given the actions of the other firms. That is, for all j,

$$\pi_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \varepsilon_{i,j}, \theta_0) \geq \pi_{i,j}(Y^*, Y_{i,-j}, X_{i,j}, \varepsilon_{i,j}, \theta_0), \forall Y^*. \tag{3}$$

In the absence of multiple equilibria, these conditions are necessary and sufficient for Y_i to be the pure-strategy Nash equilibrium. However, if multiple equilibria are possible, this condition is only necessary—the same (ε_i, X_i) might lead to another outcome.

Note that the best-reply condition in (3) can be expressed as a restriction on the unobservables. For example, when the profit function is of the form (2), then (3) is equivalent to a simple inequality restriction on $\varepsilon_{i,j}$:

$$(Y^* - Y_{i,j})\varepsilon_{i,j} \leq \bar{\pi}_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \theta_0) - \bar{\pi}_{i,j}(Y^*, Y_{i,-j}, X_{i,j}, \theta_0), \forall Y^*, \forall j. \tag{4}$$

Let $\Omega(Y_i, X_i, \theta_0)$ be the region of the ε_i's that satisfy (3). For example, when (2) holds, $\Omega(Y_i, X_i, \theta_0)$ is given by

$$\Omega(Y_i, X_i, \theta_0) = \{\varepsilon_i : (Y^* - Y_{i,j})\varepsilon_{i,j} \leq \bar{\pi}_{i,j}(Y_{i,j}, Y_{i,-j}, X_{i,j}, \theta_0) - \bar{\pi}_{i,j}(Y^*, Y_{i,-j}, X_{i,j}, \theta_0), \forall Y^*, \forall j\}. \tag{5}$$

Given θ_0 and X_i, the probability that the necessary conditions for Y_i hold is the probability (with respect to the distribution of ε_i) of $\Omega(Y_i, X_i, \theta_0)$. Because necessary and sufficient is a subset of necessary, the probability of necessary conditions for an event is greater than or equal to the probability of the event itself.

Entry models often are estimated by MLE. The MLE method typically proceeds by identifying a one-to-one mapping between the possible discrete outcomes and regions of the unobservables. Probabilities of observed events are equal to the probabilities of the associated regions of unobservables. This method fails in the presence of multiple equilibria, because the same region of unobservables can be associated with more than one outcome. As noted above, solutions have included simplifying the model to eliminate the multiple equilibria and introducing extra parameters (or even extra nonparametric functions) that choose between possible equilibria. Neither of these approaches works well in models with realistic levels of complexity. Consider, for example, the many equilibria that are possible when the model allows for multi-product firms. It is difficult to even enumerate the possible regions of multiple equilibria in such models, much less to estimate an equilibrium-choice function for every region.

It is relatively easy, however, to calculate the probability given θ that the necessary conditions in (5) hold. Let $\mathcal{Y}$ and $\mathcal{X}$ denote the supports of Y_i and X_i, respectively. By assumption, $\mathcal{Y}$ is a finite set. For any $(y, x) \in \mathcal{Y} \times \mathcal{X}$, this probability is defined to be

$$P(y \mid x, \theta) = Pr(\varepsilon_i \in \Omega(y, x, \theta)). \tag{6}$$

When (2) holds, this is a simple ε_i-orthant probability.

At the true $\theta = \theta_0$, the probabilities of the necessary conditions must be at least as large as the true probabilities of the events $y \in \mathcal{Y}$, denoted $P_0(y \mid x)$:

$$P(y \mid x, \theta_0) \geq P_0(y \mid x), \ \forall (y, x) \in \mathcal{Y} \times \mathcal{X}. \tag{7}$$

Again, the inequality follows from the fact that the outcome y implies the necessary conditions for y but the necessary condition need not imply the outcome y.

2.2 Alternative Restrictions

The prior subsection shows that it is easy to construct the probabilities of the necessary conditions for a pure-strategy Nash Equilibrium. However, one need not restrict oneself to those restrictions alone. There are both stronger and weaker conditions that can be imposed. On the side of stronger restrictions, one could consider adding sufficient and/or necessary-and-sufficient conditions for at least a subset of events. On the weaker side, for example, it may be necessary to employ only an individual firm's rationality constraints rather than the full joint set of constraints.

Sometimes one can show that a particular event cannot be associated with multiple equilibria. For example, in many entry models the $(\varepsilon_i, X_i, \theta)$ combinations that are consistent with no firm entering are never consistent with any other equilibrium. The necessary conditions for "no entry" are therefore both necessary and sufficient. This creates an "equality" constraint: the probability of the necessary condition should exactly equal the probability of the event.

One can also utilize sufficient-but-not-necessary conditions, as done in Ciliberto and Tamer (2003). The probability that a sufficient condition holds is *less than or equal to* the probability of the event. Such inequalities are incorporated into our analysis by multiplying them by -1 to obtain $\geq$ constraints.

In other cases, it may be difficult or undesirable to use the full set of necessary conditions. As one extreme, one could consider only a single firm's individual rationality constraints (that the firm is optimizing given rivals actions). This might be desirable if the profit functions of the rivals are hard to model or if some data about rivals' profit-shifters are missing.

The last example (of a single-firm's profitability constraints) can be extended to allow for the presence of "endogenous" variables in the profit function, where the model for the endogenous variables is unspecified. This is similar to the case of instrumental variables estimation in continuous models. There might be some variables in X_i that are "instruments" for endogenous variables in the profit function. For example, unlike in Berry (1992), one could treat the structure of an airline network as endogenous in a model of city-pair airline entry, without providing a full model of overall network formation.

The choice of what restrictions to use is governed by what restrictions one actually believes, but also by practical computational issues. The necessary conditions of subsection (2.1) are relatively easy-to-compute orthant probabilities when (2) holds. A wide variety of simulation methods are available to compute such orthant probabilities in cases where analytic probabilities are hard to obtain. Sufficient conditions, however, often involve harder to characterize regions of the unobservables. Ciliberto and Tamer (2003) get around this problem by using a simple frequency simulator of the probabilities of necessary conditions. For each simulated ε_i draw (given X_i and θ), one can see which necessary conditions hold. If only one necessary condition (i.e., for a single event) holds, then that draw is in the region of a sufficient condition. Thus, with the frequency simulator, the sufficient-but-not-necessary condition probabilities can be calculated at little extra cost. This solution for sufficient conditions, however, does not carry over to lower variance "smooth" simulation methods.

Because of the generality and computational ease of the Nash equilibrium necessary conditions, we emphasize $\geq$ inequality constraints in much of our write-up. However, with simple modifications, the methods apply to the other forms of restrictions described above. One can start with $\leq$ inequality constraints generated by sufficient conditions and multiply them -1 to obtain $\geq$ inequality constraints. Equality conditions can be expressed as pairs of inequality constraints—with one $\leq$ and one $\geq$ in each pair. Then, the $\leq$ constraints can be transformed into $\geq$ constraints via multiplication by -1.

2.3 Introduction to Set Identification

The inequalities in (7) are satisfied for the true θ and possibly for other values of the parameters. If only one θ satisfies the inequalities, then the model is point identified. If the necessary conditions are derived from an incorrect model, then perhaps no θ satisfies the inequalities.

In the absence of a proof, it is often difficult to rule out the case of under-identification, where more than one θ satisfies the inequalities. Note that point identification is certainly possible given multiple equilibria, but moving from equality to inequality constraints increases our concern with a lack of point identification. Conceptually, though, the use of necessary conditions and the concern with a lack of point identification are two different matters. Obviously, a model based only on equality constraints could suffer from a lack of point identification and the techniques of this paper (and related work) are of use in that context as well.

We define the asymptotically identified set of parameters, Θ_0, to be the set of parameters such that the inequality restrictions in (7) hold. Again, Θ_0 could be (i) the null set, (ii) a single point, (iii) a strict subset of the parameter space consisting of more than one point, or (iv) the entire parameter space. Correspondingly, we would say the model is (i) rejected, (ii) point identified, (iii) set identified, or (iv) completely uninformative.

2.4 Introduction to Estimation

One simple idea for estimation is to construct the sample analog of the inequality constraints in (7). To do so, we replace the true probabilities of outcomes, $P_0(y \mid x)$, with sample averages.

If both Y_i and X_i are discrete, this is particularly straightforward. For example, the sample analog of $P_0(y \mid x)$ is just the multinomial sample probability

$$\widehat{P}_n(y \mid x) = \sum_{i=1}^{n} [Y_i = y][X_i = x], \tag{8}$$

where $[\cdot]$ is the indicator function.

In this simple case, the sample inequalities just replace the true probabilities with the multinomial probabilities:

$$P(y \mid x, \theta) \geq \widehat{P}_n(y \mid x), \forall (y, x) \in \mathcal{Y} \times \mathcal{X}. \tag{9}$$

In many cases X_i has a continuous distribution. In these cases, one can replace the simple multinomial probabilities of the last paragraph with averages over cells in the $\mathcal{X}$ space. If there are many events y, then it also may be desirable to reduce sampling variance by collecting the y's into cells that are aggregates of the individual events. In constructing x and y cells, there is a trade-off between the benefits of reducing sample variance via the use of larger cells, on one hand, and the benefits of reducing the size of the asymptotically identified set via the use of more restrictive cells. We discuss the construction of cells for estimation in the detailed discussion below.

In some cases, it is easy to calculate the probabilities of model events and in other cases one needs to simulate such probabilities. The details of the simulation also are discussed below.

Once one decides on the details (cells, simulation, and so forth) of the sample analog inequalities, one can define a simple estimator as the set of the parameters that satisfy the sample inequality constraints.

One problem is that no θ may satisfy the constraints. This might be taken as a rejection of the model, but it is possible that the lack of a θ that satisfies the sample constraints is due solely to the variance in the sample inequalities. If there is no θ satisfying the inequality constraints, then we choose a single point that minimizes the sum of absolute values of the amounts by which each constraint is violated.

The resulting estimator then either provides a point, or set of points, no matter what is the sampled data.

2.5 A Brief Discussion of Consistency

In the body of the paper below, we provide a set of conditions on the sampling process and the asymptotic set Θ_0 that are sufficient for our estimator to be consistent. The proof works in the point identified case and in the set identified case when the multiple points are not isolated. In the set-identified case, we require that the set Θ_0 has some interior points and consists of those interior points plus the points that are on the boundary. (Formally, Θ_0 must equal the closure of its interior.) This rules out multiple isolated single points as well as isolated lines and curves.

The restriction on Θ_0 allows us to prove consistency without relying on a "tolerance" as in Manski and Tamer (2002). To see the problem considered by Manski and Tamer, consider a Θ_0 set that consists of two isolated points. At each point, one or more population inequality constraints must hold with equality (or else other nearby points would also satisfy the inequalities, assuming continuity in θ of the model probabilities). In a given sample, the estimator is likely to pick out only one of the points and the estimates could alternate between the two points as the sample gets increasingly large, never choosing the set consisting of both points.

Manski and Tamer (2002) solve the problem by introducing some extra slackness into either the constraints, or into the objective function itself. This new estimator picks a tolerance (or "tuning parameter"), δ_n, and then chooses all the θ's that give an objective function value that differs from the minimum by less than δ_n. One promises to let the tolerance go to zero as n increases.

The problem with this solution is that there is currently little guidance on picking the absolute level of δ_n (as opposed to the rate at which it must disappear). This gives an arbitrary quality to the size of estimated set, which reflects the choice of δ_n as much as the lack of point identification in the model and data.

Figure 1 illustrates a case with two parameters and two inequality constraints that yields a Θ_0 set (the union of the shaded regions) that is consistent with our assumptions. (In the figure, the first inequality constraint requires θ to be on the upper left of one curve and the second inequality constraint requires θ to be on the lower right of the other curve.) Figure 2 illustrates a case where Θ_0 consists of exactly two points (at tangencies of the curves) and therefore violates our assumptions. Note, however, that the constraints are typically formed as averages across x-cells. A slight change in the distribution of X_i would move the constraints in Figure 2 slightly, so they intersect rather than being tangent and this would remove the problem.

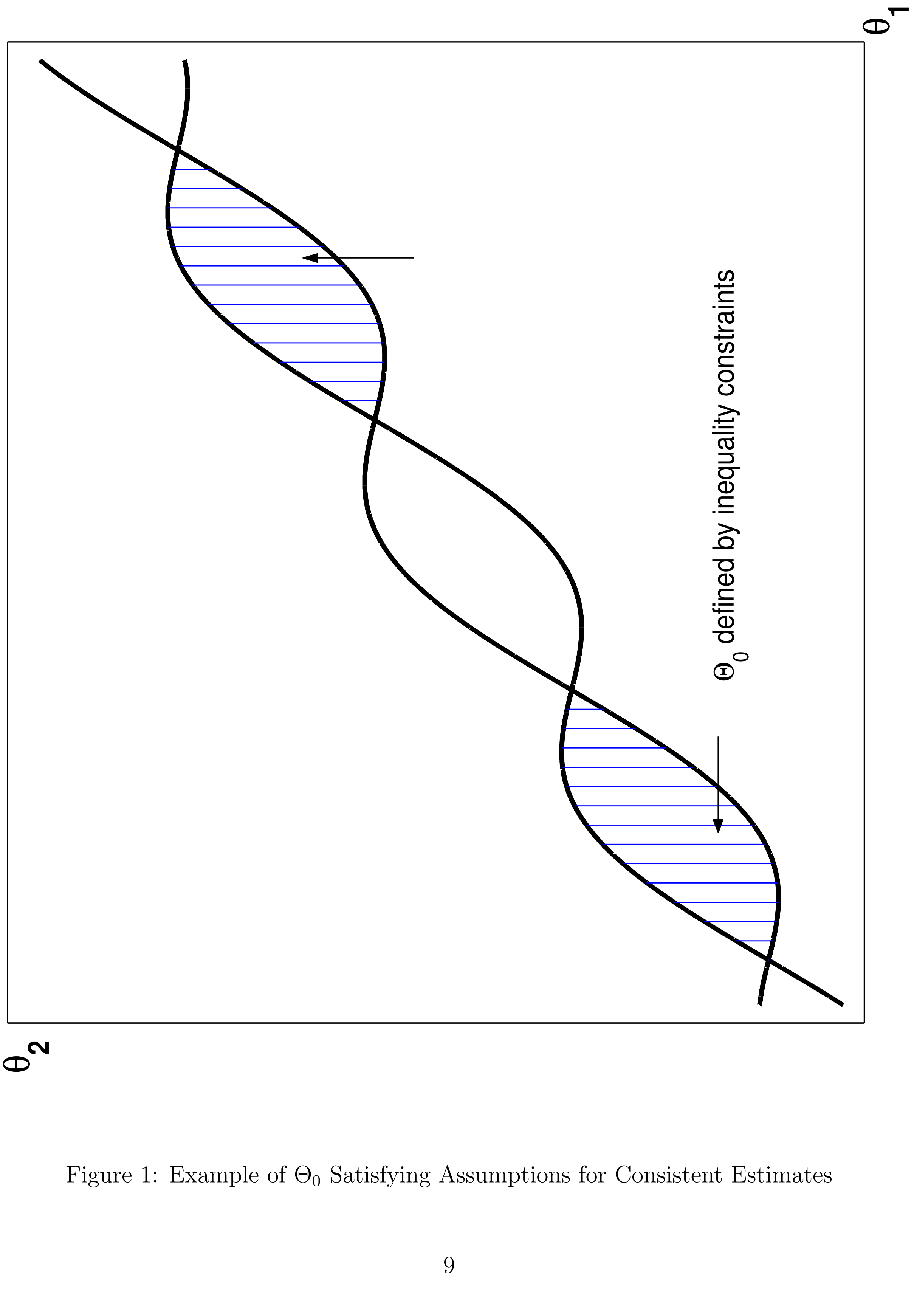


Figure 1: Example of Θ_0 Satisfying Assumptions for Consistent Estimates

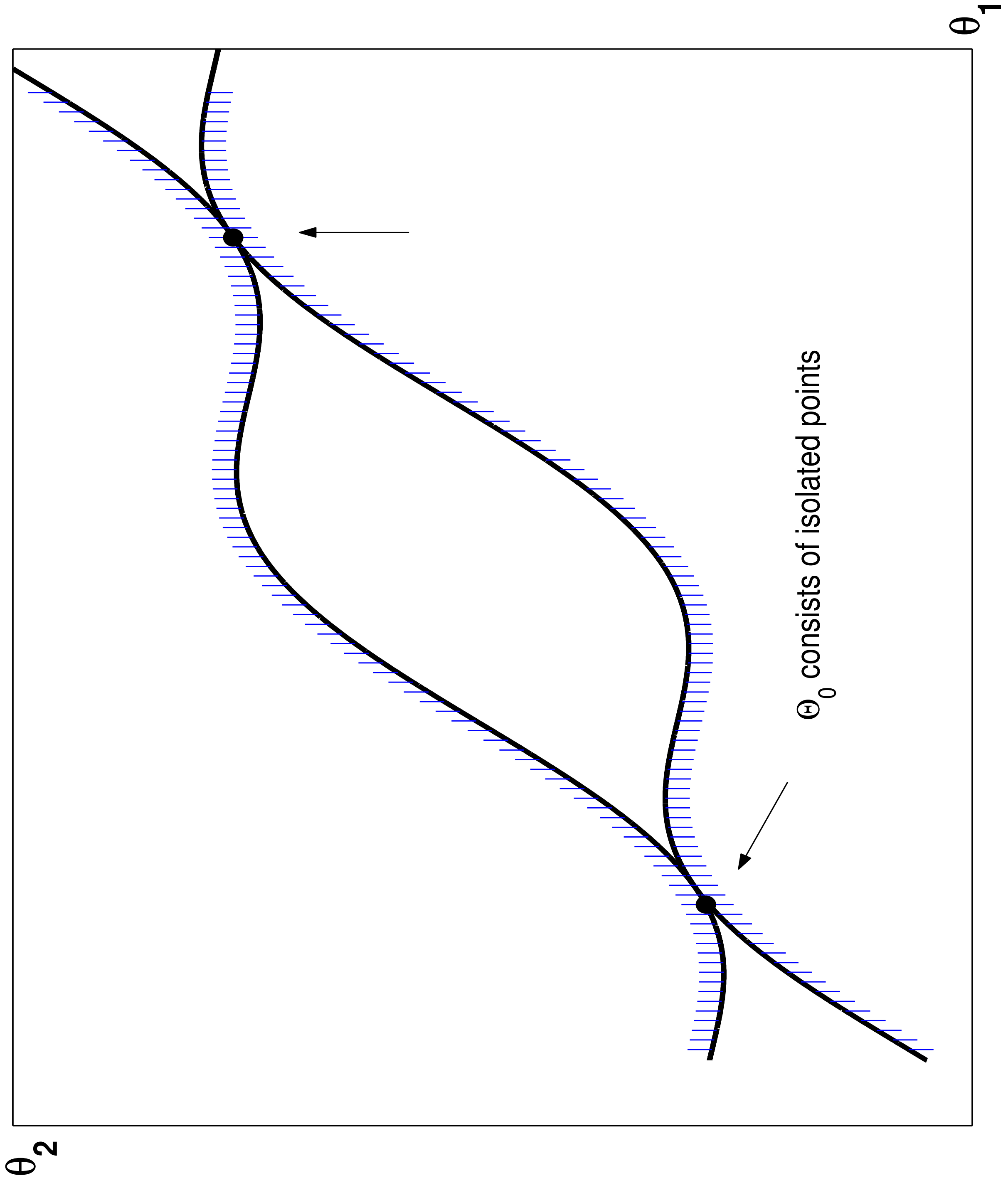


Figure 2: Example of Θ_0 Not Satisfying Assumptions for Consistent Estimates

If one is concerned about cases like that shown in Figure 2, one can still introduce a tolerance to our procedure. However, we maintain the simpler assumption throughout the paper.[2]

2.6 An Introduction to Inference on Sets

The case of set identification requires a non-standard inference procedure. One idea for constructing confidence regions would be to simply bootstrap the sample-analog estimator, calculating a new "estimate" for each bootstrapped sample. However, set-valued estimates are very hard to construct even once. Trying to compute such sets many times over will be computationally impossible in many realistic cases.

Chernozhukov, Hong, and Tamer (2003) specify a method of constructing a set that contains the true parameter θ_0, say, 95% of the time. Their method is based on the level sets of a sample objective function. The corresponding population (or asymptotic) objective function is minimized on a set, rather than at a point. They estimate the quantiles of the (normalized) sample objective function minimized over an estimate of the identified set using a subsampling technique. Given a suitable quantile of the subsample distribution of the minimum of the objective function, say $q_{.95}$, their 95% confidence region consists of all points θ that yield a value of the sample objective function within $q_{.95}$ of its minimum.

Chernozhukov, Hong, and Tamer (2003)'s method is a valuable contribution to the literature and is quite useful in a number of contexts. But, it is computationally burdensome (though still somewhat easier than a full bootstrap) in the context considered in this paper because it requires minimization of the objective function for each subsample. A principal objective of the method considered in this paper is to avoid calculations of this sort. In addition, we seek a method that yields confidence intervals for individual parameters as well as the whole parameter vector θ. Chernozhukov, Hong, and Tamer (2003)'s approach (coupled with the projection method) only yields quite conservative confidence intervals for individual parameters.

We choose an alternative approach to constructing confidence regions, based on the variance in the estimated inequalities. The basic idea is to relax the sample inequality constraints by an amount that varies with the sampling variance in the constraints.

As a simple example, consider the simple multinomial case discussed above, when there is no need to simulate the probabilities of the necessary conditions. In this case, all the variance comes from the vector of sample probabilities, $\widehat{P}_n$. We first bootstrap the distribution of the sample probabilities and from that find a new vector, $\tilde{P}_n$, such that the true probabilities P_0 are (element by element) greater than $\tilde{P}_n$ 95% of the time.

A joint confidence region for Θ_0 in this simple example is then the set of parameters that

[2] We need to provide a further explanation of consistency in the case where we make use of equality, as well as inequality constraints. It is very likely in this case that some parameters are point identified while others are not. A more complicated, but perhaps tractable, problem is the case of multiple intersections of equality constraints.

satisfy the new inequalities:

$$P(y \mid x, \theta) \geq \tilde{P}_n(y \mid x), \ \forall (y, x) \in \mathcal{Y} \times \mathcal{X}. \tag{10}$$

Because the true $P_0(y \mid x)$ is greater than $\tilde{P}_n(y|x)$ $\forall y, x$ 95% of the time, the set Θ_0 is in this confidence region at least 95% of the time.

The joint confidence region just constructed may be large because we require the entire P_0 vector to be greater than the $\tilde{P}_n$ vector 95% of the time. With a large number of cells, this can require $\tilde{P}_n$ to be far below $\widehat{P}_n$. This leads us to consider parameter-by-parameter confidence regions (or, more generally, confidence regions for any scalar function of the parameter vector, like the outcome of a policy experiment).

Furthermore, the confidence region for event probabilities in this subsection is an open rectangle, which may not be the most efficient shape. This leads us to consider methods for changing the shape of the confidence region. We provide an introduction to these important extensions in the next subsection of the paper.

2.7 Introduction to More Efficient Confidence Regions

Joint confidence regions often imply very wide confidence intervals for individual parameters and this leads to a frequent preference for parameter-specific confidence intervals. We adapt our method to the single parameter case by moving only those binding constraints that define the upper or lower bound on each parameter. Because we move fewer constraints, we do not have to move each constraint as far when ensuring that the constraints are jointly above zero with some fixed probability.

The method is easily extended to the discussion of any scalar-valued function of the parameter vector. For example, we might want to predict a change in number of competitors as a result of some exogenous policy experiment.

This still leaves the question of how exactly to compute the single-parameter confidence interval. When we relax the model's constraints according to our method, it is easy to show that we obtain *at least* a 95% confidence interval. However, we would like to get the "smallest" interval possible – that is, we would like to have an exact 95% confidence region, not a larger one.

We show that we can construct a tighter confidence region by adding a set of additional constraints that do not affect the identified set, but do affect the shape and size of the confidence regions. In particular, we note that positively weighted sums of non-negative constraints are also non-negative. Introducing such constraints does not change the identified set, because if the original constraints are positive, so are positively weighted sums of those constraints. However, the shape of the additional constraints is different from the originals, and as we relax the original and new constraints together this changes the shape of the confidence region.

When constructing the confidence interval for one parameter, we note that we can mimic the most efficient confidence interval if we can construct a constraint that is flat in the other parameters. We can, in fact, often construct a constraint that is at least locally flat in the other parameters by varying the weights in the positively weighted-sum of the original constraints. We show in Monte Carlo's that adding this constraint greatly decreases our single-parameter coverage ratio.

2.8 The Path Ahead

In the formal discussion of inference and estimation below, we modify the simple procedure presented here in several important ways. First, we do not assume discrete X_i's. We discuss the construction of cells from continuous and/or discrete X_i's.

Second, we consider confidence intervals (CIs) for individual elements of θ rather than a multi-dimensional joint confidence region for θ. This allows one to consider only the subset of inequality constraints that define the upper and lower bound for the given parameter. In turn, this mitigates the problem mentioned above of finding a joint confidence region for the whole P_0 vector.

Third, we can employ the insight of Imbens and Manski (2003), who point out that a confidence region for the true parameter θ_0 may be smaller than the confidence region for the entire identified set Θ_0. Typically, we desire a confidence region for the true parameter (or a function of the true parameter) and so we focus on this case. We do briefly discuss inference for the identified set.

Fourth, we allow for the possibility that the probabilities of necessary conditions have to be simulated. We typically focus on the case of the simple frequency simulator, but other simulations methods can be treated similarly.

Finally, and most importantly, in contrast to the informal discussion of this section, we provide a full set of sufficient conditions for the consistent estimation of Θ_0 and the construction of asymptotically valid CIs for real-valued functions of the parameters.

3 Set Estimators

We now provide further detail concerning the set estimator introduced in the preceding section and provide consistency results.

3.1 Simulation of Model Probabilities

In many cases, the model probabilities $\{P(y|X_i,\theta) : i = 1,...,n\}$ do not have closed form expressions and need to be obtained via simulation. We let $\{\widehat{P}_S(y|X_i,\theta) : i = 1,...,n\}$ denote the estimators of $\{P(y|X_i,\theta) : i = 1,...,n\}$ based on S simulation draws for each i using simulation random variables that are iid across i. If the probabilities $\{P(y|X_i,\theta) : i = 1,...,n\}$ can be computed analytically, then they are used in place of $\widehat{P}_S(y|X_i,\theta)$. For notational convenience, we still denote them by $\{\widehat{P}_S(y|X_i,\theta) : i = 1,...,n\}$.

An example of a simulation estimator is the crude frequency simulator. In this case, the model probabilities are simulated by taking nS iid draws $\{\varepsilon_i(s) : i = 1,...,n,\ s = 1,...,S\}$

each with the same distribution as $\varepsilon_i = (\varepsilon_{i,1}, ..., \varepsilon_{i,J})'$ and forming

$$\widehat{P}_S(y|X_i, \theta) = S^{-1} \sum_{s=1}^{S} [\varepsilon_i(s) \in \Omega(y, X_i, \theta)] \text{ for } i = 1, ..., n, \tag{11}$$

where $[\cdot]$ denotes the indicator function. Note that the same variables $\{\varepsilon_i(s) : s = 1, ..., S\}$ are used to simulate $\widehat{P}_S(y|X_i, \theta)$ for all values of y and θ. The function $\widehat{P}_S(y|X_i, \theta)$ is not necessarily continuous in θ even if $\Omega(y, x, \theta)$ is continuous in θ because of the indicator function. If the distribution of ε_i depends on parameters (included in θ), as it often does, then we assume $\varepsilon_i = g(\eta_i, \theta)$ for some random vector η_i with known distribution and some known function g and we take $\varepsilon_i(s) = g(\eta_i(s), \theta)$, where $\{\eta_i(s) : s = 1, ..., S\}$ are iid each with the same distribution as η_i.

In many cases, it is desirable to use a more sophisticated simulator than the crude frequency simulator, such as an importance sampling simulator.

3.2 Construction of Y Cells

Here we add some detail on the construction of the Y_i and X_i cells to be used in the estimation. This lets us move beyond simple multinomial example of the earlier section. Note that even though Y_i is discrete, it may be desirable to aggregate across different events to avoid sparsely populated cells.

Formally, let $\{\mathcal{Y}_k : k = 1, ..., K\}$ denote K distinct subsets of $\mathcal{Y}$ (none of which equals $\varnothing$ or $\mathcal{Y}$). For example, $\{\mathcal{Y}_k : k = 1, ..., K\}$ could be a partition of $\mathcal{Y}$ comprised of singleton sets. Alternatively, if $\mathcal{Y}$ contains a large number of elements, y, with $P_0(Y_i = y|x)$ being quite small for some or all elements, then it is advantageous to group the decision vectors and form sets $\{\mathcal{Y}_k : k = 1, ..., K\}$ that may contain more than one element. The sets $\{\mathcal{Y}_k : k = 1, ..., K\}$ need not be disjoint, but usually one would choose them to be so.

In practice, it is often useful to use a data-dependent method to select a random collection of sets $\{\widehat{\mathcal{Y}}_{n,k} : k = 1, ..., \widehat{K}_n\}$ in order to guarantee that sets whose probabilities are quite small are not selected. For example, suppose we wish to create sets $\{\mathcal{Y}_k : k = 1, ..., K\}$ whose probabilities are all greater than δ for some $\delta > 0$. Then, we can use the sets $\{\widehat{\mathcal{Y}}_{n,k} : k = 1, ..., \widehat{K}_n\}$, which consist of all distinct subsets of $\mathcal{Y}$ for which

$$\widehat{P}_n(Y_i \in \widehat{\mathcal{Y}}_{n,k}) = n^{-1} \sum_{i=1}^{n} [Y_i \in \widehat{\mathcal{Y}}_{n,k}] > \delta \tag{12}$$

for all $k = 1, ..., \widehat{K}_n$. Alternatively, one need not consider all subsets that satisfy (12). One could consider some smaller collection of sets subject to the restriction that each set satisfies (12).

If $\{Y_i : i \geq 1\}$ are iid, then $\widehat{P}_n(Y_i \in \mathcal{Y}_*) \to P_0(Y_i \in \mathcal{Y}_*)$ uniformly over all subsets $\mathcal{Y}_*$ of $\mathcal{Y}$ (of which there is a finite number) by the WLLN. Let $\{\mathcal{Y}_k : k = 1, ..., K\}$ denote the subsets of $\mathcal{Y}$ that satisfy

$$P_0(Y_i \in \mathcal{Y}_k) > \delta \text{ for } k = 1, ..., K. \tag{13}$$

Then, it follows that

$$P_0\left(\{\widehat{\mathcal{Y}}_{n,k} : k = 1, ..., \widehat{K}_n\} = \{\mathcal{Y}_k : k = 1, ..., K\}\right) \to 1. \tag{14}$$

For any collection of random sets $\{\widehat{\mathcal{Y}}_{n,k} : k = 1, ..., \widehat{K}_n\}$, as long as (14) holds for some non-random sets $\{\mathcal{Y}_k : k = 1, ..., K\}$, the asymptotic properties of procedures (considered below) based on $\{\widehat{\mathcal{Y}}_{n,k} : k = 1, ..., \widehat{K}_n\}$ are the same as those of procedures based on $\{\mathcal{Y}_k : k = 1, ..., K\}$.[3] In consequence, for simplicity and without loss of generality (wlog), we consider the non-random sets $\{\mathcal{Y}_k : k = 1, ..., K\}$ in the discussion below.

3.3 Construction of X Cells

Next, we discuss data-dependent construction of X_i cells from possibly continuous X_i's. We formally allow for data-dependent selection of X_i cells because this data-dependence affects the asymptotic distribution of the statistics considered below. Data-dependent grouping into X_i cells needs to be accounted for in the determination of the critical values that are employed.

Consider a set $\{h_\gamma : \gamma \in \Gamma\}$ of real-valued weight functions on the support $\mathcal{X}$ of X_i, where γ is a subset of $\mathcal{X}$ and Γ is a set of subsets of $\mathcal{X}$. In particular, for each $\mathcal{Y}_k$ set, we consider M_k subsets of $\mathcal{X}$ denoted by $\gamma_{k,m}$:

$$\begin{aligned}\Gamma &= \{\gamma_{k,m} \subset \mathcal{X} : (k, m) \in \mathcal{I}_{K,M}\}, \text{ where} \\ \mathcal{I}_{K,M} &= \{(k, m) : m = 1, ..., M_k, k = 1, ..., K\}.\end{aligned} \tag{15}$$

The functions $\{h_\gamma : \gamma \in \Gamma\}$ are used to aggregate and/or weight the necessary conditions for an equilibrium over different values of x. Doing so is useful if either X_i contains a continuous component or X_i is discrete but takes on a large number of different values, some or all of which occur with small probability.

Examples of different h_γ functions that could be employed are (i) $h_\gamma(x) = [x \in \gamma]$ and (ii) $h_\gamma(x)$ is a smoothed version of $[x \in \gamma]$ that equals one when $x \in \gamma$ and equals zero when $x \notin S(\gamma, \varepsilon)$ for some $\varepsilon > 0$, where $S(\gamma, \varepsilon) = \{x \in \mathcal{X} : ||x - x'|| < \varepsilon \ \forall x' \in \gamma\}$.

In the case where X_i is discrete and takes on a relatively small number of values, nothing is lost by considering weight functions $\{h_\gamma : \gamma \in \Gamma\}$ because one can take (iii) $h_\gamma(x) = [x = \gamma]$, where γ is a set containing a single element of $\mathcal{X}$ and Γ is a partition of $\mathcal{X}$ consisting of singleton sets, which provides a separate function for each $x \in \mathcal{X}$ and no aggregation or weighting occurs.

When X_i has a continuous component, it is desirable to consider functions h_γ based on a data-dependent collection $\widehat{\Gamma}_n$ of subsets of $\mathcal{X}$. The object is to use the data to find sets of x values in which the conditional probability of $Y_i \in \mathcal{Y}_k$ given $X_i = x$ varies as little as possible across x values. This is desirable because it leads to as little loss of information as possible when aggregating over different x values.

[3] This follows by considering the intersection of any sequence of sets of interest with the sets in (14) and using the fact that the complements of these intersected sets necessarily converge in probability to zero.

Let

$$\widehat{\Gamma}_n = \{\widehat{\gamma}_{n,k,m} \subset \mathcal{X} : (k,m) \in \mathcal{I}_{K,M}\}, \tag{16}$$

where $\widehat{\gamma}_{n,k,m}$ is a random subset of $\mathcal{X}$. We require that the random sets satisfy $\widehat{\Gamma}_n \rightarrow_p \Gamma_0$ (in a sense made precise in the following subsection), where

$$\Gamma_0 = \{\gamma_{0,k,m} \subset \mathcal{X} : (k,m) \in \mathcal{I}_{K,M}\}. \tag{17}$$

For example, one can select X_i cells based on the sample quantiles of the elements of X_i. Let $X_{i,\ell}$ denote the ℓ-th element of X_i. (This is an abuse of notation because $X_{i,\ell}$ is different from $X_{i,j}$ above.) One can group $\{X_{i,\ell} : i = 1, ..., n\}$ into the values $\leq$ its sample median and $\geq$ its sample median for each $\ell = 1, ..., d_x$. Then, 2^{d_x} cells in $\mathcal{X}$ are obtained by taking all the combinations of hi/low cells for $\ell = 1, ..., d_x$. A finer grid of X_i cells can be obtained by creating more cells for one or more $X_{i,\ell}$ variable by using two or more quantiles in place of the median. Fewer cells can be obtained by not splitting the observations into two groups for one or more $X_{i,\ell}$.

The above method yields X_i cells that contain different numbers of observations due dependence between the elements of X_i. For example, when $d_x = 2$, a hi/hi cell will have more observations than a hi/low cell when $X_{i,1}$ and $X_{i,2}$ are positively related.

A method that yields approximately equally-populated X_i cells is as follows. Create two cells for $X_{i,1}$ using its sample median. Then, for the observations in the low $X_{i,1}$ cell, create two cells for $X_{i,2}$ using the sample median of $X_{i,2}$ out of the set of X_i's for which $X_{i,1}$ is in the low cell. Do likewise for the observations in the hi $X_{i,1}$ cell. Continue this process for additional regressors $\ell = 3, ..., d_x$. This procedure yields 2^{d_x} cells. More or less cells can be obtained by the same method as above.

Given the data-dependent X_i cells, it may be necessary to aggregate some Y_i cells for certain X_i cells due to under-populated (Y_i, X_i) cells.

WE ARE STILL WORKING ON THE BEST WAY TO CHOOSE CELLS.

3.4 Definition of a Set Estimator

We define

$$\begin{aligned}
P_0(\mathcal{Y}_k|x) &= \sum_{y \in \mathcal{Y}_k} P_0(y|x), \\
P(\mathcal{Y}_k|x,\theta) &= \sum_{y \in \mathcal{Y}_k} P(y|x,\theta), \\
\widehat{P}_S(\mathcal{Y}_k|X_i,\theta) &= \sum_{y \in \mathcal{Y}_k} \widehat{P}_S(y|X_i,\theta), \\
c_0(k,\gamma,\theta) &= \int \left(P(\mathcal{Y}_k|x,\theta) - P_0(\mathcal{Y}_k|x)\right) h_\gamma(x) dG(x), \text{ and} \\
\widehat{c}_n(k,\gamma,\theta) &= n^{-1} \sum_{i=1}^{n} \left(\widehat{P}_S(\mathcal{Y}_k|X_i,\theta) - [Y_i \in \mathcal{Y}_k]\right) h_\gamma(X_i).
\end{aligned} \tag{18}$$

Note that $E\widehat{c}_n(k,\gamma,\theta) = c_0(k,\gamma,\theta)$ for all (k,γ,θ). Hence, with iid observations, $\widehat{c}_n(k,\gamma,\theta) \rightarrow_p c_0(k,\gamma,\theta)$ for all (k,γ,θ) provided $c_0(k,\gamma,\theta)$ is well-defined.

Necessary conditions for θ to be the true parameter are

$$P(y|x,\theta) - P_0(y|x) \geq 0,\ \forall (y,x) \in \mathcal{Y} \times \mathcal{X}. \tag{19}$$

These conditions imply that

$$c_0(k,\gamma_{0,k,m},\theta) \geq 0,\ \forall (k,m) \in \mathcal{I}_{K,M}. \tag{20}$$

Define

$$\begin{aligned}\Theta_0 &= \{\theta \in \Theta : (19) \text{ holds}\} \text{ and} \\ \Theta_+ &= \{\theta \in \Theta : (20) \text{ holds}\}.\end{aligned} \tag{21}$$

The set Θ_0 is the smallest set of parameter values that necessarily includes the true value θ_0. The set $\Theta_+ \supset \Theta_0$.

Suppose (a) the regressors are discrete with finite support, (b) $\{h_\gamma : \gamma \in \Gamma\}$ is as in case (iii) above, and (c) $\mathcal{Y}_k = \{y_k\}$ for $k = 1,...,K$, where $\mathcal{Y} = \{y_k\colon\ k = 1,...,K\}$. Then, the necessary conditions (19) and (20) are equivalent and $\Theta_+ = \Theta_0$.

More generally, Θ_+ is a larger set than Θ_0 and there is a loss of information when focusing on the necessary conditions in (20), rather than those in (19). However, when X_i has a continuous component or is discrete with a large support, some aggregation and/or weighting is desirable for statistical reasons which leads us to employ the necessary conditions in (20). The reason is as follows. Nonparametric estimators of $P_0(\mathcal{Y}_k|x)$ would be required in the continuous regressor case in order to focus on (19) and nonparametric estimators have a number of drawbacks which makes their use problematic in the present context. First, they are subject to the curse of dimensionality, which implies that they are infeasible in some cases and are likely to perform poorly in finite samples in other cases. Second, they are not uniformly consistent over $\mathcal{X}$ without restrictions on $\mathcal{X}$, such as truncation, and the latter typically involves a loss of information. Third, joint confidence bounds for nonparametric estimators over uncountable sets of regressors values are not available in the literature and, hence, nonparametric estimators cannot be used to construct confidence intervals for parameters in the manner described below. Fourth, nonparametric estimators require the choice of smoothing parameters, which can be somewhat arbitrary.

In the case of discrete regressors with large support, focus on the necessary conditions in (19) runs into the problem of poor finite sample performance due to the high variability of estimators of the cell probabilities $P_0(y|x)$ when the number of cells is large and, hence, the probabilities of most cells are small relative to the sample size. Thus, in this case too, it is advantageous to consider the necessary conditions in (20), rather than those in (19).

We now define a set estimator $\widehat{\Theta}_n$ of Θ_+. First, we define the estimator criterion function:

$$Q_n(\theta) = \sum_{(k,m)\in\mathcal{I}_{K,M}} |\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta)| \cdot [\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta) \leq 0]. \tag{22}$$

A sample constraint (i.e., necessary condition) is violated when $\widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) \leq 0$. Hence, $Q_n(\theta)$ is the sum over constraints of the absolute magnitude of sample constraint violations. The criterion function $Q_n(\theta)$ is the sample analogue of

$$Q(\theta) = \sum_{(k,m) \in \mathcal{I}_{K,M}} |c_0(k, \gamma_{0,k,m}, \theta)| \cdot [c_0(k, \gamma_{0,k,m}, \theta) \leq 0]. \tag{23}$$

The function $Q(\theta)$ is minimized (and equals zero) for all values θ for which the necessary conditions $c_0(k, \gamma_{0,k,m}, \theta) \geq 0$ for all $(k, m) \in \mathcal{I}_{K,M}$ hold. In consequence,

$$\Theta_+ = \{\theta \in \Theta : \theta \text{ minimizes } Q(\theta) \text{ over } \Theta\}. \tag{24}$$

For this reason, we define the set estimator $\widehat{\Theta}_n$ to be

$$\widehat{\Theta}_n = \{\theta \in \Theta : \theta \text{ minimizes } Q_n(\theta) \text{ over } \Theta\}. \tag{25}$$

It is easy to see that if there exists a value of θ for which $\widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0$, $\forall (k, m) \in \mathcal{I}_{K,M}$, then $\widehat{\Theta}_n$ equals

$$\{\theta \in \Theta : \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0, \ \forall (k, m) \in \mathcal{I}_{K,M}\}. \tag{26}$$

In this case, $\widehat{\Theta}_n$ consists of all points in Θ for which the estimated necessary condition $\widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0$ holds for all $(k, m) \in \mathcal{I}_{K,M}$.

Because of randomness in the estimator $\widehat{c}_n$ of c_0, it is possible that the set in (26) is empty. For example, if Θ_+ contains only the true parameter θ_0, this holds with probability bounded away from zero as $n \to \infty$. For this reason, we do not define $\widehat{\Theta}_n$ to be the set in (26), but rather define it as in (25). This minimization definition guarantees that $\widehat{\Theta}_n$ is not empty.[4]

We show in the next subsection that $\widehat{\Theta}_n$ converges in probability to Θ_+ in a certain sense under suitable assumptions.

Next, we consider estimation of a real function $\beta_n(\theta)$ of θ. Here, $\beta_n(\cdot)$ is a known real function of θ that may depend on $\{X_i : i \leq n\}$ and/or on n. The leading case is when $\beta_n(\theta)$ equals some element of θ. In this case, $\beta_n(\cdot)$ is non-random and does not depend on n. More generally, $\beta_n(\theta)$ can be some policy variable/parameter. GIVE MORE DETAILS RE THE LATTER. PERHAPS REFERENCE APPLICATION GIVEN BELOW.

The largest and smallest values of $\beta_n(\theta)$ across all θ values that satisfy the necessary conditions (20) are

$$\begin{aligned} \beta_{n,U} &= \sup\{\beta_n(\theta) : \theta \in \Theta_+\} \text{ and} \\ \beta_{n,L} &= \inf\{\beta_n(\theta) : \theta \in \Theta_+\}. \end{aligned} \tag{27}$$

[4] The criterion function $Q_n(\theta)$ is not necessarily continuous because the simulated probabilities $\{\widehat{P}_S(\mathcal{Y}_k|X_i, \theta) : i \geq 1\}$ are not necessarily continuous. Nevertheless, the set $\widehat{\Theta}_n$ cannot be empty because $\widehat{P}_S(\mathcal{Y}_k|X_i, \theta)$, $\widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta)$, and $Q_n(\theta)$ only take on a finite number of different values for $\theta \in \Theta$. If the probabilities $\{P(y|X_i, \theta) : i = 1, ..., n\}$ are computed analytically, then $\widehat{\Theta}_n$ cannot be empty because $Q_n(\theta)$ is a continuous function defined on a compact set under the assumption given below that $P(y|x, \theta)$ is continuous on Θ for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$.

Estimators of $\beta_{n,U}$ and $\beta_{n,L}$ based on $\widehat{\Theta}_n$ are given by

$$\begin{aligned}\widehat{\beta}_{n,U} &= \sup\{\beta_n(\theta) : \theta \in \widehat{\Theta}_n\} \text{ and} \\ \widehat{\beta}_{n,L} &= \inf\{\beta_n(\theta) : \theta \in \widehat{\Theta}_n\},\end{aligned} \tag{28}$$

respectively.

3.5 Probability Limit of the Set Estimator

In this subsection, we show that $\widehat{\Theta}_n$, $\widehat{\beta}_{n,U}$, and $\widehat{\beta}_{n,L}$ are consistent estimators of Θ_+, $\beta_{n,U}$, and $\beta_{n,L}$, respectively, under some primitive conditions and some high-level conditions that depend on the choice of $\widehat{\Gamma}_n$. In Section **??**, we verify the high-level conditions for the "binary response model" choice of $\widehat{\Gamma}_n$ discussed above.

To show that $\widehat{\Theta}_n \rightarrow_p \Theta_+$, we need to specify a measure of distance between two sets of p-vectors. We employ the Hausdorff metric, which is a strong metric. For two sets of p-vectors A and B, let the maximum distance from any point in A to B be given by

$$\begin{aligned}\rho(A|B) &= \sup_{a \in A} \rho(a|B), \text{ where for } a \in A \\ \rho(a|B) &= \inf\{||a-b|| : b \in B\}.\end{aligned} \tag{29}$$

Note that $\rho(a|B)$ is the distance from the vector a to the set B. By definition, if $A = \varnothing$ and $B \neq \varnothing$, then $\rho(A|B) = 0$. Also, if $B = \varnothing$, then $\rho(A|B) = \infty$. The Hausdorff metric distance between A and B is

$$d(A,B) = \max\{\rho(A|B), \rho(B|A)\}. \tag{30}$$

For $\varepsilon > 0$, let $S(A,\varepsilon) = \{b \in R^p : \rho(b|A) < \varepsilon\}$.

We derive conditions under which

$$\begin{aligned}\text{(i) } &\rho(\widehat{\Theta}_n|\Theta_+) \rightarrow_p 0 \text{ and} \\ \text{(ii) } &\rho(\Theta_+|\widehat{\Theta}_n) \rightarrow_p 0.\end{aligned} \tag{31}$$

These results then give $d(\widehat{\Theta}_n, \Theta_+) \rightarrow_p 0$. Result (i) ensures that $\widehat{\Theta}_n$ is not larger than Θ_+ asymptotically. Result (ii) ensures that $\widehat{\Theta}_n$ is not smaller than Θ_+ asymptotically. Result (ii) ensures that the distance of the true value θ_0 ($\in \Theta_0 \subset \Theta_+$) from the set estimator $\widehat{\Theta}_n$ satisfies $\rho(\theta_0|\widehat{\Theta}_n) \rightarrow_p 0$.

Let Γ_{all} be a class of subsets of $\mathcal{X}$ that includes all possible realizations of $\widehat{\gamma}_{n,k,m}$ for all $(k,m) \in \mathcal{I}_{K,M}$ and $n \geq 1$. For convenience, let

$$\mathcal{I}_K = \{1, ..., K\}. \tag{32}$$

To establish result (i), the following assumptions are employed:

Assumption 1. $\{(Y_i, X_i) : i \geq 1\}$ are iid.

Assumption 2. The true parameter θ_0 satisfies $P(y|x,\theta_0) - P_0(y|x) \geq 0$, $\forall (y,x) \in \mathcal{Y} \times \mathcal{X}$.

Assumption 3. (a) Θ is compact.
(b) $P(y|x, \theta)$ is continuous in θ on Θ for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$.
(c) $\int |h_\gamma(x)|dG(x) < \infty$ for all $\gamma \in \Gamma_{all}$.

Assumptions 1 and 3 are fairly standard assumptions. Assumption 2 states that the model is correctly specified. It guarantees that Θ_0 and Θ_+ are not empty.

To establish result (ii), additional assumptions are required. The following condition is appropriate when the inequality restrictions employed are not generated from equality restrictions.

Let int(A) and cl(A) denote the interior and closure of a set A, respectively.

Assumption 4. Either (a) $\Theta_+ = \{\theta_0\}$ or
(b) (i) $\Theta_+ = \mathrm{cl}(\mathrm{int}(\Theta_+))$ and (ii) $\forall \theta \in \mathrm{int}(\Theta_+)$, $\inf_{(k,m)\in\mathcal{I}_{K,M}} c_0(k, \gamma_{0,k,m}, \theta) > 0$.

Assumption 4(a) holds if the necessary conditions (20) are sufficiently strong that they identify the true parameter θ_0. This is not the situation that is of primary interest in this paper. But, the results of the paper cover this case. Assumption 4(b)(i) implies that Θ_+ has a non-empty interior and does not contain isolated points. Sets with this property are called *regular* in topology. Finite unions of closed convex sets satisfy Assumption 4(a)(i)—as do many other sets. Assumption 4(b)(i) implies that for any θ in the interior of Θ_+ the necessary conditions (20) for an equilibrium hold with a strict inequality.

When the inequality restrictions include restrictions generated from equality restrictions (via the method described above), Assumption 4(b) is not applicable because Θ_+ typically has empty interior and if it does not Assumption 4(b)(ii) is violated. With equality constraints we use the following alternative assumption.

NEED TO ADD ASSUMPTION THAT HANDLES EQUALITY CONSTRAINTS.

For any two real functions c_1 and c_2 on $\mathcal{I}_K \times \Gamma_{all} \times \Theta$, let

$$||c_1 - c_2||_U = \sup_{(k,\gamma,\theta)\in\mathcal{I}_K\times\Gamma_{all}\times\Theta} |c_1(k, \gamma, \theta) - c_2(k, \gamma, \theta)|. \tag{33}$$

The random function $\widehat{c}_n$ $(= \widehat{c}_n(k, \gamma, \theta))$ needs to be a uniformly consistent estimator of c_0 $(= c_0(k, \gamma, \theta))$:

Assumption 5. $\widehat{c}_n \rightarrow_p c_0$ under $|| \cdot ||_U$.

An equivalent way of expressing Assumption 5 is $||\widehat{c}_n - c_0||_U \rightarrow_p 0$.

Below we give sufficient conditions for Assumption 5. Note that a pointwise version of Assumption 5 holds automatically under Assumptions 1 and 3(c) by the WLLN because $\widehat{c}_n(k, \gamma, \theta)$ is a sample average of iid random variables with $E\widehat{c}_n(k, \gamma, \theta) = c_0(k, \gamma, \theta)$. To strengthen pointwise convergence to uniform convergence over $\mathcal{I}_K \times \Gamma_{all} \times \Theta$ requires some additional conditions on the model probabilities $\{P(y|x, \theta) : \theta \in \Theta\}$ and the functions $\{h_\gamma(x) : \gamma \in \Gamma_{all}\}$, such as Vapnik-Cervonenkis (VC) or metric entropy with bracketing conditions, see below.

As noted above, we take the sets $\widehat{\Gamma}_n$ such that they converge in probability to some non-random sets Γ_0. To make the notation of convergence of random sets precise, we need to specify a semi-norm or pseudo-metric on a space of subsets of $\mathcal{X}$. We specify a semi-norm that yields a pseudo-metric that is weaker than the metric $d(\cdot,\cdot)$ introduced above.[5] We define a semi-norm $||\cdot||_G$ on Γ_{all} as follows: for $\gamma_1, \gamma_2 \in \Gamma_{all}$,

$$||\gamma_1 - \gamma_2||_G^2 = \int |h_{\gamma_1}(x) - h_{\gamma_2}(x)|^2 dG(x), \tag{34}$$

where $G(\cdot)$ is the distribution of X_i. We define a corresponding semi-norm on the space of collections of $\sum_{k=1}^K M_k$ subsets of Γ_{all} via

$$\begin{aligned} ||\Gamma_1 - \Gamma_2||_G &= \max_{(k,m)\in\mathcal{I}_{K,M}} ||\gamma_{1,k,m} - \gamma_{2,k,m}||_G, \text{ where} \\ \Gamma_1 &= \{\gamma_{1,k,m} \in \Gamma_{all} : (k,m) \in \mathcal{I}_{K,M}\} \text{ and} \\ \Gamma_2 &= \{\gamma_{2,k,m} \in \Gamma_{all} : (k,m) \in \mathcal{I}_{K,M}\}. \end{aligned} \tag{35}$$

We require that $\widehat{\Gamma}_n$ satisfies the following assumption.

Assumption 6. $\widehat{\Gamma}_n \to_p \Gamma_0$ under $||\cdot||_G$.

Sufficient conditions for Assumption 6 are given below.

Using the fact that $\Theta_+ = \{\theta \in \Theta : \theta$ minimizes $Q(\theta)$ over $\Theta\}$, we obtain the following result:

Theorem 1. (a) *Under Assumptions* 2, 3, 5, *and* 6, $\rho(\widehat{\Theta}_n|\Theta_+) \to_p 0$.
(b) *Under Assumptions* 3-6, $\rho(\Theta_+|\widehat{\Theta}_n) \to_p 0$.
(c) *Under Assumptions* 2-6, $d(\widehat{\Theta}_n, \Theta_+)) \to_p 0$.

Comments. 1. The results of the Theorem hold for any sequence of random real functions $\{\widehat{c}_n : n \geq 1\}$ on $\mathcal{I}_K \times \Gamma_{all} \times \Theta$ that satisfy Assumption 5, not just for $\widehat{c}_n$ as defined in (18).

2. Under Assumption 4(b), $\widehat{\Theta}_n$ equals the set in (26) with probability that goes to one as $n \to \infty$.

Next, we use the result of Theorem 1 to establish consistency of $\widehat{\beta}_{n,U}$ for $\beta_{n,U}$ and $\widehat{\beta}_{n,L}$ for $\beta_{n,L}$. Such results only hold if $\beta_n(\theta)$ exhibits some continuity property. If $\beta_n(\cdot)$ is non-random and does not depend on n, then the requisite condition is that $\beta_n(\cdot)$ is continuous on Θ. If $\beta_n(\cdot)$ is random and/or depends on n, then the requisite condition is that $\{\beta_n(\cdot) : n \geq 1\}$ is *stochastically equicontinuous* on Θ. This condition reduces to continuity of $\beta_n(\cdot)$ on Θ if $\beta_n(\cdot)$ is non-random and does not depend on n.

By definition, $\{\beta_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous on Θ if given any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left(\sup_{||\theta_1-\theta_2||\leq\delta} |\beta_n(\theta_1) - \beta_n(\theta_2)| > \varepsilon\right) < \varepsilon. \tag{36}$$

Assumption 7. $\{\beta_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous on Θ.

Primitive sufficient conditions for stochastically equicontinuity of $\{\beta_n(\cdot) : n \geq 1\}$ are given in Andrews (1992).

Using the result of Theorem 1 that $d(\widehat{\Theta}_n, \Theta_+) \to_p 0$, we obtain:

[5] This makes it easier to verify that $\widehat{\Gamma}_n \to_p \Gamma_0$.

Theorem 2. *Under Assumptions* 2-7, $\widehat{\beta}_{n,U} - \beta_{n,U} \to_p 0$ *and* $\widehat{\beta}_{n,L} - \beta_{n,L} \to_p 0$.

PERHAPS GIVE RESULTS FOR PLIM OF A SET ESTIMATOR WITHOUT ASSUMPTION 4 BY USING POINTWISE IN θ UPPER BOUNDS THAT HOLD POINTWISE OVER (k, m).

4 Confidence Intervals

In this section, we introduce CIs both for the true value β_0 and for the identified set of β values.

4.1 Confidence Interval for β_0

THIS WRITE-UP DOES NOT INCLUDE THE USE OF FLAT CONSTRAINTS. THE LATTER WILL BE ADDED.

THIS WRITE-UP DOES NOT INCLUDE WEIGHT FUNCTION. NEED TO ADD IT BACK IN.

First, we a construct a CI for the true value $\beta_0 = \beta_n(\theta_0)$, where (as above) $\beta_n(\cdot)$ is a known real function of θ that may depend on n and $\{X_i : i \leq n\}$. (For notational simplicity, we do not make the potential dependence of β_0 on n explicit.) As in Imbens and Manski (2003), the CI in this section is for the true value β_0, not for the set of values $\beta_n(\theta)$ for which $\theta \in \Theta_+$. A CI for the latter is given in the next section.

The CI is

$$CI_n(1-\alpha) = [\widetilde{\beta}_{n,L}, \widetilde{\beta}_{n,U}]. \tag{37}$$

This CI is defined such that

$$\liminf_{n\to\infty} P(\beta_0 \subset CI_n(1-\alpha)) = \liminf_{n\to\infty} P(\widetilde{\beta}_{n,L} \leq \beta_0 \leq \widetilde{\beta}_{n,U}) \geq 1 - \alpha. \tag{38}$$

The upper and lower bounds, $\widetilde{\beta}_{n,U}$ and $\widetilde{\beta}_{n,L}$, are of the following form:

$$\begin{aligned} \widetilde{\beta}_{n,U} &= \sup\{\beta_n(\theta) : \theta \in \Theta \ \& \ \widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}\} \text{ and} \\ \widetilde{\beta}_{n,L} &= \inf\{\beta_n(\theta) : \theta \in \Theta \ \& \ \widetilde{c}_{n,L}(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,L}\}, \end{aligned} \tag{39}$$

where $\widetilde{c}_{n,U}$ is an upper bound on c_0 for those $(\mathcal{Y}_k, \widehat{\gamma}_{n,k,m})$ sets for which (k,m) is in $\widehat{\mathcal{B}}_{n,U}$ (defined below) for a particular value of θ, viz., $\theta_{n,U}$ (defined below) and, analogously, $\widetilde{c}_{n,L}$ is an upper bound on c_0 for (k,m) in $\widehat{\mathcal{B}}_{n,L}$ for $\theta = \theta_{n,L}$.

By definition, $\widetilde{c}_{n,U}$ and $\widetilde{c}_{n,L}$ are random real functions on Θ defined by

$$\begin{aligned} \widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta) &= \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) + \widehat{w}_n(k, \gamma, \theta)\lambda^*_{n,U}(k,m,\alpha)/\sqrt{n} \text{ and} \\ \widetilde{c}_{n,L}(k, \widehat{\gamma}_{n,k,m}, \theta) &= \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) + \widehat{w}_n(k, \gamma, \theta)\lambda^*_{n,L}(k,m,\alpha)/\sqrt{n}, \end{aligned} \tag{40}$$

where $\widehat{w}_n(k, \gamma, \theta)$ is a positive weight function and $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ are non-negative critical values. For example, the weight function can be defined to approximately stabilize the standard deviation of $\widehat{c}_n(k,\gamma,\theta)$ across different (k,γ,θ) values. In this case, we let[6]

$$\widehat{w}_n(k,\gamma,\theta) \;=\; \left(n^{-1}\sum_{i=1}^{n}\left((\widehat{P}_S(\mathcal{Y}_k|X_i,\theta)-[Y_i\in\mathcal{Y}_k])h_\gamma(X_i)-\widehat{c}_n(k,\gamma,\theta)\right)^2\right)^{1/2}. \tag{41}$$

The critical values $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,U}(k,m,\alpha)$ are obtained using the bootstrap, as described in the following subsection.

The sets $\widehat{\mathcal{B}}_{n,U}$ and $\widehat{\mathcal{B}}_{n,L}$ are subsets of $\mathcal{I}_{K,M}$ that consist of the constraints that are binding at two boundary points of $\widehat{\Theta}_n$, viz., $\widehat{\theta}_{n,U}$ and $\widehat{\theta}_{n,L}$, respectively. In particular,

$$\begin{aligned}\widehat{\mathcal{B}}_{n,U} &= \{(k,m)\in\mathcal{I}_{K,M}:\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,U})=0\}\text{ and}\\ \widehat{\mathcal{B}}_{n,L} &= \{(k,m)\in\mathcal{I}_{K,M}:\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,L})=0\}.\end{aligned} \tag{42}$$

The random vectors $\widehat{\theta}_{n,U}$ and $\widehat{\theta}_{n,L}$ are defined as follows. Let

$$\widehat{\Theta}_{n,U}=\{\theta\in\widehat{\Theta}_n:\beta_n(\theta)=\widehat{\beta}_{n,U}\}, \tag{43}$$

where $\widehat{\beta}_{n,U}$ is defined in (28). The set $\widehat{\Theta}_{n,U}$ is not empty because $\widehat{\Theta}_n$ is compact. (The latter holds because Θ is compact and $\widehat{\Theta}_n$ is defined using the non-strict inequality $\geq$.) We select a unique value $\widehat{\theta}_{n,U}$ from $\widehat{\Theta}_{n,U}$ by taking the value that has the smallest Euclidean norm:

$$\widehat{\theta}_{n,U}=\arg\min\{||\theta||:\theta\in\widehat{\Theta}_{n,U}\}. \tag{44}$$

(Note that the arg min is attained by a parameter value in $\widehat{\Theta}_{n,U}$ because $\widehat{\Theta}_{n,U}$ is compact.[7]) By definition, $\beta_n(\widehat{\theta}_{n,U})=\widehat{\beta}_{n,U}$.

We define $\widehat{\Theta}_{n,L}$ and $\widehat{\theta}_{n,L}$ analogously with $\widehat{\beta}_{n,L}$ in place of $\widehat{\beta}_{n,U}$.

4.2 Bootstrap Critical Values for CI for β_0

The bootstrap critical values $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ are obtained using the standard nonparametric iid bootstrap as follows. Let $\{(Y_i^*,X_i^*):i=1,...,n\}$ denote a *bootstrap sample* of iid random variables (conditional on the original sample). Each observation (Y_i^*,X_i^*)

[6] This choice of weight function is not a consistent estimator of the asymptotic variance of $n^{1/2}(\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta)-c_0(k,\gamma_{0,k,m},\theta))$ because it does not take account of the effect of the random set $\widehat{\gamma}_{n,k,m}$ on the asymptotic distribution. Nevertheless, this does not affect the asymptotic validity of the CI because any weight function yields a CI with correct asymptotic coverage probability. The weight function is employed to make $\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta)/\widehat{w}_n(k,\widehat{\gamma}_{n,k,m},\theta)$ have comparable distributions across different (k,m) values which leads to a smaller $\lambda^*_n(\alpha)$ value, which, in turn, leads to a shorter CI.

[7] If the arg min is not unique, then some rule is needed to uniquely define it. For example, $\widehat{\theta}_{n,U}$ could be defined to be the arg min whose first element is smallest. If the latter is not unique, then the arg min is defined to be the one whose first and second elements are smallest, etc.

in the bootstrap sample has a discrete distribution with probability $1/n$ of equaling each observation in the original sample $\{(Y_i, X_i) : i = 1, ..., n\}$.

Let $P_S^*(y|X_i^*, \theta)$ denote the bootstrapped value of $\widehat{P}_S(y|X_i, \theta)$ for $i = 1, ..., n$. When simulation is not required to compute $P(y|X_i, \theta)$, we define $P_S^*(y|X_i^*, \theta) = P(y|X_i^*, \theta)$. When $\widehat{P}_S(y|X_i, \theta)$ is computed via simulation, new simulation draws are generated for each bootstrap sample (iid across bootstrap samples). In this case, $P_S^*(y|X_i^*, \theta)$ is defined just as $\widehat{P}_S(y|X_i, \theta)$ is defined, but with X_i replaced by X_i^* and with new simulated rv's. For example, if the crude frequency simulator is employed and ε_i does not depend on θ, then

$$P_S^*(y|X_i^*, \theta) = S^{-1} \sum_{s=1}^{S} [\varepsilon_i^*(s) \in \Omega(y, X_i^*, \theta)], \tag{45}$$

where $\{(\varepsilon_i^*(1), ..., \varepsilon_i^*(S)) : i = 1, ..., n\}$ denotes an iid sample of nS rv's each with the same distribution as ε_i and independent of $\{(\varepsilon_i(1), ..., \varepsilon_i(S)) : i = 1, ..., n\}$. We call $\{(\varepsilon_i^*(1), ..., \varepsilon_i^*(S)) : i = 1, ..., n\}$ the *bootstrap simulation rv's.*

Define $c_n^*(k, \gamma, \theta)$, $\gamma_{n,k,m}^*$, Γ_n^*, and $w_n^*(k, \gamma, \theta)$ as $\widehat{c}_n(k, \gamma, \theta)$, $\widehat{\gamma}_{n,k,m}$, $\widehat{\Gamma}_n$, and $\widehat{w}_n(k, \gamma, \theta)$ are defined, respectively, but using the bootstrap sample $\{(Y_i^*, X_i^*) : i = 1, ..., n\}$ and the bootstrap simulation rv's in place of the original sample $\{(Y_i, X_i) : i = 1, ..., n\}$ and the original simulation rv's, respectively.[8]

Define[9]

$$\begin{aligned} D_{n,U}^*(k, m) &= n^{1/2} \left(c_n^*(k, \gamma_{n,k,m}^*, \widehat{\theta}_{n,U}) - \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \widehat{\theta}_{n,U}) / w_n^*(k, \gamma_{n,k,m}^*, \widehat{\theta}_{n,U}) \right) \text{ and} \\ D_{n,L}^*(k, m) &= n^{1/2} \left(c_n^*(k, \gamma_{n,k,m}^*, \widehat{\theta}_{n,L}) - \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \widehat{\theta}_{n,L}) / w_n^*(k, \gamma_{n,k,m}^*, \widehat{\theta}_{n,L}) \right). \end{aligned} \tag{46}$$

Let $P^*(\cdot)$ denote probability with respect to the bootstrap sample conditional on the original sample.

Let $\lambda_{n,U}^*(k, m, \alpha)$ for $(k, m) \in \widehat{\mathcal{B}}_{n,U}$ and $\lambda_{n,L}^*(k, m, \alpha)$ for $(k, m) \in \widehat{\mathcal{B}}_{n,L}$ be non-negative constants (conditional on the original sample) such that

$$\begin{aligned} & P^* \Big(D_{n,U}^*(k, m) + \lambda_{n,U}^*(k, m, \alpha) \geq 0 \text{ for } (k, m) \in \widehat{\mathcal{B}}_{n,U} \text{ and} \\ & \qquad D_{n,U}^*(k, m) + \lambda_{n,L}^*(k, m, \alpha) \geq 0 \text{ for } (k, m) \in \widehat{\mathcal{B}}_{n,L} \Big) \\ & = 1 - \alpha \end{aligned} \tag{47}$$

[8] One must compute new bootstrap regressor sets, $\gamma_{n,k,m}^*$, that differ from the original sample regressor sets, $\widehat{\gamma}_{n,k,m}$, because the randomness in the sets $\widehat{\gamma}_{n,k,m}$ contributes to the asymptotic distribution of $n^{1/2}[\widetilde{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) - c_0(k, \gamma_{0,k,m}, \theta_{n,U})]$, where $\theta_{n,U}$ and $\theta_{n,L}$ are defined below. On the other hand, one does not need to compute new bootstrap dependent variable sets $\{\mathcal{Y}_{n,k}^* : k \in \mathcal{I}_{K_n^*}\}$ in place of $\{\widehat{\mathcal{Y}}_{n,k} : k \in \mathcal{I}_K\}$ nor new bootstrap binding constraint sets $\mathcal{B}_{n,U}^*$ and $\mathcal{B}_{n,L}^*$ in place of because $\widehat{\mathcal{B}}_{n,U}$ and $\widehat{\mathcal{B}}_{n,L}$ because with probability that goes to one as $n \to \infty$ the original sample sets $\{\widehat{\mathcal{Y}}_{n,k} : k \in \mathcal{I}_K\}$, $\widehat{\mathcal{B}}_{n,U}$, and $\widehat{\mathcal{B}}_{n,L}$ equal certain non-random sets that do not depend on n and, hence, are constant.

[9] Note that $(\gamma_{n,k,m}^*, \widehat{\theta}_{n,U})$ should appear as arguments of c_n^* and $(\widehat{\gamma}_{n,k,m}, \widehat{\theta}_{n,U})$ should appear as arguments of $\widehat{c}_n$ in the definition of $D_{n,U}^*(k, m)$ and analogously for $D_{n,L}^*(k, m)$. These are not typographical errors.

and same condition holds with U and L interchanged. In (47), the inequalities for $(k,m) \in \widehat{\mathcal{B}}_{n,U}$ ensure that the CI does not miss to the left of $\beta_{n,U}$ when the true parameter is $\beta_{n,U}$ and the inequalities for $(k,m) \in \widehat{\mathcal{B}}_{n,L}$ ensure that the CI does not miss to the right of $\beta_{n,U}$ when the true parameter is $\beta_{n,U}$. Condition (47) with U and L interchanged ensures analogous results hold when $\beta_{n,L}$ is the true value. Note that the inequalities in (47) plus the same inequalities with U and L interchanged do not have to hold simultaneously with P^*-probability $1-\alpha$, because the true value $\beta_0 = \beta_n(\theta_0)$ cannot equal both $\beta_{n,U}$ and $\beta_{n,L}$ (unless $\beta_{n,U} = \beta_{n,L}$ in which case the two conditions are the same asymptotically and, hence, hold simultaneously).

Condition (47) on $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ is enough to obtain the desired asymptotic results. But it does not uniquely define these quantities. For the latter, we add additional conditions. We add the conditions that the probability that an upper constraint is satisfied is the same for all upper constraints:

$$P^*\left(D^*_{n,U}(k,m) + \lambda^*_{n,U}(k,m,\alpha) \geq 0\right) = P^*\left(D^*_{n,U}(k',m') + \lambda^*_{n,U}(k',m',\alpha) \geq 0\right) \\ \text{for all } (k,m),(k',m') \in \widehat{\mathcal{B}}_{n,U}. \tag{48}$$

We also impose this condition with U replaced by L. The combination of (47) and (48) uniquely determines $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ because the total number of conditions equals the number of quantities to be determined (which equals the number of elements in $\widehat{\mathcal{B}}_{n,U} \cup \widehat{\mathcal{B}}_{n,L}$).

The bootstrap quantities $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ do not depend on any unknowns. Hence, in principle, they can be calculated analytically. In practice, however, they need to be calculated by simulation. To do so, let $\{D^*_{n,U,r}(k,m) : r = 1,...,R\}$ be R iid copies of $D^*_{n,U}(k,m)$ based on R iid copies of $\{(Y^*_i, X^*_i) : i = 1,...,n\}$ and $\{(\varepsilon^*_i(1),...,\varepsilon^*_i(S)) : i = 1,...,n\}$. Note that new simulation draws of the $\varepsilon_i(s)$ rv's are taken for each of the r bootstrap samples. Define $\{D^*_{n,L,r}(k,m) : r = 1,...,R\}$ and $\lambda^*_{n,L,R}(k,m,\alpha)$ analogously. Then, take $\lambda^*_{n,U,R}(k,m,\alpha)$ and $\lambda^*_{n,L,R}(k,m,\alpha)$ such that conditions (47) and (48) hold with $P^*(\cdot)$, $D^*_{n,U}(k,m)$, and $\lambda^*_{n,U}(k,m,\alpha)$ replaced by $R^{-1}\sum_{r=1}^R[\cdot]$, $D^*_{n,U,r}(k,m)$, and $\lambda^*_{n,U,R}(k,m,\alpha)$, respectively, where $[\cdot]$ denotes the indicator function.

The values $\lambda^*_{n,U,R}(k,m,\alpha)$ and $\lambda^*_{n,L,R}(k,m,\alpha)$ approximate $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ arbitrarily closely for R sufficiently large. (For this reason, in our analysis of the asymptotic properties of the CI we assume that $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$, which satisfy (47), are known.[10])

In sum, one computes the CI $CI_n(1-\alpha)$ by calculating the following quantities in the following order: $\{\widehat{\mathcal{Y}}_{n,k} : k = 1,...,\widehat{K}_n\}$, $\widehat{\Gamma}_n$, $\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta)$, $\widehat{\Theta}_n$, $\widehat{\beta}_{n,U}$, $\widehat{\theta}_{n,U}$, $\widehat{\mathcal{B}}_{n,U}$, $\widehat{\beta}_{n,L}$, $\widehat{\theta}_{n,L}$, $\widehat{\mathcal{B}}_{n,L}$, $\widehat{w}_n(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,U})$, $\widehat{w}_n(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,L})$, $\lambda^*_{n,U,R}(k,m,\alpha)$ for $(k,m) \in \widehat{\mathcal{B}}_{n,U}$, $\lambda^*_{n,L,R}(k,m,\alpha)$ for $(k,m) \in \widehat{\mathcal{B}}_{n,L}$, $\widetilde{c}_{n,U}(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,U})$, $\widetilde{c}_{n,L}(k,\widehat{\gamma}_{n,k,m},\widehat{\theta}_{n,L})$, $\widetilde{\beta}_{n,U}$, and $\widetilde{\beta}_{n,L}$. The computation of the bootstrap critical values $\lambda^*_{n,U,R}(k,m,\alpha)$ and $\lambda^*_{n,L,R}(k,m,\alpha)$ involves computing the

[10] This is consistent with the extensive literature on the bootstrap, which, for the same reason, analyzes the properties of the "ideal bootstrap" rather than the bootstrap based on a finite, but large, number of bootstrap repetitions.

closed form quantities $\{(D^*_{n,U,r}(k,m), D^*_{n,L,r}(k,m)) : r = 1,...,R\}$ based on R iid copies of $\{(Y_i^*, X_i^*) : i = 1,...,n\}$ and the simulation rv's (if simulation is used to compute the model probabilities).

4.3 Confidence Interval for the Identified β Set

In this section, we a construct CI for the identified set of values of $\beta_n(\theta)$. Let

$$\begin{aligned} B_0 &= \{\beta_n(\theta) \in R : \theta \in \Theta_0\} \text{ and} \\ B_+ &= \{\beta_n(\theta) \in R : \theta \in \Theta_+\}. \end{aligned} \tag{49}$$

That is, B_0 and B_+ are the sets of $\beta_n(\theta)$ values that are consistent with the necessary conditions for an equilibrium (19) and (20), respectively. For notational simplicity, we do not make the dependence of B_0 and B_+ on n explicit.

We construct a CI of the same form as in (37), i.e., $CI_n(1-\alpha) = [\widetilde{\beta}_{n,L}, \widetilde{\beta}_{n,U}]$. But, for the set B_+ rather than for the true value β_0. The CI is defined such that

$$\liminf_{n\to\infty} P(B_+ \subset CI_n(1-\alpha)) \geq 1-\alpha. \tag{50}$$

Because $B_0 \subset B_+$, this CI also is valid for B_0, the identified set of values of $\beta_n(\theta)$.

The upper and lower bounds, $\widetilde{\beta}_{n,U}$ and $\widetilde{\beta}_{n,L}$, are the same as in (39)-(44), but the definition of the bootstrap critical values $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ is slightly different. The bootstrap critical values are non-negative constants (conditional on the original sample) such that

$$\begin{aligned} P^*\Big(&D^*_{n,U}(k,m) + \lambda^*_{n,U}(k,m,\alpha) \geq 0 \text{ for } (k,m) \in \widehat{\mathcal{B}}_{n,U} \text{ and} \\ &D^*_{n,L}(k,m) + \lambda^*_{n,L}(k,m,\alpha) \geq 0 \text{ for } (k,m) \in \widehat{\mathcal{B}}_{n,L}\Big) \\ = {}& 1-\alpha. \end{aligned} \tag{51}$$

The difference between (51) and (47) is that the inequalities for $(k,m) \in \mathcal{B}_{+,L}$ are evaluated at $\theta_{+,L}$ in (51) whereas they are evaluated at $\theta_{+,U}$ in (47). Furthermore, (51) specifies two inequalities, but (51) specifies only one—interchanging U and L in (51) does not change the condition. In (51), the inequalities for $(k,m) \in \widehat{\mathcal{B}}_{n,U}$ ensure that the CI does not miss to the left of $\beta_{n,U}$ when the true parameter is $\beta_{n,U}$ and the inequalities for $(k,m) \in \widehat{\mathcal{B}}_{n,L}$ ensure that it does not miss to the right of $\beta_{n,L}$ when the true parameter is $\beta_{n,L}$.

As in the previous section, condition (51) is enough to obtain the desired asymptotic results, but it does not uniquely define $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$. To do the latter, we add the conditions in (48) and the following condition that leads to an equal-tailed CI:

$$\begin{aligned} & P^*\left(D^*_{n,U}(k,m) + \lambda^*_{n,U}(k,m,\alpha) \geq 0 \text{ for } (k,m) \in \widehat{\mathcal{B}}_{n,U}\right) \\ = {}& P^*\left(D^*_{n,L}(k,m) + \lambda^*_{n,L}(k,m,\alpha) \geq 0 \text{ for } (k,m) \in \widehat{\mathcal{B}}_{n,L}\right). \end{aligned} \tag{52}$$

The combination of conditions in (51), (48), and (52) uniquely determines $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$.

5 Test of Model Specification

A test of model specification is obtained by constructing a confidence region (CR), denoted $CR_n(1-\alpha)$, for the whole parameter vector θ. One rejects the parametric model specification if the confidence region is the null set because such a CR indicates that no parameter value is consistent with the inequality constraints. The CR is defined such that the probability that it contains the true value θ_0 is greater than or equal to $1-\alpha$ asymptotically for $\alpha > 0$. Hence, the CR and corresponding test of model specification may be conservative. The CR considered here also has the drawback that it is noticeably more difficult to compute than the CIs introduced above because it requires a maximization over the estimated set $\widehat{\Theta}_n$ for each bootstrap repetition.

The confidence region for θ is constructed as follows:

$$\begin{aligned} CR_n(1-\alpha) &= \{\theta \in \Theta : \widetilde{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_n\}, \text{ where} \\ \widetilde{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) &= \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) + \widehat{w}_n(k, \gamma, \theta)\lambda_n^*(k, m, \alpha)/\sqrt{n}, \end{aligned} \tag{53}$$

$\widehat{\mathcal{B}}_n$ is the subset of $\mathcal{I}_{K,M}$ that contains the constraints that are binding at some boundary point of $\widehat{\Theta}_n$, i.e., $\widehat{\mathcal{B}}_n = \{(k,m) \in \mathcal{I}_{K,M} : \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta) = 0 \text{ for some } \theta \in \Theta\}$, and where $\lambda_n^*(k,m,\alpha)$ is a non-negative critical value.obtained using the bootstrap.

Define

$$D_n^*(k,m,\theta) = n^{1/2}\left(c_n^*(k, \gamma_{n,k,m}^*, \theta) - \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta)/w_n^*(k, \gamma_{n,k,m}^*, \theta)\right). \tag{54}$$

Let $\lambda_n^*(k,m,\theta,\alpha)$ for $(k,m) \in \widehat{\mathcal{B}}_n$ and $\theta \in \widehat{\Theta}_n$ be non-negative constants that satisfy:

$$P^*\left(D_n^*(k,m,\theta) + \lambda_n^*(k,m,\theta,\alpha) \geq 0 \text{ for } (k,m) \in \widehat{\mathcal{B}}_n\right) = 1-\alpha. \tag{55}$$

Define

$$\lambda_n^*(k,m,\alpha) = \max_{\theta \in \widehat{\Theta}_n} \lambda_n^*(k,m,\theta,\alpha). \tag{56}$$

The bootstrap critical values $\lambda_n^*(k,m,\alpha)$ are computed by bootstrap simulation. Due to the maximization over $\theta \in \widehat{\Theta}_n$, these critical values are much more difficult to compute than those needed for the CIs introduced above.

The CR is designed to contain the true value θ_0 with asymptotic probability at least $1-\alpha$:

$$\liminf_{n\to\infty} P(\theta_0 \subset CR_n(1-\alpha)) \geq 1-\alpha. \tag{57}$$

6 Asymptotics for CIs

6.1 High-level Assumptions

In this subsection, we state the high-level assumptions that are used to justify the CIs introduced in the previous section. These assumptions are verified for leading choices of the model probabilities $\{P(y|x,\theta):\theta\in\Theta\}$, sets $\widehat{\Gamma}_n$, and weight function $\widehat{w}_n(k,\gamma,\theta)$ below. THIS IS NOT DONE YET.

Let

$$\begin{aligned}\widehat{\nu}_n(k,\gamma,\theta) &= \sqrt{n}\left(\widehat{c}_n(k,\gamma,\theta)-c_0(k,\gamma,\theta)\right) \text{ and}\\ \widehat{Z}_n(k,m,\theta) &= \sqrt{n}\left(c_0(k,\widehat{\gamma}_{n,k,m},\theta)-c_0(k,\gamma_{0,k,m},\theta)\right).\end{aligned} \tag{58}$$

Viewed as a function of (k,γ,θ), $\widehat{\nu}_n(k,\gamma,\theta)$ is a stochastic process on $\mathcal{I}_K\times\Gamma_{all}\times\Theta$. Under suitable conditions, $\widehat{\nu}_n(\cdot,\cdot,\cdot)$ converges weakly to a mean zero Gaussian process $\nu_0(\cdot,\cdot,\cdot)$ on $\mathcal{I}_K\times\Gamma_{all}\times\Theta$ as $n\to\infty$. The covariance function of $\nu_0(\cdot,\cdot,\cdot)$ is

$$\begin{aligned}&V_0((k_1,\gamma_1,\theta_1),(k_2,\gamma_2,\theta_2))\\ &= Cov(\nu_0(k_1,\gamma_1,\theta_1),\nu_0(k_2,\gamma_2,\theta_2)),\\ &= E\left(\left(\widehat{P}_S(\mathcal{Y}_{k_1}|X_i,\theta_1)-[Y_i\in\mathcal{Y}_{k_1}]\right)h_{\gamma_1}(X_i)-c_0(k_1,\gamma_1,\theta_1)\right)\\ &\quad\times\left(\left(\widehat{P}_S(\mathcal{Y}_{k_2}|X_i,\theta_2)-[Y_i\in\mathcal{Y}_{k_2}]\right)h_{\gamma_2}(X_i)-c_0(k_2,\gamma_2,\theta_2)\right).\end{aligned} \tag{59}$$

Let $\widehat{Z}_n(\theta)$ denote the column vector whose elements are $\{\widehat{Z}_n(k,m,\theta):(k,m)\in\mathcal{I}_{K,M}\}$ with first element $\widehat{Z}_n(1,1,\theta)$, second element $\widehat{Z}_n(1,2,\theta)$, etc. The dimension of $\widehat{Z}_n(\theta)$ is $\sum_{k=1}^K M_k$.

Let $\Rightarrow$ denote weak convergence of a sequence of stochastic processes.

We make the following high-level assumptions:

Assumption CI1. $\widehat{\nu}_n(\cdot,\cdot,\cdot)\Rightarrow\nu_0(\cdot,\cdot,\cdot)$, where $\nu_0(\cdot,\cdot,\cdot)$ is a mean zero Gaussian process indexed by $(k,\gamma,\theta)\in\mathcal{I}_K\times\Gamma_{all}\times\Theta$ with bounded and continuous sample paths a.s. (with respect to $||\cdot||_G$ on Γ_{all} and the Euclidean norm on Θ) and with covariance function $V_0(\cdot,\cdot,\cdot)$ defined in (59).

Assumption CI2. $\widehat{Z}_n(\cdot)\to_d Z_0(\cdot)$, where $Z_0(\cdot)$ is a mean zero Gaussian process indexed by $\theta\in\Theta$ with bounded and continuous sample paths a.s. (with respect to the Euclidean norm on Θ) and the convergence holds jointly with that in Assumption CI1 with the joint limit being Gaussian.

Assumption CI3. (a) $\sup_{(k,\gamma,\theta)\in\mathcal{I}_K\times\Gamma_{all}\times\Theta}|\widehat{w}_n(k,\gamma,\theta)-w_0(k,\gamma,\theta)|\to_p 0$ for some non-random positive function w_0 on $\mathcal{I}_K\times\Gamma_{all}\times\Theta$ that is bounded and bounded away from zero.
(b) $w_0(k,\gamma,\theta)$ is continuous in (γ,θ) (with respect to the product of the $||\cdot||_G$ semi-norm on Γ_{all} and the Euclidean norm on Θ) at $(\gamma_{0,k,m},\theta)$ $\forall\theta\in\Theta$, $\forall(k,m)\in\mathcal{I}_{K,M}$.

Assumption CI3 is verified below for $\widehat{w}_n(k,\gamma,\theta)$ as defined in (41).

Next, we define what are essentially population analogues of $\widehat{\theta}_{n,U}$ and $\widehat{\theta}_{n,L}$. Define $\theta_{n,U}$ and $\theta_{n,L}$ just as $\widehat{\theta}_{n,U}$ and $\widehat{\theta}_{n,L}$ are defined in (43) and (44), but with Θ_+ and $\beta_{n,U}$ in place of $\widehat{\Theta}_n$ and $\widehat{\beta}_{n,U}$, respectively. Thus,

$$\theta_{n,U}=\arg\min\{||\theta||:\theta\in\Theta_+,\beta_n(\theta)=\beta_{n,U}\} \tag{60}$$

and $\theta_{n,L}$ is defined analogously with $\beta_{n,L}$ in place of $\beta_{n,U}$.[11]

Assumption CI4. (a) $\beta_n(\cdot) \to_p \beta_0(\cdot)$ uniformly over $\theta \in \Theta$ for some non-random continuous function $\beta_0(\cdot)$ on Θ.
(b) $\theta_{n,U} \to_p \theta_{+,U}$ and $\theta_{n,L} \to_p \theta_{+,L}$, where

$$\begin{aligned} \theta_{+,U} &= \arg\min\{||\theta|| : \theta \in \Theta, \beta_0(\theta) = \beta_{+,U}\}, \\ \beta_{+,U} &= \sup\{\beta_0(\theta) : \theta \in \Theta_+\}, \end{aligned}$$

and $(\theta_{+,L}, \beta_{+,L})$ are defined analogously with sup replaced by inf in the definition of $\beta_{+,L}$.

As defined, $\beta_{+,U}$ is the upper bound on the values of the asymptotic function $\beta_0(\cdot)$ over points θ that satisfy the necessary conditions (20) for profit maximization. Note that $\beta_{+,U}$ has the same definition as $\beta_{n,U}$ but with $(\beta_n(\theta), \widehat{c}_n, \widehat{\gamma}_{n,k,m})$ replaced by their probability limits $(\beta_0(\theta), c_0, \gamma_{0,k,m})$, respectively.

In the leading case in which $\beta_n(\cdot)$ is non-random and does not depend on n, Assumption CI4 automatically holds with $\beta_0(\cdot) = \beta_n(\cdot)$, $\theta_{+,U} = \theta_{n,U}$, and $\theta_{+,L} = \theta_{n,L}$ (none of which depend of n). In addition, Assumption CI2 can be replaced by the following assumption.

Assumption CI2′. $(\widehat{Z}_n(\theta_{+,U})', \widehat{Z}_n(\theta_{+,L})')' \to_d (Z_0(\theta_{+,U})', Z_0(\theta_{+,L})')'$ for some multivariate normal vector $(Z_0(\theta_{+,U})', Z_0(\theta_{+,L})')'$ and the convergence holds jointly with that in Assumption CI1 with the joint limit being Gaussian.

We assume that the random binding constraint sets $\widehat{\mathcal{B}}_{n,U}$ and $\widehat{\mathcal{B}}_{n,L}$ are contained in population analogues of these sets, $\mathcal{B}_{n,U}$ and $\mathcal{B}_{n,L}$, respectively, which in turn are contained in non-random asymptotic analogues of these sets, $\mathcal{B}_{+,U}$ and $\mathcal{B}_{+,L}$, with probability that goes to one as $n \to \infty$. (In fact, typically, $\widehat{\mathcal{B}}_{n,U} = \mathcal{B}_{n,U} = \mathcal{B}_{+,U}$ with probability that goes to one and likewise with L in place of U, but that is not needed for the results.) By definition,

$$\begin{aligned} \mathcal{B}_{n,U} &= \{(k,m) \in \mathcal{I}_{K,M} : c_0(k, \gamma_{0,k,m}, \theta_{n,U}) = 0\}, \\ \mathcal{B}_{+,U} &= \{(k,m) \in \mathcal{I}_{K,M} : c_0(k, \gamma_{0,k,m}, \theta_{+,U}) = 0\}, \end{aligned} \qquad (61)$$

and $\mathcal{B}_{n,L}$ and $\mathcal{B}_{+,L}$ are defined analogously with $\theta_{n,L}$ and $\theta_{+,L}$ in place of $\theta_{n,U}$ and $\theta_{+,U}$, respectively. If $\beta_n(\cdot)$ is non-random and does not depend on n, then $\mathcal{B}_{n,U} = \mathcal{B}_{+,U}$ and $\mathcal{B}_{n,L} = \mathcal{B}_{+,L}$.

The estimators $\widehat{\beta}_{n,U}$ and $\widehat{\beta}_{n,L}$ and the binding constraint sets $\widehat{\mathcal{B}}_{n,U}$ and $\widehat{\mathcal{B}}_{n,L}$ are assumed to satisfy:

Assumption CI5. (a) $\widehat{\beta}_{n,U} - \beta_{n,U} \to_p 0$ and $\widehat{\beta}_{n,L} - \beta_{n,L} \to_p 0$.
(b) $P(\widehat{\mathcal{B}}_{n,U} \subset \mathcal{B}_{n,U} \subset \mathcal{B}_{+,U}) \to 1$ and $P(\widehat{\mathcal{B}}_{n,L} \subset \mathcal{B}_{n,L} \subset \mathcal{B}_{+,L}) \to 1$.

Conditions under which Assumption CI5(a) holds are given in the preceding section. Assumption CI5(b) allows the estimated binding constraints sets $\widehat{\mathcal{B}}_{n,U}$ and $\widehat{\mathcal{B}}_{n,L}$ to be smaller than the population versions $\mathcal{B}_{n,U}$ and $\mathcal{B}_{n,L}$. (The use of fewer constraints in constructing the CI cannot decrease the length of the CI and, hence, cannot reduce its coverage probability.)

[11] Footnote 1 applies here and in the definition of $\theta_{+,U}$ immediately below if the arg min is not unique.

This has important consequences for the flexibility of choosing Y_i and X_i cells. It allows one to consider more constraints in the set estimation stage than in the CI construction stage. One might do this because one is "fishing" for good cells at the set estimation stage.

Next, we assume that the critical values $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ converge in probability to non-negative constants $\lambda_{0,U}(k,m,\alpha)$ and $\lambda_{0,L}(k,m,\alpha)$ that satisfy the following condition when employing the CI for β_0, the true value:

$$\begin{aligned}
&P\,(\nu_0(k,\gamma_{0,k,m},\theta_{+,U}) + Z_0(k,m,\theta_{+,U}) \\
&\quad +w_0(k,\gamma_{0,k,m},\theta_{+,U})\lambda_{0,U}(k,m,\alpha) \geq 0 \text{ for } (k,m)\in\mathcal{B}_{+,U} \text{ and} \\
&\qquad \nu_0(k,\gamma_{0,k,m},\theta_{+,U}) + Z_0(k,m,\theta_{+,U}) \\
&\quad +w_0(k,\gamma_{0,k,m},\theta_{+,U})\lambda_{0,L}(k,m,\alpha) \geq 0 \text{ for } (k,m)\in\mathcal{B}_{+,L}) \\
&= 1-\alpha \qquad (62)
\end{aligned}$$

and the same condition holds with U and L interchanged.

For the CI for the identified interval B_+, (62) is replaced by the following condition, which reflects the definition of $\lambda^*_{n,U}(k,m,\alpha)$ and $\lambda^*_{n,L}(k,m,\alpha)$ in (51), rather than in (47):

$$\begin{aligned}
&P\,(\nu_0(k,\gamma_{0,k,m},\theta_{+,U}) + Z_0(k,m,\theta_{+,U})+ \\
&\quad w_0(k,\gamma_{0,k,m},\theta_{+,U})\lambda_{0,U}(k,m,\alpha) \geq 0 \text{ for } (k,m)\in\mathcal{B}_{+,U} \text{ and} \\
&\qquad \nu_0(k,\gamma_{0,k,m},\theta_{+,L}) + Z_0(k,m,\theta_{+,L})+ \\
&\quad w_0(k,\gamma_{0,k,m},\theta_{+,L})\lambda_{0,L}(k,m,\alpha) \geq 0 \text{ for } (k,m)\in\mathcal{B}_{+,L}) \\
&= 1-\alpha. \qquad (63)
\end{aligned}$$

We assume:

Assumption CI6. $\lambda^*_{n,U}(k,m,\alpha) \to_p \lambda_{0,U}(k,m,\alpha) \geq 0$ for all $(k,m)\in\mathcal{B}_{+,U}$ and likewise with U replaced by L, where $\lambda_{0,U}(k,m,\alpha)$ and $\lambda_{0,L}(k,m,\alpha)$ satisfy (62) when considering the CI for the true value β_0 and satisfy (63) when considering the CI for the identified interval B_+.

The asymptotic critical values $\lambda_{0,U}(k,m,\alpha)$ and $\lambda_{0,L}(k,m,\alpha)$ typically also satisfy an analogue of (48) (and (52) when considering the CI for the identified set) with $P^*(\cdot)$, $D^*_{n,U}(k,m)$, $\widehat{\mathcal{B}}_{n,U}$, etc. replaced by $P(\cdot)$, $\nu_0(k,\gamma_{0,k,m},\theta_{+,U}) + Z_0(k,m,\theta_{+,U})$, $\mathcal{B}_{+,U}$, etc., respectively. But this is not necessary for the asymptotic results, so it is not included in Assumption CI6.

NEED ADD SUFFICIENT CONDITIONS FOR ASSUMPTION CI6, I.E., THE BOOTSTRAP RESULTS. WILL NEED CONDITIONS OF THE FOLLOWING SORT:

For example, to get $\widehat{Z}_n(k,m,\theta)$ to be well-behaved when evaluated at $\widehat{\theta}_{n,U}$ and $\widehat{\theta}_{n,L}$, rather than $\theta_{+,U}$ and $\theta_{+,L}$, we may need:

(i) $P(\mathcal{Y}_i|x,\theta)$ is partially differentiable in θ on a neighborhood of $\theta_{+,U}$ and $\theta_{+,L}$ for all $x\in\mathcal{X}$ and all $k\in\mathcal{I}_K$.

(ii) $n^{1/2}(\widehat{\theta}_{n,U}-\theta_{n,U}) = O_p(1)$ and $n^{1/2}(\widehat{\theta}_{n,L}-\theta_{n,L}) = O_p(1)$.

(iii) $\int \delta_k(x)\left(h_{\widehat{\gamma}_{n,k,n}}(x) - h_{\gamma_{0,k,n}}(x)\right) dG(x) \to_p 0 \; \forall (k,m)\in\mathcal{I}_{K,M}$,
where $\delta_k(x) = \sup_{\theta\in S(\theta_{+,U},\varepsilon)\cup S(\theta_{+,L},\varepsilon)} ||\frac{\partial}{\partial\theta}P(\mathcal{Y}_i|x,\theta)||$.

6.2 Asymptotic Justification of the CIs

The main result of this section is the following:

Theorem 3 *Under Assumptions* CI1-CI6, 2, 3(a), *and* 6, (38) *holds for the CI for the true value* β_0 *and* (50) *holds for the CI for the identified interval* B_+.

ADD COMMENT RE $\widetilde{\beta}_{n,U} - \beta_{n,U} \to_p 0$ AND $\widetilde{\beta}_{n,L} - \beta_{n,L} \to_p 0$ UNDER AS. CI1-CI6 USING LEMMA BELOW PARTS (a) AND (b).

TO SHOW THAT THE CI'S ARE NOT LONGER THAN NECESSARY ASYMPTOTICALLY, WE NEED ANOTHER RESULT THAT SHOWS THAT $\widetilde{\beta}_{n,U} - \beta_{n,U} \to_p 0$ AND $\widetilde{\beta}_{n,L} - \beta_{n,L} \to_p 0$. THIS REQUIRES THAT

(i) Assumption CI5 holds. I.e., $P\left(\widehat{\mathcal{B}}_{n,U} = \mathcal{B}_{+,U}\right) \to 1$ and $P\left(\widehat{\mathcal{B}}_{n,L} = \mathcal{B}_{+,L}\right) \to 1$ AND

(ii) $\sup\{\beta_n(\theta) : \theta \in \Theta \ \&\ c_0(k, \gamma_{0,k,m}, \theta) \geq 0 \ \forall (k,m) \in \mathcal{B}_{+,U}\}$

$= \sup\{\beta_n(\theta) : \theta \in \Theta \ \&\ c_0(k, \gamma_{0,k,m}, \theta) \geq 0 \ \forall (k,m) \in \mathcal{I}_{K,M}\}$ AND THE SAME WITH U REPLACED BY L.

ACTUALLY, THE CONVERGENCE IN PROBABILITY RESULTS $\widetilde{\beta}_{n,U} - \beta_{n,U} \to_p 0$ AND $\widetilde{\beta}_{n,L} - \beta_{n,L} \to_p 0$ ONLY REQUIRE CONSISTENCY OF $\widehat{\beta}_{n,U}$ AND $\widehat{\beta}_{n,L}$ BECAUSE THE CI ADJUSTMENT FACTOR IS $O_p(n^{-1/2})$! SO, DON'T NEED CONDITIONS (i) AND (ii) ABOVE.

7 Simple Monte-Carlo Examples

To illustrate how the method works, we start with a very simple example that is easy to exposit and graph. The example has no x variables at all. We then consider a much more realistic example. We begin with a nearly trivial two-parameter example that lends itself to a simple graphical exposition. We also provide some Monte Carlo results for a more complicated four parameter example. The two-parameter example would be point identified if we exploited the model fully, but we ignore this in the example. The four-parameter version is not point identified unless we make use of an assumption on the equilibrium selection mechanism.

To make the examples of this section more interesting, we note that these examples without x variables could be considered in terms of pointwise (in x) non-parametric set identification. In that case, we would need to substitute in non-parametric estimates of the sample events at each x.

7.1 A Two Parameter Model

Consider a symmetric two-firm entry model with random profits. Both firms make a simultaneous entry decision. If firm j ($j = 1, 2$) does not enter, it earns zero profit. If it enters while its rival does not, it gets a monopoly profit of

$$\Pi_j = \alpha + \varepsilon_j,$$

where ε_j is i.i.d. standard normal across both firms and markets. If firm j enters and its rival also enters, it gets a duopoly profit of

$$\Pi_j = \beta + \varepsilon_j.$$

We assume that competition reduces a firm's profit, i.e., $\alpha > \beta$.

Without any x's, the distributional assumption on ε serves only to map α and β into the probabilities of being profitable as a monopolist and as a duopolist. For simplicity, then, we treat these probabilities as the parameters of interest. Let μ denote the probability of being profitable as a monopolist, i.e. $\mu = Pr(\varepsilon > -\alpha) = \Phi(\alpha)$. Let δ denote the probability of being profitable as a duopolist, i.e. $\delta = Pr(\varepsilon > -\beta) = \Phi(\beta)$. In this section, we simulate the model, estimate the set of μ and δ defined by the Nash equilibrium conditions, and provide a confidence region for that set.

The multiple equilibria problem is discussed for more complicated versions of this example in Berry (1989), Bresnahan and Reiss (1991), Berry (1992), and Tamer (2003). The condition for the one-firm equilibrium is that at least one firm makes a profit in the monopoly case and that at least one firm does not make a profit in the duopoly case. If both firms make a positive monopoly profit, but neither makes a positive duopoly profit, then there are two equilibria in which one or the other firm (but not both) enters.

In the present example, the literature cited above points out that in spite of the potential multiple equilibria one can base an estimation routine on the total number of firms in the market (and in fact identify μ and δ exactly). But, for the present section, we ignore that strategy and make use only of the Nash equilibrium conditions for the firms.

There are four events: both firms out, one firm in, the other firm in, and both firms in. Denote these as (0,0), (0,1), (1,0), and (1,1). When the number of firms is equal to zero, the necessary condition is that both firms are unprofitable in a monopoly and the probability of this event is $(1-\mu)^2$. The other probabilities are found in a similar way and give the c_0 conditions from (18) as

$$\begin{aligned}
c_0(Y = (0,0), \mu, \delta) &\equiv (1-\mu)^2 - P_0(0,0), \\
c_0(Y = (0,1), \mu, \delta) &\equiv \mu(1-\delta) - P_0(0,1), \\
c_0(Y = (1,0), \mu, \delta) &\equiv \mu(1-\delta) - P_0(1,0), \text{ and} \\
c_0(Y = (1,1), \mu, \delta) &\equiv \delta^2 - P_0(1,1),
\end{aligned} \tag{64}$$

where the P_0's denote the true event probabilities. The identified set Θ_0 consists of those (μ, δ) such that all four conditions are non-negative.[12]

The sample analog conditions, the $\widehat{c}_n$'s from (18), simply replace the population probabilities with sample frequencies in this case:

$$\widehat{c}_n(Y = (0,0), \mu, \delta) = (1-\mu)^2 - \widehat{P}_n(0,0),$$

[12] Again, a competent theorist would point out that the first and last conditions are in fact both necessary and sufficient (there are no multiple equilibria associated with those outcomes) and so those two conditions in fact hold with equality and identify the parameters. To illustrate the method, we ignore this information.

$$\begin{aligned}\widehat{c}_n(Y=(0,1),\mu,\delta) &= \mu(1-\delta)-\widehat{P}_n(0,1),\\ \widehat{c}_n(Y=(1,0),\mu,\delta) &= \mu(1-\delta)-\widehat{P}_n(1,0), \text{ and}\\ \widehat{c}_n(Y=(1,1),\mu,\delta) &= \delta^2-\widehat{P}_n(1,1). \qquad (65)\end{aligned}$$

The set estimator is the set of (μ,δ) such that all four sample conditions are non-negative. If there are no such parameters, the estimate is then the parameter vector that minimizes the sum of squared $\widehat{c}_n$'s that are negative. In this subsection, we assume that we do not observe the identities of the two different firms, but only the number of firms and so we re-write the two middle constraints as

$$\mu(1-\delta) > \frac{1}{2}(\widehat{P}_n(0,1)+\widehat{P}_n(1,0)). \qquad (66)$$

For the Monte Carlo exercise, we assume $P_0(0,0)=0.1225$, $P_0(0,1)=P_0(1,0)=0.35875$, $P_0(1,1)=0.16$, $\mu=0.65$, and $\delta=0.4$. (The implied parameters of the profit function are $\alpha=0.385$ and $\beta=-0.253$.) We first generate a random sample of Y with 500 observations using the multinomial distribution specified by the probability vector $(P_0(0,0),P_0(0,1),P_0(1,0),P_0(1,1))$.[13]

Figure 3 displays, in the space of the parameters μ and δ, the population constraints and an example of the sample analog of those constraints. The solid lines are the three constraints and the identified set is the roughly triangular region in formed by those lines. For example, the horizontal line graphs the condition $c_0(Y=(0,0),\mu,\delta)=0$, which has the solution $\mu=\sqrt{P_0(Y=(0,0))}$. The vertical line is the $Y=(1,1)$ condition and the remaining curve is the one-firm condition.

The dashed (blue) lines are constructed from a single 500 observation draw on the true data process. The sample analog estimate is the roughly triangular region in the center of the dashed lines.

The true parameter vector is in fact at the upper-left intersection of the horizontal and vertical lines. Indeed, this problem would be point identified if we allowed ourselves to notice that fact (which we ignore only for purposes of exposition, since it does not carry over to more complicated models.)

For this very simple example, we can in fact bootstrap a single parameter confidence region, because the intersections of the constraints are trivial to compute. The top panel of Figure (4) shows a set of bootstrapped event probabilities, which then translate into the clouds of intersection points for the upper $(\bar{\mu})$ and lower $(\underline{\mu})$ bounds for μ. The horizontal lines cut off approximately 2.5% of the top and bottom points, creating an approximately 95% confidence region for the identified single-parameter interval.

¿From the viewpoint of constructing CIs with the right coverage ratio, an important bit of intuition is that the bootstrap CIs (which should have the correct asymptotic coverage ratios) are constructed from horizontal lines in the parameter space.

When bootstrapping the set estimates is too computationally difficult, we can turn to our CI method. The first idea is to return to Figure 3 and move the original constraints out.

[13] When we create the data, in the region of multiple equilibria we assume that each equilibrium occurs with equal probability, but in the estimation exercise, we treat the equilibrium selection rule as being unknown.

Figure 3: Identified and Estimated Region for 2 Parameter Example

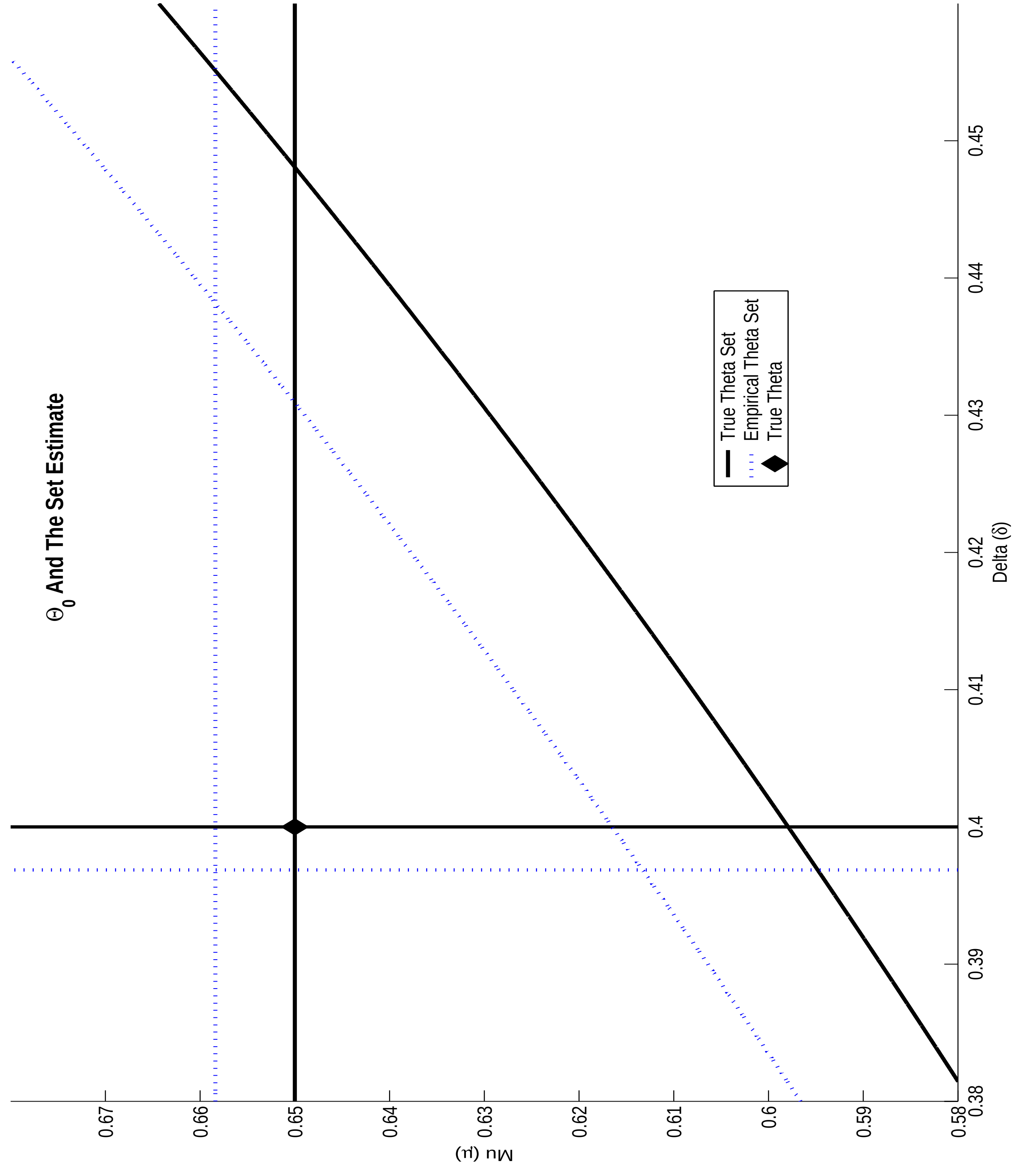
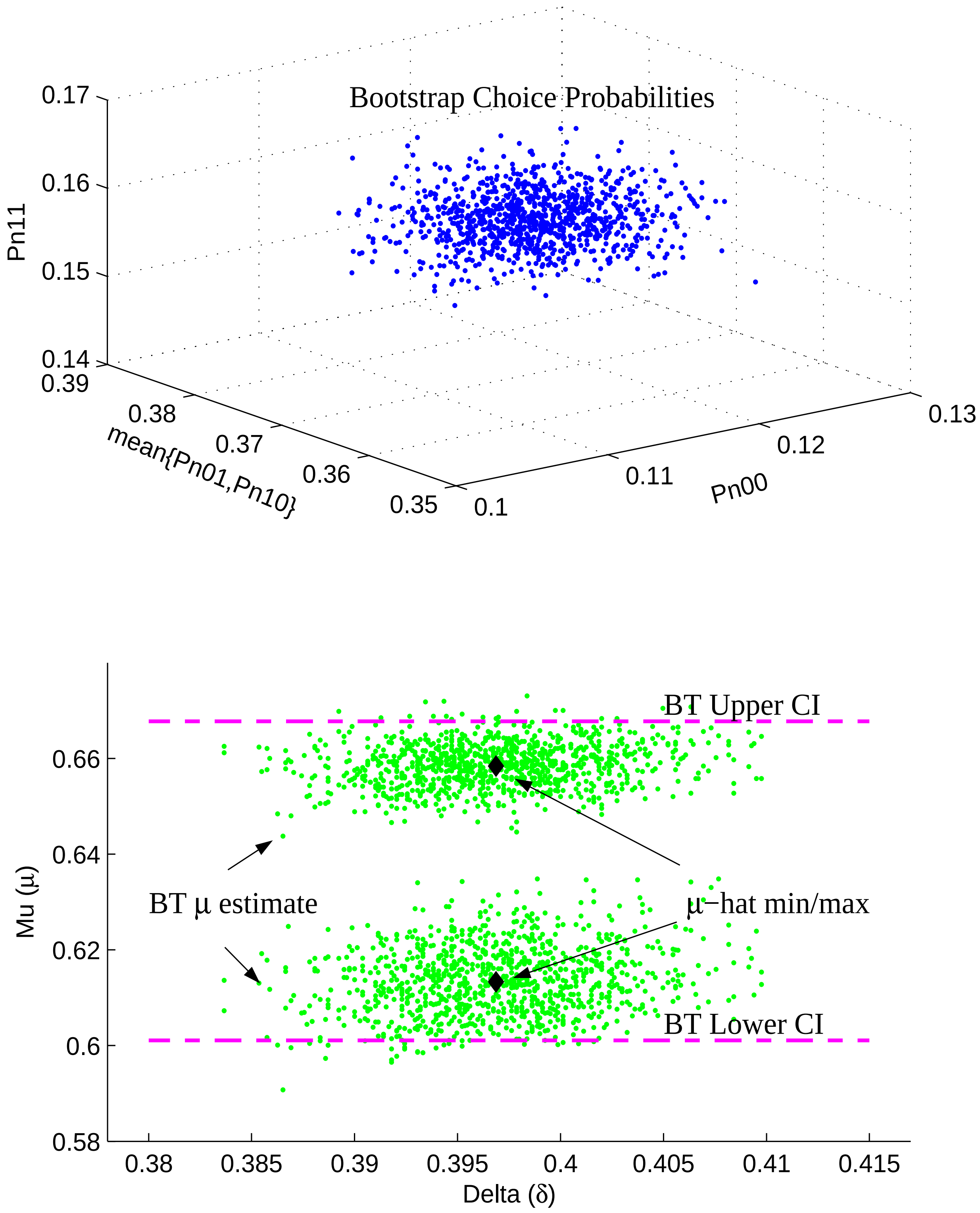


Figure 4: Bootstrapped CI for μ



Let us consider the confidence region for μ. The upper bound is defined by a horizontal line, so moving that line is very much like the bootstrap. The lower bound, $\underline{\mu}$, is defined by the intersection of the other two curves. If we add no additional constraints, the CI is defined by

$$\begin{aligned}
\widetilde{c}_n(Y = (0,0)) &= (1-\bar{\mu})^2 - \widehat{P}_{n,i}(0,0) + w(0,0)_n \lambda/\sqrt{n}, && (67)\\
\widetilde{c}_n(Y = (0,1) \text{ or } (1,0)) &= \underline{\mu}(1-\underline{\delta}) - \frac{1}{2}(\widehat{P}_{n,i}(0,1) + \widehat{P}_{n,i}(1,0)) + w(1,0)_n \lambda/\sqrt{n} && (68)\\
\widetilde{c}_n(Y = (1,1)) &= \underline{\delta}^2 - \widehat{P}_{n,i}(1,1) + w(1,1)_n \lambda/\sqrt{n} && (69)\\
& && (70)
\end{aligned}$$

where the $P_{n,i}$ are the bootstrapped sample probabilities and the w_n's are again estimates of the standard deviations of the original estimated constraints (used to scale the movement of the constraints.) In the simplest case, we choose a single λ so that the minimum of the three constraints is above zero for 95% of the bootstrapped probabilities. We could also choose different λ for the constraint defining $\bar{\mu}$ and $\underline{\mu}$, as long as the resulting number of excluded points is correct.

However, the resulting $\widetilde{c}_n$ do not provide a very tight lower bound on μ. Much of the excluded points rule out alternative values of δ rather than alternative values of μ. To help with this, we can construct a new, nearly "flat" constraint as the positive sum of the two original constraints defining $\underline{\mu}$:

$$\widetilde{c}_n^* \;=\; \widehat{c}_n(Y = (0,1) \text{ or } (1,0)) \;+\; \alpha \widehat{c}_n(Y = (1,1)) + \lambda^*/\sqrt{n}, \qquad (71)$$

where α is chosen to set the local derivative of the constraint, with respect to δ, to zero. [Further discussion of choice of λ's and the construction of α goes here?]

Figure (5) shows the CI that results from our procedure, using the nearly flat constraint in additional to the original two. [discussion of Figure goes here.]

Turning from the graphical analysis, Table 1 shows the identified set for this simple two parameter example and also provides an example of estimated upper and lower bounds for a medium sized and a very large sample.

Table 2 provides a summary of the average confidence intervals and coverage ratios for the same example. We see that the length of the CIs increases as we move from bootstrapping, or from our method with the additional nearly flat constraint, to our method without the flat constraint.. The lower panel of the table provides the Monte Carlo coverage ratios for the CIs. The higher rows gives the coverage ratios for individual bounds, while the last two rows gives the percentage of the samples in which the identified set falls entirely within the CI. We would like these last two numbers to be exactly 95%. The bootstrap is quite close and our method with the flat constraint is only very slightly higher.

7.2 A Four Parameter Example

We now discuss a four parameter example where the two firms have heterogeneous parameters. The probability that firm j ($j = 1, 2$) is profitable in a monopoly is μ_j and the

Figure 5: CI, using “flat” constraint, for μ

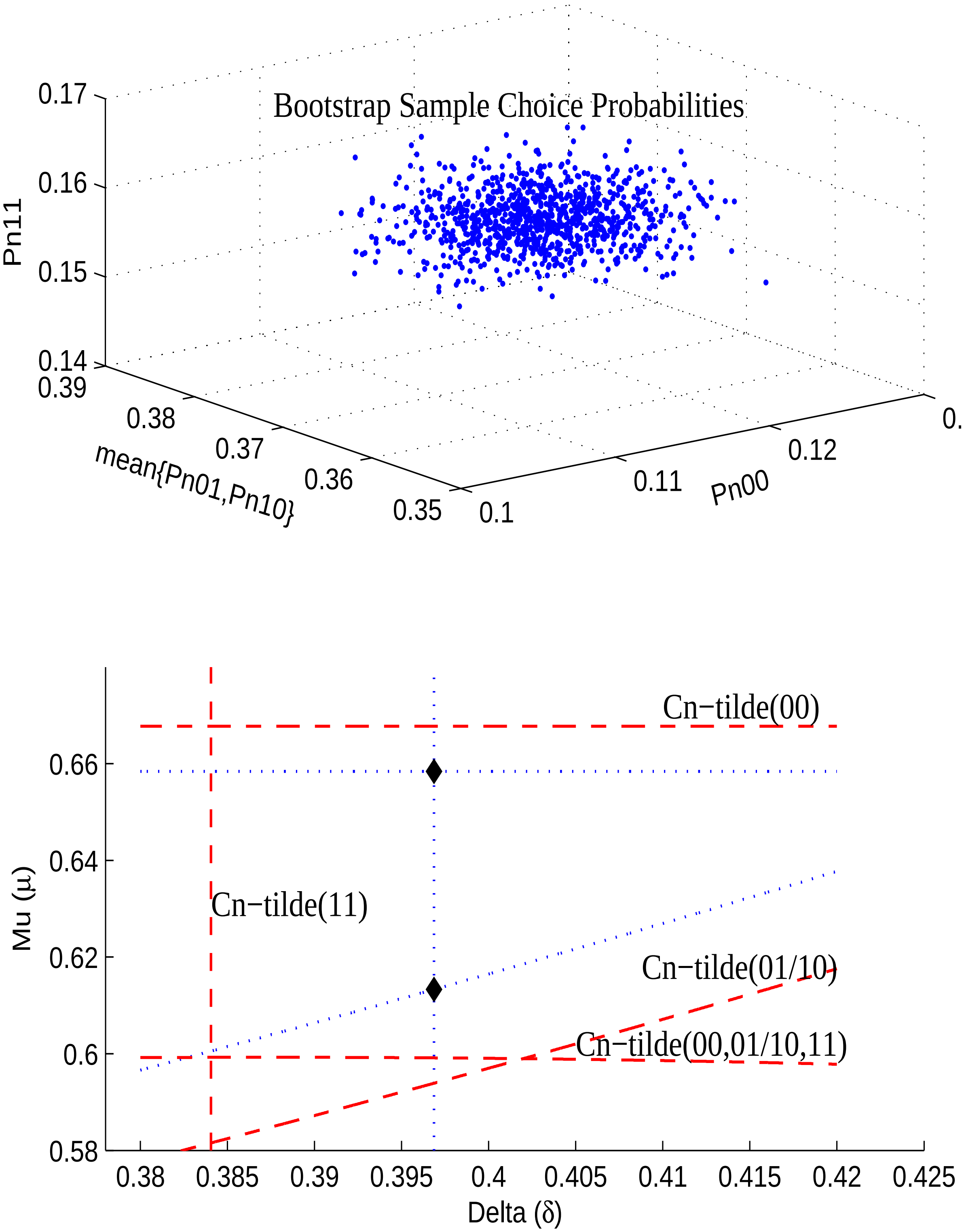


Table 1: Set Estimate for the Simple Two-Firm Case

Sample	Parm	True	Θ_0		Estimate	
Size	Name	Parm	Min	Max	Min	Max
500						
	μ	0.650	0.598	0.650	0.630	0.656
	δ	0.400	0.400	0.448	0.371	0.397
10000						
	μ	0.650	0.598	0.650	0.601	0.646
	δ	0.400	0.400	0.448	0.400	0.441

Table 2: 95% Confidence Interval for the Simple Two-Firm Case,1000 Monte-Carlo

	Θ_0	Ave Bootstrap CI	Ave CI “flat” $\widetilde{c}_n^*$	Ave CI no “flat”
$\underline{\mu}$	0.598	0.569	0.568	0.501
$\bar{\mu}$	0.650	0.693	0.692	0.692
$\underline{\delta}$	0.400	0.357	0.359	0.359
$\bar{\delta}$	0.448	0.477	0.478	0.535
Coverage Ratios:				
%(CIL< $\underline{\mu}$)		0.978	0.979	1.000
%(CIU> $\bar{\mu}$)		0.985	0.985	0.985
%(CIL< $\underline{\delta}$)		0.983	0.980	0.980
%(CIU> $\bar{\delta}$)		0.971	0.978	1.000
%(CIL< $\underline{\mu} < \bar{\mu}$ <CIU)		0.963	0.964	0.985
%(CIL< $\underline{\delta} < \bar{\delta}$ <CIU)		0.954	0.958	0.980

Table 3: 95% Confidence Interval for the Four-Parameter Case,2000 Monte-Carlo

	Θ_0	Ave Boot- strap CI	Ave CI with “flat”	Ave $\tilde{c}_n$ CI	Ave $\tilde{c}_n$ CI $\mu > \delta$
$\underline{\mu_1}$	0.204	0.169	0.170	0.114	0.147
$\bar{\mu_1}$	0.341	0.411	0.458	0.556	0.539
$\underline{\mu_2}$	0.620	0.577	0.577	0.531	0.548
$\bar{\mu_2}$	0.686	0.732	0.744	0.765	0.756
$\underline{\delta_1}$	0.119	0.084	0.073	0.059	0.070
$\bar{\delta_1}$	0.204	0.239	0.241	0.347	0.292
$\underline{\delta_2}$	0.364	0.270	0.212	0.140	0.166
$\bar{\delta_2}$	0.620	0.663	0.663	0.825	0.694
Coverage Ratios:					
%(CIL< $\underline{\mu_1}$)		0.974	0.970	1.000	1.000
%(CIU> $\bar{\mu_1}$)		0.975	1.000	1.000	1.000
%(CIL< $\underline{\mu_2}$)		0.979	0.977	1.000	1.000
%(CIU> $\bar{\mu_2}$)		0.978	0.996	1.000	1.000
%(CIL< $\underline{\delta_1}$)		0.983	0.999	1.000	1.000
%(CIU> $\bar{\delta_1}$)		0.971	0.973	1.000	1.000
%(CIL< $\underline{\delta_2}$)		0.980	1.000	1.000	1.000
%(CIU> $\bar{\delta_2}$)		0.974	0.975	1.000	1.000
%(CIL< $\underline{\mu_1} < \bar{\mu_1}$ <CIU)		0.949	0.970	1.000	1.000
%(CIL< $\underline{\mu_2} < \bar{\mu_2}$ <CIU)		0.957	0.973	1.000	1.000
%(CIL< $\underline{\delta_1} < \bar{\delta_1}$ <CIU)		0.954	0.972	1.000	1.000
%(CIL< $\underline{\delta_2} < \bar{\delta_2}$ <CIU)		0.954	0.975	1.000	1.000
Sample Size	500				

probability that it is profitable in a duopoly is δ_j. There are now four constraints. For example, the probability of the necessary condition for firm one entering while firm two stays out is $\mu_1(1-\delta_2)$.

The summary of the coverage ratios and CIs for this case is provided in Table (3). Now, even the flat CI drives the coverage ratio for our method up to 0.97 or a bit higher, while the bootstrap remains near 0.95. Without the flat constraint, our method gives very high coverage ratios. In this latter case, imposing more constraints from the model (like $\mu > \delta$) improves the size of the CI (but does not noticeably improve the coverage ratio, since they are so very conservative to start).

8 Estimating Competitive Effects Among Principal Discount Chain Retailers

The discount retail industry has been the fastest growing retail sector since the early 1970s. By the end of 1990, there were 187 companies, operating 7,937 stores, with a total revenue of \$114.9 billion, compared with \$2.1 billion in 1961.[14] Among all of the 187 firms in the sample, 66 are single-store firms. The top 30 firms account for a total of 6,822 stores with revenue sales of \$104.9 billion. Most of these stores are departmentalized retail establishments, carrying diversified general merchandise at a low mark-up (for example, Target with headquarters based in MN and ShopKo in WI).[15]

Most of the firms are regional retailers. The exceptions are Kmart and Wal-Mart.[16] By the end of 1990, Kmart had 2,270 stores in 48 states and Puerto Rico selling merchandise worth \$29.53 billion. Wal-Mart had 1,827 stores in 42 states with a total sales of \$25.81 billion.

To study the competitive effects among the different firms and how the firms strategically locate their stores, we focus on small to mid-size counties in the 42 states outside New England, Alaska, and Hawaii.[17] There are 3,043 counties in these 42 states. Roughly 10% of these counties have populations less than 5,000 and we exclude them from the sample. (They have almost no discount stores.) Another roughly 10% of the counties have populations greater than 150,000 (with a mean population of 520,000). Most of these counties are large metropolitan areas with very developed and complicated retail networks. We also exclude these counties from the sample. This leaves us with a sample of counties with populations from 5,000 to 150,000.

Table 4 gives some descriptive statistics for the counties in the sample and Table 5 describes the top 10 chains in the industry ranked by sales volume.

To simplify the analysis, we focus on estimating the competitive effects among Kmart, Wal-Mart, and all other firms.[18] Let the market competition environment be denoted by a three element vector (y_1, y_2, y_3), where $y_j \in \{0, 1\}$ for $j = 1, 2, 3$. Firm j is in the market if $y_j = 1$. Let y_1 stand for Kmart's action, y_2 other firm's action, and y_3 Wal-Mart's action. With three players, there are eight possible market configurations:

$$\{\mathcal{Y}_k\}_{k=1}^{8} = \{(y_1, y_2, y_3)\} = \{(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}.$$

[14] Data source: 1991 Directory of Discount Department Stores, published by Chain Store Guide, NY. Only stores with at least 10,000 square feet sales area are included. The retail sales for all retail sectors was \$709 billion in 1990, including building materials and garden supplies, general merchandise, food, autos and parts and services, apparel, furniture, eating and drinking places, and miscellaneous retail.

[15] The mark-up at these discount stores is typically as low as 10% to 15%, compared with 33% at traditional retailers like Sears. [cite]

[16] Target has 439 stores in 32 states by the end of 1990.

[17] The county and MSA classification is very complicated in the New England area and has changed much more frequently than other states.

[18] Grouping all other firms and treating them as a single firm allows us to focus on how K-Mart and Wal-Mart respond to entry by these chains.

Table 4: Description of the Counties in the Sample, 1990

Variable	Mean	Std	Min	Max
No. of Discount Stores	1.506	1.624	0	11
Total Sales Footage (000 ft^2)	86.379	109.801	0	703
No. of Kmart Stores	0.391	0.624	0	4
Sales Footage of Kmart Stores (000 ft^2)	25.424	44.796	0	295
No. of Wal-mart Stores	0.501	0.672	0	5
Sales Footage of Wal-mart Stores (000 ft^2)	29.830	42.643	0	303
No. of Places/Towns/Cities bigger than 5,000	0.795	1.000	0	9
Total Population in 1990	33159.65	30258.91	5013	149967
Median Household Income ($)	23037.11	5545.477	8595	54801
Sample Size (N)	2461			

Table 5: Sales and No. of Stores for the Top 10 Chains

		#	Sales	Store Size	Std Dev.	Min	Max
Firm Name	Rank	Stores	($ bill.)	(1000 ft^2)	of Size	Size	Size
Kmart	1	2270	30	77	21	40	244
Wal-Mart	2	1827	26	66	21	23	268
Target	3	438	7.5	100	15	45	169
Price	4	57	5.3	105	7	96	117
Ames	5	454	5.0	63	24	30	247
Costco	6	68	4.1	101	6	100	135
Meijer	7	55	2.8	130	55	40	275
Fred Meyer	8	93	2.3	103	39	11	210
Hills	9	153	2.1	76	8	45	113
Pace	10	79	2.1	104	16	98	190

The profit for firm j in market i is

$$\pi_{j,i}((y_1, y_2, y_3), x_i, \theta, \varepsilon_{j,i}) = \beta_0 + x_i'\beta + \delta_j(\sum_{k \neq j} y_k) + \varepsilon_{j,i}, \tag{72}$$

where $x_i'\beta$ measures the size of the market and x_i includes variables such as population, household income, percentage of population that is urban, etc. The parameter δ_j measures the competitive effects: δ_j is the decrease in firm j's profit due to an additional rival. $\varepsilon_{j,i}$ is a firm and market specific profit shock observed by all players but not by the econometrician. $\varepsilon_{j,i}$ is assumed to be i.i.d. across both firms and markets with a standard normal distribution.[19]

The Nash equilibrium condition defines eight $\widehat{c}_n(\cdot)$ statistics, one for each $\mathcal{Y}_k$:

$$\widehat{c}_n(\mathcal{Y}_k, \gamma, \theta) = n^{-1} \sum_{i=1}^{n} \{P(\mathcal{Y}_k | X_i, \theta) - [Y_i = \mathcal{Y}_k]\} h_\gamma(X_i)$$

where n is the total number of markets (one for each county), $P(\mathcal{Y}_k|X_i, \theta)$ is the probability that $\mathcal{Y}_k$ is a Nash equilibrium given X_i and θ^{20}, $[\cdot]$ denotes the indicator function, γ is the vector of X cells, and h_γ is a real-valued function of γ and X_i.[21] The estimated parameter set includes all θ such that the $\widehat{c}_n(\cdot)$ are non-negative for all $\mathcal{Y}_k$ and γ. Since it is possible that no parameters satisfy all these constraints, we adopt the estimator defined by (26).

We start by using the specification of the profit functions and the distribution of the exogenous variables in the discount store sample to carry out a Monte Carlo experiment. We then turn to estimation of the model using the real data.

9 Monte Carlo Example 2

Suppose a firm's profit in market i depends on market population, the (log of) median household income, the distance to the firm's headquarters, and the competitive effects. To generate a sample of market outcomes $\{Y_i : i \leq n\}$, a three-element vector $\varepsilon_i = \{\varepsilon_{1,i}, \varepsilon_{2,i}, \varepsilon_{3,i}\}$ is simulated for each market i. The vector Y_i is the Nash equilibrium corresponding to the simulated ε_i coupled with $x_i'\beta_0$ and $\{\delta_j : j = 1, 2, 3\}$, where x_i is from the real data set and β_0 is the parameter vector that we take to be the truth.[22] In the case of multiple equilibria, each possible equilibrium is chosen with equal probability (but the equilibrium selection rule is taken to be unknown in the estimation exercise).

[19] An implicit assumption is that there is no scale economy associated with multiple stores. Each firm makes independent entry decisions for *every* market.

[20] In regions where there are multiple equilibria, $\mathcal{Y}_k$ is not necessarily the equilibrium observed in the sample. Which one is observed depends on specific equilibrium selection rule, which is not modeled here.

[21] In the case of m cells, a simple example can be each cell containing $1/m$ of the entire sample. Different $\mathcal{Y}_k$ can have different X cells.

[22] Each element of x_i is normalized to have mean zero and standard deviation one within the sample. Hence, each element of β yields the change in profit when the corresponding element of x_i increases by one standard deviation.

Table 6: θ_0 for Example 2

Parameter	True Value
Constant	0.35
Log HH Income	0.45
Population	0.6
Distance to Headquarters	-0.2
Competition: K-Mart	-0.45
Competition: Wal-Mart	-1.05
Competition: Other	-0.75

Table 7: Choice Probabilities for Example 2

$\mathcal{Y}_k$	Ave Prob of Nec. Cond.*	True Prob of $\mathcal{Y}_k$	Sample Freq.
000	0.136	0.136	0.137
001	0.184	0.151	0.145
010	0.207	0.175	0.176
011	0.053	0.045	0.048
100	0.242	0.217	0.224
101	0.099	0.084	0.081
110	0.164	0.147	0.146
111	0.046	0.046	0.042

* Evaluated at $\hat{\theta}_n$ reported in Table 7.

Table 6 displays the true θ_0 value for the exercise. Table 7 summarizes the probabilities of the model events and the necessary conditions associated with those events. Note that the (0,0,0) and (1,1,1) events admit no multiple equilibria, and so, the probabilities of these two necessary (and sufficient) conditions equal the true choice probabilities. In estimating the model, we utilize these equality conditions and they turn out to have important effects on the estimates of β.[23]

One question for the Monte Carlo exercise is how different sets of economic restrictions on the model affect the estimated parameter set. In Table 8, we report different results for the estimated Θ_0 set as we change the restrictions imposed by the model. These are evaluated at the true parameters, so that the full model (with the correct equilibrium selection rule)

[23] An equality constraint can be expressed as two inequality constraints and so the method does not have to be altered to use the equality conditions.

Table 8: Θ_0 for the Chain Store Example

Parm	Necessary		Add Equality		Add		True Equil.	
Parm	Only ($\geq$)		(0,0,0) (1,1,1)		Sufficient		Selection	
Name	Min	Max	Min	Max	Min	Max	Min	Max
Constant	0.177	0.499	0.302	0.376	0.341	0.375	0.350	0.350
Log HH Inc	0.297	0.682	0.424	0.501	0.424	0.457	0.450	0.450
Population	0.426	0.817	0.439	0.676	0.573	0.676	0.600	0.600
Distance	-0.417	-0.059	-0.215	-0.192	-0.215	-0.195	-0.200	-0.200
δ_k	-0.594	-0.243	-0.593	-0.281	-0.541	-0.374	0.450	0.450
δ_t	-0.930	-0.538	-0.854	-0.600	-0.832	-0.684	0.750	0.750
δ_w	-1.348	-0.824	-1.157	-0.906	-1.157	-0.990	0.105	0.105

gives back the true parameters.[24] From left to right, the first set of results gives the set identified by only the inequality constraints implied by the necessary conditions. The next set impose equality on the constraints involving the events that either no firms or else all firms enter. This tightens the constraints greatly. Then, as in Ciliberto and Tamer (2003), we add the sufficient conditions for events (which imply "less-than-or-equal" restrictions). These are still fairly easy to compute in the three firm example. Finally, we add two necessary conditions for the aggregate events that [a] exactly one firm (of any identity) enters and that [b] exactly two firms (of any identify) enters. From the prior literature (e.g. Berry (1989)) we know that in a model with simple competition effects (like this one but unlike the more complicated model of the next section), the necessary conditions for the *number* of firms can point-identify the model. We see that we do obtain the true parameters.

Turning now to questions of estimation for a fixed set of restrictions, note that we can estimate the β vector (but not the competitive effects) using only binary information on whether the outcome (0,0,0) occurs or not. This is because there is no multiple equilibria problem associated with the outcome that all firms are out. The third and fourth columns of Table 9 give CIs for the elements of β using only this information.

We estimate the market size parameters by three methods. The first is MLE (which is possible using only the binary information on (0,0,0)). The second is using a set of 8 X cells, defined by the various combinations of whether each of the three non-constant X variables is above or below its median. The third method uses a first-stage probit to construct an index of the X's, which is then broken into cells.

Table 9 reports the 90% CIs for each parameter and method. The message of the table is that we definitely give up precision relative to the MLE estimator (no surprise) and that the construction of the X cells matters. The first method for constructing X cells works much

[24] There are 39 constraints, which come from interacting the 8 possible outcomes with the X cells and then aggregating somewhat to be sure that each cell has a sufficient number of observattions.

Table 9: 95% Confidence Intervals from Estimation of Market Size Parameters Only

		MLE CI		CI (Good X Cells)		CI (Bad X Cells)	
Parm	True	Min	Max	Min	Max	Min	Max
Constant	0.35	0.233	0.418	0.099	0.583	0.130	6.417
Log HH Inc	0.45	0.322	0.452	0.164	0.666	-1.297	9.000
Population	0.60	0.490	0.780	0.255	1.026	-1.167	9.000
Distance	-0.20	-0.232	-0.128	-0.397	-0.004	-9.000	3.807

Table 10: $\hat{\theta}_n$ and 90% Confidence Intervals

			Confidence Interval	
Parm	θ_0	$\hat{\theta}_n$	Min	Max
Constant	0.35	0.291	0.080	0.680
Log HH Inc	0.45	0.337	0.132	0.760
Population	0.60	0.604	0.217	1.010
Distance	-0.20	-0.151	-0.462	0.047
δ_1	-0.45	-0.406	-1.264	0.000
δ_2	-0.75	-0.677	-1.567	-0.208
δ_3	-1.05	-1.016	-1.819	-0.500

better than the last. Indeed, we constrained each of the parameters to fall in [-9,9] and the confidence bounds for the last method frequently hit those bounds.[25]

Table 10 reports CIs for all parameters, including the competitive effects, when information on all market configurations is utilized ($\{(000), (001), ..., (111)\}$). Here the usual identification argument fails and MLE estimates are not available without assuming some specific equilibrium selection rule. The CIs for the β's are slightly larger than those in Table (9).

10 Confidence Intervals Based on Real Data

Turning away from Monte Carlo examples, we now discuss preliminary estimates on the real data. We find a much better fit when we allow the market size parameters, as well as the

[25] The problem with the "bad" X cells seems partly to be an issue of scale – the restrictions do not rule out parameter vectors where each element is very large.

Table 11: Further Descriptive Statistics for Counties

Variable	Mean	Std	Min	Max
Dummy for Kmart	0.326	0.469	0	1
Dummy for Wal-mart	0.423	0.494	0	1
Dummy for Other chains	0.275	0.446	0	1
Distance to AR	2.054	0.838	0	3
Percentage of Urban (1990)	0.345	0.255	0	1
Log of Retail Sales ($000)	11.336	1.068	7.910	14.013
No. of Place $\geq$ 5k	0.795	1.000	0	9
Total No of Obs	2458			

competitive effects, to vary by firm. The profit functions for the three firms are:

$$\begin{aligned}
\Pi_{i,k} &= \beta_{0,k} + \beta_{1,k} * lnrt_i + \beta_{2,k} * urb_i + \delta_k * D_{i,w} + \varepsilon_{i,k} \\
\Pi_{i,t} &= \beta_{0,t} + \beta_{1,t} * lnrt_i + \delta_t * D_{i,w} + \varepsilon_{i,t} \\
\Pi_{i,w} &= \beta_{0,w} + \beta_{1,w} * lnrt_i + \beta_{2,w} * dhq_i + \delta_{wk} * D_{i,k} + \delta_{wt} * D_{i,t} + \varepsilon_{i,w}
\end{aligned}$$

where k refers to Kmart, w refers to Wal-Mart and t refers to other firms [26]. The unobservable in market i, $\varepsilon_i = (\varepsilon_{i,k}, \varepsilon_{i,t}, \varepsilon_{i,w})'$, is i.i.d. across firms and markets. We found that population is an insufficient measure of market size and so we also use the variable $lnrt$, which is the log of county retail sales in thousands of dollars. The variable urb is the percentage of urban population, and dhq is a discrete variable that takes value one if the county is located in Wal-Mart's headquarter state Arkansas (AR) or states contiguous to AR, two for states contiguous these states and three for all other states. 26.5% of the counties are adjacent to Wal-Mart's headquarter state AR, 38.5% of the counties are in nearby states, and 35% of the counties are in states further away from AR. $D_{i,k}, D_{i,w}$, and $D_{i,t}$ are dummies that take value one if the firm has stores in the market.

Table 11 lists some further descriptive statistics about the counties.[27]

The estimated parameter values and CIs are listed in Table 9.

Once again, the construction of the X cells is important. When we try very fine cells, we simply reject the model—there is no parameter in the confidence region. This is not surprising given the strong parametric restrictions on the profit function. We also tried a set of very "rough" cells that give the first confidence region in the table. The rough X cells are combinations of low, medium and high levels of different regressors, with some cells merged to guarantee enough observations of the Y events. The subsets $\mathcal{Y}_k$ used in the estimation are

[26] All other chains are treated as a single firm that competes against Kmart and Wal-mart. This is a parsimonious way to model how the national chains Kmart and Wal-mart compete against each other as well as against regional chains.

[27] There are 2461 counties in total. Three counties do not have retail sales number and are not included in the estimation.

Table 12: 90% Confidence Region Using Rough Cells and Fine Cells

	Rough Cells		Fine Cells	
	Max	Min	Max	Min
Wal-mart				
Constant	-39.000	-14.183	-38.987	-28.023
Log Retail Sales	1.609	4.175	2.726	3.799
Dist to HQ	-3.507	-0.943	-1.636	-1.140
Comp. w/ K-M.	-3.432	0.000	-2.676	-0.751
Comp w/ Oth.	-4.863	-0.554	-3.977	-1.925
Kmart				
Constant	-30.860	-16.857	-29.181	-17.837
Log Retail Sales	1.364	2.330	1.642	2.379
% Urban	-1.454	4.734	1.427	3.759
Comp. w/ W-M	-1.684	-0.078	-1.675	-0.300
Other				
Constant	-20.501	-11.622	-19.672	-14.300
Log Retail Sales	1.006	1.719	1.369	1.593
Comp. w/ W-M	-2.426	-0.756	-2.276	-1.177

the eight mutually exclusive entry events $\{000, 001, 010, 011, 100, 101, 110, 111\}$, plus three aggregated indicators for whether each firm enters. These last three are especially important for estimating the heterogeneous parameters on x. Each X cell is constructed to have at least 200 observations, with at least 25 of the observations associated with each event. If the cells were badly violated in initial runs,[28] then they are merged with adjacent cells. In the end, we have 14 equality constraints (associated with the (0,0,0) and (1,1,1) events) and 55 inequality constraints.

All the parameters are statistically different from zero except for the coefficient on the percentage of urban population in Kmart's profit.[29] The confidence region listed under 'fine cells' are the result of a more complicated process of "fishing" for good cells. The results should be viewed with some caution because of the very data-dependent process generating the cells. The results do show that improved cells can give improved results. Further research into how to construct the X cells would be useful.

We also estimate the model using ML, assuming equal likelihood in regions of multi-equilibria. In Table 13 we show the ML results alongside the "rough cells" result from before.

Using the CIs defined by the rough cells, two $\widehat{\beta}^{MLE}$'s and $\widehat{\delta}_{wt}^{MLE}$ are outside the CIs, where $\widehat{\delta}_{wt}^{MLE}$ is the estimated parameter for "competition with other firms." In particular, $\widehat{\delta}_{wt}^{MLE}$ is a statistically significant positive number 2.284 if it is not constrained to be non-positive. The ML estimates for the parameters in Wal-Mart's profit function are very different from that of our estimates, largely because of a very different estimate for the competitive effects. The competitive effects are very sensitive to the assumption of the equilibrium selection rule; therefore the MLE is inconsistent if the equilibrium-selection assumption is wrong.

When the estimated set becomes a point $\widehat{\theta}_n$ that minimizes all the violated $\widehat{c}_n$ constraints, we lose information as to what the constraint binding sets are for the upper bound and lower bound of a particular parameter. With many constraints there are typically many cells binding at the estimated $\widehat{\theta}_n$. The conservative approach (which is the approach used in all of the above tables) is to include all these binding cells in the constraint binding sets. Obviously this leads to a fairly big $\lambda_n^*(\alpha)$. Another approach is to include only those cells that are binding when estimating the confidence intervals. If one can show that certain cells should never bind at the boundary of a parameter and therefore should not be included in the constraint binding sets, it can help to obtain the right $\lambda_n^*(\alpha)$ and a more reasonable confidence region. Table (14) shows the difference when different constraint binding sets are adopted in calculating $\lambda_n^*(\alpha)$. The usual conservative approach is used for the first two columns, while in the second two columns, only the cells that are binding at the boundary of the confidence intervals are included in the calculation of $\lambda_n^*(\alpha)$.[30]

[28] 'Badly violated' means that the $\tilde{c}_n$ constraints at these cells can't be satisfied, which leads to a null confidence interval.

[29] One interesting empirical finding is that the first stage parameter estimate $\hat{\theta}_n$ is not necessarily in the confidence interval if no estimates can satisfy all the $\hat{c}_n$ constraints so the estimated set becomes a point that minimizes the sum of squared $\hat{c}_n$ that are violated.

[30] The second approach requires several steps. First one finds confidence region using the usual approach. Then the cells that are binding at the boundary of the confidence region are

Table 13: Comparison With MLE Estimate

		MLE* 90% CI		$\tilde{C}_n$ 90% CI	
	MLE $\hat{\theta}$	Min	Max	Min	Max
Walmart:					
Constant	-9.1594	-10.873	-7.445	-39.000	-14.183
Log Retail Sales	0.96773	0.952	0.983	1.609	4.175
Dist. to HQ	-1.0151	-1.019	-1.011	-3.507	-0.943
Comp. w/ K-M.	-0.074015	-0.117	-0.031	-3.432	0.000
Comp. w/ Oth.	0	-0.119	0.119	-4.863	-0.554
Kmart:					
Constant	-18.781	-19.883	-17.679	-30.860	-16.857
Log Retail Sales	1.5327	1.525	1.541	1.364	2.330
% Urban	1.3673	1.307	1.428	-1.454	4.734
Competition	-0.11742	-0.148	-0.087	-1.684	-0.078
Other:					
Constant	-12.02	-12.813	-11.226	-20.501	-11.622
Log Retail Sales	1.0307	1.024	1.037	1.006	1.719
Competition	-1.1309	-1.177	-1.085	-2.426	-0.756

*The MLE confidence interval is calculated according to the usual formula, which does not take into account the fact that one of the parameters is estimated to be at the boundary of zero. We need to fix this.

Table 14: Confidence Regions with Different Critical Values

	All Binding Cells		Only CI Binding Cells	
	Min	Max	Min	Max
Kmart:				
Constant	-30.860	-16.857	-28.758	-17.149
Log Retail Sales	1.364	2.330	1.521	2.062
% Urban	-1.454	4.734	-0.604	2.792
Comp w/ W-M	-1.684	-0.078	-1.413	-0.092
Other:				
Constant	-20.501	-11.622	-19.181	-12.220
Log Retail Sales	1.006	1.719	1.107	1.564
Comp w/ W-M	-2.426	-0.756	-2.104	-0.864
Walmart:				
Constant	-39.000	-14.183	-39.000	-15.330
Log Retail Sales	1.609	4.175	1.748	4.099
Dist. to HQ	-3.507	-0.943	-3.080	-1.086
Comp w/ Kmart.	-3.432	0.000	-2.260	0.000
Comp w/ Other	-4.863	-0.554	-4.169	-0.872

11 Conclusion

To be written.

12 Appendix of Proofs

For any real function c on $\mathcal{I}_K \times \Gamma_{all} \times \Theta$ and any collection of $\sum_{k=1}^{K} M_k$ subsets Γ of $\mathcal{X}$, as in (15), let

$$\Theta(c, \Gamma) = \{\theta \in \Theta : c(k, \gamma_{k,m}, \theta) \geq 0,\ \forall (k,m) \in \mathcal{I}_{K,M}\}. \tag{73}$$

Note that

$$\Theta(c_0, \Gamma_0) = \Theta_+. \tag{74}$$

In addition, if $\Theta(\widehat{c}_n, \widehat{\Gamma}_n)$ is non-empty, then $\Theta(\widehat{c}_n, \widehat{\Gamma}_n) = \widehat{\Theta}_n$. Hence, $\Theta(\widehat{c}_n, \widehat{\Gamma}_n) \subset \widehat{\Theta}_n$.

The proof of part (b) of Theorem 1 uses the following lemma.

Lemma 4 *Under Assumptions* 3 *and* 4(b), $\rho(\Theta(c_0, \Gamma_0)|\Theta(c, \Gamma)) \to 0$ *as* $c \to c_0$ *and* $\Gamma \to \Gamma_0$ *under* $||\cdot||_U$ *and* $||\cdot||_G$.

Proof of Theorem 1. To prove part (a), we use an extension of a standard method of establishing the probability limit of a sequence of extremum estimators. The extension allows for the probability limit to be a set rather than a single vector. Under Assumption 3, $Q(\theta)$ (defined in (23)) is continuous. (This holds even though an indicator function appears in $Q(\theta)$ because $|b(\theta)| \cdot [b(\theta) \leq 0]$ is continuous whenever $b(\theta)$ is continuous, which follows from the fact that $|\ |b(\theta_1)| \cdot [b(\theta_1) \leq 0] - |b(\theta_2)| \cdot [b(\theta_2) \leq 0]\ | \leq |\ |b(\theta_1)| - |b(\theta_2)|\ |$ for all θ_1, θ_2.) The continuous function $Q(\theta)$ on the compact set Θ attains its minimum value zero at points in the set Θ_+ by (24).

We claim that for all $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\inf_{\theta \notin S(\Theta_+, \varepsilon)} Q(\theta) \geq \delta > 0. \tag{75}$$

Suppose not. Then, for some $\varepsilon > 0$, there is a sequence $\{\theta_j \in \Theta / S(\Theta_+, \varepsilon) : j \geq 1\}$ for which $\lim_{j \to \infty} Q(\theta_j) = 0$. Because Θ is compact and $S(\Theta_+, \varepsilon)$ is open, the set $\Theta / S(\Theta_+, \varepsilon)$ is compact. Hence, $\{\theta_j : j \geq 1\}$ has a convergent subsequence, say $\{\theta_{j_\ell} : \ell \geq 1\}$, that converges to a point in $\Theta / S(\Theta_+, \varepsilon)$, say θ_∞. By continuity of $Q(\cdot)$, $Q(\theta_\infty) = \lim_{\ell \to \infty} Q(\theta_{j_\ell}) = 0$. This implies that θ_∞ is in Θ_+, which is a contradiction. Hence, (75) holds.

The set Θ_+ is not empty by Assumption 2 and $\widehat{\Theta}_n$ is not empty by the argument in footnote 4 [I.E., THE FOOTNOTE FOLLOWING (26)]. Let θ_+ denote some element of

included in the constraint binding sets. To reduce the uncertainty about the binding cells for the confidence region, one can bootstrap the sample, repeat the calculation several times and include all the cells that are binding in any one of these calculations. This new constraint binding set is then used to calculate $\lambda_n^*(\alpha)$ and a more reasonable confidence interval.

Θ_+. Equation (75) and the fact Θ_+ and $\widehat{\Theta}_n$ are not empty imply that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\begin{aligned}
&P(\rho(\widehat{\Theta}_n|\Theta_+) > \varepsilon) \\
&= P(\widehat{\Theta}_n \cap (\Theta/S(\Theta_+,\varepsilon)) \neq \varnothing) \\
&\leq P(\sup_{\theta\in\widehat{\Theta}_n} Q(\theta) \geq \delta) \\
&= P(\sup_{\theta\in\widehat{\Theta}_n} (Q(\theta) - Q_n(\theta) + Q_n(\theta)) \geq \delta) \\
&\leq P(\sup_{\theta\in\widehat{\Theta}_n} (Q(\theta) - Q_n(\theta) + Q_n(\theta_+) - Q(\theta_+)) \geq \delta) \\
&\leq P(2\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \geq \delta), \qquad (76)
\end{aligned}$$

where the second inequality holds because (i) $Q_n(\theta) \leq Q_n(\theta_+)$ for each $\theta \in \widehat{\Theta}_n$ because each $\theta \in \widehat{\Theta}_n$ minimizes $Q_n(\theta)$ over Θ and (ii) $Q(\theta_+) = 0$ because $\theta_+ \in \Theta_+$.

The right-hand side of (76) converges in probability to zero because

$$\begin{aligned}
\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| &\leq \sup_{\theta\in\Theta} \max_{(k,m)\in\mathcal{I}_{K,M}} |\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta) - c_0(k,\gamma_{0,k,m},\theta)| \\
&\leq \sup_{\theta\in\Theta} \max_{(k,m)\in\mathcal{I}_{K,M}} |\widehat{c}_n(k,\widehat{\gamma}_{n,k,m},\theta) - c_0(k,\widehat{\gamma}_{n,k,m},\theta)| \\
&\quad + \sup_{\theta\in\Theta} \max_{(k,m)\in\mathcal{I}_{K,M}} |c_0(k,\widehat{\gamma}_{n,k,m},\theta) - c_0(k,\gamma_{0,k,m},\theta)| \\
&\leq \sup_{(k,\gamma,\theta)\in\mathcal{I}_K\times\Gamma_{all}\times\Theta} |\widehat{c}_n(k,\gamma,\theta) - c_0(k,\gamma,\theta)| \\
&\quad + 2 \max_{(k,m)\in\mathcal{I}_{K,M}} \int |h_{\widehat{\gamma}_{n,k,m}}(x) - h_{\gamma_{0,k,m}}(x)| dG(x) \\
&\to_p 0 \qquad (77)
\end{aligned}$$

where the first inequality holds by the definitions of $Q_n(\theta)$ and $Q(\theta)$, the second inequality holds by the triangle inequality, the third inequality uses the definition of $c_0(k,\gamma,\theta)$ in (18), and the convergence to zero holds by Assumptions 5 and 6 using the definition of $||\cdot||_G$ given in (34) and (35). This completes the proof of part (a).

We now prove part (b). Suppose Assumption 4(b) holds. Then, by Lemma 4, given $\varepsilon > 0$, there exists $\delta > 0$ such that $||c - c_0||_U < \delta$ and $||\Gamma - \Gamma_0||_G < \delta$ imply that

$$\rho(\Theta(c_0,\Gamma_0)|\Theta(c,\Gamma)) < \varepsilon. \qquad (78)$$

Hence,

$$P(\rho(\Theta(c_0,\Gamma_0)|\Theta(\widehat{c}_n,\widehat{\Gamma}_n)) < \varepsilon) \geq P(||\widehat{c}_n - c_0||_U < \delta, ||\widehat{\Gamma}_n - \Gamma_0||_G < \delta) \to 1, \qquad (79)$$

where the convergence holds by Assumptions 5 and 6.

As noted just after (74), $\Theta(\widehat{c}_n,\widehat{\Gamma}_n) \subset \widehat{\Theta}_n$. Hence, we have

$$P(\rho(\Theta(c_0,\Gamma_0)|\widehat{\Theta}_n) < \varepsilon) \geq P(\rho(\Theta(c_0,\Gamma_0)|\Theta(\widehat{c}_n,\widehat{\Gamma}_n)) < \varepsilon). \qquad (80)$$

This result and (79) establish part (b) under Assumption 4(b).

Next, suppose Assumption 4(a) holds. Then,

$$\rho(\Theta_+|\widehat{\Theta}_n) = \rho(\{\theta_0\}|\widehat{\Theta}_n) \leq \rho(\widehat{\Theta}_n|\{\theta_0\}) \to_p 0, \tag{81}$$

where the inequality holds because (i) the distance from a point to a non-empty set is less than or equal to the distance from the set to the point and (ii) the set $\widehat{\Theta}_n$ is not empty by the argument in footnote 4 [I.E., THE FOOTNOTE FOLLOWING (26)] and the convergence to zero holds by part (a) of the theorem. This completes the proof of part (b) of the Theorem.

Part (c) follows immediately from parts (a) and (b). □

Proof of Lemma 4. For any $\theta \in \Theta$, we have

$$\begin{aligned} &\limsup_{(c,\Gamma)\to(c_0,\Gamma_0)} \left| \min_{(k,m)\in\mathcal{I}_{K,M}} c(k,\gamma_{k,m},\theta) - \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{k,m},\theta) \right| \\ &\leq \limsup_{c\to c_0} \sup_{k\in\mathcal{I}_K,\gamma\in\Gamma_{all}} |c(k,\gamma,\theta) - c_0(k,\gamma,\theta)| = 0, \end{aligned} \tag{82}$$

because $c \to c_0$ with respect to the uniform metric over $\mathcal{I}_K \times \Gamma_{all} \times \Theta$. Hence, for any $\theta \in \Theta$,

$$\liminf_{(c,\Gamma)\to(c_0,\Gamma_0)} \min_{(k,m)\in\mathcal{I}_{K,M}} c(k,\gamma_{k,m},\theta) = \liminf_{\Gamma\to\Gamma_0} \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{k,m},\theta). \tag{83}$$

Next, for any $\theta \in \Theta$, we have

$$\begin{aligned} &\limsup_{\Gamma\to\Gamma_0} \left| \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{k,m},\theta) - \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{0,k,m},\theta) \right| \\ &\leq \limsup_{\Gamma\to\Gamma_0} \max_{(k,m)\in\mathcal{I}_{K,M}} 2\int \left| h_{\gamma_{k,m}}(x) - h_{\gamma_{0,k,m}}(x) \right| dG(x) = 0, \end{aligned} \tag{84}$$

where the inequality holds using the definition of c_0 in (18) and the equality holds by the definition of $\Gamma \to \Gamma_0$ given in (34) and the Cauchy-Schwarz inequality. In consequence, for any $\theta \in \text{int}(\Theta_+)$,

$$\liminf_{\Gamma\to\Gamma_0} \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{k,m},\theta) = \min_{(k,m)\in\mathcal{I}_{K,M}} c_0(k,\gamma_{0,k,m},\theta) > 0, \tag{85}$$

where the inequality holds by Assumption 4(b). Equations (83) and (85) combine to show that for any $\theta \in \text{int}(\Theta_+)$, $\theta \in \Theta(c,\Gamma)$ for (c,Γ) sufficiently close to (c_0,Γ_0). In particular, this implies that $\Theta(c,\Gamma)$ is not empty for (c,Γ) sufficiently close to (c_0,Γ_0).

Now, suppose $\rho(\Theta(c_0,\Gamma_0)|\Theta(c,\Gamma)) \nrightarrow 0$ as $(c,\Gamma) \to (c_0,\Gamma_0)$. Then, since $\rho(\Theta(c_0,\Gamma_0)|\Theta(c,\Gamma)) = \inf_{\theta\in\Theta(c_0,\Gamma_0)} \rho(\theta|\Theta(c,\Gamma))$, there exists (i) a constant $\varepsilon > 0$, (ii) a sequence of functions on $\mathcal{I}_K \times \Gamma_{all} \times \Theta$, $\{c_j : j \geq 1\}$, and a sequence of collections of $\sum_{k=1}^K M_k$ sets in Γ_{all}, $\{\Gamma_j : j \geq 1\}$, such that $(c_j,\Gamma_j) \to (c_0,\Gamma_0)$, and (iii) a sequence of parameters $\{\theta_{c_j} \in \Theta(c_0,\Gamma_0) : j \geq 1\}$ such that $\rho(\theta_{c_j}|\Theta(c_j,\Gamma_j)) \geq \varepsilon$ for all $j \geq 1$. The sequence $\{\theta_{c_j} \in \Theta(c_0,\Gamma_0) : j \geq 1\}$ has a subsequence, say $\{\theta_{c_{j_\ell}} : \ell \geq 1\}$, that converges to a point in

$\Theta(c_0, \Gamma_0)$ because $\Theta(c_0, \Gamma_0)$ is compact under Assumption 3. That is, $\theta_{c_{j_\ell}} \to \theta_*$ as $\ell \to \infty$ for some $\theta_* \in \Theta(c_0, \Gamma_0)$. For all ℓ sufficiently large that $||\theta_{c_{j_\ell}} - \theta_*|| < \varepsilon/2$, we have

$$|\rho(\theta_{c_{j_\ell}}|\Theta(c_{j_\ell}, \Gamma_{j_\ell})) - \rho(\theta_*|\Theta(c_{j_\ell}, \Gamma_{j_\ell}))| \leq ||\theta_{c_{j_\ell}} - \theta_*|| < \varepsilon/2, \tag{86}$$

where the inequality holds by the triangle inequality. Hence, for all ℓ sufficiently large,

$$\rho(\theta_*|\Theta(c_{j_\ell}, \Gamma_{j_\ell})) \geq \rho(\theta_{c_{j_\ell}}|\Theta(c_{j_\ell}, \Gamma_{j_\ell})) - \varepsilon/2 \geq \varepsilon/2. \tag{87}$$

If $\theta_* \in \text{int}(\Theta_+)$, equation (87) contradicts the result given at the end of the second paragraph of the proof that for any $\theta \in \text{int}(\Theta_+)$, $\theta \in \Theta(c, \Gamma)$ for (c, Γ) sufficiently close to (c_0, Γ_0).

If $\theta_* \notin \text{int}(\Theta_+)$, then by the first part of Assumption 4(b) there exists a parameter $\theta_{int} \in \text{int}(\Theta_+)$ with $||\theta_* - \theta_{int}|| < \varepsilon/4$. By the triangle inequality and (87), $\rho(\theta_{int}|\Theta(c_{j_\ell}, \Gamma_{j_\ell})) \geq \varepsilon/4$ for all ℓ sufficiently large. This also contradicts the result that for any $\theta \in \text{int}(\Theta_+)$, $\theta \in \Theta(c, \Gamma)$ for (c, Γ) sufficiently close to (c_0, Γ_0). Hence, $\rho(\Theta(c_0, \Gamma_0))|\Theta(c, \Gamma))) \to 0$ as $(c, \Gamma) \to (c_0, \Gamma_0)$. □

Proof of Theorem 2. First, we establish the following result. For any two sets of real numbers B_1 and B_2, let $b_j^* = \sup\{b \in B_j\}$ for $j = 1, 2$. Then,

$$|b_1^* - b_2^*| \leq d(B_1, B_2). \tag{88}$$

To prove (88), suppose $b_1^* > b_2^*$. Then, $|b_1^* - b_2^*| = \rho(b_1^*|B_2) \leq \sup_{b \in B_1} \rho(b|B_2) = \rho(B_1|B_2) \leq d(B_1, B_2)$. By an analogous argument, (88) holds if $b_1^* < b_2^*$. Hence, (88) holds.

Define $\widehat{B}_n = \{\beta_n(\theta) : \theta \in \widehat{\Theta}_n\}$ and $B_{n,+} = \{\beta_n(\theta) : \theta \in \Theta_+\}$. Equation (88) implies that

$$|\widehat{\beta}_{n,U} - \beta_{n,U}| \leq d(\widehat{B}_n, B_{n,+}) \tag{89}$$

because $\widehat{\beta}_{n,U} = \sup\{\beta_n(\theta) : \theta \in \widehat{\Theta}_n\} = \sup\{\beta \in \widehat{B}_n\}$ and $\beta_{n,U} = \sup\{\beta \in B_{n,+}\}$.

Next, under Assumption 7, we claim that for all $\varepsilon > 0$ there exists $\delta > 0$ such that for all n large

$$P\left(d(\widehat{B}_n, B_{n,+}) > \varepsilon\right) \leq 2P\left(d(\widehat{\Theta}_n, \Theta_+) > \delta\right) + 2\varepsilon. \tag{90}$$

Since $\varepsilon > 0$ is arbitrary, this claim, Theorem 1(c), and (89) combine to establish the result of the Theorem for $\widehat{\beta}_{n,U}$. The proof for $\widehat{\beta}_{n,L}$ is analogous.

It remains to verify (90). Using the definition of $d(\cdot, \cdot)$, we have

$$P\left(d(\widehat{B}_n, B_{n,+}) > \varepsilon\right) \leq P\left(\rho(\widehat{B}_n|B_{n,+}) > \varepsilon\right) + P\left(\rho(B_{n,+}|\widehat{B}_n) > \varepsilon\right). \tag{91}$$

Let $h_\delta = P\left(d(\widehat{\Theta}_n, \Theta_+) > \delta\right)$. We have

$$\begin{aligned} & P\left(\rho(\widehat{B}_n|B_{n,+}) > \varepsilon\right) \\ \leq\ & P\left(\rho(\widehat{B}_n|B_{n,+}) > \varepsilon,\ d(\widehat{\Theta}_n, \Theta_+) \leq \delta\right) + h_\delta \qquad (92) \\ =\ & P\left(\sup_{\theta \in \widehat{\Theta}_n} \inf\{|\beta_n(\theta) - \beta_n(\theta_+)| : \theta_+ \in \Theta_+\} > \varepsilon,\ d(\widehat{\Theta}_n, \Theta_+) \leq \delta\right) + h_\delta \end{aligned}$$

where the equality holds by the definitions of $\rho(\cdot|\cdot)$, $\widehat{B}_n$, and $B_{n,+}$, Now, if $d(\widehat{\Theta}_n, \Theta_+) \leq \delta$, then $\rho(\widehat{\Theta}_n|\Theta_+) < \delta$ and given any $\theta \in \widehat{\Theta}_n$ there exists $\theta_{++} \in \Theta_+$ such that $||\theta - \theta_{++}|| \leq \delta$. Hence,

$$\begin{aligned} \sup_{\theta \in \widehat{\Theta}_n} \inf\{|\beta_n(\theta) - \beta_n(\theta_+)| : \theta_+ \in \Theta_+\} &\leq \sup_{\theta \in \widehat{\Theta}_n} |\beta_n(\theta) - \beta_n(\theta_{++})| \\ &\leq \sup_{||\theta_1 - \theta_2|| \leq \delta} |\beta_n(\theta_1) - \beta_n(\theta_2)|. \end{aligned} \tag{93}$$

Combining (92), (93), and Assumption 7 gives: for all n large,

$$P\left(\rho(\widehat{B}_n|B_{n,+}) > \varepsilon\right) \leq \varepsilon + P\left(d(\widehat{\Theta}_n, \Theta_+) > \delta\right). \tag{94}$$

An analogous argument gives the same result as in (94), but with $\rho(B_{n,+}|\widehat{B}_n)$ and $\rho(\Theta_+|\widehat{\Theta}_n)$ in place of $\rho(\widehat{B}_n|B_{n,+})$ and $\rho(\widehat{\Theta}_n|\Theta_+)$, respectively. These results and (91) combine to give (90) and the proof is complete. □

The proof of Theorem 3 uses the following Lemma.

Lemma 5 *Under Assumptions* CI1-CI4, CI6, 2, 3(a), *and* 6, *for all* $(k, m) \in \mathcal{I}_{K,M}$,
(a) $\widehat{\nu}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \rightarrow_d \nu_0(k, \gamma_{0,k,m}, \theta_{+,U})$,
(b) $\widehat{Z}_n(k, m, \theta_{n,U}) \rightarrow_d Z_0(k, m, \theta_{+,U})$,
(c) $\widehat{w}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \rightarrow_p w_0(k, \gamma_{0,k,m}, \theta_{+,U})$, *and*
(d) $\widehat{\nu}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) + \widehat{Z}_n(k, m, \theta_{n,U}) + \widehat{w}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U})\lambda^*_{n,U}(k, m, \alpha)$
$\rightarrow_d \nu_0(k, \gamma_{0,k,m}, \theta_{+,U}) + Z_0(k, m, \theta_{+,U}) + w_0(k, \gamma_{0,k,m}, \theta_{+,U})\lambda_{0,U}(k, m, \alpha)$.
The results of parts (a)-(d) *hold with* U *replaced by* L *and all the convergence results of the Lemma hold jointly.*

Proof of Theorem 3. First, we establish the desired result for the CI for the true value β_0. By Assumption CI4(a), $\beta_0 = \beta_n(\theta_0) \rightarrow_p \beta_0(\theta_0)$. For notational simplicity, let $\beta_{0,0} = \beta_0(\theta_0)$. Note that $\beta_{0,0}$ is the asymptotic true value.

We consider the following cases separately: (i) $\beta_{+,L} < \beta_{0,0} < \beta_{+,U}$, (ii) $\beta_{+,L} < \beta_{0,0} = \beta_{+,U}$, (iii) $\beta_{+,L} = \beta_{0,0} < \beta_{+,U}$, and (iv) $\beta_{+,L} = \beta_{0,0} = \beta_{+,U}$.

We consider case (i) first. Because $\widehat{w}_n(k, \gamma, \theta) > 0$ and $\lambda^*_{n,U}(k, m, \alpha) \geq 0$, we have $\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta) \geq \widehat{c}_n(k, \widehat{\gamma}_{n,k,m}, \theta)$ for all (k, m, θ). Hence, $\widetilde{\beta}_{n,U} \geq \widehat{\beta}_{n,U}$. Combining this with Assumption CI5(a) gives

$$\begin{aligned} \widetilde{\beta}_{n,U} - \beta_0 \geq \widehat{\beta}_{n,U} - \beta_0 \rightarrow_p \beta_{+,U} - \beta_{0,0} > 0 \text{ and} \\ P(\beta_0 \leq \widetilde{\beta}_{n,U}) \rightarrow 1. \end{aligned} \tag{95}$$

By an analogous argument, $P(\beta_0 \geq \widetilde{\beta}_{n,L}) \rightarrow 1$, which establishes the result of the Theorem for case (i).

Next, we consider case (ii). Because $\beta_{0,0} > \beta_{+,L}$, the same argument as above gives $P(\beta_0 \geq \widetilde{\beta}_{n,L}) \rightarrow 1$. Hence, it suffices to show that $\liminf_{n\rightarrow\infty} P(\beta_0 \leq \widetilde{\beta}_{n,U}) \geq 1 - \alpha$. By definition of $\theta_{n,U}$ in (60), $\beta_n(\theta_{n,U}) = \beta_{n,U}$. Hence, if $\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}$, then $\widetilde{\beta}_{n,U}$ must be at least as large as $\beta_{n,U}$ by definition of $\widetilde{\beta}_{n,U}$. In addition, $\beta_0 \leq \beta_{n,U}$ by the definition of $\beta_{n,U}$ and the fact that $\theta_0 \in \Theta_+$ by Assumption 2. These results imply that

$$P(\beta_0 \leq \widetilde{\beta}_{n,U}) \geq P(\beta_{n,U} \leq \widetilde{\beta}_{n,U}) \geq P\left(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}\right) \tag{96}$$

for all $n \geq 1$.

We have

$$\begin{aligned}
&\liminf_{n\to\infty} P\left(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}\right) \\
&= \liminf_{n\to\infty} P\left(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq c_0(k, \gamma_{0,k,m}, \theta_{n,U}) \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}\right. \\
&\qquad \left. \& \ \widehat{\mathcal{B}}_{n,U} \subset \mathcal{B}_{n,U}\right) \\
&\quad + \liminf_{n\to\infty} P\left(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq 0 \ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U} \ \& \ (\widehat{\mathcal{B}}_{n,U} \subset \mathcal{B}_{n,U})^c\right) \\
&\geq \liminf_{n\to\infty} P\left(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq c_0(k, \gamma_{0,k,m}, \theta_{n,U}) \ \forall (k,m) \in \mathcal{B}_{n,U}\right) \\
&= \liminf_{n\to\infty} Q_{n,U},
\end{aligned} \tag{97}$$

where the first equality holds because $c_0(k, \gamma_{0,k,m}, \theta_{n,U}) = 0$ for all $\forall (k,m) \in \mathcal{B}_{n,U}$ by the definition of $\mathcal{B}_{n,U}$, the inequality holds by Assumption CI5(b) and the fact that a set is no larger when it is defined using more restrictions, and the last equality defines $Q_{n,U}$.

Using the definitions of $\widetilde{c}_{n,U}(k, \gamma, \theta)$, $\widehat{\nu}_n(k, \gamma, \theta)$, and $\widehat{Z}_n(k, m, \theta)$, we have

$$\begin{aligned}
Q_{n,U} = P\Big(&\widehat{\nu}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) + \widehat{Z}_n(k, m, \theta_{n,U}) \\
&+ \widehat{w}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \lambda^*_{n,U}(k, m, \alpha) \geq 0 \ \forall (k,m) \in \mathcal{B}_{n,U}\Big).
\end{aligned} \tag{98}$$

Hence, using Lemma 5(d) and Assumption CI5(b),

$$\begin{aligned}
\liminf_{n\to\infty} Q_{n,U} \geq P(&\nu_0(k, \gamma_{0,k,m}, \theta_{+,U}) + Z_0(k, m, \theta_{+,U}) + \\
&w_0(k, \gamma_{0,k,m}, \theta_{+,U}) \lambda_{0,U}(k, m, \alpha) \geq 0 \ \forall (k,m) \in \mathcal{B}_{+,U}).
\end{aligned} \tag{99}$$

(This is an inequality only because Assumption CI5(b) allows $\mathcal{B}_{n,U}$ to be a strict subset of $\mathcal{B}_{+,U}$ with probability that goes to one.)

By the definition of $\lambda_{0,U}(k, m, \alpha)$ in (62), the right-hand side is greater than or equal to $1 - \alpha$, which completes the proof for case (ii).

The proof for case (iii) is analogous to that for case (ii).

It remains to prove the result of the Theorem for case (iv). The result of (96) holds with U replaced by L throughout and with $\beta_0 \leq \widetilde{\beta}_{n,U}$ and $\beta_{n,U} \leq \widetilde{\beta}_{n,U}$ replaced by $\widetilde{\beta}_{n,L} \leq \beta_0$ and $\widetilde{\beta}_{n,L} \leq \beta_{n,L}$, respectively (using the fact that $\widetilde{\beta}_{n,L}$ is defined by taking the inf over $\beta_n(\theta)$

values rather than the sup) by the same argument as in (96). Combining this result with (96) gives

$$\begin{aligned}
& P(\widetilde{\beta}_{n,L} \leq \beta_0 \leq \widetilde{\beta}_{n,U}) \\
\geq\ & P(\widetilde{\beta}_{n,L} \leq \beta_{n,L},\ \beta_{n,U} \leq \widetilde{\beta}_{n,U}) \\
\geq\ & P\Big(\widetilde{c}_{n,U}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \geq 0\ \forall (k,m) \in \widehat{\mathcal{B}}_{n,U}, \\
& \widetilde{c}_{n,L}(k, \widehat{\gamma}_{n,k,m}, \theta_{n,L}) \geq 0\ \forall (k,m) \in \widehat{\mathcal{B}}_{n,L}, \Big). \qquad (100)
\end{aligned}$$

Analogous results to those of (97)-(99) hold with U replaced by L throughout. Furthermore, doing the analysis for U and L simultaneously shows that the right-hand side of (100) equals the right-hand side of (99) because $\beta_{+,U} = \beta_{+,L}$ implies that $\theta_{+,U} = \theta_{+,L}$ and $\mathcal{B}_{+,U} = \mathcal{B}_{+,L}$ using the definitions of $\beta_{+,U}$ and $\theta_{+,U}$ in Assumption CI4(b) and the definition of $\mathcal{B}_{+,U}$ in (61). This completes the proof for case (iv).

Now, we establish the desired result for the CI for the identified interval B_+. The proof is the same as that of case (iv) for the CI for β_0 except that $\beta_{n,U}$ does not necessarily equal $\beta_{n,L}$, $\theta_{n,U}$ does not necessarily equal $\theta_{n,L}$, and $\mathcal{B}_{n,U}$ does not necessarily equal $\mathcal{B}_{n,L}$. In consequence, using Assumption CI6 with the condition in (63), the right-hand side of (100) equals the expression in (63) rather than the right-hand side of (99). □

Proof of Lemma 5. First, we establish part (a). Combining Assumptions CI1, CI4(b), and 6 gives

$$\begin{pmatrix} \widehat{\nu}_n(\cdot,\cdot,\cdot) \\ \widehat{\Gamma}_n \\ \theta_{n,U} \end{pmatrix} \Rightarrow \begin{pmatrix} \nu_0(\cdot,\cdot,\cdot) \\ \Gamma_0 \\ \theta_{+,U} \end{pmatrix} \qquad (101)$$

as processes indexed by $(k, \gamma, \theta) \in \mathcal{I}_K \times \Gamma_{all} \times \Theta$ (and the convergence is joint with that in Assumption CI2). The continuous mapping theorem, e.g., see Pollard (1984), now gives

$$\widehat{\nu}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) \rightarrow_d \nu_0(k, \gamma_{0,k,m}, \theta_{+,U}) \qquad (102)$$

for all $(k,m) \in \mathcal{I}_{K,M}$ and the convergence is joint. The function $g(\nu(\cdot,\cdot,\cdot), \Gamma, \theta) = \nu(k, \gamma_{k,m}, \theta)$ is continuous at $(\nu_0(\cdot,\cdot,\cdot), \Gamma_0, \theta_{+,U})$ because $\nu_0(k,\cdot,\cdot)$ has continuous sample paths a.s. with respect to the product of the $||\cdot||_G$ norm and the Euclidean norm.

Part (b) holds by Assumptions CI2 and CI4(b) and the continuous mapping theorem.

Next, we prove part (c). Using the triangle inequality and Assumption CI3, we have

$$\begin{aligned}
& |\widehat{w}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) - w_0(k, \gamma_{0,k,m}, \theta_{+,U})| \\
\leq\ & |\widehat{w}_n(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) - w_0(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U})| \\
& + |w_0(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) - w_0(k, \gamma_{0,k,m}, \theta_{+,U})| \\
\leq\ & \sup_{\gamma \in \Gamma_{all}, \theta \in \Theta} |\widehat{w}_n(k, \gamma, \theta) - w_0(k, \gamma, \theta)| + |w_0(k, \widehat{\gamma}_{n,k,m}, \theta_{n,U}) - w_0(k, \gamma_{0,k,m}, \theta_{+,U})| \\
\rightarrow_p\ & 0 \qquad (103)
\end{aligned}$$

for all $(k,m) \in \mathcal{I}_{K,M}$.

Combining parts (a)-(c) of the Lemma and Assumption CI6 gives part (d). □

References

Berry, S. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 60(4), 889–917.

Berry, S. T. (1989): "Entry in the Airline Industry," Ph.D. thesis, Univ. of Wisconsin.

Bresnahan, T., and P. Reiss (1988): "Do Entry Conditions Vary Across Markets," *Brookings Papers in Economic Activities: Microeconomic Annual*, 1, 833–882.

——— (1991): "Empirical Models of Discrete Games," *Journal of Econometrics*, 48, 57–82.

Chernozhukov, V., H. Hong, and E. Tamer (2003): "Parameter Set Inference in a Class of Econometric Models," Discussion paper, Princeton Univ.

Ciliberto, F., and E. Tamer (2003): "Market Structure and Multiple Equilibrium in Airline Markets," Working paper, Princeton University.

Imbens, G., and C. Manski (2003): "Confidence Intervals for Partially Identified Parameters," Discussion paper, UC – Berkeley.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Mazzeo, J. M. (2002): "Product Choice and Oligopoly Market Structure," *RAND Journal of Economics*, 33(2), 221–242.

Seim, K. (2001): "Spatial Differentiation and Market Structure: The Video Retail Industry," Ph.D. thesis, Yale University.

Sutton, J. (2000): *Marshall's tendencies: What can economists know?* MIT Press, Cambridge and London.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70(1), 147–167.