Journal of Econometrics 166 (2012) 17–32

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Partial identification using random set theory☆

Arie Beresteanu [a], Ilya Molchanov [b], Francesca Molinari [c,*]

[a] Department of Economics, University of Pittsburgh, United States

[b] Department of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland

[c] Department of Economics, Cornell University, United States

ARTICLE INFO

Article history:
Available online 23 June 2011

ABSTRACT

This paper illustrates how the use of random set theory can benefit partial identification analysis. We revisit the origins of Manski's work in partial identification (e.g., Manski (1989, 1990)) focusing our discussion on identification of probability distributions and conditional expectations in the presence of selectively observed data, statistical independence and mean independence assumptions, and shape restrictions. We show that the use of the Choquet capacity functional and the Aumann expectation of a properly defined random set can simplify and extend previous results in the literature. We pay special attention to explaining how the relevant random set needs to be constructed, depending on the econometric framework at hand. We also discuss limitations in the applicability of specific tools of random set theory to partial identification analysis.

1. Introduction

Overview. Partial identification predicates that econometric analysis should include the study of the *set* of values for a parameter vector (or statistical functional) of interest which are observationally equivalent, given the available data and *credible* maintained assumptions. We refer to this set as the parameter vector's *sharp identification region*.[1] This principle is perhaps best summarized in Manski's (2003) monograph on *Partial Identification of Probability Distributions*, where he states: "It has been commonplace to think of identification as a binary event – a parameter is either identified or it is not – and to view point identification as a precondition for meaningful inference. Yet there is enormous scope for fruitful inference using data and assumptions that partially identify population parameters" (p. 3). Following this basic principle, partial identification analysis, whether applied for prediction or for decision making, aims at: (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 it.

While conceptually these aims imply a fundamental shift of focus from single valued to set valued objects, in practice they have been implemented using "standard" mathematical tools, such as probability distributions, conditional and unconditional expectations, laws of large numbers and central limit theorems for (single valued) random vectors. This approach has been very productive in many contexts; see, for example, Manski (1995, 2007) and Haile and Tamer (2003) for results on identification, and Imbens and Manski (2004), Chernozhukov et al. (2007), Stoye (2009) and Andrews and Soares (2010) for results on statistical inference. However, certain aspects of the study of identification and statistical inference in partially identified models can substantially benefit from, and be simplified by, the use of mathematical tools borrowed from the *theory of random sets* (Molchanov, 2005). This literature originated in the seminal contributions of Choquet (1953–1954), Aumann (1965) and Debreu (1967), and its first self-contained treatment was given by Matheron (1975). It has been an area intensely researched in mathematics and probability ever since.

The applicability of random set theory to partial identification is due to the fact that partially identified models are often characterized by a collection of random outcomes (or covariates) which are consistent with the data and the maintained

☆ This paper was prepared for the Northwestern University/CeMMAP conference, *Identification and Decisions*, in honour of Chuck Manski on his 60th birthday, held at Northwestern University in May 2009. We thank the seminar participants there, at Penn State, UCL, the ZEW Workshop "Measurement Errors in Administrative Data", the 2010 European Meetings of Statisticians, Adam Rosen, Joerg Stoye, two anonymous referees, and a guest co-editor for comments that helped us to improve this paper significantly. We are grateful to Darcy Steeg Morris for excellent research assistance. Beresteanu gratefully acknowledges financial support from the NSF through Grants SES-0617559 and SES-0922373. Molchanov gratefully acknowledges financial support from the Swiss National Science Foundation Grants No. 200021-117606 and No. 200021-126503. Molinari gratefully acknowledges financial support from the NSF through Grants SES-0617482 and SES-0922330.

* Corresponding author.
E-mail addresses: arie@pitt.edu (A. Beresteanu), ilya@stat.unibe.ch (I. Molchanov), fm72@cornell.edu (F. Molinari).

[1] This region contains all the parameters' values that could generate the same distribution of observables as the one in the data, for some data generating process consistent with all the maintained assumptions, and no other values.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.06.003

assumptions. To fix ideas, suppose that one wants to learn a feature of the distribution of an outcome variable y conditional on covariates w. Let w be perfectly observed and y be interval measured, with $\mathbf{P}(y \in [y_L, y_U]) = 1$. In the absence of assumptions on how y is selected from $[y_L, y_U]$, the distribution $\mathbf{P}(y|w)$ is partially identified. The collection of random variables $\tilde{y}$ such that $\mathbf{P}(\tilde{y} \in [y_L, y_U]) = 1$, paired with w, gives all the random elements that are consistent with the data and the maintained assumptions; hence, the collection of random elements which are observationally equivalent. In the language of random set theory, these random elements constitute the *family of selections* of a properly specified random closed set; in this example, $[y_L, y_U] \times w$.[2] Depending on the specific econometric model at hand, different features of the observationally equivalent random elements might be of interest; for example, their distributions or their expectations. Random set theory provides probability "distributions" (capacity functionals) and conditional and unconditional (Aumann) "expectations" for random sets, which can be employed to learn the corresponding features of interest for the family of their selections, and hence for the observationally equivalent random elements of interest. The main task left to the researcher is to judiciously construct the relevant random set to which these tools need to be applied. In turn, this leads to characterizing the sharp identification region of a model's parameters in the space of sets, in a manner which is the exact analog of how point-identification arguments are constructed for point identified parameters in the space of vectors. Laws of large numbers and central limit theorems for random sets can then be used to conduct statistical inference, again in a manner which is the exact analog in the space of sets of how statistical inference is conducted for point identified parameters in the space of vectors.

The fundamental goal of this paper is to explain when and how the theory of random sets can be useful for partial identification analysis. In order to make our discussion as accessible as possible, and relate it to the origins of Manski's work on the topic (e.g., Manski (1989, 1990)), we focus our analysis on identification in the presence of interval outcome data, paying special attention to the selection problem. Statistical considerations can be addressed using the methodologies provided by Beresteanu and Molinari (2008), Galichon and Henry (2009b), Chernozhukov et al. (2007, 2009), Andrews and Shi (2009) and Andrews and Soares (2010), among others, as we discuss in Section 4 below. Some of the results that we report have already been derived by other researchers (specifically, the results in Proposition 2.2, part of 2.4, 3.2, C.2 and C.3). We rederive these basic results, as this helps make plain the connection between random set theory and standard approaches to partial identification. We then provide a number of novel results which are simple extensions of these basic findings, if derived using random set theory, but would not be as easy to obtain if using standard techniques, thereby showcasing the usefulness of our approach (specifically, the results novel to this paper appear in Proposition 2.3, part of 2.4, 2.5, 2.6, 3.3, C.1 and C.4). We also pay special attention to explaining how the relevant random closed set needs to be defined, depending on the econometric framework at hand. As it turns out, this boils down to the same careful exercise in deductive logic, based on the maintained assumptions and the available data, which characterizes all partial identification analysis. Finally, we discuss limitations in the applicability of random set theory to partial identification.

Related Literature applying random sets theory in econometrics. While sometimes applied in microeconomics, the theory of random sets has not been introduced in econometrics until recently. The first systematic use of tools from this literature in partial identification analysis appears in Beresteanu and Molinari (2006, 2008). They study a class of partially identified models in which the sharp identification region of the parameter vector of interest can be written as a transformation of the Aumann expectation of a properly defined random set. For this class of models, they propose to use the sample analog estimator given by a transformation of a Minkowski average of properly defined random sets. They use limit theorems for independent and identically distributed sequences of random sets, to establish consistency of this estimator with respect to the Hausdorff metric. They propose two Wald-type test statistics, based on the Hausdorff metric and on the lower Hausdorff hemimetric, to test hypothesis and make confidence statements about the entire sharp identification region and its subsets. And they introduce the notion of "confidence collection" for partially identified parameters as a counterpart to the notion of confidence interval for point identified parameters.

General results for identification analysis are given by Beresteanu et al. (2008, 2009, in press), who provide a tractable characterization of the sharp identification region of the parameters characterizing incomplete econometric models with convex moment predictions. Examples of such models include static, simultaneous move finite games of complete and incomplete information in the presence of multiple equilibria; random utility models of multinomial choice in the presence of interval regressors data; and best linear predictors with interval outcome and covariate data. They show that algorithms in convex programming can be exploited to efficiently verify whether a candidate parameter value is in the sharp identification region. Their results are based on an array of tools from random set theory, ranging from conditional Aumann expectations, to capacity functionals, to laws of large numbers and central limit theorems for random closed sets.

Galichon and Henry (2006, 2009b) provide a specification test for partially identified structural models. In particular, they use a result due to Artstein (1983), discussed in Section 2 below, to conclude that the model is correctly specified if the distribution of the observed outcome is dominated by the Choquet capacity functional of the random correspondence between the latent variables and the outcome variables characterizing the model. This allows them to extend the Kolmogorov–Smirnov test of correct model specification to partially identified models. They then define the notion of "core determining" classes of sets, to find a manageable class of sets for which to check that the dominance condition is satisfied. They also introduce an equivalent formulation of the notion of a correctly specified partially identified structural model, based on optimal transportation theory, which provides computational advantages for certain classes of models.[3]

Structure of the paper. In Section 2 we address the problem of characterizing the sharp identification region of probability distributions from selectively observed data, when the potential outcome of interest is statistically independent from an instrument, and when it satisfies certain shape restrictions. In doing so, we extend the existing literature by allowing the instrument to have a continuous distribution, by allowing for more than two treatments, and by deriving sharp identification regions for the entire response function both under independence assumptions and shape restrictions. The fundamental tool from random set theory used for this analysis is the capacity functional (probability distribution) of a properly specified random set. In Section 3 we address the problem of characterizing the sharp identification region of conditional expectations from selectively observed data, in the presence of mean

[2] We formally define the family of selections of a random closed set in Appendix A.

[3] For example, this occurs in finite static games of complete information where players use only pure strategies and certain monotonicity conditions are satisfied.

independence assumptions and shape restrictions. We also discuss best linear prediction, and provide a number of novel results of practical use, concerning the implications of affine transformations of covariate data (e.g., demeaning and rescaling) for the characterization of the sharp identification region of parameters of interest. The fundamental tools from random set theory used for this analysis is the Aumann expectation of a properly defined random set and its support function.

In Section 4 we outline how to estimate the sharp identification regions and conduct statistical inference. In Section 5 we discuss the issue of how one should choose whether to use the capacity functional or the Aumann expectation as the main tool to address a specific partial identification problem. Section 6 concludes. Appendix A provides basic definitions. Appendix B provides a few auxiliary Lemmas. Appendix C provides sharp identification regions for the distribution and the expectation of the response function under independence and shape restrictions.

Notation. Throughout the paper, we use capital Latin letters to denote sets and random sets.[4] We use lower case Latin letters for random vectors. We denote parameter vectors and sets of parameter vectors, respectively by θ and Θ. We let $(\Omega, \mathfrak{F}, \mathbf{P})$ denote a nonatomic probability space on which all random variables and random sets are defined.[5] We denote the Euclidean space by $\Re^d$, and equip it with the Euclidean norm (which is denoted by $\|\cdot\|$). The theory of random closed sets generally applies to the space of closed subsets of a locally compact Hausdorff second countable topological space $\mathbb{F}$, see Molchanov (2005). For the purposes of this paper it suffices to consider $\mathbb{F} = \Re^d$, which simplifies the exposition. Denote by $\mathcal{F}$ and $\mathcal{K}$, respectively, the collection of closed subsets and compact subsets of $\Re^d$. Given a set $A \subset \Re^d$, let $\mathrm{co}(A)$ denote its convex hull.

2. Usefulness of the capacity functional

2.1. Capacity functional and Artstein's inequality

Consider cases in which all the information provided by the empirical evidence and the maintained assumptions can be expressed by saying that a random vector x belongs to a properly specified *random set* X (see Definition A.1 in Appendix A) in the sense that $\mathbf{P}(x \in X) = 1$. This happens, for example, when we observe interval data. In this case the researcher is interested in a variable x which is only known to lie in an interval $X = [x_L, x_U]$, with $\mathbf{P}(x \in X) = 1$. In other words, the unobserved variable of interest is a *selection* of the observed random set X (see Definition A.2 in Appendix A). In order to utilize the information embodied in the statement that $\mathbf{P}(x \in X) = 1$, one needs to be able to relate features of the random set to corresponding features of its selections.[6]

A fundamental result in random set theory, due to Artstein (1983) and Norberg (1992), provides a necessary and sufficient condition for $\mathbf{P}(x \in X) = 1$, which relates the distribution of the random vector x to the *capacity functional* of the random set X.[7] The capacity functional is a subadditive measure which uniquely determines the distribution of a random closed set by giving the probability that the random set hits a given compact set, see Definition A.3 in Appendix A. In what follows, let "$x \overset{d}{\sim} x'$" ("$X \overset{d}{\sim} X'$") denote that two random vectors (sets) are equivalent in distribution.

Theorem 2.1 (*Artstein's Inequality*)**.** *A random vector x and a random set X can be realized on the same probability space as random elements x' and X', with $x' \overset{d}{\sim} x$ and $X' \overset{d}{\sim} X$, so that $\mathbf{P}(x' \in X') = 1$, if and only if*

$$\mathbf{P}(x \in K) \leq \mathbf{P}(X \cap K \neq \emptyset) \equiv \mathbf{T}_X(K) \quad \forall K \in \mathcal{K}. \tag{2.1}$$

Equivalently, if and only if

$$\mathbf{P}(x \in K) \geq \mathbf{P}(X \subset K) \equiv \mathbf{C}_X(K) \quad \forall K \in \mathcal{K}. \tag{2.2}$$

When condition (2.1) is satisfied, we say that x is stochastically smaller than X.[8]

Proof. The proof of this result for the capacity functional, i.e., for condition (2.1), can be found in Molchanov (2005, Corollary 1.4.44). Here we provide an argument for the equivalence between condition (2.1) and condition (2.2). Consider $K \in \mathcal{K}$. Its complement K^c can be approximated from below by a sequence of compact sets $\{K_n\}$, i.e. $K_n \uparrow K^c$. By condition (2.1),

$$\mathbf{P}(x \in K_n) \leq \mathbf{P}(X \cap K_n \neq \emptyset), \quad n \geq 1.$$

By passing to the limit as $n \to \infty$ and using the continuity of probability from below, we arrive at

$$\mathbf{P}(x \in K^c) \leq \mathbf{P}(X \cap K^c \neq \emptyset).$$

By the relationship between capacity functional and containment functional (see Eq. (A.1) in Appendix A), the above can be rephrased as

$$1 - \mathbf{P}(x \in K) \leq 1 - \mathbf{P}(X \subset K)$$

yielding exactly the dominance condition for the containment functional in (2.2). The reversed implication is similar. □

Intuition for the capacity functional dominance condition. The nature of the domination condition in inequality (2.1) can be traced to the ordering – or first order stochastic dominance – concept for random variables. Namely, a random variable x is said to be stochastically smaller than a random variable y if $\mathbf{P}(x \leq t) \geq \mathbf{P}(y \leq t)$ for all $t \in \Re$; in other words, if the cumulative distribution function of x dominates that of y. When this is the case, x and y can be realized on the same probability space as random variables $x' \overset{d}{\sim} x$ and $y' \overset{d}{\sim} y$, such that $x' \leq y'$ almost surely. This is referred to as the *ordered coupling for random variables* x and y. The stochastic dominance condition can be written also as $\mathbf{P}(x \in A) \leq \mathbf{P}(y \in A)$ for $A = [t, \infty)$ and all $t \in \Re$. Such a set A is increasing (or upper), i.e. $x \in A$ and $x \leq y$ implies $y \in A$. Using the probabilities of upper sets, this domination condition can be extended to any partially ordered space. In particular, this leads to the condition for the *ordered coupling for random closed sets* Z and X obtained by Norberg (1992); see also Molchanov (2005, Section 1.4.8). Two random

[4] The notations $\mathbf{P}$ and $\mathbf{E}$ are reserved to the probability measure on the sample space and the expectation operator taken with respect to this probability measure.

[5] Similar results to those reported here apply for the case of atomic probability spaces, see Molchanov (2005). We restrict attention to the nonatomic case to simplify the exposition, and because when one considers a sequence of i.i.d. random elements, the appropriate (product) probability space is always nonatomic.

[6] In other partial identification problems, such as for example static discrete games of complete information in the presence of multiple *pure* strategy Nash equilibria, the model predicts a random closed set of equilibrium outcomes Y. The econometrician observes an equilibrium outcome y which, if the model is correctly specified, satisfies $\mathbf{P}(y \in Y) = 1$, see Beresteanu et al. (2008).

[7] Beresteanu and Molinari (2006, 2008, Proposition 4.1) use this result to establish sharpness of the identification region of the parameters of a best linear predictor with interval outcome data. Galichon and Henry (2006) use it to define a correctly specified partially identified structural model, and derive a Kolmogorov–Smirnov test for Choquet capacities.

[8] In the statement of Artstein's inequality, compact sets $K \in \mathcal{K}$ can be replaced by closed sets $F \in \mathcal{F}$.

closed sets Z and X can be realized on the same probability space as random sets $Z' \overset{d}{\sim} Z$ and $X' \overset{d}{\sim} X$ and so that $Z' \subset X'$ almost surely, if and only if the probabilities that Z has non-empty intersection with each set from $K_1, \ldots, K_n$, $n \geq 1$, are dominated by those of X. If Z is a singleton, say $Z = \{x\}$, this condition can be substantially simplified and reduces to the one in inequality (2.1). □

In all that follows, to simplify the exposition, we refer to Artstein's inequality as a necessary and sufficient condition for $\mathbf{P}(x \in X) = 1$, with the understanding that such statement is meant up to an ordered coupling. We denote by $\mathrm{Sel}(X)$ the set of random elements x such that $x(\omega) \in X(\omega)$ $\mathbf{P}$-*a.s.*, see Definition A.2 in Appendix A. Let $\mathbb{P}_X$ denote the family of all probability measures μ_x that are dominated by $\mathbf{T}_X$, or equivalently that dominate $\mathbf{C}_X$:

$$\begin{aligned}\mathbb{P}_X &= \{\mu_x : \mu_x(K) \leq \mathbf{T}_X(K)\ \forall K \in \mathcal{K}\} \\ &= \{\mu_x : \mu_x(K) \geq \mathbf{C}_X(K)\ \forall K \in \mathcal{K}\}. \end{aligned} \tag{2.3}$$

Then the capacity functional equals the upper envelope of all probability measures that it dominates, and the containment functional equals the lower envelope of all probability measures that dominate it, see Molchanov (2005, Theorem 1.5.13):

$$\begin{aligned}\mathbf{T}_X(K) &= \sup\{\mu_x(K) : \mu_x \in \mathbb{P}_X\}, \quad K \in \mathcal{K}, \\ \mathbf{C}_X(K) &= \inf\{\mu_x(K) : \mu_x \in \mathbb{P}_X\}, \quad K \in \mathcal{K}.\end{aligned}$$

2.2. Conditional distributions and the selection problem

In this Section we illustrate how the use of the capacity functional, and in particular the application of Theorem 2.1, can simplify the task of finding the sharp identification region for probability distributions of interest, in the presence of selectively observed data, statistical independence assumptions, and shape restrictions. This problem is discussed, for example, in Manski (2003, Chapters 7 and 8), where several findings are reported. It is especially suited to explain the usefulness of the capacity functional in partial identification, because: (1) the relevant random sets to which Artstein's inequality needs to be applied have been derived by Manski, see for example Manski (1989, Eq. (3)) and Manski (2003, Proposition 8.1), and are of familiar use in partial identification[9]; and (2) statistical independence assumptions directly constrain the probability distributions of selections of these random sets, and are therefore easy to couple with Artstein's inequality.[10]

2.2.1. Basic setup and worst-case analysis

Using standard notation (e.g., Neyman (1923)), let $\mathcal{T} = \{0, \ldots, T\}$ denote a set of mutually exclusive and exhaustive treatments, let $w \in \mathcal{W}$ denote some covariates, and let $y(\cdot) : \mathcal{T} \to \mathcal{Y}$ denote a response function mapping treatments $t \in \mathcal{T}$ into outcomes $y(t) \in \mathcal{Y}$, with $\mathcal{Y}$ a compact set in $\Re$. Without loss of generality assume $\min \mathcal{Y} = 0$, and $\max \mathcal{Y} = 1$. Let $z \in \mathcal{T}$ denote the received treatment. The object of interest is to learn the probability distribution of the potential outcomes given covariates w, $\mathbf{P}(y(t)|w)$, $t \in \mathcal{T}$, and the probability distribution of the response function given covariates w, $\mathbf{P}(y(\cdot)|w)$. The identification problem arises because while for $t = z$ the outcome $y(t) \equiv y(z) \equiv y$ is realized and observable, for $t \neq z$ the outcome $y(t)$ is counterfactual and unobservable. Let the tuple $(y(\cdot), z, w)$ be defined on $(\Omega, \mathfrak{F}, \mathbf{P})$, and let the researcher observe (y, z, w). To simplify the exposition, we henceforth leave implicit the conditioning on w.

Manski (2003, Eq. (7.2)) characterizes the sharp identification region for $\mathbf{P}(y(t))$ as follows:

$$\mathrm{H}[\mathbf{P}(y(t))] = \left\{\mathbf{P}(y|z=t)\,\mathbf{P}(z=t) + \gamma\,\mathbf{P}(z \neq t),\ \gamma \in \Gamma_{\mathcal{Y}}\right\} \tag{2.4}$$

with $\Gamma_{\mathcal{Y}}$ denoting the collection of all probability measures on $\mathcal{Y}$. Here we provide an equivalent characterization, using Artstein's inequality.

Construction of the relevant random set for $y(t)$

The data alone reveal that $y(t) = y$ if $t = z$ and $y(t) \in \mathcal{Y}$ for $t \neq z$, $t \in \mathcal{T}$. Hence, for each $t \in \mathcal{T}$, *all* the information embodied in the data can be expressed by stating that $y(t) \in \mathrm{Sel}(Y(t))$, with

$$Y(t) = \begin{cases} \{y\} & \text{if } z = t, \\ \mathcal{Y} & \text{if } z \neq t. \end{cases} \tag{2.5}$$

This is the simplest example of how a random closed set can be constructed, which collects all the information given by the data and the maintained assumptions.

Characterization of the sharp identification region of $\mathbf{P}(y(t))$

Let $\mathcal{K}(\mathcal{Y})$ denote the family of compact subsets of $\mathcal{Y}$. The sharp identification region of $\mathbf{P}(y(t))$ can be obtained applying Artstein's inequality:

Proposition 2.2. *The sharp identification region for $\mathbf{P}(y(t))$ is given by*

$$\begin{aligned}\mathrm{H}[\mathbf{P}(y(t))] = \{&\mu \in \Gamma_{\mathcal{Y}} : \mu(K) \\ &\geq \mathbf{P}(y \in K|z=t)\,\mathbf{P}(z=t)\ \forall K \in \mathcal{K}(\mathcal{Y})\}.\end{aligned} \tag{2.6}$$

If $\mathcal{Y}$ is finite,

$$\begin{aligned}\mathrm{H}[\mathbf{P}(y(t))] = \{&\mu \in \Gamma_{\mathcal{Y}} : \mu(k) \\ &\geq \mathbf{P}(y=k|z=t)\,\mathbf{P}(z=t)\ \forall k \in \mathcal{Y}\}.\end{aligned}$$

If $\mathcal{Y} = [0, 1]$,

$$\begin{aligned}\mathrm{H}[\mathbf{P}(y(t))] = \{&\mu \in \Gamma_{\mathcal{Y}} : \mu([k_1, k_2]) \\ &\geq \mathbf{P}(y \in [k_1, k_2]|z=t)\mathbf{P}(z=t)\ \forall k_1, k_2 \in \mathcal{Y} : k_1 \leq k_2\}.\end{aligned}$$

Proof. By Theorem 2.1, $y(t) \in \mathrm{Sel}(Y(t))$ if and only if $\mathbf{P}(y(t) \in K) \geq \mathbf{C}_{Y(t)}(K)\ \forall K \in \mathcal{K}(\mathcal{Y})$. Simple algebra gives $\mathbf{C}_{Y(t)}(\mathcal{Y}) = 1$ and

$$\begin{aligned}&\mathbf{C}_{Y(t)}(K) = \mathbf{P}(y \in K|z=t)\,\mathbf{P}(z=t) \\ &\forall K \in \mathcal{K}(\mathcal{Y}) \text{ such that } K \neq \mathcal{Y}.\end{aligned}$$

If $\mathcal{Y}$ is a finite set, then Lemma B.1 guarantees that it suffices to check the containment functional dominance condition for all singleton sets $K = \{k\} \subset \mathcal{Y}$. If $\mathcal{Y} = [0, 1]$, $Y(t)$ is a random closed convex set, and Lemma B.2 in the Appendix guarantees that it suffices to check the containment functional dominance condition for sets $K \in \mathcal{K}(\mathcal{Y})$ which are intervals.

To see that this characterization is equivalent to the one in Eq. (2.4), let

$$\mathbb{P}_{Y(t)} = \{\mu \in \Gamma_{\mathcal{Y}} : \mu(K) \geq \mathbf{P}(y \in K|z=t)\mathbf{P}(z=t)\ \forall K \in \mathcal{K}(\mathcal{Y})\}.$$

Take a probability measure $\mu \in \mathrm{H}[\mathbf{P}(y(t))]$ as defined in Eq. (2.4). Then $\mu = \mathbf{P}(y|z=t)\,\mathbf{P}(z=t) + \gamma\mathbf{P}(z \neq t)$, for some $\gamma \in \Gamma_{\mathcal{Y}}$. Hence, for any $K \in \mathcal{K}(\mathcal{Y})$, $K \neq \mathcal{Y}$ (the inequality is trivially satisfied for $K = \mathcal{Y}$),

$$\begin{aligned}\mu(K) &= \mathbf{P}(y \in K|z=t)\,\mathbf{P}(z=t) + \gamma(K)\,\mathbf{P}(z \neq t) \\ &\geq \mathbf{P}(y \in K|z=t)\,\mathbf{P}(z=t) = \mathbf{C}_{Y(t)}(K),\end{aligned}$$

[9] Manski did not use the language of random sets. However, his analysis in Manski (1989, 1997) effectively gives the random sets which collect all the information provided by the data and the maintained assumptions, as we show below.

[10] Our formal results are written using the containment functional, as this allows us to easily characterize the class of sets for which Artstein's inequality has to be satisfied. In view of Eq. (A.1), this is equivalent to using the capacity functional.

and therefore $\mu \in \mathbb{P}_{Y(t)}$. Conversely, take a probability measure $\mu \in \mathbb{P}_{Y(t)}$. Let

$$\gamma(K) = \frac{\mu(K) - \mathbf{P}(y \in K | z = t)\,\mathbf{P}(z = t)}{\mathbf{P}(z \neq t)}.$$

Then γ is a probability measure on $\mathcal{Y}$ and therefore $\mu \in \mathrm{H}[\mathbf{P}(y(t))]$. □

Remark 1. When $\mathcal{Y}$ is a finite set, Proposition 2.2 shows that it suffices to check the containment functional dominance condition only for singletons $k \in \mathcal{Y}$. This is because the realizations of $Y(t)$ are either singletons, or the entire space $\mathcal{Y}$. Beresteanu et al. (2009, Appendix B) discuss general cases where a random set X defined on a finite space $\mathcal{X}$ takes on realizations which are proper subsets of $\mathcal{X}$ but not singletons. In these cases, one needs to check the containment functional dominance condition also for subsets of $\mathcal{X}$ which are not singletons.

Construction of the relevant random set for $y(\cdot)$

The data alone reveals that the vector $[y(0), y(1), \ldots, y(T)]$ (i.e., the response function $y(\cdot)$) has its t-th component, $t \in \mathcal{T}$, equal to y if $z = t$, and a member of $\mathcal{Y}$ otherwise. Hence, *all* the information embodied in the data can be expressed by stating that $y(\cdot) \in \mathrm{Sel}\left(Y^{\mathcal{T}}\right)$, with

$$Y^{\mathcal{T}} = \times_{t=0}^{T} Y(t). \tag{2.7}$$

Characterization of the sharp identification region of $\mathbf{P}(y(\cdot))$

Let $\mathcal{Y}^{\mathcal{T}}$ denote the Cartesian product $\mathcal{Y} \times \mathcal{Y} \times \cdots \times \mathcal{Y}$. Let $\mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)$ denote the family of compact subsets of $\mathcal{Y}^{\mathcal{T}}$. Let $\Gamma_{\mathcal{Y}^{\mathcal{T}}}$ denote the space of all probability measures on $\mathcal{Y}^{\mathcal{T}}$. Then we have the following result:

Proposition 2.3. *The sharp identification region for $\mathbf{P}(y(\cdot))$ is given by*

$$\mathrm{H}[\mathbf{P}(y(\cdot))] = \left\{\mu \in \Gamma_{\mathcal{Y}^{\mathcal{T}}} : \mu(K) \geq \mathbf{C}_{Y^{\mathcal{T}}}(K)\ \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)\right\}.$$

If $\mathcal{Y} = [0, 1]$, it suffices to check the above condition for sets $\tilde{K} = \mathrm{co}(\tilde{K}(0) \cup \tilde{K}(1) \cup \cdots \cup \tilde{K}(T))$, where for $t \in \mathcal{T}$ either $\tilde{K}(t) = \emptyset$ or $\tilde{K}(t) = \mathcal{Y} \times \cdots \times \mathcal{Y} \times \left[k_1^t, k_2^t\right] \times \mathcal{Y} \cdots \times \mathcal{Y}$, $k_1^t \leq k_2^t$, $k_1^t, k_2^t \in \mathcal{Y}$, $t \in \mathcal{T}$. For these sets, $\mathbf{C}_{Y^{\mathcal{T}}}\left(\tilde{K}\right) = \sum_{t \in \mathcal{T}: \tilde{K}(t) \neq \emptyset} \mathbf{P}(y \in [k_1^t, k_2^t] | z = t)\mathbf{P}(z = t)$.

Proof. By Theorem 2.1, $y(\cdot) \in \mathrm{Sel}\left(Y^{\mathcal{T}}\right)$ if and only if

$$\mathbf{P}(y(\cdot) \in K) \geq \mathbf{P}\left(Y^{\mathcal{T}} \subset K\right)\ \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right). \tag{2.8}$$

If $\mathcal{Y} = [0, 1]$, by Lemma B.2 it suffices to check the above inequality for convex sets $K \subset \mathcal{Y}^{\mathcal{T}}$. Observe that if more than one of the projections of K on the axes is a proper subset of $\mathcal{Y}$, then $\mathbf{P}\left(Y^{\mathcal{T}} \subset K\right) = 0$ and inequality (2.8) is trivially satisfied. For sets $K \subset \mathcal{Y}^{\mathcal{T}}$ such that their projection on all but at most one of the axis is equal to $\mathcal{Y}$, the convexity of K implies that the set of all k such that $\mathcal{Y} \times \cdots \mathcal{Y} \times \{k\} \times \mathcal{Y} \times \cdots \mathcal{Y} \subset K$ (with $\{k\}$ occupying the t-th place) is an interval denoted by $[k_1^t, k_2^t]$ as per the definition of $\tilde{K}(t)$. The convexity of K also implies that the corresponding set $\tilde{K}$ introduced in the statement of the theorem is such that $\tilde{K} \subset K$. Finally note that $Y^{\mathcal{T}} \subset K$ if and only if $Y^{\mathcal{T}}$ is a subset of $\tilde{K}(t)$ for some $t \in \mathcal{T}$. This is because the realizations of $Y^{\mathcal{T}}$ are the Cartesian product of copies of $\mathcal{Y}$ and a point in one specific position. Moreover, $\mathbf{P}(y(\cdot) \in K) \geq \mathbf{P}\left(y(\cdot) \in \tilde{K}\right)$, hence if inequality (2.8) is satisfied for $\tilde{K}$, it is satisfied also for K. For such sets $\tilde{K}$,

$$\mathbf{P}(Y^{\mathcal{T}} \subset \tilde{K}) = \sum_{t \in \mathcal{T}: \tilde{K}(t) \neq \emptyset} \mathbf{P}\left(y \in \left[k_1^t, k_2^t\right] | z = t\right)\mathbf{P}(z = t). \quad \square$$

Remark 2 (*Binary Outcomes and Fréchet Bounds*)**.** Consider the special case in which $\mathcal{Y} = \{0, 1\}$. In this case the compact subsets of $\mathcal{Y}$ are $\emptyset$, $\{0\}$, $\{1\}$ and $\{0, 1\}$. Hence we can use directly Artstein's inequality applied to the capacity functional, obtaining:

$$\mu(\{j, k\}) \leq \mathbf{P}(y = j | z = 0)\,\mathbf{P}(z = 0) + \mathbf{P}(y = k | z = 1)\,\mathbf{P}(z = 1), \text{ for } j, k = 0, 1. \tag{2.9}$$

Notice that this upper bound on $\mu(\{j, k\})$ coincides with the familiar Fréchet bound on the joint probability that $(y(0) = j, y(1) = k)$. This can be shown by observing that

$$\mathbf{P}(y(0) = j, y(1) = k) = \sum_{t=0}^{1} \mathbf{P}(y(0) = j, y(1) = k | z = t)\,\mathbf{P}(z = t)$$

and applying the Fréchet upper bound on each of $\mathbf{P}(y(0) = j, y(1) = k | z = t)$, $t = 0, 1$. Similarly, one can show that the lower bound on $\mu(\{j, k\})$ also coincides with the Fréchet bound.

2.2.2. Adding statistical independence assumptions

Suppose now that the researcher also observes a variable v defined on $(\Omega, \mathfrak{F}, \mathbf{P})$ and taking values in $\mathcal{V} \subset \Re$. We consider the following assumptions, which use the nomenclature in Manski (2003, Section 7.4).

Assumption SI (*Statistical Independence of Outcomes and Instruments*)**.**

$$\mathbf{P}(y(t) | v) = \mathbf{P}(y(t)), \quad t \in \mathcal{T}.$$

Assumption SI-RF (*Statistical Independence of Response Functions and Instruments*)**.**

$$\mathbf{P}(y(\cdot) | v) = \mathbf{P}(y(\cdot)).$$

Whereas Assumption SI is treatment-specific, Assumption SI-RF posits that the entire response function is statistically independent from v, and therefore constrains its joint distribution rather than each of its marginals. Clearly, Assumption SI-RF implies Assumption SI. It is especially credible when the data come from a randomized experiment, where treatment is randomly assigned and the instrument v corresponds to the designated treatment. In this case, the identification problem persists as described in this Section when there is non-compliance with the randomly assigned treatment, and z is the treatment actually received and may or may not coincide with v.

Manski (2003, Proposition 7.3) derives the sharp identification region for $\mathbf{P}(y(t))$ under Assumption SI. The result in Manski (2003, Corollary 2.2.1) can easily be applied to obtain a useful alternative characterization when $\mathcal{V}$ is a finite set. Balke and Pearl (1997) derive the sharp identification region for $\mathbf{P}(y(t))$ under Assumption SI-RF when treatments, outcomes and instruments are all binary. Kitagawa (2009) significantly extends their findings, by allowing the outcome variable to have a continuous distribution. Here we extend the treatment of Manski (2003, Corollary 2.2.1), Balke and Pearl (1997) and Kitagawa (2009) by allowing for continuous outcomes, more than two treatments, and continuous instruments. Our use of random set theory allows us to establish the sharpness result through proofs which are relatively simple extensions of the proofs of Propositions 2.2 and 2.3. Most importantly, the results easily extend to the case that one additionally imposes shape restrictions on the response functions, in the spirit of Manski (1997), as we show in Section 2.2.3.

Characterization of the sharp identification regions under Assumption SI

Let $Y(t)$ be defined as in Eq. (2.5). Consider first the case that Assumption SI is maintained. When $\mathcal{Y}$ and $\mathcal{V}$ are finite sets,

the following Proposition repeats the result previously given by Manski (2003), Corollary 2.2.1, applied to the distribution of the potential outcome $y(t)$. When $\mathcal{V} = \mathcal{T} = \{0, 1\}$ but $\mathcal{Y}$ is not necessarily finite, it repeats the result previously given by Kitagawa (2009, Proposition 3.1). In all other cases, it extends their results.

Proposition 2.4. *Let Assumption SI hold. Then the sharp identification region for* $\mathbf{P}(y(t))$ *is*

$$\mathrm{H}[\mathbf{P}(y(t))] = \Big\{\mu \in \Gamma_{\mathcal{Y}} : \mu(K) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{P}(y \in K | z = t, v) \times \mathbf{P}(z = t|v)\ \forall K \in \mathcal{K}(\mathcal{Y})\Big\}. \tag{2.10}$$

If $\mathcal{Y}$ *is finite,*

$$\mathrm{H}[\mathbf{P}(y(t))] = \Big\{\mu \in \Gamma_{\mathcal{Y}} : \mu(k) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{P}(y = k, z = t, v) \times \mathbf{P}(z = t|v)\ \forall k \in \mathcal{Y}\Big\}.$$

If $\mathcal{Y} = [0, 1]$,

$$\mathrm{H}[\mathbf{P}(y(t))] = \Big\{\mu \in \Gamma_{\mathcal{Y}} : \mu([k_1, k_2]) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{P}(y \in [k_1, k_2] | z = t, v) \times \mathbf{P}(z = t|v), \forall k_1, k_2 \in \mathcal{Y} : k_1 \leq k_2\Big\}.$$

Proof. Using random sets, all the information in the available data and maintained assumptions can be expressed as $(y(t), v) \in \mathrm{Sel}((Y(t), v)) \cap \mathcal{I}$, where $\mathcal{I}$ is the set of random elements $(\xi, v) \in \mathcal{Y} \times \mathcal{V}$ such that ξ is statistically independent of v. Notice that if the SI Assumption is correct, this intersection is non-empty. By Theorem 2.1, $(y(t), v) \in \mathrm{Sel}((Y(t), v))$ if and only if

$$\begin{aligned}\mathbf{P}((y(t), v) \in M) &= \int_{\mathcal{V}} \mathbf{P}(y(t) \in M_v | v)\mathbf{P}_v(\mathrm{d}v)\\ &\geq \int_{\mathcal{V}} \mathbf{P}(Y(t) \subset M_v | v)\mathbf{P}_v(\mathrm{d}v)\\ &= \int_{\mathcal{V}} \mathbf{P}((Y(t), v) \subset M_v \times \{v\})\mathbf{P}_v(\mathrm{d}v)\end{aligned}$$

for all $M \in \mathcal{K}(\mathcal{Y} \times \mathcal{V})$, where $M_v = \{k : (k, v) \in M\}$ is the section of M at level v. Since v is a singleton, the events under the integral are disjoint and the integral equals $\mathbf{P}((Y(t), v) \subset M)$. Hence, this inequality can be written as

$$\mathbf{P}(y(t) \in K | v) \geq \mathbf{P}(Y(t) \subset K | v) \quad \forall K \in \mathcal{K}(\mathcal{Y})\ v\text{-a.s.}$$

By Assumption SI, $(y(t), v)$ belongs to $\mathcal{I}$. Hence we obtain

$$\mathbf{P}(y(t) \in K) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{P}(Y(t) \subset K | v)\ \forall K \in \mathcal{K}(\mathcal{Y}). \tag{2.11}$$

Observe that for a given $v \in V$, and for any $K \in \mathcal{K}(\mathcal{Y})$, $K \neq \mathcal{Y}$

$$\begin{aligned}\mathbf{P}(Y(t) \subset K | v) &= \mathbf{P}(Y(t) \subset K | z = t, v)\mathbf{P}(z = t|v)\\ &\quad + \mathbf{P}(Y(t) \subset K | z \neq t, v)\mathbf{P}(z \neq t|v)\\ &= \mathbf{P}(y \in K | z = t, v)\mathbf{P}(z = t|v).\end{aligned}$$

If $\mathcal{Y}$ is a finite set, Lemma B.1 guarantees that for each $v \in \mathcal{V}$ it suffices to check the containment functional dominance condition for all singleton sets $K = \{k\} \in \mathcal{Y}$, and therefore it also suffices for the essential supremum of the containment functional. If $\mathcal{Y} = [0, 1]$, $Y(t)$ is a random closed convex set, and Lemma B.2 in the Appendix guarantees that for each $v \in \mathcal{V}$ it suffices to check the containment functional dominance condition for sets $K \in \mathcal{K}(\mathcal{Y})$ which are intervals. Again, this assures that it suffices also for the essential supremum of the containment functional.

In summary, any μ satisfying the condition in Eq. (2.10) is the probability distribution of a random variable $y(t)$ such that $(y(t), v) \in \mathrm{Sel}((Y(t), v))$ and $y(t)$ is statistically independent of v. Conversely, any random variable $y(t)$ such that $(y(t), v) \in \mathrm{Sel}((Y(t), v))$ and $y(t)$ is statistically independent of v has a probability distribution satisfying the condition in Eq. (2.10). □

Characterization of the sharp identification regions under Assumption SI-RF

Consider now the case that the stronger Assumption SI-RF is maintained. Let $Y^{\mathcal{T}}$ be defined as in Eq. (2.7). Then we have the following result:

Proposition 2.5. *Let Assumption SI-RF hold. Then the sharp identification region for* $\mathbf{P}(y(\cdot))$ *is*

$$\mathrm{H}[\mathbf{P}(y(\cdot))] = \Big\{\mu \in \Gamma_{\mathcal{Y}^{\mathcal{T}}} : \mu(K) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{C}_{Y^{\mathcal{T}}|v}(K)\ \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)\Big\},$$

where $\mathbf{C}_{Y^{\mathcal{T}}|v}$ *is the conditional containment functional of* $Y^{\mathcal{T}}$ *given* v. *If* $\mathcal{Y} = [0, 1]$, *it suffices to check the above condition for sets* $\tilde{K} = \mathrm{co}(\tilde{K}(0) \cup \tilde{K}(1) \cup \cdots \cup \tilde{K}(T))$, *where for* $t \in \mathcal{T}$ *either* $\tilde{K}(t) = \emptyset$ *or* $\tilde{K}(t) = \mathcal{Y} \times \cdots \times \mathcal{Y} \times \left[k_1^t, k_2^t\right] \times \mathcal{Y} \cdots \times \mathcal{Y}$, $k_1^t \leq k_2^t$, $k_1^t, k_2^t \in \mathcal{Y}$, $t \in \mathcal{T}$. *For these sets,* $\mathbf{C}_{Y^{\mathcal{T}}|v}\left(\tilde{K}\right) = \sum_{t\in\mathcal{T}:\tilde{K}(t)\neq\emptyset} \mathbf{P}\left(y \in \left[k_1^t, k_2^t\right] | z = t, v\right)\mathbf{P}(z = t|v)$.

Proof. By the same argument as in the proof of Proposition 2.4, $([y(0), \ldots, y(T)], v) \in \left(Y^{\mathcal{T}}, v\right)$ if and only if $\forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)$

$$\mathbf{P}([y(0), \ldots, y(T)] \in K | v) \geq \mathbf{P}\left(Y^{\mathcal{T}} \subset K | v\right)\ v\text{-a.s.}$$

By the SI-RF assumption, $(y(0), \ldots, y(T))$ is statistically independent of v. Hence, the above condition reduces to

$$\mathbf{P}([y(0), \ldots, y(T)] \in K | v) \geq \operatorname*{ess\,sup}_{v\in\mathcal{V}} \mathbf{P}\left(Y^{\mathcal{T}} \subset K | v\right)\ \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right).$$

The specific result for $\mathcal{Y} = [0, 1]$ follows by the same argument as in the proof of Proposition 2.3. Its proof shows that for each $v \in \mathcal{V}$ it suffices to check the containment functional dominance condition for the sets in the statement of the proposition. This assures that it suffices also for the essential supremum of the containment functional. □

Remark 3 (*Binary Outcomes and Balke–Pearl Bounds*)**.** When $\mathcal{Y} = \mathcal{T} = \mathcal{V} = \{0, 1\}$, the compact subsets of $\mathcal{Y}$ are $\emptyset$, $\{0\}$, $\{1\}$ and $\{0, 1\}$ and we can use directly Artstein's inequality applied to the capacity functional to replicate the result in Balke and Pearl (1997) concerning sharp bounds on $\mathbf{P}(y(t) = 1)$, $t = 0, 1$. To see why this is the case, observe that the inequalities $\mu(K) \geq \operatorname{ess\,sup}_{v\in\mathcal{V}} \mathbf{C}_{Y^{\mathcal{T}}|v}(K)$ are equivalent to $\mu(K) \leq \operatorname{ess\,inf}_{v\in\mathcal{V}} \mathbf{T}_{Y^{\mathcal{T}}|v}(K)$ and reduce to:

$$\mathbf{P}(y(1) = j, y(0) = j) \leq \min_{v\in\{0,1\}} \{\mathbf{P}(y = j|v)\}, \quad \text{for } j = 0, 1.$$

$$\mathbf{P}(y(1) = j, y(0) = 1 - j) \leq \min_{v\in\{0,1\}} \{\mathbf{P}(y = j, z = 1|v) + \mathbf{P}(y = 1 - j, z = 0|v)\}, \quad \text{for } j = 0, 1.$$

$$\mathbf{P}(y(i) = j) \leq \min_{v\in\{0,1\}} \{\mathbf{P}(y = j, z = i|v) + \mathbf{P}(z = 1 - i|v)\}, \quad \text{for } i, j = 0, 1.$$

Hence, the upper bound for $\mathbf{P}(y(1) = 1)$, for example, is given by

$$\mathbf{P}(y(1) = 1) \leq \min\Big\{\min_{v\in\{0,1\}} \{\mathbf{P}(y = 1, z = 1|v) + \mathbf{P}(z = 0|v)\}, \min_{v\in\{0,1\}} \{\mathbf{P}(y = 1, z = 1|v) + \mathbf{P}(y = 0, z = 0|v) + \mathbf{P}(y = 1|1 - v)\}\Big\}.$$

One can similarly obtain other bounds. Notice that these bounds can also be derived using the Artstein's inequality/Fréchet bounds in Eq. (2.9) conditional on v, along with the bounds on each marginal distribution conditional on v, and then taking the minimum over v. The connection between the bounds of Balke and Pearl (1997) and the Fréchet bounds in Eq. (2.9) was first pointed out by Pepper (2002).

2.2.3. Adding statistical independence and monotone treatment response assumptions

Consider now the case that one adds to the analysis the assumption that treatment response is monotone, as in Manski (1997). Formally,

Assumption MTR (*Monotone Treatment Response*): Let the set $\mathcal{T}$ be ordered in terms of degree of intensity. Assume that for all treatment pairs $s, t \in \mathcal{T}$

$$t \geq s \Rightarrow \mathbf{P}\left(y\left(t\right) \geq y\left(s\right)\right) = 1.$$

Construction of the relevant random set for $y(t)$ under Assumption MTR

The analysis in Manski (1997) shows that *all* the information embodied in the available data and Assumption MTR translates into the fact that, for each $t \in \mathcal{T}$, $y\left(t\right) \in \text{Sel}\left(\overrightarrow{Y}\left(t\right)\right)$, where

$$\overrightarrow{Y}\left(t\right) = \begin{cases} [0, y] \cap \mathcal{Y} & \text{if } t < z, \\ \{y\} & \text{if } z = t, \\ [y, 1] \cap \mathcal{Y} & \text{if } t > z. \end{cases} \tag{2.12}$$

Here we provide novel results, characterizing the sharp identification region for $\mathbf{P}(y(t))$ under the joint assumption of statistical independence and of monotone treatment response.

Characterization of the sharp identification region under Assumptions SI and MTR

If we jointly impose Assumptions SI and MTR, we have the following result:

Proposition 2.6. *Let Assumptions SI and MTR hold. Then the sharp identification region for $\mathbf{P}(y(t))$ is*

$$\begin{aligned} \mathrm{H}\left[\mathbf{P}(y(t))\right] = \Big\{ & \mu \in \Gamma_{\mathcal{Y}} : \mu(K) \geq \operatorname*{ess\,sup}_{v \in \mathcal{V}} [\mathbf{P}(y < \sup K, z > t | v) \\ & + \mathbf{P}(y \in K, z = t | v) + \mathbf{P}(y > \inf K, z < t | v)] \, \forall K \in \mathcal{K}(\mathcal{Y}) \Big\}. \end{aligned}$$

If $\mathcal{Y} = [0, 1]$,

$$\begin{aligned} \mathrm{H}\left[\mathbf{P}(y(t))\right] = \Big\{ & \mu \in \Gamma_{\mathcal{Y}} : \mu([k_1, k_2]) \geq \operatorname*{ess\,sup}_{v \in \mathcal{V}} [\mathbf{P}(y < k_2, z > t | v) \\ & + \mathbf{P}(y \in [k_1, k_2], z = t | v) \\ & + \mathbf{P}(y > k_1, z < t | v)] \, \forall k_1, k_2 \in \mathcal{Y} : k_1 \leq k_2 \Big\}. \end{aligned}$$

Proof. The assumptions are summarized by requiring that $(y(t), v) \in \text{Sel}((\overrightarrow{Y}(t), v)) \cap \mathcal{I}$, where $\mathcal{I}$ is the set of random elements $(\xi, v) \in \mathcal{Y} \times \mathcal{V}$ such that ξ is statistically independent of v. If Assumptions SI and MTR are correct, this intersection is non-empty. By the same argument as in the proof of Proposition 2.4, $(y(t), v) \in \text{Sel}\left(\left(\overrightarrow{Y}(t), v\right)\right)$ if and only if

$$\mathbf{P}\left(y\left(t\right) \in K | v\right) \geq \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | v\right) \quad \forall K \in \mathcal{K}\left(\mathcal{Y}\right) \; v\text{-a.s.}$$

By Assumption SI, $(y(t), v)$ belongs to $\mathcal{I}$. Hence we obtain

$$\mathbf{P}\left(y\left(t\right) \in K\right) \geq \operatorname*{ess\,sup}_{v \in \mathcal{V}} \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | v\right) \quad \forall K \in \mathcal{K}\left(\mathcal{Y}\right).$$

Observe that for a given $v \in V$,

$$\begin{aligned} \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | v\right) &= \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | z > t, v\right) \mathbf{P}\left(z > t | v\right) \\ &\quad + \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | z = t, v\right) \mathbf{P}\left(z = t | v\right) \\ &\quad + \mathbf{P}\left(\overrightarrow{Y}\left(t\right) \subset K | z < t, v\right) \mathbf{P}\left(z < t | v\right) \\ &= \mathbf{P}\left(y < \sup K | z > t, v\right) \mathbf{P}\left(z > t | v\right) \\ &\quad + \mathbf{P}\left(y \in K | z = t, v\right) \mathbf{P}\left(z = t | v\right) \\ &\quad + \mathbf{P}\left(y > \inf K | z < t, v\right) \mathbf{P}\left(z < t | v\right). \end{aligned}$$

If $\mathcal{Y} = [0, 1]$, $Y(t)$ is a random closed convex set, and Lemma B.2 in the Appendix guarantees that for each $v \in \mathcal{V}$ it suffices to check the containment functional dominance condition for sets $K \in \mathcal{K}(\mathcal{Y})$ which are intervals. This assures that it suffices also for the essential supremum of the containment functional. □

Remark 4. Using the same approach as in this section and in Section 2.2.2 one can extend these results to obtain sharp identification regions for the probability distribution of the response function under statistical independence and shape restrictions. While conceptually straightforward if using Artstein's inequality, this extension is notationally cumbersome. We provide it in Appendix C.

3. Usefulness of the Aumann expectation

3.1. Aumann expectation represented through its support function

In many partial identification problems the object of interest is a conditional expectation, or taking expectations is a crucial step towards characterizing a sharp identification region (see, e.g., Beresteanu et al. (in press)). In these cases, the information provided by the empirical evidence and the maintained assumptions can often be expressed by saying that the conditional expectation of a random vector x belongs to the conditional *Aumann expectation* of a properly defined random set X, in the sense that $\mathbf{P}\left(\mathbf{E}\left(x|\mathfrak{F}_0\right) \in \mathbb{E}\left(X|\mathfrak{F}_0\right)\right) = 1$, where $\mathfrak{F}_0 \subset \mathfrak{F}$ denotes a sub-σ-algebra, see Definitions A.4 and A.5 in Appendix A.

If X is an integrably bounded random compact set, i.e., $\sup\{\|x\| : x \in X\}$ has a finite expectation, on a nonatomic probability space, then $\mathbb{E}[X]$ is a convex set and coincides with $\mathbb{E}[\text{co}(X)]$, see Molchanov (2005, Theorem 2.1.15).[11] Moreover, because X takes its realizations in a subset of the finite dimensional space $\Re^d$, $\mathbb{E}[X]$ is closed, see Molchanov (2005, Theorem 2.1.24). By the same argument, provided that the probability space contains no $\mathfrak{F}_0$-atoms (i.e., $\forall A \in \mathfrak{F}$ having positive measure, there is a $B \subseteq A$ such that $0 < \mathbf{P}(B|\mathfrak{F}_0) < \mathbf{P}(A|\mathfrak{F}_0)$ with positive probability), $\mathbb{E}[X|\mathfrak{F}_0]$ is a closed convex set almost surely, and $\mathbb{E}[X|\mathfrak{F}_0] = \mathbb{E}[\text{co}(X)|\mathfrak{F}_0]$.[12] This result is especially useful, because it implies that $\mathbb{E}[X|\mathfrak{F}_0]$ is equal to the intersection of its supporting half-spaces (see Rockafellar (1970, Theorem 13.1) and Molchanov (2005, Theorem 2.1.49-(iii))), which in turn are determined by its *support function* $h(\mathbb{E}[X|\mathfrak{F}_0], u)$, see Definition A.6 in Appendix A. In particular,

$$\begin{aligned} \mathbb{E}[X|\mathfrak{F}_0] &= \bigcap_{u \in \Re^d} \{\eta : \langle \eta, u \rangle \leq h(\mathbb{E}[X|\mathfrak{F}_0], u)\} \\ &= \bigcap_{u : \|u\| = 1} \{\eta : \langle \eta, u \rangle \leq h(\mathbb{E}[X|\mathfrak{F}_0], u)\}, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\Re^d$, and the last equality follows from the sublinearity of the support function, see Molchanov (2005, Appendix F).

[11] Of course the same conclusion holds if X is an integrably bounded random compact set with almost surely convex realizations.

[12] We continue the discussion focusing on $\mathbb{E}[X|\mathfrak{F}_0]$ and assuming that the probability space contains no $\mathfrak{F}_0$-atoms, but of course all the results apply, with obvious modifications, to $\mathbb{E}[X]$.

The above considerations imply that a candidate η belongs to $\mathbb{E}[X|\mathfrak{F}_0]$ if and only if $\langle \eta, u\rangle \leq h(\mathbb{E}[X|\mathfrak{F}_0], u)\ \forall u : \|u\| = 1$. This gives a necessary and sufficient condition for $\mathbf{P}(\mathbf{E}(x|\mathfrak{F}_0) \in \mathbb{E}(X|\mathfrak{F}_0)) = 1$, which relates the conditional expectation of the random vector x to the conditional Aumann expectation of the random set X. Yet, the family of all selections is very rich even for simple random sets. But a fundamental simplification is possible, by relating the support function of $\mathbb{E}[X|\mathfrak{F}_0]$ to $\mathbf{E}(h(X,u)|\mathfrak{F}_0)$. This is a fundamental result in random set theory, first given by Artstein (1974) for the case of unconditional Aumann expectations.[13]

Theorem 3.1 (*Aumann Expectation and Support Function*). *Let $X \in \mathcal{F}$ be an integrably bounded random set defined on a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$. Let $\mathfrak{F}_0 \subset \mathfrak{F}$ be a sub-σ-algebra, and assume that the probability space contains no $\mathfrak{F}_0$-atoms.[14] Then the conditional Aumann expectation of X is the unique convex closed set $\mathbb{E}[X|\mathfrak{F}_0]$ satisfying*

$$\mathbf{E}(h(X,u)|\mathfrak{F}_0) = h(\mathbb{E}[X|\mathfrak{F}_0], u) \quad \text{for all } u \in \Re^d.$$

Proof. See Dynkin and Evstigneev (1976, Theorem 1.2) and Molchanov (2005, Theorems 2.1.22 and 2.1.47-iv). □

Hence, one can conclude that a random vector η belongs to $\mathbb{E}[X|\mathfrak{F}_0]$ if and only if $\langle \eta, u\rangle \leq \mathbf{E}(h(X,u)|\mathfrak{F}_0)\ \forall u : \|u\| = 1$. The latter conditional expectation is usually simple to compute.

Remark 5. A simple application of Theorem 3.1 yields immediately the sharp identification region for $\mathbf{E}(y(t))$ and $\mathbf{E}(y(\cdot))$, hence replicating results in Manski (2003, Eqs. (7.10) and (7.11)). Using the support function/Aumann expectation approach, the analysis easily extends to cases where mean independence assumptions and shape restrictions are imposed. See Propositions C.2–C.4 in Appendix C. A characterization of the sharp identification region for $\mathbf{E}(y(\cdot))$ under these various sets of assumptions is especially important if the ultimate goal of the researcher is treatment choice, see e.g. Manski (2003, Chapter 7).[15]

3.2. Best linear prediction and the selection problem

We now consider the case that one is interested in best linear prediction of $y(t)$ given covariates w (including a constant). Let θ denote the parameters of such linear prediction, let w be of dimension $d \times 1$, and let $L(y(t)|w^0) = w^{0\prime}\theta$ denote the linear prediction of $y(t)$ given a specific value of $w = w^0$. Notice that here we are not assuming a linear model in any substantive sense, nor are we assuming availability of instruments.[16] Our analysis revisits results in Beresteanu and Molinari (2008, Section 4), specializing them for specific questions of interest in empirical applications.[17] Stoye (2007) provides related findings; in particular, he derives sharp identification regions for linear combinations of coefficients of best linear predictors which coincide with those given below for a single component of the vector θ and for $L(y(t)|w^0)$.

Let $Y(t)$ be defined as in Eq. (2.5), and let $\Sigma \equiv \mathbf{E}(ww')$. Assume that Σ is finite and of full rank. Let $G(t) = \{g : g = w\psi, \psi \in \text{Sel}(Y(t))\}$. Beresteanu and Molinari (2008) show that $G(t)$ is a random closed set and the sharp identification region for θ is given by

$$\begin{aligned}\text{H}(\theta) &= \left\{\theta : \theta = \Sigma^{-1}\mathbf{E}(w\psi), \psi \in \text{Sel}(Y(t))\right\} \\ &= \left\{\theta : \theta = \Sigma^{-1}\mathbf{E}(g), g \in \text{Sel}(G(t))\right\} = \Sigma^{-1}\mathbb{E}[G(t)]. \qquad (3.1)\end{aligned}$$

They also show that the sharp identification region for each component θ_k of θ is given by

$$\begin{aligned}\text{H}(\theta_k) &= \{\theta_k : \exists\theta_{-k} \text{ such that } [\theta_k, \theta_{-k}] \in \text{H}(\theta)\} \\ &= \left[\frac{\mathbf{E}\left[\min\{\tilde{w}_k y1(z=t), \tilde{w}_k[y1(z=t)+1(z\neq t)]\}\right]}{\mathbf{E}(\tilde{w}_k^2)}, \right. \\ &\qquad \left. \frac{\mathbf{E}\left[\max\{\tilde{w}_k y1(z=t), \tilde{w}_k[y1(z=t)+1(z\neq t)]\}\right]}{\mathbf{E}(\tilde{w}_k^2)}\right],\end{aligned}$$

where, with some abuse of notation, $[\theta_k, \theta_{-k}]$ denotes a candidate value for θ, $\tilde{w}_k$ is the residual obtained after projecting w_k on the other covariates w_{-k}, and $1(\cdot)$ is the indicator function of the event in parenthesis.

Remark 6. Ponomareva and Tamer (2011) study the problem of misspecification in moment inequality models. One of the examples they use is the linear model for conditional expectations in the presence of interval outcome data. They propose a misspecification robust Least Squares Set. This set collects all parameter values giving a best linear approximation to some conditional expectation function that lies between the upper and lower conditional expectation functions corresponding to the upper and lower points in the interval data. Their Least Squares Set is equal to $\text{H}(\theta)$ in Eq. (3.1). To see this, it suffices to take Example A.1—Selections from Appendix A, and see that $\text{Sel}(Y(t))$ coincides with the set of variables for which Ponomareva and Tamer run linear projections.

Suppose that one is interested in predicting $y(t)$ for a specific value of w, denoted w^0. This amounts to obtaining

$$\text{H}\left[L(y(t)|w^0)\right] = \left\{r : r = w^{0\prime}\theta, \theta \in \text{H}(\theta)\right\}.$$

Alternatively, one might be interested in contrasts among predictions obtained for different values of w, denoted w^0 and w^1. This amounts to obtaining

$$\begin{aligned}&\text{H}\left[L(y(t)|w=w^1) - L(y(t)|w=w^0)\right] \\ &\quad = \left\{r : r = (w^1 - w^0)'\theta, \theta \in \text{H}(\theta)\right\}.\end{aligned}$$

These sets are intervals in $\Re$, hence fully described by their support functions for $u = \pm 1$. This observation leads to an extremely simple characterization:

Proposition 3.2. *The sharp identification region for $L(y(t)|w^0)$ is given by*

$$\begin{aligned}\text{H}\left[L(y(t)|w^0)\right] = [&\mathbf{E}[\min\{w^{0\prime}\Sigma^{-1}wy1(z=t), \\ &w^{0\prime}\Sigma^{-1}w(y1(z=t)+1(z\neq t))\}], \\ &\mathbf{E}[\max\{w^{0\prime}\Sigma^{-1}wy1(z=t), w^{0\prime}\Sigma^{-1}w(y1(z=t) \\ &+1(z\neq t))\}]].\end{aligned}$$

The sharp identification region for $L(y(t)|w=w^1) - L(y(t)|w=w^0)$ is given by

---

[13] The result of the following Theorem also holds if X is a random closed set with almost surely convex realizations. It is easy to see that $\sup_{u:\|u\|=1}|h(X,u)| = \sup\{\|x\| : x \in X\} = \|X\|_H$. Hence, if X is integrably bounded, then $\mathbf{E}[|h(X,u)||\mathfrak{F}_0]$ is finite for all $u \in \Re^d$.

[14] Formally, assume that $\forall A \in \mathfrak{F}$ having positive measure, there is a $B \subseteq A$ such that $0 < \mathbf{P}(B|\mathfrak{F}_0) < \mathbf{P}(A|\mathfrak{F}_0)$ with positive probability.

[15] For example, a planner who wants to maximize population mean welfare needs to work with the elements of $\text{H}[\mathbf{E}(y(\cdot))]$ rather than with the elements of $\{\text{H}[\mathbf{E}(y(t))], t \in \mathcal{T}\}$.

[16] Bontemps et al. (2008) study the related problem of best linear prediction with interval outcome data, assuming a linear model and the availability of instruments. They allow for the presence of more instruments than parameters, and extend the familiar Sargan test for overidentifying restrictions to partially identified models.

[17] Beresteanu et al. (2009, Section 5) provide a tractable characterization of the sharp identification region of θ for the more general problem of best linear prediction with interval data both on outcomes and covariates.

$$\begin{aligned}&\mathrm{H}\left[L\left(y\left(t\right)|\, w=w^1\right)-L\left(y\left(t\right)|\, w=w^0\right)\right]\\&\quad=\Big[\mathbf{E}\Big[\min\Big\{\left(w^1-w^0\right)'\Sigma^{-1}wy1\left(z=t\right),\\&\qquad\left(w^1-w^0\right)'\Sigma^{-1}w\left(y1\left(z=t\right)+1\left(z\neq t\right)\right)\Big\}\Big],\\&\qquad\mathbf{E}\Big[\max\Big\{\left(w^1-w^0\right)'\Sigma^{-1}wy1\left(z=t\right),\\&\qquad\left(w^1-w^0\right)'\Sigma^{-1}w\left(y1\left(z=t\right)+1\left(z\neq t\right)\right)\Big\}\Big]\Big].\end{aligned}$$

If $w^1=\left[w_k^0+1,\, w_{-k}^0\right]$, *then*

$$\mathrm{H}\left[L\left(y\left(t\right)|\, w=w^1\right)-L\left(y\left(t\right)|\, w=w^0\right)\right]=\mathrm{H}\left(\theta_k\right).$$

Proof. To obtain the sharp identification region, recall that $r=w^{0\prime}\theta\in\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right]$ if and only if $ur\leq h\left(\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right],u\right)$ for $u=\pm1$, so that it suffices to characterize the support function of $\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right]$. This function is equal to:

$$\begin{aligned}&h\left(\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right],u\right)=h\left(w^{0\prime}\Sigma^{-1}\mathbb{E}\left(G\left(t\right)\right),u\right)\\&\quad=\max_{g\in G(t)}uw^{0\prime}\Sigma^{-1}\mathbf{E}\left(g\right)\\&\quad=\max_{\psi\in Y(t)}uw^{0\prime}\Sigma^{-1}\mathbf{E}\left(w\psi\right)=\max_{\psi\in Y(t)}\mathbf{E}\left(uw^{0\prime}\Sigma^{-1}w\psi\right).\end{aligned}$$

Simple algebra gives the final result, observing that $Y\left(t\right)$ can be written as

$$Y\left(t\right)=\left[y1\left(z=t\right),y1\left(z=t\right)+1\left(z\neq t\right)\right]\cap\mathcal{Y}.$$

The same reasoning and algebra gives the sharp identification region for contrasts. The last result follows from observing that when $w^1=\left[w_k^0+1,\, w_{-k}^0\right]$,

$$\begin{aligned}&\mathrm{H}\left[L\left(y\left(t\right)|\, w=w^1\right)-L\left(y\left(t\right)|\, w=w^0\right)\right]\\&\quad=\left\{r\in\Re:r=\left(\left(w_k^0+1\right)\theta_k+w_{-k}^{0\prime}\theta_{-k}\right)-w^{0\prime}\theta,\theta\in\mathrm{H}\left[\theta\right]\right\}\\&\quad=\left\{\theta_k:\exists\theta_{-k}\text{ such that }\left[\theta_k,\theta_{-k}\right]\in\mathrm{H}\left(\theta\right)\right\}=\mathrm{H}\left[\theta_k\right].\quad\square\end{aligned}$$

A nice consequence of this result is that the identification regions for the best linear predictor, for its contrasts, and for each component of θ can be easily calculated by running simple linear projections on a standard statistical package such as, for example, Stata.[18]

It is also common, in empirical applications, to work with affine transformations of the covariates w. Demeaning or standardization are typical affine transformations used in practice. Here we apply them to the non-constant components of w. Let Π be a $(d-1)\times(d-1)$ matrix of full rank and let λ be a $(d-1)\times1$ vector. Let $\breve{w}_{-1}=\Pi w_{-1}+\lambda$. If for example one is interested in demeaning w, then Π is the identity matrix and $\lambda=-\mathbf{E}\left(w_{-1}\right)$. The following proposition shows how the sharp identification regions of parameters of interest change, in conjunction with these affine transformations.

Proposition 3.3. *The sharp identification region for the coefficients* $\breve{\theta}$ *of the best linear predictor of* $y\left(t\right)$ *given* $\breve{w}=\left[1\ \breve{w}_{-1}\right]$ *is*

$$\mathrm{H}\left(\breve{\theta}\right)=\begin{bmatrix}1&-\lambda'\Pi^{-1\prime}\\0&\Pi^{-1\prime}\end{bmatrix}\mathrm{H}\left(\theta\right).$$

The sharp identification region for $L\left(y\left(t\right)|\,\breve{w}^0\right)$ *is*

$$\mathrm{H}\left[L\left(y\left(t\right)|\,\breve{w}^0\right)\right]=\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right].$$

Proof. Consider first the parameters of the best linear predictor. Observe that with the non-transformed covariates, $\theta\in\mathrm{H}\left(\theta\right)$ if and only if there exists a $\psi\in\mathrm{Sel}\left(Y\left(t\right)\right)$ such that $\mathbf{E}\left(w\left(\psi-w'\theta\right)\right)=0$. Similarly, with the transformed covariate, $\breve{\theta}\in\mathrm{H}\left(\breve{\theta}\right)$ if and only if there exists a $\breve{\psi}\in\mathrm{Sel}\left(Y\left(t\right)\right)$ such that $\mathbf{E}\left(\breve{w}\left(\breve{\psi}-\breve{w}'\breve{\theta}\right)\right)=0$. Take $\theta\in\mathrm{H}\left(\theta\right)$ such that for a $\psi\in\mathrm{Sel}\left(Y\left(t\right)\right)$, $\mathbf{E}\left(w\left(\psi-w'\theta\right)\right)=0$. Let

$$\breve{\theta}=\begin{bmatrix}1&-\lambda'\Pi^{-1\prime}\\0&\Pi^{-1\prime}\end{bmatrix}\begin{bmatrix}\theta_1\\\theta_{-1}\end{bmatrix}.$$

Then

$$\begin{aligned}&\mathbf{E}\left(\breve{w}\left(\psi-\breve{w}'\theta\right)\right)=\mathbf{E}\left(\left(\Pi w_{-1}+\lambda\right)\left(\psi-\left(\Pi w_{-1}+\lambda\right)'\Pi^{-1\prime}\theta_{-1}\right.\right.\\&\qquad\left.\left.+\ \lambda'\Pi^{-1\prime}\theta_{-1}-\theta_1\right)\right)\\&\quad=\mathbf{E}\left(\left(\Pi w_{-1}+\lambda\right)\left(\psi-\left(w_{-1}'\Pi'+\lambda'\right)\Pi^{-1\prime}\theta_{-1}\right.\right.\\&\qquad\left.\left.+\ \lambda'\Pi^{-1\prime}\theta_{-1}-\theta_1\right)\right)\\&\quad=\mathbf{E}\left(\left(\Pi w_{-1}+\lambda\right)\left(\psi-w_{-1}'\theta_{-1}-\theta_1\right)\right)=0.\end{aligned}$$

Hence $\breve{\theta}\in\mathrm{H}\left(\breve{\theta}\right)$. The reverse argument follows by the same logic. Consider now the best linear predictor itself:

$$\begin{aligned}&\mathrm{H}\left[L\left(y\left(t\right)|\,\breve{w}^0\right)\right]=\left\{\breve{w}^{0\prime}\breve{\theta}:\breve{\theta}\in\mathrm{H}\left(\breve{\theta}\right)\right\}\\&\quad=\left\{\breve{w}^{0\prime}\breve{\theta}:\breve{\theta}=\begin{bmatrix}1&-\lambda'\Pi^{-1\prime}\\0&\Pi^{-1\prime}\end{bmatrix}\begin{bmatrix}\theta_1\\\theta_{-1}\end{bmatrix},\theta\in\mathrm{H}\left(\theta\right)\right\}\\&\quad=\left\{r:r=\breve{w}_{-1}^{0\prime}\Pi^{-1\prime}\theta_{-1}-\lambda'\Pi^{-1\prime}\theta_{-1}+\theta_1,\theta\in\mathrm{H}\left(\theta\right)\right\}\\&\quad=\Big\{r:r=\left(w_{-1}^{0\prime}\Pi'+\lambda'\right)\Pi^{-1\prime}\theta_{-1}-\lambda'\Pi^{-1\prime}\theta_{-1}+\theta_1,\\&\qquad\theta\in\mathrm{H}\left(\theta\right)\Big\}\\&\quad=\left\{r:r=w_{-1}^{0\prime}\theta_{-1}+\theta_1,\theta\in\mathrm{H}\left(\theta\right)\right\}\\&\quad=\mathrm{H}\left[L\left(y\left(t\right)|\, w^0\right)\right].\quad\square\end{aligned}$$

This result implies, for example, that demeaning the data will have, in the partially identified case, the same effect that it has in the point identified case. The sharp identification region of the best linear predictor itself is not affected, and neither is the sharp identification region of each slope parameter. On the other hand, the sharp identification region of the intercept parameter may change substantially. Similarly, rescaling the data leaves the sharp identification region of the best linear predictor itself and of the intercept unaffected. On the other hand, the sharp identification region of the slope parameter may change substantially. Fig. 1 illustrates graphically these changes.[19] Clearly, these changes in the size and shape of the identification region are purely the result of standardizing, so caution should be taken in interpreting the results of the analysis.

4. A note on estimation and statistical inference

The sharp identification regions derived in Sections 2 and 3 can be categorized as follows: (a) transformations of conditional or unconditional Aumann expectations; (b) sets of multinomial distributions defined by a finite number of unconditional (conditional

[18] Stata code implementing sample analog estimators of these identification regions, along with confidence sets, confidence collections, and test of hypothesis as in Beresteanu and Molinari (2008), is freely downloadable at http://www.arts.cornell.edu/econ/fmolinari/#Stata_SetBLP. This code also allows for estimation, confidence statements, and test of hypothesis concerning the identification regions of any two components of θ.

[19] These figures are for illustration only. They where created using data taken from the Health and Retirement Study on individuals' expectations of surviving to age 75, mapped into intervals as in Manski and Molinari (2010). The interval expectation data were projected on a constant and individuals' age.

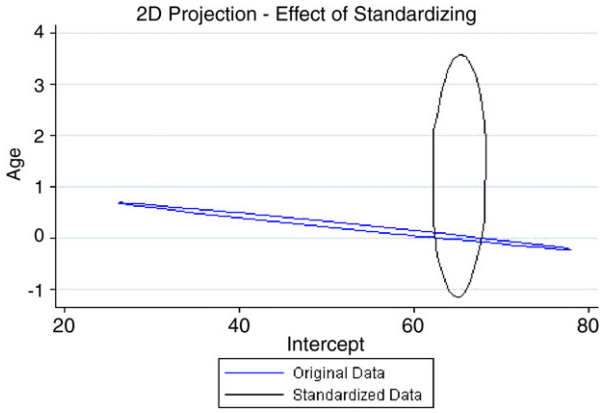

Fig. 1. $H(\theta)$ and $H(\breve{\theta})$ obtained, respectively, using x_{-1} and $\frac{x_{-1}-E(x_{-1})}{\sqrt{var(x_{-1})}}$.

in the presence of instruments v and/or covariates w) moment inequalities; (c) sets of continuous distributions defined by a continuum of unconditional (conditional in the presence of instruments v and/or covariates w) moment inequalities indexed by $k_1, k_2 \in \mathcal{Y}$. Category (a) applies to Propositions 3.2, 3.3 and C.2–C.4. Categories (b) and (c) apply to Propositions 2.2–2.6 and C.1, respectively for the case of $\mathcal{Y}$ being discrete and $\mathcal{Y} = [0, 1]$. Here we assume that one observes a random sample $(y_i, z_i, w_i)_{i=1}^n$ drawn from the same population as (y, z, w). This in turn assures that the random sets $Y_i(t)$, $Y_i^{\mathcal{T}}$, $\overrightarrow{Y}_i(t)$, and $\overrightarrow{Y}_i^{\mathcal{T}}$ defined as in Eqs. (2.5), (2.7), (2.12) and (C.1) with (y, z, w) replaced by (y_i, z_i, w_i) are independently and identically distributed, see Beresteanu and Molinari (2008, Lemmas A.3 and A.5).

Estimation of sharp identification regions of type (a) for unconditional Aumann expectations can be carried out by sample analog methods, replacing the Aumann expectation by a Minkowski average of random sets as explained in Beresteanu and Molinari (2008, Sections 3 and 4). Confidence sets and confidence collections can be constructed to cover or have as a member the sharp identification region and its subsets with a prespecified asymptotic probability using the method proposed by Beresteanu and Molinari.[20] When the relevant unconditional Aumann expectation is a subset of $\mathfrak{R}$, the methods of Imbens and Manski (2004) and Stoye (2009) can be employed to obtain confidence sets that cover each point in the sharp identification region with a prespecified asymptotic probability. For the case of conditional Aumann expectations as in Propositions C.3 and C.4, estimation and statistical inference can be carried out using the methods proposed by Andrews and Shi (2009), Chernozhukov et al. (2009) and Ponomareva (2010).

Estimation of sharp identification regions of types (b) and (c) with conditional or unconditional moment inequalities can be carried out by replacing probability distribution functions by empirical distribution functions. By Theorem 1.2.22 in Molchanov (2005) the resulting estimators of the sharp identification regions, obtained by replacing the population versions of the capacity and containment functionals with their empirical counterparts, are consistent in the Hausdorff–Prokhorov metric. In the case of sharp identification regions of type (b) with unconditional moment inequalities, test of hypothesis and confidence statements can be carried out using the methods proposed by Chernozhukov et al. (2007), Andrews and Soares (2010), Bugni (2010) and Canay (2010), among others. When sharp identification regions of type (b) are defined via conditional moment inequalities but w is discrete, estimation and statistical inference can be carried out using the methods proposed by Andrews and Shi (2009) and Ponomareva (2010), even if v has a continuous distribution.

In the case of sharp identification regions of types (b) and (c) with conditional moment inequalities indexed by a continuously distributed w, existing methods for construction of confidence sets do not readily apply, because the object of interest is not a finite dimensional parameter vector. Development of a procedure to conduct statistical inference in this case is left for future research.

5. Aumann expectation or capacity functional?

It is often the case that theoretically one can use either the "capacity functional approach" or the "Aumann expectation approach" to address a specific partial identification problem. However, there might be computational advantages to using one of these approaches rather than the other. Here we give a few examples of how to choose between them.

5.1. Limitations of the Aumann expectation approach

Consider first the case where the object of ultimate interest is the partially identified probability distribution $\mathbf{P}(x)$ of an unobservable random variable $x \in \mathcal{X} \subset \mathfrak{R}^d$. The researcher knows that $x \in \text{Sel}(X)$ for a random set X revealed by the data and taking its realizations in $\mathcal{X}$.[21] In this case, the capacity functional and Artstein's inequality allow for a simple characterization of the sharp identification region, see Eq. (2.3). On the other hand, the Aumann expectation can be used to conclude that $x \in \text{Sel}(X)$ if and only if

$$\mathbf{E}(x1(A)) \in \mathbb{E}(X1(A)) \quad \forall A \in \mathfrak{F}, \tag{5.1}$$

where $1(\cdot)$ is the indicator function of the event in parenthesis (see Molchanov (2005, Theorem 2.1.18)). Hence, one could characterize $H[\mathbf{P}(x)]$ as the set of $\mu \in \Gamma_{\mathcal{X}}$ such that μ is the probability distribution of a random element ξ satisfying Eq. (5.1). However, this characterization is much less tractable computationally than the characterization obtained through Artstein's inequality. Moreover, it is not simple, computationally, to incorporate into the Aumann expectation approach assumptions which restrict $\mathbf{P}(x)$ directly, such as for example the statistical independence conditions considered in Section 2.2.2.

Notice that there are cases in which the two approaches are equivalent, both conceptually and computationally. To clarify this claim, consider the following simple example.[22] Let X_θ be a random closed set with realizations in $\{0, 1\}$, and suppose that the specific realizations that this set takes are a known function of a parameter θ and some unobservable random variable ε. Let the distribution function of ε be known up to a parameter vector which is included in θ. Let θ be the object of ultimate interest. Assume that the researcher observes a binary random variable x and can learn its distribution, $\mathbf{P}(x = 1)$. Assume further that the informational content of the economic model is equivalent to the statement that $x \in \text{Sel}(X_\theta)$.[23] Then using Artstein's inequality one can easily characterize the sharp identification region of θ as[24]

$$H(\theta) = \left\{\theta : \mathbf{P}(x = k) \leq \mathbf{T}_{X_\theta}(\{k\}), k \in \{0, 1\}\right\}.$$

[20] Stata code implementing these procedures is freely downloadable at http://www.arts.cornell.edu/econ/fmolinari/#Stata_SetBLP.

[21] In Section 2.2, we consider two examples: (1) $\mathcal{X} = \mathcal{Y}$ and $x = y(t)$ with $X = Y(t)$; and (2) $\mathcal{X} = \mathcal{Y}^{\mathcal{T}}$ and $x = y(\cdot)$ with $X = Y^{\mathcal{T}}$.

[22] More general and complex instances of the same basic idea are studied in Beresteanu et al. (2008) and Galichon and Henry (2009a).

[23] See Beresteanu et al. (2009, Appendix B) for examples.

[24] Here it suffices to look at singletons k because the realizations of X_θ are either singletons, or the entire space $\{0, 1\}$, see Lemma B.1.

On the other hand, one can construct a random closed set Q_θ taking its realizations in $\{[1\,0]\,,[0\,1]\} \subset \Re^2$ as follows

$$Q_\theta = \begin{cases} \{[1\,0]\} & \text{if } X_\theta = \{0\}\,, \\ \{[0\,1]\} & \text{if } X_\theta = \{1\}\,, \\ \{[1\,0]\,,[0\,1]\} & \text{if } X_\theta = \{0,1\}\,. \end{cases}$$

Let $\mathbf{P}(x) = [\mathbf{P}(x=0)\,\mathbf{P}(x=1)]$. Then

$$\begin{aligned} \mathrm{H}(\theta) &= \{\theta : \mathbf{P}(x) \in \mathbb{E}(Q_\theta)\} = \{\theta : \langle \mathbf{P}(x)\,, u\rangle \\ &\leq \mathbf{E}(h(Q_\theta, u))\,, u \in \{[1\,0]\,,[0\,1]\}\}\,. \end{aligned}$$

To see this, observe that for $u = [1\,0]$,

$$\begin{aligned} &\mathbf{E}(h(Q_\theta, [1\,0])) \\ &\quad = \langle [1\,0]\,,[1\,0]\rangle \mathbf{P}(X_\theta = \{0\}) + \langle [0\,1]\,,[1\,0]\rangle \mathbf{P}(X_\theta = \{1\}) \\ &\qquad + \max\{\langle [1\,0]\,,[1\,0]\rangle\,, \langle [0\,1]\,,[1\,0]\rangle\}\mathbf{P}(X_\theta = \{0,1\}) \\ &\quad = \mathbf{P}(X_\theta = \{0\}) + \mathbf{P}(X_\theta = \{0,1\}) = \mathbf{T}_{X_\theta}(\{0\})\,. \end{aligned}$$

Similar algebra gives that $\mathbf{E}(h(Q_\theta, [0\,1])) = \mathbf{T}_{X_\theta}(\{1\})$, hence establishing equivalence of the two approaches. Notice that in this example a crucial role is played by the fact that the random variable x and the random set X_θ take on a finite number of realizations, hence replicating the familiar result that the distribution of a discrete random variable can be equivalently represented by taking the expectation of a vector of indicator functions.

5.2. Limitations of the capacity functional approach

The capacity functional approach resulting from a judicious application of Artstein's inequality may not be computationally practical for obtaining sharp identification regions of expectations, unless the problem at hand is particularly simple. To illustrate this claim, suppose first that one is interested in the expectation $\mathbf{E}(x)$ of an unobservable random variable $x \in \mathcal{X} \subset \Re^d$, and that the researcher knows that $x \in \mathrm{Sel}(X)$ for a random set X revealed by the data and taking its realizations in $\mathcal{X}$. In this case, the Aumann expectation and Theorem 3.1 allow for a simple characterization of the sharp identification region as

$$\mathrm{H}[\mathbf{E}(x)] = \left\{\eta \in \Re^d : \langle \eta, u\rangle \leq \mathbf{E}(h(X,u))\ \forall u \in \Re^d : \|u\| = 1\right\}.$$

If $d = 1$ and $X \subset \Re_+$ *a.s.*, it turns out that $\mathrm{H}[\mathbf{E}(x)]$ can be equivalently characterized using the Choquet integral with respect to the containment and capacity functionals, as

$$\mathrm{H}[\mathbf{E}(x)] = \left[\int x \mathrm{d}\mathbf{C}_X\,, \int x \mathrm{d}\mathbf{T}_X\right],$$

where $\int x \mathrm{d}\mathbf{T}_X = \int_0^\infty \mathbf{T}_X(\{x : x \geq t\})\,\mathrm{d}t$, and similarly for $\int x \mathrm{d}\mathbf{C}_X$, see Molchanov (2005, Theorem 1.5.1). When X can take on negative values, the above definition can be extended, see Molchanov (2005, p. 72). This result is the analog for random sets, of the familiar result that a nonnegative random variable x has $\mathbf{E}(x) = \int_\Omega x(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_0^{+\infty} \mathbf{P}(x > t)\,\mathrm{d}t$.

If $\mathrm{d} > 1$, it is still possible to characterize the expectation of the support function of X through the capacity functional, applying a formula similar to the one above to the function $\langle x, u\rangle$. This function takes on negative values, and therefore one needs to use the expression in Molchanov (2005, p. 72). However, this result is a mere repetition of the Aumann expectation approach. Moreover, it requires one to calculate the capacity functional of X, and then take integrals with respect to it. This task can be computationally intense. On the other hand, calculating directly the expectation of the support function of X is usually straightforward and computationally very simple.

There are additional cases in which taking expectations is a crucial step towards characterizing a sharp identification region of interest, and the Aumann expectation approach is preferable to the capacity functional approach, because it is computationally much faster as well as more intuitive. To clarify this claim, consider the following simple example.[25] Let Q_θ be a random closed set with realizations in $[0, 1]$, and suppose that the specific realizations that this set takes are a known function of a parameter vector θ and some unobservable random variable ε. Let the distribution function of ε be known up to a parameter vector which is included in θ. Let θ be the object of ultimate interest. Interpret the selections $q \in \mathrm{Sel}(Q_\theta)$ as parameters of a Bernoulli law. Assume that the researcher observes a binary random variable x and can learn its distribution, $\mathbf{P}(x = 1)$. Assume further that the informational content of the economic model is equivalent to the statement that $\mathbf{P}(x = 1) = \mathbf{E}(q^\star)$, with $q^\star \in \mathrm{Sel}(Q_\theta)$ and the expectation taken with respect to the distribution of ε. One can easily characterize the sharp identification region of θ as

$$\begin{aligned} \mathrm{H}(\theta) &= \{\theta : \mathbf{P}(x=1) \in \mathbb{E}(Q_\theta)\} \\ &= \{\theta : u\mathbf{P}(x=1) \leq \mathbf{E}(h(Q_\theta, u))\,, u = \pm 1\}\,, \end{aligned}$$

where the expectation of the support function of Q_θ is taken with respect to ε. For given θ, the support function of Q_θ is straightforward to calculate, and therefore the same is true for $\mathrm{H}(\theta)$.

Even in this stylized example, however, it is not immediate how one can use the capacity functional approach to characterize $\mathrm{H}(\theta)$. This is because in order to construct a random set to which x belongs with probability one, we would need to add an auxiliary random variable z, uniformly distributed on $[0, 1]$ and independent of ε, and define

$$X_\theta = \{\xi : \xi = 1(z < q)\,, q \in \mathrm{Sel}(Q_\theta)\}\,.$$

Such construction does not lead to a computationally feasible application of Artstein's inequality.

6. Conclusions

This paper has illustrated how the use of random set theory can benefit, and simplify, partial identification analysis. We have revisited results previously available in the literature, and established new results concerning identification of the distributions of potential outcomes and response functions and their expectation, in the presence of selectively observed data, statistical independence and mean independence assumptions, and shape restrictions. We have also derived new results concerning best linear prediction with interval outcome data.

The broad picture emerging from our analysis is the following. When a feature of a probability distribution of interest is partially identified, it is often possible to trace back the lack of point identification to the fact that either the data or the maintained assumptions yield a collection of random variables which are observationally equivalent. This collection is equal to the family of selections of a properly specified random closed set, and random set theory can be applied.

The first task that the researcher needs to carry out is to specify the relevant random closed set. In the case of incomplete data, such as the selection problem studied here, the relevant random closed set is the collection of values that the potential outcome can take — the observed (singleton) outcome when the treatment of interest is realized, and the entire outcome space otherwise.

The next task is to carefully determine how the observable variables relate to this random set. In certain partial identification problems, such as the selection problem studied here, the

[25] More general and complex instances of the same basic idea are studied in Beresteanu et al. (in press).

observable variables determine a random closed set to which the (unobservable) variable of interest belongs with probability one. In other partial identification problems, the observable variable belongs to a random closed set which is determined by the model. In other partial identification problems, the distribution of the observable variable belongs to the Aumann expectation of a random closed set which is determined by the model. See Section 5 above and Beresteanu et al. (2009) for examples.

The final task is to determine which tool of random set theory is best suited (either because computationally preferable, or more intuitive) to characterize the sharp identification region of the parameter of interest. In certain cases, working directly with probability distributions is a crucial step in describing the set of observationally equivalent parameters of interest, and the informational content of the data and the model is equivalent to saying that a random variable belongs to a properly specified random set with probability one. Hence, here the capacity functional approach based on Artstein's inequality is ideal to characterize the sharp identification region.

In other cases, taking expectations is a crucial step in describing the set of observationally equivalent parameters of interest, and the informational content of the data and the model is equivalent to saying that the expectation of a random variable, or the distribution of a random variable in the discrete case, belongs to the Aumann expectation of a properly specified random set. Hence, here the Aumann expectation approach is ideal to characterize the sharp identification region.

Appendix A. Basic definitions

Random sets and selections

As the name suggests, a random set X is a measurable mapping from a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$ to $\mathcal{F}$ that associates a set to each point in the sample space.

Definition A.1. A map $X : \Omega \to \mathcal{F}$ is called a *random closed set* (or a set valued random variable) if for every compact set K in $\Re^d$, $X^{-1}(K) = \{\omega \in \Omega : X(\omega) \cap K \neq \emptyset\} \in \mathfrak{F}$.

The measurability concept used above is different from the more familiar one for vector valued random variables because it must be restrictive enough to ensure that all functionals of interest of the random set become random variables. An example of a relevant functional of a random set which, given Definition A.1, is a random variable, is its support function, see Definition A.6 below. Definition A.1 means that a random closed set is a random element taking values in the family of closed sets equipped with the σ-algebra generated by the families of closed sets $\{F : F \cap K \neq \emptyset\}$ for all compact sets K. Two simple examples can help clarify the concept of a random set:

Example A.1 (*Random Closed Set*)**.** (a) (Trivial) If x is a random vector in $\Re^d$, then $X = \{x\}$ is a random closed set.

(b) Let x_1, x_2 be random variables in $\Re$ such that $\mathbf{P}(x_1 \leq x_2) = 1$. The interval $X = [x_1, x_2]$ is a random closed set.

Aumann's (1965) work on correspondences suggests to think of random sets as bundles of random variables — the selections of the random sets. The formal definition follows:

Definition A.2. For any random set X, a (measurable) *selection* of X is a random vector x with values in $\Re^d$ such that $x(\omega) \in X(\omega)$ $\mathbf{P}$-*a.s.* We denote by Sel (X) the set of all selections from X.

If X is a measurable closed valued almost surely non-empty random set in $\mathcal{F}$, Sel (X) is non-empty (Aumann (1965); see also Li et al. (2002, Theorem 1.2.6)).

In practice, it has been common in certain partial identification analyses to work with selections of random closed sets, although the connection with random set theory was not made. For example, when first proposing partial identification of conditional expectations from selectively observed data, Manski (1989, Eq. (3)) assumed that a partially unobservable outcome variable y belongs to a (non-stochastic) interval with probability one. This is exactly the definition of a selection of a random set.[26] The following examples further clarify this connection.

Example A.2 (*Selections*)**.** Consider the random sets in Example A.1. Then we have:

(a) (Trivial) Sel $(\{x\}) = \{x\}$.

(b) Sel$([x_1, x_2]) = \{x : x$ is $\mathfrak{F}$-measurable and $x(\omega) \in [x_1(\omega), x_2(\omega)]$ $\mathbf{P}$-*a.s.*$\}$. Note that each selection of $[x_1, x_2]$ can be represented as follows. Take a random variable r such that $\mathbf{P}(0 \leq r \leq 1) = 1$ and whose distribution is left unspecified and can be any probability distribution on $[0, 1]$. Let

$$x_r = rx_1 + (1 - r)x_2.$$

Then $x_r \in$ Sel $([x_1, x_2])$. This representation has been used, for example, by Ponomareva and Tamer (2011) and Tamer (2010).

Capacity functional and containment functional

The probability distribution of a random closed set X is uniquely determined by its capacity functional, see Molchanov (2005, Chapter 1, Sections 1.1–1.2). Here we formally define this functional, along with the containment functional.

Definition A.3. The functionals $\mathbf{T}_X : \mathcal{K} \to [0, 1]$ and $\mathbf{C}_X : \mathcal{K} \to [0, 1]$ given by

$$\mathbf{T}_X(K) = \mathbf{P}\{X \cap K \neq \emptyset\}, \quad \mathbf{C}_X(K) = \mathbf{P}\{X \subset K\}, \quad K \in \mathcal{K},$$

are said to be, respectively, the *capacity functional* and the *containment functional* of X.

The following relationship holds:

$$\mathbf{C}_X(K) = 1 - \mathbf{T}_X\left(K^c\right), \tag{A.1}$$

where K^c denotes the complement of the set K in $\Re^d$. While $\mathbf{T}_X$ is defined on compact sets and K^c might be open and not compact, the notation $\mathbf{T}_X(K^c)$ stands for the probability of the (measurable) event $\{X \cap K^c \neq \emptyset\}$, and the functional $\mathbf{T}_X$ is extended onto the family of all sets as described in Molchanov (2005, page 9, Eqs. (1.19)–(1.20); see also Theorem 1.1.12).

Example A.3 (*Capacity and Containment Functional*)**.** Consider the random sets in Example A.1. Then we have:

(a) $\mathbf{T}_X(K) = \mathbf{P}\{\{x\} \cap K \neq \emptyset\} = \mathbf{P}\{x \in K\} = \mathbf{P}\{\{x\} \subset K\}$ for all $K \in \mathcal{K}$. In the singleton case, the capacity functional and the containment functional coincide, and are equal to the probability distribution of x.

(b) In this case X is a random convex compact set taking its realizations in $\Re$. By Theorem 1.7.8 in Molchanov (2005), its distribution is determined uniquely by the values of $\mathbf{C}_X(K)$ for all K convex compact sets, i.e. for all intervals $[k_1, k_2]$ with $k_1, k_2 \in \Re : k_1 \leq k_2$. In this case, $\mathbf{C}_X([k_1, k_2]) = \mathbf{P}\{[x_1, x_2] \subset [k_1, k_2]\} = \mathbf{P}\{x_1 \geq k_1, x_2 \leq k_2\}$.

[26] In this example, the random set is especially simple because it takes on a specific realization with probability 1.

Aumann expectation and support function

Let $\mathbf{L}^1 = \mathbf{L}^1\left(\Omega, \Re^d\right)$ denote the space of $\mathfrak{F}$-measurable random variables with values in $\Re^d$ such that their $\mathbf{L}^1$-norm $\|\xi\|_1 = \mathbf{E}(\|\xi\|)$ is finite, and let the family of all integrable selections of X be given by $\mathrm{Sel}^1(X) = \mathrm{Sel}(X) \cap \mathbf{L}^1$. Then the Aumann expectation of X is defined as follows.

Definition A.4. Let X be a random closed set with $\mathrm{Sel}^1(X) \neq \emptyset$. The *Aumann expectation* of X is

$$\mathbb{E}[X] = \left\{ \int_\Omega x d\mathbf{P} : x \in \mathrm{Sel}^1(X) \right\}$$

where $\int_\Omega x d\mathbf{P}$ is taken coordinate wise. If X is integrably bounded, i.e., if $\sup\{\|x\| : x \in X\}$ has a finite expectation, then[27]

$$\mathbb{E}[X] = \left\{ \int_\Omega x d\mathbf{P} : x \in \mathrm{Sel}(X) \right\}.$$

Clearly, since $\mathrm{Sel}(X)$ is non-empty, the Aumann expectation of an integrably bounded random set is non-empty.

Example A.4 (*Aumann Expectation*)**.** Consider the random sets in Example A.1. Then we have:

(a) $\mathbb{E}[X] = \mathbb{E}[\{x\}] = \mathbf{E}(x)$, so that the Aumann expectation of a singleton coincides with the expectation taken with respect to $\mathbf{P}$.
(b) $\mathbb{E}[X] = \mathbb{E}[[x_1, x_2]] = [\mathbf{E}(x_1), \mathbf{E}(x_2)]$, see Beresteanu and Molinari (2008, Theorem 3.2-(i)).

The definition of Aumann expectation can be extended to the case where one wants to condition on a σ-algebra as follows, see Molchanov (2005, Theorem 2.1.46):

Definition A.5. Let X be an integrably bounded random closed set. For each σ-algebra $\mathfrak{F}_0 \subset \mathfrak{F}$ there exists a unique integrable $\mathfrak{F}_0$-measurable random closed set X_0, denoted by $X_0 = \mathbb{E}[X|\mathfrak{F}_0]$ and called the conditional Aumann expectation of X, such that

$$\mathrm{Sel}_{\mathfrak{F}_0}(X_0) = \mathrm{cl}\{\mathbf{E}(x|\mathfrak{F}_0) : x \in \mathrm{Sel}(X)\},$$

where the closure is taken with respect to the norm in $\mathbf{L}^1_{\mathfrak{F}_0}$. Since X is integrably bounded, so is X_0.

We conclude this section by introducing the notion of support function of a random compact convex set X.

Definition A.6. Let X be a non-empty compact random set with almost surely convex realizations. Then the *support function* of X at $u \in \Re^d$, denoted $h(X, u)$, is the random variable

$$h(X, u) = \sup_{x \in X} \langle x, u \rangle.$$

In Definition A.6, $\langle \cdot, \cdot \rangle$ denotes the inner product in $\Re^d$. To gain insight on the support function, see Fig. 2. It is well known (e.g., Rockafellar (1970, Chapter 13) and Schneider (1993, Section 1.7)) that the support function of a non-empty compact convex set is a continuous sublinear (hence convex) function. In particular, $h(X, u+v) \leq h(X, u) + h(X, v)$ for all $u, v \in \Re^d$ and $h(X, cu) = ch(X, u)$ for all $c > 0$ and for all $u \in \Re^d$. Additionally, one can show that the support function of a bounded set $X \in \Re^d$ is Lipschitz with Lipschitz constant $\sup\{\|x\| : x \in X\}$, see Molchanov (2005, Theorem F.1).

Example A.5 (*Support Function*)**.** Consider the random sets in Example A.1. Then we have:

(a) $h(X, u) = h(\{x\}, u) = \langle x, u \rangle$, $u \in \Re^d$.
(b) $h(X, u) = h([x_1, x_2], u) = \max\{ux_1, ux_2\}$, $u \in \Re$.

[27] Observe that for any $x \in \mathrm{Sel}(X)$, $\|x\| \leq \sup\{\|x\| : x \in X\}$. Hence, all selections of an integrably bounded random set are integrable and $\mathrm{Sel}^1(X) = \mathrm{Sel}(X)$.

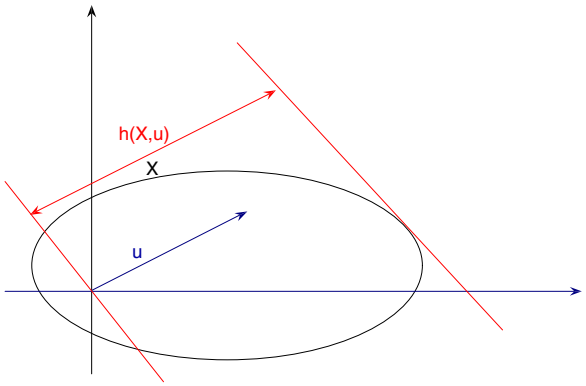

Fig. 2. Support function of X at u: $h(X, u)$ is orthogonal to the supporting hyperplane to X with exterior normal vector u.

Appendix B. Auxiliary results

Lemma B.1. *Let X be a random compact set taking its realization in a finite space $\mathcal{X} \subset \Re^d$. Assume that the probability space can be partitioned as $\Omega = \Omega_1 \cup \Omega_2$. Let $X(\omega) = \{\chi(\omega)\}$ for $\omega \in \Omega_1$ and $X(\omega) = \mathcal{X}$ for $\omega \in \Omega_2$, with χ a random vector taking its realization in $\mathcal{X}$. Then a random vector x is stochastically smaller than X if and only if*

$$\mathbf{P}(x = k) \geq \mathbf{P}\{\chi = k|\Omega_1\}\mathbf{P}(\Omega_1) = \mathbf{C}_X(k)$$

for all $k \in \mathcal{X}$.

Proof. Given that X is either a singleton or the entire space, for each $K \in \mathcal{K}(\mathcal{X})$, $K \neq \mathcal{X}$,

$$\begin{aligned}\mathbf{P}(X \subset K) &= \mathbf{P}\{X \subset K|\Omega_1\}\mathbf{P}(\Omega_1) + \mathbf{P}\{X \subset K|\Omega_2\}\mathbf{P}(\Omega_2) \\ &= \mathbf{P}\{X \subset K|\Omega_1\}\mathbf{P}(\Omega_1) = \mathbf{P}\{\chi \in K|\Omega_1\}\mathbf{P}(\Omega_1).\end{aligned}$$

Because $\mathcal{X}$ is finite, $\mathbf{P}\{\chi \in K|\Omega_1\} = \sum_{k \in K} \mathbf{P}\{\chi = k|\Omega_1\}$. Hence, if the dominance condition holds for singleton sets $K = \{k\}$ for all $k \in \mathcal{X}$, it also holds for any $K \subset \mathcal{K}(\mathcal{X})$. □

Lemma B.2. *Let X be a random compact convex set. Then a random vector x is stochastically smaller than X if and only if*

$$\mathbf{P}(x \in K) \geq \mathbf{P}\{X \subset K\} = \mathbf{C}_X(K)$$

for all compact convex sets K. Moreover, it suffices to consider all K being convex polytopes.

Proof. If a random closed set X is compact convex almost surely, its distribution is uniquely determined by the values of the containment functional $\mathbf{C}_X(K) = \mathbf{P}(X \subset K)$ on all compact convex polytopes K, see Molchanov (1993, 2005, Theorem 1.7.8). We now show that the dominance condition verified on such polytopes suffices to guarantee the condition in Theorem 2.1. Realize x and X on the same probability space; then by standard results in convex analysis (e.g., Rockafellar (1970, Theorem 13.1)), $x \in X$ if and only if the support function of x is dominated by the support function of X. By a result on ordering of stochastic processes (Kamae et al. (1977)) this is the case if and only if

$$\begin{aligned}\mathbf{P}(\langle x, u_1 \rangle \leq s_1, \ldots, \langle x, u_k \rangle \leq s_k) \geq \mathbf{P}(h(X, u_1) \\ \leq s_1, \ldots, h(X, u_k) \leq s_k)\end{aligned} \tag{B.1}$$

for all unit vectors $u_1, \ldots, u_k$, real numbers $s_1, \ldots, s_k$, and $k \geq 1$. By letting K be a convex polytope bounded by hyperplanes with normals $u_1, \ldots, u_k$ located at distances $s_1, \ldots, s_k$ from the origin, we see that the left-hand side in Eq. (B.1) becomes $\mathbf{P}(x \in K)$, while the right-hand side becomes $\mathbf{P}(X \subset K)$. If such a polytope is not bounded, one can pass to the limit in the condition written for all bounded polytopes. □

Appendix C. Partial identification of probability distributions and expectations of response functions with independence assumptions and shape restrictions

Construction of the relevant random set for $y(\cdot)$ under Assumption MTR

In this case, we need to assume that the outcomes in $\mathcal{Y}$ can be ordered, and we need to define a proper random set that contains the response function $y(\cdot)$, i.e. the vector $[y(0), \ldots, y(T)]$, and is such that this function is monotone in t. Observe that if $z = t$, the data and the MTR Assumption reveal that $y(t) = y, y(s) \in \text{Sel}([0, y])$ for each $s \in \mathcal{T} : s < t, y(s) \in \text{Sel}([y, 1])$ for each $s \in \mathcal{T} : s > t$, and $\mathbf{P}(0 \leq y(0) \leq y(1) \leq \cdots \leq y(T) \leq 1) = 1$. Hence we construct a random set $\overrightarrow{Y}^{\mathcal{T}}$ whose vertices are given in Box I with vert $(\cdot)$ the vertices of the set in parenthesis. If $\mathcal{Y} = [0, 1]$, then $\overrightarrow{Y}^{\mathcal{T}} = \text{co}\left(\text{vert}\left(\overrightarrow{Y}^{\mathcal{T}}\right)\right)$ is a simplex. If $\mathcal{Y}$ is finite, then $\overrightarrow{Y}^{\mathcal{T}}$ is the collection of points in $\mathcal{Y}^{\mathcal{T}}$ contained in $\text{co}\left(\text{vert}\left(\overrightarrow{Y}^{\mathcal{T}}\right)\right)$.

This characterization, while exact, is somewhat abstract. Hence, to illustrate, we specialize it to the case that $\mathcal{Y} = [0, 1]$ and $\mathcal{T} = \{0, 1, 2\}$. In this case,

$$\overrightarrow{Y}^{\mathcal{T}} = \text{co}\left\{\begin{bmatrix}y\\y\\y\end{bmatrix}, \begin{bmatrix}y\\y\\1\end{bmatrix}, \begin{bmatrix}y\\1\\1\end{bmatrix}\right\} \quad \text{for } z = 0.$$

$$\overrightarrow{Y}^{\mathcal{T}} = \text{co}\left\{\begin{bmatrix}0\\y\\y\end{bmatrix}, \begin{bmatrix}0\\y\\1\end{bmatrix}, \begin{bmatrix}y\\y\\y\end{bmatrix}, \begin{bmatrix}y\\y\\1\end{bmatrix}\right\} \quad \text{for } z = 1.$$

$$\overrightarrow{Y}^{\mathcal{T}} = \text{co}\left\{\begin{bmatrix}0\\0\\y\end{bmatrix}, \begin{bmatrix}0\\y\\y\end{bmatrix}, \begin{bmatrix}y\\y\\y\end{bmatrix}\right\} \quad \text{for } z = 2.$$

Characterization of the sharp identification regions for $\mathbf{P}(y(\cdot))$ under Assumptions SI-RF and MTR

Proposition C.1. *Let Assumptions SI-RF and MTR hold. Then the sharp identification region for $\mathbf{P}(y(\cdot))$ is*

$$\text{H}[\mathbf{P}(y(\cdot))] = \left\{\mu : \mu(K) \geq \operatorname*{ess\,sup}_{v \in \mathcal{V}} \mathbf{P}\left(\overrightarrow{Y}^{\mathcal{T}} \subset K | v\right) \quad \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)\right\}.$$

If $\mathcal{Y} = [0, 1]$, it suffices to check the above condition for all K being convex polytopes in $\Re^{T+1}$.

Proof. The assumptions are summarized by requiring that $(y(t), v) \in \text{Sel}((\overrightarrow{Y}^{\mathcal{T}}, v)) \cap \mathcal{U}$, where $\mathcal{U}$ is the set of random elements $(\xi, v) \in \mathcal{Y}^{\mathcal{T}} \times \mathcal{V}$ such that ξ is statistically independent of v. If Assumptions SI-RF and MTR are correct, this intersection is non-empty. By the same argument as in the proof of Proposition 2.4, $([y(0), \ldots, y(T)], v) \in (\overrightarrow{Y}^{\mathcal{T}}, v)$ if and only if $\forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right)$

$$\mathbf{P}([y(0), \ldots, y(T)] \in K | v) \geq \mathbf{P}\left(\overrightarrow{Y}^{\mathcal{T}} \subset K | v\right) \; v\text{-a.s.}$$

By the SI-RF assumption, $(y(0), \ldots, y(T))$ is statistically independent of v. Hence, the above condition reduces to

$$\mathbf{P}([y(0), \ldots, y(T)] \in K | v) \geq \operatorname*{ess\,sup}_{v \in \mathcal{V}} \mathbf{P}\left(\overrightarrow{Y}^{\mathcal{T}} \subset K | v\right) \quad \forall K \in \mathcal{K}\left(\mathcal{Y}^{\mathcal{T}}\right).$$

The last claim follows directly from Lemma B.2. □

Formal derivation of the worst-case sharp identification regions for $\mathbf{E}(y(t))$ and $\mathbf{E}(y(\cdot))$

Proposition C.2. *The sharp identification region for $\mathbf{E}(y(t))$ is given by*

$$\begin{aligned}\text{H}[\mathbf{E}(y(t))] &= \{\eta \in \Re : \langle \eta, u\rangle \leq \mathbf{E}(h(Y(t), u)), u = \pm 1\} \\ &= \{\eta \in [\mathbf{E}(y | z = t)\mathbf{P}(z = t), \\ &\quad \mathbf{E}(y | z = t)\mathbf{P}(z = t) + \mathbf{P}(z \neq t)]\}.\end{aligned}$$

The sharp identification region for $\mathbf{E}(y(\cdot))$ is given by

$$\begin{aligned}\text{H}[\mathbf{E}(y(\cdot))] &= \left\{\eta \in \Re^{T+1} : \langle \eta, u\rangle \leq \mathbf{E}\left(h\left(Y^{\mathcal{T}}, u\right)\right) \forall u \in \Re^{T+1}\right\} \\ &= \{\eta \in \times_{t \in \mathcal{T}} [\mathbf{E}(y | z = t)\mathbf{P}(z = t), \\ &\quad \mathbf{E}(y | z = t)\mathbf{P}(z = t) + \mathbf{P}(z \neq t)]\}.\end{aligned}$$

Proof. The random set $Y(t)$ collects all the information given by the data concerning $y(t)$, and therefore $y(t) \in \text{Sel}(Y(t))$. This implies that $\mathbf{E}(y(t)) \in \mathbb{E}[Y(t)]$. Conversely, if $\eta \in \mathbb{E}[Y(t)]$, then there exists a selection $\tilde{y}(t) \in \text{Sel}(Y(t))$ such that $\mathbf{E}(\tilde{y}(t)) = \eta$, and therefore η is an admissible value for the conditional expectation of a selection of $Y(t)$. The final result follows from Theorem 3.1, observing that

$$\begin{aligned}\mathbf{E}(h(Y(t), u)) &= \mathbf{E}(h(Y(t), u) | z = t)\mathbf{P}(z = t) \\ &\quad + \mathbf{E}(h(Y(t), u) | z \neq t)\mathbf{P}(z \neq t) \\ &= u\mathbf{E}(y | z = t)\mathbf{P}(z = t) + h(\mathcal{Y}, u)\mathbf{P}(z \neq t) \\ &= \begin{cases}-\mathbf{E}(y | z = t)\mathbf{P}(z = t) & \text{if } u = -1, \\ \mathbf{E}(y | z = t)\mathbf{P}(z = t) + \mathbf{P}(z \neq t) & \text{if } u = 1.\end{cases}\end{aligned}$$

A similar reasoning gives that $\mathbf{E}(y(\cdot)) \in \mathbb{E}\left[Y^{\mathcal{T}}\right]$. The final result follows from Theorem 3.1, observing that $Y^{\mathcal{T}}$ is a hyperrectangle taking its realizations in $\Re^{T+1}$, fully defined by its support function in directions $u \in U = \{u = [u_0 \cdots u_T]' : u_i \in \{-1, 1\}$ and $u_k = 0$ for $k \neq i, i = 0, \ldots, T\}$, and that

$$\begin{aligned}\mathbf{E}\left(h\left(Y^{\mathcal{T}}, u\right)\right) &= \sum_{t=0}^{T} \mathbf{E}\left(h\left(Y^{\mathcal{T}}, u\right) \middle| z = t\right)\mathbf{P}(z = t) \\ &= \sum_{t=0}^{T} \mathbf{E}\left(\max\left\{\langle \alpha, u\rangle : \alpha_s \in \{0, 1\} \right.\right. \\ &\qquad \left.\left. \text{for } s \neq t, \alpha_t = y\right\} \middle| z = t\right)\mathbf{P}(z = t). \quad \square\end{aligned}$$

Adding mean independence and monotone treatment response assumptions

Suppose now that the researcher also observes a variable v defined on $(\Omega, \mathfrak{F}, \mathbf{P})$ and taking values in $\mathcal{V} \subset \Re$. We consider the following assumption, which uses the nomenclature in Manski (2003, Section 2).

Assumption MI (*Mean Independence of Outcomes and Instruments*)**.**

$$\mathbf{E}(y(t) | v) = \mathbf{E}(y(t)), \quad t \in \mathcal{T}.$$

Notice that Assumption MI is equivalent to an assumption stating that the entire response function is mean independent of v. Manski (2003, Proposition 2.4) derives the sharp identification region for $\mathbf{E}(y(t))$ under Assumption MI. His result can be extended to obtain the sharp identification region for $\mathbf{E}(y(\cdot))$ under Assumption MI. They can further be extended by additionally imposing shape restrictions in the form of the MTS assumption. We provide these results here.

Proposition C.3. *Let Assumption MI hold. Then the sharp identification region for $\mathbf{E}(y(t))$ is*

$$\text{vert}\left(\overrightarrow{Y}^{\mathcal{T}}\right)=\left\{t\text{-th entry}\longrightarrow\begin{bmatrix}0\\0\\\vdots\\0\\y\\y\\\vdots\\y\\y\end{bmatrix},\begin{bmatrix}0\\0\\\vdots\\0\\y\\y\\\vdots\\y\\1\end{bmatrix},\ldots,\begin{bmatrix}0\\0\\\vdots\\0\\y\\1\\\vdots\\1\\1\end{bmatrix},\ldots,\begin{bmatrix}0\\0\\\vdots\\y\\1\\\vdots\\1\\1\end{bmatrix},\ldots,\begin{bmatrix}0\\y\\\vdots\\y\\1\\\vdots\\1\\1\end{bmatrix},\begin{bmatrix}y\\y\\\vdots\\y\\1\\\vdots\\1\\1\end{bmatrix}\right\}\text{ for } z=t \tag{C.1}$$

Box I.

$$\begin{aligned}\mathrm{H}[\mathbf{E}(y(t))]&=\{\eta\in\Re:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(Y\left(t\right),u\right)|\,v\right),\\&\quad u=\pm1,\,v\text{-a.s.}\}\\&=\left\{\eta\in\left[\operatorname*{ess\,sup}_{v\in\mathcal{V}}\mathbf{E}\left(y|\,z=t,\,v\right)\mathbf{P}\left(z=t|v\right),\right.\right.\\&\quad\left.\left.\operatorname*{ess\,inf}_{v\in\mathcal{V}}\left[\mathbf{E}\left(y|z=t,\,v\right)\mathbf{P}\left(z=t|v\right)+\mathbf{P}\left(z\neq t|v\right)\right]\right]\right\}.\end{aligned}$$

The sharp identification region for $\mathbf{E}(y(\cdot))$ *is given by*

$$\begin{aligned}\mathrm{H}[\mathbf{E}(y(\cdot))]&=\left\{\eta\in\Re^{T+1}:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(Y^{\mathcal{T}},u\right)\middle|\,v\right)\right.\\&\quad\left.\forall u\in\Re^{T+1},\,v\text{-a.s.}\right\}\\&=\left\{\eta\in\times_{t\in\mathcal{T}}\left[\operatorname*{ess\,sup}_{v\in\mathcal{V}}\mathbf{E}\left(y|\,z=t,\,v\right)\mathbf{P}\left(z=t|v\right),\right.\right.\\&\quad\left.\left.\operatorname*{ess\,inf}_{v\in\mathcal{V}}\left[\mathbf{E}\left(y|\,z=t,\,v\right)\mathbf{P}\left(z=t|v\right)+\mathbf{P}\left(z\neq t|v\right)\right]\right]\right\}.\end{aligned}$$

Proof. For each $v\in\mathcal{V}$, the data reveals that $\mathbf{E}(y(t)|v)\in\mathbb{E}[Y\left(t\right)|\,v]$, which holds if and only if

$$\begin{aligned}\mathbf{E}\left(h\left(y(t),u\right)|\,v\right)&=\mathbf{E}\left(\langle y(t),u\rangle|\,v\right)\\&\leq\mathbf{E}\left(h\left(Y\left(t\right),u\right)|\,v\right),\quad u=\pm1.\end{aligned}$$

Assumption MI states that $\mathbf{E}(y(t)|v)=\mathbf{E}(y(t))$, which is equivalent to $\mathbf{E}\left(\langle y(t),u\rangle|\,v\right)=\mathbf{E}\left(\langle y(t),u\rangle\right)$ for each $u=-1,1$. Hence we obtain

$$\mathbf{E}\left(\langle y(t),u\rangle\right)\leq\mathbf{E}\left(h\left(Y\left(t\right),u\right)|\,v\right),\quad u=\pm1,\,v\text{-a.s.}$$

The final expression for the bounds follows from Proposition C.2. The same reasoning gives the result for $\mathrm{H}[\mathbf{E}(y(\cdot))]$. □

Proposition C.4. *Let Assumptions MI and MTR hold. Let* $\overrightarrow{Y}\left(t\right)$ *and* $\overrightarrow{Y}^{\mathcal{T}}$ *be defined as in Eqs. (2.12) and (C.1), respectively. Then the sharp identification region for* $\mathbf{E}(y(t))$ *is*

$$\begin{aligned}\mathrm{H}[\mathbf{E}(y(t))]&=\left\{\eta\in\Re:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(\overrightarrow{Y}\left(t\right),u\right)\middle|\,v\right),\right.\\&\quad\left.u=\pm1,\,v\text{-a.s.}\right\}\\&=\left\{\eta\in\left[\operatorname*{ess\,sup}_{v\in\mathcal{V}}\mathbf{E}\left(y|\,z\leq t,\,v\right)\mathbf{P}\left(z\leq t|v\right),\right.\right.\\&\quad\left.\left.\operatorname*{ess\,inf}_{v\in\mathcal{V}}\left[\mathbf{E}\left(y|\,z\geq t,\,v\right)\mathbf{P}\left(z\geq t|v\right)+\mathbf{P}\left(z<t|v\right)\right]\right]\right\}.\end{aligned}$$

The sharp identification region for $\mathbf{E}(y(\cdot))$ *is given by*

$$\begin{aligned}\mathrm{H}[\mathbf{E}(y(\cdot))]&=\left\{\eta\in\Re^{T+1}:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(\overrightarrow{Y}^{\mathcal{T}},u\right)\middle|\,v\right)\right.\\&\quad\left.\forall u\in\Re^{T+1},\,v\text{-a.s.}\right\}.\end{aligned}$$

Proof. The same argument as in the proof of Proposition C.3 gives that

$$\begin{aligned}\mathrm{H}[\mathbf{E}(y(t))]&=\left\{\eta\in\Re:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(\overrightarrow{Y}\left(t\right),u\right)\middle|\,v\right),\right.\\&\quad\left.u=\pm1,\,v\text{-a.s.}\right\},\\\mathrm{H}[\mathbf{E}(y(\cdot))]&=\left\{\eta\in\Re^{T+1}:\langle\eta,u\rangle\leq\mathbf{E}\left(h\left(\overrightarrow{Y}^{\mathcal{T}},u\right)\middle|\,v\right)\right.\\&\quad\left.\forall u\in\Re^{T+1},\,v\text{-a.s.}\right\}.\end{aligned}$$

To get the final expressions, observe that

$$\begin{aligned}\mathbf{E}\left(h\left(\overrightarrow{Y}\left(t\right),u\right)\middle|\,v\right)&=\mathbf{E}\left(h\left([0,y],u\right)|\,z>t,\,v\right)\mathbf{P}\left(z>t|v\right)\\&\quad+\mathbf{E}\left(\langle y,u\rangle|\,z=t,\,v\right)\mathbf{P}\left(z=t|v\right)\\&\quad+\mathbf{E}\left(h\left([y,1],u\right)|\,z<t,\,v\right)\mathbf{P}\left(z<t|v\right)\\&=\begin{cases}\mathbf{E}\left(\langle y,u\rangle|\,z\leq t,\,v\right)\mathbf{P}\left(z\leq t|v\right) & \text{if } u=-1\\ \mathbf{E}\left(\langle y,u\rangle|\,z\geq t,\,v\right)\mathbf{P}\left(z\geq t|v\right)+\mathbf{P}\left(z<t|v\right) & \text{if } u=1.\end{cases}\quad\square\end{aligned}$$

While $\mathbf{E}\left(h\left(\overrightarrow{Y}^{\mathcal{T}},u\right)\middle|\,v\right)$ does not have a simple closed form expression for arbitrary $\mathcal{T}$, it is extremely simple to compute in practice. To illustrate this claim, we specialize the above result to the case that $\mathcal{Y}=[0,1]$ and $\mathcal{T}=\{0,1,2\}$. Let $u=[u_0\,u_1\,u_2]$ and let $u_{sum}=(u_0+u_1+u_2)$. Then

$$\begin{aligned}\mathbf{E}\left(h\left(\overrightarrow{Y}^{\mathcal{T}},u\right)\middle|\,v\right)&=\sum_{t=0}^{2}\mathbf{E}\left(h\left(\overrightarrow{Y}^{\mathcal{T}},u\right)\middle|\,z=t,\,v\right)\mathbf{P}\left(z=t|v\right)\\&=\mathbf{E}\left(\max\left\{yu_{sum},\,y\left(u_0+u_1\right)+u_2,\right.\right.\\&\quad\left.\left.yu_0+\left(u_1+u_2\right)\right\}|\,z=0,\,v\right)\mathbf{P}\left(z=0|v\right)\\&\quad+\mathbf{E}\left(\max\left\{y\left(u_1+u_2\right),\,yu_1+u_2,\,yu_{sum},\right.\right.\\&\quad\left.\left.+\,y\left(u_0+u_1\right)u_2\right\}|\,z=1,\,v\right)\mathbf{P}\left(z=1|v\right)\\&\quad+\mathbf{E}\left(\max\left\{yu_2,\,y\left(u_1+u_2\right),\,yu_{sum}\right\}|\,z=2,\,v\right)\mathbf{P}\left(z=2|v\right).\end{aligned}$$

References

Andrews, D.W.K., Shi, X., 2009. Inference Based on Conditional Moment Inequalities. mimeo.

Andrews, D.W.K., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.

Artstein, Z., 1974. On the calculus of closed set-valued functions. Indiana University Mathematics Journal 24 (5), 433–441.

Artstein, Z., 1983. Distributions of random sets and random selections. Israel Journal of Mathematics 46 (4), 313–324.

Aumann, R.J., 1965. Integrals of set valued functions. Journal of Mathematical Analysis and Applications 12, 1–12.

Balke, A., Pearl, J., 1997. Bounds on treatment effects from studies with imperfect compliance. Journal of the American Statistical Association 92 (439), 1171–1176.

Beresteanu, A., Molchanov, I., Molinari, F., 2008. Sharp identification regions in games, CeMMAP Working Paper CWP15/08.

Beresteanu, A., Molchanov, I., Molinari, F., 2009. Sharp identification regions in models with convex predictions: games, Individual choice, and incomplete data, CeMMAP Working Paper CWP27/09.

Beresteanu, A., Molchanov, I., Molinari, F., 2011. Sharp identification regions in models with convex moment predictions. Econometrica (in press).

Beresteanu, A., Molinari, F., 2006. Asymptotic properties for a class of partially identified models, CeMMAP Working Paper CWP10/06.
Beresteanu, A., Molinari, F., 2008. Asymptotic properties for a class of partially identified models. Econometrica 76, 763–814.
Bontemps, C., Magnac, T., Maurin, E., 2008. Set Identified Linear Models. mimeo.
Bugni, F.A., 2010. Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set. Econometrica 78, 735–753.
Canay, I.A., 2010. EL Inference for partially identified models: large deviations optimality and bootstrap validity. Journal of Econometrics 156, 408–425.
Chernozhukov, V., Hong, H., Tamer, E., 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75, 1243–1284.
Chernozhukov, V., Lee, S., Rosen, A., 2009. Intersection bounds: estimation and inference, CeMMAP Working Paper CWP19/09.
Choquet, G., 1953–1954. Theory of capacities. Annales de l'Institut Fourier 5, 131–295.
Debreu, G., 1967. Integration of correspondences. In: Proceedings of the Fifth Berkeley Symposium in Mathematical Statistic and Probability, vol. 2. University of California Press, pp. 351–372.
Dynkin, E., Evstigneev, I.V., 1976. Regular conditional expectations of correspondences. Theory of Probability and its Applications 21, 325–338.
Galichon, A., Henry, M., 2006. Inference in incomplete models, Working Paper, Columbia University.
Galichon, A., Henry, M., 2009a. Set Identification in Models with Multiple Equilibria. mimeo.
Galichon, A., Henry, M., 2009b. A test of non-identifying restrictions and confidence regions for partially identified parameters. Journal of Econometrics 152, 186–196.
Haile, P., Tamer, E., 2003. Inference with an incomplete model of english auctions. Journal of Political Economy 111, 1–51.
Imbens, G.W., Manski, C.F., 2004. Confidence intervals for partially identified parameters. Econometrica 72, 1845–1857.
Kamae, T., Krengel, U., O'Brien, G.L., 1977. Stochastic inequalities on partially ordered spaces. Annals of Probability 5, 899–912.
Kitagawa, T., 2009: Identification region of the potential outcome distributions under instrument independence, CeMMAP Working Paper CWP30/09.
Li, S., Ogura, Y., Kreinovich, V., 2002. Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables. Kluwer Academic Publishers.
Manski, C.F., 1989. Anatomy of the selection problem. Journal of Human Resources 24, 343–360.
Manski, C.F., 1990. Nonparametric bounds on treatment effects. American Economic Review Papers and Proceedings 80, 319–323.
Manski, C.F., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge, MA.
Manski, C.F., 1997. Monotone treatment response. Econometrica 65 (6), 1311–1334.
Manski, C.F., 2003. Partial Identification of Probability Distributions. Springer Verlag, New York.
Manski, C.F., 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge, MA.
Manski, C.F., Molinari, F., 2010. Rounding probabilistic expectations in surveys. Journal of Business and Economic Statistics 28, 219–231.
Matheron, G., 1975. Random Sets and Integral Geometry. Wiley, New York.
Molchanov, I., 1993. Limit theorems for convex hulls of random sets. Advances in Applied Probability 25, 395–414.
Molchanov, I., 2005. Theory of Random Sets. Springer Verlag, London.
Neyman, J.S., 1923. On the application of probability theory to agricaltural experiments. Essay on Principles. Section, Roczniki Nauk Rolniczych., 10, 1–51, Reprinted (Translated and Edited) in Statistical Science, 1990, Vol. 5, No. 4, pp. 465–480.
Norberg, T., 1992. On the existence of ordered couplings of random sets — with applications. Israel Journal of Mathematics 77, 241–264.
Pepper, J.V., 2002. Strong Instrumental Variables. mimeo.
Ponomareva, M., 2010. Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates. mimeo.
Ponomareva, M., Tamer, E., 2011. Misspecification in moment inequality models: back to moment equalities? The Econometrics Journal 14, 186–203.
Rockafellar, R., 1970. Convex Analysis. Princeton University Press.
Schneider, R., 1993. Convex Bodies: The Brunn-Minkowski Theory. Cambridge Univ. Press.
Stoye, J., 2007. Bounds on generalized linear predictors with partially identified outcomes. Reliable Computing 13, 293–302.
Stoye, J., 2009. More on confidence intervals for partially identified parameters. Econometrica 77, 1299–1315.
Tamer, E., 2010. Partial identification in econometrics. Annual Reviews of Economics 2, 167–195.