Molinari, Francesca

Working Paper

Econometrics with partial identification

cemmap working paper, No. CWP25/19

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Molinari, Francesca (2019) : Econometrics with partial identification, cemmap working paper, No. CWP25/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.2519

This Version is available at:
https://hdl.handle.net/10419/211118

cemmap
centre for microdata methods and practice

Econometrics with Partial Identification

Francesca Molinari

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP25/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Econometrics with Partial Identification

Francesca Molinari
Cornell University
Department of Economics
fm72@cornell.edu*

May 30, 2019

Abstract

Econometrics has traditionally revolved around *point identification*. Much effort has been devoted to finding the weakest set of assumptions that, together with the available data, deliver point identification of population parameters, finite or infinite dimensional that these might be. And point identification has been viewed as a necessary prerequisite for meaningful statistical inference. The research program on *partial identification* has begun to slowly shift this focus in the early 1990s, gaining momentum over time and developing into a widely researched area of econometrics. Partial identification has forcefully established that much can be learned from the available data and assumptions imposed because of their credibility rather than their ability to yield point identification. Within this paradigm, one obtains a set of values for the parameters of interest which are observationally equivalent given the available data and maintained assumptions. I refer to this set as the parameters' *sharp identification region*.

Econometrics with partial identification is concerned with: (1) obtaining a tractable characterization of the parameters' sharp identification region; (2) providing methods to estimate it; (3) conducting test of hypotheses and making confidence statements about the partially identified parameters. Each of these goals poses challenges that differ from those faced in econometrics with point identification. This chapter discusses these challenges and some of their solution. It reviews advances in partial identification analysis both as applied to learning (functionals of) probability distributions that are well-defined in the absence of models, as well as to learning parameters that are well-defined only in the context of particular models. The chapter highlights a simple organizing principle: the source of the identification problem can often be traced to a collection of random variables that are consistent with the available data and maintained assumptions. This collection may be part of the observed data or be a model implication. In either case, it can be formalized as a *random set*. Random set theory is then used as a mathematical framework to unify a number of special results and produce a general methodology to conduct econometrics with partial identification.

*I am grateful to Don Andrews, Levon Barseghyan, Federico Bugni, Ivan Canay, Joachim Freyberger, Chuck Manski, Ilya Molchanov, Áureo de Paula, Jack Porter, Seth Richards-Shubik, Adam Rosen, Shuyang Sheng, Jörg Stoye, and Elie Tamer for helpful conversations on topics appearing in this chapter, and to the National Science Foundation for financial support through grants SES-1824375 and SES-1824448. I am grateful to Louis Liu and Yibo Sun for research assistance supported by the Robert S. Hatfield Fund for Economic Education at Cornell University. Part of this research was carried out during my sabbatical leave at the Department of Economics at Duke University, whose hospitality is gratefully acknowledged.

Contents

1 Introduction

1.1 Why Partial Identification?

Knowing the population distribution that data are drawn from, what can one learn about a parameter of interest? It has long been understood that assumptions about the data generating process (DGP) play a crucial role in answering this *identification question* at the core of all empirical research. Inevitably, assumptions brought to bear enjoy a varying degree of credibility. Some are rooted in economic theory (e.g., optimizing behavior) or in information available to the researcher on the DGP (e.g., randomization mechanisms). These assumptions can be argued to be highly credible. Others are driven by concerns for tractability and the desire to answer the identification question with a certain level of precision (e.g., functional form and distributional assumptions). These are arguably less credible.

There has also been a widespread perception that in order to be useful, the answer to the identification question needs to be that the parameter of interest can be learned exactly. As a result, *point identification* has traditionally been regarded as a necessary prerequisite for meaningful statistical inference. Fundamental contributions in the econometrics and statistics literature on semiparameteric and nonparametric methods have characterized sufficient sets of assumptions, that exclude many suspect ones (sometimes as many as possible), to guarantee that point identification attains (see, e.g., Matzkin, 2007, for a recent review).

In a given application, however, some assumptions required for point identification may not be tenable. Early on, Koopmans and Reiersol (1950) cautioned against allowing the quest for point identification to drive the choice of maintained assumptions, stating (p. 169): "One might regard problems of identifiability as a necessary part of the specification problem. We would consider such a classification acceptable, provided the temptation to specify models in such a way as to produce identifiability of relevant characteristics is resisted." They then went on to recommend that restrictions should be imposed based on prior knowledge of the phenomenon under analysis and some criteria of simplicity, but not on the desire to achieve point identification of a parameter that the researcher happens to be interested in.

This principle is systematically embodied in the research program on *partial identification* analysis that was put forward by Chuck Manski and developed by several authors since the early 1990s (starting with Manski, 1989). Earlier important contributions exist, but had remained fragmented and unable to shift the point identification paradigm.[1] Manski forcefully argued that identification is not an "all or nothing" concept: much can be learned about parameters of interest from the available data and credible assumptions, even if not everything. He proposed that empirical analysis begin by asking what can the data alone reveal about the parameters of interest. This is a nonparametric approach that dispenses with all

[1] Examples include Frisch (1934), Reiersol (1941), Marschak and Andrews (1944), Fréchet (1951), Duncan and Davis (1953), Peterson (1976), Klepper and Leamer (1984), Leamer (1987), Jovanovic (1989), and Phillips (1989).

assumptions, except basic restrictions on the sampling process such that the distribution of the observable variables can be learned as data accumulate. In subsequent steps, one incorporates additional assumptions into the analysis, reporting how each assumption (or set of assumptions) affects what one can learn about the parameter of interest. Point identification may result from the process of increasingly strengthening the maintained assumptions, but it is not the goal in itself. Rather, the objective is to make transparent the relative role played by the data and the assumptions in shaping the inference that one draws.

The resulting partial identification paradigm yields a shift of focus to the parameter's *sharp identification region*: the collection of values that can generate the same distribution of observables as the one in the data for some DGP consistent with the maintained assumptions (the *observationally equivalent* values).

While the first reactions to partial identification were tepid, the paradigm gained momentum over time, developing into a widely researched area of econometrics and a valued approach to empirical research in economics and more broadly.

1.2 Goals and Structure of this Chapter

In order to carry out econometric analysis with partial identification, one needs: (1) computationally feasible characterizations of the parameters' sharp identification region; (2) methods to estimate this region; and (3) methods to test hypotheses and construct confidence sets. The goal of this chapter is to provide insights into the challenges posed by each of these desiderata, and into some of their solutions. In order to discuss these issues in some level of detail while keeping the chapter to a manageable length, I focus on a selection of papers and not on a complete survey of the literature. As a consequence, many relevant contributions are left out of the presentation and the references. I also do not discuss the important but separate topic of statistical decisions in the presence of partial identification, for which I refer to the textbook treatment in Manski (2005) and to the review by Hirano and Porter (2019, Chapter XXX in this Volume).

The presumption in identification analysis that the distribution from which the data are drawn is known allows one to keep separate the identification question from the distinct question of statistical inference from a finite sample. I use the same separation in this chapter. I assume solid knowledge of the topics covered in first year Economics PhD courses in econometrics and microeconomic theory.

I begin in Section 2 with the analysis of what can be learned about features of probability distributions that are well defined in the absence of an economic model, such as moments, quantiles, cumulative distribution functions, etc., when one faces measurement problems. Specifically, I focus on cases where the *data is incomplete*, either due to sample selection or to interval measurements. I lay out formally the maintained assumptions for several examples, and then discuss in detail what is the source of the identification problem. I conclude with

providing tractable characterizations of what can be learned about the parameters of interest, with formal proofs. I show that even in simple problems, great care may be needed to obtain the sharp identification region. It is often easier to characterize an *outer region*, i.e., a collection of values for the parameter of interest that contains the sharp one but may contain also additional values. Outer regions are useful because of their simplicity and because in certain applications they may suffice to answer questions of great interest, e.g., whether a policy intervention has a nonnegative effect. However, compared to the sharp identification region they may afford the researcher less useful predictions, and a lower ability to test for misspecification, because they do not harness all the information in the observed data and maintained assumptions.

In Section 3, I use the same approach to study what can be learned about features of parameters of structural econometric models when the *model is incomplete* (Tamer, 2003; Haile and Tamer, 2003; Ciliberto and Tamer, 2009). Specifically, I discuss single agent discrete choice models under a variety of challenging situations (interval measured as well as endogenous explanatory variables; unobserved as well as counterfactual choice sets); finite discrete games with multiple equilibria; auction models under weak assumptions on bidding behavior; and network formation models. Again I formally derive sharp identification regions for several examples.

I conclude each of these sections with a brief discussion of further theoretical advances and empirical applications that is meant to give a sense of the breadth of the approach, but not to be exhaustive. I refer to the recent survey by Ho and Rosen (2017) for a thorough discussion of empirical applications of partial identification methods.

In Section 4, I discuss finite sample inference. I limit myself to highlighting the challenges that one faces for consistent estimation when the identified object is a set, and several coverage notions and requirements that have been proposed over the last 20 years. I refer to the recent survey by Canay and Shaikh (2017) for a thorough discussion of methods to tests hypotheses and build confidence sets in moment inequality models.

In Section 5, I discuss the distinction between refutable and non-refutable assumptions, and how model misspecification may be detectable in the presence of the former, even within the partial identification paradigm. I then highlight certain challenges that model misspecification presents for the interpretation of sharp identification (as well as outer) regions, and for the construction of confidence sets.

In Section 6, I highlight that while most of the sharp identification regions characterized in Section 2 can be easily computed, many of the ones in Section 3 are more challenging. This is because the latter are obtained as level sets of criterion functions in moderately dimensional spaces, and tracing out these level sets or their boundaries is a non-trivial computational problem. In Section 7, I conclude providing some considerations on what I view as open questions for future research.

I refer to Manski (1995, 2003, 2007a) for textbook treatments of partial identification of

probability distributions, and to Lewbel (2018) for a careful presentation of the many notions of identification that are used in the literature.[2]

1.3 Random Set Theory as a Tool for Partial Identification Analysis

Throughout Sections 2 and 3, a simple organizing principle for much of partial identification analysis emerges. The cause of the identification problems discussed can be traced back to a collection of random variables that are consistent with the available data and maintained assumptions. For the problems studied in Section 2, this set is often a simple function of the observed variables. The incompleteness of the data stems from the fact that instead of observing the singleton variables of interest, one observes set-valued variables to which these belong, but one has no information on their exact value within the sets. For the problems studied in Section 3, the collection of random variables consistent with the maintained assumptions comprises what the model predicts for the endogenous variable(s). The incompleteness of the model stems from the fact that instead of making a singleton prediction for the variable(s) of interest, the model makes multiple predictions but does not specify how one is chosen.

The central role of set-valued objects, both stochastic and nonstochastic, in partial identification renders *random set theory* a natural toolkit to aid the analysis.[3] This theory originates in the seminal contributions of Choquet (1953/54), Aumann (1965), and Debreu (1967), with the first self contained treatment of the theory given by Matheron (1975). I refer to Molchanov (2017) for a textbook presentation, and to Molchanov and Molinari (2014, 2018) for a treatment focusing on its applications in econometrics.

Beresteanu and Molinari (2008) introduce the use of random set theory in econometrics to carry out identification analysis and statistical inference with incomplete data. Beresteanu, Molchanov, and Molinari (2011, 2012) propose it to characterize sharp identification regions both with incomplete data and with incomplete models. Galichon and Henry (2011) propose the use of optimal transportation methods that in some applications deliver the same characterizations as the random set methods. I do not discuss optimal transportation methods in this chapter, but refer to Galichon (2016) for a thorough treatment.

Over the last ten years, random set methods have been used to unify a number of specific results in partial identification, and to produce a general methodology for identification analysis that dispenses completely with case-by-case distinctions. In particular, as I show throughout the chapter, the methods allow for simple and tractable characterizations of sharp identification regions. The collection of these results establishes that indeed this is a useful tool to carry out econometrics with partial identification, as exemplified by its prominent role both in this chapter and in Chapter XXX in this Volume by Chesher and Rosen (2019), which focuses on general classes of instrumental variable models. The random sets approach

[2] Lewbel (2018) also provides an important historical account of how these notions developed over time.

[3] Random elements whose realizations are sets appeared a long time ago in statistics and econometrics in the form of confidence regions, which can be naturally described as random sets. Their role here is different.

complements the more traditional one, based on mathematical tools for (single valued) random vectors, that proved extremely productive since the beginning of the research program in partial identification.

This chapter shows that to fruitfully apply random set theory for identification and inference, the econometrician needs to carry out three fundamental steps. First, she needs to define the random closed set that is relevant for the problem under consideration using all information given by the available data and maintained assumptions. This is a delicate task, but one that is typically carried out in identification analysis regardless of whether random set theory is applied. Indeed, throughout the chapter I highlight how relevant random closed sets were characterized in partial identification analysis since the early 1990s, albeit the connection to the theory of random sets was not made. As a second step, the econometrician needs to determine how the observable random variables relate to the random closed set. Often, one of two cases occurs: either the observable variables determine a random set to which the unobservable variable of interest belongs with probability one, as in incomplete data scenarios; or the (expectation of the) (un)observable variable belongs to (the expectation of) a random set determined by the model, as in incomplete model scenarios. Finally, the econometrician needs to determine which tool from random set theory should be utilized. To date, new applications of random set theory to econometrics have fruitfully exploited (Aumann) expectations and their support functions, (Choquet) capacity functionals, and laws of large numbers and central limit theorems for random sets. Appendix A reports basic definitions and results from random set theory defining these concepts, as well as some useful theorems. The chapter explains in detail through applications to important identification problems how these steps can be carried out.

1.4 Notation

This chapter employs consistent notation that is summarized in Table 1.1. Some important conventions are as follows: $\boldsymbol{y}$ denotes outcome variables, $(\boldsymbol{x}, \boldsymbol{w})$ denote explanatory variables, and $\boldsymbol{z}$ denotes instrumental variables (i.e., variables that satisfy some form of independence with the outcome or with the unobservable variables, possibly conditional on $\boldsymbol{x}, \boldsymbol{w}$).

I denote by P the joint distribution of all observable variables. Identification analysis is carried out using the information contained in this distribution, and finite sample inference is carried out under the presumption that one draws a random sample of size n from P. I denote by Q the joint distribution whose features the researcher wants to learn. If Q were identified given the observed data (e.g., if it were a marginal of P), point identification of the parameter or functional of interest would attain. I denote by R the joint distribution of all variables, observable and unobservable ones; both P and Q can be obtained from it. I use S to denote any distribution that is not revealed by the data, other than Q; this also can be obtained from R. In the context of structural models, I denote by M a distribution for

Table 1.1: Notation Used

$(\Omega, \mathfrak{F}, \mathbb{P})$	Nonatomic probability space
$\mathbb{R}^d, \|\cdot\|$	Euclidean space equipped with the Euclidean norm
$\mathcal{F}, \mathcal{G}, \mathcal{K}$	Collection of closed, open, and compact subsets of $\mathbb{R}^d$ (respectively)
$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$	Unit sphere in $\mathbb{R}^d$
$\mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$	Unit ball in $\mathbb{R}^d$
$\text{conv}(A)$	Convex hull of a set $A \subset \mathbb{R}^d$
$\text{cl}(A)$	Closure of a set $A \subset \mathbb{R}^d$
$\|A\|$	Cardinality of a finite set $A \subset \mathbb{R}^d$
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$	Random vectors
$x, y, z, \ldots$	Realizations of random vectors or deterministic vectors
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$	Random sets
$X, Y, Z, \ldots$	Realizations of random sets or deterministic sets
$\epsilon, \varepsilon, \nu, \zeta$	Unobserved random variables (heterogeneity)
θ, Θ	Parameter vector and its parameter space
R	Joint distribution of all variables (observable and unobservable)
P	Joint distribution of the observable variables
Q	Joint distribution whose feature one wants to learn
S	Other distribution or functional of R, not revealed by the observable data
$\mathbb{E}_\tau$	Expectation operator associated with distribution $\tau \in \{\mathsf{R}, \mathsf{P}, \mathsf{Q}, \mathsf{S}\}$
$q_\tau(\alpha)$	Quantile function at level $\alpha \in (0,1)$ for a random variable distributed τ
$\tau(K)$	Probability that distribution τ assigns to set K
$\mathsf{T}_{\boldsymbol{X}}(K) = \mathbb{P}\{\boldsymbol{X} \cap K \neq \emptyset\}$, $K \in \mathcal{K}$	Capacity functional of random set $\boldsymbol{X}$
$\mathsf{C}_{\boldsymbol{X}}(F) = \mathbb{P}\{\boldsymbol{X} \subset F\}$, $F \in \mathcal{F}$	Containment functional of random set $\boldsymbol{X}$
$\xrightarrow{p}$	Convergence in probability
$\xrightarrow{\text{a.s.}}$	Convergence almost surely
$\boldsymbol{x} \overset{d}{=} \boldsymbol{y}$	$\boldsymbol{x}$ and $\boldsymbol{y}$ have the same distribution
$\Rightarrow$	Weak convergence
$\mathcal{H}_{\mathsf{P}}[\cdot]$	Sharp identification region of the functional in square brackets (a function of P)
$\mathcal{O}_{\mathsf{P}}[\cdot]$	An outer region of the functional in square brackets (a function of P)

the observable variables that is consistent with the model. I note that model incompleteness typically implies that M is not unique. I let $\mathcal{H}_{\mathsf{P}}[\cdot]$ denote the sharp identification region of the functional in square brackets, and $\mathcal{O}_{\mathsf{P}}[\cdot]$ an outer region. In both cases, the regions are indexed by P, because they depend on the distribution of the observed data.

2 Partial Identification of Probability Distributions

The partial identification approach to empirical research finds its genesis in Manski's analysis of what can be learned about functionals of probability distributions that are well-defined in the absence of a model. The proposed approach is nonparametric, and it is typically *constructive*, in the sense that it leads to "plug-in" formulae for the bounds on the functionals of interest. In this section I review the first partial identification problem studied by Manski, and discuss several important extensions of his original idea.

2.1 Selectively Observed Data

The basic features of Manski's nonparametric bounds are clearly put forward in the analysis of the following identification problem, carried out in Manski (1989).

IDENTIFICATION PROBLEM 2.1 (Conditional Expectation of Selectively Observed Data): Let $\boldsymbol{y} \in \mathcal{Y} \subset \mathbb{R}$ and $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ be, respectively, an outcome variable and a vector of covariates with support $\mathcal{Y}$ and $\mathcal{X}$ respectively, with $\mathcal{Y}$ a compact set. Let $\boldsymbol{d} \in \{0,1\}$. Suppose that the researcher observes a random sample of realizations of $(\boldsymbol{x}, \boldsymbol{d})$ and, in addition, observes the realization of $\boldsymbol{y}$ when $\boldsymbol{d} = 1$. Hence, the observed data is $(\boldsymbol{yd}, \boldsymbol{d}, \boldsymbol{x}) \sim \mathsf{P}$. Let G be the space of measurable functions that map $\mathcal{Y}$ into $\mathbb{R}$ and that attain their lower and upper bounds $g_0 = \min_{y \in \mathcal{Y}} g(y)$ and $g_1 = \max_{y \in \mathcal{Y}} g(y)$, and assume that $-\infty < g_0 < g_1 < \infty$. Let g be a function in G, so that there exists a $y_j \in \mathcal{Y}$ such that $g(y_j) = g_j$, $j = 0, 1$. In the absence of additional information, what can the researcher learn about $\mathbb{E}_{\mathsf{Q}}(g(\boldsymbol{y})|\boldsymbol{x} = x)$, with Q the distribution of $(\boldsymbol{y}, \boldsymbol{x})$? △

Manski's analysis of this problem begins with a simple application of the law of total probability, that yields

$$\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x) = \mathsf{P}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) + \mathsf{S}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 0)\mathsf{P}(\boldsymbol{d} = 0|\boldsymbol{x} = x). \quad (2.1)$$

Equation (2.1) lends a simple but powerful anatomy of the selection problem. While $\mathsf{P}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 1)$ and $\mathsf{P}(\boldsymbol{d}|\boldsymbol{x} = x)$ can be learned from the observable distribution $\mathsf{P}(\boldsymbol{yd}, \boldsymbol{d}, \boldsymbol{x})$, under the maintained assumptions the sampling process reveals nothing about $\mathsf{S}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 0)$. Hence, $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)$ is not point identified.

If one were to assume *exogenous selection* (or data missing at random conditional on $\boldsymbol{x}$), i.e., $\mathsf{S}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{d} = 0) = \mathsf{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{d} = 1)$, point identification would obtain. However, that assumption is non-refutable and it is well known that it may fail in applications.[4] Let $\mathcal{T}$ denote the space of all probability measures with support in $\mathcal{Y}$. The unknown functional vector is $\{\tau(x), \upsilon(x)\} \equiv \{\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x), \mathsf{S}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 0)\}$. What the researcher can learn, in the absence of additional restrictions on $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{d} = 0)$, is the region of *observationally equivalent* distributions for $\boldsymbol{y}|\boldsymbol{x} = x$, and the associated set of expectations taken with respect to these distributions.

THEOREM SIR-2.1 (Conditional Expectations of Selectively Observed Data): *Under the assumptions in Identification Problem 2.1,*

$$\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(g(\boldsymbol{y})|\boldsymbol{x} = x)] = \Big[\mathbb{E}_{\mathsf{P}}(g(\boldsymbol{y})|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) + g_0 P(\boldsymbol{d} = 0|\boldsymbol{x} = x),$$
$$\mathbb{E}_{\mathsf{P}}(g(\boldsymbol{y})|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) + g_1 \mathsf{P}(\boldsymbol{d} = 0|\boldsymbol{x} = x)\Big] \quad (2.2)$$

[4] Section 5 discusses the consequences of model misspecification (with respect to refutable assumptions).

is the sharp identification region for $\mathbb{E}_{\mathsf{Q}}(g(\boldsymbol{y})|\boldsymbol{x}=x)$.

Proof. Due to the discussion following equation (2.1), the collection of observationally equivalent distribution functions for $\boldsymbol{y}|\boldsymbol{x}=x$ is

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)] = \Big\{\tau(x)\in\mathcal{T} : \tau(x) = \mathsf{P}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x) + \upsilon(x)\mathsf{P}(\boldsymbol{d}=0|\boldsymbol{x}=x),\ \text{for some } \upsilon(x)\in\mathcal{T}\Big\}. \quad (2.3)$$

Next, observe that the lower bound in equation (2.2) is achieved by integrating $g(\boldsymbol{y})$ against the distribution $\tau(x)$ that results when $\upsilon(x)$ places probability one on y_0. The upper bound is achieved by integrating $g(\boldsymbol{y})$ against the distribution $\tau(x)$ that results when $\upsilon(x)$ places probability one on y_1. Both are contained in the set $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]$ in equation (2.3). □

There are the *worst case bounds*, so called because assumptions free and therefore representing the widest possible range of values for the parameter of interest that are consistent with the observed data. A simple "plug-in" estimator for $\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(g(\boldsymbol{y})|\boldsymbol{x}=x)]$ replaces all unknown quantities in (2.2) with consistent estimators, obtained, e.g., by kernel or sieve regression. I return to consistent estimation of partially identified parameters in Section 4. Here I emphasize that identification problems are fundamentally distinct from finite sample inference problems. The latter are typically reduced as sample size increase (because, e.g., the variance of the estimator becomes smaller). The former do not improve, unless a different and better type of data is collected, e.g. with a smaller prevalence of missing data (see Dominitz and Manski, 2017, for a discussion).

Manski (2003, Section 1.3) shows that the proof of Theorem SIR-2.1 can be extended to obtain the smallest and largest points in the sharp identification region of any parameter that respects stochastic dominance.[5] This is especially useful to bound the quantiles of $\boldsymbol{y}|\boldsymbol{x}=x$. For any given $\alpha\in(0,1)$, let $q_{\mathsf{P}}^{g(\boldsymbol{y})}(\alpha,1,x) \equiv \{\min t : \mathsf{P}(g(\boldsymbol{y})\leq t|\boldsymbol{d}=1,\boldsymbol{x}=x)\geq\alpha\}$. Then the smallest and largest admissible values for the α-quantile of $\boldsymbol{y}|\boldsymbol{x}=x$ are, respectively,

$$r(\alpha,x) \equiv \begin{cases} q_{\mathsf{P}}^{g(\boldsymbol{y})}\left(\left[1-\frac{(1-\alpha)}{\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)}\right],1,x\right) & \text{if } \mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)>1-\alpha, \\ g_0 & \text{otherwise;} \end{cases}$$

$$s(\alpha,x) \equiv \begin{cases} q_{\mathsf{P}}^{g(\boldsymbol{y})}\left(\left[\frac{\alpha}{\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)}\right],1,x\right) & \text{if } \mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)\geq\alpha, \\ g_1 & \text{otherwise.} \end{cases}$$

The lower bound on $\mathbb{E}_{\mathsf{Q}}(g(\boldsymbol{y})|\boldsymbol{x}=x)$ is informative only if $g_0>-\infty$, and the upper bound is informative only if $g_1<\infty$. By comparison, for any value of α, $r(\alpha,x)$ and $s(\alpha,x)$ are

[5] Recall that a probability distribution $\mathsf{F}\in\mathcal{T}$ stochastically dominates $\mathsf{F}'\in\mathcal{T}$ if $\mathsf{F}(-\infty,t]\leq\mathsf{F}'(-\infty,t]$ for all $t\in\mathbb{R}$. A real-valued functional $\mathsf{d}:\mathcal{T}\to\mathbb{R}$ respects stochastic dominance if $\mathsf{d}(\mathsf{F})\geq\mathsf{d}(\mathsf{F}')$ whenever F stochastically dominates F'.

generically informative if, respectively, $\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) > 1 - \alpha$ and $\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) \geq \alpha$, regardless of the range of g.

Stoye (2010) further extends partial identification analysis to the study of spread parameters in the presence of missing data (as well as interval data, data combinations, and other applications). These parameters include ones that respect second order stochastic dominance, such as the variance, the Gini coefficient, and other inequality measures, as well as other measures of dispersion which do not respect second order stochastic dominance, such as interquartile range and ratio. Stoye shows that the sharp identification region for these parameters can be obtained by fixing the mean or quantile of the variable of interest at a specific value within its sharp identification region, and deriving a distribution consistent with this value which is "compressed" with respect to the ones which bound the cumulative distribution function (CDF) of the variable of interest, and one which is "dispersed" with respect to them. Heuristically, the compressed distribution minimizes spread, while the dispersed one maximizes it (the sense in which this optimization occurs is formally defined in the paper). The intuition for this is that a compressed CDF is first below and then above any non-compressed one; a dispersed CDF is first above and then below any non-dispersed one. Second-stage optimization over the possible values of the mean or the quantile delivers unconstrained bounds. The main results of the paper are sharp identification regions for the expectation and variance, for the median and interquartile ratio, and for many other combinations of parameters.

KEY INSIGHT 2.1 (Identification is not a binary event): *Identification Problem 2.1 is mathematically simple, but it puts forward a completely new approach to empirical research. The traditional approach aims at finding a sufficient (possibly minimal) set of assumptions guaranteeing point identification of parameters, viewing identification as an "all or nothing" notion, where either the functional of interest can be learned exactly or nothing of value can be learned. The partial identification approach pioneered by Manski (1989) points out that much can be learned from combination of data and assumptions that restrict the functionals of interest to a set of observationally equivalent values, even if this set is not a singleton. Along the way, Manski (1989) points out that in Identification Problem 2.1 the observed outcome is the singleton $\boldsymbol{y}$ when $\boldsymbol{d} = 1$, and $\mathcal{Y}$ when $\boldsymbol{d} = 0$. This is a random closed set, see Definition A.1. I return to this connection in Section 2.3.*

Despite how transparent the framework in Identification Problem 2.1 is, important subtleties arise even in this seemingly simple context. For a given $t \in \mathbb{R}$, consider the function $g(\boldsymbol{y}) = \mathbf{1}(\boldsymbol{y} \leq t)$, with $\mathbf{1}(A)$ the indicator function taking the value one if the logical condition in parentheses holds and zero otherwise. Then equation (2.2) yields *pointwise-sharp* bounds

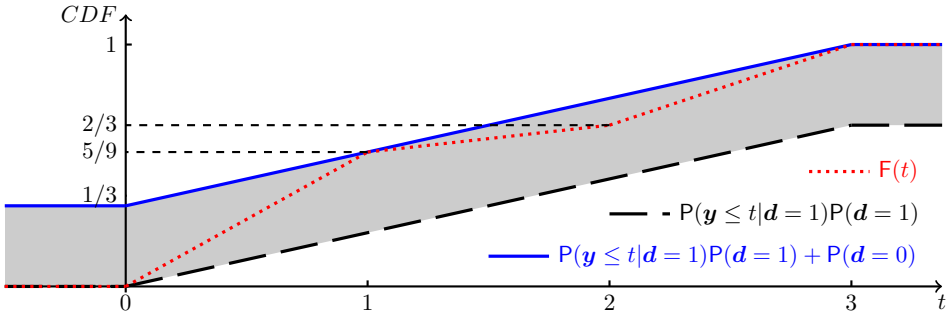


Figure 2.1: The tube defined by inequalities (2.4) in the set-up of Example 2.1, and the CDF in (2.7).

on the CDF of $\boldsymbol{y}$ at any fixed $t \in \mathbb{R}$:

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y} \leq t|\boldsymbol{x} = x)] = [\mathsf{P}(\boldsymbol{y} \leq t|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x)\ , \\ \mathsf{P}(\boldsymbol{y} \leq t|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) + \mathsf{P}(\boldsymbol{d} = 0|\boldsymbol{x} = x)]\,. \quad (2.4)$$

Yet, the collection of CDFs that belong to the band defined by (2.4) is *not* the sharp identification region for the CDF of $\boldsymbol{y}|\boldsymbol{x} = x$. Rather, it constitutes an *outer region*, as originally pointed out by Manski (1994, p. 148 and note 2).

THEOREM OR-2.1 (Cumulative Distribution Function of Selectively Observed Data): *Let $\mathcal{C}$ denote the collection of cumulative distribution functions on $\mathcal{Y}$. Then, under the assumptions in Identification Problem 2.1,*

$$\mathcal{O}_{\mathsf{P}}[\mathsf{F}(\boldsymbol{y}|\boldsymbol{x} = x)] = \{\mathsf{F} \in \mathcal{C}:\ \mathsf{P}(\boldsymbol{y} \leq t|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) \leq \mathsf{F}(t|x) \leq \\ \mathsf{P}(\boldsymbol{y} \leq t|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x) + \mathsf{P}(\boldsymbol{d} = 0|\boldsymbol{x} = x)\ \forall t \in \mathbb{R}\} \quad (2.5)$$

is an outer region for the CDF of $\boldsymbol{y}|\boldsymbol{x} = x$.

Proof. Any admissible CDF for $\boldsymbol{y}|\boldsymbol{x} = x$ belongs to the family of functions in equation (2.5). However, the bound in equation (2.5) does not impose the restriction that for any $t_0 \leq t_1$,

$$\mathsf{Q}(t_0 \leq \boldsymbol{y} \leq t_1|\boldsymbol{x} = x) \geq \mathsf{P}(t_0 \leq \boldsymbol{y} \leq t_1|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1|\boldsymbol{x} = x). \quad (2.6)$$

This restriction is implied by the maintained assumptions, but is not necessarily satisfied by all CDFs in $\mathcal{O}_{\mathsf{P}}[\mathsf{F}(\boldsymbol{y}|\boldsymbol{x} = x)]$, as illustrated in the following simple example. □

Example 2.1. Omit $\boldsymbol{x}$ for simplicity, let $\mathsf{P}(\boldsymbol{d}=1)=\frac{2}{3}$, and let

$$\mathsf{P}(\boldsymbol{y}\leq t|\boldsymbol{d}=1)\begin{cases}0 & \text{if } t<0,\\ \frac{1}{3}t & \text{if } 0\leq t<3,\\ 1 & \text{if } t\geq 3.\end{cases}$$

The bounding functions and associated tube from the inequalities in (2.4) are depicted in Figure 2.1. Consider the cumulative distribution function

$$\mathsf{F}(t)=\begin{cases}0 & \text{if} \quad t<0,\\ \frac{5}{9}t & \text{if} \quad 0\leq t<1,\\ \frac{1}{9}t+\frac{4}{9} & \text{if} \quad 1\leq t<2,\\ \frac{1}{3}t & \text{if} \quad 2\leq t<3,\\ 1 & \text{if} \quad t\geq 3.\end{cases} \tag{2.7}$$

For each $t\in\mathbb{R}$, $\mathsf{F}(t)$ lies in the tube defined by equation (2.4). However, it cannot be the CDF of $\boldsymbol{y}$, because $\mathsf{F}(2)-\mathsf{F}(1)=\frac{1}{9}<\mathsf{P}(1\leq\boldsymbol{y}\leq 2|\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1)$, directly contradicting equation (2.6). △

How can one characterize the sharp identification region for the CDF of $\boldsymbol{y}|\boldsymbol{x}=x$ under the assumptions in Identification Problem 2.1? In general, there is not a single answer to this question: different methodologies can be used. Here I use results in Manski (2003, Corollary 1.3.1) and Molchanov and Molinari (2018, Theorem 2.25), which yield an alternative characterization of $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]$ that translates directly into a characterization of $\mathcal{H}_{\mathsf{P}}[\mathsf{F}(\boldsymbol{y}|\boldsymbol{x}=x)]$.[6]

Theorem SIR-2.2 (Conditional Distribution and CDF of Selectively Observed Data): *Under the assumptions in Identification Problem 2.1,*

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]=\Big\{\tau(x)\in\mathcal{T}:\tau_K(x)\geq\mathsf{P}(\boldsymbol{y}\in K|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x),\ \forall K\subset\mathcal{Y}\Big\}, \tag{2.8}$$

where K is measurable. If $\mathcal{Y}$ is countable,

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]=\Big\{\tau(x)\in\mathcal{T}:\tau_y(x)\geq\mathsf{P}(\boldsymbol{y}=y|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x),\ \forall y\in\mathcal{Y}\Big\}. \tag{2.9}$$

[6] Whereas Manski (1994) is very clear that the collection of CDFs in (2.4) is an outer region for the CDF of $\boldsymbol{y}|\boldsymbol{x}=x$, and Manski (2003) provides the sharp characterization in (2.8), Manski (2007a, p. 39) does not state all the requirements that characterize $\mathcal{H}_{\mathsf{P}}[\mathsf{F}(\boldsymbol{y}|\boldsymbol{x}=x)]$.

If $\mathcal{Y}$ is a bounded interval,

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)] = \Big\{\tau(x) \in \mathcal{T} : \tau_{[t_0,t_1]}(x) \geq \mathsf{P}(t_0 \leq \boldsymbol{y} \leq t_1|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x),\ \forall t_0 \leq t_1, t_0, t_1 \in \mathcal{Y}\Big\}. \quad (2.10)$$

Proof. The characterization in (2.8) follows from equation (2.3), observing that if $\tau(x) \in \mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]$ as defined in equation (2.3), then there exists a distribution $\upsilon(x) \in \mathcal{T}$ such that $\tau(x) = \mathsf{P}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x) + \upsilon(x)\mathsf{P}(\boldsymbol{d}=0|\boldsymbol{x}=x)$. Hence, by construction $\tau_K(x) \geq \mathsf{P}(\boldsymbol{y} \in K|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)$, $\forall K \subset \mathcal{Y}$. Conversely, if one has $\tau_K(x) \geq \mathsf{P}(\boldsymbol{y} \in K|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)$, $\forall K \subset \mathcal{Y}$, one can define $\upsilon(x) = \frac{\tau(x)-\mathsf{P}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)}{\mathsf{P}(\boldsymbol{d}=0|\boldsymbol{x}=x)}$. The resulting $\upsilon(x)$ is a probability measure, and hence $\tau(x) \in \mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x}=x)]$ as defined in equation (2.3). When $\mathcal{Y}$ is countable, if $\tau_y(x) \geq \mathsf{P}(\boldsymbol{y}=y|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x)$ it follows that for any $K \subset \mathcal{Y}$,

$$\tau_K(x) = \sum_{y\in K}\tau_y(x) \geq \sum_{y\in K}\mathsf{P}(\boldsymbol{y}=y|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x) = \mathsf{P}(\boldsymbol{y}\in K|\boldsymbol{x}=x,\boldsymbol{d}=1)\mathsf{P}(\boldsymbol{d}=1|\boldsymbol{x}=x).$$

The result in equation (2.10) is proven in Molchanov and Molinari (2018, Theorem 2.25) using elements of random set theory, to which I return in Section 2.3. Using elements of random set theory it is also possible to show that the characterization in (2.8) requires only to check the inequalities for K the compact subsets of $\mathcal{Y}$. □

This section provides sharp identification regions and outer regions for a variety of functionals of interest. The computational complexity of these characterizations varies widely. Sharp bounds on parameters that respect stochastic dominance only require computing the parameters with respect to two probability distributions. An outer region on the CDF can be obtained by evaluating all tail probabilities of a certain distribution. A sharp identification region on the CDF requires evaluating the probability that a certain distribution assigns to all intervals. I return to computational challenges in partial identification in Section 6.

2.2 Treatment Effects with and without Instrumental Variables

The discussion of partial identification of probability distributions of selectively observed data naturally leads to the question of its implications for program evaluation. The literature on program evaluation is vast. The purpose of this section is exclusively to show how the ideas presented in Section 2.1 can be applied to learn features of treatment effects of interest, when no assumptions are imposed on treatment selection and outcomes. I also provide examples of assumptions that can be used to tighten the bounds. To keep this chapter to a manageable length, I discuss only partial identification of the average response to a treatment and of the

average treatment effect (ATE). There are many different parameters of interest. Examples include the *local average treatment effect* of Imbens and Angrist (1994) and the *marginal treatment effect* of Heckman and Vytlacil (1999, 2001, 2005). For thorough discussions of the literature on program evaluation, I refer to the textbook treatments in Manski (1995, 2003, 2007a) and Imbens and Rubin (2015), to the Handbook chapters by Heckman and Vytlacil (2007a,b) and Abbring and Heckman (2007), and to the review articles by Imbens and Wooldridge (2009) and Mogstad and Torgovitsky (2018).

Using standard notation (e.g., Neyman, 1923), let $\boldsymbol{y} : \mathbb{T} \mapsto \mathcal{Y}$ be an individual-specific response function, with $\mathbb{T} = \{0, 1, \ldots, T\}$ a finite set of mutually exclusive and exhaustive treatments, and let $\boldsymbol{s}$ denote the individual's received treatment (taking its realizations in $\mathbb{T}$).[7] The researcher observes data $(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{x}) \sim \mathsf{P}$, with $\boldsymbol{y} \equiv \boldsymbol{y}(\boldsymbol{s})$ the outcome corresponding to the received treatment $\boldsymbol{s}$, and $\boldsymbol{x}$ a vector of covariates. The outcome $\boldsymbol{y}(t)$ for $\boldsymbol{s} \neq t$ is counterfactual, and hence can be conceptualized as missing. Hence, we are in the framework of Identification Problem 2.1 and all the results from Section 2.1 apply in this context too, subject to adjustments in notation.[8] For example,

$$\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x} = x)] = \Big[\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{s} = t)\mathsf{P}(\boldsymbol{s} = t|\boldsymbol{x} = x) + y_0 P(\boldsymbol{s} \neq t|\boldsymbol{x} = x),$$
$$\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{s} = t)\mathsf{P}(\boldsymbol{s} = t|\boldsymbol{x} = x) + y_1 P(\boldsymbol{s} \neq t|\boldsymbol{x} = x)\Big], \quad (2.11)$$

where $y_0 \equiv \inf_{y \in \mathcal{Y}} y$, $y_1 \equiv \sup_{y \in \mathcal{Y}} y$. If $y_0 < \infty$ and/or $y_1 < \infty$, these *worst case bounds* are informative. When both are infinite, the data is uninformative in the absence of additional restrictions. If the researcher is interested in an Average Treatment Effect (ATE), e.g. $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_1)|\boldsymbol{x} = x) - \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_0)|\boldsymbol{x} = x)$ with $t_0, t_1 \in \mathbb{T}$, sharp worst case bounds on this quantity can be obtained by subtracting the upper bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_0)|\boldsymbol{x} = x)$ from the lower bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_1)|\boldsymbol{x} = x)$ (to get a lower bound on the ATE), and by subtracting the lower bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_0)|\boldsymbol{x} = x)$ from the upper bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_1)|\boldsymbol{x} = x)$ (to get an upper bound on the ATE). The resulting bounds have width equal to $|y_1 - y_0|$ and hence are informative only if both $y_0 < \infty$ and $y_1 < \infty$. These bounds always cover zero.

Key Insight 2.2: *How should one think about the finding on the size of the worst case bounds on the ATE? On the one hand, if both $y_0 < \infty$ and $y_1 < \infty$ the bounds are informative, because they are a strict subset of the ATE's possible realizations. On the other hand, they reveal that the data alone are silent on the sign of the ATE. This means that assumptions play a crucial role in delivering stronger conclusions about this policy relevant parameter. The partial identification approach to empirical research recommends that as assumptions are added to the analysis, one systematically reports how each contributes to shrinking the bounds,*

[7] Here the treatment response is a function only of the (scalar) treatment received by the given individual, an assumption known as *stable unit treatment value assumption* (Rubin, 1978).

[8] Beresteanu, Molchanov, and Molinari (2012) and Molchanov and Molinari (2018, Section 2.5) provide a characterization of the sharp identification region for the joint distribution of $[\boldsymbol{y}(t), t \in \mathbb{T}]$.

making transparent their role in shaping inference.

What assumptions may researchers bring to bear to learn more about treatment effects of interest? The literature has provided a wide array of well motivated and useful restrictions. Here I consider two examples. The first one entails *shape restrictions* on the treatment response function, leaving selection unrestricted. Manski (1997b) obtains bounds on treatment effects under the assumption that the response functions are monotone, semi-monotone, or concave-monotone. These restrictions are motivated by economic theory, where it is commonly presumed, e.g., that demand functions are downward sloping and supply functions are upward sloping. Let the set $\mathbb{T}$ be ordered in terms of degree of intensity. Then Manski's *monotone treatment response* assumption requires that

$$t_1 \geq t_0 \Rightarrow \mathsf{Q}(\boldsymbol{y}(t_1) \geq \boldsymbol{y}(t_0)) = 1 \quad \forall t_0, t_1 \in \mathbb{T}.$$

Under this assumption, one has that

$$\boldsymbol{y}(t) \in \begin{cases} (-\infty, \boldsymbol{y}] \cap \mathcal{Y} & \text{if } t < \boldsymbol{s}, \\ \{\boldsymbol{y}\} & \text{if } t = \boldsymbol{s}, \\ [\boldsymbol{y}, \infty) \cap \mathcal{Y} & \text{if } t > \boldsymbol{s}. \end{cases} \tag{2.12}$$

Using this information, the sharp bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x} = x)$ are

$$\begin{aligned} \mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x} = x)] = \Big[& \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{s} \leq t)\mathsf{P}(\boldsymbol{s} \leq t|\boldsymbol{x} = x) + y_0 P(\boldsymbol{s} > t|\boldsymbol{x} = x), \\ & \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{s} \geq t)\mathsf{P}(\boldsymbol{s} \geq t|\boldsymbol{x} = x) + y_1 P(\boldsymbol{s} < t|\boldsymbol{x} = x) \Big]. \end{aligned} \tag{2.13}$$

This finding highlights some important facts. Under the monotone treatment response assumption, the bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x} = x)$ are obtained using information from all $(\boldsymbol{y}, \boldsymbol{s})$ pairs (given $\boldsymbol{x} = x$), while the bounds in (2.11) only use the information provided by $(\boldsymbol{y}, \boldsymbol{s})$ pairs for which $\boldsymbol{s} = t$ (given $\boldsymbol{x} = x$). As a consequence, the bounds in (2.13) are informative even if $\mathsf{P}(\boldsymbol{s} = t|\boldsymbol{x} = x) = 0$, whereas the worst case bounds are not.

Concerning the ATE with $t_1 > t_0$, under monotone treatment response its lower bound is zero, and its upper bound is obtained by subtracting the lower bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_0)|\boldsymbol{x} = x)$ from the upper bound on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t_1)|\boldsymbol{x} = x)$, where both bounds are obtained as in (2.13).

The second example entails *exclusion restrictions*, as in, e.g., Manski (1990). Suppose the

researcher observes a random variable $\boldsymbol{z}$, taking its realizations in $\mathcal{Z}$, such that[9]

$$\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{z},\boldsymbol{x}) = \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x}) \quad \forall t \in \mathbb{T}, \ \boldsymbol{x}\text{-a.s.}. \tag{2.14}$$

This assumption is treatment-specific, and requires that the treatment response to t is mean independent with $\boldsymbol{z}$. It is easy to show that under the assumption in (2.14), the bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x}=x)$ become

$$\begin{aligned}\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x}=x)] = \Big[&\operatorname*{ess\,sup}_{\boldsymbol{z}} \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{s}=t,\boldsymbol{z})\mathsf{P}(\boldsymbol{s}=t|\boldsymbol{x}=x,\boldsymbol{z}) + y_0 P(\boldsymbol{s}\neq t|\boldsymbol{x}=x,\boldsymbol{z}),\\ &\operatorname*{ess\,inf}_{\boldsymbol{z}} \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{s}=t,\boldsymbol{z})\mathsf{P}(\boldsymbol{s}=t|\boldsymbol{x}=x,\boldsymbol{z}) + y_1 P(\boldsymbol{s}\neq t|\boldsymbol{x}=x,\boldsymbol{z})\Big]. \end{aligned} \tag{2.15}$$

These are called *intersection bounds* because they are obtained as follows. Given $\boldsymbol{x}$ and $\boldsymbol{z}$, one uses (2.11) to obtain sharp bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{z}=z,\boldsymbol{x}=x)$. Due to the mean independence assumption in (2.14), $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x}=x)$ must belong to each of these bounds $\boldsymbol{z}$-a.s., hence to their intersection. The expression in (2.15) follows. If the instrument affects the probability of being selected into treatment, or the average outcome for the subpopulation receiving treatment t, the bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x}=x)$ shrink. If the bounds are empty, the mean independence assumption can be refuted (see Section 5 for a discussion of misspecification in partial identification). Manski and Pepper (2000, 2009) generalize the notion of instrumental variable to *monotone* instrumental variable, and show how these can be used to obtain tighter bounds on treatment effect parameters.[10] They also show how shape restrictions and exclusion restrictions can jointly further tighten the bounds. Manski (2013) generalizes these findings to the case where treatment response may have social interactions – that is, each individual's outcome depends on the treatment received by all other individuals.

2.3 Interval Data

Identification Problem 2.1, as well as the treatment evaluation problem in Section 2.2, is an instance of the more general question of what can be learned about (functionals of) probability distributions of interest, in the presence of interval valued outcome and/or covariate data. Such data have become commonplace in Economics. For example, since the early 1990s the Health and Retirement Study collects income data from survey respondents in the form of brackets, with degenerate (singleton) intervals for individuals who opt to fully reveal their income (see, e.g., Juster and Suzman, 1995). Due to concerns for privacy, public use tax data are recorded as the number of tax payers which belong to each of a finite number of

[9] Stronger exclusion restrictions include statistical independence of the response function at each t with $\boldsymbol{z}$: $\mathsf{Q}(\boldsymbol{y}(t)|\boldsymbol{z},\boldsymbol{x}) = \mathsf{Q}(\boldsymbol{y}(t)|\boldsymbol{x})$ $\forall t \in \mathbb{T}$, $\boldsymbol{x}$-a.s.; and statistical independence of the entire response function with $\boldsymbol{z}$: $\mathsf{Q}([\boldsymbol{y}(t), t \in \mathbb{T}]|\boldsymbol{z},\boldsymbol{x}) = \mathsf{Q}([\boldsymbol{y}(t), t \in \mathbb{T}]|\boldsymbol{x})$, $\boldsymbol{x}$-a.s. Examples of partial identification analysis under these conditions can be found in Balke and Pearl (1997), Manski (2003), Kitagawa (2009), Beresteanu, Molchanov, and Molinari (2012), Machado, Shaikh, and Vytlacil (2018), and many others.

[10] See Chesher and Rosen (2019, Chapter XXX in this Volume) for further discussion.

cells (see, e.g., Picketty, 2005). The Occupational Employment Statistics (OES) program at the Bureau of Labor Statistics (Bureau of Labor Statistics, 2018) collects wage data from employers as intervals, and uses these data to construct estimates for wage and salary workers in more than 800 detailed occupations. Manski and Molinari (2010) and Giustinelli, Manski, and Molinari (2019b) document the extensive prevalence of rounding in survey responses to probabilistic expectation questions, and propose to use a person's response pattern across different questions to infer his rounding practice, the result being interpretation of reported numerical values as interval data. Other instances abound. Here I focus first on the case of interval outcome data.

Identification Problem 2.2 (Interval Outcome Data): Assume that in addition to being compact, either $\mathcal{Y}$ is countable or $\mathcal{Y} = [y_0, y_1]$, with $y_0 = \min_{y \in \mathcal{Y}} y$ and $y_1 = \max_{y \in \mathcal{Y}} y$. Let $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x}) \sim \mathsf{P}$ be observable random variables and $\boldsymbol{y}$ be an unobservable random variable whose distribution (or features thereof) is of interest. Suppose that $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{y})$ are such that $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$.[11] In the absence of additional information, what can the researcher learn about features of $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)$, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x} = x$?

It is immediate to obtain the sharp identification region

$$\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)] = [\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x} = x), \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x} = x)]\,.$$

Similarly to the discussion in the previous section, it is also easy to obtain sharp bounds on parameters that respect stochastic dominance, and pointwise-sharp bounds on the CDF of $\boldsymbol{y}$ at any fixed $t \in \mathbb{R}$:

$$\mathbb{P}(\boldsymbol{y}_{\mathrm{U}} \leq t|\boldsymbol{x} = x) \leq \mathbb{P}(\boldsymbol{y} \leq t|\boldsymbol{x} = x) \leq \mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq t|\boldsymbol{x} = x). \tag{2.16}$$

In this case too, however, as in Theorem OR-2.1, the tube of CDFs satisfying equation (2.16) for all $t \in \mathbb{R}$ is an outer region for the CDF of $\boldsymbol{y}|\boldsymbol{x} = x$, rather than its sharp identification region. Indeed, also in this context it is easy to construct examples similar to Example 2.1.

How can one characterize the sharp identification region for the probability distribution of $\boldsymbol{y}|\boldsymbol{x}$ when one observes $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x})$ and assumes $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$? Again, there is not a single answer to this question. Depending on the specific problem at hand, e.g., the specifics of the interval data and whether $\boldsymbol{y}$ is assumed discrete or continuous, different methods can be applied. I use *random set theory* to provide a characterization of $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)]$. Let

$$\boldsymbol{Y} \equiv [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}] \cap \mathcal{Y}.$$

[11] In Identification Problem 2.1 the observable variables are $(\boldsymbol{yd}, \boldsymbol{d}, \boldsymbol{x})$, and $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}})$ are determined as follows: $\boldsymbol{y}_{\mathrm{L}} = \boldsymbol{yd} + y_0(1 - \boldsymbol{d})$, $\boldsymbol{y}_{\mathrm{U}} = \boldsymbol{yd} + y_1(1 - \boldsymbol{d})$. For the analysis in Section 2.2, the data is $(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{x})$ and $\boldsymbol{y}_{\mathrm{L}} = \boldsymbol{y}(t)\mathbf{1}(\boldsymbol{s} = t) + y_0\mathbf{1}(\boldsymbol{s} \neq t)$, $\boldsymbol{y}_{\mathrm{U}} = \boldsymbol{y}(t)\mathbf{1}(\boldsymbol{s} = t) + y_1\mathbf{1}(\boldsymbol{s} \neq t)$. Hence, $\mathsf{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$ by construction.

Then $\boldsymbol{Y}$ is a random closed set according to Definition A.1.[12] The requirement $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$ can be equivalently expressed as

$$\boldsymbol{y} \in \boldsymbol{Y} \quad \text{almost surely.} \tag{2.17}$$

Equation (2.17), together with knowledge of P, exhausts all the information in the data and maintained assumptions. In order to harness such information to characterize the set of observationally equivalent probability distributions for $\boldsymbol{y}$, one can leverage a result due to Artstein (1983) (and Norberg, 1992), reported in Theorem A.1 in Appendix A, which allows one to translate (2.17) into a collection of conditional moment inequalities. Specifically, let $\mathcal{T}$ denote the space of all probability measures with support in $\mathcal{Y}$.

Theorem SIR-2.3 (Conditional Distribution of Interval-Observed Outcome Data): *Under the assumptions in Identification Problem 2.2, the sharp identification region for* $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)$ *is*

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)] = \Big\{\tau(x) \in \mathcal{T} : \tau_K(x) \geq \mathsf{P}(\boldsymbol{Y} \subset K|\boldsymbol{x} = x),\ \forall K \subset \mathcal{Y},\ K \text{ compact}\Big\} \tag{2.18}$$

When $\mathcal{Y} = [y_0, y_1]$, *equation* (2.18) *becomes*

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)] = \Big\{\tau(x) \in \mathcal{T} : \tau_{[t_0,t_1]}(x) \geq \mathsf{P}(\boldsymbol{y}_{\mathrm{L}} \geq t_0, \boldsymbol{y}_{\mathrm{U}} \leq t_1|\boldsymbol{x} = x),\ \forall t_0 \leq t_1,\ t_0, t_1 \in \mathbb{R}\Big\}. \tag{2.19}$$

Proof. Theorem A.1 yields (2.18). If $\mathcal{Y} = [y_0, y_1]$, Molchanov and Molinari (2018, Theorem 2.25) show that it suffices to verify the inequalities in (2.19) for sets K that are intervals. □

Compare equation (2.18) with equation (2.8). Under the set-up of Identification Problem 2.1, when $\boldsymbol{d} = 1$ we have $\boldsymbol{Y} = \{\boldsymbol{y}\}$ and when $\boldsymbol{d} = 0$ we have $\boldsymbol{Y} = \mathcal{Y}$. Hence, for any $K \subsetneq \mathcal{Y}$, $\mathsf{P}(\boldsymbol{Y} \subset K|\boldsymbol{x} = x) = \mathsf{P}(\boldsymbol{y} \in K|\boldsymbol{x} = x, \boldsymbol{d} = 1)\mathsf{P}(\boldsymbol{d} = 1)$.[13] It follows that the characterizations in (2.18) and (2.8) are equivalent. If $\mathcal{Y}$ is countable, it is easy to show that (2.18) simplifies to (2.8) (see, e.g., Beresteanu, Molchanov, and Molinari, 2012, Proposition 2.2).

Key Insight 2.3 (Random set theory and partial identification): *The mathematical framework for the analysis of random closed sets embodied in random set theory is naturally suited to conduct identification analysis and statistical inference in partially identified models. This is because, as argued by Beresteanu and Molinari (2008) and Beresteanu, Molchanov, and Molinari (2011, 2012), lack of point identification can often be traced back to a collection of random variables that are consistent with the available data and maintained assumptions. In turn, this collection of random variables is equal to the family of selections of a properly*

[12] For a proof of this statement, see Molchanov and Molinari (2018, Example 1.11).
[13] For $K = \mathcal{Y}$, both (2.18) and (2.8) hold trivially.

specified random closed set, so that random set theory applies. The interval data case is a simple example that illustrates this point. More examples are given throughout this chapter. As mentioned in the Introduction, the exercise of defining the random closed set that is relevant for the problem under consideration is routinely carried out in partial identification analysis, even when random set theory is not applied. For example, in the case of treatment effect analysis with monotone response function, Manski (1997b) derived the set in the right-hand-side of 2.12, which satisfies Definition (A.1).

An attractive feature of the characterization in (2.18) is that it holds regardless of the specific assumptions on $\boldsymbol{y}_{\mathrm{L}}$, $\boldsymbol{y}_{\mathrm{U}}$, and $\mathcal{Y}$. Later sections in this chapter illustrate how Theorem A.1 delivers the sharp identification region in other more complex instances of partial identification of probability distributions, as well as in structural models. In Chapter **XXX** in this Volume, Chesher and Rosen (2019) apply Theorem A.1 to obtain sharp identification regions for functionals of interest in the important class of *generalized instrumental variable models*. To avoid repetitions, I do not systematically discuss that class of models in this chapter.

It is possible to relate the random set theory approach to partial identification to the *selection mechanism approach* in Tamer (2010) and Ponomareva and Tamer (2011).[14] Take a random variable $\boldsymbol{u}$ with values in $[0,1]$ whose distribution conditional on $\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}$ is left completely unspecified and can be any probability distribution on $[0,1]$. Define

$$\boldsymbol{y}_{\boldsymbol{u}} = \boldsymbol{u}\boldsymbol{y}_{\mathrm{L}} + (1-\boldsymbol{u})\boldsymbol{y}_{\mathrm{U}}. \tag{2.20}$$

The set of admissible distributions for $\boldsymbol{y}$ is given by the collection of distributions of all possible random variables $\boldsymbol{y}_{\boldsymbol{u}}$ as defined in (2.20). This is because each $\boldsymbol{y}_{\boldsymbol{u}}$ is a (stochastically) convex combination of $\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}$, hence each of these random variables satisfies $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y}_s \leq \boldsymbol{y}_{\mathrm{U}}) = 1$ and has a distribution that is a mixture of the distributions of $\boldsymbol{y}_{\mathrm{L}}$ and $\boldsymbol{y}_{\mathrm{U}}$. At the same time, the collection of random variables $\boldsymbol{y}_{\boldsymbol{u}}$ equals the collection of *measurable selections* of the random closed set $\boldsymbol{Y} \equiv [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ (see Definition A.3). Theorem A.1 provides a characterization of the distribution of any $\boldsymbol{y}_{\boldsymbol{u}}$ that satisfies $\boldsymbol{y}_{\boldsymbol{u}} \in \boldsymbol{Y}$ a.s., based on a dominance condition that relates the distribution of $\boldsymbol{y}_{\boldsymbol{u}}$ to the distribution of the random set $\boldsymbol{Y}$. Such dominance condition is given by the inequalities in (2.18).

Horowitz and Manski (1998, 2000) study nonparametric conditional prediction problems with missing outcome and/or missing covariate data. Their analysis shows that this problem is considerably more pernicious than the case where only outcome data are missing. For the case of interval covariate data, Manski and Tamer (2002) provide a set of sufficient conditions under which simple and elegant sharp bounds on functionals of $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x})$ can be obtained, even in this substantially harder identification problem. Their assumptions are listed in Identification Problem 2.3, and their result (with proof) in Theorem SIR-2.4.

[14] Berry and Tamer (2006) and Ciliberto and Tamer (2009) use this approach in the context of structural models of entry. I discuss these models in Section 3

IDENTIFICATION PROBLEM 2.3 (Interval Covariate Data): Let $(\boldsymbol{y}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \sim \mathsf{P}$ be observable random variables in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}$ be an unobservable random variable. Suppose that R, the joint distribution of $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$, is such that: (I) $\mathsf{R}(\boldsymbol{x}_{\mathrm{L}} \leq \boldsymbol{x} \leq \boldsymbol{x}_{\mathrm{U}}) = 1$; (M) $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)$ is weakly increasing in x; and (MI) $\mathbb{E}_{\mathsf{R}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x})$. In the absence of additional information, what can the researcher learn about $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)$ for given $x \in \mathcal{X}$?

Compared to the earlier discussion for the interval outcome case, here there are two additional assumptions. The monotonicity condition (M) is a simple shape restrictions, which however requires some prior knowledge about the joint distribution of $\boldsymbol{y}|\boldsymbol{x}$. The mean independence restriction (MI) requires that if $\boldsymbol{x}$ were observed, knowledge of $(\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ would not affect the conditional expectation of $\boldsymbol{y}|\boldsymbol{x}$. The assumption is not innocuous, as pointed out by the authors. For example, it may fail if censoring is endogenous.[15]

THEOREM SIR-2.4 (Conditional Expectation with Interval-Observed Covariate Data): *Under the assumptions of Identification Problem 2.3, the sharp identification region for* $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x})$ *for given* $x \in \mathcal{X}$ *is*

$$\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)] = \left[\sup_{\boldsymbol{x}_{\mathrm{U}} \leq x} \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}), \inf_{\boldsymbol{x}_{\mathrm{L}} \geq x} \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \right]. \tag{2.21}$$

Proof. The law of iterated expectations and the independence assumption yield $\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \int \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x}) d\mathsf{R}(\boldsymbol{x}|\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$. For all $\underline{x} \leq \bar{x}$, the monotonicity assumption and the fact that $\boldsymbol{x} \in [\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}]$-a.s. yield $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = \underline{x}) \leq \int \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x}) d\mathsf{R}(\boldsymbol{x}|\boldsymbol{x}_{\mathrm{L}} = \underline{x}, \boldsymbol{x}_{\mathrm{U}} = \bar{x}) \leq \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = \bar{x})$. Putting this together with the previous result, $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = \underline{x}) \leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}_{\mathrm{L}} = \underline{x}, \boldsymbol{x}_{\mathrm{U}} = \bar{x}) \leq \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = \bar{x})$. Then (using again the monotonicity assumption) for any $x \geq \bar{x}$, $\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{x}_{\mathrm{L}} = \underline{x}, \boldsymbol{x}_{\mathrm{U}} = \bar{x}) \leq \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)$ so that the lower bound holds. The bound is weakly increasing as a function of x, so that the monotonicity assumption on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)$ holds and the bound is sharp. The argument for the upper bound can be concluded similarly. □

Learning about functionals of $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x} = x)$ naturally implies learning about predictors of $\boldsymbol{y}|\boldsymbol{x} = x$. For example, $\mathcal{H}_{\mathsf{P}}[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)]$ yields the collection of values for the best predictor under square loss; $\mathcal{H}_{\mathsf{P}}[\mathbb{M}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x} = x)]$, with $\mathbb{M}_{\mathsf{Q}}$ the median with respect to distribution Q, yields the collection of values for the best predictor under absolute loss. And so on. A related but distinct problem is that of *parametric* conditional prediction. Often researchers specify not only a loss function for the prediction problem, but also a parametric family of predictor functions, and wish to learn the member of this family that minimizes expected loss. To avoid confusion, let me specify that here I am not referring to a parametric assumption on the best predictor, e.g., that $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x})$ is a linear function of $\boldsymbol{x}$. In the example of linearity

[15] For the case of missing covariate data, which is a special case of interval covariate data similarly to arguments in footnote 11, Aucejo, Bugni, and Hotz (2017) show that the MI restriction implies the assumption that data is missing at random.

and square loss, I am referring to best linear prediction, i.e., best linear approximation to $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x})$. Manski (2003, pp. 56-58) discusses what can be learned about the best linear predictor of $\boldsymbol{y}$ conditional on $\boldsymbol{x}$, when only interval data on $(\boldsymbol{y}, \boldsymbol{x})$ is available.

I treat first the case of interval outcome and perfectly observed covariates.

Identification Problem 2.4 (Parametric Prediction with Interval Outcome Data): Maintain the same assumptions as in Identification Problem 2.2. Let $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x}) \sim \mathsf{P}$ be observable random variables and $\boldsymbol{y}$ be an unobservable random variable, with $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$. In the absence of additional information, what can the researcher learn about the best linear predictor of $\boldsymbol{y}$ given $\boldsymbol{x} = x$?

For simplicity suppose that $\boldsymbol{x}$ is a scalar, and let $\theta = [\theta_0 \ \theta_1]^\top \in \Theta \subset \mathbb{R}^2$ denote the parameter vector of the best linear predictor of $\boldsymbol{y}|\boldsymbol{x}$. Assume that $Var(\boldsymbol{x}) > 0$. Combining the definition of best linear predictor with a characterization of the sharp identification region for the joint distribution of $(\boldsymbol{y}, \boldsymbol{x})$, we have that

$$\mathcal{H}_{\mathsf{P}}[\theta] = \left\{ \vartheta = \arg\min \int (y - \theta_0 - \theta_1 x)^2 \, d\eta, \ \eta \in \mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}, \boldsymbol{x})] \right\}, \tag{2.22}$$

where, using an argument similar to the one in Theorem SIR-2.3,

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}, \boldsymbol{x})] = \Big\{ \eta : \eta_{([t_0,t_1],(-\infty,s])} \geq \mathsf{P}(\boldsymbol{y}_{\mathrm{L}} \geq t_0, \boldsymbol{y}_{\mathrm{U}} \leq t_1, \boldsymbol{x} \leq s) \\ \forall t_0 \leq t_1, t_0, t_1 \in \mathbb{R}, \forall s \in \mathbb{R} \Big\}. \tag{2.23}$$

Beresteanu and Molinari (2008, Proposition 4.1) show that (2.22) can be re-written in an intuitive way that generalizes the well-known formula for the best linear predictor that arises when $\boldsymbol{y}$ is perfectly observed. Define the random segment $\boldsymbol{G}$ and the random matrix Σ_{P} as

$$\boldsymbol{G} = \left\{ \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}\boldsymbol{x} \end{pmatrix} : \ \boldsymbol{y} \in \mathrm{Sel}(\boldsymbol{Y}) \right\} \subset \mathbb{R}^2, \quad \text{and} \quad \Sigma_{\mathsf{P}} = \mathbb{E}_{\mathsf{P}} \begin{pmatrix} 1 & \boldsymbol{x} \\ \boldsymbol{x} & \boldsymbol{x}^2 \end{pmatrix}, \tag{2.24}$$

where $\mathrm{Sel}(\boldsymbol{Y})$ is the set of all measurable selections from $\boldsymbol{Y}$, see Definition A.3. Then,

Theorem SIR-2.5 (Best Linear Predictor with Interval Outcome Data): *Under the assumptions of Identification Problem 2.4, the sharp identification region for the parameters of the best linear predictor of $\boldsymbol{y}|\boldsymbol{x}$ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \Sigma_{\mathsf{P}}^{-1} \mathbb{E}_{\mathsf{P}} \boldsymbol{G}, \tag{2.25}$$

where $\mathbb{E}_{\mathsf{P}} \boldsymbol{G}$ is the (Aumann or selection) expectation of the random closed set $\boldsymbol{G}$ as in Definition A.4.

Proof. By Theorem A.1, $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \in (\boldsymbol{Y} \times \boldsymbol{x})$ (up to an ordered coupling as discussed in Appendix A), if and only if the distribution of $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}})$ belongs to $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}, \boldsymbol{x})]$. The result follows. □

In either representation (2.22) or (2.25), $\mathcal{H}_{\mathsf{P}}[\theta]$ is the collection of best linear predictors for each selection of $\boldsymbol{Y}$.[16] Why should one bother with the representation in (2.25)? The reason is that $\mathcal{H}_{\mathsf{P}}[\theta]$ is a convex set, as it can be clearly evinced from representation (2.25): the Aumann expectation of a convex set is convex, and $\boldsymbol{G}$ has almost surely convex realizations that are segments.[17] Hence, it can be equivalently represented through its *support function* $h_{\mathcal{H}_{\mathsf{P}}[\theta]}$, see Definition A.5 and equation (A.2). In particular,

$$h_{\Sigma_{\mathsf{P}}^{-1}\mathbb{E}_{\mathsf{P}}\boldsymbol{G}}(u) = \mathbb{E}_{\mathsf{P}}[(\boldsymbol{y}_{\mathrm{L}}\mathbf{1}(f(\boldsymbol{x}, u) < 0) + \boldsymbol{y}_{\mathrm{U}}\mathbf{1}(f(\boldsymbol{x}, u) \geq 0))f(\boldsymbol{x}, u)], \quad u \in \mathbb{S}, \tag{2.26}$$

where $f(\boldsymbol{x}, u) \equiv [1 \ \boldsymbol{x}]\Sigma_{\mathsf{P}}^{-1}u$.[18] The characterization in (2.26) results from Theorem A.2, which yields $h_{\Sigma_{\mathsf{P}}^{-1}\mathbb{E}_{\mathsf{P}}\boldsymbol{G}}(u) = \mathbb{E}_{\mathsf{P}}h_{\Sigma_{\mathsf{P}}^{-1}\boldsymbol{G}}(u)$, and the fact that $\mathbb{E}_{\mathsf{P}}h_{\Sigma_{\mathsf{P}}^{-1}\boldsymbol{G}}(u)$ equals the expression in (2.26). As I discuss in Section 4 below, because the support function fully characterizes the boundary of $\mathcal{H}_{\mathsf{P}}[\theta]$, (2.26) allows for a simple sample analog estimator, and for inference procedures with desirable properties. It also immediately yields sharp bounds on linear combinations of θ by judicious choice of u.[19] Stoye (2007) and Magnac and Maurin (2008) provide the same characterization as in (2.26) using, respectively, direct optimization and the Frisch-Waugh-Lovell theorem.

I conclude this section by discussing a generalization of Identification Problem 2.4.

Identification Problem 2.5 (Parametric Prediction with Interval Outcome and Covariate Data): Maintain the same assumptions as in Identification Problem 2.4, but with $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}$ unobservable. Suppose the researcher observes $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ such that $\mathbb{P}\{\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x}_{\mathrm{L}} \leq \boldsymbol{x} \leq \boldsymbol{x}_{\mathrm{U}}\} = 1$. Let $\boldsymbol{X} \equiv [\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}]$ and let $\mathcal{X}$ be bounded. In the absence of additional information, what can the researcher learn about the best linear predictor of $\boldsymbol{y}$ given $\boldsymbol{x} = x$?

Abstractly, $\mathcal{H}_{\mathsf{P}}[\theta]$ is as given in (2.22), with

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}, \boldsymbol{x})] = \{\eta : \eta_K \geq \mathsf{P}((\boldsymbol{Y} \times \boldsymbol{X}) \subset K) \ \forall \text{compact } K \subset \mathcal{Y} \times \mathcal{X}\}$$

replacing (2.23) by an application of Theorem A.1. While this characterization is sharp, it is

[16] Under our assumption that $\mathcal{Y}$ is a bounded interval, all the selections of $\boldsymbol{Y}$ are integrable. Beresteanu and Molinari (2008) consider the more general case where $\mathcal{Y}$ is not required to be bounded.

[17] In $\mathbb{R}^2$ in our example, in $\mathbb{R}^d$ if $\boldsymbol{x}$ is a $d-1$ vector and the predictor includes an intercept.

[18] This result appears in Beresteanu and Molinari (2008, p. 808) and Bontemps, Magnac, and Maurin (2012, p. 1136).

[19] For example, in the case that $\boldsymbol{x}$ is a scalar, sharp bounds on θ_1 can be obtained by choosing $u = [0 \ 1]^\top$ and $u = [0 \ -1]^\top$, which yield $\theta_1 \in [\theta_{1L}, \theta_{1U}]$ with $\theta_{1L} = \min_{\boldsymbol{y} \in [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]} \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{Var(\boldsymbol{x})} = \frac{\mathbb{E}_{\mathsf{P}}[(\boldsymbol{x} - \mathbb{E}_{\mathsf{P}}\boldsymbol{x})(\boldsymbol{y}_{\mathrm{L}}\mathbf{1}(\boldsymbol{x} > \mathbb{E}_{\mathsf{P}}\boldsymbol{x}) + \boldsymbol{y}_{\mathrm{U}}\mathbf{1}(\boldsymbol{x} \leq \mathbb{E}\boldsymbol{x}))]}{\mathbb{E}_{\mathsf{P}}\boldsymbol{x}^2 - (\mathbb{E}_{\mathsf{P}}\boldsymbol{x})^2}$ and $\theta_{1U} = \max_{\boldsymbol{y} \in [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]} \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{Var(\boldsymbol{x})} = \frac{\mathbb{E}_{\mathsf{P}}[(\boldsymbol{x} - \mathbb{E}_{\mathsf{P}}\boldsymbol{x})(\boldsymbol{y}_{\mathrm{L}}\mathbf{1}(\boldsymbol{x} < \mathbb{E}_{\mathsf{P}}\boldsymbol{x}) + \boldsymbol{y}_{\mathrm{U}}\mathbf{1}(\boldsymbol{x} \geq \mathbb{E}\boldsymbol{x}))]}{\mathbb{E}_{\mathsf{P}}\boldsymbol{x}^2 - (\mathbb{E}_{\mathsf{P}}\boldsymbol{x})^2}$.

cumbersome to apply in practice.

On the other hand, when both $\boldsymbol{y}$ and $\boldsymbol{x}$ are perfectly observed, the best linear predictor is simply equal to the parameter vector that yields a mean zero prediction error that is uncorrelated with $\boldsymbol{x}$. How can this basic observation help in the case of interval data? The idea is that one can use the same insight applied to the set-valued data, and obtain $\mathcal{H}_{\mathsf{P}}[\theta]$ as the collection of θ's for which there exists a selection $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \in \text{Sel}(\boldsymbol{Y} \times \boldsymbol{X})$, and associated prediction error $\varepsilon_\theta = \tilde{\boldsymbol{y}} - \theta_0 - \theta_1 \tilde{\boldsymbol{x}}$, satisfying $\mathbb{E}_{\mathsf{P}}\varepsilon_\theta = 0$ and $\mathbb{E}_{\mathsf{P}}(\varepsilon_\theta \tilde{\boldsymbol{x}}) = 0$ (as shown by Beresteanu, Molchanov, and Molinari, 2011).[20] To obtain the formal result, define the θ-dependent set[21]

$$\mathcal{E}_\theta = \left\{ \begin{pmatrix} \tilde{\boldsymbol{y}} - \theta_0 - \theta_1 \tilde{\boldsymbol{x}} \\ (\tilde{\boldsymbol{y}} - \theta_0 - \theta_1 \tilde{\boldsymbol{x}}) \tilde{\boldsymbol{x}} \end{pmatrix} : (\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \in \text{Sel}(\boldsymbol{Y} \times \boldsymbol{X}) \right\}.$$

THEOREM SIR-2.6 (Best Linear Predictor with Interval Outcome and Covariate Data): *Under the assumptions of Identification Problem 2.5, the sharp identification region for the parameters of the best linear predictor of $\boldsymbol{y}|\boldsymbol{x}$ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \{\theta \in \Theta : \mathbf{0} \in \mathbb{E}_{\mathsf{P}}\mathcal{E}_\theta\} = \left\{ \theta \in \Theta : \min_{u \in \mathbb{B}^d} \mathbb{E}_{\mathsf{P}} h_{\mathcal{E}_\theta}(u) = 0 \right\}, \tag{2.27}$$

where $h_{\mathcal{E}_\theta}(u) = \max_{y \in \boldsymbol{Y}, x \in \boldsymbol{X}} [u_1(y - \theta_0 - \theta_1 x) + u_2(yx - \theta_0 x - \theta_1 x^2)]$ is the support function of the set $\mathcal{E}_\theta$ in direction $u \in \mathbb{S}^{d-1}$, see Definition A.5.

Proof. By Theorem A.1, $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \in (\boldsymbol{Y} \times \boldsymbol{X})$ (up to an ordered coupling as discussed in Appendix A), if and only if the distribution of $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}})$ belongs to $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}, \boldsymbol{x})]$. For given θ, one can find $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \in (\boldsymbol{Y} \times \boldsymbol{X})$ such that $\mathbb{E}_{\mathsf{P}}\varepsilon_\theta = 0$ and $\mathbb{E}_{\mathsf{P}}(\varepsilon_\theta \tilde{\boldsymbol{x}}) = 0$ with $\varepsilon_\theta \in \mathcal{E}_\theta$ if and only if the zero vector belongs to $\mathbb{E}_{\mathsf{P}}\mathcal{E}_\theta$. By Theorem A.2, $\mathbb{E}_{\mathsf{P}}\mathcal{E}_\theta$ is a convex set and by (A.9), $\mathbf{0} \in \mathbb{E}_{\mathsf{P}}\mathcal{E}_\theta$ if and only if $\langle 0, u \rangle \leq h_{\mathbb{E}_{\mathsf{P}}\mathcal{E}_\theta}(u) \quad \forall\, u \in \mathbb{B}^d$. The final characterization follows from (A.7). □

The support function $h_{\mathcal{E}_\theta}(u)$ is an easy to calculate convex sublinear function of u, regardless of whether the variables involved are continuous or discrete. The optimization problem in (2.27), determining whether $\theta \in \mathcal{H}_{\mathsf{P}}[\theta]$, is a convex program, hence easy to solve. See for example the CVX software by Grant and Boyd (2010). It should be noted, however, that the set $\mathcal{H}_{\mathsf{P}}[\theta]$ itself is not necessarily convex. Hence, tracing out its boundary is non-trivial. I discuss computational challenges in partial identification in Section 6.

[20] Here for simplicity I suppose that both $\boldsymbol{x}_{\text{L}}$ and $\boldsymbol{x}_{\text{U}}$ have bounded support. Beresteanu, Molchanov, and Molinari (2011) do not make this simplifying assumption.

[21] Note that while $\boldsymbol{G}$ is a convex set, $\mathcal{E}_\theta$ is not.

2.4 Measurement Error and Data Combination

One of the first examples of bounding analysis appears in Frisch (1934), to assess the impact in linear regression of covariate measurement error. The more recent literature in partial identification has provided important advances to learn features of probability distributions when the observed variables are error-ridden measures of the variables of interest. Here I briefly mention some of the papers in this literature, and refer to Chapter **XXX** in this Volume by Schennach (2019) for a thorough treatment of identification and inference with mismeasured and unobserved variables. In an influential paper, Horowitz and Manski (1995) study what can be learned about features of the distribution of $\boldsymbol{y}|\boldsymbol{x}$ in the presence of contaminated or corrupted outcome data. Whereas a contaminated sampling model assumes that data errors are statistically independent of sample realizations from the population of interest, the corrupted sampling model does not. These models are regularly used in the important literature on robust estimation (e.g., Huber, 1964, 2004; Hampel, Ronchetti, Rousseeuw, and Stahel, 2011). However, the goal of that literature is to characterize how point estimators of population parameters behave when data errors are generated in specified ways. As such, the inference problem is approached ex-ante: before collecting the data, one looks for point estimators that are not greatly affected by error. The question addressed by Horowitz and Manski (1995) is conceptually completely distinct. It asks what can be learned about specific population parameters ex-post, that is, after the data has been collected. For example, whereas the mean is well known not to be a robust estimator in the presence of contaminated data, Horowitz and Manski (1995) show that it can be (non-trivially) bounded provided the probability of contamination is strictly less than one. Dominitz and Sherman (2004, 2005) extend the results os Horowitz and Manski's 1995 to allow for (partial) verification of the distribution from which the data are drawn. They apply the resulting sharp bounds to learn about school performance when the observed test scores may not be valid for all students. Molinari (2008) provides sharp bounds on the distribution of a misclassified outcome variable under an array of different assumptions on the extent and type of misclassification.

A completely different problem is that of data combination. Applied economists often face the problem that no single data set contains all the variables that are necessary to conduct inference on a population of interest. When this is the case, they need to integrate the information contained in different samples; for example, they might need to combine survey data with administrative data (see Ridder and Moffitt, 2007, for a survey of the econometrics of data combination). From a methodological perspective, the problem is that while the samples being combined might contain some common variables, other variables belong only to one of the samples. When the data is collected at the same aggregation level (e.g., individual level, household level, etc.), if the common variables include a unique (and correctly recorded) identifier of the units constituting each sample, and there is a substantial overlap of units across all samples, then exact matching of the data sets is relatively straightforward,

and the combined data set provides all the relevant information to identify features of the population of interest. However, it is rather common that there is a limited overlap in the units constituting each sample, or that variables that allow identification of units are not available in one or more of the input files, or that one sample provides information at the individual or household level (e.g., survey data) while the second sample provides information at a more aggregate level (e.g., administrative data providing information at the precinct or district level). Formally, the problem is that one observes data that identify the joint distributions $\mathsf{P}(\boldsymbol{y},\boldsymbol{x})$ and $\mathsf{P}(\boldsymbol{x},\boldsymbol{w})$, but not data that identifies the joint distribution $\mathsf{Q}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{w})$ whose features one wants to learn. The literature on *statistical matching* has aimed at using the common variable(s) $\boldsymbol{x}$ as a bridge to create synthetic records containing $(\boldsymbol{y},\boldsymbol{xw})$ (see, e.g., Okner, 1972, for an early contribution). As Sims (1972) points out, the inherent assumption at the base of statistical matching is that conditional on $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{w}$ are independent. This conditional independence assumption is strong and untestable. While it does guarantee point identification of features of the conditional distributions $\mathsf{Q}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$, it often finds very little justification in practice. Early on, Duncan and Davis (1953) provided numerical illustrations on how one can bound the object of interest, when both $\boldsymbol{y}$ and $\boldsymbol{w}$ are binary variables. Cross and Manski (2002) provide a general analysis of the problem. They obtain bounds on the long regression $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$, under the assumption that $\boldsymbol{w}$ has finite support. They show that sharp bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}=w)$ can be obtained using the results in Horowitz and Manski (1995), thereby establishing a connection with the analysis of contaminated data. They then derive sharp identification regions for $[\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{w}=w), x\in\mathcal{X}, w\in\mathcal{W}]$. They show that these bounds are sharp when $\boldsymbol{y}$ has finite support, and Molinari and Peski (2006) establish sharpness without this restriction. Fan, Sherman, and Shum (2014) address the question of what can be learned about counterfactual distributions and treatment effects under the data scenario just described, but with $\boldsymbol{x}$ replaced by $\boldsymbol{s}$, a binary indicator for the received treatment (using the notation of the previous section). In this case, the exogenous selection assumption (conditional on $\boldsymbol{w}$) does not suffice for point identification of the objects of interest. The authors derive, however, sharp bounds on these quantities using monotone rearrangement inequalities. Pacini (2017) provides partial identification results for the coefficients in the linear projection of $\boldsymbol{y}$ on $(\boldsymbol{x},\boldsymbol{w})$.

2.5 Further Theoretical Advances and Empirical Applications

In order to discuss the partial identification approach to learning features of probability distributions in some level of detail while keeping this chapter to a manageable length, I have focused on a selection of papers. In this section I briefly mention several other excellent theoretical contributions that could be discussed more closely, as well as several papers that have applied partial identification analysis to answer important empirical questions.

While selectively observed data are commonplace in observational studies, in randomized

experiments subjects are randomly placed in designated treatment groups conditional on $\boldsymbol{x}$, so that the assumption of exogenous selection is satisfied with respect to the assigned treatment. Yet, identification of some highly policy relevant parameters can remain elusive in the absence of strong assumptions. One challenge results from noncompliance, where individuals' received treatments differs from the randomly assigned ones. Balke and Pearl (1997) derive sharp bounds on the ATE in this context, when $\mathcal{Y} = \mathbb{T} = \{0, 1\}$. Even if one is interested in the intention-to-treat parameter, selectively observed data may continue to be a problem. For example, Lee (2009) studies the wage effects of the Job Corps training program, which randomly assigns eligibility to participate in the program. Individuals randomized to be eligible were not compelled to receive treatment, hence Lee (2009) focuses on the intention-to-treat effect. Because wages are only observable when individuals are employed, a selection problem persists despite the random assignment of eligibility to treatment, as employment status may be affected by the training program. Lee obtains sharp bounds on the intention-to-treat effect, through a trimming procedure that leverages results in Horowitz and Manski (1995). Molinari (2010) analyzes the problem of identification of the ATE and other treatment effects, when the received treatment is unobserved for a subset of the population. Missing treatment data may be due to item or survey nonresponse in observational studies, or noncompliance with randomly assigned treatments that are not directly monitored. She derives sharp worst case bounds leveraging results in Horowitz and Manski (1995), and she shows that these are a function of the available prior information on the distribution of missing treatments. If the response function is assumed monotone as in (2.13), she obtains informative bounds without restrictions on the distribution of missing treatments.

Even randomly assigned treatments and perfect compliance with no missing data may not suffice for point identification of all policy relevant parameters. Important examples are given by Heckman, Smith, and Clements (1997) and Manski (1997a). Heckman, Smith, and Clements show that features of the joint distribution of the potential outcomes of treatment and control, including the distribution of treatment effects impacts, cannot be point identified in the absence of strong restrictions. This is because although subjects are randomized to treatment and control, nobody's outcome is observed under both states. Nonetheless, the authors obtain bounds for the functionals of interest. Manski shows that features of outcome distributions under treatment rules in which treatment may vary within groups cannot be point identified in the absence of strong restrictions. This is because data resulting from randomized experiments with perfect compliance allow for point identification of the outcome distributions under treatment rules that assign all persons with the same $\boldsymbol{x}$ to the same treatment group. However, such data only allow for partial identification of outcome distributions under rules in which treatment may vary within groups. Manski derives sharp bounds for functionals of these distributions.

Analyses of data resulting from natural experiments also face identification challenges. Hotz, Mullin, and Sanders (1997) study what can be learned about treatment effects when

one uses a contaminated instrumental variable, i.e. when a mean-independence assumption holds in a population of interest, but the observed population is a mixture of the population of interest and one in which the assumption doesn't hold. They extend the results of Horowitz and Manski (1995) to learn about the causal effect of teenage childbearing on a teen mother's subsequent outcomes, using the natural experiment of miscarriages to form an instrumental variable for teen births. This instrument is contaminated because miscarriges may not occur randomly for a subset of the population (e.g., higher miscarriage rates are associated with smoking and drinking, and these behaviors may be correlated with the outcomes of interest).

Of course, analyses of selectively observed data present many challenges, including but not limited to the ones described in Section 2.1. Motivated by the question of whether the age-adjusted mortality rate from cancer in 2000 was lower than that in the early 1970s, Honoré and Lleras-Muney (2006) study partial identification of competing risk models (see Peterson, 1976, for earlier partial identification results). To answer this question, they need to contend with the fact that mortality rate from cardiovascular disease declined substantially over the same period of time, so that individuals that in the early 1970s might have died from cardiovascular disease before being diagnosed with cancer, do not in 2000. In this context, it is important to carry out the analysis without assuming that the underlying risks are independent. Honoré and Lleras-Muney show that bounds for the parameters of interest can be obtained as the solution to linear programming problems. The estimated bounds suggest much larger improvements in cancer than previously estimated.

Blundell, Gosling, Ichimura, and Meghir (2007) use UK data to study changes over time in the distribution of male and female wages, and in wage inequality. Because the composition of the workforce changes over time, it is difficult to disentangle that effect from changes in the distribution of wages, given that the latter are observed only for people in the workforce. Blundell, Gosling, Ichimura, and Meghir begin their empirical analysis by reporting worst case bounds (as in Manski, 1994) on the CDF of wages conditional on covariates. They then consider various restrictions on treatment selection, e.g., a first order stochastic dominance assumption according to which people with higher wages are more likely to work, and derive tighter bounds under this assumption (and under weaker ones). Finally, they bring to bear shape restrictions. At each step of the analysis, they report the resulting bounds, thereby illuminating the role played by each assumption in shaping the inference. Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2018) provide best linear approximations to the identification region for the quantile gender wage gap using CPS repeated cross-sections data from 1975-2001, using treatment selection assumptions in the spirit of Blundell, Gosling, Ichimura, and Meghir (2007) as well as exclusion restrictions.

Bhattacharya, Shaikh, and Vytlacil (2012) study the effect of Swan-Ganz catheterization on subsequent mortality.[22] Previous research had shown, using propensity score matching

[22]The Swan-Ganz catheter is a device placed in patients in the intensive care unit to guide therapy.

(assuming that there are no unobserved differences between catheterized and non catheterized patients) that Swan-Ganz catheterization increases the probability that patient die within 180 days from admission to the intensive care unit. Bhattacharya, Shaikh, and Vytlacil re-analyze the data using (and extending) bounds results obtained by Shaikh and Vytlacil (2011). These results are based on exclusion restrictions combined with a threshold crossing structure for both the treatment and the outcome variables in problems where $\mathcal{Y} = \mathcal{T} = \{0, 1\}$. Bhattacharya, Shaikh, and Vytlacil use as instrument for Swan-Ganz catheterization the day of the week that the patient was admitted to the intensive care unit. The reasoning is that patients are less likely to be catheterized on the weekend, but the admission day to the intensive care unit is plausible uncorrelated with subsequent mortality. Their results confirm that for some diagnoses, Swan-Ganz catheterization increases mortality at 30 days after catheterization and beyond.

Manski and Pepper (2018) use data from Maryland, Virginia and Illinois to learn about the impact of laws allowing individuals to carry concealed handguns (right-to-carry laws) on violent and property crimes. Point identification of these treatment effects is possible under invariance assumptions that certain features of treatment response are constant across states and years. Manski and Pepper propose the use of weaker but more credible restrictions according to which these features exhibit bounded variation – the invariance case being the limit where the bound equals zero. They carry out their analysis under different combinations of the bounded variation assumptions, and at each step they report the resulting bounds, thereby illuminating the role played by each assumption in shaping the inference.

Mourifié, Henry, and Méango (2018) provide sharp bounds on the joint distribution of potential (binary) outcomes in a Roy model with sector specific unobserved heterogeneity and self selection based on potential outcomes. The key maintained assumption is that the researcher has access to data that includes a stochastically monotone instrumental variable. This is a selection shifter that is restricted to affect potential outcomes monotonically. An example is parental education, which may not be independent from potential wages, but plausibly does not negatively affect future wages. Under this assumption, Mourifié, Henry, and Méango show that all observable implications of the model are embodied in the stochastic monotonicity of observed outcomes in the instrument, hence Roy selection behavior can be tested by checking this stochastic monotonicity. They apply the method to estimate a Roy model of college major choice in Canada and Germany, with special interest in the under-representation of women in STEM.

Mogstad, Santos, and Torgovitsky (2018) provide a general method to obtain sharp bounds on a certain class of treatment effects parameters. This class is comprised of parameters that can be expressed as weighted averages of marginal treatment effects (Heckman and Vytlacil, 1999, 2001, 2005). Torgovitsky (2019b) provides a general method, based on copulas, to obtain sharp bounds on treatment effect parameters in semiparametric binary models. A notable feature of both Mogstad, Santos, and Torgovitsky (2018) and Torgovitsky

(2019b) is that the bounds are obtained as solutions to convex (even linear) optimization problems, rendering them computationally attractive.

3 Partial Identification of Structural Models

The partial identification approach to learning structural parameters in economic models is often semiparametric: the underlying models are specified up to parameters that are finite dimensional (often preference parameters) along with parameters that are infinite dimensional (often distribution functions). Contrary to the nonparametric bounds results discussed in Section 2, structural partial identification usually yields an identification region that is *not* constructive, in the sense that the boundary of the set is not obtained in closed form as a functional of the distribution of the observable data. Rather, the identification region can often be characterized as a *level set* of a properly specified criterion function.

While several early examples of set identification (e.g., Frisch, 1934; Marschak and Andrews, 1944; Klepper and Leamer, 1984; Jovanovic, 1989; Phillips, 1989; Hansen, Heaton, and Luttmer, 1995) were concerned with learning about parameters of structural models, the systematic research program on partial identification of structural economic models finds its genesis in the work of Manski and Tamer (2002), Tamer (2003) and Ciliberto and Tamer (2009), and Haile and Tamer (2003). Each of these papers has advanced the literature in fundamental ways, studying conceptually very distinct problems.

Manski and Tamer (2002) are concerned with partial identification (and estimation) of nonparametric, semiparametric, and parametric conditional expectation functions when one of the conditioning variables is interval valued. Hence, in their analysis, the root cause of the identification problem is that the *data is incomplete*.

Tamer (2003) and Ciliberto and Tamer (2009) are concerned with identification (and estimation) of simultaneous equation models with dummy endogeneous variables which are representations of two-player entry games with multiple equilibria.[23] Haile and Tamer (2003) are concerned with nonparametric identification and estimation of the distribution of valuations in a model of English auctions under weak assumptions on bidders' behavior. In both cases, the root cause of the identification problem is that the *structural model is incomplete*. This is because the model makes multiple predictions for the observed outcome variables, but does not specify a selection mechanism to pick one of them.[24]

Set-valued predictions for the observable outcome (endogenous variables) are a key feature of partially identified structural models. The goal of this section is to explain how they result in a wide array of theoretical frameworks, and how sharp identification regions can be characterized using a unified approach based on random set theory. Although the work

[23] Ciliberto and Tamer (2009) consider more general multi-player entry games.

[24] In Haile and Tamer (2003), the observable outcome variables are the bidders' bids. In Tamer (2003) and Ciliberto and Tamer (2009), the observable outcome variables are the players' actions.

of Manski and Tamer (2002), Tamer (2003) and Ciliberto and Tamer (2009), and Haile and Tamer (2003) has spurred many of the developments discussed in this section, for pedagogical reasons I organize the presentation based on application topic rather than chronologically. The work of Pakes (2010) and Pakes, Porter, Ho, and Ishii (2015) further stimulated a large empirical literature that applies partial identification methods to a wide array of questions of substantive economic importance, to which I return in Section 3.5.

3.1 Discrete Choice in Single Agent Random Utility Models

Let $\mathcal{I}$ denote a population of decision makers and $\mathcal{Y} = \{c_1, \ldots, c_{|\mathcal{Y}|}\}$ a finite universe of potential alternatives (*feasible set* henceforth). Let $\mathfrak{U}$ be a family of real valued functions defined over the elements of $\mathcal{Y}$. Let $\in^*$ denote "is chosen from." Then observed choice is consistent with a *random utility model* if there exists a function π_i drawn from $\mathfrak{U}$ according to some probability distribution, such that $\mathbb{P}(c \in^* C) = \mathbb{P}(\pi_i(c) \geq \pi_i(b)\ \forall b \in C)$ for all $c \in C$, all non empty sets $C \subset \mathcal{Y}$, and all $i \in \mathcal{I}$ (Block and Marschak, 1960). See Manski (2007a, Chapter 13) for a textbook presentation of this class of models, and Matzkin (2007) for a review of sufficient conditions for point identification of nonparametric and semiparametric limited dependent variables models.

As in the seminal work of McFadden (1973), assume that the decision makers and alternatives are characterized by observable and unobservable vectors of real valued attributes. Denote the observable attributes by $\boldsymbol{x}_i \equiv \{\boldsymbol{x}_i^1, (\boldsymbol{x}_{ic}^2, c \in \mathcal{Y})\}, i \in \mathcal{I}$. These include attribute vectors $\boldsymbol{x}_i^1$ that are specific to the decision maker, as well as attribute vectors $\boldsymbol{x}_{ic}^2$ that include components that are specific to the alternative and components that are indexed by both. Denote the unobservable attributes (preferences) by $\nu_i \equiv (\zeta_i, \{\epsilon_{ic},\ c \in \mathcal{Y}\}), i \in \mathcal{I}$. These are idiosyncratic to the decision maker and similarly may include alternative and decision maker specific terms. Denote $\mathcal{X}, \mathcal{V}$ the supports of $\boldsymbol{x}, \nu$, respectively.

In what follows, I label "standard" a random utility model that maintains some form of exogeneity for $\boldsymbol{x}_i$ (e.g., mean or quantile or statistical independence with ν_i) and presupposes observation of data that include $\{(\boldsymbol{C}_i, \boldsymbol{y}_i, \boldsymbol{x}_i) : \boldsymbol{y}_i \in^* \boldsymbol{C}_i\}, i = 1, \ldots, n$, with $|\boldsymbol{C}_i| \geq 2$ (e.g., Manski, 1975, Assumption 1). Often it is also assumed that all members of the population face the same choice set, so that $\boldsymbol{C}_i = D$ for all $i \in \mathcal{I}$ and some known $D \subseteq \mathcal{Y}$, although this requirement is not critical to identification analysis.

3.1.1 Semiparametric Binary Choice Models with Interval Valued Covariates

Manski and Tamer (2002) provide inference methods for nonparametric, semiparametric, and parametric conditional expectation functions when one of the conditioning variables is interval valued. I have discussed their nonparametric sharp bounds on conditional expectations with interval valued covariates in Identification Problem 2.3 and Theorem SIR-2.4. Here I focus on their analysis of semiparametric binary choice models. Compared to the generic notation

set forth at the beginning of Section 3.1, I let $\boldsymbol{C}_i = \mathcal{Y} = \{0,1\}$ for all $i \in \mathcal{I}$, and with some abuse of notation I denote the vector of observed covariates $(\boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}, \boldsymbol{w})$.

Identification Problem 3.1 (Semiparametric Binary Regression with Interval Covariate Data): Let $(\boldsymbol{y}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}, \boldsymbol{w}) \sim \mathsf{P}$ be observable random variables in $\{0,1\} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, $d < \infty$, and let $\boldsymbol{x} \in \mathbb{R}$ be an unobservable random variable. Suppose that $\boldsymbol{y} = \mathbf{1}(\boldsymbol{w}\theta + \delta\boldsymbol{x} + \epsilon > 0)$. Here ϵ is an unobserved heterogeneity term with continuous distribution conditional on $(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$, $(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$-a.s., and $\theta \in \Theta \subset \mathbb{R}^d$ is a parameter vector with compact parameter space Θ. Suppose $\delta > 0$, and further normalize $\delta = 1$ because the threshold-crossing condition is invariant to the scale of the parameters. Assume that R, the joint distribution of $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}, \boldsymbol{w}, \epsilon)$, is such that $\mathsf{R}(\boldsymbol{x}_\mathrm{L} \leq \boldsymbol{x} \leq \boldsymbol{x}_\mathrm{U}) = 1$; $\mathsf{R}(\epsilon | \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}) = \mathsf{R}(\epsilon | \boldsymbol{w}, \boldsymbol{x})$; and for a specified $\alpha \in (0,1)$, $q^\epsilon_\mathsf{R}(\alpha, \boldsymbol{w}, \boldsymbol{x}) = 0$ and $\mathsf{R}(\epsilon \leq 0 | \boldsymbol{w}, \boldsymbol{x}) = \alpha$. In the absence of additional information, what can the researcher learn about θ?

Compared to Identification Problem 2.3 (see p. 21), here one continues to impose $\boldsymbol{x} \in [\boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}]$ a.s. The sign restriction on δ replaces the monotonicity restriction (M) in Identification Problem 2.3, but does not imply it unless the distribution of ϵ is independent of $\boldsymbol{x}$ conditional on $\boldsymbol{w}$. The quantile independence restriction is inspired by Manski (1985).

For given $\theta \in \Theta$, this model yields set valued predictions because $\boldsymbol{y} = 1$ can occur whenever $\epsilon > -\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{U}$, whereas $\boldsymbol{y} = 0$ can occur whenever $\epsilon \leq -\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{L}$, and $-\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{U} \leq -\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{L}$. Conversely, observation of $\boldsymbol{y} = 1$ allows one to conclude that $\epsilon \in (-\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{U}, +\infty)$, whereas observation of $\boldsymbol{y} = 0$ allows one to conclude that $\epsilon \in (-\infty, -\boldsymbol{w}\theta - \boldsymbol{x}_\mathrm{L}]$, and these regions of possible realizations of ϵ overlap. In contrast, when $\boldsymbol{x}$ is observed the prediction is unique because the value $-\boldsymbol{w}\theta - \boldsymbol{x}$ partitions the space of realizations of ϵ in two disjoint sets, one associated with $\boldsymbol{y} = 1$ and the other with $\boldsymbol{y} = 0$. Figure 3.1[25] depicts the model's set-valued predictions for $\boldsymbol{y}$ given $(\boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$ as a function of ϵ, and the model's set valued predictions for ϵ given $(\boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$ as a function of $\boldsymbol{y}$.[26]

Why does this set-valued prediction hinder point identification? The reason is that the distribution of the observable data relates to the model structure in an *incomplete* manner: $\mathsf{P}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}) = \int \mathsf{R}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}) d\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U}) = \int \mathsf{R}(\epsilon > -\boldsymbol{w}\theta - \boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}) d\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$, $(\boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$-a.s. Because the distribution $\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$ is left completely unspecified, one can find multiple values for θ, for $\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$, and for $\mathsf{R}(\epsilon | \boldsymbol{w}, \boldsymbol{x})$, satisfying the assumptions in Identification Problem 3.1, such that they yield the observed value of $\mathsf{P}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$, $(\boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$-a.s. Nonetheless, in general, not all values of $\theta \in \Theta$ can be paired with some $\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$ and $\mathsf{R}(\epsilon | \boldsymbol{w}, \boldsymbol{x})$ so that they are compatible with $\mathsf{P}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$, $(\boldsymbol{w}, \boldsymbol{x}_\mathrm{L}, \boldsymbol{x}_\mathrm{U})$-a.s. Hence, θ can be partially identified using the informa-

[25] This figure is based on Figure 1 in Manski and Tamer (2002).

[26] See Chesher and Rosen (2019, Chapter XXX in this Volume) for an extensive discussion of the duality between the model's set valued predictions for $\boldsymbol{y}$ as a function of ϵ and for ϵ as a function of $\boldsymbol{y}$, in both cases given the observed covariates.

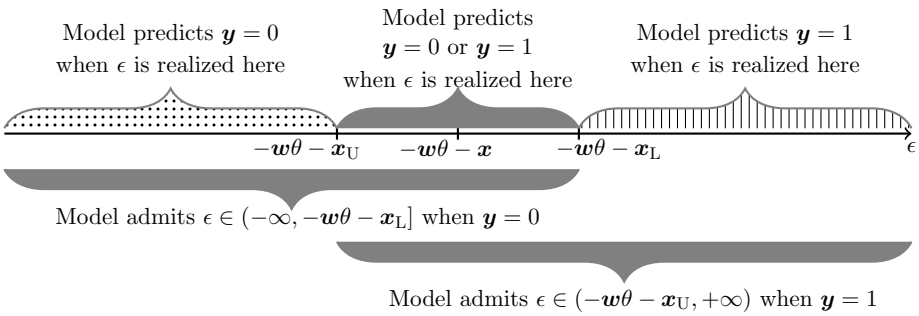


Figure 3.1: Predicted value of $\boldsymbol{y}$ as a function of ϵ, and admissible values of ϵ for each realization of $\boldsymbol{y}$, in Identification Problem 3.1, conditional on $(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$.

tion in the model and observed data.

THEOREM SIR-3.1 (Semiparametric Binary Regression with Interval Covariate Data): *Under the Assumptions of Identification Problem 3.1, the sharp identification region for θ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \Big\{\vartheta \in \Theta : \mathsf{P}\Big((\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) : \{0 \leq \boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{L}} \cap \mathsf{P}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \leq 1 - \alpha\} \\ \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{U}} \leq 0 \cap \mathsf{P}(\boldsymbol{y} = 1 | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \geq 1 - \alpha\}\Big) = 0\Big\}. \quad (3.1)$$

Proof. Define the set of possible values for the unobservable associated with the possible realizations of $(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$, illustrated in Figure 3.1, as[27]

$$\mathcal{E}_\theta(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \begin{cases} (-\infty, -\boldsymbol{w}\theta - \boldsymbol{x}_{\mathrm{L}}] & \text{if } \boldsymbol{y} = 0, \\ [-\boldsymbol{w}\theta - \boldsymbol{x}_{\mathrm{U}}, +\infty) & \text{if } \boldsymbol{y} = 1. \end{cases}$$

Then $\mathcal{E}_\theta(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ is a random closed set. To simplify notation, I write $\mathcal{E}_\theta(\boldsymbol{y}) \equiv \mathcal{E}_\theta(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$, suppressing this set's dependence on $(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$. Let $(\mathcal{E}_\theta(\boldsymbol{y}), \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \mathcal{E}_\theta(\boldsymbol{y}) \times (\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \{(\mathbf{e}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) : \mathbf{e} \in \mathcal{E}_\theta(\boldsymbol{y})\}$. If the model is correctly specified, $(\epsilon, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \in (\mathcal{E}_\theta(\boldsymbol{y}), \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ a.s. for the data generating value of θ. By Theorem A.1 and Theorem 2.33 in Molchanov and Molinari (2018), this occurs if and only if

$$\mathsf{R}(\epsilon \in C | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \geq \mathsf{P}(\mathcal{E}_\theta(\boldsymbol{y}) \subset C | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}), \ (\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})\text{-a.s. } \forall C \in \mathsf{F},$$

where F here denotes the collection of closed subsets of $\mathbb{R}$. The above condition can be rewritten as

$$\int \mathsf{R}(\epsilon \in C | \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) d\mathsf{R}(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \geq \mathsf{P}(\mathcal{E}_\theta(\boldsymbol{y}) \subset C | \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}), \ (\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})\text{-a.s. } \forall C \in \mathsf{F}.$$

[27] In the definition of $\mathcal{E}_\theta(1, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ I exploit the fact that under the maintained assumptions $\mathbb{P}(\epsilon = -\boldsymbol{w}\theta - \boldsymbol{x}_{\mathrm{U}} | \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = 0$ to enforce its closedness.

The assumption that $\mathsf{R}(\epsilon|\boldsymbol{w},\boldsymbol{x},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) = \mathsf{R}(\epsilon|\boldsymbol{w},\boldsymbol{x})$ yields that the above system of inequalities reduces to

$$\int \mathsf{R}(\epsilon \in C|\boldsymbol{w},\boldsymbol{x}) d\mathsf{R}(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \geq \mathsf{P}(\mathcal{E}_\theta(\boldsymbol{y}) \subset C|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}),\ \ (\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}})\text{-a.s. } \forall C \in \mathsf{F}.$$

Next, note that given the possible realizations of $\mathcal{E}_\theta(\boldsymbol{y})$, the above inequality is trivially satisfied unless $C = (-\infty, t]$ or $C = [t, \infty)$ for some $t \in \mathbb{R}$. Finally, the only restriction on the distribution of ϵ is the quantile independence condition, hence it suffices to consider $t = 0$. That, together with the definition of $\mathcal{E}_\theta(\boldsymbol{y})$, reduces the above inequalities to

$$1-\alpha \geq \mathsf{P}(\boldsymbol{y}=1|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \ \text{ for all } (\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \text{ such that } \boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{U}} \leq 0, \tag{3.2}$$

$$1-\alpha \leq \mathsf{P}(\boldsymbol{y}=1|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \ \text{ for all } (\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \text{ such that } \boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{L}} \geq 0. \tag{3.3}$$

Any given $\vartheta \in \Theta$, $\vartheta \neq \theta$, violates the above conditions if and only if $\mathsf{P}\big((\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) : \{0 \leq \boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{L}} \cap \mathsf{P}(\boldsymbol{y}=1|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \leq 1-\alpha\} \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{U}} \leq 0 \cap \mathsf{P}(\boldsymbol{y}=1|\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}) \geq 1-\alpha\}\big) > 0.$ □

KEY INSIGHT 3.1: *The analysis in Manski and Tamer (2002) systematically studies what can be learned under increasingly strong sets of assumptions. These include both assumptions that constrain the model from fully nonparametric to semiparametric to parametric, as well as assumptions that constrain the distribution of the observable covariates. For example, Manski and Tamer (2002, Corollary to Proposition 2) provide sufficient conditions on the joint distribution of* $(\boldsymbol{w},\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}})$ *that allow for identification of the sign of components of* θ*, as well as for point identification of* θ*.*[28] *The careful analysis of the identifying power of increasingly stronger assumptions is the pillar of the partial identification approach to empirical research put forward by Manski, as illustrated in Section 2. The work of Manski and Tamer (2002) was the first example of this kind in semiparametric structural models.*

Revisiting Manski and Tamer's 2002 study of Identification Problem 3.1 nearly 20 years later yields important insights on the differences between point and partial identification analysis. It is instructive to take as a point of departure the analysis of Manski (1985), which under the additional assumption that $(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})$ is observed implies that

$$\boldsymbol{w}\theta + \boldsymbol{x} > 0 \Leftrightarrow \mathsf{P}(\boldsymbol{y}=1|\boldsymbol{w},\boldsymbol{x}) > 1-\alpha.$$

Hence, θ is identified relative to $\vartheta \in \Theta$ if

$$\mathsf{P}\left((\boldsymbol{w},\boldsymbol{x}) : \{\boldsymbol{w}\theta + \boldsymbol{x} \leq 0 < \boldsymbol{w}\vartheta + \boldsymbol{x}\} \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x} \leq 0 < \boldsymbol{w}\theta + \boldsymbol{x}\}\right) > 0. \tag{3.4}$$

Manski and Tamer extend this reasoning to the case that $\boldsymbol{x}$ is unobserved, but known to satisfy $\boldsymbol{x} \in [\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$ a.s. The first part of their analysis, collected in their Proposition 2,

[28] This Corollary is related in spirit to the analysis in Manski (1988).

characterizes the collection of values that cannot be distinguished from θ on the basis of $\mathsf{P}(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ alone, through a clear generalization of (3.4):

$$\{\vartheta \in \Theta : \mathsf{P}\left((\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) : \{\boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{U}} \leq 0 < \boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{L}}\} \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{U}} \leq 0 < \boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{L}}\}\right) = 0\}. \tag{3.5}$$

It is worth emphasizing that the characterization in (3.5) depends on θ, and makes no use of the information in $\mathsf{P}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$. The Corollary to Proposition 2 yields conditions on $\mathsf{P}(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ under which either the sign of components of θ, or θ itself, can be identified, regardless of the distribution of $\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}$.

Manski and Tamer (2002, Lemma 1) provide a second characterization, which presupposes knowledge of $\mathsf{P}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$, yields a set smaller than the one in (3.5), and coincides with the result in Theorem SIR-3.1. Manski and Tamer (2002) use the same notation for the two sets, although the sets are conceptually and mathematically distinct.[29] The result in Theorem SIR-3.1 is due to Manski and Tamer (2002, Lemma 1), but the proof provided here is new, as is the use of random set theory in this application.[30]

KEY INSIGHT 3.2: *The preceding discussion allows me to draw a novel connection between the two characterizations in Manski and Tamer (2002), and the distinction put forward by Chesher and Rosen (2017b) and Chesher and Rosen (2019, Chapter XXX in this Volume, Definition 2) between* potential observational equivalence *and* observational equivalence *with partial identification. Applying Chesher and Rosen's definition, parameter vectors θ and ϑ are* potentially *observationally equivalent if there exists* some *distribution of $\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}$ for which conditions (3.2)-(3.3) hold. Simple algebra confirms that this yields the region in (3.5). This notion of potential observational equivalence parallels one of the notions used to obtain sufficient conditions for point identification in the semiparametric literature (as in, e.g. Manski, 1985). Both notions, as explained in Chesher and Rosen (2019, Section 4.1), make no reference to the conditional distribution of outcomes given covariates delivered by the process being studied. To obtain that parameters θ and ϑ are observationally equivalent one requires instead that conditions (3.2)-(3.3) hold for the* observed *distribution $\mathsf{P}(\boldsymbol{y} = 1|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ (as opposed to "for some distribution" as in the case of potential observational equivalence). This yields the sharp identification region in (3.1).*

Magnac and Maurin (2008) consider a different but closely related model to the one studied by Manski and Tamer. They assume that an instrumental variable $\boldsymbol{z}$ is available, that ϵ is independent of $\boldsymbol{x}$ conditional on $(\boldsymbol{w}, \boldsymbol{z})$, and that $Corr(\boldsymbol{z}, \epsilon) = 0$. They assume

[29] This was confirmed in personal communication with Chuck Manski and Elie Tamer.

[30] The proof closes a gap in the argument in Manski and Tamer (2002) connecting their Proposition 2 and Lemma 1, due to the fact that for a given ϑ the sets $\{(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) : \{\boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{U}} \leq 0 < \boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{L}}\} \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{U}} \leq 0 < \boldsymbol{w}\theta + \boldsymbol{x}_{\mathrm{L}}\}\}$ and $\{(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) : \{0 < \boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{L}} \cap \mathsf{P}(\boldsymbol{y} = 1|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \leq 1 - \alpha\} \cup \{\boldsymbol{w}\vartheta + \boldsymbol{x}_{\mathrm{U}} \leq 0 \cap \mathsf{P}(\boldsymbol{y} = 1|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) > 1 - \alpha\}\}$ need not coincide, with the former being a subset of the latter due to part (c) of the proof of Proposition 2 in Manski and Tamer (2002).

that the distribution of $\boldsymbol{x}$ is absolutely continuous with support $[v_1, v_k]$, and that $\boldsymbol{x}$ is not a deterministic linear function of $(\boldsymbol{w}, \boldsymbol{z})$. They consider the case that $\boldsymbol{x}$ is unobserved but known to belong to one of the fixed (and known) intervals $[v_i, v_{i+1})$, $i = 1, \ldots, k-1$, with $\mathbb{P}[\boldsymbol{x} \in [v_i, v_{i+1})|\boldsymbol{w}, \boldsymbol{z}] > 0$ almost surely for all i. Finally, they assume that $(-\boldsymbol{w}\theta - \epsilon) \in [v_1, v_k]$ with probability one. They do not, however, make quantile independence assumptions.

Their point of departure is the fact that under these conditions, if $\boldsymbol{x}$ were observed, one could employ a transformation proposed by Lewbel (2000) for the binary outcome $\boldsymbol{y}$, such that θ can be identified through a simple linear moment condition. Specifically, let

$$\tilde{\boldsymbol{y}} = \frac{\boldsymbol{y} - \mathbf{1}_{\boldsymbol{x}>0}}{f_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{z})},$$

where $f_{\boldsymbol{x}}(\cdot|\boldsymbol{w}, \boldsymbol{z})$ is the conditional density function of $\boldsymbol{x}$. Then, using the assumption that $\boldsymbol{z}$ and ϵ are uncorrelated, one has

$$\mathbb{E}_{\mathsf{P}}(\boldsymbol{z}\tilde{\boldsymbol{y}}) - \mathbb{E}_{\mathsf{P}}(\boldsymbol{z}\boldsymbol{w}^\top)\theta = 0. \tag{3.6}$$

With interval valued $\boldsymbol{x}$, Magnac and Maurin (2008) denote by $\boldsymbol{x}^*$ the random variable that takes value $i \in \{1, \ldots, k-1\}$ if $\boldsymbol{x} \in [v_i, v_{i+1})$, so that the observed data are draws from the joint distribution of $(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{x}^*)$. They let $\delta(\boldsymbol{x}^*) = v_{\boldsymbol{x}^*+1} - v_{\boldsymbol{x}^*}$ denote the length of the $\boldsymbol{x}^*$-th interval, and define the transformed outcome variable:

$$\boldsymbol{y}^* = \frac{\delta(\boldsymbol{x}^*)}{\mathsf{P}(\boldsymbol{x}^* = i|\boldsymbol{w}, \boldsymbol{z})}\boldsymbol{y} - v_k.$$

The assumptions on $\boldsymbol{x}$ yield that, given $\boldsymbol{z}$ and $\boldsymbol{w}$, ϵ does not depend on $\boldsymbol{x}^*$. Moreover, $\mathbb{P}(\boldsymbol{y} = 1|\boldsymbol{x}^*, \boldsymbol{w}, \boldsymbol{z})$ is non-decreasing in $\boldsymbol{x}^*$ and $F_\epsilon(\cdot|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}^*) = F_\epsilon(\cdot|\boldsymbol{z}, \boldsymbol{w})$. Magnac and Maurin (2008) show that the sharp identification region for θ is

$$\mathcal{H}_{\mathsf{P}}[\theta] = \mathbb{E}_{\mathsf{P}}(\boldsymbol{z}\boldsymbol{w}^\top)^{-1}\mathbb{E}_{\mathsf{P}}(\boldsymbol{z}\boldsymbol{y}^* + \boldsymbol{z}\boldsymbol{U}), \tag{3.7}$$

where $\mathbb{E}_{\mathsf{P}}(\boldsymbol{z}\boldsymbol{y}^* + \boldsymbol{z}\boldsymbol{U})$ is the (Aumann or selection) expectation of the random interval $\boldsymbol{z}\boldsymbol{y}^* + \boldsymbol{z}\boldsymbol{U}$, see Definition A.4, with

$$\boldsymbol{U} = \left[-\sum_{i=1}^{k-1} (r_i(\boldsymbol{w}, \boldsymbol{z}) - r_{i-1}(\boldsymbol{w}, \boldsymbol{z}))(v_{i+1} - v_i), \sum_{i=1}^{k-1} (r_{i+1}(\boldsymbol{w}, \boldsymbol{z}) - r_i(\boldsymbol{w}, \boldsymbol{z}))(v_{i+1} - v_i) \right].$$

In this expression, $r_{\boldsymbol{x}^*}(\boldsymbol{w}, \boldsymbol{z}) \equiv \mathsf{P}(\boldsymbol{y} = 1|\boldsymbol{x}^*, \boldsymbol{w}, \boldsymbol{z})$ and by convention $r_0(\boldsymbol{w}, \boldsymbol{z}) = 0$ and $r_K(\boldsymbol{w}, \boldsymbol{z}) = 1$, see Magnac and Maurin (2008, Theorem 4). If $r_i(\boldsymbol{w}, \boldsymbol{z}), i = 0, \ldots, k$, were observed, this characterization would be very similar to the one provided by Beresteanu and Molinari (2008) for Identification Problem 2.4, see equation (2.25). However, these random functions need to be estimated. While the first-stage estimation of $r_i(\boldsymbol{w}, \boldsymbol{z}), i = 0, \ldots, k$,

does not affect the identification arguments, it does complicate inference, see Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2018) and the discussion in Section 4.

Manski and Tamer (2002) also study identification of parametric regression models under the assumptions in Identification Problem 3.2; Theorem SIR-3.2 below reports the result.[31] The proof is omitted because it follows immediately from the proof of Theorem SIR-2.4.[32]

IDENTIFICATION PROBLEM 3.2 (Parametric Regression with Interval Covariate Data): Let $(\boldsymbol{y}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}, \boldsymbol{w}) \sim \mathsf{P}$ be observable random variables in $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, $d < \infty$, and let $\boldsymbol{x} \in \mathbb{R}$ be an unobservable random variable. Assume that the joint distribution R of $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})$ is such that $\mathsf{R}(\boldsymbol{x}_{\mathrm{L}} \leq \boldsymbol{x} \leq \boldsymbol{x}_{\mathrm{U}}) = 1$ and $\mathbb{E}_{\mathsf{R}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) = \mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x})$. Suppose that $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}) = f(\boldsymbol{w}, \boldsymbol{x}; \theta)$, with $f : \mathbb{R}^d \times \mathbb{R} \times \Theta \mapsto \mathbb{R}$ a known function such that for each $w \in \mathbb{R}$ and $\theta \in \Theta$, $f(w, x; \theta)$ is weakly increasing in x. In the absence of additional information, what can the researcher learn about θ?

THEOREM SIR-3.2 (Parametric Regression with Interval Covariate Data): *Under the Assumptions of Identification Problem 3.2, the sharp identification region for θ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \big\{\vartheta \in \Theta : f(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}; \vartheta) \leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}) \leq f(\boldsymbol{w}, \boldsymbol{x}_{\mathrm{U}}; \vartheta),\ (\boldsymbol{w}, \boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}})\text{-}a.s.\big\}. \quad (3.8)$$

Aucejo, Bugni, and Hotz (2017) study Identification Problem 3.2 for the case of missing covariate data *without* imposing the mean independence restriction of Manski and Tamer (2002) (Assumption MI in Identification Problem 2.3). As discussed in Section 2.3, restriction MI is undesirable in this context because it implies the assumption that data are missing at random. Aucejo, Bugni, and Hotz (2017) characterize $\mathcal{H}_{\mathsf{P}}[\theta]$ under the weaker assumptions, but face the problem that this characterization is usually too complex to compute or to use for inference. They therefore provide outer regions that are easier to compute, and they show that these regions are informative and relatively easy to use.

Manski (2010) studies random *expected* utility models, where agents choose the alternative that maximizes their expected utility. The core difference with standard models is that Manski does not fully specify the subjective beliefs that agents use to form their expectations, but only a *set* of such beliefs. Manski shows that the resulting, partially identified, discrete choice model can be formulated similarly to how Manski and Tamer (2002)'s treat interval valued covariates, and leverages their results to obtain bounds on preference parameters.

[31] Their analysis applies to parametric regression models that go beyond the ones for binary choice.

[32] Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix F) extend the analysis of Manski and Tamer (2002) to multinomial choice models with interval covariates.

3.1.2 Endogenous Explanatory Variables

Whereas the standard random utility model presumes some form of exogeneity for $\boldsymbol{x}_i$, in practice often some explanatory variables are endogenous. This problem has been addressed in the literature to obtain point identification of the model through a combination of several assumptions, including large support conditions, special regressors, control function restrictions, and more (see, e.g., Matzkin, 1993; Berry, Levinsohn, and Pakes, 1995; Lewbel, 2000; Petrin and Train, 2010). Here I discuss how to carry out identification analysis in the absence of such assumptions when instrumental variables $\boldsymbol{z}$ are available, as proposed by Chesher, Rosen, and Smolinski (2013).[33]

Identification Problem 3.3 (Discrete Choice with Endogenous Explanatory Variables): Let $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) \sim \mathsf{P}$ be observable random variables in $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$. Suppose that all members of the population face the same choice set $\mathcal{Y}$. Suppose that each alternative has one unobservable attribute $\epsilon_c, c \in \mathcal{Y}$ and let $\nu \equiv (\epsilon_{c_1}, \ldots, \epsilon_{c_{|\mathcal{Y}|}})$.[34] Let $\nu \sim \mathsf{Q}$ and assume that $\nu \perp\!\!\!\perp \boldsymbol{z}$. Suppose Q belongs to a specified family of distributions $\mathcal{T}$, and that the conditional distribution of $\nu|\boldsymbol{x}, \boldsymbol{z}$, denoted $\mathsf{S}(\nu|\boldsymbol{x}, \boldsymbol{z})$, is absolutely continuous with respect to Lebesgue measure with everywhere positive density on its support, $(\boldsymbol{x}, \boldsymbol{z})$-a.s. Suppose utility is separable in unobservables and has a functional form known up to finite dimensional parameter vector $\delta \in \Delta \subset \mathbb{R}^m$, so that $\pi_i(c) = g(\boldsymbol{x}_c; \delta) + \epsilon_c$, $(\boldsymbol{x}_c, \epsilon_c)$-a.s., for all $c \in \mathcal{Y}$. Maintain the normalizations $g(\boldsymbol{x}_{c_{|\mathcal{Y}|}}; \delta) = 0$ for all $\delta \in \Delta$ and all $\boldsymbol{x} \in \mathcal{X}$, and $g(x_c^0; \delta) = \bar{g}$ for known $(x_c^0, \bar{g})$ for all $\delta \in \Delta$ and $c \in \mathcal{Y}$.[35] Given $(\boldsymbol{x}, \boldsymbol{z}, \nu)$, suppose $\boldsymbol{y}$ is the utility maximizing choice in $\mathcal{Y}$. In the absence of additional information, what can the researcher learn about (δ, Q)?

The key challenge to identification here results because the distribution of ν can vary across different values of $\boldsymbol{x}$, both conditional and unconditional on $\boldsymbol{z}$. Why does this fact hinder point identification? For any $c \in \mathcal{Y}$ and $x \in \mathcal{X}$, the model yields that c is optimal, and hence chosen, if and only if ν realizes in the set

$$\mathcal{E}_\delta(c, x) = \{e \in \mathcal{V} : g(x_c; \delta) + e_c \geq g(x_d; \delta) + e_d \ \forall d \in \mathcal{Y}\}. \tag{3.9}$$

Figure 3.2 plots the set $\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$ in a stylized example with $\mathcal{Y} = \{1, 2, 3\}$ and $\mathcal{X} = \{x^1, x^2\}$, as a function of $(\epsilon_1 - \epsilon_3, \epsilon_2 - \epsilon_3)$.[36] Consider the model implied distribution, denoted M below,

[33] Chesher, Rosen, and Smolinski (2013) consider a more general case than I do here, with utility function that is not parametrically specified and not restricted to be separable in the unobservables. Even in that more general case, the identification analysis follows through similar steps as reported here.

[34] Compared to the general model put forward in Section 3.1, in this model there are no preference heterogeneity terms ζ (random coefficients) that vary only across decision makers.

[35] Of course, under these conditions one can work directly with utility differences. To try and economize on notation, I do not explicitly do so here.

[36] This figure is based on Figures 1-3 in Chesher, Rosen, and Smolinski (2013).

of the optimal choice. Then, recalling the restriction $\boldsymbol{z} \perp\!\!\!\perp \nu$, we have

$$\mathsf{M}(c|\boldsymbol{x} \in R_z, \boldsymbol{z} = z; \delta) = \int_{x \in R_z} \mathsf{S}(\mathcal{E}_\delta(c, \boldsymbol{x})|\boldsymbol{x} = x, \boldsymbol{z} = z) d\mathsf{P}(x|z), \ \forall\ R_z \subseteq \mathcal{X},\ \boldsymbol{z}\text{-a.s.} \tag{3.10}$$

$$\mathsf{Q}(F) = \int_{x \in \mathcal{X}} \mathsf{S}(F|\boldsymbol{x} = x, \boldsymbol{z} = z) d\mathsf{P}(x|z), \quad \forall\ F \subseteq \mathcal{Z},\ \boldsymbol{z}\text{-a.s.}, \tag{3.11}$$

Because the joint distribution of $(\boldsymbol{x}, \nu)$ conditional on $\boldsymbol{z}$ is left completely unrestricted (except for (3.11)), one can find multiple triplets $(\delta, \mathsf{Q}, \mathsf{S})$ satisfying the maintained assumptions and such that $\mathsf{M}(c|\boldsymbol{x} \in R_v, \boldsymbol{z} = z; \delta) = \mathsf{P}(c|\boldsymbol{x} \in R_z, \boldsymbol{z} = z)$ for all $c \in \mathcal{Y}$ and $R_z \subseteq \mathcal{X}$, $\boldsymbol{z}$-a.s.

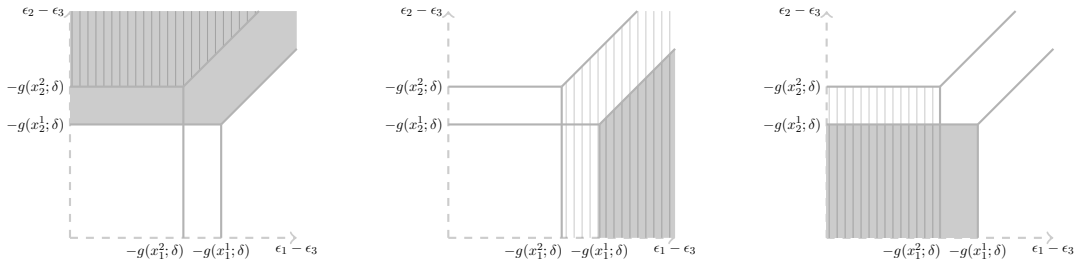
It is instructive to compare (3.10)-(3.11) with McFadden's 1973 conditional logit. Under the standard assumptions, $\boldsymbol{x} \perp\!\!\!\perp \nu$ so that no instrumental variables are needed. This yields $\mathsf{S} = \mathsf{Q}$ $\boldsymbol{x}$-a.s., and in addition Q is typically known, with corresponding simplifications in (3.10). The resulting system of equalities can be inverted under standard order and rank conditions to yield point identification of δ.

Further insights can be gained by looking at Figure 3.2. As the value of $\boldsymbol{x}$ changes from x_1 to x_2, the region of values where, say, alternative 1 is optimal changes. When $\boldsymbol{x}$ is exogenous, say independent of ν, this yields a system of inequalities relating (δ, Q) to the observed distribution $\mathsf{P}(\boldsymbol{y}, \boldsymbol{x})$ which, as stated above, can be inverted to obtain point identification. When $\boldsymbol{x}$ is endogenous, this reasoning breaks down because the conditional distribution $\mathsf{S}(\nu|\boldsymbol{x}, \boldsymbol{z})$ may change across realizations of $\boldsymbol{x}$. Figure 3.2 also offers an instructive way to connect Identification Problem 3.3 with the identification problem studied in the previous Section 3.1.1 (as well as with those in Sections 3.2-3.3 below). In the latter, the model has set-valued predictions for the *outcome variable* given realizations of the covariates and unobserved heterogeneity terms. In the problem studied here, the model has singleton-valued predictions for the outcome variable of interest $\boldsymbol{y}$ as a function of the observable explanatory variables $\boldsymbol{x}$ and unobservables ν. However, for given realization of ν, the model admits *sets* of values for the *endogenous variables* $(\boldsymbol{y}, \boldsymbol{x})$. Because the model is silent on the joint distribution of $(\boldsymbol{x}, \nu)$ (except for requiring that the marginal distribution of ν does not depend on $\boldsymbol{z}$), partial identification results.

It is possible to couple the maintained assumptions with the observed data to learn features of (δ, Q). Because the observed choice $\boldsymbol{y}$ is assumed to maximize utility, for the data generating (δ, Q) the model yields

$$\nu \in \mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})\text{-a.s.}, \tag{3.12}$$

with $\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$ a random closed set as per Definition A.1. Equation (3.12) exhausts the modeling content of Identification Problem 3.3. Theorem A.1 (as expressed in (A.5)) can then be leveraged to extract its empirical content from the observed distribution $\mathsf{P}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$. As a preparation for doing so, note that for given $F \in \mathcal{F}$ (with $\mathcal{F}$ the collection of closed subsets



(a) $\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$ for $\boldsymbol{y} = 1$ (b) $\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$ for $\boldsymbol{y} = 2$ (c) $\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$ for $\boldsymbol{y} = 3$

Figure 3.2: The set $\mathcal{E}_\delta$ in equation (3.9) and the corresponding admissible values for $(\boldsymbol{y}, \boldsymbol{x})$ as a function of $(\epsilon_1 - \epsilon_3, \epsilon_2 - \epsilon_3)$ under the simplifying assumption that $\mathcal{X} = \{x^1, x^2\}$ and $\mathcal{Y} = \{1, 2, 3\}$. The admissible values for $(\boldsymbol{y}, \boldsymbol{x})$ are $\{(c, x^1)\}$ in the gray area, and $\{(c, x^2)\}$ in the area with vertical lines. Because the two areas overlap, the model has set-valued predictions for $(\boldsymbol{y}, \boldsymbol{x})$.

of $\mathcal{V}$) and $\delta \in \Delta^m$, we have

$$\mathsf{P}(\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x}) \subseteq F | \boldsymbol{z}) = \int_{x \in \mathcal{X}} \sum_{c \in \mathcal{Y}} \mathbf{1}(\mathcal{E}_\delta(c, x) \subseteq F) \mathsf{P}(\boldsymbol{y} = c | \boldsymbol{x} = x, \boldsymbol{z}) d\mathsf{P}(x | \boldsymbol{z}),$$

so that this probability can be learned from the observed data.

THEOREM SIR-3.3 (Discrete Choice with Endogenous Explanatory Variables): *Under the assumptions of Identification Problem 3.3, the sharp identification region for* (δ, Q) *is*

$$\mathcal{H}_{\mathsf{P}}[\delta, \mathsf{Q}] = \{\delta \in \Delta, \mathsf{Q} \in \mathcal{T} : \mathsf{Q}(F) \geq \mathsf{P}(\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x}) \subseteq F | \boldsymbol{z}),\ \forall F \in \mathcal{F},\ \boldsymbol{z}\text{-a.s.}\}\,. \tag{3.13}$$

Proof. To simplify notation, I write $\mathcal{E}_\delta \equiv \mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x})$. Let $(\mathcal{E}_\delta, \boldsymbol{x}, \boldsymbol{z}) = \{(\mathbf{e}, \boldsymbol{x}, \boldsymbol{z}) : \mathbf{e} \in \mathcal{E}_\delta\}$. If the model is correctly specified, $(\nu, \boldsymbol{x}, \boldsymbol{z}) \in (\mathcal{E}_\delta, \boldsymbol{x}, \boldsymbol{z})$-a.s. for the data generating value of (δ, Q). By Theorem A.1 and Theorem 2.33 in Molchanov and Molinari (2018), this occurs if and only if

$$\mathsf{S}(F | \boldsymbol{x}, \boldsymbol{z}) \geq \mathsf{P}(\mathcal{E}_\delta(\boldsymbol{y}, \boldsymbol{x}) \subseteq F | \boldsymbol{x}, \boldsymbol{z}),\ \forall F \in \mathcal{F},\ (\boldsymbol{x}, \boldsymbol{z})\text{-a.s.}$$

Since the distribution of ν is only restricted so that $\boldsymbol{z} \perp\!\!\!\perp \nu$, one can integrate both sides of the inequality with respect to $\boldsymbol{x}$. The final result follows because Q does not depend on $\boldsymbol{z}$. □

While Theorem SIR-3.3 relies on checking inequality (3.13) for all $F \in \mathcal{F}$, the results in Chesher, Rosen, and Smolinski (2013, Theorem 2) and Molchanov and Molinari (2018, Chapter 2) can be used to obtain a smaller collection of sets over which to verify it.

KEY INSIGHT 3.3: *A conceptual contribution of Chesher, Rosen, and Smolinski (2013) is to show that one can frame models with endogenous explanatory variables as* incomplete *models. Incompleteness here results from the fact that the model does not specify how the endogenous variables* $\boldsymbol{x}$ *are determined. One can then think of these as models with set-*

valued predictions for the endogeneous variables ($\boldsymbol{y}$ and $\boldsymbol{x}$ in this application), even though the outcome of the model ($\boldsymbol{y}$) is uniquely predicted by the realization of the observed explanatory variables ($\boldsymbol{x}$) and the unobserved heterogeneity terms (ν). Random set theory can again be leveraged to characterize sharp identification regions.

Chesher and Rosen (2019, Chapter XXX in this Volume) discuss related generalized instrumental variables models where random set methods are used to obtain characterizations of sharp identification regions in the presence of endogenous explanatory variables.

3.1.3 Unobserved Heterogeneity in Choice Sets and/or Consideration Sets

As pointed out in Manski (1977), often the researcher observes $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ but not $\boldsymbol{C}_i$, $i = 1, \ldots, n$. Even when $\boldsymbol{C}_i$ is observable, the researcher may be unaware of which of its elements the decision maker actually evaluates before selecting one. In what follows, to shorten expressions, I refer to both the measurement problem of unobserved choice sets and the (cognitive) problem of limited consideration as "unobserved heterogeneity in choice sets."

Learning features of preferences using discrete choice data in the presence of unobserved heterogeneity in choice sets is a formidable task. When a decision maker chooses an alternative, this may be because her choice set equals the feasible set and the chosen alternative is the one yielding the highest utility. Then observed choice reveals preferences. But it can also be that the decision maker has access to/considers only the chosen alternative (e.g., Block and Marschak, 1960, p. 99). Then observed choice is driven entirely by choice set composition, and is silent about preferences. A plethora of scenarios between these extremes is possible, but the researcher does not know which has generated the observed data. This fundamental identification problem calls either for restrictions on the random utility model and consideration set formation process, or for collection of richer data that eliminates unobserved heterogeneity in $\boldsymbol{C}_i$ or allows for enhanced modeling of it (see, e.g., Caplin, 2016).

A sizable literature spanning behavioral economics, econometrics, experimental economics, marketing, microeconomics, and psychology, has put forward different models to formalize the complex process that leads to the formation of the set of alternatives that the agent considers or can choose from (see, e.g., Simon, 1959; Howard, 1963; Tversky, 1972, for early contributions). Manski (1977) proposes both a general econometric model where decision makers draw choice sets from an unknown distribution, as well as a specific model of choice set formation, independent from preferences, and studies their implications for the distributional structure of random utility models.[37]

However, assumptions about the choice set formation process are often rooted in a desire to achieve point identification rather than in information contained in the model or observed

[37] The specific model in Manski (1977, Section II-A) is often used in applications. It posits that each alternative $c \in \mathcal{Y}$ enters the decision maker's choice set with probability ϕ_c, independently of the other alternatives. The probability ϕ_c may depend on observable individual characteristics, and $\phi_c = 1$ for at least one option $c \in \mathcal{Y}$ (the "default" good).

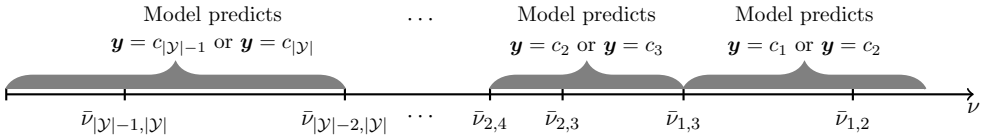


Figure 3.3: Predicted value of $\boldsymbol{y}$ in Identification Problem 3.4 as a function of ν for $\kappa = |\mathcal{Y}| - 1$. In this case, $\boldsymbol{C} = \mathcal{Y} \setminus \{c\}$ for some $c \in \mathcal{Y}$, and the model predicts either the first or the second best alternative in $\mathcal{Y}$.

data.[38] It is then important to ask what can be learned about decision maker's preferences under minimal assumptions on the choice set formation process. Allowing for unrestricted dependence between choice sets and preferences, while challenging for identification analysis, is especially relevant. Indeed, decision makers' unobserved attributes may determine both their preferences and which items in the feasible set they pay attention to or are available to them (e.g., through unobserved liquidity constraints, unobserved characteristics such as religious preferences in the context of school choice, or behavioral phenomena such as aversion to extremes, salience, etc.). Here I use the framework put forward by Barseghyan, Coughlin, Molinari, and Teitelbaum (2019) to study identification of discrete choice models with unobserved heterogeneity in choice sets and preferences.

Identification Problem 3.4 (Discrete Choice with Unobserved Heterogeneity in Choice Sets and Preferences): Let $(\boldsymbol{y}, \boldsymbol{x}) \sim \mathsf{P}$ be observable random variables in $\mathcal{Y} \times \mathcal{X}$. Assume that there exists a real valued function g, which for simplicity I posit known up to parameter $\delta \in \Delta \subset \mathbb{R}^m$ and continuous in its second argument, such that $\pi_i(c) = g(\boldsymbol{x}_{ic}, \nu_i; \delta)$, $(\boldsymbol{x}_{ic}, \nu_i)$-a.s., for all $c \in \mathcal{Y}, i \in \mathcal{I}$, where $\boldsymbol{x}_{ic}$ denotes the vectors of attributes relevant to alternative c, and includes attributes that are alternative invariant ($\boldsymbol{x}_i^1$ in the general notation laid out in Section 3.1). Suppose that $\boldsymbol{y} = \arg\max_{c \in \boldsymbol{C}} g(\boldsymbol{x}_c, \nu; \delta)$, where ties are assumed to occur with probability zero and $\boldsymbol{C}$ is an unobservable choice set drawn from the subsets of $\mathcal{Y}$ according to some unknown probability distribution. Suppose $\mathbb{P}(|\boldsymbol{C}| \geq \kappa) = 1$ for some known constant $\kappa \geq 2$. Let Q denote the distribution of ν, and assume that it is known up to a finite dimensional parameter $\gamma \in \Gamma \subset \mathbb{R}^k$. For simplicity, assume that $\nu \perp\!\!\!\perp \boldsymbol{x}$.[39] In the absence of additional information, what can the researcher learn about $\theta \equiv [\delta; \gamma]$?

The model just laid out has set valued predictions for the decision maker's optimal choice, because different alternatives might be optimal depending on which choice set the decision maker draws. Figure 3.3, which is based on the analysis in Barseghyan, Coughlin, Molinari, and Teitelbaum (2019), illustrates the set valued predictions in a stylized example. In the

[38] These assumptions are akin to assumptions about selection mechanisms in models with multiple equilibria. The latter are discussed further below in Section 3.2.1, along with their criticisms.

[39] This assumption can be relaxed as discussed in Matzkin (2007). The procedure proposed here can also be adapted to allow for endogenous explanatory variables as in Section 3.1.2 by combining the results in Barseghyan, Coughlin, Molinari, and Teitelbaum (2019) with those in Chesher, Rosen, and Smolinski (2013).

figure ν is assumed to be a scalar; $\bar{\nu}_{j,m}$ denotes the threshold value of ν above which c_j yields higher utility than c_m and below which c_m yields higher utility than c_j (the threshold's dependence on $(\boldsymbol{x};\delta)$ is suppressed for notational convenience). Consider the case that $\nu \in [\bar{\nu}_{2,3}, \bar{\nu}_{1,2}]$, so that c_2 is the option yielding the highest utility among all options in $\mathcal{Y}$. When $\kappa = |\mathcal{Y}| - 1$, the agent may draw a choice set that does not include one of the alternatives in $\mathcal{Y}$. If the excluded alternative is not c_2 (or if $\boldsymbol{C}$ realizes equal to $\mathcal{Y}$), the model predicts that the decision maker chooses c_2. If $\boldsymbol{C}$ realizes equal to $\mathcal{Y} \setminus \{c_2\}$, the model predicts that the decision maker chooses the second best: c_1 if $\nu \in [\bar{\nu}_{1,3}, \bar{\nu}_{1,2}]$, and c_3 if $\nu \in [\bar{\nu}_{2,3}, \bar{\nu}_{1,3}]$.

Why does this set valued prediction hinder point identification? The reason is similar to the explanation given for Identification Problem 3.1: the distribution of the observable data relates to the model structure in an *incomplete* manner, because the distribution of the (unobserved) choice sets is left completely unspecified. Barseghyan, Coughlin, Molinari, and Teitelbaum (2019) show that one can find multiple candidate distributions for $\boldsymbol{C}$ and parameter vectors ϑ, such that together they yield a model implied distribution for $\boldsymbol{y}|\boldsymbol{x}$ that matches $\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})$, $\boldsymbol{x}$-a.s.

Barseghyan, Coughlin, Molinari, and Teitelbaum propose to work directly with the set of model implied optimal choices given $(\boldsymbol{x}, \nu)$ associated with each possible realization of $\boldsymbol{C}$, which is depicted in Figure 3.3 for a specific example. The key idea is that, according to the model, the observed choice maximizes utility among the alternatives in $\boldsymbol{C}$. Hence, for the data generating value of θ, it belongs to the set of model implied optimal choices. With this, the authors are able to characterize $\mathcal{H}_{\mathsf{P}}[\theta]$ through Theorem A.1 as the collection of parameter vectors that satisfy a finite number of conditional moment inequalities.

Key Insight 3.4: *Barseghyan, Coughlin, Molinari, and Teitelbaum (2019) show that working directly with the set of model implied optimal choices given $(\boldsymbol{x}, \nu)$ allows one to dispense with considering all possible distributions of choice sets that are allowed for in Identification Problem 3.4 to complete the model. Such distributions may depend on ν even after conditioning on observables and may constitute an infinite dimensional nuisance parameter, which creates great difficulties for the computation of $\mathcal{H}_{\mathsf{P}}[\theta]$ and for inference.*

Identification Problem 3.4 sets up a structure where preferences include idiosyncratic components ν that are decision maker specific and can depend on $\boldsymbol{C}$, and where heterogeneity in $\boldsymbol{C}$ can be driven either by a measurement problem, or by the decision maker's limited attention to the options available to her. However, it restricts the family of utility functions to be known up to a finite dimensional parameter vector δ.

A rich literature in decision theory has analyzed a different framework, where the decision maker's choice set is observable to the researcher, but the decision maker does not consider all alternatives in it (for recent contributions see, e.g., Masatlioglu, Nakajima, and Ozbay, 2012; Manzini and Mariotti, 2014). In this literature, the utility function is left completely unspecified, so that interest focuses on identification of preference rankings of the available options.

Unobserved heterogeneity in preferences is assumed away, so that heterogeneous choice is driven by randomness in consideration sets. If the consideration set formation process is left unspecified or is subject only to weak restrictions, point identification of the preference orderings is not possible even if preferences are homogeneous and the researcher observes a representative agent facing multiple distinct choice problems with varying choice sets. Cattaneo, Ma, Masatlioglu, and Suleymanov (2017) propose a general model for the consideration set formation process where the only restriction is a weak and intuitive monotonicity condition: the probability that any particular consideration set is drawn does not decrease when the number of possible consideration sets decreases. Within this framework, they provide revealed preference theory and testable implications for observable choice probabilities.

IDENTIFICATION PROBLEM 3.5 (Homogeneous Preferences Ranking in Random Attention Models): Let $(\boldsymbol{y}, \boldsymbol{C}) \sim \mathsf{P}$ be a pair of observable random variable and random set in $\mathcal{Y} \times \mathfrak{D}$, where $\mathfrak{D} = \{D : D \subseteq \mathcal{Y}\} \setminus \emptyset$.[40] Let $\mu : \mathfrak{D} \times \mathfrak{D} \to [0,1]$ denote an *attention rule* such that $\mu(A|G) \geq 0$ for all $A \subseteq G$, $\mu(A|G) = 0$ for all $A \nsubseteq G$, and $\sum_{A \subset G} \mu(A|G) = 1$, $A, G \in \mathfrak{D}$. Assume that for any $b \in G \setminus A$,

$$\mu(A|G) \leq \mu(A|G \setminus \{b\}), \tag{3.14}$$

and that the decision maker has a strict preference ordering $\succ$ on $\mathcal{Y}$ (but no other restriction is placed on it).[41] In the absence of additional information, what can the researcher learn about $\succ$?

Cattaneo, Ma, Masatlioglu, and Suleymanov (2017) posit that an observed distribution of choice $\mathsf{P}(\boldsymbol{y}|\boldsymbol{C})$ has a random attention representation, and hence they name it a *random attention model*, if there exists a preference ordering $\succ$ over $\mathcal{Y}$ and a monotonic attention rule μ such that

$$\mathsf{p}(c|G) \equiv \mathsf{P}(\boldsymbol{y} = c|\boldsymbol{C} = G) = \sum_{A \subseteq G} \mathbf{1}(c \text{ is } \succ\text{-best in } A)\mu(A|G), \quad \forall c \in G, \ \forall G \in \mathfrak{D}. \tag{3.15}$$

Hence, the sharp identification region for the preference ordering, denoted $\mathcal{H}_{\mathsf{P}}[\succ]$, is given by the collection of preference orderings for which one can find a monotonic attention rule to pair it with, so that (3.15) holds. Of course, an observed distribution of choice can be represented by multiple preference orderings and attention rules. The authors, however, show that if for *some* $G \in \mathfrak{D}$ with $\{b, c\} \in G$,

$$\mathsf{p}(c|G) > \mathsf{p}(c|G \setminus \{b\}), \tag{3.16}$$

[40] Here I omit observable covariates $\boldsymbol{x}$ for simplicity.
[41] Specifically, $\succ$ is an asymmetric, transitive and complete binary relation.

then $c \succ b$ for any $\succ$ for which one can find a monotonic attention rule μ such that (3.15) holds. Because of preference transitivity, one can also learn $a \succ b$ if in addition to the above condition one has $\mathsf{p}(a|G) > \mathsf{p}(a|G \setminus \{c\})$ for some $c \in G$. This yields a system of linear inequalities in $\mathsf{P}(\boldsymbol{y}|\boldsymbol{C})$ that fully characterize $\mathcal{H}_{\mathsf{P}}[\succ]$. Let $\vec{\mathsf{p}}$ denote the vector with elements $[\mathsf{p}(c|G) : c \in G, G \in \mathfrak{D}]$ and $\Pi_{\succ}$ denote a conformable matrix collecting the constraints on $\mathsf{P}(\boldsymbol{y}|\boldsymbol{C})$ embodied in (3.16) and its generalizations based on transitive closure. Then

$$\mathcal{H}_{\mathsf{P}}[\succ] = \{\succ\colon \Pi_{\succ}\vec{\mathsf{p}} \leq 0\}. \tag{3.17}$$

The authors show that for any given preference ordering $\succ$, the matrix $\Pi_{\succ}$ characterizing whether $\succ \in \mathcal{H}_{\mathsf{P}}[\succ]$ through the system of linear inequalities in (3.17) is unique, and they provide a simple algorithm to compute it.

KEY INSIGHT 3.5: *Cattaneo, Ma, Masatlioglu, and Suleymanov (2017) show that learning features of preference orderings in Identification Problem 3.5 requires the existence in the data of choice problems where the choice probabilities satisfy* (3.16)*. The latter is a violation of the principle of "regularity" (Luce and Suppes, 1965) according to which the probability of choosing an alternative from any set is at least as large as the probability of choosing it from any of its supersets. Regularity is a monotonicity property of choice probabilities, and it is implied by a wide array of models of decision making. The monotonicity of attention rules in* (3.14) *can be viewed as regularity of the process that chooses a consideration set from the subsets of the choice set. Cattaneo, Ma, Masatlioglu, and Suleymanov (2017) show that it is implied by various models of limited attention. While the violation required in* (3.16) *is weak in that it needs only to occur for some G, it sheds a different light on the severity of the identification problem described at the beginning of this section. Regularity of choice probabilities and (partial) identification of preference orderings can co-exist only under restrictions on the consideration set formation process that are stronger than the regularity of attention rules in* (3.14)*.*

Abaluck and Adams (2018) and Barseghyan, Molinari, and Thirkettle (2019) provide different sets of sufficient conditions for point identification of models of limited consideration. In both cases, the authors assume that unobserved heterogeneity in preferences and in consideration sets are independent. Both also posit specific models of consideration set formation. With that structure in place, they show semi-nonparametric identification of the distribution of preferences and consideration under exclusion and large support assumptions. To do so, Abaluck and Adams (2018) exploit violations of Slutsky symmetry that result from inattention, assuming that for each alternative there is an observable characteristic with large support that does not affect the consideration probability of the other options. Barseghyan, Molinari, and Thirkettle (2019) exploit a requirements of standard economic theory –the single crossing property of utility functions– coupled with a slight strengthening of the classic

conditions for semi-nonparametric identification of discrete choice models with full consideration and identical choice sets (see, e.g., Matzkin, 2007), assuming that there is a single decision maker-specific characteristic with large support that does not affect consideration.

3.1.4 Prediction of Choice Behavior with Counterfactual Choice Sets

Manski (2007b) studies a question related but distinct from those in Identification Problems 3.4-3.5. He is concerned with prediction of choice behavior when decision makers face counterfactual choice sets. Manski frames this question as one of predicting treatment response (see Section 2.2). Here the collection of potential treatments is given by $\mathfrak{D}$, the nonempty subsets of the universe of feasible alternatives $\mathcal{Y}$, and the response function specifies the alternative chosen by a decision makes when facing choice set $G \in \mathfrak{D}$. Manski assumes that the researcher observes realized choice sets and chosen alternatives, $(\boldsymbol{y}, \boldsymbol{C}) \sim \mathsf{P}$.[42] Under the standard assumptions laid out at the beginning of Section 3.1, specifically if utility functions are (say) linear in ϵ_{ic} and the distribution of ϵ_{ic} is (say) Type I extreme value or multivariate normal, prediction of choice behavior with counterfactual choice sets is immediate (and point identified). Manski, however, leaves utility functions completely unspecified, and in fact works directly with preference orderings, which he labels decision maker's *types*. He places no restriction on the distribution of preference types, except requiring that they are independent of the observed choice sets. Manski shows that under these rather weak assumptions, the distribution of predicted choices from counterfactual choice sets can be partially identified, and characterized as the solution to linear programs.

Specifically, let $\boldsymbol{y}^*(G)$ denote the decision maker's optimal choice when facing choice set $G \in \mathfrak{D}$. Assume $\boldsymbol{y}^*(\cdot) \perp\!\!\!\perp \boldsymbol{C}$, and let y_k denote the choice function for a decision maker of type k –that is, a decision maker with a specific preference ordering labeled k. One example of such preference ordering might be $c_1 \succ c_2 \succ \cdots \succ c_{|\mathcal{Y}|}$. If a decision maker of this type faces, say, choice set $G = \{c_2, c_3, c_4\}$, then she chooses alternative c_2. Let K denote the set of logically possible types, and θ_k the probability that a decision maker in the population is of type k. Suppose that the researcher posits a behavioral model that can be expressed as an assumption that θ lies in some specified set of distributions, and let Θ denote the values of ϑ that satisfy this requirement plus the conditions $\vartheta_k \geq 0$ for all $k \in K$ and $\sum_{k \in K} \theta_k = 1$. Then for any $c \in \mathcal{Y}$ and $\vartheta \in \Theta$, the model predicts

$$\mathsf{Q}(\boldsymbol{y}^*(G) = c) = \sum_{k \in K} \mathbf{1}(y_k(G) = c)\vartheta_k.$$

How can one partially identify this probability based on the observed data? Suppose $\boldsymbol{C}$ is

[42] Here I suppress covariates for simplicity.

observed to take realizations $D_1, \ldots, D_m$. Then the data reveal

$$\mathsf{P}(\boldsymbol{y}(D_j) = d_j) = \sum_{k \in K} \mathbf{1}(y_k(D_j) = d_j)\theta_k \ \forall d_j \in D_j, j = 1, \ldots, m.$$

This yields that the sharp identification region for θ is

$$\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : \ \mathsf{P}(\boldsymbol{y}(D_j) = d_j) = \sum_{k \in K} \mathbf{1}(y_k(D_j) = d_j)\vartheta_k \ \forall d_j \in D_j, j = 1, \ldots, m\}.$$

If the behavioral model is correctly specified, $\mathcal{H}_{\mathsf{P}}[\theta]$ is non-empty. In turn, the sharp identification region for each choice probability is

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}(\boldsymbol{y}^*(G) = c)] = \left\{ \sum_{k \in K} \mathbf{1}(y_k(G) = c)\vartheta_k : \vartheta \in \mathcal{H}_{\mathsf{P}}[\theta] \right\},$$

and its extreme points can be obtained by solving linear programs.

Kitamura and Stoye (2019) provide closely related sharp bounds on features of counterfactual choices in the nonparametric random utility model of demand, where observable choices are repeated cross-sections and one allows for unrestricted, unobserved heterogeneity. Their approach builds on the work of Kitamura and Stoye (2018), who test weather agents' behavior is consistent with the Axiom of Revealed Stochastic Preference (SARP) in a random utility model in which the utility function of each consumer over commodity bundles is assumed to satisfy only the basic restriction that "more is better" with no satiation. Because the testing exercise is to be carried out using repeated cross-sections data, the authors maintain the assumption that multiple populations of consumers who face distinct choice sets have the same distribution of preferences. With this structure in place, de facto the task is to test the full implications of rationality without functional form restrictions. Kitamura and Stoye's approach is based on several novel and insightful ideas. As a first step, they leverage an earlier insight of McFadden (2005) to discretize the data without loss of information, so that they can define a large but finite set of rational preferences types. As a second step, they show that this implies that rationality can be tested by checking whether observed behavior lies in a cone corresponding to positive linear combinations of preference types. While the problem is discrete, its dimension is at first sight prohibitive. Nonetheless, Kitamura and Stoye are able to develop novel computational methods that render the problem tractable. They apply their method to the U.K. Household Expenditure Survey, adapting to their framework results on nonparametric instrumental variable analysis by Imbens and Newey (2009) so that they can handle price endogeneity.

Kamat (2018) builds on Manski (2007b) to learn program effects when agents are randomly assigned to control or treatment. The treatment group is provided access to the program, while the control group is not. However, members of the control group may receive

access to the program from outside the experiment, leading to noncompliance with the randomly assigned treatment. The researcher wants to learn about the average effect of program access on the decision to participate in the program and on the subsequent outcome. While sufficiently rich data may allow the researcher to learn these effects, Kamat is concerned with the identification problem that arises when the researcher only observes the treatment assignment status, the program participation decision, and the outcome, but not the receipt of program access for every agent. Kamat formalizes this problem as one where the received treatment is selected from a choice set that depends on the assigned treatment and is unobservable to the researcher, and the agents optimally choose whether to participate in the program by maximizing their utility function over their choice set. Importantly, the utility functions are not subject to parametric restrictions, as in Manski (2007b). But while Manski assumed independence of choice sets and preference types, Kamat allows them to be arbitrarily dependent on each other, as in Barseghyan, Coughlin, Molinari, and Teitelbaum (2019). Kamat's 2018 approach leverages specific assumptions on random assignment of treatments and on compliance (or lack thereof) of participants to obtain nonparametric bounds on the treatment effects of interest that can be characterized using tractable linear programs.

3.2 Static, Simultaneous-Move Finite Games with Multiple Equilibria

3.2.1 An Inference Approach Robust to the Presence of Multiple Equilibria

Tamer (2003) and Ciliberto and Tamer (2009) substantially enlarge the scope of partial identification analysis of structural models by showing how to apply it to learn features of payoff functions in static, simultaneous-move finite games of complete with multiple equilibria. The approach and considerations that follow can be extended to games of incomplete information, as shown in Berry and Tamer (2006). To simplify notation here I focus on two-player entry games with complete information.[43]

Identification Problem 3.6 (Complete Information Two Player Entry Game): Let $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathsf{P}$ be observable random variables in $\{0,1\} \times \{0,1\} \times \mathbb{R}^d \times \mathbb{R}^d$, $d < \infty$. Suppose that $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ result from simultaneous move, pure strategy Nash play (PSNE) in a game where the payoffs are $\pi_j(\boldsymbol{y}_j, \boldsymbol{y}_{3-j}, \boldsymbol{x}_j; \beta_j, \delta_j) \equiv \boldsymbol{y}_j(\boldsymbol{x}_j\beta_j + \delta_j \boldsymbol{y}_{3-j} + \varepsilon_j)$, $j = 1, 2$ and the strategies are "enter" $(\boldsymbol{y}_j = 1)$ or "stay out" $(\boldsymbol{y}_j = 0)$. Here $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are observable payoff shifters, $(\varepsilon_1, \varepsilon_2)$ are payoff shifters observable to the players but not to the econometrician, $\delta_1 \leq 0, \delta_2 \leq 0$ are interaction effect parameters, and β_1, β_2 are parameter vectors in $B \subset \mathbb{R}^d$ reflecting the effect of the observable covariates on payoffs. Each player enters the market if and only if entering yields non-negative payoff, so that $\boldsymbol{y}_j = \mathbf{1}(\boldsymbol{x}_j\beta_j + \delta_j \boldsymbol{y}_{3-j} + \varepsilon_j \geq 0)$. For simplicity, assume that $(\varepsilon_1, \varepsilon_2)$ are independent of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and have jointly Normal

[43] Completeness of information is motivated by the idea that firms in the industry have settled in a long-run equilibrium, and have detailed knowledge of both their own and their rivals' profit functions.

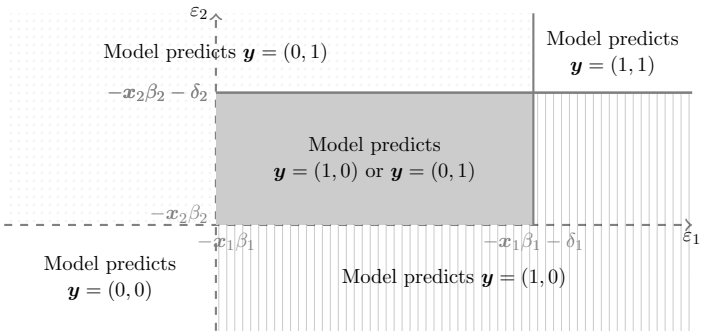


Figure 3.4: PSNE outcomes of the game in Identification Problem 3.6 as a function of $(\varepsilon_1, \varepsilon_2)$.

distribution with mean vector zero, variances equal to one (a normalization required by the threshold crossing nature of the model), and correlation $\rho \in [-1, 1]$. In the absence of additional information, what can the researcher learn about $\theta = [\delta_1 \ \delta_2 \ \beta_1 \ \beta_2 \ \rho]$?

From the econometric perspective, this is a generalization of a standard discrete choice model to a bivariate simultaneous response model which yields a stochastic representation of equilibria in a two player, two action game. Generically, for a given value of θ and realization of the payoff shifters, the model just laid out admits multiple equilibria (existence of PSNE is guaranteed because the interaction parameters are non-negative). In other words, it yields set valued predictions as depicted in Figure 3.4.[44]

Why does this set valued prediction hinder point identification? Intuitively, the challenge can be traced back to the fact that for different values of $\theta \in \Theta$, one may find different ways to assign the probability mass in $[-\boldsymbol{x}_1\beta_1, -\boldsymbol{x}_1\beta_1 - \delta_1) \times [-\boldsymbol{x}_2\beta_2, -\boldsymbol{x}_2\beta_2 - \delta_2)$ to $(0,1)$ and $(1,0)$, so as to match the observed distribution $\mathsf{P}(\boldsymbol{y}_1, \boldsymbol{y}_2 | \boldsymbol{x}_1, \boldsymbol{x}_2)$.

More formally, multiplicity of equilibria implies that the mapping from the model's exogenous variables $(\boldsymbol{x}_1, \boldsymbol{x}_2, \varepsilon_1, \varepsilon_2)$ to outcomes $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is a correspondence rather than a function. This violates the classical "principal assumptions" or "coherency conditions" for simultaneous discrete response models discussed extensively in the econometrics literature (e.g., Heckman, 1978; Gourieroux, Laffont, and Monfort, 1980; Schmidt, 1981; Maddala, 1983; Blundell and Smith, 1994). Such coherency conditions require the existence of a unique reduced form, mapping the model's exogenous variables and parameters to a unique realization of the endogenous variable; hence, they constrain the model to be recursive or triangular in nature. As pointed out by Bjorn and Vuong (1984), however, the coherency conditions shut down exactly the social interaction effect of interest by requiring, e.g., that $\delta_1\delta_2 = 0$, so that at least one player's action has no impact on the other player's payoff.

[44] This figure is based on Figure 1 in Tamer (2003).

The desire to learn about interaction effects coupled with the difficulties generated by multiplicity of equilibria prompted the earlier literature to provide at least two different ways to achieve point identification. The first one relies on imposing simplifying assumptions that shift focus to outcome features that are common across equilibria (e.g. Bresnahan and Reiss, 1988, 1990, 1991; Berry, 1992, who study entry games where the number, though not the identities, of entrants is uniquely predicted by the model in equilibrium). Unfortunately, however, these simplifying assumptions substantially constrain the amount of heterogeneity in player's payoffs that the model allows for. The second one relies on explicitly modeling a selection mechanism which specifies the equilibrium played in the regions of multiplicity (e.g. Bjorn and Vuong, 1984; Berry, 1992; Bajari, Hong, and Ryan, 2010, with Bjorn and Vuong using a constant and Bajari et al. assuming a more flexible, covariate dependent parametrization; while Berry considering two possible selection mechanism specifications, one where the incumbent moves first, and the other where the most profitable player moves first). Unfortunately, however, the chosen selection mechanism can have non-trivial effects on inference, and the data and theory might be silent on which is more appropriate. A nice example of this appears in Berry (1992, Table VII). Berry and Tamer (2006) review and extend a number of results on the identification of entry models extensively used in the empirical literature. Jovanovic (1989) discusses the observable implications of models with multiple equilibria, and within the analysis of a model with homogeneous preferences shows that partial identification is possible (see p. 1435). I refer to de Paula (2013) for a review of the literature on econometric analysis of games with multiple equilibria.

Ciliberto and Tamer (2009) show, on the other hand, that it is possible to partially identify entry models that allow for rich heterogeneity in payoffs and for any possible selection mechanism (even ones that are arbitrarily dependent on the unobservable payoff shifters after conditioning on the observed payoff shifters). In addition, Tamer (2003) provides sufficient conditions for point identification based on exclusion restrictions and large support assumptions.[45]

KEY INSIGHT 3.6: *An important conceptual contribution of Tamer (2003) is to clarify the distinction between a model which is* incoherent*, so that no reduced form exists, and a model which is* incomplete*, so that multiple reduced forms may exist. Models with multiple equilibria belong to the latter category. Whereas the earlier literature in partial identification had been motivated by* measurement problems*, e.g., missing or interval data, the work of Tamer (2003) and Ciliberto and Tamer (2009) is motivated by the fact that economic theory often does not specify how an equilibrium is selected in the regions of the exogenous variables which admit multiple equilibria. This is a conceptually completely distinct identification problem.*

Ciliberto and Tamer (2009) propose to use simple and tractable implications of the model

[45] Kline and Tamer (2012) analyze partial identification of nonparametric models of entry in a two-player model, drawing connections with the program evaluation literature.

to learn features of the structural parameters of interest. Specifically, they point out that the probability of observing any outcome of the game cannot be smaller than the model's implied probability that such outcome is the *unique* equilibrium of the game, and cannot be larger than the model's implied probability that such outcome is *one of the possible* equilibria of the game. Looking at Figure 3.4 this means, for example, that the observed $\mathsf{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,1)|\boldsymbol{x}_1, \boldsymbol{x}_2)$ cannot be smaller than the probability that $(\varepsilon_1, \varepsilon_2)$ realizes in the dotted region, and cannot be larger than the probability that it realizes either in the dotted region or in the gray region. Denote by $\Phi(A_1, A_2; \rho)$ the probability that the bivariate normal with mean vector zero, variances equal to one, and correlation ρ assigns to the event $\{\varepsilon_1 \in A_1, \varepsilon_2 \in A_2\}$. Then Ciliberto and Tamer (2009) show that any $\vartheta = [d_1, d_2, b_1, b_2, r]$ that is observationally equivalent to the data generating value θ satisfies, $(\boldsymbol{x}_1, \boldsymbol{x}_2)$-a.s.,

$$\mathsf{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,0)|\boldsymbol{x}_1, \boldsymbol{x}_2) = \Phi((-\infty, -\boldsymbol{x}_1 b_1), (-\infty, -\boldsymbol{x}_2 b_2); r) \tag{3.18}$$

$$\mathsf{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (1,1)|\boldsymbol{x}_1, \boldsymbol{x}_2) = \Phi([-\boldsymbol{x}_1 b_1 - d_1, \infty), [-\boldsymbol{x}_2 b_2 - d_2, \infty); r) \tag{3.19}$$

$$\mathsf{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,1)|\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \Phi((-\infty, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, \infty); r) \tag{3.20}$$

$$\begin{aligned}\mathsf{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,1)|\boldsymbol{x}_1, \boldsymbol{x}_2) \geq \Big\{ &\Phi((-\infty, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, \infty); r) \\ &- \Phi((-\boldsymbol{x}_1 b_1, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, -\boldsymbol{x}_2 b_2 - d_2); r) \Big\}\end{aligned} \tag{3.21}$$

While the approach of Ciliberto and Tamer (2009) is summarized here for a two player entry game, it extends without difficulty to any finite number of players and actions and to solution concepts other than pure strategy Nash equilibrium.

Aradillas-Lopez and Tamer (2008) build on the insights of Ciliberto and Tamer (2009) to study what is the identification power of equilibrium in games. To do so, they compare the set-valued model predictions and what can be learned about θ when one assumes only level-k rationality as opposed to Nash play. In static entry games of complete information, they find that the model's predictions when $k \geq 2$ are similar to those obtained with Nash behavior and allowing for multiple equilibria and mixed strategies.[46]

The collections of parameter vectors satisfying (in)equalities (3.18)-(3.21) yields the sharp identification region $\mathcal{H}_{\mathsf{P}}[\theta]$ in the case of two player entry games with pure strategy Nash equilibrium as solution concept, as shown by Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix D, Corollary D.4). When there are more than two players or more than two actions (or with different solutions concepts, such as, e.g., mixed strategy Nash equilibrium; correlated equilibrium; or rationality of level k as in Aradillas-Lopez and Tamer, 2008), the characterization in Ciliberto and Tamer (2009) obtained by extending the reasoning just laid out yields an outer region. Beresteanu, Molchanov, and Molinari (2011) use elements of random set theory to provide a general and computationally tractable characterization of

[46] Molinari and Rosen (2008) extend the analysis of Aradillas-Lopez and Tamer (2008) to the class of supermodular games.

the identification region that is sharp, regardless of the number of players and actions, or the solution concept adopted. For the case of PSNE with any finite number of players or actions, Galichon and Henry (2011) provide a computationally tractable sharp characterization of the identification region using elements of optimal transportation theory.

3.2.2 Characterization of Sharpness through Random Set Theory

Beresteanu, Molchanov, and Molinari (2011) provide a general approach based on random set theory that delivers sharp identification regions on parameters of structural semiparametric models with set valued predictions. Here I summarize it for the case of static, simultaneous move finite games of complete information, first with PSNE as solution concept and then with mixed strategy Nash equilibrium. Then I discuss games of incomplete information.

For a given $\vartheta \in \Theta$, denote the set of pure strategy Nash equilibria (depicted in Figure 3.4) as $\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \varepsilon)$. It is easy to show that $\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \varepsilon)$ is a random closed set as in Definition A.1. Under the assumption in Identification Problem 3.6 that $\boldsymbol{y}$ results from simultaneous move, pure strategy Nash play, at the true DGP value of $\theta \in \Theta$, one has

$$\boldsymbol{y} \in \boldsymbol{Y}_\theta \text{ a.s.} \tag{3.22}$$

Equation (3.22) exhausts the modeling content of Identification Problem 3.6. Theorem A.1 can be leveraged to extract its empirical content from the observed distribution $\mathsf{P}(\boldsymbol{y}, \boldsymbol{x})$. For a given set $K \subset \mathcal{Y}$, let $\mathsf{T}_{\boldsymbol{Y}_\vartheta(\boldsymbol{x},\varepsilon)}(K; \Phi_r)$ denote the probability of the event $\{\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \varepsilon) \cap K \neq \emptyset\}$ implied when $\varepsilon \sim \Phi_r$, $\boldsymbol{x}$-a.s., with Φ_r the bivariate Normal distribution with mean vector zero, variances equal to one, and covariance equal to r.

Theorem SIR-3.4 (Structural Parameters in Static, Simultaneous Move Finite Games of Complete Information with PSNE): *Under the assumptions of Identification Problem 3.6, the sharp identification region for θ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : \mathsf{P}(\boldsymbol{y} \in K | \boldsymbol{x}) \leq \mathsf{T}_{\boldsymbol{Y}_\vartheta(\boldsymbol{x},\varepsilon)}(K; \Phi_r) \, \forall K \subset \mathcal{Y}, \, \boldsymbol{x}\text{-a.s.}\}. \tag{3.23}$$

Proof. To simplify notation, let $\boldsymbol{Y}_\vartheta \equiv \boldsymbol{Y}_\vartheta(\boldsymbol{x}, \varepsilon)$. In order to establish sharpness, it suffices to show that $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ if and only if one can complete the model with an admissible selection mechanism, so that the probability distribution over outcome profiles implied by the model with that selection mechanism is equal to the probability distribution of $\boldsymbol{y}$ observed in the data. An admissible selection mechanism is a probability distribution conditional on $(\boldsymbol{x}, \varepsilon)$ and possibly dependent on ϑ, with support contained in $\boldsymbol{Y}_\vartheta$, and with no further assumptions placed on it (see, e.g., Berry and Tamer, 2006, for a formal definition). Suppose first that ϑ is such that a selection mechanism with these properties is available. Then there exists a selection of $\boldsymbol{Y}_\vartheta$ which is equal to the prediction selected by the selection mechanism and whose conditional distribution is equal to $\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})$, $\boldsymbol{x}$-a.s., and therefore $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$. Next take

$\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$. Then by Theorem A.1, $\boldsymbol{y}$ and $\boldsymbol{Y}_\vartheta$ can be realized on the same probability space as random elements $\boldsymbol{y}'$ and $\boldsymbol{Y}'_\vartheta$, so that $\boldsymbol{y}'$ and $\boldsymbol{Y}'_\vartheta$ have the same distributions, respectively, as $\boldsymbol{y}$ and $\boldsymbol{Y}_\vartheta$, and $\boldsymbol{y}' \in \text{Sel}(\boldsymbol{Y}'_\vartheta)$, where $\text{Sel}(\boldsymbol{Y}'_\vartheta)$ is the set of all measurable selections from $\boldsymbol{Y}'_\vartheta$), see Definition A.3. One can then complete the model with a selection mechanism that picks $\boldsymbol{y}'$ with probability 1, and the result follows. □

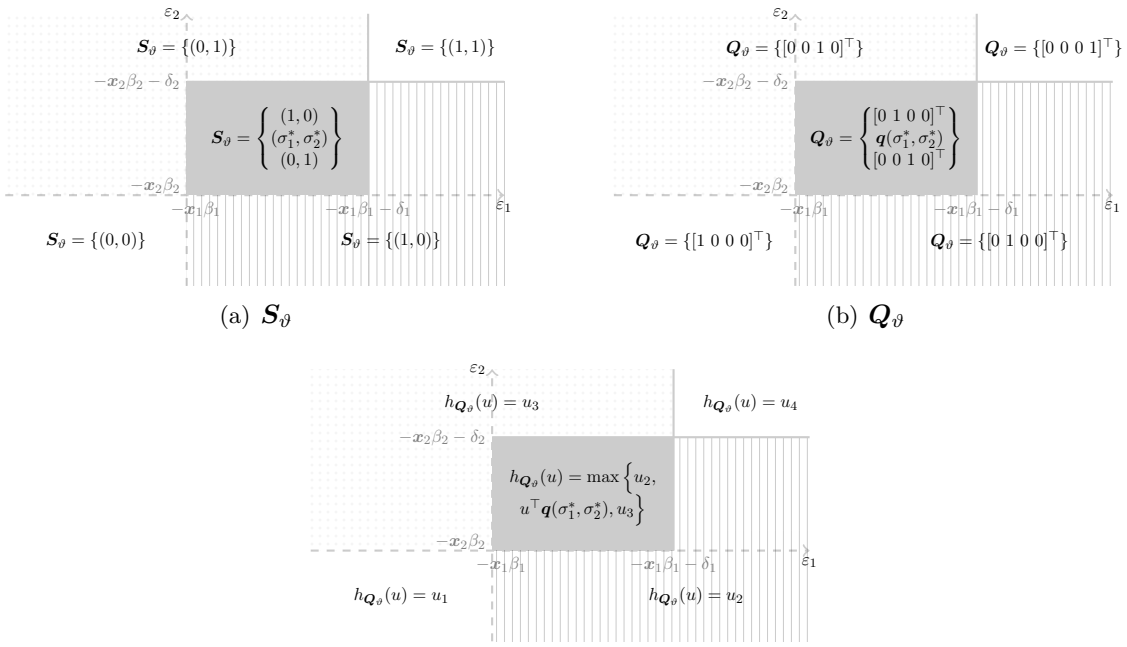
The characterization provided in Theorem SIR-3.4 for games with multiple PSNE, taken from Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix D), is equivalent to the one provided by Galichon and Henry (2011). When $J = 2$ and $\mathcal{Y} = \{0, 1\} \times \{0, 1\}$, the inequalities in (3.23) reduce to (3.18)-(3.21). With more players and/or more actions, the inequalities in (3.23) are a superset of those in (3.18)-(3.21), and are more informative.

KEY INSIGHT 3.7: *(Random set theory and partial identification – continued) In Identification Problem 3.6 lack of point identification can be traced back to the set valued predictions delivered by the model, which in turn derive from the model incompleteness defined by Tamer (2003). As stated in the Introduction, constructing the (random) set of model predictions delivered by the maintained assumptions is an exercise typically carried out in identification analysis, regardless of whether random set theory is applied. Indeed, for the problem studied in this section, Tamer (2003, Figure 1) put forward the set of admissible outcomes of the game. Beresteanu, Molchanov, and Molinari (2011) propose to work directly with this random set to characterize $\mathcal{H}_{\mathsf{P}}[\theta]$. The fundamental advantage of this approach is that it dispenses with considering the possible selection mechanisms that may complete the model. Selection mechanisms may depend on the model's unobservables even after conditioning on observables and may constitute an infinite dimensional nuisance parameter, which creates great difficulties for the computation of $\mathcal{H}_{\mathsf{P}}[\theta]$ and for inference.*

Next, I discuss the case that the outcome of the game results from simultaneous move, mixed strategy Nash play.[47] When mixed strategies are allowed for, the model predicts multiple mixed strategy Nash equilibria (MSNE). But whereas when only pure strategies are allowed for, if the model is correctly specified, the observed outcome of the game is one of the predicted PSNE, with mixed strategy it is only the result of a random mixing draw from one of the predicted MSNE. Hence, the identification problem is more complex, and in order to obtain a tractable characterization of θ's sharp identification region one needs to use different tools from random set theory.

To keep the treatment simple here I continue to consider the case of two players with two strategies, as in Identification Problem 3.6, with mixed strategies allowed for, and refer to Molchanov and Molinari (2018, Section 3.4) for the general case. Fix $\vartheta \in \Theta$. Let $\sigma_j : \{0, 1\} \to [0, 1]$ denote the probability that player j enters the market, with $1 - \sigma_j$ the probability that

[47] The same reasoning given here applies if instead of mixed strategy Nash the solution concept is correlated equilibrium, by replacing the set of MSNE below with the set of correlated equilibria.



(a) $\boldsymbol{S}_\vartheta$ (b) $\boldsymbol{Q}_\vartheta$

(c) $h_{\boldsymbol{Q}_\vartheta}$

Figure 3.5: MSNE strategies ($\boldsymbol{S}_\vartheta$), set of multinomial distributions over outcomes of the game ($\boldsymbol{Q}_\vartheta$), and its support function ($h_{\boldsymbol{Q}_\vartheta}$), as a function of $(\varepsilon_1, \varepsilon_2)$, where $\sigma_1^* \equiv \frac{-\varepsilon_2 - \boldsymbol{x}_2\beta_2}{\vartheta_2}$, $\sigma_2^* \equiv \vartheta_2$, $\frac{-\varepsilon_1 - \boldsymbol{x}_1\beta_1}{\vartheta_1}$.

she stays out. With some abuse of notation, let $\pi_j(\sigma_j, \sigma_{-j}, \boldsymbol{x}_j, \varepsilon_j, \vartheta)$ denote the expected payoff associated with the mixed strategy profile $\sigma = (\sigma_1, \sigma_2)$. For a given realization (w, e) of $(\boldsymbol{x}, \varepsilon)$ and a given value of $\vartheta \in \Theta$, the set of mixed strategy Nash equilibria is

$$S_\vartheta(w,e) = \left\{ \sigma \in [0,1]^2 : \ \pi_j(\sigma_j, \sigma_{-j}, w_j, e_j; \vartheta) \geq \pi_j(\tilde{\sigma}_j, \sigma_{-j}, w_j, e_j; \vartheta) \ \forall \tilde{\sigma}_j \in [0,1] \ j = 1,2 \right\}.$$

Beresteanu, Molchanov, and Molinari (2011) show that $\boldsymbol{S}_\vartheta \equiv S_\vartheta(\boldsymbol{x}, \varepsilon)$ is a random closed set in $[0,1]^2$. Its realizations are illustrated in Panel (a) of Figure 3.5 as a function of $(\varepsilon_1, \varepsilon_2)$.[48]

Define the set of possible multinomial distributions over outcomes of the game associated with the selections σ of each possible realization of $\boldsymbol{S}_\vartheta$ as

$$\boldsymbol{Q}_\vartheta = \left\{ \boldsymbol{q}(\sigma) \equiv \begin{bmatrix} (1-\sigma_1)(1-\sigma_2) \\ \sigma_1(1-\sigma_2) \\ (1-\sigma_1)\sigma_2 \\ \sigma_1\sigma_2 \end{bmatrix} : \sigma \in \boldsymbol{S}_\vartheta \right\}. \tag{3.24}$$

As $\boldsymbol{Q}_\vartheta$ is the image of a continuous map applied to the random compact set $\boldsymbol{S}_\vartheta$, it is a random compact set. Its realizations are plotted in Panel (b) of Figure 3.5 as a function of

[48] This figure is based on Figure 1 in Beresteanu, Molchanov, and Molinari (2011).

$(\varepsilon_1, \varepsilon_2)$.

The multinomial distribution over outcomes of the game determined by a given $\sigma \in \boldsymbol{S}_\vartheta$ is a function of ε. In order to obtain the predicted distribution over outcomes of the game conditional on observed payoff shifters, one needs to integrate out the unobservable payoff shifters ε. Doing so requires care, as it needs to be done for each $\boldsymbol{q}(\sigma) \in \boldsymbol{Q}_\vartheta$. First, observe that all the $\boldsymbol{q}(\sigma) \in \boldsymbol{Q}_\vartheta$ are contained in the 3 dimensional unit simplex, and are therefore integrable. Next, define the conditional selection expectation (see Definition A.4) of $\boldsymbol{Q}_\vartheta$ as

$$\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x}) = \left\{\mathbb{E}_{\Phi_r}(\boldsymbol{q}(\sigma)|\boldsymbol{x}) : \ \sigma \in \text{Sel}(\boldsymbol{S}_\vartheta)\right\},$$

where $\text{Sel}(\boldsymbol{S}_\vartheta)$ is the set of all measurable selections from $\boldsymbol{S}_\vartheta$, see Definition A.3. By construction, $\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$ is the set of probability distributions over action profiles conditional on $\boldsymbol{x}$ which are consistent with the maintained modeling assumptions, i.e. with *all* the model's implications (including the assumption that $\varepsilon \sim \Phi_r$). If the model is correctly specified, there exists at least one vector $\theta \in \Theta$ such that the observed conditional distribution $p(\boldsymbol{x}) = [\mathsf{P}(\boldsymbol{y} = y^1|\boldsymbol{x}), \ldots, \mathsf{P}(\boldsymbol{y} = y^4|\boldsymbol{x})]^\top$ almost surely belongs to the set $\mathbb{E}_{\Phi_\rho}(\boldsymbol{Q}_\theta|\boldsymbol{x})$. Indeed, by the definition of $\mathbb{E}_{\Phi_\rho}(\boldsymbol{Q}_\theta|\boldsymbol{x})$, $p(\boldsymbol{x}) \in \mathbb{E}_{\Phi_\rho}(\boldsymbol{Q}_\theta|\boldsymbol{x})$ almost surely if and only if there exists $\boldsymbol{q} \in \text{Sel}(\boldsymbol{Q}_\theta)$ such that $\mathbb{E}_{\Phi_\rho}(\boldsymbol{q}|\boldsymbol{x}) = p(\boldsymbol{x})$ almost surely, with $\text{Sel}(\boldsymbol{Q}_\theta)$ the set of all measurable selections from $\boldsymbol{Q}_\theta$. Hence, the collection of parameter vectors $\vartheta \in \Theta$ that are observationally equivalent to the data generating value θ is given by the ones that satisfy $p(\boldsymbol{x}) \in \mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$ almost surely. In turn, observing that by Theorem A.2 the set $\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$ is convex, we have that $p(\boldsymbol{x}) \in \mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$ if and only if $u^\top p(\boldsymbol{x}) \leq h_{\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})}(u)$ for all u in the unit ball (see, e.g., Rockafellar, 1970, Theorem 13.1), where $h_{\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})}(u)$ is the support function of $\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$, see Definition A.5.

THEOREM SIR-3.5 (Structural Parameters in Static, Simultaneous Move Finite Games of Complete Information with MSNE): *Under the assumptions in Identification Problem 3.6, allowing for mixed strategies and with the observed outcomes of the game resulting from mixed strategy Nash play, the sharp identification region for θ is*

$$\mathcal{H}_\mathsf{P}[\theta] = \left\{\vartheta \in \Theta : \ \left(\max_{u:\, \|u\| \leq 1} u^\top p(\boldsymbol{x}) - \mathbb{E}_{\Phi_r}[h_{\boldsymbol{Q}_\vartheta}(u)|\boldsymbol{x}]\right) = 0, \ \boldsymbol{x}\text{-}a.s.\right\} \tag{3.25}$$

$$= \left\{\vartheta \in \Theta : \ \int_{\|u\| \leq 1} (u^\top p(\boldsymbol{x}) - \mathbb{E}_{\Phi_r}[h_{\boldsymbol{Q}_\vartheta}(u)|\boldsymbol{x}])_+ \mathrm{d}\mu(u) = 0, \ \boldsymbol{x}\text{-}a.s.\right\}, \tag{3.26}$$

where μ is any probability measure on the unit ball in $\mathbb{R}^4$.

Proof. Theorem A.2 (equation (A.10)) yields (3.25), because by the arguments given before the theorem, $\mathcal{H}_\mathsf{P}[\theta] = \{\vartheta \in \Theta : \ p(\boldsymbol{x}) \in \mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x}), \ \boldsymbol{x}\text{-a.s.}\}$. The result in (3.26) follows because the integrand in (3.26) is continuous in u and both conditions inside the curly brackets are satisfied if and only if $u^\top p(\boldsymbol{x}) - \mathbb{E}_{\Phi_r}[h_{\boldsymbol{Q}_\vartheta}(u)|\boldsymbol{x}] \leq 0 \ \forall u \in \mathbb{B}^4 \ \boldsymbol{x}\text{-a.s.}$ □

For a fixed $u \in \mathbb{B}^4$, the possible realizations of $h_{\boldsymbol{Q}_\vartheta}(u)$ are plotted in Panel (c) of Figure 3.5 as a function of $(\varepsilon_1, \varepsilon_2)$. The expectation of $h_{\boldsymbol{Q}_\vartheta}(u)$ is quite straightforward to compute, whereas calculating the set $\mathbb{E}_{\Phi_r}(\boldsymbol{Q}_\vartheta|\boldsymbol{x})$ is computationally prohibitive in many cases. Hence, the characterization in (3.25) is computationally attractive, because for each $\vartheta \in \Theta$ it requires to maximize an easy-to-compute superlinear, hence concave, function over a convex set, and check if the resulting objective value vanishes. This problem is computationally tractable and several efficient algorithms in convex programming are available to solve it, see for example the MatLab software for disciplined convex programming CVX (Grant and Boyd, 2010). Nonetheless, $\mathcal{H}_{\mathsf{P}}[\theta]$ itself is not necessarily convex, hence tracing out its boundary is non-trivial. I return to computational challenges in partial identification in Section 6.

KEY INSIGHT 3.8 (Random set theory and partial identification – continued): *Beresteanu, Molchanov, and Molinari (2011) provide a general characterization of sharp identification regions for models with* convex moment predictions. *These are models that for a given $\vartheta \in \Theta$ and realization of observable variables, predict a set of values for a vector of variables of interest. This set is* not *necessarily convex, as exemplified by $\boldsymbol{Y}_\vartheta$ and $\boldsymbol{Q}_\vartheta$, which are finite. No restriction is placed on the manner in which, in the DGP, a specific model prediction is selected from this set. When the researcher takes conditional expectations of the resulting elements of this set, the unrestricted process of selection yields a convex set of moments for the model variables (all possible mixtures). This is the model's convex set of moment predictions. If this set were almost surely single valued, the researcher would learn (features of) θ by solving moment equality conditions involving the observed variables and predicted ones. The approach reviewed in this section is a set-valued method of moments that extends the singleton-valued one commonly used in econometrics.*

I conclude this section discussing the case of static, simultaneous move finite games of incomplete information, using the results in Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix C).[49] For clarity, I formalize the maintained assumptions.

IDENTIFICATION PROBLEM 3.7 (Structural Parameters in Static, Simultaneous Move Finite Games of Incomplete Information with multiple BNE): Impose the same structure on payoffs, entry decision rule, outcome space, parameter space, and observable variables as in Identification Problem 3.6. Assume that the observed outcome of the game results from simultaneous move, pure strategy Bayesian Nash play. Both players and the researcher observe $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. However, ε_j is private information to player $j = 1, 2$ and unobservable to the researcher, with $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 | (\boldsymbol{x}_1, \boldsymbol{x}_2)$. Assume that players have correct common prior F_γ

[49] See Berry and Tamer (2006, Section 3) and Grieco (2014) for a thorough discussion of the literature on identification problems in games of incomplete information with multiple Bayesian Nash equilibria (BNE). Berry and Tamer (2006) explain how to extend the approach proposed by Ciliberto and Tamer (2009) to obtain outer regions on θ when no restrictions are imposed on the equilibrium selection mechanism that chooses among the multiple BNE.

on the distribution of $(\varepsilon_1, \varepsilon_2)$ and the researcher knows this distribution up to γ, a finite dimensional parameter vector. Under these assumptions, multiple Bayesian Nash equilibria (BNE) may result.[50] In the absence of additional information, what can the researcher learn about $\theta = [\delta_1 \ \delta_2 \ \beta_1 \ \beta_2 \ \gamma]$?

With incomplete information, players' strategies are decision rules that map the support of $(\varepsilon, \boldsymbol{x})$ into $\{0, 1\}$. The non-negativity condition on expected payoffs that determines each player's decision to enter the market results in equilibrium mappings (decision rules) that are step functions determined by a threshold: $y_j(\varepsilon_j) = \mathbf{1}(\varepsilon_j \geq t_j), j = 1, 2$. As a result, player j's beliefs about player $3 - j$'s probability of entry under the common prior assumption is $\int y_{3-j}(\varepsilon_{3-j}) d\mathsf{F}_\gamma(\varepsilon_{3-j}|\boldsymbol{x}) = 1 - \mathsf{F}_\gamma(t_{3-j}|\boldsymbol{x})$, and therefore player j's best response cutoff is

$$t_j^b(t_{3-j}, \boldsymbol{x}; \theta) = -\boldsymbol{x}_j\beta_j - \delta_j(1 - \mathsf{F}_\gamma(t_{3-j}|\boldsymbol{x})).$$

Hence, the set of equilibria can be defined as the set of cutoff rules:

$$\boldsymbol{T}_\theta(\boldsymbol{x}) = \{(t_1, t_2) : t_j = t_j^b(t_{3-j}, \boldsymbol{x}; \theta), \ j = 1, 2\}.$$

The equilibrium thresholds are functions of $\boldsymbol{x}$ and θ only. The set $\boldsymbol{T}_\theta(\boldsymbol{x})$ might contain a finite number of equilibria (e.g., if the common prior is the Normal distribution), or a continuum of equilibria. For ease of notation I suppress its dependence on $\boldsymbol{x}$ in what follows.

Given the equilibrium decision rules (the selections of the set $\boldsymbol{T}_\theta$), it is possible to determine their associated action profiles. Because in the simple two-player entry game that I consider actions and outcomes coincide, I denote the set of admissible action profiles by $\boldsymbol{Y}_\theta$:

$$\boldsymbol{Y}_\theta = \left\{ \boldsymbol{y}(\boldsymbol{t}) \equiv \begin{bmatrix} \mathbf{1}(\varepsilon_1 < \boldsymbol{t}_1, \varepsilon_2 < \boldsymbol{t}_2) \\ \mathbf{1}(\varepsilon_1 \geq \boldsymbol{t}_1, \varepsilon_2 < \boldsymbol{t}_2) \\ \mathbf{1}(\varepsilon_1 < \boldsymbol{t}_1, \varepsilon_2 \geq \boldsymbol{t}_2) \\ \mathbf{1}(\varepsilon_1 \geq \boldsymbol{t}_1, \varepsilon_2 \geq \boldsymbol{t}_2) \end{bmatrix} : \boldsymbol{t} \in \text{Sel}(\boldsymbol{T}_\theta) \right\}, \tag{3.27}$$

with $\text{Sel}(\boldsymbol{T}_\theta)$ the set of all measurable selections from $\boldsymbol{T}_\theta$, see Definition A.3. To obtain the predicted set of multinomial distributions for the outcomes of the game, one needs to integrate out ε conditional on $\boldsymbol{x}$. Again this can be done by using the conditional Aumann expectation:

$$\mathbb{E}_{\mathsf{F}_\gamma}(\boldsymbol{Y}_\theta|\boldsymbol{x}) = \{\mathbb{E}_{\mathsf{F}_\gamma}(\boldsymbol{y}(\boldsymbol{t})|\boldsymbol{x}) : \boldsymbol{t} \in \text{Sel}(\boldsymbol{T}_\theta)\}.$$

This set is closed and convex. Regardless of whether $\boldsymbol{T}_\theta$ contains a finite number of equilibria or a continuum, $\boldsymbol{Y}_\theta$ can take on only a finite number of realizations corresponding to each of the vertices of the three dimensional simplex, because the vectors $\boldsymbol{y}(\boldsymbol{t})$ in (3.27) collect

[50] Both the independence assumption and the correct common prior assumption are maintained here to simplify exposition. Both could be relaxed with no conceptual difficulty, though computation of the set of Bayesian Nash equilibria, for example, would become more cumbersome.

threshold decision rules. This implies that $\mathbb{E}_{\mathsf{F}_\gamma}(\boldsymbol{Y}_\theta|\boldsymbol{x})$ is a closed convex polytope $\boldsymbol{x}$-a.s., fully characterized by a finite number of supporting hyperplanes. Hence, it is possible to determine whether $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ using efficient algorithms in linear programming.

THEOREM SIR-3.6 (Structural Parameters in Static, Simultaneous Move Finite Games of Incomplete Information with BNE): *Under the assumptions in Identification Problem 3.7, the sharp identification region for θ is*

$$\mathcal{H}_{\mathsf{P}}[\theta] = \left\{ \vartheta \in \Theta : \max_{u:\, \|u\| \leq 1} u^\top p(\boldsymbol{x}) - \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}] = 0,\ \boldsymbol{x}\text{-a.s.} \right\} \tag{3.28}$$

$$= \left\{ \vartheta \in \Theta : \ u^\top p(\boldsymbol{x}) \leq \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}] = 0,\ \forall u \in D, \boldsymbol{x}\text{-a.s.} \right\}, \tag{3.29}$$

$$= \left\{ \vartheta \in \Theta : \ \mathsf{P}(\boldsymbol{y} \in K|\boldsymbol{x}) \leq \mathsf{T}_{\boldsymbol{Y}_\vartheta(\boldsymbol{x},\varepsilon)}(K; \mathsf{F}_{\tilde{\gamma}})\, \forall K \subset \mathcal{Y},\ \boldsymbol{x}\text{-a.s.} \right\}, \tag{3.30}$$

where $D = \{u = [u_1, \ldots, u_{|\mathcal{Y}|}]^\top : u_i \in \{0,1\}, i = 1, ..., |\mathcal{Y}|\}$, $\vartheta = [d_1, d_2, b_1, b_2, \tilde{\gamma}]$, and $\mathsf{T}_{\boldsymbol{Y}_\vartheta(\boldsymbol{x},\varepsilon)}(K; \mathsf{F}_{\tilde{\gamma}})$ denotes the probability of the event $\{\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \varepsilon) \cap K \neq \emptyset\}$ implied when $\varepsilon \sim \mathsf{F}_{\tilde{\gamma}}$, $\boldsymbol{x}$-a.s.

Proof. The result in (3.28) follows by the same argument as in the proof of Theorem SIR-3.5. Next I show equivalence of the conditions

$$(i)\ p(\boldsymbol{x}) \leq \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}]\ \forall u : \|u\| \leq 1,$$
$$(ii)\ p(\boldsymbol{x}) \leq \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}]\ \forall u \in D.$$

By the positive homogeneity of the support function, condition (i) is equivalent to $p(\boldsymbol{x}) \leq \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}]\ \forall u \in \mathbb{R}^{|\mathcal{Y}|}$, which implies condition (ii). Next I show that condition (ii) implies condition (i). As explained before, the set $\boldsymbol{Y}_\theta$, and hence also its convex hull $\text{conv}(\boldsymbol{Y}_\theta)$, can take on only a finite number of realizations. Let $Y_1, \ldots, Y_m$ be convex compact sets in the simplex of dimension $|\mathcal{Y}| - 1$ equal to the possible realizations of $\text{conv}(\boldsymbol{Y}_\theta)$, and let $\varpi_1(\boldsymbol{x}), \ldots, \varpi_m(\boldsymbol{x})$ denote the probability of each of these realizations conditional on $\boldsymbol{x}$. Then by Theorem 2.1.34 in Molchanov (2017), $\mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}(\boldsymbol{Y}_\theta|\boldsymbol{x}) = \sum_{j=1}^m Y_j \varpi_j(\boldsymbol{x})$. By the properties of the support function (see, e.g., Schneider, 1993, Theorem 1.7.5), $h_{\mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}(\boldsymbol{Y}_\theta|\boldsymbol{x})}(u) = \sum_{j=1}^m \varpi_j(\boldsymbol{x}) h_{Y_j}(u)$. For each $j = 1, ..., m$, the vertices of Y_j are a subset of the vertices of the $(|\mathcal{Y}| - 1)$-dimensional simplex. Hence the supporting hyperplanes of $Y_j, j = 1, ..., m$, are a subset of the supporting hyperplanes of that simplex, which in turn are obtained through its support function evaluated in directions $u \in D$. Finally, I show equivalence with the result in (3.30). Because the vertices of Y_j are a subset of the vertices of the $(|\mathcal{Y}| - 1)$-dimensional simplex, each direction $u \in D$ determines a set $K_u \subset \mathcal{Y}$. Given the choice of u, the value of

$u^\top \boldsymbol{y}(\boldsymbol{t})$ equals one if $\boldsymbol{y}(\boldsymbol{t}) \in Y_u$ and zero otherwise. Hence, condition (3.29) reduces to

$$\begin{aligned}\mathsf{P}(\boldsymbol{y} \in K_u|\boldsymbol{x}) = u^\top p(\boldsymbol{x}) \leq \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[h_{\boldsymbol{Y}_\vartheta}(u)|\boldsymbol{x}] = \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}\left[\sup_{\boldsymbol{y}(\boldsymbol{t})\in\boldsymbol{Y}_\vartheta} u^\top \boldsymbol{y}(\boldsymbol{t})|\boldsymbol{x}\right] \\ = \mathbb{E}_{\mathsf{F}_{\tilde{\gamma}}}[\mathbf{1}(\boldsymbol{Y}_\vartheta \cap K_u \neq \emptyset)|\boldsymbol{x}] = \mathsf{T}_{\boldsymbol{Y}_\vartheta(\boldsymbol{x},\varepsilon)}(K_u; \mathsf{F}_{\tilde{\gamma}}).\end{aligned}$$

Observing that the collection D comprises the $2^{|\mathcal{Y}|}$ vectors with entries equal to either 1 or 0, and that these determine all possible subsets K_u of $\mathcal{Y}$, yields condition (3.30). □

One can use the same argument as in the proof of Theorem SIR-3.6, to show that the Aumann expectation/support function characterization of the sharp identification region in Theorem SIR-3.5 coincides with the characterization based on the capacity functional in Theorem SIR-3.4, when only pure strategies are allowed for. This shows that in this class of models, the capacity functional based characterization is a special case of the Aumann expectation/support function based one.

Aradillas-Lopez and Tamer (2008) study what is the identification power of equilibrium also in the case of static entry games with incomplete information. They show that in the presence of multiple equilibria, assuming Bayesian Nash behavior yields more informative regions for the parameter vector θ than assuming only rational behavior, but at the price of a higher computational cost.

de Paula and Tang (2012) propose a procedure to test for the sign of the interaction effects (which here I have assumed to be non-positive) in discrete simultaneous games with incomplete information and (possibly) multiple equilibria. As a by-product of this procedure, they also provide a test for the presence of multiple equilibria in the DGP. The test does not require parametric specifications of players' payoffs, the distributions of their private signals, or the equilibrium selection mechanism. Rather, the test builds on the commonly invoked assumption that players' private signals are independent conditional on observed states.

Grieco (2014) introduces an important class of models with flexible information structure. Each player is assumed to have a vector of payoff shifters unobservable by the researcher composed of elements that are private information to the player, and elements that are known to all players. The results of Beresteanu, Molchanov, and Molinari (2011) reported in this section apply to this set-up as well.

3.3 Auction Models with Independent Private Values

3.3.1 An Inference Approach Robust to Bidding Behavior Assumptions

Haile and Tamer (2003) study what can be learned about the distribution of valuations in an open outcry English auction where symmetric bidders have independent private values for the object being auctioned. The standard theoretical model (Milgrom and Weber, 1982), called "button auction" model, posits that each bidder holds down a button while the object's

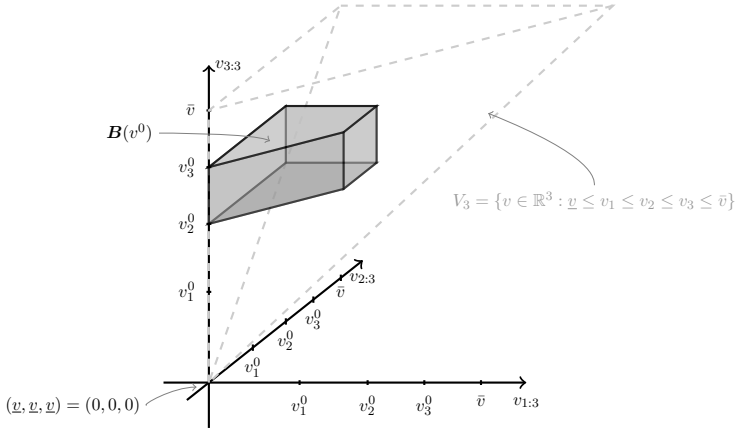


Figure 3.6: A realization of the model predicted ordered bids $\boldsymbol{B}(\vec{\boldsymbol{v}}_n)$ in (3.31) for $n = 3, \vec{\boldsymbol{v}}_n = v^0, \delta = 0$.

price rises continuously and exogenously, releasing it (in the dominant strategy equilibrium) when it reaches her valuation or all her opponents have left. In this case, the distribution of bidder's valuation can be learned exactly. Haile and Tamer (2003) show that much can be learned about the distribution of valuations, even allowing for the fact that real-life auctions may depart from this stylized framework, as in the following identification problem.[51]

Identification Problem 3.8 (Incomplete Auction Model with Independent Private Values): For a given auction with $n < \infty$ participating bidders, let $\boldsymbol{v}_i \sim \mathsf{Q}, i = 1, \ldots, n$, be bidder i's valuation for the object being auctioned and assume that $\boldsymbol{v}_i \perp\!\!\!\perp \boldsymbol{v}_j$ for all $i \neq j$. Assume that the support of Q is $[\underline{v}, \bar{v}]$ and that each bidder knows her own valuation but not that of her opponents. Let the auctioneer set a minimum bid increment $\delta \in [0, \bar{v})$, and for simplicity suppose there is no reserve price.[52] Suppose the researcher observes order statistics of the bids, $\vec{\boldsymbol{b}}_n \equiv (\boldsymbol{b}_{1:n}, \ldots, \boldsymbol{b}_{n:n}) \sim \mathsf{P}$ in $\mathbb{R}^n_+$, with $\boldsymbol{b}_{i:n}$ the i-th lowest of the n bids. Assume that: (1) Bidders do not bid more than they are willing to pay; (2) Bidders do not allow an opponent to win at a price they are willing to beat. In the absence of additional information, what can the researcher learn about Q?

The model in Identification Problem 3.8 delivers set valued predictions because given valuations $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, the two fundamental assumptions about bidder's behavior yield

$$\vec{\boldsymbol{b}}_n \in \boldsymbol{B}(\vec{\boldsymbol{v}}_n) \equiv \left[\left\{ \prod_{i=1}^{n-1} [\underline{v}, \boldsymbol{v}_{i:n}] \right\} \times [\boldsymbol{v}_{n-1:n} - \delta, \boldsymbol{v}_{n:n}] \right] \cap V_n, \tag{3.31}$$

where $\vec{\boldsymbol{v}}_n \equiv (\boldsymbol{v}_{1:n}, \ldots, \boldsymbol{v}_{n:n})$ denotes the vector of order statistics of the valuations, and

[51] Examples of departures from the standard model include the case where active bidding by a player's opponents may eliminate her incentives to bid close to her valuation or at all; the econometrician does not precisely observe the point at which each bidder drops out; there are discrete bid increments; etc.

[52] If there is a reserve price $r > \underline{v}$, nothing can be learned about $\mathsf{Q}(\boldsymbol{v} \in [\underline{v}, v])$ for any $v < r$. In that case, one can learn features of the truncated distribution of valuations using the same insights summarized here.

$V_n = \{v \in \mathbb{R}^n : \underline{v} \leq v_1 \leq v_2 \leq \cdots \leq v_n \leq \bar{v}\}$.[53] Figure 3.6 provides a stylized depiction of a realization of this set for $\vec{\boldsymbol{v}}_n = v^0$ when there are three bidders ($n = 3$), $\underline{v} = 0$, and $\delta = 0$. In words, $\boldsymbol{B}(\vec{\boldsymbol{v}}_n)$ collects the model predicted values of ordered bids. The fact that $\boldsymbol{b}_{i:n} \leq \boldsymbol{v}_{i:n}$ for all i results from assumption (1): since each bidder bids at most an amount equal to her valuation, the i-th highest bid cannot exceed the i-th highest valuation (Haile and Tamer, 2003, Lemma 1).[54] The fact that $\boldsymbol{b}_{n:n} \geq \boldsymbol{v}_{n-1,n} - \delta$ follows immediately from assumption (2) (Haile and Tamer, 2003, Lemma 3). The fact that $\vec{\boldsymbol{b}}_n$ has to lie in V_n follows because it is a vector of *ordered* bids.

Why does this set-valued prediction hinder point identification? The reason is that the distribution of the observable data relates to the model structure in an *incomplete* manner.[55] Define a bidding rule $\mathsf{B}(\boldsymbol{b}_{1:n}, \ldots, \boldsymbol{b}_{n:n} | \boldsymbol{v}_{1:n}, \ldots, \boldsymbol{v}_{n:n})$ to be a conditional joint distribution for the order statistics of the bids conditional on the order statistics of the valuations. Then, for a given realization of the valuations $\boldsymbol{v}_{1:n} = v_1, \ldots, \boldsymbol{v}_{n:n} = v_n$, the model requires that the support of Q is in $B(\vec{v})$ as defined in (3.31) with $\boldsymbol{v}_{1:n} = v_1, \ldots, \boldsymbol{v}_{n:n} = v_n$, but imposes no other restriction on it. Hence, the model implied joint distribution of ordered bids is

$$\mathsf{M}_{1,\ldots,n:n}(\cdot; \mathsf{B}, \mathsf{Q}) \equiv \int_{B(\vec{v})} \mathsf{B}(\cdot | v_1, \ldots, v_n) \mathsf{Q}_{1,\ldots,n:n}(dv_1, \ldots, dv_n), \tag{3.32}$$

where $\mathsf{Q}_{1,\ldots,n:n}$ is the joint distribution of order statistics of the valuations implied by Q. Since the bidding rule B is left completely unspecified (other than requiring it to be a valid joint conditional probability distribution with support in $\boldsymbol{B}$), one can find multiple pairs (B, Q) satisfying the assumptions of Identification Problem 3.8, such that $\mathsf{M}_{1,\ldots,n:n}(\cdot; \mathsf{B}, \mathsf{Q}) = \mathsf{G}_{1,\ldots,n:n}(\cdot)$, with $\mathsf{G}_{1,\ldots,n:n}$ the observed joint distribution of the order statistics of the bids.

Haile and Tamer (2003) propose to use simple and tractable implications of the model to learn features of Q. Recall that with i.i.d. valuations, the distribution of each order statistic uniquely determines $\mathsf{Q}(v)$, with $\mathsf{Q}(v) \equiv \mathsf{Q}(\boldsymbol{v} \leq v)$ for any $v \geq \underline{v}$, through:

$$\mathsf{Q}(v) = q_{\mathcal{B}}(\mathsf{Q}_{i:n}(v); i, n - i + 1), \tag{3.33}$$

where $\mathsf{Q}_{i:n}$ is the CDF of $\boldsymbol{v}_{i:n}$ and $q_{\mathcal{B}}(\cdot; i, n-i+1)$ is the quantile function of a Beta-distributed random variable with parameters i and $n - i + 1$. Using this, their Lemmas 1 and 3 yield,

[53] Using the same convention as for the bids, $\boldsymbol{v}_{i:n}$ denotes the i-th lowest of the n valuations.

[54] Note that $\boldsymbol{b}_{i:n}$ needs not be the bid made by the bidder with valuation $\boldsymbol{v}_{i:n}$.

[55] Haile and Tamer (2003, Appendix D) provide the discussion summarized here. Additionally, in their Appendix B, they give a simple example of a two-bidder auction satisfying all assumptions in Identification Problem 3.8, where two different distributions Q and $\tilde{\mathsf{Q}}$ yield the same distribution of ordered bids.

respectively,

$$\mathsf{Q}(v) \leq \min_{n,i} q_{\mathcal{B}}(\mathsf{G}_{i:n}(v); i, n-i+1), \ \forall v \in [\underline{v}, \bar{v}], \tag{3.34}$$

$$\mathsf{Q}(v) \geq \max_{n} q_{\mathcal{B}}(\mathsf{G}_{n:n}(v-\delta); i, n-i+1), \ \forall v \in [\underline{v}, \bar{v}], \tag{3.35}$$

where, for any $v \geq \underline{v}$, $\mathsf{G}_{i:n}(v) \equiv \mathsf{P}(\boldsymbol{b}_{i:n} \leq v)$ denotes the observed CDF of $\boldsymbol{b}_{i:n}$ for $i = 1, \ldots, n$.

KEY INSIGHT 3.9: *The model and analysis put forward by Haile and Tamer (2003) trade point identification of the distribution of valuation under stringent assumptions on the bidding rule, for a* robust *inference approach that yields informative bounds under weak and widely credible assumptions on bidding behavior. Remarkably, "nothing is lost" due to the use of their robust approach: point identification is recovered when the standard assumptions of the button auction model hold.*[56] *This is because in the dominant strategy equilibrium the top losing bidder exits at her valuation, followed immediately by the winning bidder. Hence,* $\boldsymbol{b}_{n-1:n} = \boldsymbol{v}_{n-1:n} = \boldsymbol{b}_{n:n}$ *and* $\delta = 0$*, so that the upper and the lower bound in* (3.34)-(3.35) *coincide and point identify the distribution of valuations.*

Haile and Tamer (2003) also provide sharp bounds on the optimal reserve price, which I do not discuss here. However, they leave open the question of whether the collection of CDFs satisfying (3.34)-(3.35) yields the sharp identification region for Q. As discussed in Sections 2.1-2.3, pointwise bounds on the CDF deliver tubes of admissible CDFs that in general yield outer regions on the CDF of interest. But in this identification problem, the issue of sharpness is even more subtle, and therefore addressed in the following subsection.

Before moving on to that discussion, I note that the work of Haile and Tamer (2003) spurred a rich literature applying partial identification analysis to the study of auction models. Tang (2011) studies first price sealed bid auctions with equilibrium behavior, where affiliated valuations prevent –in the absence of parametric restrictions on the distribution of the model primitives– point identification of the model. He derives bounds on seller revenue under various counterfactual scenarios on reserve prices and auction formats. Armstrong (2013) also studies first price sealed bid auctions with equilibrium behavior, but relaxes the independence assumptions on symmetric valuations by requiring it to hold only conditional on unobserved heterogeneity. He derives bounds on various functionals of the distributions of interest, including the mean bid and mean valuation. Aradillas-López, Gandhi, and Quint (2013) analyze second price auctions with correlated private values. In this case, the distribution of valuations is not point identified even under the assumptions of the button auction model (Athey and Haile, 2002, Theorem 4). Nonetheless, Aradillas-López, Gandhi, and Quint (2013) show that interesting functionals of it (seller profits and bidder surplus) can be bounded, if one assumes that transaction prices are determined by the second highest valuation and imposes

[56] The button auction model yields bidding behavior consistent with Identification Problem 3.8.

some restrictions on the joint distribution of the number of bidders and distribution of the valuations. Komarova (2013) studies a related model of second-price ascending auctions with arbitrary dependence in bidders' private values. She provides partial identification results for the joint distribution of values for any subset of bidders under various assumptions about what data the researcher observes. While in her framework the highest bid is never observed, she considers the case where only the winner's identity and the winning price are observed, and the case where all the identities and all the bids except for the highest bid are known. She also investigates the informational content of assuming positive dependence in bidders' values. Gentry and Li (2014) are concerned with nonparametric identification of a two-stage entry and bidding game. Potential bidders are assumed to have private valuations and observe private signals before deciding whether to enter the auction. The dependence between signals and valuations is only minimally restricted. Hence, even with some excluded instruments that affect selection into the auction, the model primitives are only partially identified. The authors derive bounds on these primitives, and provide conditions under which point identification is restored.

3.3.2 Characterization of Sharpness through Random Set Theory

Haile and Tamer's 2003 bounds exploit the information contained in the *marginal* distributions $\mathsf{G}_{i:n}$ for each i and n. However, in Identification Problem 3.8 additional information can be extracted from the *joint* distribution of ordered bids. Chesher and Rosen (2017b) obtain the sharp identification region $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}]$ using random set methods (Artstein's characterization in Theorem A.1) applied to a quantile function representation of the order statistics. Here I provide an equivalent characterization that uses equation (3.31) directly. Let $\mathcal{T}$ denote the space of probability distributions with support on $[\underline{v}, \bar{v}]$, so that $\mathsf{Q} \in \mathcal{T}$. For a candidate distribution $\tilde{\mathsf{Q}} \in \mathcal{T}$, let $\tilde{\mathsf{Q}}_{1,\ldots,n:n}$ denote the implied distribution of order statistics of n i.i.d. random variables distributed $\tilde{\mathsf{Q}}$. Let $\tilde{\boldsymbol{B}}$ be a random closed set defined as in (3.31) with respect to order statistics of i.i.d. random variables with distribution $\tilde{\mathsf{Q}}$. For a given set $K \in \mathcal{K}$, with $\mathcal{K}$ the collection of compact subsets of $\mathbb{R}^n$, let $\mathsf{T}_{\tilde{\boldsymbol{B}}}(K; \tilde{\mathsf{Q}})$ denote the probability of the event $\{\tilde{\boldsymbol{B}} \cap K \neq \emptyset\}$ implied by $\tilde{\mathsf{Q}}$.

Theorem SIR-3.7 (Distribution of Valuations in Incomplete Auction Model with Independent Private Values): *Under the assumptions of Identification Problem 3.8, the sharp identification region for* Q *is*

$$\mathcal{H}_{\mathsf{P}}[\mathsf{Q}] = \left\{ \tilde{\mathsf{Q}} \in \mathcal{T} : \mathsf{P}(\vec{\boldsymbol{b}}_n \in K) \leq \mathsf{T}_{\tilde{\boldsymbol{B}}}(K; \tilde{\mathsf{Q}}) \ \forall K \in \mathcal{K} \right\}. \tag{3.36}$$

Proof. The sharp identification region for Q is given by the collection of probability distributions $\tilde{\mathsf{Q}} \in \mathcal{T}$ for which one can find a bidding rule $\mathsf{B}(\cdot|\cdot)$ with support in $\tilde{\boldsymbol{B}}$ a.s. such that $\mathsf{G}_{1,\ldots,n:n}(\cdot) = \mathsf{M}_{1,\ldots,n:n}(\cdot; \mathsf{B}, \tilde{\mathsf{Q}})$. Here $\mathsf{M}_{1,\ldots,n:n}(\cdot; \mathsf{B}, \tilde{\mathsf{Q}})$ is defined as in (3.32) with $\tilde{\mathsf{Q}}$ replacing

Q. Take a distribution $\tilde{\mathsf{Q}}$ satisfying this definition of sharpness. Then there exists a selection of $\tilde{\boldsymbol{B}}$ determined by the bidding rule associated with $\tilde{\mathsf{Q}}$, such that its distribution matches that of $\vec{\boldsymbol{b}}_n$. But then Theorem A.1 implies that the inequalities in (3.36) hold. Conversely, take $\tilde{\mathsf{Q}}$ satisfying the inequalities in (3.36). Then, by Theorem A.1, $\vec{\boldsymbol{b}}_n$ and $\tilde{\boldsymbol{B}}$ can be realized on the same probability space as random elements $\vec{\boldsymbol{b}}'_n$ and $\tilde{\boldsymbol{B}}'$, $\vec{\boldsymbol{b}}_n \overset{d}{=} \vec{\boldsymbol{b}}'_n$, $\tilde{\boldsymbol{B}} \overset{d}{=} \tilde{\boldsymbol{B}}'$, such that $\vec{\boldsymbol{b}}'_n \in \tilde{\boldsymbol{B}}'$ a.s. One can then complete the auction model with a bidding rule that picks $\vec{\boldsymbol{b}}'_n$ with probability 1, and the result follows. □

In (3.36), $\mathsf{P}(\vec{\boldsymbol{b}}_n \in K)$ is determined by the joint distribution of the ordered bids and hence can be learned from the data. On the other side, $\mathsf{T}_{\tilde{\boldsymbol{B}}}(K; \tilde{\mathsf{Q}})$ is a function of the model and $\tilde{\mathsf{Q}} \in \mathcal{T}$. Hence, it can be computed using (3.31), with $\tilde{\boldsymbol{B}}$ defined with respect to order statistics of i.i.d. random variables with distribution $\tilde{\mathsf{Q}} \in \mathsf{T}$. To gain insights in the characterization of $\mathcal{H}_{\mathsf{P}}[\mathsf{Q}]$, consider for example the set $K = \{\prod_{i=1}^{n-1}(-\infty, +\infty)\} \times (-\infty, v]$. Plugging it in the inequalities in (3.36), one obtains

$$\mathsf{G}_{n:n}(v) \leq \mathsf{Q}_{n-1,n}(v), \text{ for all } n,$$

which, using (3.33), yields (3.35). Similarly, plugging in the sets $K_j = \{\prod_{i=1}^{j-1}(-\infty, +\infty)\} \times [v, \infty) \times \{\prod_{j+1}^{n}(-\infty, +\infty)\}$, $j = 1, \ldots, n$, yields (3.34). So the inequalities proposed by Haile and Tamer (2003) are a subset of the inequalities yielding the sharp identification region in Theorem SIR-3.7. More information can be obtained by using additional sets K. For instance, the set $K = [v_1, \infty) \times [v_2, \infty) \times \{\prod_{i=1}^{n}(-\infty, +\infty)\}$, $v_2 \geq v_1$, yields $\mathsf{P}(\boldsymbol{b}_{1:n} \geq v_1, \boldsymbol{b}_{2:n} \geq v_2) \leq \mathsf{Q}_{1,2:n}([v_1, \infty) \times [v_2, \infty))$, which further restricts Q. Numerous examples can be given.

Characterization (3.36) is stated using inequality (A.4) for the collection of compact subsets of $\mathbb{R}^n$. One can instead use the (equivalent) inequality (A.5), and show that in fact it suffices to check it for a much smaller collection of sets, as shown by Chesher and Rosen (2017b) (see also Molchanov and Molinari, 2018, Section 2.2). Nonetheless, this collection remains uncountable.

KEY INSIGHT 3.10 (Random set theory and partial identification – continued): *As stated in the Introduction, constructing the (random) set of model predictions delivered by the maintained assumptions is an exercise typically carried out in identification analysis, regardless of whether random set theory is applied. Indeed, for the problem studied in this section, Haile and Tamer (2003, equation D1) put forward the set of admissible bids in (3.31).*[57] *With this set in hand, the tools of random set theory (in this case, Theorem A.1) immediately deliver the sharp identification region of interest.*

Chesher and Rosen (2017a) further generalize the analysis in this section by dropping the requirement of independent private values. This allows them, for example, to consider

[57] Equations D1 in Haile and Tamer and (3.31) here differ in that the latter also requires bids to be ordered. This observation was besides the point in Haile and Tamer's 2003 discussion that led to equation D1.

affiliated private values. They show that even in this significantly more complex context, the key behavioral restrictions imposed by Haile and Tamer (2003) to relate bids to valuations can be coupled with the use of random set theory, to characterize sharp identification regions.

3.4 Network Formation Models

Strategic models of network formation generalize the frameworks of single agents and multiple agents discrete choice models reviewed in Sections 3.1 and 3.2. They posit that pairs of agents (nodes) form, maintain, or sever connections (links) according to an explicit equilibrium notion and utility structure. Each individual's utility depends on the links formed by others (the network) and on utility shifters that may be pair-specific.

One may conjecture that the results reported in Sections 3.1-3.2 apply in this more general context too. While of course lessons can be carried over, network formation models present challenges that combined cannot be overcome without the development of new tools. These include the issue of equilibrium existence and the possibility of multiple equilibria when they exist, due to the interdependence in agents' choices (this problem was already discussed in Section 3.2). Another challenge is the degree of correlation between linking decisions, which interacts with how the observable data is generated: one may observe a growing number of independent networks, or a growing number of agents on a single network. Yet another challenge, which substantially increases the difficulties associated with the previous two, is the combinatoric complexity of network formation problems. The purpose of this section is exclusively to discuss some recent papers that have made important progress to address these specific challenges and carry out partial identification analysis. For a thorough treatment of the literature on network formation, I refer to the reviews in Graham (2015), Chandrasekhar (2016), de Paula (2017), and Graham (2019, Chapter XXX in this Volume).[58]

Depending on whether the researcher observes data from a single network or multiple independent networks, the underlying population of agents may be represented as a continuum or as a countably infinite set in the first case, or as a finite set in the second case. Henceforth, I denote generic agents as i, j, k, and m. I consider static models of undirected network formation with non-transferable utility.[59] The collection of all links among nodes forms the network, denoted $\boldsymbol{y}$. For any pair (i,j) with $i \neq j$, $\boldsymbol{y}_{ij} = 1$ if they are linked, and $\boldsymbol{y}_{ij} = 0$ otherwise ($\boldsymbol{y}_{ii} = 0$ for all i by convention). The notation $\boldsymbol{y} - \{ij\}$ denotes the network that results if a link present between nodes i and j is deleted, while $\boldsymbol{y} + \{ij\}$ denotes the network that results if a link absent between nodes i and j is added. Denote agent i's payoff by $\pi_i(\boldsymbol{y}, \boldsymbol{x}, \epsilon)$. This payoff depends on the network $\boldsymbol{y}$ and the payoff shifters $(\boldsymbol{x}, \epsilon)$, with $\boldsymbol{x}$ observable both to the agents and to the researcher, ϵ only to the agents, and $(\boldsymbol{x}, \epsilon)$ collecting

[58] For a review of the literature on peer group effect analysis, see, e.g., Brock and Durlauf (2001), Blume, Brock, Durlauf, and Ioannides (2011), de Paula (2017), and Graham (2019).

[59] These are models where if a link from node i to node j exists, then the link from j to i exists. The discussion that follows can be generalized to the case of models with transferable utility.

$(\boldsymbol{x}_{ij}, \epsilon_{ij})$ for all i and j.[60]

Following much of the literature, I employ *pairwise stability* (Jackson and Wolinsky, 1996) as equilibrium notion: $\boldsymbol{y}$ is a pairwise stable network if all linked agents prefer not to sever their links, and all non-existing links are damaging to at least one agent. Formally,

$$\forall (i,j) : \boldsymbol{y}_{ij} = 1,\ \pi_i(\boldsymbol{y}, \boldsymbol{x}, \epsilon) \geq \pi_i(\boldsymbol{y} - \{ij\}, \boldsymbol{x}, \epsilon) \text{ and } \pi_j(\boldsymbol{y}, \boldsymbol{x}, \epsilon) \geq \pi_j(\boldsymbol{y} - \{ij\}, \boldsymbol{x}, \epsilon),$$
$$\forall (i,j) : \boldsymbol{y}_{ij} = 0,\ \text{if } \pi_i(\boldsymbol{y} + \{ij\}, \boldsymbol{x}, \epsilon) > \pi_i(\boldsymbol{y}, \boldsymbol{x}, \epsilon) \text{ then } \pi_j(\boldsymbol{y} + \{ij\}, \boldsymbol{x}, \epsilon) < \pi_j(\boldsymbol{y}, \boldsymbol{x}, \epsilon).$$

Under this equilibrium notion if equilibria exist multiplicity is likely; see, among others, the examples in Graham (2015, p. 475), de Paula (2017, p. 301), and Sheng (2018, example 3.1). The model is therefore *incomplete*, because it does not specify how an equilibrium is selected in the region of multiplicity. For the same reasons as discussed in the context of finite games in Section 3.2, partial identification results (unless one is willing to impose restrictions on the equilibrium selection mechanism). However, as I explain below, an immediate application of the identification analysis carried out there presents enormous practical challenges because there are $2^{n(n-1)/2}$ possible network configurations to be checked for stability (and the dimensionality of the space of unobservables is also very large).

In what follows I consider two distinct frameworks that make different assumptions about the utility function and how the data is generated, and discuss what can be learned about the parameters of interest in these cases.

3.4.1 Data from Multiple Independent Networks

I first consider the case that the researcher observes data from multiple independent networks. I follow the set-up put forward by Sheng (2018).

Identification Problem 3.9 (Network Formation Model with Multiple Independent Networks): Let there be $n \in \mathbb{N}, n < \infty$ agents, and let $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathsf{P}$ be observable random variables in $\times_{j=1}^{n} \mathbb{R}^d \times \{0,1\}^{n(n-1)/2}$, $d < \infty$. Suppose that $\boldsymbol{y}$ is a pairwise stable network. For each agent i, let the utility function be known up to finite dimensional parameter vector $\delta \in \Delta \subset \mathbb{R}^p$, and given by

$$\pi_i(\boldsymbol{y}, \boldsymbol{x}, \epsilon; \delta) = \sum_{j=1}^{n} \boldsymbol{y}_{ij}(f(\boldsymbol{x}_i, \boldsymbol{x}_j; \delta_1) + \epsilon_{ij}) + \delta_2 \frac{\sum_{j=1}^{n}\sum_{k\neq i,k=1}^{n} \boldsymbol{y}_{ij}\boldsymbol{y}_{jk}}{n-2} + \delta_3 \frac{\sum_{j=1}^{n}\sum_{k=j+1}^{n} \boldsymbol{y}_{ij}\boldsymbol{y}_{ik}\boldsymbol{y}_{jk}}{n-2} \quad (3.37)$$

with $f(\cdot,\cdot;\cdot)$ a continuous function of its arguments.[61] Suppose that ϵ_{ij} are independent for

[60] Here I consider a framework where the agents have complete information.

[61] The effects of having friends in common and of friends of friends in (3.37) are normalized by $n-2$. This enforces that the marginal utility that i receives from linking with j is affected by j having an additional link

all $i \neq j$ and identically distributed with CDF known up to parameter vector $\gamma \in \Gamma \subset \mathbb{R}^m$, denoted F_γ. Assume that the support of F_γ is $\mathbb{R}$, that F_γ is absolutely continuous with respect to Lebesgue measure, and continuously differentiable with respect to $\gamma \in \Gamma$. Let $\Theta = \Delta \times \Gamma$. Assume that the researcher observes a random sample of networks and observable payoff shifters drawn from P. In the absence of additional information, what can the researcher learn about $\theta \equiv [\delta_1 \ \delta_2 \ \delta_3 \ \gamma]$?

Sheng (2018) analyzes this problem. She establishes equilibrium existence provided that $\delta_2 \geq 0$ and $\delta_3 \geq 0$ (Sheng, 2018, Proposition 2.2).[62] Given payoff shifters $(\boldsymbol{x}, \epsilon)$ and parameters $\vartheta \equiv [\tilde{\delta}_1 \ \tilde{\delta}_2 \ \tilde{\delta}_3 \ \tilde{\gamma}] \in \Theta$, let $\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \epsilon)$ denote the collection of pairwise stable networks implied by the model. It is easy to show that $\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \epsilon)$ is a random closed set as in Definition A.1. The networks in $\boldsymbol{Y}_\vartheta(\boldsymbol{x}, \epsilon)$ are $n \times n$ symmetric adjacency matrices with diagonal elements equal to zero and off diagonal elements in $\{0, 1\}$. To ease notation, I omit $\boldsymbol{Y}_\vartheta$'s dependence on $(\boldsymbol{x}, \epsilon)$ in what follows. Under the assumption that $\boldsymbol{y}$ is a pairwise stable network, at the true data generating value of $\theta \in \Theta$, one has

$$\boldsymbol{y} \in \boldsymbol{Y}_\theta \text{ a.s.} \tag{3.38}$$

Equation (3.38) exhausts the modeling content of Identification Problem 3.9. Theorem A.1 can be leveraged to extract its empirical content from the observed distribution $\mathsf{P}(\boldsymbol{y}, \boldsymbol{x})$. Let $\mathcal{Y}$ be the collection of $n \times n$ symmetric matrices with diagonal elements equal to zero and all other entries in $\{0, 1\}$, so that $|\mathcal{Y}| = 2^{n(n-1)/2}$. For a given set $K \subset \mathcal{Y}$, let $\mathsf{T}_{\boldsymbol{Y}_\vartheta}(K; \mathsf{F}_\gamma)$ denote the probability of the event $\{\boldsymbol{Y}_\vartheta \cap K \neq \emptyset\}$ implied when $\epsilon \sim \mathsf{F}_\gamma$, $\boldsymbol{x}$-a.s.

Theorem SIR-3.8 (Structural Parameters in Network Formation Models with Multiple Independent Networks): *Under the assumptions of Identification Problem 3.9, the sharp identification region for θ is*

$$\mathcal{H}_\mathsf{P}[\theta] = \{\vartheta \in \Theta : \mathsf{P}(\boldsymbol{y} \in K | \boldsymbol{x}) \leq \mathsf{T}_{\boldsymbol{Y}_\vartheta}(K; \mathsf{F}_{\tilde{\gamma}}) \, \forall K \subset \mathcal{Y}, \, \boldsymbol{x}\text{-a.s.}\}. \tag{3.39}$$

Proof. Follows from similar arguments as for the proof of Theorem 3.4 on p. 52. □

The characterization of $\mathcal{H}_\mathsf{P}[\theta]$ in Theorem SIR-3.8 is new to this chapter.[63] While technically it entails a finite number of conditional moment inequalities, in practice their number can be prohibitive as it can be as large as $2^{2^{n(n-1)/2}} - 2$.[64] Even using only a subset of the

with k to a smaller degree as n grows. This does not result in diminishing network effects.

[62] With transferable utility, Sheng (2018, Proposition 2.1) establishes existence for any $\delta_2, \delta_3 \in \mathbb{R}$. See Hellmann (2013) for an earlier analysis of existence and uniqueness of pairwise stable networks.

[63] Gualdani (2019) has previously used Theorem D.1 in Beresteanu, Molchanov, and Molinari (2011), as I do here, to characterize sharp identification regions in unilateral and bilateral directed network formation games.

[64] This number may be reduced drastically using the notion of *core determining class* of sets, see Definition A.8 and the discussion on p. 106. Nonetheless, even with relatively few agents, the number of inequalities in (3.39) may remain overwhelming.

inequalities in (3.39) to obtain an outer region, for example applying the insights in Ciliberto and Tamer (2009), may not be practical (with $n = 20$, $|\mathcal{Y}| \approx 10^{57}$). Moreover, computation of $\mathsf{T}_{\boldsymbol{Y}_\vartheta}(K; \mathsf{F}_\gamma)$ may require (depending on the set K) evaluation of rather complex integrals.

To circumvent these challenges, Sheng (2018) proposes to analyze network formation through *subnetworks*. A subnetwork is the restriction of a network to a subset of the agents (i.e., a subset of nodes and the links between them). For given $A \subset \{1, 2, \ldots, n\}$, let $\boldsymbol{y}^A = \{\boldsymbol{y}_{ij}\}_{i,j \in A, i \neq j}$ be the submatrix in $\boldsymbol{y}$ with rows and columns in A, and let $\boldsymbol{y}^{-A}$ be the remaining elements of $\boldsymbol{y}$ after $\boldsymbol{y}^A$ is deleted. With some abuse of notation, let $(\boldsymbol{y}^A, \boldsymbol{y}^{-A})$ denote the composition of $\boldsymbol{y}^A$ and $\boldsymbol{y}^{-A}$ that returns $\boldsymbol{y}$. Let

$$\boldsymbol{Y}_\vartheta^A = \{\boldsymbol{y}^A \in \{0,1\}^{|A|} : \exists \boldsymbol{y}^{-A} \in \{0,1\}^{|-A|} \text{ such that } (\boldsymbol{y}^A, \boldsymbol{y}^{-A}) \in \boldsymbol{Y}_\vartheta\}$$

denote the collection of subnetworks with rows and columns in A that can be part of a pairwise stable network in $\boldsymbol{Y}_\vartheta$. Let $\boldsymbol{x}_A$ denote the subset of $\boldsymbol{x}$ collecting $\boldsymbol{x}_{ij}$ for $i, j \in A$. For a given $y^A \in \{0,1\}^{|A|}$, let $\mathsf{C}_{\boldsymbol{Y}_\vartheta^A}(y^A; \mathsf{F}_\gamma)$ and $\mathsf{T}_{\boldsymbol{Y}_\vartheta^A}(y^A; \mathsf{F}_\gamma)$ denote, respectively, the probability of the events $\{\boldsymbol{Y}_\vartheta^A = \{y^A\}\}$ and $\{\{y^A\} \in \boldsymbol{Y}_\vartheta^A\}$ implied when $\epsilon \sim \mathsf{F}_\gamma$, $\boldsymbol{x}$-a.s. The first event means that only the subnetwork y^A is part of a pairwise stable network, while the second event means that y^A is a possible subnetwork that is part of a pairwise stable network but other subnetworks may be part of it too. Sheng (2018, Section 4.3) provides the following outer region for θ by adapting the insight in Ciliberto and Tamer (2009) to subnetworks.

THEOREM OR-3.1 (Subnetworks-based Outer Region on Structural Parameters in Network Formation Models with Multiple Independent Networks): *Under the assumptions of Identification Problem 3.9,*

$$\mathcal{O}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : \mathsf{C}_{\boldsymbol{Y}_\vartheta^A}(y^A; \mathsf{F}_{\tilde{\gamma}}) \leq \mathsf{P}(\boldsymbol{y}^A = y^A | \boldsymbol{x}_A) \leq \mathsf{T}_{\boldsymbol{Y}_\vartheta^A}(y^A; \mathsf{F}_{\tilde{\gamma}}) \, \forall y^A \subset \mathcal{Y}^A, \, \boldsymbol{x}_A\text{-}a.s.\}, \tag{3.40}$$

where $\mathcal{Y}^A$ is the collection of $|A| \times |A|$ symmetric matrices with diagonal elements equal to zero and all other elements in $\{0,1\}$ so that $|\mathcal{Y}^A| = 2^{|A|(|A|-1)/2}$.

Proof. Let $\boldsymbol{u}(\tilde{\boldsymbol{y}} | \boldsymbol{Y}_\vartheta)$ be a random variable in the unit simplex in $\mathbb{R}^{n(n-1)/2}$ which assigns to each possible pairwise stable network $\tilde{\boldsymbol{y}}$ that may realize given $(\boldsymbol{x}, \epsilon)$ and $\vartheta \in \Theta$ the probability that it is selected from $\boldsymbol{Y}_\vartheta$. Given $y \in \mathcal{Y}$, denote by $\mathsf{M}(y|\boldsymbol{x})$ the model predicted probability that the network realizes equal to y. Then the model yields

$$\mathsf{M}(y|\boldsymbol{x}) = \int \boldsymbol{u}(y|Y_\vartheta) d\mathsf{F}_\gamma = \int_{y \in Y_\vartheta, |Y_\vartheta| = 1} d\mathsf{F}_\gamma + \int_{y \in Y_\vartheta, |Y_\vartheta| \geq 2} \boldsymbol{u}(y|Y_\vartheta) d\mathsf{F}_\gamma. \tag{3.41}$$

The model implied distribution for subnetwork $\tilde{\boldsymbol{y}}^A$ is obtained by taking the marginal of

expression (3.41) with respect to $\tilde{\boldsymbol{y}}^{-A}$

$$\mathsf{M}(y^A|\boldsymbol{x}) = \sum_{y^{-A}} \mathsf{M}((y^A, y^{-A})|\boldsymbol{x}) = \int_{y^A \in Y_\vartheta^A, |Y_\vartheta^A|=1} d\mathsf{F}_\gamma + \int_{y^A \in Y_\vartheta^A, |Y_\vartheta^A|\geq 2} \sum_{y^{-A}} \boldsymbol{u}((y^A, y^{-A})|Y_\vartheta) d\mathsf{F}_\gamma. \tag{3.42}$$

Replacing $\boldsymbol{u}$ in (3.42) with zero and one yields the bounds in (3.40). Sheng (2018, Section 4.2) shows that under the maintained assumptions on ϵ, these inequalities are invariant under permutations of labels, so subnetworks in any two subsets $A, A' \subseteq \{1, 2, \ldots, n\}$ with $|A| = |A'|$ and $\boldsymbol{x}_A = \boldsymbol{x}_{A'}$ yield the same inequalities for all $y^A = y^{A'}$. It is therefore sufficient to consider subnetwork A and the inequalities in (3.40) associated with it. □

As long as the subnetworks are chosen to be small, e.g., $|A| = 2, 3, 4$, the inequalities in (3.40) can be computed even if the network is large. Moreover, Sheng (2018) shows that the inequalities in (3.40) remain informative even as n grows. This fact highlights the importance of working with subnetworks. One could have applied the insight of Ciliberto and Tamer (2009) directly to the full network by setting $\boldsymbol{u}$ equal to zero and to one in (3.41). The resulting bounds, however, would vanish to zero as n grows and become uninformative for θ. The characterization in Theorem OR-3.1 can be refined to obtain a smaller region, adapting the results in Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix Theorem D.1) to subnetworks. The size of this refined region is weakly decreasing in $|A|$.[65] However, the refinement does not yield $\mathcal{H}_\mathsf{P}[\theta]$ because it is applied only to subnetworks.

KEY INSIGHT 3.11: *At the beginning of this section I highlighted some key challenges to inference in network formation models. Identification Problem 3.9 bypasses the concern on the dependence among linking decisions through the independence assumption on ϵ_{ij} and the presumption that the researcher observes data from multiple independent networks. Sheng (2018) takes on the remaining challenges by formally establishing equilibrium existence and allowing for unrestricted selection among multiple equilibria. In order to overcome the computational complexity of the problem, she puts forward the important idea of inference based on subnetworks. While of course information is left on the table, the approach remains feasible even with large networks.*

Miyauchi (2016) considers a framework similar to the one laid out in Identification Problem 3.9. He assumes non-negative externalities, and shows that in this case the set of pairwise stable equilibria is a complete lattice with a smallest and a largest equilibrium.[66] He then uses moment functions that are monotone in the pairwise stable network (so that they take their extreme values at the smallest and largest equilibria), to obtain moment conditions that

[65] The idea of using random set methods on subnetworks to obtain the refined region was put forward in an earlier version of Sheng (2018). She provided a proof that the refined region's size decreases weakly in $|A|$.

[66] This approach exploits supermodularity, and is related to Jia (2008) and Echenique (2005).

restrict θ. Examples of the moment functions used include the proportion of pairs with a link, the proportion of links belonging to traingles, and many more (see Miyauchi, 2016, Table 1).

Gualdani (2019) considers unilateral and bilateral directed network formation games, still under a sampling framework where the researcher observes many independent networks. The equilibrium notion that she uses is pure strategy Nash. She assumes that the payoff that player i receives from forming link ij is allowed to depend on the number of additional players forming a link pointing to j (but rules out other spillover effects). Under this assumption and some regularity conditions, Gualdani shows that the network formation game can be decomposed into local games (i.e., games whose sets of players and strategy profiles are subsets of the network formation game's ones), so that the network formation game is in equilibrium if and only if each local game is in equilibrium. Thanks to this result, she obtains a computationally feasible characterization of $\mathcal{H}_{\mathsf{P}}[\theta]$ using elements of random set theory.

3.4.2 Data From a Single Network

When the researcher observes data from a single network, extra care has to be taken to restrict the dependence among linking decisions. This can be done in various ways (see, e.g., Chandrasekhar, 2016, for some examples). Here I consider a framework proposed by de Paula, Richards-Shubik, and Tamer (2018).

Identification Problem 3.10 (Network Formation Model with a Single Network): Let there be a continuum of agents $j \in \mathcal{I} = [0, \mu]$, with $\mu > 0$ their total measure, who choose whom to link to based on a utility function specified below.[67] Let $y : \mathcal{I} \times \mathcal{I} \to \{0, 1\}$ be an adjacency mapping with $y_{jk} = 1$ if nodes j and k are linked, and $y_{jk} = 0$ otherwise. Assume that only connections up to distance $\bar{d}$ affect utility and that preferences are such that agents never choose to form more than a total of $\bar{l}$ links.[68] To simplify exposition, let $\bar{d} = 2$. Let each agent j be endowed with characteristics $\boldsymbol{x}_j \in \mathcal{X}$, with $\mathcal{X}$ a finite set in $\mathbb{R}^p$, that are observable to the researcher. Additionally, let each agent j be endowed with $\bar{l} \times |\mathcal{X}|$ preference shocks $\epsilon_{j\ell}(x) \in \mathbb{R}, \ell = 1, \ldots, \bar{l}, x \in \mathcal{X}$, that are unobservable to the researcher and correspond to the possible direct connections and their characteristics.[69] Suppose that the vector of preference shocks is independent of $\boldsymbol{x}$ and has a distribution known up to parameter vector $\gamma \in \Gamma \subset \mathbb{R}^m$, denoted Q_γ. Let $\mathcal{I}(j) = \{k : y_{jk} = 1\}$. Assume that agents with characteristics and preference

[67] This is an approximation to a framework with a large but finite number of agents. The utility function can be less restrictive than the one considered here (see Assumptions 1 and 2 in de Paula, Richards-Shubik, and Tamer, 2018).

[68] The distance measure used here is the shortest path between two nodes.

[69] Under this assumption, the preference shocks do not depend on the individual identities of the agents. Hence, it agents k and m have the same observable characteristics, then j is indifferent between them.

shocks (x, e) value links according to the utility function

$$\pi_j(y, x, e) = \sum_{k \in \mathcal{I}(j)} (f(x_j, x_k) + e_{j\ell(k)}(x_k)) + \delta_1 \left| \bigcup_{k \in \mathcal{I}(j)} \mathcal{I}(k) - \mathcal{I}(j) - \{j\} \right| + \delta_2 \sum_{k \in \mathcal{I}(j)} \sum_{m \in \mathcal{I}(j): m > k} y_{km} - \infty \mathbf{1}(|\mathcal{I}(k)| > \bar{l}) \quad (3.43)$$

Assume that the network $\boldsymbol{y}$ formed by agents with characteristics and shocks $(\boldsymbol{x}, \epsilon)$ is pairwise stable. Let $\Theta \equiv \Upsilon \times \Delta \times \Gamma$, with Υ the parameter space for $\mathsf{f} \equiv \{f(x, w) : x \in \mathcal{X}, w \in \mathcal{X}\}$. In the absence of additional information, what can the researcher learn about $\theta \equiv [\mathsf{f}\ \delta_1\ \delta_2\ \gamma]$?

Identification Problem 3.10 enforces dimension reduction through the restrictions on depth and degree (the bounds $\bar{d}$ and $\bar{l}$), so that it is applicable to frameworks with networks that have limited degree distribution (e.g., close friendships network, but not Facebook network). It also requires that individual identities are irrelevant. This substantially reduces the richness of unobserved heterogeneity allowed for and the dimensionality of the space of unobservables. While the latter feature narrows the domain of applicability of the model, it is very beneficial to obtain a tractable characterization of what can be learned about θ, and yields equilibria that may include isolated nodes, a feature often encountered in networks data.

de Paula, Richards-Shubik, and Tamer (2018) study Identification Problem 3.10 focusing on the payoff-relevant local subnetworks that result from the maintained assumptions. These are distinct from the subnetworks used by Sheng (2018): whereas Sheng looks at subnetworks formed by arbitrary individuals and whose size is chosen by the researcher on the base of computational tractability, de Paula, Richards-Shubik, and Tamer look at subnetworks among individuals that are within a certain distance of each other, as determined by the structure of the preferences. On the other hand, Sheng's 2018 analysis does not require that agents have a finite number of types nor bounds the number of links that they may form.

To characterize the local subnetworks relevant for identification analysis in their framework, de Paula, Richards-Shubik, and Tamer (2018) propose the concepts of *network type* and *preference class*. A network type $t = (a, v)$ describes the local network up to distance $\bar{d}$ from the reference node. Here a is a square matrix of size $1 + \bar{l} \sum_{d=1}^{\bar{d}} (\bar{l} - 1)^{d-1}$ that describes the local subnetwork that is utility relevant for an agent of type t. It consists of the reference node, its direct potential neighbors ($\bar{l}$ elements), its second order neighbors ($\bar{l}(\bar{l}-1)$ elements), through its $\bar{d}$-th order neighbors ($\bar{l}(\bar{l}-1)^{\bar{d}-1}$ elements). The other component of the type, v, is a vector of length equal to the size of a that contains the observable characteristics of the reference node and her alters. The bounds $\bar{d}$ and $\bar{l}$ enforce dimension reduction by bounding the number of network types. The partial identification approach of de Paula, Richards-Shubik, and Tamer depends on this number, rather than on the number of agents. For example, the number of moment inequalities is determined by the number of network

types, not by the number of agents. As such, it yields its highest dividends for dimension reduction in large networks.

Let $\mathcal{T}$ denote the collection of network types generated from a preference structure π and set of characteristics $\mathcal{X}$. For given realization (x, e) of the observable characteristics and preference shocks of a reference agent, and for given $\vartheta \in \Theta$, define the collection of network types for which no agent wants to drop a link by

$$H_\vartheta(x, e) = \{(a, v) \in \mathcal{T} : v_1 = x \text{ and } \pi(a, v, e) \geq \pi(a_{-\ell}, v, e) \ \forall \ell = 1, \ldots, \bar{l}\},$$

where $a_{-\ell}$ is equal to the local adjacency matrix a but with the ℓ-th link removed (that is, it sets the $(1, \ell + 1)$ and $(\ell + 1, 1)$ elements of a equal to zero). In what follows I omit the dependence of H_ϑ on (x, e). Because $(\boldsymbol{x}, \epsilon)$ are random vectors, $\boldsymbol{H}_\vartheta = H_\vartheta(\boldsymbol{x}, \epsilon)$ is a random closed set as per Definition A.1. This random set takes on a finite number of realizations (equal to the possible subsets of $\mathcal{T}$), so that its distribution is completely determined by the probability with which it takes on each of these realizations. A preference class $H \subset \mathcal{T}$ is one of the possible realizations of $\boldsymbol{H}_\vartheta$ for some $\vartheta \in \Theta$. The model implied probability that $\boldsymbol{H}_\vartheta = H$ is given by

$$\mathsf{M}(H|\boldsymbol{x}; \vartheta) \equiv \mathsf{Q}_{\tilde{\gamma}}(\epsilon : \boldsymbol{H}_\vartheta = H|\boldsymbol{x}). \tag{3.44}$$

Observation of data from one network allows the researcher, under suitable restrictions on the sampling process, to learn the distribution of network types in the data (type shares), denoted $\mathsf{P}(t)$.[70] For example, in a network of best friends with $\bar{l} = 1$ and $\bar{d} = 2$, and $\mathcal{X} = \{x^1, x^2\}$ (e.g., a simplified framework with only two possible races), agents are either isolated or in a pair. Network types are pairs for the agents' race and the best friend's race (with second element equal zero if the agent is isolated). Type shares are the fraction of isolated blacks, the fraction of isolated whites, the fraction of blacks with a black best friend, the fraction of whites with a black best friend, and the fraction of whites with a white best friend. The preference classes for a black agent are $H^1(b, e) = \{(b, 0)\}$, $H^2(b, e) = \{(b, 0), (b, b)\}$, $H^3(b, e) = \{(b, 0), (b, w)\}$, $H^4(b, e) = \{(b, 0), (b, w), (b, b)\}$ (and similarly for whites). In each case, being alone is part of the preference class, as there are no links to sever. In the second class the agent has a preference for having a black friend, in the third class for a white friend, and in the last class for a friend of either race. It is easy to see that the model is *incomplete*, as for a given realization of ϵ it makes multiple predictions on the agent's preference type.

de Paula, Richards-Shubik, and Tamer propose to map the distribution of preference classes into the observed distribution of preference types in the data through the use of *allocation parameters*, denoted $\alpha_H(t) \in [0, 1]$. These are distinct from but play the same role as a selection mechanism. The model, augmented with them, implies a probability that an

[70] Full observation of the network is not required (and in practice it often does not occur). Sampling uncertainty results from it because in this model there is a continuum of agents.

agent with preferences in class H is of network type t given by:

$$\mathsf{M}(t;\vartheta,\alpha) = \frac{1}{\mu}\sum_{H\subset\mathcal{T}} \mu_{v_1(t)}\mathsf{M}(H|v_1(t);\vartheta)\alpha_H(t), \tag{3.45}$$

where $\mu_{v_1(t)}$ is the measure of reference agents with characteristics equal to the second component of the preference type t, $\boldsymbol{x} = v_1(t)$, and $\alpha \equiv \{\alpha_H(t) : t \in \mathcal{T}, H \subset \mathcal{T}\}$.

de Paula, Richards-Shubik, and Tamer provide a characterization of an outer region for θ based on two key implications of pairwise stability that deliver restrictions on α. They also show that under some additional assumptions, this characterization yields $\mathcal{H}_{\mathsf{P}}[\theta]$ (de Paula, Richards-Shubik, and Tamer, 2018, Appendix B). Here I focus on their more general result.

The first implication that they use is that existing links should not be dropped:

$$t \notin H \Rightarrow \alpha_H(t) = 0. \tag{3.46}$$

The condition in (3.46) is embodied in $\bar{\alpha} \equiv \{\alpha_H(t) : t \in H, H \subset \mathcal{T}\}$.

The second implication is that it should not be possible to establish mutually beneficial links among nodes that are far from each other. Let t' and s' denote the network types that are generated if one adds a link in networks of types t and s among two nodes that are at distance at least $2\bar{d}$ from each other and each have less than $\bar{l}$ links. Then the requirement is

$$\left(\sum_{H\subset\mathcal{T}} \mu_{v_1(t)}\mathsf{M}(H|v_1(t);\vartheta)\alpha_H(t)\mathbf{1}(t' \in H)\right)\left(\sum_{H\subset\mathcal{T}} \mu_{v_1(s)}\mathsf{M}(H|v_1(s);\vartheta)\alpha_H(s)\mathbf{1}(s' \in H)\right) = 0 \tag{3.47}$$

In words, if a positive measure of agents of type t prefer t' (i.e., $\alpha_H(t) > 0$ for some H such that $t' \in H$), there must be zero measure of type s individuals who prefer s', because otherwise the network is unstable. de Paula, Richards-Shubik, and Tamer show that the conditions in (3.47) can be embodied in a square matrix q of size equal to the length of $\bar{\alpha}$. The entries of q are constructed as follows. Let H and $\tilde{H}$ be two preference classes with $t \in H$ and $s \in \tilde{H}$. With some abuse of notation, let $q_{\alpha_H(t),\alpha_{\tilde{H}}(s)}$ denote the element of q corresponding to the index of the entry in $\bar{\alpha}$ equal to $\alpha_H(t)$ for the row, and to $\alpha_{\tilde{H}}(s)$ for the column. Then set $q_{\alpha_H(t),\alpha_{\tilde{H}}(s)}(\vartheta) = \mathbf{1}(t' \in H)\mathbf{1}(s' \in \tilde{H})$. It follows that this element yields the term $\big(\alpha_H(t)\mathbf{1}(t' \in H)\big)\big(\alpha_{\tilde{H}}(s)\mathbf{1}(s' \in \tilde{H})\big)$ in the quadratic form $\bar{\alpha}^\top q\bar{\alpha}$. As long as $\mu_{v_1(\cdot)}$ and $\mathsf{M}(\cdot|\boldsymbol{x};\vartheta)$ in (3.44) are strictly positive, this term is equal to zero if and only if condition (3.47) holds for types t and s.[71]

With this background, Theorem OR-3.2 below provides an outer region for θ. The proof of this result follows from the arguments laid out above (see de Paula, Richards-Shubik, and

[71] The possibility that $\mu_{v_1(\cdot)}$ or $\mathsf{M}(\cdot|\boldsymbol{x};\vartheta)$ are equal to zero can be accommodated by setting $q_{\alpha_H(t),\alpha_{\tilde{H}}(s)}(\vartheta) = (\mu_{v_1(t)}\mathsf{M}(H|v_1(t);\vartheta)\mathbf{1}(t' \in H))(\mu_{v_1(s)}\mathsf{M}(H|v_1(s);\vartheta)\mathbf{1}(s' \in \tilde{H}))$. However, in that case q depends on ϑ and its computational cost increases.

Tamer, 2018, Theorems 1 and 2, for the full details).

THEOREM OR-3.2 (Outer Region on Parameters of a Network Formation Model with a Single Network): *Under the assumptions of Identification Problem 3.10,*

$$\mathcal{O}_{\mathsf{P}}[\theta] = \left\{ \vartheta \in \Theta : \left(\begin{array}{ll} \min_{\bar{\alpha}} \bar{\alpha}^{\top} q \bar{\alpha} & \\ s.t. & \mathsf{M}(t;\vartheta,\bar{\alpha}) = \mathsf{P}(t) \ \forall \ t \in \mathcal{T} \\ & \sum_{t \in H} \bar{\alpha}_H(t) = 1 \ \forall H \subset \mathcal{T} \\ & \alpha_H(t) \geq 0 \ \forall t \in H, \forall H \subset \mathcal{T} \end{array} \right) = 0 \right\}. \tag{3.48}$$

The set in (3.48) does not equal $\mathcal{H}_{\mathsf{P}}[\theta]$ in all models allowed for in Identification Problem 3.10 because condition (3.47) does not embody all implications of pairwise stability on non-existing links. While the optimization problem in (3.48) is quadratic, it is not necessarily convex because q may not be positive definite. Nonetheless, the simulations reported by de Paula, Richards-Shubik, and Tamer suggest that $\mathcal{O}_{\mathsf{P}}[\theta]$ can be computed rapidly, as least for the examples they considered.

KEY INSIGHT 3.12: *At the beginning of this section I highlighted some key challenges to inference in network formation models. When data is observed from a single network, as in Identification Problem 3.10, de Paula, Richards-Shubik, and Tamer's 2018 proposal to base inference on local networks achieves two main benefits. First, it delivers consistently estimable features of the game, namely the probability that an agent belongs to one of a finite collection of network types. Second, it achieves dimension reduction, so that computation of outer regions on θ remains feasible even with large networks and allowing for unrestricted selection among multiple equilibria.*

3.5 Further Theoretical Advances and Empirical Applications

In order to discuss the partial identification approach to learning structural parameters of economic models in some level of detail while keeping this chapter to a manageable length, I have focused on a selection of papers. In this section I briefly mention several other excellent theoretical contributions that could be discussed more closely, as well as several empirical papers that have applied partial identification analysis of structural models to answer a wide array of questions of substantive economic importance.

Pakes (2010) and Pakes, Porter, Ho, and Ishii (2015) propose to embed revealed preference-based inequalities into structural models of both demand and supply in markets where firms face discrete choices of product configuration or of location. Whereas using revealed preference arguments is a trademark of the entire literature on discrete choice, Pakes (2010) and Pakes, Porter, Ho, and Ishii (2015) propose to use a subset of the model's implications to obtain easy-to-compute moment inequalities. For example, in the context of entry games such as the ones discussed in Section 3.2, they propose to base inference on the implication

that a player enters the market if and only if (s)he expects to make non-negative profits. This condition can be exploited even when players have heterogeneous (unobserved to the researcher) information sets, and it implies that the expected profits for entrants should be non-negative. Nonetheless, the condition does not suffice to obtain moment inequalities that include only observed payoff shifters and preference parameters. This is because the expected value of unobserved payoff shifters for entrants is not equal to zero, as the group of entrants is selected. The authors require the availability of valid (monotone) instrumental variables to solve this problem (see, e.g., Manski, 1990; Manski and Pepper, 2000, for uses of instrumental variables and monotone instrumental variables in the analysis of treatment effects). Interesting features of their approach include that the researcher does not need to solve for the set of equilibria, nor to require that the distribution of unobservable payoff shifters is known up to finite dimensional parameter vector. Moreover, the same basic ideas can be applied to single agent models (with or without heterogeneous information sets). A shortcoming of the method is that the set of parameter vectors satisfying the moment inequalities may be wider than the sharp identification region under the maintained assumptions.

The breadth of applications of the approach proposed by Pakes (2010) and Pakes, Porter, Ho, and Ishii (2015) is vast.[72] For example, Ho (2009) uses it to model the formation of the hospital networks offered by US health insurers, and Ho, Ho, and Mortimer (2012) and Lee (2013) use it to obtain bounds on firm fixed costs as an input to modeling product choices in the movie industry and in the US video game industry, respectively. Holmes (2011) estimates the effects of Wal-Mart's strategy of creating a high density network of stores. While the close proximity of stores implies cannibalization in sales, Wal-Mart is willing to bear it to achieve density economies, which in turn yield savings in distribution costs. His results suggest that Wal-Mart substantially benefits from high store density. Ellickson, Houghton, and Timmins (2013) measure the effects of chain economies, business stealing, and heterogeneous firms' comparative advantages in the discount retail industry. Kawai and Watanabe (2013) estimate a model of strategic voting and quantify the impact it has on election outcomes. As in other models analyzed in this section, the one they study yields multiple predicted outcomes, so that partial identification methods are required to carry out the empirical analysis if one does not assume a specific selection mechanism to resolve the multiplicity. They estimate their model on Japanese general-election data, and uncover a sizable fraction of strategic voters. They also estimate that only a small fraction of voters are misaligned (voting for a candidate other than their most preferred one). Eizenberg (2014) studies whether the rapid removal in the market for personal computers of existing central processing units upon creation of new ones through innovation reduces surplus. He finds that a limited group of price-insensitive

[72] Statistical inference in these papers is often carried out using the methods proposed by Chernozhukov, Hong, and Tamer (2007), Beresteanu and Molinari (2008), and Andrews and Soares (2010). Model specification tests, if carried out, are based on the method proposed by Bugni, Canay, and Shi (2015). See Sections 4.3 and 5, respectively, for a discussion of confidence sets and specification tests.

consumers enjoys the largest share of the welfare gains from innovation. A policy that kept older technologies on the shelf would allow for the benefits from innovation to reach price-sensitive consumers thanks to improved access to mobile computing, but total welfare would not increase because consumer welfare gains would be largely offset by producer losses. Ho and Pakes (2014) analyze hospital referrals for labor and birth episodes in California in 2003, for patients enrolled with six health insurers that use to a different extent incentives to referring physicians groups to reduce hospital costs (capitation contracts). The aim is to learn whether enrollees with high-capitation insurers tend to be referred to lower-priced hospitals (ceteris paribus) compared to other patients with same-severity conditions, and whether quality of care was affected. Their model allows for an insurer-specific preference function that is additively separable in the hospital price paid by the insurer (which is allowed to be measured with error), the distance traveled, and plan and severity-specific hospital fixed effects. Importantly, unobserved heterogeneity entering the preference function is not assumed to be drawn from a distribution known up to finite dimensional parameter vector. The results of the empirical analysis indicate that the price paid by insurers to hospitals has an impact on referrals, with higher elasticity to price for insurers whose physicians groups are more highly capitated. Dickstein and Morales (2018) study how the information that potential exporters have to predict the profits they will earn when serving a foreign market influences their decisions to export. They propose a model where the researcher specifies and observes a subset of the variables that agents use to form their expectations, but may not observe other variables that affect firms' expectations heterogeneously (across firms and markets, and over time). Because only a subset of the variables entering the firms' information set is observed, partial identification results. They show that, under rational expectations, they can test whether potential exporters know and use specific variables to predict their export profits. They also use their model's estimates to quantify the value of information. Wollmann (2018) studies the implications of the $85 billion automotive industry bailout in 2009 on the commercial vehicle segment. He finds that had Chrysler and GM been liquidated (or aquired by a major competitor) rather than bailed out, the surviving firms would have experienced a rise in profits high enough to induce them to introduce new products.

A different use of revealed preference arguments appears in the contributions of Blundell, Browning, and Crawford (2008), Blundell, Kristensen, and Matzkin (2014), Hoderlein and Stoye (2014), Manski (2014), Barseghyan, Molinari, and Teitelbaum (2016), Hausman and Newey (2016), and many others. For example, Manski (2014) proposes a method to partially identify income-leisure preferences and to evaluate the associated effects of tax policies. He starts from basic revealed-preference analysis performed under the assumption that individuals prefer more income and leisure, and no other restriction. The analysis shows that observing an individual's time allocation under a status quo tax policy yields bounds on his allocation that may or may not be informative, depending on how the person allocates his time under the status quo policy and on the tax schedules. He then explores what more can

be learned if one additionally imposes restrictions on the distribution of income-leisure preferences, using the method put forward by Manski (2007b). One assumption restricts groups of individuals facing different choice sets to have the same distribution of preferences. The other assumption restricts this distribution to a parametric family. Kline and Tartari (2016) build on and expand Manski (2014)'s framework to evaluate the effect of Connecticut's Jobs First welfare reform experiment on women' labor supply and welfare participation decisions.

Barseghyan, Molinari, and Teitelbaum (2016) propose a method to learn features of households' risk preferences in a random utility model that nests expected utility theory plus a range of non-expected utility models.[73] They allow for unobserved heterogeneity in preferences (that may enter the utility function non-separably) and leave completely unspecified their distribution. The authors use revealed preference arguments to infer, for each household, a set of values for its unobserved heterogeneity terms that are consistent with the household's choices in the three lines of insurance coverage. As their core restriction, they assume that each household's preferences are *stable* across contexts: the household's utility function is the same when facing distinct but closely related choice problems. This allows them to use the inferred set valued data to partially identify features of the distribution of preferences, and to classify households into preference types. They apply their proposed method to analyze data on households' deductible choices across three lines of insurance coverage (home all perils, auto collision, and auto comprehensive).[74] Their results show that between 70 and 80 percent of the households make choices that can be rationalized by a model with linear utility and monotone, quadratic, or even linear probability distortions. These probability distortions substantially overweight small probabilities. By contrast, fewer than 40 percent can be rationalized by a model with concave utility but no probability distortions.

Hausman and Newey (2016) propose a method to carry out demand analysis while allowing for general forms of unobserved heterogeneity. Preferences and linear budget sets are assumed to be statistically independent (conditional on covariates and control functions). Hausman and Newey show that for continuous demand, average surplus is generally not identified from the distribution of demand for a given price and income, and therefore propose a partial identification approach. They use bounds on income effects to derive bounds on average surplus. They apply the bounds to gasoline demand, using data from the 2001 U.S. National Household Transportation Survey.

Another strand of empirical applications pertains to the analysis of discrete games. Ciliberto and Tamer (2009) use the method they develop, described in Section 3.2.1, to study

[73] Their model is based on the one put forward by Barseghyan, Molinari, O'Donoghue, and Teitelbaum (2013). See Barseghyan, Molinari, O'Donoghue, and Teitelbaum (2018) for a review of these and other non-expected utility models in the context of estimation of risk preferences.

[74] Auto collision coverage pays for damage to the insured vehicle caused by a collision with another vehicle or object, without regard to fault. Auto comprehensive coverage pays for damage to the insured vehicle from all other causes, without regard to fault. Home all perils (or simply home) coverage pays for damage to the insured home from all causes, except those that are specifically excluded (e.g., flood, earthquake, or war).

market structure in the US airline industry and the role that firm heterogeneity plays in shaping it. Their findings suggest that the competitive effects of each carrier increase in that carrier's airport presence, but also that the competitive effects of large carriers (American, Delta, United) are different from those of low cost ones (Southwest). They also evaluate the effect of a counterfactual policy repealing the Wright Amendment, and find that doing so would see an increase in the number of markets served out of Dallas Love.

Grieco (2014) proposes a model of static entry that extends the one in Section 3.2 by allowing individuals to have flexible information structures, where players's payoffs depend on both a common-knowledge unobservable payoff shifter, and a private-information one. His characterization of $\mathcal{H}_{\mathsf{P}}[\theta]$ is based on using an unrestricted selection mechanism, as in Berry and Tamer (2006) and Ciliberto and Tamer (2009). He applies the model to study the impact of supercenters such as Wal-Mart, that sell both food and groceries, on the profitability of rural grocery stores. He finds that entry by a supercenter outside, but within 20 miles, of a local monopolist's market has a smaller impact on firm profits than entry by a local grocer. Their entrance has a small negative effect on the number of grocery stores in surrounding markets as well as on their profits. The results suggest that location and format-based differentiation partially insulate rural stores from competition with supercenters.

A larger class of information structures is considered in the analysis of static discrete games carried out by Magnolfi and Roncoroni (2017). They allow for all information structures consistent with the players knowing their own payoffs and the distribution of opponents' payoffs. As solution concept they adopt the Bayes Correlated Equilibrium recently developed by Bergemann and Morris (2016). Also with this solution concept multiple equilibria are possible. The authors leave completely unspecified the selection mechanism picking the equilibrium played in the regions of multiplicity, so that partial identification attains. Magnolfi and Roncoroni use the random sets approach to characterize $\mathcal{H}_{\mathsf{P}}[\theta]$. They apply the method to estimate a model of entry in the Italian supermarket industry and quantify the effect of large malls on local grocery stores. Berry and Compiani (2019) use the random sets approach to partially identify and estimate dynamic discrete choice models with serially correlated unobservables, under instrumental variables restrictions. They extend two-step dynamic estimation methods to characterize a set of structural parameters that are consistent with the dynamic model, the instrumental variables restrictions and the data. Gualdani (2019) uses the random sets approach and a network formation model, to learn about Italian firms' incentives for having their executive directors sitting on the board of their competitors.

Barseghyan, Coughlin, Molinari, and Teitelbaum (2019) use the method described in Section 3.1.3 to partially identify the distribution of risk preferences using data on deductible choices in auto collision insurance.[75] They posit an expected utility theory model and allow for unobserved heterogeneity in households' risk aversion and choice sets, with unre-

[75] Statistical inference on projections of the partially identified parameters is carried out using the method proposed by Kaido, Molinari, and Stoye (2019a).

stricted dependence between them. Motivation for why unobserved heterogeneity in choice sets might be an important factor in this empirical framework comes from the earlier analysis of Barseghyan, Molinari, and Teitelbaum (2016) and novel findings that are part of Barseghyan, Coughlin, Molinari, and Teitelbaum's 2019 contribution. They show that commonly used models that make strong assumptions about choice sets (e.g., the mixed logit model with each individual's choice set assumed equal to the feasible set, and various models of choice set formation) can be rejected in their data. With regard to risk aversion, their key finding is that their estimated lower bounds are an order of magnitude less than the point estimates obtained in the related literature. This suggests that the data can be explained by expected utility theory with lower and more homogeneous levels of risk aversion than it had been uncovered before. This provides new evidence on the importance of developing models that differ in their specification of *which* alternatives agents evaluate (rather than or in addition to models focusing on *how* they evaluate them), and to data collection efforts that seek to directly measure agents' heterogeneous choice sets (Caplin, 2016).

Iaryczower, Shi, and Shum (2018) study the effect of pre-vote deliberation on the decisions of US appellate courts. The question of interest is weather deliberation increases or reduces the probability of an incorrect decision. They use a model where communication equilibrium is the solution concept, and only observed heterogeneity in payoffs is allowed for. In the model, multiple equilibria are again possible, and the authors leave the selection mechanism completely unspecified. They characterize $\mathcal{H}_{\mathsf{P}}[\theta]$ through an optimization problem, and structurally estimate the model on US Courts of Appeal data. Iaryczower, Shi, and Shum compare the probability of making incorrect decisions under the pre-vote deliberation mechanism, to that in a counterfactual environment where no deliberation occurs. The results suggest that there is a range of parameters in $\mathcal{H}_{\mathsf{P}}[\theta]$, for which judges have ex-ante disagreement of imprecise prior information, for which deliberation is beneficial. Otherwise deliberation leads to lower effectiveness for the court.

D'Haultfoeuille, Gaillac, and Maurel (2018) show that inference methods developed for partially identified models can be useful even outside this context. They are concerned with testing the hypothesis of rational expectations when one observes only the marginal distributions of realizations and subjective beliefs, but not their joint distribution (e.g., when subjective beliefs are observed in one dataset, and realizations in a different one, and the two cannot be matched). They establish that the hypothesis of rational expectations can be expressed as testing that a continuum of moment inequalities is satisfied, and they leverage the results in Andrews and Shi (2017) to provide a simple-to-compute test for this hypothesis. They apply their method to test for and quantify deviations from rational expectations about future earnings, and examine the consequences of such departures in the context of a life-cycle model of consumption.

Another important strand of theoretical literature is concerned with partial identification of panel data models. Honoré and Tamer (2006) consider a dynamic random effects probit

model, and use partial identification analysis to obtain bounds on the model parameters that circumvent the initial conditions problem. Rosen (2012) considers a fixed effect panel data model where he imposes a conditional quantile restriction on time varying unobserved heterogeneity. Differencing out inequalities resulting from the conditional quantile restriction delivers inequalities that depend only on observable variables and parameters to be estimated, but not on the fixed effects, so that they can be used for estimation. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) obtain bounds on average and quantile treatment effects in nonparametric and semiparametric nonseparable panel data models. Torgovitsky (2019a) provides a method to partially identify state dependence in panel data models where individual unobserved heterogeneity needs not be time invariant.

4 Estimation and Inference

4.1 Framework and Scope of the Discussion

The identification analysis carried out in Sections 2-3 presumes knowledge of the joint distribution P of the observable variables. That is, it presumes that P can be learned with certainty from observation of the entire population. In practice, one observes a sample of size n drawn from P. For simplicity I assume it to be a random sample.[76]

Statistical inference on $\mathcal{H}_{\mathsf{P}}[\theta]$ needs to be conducted using knowledge of P_n, the empirical distribution of the observable outcomes and covariates. Because $\mathcal{H}_{\mathsf{P}}[\theta]$ is not a singleton, this task is particularly delicate. To start, care is required to choose a proper notion of consistency for a set estimator $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ and to obtain palatable conditions under which such consistency attains. Next, the asymptotic behavior of statistics designed to test hypothesis or build confidence sets for $\mathcal{H}_{\mathsf{P}}[\theta]$ or for $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ might change with ϑ, creating technical challenges for the construction of confidence sets that are not encountered when θ is point identified. Many of the sharp identification regions derived in Sections 2-3 can be written as collections of vectors $\vartheta \in \Theta$ that satisfy conditional or unconditional moment (in)equalities. For simplicity, I assume that Θ is a compact and convex subset of $\mathbb{R}^d$, and I use the formalization for the case of a finite number of unconditional moment (in)equalities:

$$\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : \mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i;\vartheta)) \leq 0 \ \forall j \in \mathcal{J}_1, \ \mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i;\vartheta)) = 0 \ \forall j \in \mathcal{J}_2\}. \tag{4.1}$$

In (4.1), $\boldsymbol{w}_i \in \mathcal{W} \subseteq \mathbb{R}^{d_{\mathcal{W}}}$ is a random vector collecting the observable variables, with $\boldsymbol{w} \sim \mathsf{P}$;

[76] This assumption is often maintained in the literature. See, e.g., Andrews and Soares (2010) for a treatment of inference with dependent observations. Epstein, Kaido, and Seo (2016) study inference in games of complete information as in Identification Problem 3.6, imposing the i.i.d. assumption on the unobserved payoff shifters $\{\varepsilon_{i1}, \varepsilon_{i2}\}_{i=1}^n$. The authors note that because the selection mechanism picking the equilibrium played in the regions of multiplicity (see Section 3.2) is left completely unspecified and may be arbitrarily correlated across markets, the resulting observed variables $\{\boldsymbol{w}_i\}_{i=1}^n$ may not be independent and identically distributed, and they propose an inference method to address this issue.

$m_j : \mathcal{W} \times \Theta \to \mathbb{R}$, $j \in \mathcal{J} \equiv \mathcal{J}_1 \cup \mathcal{J}_2$, are known measurable functions characterizing the model; and $\mathcal{J}$ is a finite set equal to $\{1, \ldots, |\mathcal{J}|\}$.[77] Instances where $\mathcal{H}_{\mathsf{P}}[\theta]$ is characterized through a finite number of conditional moment (in)equalities and the conditioning variables have finite support can easily be recast as in (4.1).[78] Consider, for example, the two player entry game model in Identification Problem 3.6 on p. 48, where $\boldsymbol{w} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{x}_1, \boldsymbol{x}_2)$. Using (in)equalities (3.18)-(3.21) and assuming that the distribution of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ has $\bar{k}$ points of support, denoted $(x_{1,k}, x_{2,k}), k = 1, \ldots, \bar{k}$, we have $|\mathcal{J}| = 4\bar{k}$ and for $k = 1, \ldots, \bar{k}$,[79]

$$
\begin{aligned}
m_{4k-3}(\boldsymbol{w}_i; \vartheta) &= [\mathbf{1}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,0)) - \Phi((-\infty, -\boldsymbol{x}_1 b_1), (-\infty, -\boldsymbol{x}_2 b_2); r)]\mathbf{1}((\boldsymbol{x}_1, \boldsymbol{x}_2) = (x_{1,k}, x_{2,k})) \\
m_{4k-2}(\boldsymbol{w}_i; \vartheta) &= [\mathbf{1}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (1,1)) - \Phi([-\boldsymbol{x}_1 b_1 - d_1, \infty), [-\boldsymbol{x}_2 b_2 - d_2, \infty); r)]\mathbf{1}((\boldsymbol{x}_1, \boldsymbol{x}_2) = (x_{1,k}, x_{2,k})) \\
m_{4k-1}(\boldsymbol{w}_i; \vartheta) &= [\mathbf{1}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,1)) - \Phi((-\infty, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, \infty); r)]\mathbf{1}((\boldsymbol{x}_1, \boldsymbol{x}_2) = (x_{1,k}, x_{2,k})) \\
m_{4k}(\boldsymbol{w}_i; \vartheta) &= \Big[\mathbf{1}((\boldsymbol{y}_1, \boldsymbol{y}_2) = (0,1)) - \Big\{\Phi((-\infty, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, \infty); r) \\
&\qquad - \Phi((-\boldsymbol{x}_1 b_1, -\boldsymbol{x}_1 b_1 - d_1), (-\boldsymbol{x}_2 b_2, -\boldsymbol{x}_2 b_2 - d_2); r)\Big\}\Big]\mathbf{1}((\boldsymbol{x}_1, \boldsymbol{x}_2) = (x_{1,k}, x_{2,k})).
\end{aligned}
$$

In point identified moment equality models it has been common to conduct estimation and inference using a criterion function that aggregates moment violations (Hansen, 1982). Manski and Tamer (2002) generalize this approach by proposing the use of a criterion function $q_{\mathsf{P}} : \Theta \to \mathbb{R}_+$ such that $q_{\mathsf{P}}(\vartheta) = 0$ if and only if $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$. Many criterion functions can be used (see, e.g. Manski and Tamer, 2002; Chernozhukov, Hong, and Tamer, 2007; Romano and Shaikh, 2008; Rosen, 2008; Galichon and Henry, 2009; Andrews and Guggenberger, 2009; Andrews and Soares, 2010; Canay, 2010; Romano and Shaikh, 2010). Some simple and commonly employed ones include

$$
q_{\mathsf{P},\text{sum}}(\vartheta) = \sum_{j \in \mathcal{J}_1} \left[\frac{\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i; \vartheta))}{\sigma_{\mathsf{P},j}(\vartheta)}\right]_+^2 + \sum_{j \in \mathcal{J}_2} \left[\frac{\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i; \vartheta))}{\sigma_{\mathsf{P},j}(\vartheta)}\right]^2, \tag{4.2}
$$

$$
q_{\mathsf{P},\max}(\vartheta) = \max\left\{\max_{j \in \mathcal{J}_1}\left[\frac{\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i; \vartheta))}{\sigma_{\mathsf{P},j}(\vartheta)}\right]_+, \max_{j \in \mathcal{J}_2}\left|\frac{\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i; \vartheta))}{\sigma_{\mathsf{P},j}(\vartheta)}\right|\right\}^2, \tag{4.3}
$$

where $[x]_+ = \max\{x, 0\}$ and $\sigma_{P,j}$ is the population standard deviation of $m_j(\boldsymbol{w}_i; \vartheta)$. In (4.2)-(4.3) the moment functions are studentized, as doing so is important for statistical power

[77] Examples where the set $\mathcal{J}$ is a compact set (e.g., a unit ball) rather than a finite set include the case of best linear prediction with interval outcome and covariate data, see characterization (2.27) on p. 24, and the case of entry games with multiple mixed strategy Nash equilibria, see characterization (3.25) on p. 55. A more general continuum of inequalities is also possible, as in the case of discrete choice with endogenous explanatory variables, see characterization (3.13) on p. 40. I refer to Andrews and Shi (2017) and Beresteanu, Molchanov, and Molinari (2011, Supplementary Appendix B) for inference methods in the presence of a continuum of conditional moment (in)equalities.

[78] I refer to Andrews and Shi (2013), Chernozhukov, Lee, and Rosen (2013), Lee, Song, and Whang (2013), Armstrong (2014, 2015), Armstrong and Chan (2016), Chernozhukov, Chetverikov, and Kato (2018), and Chetverikov (2018), for inference methods in the case that the conditioning variables have a continuous distribution.

[79] In these expressions an index of the form jk not separated by a comma equals the product of j with k.

(see, e.g., Andrews and Soares, 2010, p. 127). To simplify notation, I omit the label and simply use $q_{\mathsf{P}}(\vartheta)$. Given the criterion function, one can rewrite (4.1) as

$$\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : q_{\mathsf{P}}(\vartheta) = 0\}. \tag{4.4}$$

To keep this chapter to a manageable length, I focus my discussion of statistical inference *exclusively* on consistent estimation and on different notions of coverage that a confidence set may be required to satisfy and that have proven useful in the literature.[80] The topics of test of hypotheses and construction of confidence sets in partially identified models are covered in Canay and Shaikh (2017), who provide a comprehensive survey devoted entirely to them in the context of moment inequality models. Molchanov and Molinari (2018, Chapters 4 and 5) provide a thorough discussion of related methods based on the use of random set theory.

4.2 Consistent Estimation

When the identified object is a set, it is natural that its estimator is also a set. In order to discuss statistical properties of a set-valued estimator $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ (to be defined below), and in particular its consistency, one needs to specify how to measure the distance between $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ and $\mathcal{H}_{\mathsf{P}}[\theta]$. Several distance measures among sets exist (see, e.g., Molchanov, 2017, Appendix D). A natural generalization of the commonly used Euclidean distance is the *Hausdorff distance*, which for $A, B \subset \mathbb{R}^d$ is defined as

$$\mathbf{d}_H(A, B) = \max\left\{\sup_{a\in A} \mathbf{d}(a, B), \sup_{b\in B} \mathbf{d}(b, A)\right\},$$

with $\mathbf{d}(a, B) \equiv \inf_{b\in B} \|a - b\|$.[81] In words, the Hausdorff distance between two sets measures the furthest distance from an arbitrary point in one of the sets to its closest neighbor in the other set. It is easy to verify that $\mathbf{d}_H$ metrizes the family of non-empty compact sets; in particular, given non empty compact sets $A, B \subset \mathbb{R}^d$, $\mathbf{d}_H(A, B) = 0$ if and only if $A = B$. If either A or B is empty, $\mathbf{d}_H(A, B) = \infty$.

The use of the Hausdorff distance to conceptualize consistency of set valued estimators in econometrics was proposed by Hansen, Heaton, and Luttmer (1995, Section 2.4) and Manski and Tamer (2002, Section 3.2).[82]

[80] Using the well known duality between tests of hypotheses and confidence sets, the discussion could be re-framed in terms of size of the test.

[81] The definition of the Hausdorff distance can be generalized to an arbitrary metric space by replacing the Euclidean metric by the metric specified on that space.

[82] It was previously used in the mathematical literature on random set theory, for example to formalize laws of large numbers and central limit theorems for random sets such as the ones in Theorems A.3 and A.4 (Artstein and Vitale, 1975; Giné, Hahn, and Zinn, 1983).

DEFINITION 4.1 (Hausdorff Consistency): *An estimator* $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ *is consistent for* $\mathcal{H}_{\mathsf{P}}[\theta]$ *if*

$$\mathbf{d}_H(\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta], \mathcal{H}_{\mathsf{P}}[\theta]) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Molchanov (1998) establishes Hausdorff consistency of a plug-in estimator of the set $\{\vartheta \in \Theta : g_{\mathsf{P}}(\vartheta) \leq 0\}$, with $g_{\mathsf{P}} : \mathcal{W} \times \Theta \to \mathbb{R}$ a lower semicontinuous function of $\vartheta \in \Theta$ that can be consistently estimated by a lower semicontinuous function g_n. The set estimator is $\{\vartheta \in \Theta : g_n(\vartheta) \leq 0\}$. The fundamental assumption in Molchanov (1998) is that $\{\vartheta \in \Theta : g_{\mathsf{P}}(\vartheta) \leq 0\} \subseteq \text{cl}(\{\vartheta \in \Theta : g_{\mathsf{P}}(\vartheta) < 0\})$ (see Molchanov and Molinari, 2018, Section 5.2, for a discussion). There are important applications where this condition holds. Chernozhukov, Kocatulum, and Menzel (2015) provide results related to Molchanov (1998), as well as important extensions for the construction of confidence sets, and show that these can be applied to carry out statistical inference on the Hansen–Jagannathan sets of admissible stochastic discount factors (Hansen and Jagannathan, 1991), the Markowitz–Fama mean–variance sets for asset portfolio returns (Markowitz, 1952), and the set of structural elasticities in Chetty (2012)'s analysis of demand with optimization frictions. However, these methods are not broadly applicable in the general moment (in)equalities framework of this section, as Molchanov's key condition generally fails for the set $\mathcal{H}_{\mathsf{P}}[\theta]$ in (4.4).

4.2.1 Criterion Function Based Estimators

Manski and Tamer (2002) extend the standard theory of extremum estimation of point identified parameters to partial identification, and propose to estimate $\mathcal{H}_{\mathsf{P}}[\theta]$ using the collection of values $\vartheta \in \Theta$ that approximately minimize a sample analog of q_{P}:

$$\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta] = \left\{ \vartheta \in \Theta : q_n(\vartheta) \leq \inf_{\tilde{\vartheta} \in \Theta} q_n(\tilde{\vartheta}) + \tau_n \right\}, \tag{4.5}$$

with τ_n a sequence of non-negative random variables such that $\tau_n \xrightarrow{p} 0$. In (4.5), $q_n(\vartheta)$ is a sample analog of $q_{\mathsf{P}}(\vartheta)$ that replaces $\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i; \vartheta))$ and $\sigma_{\mathsf{P},j}(\vartheta)$ in (4.2)-(4.3) with properly chosen estimators, e.g.,

$$\bar{m}_{n,j}(\vartheta) \equiv \frac{1}{n} \sum_{i=1}^{n} m_j(\boldsymbol{w}_i, \vartheta), \ j = 1, \ldots, |\mathcal{J}|$$

$$\hat{\sigma}_{n,j}(\vartheta) \equiv \left(\frac{1}{n} \sum_{i=1}^{n} [m_j(\boldsymbol{w}_i, \vartheta)]^2 - [\bar{m}_{n,j}(\vartheta)]^2 \right)^{1/2}, \ j = 1, \ldots, |\mathcal{J}|.$$

It can be shown that as long as $\tau_n = o_p(1)$, under the same assumptions used to prove consistency of extremum estimators of point identified parameters (e.g., with uniform con-

vergence of q_n to q_{P} and continuity of q_{P} on Θ),

$$\sup_{\vartheta \in \hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]} \inf_{\tilde{\vartheta} \in \mathcal{H}_{\mathsf{P}}[\theta]} \|\vartheta - \tilde{\vartheta}\| \xrightarrow{p} 0 \text{ as } n \to \infty. \tag{4.6}$$

This yields that asymptotically each point in $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ is arbitrarily close to a point in $\mathcal{H}_{\mathsf{P}}[\theta]$. I refer to (4.6) as *inner consistency* henceforth.[83] But Hausdorff consistency requires also that

$$\sup_{\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]} \inf_{\tilde{\vartheta} \in \hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]} \|\vartheta - \tilde{\vartheta}\| \xrightarrow{p} 0 \text{ as } n \to \infty,$$

i.e., that each point in $\mathcal{H}_{\mathsf{P}}[\theta]$ is arbitrarily close to a point in $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$. To establish this result for the sharp identification regions in Theorem SIR-3.1 (semiparametric binary model with interval covariates) and Theorem SIR-3.2 (parametric regression with interval covariate), Manski and Tamer (2002, Propositions 3 and 5) require the rate at which $\tau_n \xrightarrow{p} 0$ to be slower than the rate at which q_n converges uniformly to q_{P} over Θ.

What might go wrong in the absence of such a restriction? A simple stylized example can help understand the issue. Consider a model with linear inequalities of the form

$$\begin{aligned}
\theta_1 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_1), \\
-\theta_1 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_2), \\
\theta_2 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_3) + \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_4)\theta_1, \\
-\theta_2 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_5) + \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_6)\theta_1.
\end{aligned}$$

Suppose $\boldsymbol{w} \equiv (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_6)$ is distributed multivariate normal, with $\mathbb{E}_{\mathsf{P}}(\boldsymbol{w}) = [6\ 0\ 2\ 0\ -2\ 0]^\top$ and $\mathrm{Cov}_{\mathsf{P}}(\boldsymbol{w})$ equal to the identity matrix. Then $\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta = [\vartheta_1\ \vartheta_2]^\top \in \Theta : \vartheta_1 \in [0, 6] \text{ and } \vartheta_2 = 2\}$. However, with positive probability in any finite sample $q_n(\vartheta) = 0$ for ϑ in a (random) triangle that only includes points that are close to a subset of the points in $\mathcal{H}_{\mathsf{P}}[\theta]$. Hence, with positive probability the minimizer of q_n cycles between consistent estimators of subsets of $\mathcal{H}_{\mathsf{P}}[\theta]$, but does not estimate the entire set. Enlarging the estimator to include all points that are close to minimizing q_n up to a tolerance that converges to zero sufficiently slowly removes this problem.

Chernozhukov, Hong, and Tamer (2007) significantly generalize the consistency results in Manski and Tamer (2002). They work with a normalized criterion function equal to $q_n(\tilde{\vartheta}) - \inf_{\vartheta \in \Theta} q_n(\vartheta)$, but to keep notation light I simply refer to it as q_n.[84] Under suitable regularity conditions, they establish consistency of an estimator that can be a smaller set

[83] See Redner (1981) for an early contribution establishing this type of inner consistency for maximum likelihood estimators when the true parameter is not point identified, and Blevins (2015, Theorem 1) for a pedagogically helpful proof for a semiparametric binary model.

[84] Using the normalized criterion function $q_n(\tilde{\vartheta}) - \inf_{\vartheta \in \Theta} q_n(\vartheta)$ is especially important in light of possible model misspecification, see Section 5.

than the one proposed by Manski and Tamer (2002), and derive its convergence rate. Some of the key conditions required by Chernozhukov, Hong, and Tamer (2007, Conditions C1 and C2) to study convergence rates include that q_n is lower semicontinuous in ϑ, satisfies various convergence properties among which $\sup_{\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]} q_n = O_p(1/a_n)$ for a sequence of normalizing constants $a_n \to \infty$, that $\tau_n \geq \sup_{\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]} a_n q_n$ with probability approaching one, and that $\tau_n / a_n \to 0$. They also require that there exist positive constants (δ, κ, γ) such that for any $\epsilon \in (0,1)$ there are $(\kappa_\epsilon, n_\epsilon)$ such that for all $n \geq n_\epsilon$, $q_n(\vartheta) \geq \kappa[\min\{\delta, \mathbf{d}(\vartheta, \mathcal{H}_{\mathsf{P}}[\theta])\}]^\gamma$ uniformly on $\{\vartheta \in \Theta : \mathbf{d}(\vartheta, \mathcal{H}_{\mathsf{P}}[\theta]) \geq (\kappa_\epsilon / a_n)^{1/\gamma}\}$ with probability at least $1-\epsilon$. In words, the assumption, referred to as *polynomial minorant* condition, rules out that q_n can be arbitrarily "flat" outside $\mathcal{H}_{\mathsf{P}}[\theta]$. It posits that q_n changes as at least a polynomial γ in the distance of ϑ from $\mathcal{H}_{\mathsf{P}}[\theta]$. Under some additional regularity conditions, Chernozhukov, Hong, and Tamer (2007) establish that

$$\mathbf{d}_H(\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta], \mathcal{H}_{\mathsf{P}}[\theta]) = O_p(\max\{1, \tau_n\}/a_n)^{1/\gamma}. \tag{4.7}$$

What is the role played by the polynomial minorant condition for the result in (4.7)? In moment (in)equalities models, Chernozhukov, Hong, and Tamer require $\gamma = 2$.[85] Consider a simple stylized example with (in)equalities of the form

$$\begin{aligned} -\theta_1 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_1), \\ -\theta_2 &\leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_2), \\ \theta_1 \theta_2 &= \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_3), \end{aligned}$$

with $\mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_1) = \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_2) = \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_3) = 0$, and note that the sample means $(\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{w}}_2, \bar{\boldsymbol{w}}_3)$ are $\sqrt{n}$-consistent estimators of $(\mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_1), \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_2), \mathbb{E}_{\mathsf{P}}(\boldsymbol{w}_3))$. Suppose $(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3)$ are distributed multivariate standard normal. Consider a sequence $\vartheta_n = [\vartheta_{1n} \ \vartheta_{2n}]^\top = [n^{-1/4} \ n^{-1/4}]^\top$. Then $[\mathbf{d}(\vartheta_n, \mathcal{H}_{\mathsf{P}}[\theta])]^\gamma = O_p(n^{-1/2})$. On the other hand, with positive probability $q_n(\vartheta_n) = (\bar{\boldsymbol{w}}_3 - \vartheta_{1n}\vartheta_{2n})^2 = O_p\left(n^{-1}\right)$, so that for n large enough $q_n(\vartheta_n) < [\mathbf{d}(\vartheta_n, \mathcal{H}_{\mathsf{P}}[\theta])]^\gamma$, violating the assumption. This occurs because the gradient of the moment equality vanishes as ϑ approaches zero, rendering the criterion function flat in a neighborhood of $\mathcal{H}_{\mathsf{P}}[\theta]$. As intuition would suggest, rates of convergence are slower the flatter q_n is outside $\mathcal{H}_{\mathsf{P}}[\theta]$.

Kaido, Molinari, and Stoye (2019b) show that in moment inequality models with smooth moment conditions, the polynomial minorant assumption with $\gamma = 2$ implies the Abadie constraint qualification (ACQ); see, e.g., Bazaraa, Sherali, and Shetty (2006, Chapter 5) for a definition and discussion of ACQ. The example just given to discuss the role of the polynomial minorant condition is in fact a known example where ACQ fails at $\vartheta = [0 \ 0]^\top$. The connection with ACQ is somewhat of a cautionary tale, in that such constraint qualifications

[85] Chernozhukov, Hong, and Tamer (2007, equation (4.1) and equation (4.6)) set $\gamma = 1$ because they report the assumption for a criterion function that does not square the moment violations.

are notoriously difficult to verify.

Chernozhukov, Hong, and Tamer (2007, Condition C.3, referred to as *degeneracy*) also consider the case that q_n vanishes on subsets of Θ that converge in Hausdorff distance to $\mathcal{H}_{\mathsf{P}}[\theta]$ at rate $a_n^{-1/\gamma}$. While it might be difficult to verify degeneracy in practice, Chernozhukov, Hong, and Tamer show that if it holds, τ_n can be set to a constant or zero.

Menzel (2014) studies estimation of $\mathcal{H}_{\mathsf{P}}[\theta]$ when the number of moment inequalities is large relative to sample size (possibly infinite). He provides a consistency result for criterion-based estimators that use a number of unconditional moment inequalities that grows with sample size. He also considers estimators based on conditional moment inequalities, and derives the fastest possible rate for estimating $\mathcal{H}_{\mathsf{P}}[\theta]$ under smoothness conditions on the conditional moment functions. He shows that the rates achieved by the procedures in Armstrong (2014, 2015) are (minimax) optimal, and cannot be improved upon.

KEY INSIGHT 4.1: *Manski and Tamer (2002) extend the notion of extremum estimation from point identified to partially identified models. They do so by putting forward a generalized criterion function whose zero-level set can be used to define $\mathcal{H}_{\mathsf{P}}[\theta]$ in partially identified structural semiparametric models. It is then natural to define the set valued estimator $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ as the collection of approximate minimizers of the sample analog of this criterion function. Manski and Tamer's analysis of statistical inference focuses exclusively on providing consistent estimators. Chernozhukov, Hong, and Tamer (2007) substantially generalize the analysis of consistency of criterion function-based set estimators. They provide a comprehensive study of convergence rates in partially identified models. Their work highlights the challenges a researcher faces in this context, and puts forward possible solutions in the form of assumptions under which specific rates of convergence attain.*

4.2.2 Support Function Based Estimators

Beresteanu and Molinari (2008) introduce to the econometrics literature inference methods for set valued estimators based on random set theory. They study the class of models where $\mathcal{H}_{\mathsf{P}}[\theta]$ is convex and can be written as the Aumann (or selection) expectation of a properly defined random closed set.[86] They propose to carry out estimation and inference leveraging the representation of convex sets through their *support function* (given in Definition A.5), as it is done in random set theory; see Molchanov (2017, Chapter 3) and Molchanov and Molinari (2018, Chapter 4). Because the support function fully characterizes the boundary of $\mathcal{H}_{\mathsf{P}}[\theta]$, it allows for a simple sample analog estimator, and for inference procedures with desirable properties.

[86] By Theorem A.2, the Aumann expectation of a random closed set defined on a nonatomic probability space is convex. In this chapter I am assuming nonatomicity of the probability space. Even if I did not make this assumption, however, when working with a random sample the relevant probability space is the product space with $n \to \infty$, hence nonatomic (Artstein and Vitale, 1975). If $\mathcal{H}_{\mathsf{P}}[\theta]$ is not convex, Beresteanu and Molinari's analysis can be applied to its convex hull.

An example of a framework where the approach of Beresteanu and Molinari can be applied is that of best linear prediction with interval outcome data in Identification Problem 2.4.[87] Recall that in that case, the researcher observes random variables $(\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}, \boldsymbol{x})$ and wishes to learn the best linear predictor of $\boldsymbol{y}|\boldsymbol{x}$, with $\boldsymbol{y}$ unobserved and $\mathbb{P}(\boldsymbol{y}_{\mathrm{L}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{U}}) = 1$. For simplicity let $\boldsymbol{x}$ be a scalar. Given a random sample $\{\boldsymbol{y}_{\mathrm{L}i}, \boldsymbol{y}_{\mathrm{U}i}, \boldsymbol{x}_i\}_{i=1}^n$ from P, the researcher can construct a random segment $\boldsymbol{G}_i$ for each i and a consistent estimator $\hat{\Sigma}_n$ of the random matrix Σ_{P} in (2.24) as

$$\boldsymbol{G}_i = \left\{ \begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{y}_i \boldsymbol{x}_i \end{pmatrix} : \ \boldsymbol{y}_i \in \mathrm{Sel}(\boldsymbol{Y}_i) \right\} \subset \mathbb{R}^2, \quad \text{and} \quad \hat{\Sigma}_n = \begin{pmatrix} 1 & \overline{\boldsymbol{x}} \\ \overline{\boldsymbol{x}} & \overline{\boldsymbol{x}^2} \end{pmatrix},$$

where $\boldsymbol{Y}_i = [\boldsymbol{y}_{\mathrm{L}i}, \boldsymbol{y}_{\mathrm{U}i}]$ and $\overline{\boldsymbol{x}}, \overline{\boldsymbol{x}^2}$ are the sample means of $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^2$ respectively. Because in this problem $\mathcal{H}_{\mathsf{P}}[\theta] = \Sigma_{\mathsf{P}}^{-1} \mathbb{E}_{\mathsf{P}} \boldsymbol{G}$ (see Theorem SIR-2.5 on p. 22), a natural sample analog estimator replaces Σ_{P} with $\hat{\Sigma}_n$, and $\mathbb{E}_{\mathsf{P}} \boldsymbol{G}$ with a Minkowski average of $\boldsymbol{G}_i$ (see Appendix A, p. 107 for a formal definition), yielding

$$\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta] = \hat{\Sigma}_n^{-1} \frac{1}{n} \sum_{i=1}^n \boldsymbol{G}_i. \tag{4.8}$$

The support function of $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ is the sample analog of that of $\mathcal{H}_{\mathsf{P}}[\theta]$ provided in (2.26):

$$h_{\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]}(u) = \frac{1}{n} \sum_{i=1}^n [(\boldsymbol{y}_{\mathrm{L}i} \mathbf{1}(f(\boldsymbol{x}_i, u) < 0) + \boldsymbol{y}_{\mathrm{U}i} \mathbf{1}(f(\boldsymbol{x}_i, u) \geq 0)) f(\boldsymbol{x}_i, u)], \quad u \in \mathbb{S},$$

where $f(\boldsymbol{x}_i, u) = [1 \ \boldsymbol{x}_i] \hat{\Sigma}_n^{-1} u$. Beresteanu and Molinari (2008) use the Law of Large Numbers for random sets reported in Theorem A.3 to show that $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ in (4.8) is $\sqrt{n}$-consistent under standard conditions on the moments of $(\boldsymbol{y}_{\mathrm{L}i}, \boldsymbol{y}_{\mathrm{U}i}, \boldsymbol{x}_i)$.

Bontemps, Magnac, and Maurin (2012) and Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2018) significantly expand the applicability of Beresteanu and Molinari's 2008 estimator. Bontemps, Magnac, and Maurin show that it can be used in a large class of partially identified linear models, including ones that allow for the availability of instrumental variables. Chandrasekhar, Chernozhukov, Molinari, and Schrimpf show that it can be used for best linear approximation of any function $f(x)$ that is known to lie within two identified bounding functions. The lower and upper functions defining the band are allowed to be any functions, including ones carrying an index, and can be estimated parametrically or nonparametrically. The method allows for estimation of the parameters of the best linear approximations to the set identified functions in many of the identification problems described in Section 2. It can also be used to estimate the sharp identification region for the parameters

[87] Kaido, Molinari, and Stoye (2019a, Supplementary Appendix F) establish that if $\boldsymbol{x}$ has finite support, $\mathcal{H}_{\mathsf{P}}[\theta]$ in Theorem SIR-2.5 can be written as the collection of $\vartheta \in \Theta$ that satisfy a finite number of moment inequalities, as posited in this section.

of a binary choice model with interval or discrete regressors under the assumptions of Magnac and Maurin (2008), characterized in (3.7) in Section 3.1.1.

Kaido and Santos (2014) develop a theory of efficiency for estimators of sets $\mathcal{H}_{\mathsf{P}}[\theta]$ as in (4.1) under the additional requirements that the inequalities $\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w},\theta))$ are convex in $\theta \in \Theta$ and smooth as functionals of the distribution of the data. Because of the convexity of the moment inequalities, $\mathcal{H}_{\mathsf{P}}[\theta]$ is convex and can be represented through its support function. Using the classic results in Bickel, Klaassen, Ritov, and Wellner (1993), Kaido and Santos show that under suitable regularity conditions, the support function admits for $\sqrt{n}$-consistent regular estimation. They also show that a simple plug-in estimator based on the support function attains the semiparametric efficiency bound, and the corresponding estimator of $\mathcal{H}_{\mathsf{P}}[\theta]$ minimizes a wide class of asymptotic loss functions based on the Hausdorff distance. As they establish, this efficiency result applies to the estimators proposed by Beresteanu and Molinari (2008), including that in (4.8), and by Bontemps, Magnac, and Maurin (2012).

Kaido (2016) further enlarges the applicability of the support function approach by establishing its duality with the criterion function approach, for the case that q_{P} is a convex function and q_n is a convex function almost surely. This allows one to use the support function approach also when a representation of $\mathcal{H}_{\mathsf{P}}[\theta]$ as the Aumann expectation of a random closed set is not readily available. Kaido considers $\mathcal{H}_{\mathsf{P}}[\theta]$ and its level set estimator $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ as defined, respectively, in (4.4) and (4.5), with Θ a convex subset of $\mathbb{R}^d$. Because q_{P} and q_n are convex functions, $\mathcal{H}_{\mathsf{P}}[\theta]$ and $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ are convex sets. Under the same assumptions as in Chernozhukov, Hong, and Tamer (2007), including the polynomial minorant and the degeneracy conditions, one can set $\tau_n = \tau$ and have $\mathbf{d}_H(\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta], \mathcal{H}_{\mathsf{P}}[\theta]) = O_p(a_n^{-1/\gamma})$. Moreover, due to its convexit, $\mathcal{H}_{\mathsf{P}}[\theta]$ is fully characterized by its support function $h_{\mathcal{H}_{\mathsf{P}}[\theta]}(u) = \max_{a_n q_n(\vartheta) \leq \tau} u^\top \vartheta$, which can be computed via convex programming.

Kitagawa and Giacomini (2018) consider consistent estimation of $\mathcal{H}_{\mathsf{P}}[\theta]$ in the context of Bayesian inference. They focus on partially identified models where $\mathcal{H}_{\mathsf{P}}[\theta]$ depends on a "reduced form" parameter ϕ (e.g., a vector of moments of random variables). They observe that while a prior on ϕ can be revised in light of the data, a prior on θ cannot, due to the lack of point identification. As such they propose to choose a single prior for the revisable parameters, and a set of priors for the unrevisable ones. The latter is the collection of priors such that the distribution of $\theta|\phi$ places probability one on $\mathcal{H}_{\mathsf{P}}[\theta]$. A crucial observation in Kitagawa and Giacomini is that once ϕ is viewed as a random vector, as in the Bayesian paradigm, under mild regularity conditions $\mathcal{H}_{\mathsf{P}}[\theta]$ is a random closed set, and Bayesian inference on it can be carried out using elements of random set theory. In particular, they show that the set of posterior means of $\theta|\boldsymbol{w}$ equals the Aumann expectation of $\mathcal{H}_{\mathsf{P}}[\theta]$ (with the underlying probability measure of $\phi|\boldsymbol{w}$). They also show that this Aumann expectation converges in Hausdorff distance to the "true" identified set if the latter is convex, or otherwise to its convex hull.

KEY INSIGHT 4.2: *Beresteanu and Molinari (2008) show that elements of random set theory can be employed to obtain inference methods for partially identified models that are easy to implement and have desirable statistical properties. Whereas they apply their findings to a specific class of models based on the Aumann expectation, the ensuing literature has demonstrated that random set methods are widely applicable to obtain estimators of sharp identification regions and establish their consistency.*

Chernozhukov, Lee, and Rosen (2013) propose an alternative to the notion of consistent estimator. Rather than asking that $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ satisfies the requirement in Definition 4.1, they propose the notion of *half-median-unbiased* estimator. This notion is easiest to explain in the case of interval identified scalar parameters. Take, e.g., the bound in Theorem SIR-2.1 for the conditional expectation of selectively observed data. Then an estimator of that interval is half-median-unbiased if the estimated upper bound exceeds the true upper bound, and the estimated lower bound falls below the true lower bound, each with probability at least 1/2 asymptotically. More generally, one can obtain a half-median-unbiased estimator as

$$\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta] = \left\{\vartheta \in \Theta : nq_n(\vartheta) \leq c_{1/2}(\vartheta)\right\}, \tag{4.9}$$

where $c_{1/2}(\vartheta)$ is a critical value chosen so that $\hat{\mathcal{H}}_{\mathsf{P}_n}[\theta]$ asymptotically contains $\mathcal{H}_{\mathsf{P}}[\theta]$ (or any fixed element in $\mathcal{H}_{\mathsf{P}}[\theta]$; see the discussion in Section 4.3.1 below) with at least probability 1/2. As discussed in the next section, $c_{1/2}(\vartheta)$ can be further chosen so that this probability is uniform over $\mathsf{P} \in \mathcal{P}$.

The requirement of half-median unbiasedness has the virtue that, by construction, an estimator such as (4.9) is a subset of a $1-\alpha$ confidence set as defined in (4.10) below for any $\alpha < 1/2$, provided $c_{1-\alpha}(\vartheta)$ is chosen using the same criterion for all $\alpha \in (0,1)$. In contrast, a consistent estimator satisfying the requirement in Definition 4.1 needs not be a subset of a confidence set. This is because the sequence τ_n in (4.5) may be larger than the critical value used to obtain the confidence set, see equation (4.10) below, unless regularity conditions such as degeneracy or others allow one to set τ_n equal to zero. Moreover, choice of the sequence τ_n is not data driven, and hence can be viewed as arbitrary. This raises a concern for the scope of consistent estimation in general settings.

However, reporting a set estimator together with a confidence set is arguably important to shed light on how much of the volume of the confidence set is due to statistical uncertainty and how much is due to a large identified set. One can do so by either using a half-median unbiased estimator as in (4.9), or the set of minimizers of the criterion function in (4.5) with $\tau_n = 0$ (which, as previously discussed, satisfies the inner consistency requirement in (4.6) under weak conditions, and is Hausdorff consistent in some well behaved cases).

4.3 Confidence Sets Satisfying Various Coverage Notions

4.3.1 Coverage of $\mathcal{H}_{\mathsf{P}}[\theta]$ vs. Coverage of θ

I first discuss confidence sets $C_n \subset \mathbb{R}^d$ defined as level sets of a criterion function:

$$C_n = \{\vartheta \in \Theta : nq_n(\vartheta) \leq c_{1-\alpha}(\vartheta)\}. \tag{4.10}$$

In (4.10), $c_{1-\alpha}(\vartheta)$ may be constant or vary in $\vartheta \in \Theta$. It is chosen to that C_n satisfies (asymptotically) a certain coverage property with respect to either $\mathcal{H}_{\mathsf{P}}[\theta]$ or each $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$. Correspondingly, different appearances of $c_{1-\alpha}(\vartheta)$ may refer to different critical values associated with different coverage notions. The challenging theoretical aspect of inference in partial identification is the determination of $c_{1-\alpha}$ and of methods to approximate it.

A first classification of coverage notions pertains to whether the confidence set should cover $\mathcal{H}_{\mathsf{P}}[\theta]$ or each of its elements with a prespecified asymptotic probability. Early on, within the study of interval-identified parameters, Horowitz and Manski (1998, 2000) put forward a confidence interval that expands each of the sample analogs of the extreme points of the population bounds by an amount designed so that the confidence interval asymptotically covers the population bounds with prespecified probability.

Chernozhukov, Hong, and Tamer (2007) study the general problem of inference for a set $\mathcal{H}_{\mathsf{P}}[\theta]$ defined as the zero-level set of a criterion function. The coverage notion that they propose is *pointwise coverage of the set*, whereby $c_{1-\alpha}$ is chosen so that:

$$\liminf_{n\to\infty} \mathsf{P}(\mathcal{H}_{\mathsf{P}}[\theta] \subseteq C_n) \geq 1-\alpha \text{ for all } \mathsf{P} \in \mathcal{P}. \tag{4.11}$$

Chernozhukov, Hong, and Tamer (2007) provide conditions under which C_n satisfies (4.11) with $c_{1-\alpha}$ constant in ϑ, yielding the so called *criterion function approach* to statistical inference in partial identification. Under the same coverage requirement, Bugni (2010) and Galichon and Henry (2013) introduce novel bootstrap methods for inference in moment inequality models. Henry, Méango, and Queyranne (2015) propose an inference method for finite games of complete information that exploits the structure of these models.

Beresteanu and Molinari (2008) propose a method to test hypotheses and build confidence sets satisfying (4.11) based on random set theory, the so called *support function approach*, which yields simple to compute confidence sets with asymptotic coverage equal to $1-\alpha$ when $\mathcal{H}_{\mathsf{P}}[\theta]$ is strictly convex. The reason for the strict convexity requirement is that in its absence, the support function of $\mathcal{H}_{\mathsf{P}}[\theta]$ is not fully differentiable, but only directionally differentiable, rendering the classical delta method inapplicable and complicating inference. Fang and Santos (2018) provide bootstrap methods that remain valid even with only directional differentiability. Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2018) propose a data jittering method that enforces full differentiability at the price of a small conservative

distortion. Kaido and Santos (2014) extend the applicability of the support function approach to other moment inequality models and establish important efficiency results. Chernozhukov, Kocatulum, and Menzel (2015) show that an Hausdorff distance-based test statistic can be weighted to enforce either exact or first-order equivariance to transformations of parameters. Adusumilli and Otsu (2017) provide empirical likelihood based inference methods for the support function approach. The test statistics employed in the criterion function approach and in the support function approach are asymptotically equivalent in specific moment inequality models (Beresteanu and Molinari, 2008; Kaido, 2016), but the criterion function approach is more broadly applicable.

The field's interest changed to a different notion of coverage when Imbens and Manski (2004) pointed out that often there is one "true" data generating θ, even if it is only partially identified. Hence, they proposed confidence sets that cover each $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ with a prespecified probability. For pointwise coverage, this leads to choosing $c_{1-\alpha}$ so that:

$$\liminf_{n\to\infty} \mathsf{P}(\vartheta \in C_n) \geq 1-\alpha \text{ for all } \mathsf{P} \in \mathcal{P} \text{ and } \vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]. \tag{4.12}$$

If $\mathcal{H}_{\mathsf{P}}[\theta]$ is a singleton then (4.11) and (4.12) both coincide with the pointwise coverage requirement employed for point identified parameters. However, as shown in Imbens and Manski (2004, Lemma 1), if $\mathcal{H}_{\mathsf{P}}[\theta]$ contains more than one element, the two notions differ, with confidence sets satisfying (4.12) being weakly smaller than ones satisfying (4.11). Rosen (2008) provides confidence sets for general moment (in)equalities models that satisfy (4.12) and are easy to compute.

Although confidence sets that take each $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ as the object of interest (and which satisfy the *uniform coverage* requirements described in Section 4.3.2 below) have received the most attention in the literature on inference in partially identified models, this choice merits some words of caution. First, Henry and Onatski (2012) point out that if confidence sets are to be used for decision making, a policymaker concerned with robust decisions might prefer ones satisfying (4.11) (respectively, (4.13) below once uniformity is taken into account) to ones satisfying (4.12) (respectively, (4.14) below with uniformity). Second, while in many applications a "true" data generating θ exists, in others it does not. For example, Manski and Molinari (2010) and Giustinelli, Manski, and Molinari (2019a) query survey respondents (in the American Life Panel and in the Health and Retirement Study, respectively) about their subjective beliefs on the probability chance of future events. A large fraction of these respondents, when given the possibility to do so, report imprecise beliefs in the form of intervals. In this case, there is no "true" point-valued belief: the "truth" is interval-valued. If one is interested in (say) average beliefs, the sharp identification region is the (Aumann) expectation of the reported intervals, and the appropriate coverage requirement for a confidence set is that in (4.11) (respectively, (4.13) below with uniformity).

4.3.2 Pointwise vs. Uniform Coverage

In the context of interval identified parameters, such as, e.g., the mean with missing data in Theorem SIR-2.1 with $\theta \in \mathbb{R}$, Imbens and Manski (2004) pointed out that extra care should be taken in the construction of confidence sets for partially identified parameters, as otherwise they may be asymptotically valid only pointwise (in the distribution of the observed data) over relevant classes of distributions.[88] For example, consider a confidence interval that expands each of the sample analogs of the extreme points of the population bounds by a one-sided critical value. This confidence interval controls the asymptotic coverage probability pointwise for any DGP at which the width of the population bounds is positive. This is because the sampling variation becomes asymptotically negligible relative to the (fixed) width of the bounds, making the inference problem essentially one-sided. However, for every n one can find a distribution $\mathsf{P} \in \mathcal{P}$ and a parameter $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$ such that the width of the population bounds (under P) is small relative to n and the coverage probability for ϑ is below $1-\alpha$. This happens because the proposed confidence interval does not take into account the fact that for some $\mathsf{P} \in \mathcal{P}$ the problem has a two-sided nature.

This observation naturally leads to a more stringent requirement of *uniform coverage*, whereby (4.11)-(4.12) are replaced, respectively, by

$$\liminf_{n\to\infty} \inf_{\mathsf{P}\in\mathcal{P}} \mathsf{P}(\mathcal{H}_{\mathsf{P}}[\theta] \subseteq C_n) \geq 1-\alpha, \tag{4.13}$$

$$\liminf_{n\to\infty} \inf_{\mathsf{P}\in\mathcal{P}} \inf_{\vartheta\in\mathcal{H}_{\mathsf{P}}[\theta]} \mathsf{P}(\vartheta \in C_n) \geq 1-\alpha, \tag{4.14}$$

and $c_{1-\alpha}$ is chosen accordingly, to obtain either (4.13) or (4.14). Sets satisfying (4.13) are referred to as confidence regions for $\mathcal{H}_{\mathsf{P}}[\theta]$ that are uniformly consistent in level (over $\mathsf{P} \in \mathcal{P}$). Romano and Shaikh (2010) propose such confidence regions, study their properties, and provide a step-down procedure to obtain them.

Sets satisfying (4.14) are referred to as confidence regions for points in $\mathcal{H}_{\mathsf{P}}[\theta]$ that are uniformly consistent in level (over $\mathsf{P} \in \mathcal{P}$). Within the framework of Imbens and Manski (2004), Stoye (2009) shows that one can obtain a confidence interval satisfying (4.14) by pre-testing whether the lower and upper population bounds are sufficiently close to each other. If so, the confidence interval expands each of the sample analogs of the extreme points of the population bounds by a two-sided critical value; otherwise, by a one-sided. Stoye provides important insights clarifying the connection between superefficient (i.e., faster than $O_p(1/\sqrt{n})$) estimation of the width of the population bounds when it equals zero, and certain challenges in Imbens and Manski's proposed method.[89] Bontemps, Magnac, and Maurin (2012) leverage Stoye (2009)'s results to obtain confidence sets satisfying (4.14) using

[88] This discussion draws on many conversations with Jörg Stoye, as well as notes that he shared with me, for which I thank him.

[89] Indeed, the confidence interval proposed by Stoye (2009) can be thought of as using a Hodges-type shrinkage estimator (see, e.g., van der Vaart, 1997) for the width of the population bounds.

the support function approach for set identified linear models.

Obtaining confidence sets that satisfy the requirement in (4.14) becomes substantially more complex in the context of general moment (in)equalities models. One of the key challenges to uniform inference stems from the fact that the behavior of the limit distribution of the test statistic depends on $\sqrt{n}\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w}_i;\vartheta))$, $j=1,\ldots,|\mathcal{J}|$, which cannot be consistently estimated. Romano and Shaikh (2008); Andrews and Guggenberger (2009); Andrews and Soares (2010); Canay (2010); Andrews and Barwick (2012); Romano, Shaikh, and Wolf (2014), among others, make significant contributions to circumvent these difficulties in the context of a finite number of unconditional moment (in)equalities. Andrews and Shi (2013); Chernozhukov, Lee, and Rosen (2013); Lee, Song, and Whang (2013); Armstrong (2014, 2015); Armstrong and Chan (2016); Chetverikov (2018), among others, make significant contributions to circumvent these difficulties in the context of a finite number of conditional moment (in)equalities (with continuously distributed conditioning variables). Chernozhukov, Chetverikov, and Kato (2018) and Andrews and Shi (2017) study, respectively, the challenging frameworks where the number of moment inequalities grows with sample size and where there is a continuum of conditional moment inequalities.

I refer to Canay and Shaikh (2017, Section 4) for a thorough discussion of these methods and a comparison of their relative (de)merits (see also Bugni, Canay, and Guggenberger, 2012; Bugni, 2016).

4.3.3 Coverage of the Vector θ vs. Coverage of a Component of θ

The coverage requirements in (4.13)-(4.14) refer to confidence sets in $\mathbb{R}^d$ for the entire θ or $\mathcal{H}_{\mathsf{P}}[\theta]$. Often empirical researchers are interested in inference on a specific component or (smooth) function of θ (e.g., the returns to education; the effect of market size on the probability of entry; the elasticity of demand for insurance to price, etc.). For simplicity, here I focus on the case of a component of θ, which I represent as $u^\top\theta$, with u a standard basis vector in $\mathbb{R}^d$. In this case, the (sharp) identification region of interest is

$$\mathcal{H}_{\mathsf{P}}[u^\top\theta] = \{s \in [-h_\Theta(-u), h_\Theta(u)] : s = u^\top\theta \text{ and } \theta \in \mathcal{H}_{\mathsf{P}}[\theta]\}.$$

One could report as confidence interval for $u^\top\theta$ the projection of C_n in direction $\pm u$. The resulting confidence interval is asymptotically valid but typically conservative. The extent of the conservatism increases with the dimension of θ and is easily appreciated in the case of a point identified parameter. Consider, for example, a linear regression in $\mathbb{R}^{10}$, and suppose for simplicity that the limiting covariance matrix of the estimator is the identity matrix. Then a 95% confidence interval for $u^\top\theta$ is obtained by adding and subtracting 1.96 to that component's estimate. In contrast, projection of a 95% confidence ellipsoid for θ on each component amounts to adding and subtracting 4.28 to that component's estimate.

It is therefore desirable to provide confidence intervals CI_n specifically designed to cover

$u^\top\theta$ rather then the entire θ. Natural counterparts to (4.13)-(4.14) are

$$\liminf_{n\to\infty} \inf_{\mathsf{P}\in\mathcal{P}} \mathsf{P}(\mathcal{H}_\mathsf{P}[u^\top\theta] \subseteq CI_n) \geq 1-\alpha, \tag{4.15}$$

$$\liminf_{n\to\infty} \inf_{\mathsf{P}\in\mathcal{P}} \inf_{\vartheta\in\mathcal{H}_\mathsf{P}[\theta]} \mathsf{P}(u^\top\vartheta \in CI_n) \geq 1-\alpha. \tag{4.16}$$

As shown in Beresteanu and Molinari (2008) and Kaido (2016) for the case of pointwise coverage, obtaining asymptotically valid confidence intervals is simple if the identified set is convex and one uses the support function approach. This is because it suffices to base the test statistic on the support function in direction u, and it is often possible to easily characterize the limiting distribution of this test statistic. See Molchanov and Molinari (2018, Chapters 4 and 5) for details.

The task is significantly more complex in general moment inequality models when $\mathcal{H}_\mathsf{P}[\theta]$ is non-convex and one wants to satisfy the criterion in (4.15) or that in (4.16). Romano and Shaikh (2008) and Bugni, Canay, and Shi (2017) propose confidence intervals of the form

$$CI_n = \left\{ s \in [-h_\Theta(-u), h_\Theta(u)] : \inf_{\vartheta\in\Theta(s)} nq_n(\vartheta) \leq c_{1-\alpha}(s) \right\}, \tag{4.17}$$

where $\Theta(s) = \{\vartheta \in \Theta : u^\top\vartheta = s\}$ and $c_{1-\alpha}$ is such that (4.16) holds. An important idea in this proposal is that of *profiling* the test statistic $nq_n(\vartheta)$ by minimizing it over all ϑs such that $u^\top\vartheta = s$. One then includes in the confidence interval all values s for which the profiled test statistic's value is not too large. Romano and Shaikh (2008) propose the use of subsampling to obtain the critical value $c_{1-\alpha}(s)$ and provide high-level conditions ensuring that (4.16) holds. Bugni, Canay, and Shi (2017) substantially extend and improve the *profiling approach* by providing a bootstrap-based method to obtain $c_{1-\alpha}$ so that (4.16) holds which is more powerful than subsampling (for reasonable choices of subsample size).

Kaido, Molinari, and Stoye (2019a) propose a bootstrap-based *calibrated projection approach* where

$$CI_n = [-h_{\boldsymbol{C}_n(c_{1-\alpha})}(-u), h_{\boldsymbol{C}_n(c_{1-\alpha})}(u)], \tag{4.18}$$

with

$$h_{\boldsymbol{C}_n(c_{1-\alpha})}(u) \equiv \sup_{\vartheta\in\Theta} u^\top\vartheta \text{ s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leq c_{1-\alpha}(\vartheta),\ j = 1,\ldots,|\mathcal{J}| \tag{4.19}$$

and $c_{1-\alpha}$ a critical level function calibrated so that (4.16) holds. Compared to the simple projection of C_n mentioned at the beginning of the discussion of inference for components of θ, calibrated projection (weakly) reduces the value of $c_{1-\alpha}$ so that the projection of θ, rather than θ itself, is asymptotically covered with the desired probability uniformly.

4.3.4 A Brief Note on Bayesian Methods

The confidence sets discussed in this section are based on the frequentist approach to inference. It is natural to ask whether in partially identified models, as in well behaved point identified models, one can build Bayesian credible sets that at least asymptotically coincide with frequentist confidence sets. This question was first addressed by Moon and Schorfheide (2012), with a negative answer for the case that the coverage in (4.14) is sought out. In particular, they showed that the resulting Bayesian credible sets are a subset of $\mathcal{H}_{\mathsf{P}}[\theta]$, and hence too narrow from the frequentist perspective. This discrepancy can be ameliorated when inference is sought out for $\mathcal{H}_{\mathsf{P}}[\theta]$ rather than for each $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$. Kline and Tamer (2016) and Kitagawa and Giacomini (2018) propose Bayesian credible regions that are valid for frequentist inference in the sense of (4.11). Chen, Christensen, and Tamer (2018) propose both Bayesian credible sets that satisfy the (frequentist) criterion in (4.13), and ones that satisfy the (frequentist) criterion in (4.15). Their method is based on level sets of criterion functions constructed using cutoffs that are computed via Monte Carlo simulations from the quasi-posterior distribution of the criterion, and is shown to have asymptotically valid frequentist coverage.

5 Misspecification in Partially Identified Models

Although partial identification often results from reducing the number of assumptions maintained in counterpart point identified models, care still needs to be taken in assessing the possible consequences of misspecification. This section's goal is to discuss the existing literature on the topic, and to provide some additional observations. To keep the notation light, I refer to the functional of interest as θ throughout, without explicitly distinguishing whether it belongs to an infinite dimensional parameter space (as in the nonparametric analysis in Section 2), or to a finite dimensional one (as in the semiparametric analysis in Section 3).

The original nonparametric "worst-case" bounds proposed by Manski (1989) for the analysis of selectively observed data and discussed in Section 2 are not subject to the risk of misspecification, because they are based on the empirical evidence alone. However, often researchers are willing and eager to maintain additional assumptions that can help shrink the bounds, so that one can learn more from the available data. Indeed, early on Manski (1990) proposed the use of exclusion restrictions in the form of mean independence assumptions. Section 2.2 discusses related ideas within the context of nonparametric bounds on treatment effects, and Manski (2003, Chapter 2) provides a thorough treatment of other types of exclusion restriction. The literature reviewed throughout this chapter provides many more examples of assumptions that have proven useful for empirical research.

Broadly speaking, assumptions can be classified in two types (Manski, 2003, Chapter 2). The first type is *non-refutable*: it may reduce the size of $\mathcal{H}_{\mathsf{P}}[\theta]$, but cannot lead to it being

empty. An example in the context of selectively observed data is that of exogenous selection, or data missing at random conditional on covariates and instruments (see Section 2.1 on p. 9): under this assumption $\mathcal{H}_{\mathsf{P}}[\theta]$ is a singleton, but the assumption cannot be refuted because it poses a distributional (independence) assumption on unobservables.

The second type is *refutable*: it may reduce the size of $\mathcal{H}_{\mathsf{P}}[\theta]$, and it may result in $\mathcal{H}_{\mathsf{P}}[\theta] = \emptyset$ if it does not hold in the DGP. An example in the context of treatment effects is the assumption of mean independence between response function at treatment t and instrumental variable $\boldsymbol{z}$, see (2.14) in Section 2.2. There the sharp bounds on $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}(t)|\boldsymbol{x} = x)$ are intersection bounds as in (2.15). If the instrument is invalid, the bounds can be empty.

Ponomareva and Tamer (2011) consider the impact of misspecification on semiparametric partially identified models. One of their examples concerns a linear regression model of the form $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x}) = \theta^{\top}\boldsymbol{x}$ when only interval data is available for $\boldsymbol{y}$ (as in Section 2.3). In this context, $\mathcal{H}_{\mathsf{P}}[\theta] = \{\vartheta \in \Theta : \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x}) \leq \vartheta^{\top}\boldsymbol{x} \leq \mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x}),\ \boldsymbol{x}\text{-a.s.}\}$. The concern is that the conditional expectation might not be linear. Ponomareva and Tamer make two important observations. First, they argue that the set $\mathcal{H}_{\mathsf{P}}[\theta]$ is of difficult interpretation when the model is misspecified. When $\boldsymbol{y}$ is perfectly observed, if the conditional expectation is not linear, the output of ordinary least squares can be readily interpreted as the best linear approximation to $\mathbb{E}_{\mathsf{Q}}(\boldsymbol{y}|\boldsymbol{x})$. This is not the case for $\mathcal{H}_{\mathsf{P}}[\theta]$ when only the interval data $[\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ is observed. They therefore propose to work with the set of best linear predictors for $\boldsymbol{y}|\boldsymbol{x}$ even in the partially identified case (rather than fully exploit the linearity assumption). The resulting set is the one derived by Beresteanu and Molinari (2008) and reported in Theorem SIR-2.5. Ponomareva and Tamer work with projections of this set, which coincide with the bounds in Stoye (2007).

Ponomareva and Tamer also point out that depending on the DGP, misspecification can cause $\mathcal{H}_{\mathsf{P}}[\theta]$ to be spuriously tight. This can happen, for example, if $\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x})$ and $\mathbb{E}_{\mathsf{P}}(\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x})$ are sufficiently nonlinear, even if they are relatively far from each other (e.g., Ponomareva and Tamer, 2011, Figure 1). Hence, caution should be taken when interpreting very tight partial identification results as indicative of a highly informative model and empirical evidence, as the possibility of model misspecification has to be taken into account. These observations naturally lead to the questions of how to test for model misspecification in the presence of partial identification, and of what are the consequences of misspecification for the confidence sets discussed in Section 4.3.

With partial identification, a null hypothesis of correct model specification (and its alternative) can be expressed as

$$H_0 : \mathcal{H}_{\mathsf{P}}[\theta] \neq \emptyset; \quad H_1 : \mathcal{H}_{\mathsf{P}}[\theta] = \emptyset.$$

Tests for this hypothesis have been proposed both for the case of nonparametric as well as semiparametric partially identified models. I refer to Santos (2012) for specification tests

in a partially identified nonparametric instrumental variable model; to Kitamura and Stoye (2018) for a nonparametric test in random utility models that checks whether a repeated cross section of demand data might have been generated by a population of rational consumers (thereby testing for the Axiom of Revealed Stochastic Preference); and to Guggenberger, Hahn, and Kim (2008) and Bontemps, Magnac, and Maurin (2012) for specification tests in linear moment (in)equality models.

For the general class of moment inequality models discussed in Section 4, Romano and Shaikh (2008), Andrews and Guggenberger (2009), and Andrews and Soares (2010) propose a specification test that rejects the model if C_n in (4.10) is empty, where C_n is defined with $c_{1-\alpha}(\vartheta)$ determined so as to satisfy (4.14) and approximated according to the methods proposed in the respective papers. The resulting test, commonly referred to as *by-product* test because obtained as a by-product to the construction of a confidence set, takes the form

$$\phi = \mathbf{1}(C_n = \emptyset) = \mathbf{1}\left(\inf_{\vartheta\in\Theta} nq_n(\vartheta) > c_{1-\alpha}(\vartheta)\right).$$

Denoting by $\mathcal{P}_0$ the collection of $\mathsf{P} \in \mathcal{P}$ such that $\mathcal{H}_{\mathsf{P}}[\theta] \neq \emptyset$, one then has that test BP achieves uniform size control (Bugni, Canay, and Shi, 2015, Theorem C.2):

$$\limsup_{n\to\infty} \sup_{\mathsf{P}\in\mathcal{P}_0} \mathbb{E}_{\mathsf{P}}(\phi) \leq \alpha. \tag{5.1}$$

An important feature of test BP is that the critical value $c_{1-\alpha}(\vartheta)$ is not obtained to test for model misspecification, but it is obtained to insure the coverage requirement in (4.14); hence, it is obtained by working with the asymptotic distribution of $nq_n(\vartheta)$. Bugni, Canay, and Shi (2015) propose more powerful model specification tests, using a critical value $c_{1-\alpha}$ that they obtain to ensure that (5.1), rather than (4.14), holds. In particular, Bugni, Canay, and Shi (2015) show that their tests dominate test BP in terms of power in any finite sample and in the asymptotic limit. Their critical value is obtained by working with the asymptotic distribution of $\inf_{\vartheta\in\Theta} nq_n(\vartheta)$. As such, their proposal resembles the classic approach to model specification testing (J-test) in point identified generalized method of moments models.

While it is possible to test for misspecification also in partially identified models, a word of caution is due on what might be the effects of misspecification on confidence sets constructed as in (4.10) with $c_{1-\alpha}$ determined to insure (4.14), as it is often done in empirical work. Bugni, Canay, and Guggenberger (2012) show that in the presence of local misspecification, confidence sets C_n designed to satisfy (4.14) fail to do so. In practice, the concern is that when the model is misspecified C_n might be spuriously small. Indeed, we have seen that it can be empty if the misspecification is sufficiently severe. If it is less severe but still present, it may lead to inference that is erroneously interpreted as precise.

It is natural to wonder how this compares to the effect of misspecification on inference

in point identified models.[90] In that case, the rich set of tools available for inference allows one to avoid this problem. Consider for example a point identified generalized method of moments model with moment conditions $\mathbb{E}_{\mathsf{P}}(m_j(\boldsymbol{w};\theta)) = 0$, $j = 1,\ldots,|\mathcal{J}|$. Let m denote the vector that stacks each of the m_j functions, and let the estimator of θ be

$$\hat{\theta}_n = \arg\min_{\vartheta\in\Theta} n\bar{m}_n(\vartheta)\hat{\Xi}^{-1}\bar{m}_n(\vartheta), \tag{5.2}$$

with $\hat{\Xi}$ a consistent estimator of $\Xi = \mathbb{E}_{\mathsf{P}}[m(\boldsymbol{w};\theta)^\top m(\boldsymbol{w};\theta)]$ and $\bar{m}_n(\vartheta)$ the sample analog of $\mathbb{E}_{\mathsf{P}}(m(\boldsymbol{w};\vartheta))$. Let $\Lambda = \mathbb{E}_{\mathsf{P}}[\partial(m(\boldsymbol{w};\theta))^\top/\partial\vartheta]$, and let $\hat{\Sigma}$ be a consistent and robust estimator of $\Sigma = (\Lambda^\top\Xi^{-1}\Lambda)^{-1}$. Define a Wald-statistic based confidence ellipsoid as

$$\{\vartheta\in\Theta : n(\hat{\theta}_n - \vartheta)\hat{\Sigma}^{-1}(\hat{\theta}_n - \vartheta) \leq c_{d,1-\alpha}\}, \tag{5.3}$$

with $c_{d,1-\alpha}$ the $1-\alpha$ critical value of a χ^2_d (chi-squared random variable with d degrees of freedom). Under standard regularity conditions the confidence set in (5.3) covers θ with asymptotic probability $1-\alpha$ if the model is correctly specified. If the model is incorrectly specified, it covers a pseudo-true vector (the probability limit of (5.2)) with asymptotic probability $1-\alpha$. In either case, (5.3) is never empty and its volume depends on $\hat{\Sigma}$.

Even in the point identified case a confidence set constructed similarly to (4.10), i.e.,

$$\{\vartheta\in\Theta : n\bar{m}_n(\vartheta)\hat{\Xi}^{-1}\bar{m}_n(\vartheta) \leq c_{|\mathcal{J}|,1-\alpha}\}, \tag{5.4}$$

where $c_{|\mathcal{J}|,1-\alpha}$ is the $1-\alpha$ critical value of a $\chi^2_{|\mathcal{J}|}$, incurs the same problems as its partial identification counterpart. Under standard regularity conditions, if the model is correctly specified, the confidence set in (5.4) covers θ with asymptotic probability $1-\alpha$, because $n\bar{m}_n(\vartheta)\hat{\Xi}^{-1}\bar{m}_n(\vartheta) \Rightarrow \chi^2_{|\mathcal{J}|}$. However, this confidence set is empty with asymptotic probability $\mathbb{P}(\chi^2_{|\mathcal{J}|-d} > c_{|\mathcal{J}|,1-\alpha})$ (due to the facts that $\mathbb{P}(C_n = \emptyset) = \mathbb{P}(\hat{\theta}_n \notin C_n)$ and that $n\bar{m}_n(\hat{\theta}_n)\hat{\Xi}^{-1}\bar{m}_n(\hat{\theta}_n) \Rightarrow \chi^2_{|\mathcal{J}|-d}$), and it can be arbitrarily small.

In the very special case of a linear regression model with interval outcome data studied by Ponomareva and Tamer (2011), the procedure proposed by Beresteanu and Molinari (2008) yields confidence sets satisfying (4.11) that are always non-empty and robust to misspecification. The test statistic they use is based on the Hausdorff distance between the estimator and the hypothesized set, and as such is a generalization of the standard Wald-statistic to the set-valued case. For some related moment inequality models, Kaido and White (2013) propose to build a pseudo-true set $\mathcal{H}^*_{\mathsf{P}}[\theta]$ that is obtained through a two-step procedure. In the first step one obtains a nonparametric estimator of the function(s) for which the researcher wants to impose a parametric structure. In the second step one obtains the set $\mathcal{H}^*_{\mathsf{P}}[\theta]$ as the collection of least squares projections of the set in the first step, on the parametric class

[90] The considerations that I report here are based on a conversation with Joachim Freyberger and notes that he subsequently shared with me, for which I thank him.

imposed. Kaido and White show that under regularity conditions the pseudo-true set can be consistently estimated, and derive rates of convergence for the estimator; however, they do not provide methods to obtain confidence sets. While conceptually valuable, their construction appears to be computationally difficult. Masten and Poirier (2018) propose that when a model is falsified (in the sense that $\mathcal{H}_{\mathsf{P}}[\theta]$ is empty) one should report the *falsification frontier*: the boundary between the set of assumptions which falsify the model and those which do not, obtained through continuous relaxations of the baseline assumptions of concern. The researcher can then present the set $\mathcal{H}_{\mathsf{P}}[\theta]$ that results if the true model lies somewhere on this frontier. This set can be interpreted as a pseudo-true set. However, Masten and Poirier do not provide methods for inference.

How to build confidence sets that are robust to model misspecification and that cannot be empty or arbitrarily small in general moment inequality models remains an open and important question in the literature.

6 Computational Challenges

As a rule of thumb, the difficulty in computing estimators of identification regions and confidence sets depends on whether a closed form expression is available for the boundary of the set. For example, often nonparametric bounds on functionals of partially identified distribution are known functionals of observed conditional distributions, as in Section 2. In this case "plug in" estimation is possible, and the computational cost is the same as for estimation and construction of confidence intervals (or confidence bands) for point-identified nonparametric regressions (incurred twice, once for the lower bound and once for the upper bound).

Similarly, support function based inference is easy to implement when $\mathcal{H}_{\mathsf{P}}[\theta]$ is convex. Sometimes the extreme points of $\mathcal{H}_{\mathsf{P}}[\theta]$ can be expressed as known functions of observed distributions. Even if not, level sets of convex functions are easy to compute.

But as it was shown in Section 3, many problems of interest yield a set $\mathcal{H}_{\mathsf{P}}[\theta]$ that is *not* convex. In this case, $\mathcal{H}_{\mathsf{P}}[\theta]$ is obtained as a level set of a criterion function. Because $\mathcal{H}_{\mathsf{P}}[\theta]$ (or its associated confidence set) is often a subset of $\mathbb{R}^d$ (rather than $\mathbb{R}$), even a moderate value for d, e.g., 8 or 10, can lead to extremely challenging computational problems. This is because if one wants to compute $\mathcal{H}_{\mathsf{P}}[\theta]$ or a set that covers it or its elements with a prespecified asymptotic probability (possibly uniformly over $\mathsf{P} \in \mathcal{P}$), one has to map out a level set in $\mathbb{R}^d$. If one is interested in scalar projections or smooth functions of $\vartheta \in \mathcal{H}_{\mathsf{P}}[\theta]$, one needs to solve nonlinear optimization problems, as for example in (4.19). The latter can be difficult to do, especially because $c_{1-\alpha}(\vartheta)$ is typically an unknown function of ϑ for which gradients are not available in closed form.

Mirroring the fact that computation is easier when the boundary of $\mathcal{H}_{\mathsf{P}}[\theta]$ is a known function of observed conditional distributions, several portable software packages are available to carry out estimation and inference in this case. For example, Beresteanu and Manski

(2000) provide STATA and MatLab packages implementing the methods proposed by Manski (1989, 1990, 1994, 1995, 1997b); Horowitz and Manski (1998, 2000); Manski and Pepper (2000). Tauchmann (2014) provides a STATA package to implement the bounds proposed by Lee (2009). McCarthy, Millimet, and Roy (2015) provide a STATA package to implement bounds on treatment effects with endogenous and misreported treatment assignment and under the assumptions of monotone treatment selection, monotone treatment response, and monotone instrumental-variables as in Manski (1997b), Manski and Pepper (2000), Kreider and Pepper (2007), Gundersen, Kreider, and Pepper (2012), Kreider, Pepper, Gundersen, and Jolliffe (2012). The code computes the confidence intervals proposed by Imbens and Manski (2004). In the more general context of inference for a one-dimensional parameter defined by intersection bounds, as for example the one in (2.15), Chernozhukov, Kim, Lee, and Rosen (2015) and Andrews, Kim, and Shi (2017) provide portable STATA code implementing, respectively, methods to test hypotheses and build confidence intervals in Chernozhukov, Lee, and Rosen (2013) and Andrews and Shi (2013).

Beresteanu, Molinari, and Morris (2010) provide portable STATA code implementing Beresteanu and Molinari (2008)'s method for estimation and inference for best linear prediction with interval outcome data as in Identification Problem 2.4. Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2012) provide R code implementing Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2018)'s method for estimation and inference for best linear approximations of set identified functions.

On the other hand, there is a paucity of portable software implementing the theoretical methods for inference in structural partially identified models discussed in Section 4. Ciliberto and Tamer (2009) compute Chernozhukov, Hong, and Tamer (2007) confidence sets for a parameter vector in $\mathbb{R}^d$, with d in the order of 20 and with tens of thousands of inequalities, through a "guess and verify" algorithm based on simulated annealing (with no cooling) that visits many candidate values of $\vartheta \in \Theta$, evaluates $q_n(\vartheta)$, and builds C_n as in (4.10) with $c_{1-\alpha}$ defined to satisfy (4.12). This is a tremendously hard task, due to the dimension of θ and the number of moment inequalities.

In terms of general purpose portable code, I am only aware of the MatLab package provided by Kaido, Molinari, Stoye, and Thirkettle (2017) to implement the inference method of Kaido, Molinari, and Stoye (2019a) for projections of parameter vectors in models defined by a finite number of unconditional moment (in)equalities. Their framework can be used to further highlight why the computational task is challenging even in the case of projections. The confidence interval in (4.18)-(4.19) requires solving two nonlinear programs, each with a linear objective and nonlinear constraints involving a critical value which in general is an unknown function of ϑ, with unknown gradient. When the dimension of the parameter vector is large, directly solving optimization problems with such constraints can be expensive even if

evaluating the critical value at each ϑ is cheap.[91] Hence, Kaido, Molinari, and Stoye propose an algorithm (called E-A-M for Evaluation-Approximation-Maximization) to approximate $c_{1-\alpha}$ through a Gaussian process, and determine evaluation points of the nonlinear program in (4.19) based on the *expected improvement* for the value of the program assessed based on the approximating surface. Their algorithm belongs to the family of *expected improvement algorithms* (see e.g. Jones, Schonlau, and Welch, 1998; Schonlau, Welch, and Jones, 1998; Jones, 2001, and references therein). Bull (2011) establishes convergence, as the number of evaluation points increases, of an expected improvement algorithm for unconstrained optimization problems where the objective is a "black box" function. The rate of convergence that he derives depends on the smoothness of the black box objective function. Kaido, Molinari, and Stoye (2019a) substantially extend his results to show convergence, at a slightly slower rate, of their related algorithm for constrained optimization problems in which the constraints are sufficiently smooth "black box" functions. Their Monte Carlo experiments suggest that the E-A-M algorithm is fast and accurate at computing the confidence intervals in Kaido, Molinari, and Stoye (2019a). The E-A-M algorithm also allows for very rapid computation of projections of the confidence set proposed by Andrews and Soares (2010), and for a substantial improvement in the computational time of the profiling-based confidence intervals proposed by Bugni, Canay, and Shi (2017).[92] In all cases, the speed improvement results from a reduced number of evaluation points required to approximate the optimum.

7 Conclusions

This chapter provides a discussion of the econometrics literature on partial identification. It first reviews what can be learned about (functionals of) probability distributions in the absence of parametric restrictions, under various scenarios of *data incompleteness*. It then reviews what can be learned about functionals characterizing semiparametric structural economic models, under various scenarios of *model incompleteness*. It then discusses finite sample inference, the consequences of misspecification, and the computational challenges that a researcher needs to face when implementing partial identification methods.

Taking stock, I argue that several areas emerge where more progress is needed to bring the partial identification approach to empirical research to full fruition. Whereas the last twenty years have seen the development of a burgeoning theoretical literature on the topic, empirical applications of the methods still lag behind. I conjecture that part of the reason for this discrepancy is due to the lack of easy-to-implement procedures for computation of estimators and confidence sets (or intervals) in complex structural models. While the literature so far

[91] Kaido, Molinari, and Stoye (2019a) propose a linearization method whereby $c_{1-\alpha}$ is calibrated through repeatedly solving bootstrap linear programs, hence it is reasonably cheap to compute.

[92] Bugni, Canay, and Shi (2017)'s method does not require solving a nonlinear program such as the one in (4.19). Rather it obtains CI_n as in (4.17). However, it approximates $c_{1-\alpha}$ by repeatedly solving bootstrap nonlinear programs, thereby incurring a very high computational cost at that stage.

has aimed at developing methods that have desirable asymptotic properties for very general classes of models, there is arguably scope for more problem-specific methods that exploit the particularities of a certain model to obtain easy to implement statistical procedures. It would also seem desirable that portable software accompanies the proposed methodologies, perhaps more in line with the current practice in the Statistics literature.

However, computational concerns cannot be the cause of the relative paucity of applications of partial identification methods as the ones reviewed in Section 2, e.g., bounds on treatment effects. These bounds are extremely easy to estimate and confidence intervals covering them can readily be computed. I therefore conjecture that the lack of applications might be due to a misconception, whereby nonparametric bounds are perceived as "always too wide to learn anything". While it is true that, for example, worst-case nonparametric bounds on the average treatment effect cover zero by construction, the partial identification approach to empirical research proposes a wide array of assumptions that can be brought to bear to augment the empirical evidence and tighten the bounds. The philosophy of the method is that the systematic reporting of bounds obtained under an increasingly strong set of assumptions illuminates the relative role of assumptions and data in shaping the conclusions that the researcher draws. Point identification is the limit of this process, and carefully assessing how this limit is reached is key to learning about the quantities of interest.

In Sections 2 and 3, special attention is devoted to characterizing *sharp* identification regions. Sharpness often requires *many* moment inequalities, the number of which can exceed the available sample size. Hence, there is a need of appropriate statistical inference methods. As briefly mentioned in Section 4, some methods designed to provide valid test of hypotheses and confidence sets in this scenario already exist. However, I would argue that there is a need to better understand the trade-off between sharpness of the population identification region, and statistical efficiency, especially in the context of conditional moment inequalities where instrument functions are needed to transform the inequalities in unconditional ones. Similarly, there is a need of more research on data driven procedures for the choice of tuning parameters for the construction of confidence sets, in particular in the case of projection inference where the question has not yet been addressed. Another open and arguably important question in the literature, is how to build confidence sets for moment inequality models that are robust to model misspecification and that cannot be empty or arbitrarily small.

A Basic Definitions and Facts from Random Set Theory

This appendix provides basic definitions and results from random set theory that are used throughout this chapter.[93] I refer to Molchanov (2017) for a textbook presentation of random set theory, and to Molchanov and Molinari (2018) for a discussion focusing on its applications to econometrics.

The theory of random closed sets generally applies to the space of closed subsets of a locally compact Hausdorff second countable topological space $\mathfrak{X}$, see Molchanov (2017). In this chapter I let $\mathfrak{X} = \mathbb{R}^d$ to simplify the exposition. Closedness is a property satisfied by random points (singleton sets), so that the theory of random closed sets includes the classical case of random points or random vectors as a special case. A random closed set is a measurable map $\boldsymbol{X} : \Omega \mapsto \mathcal{F}$, where measurability is defined by specifying the family of functionals of $\boldsymbol{X}$ that are random variables.

DEFINITION A.1 (Random closed set): *A map $\boldsymbol{X}$ from a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ to the family $\mathcal{F}$ of closed subsets of $\mathbb{R}^d$ is called a* random closed set *if*

$$\boldsymbol{X}^{-}(K) = \{\omega \in \Omega : \ \boldsymbol{X}(\omega) \cap K \neq \emptyset\} \tag{A.1}$$

belongs to the σ-algebra $\mathfrak{F}$ on Ω for each compact set K in $\mathbb{R}^d$.

A random *compact* set is a random closed set which is compact with probability one, so that almost all values of $\boldsymbol{X}$ are compact sets. A random *convex* closed set is defined similarly, so that $\boldsymbol{X}(\omega)$ is a convex closed set for almost all ω.

Definition A.1 means that $\boldsymbol{X}$ is explored by its hitting events, that is the events where $\boldsymbol{X}$ hits a compact set K. The corresponding hitting probabilities are very important in random set theory, because they uniquely determine the probability distribution of a random closed set $\boldsymbol{X}$, see Molchanov (2017, Ch. 1, Sec. 1.1.3). The formal definition of the hitting probabilities, and the closely related containment probabilities, follows.

DEFINITION A.2 (Capacity functional and containment functional):

1. *A functional $\mathsf{T}_{\boldsymbol{X}}(K) : \mathcal{K} \mapsto [0,1]$ given by*

$$\mathsf{T}_{\boldsymbol{X}}(K) = \mathbb{P}\{\boldsymbol{X} \cap K \neq \emptyset\}, \quad K \in \mathcal{K},$$

   is called capacity (or hitting) functional *of $\boldsymbol{X}$.*

2. *A functional $\mathsf{C}_X(F) : \mathcal{F} \mapsto [0,1]$ given by*

$$\mathsf{C}_{\boldsymbol{X}}(F) = \mathbb{P}\{\boldsymbol{X} \subset F\}, \quad F \in \mathcal{F},$$

[93] The treatment here summarizes a few of the topics presented in Molchanov and Molinari (2018).

is called the containment functional *of* $\boldsymbol{X}$.

I write $\mathsf{T}(K)$ *instead of* $\mathsf{T}_{\boldsymbol{X}}(K)$ *and* $\mathsf{C}(K)$ *instead of* $\mathsf{C}_{\boldsymbol{X}}(K)$ *where no ambiguity occurs.*

Ever since the seminal work of Aumann (1965), it has been common to think of random sets as bundles of random variables – the selections of the random sets.

Definition A.3 (Measurable selection): *For any random set* $\boldsymbol{X}$*, a (measurable)* selection *of* $\boldsymbol{X}$ *is a random element* $\boldsymbol{x}$ *with values in* $\mathbb{R}^d$ *such that* $\boldsymbol{x}(\omega) \in \boldsymbol{X}(\omega)$ *almost surely. I denote by* $\mathrm{Sel}(\boldsymbol{X})$ *the set of all selections from* $\boldsymbol{X}$.

The space of closed sets is not linear, which causes substantial difficulties in defining the expectation of a random set. One approach, inspired by Aumann (1965) and pioneered by Artstein and Vitale (1975), relies on representing a random set using the family of its selections, and considering the set formed by their expectations. If $\boldsymbol{X}$ possesses at least one integrable selection, then $\boldsymbol{X}$ is called *integrable*. The family of all integrable selections of $\boldsymbol{X}$ is denoted by $\mathrm{Sel}^1(\boldsymbol{X})$.

Definition A.4 (Unconditional and conditional Aumann –or selection– expectation): *The* (selection or Aumann) expectation *of an integrable random closed set* $\boldsymbol{X}$ *is given by*

$$\mathbb{E}\boldsymbol{X} = \mathrm{cl}\left\{\int_\Omega \boldsymbol{x} d\mathbb{P} : \ \boldsymbol{x} \in \mathrm{Sel}^1(\boldsymbol{X})\right\}.$$

For each sub-σ*-algebra* $\mathfrak{B} \subset \mathfrak{F}$*, the* conditional selection or Aumann expectation *of* $\boldsymbol{X}$ *given* $\mathfrak{B}$ *is the* $\mathfrak{B}$*-measurable random closed set* $\boldsymbol{Y} = \mathbb{E}(\boldsymbol{X}|\mathfrak{B})$ *such that the family of* $\mathfrak{B}$*-measurable integrable selections of* $\boldsymbol{Y}$*, denoted* $\mathrm{Sel}^1_{\mathfrak{B}}(\boldsymbol{Y})$*, satisfies*

$$\mathrm{Sel}^1_{\mathfrak{B}}(\boldsymbol{Y}) = \mathrm{cl}\left\{\mathbb{E}(\boldsymbol{x}|\mathfrak{B}) : \ \boldsymbol{x} \in \mathrm{Sel}^1(\boldsymbol{X})\right\},$$

where the closure in the right-hand side is taken in $\mathbf{L}^1$.

If $\boldsymbol{X}$ is almost surely non-empty and its norm $\|\boldsymbol{X}\| = \sup\{\|\boldsymbol{x}\| : \ \boldsymbol{x} \in \boldsymbol{X}\}$ is an integrable random variable, then $\boldsymbol{X}$ is said to be *integrably bounded* and all its selections are integrable. In this case the family of expectations of these integrable selections is already closed and there is no need to take an additional closure as required in Definition A.4, see Molchanov (2017, Theorem 2.1.37). The selection expectation depends on the probability space used to define $\boldsymbol{X}$, see Molchanov (2017, Section 2.1.2) and Molchanov and Molinari (2018, Section 3.1). In particular, if the probability space is non-atomic and $\boldsymbol{X}$ is integrably bounded, the selection expectation $\mathbb{E}\boldsymbol{X}$ is a convex set regardless of whether or not $\boldsymbol{X}$ might be non-convex itself Molchanov and Molinari (2018, Theorem 3.4). This convexification property of the selection expectation implies that the expectation of the closed convex hull of $\boldsymbol{X}$ equals the closed convex hull of $\mathbb{E}\boldsymbol{X}$, which in turn equals $\mathbb{E}\boldsymbol{X}$. It is then natural to describe the Aumann

expectation through its support function, because this function traces out a convex set's boundary and therefore knowing the support function is equivalent to knowing the set itself, see equation (A.2) below.

DEFINITION A.5 (Support function): *Let K be a convex set. The* support function *of K is*

$$h_K(u) = \sup\{k^\top u : \ k \in K\}, \qquad u \in \mathbb{R}^d,$$

where $k^\top u$ denotes the scalar product.

The support function is finite for all u if K is bounded, and is sublinear (positively homogeneous and subadditive) in u. Hence, it can be considered only for $u \in \mathbb{B}^d$ or $u \in \mathbb{S}^{d-1}$. Moreover, one has

$$K = \cap_{u\in\mathbb{B}^d}\{k : k^\top u \le h_K(u)\} = \cap_{u\in\mathbb{S}^{d-1}}\{k : k^\top u \le h_K(u)\}. \tag{A.2}$$

Next, I define the Hausdorff metric, a distance on the family $\mathcal{K}$ of compact sets:

DEFINITION A.6 (Hausdorff metric): *Let $K, L \in \mathcal{K}$. The* Hausdorff distance *between K and L is*

$$\mathbf{d}_{\mathrm{H}}(K, L) = \inf\Big\{r > 0 : \ K \subseteq L^r, \ L \subseteq K^r\Big\},$$

where $K^r = \{x : \dot{(}x, K) \le r\}$ is the r-envelope of K.

Since $K \subseteq L$ if and only if $h_K(u) \le h_L(u)$ for all $u \in \mathbb{S}^{d-1}$ and $h_{K^r}(u) = h_K(u) + r$, the uniform metric for support functions on the sphere turns into the Hausdorff distance between compact convex sets. Namely,

$$\mathbf{d}_{\mathrm{H}}(K, L) = \sup\Big\{|h_K(u) - h_L(u)| : \ \|u\| = 1\Big\}. \tag{A.3}$$

It follows that

$$\|K\| = \mathbf{d}_{\mathrm{H}}(K, \{0\}) = \sup\big\{|h_K(u)| : \ \|u\| = 1\big\}.$$

Finally, I define independently and identically distributed random closed sets (see Molchanov, 2017, Proposition 1.1.40 and Theorem 1.3.20, respectively):

DEFINITION A.7 (i.i.d. random closed sets): *Random closed sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^d$ are independent if and only if $\mathbb{P}\{\boldsymbol{X}_1 \cap K_1 \neq \emptyset, \ldots, \boldsymbol{X}_n \cap K_n \neq \emptyset\} = \prod_{i=1}^n \mathsf{T}_{\boldsymbol{X}_i}(K_i)$ for all $K_1, \ldots, K_n \in \mathcal{K}$. They are identically distributed if and only if for each open set G, $\mathbb{P}\{\boldsymbol{X}_1 \cap G \neq \emptyset\} = \mathbb{P}\{\boldsymbol{X}_2 \cap G \neq \emptyset\} = \cdots = \mathbb{P}\{\boldsymbol{X}_n \cap G \neq \emptyset\}$.*

With these definitions in hand, I can state the theorems used throughout the chapter. The first is a dominance condition due to Artstein (1983) (and Norberg, 1992) that characterizes probability distributions of selections (see Molchanov and Molinari, 2018, Section 2.2):

Theorem A.1 (Artstein). *A probability distribution μ on $\mathbb{R}^d$ is the distribution of a selection of a random closed set $\boldsymbol{X}$ in $\mathbb{R}^d$ if and only if*

$$\mu(K) \leq \mathsf{T}(K) = \mathbb{P}\{\boldsymbol{X} \cap K \neq \emptyset\} \tag{A.4}$$

for all compact sets $K \subseteq \mathbb{R}^d$. Equivalently, if and only if

$$\mu(F) \geq \mathsf{C}(F) = \mathbb{P}\{\boldsymbol{X} \subset F\} \tag{A.5}$$

for all closed sets $F \subset \mathbb{R}^d$. If $\boldsymbol{X}$ is a compact random closed set, it suffices to check (A.5) for compact sets F only.

If μ from Theorem A.1 is the distribution of some random vector $\boldsymbol{x}$, then it is not guaranteed that $\boldsymbol{x} \in \boldsymbol{X}$ a.s., e.g. $\boldsymbol{x}$ can be independent of $\boldsymbol{X}$. Theorem A.1 means that for each such μ, it is possible to construct $\boldsymbol{x}$ with distribution μ that belongs to $\boldsymbol{X}$ almost surely. In other words, $\boldsymbol{x}$ and $\boldsymbol{X}$ can be realized on the same probability space (coupled) as random elements $\boldsymbol{x}'$ and $\boldsymbol{X}'$ such that $\boldsymbol{x} \stackrel{d}{=} \boldsymbol{x}'$ and $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{X}'$ with $\boldsymbol{x}' \in \boldsymbol{X}'$ a.s.

The definition of the distribution of a random closed set (Definition A.2) and the characterization results for its selections in Theorem A.1 require working with functionals defined on the family of all compact sets, which in general is very rich. It is therefore important to reduce the family of all compact sets required to describe the distribution of the random closed set or to characterize its selections.

Definition A.8: *A family of compact sets $\mathcal{M}$ is said to be a* core determining class *for a random closed set $\boldsymbol{X}$ if any probability measure μ satisfying the inequalities*

$$\mu(K) \leq \mathbb{P}\{\boldsymbol{X} \cap K \neq \emptyset\} \tag{A.6}$$

for all $K \in \mathcal{M}$, is the distribution of a selection of $\boldsymbol{X}$, implying that (A.6) holds for all compact sets K.

The notion of a core determining class was introduced by Galichon and Henry (2006). A simple and general, but still mostly too rich, core determining class is obtained as a subfamily of all compact sets that is dense in a certain sense in the family $\mathcal{K}$. For instance, in the Euclidean space, it suffices to consider compact sets obtained as finite unions of closed balls with rational centers and radii (e.g., Galichon and Henry, 2006, Theorem 3c). For the case that $\boldsymbol{X}$ is a subset of a finite space, Beresteanu, Molchanov, and Molinari (2008, Algorithm 5.1) propose a simple algorithm to compute core determining classes. Chesher and Rosen (2012) provide a related algorithm. Throughout this chapter, several results are mentioned where the class of sets over which (A.4) is verified is reduced from the class of compact subsets of the carrier space, to a (significantly) smaller collection.

The next result characterizes a dominance condition that can be used to verify the existence of selections of $\boldsymbol{X}$ with specific properties for their means (see Molchanov and Molinari, 2018, Sections 3.2-3.3)

Theorem A.2 (Convexification in $\mathbb{R}^d$)**.** *Let $\boldsymbol{X}$ be an integrable random set. If $\boldsymbol{X}$ is defined on a non-atomic probability space, or if $\boldsymbol{X}$ is almost surely convex, then $\mathbb{E}\boldsymbol{X} = \mathbb{E}\operatorname{conv}\boldsymbol{X}$ and*

$$\mathbb{E}h_{\boldsymbol{X}}(u) = h_{\mathbb{E}\boldsymbol{X}}(u), \quad u \in \mathbb{R}^d. \tag{A.7}$$

If $\mathbb{P}$ is atomless over $\mathfrak{B}$,[94] *then $\mathbb{E}(\boldsymbol{X}|\mathfrak{B})$ is convex and*

$$\mathbb{E}(h_{\boldsymbol{X}}(u)|\mathfrak{B}) = h_{\mathbb{E}(\boldsymbol{X}|\mathfrak{B})}(u), \quad u \in \mathbb{R}^d. \tag{A.8}$$

Hence, for any vector $b \in \mathbb{R}^d$, it holds that

$$b \in \mathbb{E}\boldsymbol{X} \Leftrightarrow b^\top u \le \mathbb{E}h_{\boldsymbol{X}}(u) \ \ \forall u \in \mathbb{S}^{d-1}, \tag{A.9}$$

$$b \in \mathbb{E}(\boldsymbol{X}|\mathfrak{B}) \Leftrightarrow b^\top u \le \mathbb{E}(h_{\boldsymbol{X}}(u)|\mathfrak{B}) \ \ \forall u \in \mathbb{S}^{d-1}. \tag{A.10}$$

An important consequence of Theorem A.2 is that it allows one to verify whether $b \in \mathbb{E}\boldsymbol{X}$ without having to compute $\mathbb{E}\boldsymbol{X}$ but only $\mathbb{E}h_{\boldsymbol{X}}(u)$ (and similarly for the conditional case), a substantially easier task.

Finally, i.i.d. random closed sets satisfy a law of large numbers and a central limit theorem that are similar to these for random singletons. Recall that the *Minkowski sum* of two sets K and L in a linear space (which in this chapter I assume to be the Euclidean space $\mathbb{R}^d$) is obtained by adding each point from K to each point from L, formally,

$$K + L = \{x + y : \ x \in K, \ y \in L\}.$$

Below, $\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$ denotes the Minkowski sum of the random closed sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and $(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n)/n$ denotes their *Minkowski average*.

Theorem A.3 (Law of large numbers for integrably bounded random sets)**.** *Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. integrably bounded random compact sets. Define $\boldsymbol{S}_n = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. Then*

$$\mathbf{d}_{\mathrm{H}}\left(\frac{\boldsymbol{S}_n}{n}, \mathbb{E}\boldsymbol{X}\right) \to 0 \quad a.s. \ as \ \ n \to \infty. \tag{A.11}$$

The support function of a random closed set $\boldsymbol{X}$ such that $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$, is a random continuous function $h_{\boldsymbol{X}}(u)$ on $\mathbb{S}^{d-1}$ with square integrable values. Define its covariance

[94] An event $A' \in \mathfrak{B}$ is called a $\mathfrak{B}$-atom if $\mathbb{P}\{0 < \mathbb{P}(A|\mathfrak{B}) < \mathbb{P}(A'|\mathfrak{B})\} = 0$ for all $A \subset A'$ such that $A \in \mathfrak{F}$.

function as

$$\Gamma_{\boldsymbol{X}}(u,v) \equiv \mathbb{E}\left[(h_{\boldsymbol{X}}(u) - h_{\mathbb{E}\boldsymbol{X}}(u))(h_{\boldsymbol{X}}(v) - h_{\mathbb{E}\boldsymbol{X}}(v))\right], \quad u,v \in \mathbb{S}^{d-1}. \tag{A.12}$$

Let $\zeta(u)$ be a centered Gaussian random field on $\mathbb{S}^{d-1}$ with the same covariance structure as $\boldsymbol{X}$, i.e. $\mathbb{E}\big[\zeta(u)\zeta(v)\big] = \Gamma_{\boldsymbol{X}}(u,v),\ u,v \in \mathbb{S}^{d-1}$. Since the support function of a compact set is Lipschitz, it is easy to show that the random field ζ has a continuous modification by bounding the moments of $|\zeta(u) - \zeta(v)|$.

Theorem A.4 (Central limit theorem)**.** *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. copies of a random closed set $\boldsymbol{X}$ in $\mathbb{R}^d$ such that $\mathbb{E}\|\boldsymbol{X}\|^2 < \infty$, and let $\boldsymbol{S}_n = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. Then as $n \to \infty$,*

$$\sqrt{n}\Big(h_{\frac{\boldsymbol{S}_n}{n}}(u) - h_{\mathbb{E}\boldsymbol{X}}(u)\Big) \Rightarrow \zeta \tag{A.13}$$

in the space of continuous functions on the unit sphere with the uniform metric. Furthermore,

$$\sqrt{n}\mathbf{d}_{\mathrm{H}}\left(\frac{\boldsymbol{S}_n}{n}, \mathbb{E}\boldsymbol{X}\right) \Rightarrow \|\zeta\|_\infty = \sup\big\{|\zeta(u)| :\ u \in \mathbb{S}^{d-1}\big\}. \tag{A.14}$$

References

Abaluck, J., and A. Adams (2018): "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," available at https://abiadams.com/wp-content/uploads/2018/06/DiscreteChoiceInattention_master.pdf.

Abbring, J. H., and J. J. Heckman (2007): "Chapter 72 Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, pp. 5145 – 5303. Elsevier.

Adusumilli, K., and T. Otsu (2017): "Empirical Likelihood for Random Sets," *Journal of the American Statistical Association*, 112(519), 1064–1075.

Andrews, D. W. K., and P. J. Barwick (2012): "Inference for parameters defined by moment inequalities: a recommended moment selection procedure," *Econometrica*, 80(6), 2805–2826.

Andrews, D. W. K., and P. Guggenberger (2009): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25(3), 669–709.

Andrews, D. W. K., W. Kim, and X. Shi (2017): "Commands for testing conditional moment inequalities and equalities," *Stata Journal*, 17(1), 56–72.

Andrews, D. W. K., and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81(2), 609–666.

——— (2017): "Inference based on many conditional moment inequalities," *Journal of Econometrics*, 196(2), 275 – 287.

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

Aradillas-Lopez, A., and E. Tamer (2008): "The Identification Power of Equilibrium in Simple Games," *Journal of Business & Economic Statistics*, 26(3), 261–283.

Aradillas-López, A., A. Gandhi, and D. Quint (2013): "Identification and Inference in Ascending Auctions With Correlated Private Values," *Econometrica*, 81(2), 489–534.

Armstrong, T. B. (2013): "Bounds in auctions with unobserved heterogeneity," *Quantitative Economics*, 4(3), 377–415.

——— (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181(2), 92 – 116.

——— (2015): "Asymptotically exact inference in conditional moment inequality models," *Journal of Econometrics*, 186(1), 51 – 65.

Armstrong, T. B., and H. P. Chan (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194(1), 24 – 43.

Artstein, Z. (1983): "Distributions of random sets and random selections," *Israel Journal of Mathematics*, 46, 313–324.

Artstein, Z., and R. A. Vitale (1975): "A strong law of large numbers for random compact sets," *Annals of Probabability*, 3, 879–882.

Athey, S., and P. A. Haile (2002): "Identification of Standard Auction Models," *Econometrica*, 70(6), 2107–2140.

Aucejo, E. M., F. A. Bugni, and V. J. Hotz (2017): "Identification and inference on regressions with missing covariate data," *Econometric Theory*, 33(1).

Aumann, R. J. (1965): "Integrals of set-valued functions," *Journal of Mathematical Analysis and Applications*, 12(1), 1–12.

Bajari, P., H. Hong, and S. P. Ryan (2010): "Identification and estimation of a discrete game of complete information," *Econometrica*, 78(5), 1529–1568.

Balke, A., and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2019): "Heterogeneous Choice Sets and Preferences," *Work in progress*.

Barseghyan, L., F. Molinari, T. O'Donoghue, and J. C. Teitelbaum (2013): "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103(6), 2499–2529.

——— (2018): "Estimating Risk Preferences in the Field," *Journal of Economic Literature*, 56(2).

Barseghyan, L., F. Molinari, and J. C. Teitelbaum (2016): "Inference under stability of risk preferences," *Quantitative Economics*, 7(2), 367–409.

Barseghyan, L., F. Molinari, and M. Thirkettle (2019): "Discrete Choice under Risk with Limited Consideration," CeMMAP working paper CWP08/19, available at https://www.cemmap.ac.uk/publication/id/13912.

Bazaraa, M. S., H. D. Sherali, and C. Shetty (2006): *Nonlinear programming: theory and algorithms*. Hoboken, N.J. : Wiley-Interscience, 3rd edn.

Beresteanu, A., and C. F. Manski (2000): "Bounds for STATA and Bounds for MatLab," available at http://faculty.wcas.northwestern.edu/~cfm754/bounds_stata.pdf.

Beresteanu, A., I. Molchanov, and F. Molinari (2008): "Sharp Identification Regions in Games," CeMMAP working paper CWP15/08, available at https://www.cemmap.ac.uk/publication/id/4264.

——— (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79(6), 1785–1821.

——— (2012): "Partial identification using random set theory," *Journal of Econometrics*, 166(1), 17 – 32, with errata at http://economics.cornell.edu/fmolinari/NOTE_BMM2012_v3.pdf.

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

Beresteanu, A., F. Molinari, and D. S. Morris (2010): "Asymptotics for Partially Identified Models in STATA," available at https://molinari.economics.cornell.edu/programs/Stata_SetBLP.zip.

Bergemann, D., and S. Morris (2016): "Bayes correlated equilibrium and the comparison of information structures in games," *Theoretical Economics*, 11(2), 487–522.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Berry, S., and E. Tamer (2006): "Identification in Models of Oligopoly Entry," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. K. Newey, and T. E. Persson, vol. 2 of *Econometric Society Monographs*, p. 46–85. Cambridge University Press.

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

Berry, S. T., and G. Compiani (2019): "An Instrumental Variable Approach to Dynamic Models," available at https://drive.google.com/file/d/1pl1PW1w8eh3gnrTMKUBuS6T6TIKtvf9c/view.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012): "Treatment effect bounds: An application to Swan–Ganz catheterization," *Journal of Econometrics*, 168(2), 223 – 243.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

Bjorn, P. A., and Q. H. Vuong (1984): "Simultaneous Equations Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Labor Force Participation," CIT working paper SSWP 537, California Institute of Technology, available at http://resolver.caltech.edu/CaltechAUTHORS:20170919-140310752.

Blevins, J. R. (2015): "Non-Standard Rates of Convergence of Criterion-Function-Based Set Estimators," *Econometrics Journal*, 18, 172–199.

Block, H. D., and J. Marschak (1960): "Random Orderings and Stochastic Theories of Responses," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. by I. Olkin, pp. 97–132. Stanford University Press.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011): "Identification of Social Interactions," in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. O. Jackson, vol. 1, pp. 853 – 964. North-Holland.

Blundell, R., M. Browning, and I. Crawford (2008): "Best Nonparametric Bounds on Demand Responses," *Econometrica*, 76(6), 1227–1262.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75(2), 323–363.

Blundell, R., D. Kristensen, and R. Matzkin (2014): "Bounding quantile demand functions using revealed preference inequalities," *Journal of Econometrics*, 179(2), 112 – 127.

Blundell, R., and J. R. Smith (1994): "Coherency and Estimation in Simultaneous Models with Censored or Qualitative Dependent Variables," *Journal of Econometrics*, 64, 355–373.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80(3), 1129–1155.

Bresnahan, T. F., and P. C. Reiss (1988): "Do Entry Conditions Vary Across Markets?," *Brookings Papers on Economic Activity*, pp. 833–871.

——— (1990): "Entry in Monopoly Markets," *The Review of Economic Studies*, 57(4), 531–553.

——— (1991): "Empirical models of discrete games," *Journal of Econometrics*, 48(1), 57–81.

Brock, W. A., and S. N. Durlauf (2001): "Chapter 54 Interactions-Based Models," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 5, pp. 3297–3380. Elsevier.

Bugni, F. A. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set," *Econometrica*, 78(2), 735–753.

——— (2016): "Comparison of inferential methods in partially identifies models in terms of error in coverage probability," *Econometric Theory*, 32(1), 187–242.

Bugni, F. A., I. A. Canay, and P. Guggenberger (2012): "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80(4), 1741–1768.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259 – 282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Bull, A. D. (2011): "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, 12(Oct), 2879–2904.

Bureau of Labor Statistics (2018): "Occupational Employment Statistics," U.S. Department of Labor, online www.bls.gov/oes/; accessed 1/28/2018.

Canay, I. A. (2010): "EL inference for partially identified models: Large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156(2), 408 – 425.

Canay, I. A., and A. M. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 2 of *Econometric Society Monographs*, p. 271–306. Cambridge University Press.

Caplin, A. (2016): "Measuring and Modeling Attention," *Annual Review of Economics*, 8(1), 379–403.

Cattaneo, M. D., X. Ma, Y. Masatlioglu, and E. Suleymanov (2017): "A Random Attention Model," available at https://arxiv.org/abs/1712.03448.

Chandrasekhar, A. (2016): "Econometrics of Network Formation," in *Oxford Handbook on the Economics of Networks*, ed. by Y. Bramoulle, A. Galeotti, and B. Rogers, chap. 13. Oxford University Press.

Chandrasekhar, A., V. Chernozhukov, F. Molinari, and P. Schrimpf (2012): "R code implementing best linear approximations to set identified functions," available at https://bitbucket.org/paulschrimpf/mulligan-rubinstein-bounds.

——— (2018): "Best linear approximations to set identified functions: with an application to the gender wage gap," CeMMAP working paper CWP09/19, available at https://www.cemmap.ac.uk/publication/id/13913.

Chen, X., T. M. Christensen, and E. Tamer (2018): "MCMC Confidence Sets for Identified Sets," *Econometrica*, 86(6), 1965–2018.

Chernozhukov, V., D. Chetverikov, and K. Kato (2018): "Inference on causal and structural parameters using many moment inequalities," *Review of Economic Studies*, forthcoming.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81(2), 535–580.

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

Chernozhukov, V., W. Kim, S. Lee, and A. Rosen (2015): "Implementing intersection bounds in Stata," *Stata Journal*, 15(1), 21–44.

Chernozhukov, V., E. Kocatulum, and K. Menzel (2015): "Inference on sets in finance," *Quantitative Economics*, 6(2), 309–358.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: estimation and inference," *Econometrica*, 81(2), 667–737.

Chesher, A., and A. Rosen (2012): "Simultaneous equations for discrete outcomes: coherence, completeness, and identification," CeMMAP working paper CWP21/12, available at https://www.cemmap.ac.uk/publication/id/6297.

——— (2017a): "Incomplete English auction models with heterogeneity," CeMMAP working paper CWP27/17, available at https://www.cemmap.ac.uk/publication/id/9277.

Chesher, A., and A. M. Rosen (2017b): "Generalized instrumental variable models," *Econometrica*, 85, 959–989.

——— (2019): "Generalized instrumental variable models, methods, and applications," in *Handbook of Econometrics*. Elsevier.

Chesher, A., A. M. Rosen, and K. Smolinski (2013): "An instrumental variable model of multiple discrete choice," *Quantitative Economics*, 4(2), 157–196.

Chetty, R. (2012): "Bounds on elasticities with optimization frictions: a synthesis of micro and macro evidence in labor supply," *Econometrica*, 80(3), 969–1018.

Chetverikov, D. (2018): "Adaptive Test of Conditional Moment Inequalities," *Econometric Theory*, 34(1), 186–227.

Choquet, G. (1953/54): "Theory of capacities," *Annales de l'Institut Fourier (Grenoble)*, 5, 131–295.

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

Cross, P. J., and C. F. Manski (2002): "Regressions, Short and Long," *Econometrica*, 70(1), 357–368.

Debreu, G. (1967): "Integration of correspondences," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, vol. 2, pp. 351–372. University of California Press.

D'Haultfoeuille, X., C. Gaillac, and A. Maurel (2018): "Rationalizing Rational Expectations? Tests and Deviations," NBER working paper 25274, available at https://www.nber.org/papers/w25274.

Dickstein, M. J., and E. Morales (2018): "What do Exporters Know?," *The Quarterly Journal of Economics*, 133(4), 1753–1801.

Dominitz, J., and C. F. Manski (2017): "More Data or Better Data? A Statistical Decision Problem," *The Review of Economic Studies*, 84(4), 1583–1605.

Dominitz, J., and R. P. Sherman (2004): "Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data," *Journal of Agricultural, Biological, and Environmental Statistics*, 9(3), 319–338.

——— (2005): "Identification and estimation of bounds on school performance measures: a nonparametric analysis of a mixture model with verification," *Journal of Applied Econometrics*, 21(8), 1295–1326.

Duncan, O. D., and B. Davis (1953): "An Alternative to Ecological Correlation," *American Sociological Review*, 18(6), 665–666.

Echenique, F. (2005): "A short and constructive proof of Tarski's fixed-point theorem," *International Journal of Game Theory*, 33(2), 215–218.

Eizenberg, A. (2014): "Upstream Innovation and Product Variety in the U.S. Home PC Market," *The Review of Economic Studies*, 81(3 (288)), 1003–1045.

Ellickson, P. B., S. Houghton, and C. Timmins (2013): "Estimating network economies in retail chains: a revealed preference approach," *The RAND Journal of Economics*, 44(2), 169–193.

Epstein, L. G., H. Kaido, and K. Seo (2016): "Robust confidence regions for incomplete models," *Econometrica*, 84, 1799–1838.

Fan, Y., R. Sherman, and M. Shum (2014): "Identifying Treatment Effects Under Data Combination," *Econometrica*, 82(2), 811–822.

Fang, Z., and A. Santos (2018): "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86(1), 377–412.

Fréchet, M. R. (1951): "Sur les tableaux de correlation dont les marges sont donnes," *Annales de I'Université de Lyon A*, 3, 53–77.

Frisch, R. (1934): *Statistical Confluence Analysis by Means of Complete Regression Systems*, Okonomiske Institutt Oslo: Publikasjon. Universitetets Økonomiske Instituut.

Galichon, A. (2016): *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A., and M. Henry (2006): "Inference in Incomplete Models," available at http://dx.doi.org/10.2139/ssrn.886907.

——— (2009): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152(2), 186 – 196, Nonparametric and Robust Methods in Econometrics.

——— (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78(4), 1264–1298.

——— (2013): "Dilation bootstrap," *Journal of Econometrics*, 177(1), 109 – 115.

Gentry, M., and T. Li (2014): "Identification in auctions with selective entry," *Econometrica*, 82(1), 315–344.

Giné, E., M. G. Hahn, and J. Zinn (1983): "Limit theorems for random sets: An application of probability in banach space results," in *Probability in Banach Spaces IV*, ed. by A. Beck, and K. Jacobs, pp. 112–135, Berlin, Heidelberg. Springer Berlin Heidelberg.

Giustinelli, P., C. F. Manski, and F. Molinari (2019a): "Precise or Imprecise Probabilities? Evidence from survey response on dementia and long-term care," working paper, available at TBA.

——— (2019b): "Tail and Center Rounding of Probabilistic Expectations in the Health and Retirement Study," available at http://faculty.wcas.northwestern.edu/~cfm754/gmm_rounding.pdf.

Gourieroux, C., J. J. Laffont, and A. Monfort (1980): "Coherency Conditions in Simultaneous Linear Equation Models with Endogenous Switching Regimes," *Econometrica*, 48, 675–695.

Graham, B. S. (2015): "Methods of Identification in Social Networks," *Annual Review of Economics*, 7(1), 465–485.

——— (2019): "The Econometric Analysis of Networks," in *Handbook of Econometrics*. Elsevier.

Grant, M., and S. Boyd (2010): "CVX: Matlab Software for Disciplined Convex Programming, Version 1.21," available at http://cvxr.com/cvx.

Grieco, P. L. E. (2014): "Discrete games with flexible information structures: an application to local grocery markets," *The RAND Journal of Economics*, 45(2), 303–340.

Gualdani, C. (2019): "An Econometric Model of Network Formation with an Application to Board Interlocks Between Firms," available at http://docs.wixstatic.com/ugd/063589_b751c9f9c4e34d51b4da7ed7e007080a.pdf.

Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification testing under moment inequalities," *Economics Letters*, 99(2), 375 – 378.

Gundersen, C., B. Kreider, and J. Pepper (2012): "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis," *Journal of Econometrics*, 166(1), 79–91.

Haile, P. A., and E. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111(1), 1–51.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011): *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Hansen, L. P., J. Heaton, and E. G. J. Luttmer (1995): "Econometric Evaluation of Asset Pricing Models," *The Review of Financial Studies*, 8(2), 237–274.

Hansen, L. P., and R. Jagannathan (1991): "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy*, 99(2), 225–262.

Hausman, J. A., and W. K. Newey (2016): "Individual Heterogeneity and Average Welfare," *Econometrica*, 84(3), 1225–1248.

Heckman, J. J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(4), 931–959.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64(4), 487–535.

Heckman, J. J., and E. J. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96(8), 4730–4734.

——— (2001): "Instrumental variables, selection models, and tight bounds on the average treatment effect," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 1–15, Heidelberg. Physica-Verlag HD.

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73(3), 669–738.

——— (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, pp. 4779 – 4874. Elsevier.

——— (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, pp. 4875 – 5143. Elsevier.

Hellmann, T. (2013): "On the existence and uniqueness of pairwise stable networks," *International Journal of Game Theory*, 42(1), 211–237.

Henry, M., R. Méango, and M. Queyranne (2015): "Combinatorial approach to inference in partially identified incomplete structural models," *Quantitative Economics*, 6(2), 499–529.

Henry, M., and A. Onatski (2012): "Set coverage and robust policy," *Economics Letters*, 115(2), 256 – 257.

Hirano, K., and J. R. Porter (2019): "Statistical Decision Rules in Econometrics," in *Handbook of Econometrics*. Elsevier.

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

Ho, K., J. Ho, and J. H. Mortimer (2012): "The Use of Full-Line Forcing Contracts in the Video Rental Industry," *The American Economic Review*, 102(2), 686–719.

Ho, K., and A. Pakes (2014): "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians," *The American Economic Review*, 104(12), 3841–3884.

Ho, K., and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 1 of *Econometric Society Monographs*, pp. 307–359. Cambridge University Press.

Hoderlein, S., and J. Stoye (2014): "Revealed Preferences in a Heterogeneous Population," *Review of Economics and Statistics*, 96(2), 197–213.

Holmes, T. J. (2011): "The diffusion of Wal-mart and economies of density," *Econometrica*, 79(1), 253–302.

Honoré, B. E., and A. Lleras-Muney (2006): "Bounds in Competing Risks Models and the War on Cancer," *Econometrica*, 74(6), 1675–1698.

Honoré, B. E., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281–302.

——— (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84(1), 37 – 58.

——— (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

Hotz, V. J., C. H. Mullin, and S. G. Sanders (1997): "Bounding Causal Effects Using Data From a Contaminated Natural Experiment: Analysis the Effects of Teenage Chilbearing," *The Review of Economic Studies*, 64(4), 575–603.

Howard, J. A. (1963): *Consumer behavior: application of theory*. New York: McGraw-Hill, Includes indexes.

Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35(1), 73–101.

——— (2004): *Robust Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Iaryczower, M., X. Shi, and M. Shum (2018): "Can Words Get in the Way? The Effect of Deliberation in Collective Decision Making," *Journal of Political Economy*, 126(2), 688–734.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Imbens, G. W., and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Jackson, M. O., and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71(1), 44 – 74.

Jia, P. (2008): "What Happens When Wal-Mart Comes to Town: An Empirical Analysis of the Discount Retailing Industry," *Econometrica*, 76(6), 1263–1316.

Jones, D. R. (2001): "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization*, 21(4), 345–383.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13(4), 455–492.

Jovanovic, B. (1989): "Observable Implications of Models with Multiple Equilibria," *Econometrica*, 57(6), 1431–1437.

Juster, F. T., and R. Suzman (1995): "An Overview of the Health and Retirement Study," *Journal of Human Resources*, 30 (Supplement), S7–S56.

Kaido, H. (2016): "A dual approach to inference for partially identified econometric models," *Journal of Econometrics*, 192(1), 269 – 290.

Kaido, H., F. Molinari, and J. Stoye (2019a): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, forthcoming.

——— (2019b): "Constraint Qualifications in Partial Identification," working paper, available at TBA.

Kaido, H., F. Molinari, J. Stoye, and M. Thirkettle (2017): "Calibrated Projection in MATLAB," documentation available at https://arxiv.org/abs/1710.09707 and code available at https://github.com/MatthewThirkettle/calibrated-projection-MATLAB.

Kaido, H., and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82(1), 387–413.

Kaido, H., and H. White (2013): "Estimating Misspecified Moment Inequality Models," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, ed. by X. Chen, and N. R. Swanson, pp. 331–361, Springer, New York, NY.

Kamat, V. (2018): "Identification with Latent Choice Sets," available at https://arxiv.org/abs/1711.02048.

Kawai, K., and Y. Watanabe (2013): "Inferring Strategic Voting," *American Economic Review*, 103(2), 624–62.

Kitagawa, T. (2009): "Identification region of the potential outcome distributions under instrument independence," CeMMAP working paper CWP30/09, available at https://www.cemmap.ac.uk/wps/cwp3009.pdf.

Kitagawa, T., and R. Giacomini (2018): "Robust Bayesian inference for set-identified models," CeMMAP working paper CWP61/18, available at https://www.cemmap.ac.uk/publication/id/13675.

Kitamura, Y., and J. Stoye (2018): "Nonparametric Analysis of Random Utility Models," *Econometrica*, 86(6), 1883–1909.

——— (2019): "Nonparametric Counterfactuals in Random Utility Models," available at https://arxiv.org/abs/1902.08350.

Klepper, S., and E. E. Leamer (1984): "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica*, 52(1), 163–183.

Kline, B., and E. Tamer (2012): "Bounds for best response functions in binary games," *Journal of Econometrics*, 166(1), 92 – 105.

——— (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7(2), 329–366.

Kline, P., and M. Tartari (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106(4), 972–1014.

Komarova, T. (2013): "Partial identification in asymmetric auctions in the absence of independence," *The Econometrics Journal*, 16(1), S60–S92.

Koopmans, T. C., and O. Reiersol (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21(2), 165–181.

Kreider, B., and J. V. Pepper (2007): "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102(478), 432–441.

Kreider, B., J. V. Pepper, C. Gundersen, and D. Jolliffe (2012): "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, 107(499), 958–975.

Leamer, E. E. (1987): "Errors in Variables in Linear Systems," *Econometrica*, 55(4), 893–909.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76(3), 1071–1102.

Lee, R. S. (2013): "Vertical Integration and Exclusivity in Platform and Two-Sided Markets," *The American Economic Review*, 103(7), 2960–3000.

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing functional inequalities," *Journal of Econometrics*, 172(1), 14 – 32.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97(1), 145 – 177.

——— (2018): "The Identification Zoo - Meanings of Identification in Econometrics," *Journal of Economic Literature*, forthcoming.

Luce, R. D., and P. Suppes (1965): "Chapter 19: Preference, Utility, and Subjective Probability," in *Handbook of Mathematical Psychology*, vol. 3, pp. 249–410.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2018): "Instrumental Variables and the Sign of the Average Treatment Effect," *Journal of Econometrics*, forthcoming.

Maddala, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

Magnac, T., and E. Maurin (2008): "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data," *The Review of Economic Studies*, 75(3), 835–864.

Magnolfi, L., and C. Roncoroni (2017): "Estimation of Discrete Games with Weak Assumptions on Information," available at http://lorenzomagnolfi.com/estimdiscretegames.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), 205 – 228.

——— (1977): "The structure of random utility models," *Theory and Decision*, 8(3), 229–254.

——— (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313 – 333.

——— (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83(403), 729–738.

——— (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24(3), 343–360.

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review Papers and Proceedings*, 80(2), 319–323.

——— (1994): "The selection problem," in *Advances in Econometrics: Sixth World Congress*, ed. by C. A. Sims, vol. 1 of *Econometric Society Monographs*, pp. 143–170. Cambridge University Press.

——— (1995): *Identification Problems in the Social Sciences*. Harvard University Press.

——— (1997a): "The Mixing Problem in Programme Evaluation," *The Review of Economic Studies*, 64(4), 537–553.

——— (1997b): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

——— (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics. Springer.

——— (2005): *Social Choice with Partial Knowledge of Treatment Response*. Princeton University Press.

——— (2007a): *Identification for Prediction and Decision*. Harvard University Press.

——— (2007b): "Partial Indentification of Counterfactual Choice Probabilities," *International Economic Review*, 48(4), 1393–1410.

——— (2010): "Random Utility Models with Bounded Ambiguity," in *Structural Econometrics*, ed. by B. Dutta, pp. 272–284. Oxford University Press, 1 edn.

——— (2013): "Identification of treatment response with social interactions," *The Econometrics Journal*, 16(1), S1–S23.

——— (2014): "Identification of income–leisure preferences and evaluation of income tax policy," *Quantitative Economics*, 5(1), 145–174.

Manski, C. F., and F. Molinari (2010): "Rounding Probabilistic Expectations in Surveys," *Journal of Business and Economic Statistics*, 28(2), 219–231.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12(s1), S200–S216.

——— (2018): "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *The Review of Economics and Statistics*, 100(2), 232–244.

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

Manzini, P., and M. Mariotti (2014): "Stochastic Choice and Consideration Sets," *Econometrica*, 82(3), 1153–1176.

Markowitz, H. (1952): "Portfolio selection," *Journal of Finance*, 7, 77–91.

Marschak, J., and W. H. Andrews (1944): "Random Simultaneous Equations and the Theory of Production," *Econometrica*, 12(3/4), 143–205.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012): "Revealed Attention," *American Economic Review*, 102(5), 2183–2205.

Masten, M. A., and A. Poirier (2018): "Salvaging Falsified Instrumental Variable Models," available at https://arxiv.org/abs/1812.11598.

Matheron, G. (1975): *Random Sets and Integral Geometry*. Wiley, New York.

Matzkin, R. L. (1993): "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 58(1), 137 – 168.

——— (2007): "Chapter 73 Nonparametric identification," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, chap. 73, pp. 5307 – 5368. Elsevier.

McCarthy, I., D. L. Millimet, and M. Roy (2015): "Bounding treatment effects: A command for the partial identification of the average treatment effect with endogenous and misreported treatment assignment," *Stata Journal*, 15(2), 411–436.

McFadden, D. L. (1973): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka. Academic Press.

——— (2005): "Revealed Stochastic Preference: A Synthesis," *Economic Theory*, 26(2), 245–264.

Menzel, K. (2014): "Consistent estimation with many moment inequalities," *Journal of Econometrics*, 182(2), 329 – 350.

Milgrom, P. R., and R. J. Weber (1982): "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(5), 1089–1122.

Miyauchi, Y. (2016): "Structural estimation of pairwise stable networks with nonnegative externality," *Journal of Econometrics*, 195(2), 224 – 235.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86(5), 1589–1619.

Mogstad, M., and A. Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10(1), 577–613.

Molchanov, I. (1998): "A limit theorem for solutions of inequalities," *Scandinavian Journal of Statistics*, 25, 235–242.

Molchanov, I. (2017): *Theory of Random Sets*. Springer, London, 2 edn.

Molchanov, I., and F. Molinari (2014): "Applications of Random Set Theory in Econometrics," *Annual Review of Economics*, 6(1), 229–251.

——— (2018): *Random Sets in Econometrics*. Econometric Society Monograph Series, Cambridge University Press, Cambridge UK.

Molinari, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144(1), 81 – 117.

——— (2010): "Missing Treatments," *Journal of Business and Economic Statistics*, 28(1), 82–95.

Molinari, F., and M. Peski (2006): "Generalization of a Result on "Regressions, short and long"," *Econometric Theory*, 22(1), 159–163.

Molinari, F., and A. M. Rosen (2008): "The Identification Power of Equilibrium in Games: The Supermodular Case (Comment on Aradillas-Lopez and Tamer, 2008)," *Journal of Business and Economic Statistics*, 26(3), 297–302.

Moon, H. R., and F. Schorfheide (2012): "Bayesian and frequentist inference in partially identified models," *Econometrica*, 80(2), 755–782.

Mourifié, I., M. Henry, and R. Méango (2018): "Sharp Bounds and Testability of a Roy Model of STEM Major Choices," available at https://ssrn.com/abstract=2043117.

Neyman, J. S. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Roczniki Nauk Rolniczych*, X, 1–51, reprinted in *Statistical Science*, 5(4), 465-472, translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original.

Norberg, T. (1992): "On the existence of ordered couplings of random sets — with applications," *Israel Journal of Mathematics*, 77, 241–264.

Okner, B. (1972): "Constructing A New Data Base From Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, 1(3), 325–362.

Pacini, D. (2017): "Two-sample least squares projection," *Econometric Reviews*, 0(0), 1–29.

Pakes, A. (2010): "Alternative models for moment inequalities," *Econometrica*, 78(6), 1783–1822.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015): "Moment Inequalities and Their Application," *Econometrica*, 83(1), 315–334.

de Paula, A. (2013): "Econometric Analysis of Games with Multiple Equilibria," *Annual Review of Economics*, 5(1), 107–131.

——— (2017): "Econometrics of Network Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 1 of *Econometric Society Monographs*, p. 268–323. Cambridge University Press.

de Paula, A., S. Richards-Shubik, and E. Tamer (2018): "Identifying Preferences in Networks With Bounded Degree," *Econometrica*, 86(1), 263–288.

de Paula, A., and X. Tang (2012): "Inference of Signs of Interaction Effects in Simultaneous Games With Incomplete Information," *Econometrica*, 80(1), 143–172.

Peterson, A. V. (1976): "Bounds for a Joint Distribution Function with Fixed Sub-Distribution Functions: Application to Competing Risks," *Proceedings of the National Academy of Sciences of the United States of America*, 73(1), 11–13.

Petrin, A., and K. Train (2010): "A Control Function Approach to Endogeneity in Consumer Choice Models," *Journal of Marketing Research*, 47(1), 3–13.

Phillips, P. C. B. (1989): "Partially Identified Econometric Models," *Econometric Theory*, 5(2), 181–240.

Picketty, T. (2005): "Top Income Shares in the Long Run: An Overview," *Journal of the European Economic Association*, 3, 382–392.

Ponomareva, M., and E. Tamer (2011): "Misspecification in moment inequality models: back to moment equalities?," *The Econometrics Journal*, 14(2), 186–203.

Redner, R. (1981): "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions," *The Annals of Statistics*, 9(1), 225–228.

Reiersol, O. (1941): "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, 9(1), 1–24.

Ridder, G., and R. Moffitt (2007): "Chapter 75 The Econometrics of Data Combination," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, pp. 5469 – 5547. Elsevier.

Rockafellar, R. (1970): *Convex Analysis*, Princeton landmarks in mathematics and physics. Princeton University Press.

Romano, J. P., and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786 – 2807.

——— (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A practical two-step method for testing moment inequalities," *Econometrica*, 82(5), 1979–2002.

Rosen, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146(1), 107 – 117.

——— (2012): "Set identification via quantile restrictions in short panels," *Journal of Econometrics*, 166(1), 127 – 137.

Rubin, D. B. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6(1), 34–58.

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80(1), 213–275.

Schennach, S. M. (2019): "Mismeasured and unobserved variables," in *Handbook of Econometrics*. Elsevier.

Schmidt, P. (1981): "Constraints on the Parameters in Simultaneous Tobit and Probit Models," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski, and D. McFadden, chap. 12, pp. 422–434. MIT Press.

Schneider, R. (1993): *Convex Bodies: The Brunn-Minkowski Theory*, Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1 edn.

Schonlau, M., W. J. Welch, and D. R. Jones (1998): "Global versus Local Search in Constrained Optimization of Computer Models," *Lecture Notes-Monograph Series*, 34, 11–25.

Shaikh, A. M., and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79(3), 949–955.

Sheng, S. (2018): "A structural econometric analysis of network formation games through subnetworks," *Econometrica*, accepted for publication.

Simon, H. A. (1959): "Theories of Decision-Making in Economics and Behavioral Science," *The American Economic Review*, 49(3), 253–283.

Sims, C. A. (1972): "Comments and Rejoinder On Okner (1972)," *Annals of Economic and Social Measurement*, 1(3), 343–345 and 355–357.

Stoye, J. (2007): "Bounds on Generalized Linear Predictors with Incomplete Outcome Data," *Reliable Computing*, 13(3), 293–302.

——— (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77(4), 1299–1315.

——— (2010): "Partial identification of spread parameters," *Quantitative Economics*, 1(2), 323–357.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70(1), 147–165.

——— (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

Tang, X. (2011): "Bounds on revenue distributions in counterfactual auctions with reserve prices," *The RAND Journal of Economics*, 42(1), 175–203.

Tauchmann, H. (2014): "Lee (2009) treatment-effect bounds for nonrandom sample selection," *Stata Journal*, 14(4), 884–894.

Torgovitsky, A. (2019a): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, forthcoming.

——— (2019b): "Partial identification by extending subdistributions," *Quantitative Economics*, 10(1), 105–144.

Tversky, A. (1972): "Elimination by aspects: A theory of choice," *Psychological review*, 79(4), 281.

van der Vaart, A. (1997): "Superefficiency," in *Festschrift for Lucien Le Cam*, ed. by D. Pollard, E. Torgersen, and G. L. Yang, chap. 27, pp. 397–410. Springer.

Wollmann, T. G. (2018): "Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," *American Economic Review*, 108(6), 1364–1406.