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.

© 2011 Elsevier B.V. All rights reserved.

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.

Download English Version:

https://daneshyari.com/en/article/5096566

Download Persian Version:

https://daneshyari.com/article/5096566

Daneshyari.com